MULTIVARIATE FINITE MIXTURE GROUP-BASED TRAJECTORY MODELING WITH APPLICATION TO MENTAL HEALTH STUDIES

A Thesis submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in the Collaborative Biostatistics Program
within the School of Public Health
University of Saskatchewan
Saskatoon, Saskatchewan, Canada

By

Yanzhao Cheng

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Department Head of School of Public Health

College of Medicine

University of Saskatchewan

Health Sciences Building E-Wing, 104 Clinic Place

Saskatoon, Saskatchewan S7N 2Z4 Canada

OR

Dean

College of Graduate and Postdoctoral Studies

University of Saskatchewan

116 Thorvaldson Building, 110 Science Place

Saskatoon, Saskatchewan S7N 5C9 Canada

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Dr. Hyun J. Lim for her intelligent supervising, motivation, encouragement and also consistent support. I really appreciate her time, patience and advice during the progress of completing my PhD thesis. In the last 5 years, she has not only taught me knowledge of statistics, research and work, but also how to be a human being, which is more important. I will never forget it.

I would like to thank my research committee: Dr. Cindy Feng, Dr. Bonnie Janzen, Dr. Christian Boudreau, Dr. Lloyd Cenon Balbuena and Dr. Longhai Li for their encouragement, the time dedicated to reading my thesis and their expert-level suggestions.

I am particularly grateful to Dr. Prosanta Mondal for his precious time and comments on this thesis.

My sincere thanks also go to the University of Saskatchewan, the School of Public Health, and the Department of Surgery for giving me the opportunity to enter the PhD Biostatistics Program and offering financial support during the program. I would like to thank the Korea Institute for Health and Social Affairs and National Health Insurance Service for allowing me to use the Korea Health Panel Study dataset in my study.

Finally, I am very much thankful to my family: my wife, Xiaoqian Sun and my daughter, Shelby Cheng for their love, understanding and continuing support to complete this research work. I am grateful to my parents Ping Sui, Zhenquan Cheng, Shujun Liu and Benrong Sun for their love and encouragement, the sacrifices they made for my education, and for preparing me for my future. This thesis is dedicated to all of them.

ABSTRACT

Traditionally, two kinds of methods are applied in trajectory analysis: 1) hierarchical modeling based on a multilevel structure, or 2) latent growth curve modeling (LGCM) based on a covariance structure (Raudenbush & Bryk, 2002; Bollen & Curran, 2006). However, this thesis used a third trajectory analysis method: group-based trajectory modeling (GBTM). GBTM was an extension of the finite mixture modeling (FMM) method that has been widely used in various fields of trajectory analysis in the last 25 years (Nagin & Odgers, 2010). GBTM was able to detect unobserved subgroups based on the multinomial logit function (Nagin, 1999). As an extended form of FMM, GBTM parameters could be estimated using maximum likelihood estimation (MLE) procedures. Since FMMs had no closed-form solution to the maximum likelihood, the Expectation-Maximization (EM) algorithm would often be applied to find maximized likelihood (Schlattmann, 2009). However, GBTM used a different optimization method called the Quasi-Newton procedure to perform the maximization.

This thesis studied both GBTM with a single outcome and trajectory modeling with multiple outcomes. Nagin constructed two extended trajectory models that can involve multiple outcomes. Group-based dual trajectory modeling (GBDTM) deals with two outcomes combined with comorbidity or heterotypic continuity, while group-based multi-trajectory modeling (GBMTM) could include more than two outcomes in one model with the same subgroup weights among the outcomes (Nagin, 2005; Nagin, Jones, Passos, & Tremblay, 2018; Nagin & Tremblay, 2001).

The methodology was applied to the Korea Health Panel Survey (KHPS) data, which included 3983 individuals who were 65 years old or older at the baseline. GBTM, GBDTM, and GBMTM were three approaches performed with two binary longitudinal outcomes - depression and anxiety. GBDTM was selected as the best model with this data set because it is more flexible than GBMTM when handling group membership, and unlike GBTM, GMDTM addressed the interrelationship between the outcomes based on conditional probability. Four depression trajectories were identified across eight years of follow-up: “low-flat” (n = 3641; 87.0%), “low-to-middle” (n = 205; 8.8%), “low-to-high” (n = 33; 1.3%) and “high-curve” (n = 104; 2.8%). Also, four anxiety trajectories were identified with: “low-flat” (n =3785; 92.5%), “low-to-middle” (n = 96; 4.7%), “high-to-low” (n =89; 2.2%) and “high-curve” (n = 13; 0.6%) trajectory groups. Female sex, the presence of more than three chronic diseases, and income-generating activity were significant risk factors for depression trajectory groups. Anxiety trajectory groups had the same risk factors except for the presence of more than three chronic diseases.

To further study the GBTM, GBDTM and GBMTM approach, the simulation study was also performed based on two correlated repeatedly measured binary outcomes. Compared based on these two outcomes with different correlation levels (ρ = 0.1, 0.2, 0.4, 0.6). GBDTM was always a better model than GBTM when we were interested in the association between the two outcomes. GBMTM could be used instead of GBDTM when the correlation coefficients between two longitudinal outcomes were high.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC Akaike information criterion

ALT Autoregressive latent trajectory

BIC Bayesian Information Criterion

CBT Cognitive-behavioral therapy

DA Trajectory groups for anxiety in group-based dual trajectory modeling

DD Trajectory groups for depression in group-based dual trajectory modeling

EM Expectation-maximization

FMM Finite mixture modeling

GAD Generalized anxiety disorder

GBDTM Group-based dual trajectory modeling

GBMTM Group-based multi-trajectory modeling

GBTM Group-based trajectory modeling

GCM Growth curve modeling

GLMs Generalized linear models

IPT Interpersonal therapy

KHPS Korea health panel survey

LGCM Latent growth curve modeling

MA Trajectory groups for anxiety in group-based multi-trajectory modeling

MAOIs Monoamine oxidase inhibitors

MD Trajectory groups for depression in group-based multi-trajectory modeling

MLE Maximum likelihood estimation

MMD Major depressive disorder

NR Newton-Raphson

OECD Organization for Economic Co-operation and Development

SEM Structure equation modeling

SNRIs Seotonin and norepinephrine reuptake inhibitors

SSRIs Selective serotonin reuptake inhibitors

TA Trajectory groups for anxiety in group-based trajectory modeling

TCAs Tricyclic antidepressants

TD Trajectory groups for depression in group-based trajectory modeling

TICs Time-invariant covariates

TVCs Time-variant covariates

VIF Variance inflation factor

ZIF Zero-Inflated Poisson

TABLE OF CONTENTS

Chapter 1 INTRODUCTION

1.1 Rationale

A trajectory is defined as: "The curved path that an object follows after it has been thrown or shot in the air" (Cambridge Dictionary, 2018). However, in statistics, a trajectory specifically refers to evaluating one or more outcomes over age or time, as in a repeat measurement from a longitudinal study (Nagin, 2005). Hierarchical modeling and latent growth curve modeling (LGCM) are two methods applied in trajectory analysis (Raudenbush & Bryk, 2002; Bollen & Curran, 2006). Hierarchical modeling focuses on individual variations with random effects, which are called growth curves. LGCM uses covariance structure modeling to generate trajectories (Jones, Nagin, & Roeder, 2001).

Hierarchical modeling and LGCM were used to study trajectory average estimation or using covariates to explain the mean variability. However, another trajectory analysis method, called Group-based trajectory modeling (GBTM) used unobserved latent class variables to identify the underlying subgroup trajectories from the population (Nagin, 2005). GBTM assumes that the population contains a mixture of unobserved groups with unique development trajectories (Nagin & Odgers, 2010). GBTM identifies the distinct groups with correlated physical or biological characteristics (Nagin & Odgers, 2010). GBTM was developed to investigate the number of criminal offenses (count outcomes) an individual had committed and apply this to the study of criminal careers (Nagin & Land, 1993). This criminology study included more than 400 ten-year-old boys who were followed biannually until they were 32-years-old and involved in up to 11 criminal offense measurements. This criminal offense study identified four subgroups in the studied population (Nagin & Land, 1993). Also, GBTM was extended to combine both a Poisson model and a multinomial logic function as a mixture model at the individual level, which is useful for revealing heterogeneity in a population of interest (Nagin & Land, 1993).

GBTM is a semi-parametric model since it lies between parametric and non-parametric models (Nagin, 1999). GBTM allows us to estimate the trajectory of the population's prototypical development and can clearly reveal the uncertainty of latent group membership based on multiple risk factors that may influence decision-making about group membership (Roeder, Lynch, & Nagin, 1999). Three types of repeat measurement outcomes can be applied in the model: (i) continuous data, following a normal distribution; (ii) count data, following a Poisson distribution or a zero-inflated Poisson distribution; or (iii) binary data, following a binary distribution (Nagin, 2005). GBTM has three main strengths. First, the population is assumed to be constituted of distinct groups, each with a different latent tendency. Second, GBTM estimation affects the covariates not only for trajectory shape, but also for group membership. Time-independent covariates impact the trajectory of group portions. Time-dependent covariates explain variations in trajectory shapes. Third, besides the time variables, the trajectory groups can be identified by GBTM without any other covariates. Fourth, GBTM handles non-monotonic trajectories and irregular trajectories in the population (Jones et al., 2001).

GBTM can only handle a single outcome with multiple measurement times for each individual. What if there are two or more related outcomes that interest us? One type of correlation between several outcomes is called "comorbidity". Here, "comorbidity" means that undesirable conditions such as anxiety and depression occurred contemporaneously more than once (Kessler et al., 1994). Another type of correlation is called "heterotypic continuity", which recognizes that two outcomes may be linked in an individual but do not co-occur (Caspi & Roberts, 2001), for example, being physically aggressive during adulthood and committing crimes as adults. Nagin developed two kinds of trajectory modeling methods that are able to deal with two or more outcomes with "comorbidity" or "heterotypic continuity" correlations. First, group-based dual trajectory modeling (GBDTM) was adopted as an extension of the group-based trajectory modeling using joint probability to link two outcomes into one model (Nagin & Tremblay, 2001). Another recently introduced model called group-based multi-trajectory modeling (GBMTM) combines trajectory modeling with two or more outcomes into one model. However, GBMTM assumes that every outcome should have the same number of trajectory groups, and that each trajectory group includes the same group memberships among the outcomes (Nagin et al., 2018).

Compared to a single outcome GBTM, GBDTM or GBMTM offers multivariate analysis to study correlated outcomes. For example, in psychiatric studies, using a single outcome is insufficient to characterize a disease's complexity (Teixeira-Pinto, Siddique, Gibbons, & Normand, 2009). Therefore, GBDTM and GBMTM are necessary for multiple outcomes, such as those on the depression and anxiety scales. Teixeira-Pinto and his colleagues (2009) did a psychiatric study with three outcomes to compare the difference between single outcome regression and the mixture model using a latent variable to link the three outcomes together as a multivariate model. He showed that multivariate mixture modeling has several advantages compared to single outcome modeling. (i) It is clear that, if only a single outcome is considered in each model, the relationship between the outcomes is effectively omitted compared to the multivariate model. This could trigger a lack of efficiency in the analysis and a failure to identify covariates' effects; (ii) if there are missing values in the outcomes, single outcome analyses may produce biased covariate estimations, especially when these values are not missing at random; and (iii) multivariate mixture modeling provides covariates' overall effect while single regression modeling does not (Teixeira-Pinto et al., 2009).

Nagin's study includes two varieties of multiple outcomes trajectory modeling: GBDTM and GBMTM. These two varieties are mostly applied to criminology (Nagin, 2005). In psychiatric and mental health studies, on the other hand, multiple published papers use single group-based trajectory modeling. Especially in the years immediately following their development, GBDTM and GBMTM were rarely used. For example, in Mustillo's study of obesity and psychiatric disorders, they studied only a single outcome (obesity) and used psychiatric disorders as time-dependent covariates (Mustillo et al., 2003). More recent studies suggest that for outcomes with "comorbidity" or "heterotypic continuity", it would be better to use GBDTM (Nagin & Tremblay, 2001). Thus, GBDTM has been more frequently applied in psychiatric and mental health studies. For example, one study uses GBDTM to demonstrate the causal relationship between substance use during adolescence and obesity in young adulthood (Huang et al., 2013). Another study applied GBDTM to study trajectories and the relationship between co-occurring delinquency and depressive symptoms among male and female adolescents separately (Wiesner & Kim, 2006).

Trajectory analysis of depression and anxiety were usually developed with GBDTM based on youth and adolescence (Côté et al., 2009; Feng, Shaw, & Silk, 2008; McLaughlin & King, 2015; Olino, Klein, Lewinsohn, Rohde, & Seeley, 2010). To study the trajectory classes in older adults, researchers often focused on only depression using GBTM (Liang, Xu, Quiñones, Bennett, & Ye, 2011; Kuo, Lin, Chen, Chuang, & Chen, 2011; Byers et al., 2012; Hsu, 2012; Montagnier et al., 2014; Kuchibhatla, Fillenbaum, Hybels, & Blazer, 2012). However, depression and anxiety trajectories have been rarely studied simultaneously. Three studies were found with the development of depression and anxiety trajectories in older adults (Holmes et al., 2018; Rzewuska, Mallen, Strauss, Belcher, & Peat, 2015; Spinhoven, van der Veen, Voshaar, & Comijs, 2017). Holmes et al. and Spinhoven et al. used latent growth mixture modeling to identify different course trajectories, where latent growth mixture modeling assumes that there are unique and different subgroups under the population (Frankfurt, Frazier, Syed, & Jung, 2016). Rzewuska et al. used latent class analysis to find depression and anxiety trajectories separately. None of them used GBDTM or GBMTM to study depression and anxiety simultaneously in older adults. Therefore, I am encouraged to use all three group-based trajectory models to study depression and anxiety in older adults.

Moreover, there are no studies that systematically explain if GBDTM developed for two distinct, but correlated outcomes (Nagin & Tremblay, 2001) or GBMTM developed for similar outcomes (Nagin et al., 2018) should be used instead of GBTM. If the correlation is too low, can we still use GBDTM? How can we demonstrate that the outcomes are similar? To compare these three trajectory modeling methods and answer these questions, this thesis uses real data analysis to evaluate depression and anxiety outcomes within the Korea Health

Panel Study (KHPS). A simulation study was conducted to identify the most useful model when different correlation levels between two longitudinal outcomes were present.

1.2 Research objectives

My thesis has three study objectives:

Objective 1: To model the trajectories using group-based trajectory modeling (GBTM), group-based dual trajectory modeling (GBDTM), and group-based multi-trajectory modeling (GBMTM) approaches using Korea Health Panel Survey (KHPS) data. Compare the trajectory shape and membership differences from those three approaches, and determine the best modeling method for depression and anxiety outcomes.

Objective 2: Based on the best modeling method, identify relevant risk factors that may influence the trajectory groups for depression and anxiety outcomes in the KHPS dataset.

Objective 3: To determine in what kind of situations we should use GBDTM or GBMTM rather than separate single GBTM and to also provide strategies to simulate various scenarios.

This thesis is organized as follows. The literature review is featured in Chapter 2, while statistical methods are discussed in Chapter 3. I describe different trajectory analysis methods and depression and anxiety among older people in Chapter 2. The theory of finite mixture models is introduced in Section 3.2. GBTM, GBDTM and GBMTM are presented in Section 3.3, Section 3.4 and Section 3.5, respectively. The aforementioned trajectory methods are applied to real KHPS data in Chapter 4. The results from simulation studies are presented in Chapter 5. The discussion is presented in Chapter 6, and the conclusion is in Chapter 7.

Chapter 2 LITERATURE REVIEW

2.1 Review of trajectory models

The key methods of trajectory analysis will be reviewed in this section. They are growth curve modeling (GCM), hierarchical modeling, latent growth curve modeling (LGCM), and group-based trajectory modeling (GBTM).

2.1.1 Growth curve modeling

Trajectory analysis has a long history in statistics. In the early nineteenth century, trajectory analysis was focused on changes in whole groups or collectivities rather than individuals. The method predicted continuous mortality rate over time for a sample of individuals by estimating a single trajectory (Gompertz, 1833). Later, multiple logistic curves were applied to examine human development based on different characteristics, such as national food consumption, influenza, etc. (Robertson, 1908). The ANOVA model predicting the rate of growth from characteristics such as experimental conditions and sex was the first attempt to analyze both the trajectory of a group of individuals and to set up an individual's trajectories varied at random (Wishart, 1938). Subsequently, ANCOVA and MANOVA were applied in trajectory analysis to calculate the individual trajectories. The difference in averages from the group membership was examined later (Gilriches, 1957).

More recently, growth curve modeling (GCM) has been implemented in trajectory analysis. GCM can estimate the difference among different individuals based on the variation between them (Bollen & Curran, 2006). Usually, the variation within individuals, called latent trajectories, relates to time trends. These time trends can be polymorphic between different individuals as a result of each person's characteristics, such as age, etc. The GCM could contain both fixed and random effects. The fixed effect provides a population's overall average trajectory. The random effect is the variation in the individual trajectories in relation to the overall mean (Gardiner, Luo, & Roman, 2009). We need to consider random effects because random effects are a way to measure trajectory parameters. Small random effects mean that trajectories are similar among individuals. On the contrary, large random effects indicate a large difference in trajectories between individuals (Verbeke, Molenberghs, & Rizopoulos, 2010). These fixed and random effects allow us to determine growth characteristics for the whole group and for individuals within the group (Curran, Obeidat, & Losardo, 2010). Compared to the conditional longitudinal method, and based on current approaches, GCM is "more flexible to deal with partially missing data, unequally spaced time points, non-normally distributed or discretely scaled repeated measures, complex nonlinear or compound-shaped trajectories, time-variant covariates (TVCs), and multivariate growth processes" (Curran et al., 2010).

There are two commonly used approaches for GCM: (i) hierarchical modeling structure, also called multilevel modeling for the growth curve, which is fitted using a multilevel modeling framework (Bryk & Raudenbush, 1987), and (ii) structural equation modeling (SEM) framework, which is a method using observed repeated measurements to indicate unobserved time trends that rely on one or more latent factors (Bollen & Curran, 2006). Compared to these two approaches, the multilevel modeling framework is better at estimating higher levels of nesting characteristics. Nevertheless, the SEM framework is specialized in evaluating latent variables and shrinks down the error of measures from both the outcome variable and covariates (Bollen & Curran, 2006).

Defining the sample size for GCM estimation is also critical. The required sample size for GCM may be different for different studies. For most studies, the sample size should not be fewer than 100 (Bollen & Curran, 2006). However, Huttenlocher et al. did a trajectory analysis of children's early vocabulary growth with a small sample size n=22 (Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991). On account of the relationship between the subject numbers and the number of repeat measurements for each individual, the total observations from the longitudinal data must be considered for model estimation and statistical power. The number of repeat measurements for each individual may also influence model estimation.

Each individual should have at least three measurement points during the study. However, most studies are not able to reach this goal because of the missing values from dropping off. Most studies consider that at least 80%-90% of the data should have more than three measurements (Curran et al., 2010). The repeated measurements should be continuous and follow a normal distribution from the initial GCM. Researchers often use the maximum likelihood estimation (MLE) method for estimations. Nevertheless, in the case of continuous data that is not normally distributed or discrete, or ordinal, alternative approaches such as three-level hierarchical modeling, can estimate the model (Satorra, 1990; P. D. Mehta, Neale, & Flay, 2004). Discrete data can be managed by exponential family trajectories, while piecewise linear modeling is often implemented for nonlinear functions (Cudeck & Harring, 2007; Bollen & Curran, 2006). In general, selecting a reasonable sample size and applying the right model that relies on the appropriate data format is vital for model fitting.

There are two approaches to dealing with missing data in GCM. The first option is to use the MLE method directly (Arbuckle, 1996). In the MLE method, the data points can be weighted for estimations. The second approach is the imputation approach. In this method, missing data is replaced based on observed data, after which data analysis is carried out on the imputed dataset (Schafer, 1999). Both approaches can be used under the assumption that data are missing completely at random or missing at random. However, for data that is missing not at random, neither the MLE approach nor the imputation approach is suitable.

2.1.2 Hierarchical modeling

Hierarchical modeling is used in various studies, though it has different names in different fields. Sociologists define this model as the “multilevel linear model” (Goldstein, 1991), while “mixed-effects model” and “random-effects model” are often applied in biometric research (Elston & Grizzle, 1962). In statistics, this model is also called the “covariance components model” (Dempster, Rubin, & Tsutakawa, 1981). “Hierarchical modeling”, however, is the most commonly used name because it highlights the importance of the data structure. This model can be applied not only in longitudinal studies to generate an individual's growth trajectory, but also in organizational studies to investigate workplace characteristics or in cross-national studies to generate the difference in characteristics from demographers such as temperature and elevation, etc. (McCoach, 2010). Hierarchical modeling was first understood as a portion of the Bayesian estimation of linear models (Lindley & Smith, 1972). Lindley and Smith developed a basic framework for the complex error structure for nested data. However, this model is too weak to handle the estimation of covariance components from unbalanced data (Lindley & Smith, 1972). Therefore, no estimation method was proved to be useful until the expectation-maximization (EM) algorithm was developed (Dempster, Laird, & Rubin, 1977). Dempster suggests that the EM algorithm is suitable for estimating the covariance structure from hierarchical modeling (Dempster et al., 1981). After Strenio used hierarchical modeling in longitudinal trajectory for the first time (Strenio, Weisberg, & Bryk, 1983), two methods for covariance component estimation were proposed: reweighted generalized least squares (Goldstein, 1986) and a Fisher score algorithm (Longford, 1987).

In general, hierarchical modeling contains two stages. The first stage provides a function to measure an individual's growth with their random error, which is called a within-subjects model. The second stage defines the different subjects based on personal characteristics that influence individual growth parameters, such as sex, race, and so on (Lunn, Barrett, Sweeting, & Thompson, 2013). Thus, for trajectory analysis, hierarchical models are used because: (i) they focus on individual growth and estimate the trending trajectories' properties; (ii) individual development could be impacted by different factors that are particular to each individual, which is a way to find changing correlates; (iii) it allows hypothesis tests for effects from experimental treatments on repeat measurements to be implemented (Bryk & Raudenbush, 1987). The hierarchical model assumes normality for growth parameters (Bryk & Raudenbush, 1987). However, when the sample size is too small, variance and covariance based on the normality assumption may not be applicable. Suppose the normality criterion is not met or the sample size is too small. In that case, the correlation between baseline status, the rate of follow-up changing, and the growth parameters' reliability could be less (Bryk & Raudenbush, 1987).

When it was initially developed, hierarchical models could only be applied with continuous data with a normal distribution. However, this model is not easy to deal with discrete outcomes such as binary outcomes, count data, or ordered categorical outcomes. For these outcomes, it is impossible to assume linear models and normality at level one. To deal with this problem, software was developed to access discrete outcomes based on two-level or three-level hierarchical modeling (Goldstein, 1991). However, software approximations could be inaccurate in certain conditions (Gelman, Carlin, Stern, & Rubin, 1995). Therefore, two improved approximations for the maximum likelihood for the two-level model were developed based on the Gauss-Hermite quadrature and the high-order Laplace transform method (Pinheiro & Bates, 1995; Raudenbush, Yang, & Yosef, 2000).

Researchers can use the logit of hierarchical models to deal with discrete outcomes because of statistical and computational developments. For example, a longitudinal study was done with a binary outcome as high-rate offenders related to the changes in life circumstances (Horney, Osgood, & Marshall, 1995). In this study, the logistic regression model for individual change was defined in level 1; in level 2, researchers defined the variation of the individual parameter change. In another example, a longitudinal study was used to check the relationship between violent crime and neighborhood. In this study, the count outcomes (defined as the number of homicides) were applied to the two-level hierarchical model (Sampson, Raudenbush, & Earls, 1997). These models, called hierarchical generalized models, are widely used to deal with discrete structure outcomes. The two-level model is a nested model structure; for example, individuals (level 1) nested within households (level 2). Three steps need to be implemented in level 1: sampling model, finding link function, and variable prediction through structure model (McCoach, 2010).

Two other methodological developments in hierarchical models are commonly used: latent variable hierarchical modeling and Bayesian inference. Latent variable hierarchical models deal with unbalanced data. This approach uses observed, incomplete data to approximate the correlation among the latent variables (Robins, Rotnitzky, & Zhao, 1995). Another approach, the Bayesian inference, provides a more realistic standard error, creates various graphs and summary results, and compares with the MLE method. The Monte Carlo approximation, which contains data augmentation, and Gibbs sampling are often implemented to approximate the posterior for Bayesian inference (Tanner & Wong, 1987; Gelfand & Smith, 1990). Overall, hierarchical models can be used for trajectory analysis and rely on the multilevel modeling framework.

2.1.3 Latent growth curve modeling

Latent growth curve modeling (LGCM) is a covariance structure modeling for trajectory analysis. Baker used latent variables to deal with trajectory modeling in the factor analytic framework by introducing loadings based on a factor pattern matrix with differential stages of growth (Baker et al., 1954). Baker found that using factor analysis to reduce complicated repeat measurements to fewer relevant latent factors could help us better comprehend variations from the pattern. Baker's study is based on an unrestricted factor analytic model because he only selected four factors with 20 repeated measures. Unrestricted factor analysis, also called exploratory factor analysis, is a method for examining the internal reliability of a measure. Unrestricted factor analysis can reproduce correlations between observed variables based on the loading factors (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Loading factors represent the correlation coefficient between the factor and variable (Shevlin & Miles, 1998). The formal function of identifying variation with respect to time was developed using latent variable factor analysis for individual estimations. However, estimating particular functional types of growth is based on the parametrization of factors (Tucker & Lewis, 1973). With time, the confirmatory latent variable framework developed to embed trajectory modeling and the trajectory modeling framework. These new frameworks have the same power as structural equation models (SEMs) to estimate and test the variation of latent curved models (Meredith & Tisak, 1990). Instead of focusing on observed factors' interest, unobserved latent factors were also considered to promote the relationship between observed factors.

The LGCM can measure not only linear trajectories but also nonlinear trajectories. Based on the higher power of time measurements, such as quadratic and cubic polynomials, polynomial functions are usually used to deal with the nonlinear trajectory from LGCM (Cohen, 1978). LGCM can generally be divided into two kinds of modeling: the unconditional latent growth curve model and the conditional latent growth curve model. The unconditional latent growth curve model involves none of the covariates that may affect the trajectory. LGCM comes from the SEM perspective that uses the latent variables to determine the trajectories (Meredith & Tisak, 1990). On the other hand, the conditional latent growth curve model includes covariates or explanatory variables as a way to directly influence the random intercepts and slopes. Thus, individual trajectories would change when different covariates are included (Bollen & Curran, 2006). Two kinds of predictors can also be incorporated into the model: time-invariant covariates (TICs), which do not change over time, and time-variant covariates (TVCs), which do change over time. TICs should be independent of time, which means the TICs are consistent at each time point measurement (Curran, Bauer, & Willoughby, 2004). Nevertheless, TVCs are also easily expanded into the growth model. Unlike TICs, which directly predict group factors, TVCs indicate repeated measurements to control the effect of growth factors (Singer et al., 2003). The TVC model can include not only the interaction between time and TVCs, but also the interaction between TICs and TVCs. In other words, TICs are used to evaluate between-person effects, but TVCs assess within-person effects (Bollen & Curran, 2006).

Several extensions of the LGCM have developed since it was first used. The multivariate latent curve model estimates growth curves using two outcomes of repeat measurements in one model (McArdle, 2014). However, the multivariate latent curve model can only include TICs. If TVCs are included in the multivariate latent curve model, the relationship of time-specific structural should be generated among the repeated measures. To combine elements for both TVCs and the multivariate latent curve model, the autoregressive latent trajectory (ALT) model was developed (Curran & Bollen, 2001). The ALT model is suitable for modeling time-specific and random curve components at the same time. Moreover, except for handling continuous repeat measurements, the ALT model can also apply to other formats of the response data. The auxiliary threshold model is a way to estimate ordinal or dichotomous variables with polychoric moment structures based on the maximum likelihood estimation method (Olsson, 1979). The polychoric moment structure is an MLE method for the polychoric correlation between a pair of ordinal variables. Additionally, polychoric correlation is a technique that uses two observed ordinal variables to estimate the correlation between two theoretically normally distributed continuous latent variables (Ekström, 2011). The models for nominal or count variables are designed with the SEM approach but have some limitations such as over-dispersion (Rabe-Hesketh & Skrondal, 2004). However, alternative methods have been developed later, such as LGCM to handle Zero-Inflated Count Data (Liu, 2007; Yoon, Brown, Bowers, Sharkey, & Horn, 2015).

2.1.4 Group-based trajectory modeling

Group-based trajectory modeling (GBTM) is a trajectory analysis method that applies finite mixture modeling (FMM). Unlike other trajectory modelings such as hierarchical modeling and LGCM, parameters from GBTM are not estimated by cluster analysis. Instead, they are estimated by maximum likelihood estimation (MLE) (Nagin, 1999). As well, GBTM uses a multinomial modeling strategy, while hierarchical and latent growth curve modeling employs multilevel models and covariance structure models (Jones et al., 2001). GBTM using sandwich estimator is not influenced by covariance structure, random components are not permitted at any level of latent factors (Nagin & Tremblay, 2001). Though both hierarchical modeling and LGCM are focused on finding trajectories within the overall population or within individuals, GBTM is inclined to find sub-group trajectories based on defined unknown latent variables (Nagin, 1999).

As mentioned in Section 1.1, GBTM was first used in criminal offense studies (Nagin & Land, 1993) and has since been applied in more than 80 criminology studies (Piquero, 2008). Yet many studies outside of criminology also use this model (Bushway & Weisburd, 2006). Trajectory models are often built to study the relationship between etiology and mental health-related disorders. For example, a trajectory analysis of depressive symptoms during childhood and adolescence related to sex and depression outcomes when individuals become adults (Dekker et al., 2007). In another example, GBTM was used to investigate post-traumatic stress in veterans after the Gulf War (Orcutt, Erickson, & Wolfe, 2004). Furthermore, GBTM is frequently used in randomized clinical trials to discover heterogeneity in responses to treatment. It has also been used to handle causal inference from epidemiological observation studies when the outcomes are not randomized (Nagin et al., 2018)

In recent years, some model extensions have been incorporated into GBTM. Nagin introduced a way to include time-dependent covariance to influence within-subject effects and group-based dual trajectory modeling (GBDTM) to deal with a trajectory model containing two correlated outcomes (Nagin, 2005). Group-based multi-trajectory modeling (GBMTM) can include more than two associated outcomes, but the number of trajectory groups for each outcome must be the same (Nagin et al., 2018). Nagin wrote that, “By segmenting the data into trajectory groups, the group-based approach to studying development, provides an empirical means of summarizing large amounts of data in an easily comprehensible fashion and for testing long standing developmental theories with a taxonomic dimension” (Nagin & Odgers, 2010).

2.2 Review of depression and anxiety

2.2.1 Introduction to depression and anxiety

Mental health disorders, or mental illness, are defined as the various mental health conditions that may influence individuals' thoughts, perceptions, emotions, behavior, and relationships with others (WHO, 2020b). Compared to mood fluctuations and emotional responses to the challenges in daily life, depression and anxiety are more severe health conditions that may influence our work, studies, and relationships with our families (WHO, 2020a). Furthermore, depression and anxiety may lead to disability, or in extreme cases, suicide (Isometsä et al., 1994; Bruce, 2001). By WHO 2018 report, the number of people who have suffered from depression and anxiety at some point in their lives grew from 416 million to 615 million between 1990 and 2013, an increase of around 50%. Thus, almost 10% of the world's population has suffered from depression or anxiety (WHO, 2018). Globally, in 2015, the estimated prevalence was 4.4% for depression and 3.6% for anxiety (WHO, 2017).

The symptoms of depression include deep sadness and depressed mood; loss of interest in activities; appetite change; sleeping too little or too much; a lack of energy and increased fatigue; doing more in purposeless physical activities; speaking and moving more slowly; feeling worthless or guilty; having difficulty thinking, concentrating and decision making; and a willingness to die or commit suicide (WHO, 2017; Comstock & Helsing, 1977). The symptoms of anxiety include persistent and excessive worry, rapid breathing, feeling weak or tired, sleeping troubles, muscle tension, sweating, trembling, trouble concentrating, and being easily fatigued (Juson, 2018; Kawachi, Sparrow, Vokonas, & Weiss, 1994; Himmelfarb & Murrell, 1984). There is some overlap among symptoms in depression and anxiety, such as trouble concentrating or sleeping disorder and being more fatigued (Smith, 2018). However, two factors can help distinguish depression from anxiety. First, patients with depression usually move slowly, which means their reactions are listless or dull (Zigmond & Snaith, 1983). In contrast, patients with anxiety feel more keyed up to manage their random thoughts.

Second, patients with anxiety tend to be deeply worried about the future, while depressed patients tend to be listless and hopeless and do not much care about the events of the future (Kendall & Watson, 1989).

2.2.2 Depression and anxiety in South Korea

Depression and anxiety are increasing worldwide. Compared to most other areas, South-east Asia has higher rates of depression and anxiety. In 2015, 85.76 million in South-east Asia were suffering from depression, while 60.05 million were suffering from anxiety (WHO, 2017). These accounts were made up for 27% and 21% of the global depression and anxiety cases, respectively (WHO, 2017). In South Korea, the prevalence of depression disorders have consistently increased since 2001 (4% in 2001; 5.6% in 2006; 6.7% in 2011) (Cho & Lee, 2005; Cho et al., 2015). Similarly, the prevalence of anxiety disorders increased from 5% in 2006 to 6.8% in 2011, which is higher than any other East Asian countries (2.7% in China and 4.8% in Japan) (Cho et al., 2010, 2015; Shen et al., 2006). Korea has the highest suicide mortality rate (24.6 per 100000 individuals) among Organization for Economic Co-operation and Development (OECD) countries (OECD-data, 2019). Koo points to the country's high-speed economic growth and changing social values over the last 50 years as responsible for the increase in mental health problems among South Koreans (Koo, 2018).

The patterning of depression and depression by age is complex and highly affected by different cultural (Kirmayer et al., 2001; Lenze & Wetherell, 2011). Usually, depression and anxiety have less prevalence in older adults than young adults (Fiske, Wetherell, & Gatz, 2009; Lenze & Wetherell, 2011; Sutin et al., 2013). However, one study shows the prevalence of depression and anxiety among older people is higher than the overall population in South Korea (Cho, Lee, Kim, Lee, & Sohn, 2011). One study reported that 10% to 20% of the older people suffer from depression disorder; additionally, 9.1% to 33% of the older adults have clinically significant depressive symptoms (Cho et al., 2011). A prospective community-based study shows that the prevalence of anxiety symptoms among older adults is 38.1% (Kang et al., 2016). Another study of 1204 older individuals found that 10.2% suffered depression and 15.3% suffered anxiety. Moreover, 22.8% were identified with comorbid anxiety and depression (Kang et al., 2017). Given that South Korea's population is aging, the older people's tendency to underestimate or underreport mental illness will become problematic (Watkins, 2018). Some research suggests that older Korean adults may be less likely to admit to being depressed or anxious because most (78%) see depression or anxiety as a sign of weakness. In contrast, only 6% of older American people have the same perspective (Watkins, 2018). Therefore, it is necessary to pay more attention to depression and anxiety in the older population.

2.2.3 Comorbidity of depression and anxiety in older adults

Depression and anxiety are two frequently concurrent mental health problems in the older population (Lenze et al., 2001). Comorbid depression and anxiety is defined as patients with both depression and anxiety disorders (Lenze et al., 2001). An increasing number of researchers focus on studying the comorbidity of depression and anxiety in older adults. There are a few reasons why this is so. The first reason is that, in almost every country in the world, the population is aging. In South Korea, this is a more serious issue than in other developed countries (Isabella, 2017). People aged 65 or older in South Korea made up 3.8% of the population in 1980; this rose to 14.2% in 2020, a number that is more than double the number of people aged 14 or younger (Cho et al., 2011). Another reason why this topic has received increased scientific attention is that comorbid depression and anxiety in the older population involves different risk factors, presentation, comorbidity, and the course of the illness compared with youth. For example, older people who suffer from one or more chronic diseases have a greater chance of developing late-onset depression and/or anxiety (Manela, Katona, & Livingston, 1996; Krishnan, Hays, & Blazer, 1997).

In the 1990s, the comorbidity between depression and anxiety was less in older adults compared to young adults (Flint, 1994). However, as time goes by, a number of studies have found that the prevalence of comorbid depression and anxiety in geriatric populations is similar to the young adult population (Beekman et al., 1998). Furthermore, a study demonstrated that 47.5% of age 65 or older people who suffer depression also had comorbid anxiety disorders (Beekman et al., 2000). In South Korea, one study discovered that 69.3% of the older people with depression also have anxiety disorders, and 59.9% of older people with anxiety also have depression disorders (Kang et al., 2017). This significant change in prevalence can be explained both by the fact that anxiety is more common and by the fact that diagnostic instruments among older adults have improved (Lenze et al., 2001). In psychopathology, older patients suffering both depression and anxiety are regarded as more severe cases compared to the patients with only one disorder (Lenze et al., 2001). A study shows that depression patients with anxiety symptoms have more severe somatic symptoms compared to the patients with just depression (Flint & Rifat, 1997b). Gould et al. also found that compared with elevated depressive symptoms, anxiety is associated with greater multimorbidity in older adults in Health and Retirement Study (Gould, O'Hara, Goldstein, & Beaudreau, 2016). Moreover, lower social function and higher suicide rates persisted in comorbid depression and anxiety patients compared to patients with only depression or anxiety (Lenze et al., 2000; Allgulander & Lavori, 1993).

2.2.4 Risk factors for depression and anxiety in older adults

Being female, unmarried, and having a lower income are demographic risk factors that commonly relate to depression and/or anxiety (Blazer, Burchett, Service, & George, 1991; Heun, Papassotiropoulos, & Ptok, 2000). Depression has a higher prevalence in older people, especially ages 55-74, but anxiety does not change substantially according to age (WHO, 2017). Indeed, the WHO (2017) found that older people are slightly less anxious than younger segments of the population. Another study found that symptoms of depression declined with age for both males and females and that symptoms of anxiety had a significant decrease as women aged but not men (Henderson et al., 1998). However, anxiety disorders may be underdiagnosed later in life because of complications of medical comorbidity, cognitive decline and different symptoms compared to young adults (Wolitzky-Taylor, Castriotta, Lenze, Stanley, & Craske, 2010).

Common risk factors for both depression and anxiety in older adults have been divided into three categories: biological, social and psychological. These include physical illness, disability, bereavement, chronic disease, etc. (Vink, Aartsen, & Schoevers, 2008). Nevertheless, in longitudinal studies, risk factors for depression differ from those for anxiety among older people. For biological risk factors, cognitive functional impairment and visual defects are risk factors for depression but not anxiety, whereas hypertension is only a risk factor in anxiety (Beekman et al., 2000; Acierno et al., 2002; De Beurs et al., 2001; Forsell, 2000; Paterniti et al., 1999; R. A. Schoevers, Deeg, Van Tilburg, & Beekman, 2005; R. Schoevers, Beekman, Deeg, Jonker, & Tilburg, 2003). However, depression and anxiety are associated with cognitive functional impairments in other studies. Multiple studies showed that depressive disorder was associated with cognitive functional impairments such as deficits in verbal and nonverbal learning, memory, attention, visual and auditory processing, everyday problem-solving ability directly and indirectly, executive function, processing speed, and reasoning (Weisenbach, Boore, & Kales, 2012; Zuckerman et al., 2018; Yen, Rebok, Gallo, Jones, & Tennstedt, 2011). Anxious subjects did not differ significantly from depressed subjects in any measure of cognitive function (Mantella et al., 2007). However, anxiety was more likely associated with short-term and delayed memory, blackouts/memory loss, complex visuospatial performance and visual learning, poorer performance on verbal working memory, poor global cognitive functioning, working memory, inhibition, information processing speed, problem-solving including concept formation and mental flexibility (Mantella et al., 2007; Butters et al., 2011). For social risk factors, marital status and network size were only correlated with depression, while risk factors related to anxiety but not depression include being childless, traumatic life events and having a low income (Beekman et al., 2000; Acierno et al., 2002; De Beurs et al., 2001; Forsell, 2000; R. A. Schoevers et al., 2005; R. Schoevers et al., 2003;

Heun et al., 2000; Russo, Vitaliano, Brewer, Katon, & Becker, 1995). Social risk factors are any risks related to social support or social ties/isolation, such as the size or density of one's social network and frequency of contact with relatives and friends (Pirlich et al., 2005). For psychological risk factors, such as organizational culture and psychological and social support, are identical for both depression and anxiety (Beekman et al., 2000; De Beurs et al., 2001; R. Schoevers et al., 2003).

2.2.5 Treatments for depression and anxiety in the older adults

Treatments for depression and anxiety involve a pharmacological aspect and a psychosocial aspect (Diefenbach & Goethe, 2006). For pharmacological treatments, the medications used to treat depression and anxiety in general adults can also be applied in the aged (Doraiswamy, 2001). Usually, antidepressants widely available on the market may also be effective for treating one or more anxiety disorders. For example, Paroxetine as a selective serotonin reuptake inhibitors (SSRIs) is effective for major depressive disorder (MMD), generalized anxiety disorder (GAD), panic disorder, etc. (Diefenbach & Goethe, 2006). As seen in the literature, SSRIs and dual serotonin and norepinephrine reuptake inhibitors (SNRIs) are more suitable for treating comorbid depression and anxiety in older adults compared to the antianxiety agents, (i.e., benzodiazepines or tricyclic antidepressants (TCAs) and the monoamine oxidase inhibitors (MAOIs)) (Diefenbach & Goethe, 2006; Doraiswamy, 2001). Despite the availability of effective medication, treatment for comorbid depression and anxiety remains challenging (Diefenbach & Goethe, 2006). For example, some studies showed that nortriptyline as a TCAs used in depressed older individuals with anxiety symptoms had a lower response rate, a higher drop-off rate and delayed response compared to the older people who only have depression (Flint & Rifat, 1997a; Dew et al., 1997). However, a study found no significant difference in drop-off rate, treatment response, and side effect change between older depressive patients with anxiety symptoms or not (Lenze et al., 2003). Venlafaxine XR, an SNRI, was shown to be an appropriate treatment for older patients with depression and/or anxiety because it is effective for both depression and anxiety, has few side effects, and minimal risk of drug interactions (Doraiswamy, 2001).

Psychosocial interventions are another type of treatment to treat depression and anxiety. They can treat depression and anxiety alone or in combination with pharmacological intervention (Lebowitz et al., 1997). Cognitive-behavioral therapy (CBT), as a short-term, problem-focused treatment, focuses on teaching and/or strengthening coping skills (Diefenbach & Goethe, 2006). CBT is effective and widely used to treat both depression and anxiety (Areán & Cook, 2002; Stanley et al., 2003). Interpersonal therapy (IPT), another therapy, focuses on patients' interpersonal problem solving and ability to process emotional distress (Diefenbach & Goethe, 2006). IPT combined with pharmacological interventions will have superior treatment results compared to IPT alone (Areán & Cook, 2002). There is no specific psychosocial intervention developed for comorbid depression and anxiety in older patients (Diefenbach & Goethe, 2006). Since CBT has the ability to treat depression or anxiety, it can be used to treat older patients with comorbid depression and anxiety as well (Wetherell, Sorrell, Thorp, & Patterson, 2005).

Chapter 3 STATISTICAL METHODS

3.1 Introduction

Group-based trajectory modeling is different from other traditional longitudinal modeling methods for trajectory analysis. The finite mixture model, the basic theory for group-based trajectory modeling, will be introduced in Section 3.2 in details. Group-based trajectory modeling with model structures for different types of outcome data and model selection methods will be discussed in Section 3.3. In Section 3.4, an extension of group-based trajectory modeling, multivariate group-based modeling based on conditional probabilities will be presented.

3.2 Finite mixture modeling

3.2.1 Introduction

The finite mixture model (FMM) is a statistical model widely applied in biology, genetics, psychiatry, and marketing, among other disciplines (Geoffrey & Peel, 2000). Many researchers use FMM as a tool for analysis because of the model's flexibility. Not only can FMM be applied in different areas of study, but it can also be used for different kinds of statistical analysis, including cluster analysis, image analysis, latent class analysis, and even survival analysis (Geoffrey & Peel, 2000). Medical image analysis increasingly uses FMM to model pixel values to combine the various mixed portions of different populations (Frosio, Ferrigno, & Borghese, 2006). For example, a study of retinal image analysis used mixture models to find hard exudates (Sánchez, García, Mayo, López, & Hornero, 2009). FMM aims to generate the heterogeneity that characterizes unobserved clusters from the overall population. Furthermore, it provides a convenient semiparametric framework for solving unknown distribution shapes instead of the variance or covariance structure (Geoffrey & Peel, 2000).

FMMs have a relatively long history of applications in statistics. The first time this model was applied was more than 120 years ago when Pearson used the mixture model for sub-normal distributions (Pearson, 1894). In the last fifty years, the method of maximum likelihood was recognized for fitting FMMs. Wolfe was the first to apply the MLE method to fit FMMs (Wolfe, 1967). The first time the EM algorithm was used to simulate an FMM was ten years later (Dempster et al., 1977).

3.2.2 Definition of a finite mixture model

If $Y_1, Y_2, ..., Y_n$ are defined as a random sample of size n, so T-dimensional random vector Y_i follows the probability density function $f(y_i)$ on R^t. Y_i denotes the $i's$ individual including random variables amounting to t measurements. If $\top$ is defined as the vector transpose, $Y = (Y_1^\top, Y_2^\top, ..., Y_n^\top)^\top$, where Y is defined as an n-tuple of points in R^t, represents the sample of interest. $y = (y_1^\top, y_2^\top, ..., y_n^\top)^\top$ is defined as an observed random sample, where y_i is the observed value of the random vector for person i.

$f(y_i)$ can be viewed as a density where Y_i is discrete by adopting counting measurements, even though the feature of vector Y_i is a continuous random vector. The form of density $f(y_i)$ of Y_i can be written as (Geoffrey & Peel, 2000):

$$f(y_i) = \sum_{j=1}^{J} \pi_j f_j(y_i), \tag{3.1}$$

where

$$0 \leq \pi_j \leq 1$$

and

$$\sum_{j=1}^{J} \pi_j = 1.$$

$\pi_1, \pi_2, ..., \pi_J$ are defined for mixing weights (proportions) and stand for the number of distributional sub-populations. J is the size of mixture components or weights. $f_j(y_i)$ denotes the j^{th} component densities of the mixture $j = 1, \ldots, J$. Since a finite mixture of distributions is focused on in most situations, we consider FMMs instead of just mixture models.

The size of the components J is fixed in equation (3.1). However, J is unknown and will depend on the real data, together with the mixing weights and the component density parameters from the specified forms. The number of mixture components can be increasing when the sample size is large enough. This model is called a Gaussian mixture sieve (Geman & Hwang, 1982).

3.2.3 Finite mixture model with parameters

If we assume $f_j(y_i)$ as the component density belongs to the specific parametric family, then, $f_j(y_i)$ can be written as $f_j(y_i; \Theta_j)$. Θ_j is denoted as a vector of unknown parameters from the form of the j^{th} component density assumed within the mixture. Instead of $f(y_i)$, the finite mixture model will be changed to:

$$f(y_i; \Psi) = Pr(y_i) = \sum_{j=1}^{J} \pi_j f_j(y_i; \Theta_j), \tag{3.2}$$

where Ψ means the vector involving all the unknown parameters from the mixture model. Ψ is defined as:

$$\Psi = (\pi_1, \ldots, \pi_{J-1}, \xi^\top)^\top.$$

$\xi = (\Theta_1, \ldots, \Theta_J)$ is defined as a vector that obtains all the parameters of the density in different components. $\top$ is the vector transpose. The prior numbers of these vectors are given to distinguish them from one another. We define Ω as the specified parameter space for Ψ and allow

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)^\top$$

as the vector for every mixture proportion. π_J is redundant because the sum of all the proportions π_j from the mixture model is equal to 1. Therefore, π_J is not included in the vector Ψ.

Usually, each of the component density of $f_j(y_i;\Theta_j)$ should belong to the same parametric family in the finite mixture model. Therefore, equation (3.2) can be reducible to

$$f(y_i;\Psi) = Pr(y_i) = \sum_{j=1}^{J} \pi_j f(y_i;\Theta_j), \tag{3.3}$$

3.2.4 Likelihood of a finite mixture model

If we define Ψ as the maximum likelihood function to estimate the parameters from a mixture distribution, then a sample is given as:

$$y_i \overset{iid}{\sim} f(y|\Psi), \quad i = 1, ..., n. \tag{3.4}$$

What we are interested in is generating the maximum likelihood estimation (MLE) of Ψ, where $\hat{\Psi}$ should be equal to (Schlattmann, 2009):

$$\hat{\Psi} = \underset{\Psi}{argmax} L(\Psi).$$

argmax is defined as arguments of the maxima, which means the set of inputs y from the domain D that achieves the highest function value

$$L(\Psi) = \prod_{i=1}^{n} \sum_{j=1}^{J} f(y_i;\Theta_j)\pi_j, \tag{3.5}$$

The alternative method is usually used to find estimates of Ψ, which is a log-likelihood function:

$$\ell(\Psi) = logL(\Psi) = \sum_{i=1}^{n} log \sum_{j=1}^{J} f(y_i;\Theta_j)\pi_j. \tag{3.6}$$

To estimate Ψ, we can assume:

$$S(y;\Psi) = \frac{\partial \ell(\Psi)}{\partial \Psi} = 0, \tag{3.7}$$

$S(y;\Psi)$ is the first derivative of the log-likelihood function defined as the score function, but there is no closed-form to find its solution in most cases. Therefore, several other alternative methods, such as EM algorithm, quasi-likelihood, etc., to estimate the likelihood of the finite mixture modeling were developed.

The expected Fisher information matrix for the vector of parameters Ψ is given by:

$$\mathcal{I}(\Psi) = E_{\Psi}\{S(y;\Psi)S^{\top}(y;\Psi)\} \tag{3.8}$$

where $S(y;\Psi)$ is the score function with observed data y and $E_{\Psi}\{S(y;\Psi)S^{\top}(y;\Psi)\}$ is the expectation of $S(y;\Psi)S^{\top}(y;\Psi)$. Usually, $\mathcal{I}(\Psi)$ can be also written as:

$$\mathcal{I}(\Psi) = E_{\Psi}I(\Psi;Y), \tag{3.9}$$

where

$$I(\Psi;Y) = -\frac{\partial^2 logL(\Psi)}{\partial\Psi\partial\Psi^{\top}}, \tag{3.10}$$

which is the negative of the Hessian matrix. Therefore, the observed Fisher information matrix is expressed as $I(\hat{\Psi};Y)$ (Geoffrey & Peel, 2000).

The asymptotic covariance matrix of the MLE $\hat{\Psi}$ is the inverse of the expected Fisher information matrix $\mathcal{I}(\Psi)$ and the approximation is $\mathcal{I}(\hat{\Psi})$. In common practice, the observed Fisher information matrix $I(\hat{\Psi};y)$ is used to estimate the covariance matrix of the MLE instead of the expected Fisher information matrix because it is easier to use without expectations when $\Psi = \hat{\Psi}$. Therefore, the standard error can be approximated as:

$$SE(\hat{\Psi}_r) \approx (I^{-1}(\hat{\Psi};y))^{1/2}_{rr} \quad r = 1,\dots,d, \tag{3.11}$$

where rr the rows and columns from the covariance matrix and d is the number of parameters from the matrix (Geoffrey & Peel, 2000).

3.2.5 Estimation of a finite mixture model

If the number of components is defined as support size, the estimation of a finite mixture model contains two cases (Schlattmann, 2009):

> Case 1: Support size is flexible, which means there is no assumption of how many components J is determined.
>
> Case 2: Support size is fixed, which means the number of components J is assumed to be known. Therefore, the unknown parameter should be the mixing proportions of π_j and parameters Θ_j from the sub-population.

In the flexible support size case, the algorithms require knowledge from convex geometry, and optimization (Boyd & Vandenberghe, 2004). Convex sets and functions provided the necessary background from which to derive the theory of semiparametric finite mixture models. Based on convex geometry, numerous converging algorithms were developed to solve the problem of directional derivatives for the flexible support size case (Böhning, 1995; Lindsay & Lesperance, 1995). Usually, two methods are used for the flexible support size case: the vertex direction method and the vertex exchange method (Schlattmann, 2009). These two methods will not be introduced in detail here since this thesis focuses on the fixed support size case.

3.2.5.1 Newton-Raphson for fixed support size

The Newton–Raphson (NR) method is one method that can be used for a maximum likelihood estimation when there is no closed-form available for the solution in equation (3.6). NR can be used to solve the likelihood equation (3.6) using a Taylor series expansion to approximate Ψ by the current fit $\Psi^{(l)}$ at the l_{th} procedure. The approach based on the Taylor series provides:

$$S(y; \Psi) \approx S(y; \Psi^{(l)}) + I(y; \Psi^{(l)})(\Psi - \Psi^{(l)}), \tag{3.12}$$

where $I(y;\Psi)$ is the information matrix. If we need to find a new fit $\Psi^{(l+1)}$, the right part of equation (3.12) needs to be assumed to be zero, and $\Psi^{(l+1)}$ is solved as:

$$\Psi^{(l+1)} = \Psi^{(l)} - I^{-1}(y;\Psi^{(l)})S(y;\Psi^{(l)}) \tag{3.13}$$

where $I^{-1}(.)$ is the inverse of the information matrix $I(.)$.

The benefit of the NR algorithm for approximation is that the convergence speed is the fastest compared to other algorithms (Everitt, 1984). However, there are two major issues in applying the NR method. First, the Fisher information matrix $I(y;\Psi^{(l)})$ is a $d \times d$ matrix, and when d is large, the computation of this matrix is complicated. The other problem is that this method may not converge to the maximum when the prior of Ψ is not guessed correctly (Titterington, Smith, & Makov, 1985).

3.2.6 EM algorithm

The expectation-maximization (EM) algorithm is another method for estimating maximum likelihood, especially for a model with unobserved latent variables (Pilla & Lindsay, 2001; Vlassis & Likas, 2002). In general, the particular model structure needed for the EM algorithm and data augmentation is the most important point of the EM algorithm (Schlattmann, 2009). The general description of the EM algorithm is defined as a model with parameters Ψ not only for the observed data y but also the missing data z. If we decide on maximizing only the observed data y, defined as $L_y(\Psi)$, it is not easy to maximize the likelihood, especially from the mixture cases (Geoffrey & Peel, 2000). On the other hand, if the unobserved data z is assumed to be known, maximizing the complete likelihood $L_c(\Psi) = L_{y,z}(\Psi)$ would be much easier for finding the maximization. The "missing data" may be completely imaginary and should have the same marginal distribution with variable y (McLachlan & Krishnan, 2008).

If we assume the current parameter value is $\Psi^{(0)}$ and wish to find $Q(\Psi,\Psi^{(0)})$, which means the conditional expectation of the complete data $logL_c(\Psi)$ when the observed data is provided, then, the E-step of the EM algorithm is (Dempster et al., 1977):

$$Q(\Psi, \Psi^{(0)}) = E_{\Psi^{(0)}}\left[(\Psi|y)\right]. \tag{3.14}$$

To maximize $Q(\Psi, \Psi^{(0)})$ with respect to Ψ based on the parameter space Ω, the M-step means selecting $\Psi^{(1)}$ for

$$Q(\Psi^{(1)}, \Psi^{(0)}) \geq Q(\Psi, \Psi^{(0)}), \quad \forall \Psi \in \Omega. \tag{3.15}$$

Then, the E-step and M-step will keep going with $\Psi^{(1)}$ instead of $\Psi^{(0)}$. The E-step and M-step for the $(l+1)th$ iteration can be denoted as:
E-step:

$$Q(\Psi, \Psi^{(l)}) = E_{\Psi^{(l)}}\left[(\Psi|y)\right]. \tag{3.16}$$

M-step:

$$Q(\Psi^{(l+1)}, \Psi^{(l)}) \geq Q(\Psi, \Psi^{(l)}), \quad \forall \Psi \in \Omega. \tag{3.17}$$

More detailed theories for EM algorithms can be found in several books or articles (Dempster et al., 1977; McLachlan & Krishnan, 2008). I will not introduce it any further since it has been fully developed and widely used for a number of applications. Instead, I am only interested in the mechanism of how the EM algorithm works in estimating finite mixture models.

3.2.7 Latent variables in finite mixture modeling

Models involving latent variables are regarded as a probability model with unobserved certain variables (Bartholomew, Knott, & Moustaki, 2011). As a by-product of the analysis, the EM algorithm is able to estimate parameters and latent variables (Sammel, Ryan, & Legler, 1997). Recalling equation (3.3), let us define $\boldsymbol{z}_i$ which is a J-dimensional binary random variable, as the latent variable with only one element $z_{ij} = 1$ and the rest of elements are 0. Therefore, the values of z_{ij} meet the conditions of $z_{ij} \in \{0, 1\}$, $\sum_j (z_{ij}) = 1$ and J possible states for $\boldsymbol{z}_i = z_{i1}, z_{i2}, \ldots z_{iJ}$. If person belong to j^{th} group, then, $z_{ij} = 1$, otherwise, $z_{ij} = 0$. The joint distribution is denoted as $Pr(y_i, \boldsymbol{z}_i) = Pr(\boldsymbol{z}_i)Pr(y_i|\boldsymbol{z}_i)$ with marginal distribution $Pr(\boldsymbol{z}_i)$ and conditional distribution $Pr(y_i|\boldsymbol{z}_i)$. The marginal distribution can be written as:

$$Pr(z_{ij} = 1) = \pi_j.$$

Since 1-of-J presentation is used for $\boldsymbol{z}_i$, the marginal distribution can also be given as:

$$Pr(\boldsymbol{z}_i) = \prod_{j=1}^{J} \pi_j^{z_{ij}}. \tag{3.18}$$

With the same method, the conditional distribution can be presented as:

$$Pr(y_i|z_{ij} = 1) = f(y_i; \Theta_j),$$

or

$$Pr(y_i|\boldsymbol{z}_i) = \prod_{j=1}^{J} f(y_i; \Theta_j)^{z_{ij}}. \tag{3.19}$$

Therefore, the margin distribution can be observed by the sum of $\boldsymbol{z}_i$:

$$Pr(y_i) = \sum_{\boldsymbol{z}_i} Pr(y_i, \boldsymbol{z}_i) = \sum_{\boldsymbol{z}_i} Pr(\boldsymbol{z}_i)Pr(y_i|\boldsymbol{z}_i) = \sum_{j=1}^{J} \pi_j f(y_i; \Theta_j). \tag{3.20}$$

This means that each observed data point y_i corresponds with a latent variable $\boldsymbol{z}_i$. Latent variables are essential for using the EM algorithm to find the closed form of the parameters of the mixture model (Geoffrey & Peel, 2000).

3.2.8 EM algorithm of one-dimensional two-Gaussian mixture model

Let us start with the simplest common case, a one-dimensional two-Gaussian mixture model, to estimate the FMM with the EM algorithm. In this model, the dimensional $T = 1$, the proportions are assumed to be $J = 2$ and $f(y_i; \Theta_j)$ should follow normal distribution $\mathcal{N}(y_i|\mu_j, \sigma_j^2)$. Then, the likelihood is presented as:

$$L(\mu, \sigma; y) = \prod_{i=1}^{n} (1 - \pi)\mathcal{N}(y_i|\mu_1, \sigma_1^2) + \pi\mathcal{N}(y_i|\mu_2, \sigma_2^2).$$

The log-likelihood can be derived as:

$$\ell(\mu, \sigma; y) = \sum_{i=1}^{n} log[(1 - \pi)\mathcal{N}(y_i|\mu_1, \sigma_1^2) + \pi\mathcal{N}(y_i|\mu_2, \sigma_2^2)]. \tag{3.21}$$

Since the sum inside the logarithm is hard to calculate from the marginal likelihood based on the observed data, the completed likelihood containing latent variables will be applied for estimation:

$$\ell(\mu, \sigma; y, z) = \sum_{i=1}^{n} [(1 - z_i)log\mathcal{N}(y_i|\mu_1, \sigma_1^2) + z_i log\mathcal{N}(y_i|\mu_2, \sigma_2^2) + (1 - z_i)log\pi + z_i log(1 - \pi)], \tag{3.22}$$

where z_i is the unobserved latent variable for subject i with values 0 or 1. If $z_i = 1$, y_i is from the second Gaussian model, otherwise, y_i is from the first Gaussian model (Friedman, Hastie, & Tibshirani, 2001).

If the initial values of parameters $(\hat{\pi}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2)$ is given, then for E step, the unknown latent variable z_i will be substituted with the expected value $\gamma(z_i)$ based on Bayes' rule, defined as:

$$\begin{aligned}\gamma(z_i) &= E(z_i|y_i) = Pr(z_i = 1|y_i) \\ &= \frac{Pr(z_i = 1)Pr(y_i|z_i = 1)}{Pr(z_i = 0)Pr(y_i|z_i = 0) + Pr(z_i = 1)Pr(y_i|z_i = 1)} \\ &= \frac{\hat{\pi}\mathcal{N}(y_i|\hat{\mu}_2, \hat{\sigma}_2^2)}{(1-\hat{\pi})\mathcal{N}(y_i|\hat{\mu}_1, \hat{\sigma}_1^2) + \hat{\pi}\mathcal{N}(y_i|\hat{\mu}_2, \hat{\sigma}_2^2)},\end{aligned} \tag{3.23}$$

where $i = 1, 2, \ldots, n$ is the number of observations. The expected complete likelihood will then be developed with the expected value of $\gamma(z_i)$. The M-step updates the parameters based on the expected complete likelihood maximization (Friedman et al., 2001). The parameters are estimated as:

$$\begin{aligned}\hat{\mu}_1 &= \frac{\sum_{i=1}^n (1-\gamma(z_i))y_i}{\sum_{i=1}^n (1-\gamma(z_i))} \\ \hat{\mu}_2 &= \frac{\sum_{i=1}^n \gamma(z_i)y_i}{\sum_{i=1}^n \gamma(z_i)} \\ \hat{\sigma}_1^2 &= \frac{\sum_{i=1}^n (1-\gamma(z_i))(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^n (1-\gamma(z_i))} \\ \hat{\sigma}_2^2 &= \frac{\sum_{i=1}^n \gamma(z_i)(y_i - \hat{\mu}_2)^2}{\sum_{i=1}^n \gamma(z_i)} \\ \hat{\pi} &= \frac{\sum_{i=1}^n \gamma(z_i)}{n}\end{aligned} \tag{3.24}$$

The E and M step will be iterated until convergence.

3.2.9 EM algorithm for Gaussian mixture models

For finite mixture modeling, if we assume component membership as the missing data part, the vector of observed data is defined as

$$\boldsymbol{y} = (y_1^\top, y_2^\top, \ldots, y_n^\top)^\top,$$

which is incomplete because the component-label vectors are not involved. Therefore, unknown indicator variables should be defined as $\boldsymbol{z}$, which correspond to $\boldsymbol{y}$. $\boldsymbol{z}_i$ from $\boldsymbol{z}$ is a J-dimensional vector with $z_{ij} = (z_i)_j \in \{0, 1\}$, relate to if y_i belong to j^{th} component of the mixture $(j = 1, \ldots, J; i = 1, \ldots, n)$. Here again, J is the number of mixture components and n is the study sample size. Thus, the complete data vector can be denoted as

$$\boldsymbol{y}_c = (\boldsymbol{y}^T, \boldsymbol{z}^T)^T, \tag{3.25}$$

where

$$\boldsymbol{z} = (\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, \ldots, \boldsymbol{z}_n^T)^T. \tag{3.26}$$

The likelihood for $(y_i, z_{i1}, z_{i2}, \ldots, z_{iJ})^T$ of the i^{th} person observation is declared as (Schlattmann, 2009):

$$\begin{aligned} &Pr(Y_i = y_i, Z_{i1} = z_{i1}, \ldots, Z_{iJ} = z_{iJ}) \\ &\qquad = Pr(Y_i = y_i | Z_{i1} = z_{i1}, \ldots, Z_{iJ} = z_{iJ}) Pr(Z_{i1} = z_{i1}, \ldots, Z_{iJ} = z_{iJ}) \\ &\qquad = \prod_{j=1}^{J} \pi_j^{z_{ij}} f(y_i, \Theta_j)^{z_{ij}}, \end{aligned} \tag{3.27}$$

where $z_{ij} = 0$ if y_i is not observed and $z_{ij} = 1$ if y_i is observed. Θ_j is a vector of unknown parameters from the postulated form for the j^{th} component density in the mixture model.

Additionally, the complete likelihood with all subjects is written as:

$$L_c(\Psi) = \prod_{i=1}^{n} \prod_{j=1}^{J} \pi_j^{z_{ij}} f(y_i, \Theta_j)^{z_{ij}}, \tag{3.28}$$

and the complete log-likelihood is given by:

$$\ell_c(\Psi) = \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} log f(y_i, \Theta_j). \tag{3.29}$$

Depending on the complete log-likelihood, the E-step calculates the current conditional expectation of z_{ij} based on the observed y (Geoffrey & Peel, 2000).

For the univariate mixture Gaussian model in a single dimension ($D = 1$), the proportions are assumed to be J and $f(y_i; \Theta_j)$ should follow normal distribution $\mathcal{N}(y_i|\mu_j, \sigma_j^2)$. Using Bayes rule, we can get:

$$\begin{aligned} Pr(Z_{ij} = 1|Y_i = y_i) &= \frac{Pr(Y_i = y_i|Z_{ij} = 1)Pr(Z_{ij} = 1)}{\sum_{m=1}^{J} Pr(Y_i = y_i|Z_{im} = 1)Pr(Z_{im} = 1)} \\ &= \frac{\sum_{z_{ij}} z_{ij}[\pi_j \mathcal{N}(y_i|\mu_j, \sigma_j^2)]^{z_{ij}}}{\sum_{z_{im}}[\pi_m \mathcal{N}(y_i|\mu_m, \sigma_m^2)]^{z_{im}}} \\ &= \frac{\pi_j(y_i|\mu_j, \sigma_j^2)}{\sum_{m=1}^{J} \pi_m(y_i|\mu_m, \sigma_m^2)} = \gamma(z_{ij}), \end{aligned} \tag{3.30}$$

where $\gamma(z_{ij})$ stands for the posterior probability. Thus, using $\gamma(z_{ij})$ instead of z_{ij}, the E-step of the mixture model is:

$$\begin{aligned} Q(\Psi, \Psi^{(l)}) &= E_{\Psi^{(l)}}(logL_c(\Psi)) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{J} \gamma(z_{ij}) log\pi_j + \sum_{i=1}^{n} \sum_{j=1}^{J} \gamma(z_{ij}) log\mathcal{N}(y_i|\mu_j, \sigma_j^2). \end{aligned} \tag{3.31}$$

With the M-step, we need to maximize the likelihood of equation (3.31). For this likelihood, the left and right part can be considered separately for maximization (Geoffrey & Peel, 2000). If we assume:

$$n_j = \sum_{i=1}^{n} \gamma(z_{ij}) \tag{3.32}$$

$\pi_j^{(l+1)}$ could be maximized when the left part is equal to 0:

$$\pi_j^{(l+1)} = \frac{n_j}{n} = \frac{\sum_{i=1}^{n} \gamma(z_{ij})}{n} = \sum_{i=1}^{n} \frac{\pi_j(y_i|\mu_j, \sigma_j^2)}{n \sum_{m=1}^{J} \pi_m(y_i|\mu_m, \sigma_m^2)}. \tag{3.33}$$

The right part of (3.33) is implemented for maximizing the unknown parameters. For the one-dimensional Gaussian case, $\mu_j^{(l+1)}$ is equal to

$$\mu_j^{(l+1)} = \frac{\sum_{i=1}^{n} \gamma(z_{ij}) y_i}{\sum_{i=1}^{n} \gamma(z_{ij})} = \frac{\sum_{i=1}^{n} \gamma(z_{ij}) y_i}{n_j}. \tag{3.34}$$

Consider $f(y; \mu_j, \sigma^2)$ as normal density:

$$f(y; \mu_j, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y-\mu_j)^2}{2\sigma^2}), \tag{3.35}$$

where σ^2 is the common variance for all weights to simplify the likelihood of component-specific variance σ_j^2. To maximize σ^2, we need to fit equations (3.33), (3.34), and (3.35) into equation (3.31) and obtain

$$\sum_{i=1}^{n} \sum_{j=1}^{J} \frac{\gamma(z_{ij})(y_i - \lambda_j)^2}{\sigma^4} + \sum_{i=1}^{n} \sum_{j=1}^{J} \frac{\gamma(z_{ij})}{\sigma^2} = 0. \tag{3.36}$$

By calculation, the maximized σ^2 should be equal to

$$\sigma_{l+1}^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \gamma(z_{ij})(y_i - \mu_j)^2, \tag{3.37}$$

If we use the EM algorithm to maximize the variance of each component, we need to follow the method developed by Böhning (Böhning, 1995). The benefit of using the EM algorithm for estimations is that it will undoubtedly converge. However, it can only converge to a local maximum, and the speed of convergence might be slow for some cases (Geoffrey & Peel, 2000).

For multivariate Gaussian mixture models, the EM algorithm to handle the MLE is similar to the Gaussian mixture model with only one variable. The difference is the multivariate Gaussian mixture models involve the D multidimensional dataset now, which means the data can be presented as an $n * D$ matrix. Therefore, $\boldsymbol{\mu}_j$ is a vector of means and Σ_j is the covariance matrix from the j^{th} component (Bishop, 2006). The complete likelihood for multivariate Gaussian mixture models is presented as:

$$L_c(\Psi) = \prod_{i=1}^{n} \prod_{j=1}^{J} \pi_j^{z_{ij}} \mathcal{N}(y_i|\boldsymbol{\mu}_j, \Sigma_j)^{z_{ij}}, \tag{3.38}$$

With a logarithm, we can get:

$$\ell_c(\Psi) = logL_c(\Psi) = \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} log \mathcal{N}(y_i|\boldsymbol{\mu}_j, \Sigma_j). \tag{3.39}$$

Based on the current parameter, the responsibilities are able to be evaluated with E-step:

$$\gamma(z_{ij}) = \frac{\pi_j(y_i|\boldsymbol{\mu}_j, \Sigma_j)}{\sum_{m=1}^{J} \pi_m(y_i|\boldsymbol{\mu}_m, \Sigma_m)} \tag{3.40}$$

Then, using the current responsibilities, the parameters are re-estimated with M-step:

$$\boldsymbol{\mu}_j^{(l+1)} = \frac{\sum_{i=1}^{n} \gamma(z_{ij}) y_i}{n_j}.$$

$$\Sigma_j^{l+1} = \frac{1}{n_j} \sum_{i=1}^{n} \sum_{j=1}^{J} \gamma(z_{ij})(y_i - \boldsymbol{\mu}_j^{(l+1)})(y_i - \boldsymbol{\mu}_j^{(l+1)})^\top,$$

$$\pi_j^{(l+1)} = \frac{n_j}{n} \tag{3.41}$$

where $n_j = \gamma(z_{ij})$. The EM algorithm with the mixture binary dataset, called latent class analysis, is described by Bernoulli distributions (Geoffrey & Peel, 2000). The EM algorithm progress is similar to the Gaussian mixture models. Therefore, it will not be introduced in detail here.

3.2.10 EM algorithm for a generalized mixture model

3.2.10.1 Generalized linear model

Before adding covariates into the mixture models, we will first introduce generalized linear models (GLMs). GLMs assume the response y_i is from the exponential distribution family, meaning the response may not only be continuous (Wedderburn, 1974). The general density function of distribution for exponential family is written as (Nelder & Wedderburn, 1972):

$$f(y_i;\theta_i,\phi) = exp(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)), \tag{3.42}$$

where θ represents the natural or canonical parameter, ϕ represents the dispersion parameter and $a(.), b(.)\,\&\, c(.)$ are known special forms of functions, $a_i(\phi)$ is the form of

$$a_i(\phi) = \frac{\phi}{\tau_i}, \tag{3.43}$$

where τ_i stands for the prior weight usually equal to one, the log-likelihood function can be denoted as:

$$logf(y_i;\theta_i,\phi) = \ell(y_i,\theta_i,\phi) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi). \tag{3.44}$$

The mean and variance of random variable Y_i are given by:

$$E(Y_i) = \mu_i = b'(\theta_i), \tag{3.45}$$

and

$$var(Y_i) = -b''(\theta_i)a_i(\phi). \tag{3.46}$$

The $b'(.)$ and $b''(.)$ are the first and second derivative of $b(\theta_i)$.

To provide the relationship between linear predictors and means in GLMs, we need to use link functions, to calculate the mean using a one-to-one continuous differentiable transformation, denoted as (Schlattmann, 2009):

$$\eta_i = g(\mu_i). \tag{3.47}$$

The transformed mean should follow a linear relationship as:

$$\eta_i = x_i\beta, \tag{3.48}$$

where η_i is the estimated linear predictor, $x_i = (x_{i1}, \ldots, x_{im})$ is the vector of covariates, m is the number of covariates that affect η_i and β is the unknown vector of parameters (Schlattmann, 2009). The inverse function can be generated since it is a one-to-one transformation

$$\mu_i = g^{-1}(x_i\beta). \tag{3.49}$$

The maximum likelihood function estimation of GLMs is provided as

$$\ell(y;\theta_i,\phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i,\phi). \tag{3.50}$$

Therefore, the estimate of β should be written as (McCullagh & Nelder, 1989):

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{a_i(\phi)} \frac{d\eta_i}{d\mu_i} x_{ij} = 0, \tag{3.51}$$

w_j is the weight function declared as:

$$w_i = \tau / \left[b''(\theta_i)(\frac{d\eta_i}{d\mu_i})^2 \right], \tag{3.52}$$

where τ is the weight from $a_i(\phi)$ shown in equation (3.43).

The Fisher's score method could be used for solving the likelihood from equation (3.51) (Nelder & Baker, 2004). With an estimated $\hat{\eta}_i = x_i\beta$ and $\hat{\mu}_i = g^{-1}(\mu_i)$, on the next iteration, the adjusted response variable $\hat{y_i}$ is calculated as

$$\hat{y_i} = \hat{\eta_i} + (y_i - \hat{\mu_i})\frac{d\eta_i}{d\mu_i}. \tag{3.53}$$

We can then recalculate the weight w_i based on the new response $\hat{y_i}$. Finally, the renewed estimation of β can be obtained from a matrix notation:

$$\hat{\beta} = (X^T W X)^{-1} X^T W y, \tag{3.54}$$

where X represents the design matrix, W is the diagonal matrix entered by w_i, and y is the response vector with entries $\hat{y_i}$ from equation (3.53), $y = (y_1, y_2, \ldots, y_n)$ (Schlattmann, 2009).

3.2.10.2 Maximum likelihood estimation of a finite mixture GLM using EM algorithm

The GLM mixture model specifies that we replace the Gaussian density function to the GLMs, which is changed by (Jansen, 1993):

$$f(y_i; \Psi) = \sum_{j=1}^{J} \pi_j f(y_i, \theta_i, \phi_j), \qquad j = 1, \ldots, J, \tag{3.55}$$

where the log density of the j^{th} component is defined as

$$\ell(y_i; \theta_{ij}, \phi_j) = \frac{y_i \theta_{ij} - b(\theta_{ij})}{a_i(\phi_j)} + c(y_i, \phi_j). \tag{3.56}$$

For the j^{th} component GLMs, μ_{ij} is the mean of Y_i and the link function provides $\eta_j = g(\mu_{ij}) = x_i^T \beta_j$ as a linear predictor. Therefore, GLMs with a different linear covariate describes each weight. The semiparametric mixture distribution can be denoted with the mixture kernel:

$$\Psi \equiv \begin{bmatrix} \mu_1 \dots \mu_J \\ \pi_1 \dots \pi_J \end{bmatrix} \tag{3.57}$$

where the components $\pi_1, \dots, \pi_J$ are offered to parameters $\mu_1, \dots, \mu_J$ based on distributions from the exponential family. Nevertheless, $\mu_1 \dots \mu_J$ will change from scalar quantities to vectors, such as (Schlattmann, 2009):

$$\mu_1 = (\beta_{01}, \beta_{11}, \dots, \beta_{h1}), \tag{3.58}$$

where h represents the number of covariates in the model. Except for the β parameters, we may also imply covariates for the mixture weights π_j, if we assume the covariate vector as $\boldsymbol{x_i}$, then:

$$\pi_{ij} = \pi_j(\boldsymbol{x_i}; \alpha) = \frac{exp(w_j^\top \boldsymbol{x_i})}{1 + \sum_{m=1}^{J-1} exp(w_m^\top \boldsymbol{x_i})} \quad j = 1, \dots, J, \tag{3.59}$$

and

$$\alpha = (w_1^\top, \dots, w_{J-1}^\top)^\top, \tag{3.60}$$

provides the multinomial logit regression coefficients, note that $w_J = 0$. Therefore, the unknown vector of parameters is described as:

$$(\alpha^\top, \beta^\top)^\top. \tag{3.61}$$

The detailed method for including the covariates into the finite mixture model is introduced by Wang (Wang, 1994).

The complete log-likelihood for finite mixture GLMs is written as

$$logL_c(\Psi) = \sum_{i=1}^{n}\sum_{j=1}^{J} z_{ij} log\pi_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{J} z_{ij} logf(y_i, \theta_{ij}, \phi_j). \tag{3.62}$$

The development of the E-step is applied using the same method. The expected value of z_{ij} is equal to

$$\gamma(z_{ij}) = \frac{\pi_{ij} f(y_i; \theta_{ij}, \phi_j)}{\sum_{l=1}^{J} \pi_{il} f(y_i; \theta_{ij}, \phi_j)}. \tag{3.63}$$

The M-step of $(l+1)^{th}$ iteration is provided for α denoted by (Geoffrey & Peel, 2000):

$$\sum_{i=1}^{n}\sum_{j=1}^{J} \gamma(z_{ij}) \frac{\partial log\pi_{ij}}{\partial \alpha} = 0, \tag{3.64}$$

and the M-step of $(l+1)^{th}$ iteration is provided for β is

$$\sum_{i=1}^{n}\sum_{j=1}^{J} \gamma(z_{ij}) \frac{\partial log f(y_i, \theta_{ij}, \phi_j)}{\partial \beta} = 0. \tag{3.65}$$

Commonly, α and β are defined as a prior without any elements. The parameters from GBTM with the continuous and binary dataset can be estimated by the EM algorithm from the exponential family. However, this would not be tenable if the density of the component is defined as a zero-inflated Poisson (ZIF) model (Lambert, 1992). The EM algorithm with zero-inflated data in GBTM is explained in Roeder's paper (Roeder et al., 1999). Since this thesis does not focus on count outcomes, this issue will not be explored in deep.

3.3 Group-based trajectory modeling

3.3.1 Introduction

Group-based trajectory modeling is an extension of finite mixture models involving time or age variables in polynomial functions (Nagin & Tremblay, 2001). This model identifies subgroup trajectories as time-varying within the population using the correct mixture probability distributions (Nagin, 1999). Three types of datasets can be fitted with this model:

continuous outcomes following a normal distribution, binary outcomes following the logistic distribution, and count outcomes following the Poisson or zero-inflated Poisson distribution (Jones et al., 2001).

3.3.2 Unconditional group-based trajectory modeling

Let Y_i be discrete random variables for the i_{th} subject with measurements t, $t = 1, 2, \ldots, T$. Then, $Y_i = \{y_{i1}, y_{i2}, y_{i3} \ldots, y_{iT}\}$ are the repeat measurements from individual i over the measurement of T. $P(Y_i)$ represents the probability of Y_i. If there is a mixture of J groups of trajectories from the population, the unconditional group-based trajectory modeling could be written as (Nagin, 1999, 2005; Jones et al., 2001):

$$\begin{aligned} P(Y_i) &= f(Y_i) \\ &= \sum_{j=1}^{J} Pr(\boldsymbol{Y} = Y_i, C = j) \\ &= \sum_{j=1}^{J} Pr(C = j) Pr(\boldsymbol{Y} = Y_i | C = j) \\ &= \sum_{j=1}^{J} \pi_j f(Y_i, \Theta_j) \\ &= \sum_{j=1}^{J} \pi_j P^j(Y_i), \end{aligned} \tag{3.66}$$

where $j = 1, 2, 3, \ldots, J$ are the unobserved trajectory groups, C is latent class, $f(Y_i)$ is the marginal probability mass function of individual Y_i, and π_j stands for the probability of the i^{th} subject belonging to group j based on a multinomial logit function for $j = 1, 2, 3, \ldots, J$. Θ_j is a vector of parameters for the density of component j. The sum of π_j is equal to one. $P^j(Y_i)$ is the probability of Y_i given the i^{th} person in trajectory group j.

The overall likelihood function should be described as (Nagin, 2005):

$$L = \prod_{i=1}^{n} P(Y_i), \tag{3.67}$$

where n is the sample size.

3.3.3 Multinomial logit function for components

Multinomial logit models usually use fitting discrete outcome models for classification (So & Kuhfeld, 1995). The probability π_j from a multinomial logit function is:

$$\pi_j = P(C = j) = \frac{e^{\vartheta_j}}{\sum_{j=1}^{J} e^{\vartheta_j}}. \tag{3.68}$$

ϑ_j is a scalar, and ϑ_1 is normalized to zero. C is defined as the unobserved discrete variable indicating the latent class of the i^{th} individual (Nagin, 1999). From this multinomial logit model, the risk factors have the same meaning as the time-invariant covariates (TICs) from latent growth curved modeling (Roeder et al., 1999). If we include risk factors into the multinomial logit function, then the model can be expanded as:

$$\pi_j = P(C = j|\boldsymbol{X} = \boldsymbol{x}) = \frac{exp(\vartheta_j + w_j' x)}{\sum_{j=1}^{J} exp(\vartheta_j + w_j' x)} \tag{3.69}$$

where $\boldsymbol{X} = \{x_1, x_2, ..., x_r\}$ is a vector of covariates for risk factors and their interactions. w_j is a vector of parameters representing the coefficients of risk factor x; to define the reference group of risk factors, w_1 should be identified as zero. These risk factors are the characteristics of each individual from the baseline, and have the ability to vary group membership probabilities (Roeder et al., 1999).

3.3.4 Group-based trajectory modeling for continuous outcomes

y_{it} is the random variables for the i^{th} person at the measurement t. y_{it} should be independent: $P^j(Y_i) = \prod_{t=1}^{T} p^{jt}(y_{it})$, where T is the maximum number of measures $t = 1, 2, \ldots, T$. Let $p^{jt}(y_{it})$ be the probability density function of y_{it} given the i^{th} subject in the group j at time t, which is selected to conform to the type of outcome under analysis (Jones & Nagin, 2007). For continuous outcomes, we consider that all y_{it} follow normal distributions, so

$P^j(Y_i)$ is denoted as (Jones et al., 2001):

$$Pr(\boldsymbol{Y} = Y_i | C = j, V = v_i) = \prod_{y_{it}} \frac{1}{\sigma} \phi \left[\frac{(y_{it} - \mu_{itj})}{\sigma} \right]. \tag{3.70}$$

In equation (3.70), ϕ is the probability density function for standard normal distribution and is scaled by $\frac{1}{\sigma}$ to make sure the integral is still equal to 1. μ_{itj} and σ are parameters representing the mean and standard error, respectively, for the i^{th} subject in the group j at time t in normal distribution. $C = 1, 2, \ldots, J$ is defined as the unobserved latent variable. $V = v_{i1}, v_{i2}, \ldots, v_{iT}$ are the time-dependent variables. These are the same as time-variant covariates from the latent growth curve modeling and can influence within-subject effects shown in Section 2.1.3. Nagin (1999) used a truncated normal distribution bound to the random variable with a minimum and a maximum number, such as a psychometric scale.

For the corresponding link function of the normal distribution, μ_{itj}, defined as the mean of trajectory over age, will be:

$$\mu_{itj} = \beta_{0j} + (age_{it})\beta_{1j} + (age_{it})^2 \beta_{2j} + ... + v_{it}\delta_j + \epsilon_{it}. \tag{3.71}$$

μ_{itj} is the mean of the i^{th} subject in the group j at time t. In most research, trajectories are defined by age. However, sometimes, age will be replaced by elapsed time (Nagin, 1999). For example, in clinical trials, the age will be changed to days, months, or years. ϵ is the error with the normal distribution assumption with mean zero and constant standard deviation of σ. $\beta's$ are the parameters of age or time. $\delta_j = (\delta_1, \delta_2, \ldots, \delta_J)$ is a vector of parameters representing the coefficients of time-dependent variables v_{it}. $\beta's$ and δ' can control the shapes of the polynomial function.

3.3.5 Group-based trajectory modeling for binary outcomes

Binary outcome data is also quite often derived from longitudinal studies. For example, we may be interested in if each individual has depression or not. Thus, the outcome of interest will be yes or no. In this case, y_{it} will be assumed to be a binary outcome, $P^j(Y_i)$ follows Bernoulli distribution (Jones et al., 2001):

$$Pr(\boldsymbol{Y} = Y_i|C = j, V = v_i) = \prod_{y_{it}=1} \rho_{itj} \prod_{y_{it}=0} (1 - \rho_{itj}) \tag{3.72}$$

where ρ_{itj} denotes the probability when $y_{it} = 1$ provided i^{th} subject in group j at time t, ρ_{itj} follows the probit function (Ziegel, 2004). The link function for ρ_{itj} can be written as:

$$\rho_{itj} = \frac{exp\,(\beta_{0j} + (age_{it})\beta_{1j} + (age_{it})^2\beta_{2j} + ... + v_{it}\delta_j)}{1 + exp\,(\beta_{0j} + (age_{it})\beta_{1j} + (age_{it})^2\beta_{2j} + ... + v_{it}\delta_j)} \tag{3.73}$$

3.3.6 Group-based trajectory modeling for count outcomes

Count data is another kind of outcome often used in epidemiology studies, such as the number of questions on which patients report feeling satisfied with the treatment on satisfaction questionnaires. Usually, $P^j(Y_i)$ should be defined by the Poisson distribution:

$$Pr(\boldsymbol{Y} = Y_i|C = j, V = v_i) = \frac{exp(-\lambda_{itj}) * -\lambda_{itj}^{y_{it}}}{y_{it}!} \tag{3.74}$$

From equation (3.74), λ_{itj} is a parameter measuring the mean rate of events that occurred for the i^{th} subject in the group j at time t. As λ_{itj} increases, the Poisson model approaches the shape of normal distribution. When λ_{itj} is large enough, the analysis results based on Poisson and normal distribution will be quite similar. $y_{it}!$ is the factorial function defined as $y_{it}! = y_{i1}y_{i2}\dots y_{it}$.

The Poisson distribution can be adapted to deal with most cases of count data. However, sometimes, using the Poisson distribution for count data with many zeros will underestimate the probability of the zero part. A zero-inflated Poisson distribution (ZIP) is one of the methods that can solve this kind of problem. Thus, for these cases, $P^j(Y_i)$ should be denoted by a ZIP distribution (Jones et al., 2001):

$$Pr(\boldsymbol{Y} = Y_i | C = j, V = v_i) = \prod_{y_{it}=0} [\rho_{itj} + (1 - \rho_{itj})\, exp(-\lambda_{itj})] \prod_{y_{it}>0} \frac{exp(-\lambda_{itj}) * -\lambda_{itj}^{y_{it}}}{y_{it}!} \quad (3.75)$$

where ρ_{itj} represents the probability when the outcome count is zero and $(1 - \rho_{itj})$ stands for the probability when the outcome count is larger than zero. λ_{itj} is the parameter with the same meaning as the Poisson distribution.

A link function for connecting the trajectory with time relying on both equation (3.74) and (3.75) are provided as:

$$ln(\lambda_{itj}) = \beta_{0j} + (age_{it})\beta_{1j} + (age_{it})^2\beta_{2j} + ... + v_{it}\delta_j \quad (3.76)$$

The reason behind using $ln(\lambda_{itj})$ instead of λ_{itj} is that maximum likelihood estimates of β may be negative values. In such cases, the estimation process will fail.

As a general method to estimate the covariance matrix of parameter estimates, the sandwich estimator, also named the robust covariance matrix estimator, could keep the covariance matrix estimates asymptotically and consistently. Based on the sandwich estimator, there was no requirement for the assumption of the covariance matrix structure, and even the assumed covariance structure was wrong (Carroll, Wang, Simpson, Stromberg, & Ruppert, 1998). Thus, the structure of the covariance matrix of the parameter estimates in GBTM was not a concern for us because the sandwich estimator was used to weight the likelihood function in the group-based trajectory modeling methods (Jones & Nagin, 2007, 2011). Figure 3.1 displays the overall structure of GBTM. From Figure 3.1, we see that the observed trajectory depends on group membership and on the time-dependent covariates. Group membership also depends on the time-invariant covariates (Jones et al., 2001).

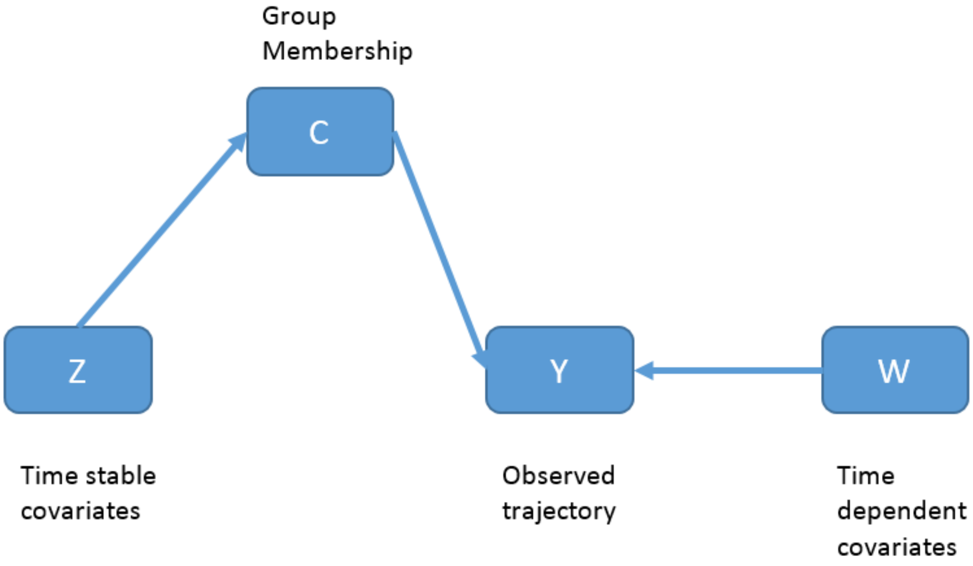


Figure 3.1: Group-based trajectory modeling framework

3.4 Maximizing likelihood for GBTM

The EM algorithm is a suitable method to use to maximize the likelihood of GBTM because GBTM is extended from FMM. In Section 3.2.10, we derived the EM algorithm to find the maximum likelihood estimator for the generalized mixture models. Since most of the GBTMs belong to generalized mixture models, this method can be used to find the maximization from them. The only particular case is the GBTM with the ZIP model, which was not involved in the exponential family. However, in Roeder's paper, using the EM algorithm to maximize this model's likelihood was introduced with detailed information (Roeder et al., 1999).

The Quasi-Newton method is considered an alternative to the Newton method for identifying functions' local maxima. Instead of calculating the Hessian matrix directly as in the Newton method, the Quasi-Newton method uses the successfully analyzed gradient vectors to update the Hessian matrix (Gower & Richtárik, 2017). The general procedure for the Quasi-Newton method is as follows (Cericola, 2015):

1. Select the starting points of the parameters;

2. Use gradient to approximate the parameters' inverse Hessian matrix;
3. Calculate changes to the parameters;
4. Determine the new parameters;
5. Determine if parameters converged;
6. If converged stop, otherwise, repeat from step 2.

Victor (2014) implemented a simulation study to compare the EM algorithm and Quasi-Newton procedure to maximize the GBTM. Compared to the EM algorithm, the Quasi-Newton procedure has a higher demand for the parameters' starting values. On the other hand, the Quasi-Newton procedure requires fewer iterations than the EM algorithm to get the values of the maximum likelihood (Victor, 2014). The Proc Traj package running in SAS 9.4 for identifying group-based trajectory models applies Quasi-Newton procedure to maximize the estimators (Jones & Nagin, 2007). Since we applied the Proc Traj package directly in the application and simulation in this thesis, the Quasi-Newton procedure will be used for maximizing the likelihood. Therefore, defining correct initial starting values is a crucial step.

Note that the maximum likelihood estimations provided parameter estimates that are asymptotically unbiased under the assumption of "data are missing at random (MAR)". When data are MAR, information from the dataset can be used to impute missing data prior to input into the trajectory model (Nagin & Odgers, 2010). SAS programming 9.4 and Proc Traj package was used to fit trajectory modeling, which employs an imputation technique to assign values for missing data.

3.5 Group-based trajectory model selection

There are usually two parts to select group-based trajectory models: choosing the right number of groups and determining the correct order of the polynomial equation to describe the proper shape of the trajectories (Nagin, 1999). The χ^2 goodness of fit test is one of the most popular and widely used methods for model selection in a longitudinal study (Erdfelder, 1990). However, this method cannot be used in the finite mixture model with J number of components (Ghosh & Sen, 1984). Therefore, we used another criterion for a model selection called the Bayesian Information Criterion (BIC) (Raftery, 1995). The model with maximum BIC will often be selected when prior information for the right model is limited (Kass & Raftery, 1995). This method can be applied to extensive statistics model selections that involve group-based trajectory modeling with a fixed number of component groups (Nagin, 1999). The way to select the model based on BIC is to find the model with the largest BIC value. The formula to calculate the BIC score from a provided model is written as:

$$BIC = log(L) - 0.5klog(N), \tag{3.77}$$

where L is defined as the maximized likelihood of the model and k denotes the number of parameters in the model. N is the sample size, which is slightly different since our data is longitudinal. Thus, in longitudinal data, N should be the number of individuals times the number of measurements (D'Unger, Land, & Nagin, 1998). Model selection requires selecting the best models from all possible models. However, it is impossible to try all models, so it is necessary to reduce the scope of the models by the sample size N to determine the largest number of trajectories J that can be considered. After this, model selection has two stages. The first stage is to make a decision about how many groups of the model will be selected, which means to screen the number of groups from one to a preset maximum. We must define the preset rule for the order of polynomials for each group's trajectory (Keribin, 2000). For example, assuming all trajectories are linear, we will find the groups of the model with the largest BIC value. The second stage is to find out the preferred order of polynomials from each trajectory based on the number of J groups from the first stage. Selecting the right order means not only relying on BIC scores and significance level α, but also on the mechanism of each trajectory's subgroup population. The reason we consider the group number of trajectories first is that selecting the group number is more critical compared to the order of trajectories (D'Unger et al., 1998).

To interpret BIC, we usually consider using the Bayes factors (B_{ij}), which means the different odds ratio for the probability of the model i is the right model compared to the probability of model j being the right model, $i < j$. The Bayes factors are evaluated by Jeffreys's scale, as shown in Table 3.1 (Wasserman, 2000).

Table 3.1: Jeffreys's scale of evidence for Bayes factors

Bayes factor	**Interpretation**
$B_{ij} < \frac{1}{10}$	Strong evidence with model j
$\frac{1}{10} < B_{ij} < \frac{1}{3}$	Moderate evidence with model j
$\frac{1}{3} < B_{ij} < 1$	Weak evidence with model j
$1 < B_{ij} < 3$	Weak evidence with model i
$3 < B_{ij} < 10$	Moderate evidence with model i
$B_{ij} > 10$	Strong evidence with model i

However, the Bayes factor is difficult and sometimes not possible to calculate (Schwarz et al., 1978). Thus, $e^{BIC_i - BIC_j}$ is an excellent method to approximate the Bayes factor $B_{ij} \approx e^{BIC_i - BIC_j}$, where i is the lower group number, and j is the upper group number (Kass & Wasserman, 1995). If the value is smaller than one, the j number group model is preferred. On the other side, if the value is larger than one, the i number group model is preferred. An alternative approach to interpreting BIC is computing the probability that a model with j groups is the right model from a number of J other models defined as p_j. This could be approximated by

$$p_j = \frac{e^{BIC_j - BIC_{max}}}{\sum_{j=1}^{J} e^{BIC_j - BIC_{max}}}, \tag{3.78}$$

where BIC_{max} is the largest BIC score from J models, and the model with the largest p_j will be the correct model (Kass & Wasserman, 1995). The Akaike information criterion (AIC) selection method can also be applied, but it is quite similar to the BIC model. The only difference between these two methods is that sample size changes will not influence the AIC method. For exceptional cases, the BIC value will keep increasing when we keep adding trajectory groups. To solve this problem, we only need the model with enough groups to reach the distinct features of the data (Nagin, 2005).

3.6 Multivariate group-based trajectory modeling

3.6.1 Group-based dual trajectory models

Group-based dual trajectory modeling (GBDTM) is extended from group-based trajectory modeling (GBTM) and includes two distinct, but correlated outcomes (Nagin & Tremblay, 2001). GBDTM provides the probability link functions that will link two related outcomes together. Compared with single outcome GBTM, this model offers a way to handle two prominent outcomes simultaneously. The constrained model and general model are two conceptual models used to link two correlated trajectory outcomes in GBDTM (Nagin & Tremblay, 2001). In the constrained model, the number of trajectory groups is assumed to be the same J to combine the outcomes Y_i^1 and Y_i^2. $i = 1, 2, \ldots, n$, Y_i^1 is the first outcome, and Y_i^2 is the second outcome. Y_i^1 and Y_i^2 should be independently distributed (Nagin, 2005). Therefore, the joint probability for the constrained model weighted by π_j is given by:

$$P(Y_i^1, Y_i^2) = \sum_{j=1}^{J} \pi_j p^j(Y_i^1) * p^j(Y_i^2), \tag{3.79}$$

where $j = 1, 2, \ldots, J$ is the number of trajectory groups of both Y_i^1 and Y_i^2. π_j represent the shared proportion of both Y_i^1 and Y_i^2.

The other model is the general model, which has different numbers of groups for the two outcomes. Here we assume there are J group of trajectories for Y_i^1 with probability link to L group of trajectories for Y_i^2 (Nagin & Tremblay, 2001). Therefore, the likelihood function weighted by π_{jl} is updated to

$$P(Y_i^1, Y_i^2) = \sum_{j=1}^{J}\sum_{l=1}^{L} \pi_{jl} p^j(Y_i^1) * p^l(Y_i^2) = \sum_{j=1}^{J}\sum_{l=1}^{L} \pi_j \pi_{l|j} p^j(Y_i^1) * p^l(Y_i^2)$$
$$= \sum_{j=1}^{J} \pi_j p^j(Y_i^1) \sum_{l=1}^{L} \pi_{l|j} p^l(Y_i^2) \tag{3.80}$$

where $j = 1, 2, \ldots, J$ is the number of trajectory groups of Y_i^1 and $l = 1, 2, \ldots, L$ is the number of trajectory groups of Y_i^2. π_{jl} is the joint probability for both outcomes Y_i^1 and Y_i^2. $\pi_{l|j}$ is the conditional probability to link group j of Y_i^1 to group of l of Y_i^2.

In GBDTM, having risk factors in the model will only influence the proportions of the first outcome π_j, but not the conditional probability $\pi_{l|j}$ (Nagin, 2005). In this way, the effects of the risk factors are able to be calculated based on the same formula from equation (3.69).

3.6.2 Group-based multi-trajectory models

We may also be interested in two or more outcomes, called multiple correlated outcomes. For example, a study will contain multiple biomarkers from a disease or multiple mental health disorders in order to generalize the overall population's mental health situation. In these cases, if we still consider using the general GBDTM, the work will be complicated because every two outcomes need to combine and build a GBDTM (Jones & Nagin, 2007). Group-based multi-trajectory modeling (GBMTM) was developed to discover the latent clusters of individuals who follow similar trajectories based on multiple outcomes of interest (Nagin et al., 2018). GBMTM is a new method that can be used to describe the inter-relationship of multiple outcomes (Nagin et al., 2018). This model is an extension of the constrained dual trajectory model (see Section 3.6.1). It includes more than two outcomes but is weighted by the same probability π_j. Therefore, this model requires a high similarity of group memberships for the individuals with each outcome. In the GBMTM, let Y_i^k denote i^{th} individual with the k^{th} outcome, $k = 1, 2, \ldots, K$. As in dual trajectory modeling, Y_i^k are independently distributed with $P^j(Y_i^1, Y_i^2, \ldots, Y_i^K) = P_1^j(Y_i^1)P_2^j(Y_i^2), \ldots, P_K^j(Y_i^K)$, where $P^j(Y_i^k) = \prod_{t=1}^{T^{(k)}}(y_{it}^k)$. Therefore, the likelihood for group j is developed by (Nagin et al., 2018):

$$P(Y_i^1, Y_i^2, \ldots, Y_i^K) = \sum_{j=1}^{J} \pi_j P^j(Y_i^1, Y_i^2, \ldots, Y_i^K) = \sum_{j=1}^{J} \pi_j \prod_{k=1}^{K} P^j(Y_i^k) \tag{3.81}$$

$T^{(k)}$ means the k^{th} outcome with the T^{th} measurement. Nagin mentioned that multi-outcomes must have the same number of groups of trajectories j with the same probabilities in GBMTM (Nagin et al., 2018).

3.7 Summary of trajectory models

Table 3.2 shows summary of the of each trajectory modeling and their property differences.

Table 3.2: Comparison of group-based trajectory models

Model	**Group-based trajectory modeling**	**Group-based dual modeling**	**Group-based multi-trajectory modeling**
Outcomes	Single	Two	Two or more
Statistical Method	Conditional likelihood distribution weighted by single probability	Conditional likelihood distribution weighted by joint probability	Conditional likelihood distribution weighted by single probability
Difference	Number of trajectories are independent	Number of trajectories will be different for different outcomes	Number of trajectories will be remain the same for different outcomes
Group membership	Independent	Remains different	Remains the same
Trajectory shape	Independent	Different for different outcomes	Similar for different outcomes
Example	Depression only or anxiety only	Depression and anxiety, 2 outcomes together to create trajectories with different group membership for each outcome	Depression and anxiety, use 2 outcomes to create trajectories with common group membership for each outcome

Chapter 4 APPLICATION FOR DEPRESSION AND ANXIETY

4.1 Introduction

In this Chapter, the three trajectory models discussed in Chapter 3 were applied using a real dataset on depression and anxiety outcomes. The data set was described in Section 4.2. Two group-based trajectory modeling (GBTM) were independently fitted by depression and anxiety outcomes separately, described in Section 4.3. Group-based dual trajectory modeling (GBDTM) and group-based multi-trajectory modeling (GBMTM) were implemented with joint depression and anxiety outcomes, as discussed in Sections 4.4 and 4.5.

All the analyses were performed using SAS 9.4. Trajectory models were developed with PROC TRAJ, a package running under SAS 9.4 (Jones, 2020). Figures were redesigned based on the SAS outputs with Excel. For this thesis, $\alpha = 0.05$ was set as the significance level. This study data set of the analysis was approved by the Behavioural Research Ethics Board, University of Saskatchewan (ID: 1759).

4.2 Data structure and study population

This study utilized a subset of an eight-year longitudinal survey called the Korea Health Panel Study (KHPS). The KHPS collected by the Korean Institute for Health and Social Affairs, in conjunction with the National Health Insurance Service, used a stratified sampling frame taken from the Korean Population and Housing Census in 2000 (KHPS, 2020). Based on this dataset, sample weights for the KHPS were calculated after going through the process of adjusting for unequal selection probabilities and non-responses and making a population distribution disclosure via post-stratification corresponding to the sample distribution (Lim, Cheng, Kabir, & Thorpe, 2020). KHPS aims to improve the national health system's responsiveness and accessibility and provide necessary information regarding the efficiency of policy implementation by identifying factors that directly or indirectly affect healthcare services, spend on financial resources, and continuously observe the trends (KHPS, 2020). The KHPS began in 2008 and incorporated a total of 24616 participants from 7387 households. In 2014, KHPS was expanded to mitigate attrition with the additional 2520 families. Using computer-assisted personal interviews, trained staff collected data with three aspects: household, individual, and case-based sections. Comprehensive assessments on the use of healthcare services, healthcare costs, and other potentially influential factors have been conducted annually since 2008. The survey's core questions involved 13 essential sectors and 10 other sectors, including household items data, household member items data, health insurance data, chronic disease data, drug use data, long-term care data for adult household members, and emergency medical use data. For medical data, the annual data disease (diagnosis) code and the Korean standard disease classifications were used.

For data collection, investigators visited the target households and used a computer (CAPI) to investigate. Baseline covariates were measured in 2008. They involved sex, age, education, marital status, residential area, number of household members, household composition type, housing type, current chronic disease status, health insurance type, household income quantile, and household expenses. Age was categorized as 65 - 69, 70 - 74, 75 - 79, and 80 years and older. Sex was coded with 0 = male and 1 = female. Education was coded as 0 = no education, 1 = Grade 1 - 6, and 2 = Grade 7 or higher. Residential area was categorized into two areas: metro-city was coded as 0, and not-metro-city was coded as 1. Household composition type was categorized as 1 = living alone, 2 = living with a spouse, and 3 = other mixed arrangements. Housing type was categorized as 1 = detached house, 2 = apartment, and 3 = other types.

Exercise and walking were evaluated based on responses to the following question: "During the past week, how many days did you do intensive/moderate physical activity, or walk more than 10 minutes a day". Responses were evaluated on an 8-point Likert scale ranging from 0 to 7 (none = 0, once a week = 1, two days a week = 2, three days a week = 3, four days a week = 4, five days a week = 5, six days a week = 6, seven days a week = 7). Alcohol consumption was also scored on an 8-point Likert scale in response to the question, “Over the past year, how often did you drink?” (never = 0; recently non-drink = 1, less than once per month = 2, once per month = 3, 2 - 3 times per month = 4; once per week = 4; 2 - 3 times a week = 6; almost daily = 7). In our study, exercise and walking variables were categorized as “none”, “less or equal to 3 days/week”, and “more than 3 days/week”. Drinking variable was categorized as “none”, “less than twice/week”, “2 - 4 times/week” and “almost daily”. The main outcomes for depression and anxiety were identified in the medical data by the disease diagnosis code: 0 = no depression/anxiety and 1 = depression/anxiety. The main dichotomous binary outcomes of depression and anxiety in each year were collected from medical expenses, including prescription drug receipts or medical institutions/pharmacies, potentially leading to inadequate recognition of our sample outcomes. Diagnostic criteria for depression and anxiety disorder was based on DSM-5 (The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition) (Lim et al., 2020).

For this thesis, baseline responses from individuals aged 65 or older in the initial 2008 households and in the additional 2014 households were examined, as were their responses for each subsequent wave, if depression or anxiety answers were provided. A total of 3983 individuals met our study criteria. Demographic and other data were extracted at each time point over eight years from 2008 to 2015. Figure 4.1 described the structure of depression and anxiety. The dataset had two parts. The first part was the original 2946 individuals aged 65 or older from 2008, for whom all the measurements over eight years were available. The retention rates were 96.7%, 90.1%, 85.7%, 81.2%, 76.4%, 72.1%, 67.6% and 62.8% from 2008 to 2015, respectively. The second part involved the additional 1137 individuals from 2014 to 2015. These were moved to the baseline (the year 2008) and second measurement (the year 2009), and considered the rest as missing measurements. The retention rates of the additional participants were 99.7% in 2014 and 85.5% in 2015. A total of 1785 (44.8%) individuals had complete data measurements for depression and anxiety.

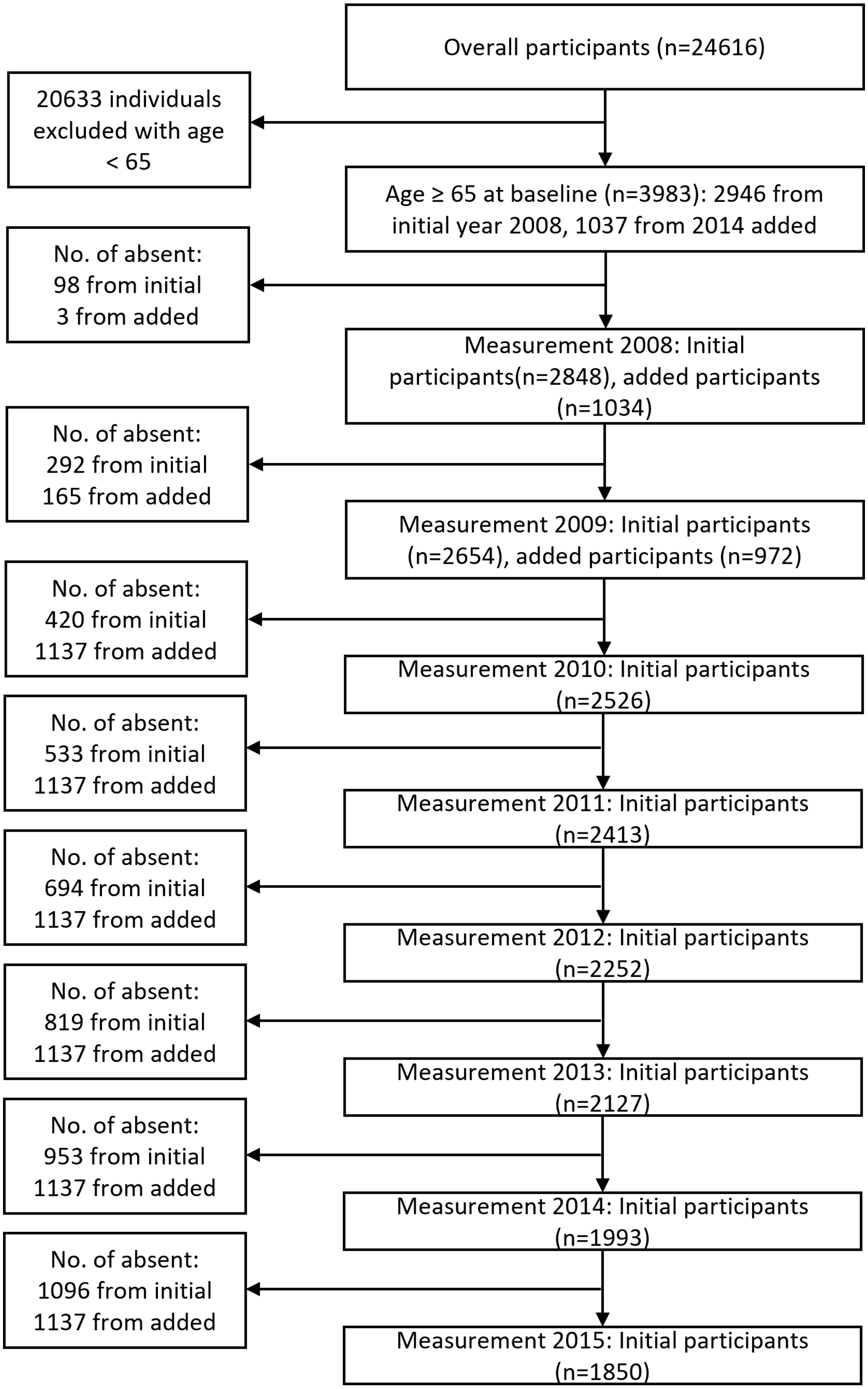

Overall participants (n=24616)
20633 individuals excluded with age < 65
Age ≥ 65 at baseline (n=3983): 2946 from initial year 2008, 1037 from 2014 added
No. of absent: 98 from initial 3 from added
Measurement 2008: Initial participants(n=2848), added participants (n=1034)
No. of absent: 292 from initial 165 from added
Measurement 2009: Initial participants (n=2654), added participants (n=972)
No. of absent: 420 from initial 1137 from added
Measurement 2010: Initial participants (n=2526)
No. of absent: 533 from initial 1137 from added
Measurement 2011: Initial participants (n=2413)
No. of absent: 694 from initial 1137 from added
Measurement 2012: Initial participants (n=2252)
No. of absent: 819 from initial 1137 from added
Measurement 2013: Initial participants (n=2127)
No. of absent: 953 from initial 1137 from added
Measurement 2014: Initial participants (n=1993)
No. of absent: 1096 from initial 1137 from added
Measurement 2015: Initial participants (n=1850)

Figure 4.1: Study flow diagram

]

Table 4.1 provided the baseline characteristics for 3983 participants aged 65 or older in the study between 2008 and 2015. 57% of the participants were female, and the average age of their baseline measurement was 72.4 years (SD ± 6). Of these participants, 62.9% had never received any education or had only finished elementary school; 65.1% lived with their spouse, while 1.6% lived alone; the majority (83%) reported their income level was lower than the median income level; and 36.7% were still attending income-generating activities. Only 38.2% lived in metro-cities, 57.4% lived in a detached house, and 23.6% rented their home. 27.5% of the participants currently smoked, and 15.4% drank two times or more per week. 31.4% could not walk more than three days per week, and only 33% engaged in physical activities. Moreover, most of them (88%) had more than three chronic diseases, and 19.9% were suffering from physical or mental disabilities. From the baseline outcomes of depression and anxiety, 107 (2.8%) of participants were diagnosed with depression, and 73 (1.9 %) had anxiety.

Table 4.1: Baseline characteristics of participants (N=3983)

Variable name	Number (%)
Sex	
Male	1714 (43)
Female	2269 (57)
Age (Continuous)	
Mean ± SD	72.4 ± 6.0
Median (IQR)	71 (68 - 76)
Age	
65-69	1534 (38.5)
70-74	1215 (30.5)
75-79	740 (18.6)
≥ 80	494 (12.4)
Marriage status	
Married	2592 (65.1)
Single/divorce/widower	1390 (34.9)
Missing	1
Education	
None	799 (20.1)
Elementary	1705 (42.8)
Middle/High	1181 (29.7)
University	298 (7.5)

Variable name	Number (%)
Smoking	
No	2218 (59.9)
Previous	467 (12.6)
Current	1017 (27.5)
Missing	281
Drinking	
No	1495 (40.4)
< 2 days/week	1640 (44.3)
2-4 days/week	278 (7.5)
Almost daily	291 (7.9)
Missing	279
Residential area	
No Metro-city	2463 (61.8)
Metro-city	1520 (38.2)
Housing	
Detached House	2286 (57.4)
Apartment	556 (14.0)
Others	1141 (28.6)
Home ownership	
Own	3042 (76.4)
Lease	941 (23.6)
Living	
Alone	62 (1.6)
Couple only	2466 (61.9)
Others	1455 (36.5)
Disability	
No	3191 (80.1)
Yes	792 (19.9)
Walking	
None	690 (18.6)
≤ 3days/week	473 (12.8)
>3 days/week	2541 (68.6)
Missing	279
Medium/Intensive Physical activity	
none	2482 (67.0)
≤ 3days/week	357 (9.6)
>3 days/week	865 (23.4)
Missing	279
More than 3 chronic diseases	
Yes	3507 (88)
No	476 (12)
Economic Activity	
Yes	1461 (36.7)
No	2522 (63.3)

Variable name	Number (%)
Income quantile	
< 20	1309 (44.8)
20 - 40	666 (22.8)
40 - 60	471 (16.1)
60 - 80	260 (8.9)
80 - 100	216 (7.4)
Baseline depression diagnosed	
Yes	107 (2.8)
No	3778 (97.2)
Missing	98
Baseline anxiety diagnosed	
Yes	73 (1.9)
No	3812 (98.1)
Missing	98

The diagnosed proportions of depression and anxiety in each year were presented in Figure 4.2. More participants were diagnosed with depression than with anxiety. The participants with depression increased until 2014, after which fewer proportion of individuals were diagnosed with depression. The anxiety patients had a higher rate in the first two years, and then decreased in the third year. However, after 2010, the proportion of anxiety increased every year.

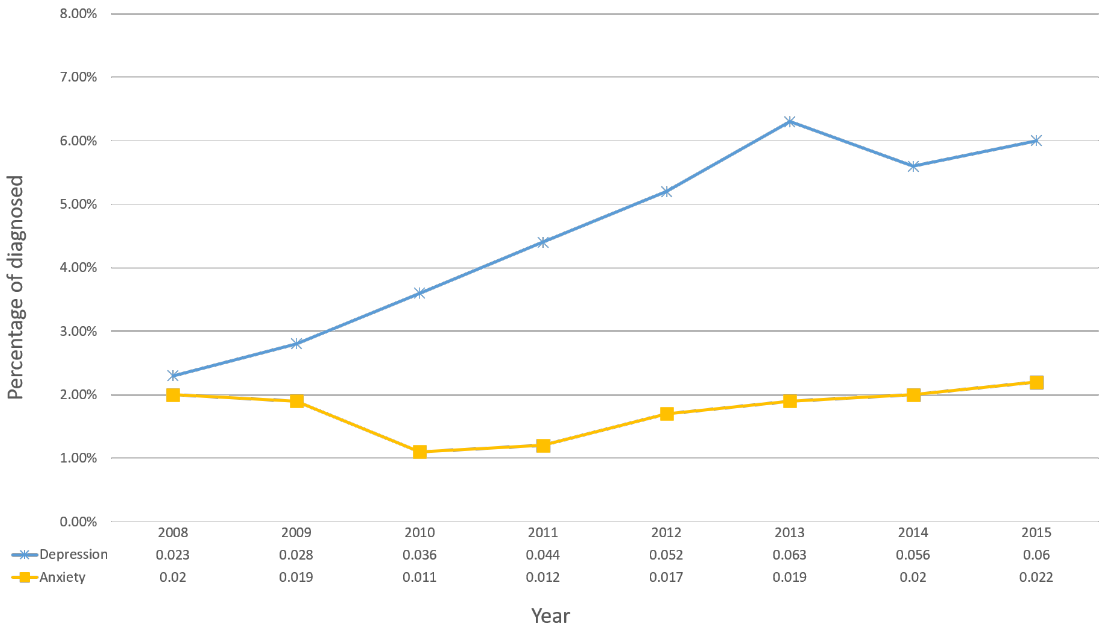


Figure 4.2: Proportions of diagnosed depression and anxiety during the study period

4.3 Selection of trajectory groups

Before building the trajectory models, we need to determine the best fitness number of trajectories for depression and anxiety outcomes. From the literature, depression and anxiety trajectories usually followed linear or quadratic shapes from polynomial functions (Hybels et al., 2016; Chui, Gerstorf, Hoppmann, & Luszcz, 2015; Hsu, 2012; Andreescu, Chang, Mulsant, & Ganguli, 2008; Holmes et al., 2018; Rzewuska et al., 2015; Spinhoven et al., 2017; Wiesner & Kim, 2006). Particularly in binary outcomes, trajectories are generated as linear most of the time (Huang et al., 2013).

I first tried GBTM with no starting points, with the number of trajectories from two to five. All trajectories from the models were assumed to be linear. The goodness-of-fit tests to select the right number of trajectories for depression and anxiety were described in Table 4.2. The number of groups with large BIC, AIC and posterior probability close to 1.0 will be better fit (Nagin, 2005). For depression, four-trajectory model had the largest BIC = -2363.7, AIC= -2354.7 and highest posterior probability 0.83. Similarly, a four-trajectory model with the largest BIC= -1476.3, AIC= -1467.2, and the posterior probability close to 1 were found with anxiety. Therefore, four groups of trajectories were assumed to be the best fit for both depression and anxiety.

Table 4.2: Goodness of model fit to select the optimal number of trajectory group for depression and anxiety

Number of trajectories	**Depression**			**Anxiety**		
	BIC	**AIC**	**PP**	**BIC**	**AIC**	**PP**
2	-2534.8	-2530.7	0	-1525.8	-1521.7	0
3	-2403.8	-2397.2	0	-1484.2	-1477.6	0
4	-2363.7	-2354.7	0.83	-1476.3	-1467.2	1
5	-2365.3	-2353.8	0.17	-1489.8	-1478.3	0

BIC = Bayesian information criterion, AIC =Akaike information criterion, PP = Posterior probability

4.4 Analysis of group-based trajectory modeling (GBTM)

4.4.1 GBTM for depression

4.4.1.1 Development of GBTM

Based on the GBTM for binary outcomes from Section 3.3.5, GBTM for a depression outcome was developed with four trajectory groups: “low-flat” (TD1), “low-to-middle” (TD2), “low-to-high” (TD3) and “high-curve” (TD4). One flat trajectory, two linear trajectories, and one quadratic trajectory are presented in Figure 4.3. The solid lines represent each trajectory group means, and the dashed lines showed predictions.

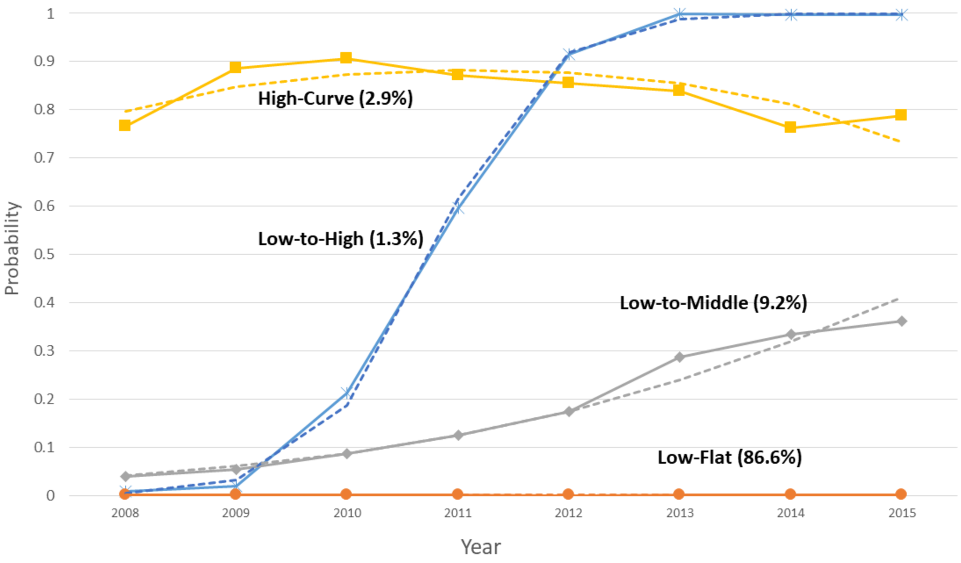


Figure 4.3: Depression trajectories for GBTM. The solid line indicates observed depression; the dashed line indicates predicted depression.

The first trajectory, TD1 (n=3636; 86.6%), was low-flat, showing the probability was close to zero. It indicated that most participants were not diagnosed with depression over time. The second trajectory, TD2 (n=214; 9.2%), was low-to-middle, meaning the probability of depression started low but increased slowly over time. The third trajectory, TD3 (n=31; 1.3%), started with the low depression but rapidly increased over time. In the last two years, depression probability was close to one for this TD3 group. The fourth trajectory, TD4 (n=102; 2.9%), was high-curve; remaining a high level of depression probability throughout the study period. Overall, TD1 contained the majority of participants (86.6%). TD3 had the lowest proportion (1.3%).

As we discussed in Section 3.3.5, each trajectory group had a probability that followed the logistic regression with time variables. The logit function was defined as:

$$logit(\rho_{itj}) = log(\frac{\rho_{itj}}{1-\rho_{itj}}) = \beta_{0j} + time_{it}\beta_{1j} + time_{it}^2\beta_{2j} + ... + \epsilon_{it}, \tag{4.1}$$

where ρ_{itj} was the probability of y_{it} belonging to group j equal to 1. The parameters for trajectory shapes and group memberships were identified in Table 4.3. Table 4.3 showed that the p-value was significant (p-value $<$ 0.05) with logistic polynomial regression in the TD1 - TD3. In TD4, the intercept and linear predictor were not significant (p-value = 0.1077 and 0.0502). However, the quadratic predictor was significant (p-value = 0.0262).

Table 4.3: Parameter estimates for trajectory shapes in depression GBTM

Group	Parameter	Estimate	Standard Error	p-value
Low-flat (TD1)	Intercept	-6.35359	0.56599	<0.0001
Low-to-middle (TD2)	Intercept	-3.53362	0.27639	<0.0001
	Linear	0.39656	0.05537	<0.0001
Low-to-high (TD3)	Intercept	-7.32440	1.89800	0.0001
	Linear	1.95051	0.52360	0.0002
High-curve (TD4)	Intercept	0.88200	0.54823	0.1077
	Linear	0.55127	0.28149	0.0502
	Quadratic	-0.06694	0.03011	0.0262

Proportions were generated from the multinomial function based on equation (3.68). All the p-values from the group memberships were significant (Table 4.4). Based on the parameters of polynomial functions, we could predict four depression trajectories with logit functions.

Table 4.4: Estimates of group membership proportions in depression GBTM

Group membership proportions			
Group	Estimate Proportion (%)	Standard Error	p-value*
Low-flat (TD1)	86.62274	1.42353	<0.0001
Low-middle (TD2)	9.16961	1.33216	<0.0001
Low-high (TD3)	1.32148	0.30925	<0.0001
High-curve (TD4)	2.88617	0.33508	<0.0001

*H_0: Proportion = 0 vs $H_{a:}$ Proportion $\neq$ 0

4.4.1.2 Characteristics of trajectory groups

Baseline characteristics across the four trajectory groups were described and compared. The data showed that there were no differences among the four depression trajectory groups in levels of education, residential area, housing type, marriage status, living alone or not, physical activity or walking, disability, and income-quantiles.

On the other hand, sex (p-value $<$ 0.0001), age levels (p-value = 0.004), smoking (p-value = 0.001), alcohol consumption (p-value = 0.001), homeownership (p-value = 0.045), more than three chronic diseases (p-value $<$ 0.0001) and current economic activities (p-value = 0.001) were significantly different among the groups from the overall chi-square test. Compared to other trajectory groups, the “low-flat” depression trajectory group (TD1) had the lowest proportion of females, non-smokers, non-drinkers, not mentally or physically disabled, involved in income-generating activities, with more than three chronic diseases. This trajectory also included the highest percentage of individuals who are 80 years old or older. On the other hand, the “high-curve” depression trajectory group (TD4) was found to contain the highest proportion of females aged 65 - 69, non-smokers, non-drinkers, living in a rental home, and with more than three chronic diseases. More detailed information could be found in Table 4.5.

Table 4.5: Distribution of baseline characteristics by GBTM depression trajectory groups (N, %)

	Trajectory groups				
	Low-flat (TD1) N=3636	Low-middle (TD2) N=214	Low-high (TD3) N=31	High-curve (TD4) N=102	p-value
Sex					
Male	1620 (44.6)	60 (28.0)	8 (25.8)	26 (25.5)	<0.0001
Female	2016 (55.5)	154 (72.0)	23 (74.2)	76 (74.5)	
Age					
65-69	1400 (38.5)	93 (43.5)	11 (35.5)	30 (29.4)	0.004
70-74	1098 (30.2)	74 (34.6)	10 (32.3)	33 (32.4)	
75-79	666 (18.3)	39 (18.2)	8 (25.8)	27 (26.5)	
≥ 80	472 (13.0)	8 (3.7)	2 (6.5)	12 (11.8)	
Marriage status					
Married	2363 (65.0)	143 (66.8)	20 (64.5)	66 (64.7)	0.962
Single/divorce/widower	1272 (35.0)	71 (33.2)	11 (35.5)	36 (35.3)	
Education					
None	727 (20)	40 (18.7)	6 (19.4)	26 (25.5)	0.275
Elementary	1547 (42.6)	105 (49.1)	13 (41.9)	40 (39.2)	
Middle/High	1078 (29.7)	60 (28.0)	10 (32.3)	33 (32.4)	
University	284 (7.8)	9 (4.2)	2 (6.5)	3 (2.9)	
Smoking					
No	1975 (58.7)	152 (71.0)	21 (67.7)	70 (74.5)	0.001
Previous	441 (13.1)	15 (7.0)	2 (6.5)	9 (9.6)	
Current	947 (28.2)	47 (22.0)	8 (28.8)	15 (16)	
Drinking					
No	1322 (39.3)	103 (48.1)	17 (54.8)	53 (56.4)	0.001
< 2 days/week	1502 (44.6)	90 (42.1)	12 (38.7)	36 (38.3)	
2-4 days/week	267 (7.9)	10 (4.7)	0 (0)	1 (1.1)	
Almost daily	274 (8.1)	11 (5.1)	2 (6.5)	4 (4.3)	
Residential area					
No Metro-city	2247 (61.8)	135 (63.1)	18 (58.1)	63 (61.8)	0.950
Metro-city	1389 (38.2)	79 (36.9)	13 (41.9)	39 (38.2)	
Housing					
Detached House	2091 (57.5)	124 (57.9)	16 (51.6)	55 (53.9)	0.786
Apartment	512 (14.1)	23 (10.8)	5 (16.1)	16 (15.7)	
Others	1033 (28.4)	67 (32.3)	10 (32.3)	31 (30.4)	
Home ownership					
Own	2783 (76.5)	168 (78.5)	25 (80.6)	66 (64.7)	0.045
Lease	853 (23.5)	46 (21.5)	6 (19.4)	36 (35.3)	
Living					
Alone	59 (1.6)	1 (0.5)	0 (0)	2 (2.0)	0.183
Couple only	2229 (61.3)	150 (70.1)	21 (67.7)	66 (64.7)	
Others	1348 (37.1)	63 (29.4)	10 (32.3)	34 (33.3)	
Disability					
No	2929 (80.6)	166 (77.6)	24 (77.4)	72 (70.6)	0.066
Yes	707 (19.4)	48 (22.4)	7 (22.6)	30 (29.4)	
Walking					
None	624 (18.5)	35 (16.4)	5 (16.1)	26 (27.7)	0.059
≤ 3days/week	416 (12.4)	36 (16.8)	6 (19.4)	15 (16.0)	
>3 days/week	2325 (69.1)	143 (66.8)	20 (64.5)	53 (56.4)	
Medium/Intensive Physical activity					
none	2246 (66.8)	137 (64.0)	25 (80.7)	74 (78.7)	0.109
≤ 3days/week	325 (9.7)	22 (10.3)	3 (9.7)	7 (7.5)	
>3 days/week	794 (23.6)	55 (25.7)	3 (9.7)	13 (13.8)	

	Trajectory groups				
	Low-flat (TD1) N=3636	Low-middle (TD2) N=214	Low-high (TD3) N=31	High-curve (TD4) N=102	p-value
More than 3 chronic diseases					
No	466 (12.8)	7 (3.3)	1 (3.2)	2 (2.0)	<0.0001
Yes	3170 (87.2)	207 (96.7)	30 (96.8)	100 (98.0)	
Economic Activity					
No	2275 (62.6)	142 (66.4)	25 (80.7)	80 (78.4)	0.001
Yes	1361 (37.4)	72 (33.6)	6 (19.4)	22 (21.6)	
Income quantile					
< 20	1162 (44.1)	93 (50.8)	14 (45.2)	40 (55.6)	0.117
20 - 40	603 (22.9)	39 (21.3)	8 (25.8)	16 (22.2)	
40 - 60	424 (16.1)	33 (18.0)	6 (19.4)	8 (11.1)	
60 - 80	238 (9.0)	12 (6.6)	3 (9.7)	7 (9.7)	
80 - 100	209 (7.9)	6 (3.3)	0 (0)	1 (1.4)	

Logistic regression was applied to compare each of the three depression trajectory groups (TD2 - TD4) to the “low-flat” depression trajectory group (TD1). Univariate logistic regression models were developed with TD1 as the reference trajectory. The odds ratios with 95% CI and p-value were shown in Table 4.6. The multivariate logistic analysis was then facilitated by including all the variables with a p-value smaller than 0.1. Multicollinearity was checked based on the variance inflation factor (VIF). If the VIF score exceeded 10, the variable would be excluded from the model. The backward selection method excluded the variables that were not significant from the multivariate logistic regression.

From the univariate logistic analysis with TD1 as the reference group, females, aged 80 or more, smoking now or in the past, drinking two or more days per week, and with more than three chronic diseases were significant in the “low-to-middle” depression trajectory group (TD2) (Table 4.6). Compared to the “low-to-high” depression trajectory group (TD3), females and not involved in income-generating activity had significantly higher odds. Significant predictors for the “high-curve” depression trajectory group (TD4) (Table 4.6) are the female sex, aged 75 - 79, having a university degree, currently smoking, drinking less than daily, staying in a rental house, having physical or mental disabilities, walking at least 10 minutes for more than three days per week, doing physical activities more than three days per week, having more than three chronic diseases, and not taking economic activities.

Table 4.6: Univariate Logistic Regression Analysis. Estimation of odds ratio (OR) and 95% confidence interval (C.I). Low-flat depression as the reference group.

Variable	**Low-to-Middle (n=214)**		**Low-to-High (n=31)**		**High-Curve (n=102)**	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
Sex						
Male	-	-	-	-	-	-
Female	2.06 (1.52-2.80)	<0.0001	2.31 (1.03-5.18)	0.042	2.35 (1.50-3.68)	<0.0001
Age						
65-69	-	-	-	-	-	-
70-74	1.01(0.74-1.39)	0.929	1.16 (0.49-2.74)	0.737	1.40 (0.85-2.31)	0.185
75-79	0.88 (0.60-1.30)	0.521	1.53 (0.61-3.82)	0.363	1.89 (1.12-3.21)	0.018
≥ 80	0.26 (0.12-0.53)	<0.0001	0.54 (0.12-2.44)	0.423	1.19 (0.60-2.34)	0.621
Marriage status						
Married	-	-	-	-	-	-
Single/divorce/widower	0.92 (0.69-1.24)	0.588	1.02 (0.49-2.14)	0.954	1.01 (0.67-1.53)	0.950
Education						
None	-	-	-	-	-	-
Elementary	1.23 (0.85,1.79)	0.272	1.02(0.39-2.69)	0.971	0.72 (0.44-1.19)	0.205
Middle/High	1.01 (0.67-1.53)	0.956	1.12 (0.41-3.11)	0.822	0.86 (0.51-1.44)	0.560
University	0.58 (0.28-1.20)	0.142	0.85 (0.17-4.25)	0.847	0.30 (0.09-0.98)	0.047
Smoking						
No	-	-	-	-	-	-
Previous	0.44 (0.26-0.76)	0.003	0.43 (0.10-1.83)	0.251	0.58 (0.29-1.16)	0.123
Current	0.65 (0.46-0.90)	0.011	0.79 (0.35-1.80)	0.582	0.45 (0.26-0.79)	0.005
Drinking						
No	-	-	-	-	-	-
< 2 days/week	0.77 (0.57-1.03)	0.078	0.62 (0.30-1.31)	0.209	0.60 (0.39-0.92)	0.019
2-4 days/week	0.48 (0.25-0.93)	0.030	0.31 (0.04-2.34)	0.975	0.09 (0.01-0.68)	0.019
Almost daily	0.52 (0.27-0.97)	0.041	0.57 (0.13-2.47)	0.451	0.36 (0.13-1.01)	0.053
Residential area						
No Metro-city	-	-	-	-	-	-
Metro-city	0.95 (0.71-1.26)	0.707	1.17 (0.57-2.39)	0.670	1.00 (0.67-1.50)	0.994
Housing						
Detached House	-	-	-	-	-	-
Apartment	0.76 (0.48-1.19)	0.232	1.28 (0.47-3.50)	0.636	1.19 (0.68-2.09)	0.550
Others	1.09 (0.81-1.49)	0.567	1.27 (0.57-2.80)	0.561	1.14 (0.73-1.78)	0.563
Home ownership						
Own	-	-	-	-	-	-
Lease	0.89 (0.64-1.25)	0.509	0.78 (0.32-1.91)	0.592	1.78 (1.18-2.69)	0.006
Living						
Alone	-	-	-	-	-	-
Couple only	0.25 (0.04-1.83)	0.173	1.89 (0.25-14.31)	0.962	1.14 (0.27-4.79)	0.853
Others	0.69 (0.51-0.94)	0.018	0.79 (0.37-1.68)	0.536	0.85 (0.56-1.29)	0.454
Disability						
No	-	-	-	-	-	-
Yes	1.20 (0.86-1.67)	0.286	1.21 (0.52-2.82)	0.661	1.73 (1.12-2.66)	0.014
Walking						
None	-	-	-	-	-	-
≤ 3days/week	1.54 (0.95-2.50)	0.078	1.80 (0.55-5.94)	0.334	0.87 (1.43-1.65)	0.662
>3 days/week	1.10 (0.75-1.60)	0.635	1.07 (0.40-2.87)	0.888	0.55 (0.34-0.88)	0.013
Medium/Intensive Physical activity						
none	-	-	-	-	-	-
≤ 3days/week	1.11 (0.70-1.77)	0.661	0.83 (0.25-2.76)	0.760	0.65 (0.30-1.43)	0.288
>3 days/week	1.14 (0.82-1.57)	0.441	0.34 (0.10-1.13)	0.078	0.50 (0.27-0.90)	0.021

Variable	Low-to-Middle (n=214)		Low-to- High (n=31)		High-Curve (n=102)	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
More than 3 chronic diseases						
No	-	-	-	-	-	-
Yes	4.35 (2.03-9.29)	<0.001	4.41 (0.60-32.41)	0.145	7.35 (1.81-29.9)	0.005
Economic Activity						
Yes	-	-	-	-	-	-
No	1.18 (0.88-1.58)	0.266	2.49 (1.02-6.09)	0.045	2.18 (1.35-3.50)	0.001
Income quantile						
< 20	-	-	-	-	-	-
20 - 40	0.81 (0.55-1.19)	0.280	1.10 (0.46-2.64)	0.829	0.77 (0.43-1.39)	0.386
40 - 60	0.97 (0.64-1.47)	0.894	1.18 (0.45-3.08)	0.743	0.55 (0.25-1.18)	0.125
60 - 80	0.63 (0.34-1.17)	0.142	1.05 (0.30-3.67)	0.944	0.85 (0.38-1.93)	0.705
80 - 100	0.36 (0.16-0.83)	0.017	0.43 (0.06-3.29)	0.976	0.14 (0.02-1.02)	0.052

In multivariate logistic regression analysis (Table 4.7), compared to the "low-flat" depression trajectory group (TD1), the members from "low-to-middle" depression trajectory group (TD2) were more likely to be females (OR = 1.82, 95% CI: 1.31 - 2.53, p-value < 0.0001) and to have more than three chronic diseases (OR = 4.15, 95% CI: 1.93 - 8.93, p-value < 0.0001), with age 80 or more (OR = 0.33, 95% CI: 0.15 - 0.73, p-value = 0.006) as an adjusted covariate. Being female was the only significant factor for comparing the "low-to-high" depression trajectory group (TD3) to TD1 (OR = 2.31, 95% CI: 1.03 - 5.18, p-value = 0.042). Individuals from the "high-curve" depression trajectory group (TD4) were more likely to be females (OR = 2.02, 95% CI: 1.16 - 3.54, p-value = 0.014), to have more

Table 4.7: Multivariate Logistic Regression Analysis. Estimation of odds ratio (OR) and 95% confidence interval (C.I). Low-flat depression as the reference group

Variable	Low-to-Middle (n=214)		Low-to-High (n=31)		High-Curve (n=102)	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
Sex						
Male	-	-	-	-	-	-
Female	1.82 (1.31-2.53)	<0.0001	2.31 (1.03-5.18)	0.042	2.02 (1.16-3.54)	0.014
Age						
65-69	-	-				
70-74	1.05 (0.74-1.49)	0.782				
75-79	0.96 (0.63-1.47)	0.865				
≥ 80	0.33 (0.15-0.73)	0.006				
More than 3 chronic diseases						
No	-	-			-	-
Yes	4.15 (1.93-8.93)	<0.0001			5.18 (1.26-21.3)	0.023
Home ownership						
Own					-	-
Lease					2.06 (1.25-3.40)	0.005

than three chronic diseases (OR = 5.18, 95% CI: 1.26 - 21.29, p-value = 0.023) and to live in a rental home (OR = 2.06, 95% CI: 1.25 - 3.40, p-value = 0.005).

4.4.1.3 Risk factors for depression using GBTM

As we mentioned in Section 3.3.3, adding risk factors to group memberships would vary the probabilities of trajectory groups, but rarely change the trajectory proportion from the overall population and trajectory shapes. In our study, all variables from multivariate logistic regression analysis (Table 4.7) were considered risk factors for depression trajectory groups. After adding all the risk factors (female sex, age 65-69, having more than three chronic diseases, and living in a rental house), each group's trajectory shapes (Figure 4.4) had tiny changes compared to Figure 4.3. In Figure 4.4, the proportion of "low-flat" (TD1), "low-to-middle" (TD2), and "low-to-high" (TD3) depression trajectory groups had a very small percentage change. For example, the "low-flat" depression trajectory group (TD1) moved 1.8% to the "low-to-middle" depression trajectory group (TD2) and 0.2% to the "low-to-

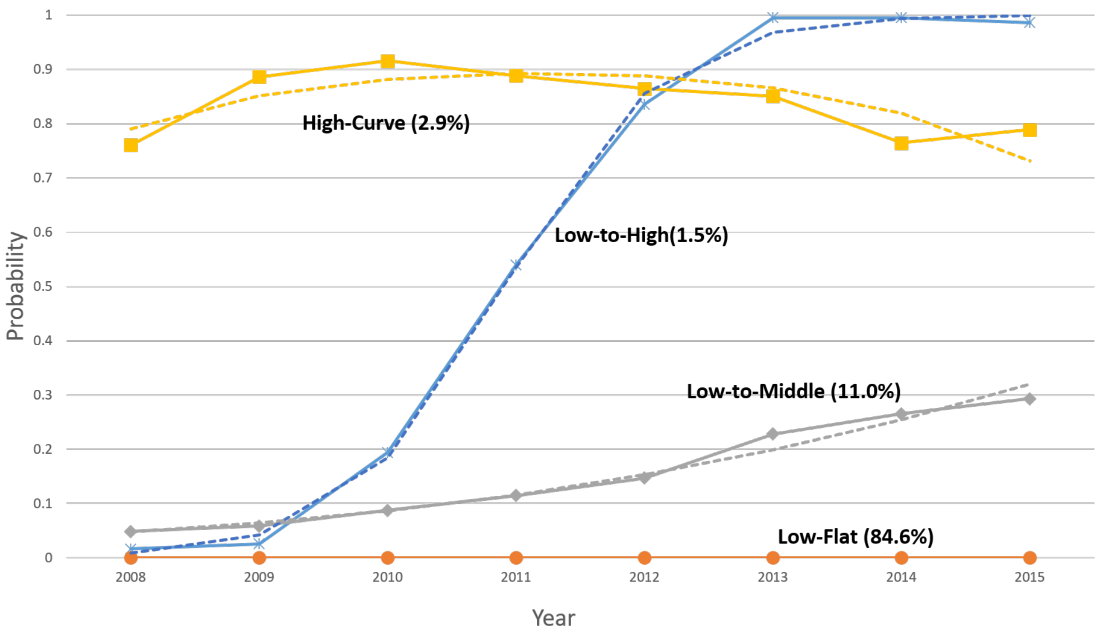


Figure 4.4: Depression trajectories with four risk factors for GBTM. The solid line indicates observed depression; the dashed line indicates predicted depression.

high" depression trajectory group (TD3). The "high-curve" depression trajectory group (TD4) remained the same.

Table 4.8 showed the parameters that allowed the trajectory groups' probabilities to vary as a function of the four risk factors mentioned above in Table 4.7. The p-values of constants were all significant in TD2 - TD4. Being female and having more than three chronic diseases were influential in TD2 and TD4 with TD1 as the reference group. Age 65-69 was significant in TD2. Living in a rental house was only significant in TD4 (Table 4.8).

Table 4.8: Parameter estimates for risk factors by depression trajectory group

Group	Parameter	Estimate	Standard Error	p-value
Low-flat (TD1)	Baseline	0	-	-
Low-middle (TD2)	Constant	-4.64259	0.57568	<0.0001
	Female	0.64925	0.17733	0.0003
	Age 65-69	0.91592	0.42111	0.0296
	> 3 chronic disease	1.50048	0.39844	0.0002
	Living in a rental house	-0.10234	0.19675	0.6030
Low-high (TD3)	Constant	-6.07875	1.45121	<0.0001
	Female	0.71159	0.44158	0.1071
	Age 65-69	0.24331	0.82911	0.7692
	> 3 chronic disease	1.60264	1.21514	0.1872
	Living in a rental house	-0.43588	0.52016	0.4021
High-curve (TD4)	Constant	-6.55316	0.98470	<0.0001
	Female	0.92573	0.24877	0.0002
	Age 65-69	0.46373	0.34837	0.1832
	> 3 chronic disease	2.09969	0.90960	0.0210
	Living in a rental house	0.54051	0.21838	0.0133

Based on the parameters from Table 4.8, the probability of group membership with the influence of the risk factors can be calculated. Table 4.9 listed situations for estimating group membership probabilities (No risk factors, female only, age 65-69 only, chronic disease only, living in a rental house only, and all risk factors).

As seen in Table 4.9, the percentage of probabilities increased with significant risk factors in depression trajectory groups TD2 - TD4. For instance, the individuals with no risk factors had a probability percentage of 98.683% in TD1, 0.951% in TD2, 0.226% in TD3, and 0.143% in TD4, respectively. If we considered the individuals with more than three chronic diseases, the probability of TD1 would decline to 93.791% from 98.683%; additionally, the probability in TD2, TD3, and TD4 rose to 4.051%, 1.067%, and 1.092%, respectively.

Table 4.9: Percentage of group membership probability with risk factors

	Trajectory groups (%)			
Risk factors	Low-Flat (TD1)	Low-to-middle (TD2)	Low-to-high (TD3)	High-curve (TD4)
No risk factors	98.683	0.951	0.226	0.143
Female only	97.399	1.796	0.455	0.35
Age 65-69 only	96.993	1.788	0.283	0.27
Chronic disease only	93.791	4.051	1.067	1.092
Living in a rental house	98.876	0.859	0.148	0.245
All factors	77.773	14.505	1.4869	6.236

An alternative way to check group membership probability with different risk factors could be seen in Figure 4.5. The bar plot in Figure 4.5 showed that individuals with only one risk factor (female only, age 65-69 only, chronic disease only, living in a rental house only) had only a small proportion change compared to individuals with no risk factors. Nevertheless, individuals with all risk factors had prodigious probability variation in each trajectory group. Compared to the individuals having no risk factors, the probability proportion for individuals with all risk factors in TD1 decreased 20.9% and increased 13.5% in TD2, 1.3% in TD3, and 6.1% in TD4.

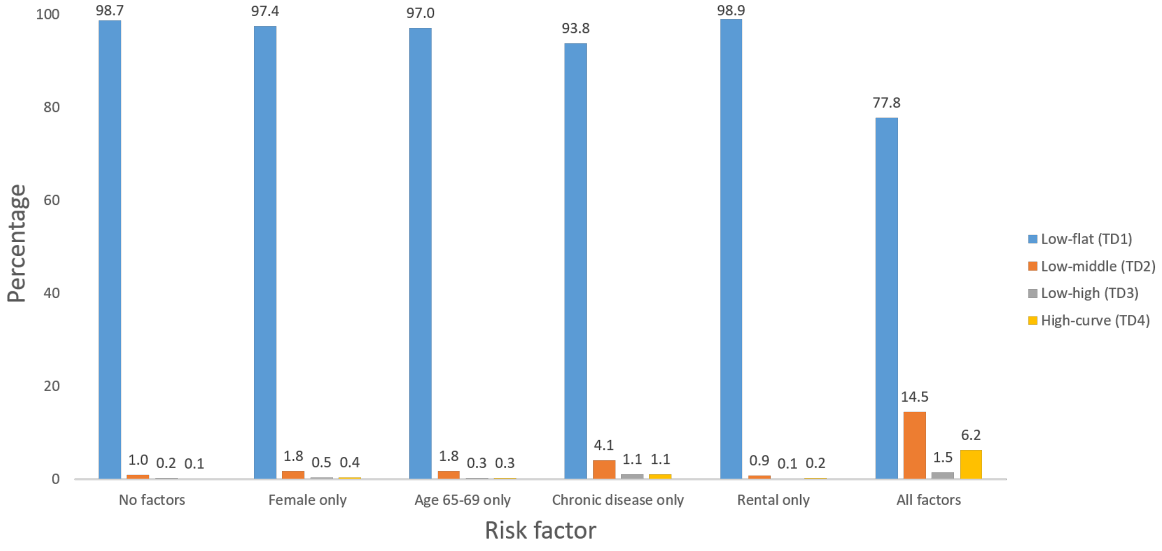


Figure 4.5: Bar plot for percentage of depression group membership in GBTM

4.4.2 GBTM for anxiety

4.4.2.1 Development of GBTM

Using the same procedures as GBTM with depression outcomes, GBTM with anxiety outcomes also identified with four trajectory groups: "low-flat" (TA1), "low-to-middle" (TA2), "high-to-low" (TA3) and "high-curve" (TA4). The four trajectory groups were constituted with one flat trajectory, two linear trajectories, and one curve shape trajectory (Table 4.10). Figure 4.6 showed the four trajectory groups, represented by solid lines for the accurate averages and dashed lines for predicted values.

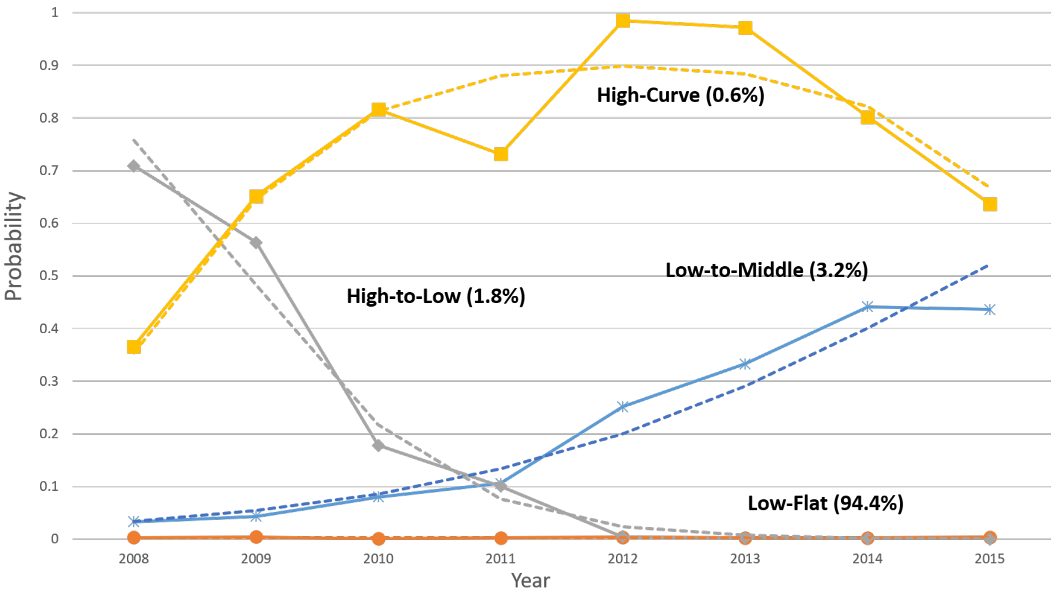


Figure 4.6: Anxiety trajectories for GBTM. The solid line indicates observed depression; the dashed line indicates predicted depression.

The "low-flat" anxiety trajectory group, TA1 (n=3843, 94.4%), had a low-flat probability of anxiety close to zero. This group included most of the study participants. The "low-to-middle" anxiety trajectory group, TA2 (n=59, 3.2%), started with low anxiety probability (around 0.05) in 2008 and increased to more than 0.4 in 2015. The "high-to-low" anxiety trajectory group, TA3 (n=68, 1.8%), began with an anxiety probability of around 0.7 but fell to zero in 2013. The "high-curve" trajectory group, TA4 (n=13, 0.6%), began with a likelihood of more than 0.35 in 2008. This rose to nearly 1.0 in 2012, then declined to 0.65 in 2015. Table 4.10 presented the parameter estimates of anxiety trajectory shapes in GBTM. All the p-values were significant at $\alpha = 0.05$.

Table 4.10: Parameter estimates for trajectory shapes in anxiety GBTM

Group	Parameter	Estimate	Standard Error	p-value
Low-flat (TA1)	Intercept	-5.80093	0.30180	<0.0001
Low-to-middle (TA2)	Intercept	-3.83411	0.52631	<0.0001
	Linear	0.49000	0.09159	<0.0001
High-to-low (TA3)	Intercept	2.35013	0.83613	0.0049
	Linear	-1.21053	0.27995	<0.0001
High-curve (TA4)	Intercept	-2.14626	0.92160	0.0199
	Linear	1.71445	0.51304	0.0008
	Quadratic	-0.16987	0.05674	0.0028

Table 4.11 showed the estimated group membership proportions. All the proportions were highly significant.

Table 4.11: Estimates of group membership proportions in anxiety GBTM

Group membership proportions			
Group	Estimate Proportion (%)	Standard Error	p-value*
Low-flat (TA1)	94.37189	0.98638	<0.0001
Low-middle (TA2)	3.23915	0.75668	<0.0001
High-low (TA3)	1.82509	0.47193	0.0001
High-curve (TA4)	0.56386	0.17349	0.0012

*H_0: Proportion = 0 vs $H_{a:}$ Proportion $\neq$ 0

4.4.2.2 Characteristics of trajectory groups

Baseline characteristics were compared using chi-square tests (Table 4.12). Females presented the lowest percentage in the “low-flat” anxiety trajectory group (TA1), but highest in the “high-to-low” anxiety trajectory group (TA3). TA1 had the lowest proportion of non-smokers, while TA3 included the highest percentage of current smokers. No overall difference was shown among age groups, marriage status, level of education, drinking habits, residential area, housing type, homeownership, living alone or not, mental or physical disability status, daily walking, involvement in physical activities, having more than three chronic diseases, taking income-generating activities and different income quantile.

Table 4.12: Distribution of baseline characteristics by anxiety trajectory groups from GBTM (N, %)

	Trajectory groups				
	Low-flat (TA1) N=3843	Low-middle (TA2) N=59	High-low (TA3) N=68	High-curve (TA4) N=13	p-value
Sex					
Male	1677 (43.6)	19 (32.2)	14 (20.6)	4 (30.8)	<0.0001
Female	2166 (56.4)	40 (67.8)	54 (79.4)	9 (69.2)	
Age					
65-69	1481 (38.5)	24 (40.7)	25 (36.8)	4 (30.8)	0.983
70-74	1169 (30.4)	20 (33.9)	22 (32.4)	4 (30.8)	
75-79	713 (18.6)	11 (18.6)	13 (19.1)	3 (23.1)	
≥ 80	480 (12.5)	4 (6.8)	8 (11.8)	2 (15.4)	
Marriage status					
Married	2506 (65.2)	39 (66.1)	38 (55.9)	9 (69.2)	0.441
Single/divorce/widower	1336 (34.8)	20 (33.9)	30 (44.1)	4 (30.8)	
Education					
None	768 (20.0)	10 (16.7)	19 (27.9)	2 (15.4)	0.278
Elementary	1637 (42.6)	32 (54.2)	27 (39.7)	9 (69.2)	
Middle/High	1148 (29.9)	13 (22.0)	19 (27.9)	1 (7.7)	
University	290 (7.6)	4 (6.8)	3 (4.4)	1 (7.7)	
Smoking					
No	2122 (59.5)	38 (65.5)	50 (76.9)	8 (61.5)	0.035
Previous	458 (12.8)	3 (5.2)	6 (9.2)	0 (0.0)	
Current	986 (27.7)	17 (29.3)	9 (13.9)	5 (38.5)	
Drinking					
No	1435 (40.2)	20 (34.5)	34 (52.3)	6 (46.2)	0.387
< 2 days/week	1577 (44.2)	29 (50.0)	28 (43.1)	6 (46.2)	
2-4 days/week	272 (7.6)	4 (6.9)	1 (1.5)	1 (7.7)	
Almost daily	284 (8.0)	5 (8.6)	2 (3.1)	0 (0.0)	
Residential area					
No Metro-city	2376 (61.8)	35 (59.3)	44 (64.7)	8 (61.5)	0.941
Metro-city	1467 (38.2)	24 (40.7)	24 (35.3)	5 (38.5)	
Housing					
Detached House	2204 (57.4)	38 (64.4)	37 (54.4)	7 (53.9)	0.881
Apartment	537 (14.0)	7 (11.9)	11 (16.2)	1 (7.7)	
Others	1102 (28.7)	14 (23.7)	20 (29.4)	5 (38.5)	
Home ownership					
Own	2937 (76.4)	49 (83.1)	49 (72.1)	7 (53.9)	0.121
Lease	906 (23.6)	10 (16.9)	19 (27.9)	6 (46.1)	
Living					
Alone	61 (1.6)	1 (1.7)	0 (0)	0 (0.0)	0.892
Couple only	2379 (61.9)	39 (66.1)	40 (58.8)	8 (61.5)	
Others	1403 (36.5)	19 (32.2)	28 (41.2)	5 (38.5)	
Disability					
No	3076 (80.0)	50 (84.8)	52 (76.5)	13 (100)	0.204
Yes	767 (20.0)	9 (15.2)	16 (23.5)	0 (0)	
Walking					
None	662 (18.6)	7 (12.1)	18 (27.7)	3 (23.1)	0.390
≤ 3days/week	458 (12.8)	8 (13.8)	5 (7.7)	2 (15.4)	
>3 days/week	2448 (68.6)	43 (74.1)	42 (64.6)	8 (61.5)	
Medium/Intensive Physical activity					
none	2390 (67.0)	32 (55.2)	49 (75.4)	11 (84.6)	0.133
≤ 3days/week	344 (9.6)	6 (10.3)	7 (10.8)	0 (0.0)	
>3 days/week	834 (23.4)	20 (34.5)	9 (13.9)	2 (15.4)	

	Trajectory groups				
	Low-flat (TA1) N=3843	Low-middle (TA2) N=59	High-low (TA3) N=68	High-curve (TA4) N=13	p-value
More than 3 chronic diseases					
No	466 (12.1)	8 (13.6)	2 (2.9)	0 (0.0)	0.064
Yes	3377 (87.9)	51 (86.4)	66 (97.1)	13 (100)	
Economic Activity					
No	2425 (63.1)	37 (62.7)	51 (75.0)	9 (69.3)	0.233
Yes	1418 (36.9)	22 (37.3)	17 (25.0)	4 (30.7)	
Income quantile					
< 20	1249 (44.6)	28 (47.5)	22 (43.1)	10 (76.9)	0.503
20 - 40	636 (22.7)	15 (25.4)	14 (27.5)	1 (7.7)	
40 - 60	452 (16.2)	7 (11.9)	11 (21.6)	1 (7.7)	
60 - 80	252 (9.0)	6 (10.2)	2 (3.9)	0 (0.0)	
80 - 100	210 (7.5)	3 (5.1)	2 (3.9)	1 (7.7)	

The univariate logistic regression analysis showed that doing physical activity more than three days per week was the only significant variable for comparing the "low-to-middle" anxiety trajectory group (TA2) to the "low-flat" anxiety trajectory group (TA1). Compared to TA1, the female sex, current smoking, having more than three chronic diseases, and doing income-generating activities were significant in the "high-to-low" anxiety trajectory group (TA3) in the univariate analysis (Table 4.13). Comparing the "high-curve" anxiety trajectory group (TA4) to TA1, none of the variables were significant because the sample size is too small in TA4 (n=13).

Table 4.13: Univariate Logistic Regression Analysis. Estimation of odds ratio (OR) and 95% confidence interval (C.I). Low-flat anxiety as the reference group.

Variable	**Low-to-Middle (n=59)**		**High-to-Low (n=68)**		**High-Curve (n=13)**	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
Sex						
Male	-	-	-	-	-	-
Female	1.63 (0.94-2.83)	0.082	2.99 (1.65-5.39)	<0.0001	1.74 (0.54-5.67)	0.356
Age						
65-69	-	-	-	-	-	-
70-74	1.06 (0.58-1.92)	0.859	1.12 (0.63-1.99)	0.712	1.27 (0.32-5.08)	0.739
75-79	0.95 (0.46-1.95)	0.893	1.08 (0.55-2.12)	0.823	1.56 (0.35-6.98)	0.562
≥ 80	0.51 (0.18-1.49)	0.220	0.99 (0.44-2.20)	0.975	1.54 (0.28-8.45)	0.617
Marriage status						
Married	-	-	-	-	-	-
Single/divorce/widower	0.96 (0.56-1.66)	0.889	1.48 (0.91-2.40)	0.111	0.83 (0.26-2.71)	0.763
Education						
None	-	-	-	-	-	-
Elementary	1.50 (0.73-3.07)	0.266	0.67 (0.37-1.21)	0.180	2.11 (0.46-9.79)	0.340
Middle/High	0.87 (0.38-1.99)	0.742	0.67 (0.35-1.27)	0.220	0.34 (0.03-3.70)	0.372
University	1.06 (0.33-3.40)	0.923	0.42 (0.12-1.42)	0.163	1.32 (0.12-14.6)	0.819

Variable	**Low-to-Middle (n=59)**		**High-to-Low (n=68)**		**High-Curve (n=13)**	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
Smoking						
No	-	-	-	-	-	-
Previous	0.37 (0.11-1.19)	0.095	0.56 (0.24-1.30)	0.177	0.01 (0.01-999)	0.969
Current	0.96 (0.54-1.71)	0.898	0.39 (0.19-0.79)	0.001	1.35 (0.44-4.12)	0.604
Drinking						
No	-	-	-	-	-	-
< 2 days/week	1.32 (0.74-2.43)	0.344	0.75 (0.45-1.24)	0.263	0.91 (0.29-2.83)	0.870
2-4 days/week	1.06 (0.36-3.11)	0.923	0.16 (0.02-1.14)	0.067	0.88 (0.11-7.33)	0.905
Almost daily	1.26 (0.47-3.39)	0.643	0.30 (0.07-1.24)	0.097	0.01 (0.01-999)	0.976
Residential area						
No Metro-city	-	-	-	-	-	-
Metro-city	1.11 (0.66-1.88)	0.695	0.88 (0.54-1.46)	0.628	1.01 (0.33-3.10)	0.983
Housing						
Detached House	-	-	-	-	-	-
Apartment	0.76 (0.34-1.70)	0.500	1.22 (0.62-2.41)	0.566	0.59 (0.07-4.78)	0.618
Others	0.74 (0.40-1.37)	0.332	1.08 (0.63-1.87)	0.781	1.43 (0.45-4.51)	0.543
Home ownership						
Own	-	-	-	-	-	-
Lease	0.66 (0.33-1.31)	0.237	1.26 (0.74-2.15)	0.402	2.78 (0.93-8.29)	0.067
Disability						
No	-	-	-	-	-	-
Yes	0.72 (0.35-1.48)	0.372	1.23 (0.70-2.17)	0.466	0.01 (0.01-999)	0.961
Walking						
None	-	-	-	-	-	-
≤ 3days/week	1.65 (0.60-4.59)	0.336	0.40 (0.15-1.09)	0.073	0.96 (0.16-5.79)	0.968
>3 days/week	1.66 (0.74-3.71)	0.216	0.63 (0.36-1.10)	0.106	0.72 (0.19-2.73)	0.630
Medium/Intensive Physical activity						
none	-	-	-	-	-	-
≤ 3days/week	1.30 (0.54-3.14)	0.661	0.99 (0.45-2.21)	0.985	0.01 (0.01-999)	0.973
>3 days/week	1.79 (1.02-3.15)	0.043	0.53 (0.26-1.08)	0.079	0.52 (0.12-2.36)	0.397
More than 3 chronic diseases						
No	-	-	-	-	-	-
Yes	0.88 (0.42-1.87)	0.738	4.55 (1.11-18.65)	0.035	999 (0.01-999)	0.970
Economic Activity						
Yes	-	-	-	-	-	-
No	0.98 (0.58-1.67)	0.951	1.75 (1.01-3.05)	0.046	1.32 (0.40-4.28)	0.649
Income quantile						
< 20	-	-	-	-	-	-
20 - 40	1.05 (0.56-1.98)	0.875	1.25 (0.64-2.46)	0.519	0.20 (0.03-1.54)	0.121
40 - 60	0.69 (0.30-1.59)	0.385	1.38 (0.67-2.87)	0.387	0.28 (0.04-2.17)	0.221
60 - 80	1.06 (0.44-2.59)	0.895	0.45 (0.11-1.93)	0.283	0.01 (0.01-999)	0.973
80 - 100	0.64 (0.19-2.12)	0.462	0.54 (0.13-2.32)	0.408	0.60 (0.08-4.67)	0.621

In the multivariate logistic analysis (Table 4.14), compared to the “low-flat” anxiety trajectory group (TA1), the female sex was the only significant variable in “high-to-low” anxiety trajectory group (TA3) (OR = 2.99, 95% CI: 1.65 - 5.39, p-value < 0.0001).

Table 4.14: Multivariate Logistic Regression Analysis. Estimation of odds ratio (OR) and 95% confidence interval (C.I). Low-flat anxiety as the reference group

Variable	**Low-to-Middle (n=59)**		**High-to-Low (n=68)**		**High-Curve (n=13)**	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
Sex						
Male			-	-		
Female			2.99 (1.65-5.39)	<0.0001		

4.4.2.3 Risk factors for anxiety using GBTM

Since the female sex was the only significant variable found from the multivariate logistic analysis, it was considered a risk factor for anxiety in GBTM. After it was added to the model, each group's trajectory shape rarely changed, but the overall proportion of each trajectory group changed (Figure 4.7). The "low-flat" anxiety trajectory group (TA1) decreased by 1.1%. The "low-to-middle" (TA2), "high-to-low" (TA3), and "high-curve" (TA4) anxiety trajectory groups increased by 0.6%, 0.5%, and 0.1%, respectively.

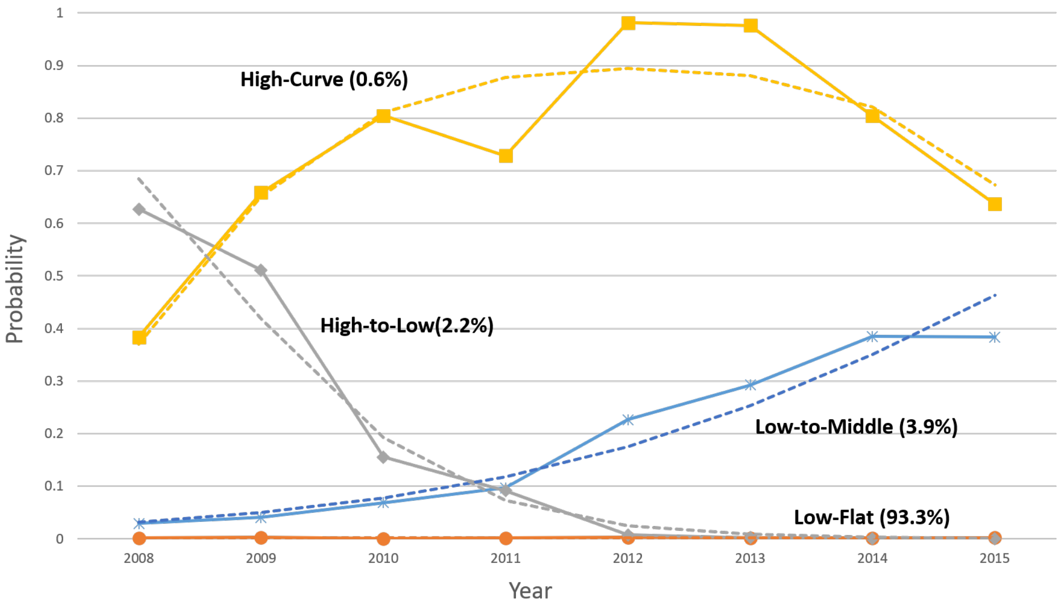


Figure 4.7: Anxiety trajectories with female as a risk factor for GBTM. The solid line indicates observed depression; the dashed line indicates predicted depression.

With TA1 as the reference trajectory group, the p-values of the intercept were all significant in TA1, TA2 and TA3. The p-values were significant for the female sex in TA2 and

TA3, but not in TA4 (Table 4.15).

Table 4.15: Parameter estimates for risk factors by anxiety trajectory group

Group	Parameter	Estimate	Standard Error	p-value
Low-flat (TA1)	Baseline	0	-	-
Low-middle (TA2)	Constant Female	-4.58415 0.84928	0.64830 0.33019	<0.0001 0.0101
High-low (TA3)	Constant Female	-6.38908 1.53314	0.95954 0.47305	<0.0001 0.0012
High-curve (TA4)	Constant Female	-5.58501 0.33610	1.05165 0.60032	<0.0001 0.5756

The proportion difference between males and females in each trajectory group was presented in Figure 4.8. Males were 2.0% more in TA1 compared to females. On the other hand, females were 1.3%, 0.5%, and 0.2% more in TA2 - TA4 than males.

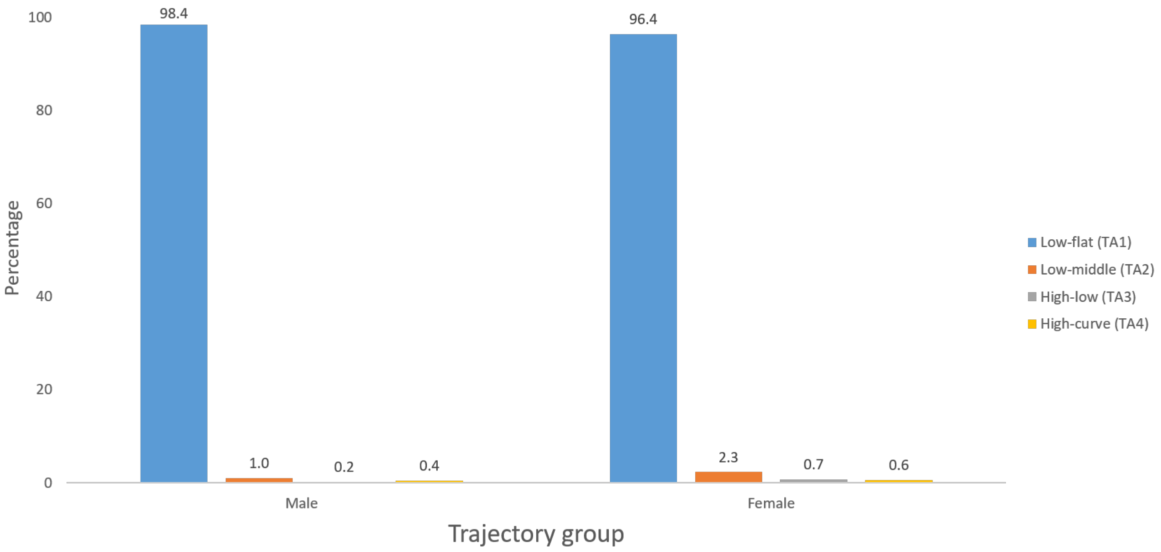


Figure 4.8: Bar plot for percentage of group membership in males and females

4.5 Analysis of group-based dual trajectory modeling (GBDTM)

4.5.1 Development of GBDTM

Since depression and anxiety were the two outcomes we were interested in, group-based dual trajectory modeling (GBDTM) was used to develop the trajectories of depression and anxiety jointly. The GBDTM diagram of this study between depression and anxiety was shown in Figure 4.9.

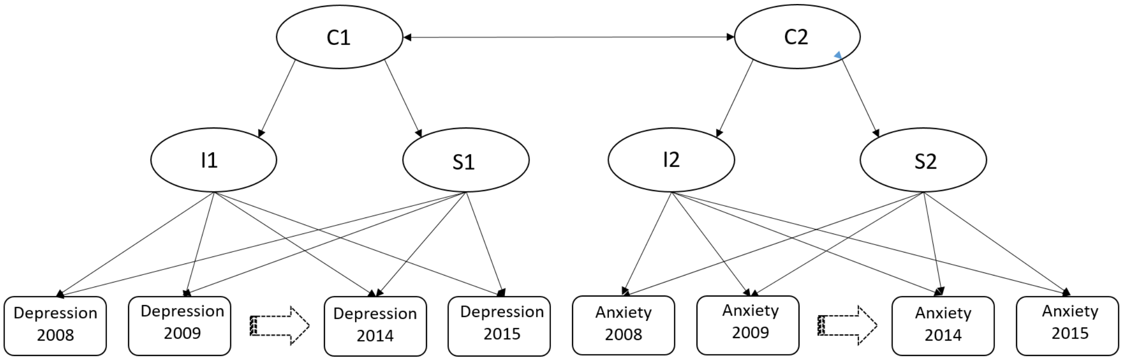


Figure 4.9: Diagram of GBDTM with depression and anxiety from 2008-2015. I1, I2 and S1, S2 are the latent intercepts and slopes for depression and anxiety. C1 and C2 are the depression and anxiety trajectory groups associated each other (Huang et al., 2013).

Group-based dual trajectory modeling (GBDTM) provided four trajectories of both depression and anxiety (Figure 4.10 and Figure 4.11). The shapes of the depression and anxiety trajectories were very similar to the trajectories from GBTM. However, group memberships for each trajectory group were changed. Compared to the depression trajectories of GBTM,

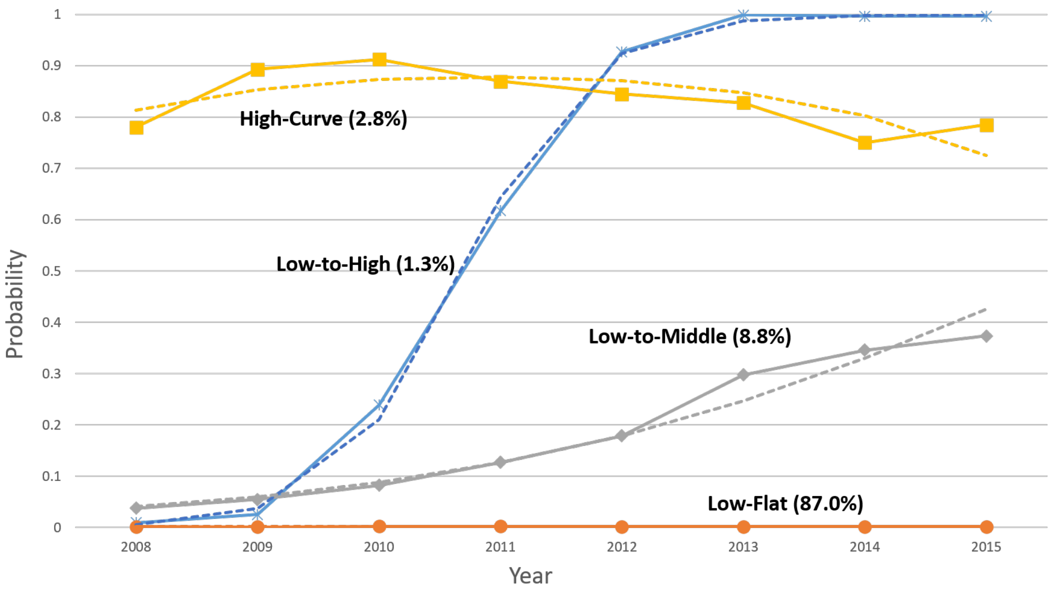


Figure 4.10: Depression trajectories for GBDTM. The solid line indicates observed depression; the dashed line indicates predicted depression.

the “low-flat” trajectory group (DD1) in GBTDM slightly increased with members N=3641 (87%) (Figure 4.10). The “low-to-middle” trajectory group (DD2) was reduced with members N=205 (8.8%). The “low-to-high” trajectory group (DD3) increased with members N=33. Still, the percentage stayed the same at 1.3%. The “high-curve” trajectory group (DD4) grew with members N=104 but decreased with a rate of 2.8% .

In anxiety trajectories of GBDTM (Figure 4.11), the “low-flat” trajectory group (DA1) was reduced with members N=3785 (92.5%). The “low-to-middle” trajectory group (DA2) grew with members N=96 (4.7%). The “high-to-low” trajectory group (DA3) increased with members N=89 (2.2%). The “high-curve” trajectory group (DA4) remained unchanged with members N=13 (0.6%).

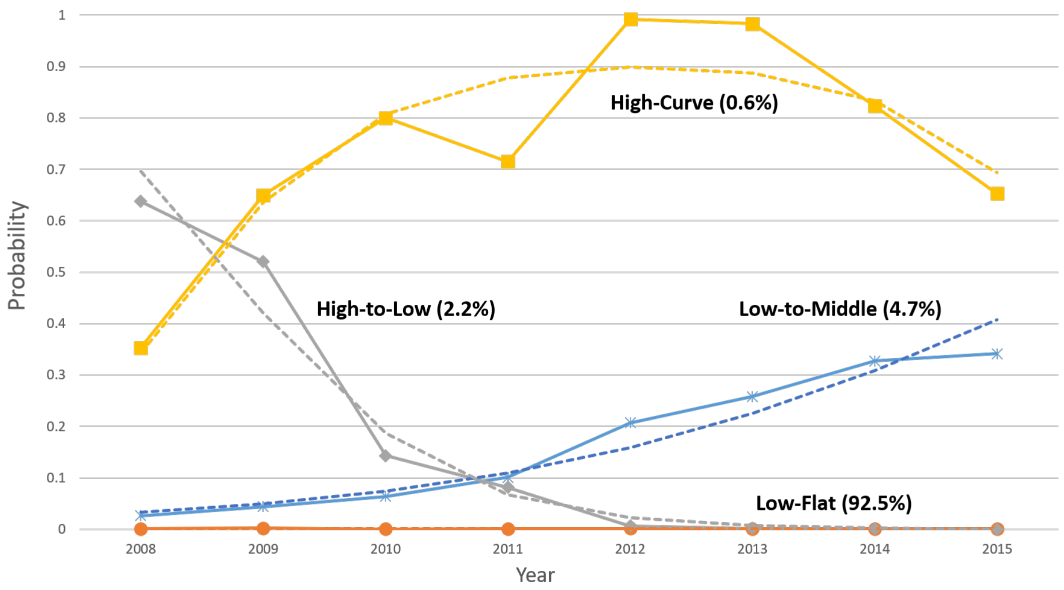


Figure 4.11: Anxiety trajectories for GBDTM. The solid line indicates observed anxiety; the dashed line indicates predicted anxiety.

Table 4.16 showed the parameters of depression trajectory shapes from GBDTM. DD1, DD2, and DD3 all had significant intercept and linear polynomial functions. The intercept and linear function were not significant with DD4, but the quadratic function was significant. Compared to the standard error of estimates in GBTM (Table 4.3), the standard errors of the estimates in GBDTM for depression were decreased in the intercept of DD1, the linear function of DD2, and the intercept and linear function of DD3, but the rest of the standard errors were increased (Table 4.16).

Table 4.16: Parameter estimates for trajectory shapes in depression GBDTM

Group	Parameter	Estimate	Standard Error*	SE Difference vs GBTM$	p-value
Low-flat (DD1)	Intercept	-6.22889	0.42701	-0.13898	<0.0001
Low-middle (DD2)	Intercept	-3.57016	0.28545	0.00906	<0.0001
	Linear	0.40901	0.05255	-0.00312	<0.0001
Low-high (DD3)	Intercept	-7.05625	1.81855	-0.07945	0.0001
	Linear	1.91196	0.50078	-0.02282	0.0001
High-curve (DD4)	Intercept	1.06733	0.55398	0.00575	0.0540
	Linear	0.46861	0.28271	0.00122	0.0974
	Quadratic	-0.06008	0.03029	0.00018	0.0473

* Standard Error of GBDTM with depression

$ Standard Error difference between GBDTM and GBTM with depression

Table 4.17 showed the estimates of depression trajectory group membership proportions in GBDTM. All the p-values were significant. Compared to the depression proportions in GBTM (Table 4.4), the standard errors of the depression proportions for GBDTM were decreased in DD1 and DD2, but increased in DD3 and DD4 (Table 4.17).

Table 4.17: Estimates of group membership proportions in depression GBDTM

Group membership proportions				
Group	Estimate Proportion (%)	Standard Error*	SE Difference vs GBTM$	p-value#
Low-flat (DD1)	86.95774	1.22352	-0.20001	<0.0001
Low-middle (DD2)	8.84487	1.15473	-0.17743	<0.0001
Low-high (DD3)	1.34794	0.30174	0.00751	<0.0001
High-curve (DD4)	2.84945	0.32924	0.00584	<0.0001

* Standard Error of GBDTM with depression

$ Standard Error difference between GBDTM and GBTM with depression

H_0: Proportion = 0 vs $H_{a:}$ Proportion $\neq$ 0

Table 4.18 showed the parameters of anxiety trajectory shapes from GBDTM. All the p-values were significant. Compared to the standard error of estimates in GBTM (Table 4.10), the standard errors of the estimates in GBDTM for anxiety were increased in the intercept of DA1, and linear and quadratic function of DA, but the rest of the standard errors were decreased (Table 4.18).

Table 4.18: Parameter estimates for trajectory shapes in anxiety GBDTM

Group	Parameter	Estimate	Standard Error*	SE Difference vs GBTM$^{\$}$	P-value
Low-flat (DA1)	Intercept	-6.53934	0.53068	0.22888	<0.0001
Low-middle (DA2)	Intercept	-3.81124	0.40827	-0.11804	<0.0001
	Linear	0.42936	0.07375	-0.01784	<0.0001
High-low (DA3)	Intercept	1.98190	0.68186	-0.15427	0.0037
	Linear	-1.15063	0.23592	-0.04403	<0.0001
High-curve (DA4)	Intercept	-2.19597	0.90875	-0.01285	0.0157
	Linear	1.71044	0.51611	0.00307	0.0009
	Quadratic	-0.16671	0.05781	0.00107	0.0039

* Standard Error of GBDTM with anxiety

$ Standard Error difference between GBDTM and GBTM with anxiety

Table 4.19 showed the estimates of anxiety trajectory group membership proportions in GBDTM. All the p-values were significant. Compared to the anxiety proportions in GBTM (Table 4.11), the standard errors of the anxiety proportions for GBDTM were decreased in DA1 and DA4, but increased in DA2 and DA3 (Table 4.19).

Table 4.19: Estimates of group membership proportions anxiety GBDTM

Group membership proportions				
Group	Estimate Proportion (%)	Standard Error*	SE Difference vs GBTM$^{\$}$	p-value$^{\#}$
Low-flat (DA1)	92.47600	0.91558	-0.0708	<0.0001
Low-middle (DA2)	4.72011	1.09933	0.34265	<0.0001
High-low (DA3)	2.22835	0.49703	0.02510	<0.0001
High-curve (DA4)	0.57554	0.17263	-0.00086	0.0009

* Standard Error of GBDTM with anxiety

$ Standard Error difference between GBDTM and GBTM with anxiety

H_0: Proportion = 0 vs $H_{a:}$ Proportion $\neq$ 0

4.5.2 Conditional probability using GBDTM

For this study, GBDTM linked trajectory groups for depression and anxiety based on conditional probabilities. These probabilities can be calculated with the Bayes rule (Nagin, 2005). Since depression and anxiety are diagnosed as co-current events in our study, the conditional probability both for depression given anxiety and anxiety given depression should be considered (Wiesner & Kim, 2006). The conditional probability represented the likelihood of a person having depression if they had already been diagnosed with anxiety or vice versa (Nagin, 2005). Thus, conditional probabilities from GBDTM provided a clear view of the association between depression and anxiety.

Based on the conditional probability of anxiety given depression (Figure 4.12A), the older adults in the "low-flat" depression trajectory (DD1) were more likely to belong to the "low-flat" anxiety trajectory (DA1) compared to the older adults in the "high-curve" depression trajectory (DD4) (95.7% vs. 68.5%). Also, the older adults in the "low-to-middle" depression trajectory (DD2) were more likely to belong to the "low-to-middle" anxiety trajectory (DA2) than the "low-flat" depression trajectory (DD1) (22.9% vs. 2.7%). Furthermore, the older people in the "high-curve" depression trajectory (DD4) had a greater chance of belonging to the "high-to-low" anxiety trajectory (DA3) (21.1% vs. 1.5%) compared to the older adults in the "low-flat" depression trajectory (DD1).

Based on the conditional probability of depression given anxiety (Figure 4.12B), the older people in the "low-flat" anxiety trajectory (DA1) were more likely to belong to the "low-flat" depression trajectory (DD1) compared to the older people in the "high-curve" anxiety trajectory (DA4) (90.0% vs. 23.9%). The older adults belonging to the "low-to-middle" and "high-curve" anxiety trajectory (DA2 and DA4) were more likely to belong to the "low-to-middle" depression trajectory (DD2) than the older people in the "low-flat" anxiety trajectory (DA1) (43.0%, 59.5% vs. 6.7%). Moreover, the older adults in "high-to-low" and "high-curve" anxiety trajectories (DA3 and DA4) were more likely to belong to the "high-curve" depression trajectory (DD4) than the older adults in the "low-flat" anxiety trajectory (DA1) (27.0%, 16.6% vs. 6.7%).

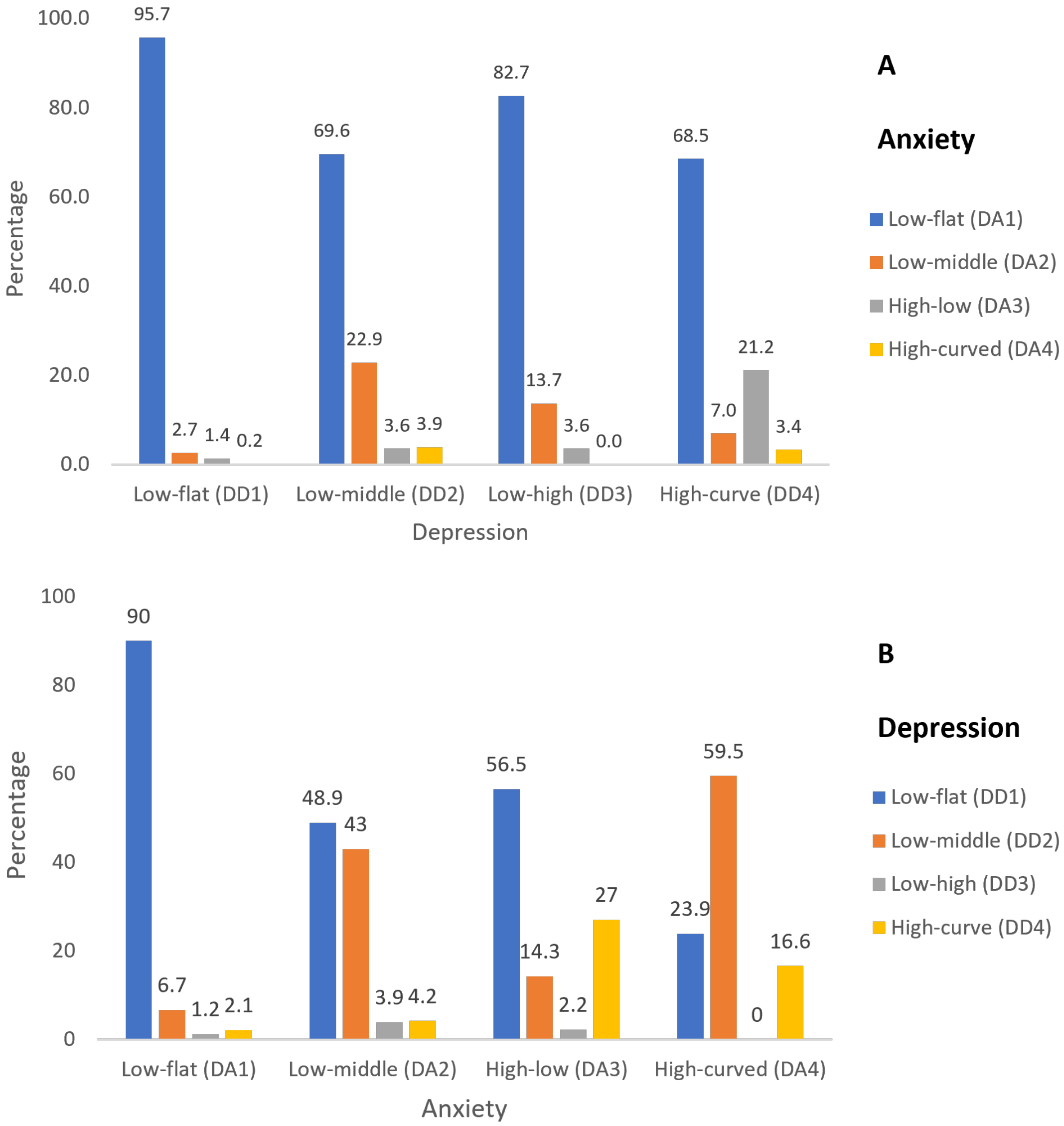


Figure 4.12: Conditional probability of anxiety given depression (A). Conditional probability of depression given anxiety (B).

4.5.3 Characteristics of Trajectory groups

Baseline characteristics and chi-square test were checked for both depression trajectory groups and anxiety groups in GBDTM. From Table 4.20, anxiety was significantly associated with depression. Compared to other anxiety trajectory groups, "low-flat" anxiety trajectory group (DA1) has the highest percentage of the subjects (96.7%) belong to “low-flat” depression trajectory group (DD1); the “low-to-middle” anxiety trajectory group (DA2) had the highest rate (14.2%) of “low-to-middle” depression trajectory group (DD2); the “high-to-low” anxiety group (DA3) had the highest percentage (22.1%) of the “high-curve” depression trajectory group (DD4). Sex, age group, smoking status, alcohol consumption, homeownership, frequency walking, with more than three chronic diseases and involvement in income-generating activity were also found to have a significant association with the depression trajectory groups (Table 4.20).

Table 4.20: Distribution of baseline characteristics by depression trajectory groups from GBDTM (N, %)

	Trajectory groups				
	Low-flat (DD1) N=3641	Low-middle (DD2) N=205	Low-high (DD3) N=33	High-curve (DD4) N=104	p-value
Anxiety					
Low-flat (DA1)	3519 (96.7)	162 (79.0)	28 (84.9)	76 (73.1)	<0.0001
Low-middle (DA2)	61 (1.7)	29 (14.2)	3 (9.1)	3 (2.9)	
High-low (DA3)	57 (1.6)	7 (3.4)	2 (6.1)	23 (22.1)	
High-curve (DA4)	4 (0.11)	7 (3.4)	0 (0.0)	2 (1.9)	
Sex					
Male	1619 (44.5)	59 (28.8)	9 (27.3)	27 (26.0)	<0.0001
Female	2022 (55.5)	146 (71.2)	24 (72.7)	77 (74.0)	
Age					
65-69	1405 (38.6)	86 (42.0)	11 (33.3)	32 (30.8)	0.009
70-74	1099 (30.2)	73 (35.6)	11 (33.3)	32 (30.8)	
75-79	666 (18.3)	37 (18.1)	8 (24.2)	29 (27.9)	
≥ 80	471 (12.9)	9 (4.4)	3 (9.1)	11 (10.6)	
Marriage status					
Married	1272 (35.0)	70 (34.2)	12 (36.4)	36 (34.6)	0.993
Single/divorce/widower	2368 (65.0)	135 (65.8)	21 (63.6)	68 (65.4)	
Education					
None	728 (20.0)	39 (19.0)	8 (24.2)	24 (23.1)	0.308
Elementary	1551 (42.6)	101 (49.3)	13 (39.4)	40 (38.5)	
Middle/High	1079 (29.6)	55 (26.8)	10 (30.3)	37 (35.6)	
University	283 (7.8)	10 (4.9)	2 (6.1)	3 (2.9)	
Smoking					
No	1982 (58.9)	143 (69.8)	22 (66.7)	71 (74.0)	<0.0001
Previous	442 (13.1)	12 (5.9)	2 (6.1)	11 (11.5)	
Current	944 (28.0)	50 (24.4)	9 (27.3)	14 (14.6)	
Drinking					
No	1323 (39.3)	101 (49.3)	19 (57.6)	52 (54.2)	0.002
< 2 days/week	1507 (44.7)	84 (41.0)	12 (36.4)	37 (38.5)	
2-4 days/week	266 (7.9)	10 (4.9)	0 (0.0)	2 (2.1)	
Almost daily	274 (8.1)	10 (4.9)	2 (6.1)	5 (5.2)	
Residential area					
No Metro-city	2249 (61.8)	130 (63.4)	19 (57.6)	65 (62.5)	0.920
Metro-city	1392 (38.2)	75 (36.6)	14 (42.4)	39 (37.5)	

	Trajectory groups				
	Low-flat (DD1) N=3641	Low-middle (DD2) N=205	Low-high (DD3) N=33	High-curve (DD4) N=104	p-value
Housing					
Detached House	2094 (57.5)	119 (58.1)	17 (51.5)	56 (53.9)	0.770
Apartment	512 (14.1)	22 (10.7)	5 (15.2)	17 (16.4)	
Others	1035 (28.4)	64 (31.2)	11 (33.3)	31 (29.8)	
Home ownership					
Own	2790 (76.6)	159 (77.6)	27 (81.8)	66 (63.5)	0.015
Lease	851 (23.4)	46 (22.4)	6 (18.2)	38 (36.5)	
Living					
Alone	59 (1.48)	1 (0.5)	0 (0.0)	2 (1.9)	0.137
Couple only	2231 (61.3)	145 (70.7)	21 (63.6)	69 (66.4)	
Others	1351 (37.1)	59 (28.8)	12 (36.4)	33 (31.7)	
Disability					
No	2931 (80.5)	161 (78.5)	26 (79.8)	73 (70.2)	0.068
Yes	710 (19.5)	44 (21.5)	7 (21.2)	31 (29.8)	
Walking					
None	625 (18.6)	33 (16.1)	5 (15.2)	27 (28.1)	0.044
≤ 3days/week	417 (12.4)	34 (16.6)	6 (18.2)	16 (16.7)	
>3 days/week	2328 (69.1)	138 (67.3)	22 (66.7)	53 (55.2)	
Medium/Intensive Physical activity					
none	2247 (66.7)	134 (65.4)	27 (81.8)	74 (77.1)	0.181
≤ 3days/week	326 (9.7)	21 (10.2)	3 (9.1)	7 (7.3)	
>3 days/week	797 (23.7)	50 (24.4)	3 (9.1)	15 (15.6)	
More than 3 chronic diseases					
No	466 (12.8)	7 (3.4)	1 (3.0)	2 (1.9)	<0.0001
Yes	3175 (87.2)	198 (96.6)	32 (97.0)	102 (98.1)	
Economic Activity					
No	2278(62.6)	135 (65.9)	27 (81.8)	82 (78.9)	<0.0001
Yes	1363 (37.4)	70 (34.1)	6 (18.2)	22 (21.1)	
Income quantile					
< 20	1164 (44.1)	91 (50.8)	15 (45.5)	39 (55.7)	0.100
20 - 40	603 (22.8)	39 (21.8)	9 (27.3)	15 (21.4)	
40 - 60	425 (16.1)	32 (17.9)	6 (18.2)	8 (11.4)	
60 - 80	238 (9.0)	12 (6.7)	3 (9.1)	7 (10.0)	
80 - 100	210 (8.0)	5 (2.8)	0 (0.0)	1 (1.4)	

On the other hand, only sex, smoking status, homeownership, having more than three chronic diseases and involvement in economic activities had a significant association with the anxiety trajectory groups (Table 4.21).

Table 4.21: Distribution of baseline characteristics by anxiety trajectory groups from GBDTM (N, %)

	Trajectory groups				
	Low-flat (DA1) N=3785	Low-middle (DA2) N=96	High-low (DA3) N=89	High-curve (DA4) N=13	p-value
Depression					
Low-flat (DD1)	3519 (93.0)	61 (63.5)	57 (64.0)	4 (30.8)	<0.0001
Low-middle (DD2)	162 (4.3)	29 (30.2)	7 (7.9)	6 (46.2)	
Low-high (DD3)	28 (0.7)	3 (3.1)	2 (2.2)	0 (0.0)	
High-curve (DD4)	76 (2.0)	3 (3.1)	23 (25.8)	2 (15.4)	
Sex					
Male	1664 (44.0)	26 (27.1)	20 (22.5)	4 (30.8)	<0.0001
Female	2121 (56.0)	70 (72.9)	69 (77.5)	9 (69.2)	
Age					
65-69	1454 (38.4)	39 (40.6)	37 (41.6)	4 (30.8)	0.995
70-74	1156 (30.5)	29 (30.2)	26 (29.2)	4 (30.8)	
75-79	702 (18.6)	19 (19.8)	16 (18.0)	3 (23.1)	
≥ 80	473 (12.5)	9 (9.4)	10 (11.2)	2 (15.4)	
Marriage status					
Married	2475 (65.4)	59 (61.5)	49 (55.1)	9 (69.2)	0.190
Single/divorce/widower	1309 (34.6)	37 (38.5)	40 (44.9)	4 (30.8)	
Education					
None	753 (19.9)	20 (20.8)	24 (27.0)	2 (15.4)	0.194
Elementary	1611 (42.6)	50 (52.1)	35 (39.3)	9 (69.2)	
Middle/High	1133 (29.9)	22 (22.9)	25 (28.1)	1 (7.7)	
University	288 (7.6)	4 (4.2)	5 (5.6)	1 (7.7)	
Smoking					
No	2083 (59.4)	65 (68.4)	62 (72.1)	8 (61.5)	0.042
Previous	451 (12.9)	6 (6.3)	10 (11.6)	0 (0)	
Current	974 (27.8)	24 (25.3)	14 (16.3)	5 (38.5)	
Drinking					
No	1408 (40.1)	37 (39.0)	44 (51.2)	6 (46.2)	0.433
< 2 days/week	1554 (44.3)	45 (47.4)	35 (40.7)	6 (46.2)	
2-4 days/week	269 (7.7)	4 (4.2)	4 (4.7)	1 (7.7)	
Almost daily	279 (8.0)	9 (9.5)	3 (3.5)	0 (0.0)	
Residential area					
No Metro-city	2338 (61.8)	60 (62.5)	57 (64.0)	8 (61.5)	0.976
Metro-city	1447 (38.2)	36 (37.5)	32 (36.0)	5 (38.5)	
Housing					
Detached House	2173 (57.4)	59 (61.5)	47 (52.8)	7 (53.9)	0.811
Apartment	528 (14.0)	11 (11.5)	16 (18.0)	1 (7.7)	
Others	1084 (28.6)	26 (27.1)	26 (29.2)	5 (38.5)	
Home ownership					
Own	2898 (76.6)	78 (81.2)	59 (66.3)	7 (53.9)	0.019
Lease	887 (23.4)	18 (18.8)	30 (33.7)	6 (46.1)	
Living					
Alone	61 (1.6)	1 (1.0)	0 (0)	0 (0)	0.631
Couple only	2336 (61.7)	67 (69.8)	55 (61.8)	8 (61.5)	
Others	1388 (36.7)	28 (29.2)	34 (38.2)	5 (38.5)	
Disability					
No	3034 (80.2)	79 (82.3)	65 (73.0)	13 (100.0)	0.097
Yes	751 (19.8)	17 (17.7)	24 (27.0)	0 (0.0)	
Walking					
None	650 (18.5)	15 (15.8)	22 (25.6)	3 (23.1)	0.562
≤ 3days/week	453 (12.9)	11 (11.6)	7 (8.1)	2 (15.4)	
>3 days/week	2407 (68.6)	69 (72.6)	57 (66.3)	8 (61.5)	

	Trajectory groups				
	Low-flat (DA1) N=3785	Low-middle (DA2) N=96	High-low (DA3) N=89	High-curve (DA4) N=13	p-value
Medium/Intensive Physical activity					
none	2349 (66.9)	59 (62.1)	63 (73.3)	11 (84.6)	0.493
≤ 3days/week	340 (9.7)	9 (27)	8 (9.3)	0 (0.0)	
>3 days/week	821 (23.4)	27 (28.4)	15 (17.4)	2 (15.4)	
More than 3 chronic diseases					
No	466 (12.3)	8 (8.3)	2 (2.3)	0 (0.0)	0.001
Yes	3319 (87.7)	88 (91.7)	87 (97.7)	13 (100)	
Economic Activity					
No	2382 (62.9)	61 (63.5)	70 (78.7)	9 (69.3)	0.204
Yes	1403 (37.1)	35 (36.5)	19 (21.4)	4 (30.7)	
Income quantile					
< 20	1224 (44.4)	44 (51.2)	31 (44.9)	31 (44.9)	0.410
20 - 40	628 (22.8)	19 (22.1)	18 (26.1)	18 (26.1)	
40 - 60	446 (16.2)	11 (12.8)	13 (18.8)	13 (18.8)	
60 - 80	246 (8.9)	9 (10.5)	5 (7.3)	5 (7.3)	
80 - 100	210 (7.6)	3 (3.5)	2 (2.9)	2 (2.9)	

Table 4.22 showed the univariate logistic regression for depression in GBDTM. With "low-flat" depression trajectory (DD1) as the reference group,"low-to-middle" anxiety trajectory (DA2),"high-curve" anxiety trajectory (DA4), females, 80 years old or older, smoking, drinking, living with two or three generations in the household, and being in the 80 - 100 income quartile were significant predictors for "low-to-middle" depression trajectory (DD2); "low-to-middle" anxiety trajectory (DA2), "high-to-low" anxiety trajectory (DA3) and involving in income-generating activities were significant predictors for "low-to-high" depression trajectory (DD3); "high-to-low" anxiety trajectory (DA3), "high-curve" anxiety trajectory group (DA4), females, being between 75 and 79 years old, smoking, drinking, living in a rental house, having disability, walking more than 10 minutes per day, having more than three chronic diseases and involving in income-generating activities were significant predictors for "high-curve" depression trajectory (DD4).

Table 4.22: Univariate Logistic Regression Analysis of Depression GBDTM. Estimation of odds ratio (OR) and 95% confidence interval (C.I). Low-flat anxiety as the reference group.

Variable	**Low-to-Middle (n=205)**		**Low-to-High (n=33)**		**High-Curve (n=104)**	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
Anxiety						
Low-flat (DA1)	-	-	-	-	-	-
Low-middle (DA2)	10.3 (6.46-16.5)	<0.0001	6.18 (1.83-20.9)	0.003	2.28 (0.70-7.42)	0.172
High-low (DA3)	2.67 (1.20-5.94)	0.016	4.41 (1.03-19.0)	0.046	18.7 (10.9-31.9)	<0.0001
High-curve (DA4)	38.0 (11.0-131)	<0.0001	0.01 (0.01-999)	0.991	23.2 (4.18-128.)	<0.0001

Variable	**Low-to-Middle (n=205)**		**Low-to- High (n=33)**		**High-Curve (n=104)**	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
Sex						
Male	-	-	-	-	-	-
Female	1.98 (1.45-2.70)	<0.0001	2.14 (0.99-4.61)	0.053	2.28 (1.47-3.56)	<0.0001
Age						
65-69	-	-	-	-	-	-
70-74	1.09 (0.79-1.50)	0.619	1.28 (0.55-2.96)	0.566	1.28 (0.78-2.10)	0.332
75-79	0.91 (0.61-1.35)	0.632	1.53 (0.61-3.83)	0.359	1.91 (1.15-3.19)	0.013
≥ 80	0.31 (0.16-0.63)	0.001	0.81 (0.23-2.93)	0.752	1.03 (0.51-2.05)	0.943
Marriage status						
Married	-	-	-	-	-	-
Single/divorce/widower	0.97 (0.72-1.30)	0.816	1.06 (0.52-2.17)	0.865	0.99 (0.65-1.49)	0.945
Education						
None	-	-	-	-	-	-
Elementary	1.22 (0.83-1.78)	0.314	0.76 (0.32-1.85)	0.548	0.78 (0.47-1.31)	0.349
Middle/High	0.95 (0.63-1.45)	0.817	0.84 (0.33-2.15)	0.721	1.04 (0.62-1.75)	0.883
University	0.66 (0.33-1.34)	0.250	0.64 (0.14-3.05)	0.578	0.32 (0.10-1.08)	0.066
Smoking						
No	-	-	-	-	-	-
Previous	0.38 (0.21-0.68)	0.001	0.41 (0.10-1.74)	0.226	0.70 (0.37-1.32)	0.267
Current	0.73 (0.53-1.02)	0.067	0.86 (0.39-1.87)	0.702	0.41 (0.23-0.74)	0.003
Drinking						
No	-	-	-	-	-	-
< 2 days/week	0.73 (0.54-0.98)	0.039	0.55 (0.27-1.15)	0.112	0.63 (0.41-0.96)	0.031
2-4 days/week	0.49 (0.25-0.96)	0.036	0.28 (0.04-2.08)	0.974	0.19 (0.05-0.79)	0.022
Almost daily	0.48 (0.25-0.93)	0.029	0.51 (0.12-2.20)	0.365	0.46 (0.18-1.17)	0.105
Residential area						
No Metro-city	-	-	-	-	-	-
Metro-city	0.93 (0.70-1.25)	0.637	1.19 (0.60-2.38)	0.622	0.97 (0.65-1.45)	0.880
Housing						
Detached House	-	-	-	-	-	-
Apartment	0.76 (0.48-1.20)	0.239	1.20 (0.44-3.28)	0.718	1.24 (0.72-2.16)	0.442
Others	1.09 (0.80-1.49)	0.597	1.31 (0.61-2.81)	0.488	1.12 (0.72-1.75)	0.618
Home ownership						
Own	-	-	-	-	-	-
Lease	0.95 (0.68-1.33)	0.758	0.73 (0.30-1.77)	0.485	1.89 (1.26-2.83)	0.002
Living						
Alone	-	-	-	-	-	-
Couple only	0.26 (0.04-1.90)	0.184	1.89 (0.25-14.32)	0.989	1.10 (0.26-4.58)	0.900
Others	0.67 (0.49-0.92)	0.012	0.94 (0.46-1.92)	0.873	0.79 (0.52-1.20)	0.271
Disability						
No	-	-	-	-	-	-
Yes	1.13 (0.80-1.59)	0.491	1.11 (0.48-2.57)	0.805	1.75 (1.14-2.69)	0.010
Walking						
None	-	-	-	-	-	-
≤ 3days/week	1.54 (0.94-2.53)	0.085	1.80 (0.55-5.93)	0.335	0.89 (0.47-1.67)	0.713
>3 days/week	1.12 (0.76-1.66)	0.561	1.18 (0.45-3.13)	0.738	0.53 (0.33-0.85)	0.008
Medium/Intensive Physical activity						
none	-	-	-	-	-	-
≤ 3days/week	1.08 (0.67-1.74)	0.750	0.77 (0.23-2.54)	0.663	0.65 (0.30-1.43)	0.285
>3 days/week	1.05 (0.75-1.47)	0.767	0.31 (0.10-1.04)	0.057	0.57 (0.33-1.00)	0.051

Variable	Low-to-Middle (n=205)		Low-to- High (n=33)		High-Curve (n=104)	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
More than 3 chronic diseases						
No	-	-	-	-	-	-
Yes	4.15 (1.94-8.88)	<0.0001	4.70 (0.64-34.45)	0.128	7.49 (1.84-30.4)	0.005
Economic Activity						
Yes	-	-	-	-	-	-
No	1.15 (0.86-1.55)	0.344	2.69 (1.11-6.54)	0.029	2.23 (1.39-3.59)	0.001
Income quantile						
< 20	-	-	-	-	-	-
20 - 40	0.83 (0.56-1.22)	0.338	1.16 (0.50-2.66)	0.729	0.74 (0.41-1.36)	0.334
40 - 60	0.96 (0.63-1.46)	0.860	1.10 (0.42-2.84)	0.851	0.56 (0.26-1.21)	0.142
60 - 80	0.65 (0.35-1.20)	0.164	0.98 (0.28-3.41)	0.972	0.88 (0.39-1.99)	0.754
80 - 100	0.31 (0.12-0.76)	0.011	0.40 (0.05-3.03)	0.975	0.14 (0.02-1.04)	0.055

Table 4.23 showed the multivariate logistic regression for depression in GBDTM. Compared "low-to-high" depression trajectory (DD3) to the "low-flat" depression trajectory (DD1), the Individuals were less likely to be involved in income-generating activity (OR = 2.79, 95% CI: 1.15 – 6.80, p-value = 0.024), with anxiety adjusted. Compared to the "low-flat" depression group (DD1), after controlling for anxiety, individuals in the "low-to-middle" depression group (DD2) had a higher chance of being females (OR = 1.51, 95% CI: 1.07 – 2.12, p-value = 0.018) and having more than three chronic diseases (OR = 3.96, 95% CI: 1.83 – 8.59, p-value = 0.0005). Compared to the "low-flat" depression group (DD1),"high-curve" depression group (DD4) indicated greater chance of being female (OR = 1.87, 95% CI: 1.02

Table 4.23: Multivariate Logistic Regression Analysis of Depression GBDTM. Estimation of odds ratio (OR) and 95% confidence interval (C.I). Low-flat depression as the reference group

Variable	Low-to-Middle (n=205)		Low-to-High (n=33)		High-Curve (n=104)	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
Anxiety						
Low-flat	-	-	-	-	-	-
Low-middle	9.26 (5.63-15.2)	<0.0001	5.80 (1.71-19.7)	0.005	2.67 (0.80-8.95)	0.111
High-low	2.19 (0.96-4.98)	0.061	3.64 (0.84-15.7)	0.084	10.7 (5.30-21.4)	<0.0001
High-curve	25.9 (7.45-90.1)	<0.0001	0.01 (0.01-999)	0.992	17.3 (3.04-98.7)	<0.0001
Sex						
Male	-	-			-	-
Female	1.51 (1.07-2.12)	0.018			1.87 (1.02-3.41)	0.042
More than 3 chronic diseases						
No	-	-			-	-
Yes	4.15 (1.93-8.93)	<0.0001			5.18 (1.26-21.3)	0.023
Economic Activity						
Yes			-	-	-	-
No			2.79 (1.15-6.80)	0.024	1.91 (1.02-3.59)	0.044

– 3.14, p-value = 0.042), having more than three chronic diseases (OR = 4.18, 95% CI: 1.01 – 17.31, p-value = 0.049) and being less likely to be involved in an income-generating activity (OR = 1.91, 95% CI: 1.02 – 3.59, p-value = 0.044), with anxiety adjusted.

In the univariate logistic regression analysis for anxiety trajectory groups in GBDTM (Table 4.24), "low-to-middle" depression trajectory group (DD2), "low-to-high" depression trajectory group (DD3), female sex and smoking were significant for the "low-to-middle" anxiety trajectory group (DA2) compared to the "low-flat" anxiety trajectory group (DA1). The "high-to-low" anxiety trajectory group (DA3) compared to the "low-flat" anxiety trajectory group (DA1), "low-to-middle" depression trajectory group (DD2), "low-to-high" depression trajectory group (DD3), "high-curve" depression trajectory group (DD4), female sex, marriage status, homeownership, having more than three chronic diseases and involving in income-generating activities were significant variables. As GBTM with an anxiety outcome, except depression trajectory groups, no other significant factors were found when comparing the "high-curve" anxiety trajectory group (DA4) to the "low-flat" anxiety trajectory group (DA1).

Table 4.24: Univariate Logistic Regression Analysis of Anxiety GBDTM. Estimation of odds ratio (OR) and 95% confidence interval (C.I). Low-flat anxiety as the reference group.

Variable	**Low-to-Middle (n=96)**		**High-to-Low (n=89)**		**High-Curve (n=13)**	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
Depression						
Low-flat (DD1)	-	-	-	-	-	-
Low-middle (DD2)	10.3 (6.46-16.5)	<0.0001	2.67 (1.20-5.94)	0.016	38.0 (11.0-131)	0.001
Low-high (DD3)	6.18 (1.83-20.9)	0.003	4.41 (1.03-19.0)	0.046	0.01 (0.01-999)	0.991
High-curve (DD4)	2.28 (0.70-7.42)	0.172	18.7 (10.9-31.90)	<0.0001	23.2 (4.18-128)	0.001
Sex						
Male	-	-	-	-	-	-
Female	2.11 (1.34-3.33)	0.001	2.71 (1.64-4.47)	<0.0001	1.77 (0.54-5.74)	0.345
Age						
65-69	-	-	-	-	-	-
70-74	0.94 (0.58-1.52)	0.788	0.88 (0.53-1.47)	0.634	1.26 (0.31-5.04)	0.746
75-79	1.01 (0.58-1.76)	0.975	0.90 (0.50-1.62)	0.716	1.55 (0.35-6.96)	0.565
≥ 80	0.71 (0.34-1.48)	0.359	0.83 (0.41-1.68)	0.607	1.54 (0.28-8.42)	0.620
Marriage status						
Married	-	-	-	-	-	-
Single/divorce/widower	1.19 (0.78-1.80)	0.423	1.54 (1.01-2.36)	0.044	0.84 (0.26-2.73)	0.773

Variable	**Low-to-Middle (n=96)**		**High-to-Low (n=89)**		**High-Curve (n=13)**	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
Education						
None	-	-	-	-	-	-
Elementary	1.17 (0.69-1.98)	0.562	0.68 (0.40-1.15)	0.154	2.10 (0.45-9.76)	0.342
Middle/High	0.73 (0.40-1.35)	0.316	0.69 (0.39-1.22)	0.204	0.33 (0.03-3.67)	0.369
University	0.52 (0.18-1.54)	0.240	0.55 (0.21-1.44)	0.221	1.31 (0.12-14.5)	0.827
Smoking						
No	-	-	-	-	-	-
Previous	0.43 (0.18-0.99)	0.047	0.75 (0.38-1.46)	0.393	0.01 (0.01-999)	0.969
Current	0.79 (0.49-1.27)	0.329	0.48 (0.27-0.87)	0.015	1.34 (0.44-4.10)	0.612
Drinking						
No	-	-	-	-	-	-
< 2 days/week	1.10 (0.71-1.71)	0.666	0.72 (0.46-1.13)	0.154	0.91 (0.29-2.82)	0.865
2-4 days/week	0.57 (0.20-1.60)	0.283	0.48 (0.17-1.34)	0.158	0.87 (0.11-7.28)	0.900
Almost daily	1.23 (0.59-2.57)	0.587	0.34 (0.11-1.12)	0.076	0.01 (0.01-999)	0.976
Residential area						
No Metro-city	-	-	-	-	-	-
Metro-city	0.97 (0.64-1.47)	0.885	0.91 (0.59-1.41)	0.663	1.01 (0.33-3.09)	0.986
Housing						
Detached House	-	-	-	-	-	-
Apartment	0.77 (0.40-1.47)	0.425	1.40 (0.79-2.49)	0.250	0.59 (0.07-4.79)	0.620
Others	0.88 (0.55-1.41)	0.603	1.10 (0.68-1.80)	0.676	1.43 (0.45-4.52)	0.541
Home ownership						
Own	-	-	-	-	-	-
Lease	0.75 (0.45-1.27)	0.286	1.66 (1.06-2.60)	0.026	2.80 (0.94-8.35)	0.065
Living						
Alone	-	-	-	-	-	-
Couple only	0.57 (0.08-4.19)	0.582	0.01 (0.01-999)	0.982	0.01 (0.01-999)	0.989
Others	0.70 (0.45-1.10)	0.122	1.04 (0.68-1.60)	0.858	1.05 (0.34-3.22)	0.929
Disability						
No	-	-	-	-	-	-
Yes	0.87 (0.51-1.48)	0.605	1.49 (0.93-2.40)	0.099	0.01 (0.01-999)	0.970
Walking						
None	-	-	-	-	-	-
≤ 3days/week	1.05 (0.48-2.31)	0.899	0.46 (0.19-1.08)	0.074	0.96 (0.16-5.75)	0.961
>3 days/week	1.24 (0.71-2.19)	0.452	0.70 (0.43-1.15)	0.161	0.72 (0.19-2.72)	0.628
Medium/Intensive Physical activity						
none	-	-	-	-	-	-
≤ 3days/week	1.05 (0.52-2.15)	0.885	0.88 (0.42-1.85)	0.730	0.01 (0.01-999)	0.973
>3 days/week	1.31 (0.83-2.09)	0.253	0.68 (0.39-1.20)	0.186	0.52 (0.12-2.35)	0.396
More than 3 chronic diseases						
No	-	-	-	-	-	-
Yes	1.54 (0.74-3.21)	0.243	6.10 (1.50-24.87)	0.012	999 (0.01-999)	0.970
Economic Activity						
Yes	-	-	-	-	-	-
No	1.03 (0.67-1.56)	0.903	2.17 (1.30-3.62)	0.003	1.33 (0.41-4.31)	0.640
Income quantile						
< 20	-	-	-	-	-	-
20 - 40	0.84 (0.49-1.45)	0.545	1.13 (0.63-2.04)	0.680	0.20 (0.03-1.53)	0.119
40 - 60	0.69 (0.35-1.34)	0.274	1.15 (0.60-2.22)	0.675	0.27 (0.04-2.15)	0.218
60 - 80	1.02 (0.49-2.11)	0.961	0.80 (0.31-2.08)	0.651	0.01 (0.01-999)	0.973
80 - 100	0.40 (0.12-1.29)	0.135	0.38 (0.09-1.58)	0.183	0.58 (0.07-4.58)	0.607

In the multivariate logistic regression analysis for anxiety trajectory in GBDTM (Table 4.25), adjusted by depression, sex was a significant predictor in the “low-to-middle” anxiety trajectory group (DA2) (OR = 1.73, 95% CI: 1.09 – 2.76, p-value = 0.021) and the “high-to-low” anxiety group (DA3) (OR = 2.17, 95% CI: 1.28 – 3.69, p-value = 0.025); involvement in income-generating activity was another predictor in the “high-to-low” (DA3) anxiety group (OR=2.17, 95% CI: 1.28 – 3.69, p-value=0.025) with “low-flat” anxiety trajectory group (DA1) as reference group.

Table 4.25: Multivariate Logistic Regression Analysis of Anxiety GBDTM. Estimation of odds ratio (OR) and 95% confidence interval (C.I). Low-flat anxiety as the reference group

Variable	**Low-to-Middle (n=96)**		**High-to-Low (n=89)**		**High-Curve (n=13)**	
	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value	O.R (95% C.I)	p-value
Depression						
Low-flat	-	-	-	-	-	-
Low-middle	8.98 (5.59-14.4)	<0.0001	2.21 (0.99-4.94)	0.054	38.0 (11.0-131)	0.001
Low-high	5.38 (1.59-18.3)	0.007	3.43 (0.79-14.9)	0.100	0.01 (0.01-999)	0.991
High-curve	2.10 (0.64-6.89)	0.220	14.8 (8.45-26.0)	<0.0001	23.2 (4.18-128)	0.001
Sex						
Male	-	-	-	-		
Female	1.73 (1.09-2.76)	0.021	2.17 (1.28-3.69)	0.025		
Economic Activity						
Yes			-	-		
No			1.86 (1.08-3.18)	0.025		

4.5.4 Risk factors for depression and anxiety using GBDTM

Before adding risk factors into GBDTM, we checked the feature of the GBDTM without any risk factors. I switched the position of depression or anxiety as the first outcome, and the parameter estimates were unchanged in the trajectory polynomial functions and trajectory group memberships for depression and anxiety outcomes. Since the risk factors only influenced the first outcome’s proportions in GBDTM, I built the GBDTM twice to add the risk factors for depression and anxiety in turn.

4.5.4.1 GBDTM with depression as the first outcome and anxiety as the second outcome

The significant covariates from the multivariate logistic regression analysis in Table 4.23 were considered risk factors for depression: female sex, involved in income-generating activities and having more than three chronic diseases. Similar to the trajectories adjusted by risk factors in the GBTMs, the polynomial trajectory shapes remained unchanged, but the group memberships showed some variation (Figure 4.13). Compared the depression trajectory group memberships of GBDTM without risk factors (Figure 4.10), the model with risk factors decreased 3.1% in the "low-flat" depression trajectory group (DD1) but increased 2.9% and 0.3% in the "low-to-middle" depression trajectory (DD2) and the "low-to-high" depression trajectory (DD3), respectively. The "high-curve" depression trajectory (DD4) did not have any variation.

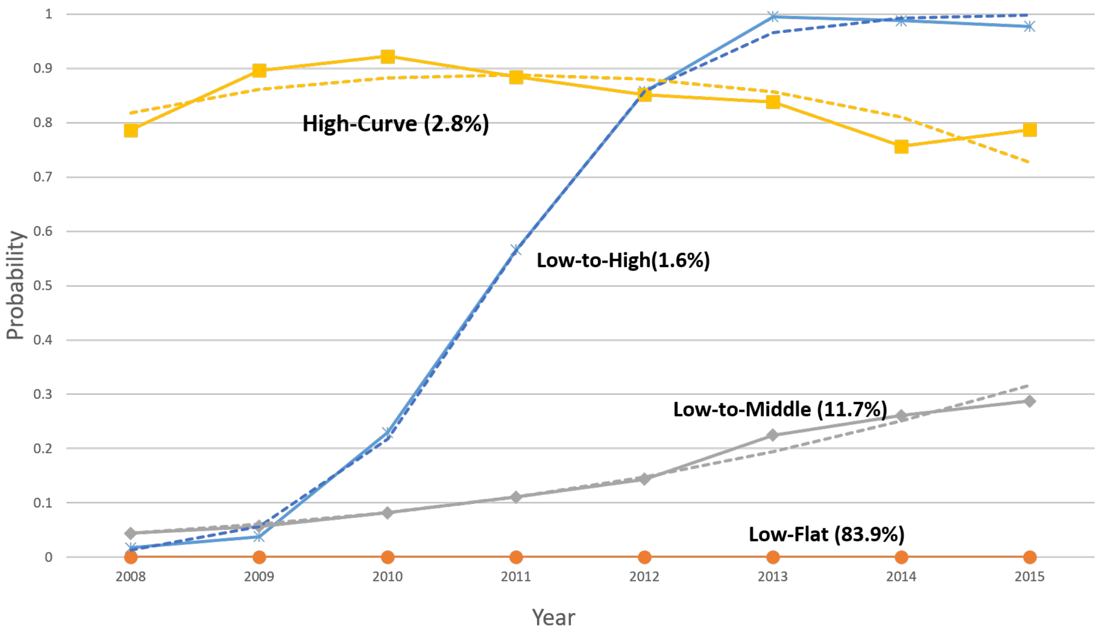


Figure 4.13: Depression trajectories adjusted by risk factors in GBDTM. The solid line indicates observed averages; the dashed line indicates predictions.

Table 4.26 showed the estimated parameters for risk factors influencing the depression trajectory memberships of GBDTM. Using DD1 as the reference group, intercepts were all significant in DD2 - DD4. Being female was a significant risk factor in DD2 and DD4. Not being involved in income-generating activity affected DD4. Having more than three chronic diseases was significant in DD2 and DD4.

Table 4.26: Parameter estimates for risk factors by depression trajectory groups in GBDTM

Group	Parameter	Estimate	Standard Error	p-value
Low-flat (DD1)	Baseline	0	-	-
Low-to-middle (DD2)	Constant	-4.57194	0.47620	<0.0001
	Female	0.64408	0.17465	0.0002
	No economy activity	0.17202	0.16732	0.3039
	> 3 chronic disease	1.55354	0.39780	0.0001
Low-to-high (DD3)	Constant	-6.95896	1.29033	<0.0001
	Female	0.70036	0.43210	0.1051
	No economy activity	0.78273	0.46199	0.0902
	> 3 chronic disease	1.40041	1.06897	0.1902
High-curve (DD4)	Constant	-7.37688	1.09993	<0.0001
	Female	0.83829	0.25005	0.0008
	No economy activity	0.68433	0.26056	0.0086
	> 3 chronic disease	2.21581	1.00571	0.0276

Figure 4.14 showed the effect of the risk factors for five situations (no risk factors, female only, Without income-generating activities only, more than three chronic diseases only, and all risk factors together). The risk factors can affect the proportion of depression in GBDTM.

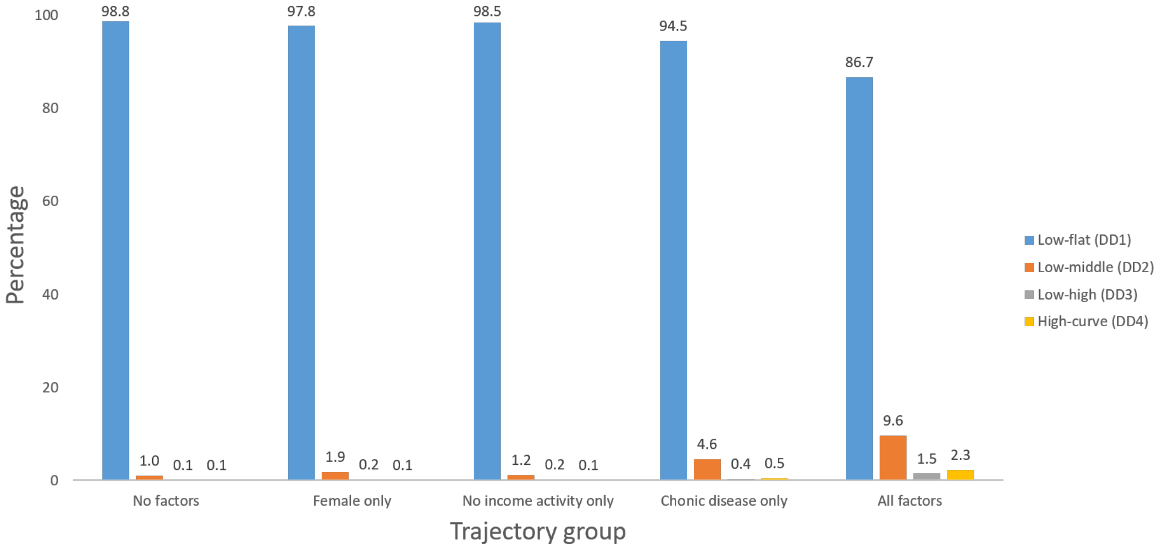


Figure 4.14: Bar plot of risk factor effects on depression group membership in GBDTM

Comparing the subjects without any risk factors, the percentage for subjects with all risk factors in TD1 decreased 12.1% and increased 8.6% in DD2, 1.4% in DD3 and 2.2% in DD4.

4.5.4.2 GBDTM with anxiety as the first outcome and depression as the second outcome

This time, GBDTM was developed with the variables of the multivariate logistic regression in Table 4.25 as risk factors that influence only anxiety parameters. After adding the risk factors into the anxiety proportions, the GBDTM was observed in Figure 4.15. Compared to the original GBDTM (Figure 4.11), the "low-flat" anxiety trajectory group (DA1) moved 1.1% and 0.7% to the "low-to-middle" anxiety trajectory group (DA2) and the "high-to-low" anxiety trajectory group (DA3). No noticeable variation was found among the "high-curve" trajectory group (DA4)and the polynomial trajectory shapes.

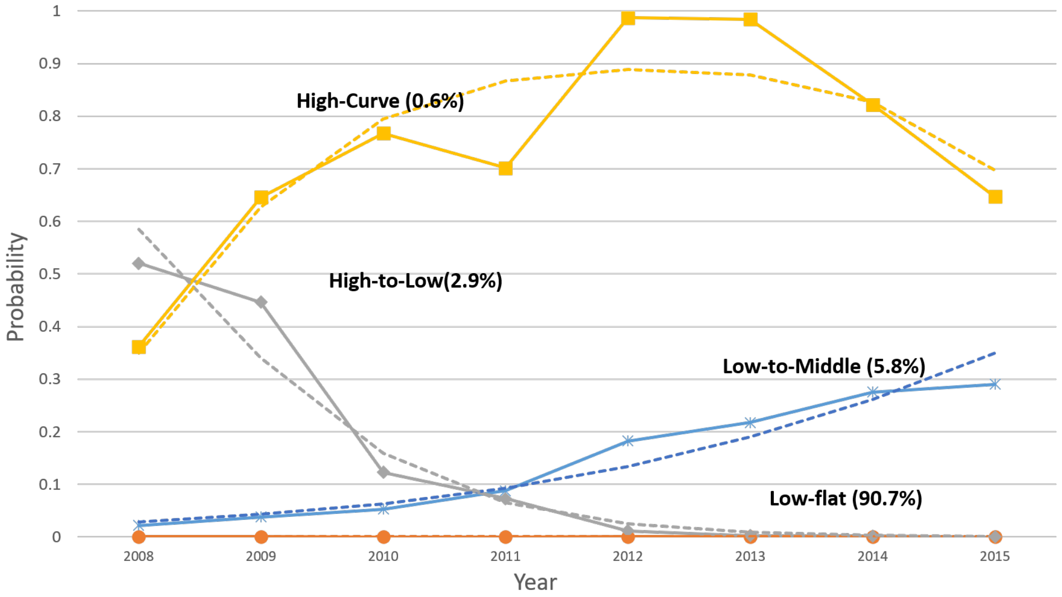


Figure 4.15: Anxiety trajectories adjusted by risk factors in GBDTM. The solid line indicates observed averages; the dashed line indicates predictions.

Being female and not being involved in income-generating activities were obtained as risk factors. Estimations of the parameters were presented in Table 4.27. Based on the baseline DA1, all the intercepts were significant for DA2 - DA4. Being female had substantial effects on DA2 and DA3. Not being involved in income-generating activities had a significant impact on DA3.

Table 4.27: Parameter estimates for risk factors by anxiety trajectory groups in GBDTM

Group	Parameter	Estimate	Standard Error	p-value
Low-flat (DA1)	Baseline	0	-	-
Low-to-middle (DA2)	Constant	-4.36378	0.56637	<0.0001
	Female	0.93107	0.28311	0.0010
	No economy activity	0.11121	0.25923	0.6679
High-to-low (DA3)	Constant	-6.18236	0.77332	<0.0001
	Female	1.29700	0.37606	0.0006
	No economy activity	0.72090	0.32192	0.0251
High-curve (DA4)	Constant	-5.44936	1.00924	<0.0001
	Female	0.28646	0.60292	0.6347
	No economy activity	0.05996	0.05996	0.9227

Figure 4.16 presented four distinct situations in which risk factors affected the proportion of the anxiety trajectory groups (no risk factors, female only, not involved in income-generating activities only, and both risk factors combined). When comparing the trajectory proportions for no risk factors to all risk factors, the “low-flat” anxiety trajectory group (DA1) declined 4.6%. However, the “low-to-middle” (DA2), “high-to-low” (DA3), and “high-curve” (DA4) anxiety trajectory group increased 2.4%, 1.3%, and 0.2%, respectively.

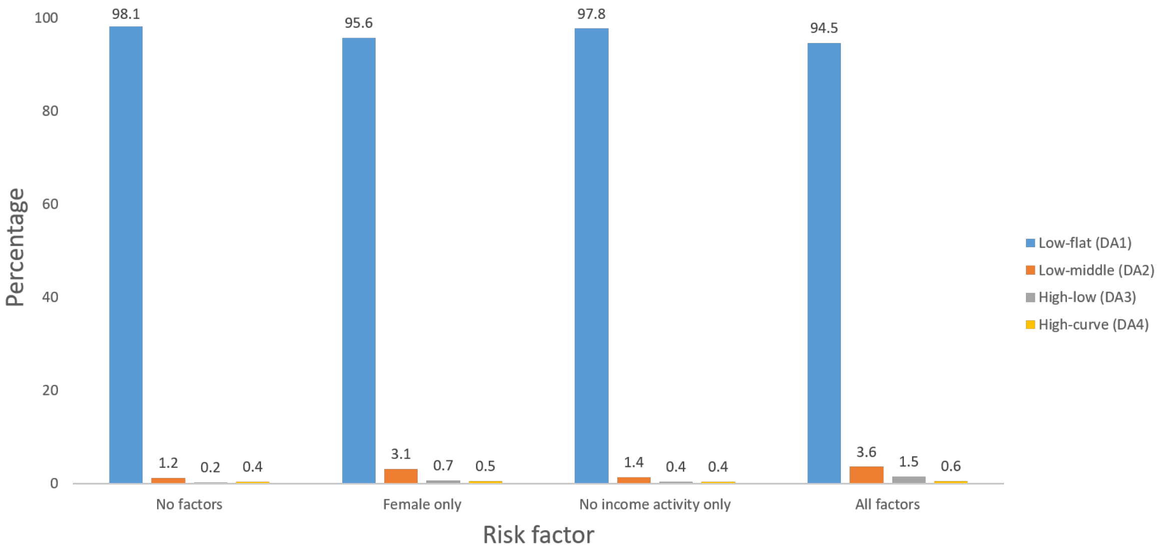


Figure 4.16: Bar plot of risk factor effects on anxiety group membership in GBDTM

4.6 Analysis of group-based multi-trajectory modeling (GBMTM)

As discussed in Section 3.5.2, group-based multi-trajectory modeling (GBMTM) could include two or more outcomes in the model simultaneously. However, the restriction for GBMTM was that the outcomes shared the same proportion of the group memberships. To make it more convenient to compare GBTMs and GBDTM, GBMTM was developed by including only two outcomes (i.e., depression and anxiety), which was the same as the constrained model of GBDTM in Equation (3.79). After the GBMTM developed, trajectory shape, group membership, and their relationship to the outcomes of depression (Figure 4.17) and anxiety (Figure 4.18) were identified. The "low-flat" depression trajectory group (MD1) and "low-flat" anxiety trajectory group (MA1) shared the same trajectory proportion with 86.9%. The "low-to-middle" depression trajectory group (MD2) and "low-to-middle" anxiety trajectory group (MA2) had the same trajectory proportion with 7.9%. The "low-mild" depression trajectory group (MD3) and the "high-to-low" anxiety trajectory group (MA3)

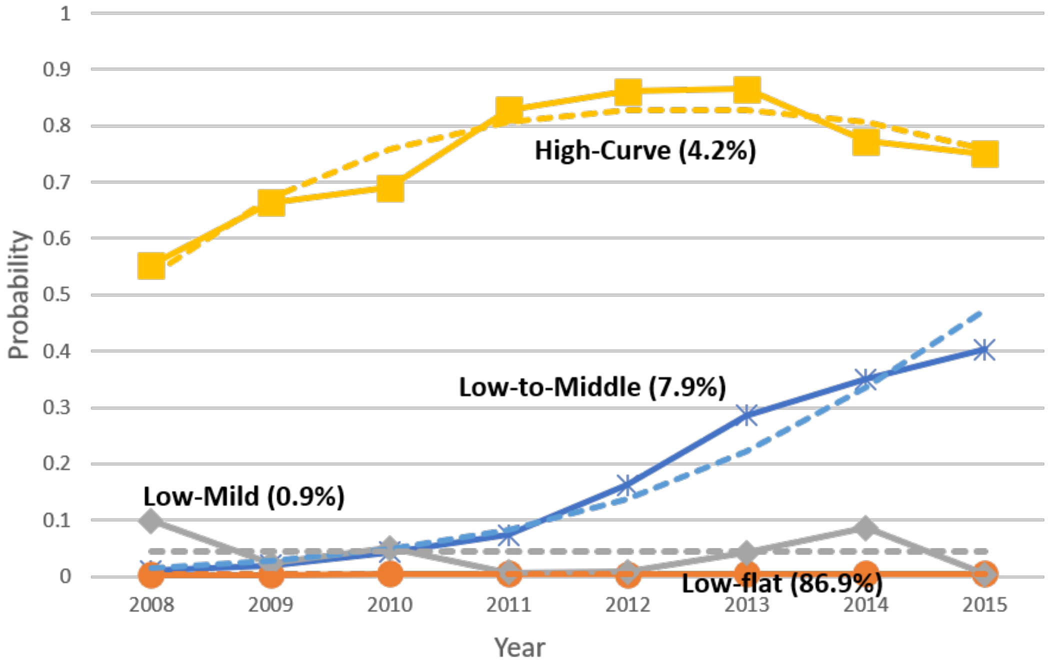


Figure 4.17: Depression trajectories for GBMTM. The solid line indicates the observed averages; the dot line the dashed line indicates the predictions.

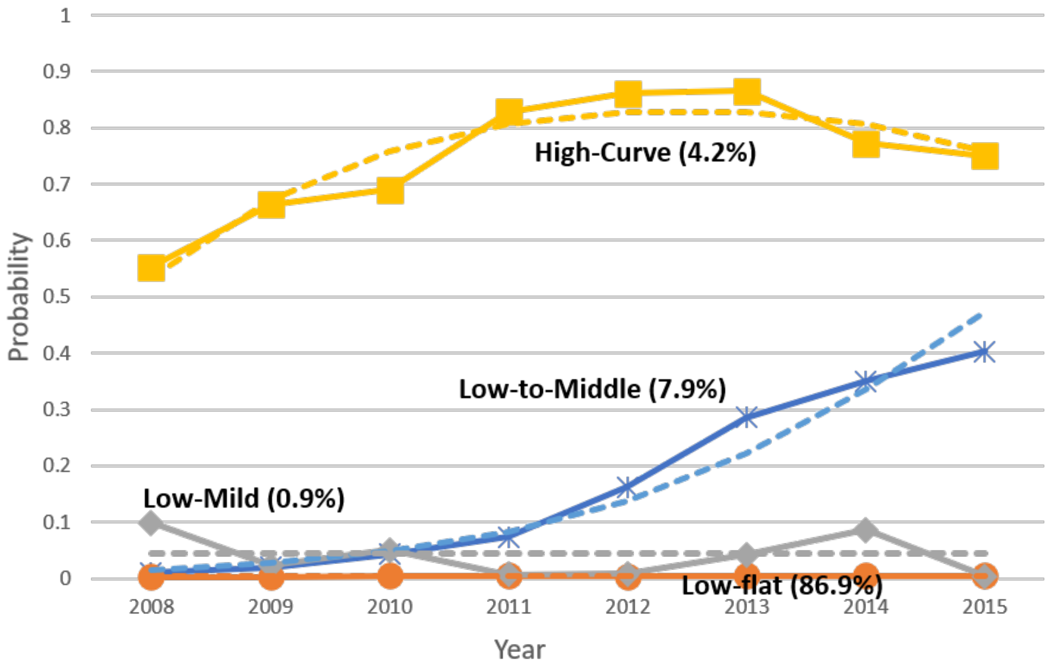


Figure 4.18: Anxiety trajectories for GBMTM. The solid line indicates the observed averages; the dot line the dashed line indicates the predictions.

shared the same trajectory proportion with 0.9%. The "high-curve" depression trajectory group (MD4) and the "low-mild" anxiety trajectory group (MA4) shared the same trajectory proportion with 4.2%.

The result indicated that individuals with a high probability of depression in MD4 had a "low-mild" possibility of anxiety over time (MA4). Individuals with a high probability of anxiety at the beginning that declined over time in DA3 had a "low-mild" depression probability. The "low-mild" trajectory group in depression (MD3) and anxiety (MA4) were close to the "low-flat" trajectories (MD1 and MA1), which could be replaceable. Compared the trajectory shapes to GBTMs or GBDTM, the "low-to-high" depression trajectory and "high-curve" anxiety trajectory disappeared in GBMTM. In general, GBMTM was applied to outcomes that shared similar proportions. For example, individuals with a high probability of depression should generally have high anxiety levels. However, this was not consistent with our analysis of GBMTM. There were two possible reasons for this. One was that the correlation between the two outcomes in each time measurement was too low. In the KHPS data, the correlations between depression and anxiety in older people at the time measurements were between 0.05 to 0.16 (correlations: 0.162, 0.154, 0.081, 0.053, 0.061, 0.104, 0.087 and 0.136 from 2008 to 2015, respectively). Another reason was that the distributed clusters of the two outcomes were different, which meant that the polynomial trajectory shape was different for outcomes sharing the same proportion. Because of these problems, GBMTM was not able to identify the "low-to-high" depression trajectory and "high-curve" anxiety trajectory. The "low-to-high" depression trajectory and "high-curve" anxiety trajectory represented the highest probability of having depression and anxiety, were vitally important, even though they only involved a small portion of the individuals from the overall population.

Chapter 5 SIMULATION STUDY

5.1 Introduction

The simulation study was performed to accomplish Study Objective 3. This objective was to examine the characteristics of three group-based trajectory models and to select the best one based on repeated measurements of two binary outcomes. These two outcomes were assumed to be associated with one another. The level of association between two outcomes

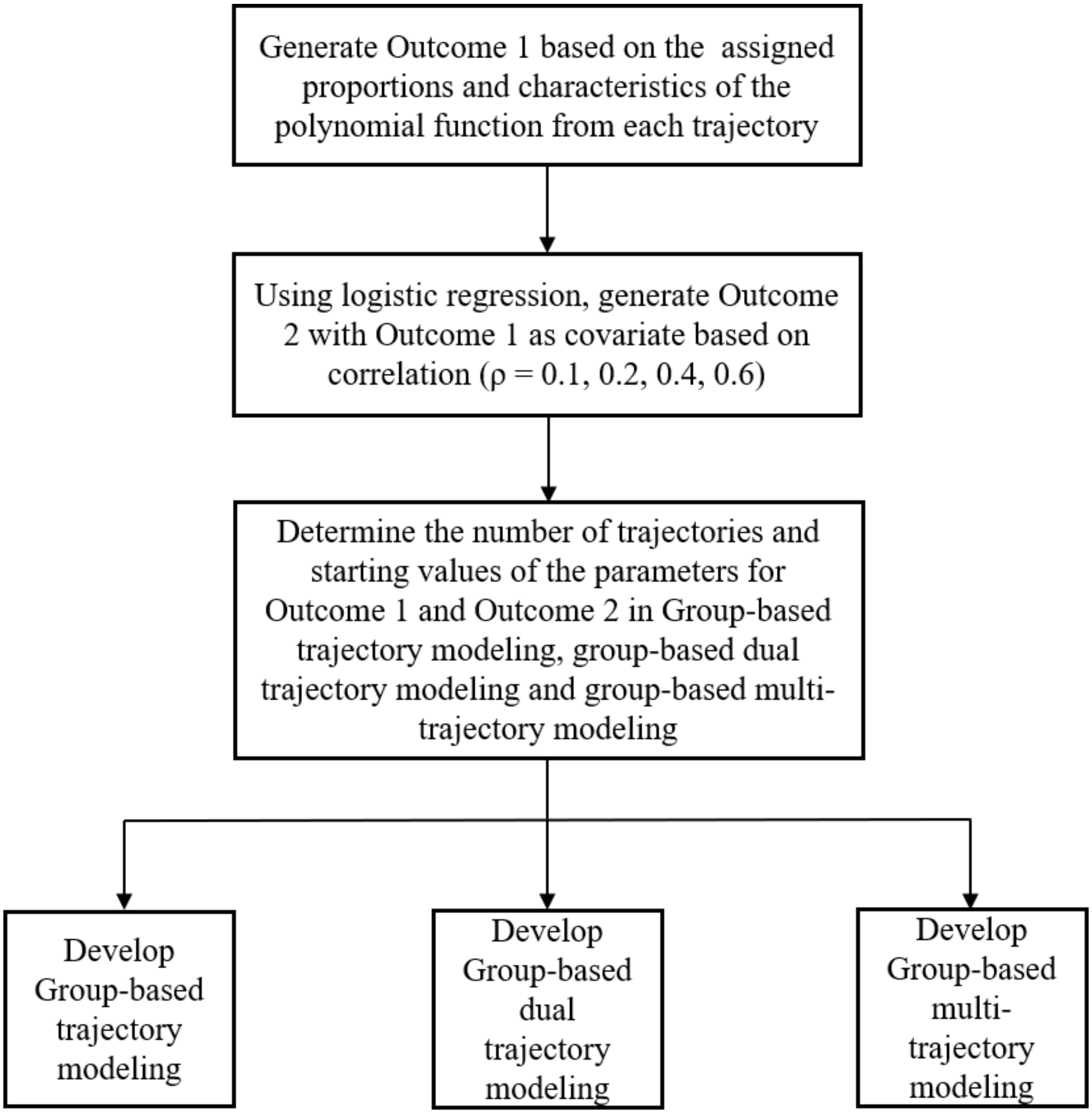

Generate Outcome 1 based on the assigned proportions and characteristics of the polynomial function from each trajectory
Using logistic regression, generate Outcome 2 with Outcome 1 as covariate based on correlation (ρ = 0.1, 0.2, 0.4, 0.6)
Determine the number of trajectories and starting values of the parameters for Outcome 1 and Outcome 2 in Group-based trajectory modeling, group-based dual trajectory modeling and group-based multi-trajectory modeling
Develop Group-based trajectory modeling
Develop Group-based dual trajectory modeling
Develop Group-based multi-trajectory modeling

Figure 5.1: Flow Chart of Simulation

was defined using selected correlation coefficient levels. In this simulation study, we generated Outcome 1 based on assigned trajectory group memberships and their trajectory shapes. Then, Outcome 2 was generated based on Outcome 1 with various correlation coefficients. After that, GBTM, GBDTM, and GBMTM were fitted from these simulated datasets. The levels of correlation coefficient between Outcome 1 and Outcome 2 were chosen to be $\rho =$ 0.1, 0.2, 0.4, and 0.6. The sample sizes N = 500, 2000, and 4000 were used in each scenario. A total of 500 simulations were performed for each scenario. The flow chart (Figure 5.1) showed the procedure used for data generation and analysis in this simulation.

5.2 Data-generation for Simulation Study

5.2.1 Generation of Outcome 1

Repeated measurements of continuous outcomes were generated based on polynomial trajectory group characteristics from GBTM (Haviland, Jones, & Nagin, 2011). In our simulation, the process of generating repeated measurement Outcome 1 was similar to Haviland's study (Haviland et al., 2011). The first step in generating Outcome 1 was to determine the number of trajectory groups. To mimic the real data shown in Chapter 4, the GBTM included four groups for depression and anxiety, so we assigned four trajectory groups for Outcome 1. The second step was to decide the proportion of each trajectory group. In the real data analysis of Chapter 4, the portions of event groups for depression and anxiety might be too small. For example, the anxiety trajectory group (TA4) in GBTM only included individuals with 0.6% (n=13) shown in Section 4.5.1 (High-curve anxiety trajectory group (DA2)). Our simulation study also used a smaller sample size with N=500. If the group proportion was arranged too low, it was possible that no individuals would be identified from the event trajectories. Therefore, the proportions of Group 1, 2, 3, and 4 were assigned with $\pi_1 = 60\%$, $\pi_2 = 20\%$, $\pi_3 = 10\%$, and $\pi_4 = 10\%$, respectively. After that, Outcome 1's trajectory shape in each group was generated based on the polynomial trajectory group characteristics from GBTM. In the real data shown in Section 4.5.1 (Development of

GBDTM), each trajectory group followed a polynomial regression with binary outcomes for depression and anxiety. The logit link function η_{itj} followed by polynomial regression was denoted as:

$$\eta_{itj} = ln\frac{p_{itj}}{1 - p_{itj}} \tag{5.1}$$

where p_{itj} is the probability of Outcome 1 equal to one as the event (for example, depression); $i = 1 \ \ldots \ N$ is the number of study subjects; $t = 1 \ \ldots \ T$ is the number of repeated measurements; and $j = 1 \ \ldots \ J$ is the number of trajectory groups. Four groups' logit link functions were applied to determine the shape of the trajectories from Outcome 1 with equations:

$$\begin{aligned} Group\,1 \ &: \eta_{it1} = -4.5; \ \ \pi_1 = 0.6 \\ Group\,2 \ &: \eta_{it2} = -4 + t; \ \ \pi_2 = 0.2 \\ Group\,3 \ &: \eta_{it3} = 3.5 - t; \ \ \pi_3 = 0.1 \\ Group\,4 \ &: \eta_{it4} = 4; \ \ \pi_4 = 0.1 \\ t &= 1, 2, \ldots, 5 \end{aligned} \tag{5.2}$$

where η_{itj} (j = 1, 2, 3, 4) were the simulated value of the polynomial link functions with individual i at time t for Group 1 to Group 4, respectively. π_j stood for the probability of group membership in Group j. The property of trajectory shapes in each group was:

Group 1: Constant polynomial logit function assigned with negative constant, which represented the constant non-event trajectory group.

Group 2: An increased linear polynomial logit function presented an increased probability of event trajectory group.

Group 3: A linear polynomial logit function that declined over time stood for the decreased probability of event trajectory group.

Group 4: Constant polynomial logit function assigned with positive constant, which represented the constant high-event trajectory group.

Then, using the logistic transformation of η_{it}, which was also called the transformation of inverse logit function in each group, the probability p_{itj} could be generated as:

$$p_{itj} = \frac{e^{\eta_{itj}}}{1 + e^{\eta_{itj}}} \tag{5.3}$$

where p_{itj} was the probability of the outcome y_{itj} with the Bernoulli random variable taking the value, $P(y_{itj} = 1)$. Finally, the variable of repeated measurement of Outcome 1 could be generated using logistic regression as a Bernoulli variable, $y_{itj} \sim Bernoulli(p_{itj})$ (Wicklin, 2013).

5.2.2 Generation of Outcome 2

Outcome 2 was generated based on Outcome 1 obtained in Section 5.2.1. It was generated with different levels of correlations between the two outcomes. To simulate one binary outcome relevant to another with correlation coefficient (ρ), the logistic regression method was used (le Cessie & Van Houwelingen, 1994; Ocram, 2014; Touloumis, 2016). The correlation level of the two outcomes was selected with $\rho = 0.1$, 0.2, 0.4, and 0.6, respectively. The structure of generating Outcome 2 was described in Figure 5.2, where Y_t was Outcome 1 at time t $(t = 1, 2, 3, 4, 5)$ and Z_t was Outcome 2 at time t $(t = 1, 2, 3, 4, 5)$. i.e.; Z_t (Outcome 2)

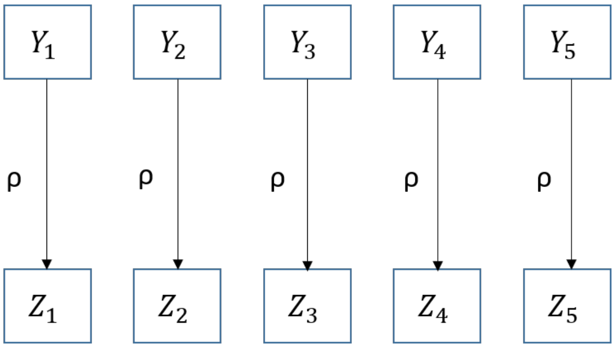


Figure 5.2: Structure of generating Outcome 2 ($\rho = 0.1$, 0.2, 0.4, 0.6)

was generated based on Y_t (Outcome 1) with the correlation coefficient closed to the assumed ρ (Figure 5.2). Based on the logistic regression, Z_t was assumed to be a response variable with Y_t as a covariate:

$$log\frac{P(Z_t = 1)}{1 - P(Z_t = 1)} = \beta_0 + \beta_1 * Y_t \tag{5.4}$$

Y_t was known as Outcome 1, but the parameters β_0 and β_1 were unknown (equation (5.4)). To make sure Outcome 1 (Y_t) and Outcome 2 (Z_t) have a certain correlation ρ in each simulation, we needed to try different values for β_0 and β_1. For example, we wanted to find the value of β_0 and β_1 to ensure the correlation between Outcome 1 (Y_t) and Outcome 2 (Z_t) was close to $\rho = 0.1$. To make sure Y_t and Z_t had the correlation close to an assumed $\rho = 0.1$, we first generated five preliminary simulated datasets for Y_t with sample size N = 4000. Then, using these preliminary simulated datasets of Y_t as a covariate, we tried different values of β_0 and β_1 in equation (5.4) to simulate a corresponding Z_t. When each measure of the preliminary simulated Z_t had a similar correlation $\rho = 0.1$ to each measure of Y_t, we stopped trying β_0 and β_1. Thus, the last-tried values of β_0 and β_1 were used in all 500 full simulations to guarantee the simulated Y_t and Z_t had a correlation level of $\rho = 0.1$. The same method was applied with a correlation level of $\rho = 0.2$, 0.4, and 0.6 between Y_t and Z_t. Note that our data simulation methods were is an adaptation of Haviland (2011), Ceossi & Houwelingen (1994), Ocram, (2014) and Touloumis (2016). With this method, we also got four groups of five preliminary simulated datasets of Z_t that had a close correlation level to Y_t. These simulated data were also used as test datasets to define the number of trajectory groups and parameters' initial values for Z_t in the simulation of GBTM and GBDTM.

5.2.3 Determining trajectory groups and initial values of parameters

After generating two outcomes, the next step was to determine the number of trajectory groups and the initial values of parameters for the two outcomes for all 500 simulation datasets.

For Outcome 1 (Y_t), the data were generated with four trajectory groups from the polynomial logistic regression with two constant and two linear. The total numbers of trajectory groups were four (equation (5.2)). The initial value of parameters was from the assigned values of the polynomial functions and the proportion of the memberships into all trajectory models (GBTM, GBDTM, and GBMTM) in the simulation.

For Outcome 2 (Z_t), we selected the number of trajectory groups and the starting values of parameters based on the test datasets (five preliminary simulated datasets of both Y_t and Z_t) in Section 5.2.2. Initial values of the parameters were set to 0 with Outcome 2 in GBTM. The proportion parameters were equally distributed. i.e., when Z_t had two trajectory groups, the proportions were assumed to be 0.5 and 0.5, respectively. To select the number of trajectory groups for Z_t, GBTM and GBDTM were based on the largest BIC in the test datasets. The initial values of parameters for Z_t were also defined by using the test dataset from GBTM and GBDTM. In GBMTM, Z_t was forced with four trajectory groups because it must have the same number of trajectory groups between two outcomes. The initial values of the polynomial parameters were all assigned as 0 in Z_t in GBMTM.

5.2.4 Analysis of simulations

After the number of trajectories and initial values had been determined, the data sets were generated for Y_t and Z_t in each correlation level (ρ) with sample sizes N = 500, 2000, and 4000. Using these simulated datasets, GBTM, GBDTM and GBMTM modeling were developed as follows:

(i) GBTM was first fitted for Y_t and Z_t. Since Y_t was generated based on polynomial trajectory group characteristics, the simulation result would only be influenced by sample size, but not by correlation level.

(ii) Z_t was generated based on a different level of correlation (ρ) from Y_t; the GBTM with Z_t had a different result for each sample size and correlation level.

(iii) GBDTM and GBMTM were fitted.

(iv) The results for Y_t and Z_t in GBDTM and GBMTM were different depending on the different correlation and sample size. This was because Z_t was impacted by Y_t in these two models.

For each case of correlation and sample size, a total of 500 simulations were performed. The process of the simulation is referenced by the SAS code in Appendix B.

5.3 Results of Simulations

5.3.1 Correlation between two outcomes

After simulating two longitudinal binary outcomes, the average correlation between the Y_t and Z_t at each measurement time was calculated, as seen in Table 5.1. The first two columns presented the assumed true correlation and sample size. The other five columns presented the average correlation for each measurement from the 500 simulations. The mean of correlation was close to the assumed correlation value in the first column. For example, when Y_t and Z_t with ρ =0.1, N = 2000, and measurement time t = 2, we have the average

Table 5.1: Average correlation between two outcomes at each time measurement based on 500 simulated datasets

		Time of measurement (t)				
Correlation (ρ)	Sample Size (N)	1	2	3	4	5
0.1	500	0.1033558	0.100965	0.095602	0.097801	0.101853
	2000	0.10077	0.094226	0.097238	0.102331	0.102057
	4000	0.1016155	0.096502	0.094763	0.100588	0.103589
0.2	500	0.1956968	0.200326	0.196183	0.206194	0.208386
	2000	0.1931838	0.200822	0.20011	0.202665	0.212142
	4000	0.1918279	0.201174	0.20077	0.203193	0.212975
0.4	500	0.3979156	0.395982	0.39863	0.408899	0.413967
	2000	0.3966869	0.396343	0.400705	0.405932	0.416253
	4000	0.3978632	0.399391	0.401439	0.405484	0.414918
0.6	500	0.6033358	0.589772	0.577329	0.60474	0.594062
	2000	0.604765	0.591468	0.580312	0.604033	0.596811
	4000	0.6045834	0.591625	0.579853	0.604802	0.597901

correlation = 0.094226. This was close to the assumed correlation ρ =0.1. Different sample sizes rarely affected the value of the mean correlation.

5.3.2 Estimates of parameters for Outcome 1 using GBTM

Three trajectory models were fitted for each simulation for two outcomes: (i) Group-based trajectory modeling (GBTM) of each outcome; (ii) Group-based dual trajectory modeling (GBDTM) with both outcomes; (iii) Group-based multi-trajectory modeling (GBMTM) with both outcomes.

Using different correlation coefficients had no effect on Y_t in GBTM. The average parameter estimates from each polynomial trajectory were presented in three situations with varying sample sizes in Table 5.2. In GBTM, each trajectory followed a polynomial function. Since Y_t was assumed to be a longitudinal binary outcome in the simulation, each path followed

Table 5.2: Estimation of parameters of Outcome 1 on each polynomial trajectory in group-based trajectory modeling (GBTM) based on 500 simulated datasets

N	Parameters	True parameter value	Mean estimates	Mean SE*	Bias	p-value
500	Intercept1	-4.5	-5.22295	141.6547	-0.72295	0.971
	Intercept2	-4	-4.12479	0.593149	-0.12479	<0.0001
	linear2	1	1.028998	0.170404	0.028998	<0.0001
	Intercept3	3.5	3.729137	0.859459	0.229137	<0.0001
	linear3	-1	-1.07216	0.249433	-0.07216	<0.0001
	Intercept4	4	7.144587	1193.6	3.144587	0.995
2000	Intercept1	-4.5	-4.51585	0.266121	-0.01585	<0.0001
	Intercept2	-4	-4.03443	0.287699	-0.03443	<0.0001
	linear2	1	1.010456	0.082816	0.010456	<0.0001
	Intercept3	3.5	3.517332	0.395721	0.017332	<0.0001
	linear3	-1	-1.00306	0.114991	-0.00306	<0.0001
	Intercept4	4	4.645703	55.4216	0.645703	0.933
4000	Intercept1	-4.5	-4.51383	0.183494	-0.01383	<0.0001
	Intercept2	-4	-4.01492	0.200848	-0.01492	<0.0001
	linear2	1	1.004075	0.057852	0.004075	<0.0001
	Intercept3	3.5	3.520556	0.278816	0.020556	<0.0001
	linear3	-1	-1.00488	0.081164	-0.00488	<0.0001
	Intercept4	4	4.067581	0.444766	0.067581	<0.0001

[*] SE = Standard Error
Note: p-values are calculated based on the average mean and SE

a logistic regression with time as the covariate. When the data sample size was small, and there was a rarity of events or no events in the dataset, the maximum likelihood estimation of the logistic model was commonly biased (King & Zeng, 2001). This is called quasi-complete separation. Even if quasi-complete separation did not occur, separation might be nearly complete, so the standard error for a parameter estimate can become very large (Vassiliadis, Spyroglou, Rigas, Rosenberg, & Lindsay, 2019). Therefore, the biased estimate and large standard error of GBTM from sample sizes 500 and 2000 was caused by the small sample size and rare event or no event cases in the subgroup of the trajectories. Thus, the size of biases and influence of the trajectory shape were limited. Figures 5.3 presented the trajectory of Y_t in GBTM with N = 4000.

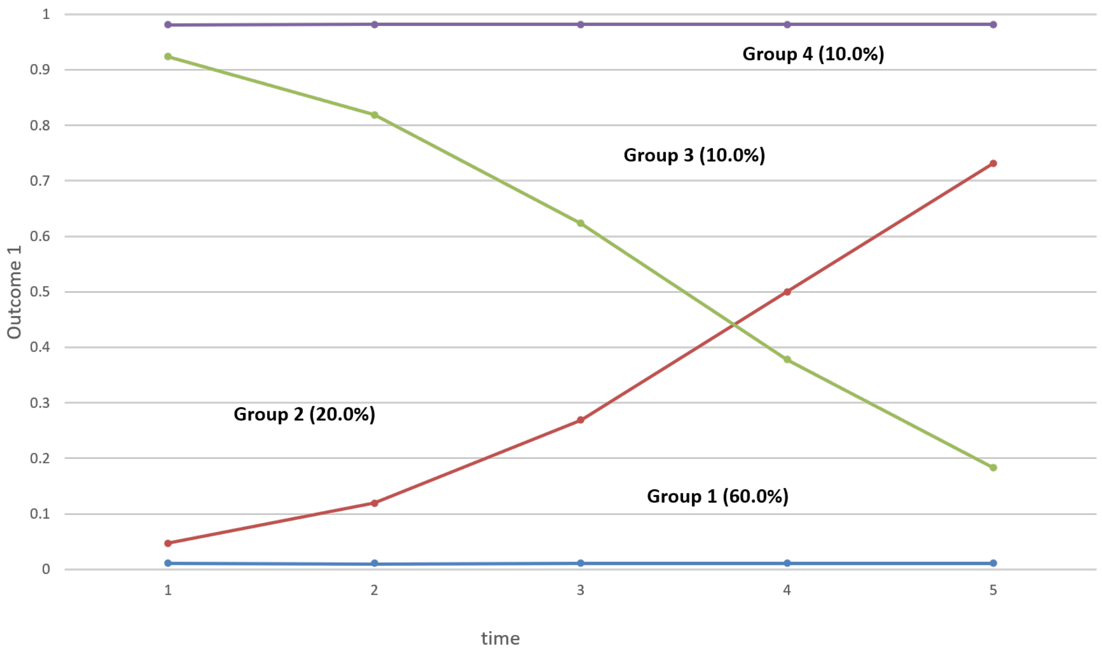


Figure 5.3: Simulation trajectory shape of Outcome 1 in GBTM with N = 4000

Table 5.2 presented the mean estimates and standard error of the parameters from the polynomial functions with Y_t in GBTM. The bias was calculated based on the difference between the actual parameter values and the mean estimates. Therefore, only the sample size (N) showed the impact of variation among parameters. The p-values were used to check whether the average parameter estimates from the polynomial functions were significant or not. The key findings included:

- The biases of most estimates for the parameters was small. However, three rows of parameters had large bias values because of quasi-complete separation problems.

- Beside the three p-values with quasi-complete separation problems, the p-value of the polynomial functions' parameters were all highly significant (p-value < 0.0001) because Outcome 1 was defined from the assumed polynomial functions.

- The average standard error got smaller for corresponding parameters as the sample size increased. The large standard errors from quasi-complete separation problems improved as the sample size increased.

5.3.3 Estimates of parameters for Outcome 1 using GBDTM and GBMTM

GBDTM and GBMTM were constructed with two outcomes jointly. Thus, Y_t and Z_t affected by one another. Table 5.3 presented the estimates and standard error (SE) of parameters of Y_t from GBDTM and GBMTM. The estimated value of the parameters was generated separately for Outcome 1 based on four different correlation levels ($\rho = 0.1$, 0.2, 0.4, 0.6) between the two outcomes with N = 4000 (The results for N = 500 and 2000 are shown in Appendix A).

Three key findings should be emphasized:

- The standard errors were small; additionally, the p-values of parameter estimates for the polynomial function were highly significant in both GBDTM and GBMTM.

- When the correlation between the two outcomes was $\rho = 0.1$, the estimates of the parameters in Y_t were very close to the real parameter value with a small bias in GBDTM and GBMTM.

Table 5.3: Estimation of parameters for Outcome 1 on each polynomial trajectory in group-based dual trajectory modeling (GBDTM) and group-based multi-trajectory modeling (GBMTM) with sample size N = 4000 based on 500 simulated datasets

			GBDTM[$]				GBMTM[&]			
ρ**	Parameter	TPV[#]	Mean Estimates	Mean SE*	Bias	p-value	Mean Estimates	Mean SE*	Bias	p-value
0.1	Intercept1	-4.5	-4.49916	0.17950	0.00084	<0.0001	-4.49260	0.17772	0.00740	<0.0001
	Intercept2	-4	-4.02252	0.20104	-0.02252	<0.0001	-4.03660	0.20026	-0.03660	<0.0001
	Linear2	1	1.00682	0.05786	0.00682	<0.0001	1.01131	0.05757	0.01131	<0.0001
	Intercept3	3.5	3.52204	0.27861	0.02204	<0.0001	3.49891	0.27550	-0.00109	<0.0001
	Linear3	-1	-1.00633	0.08112	-0.00633	<0.0001	-1.00143	0.08013	-0.00143	<0.0001
	Intercept4	4	4.05449	0.46618	0.05449	<0.0001	4.01563	0.43527	0.01563	<0.0001
0.2	Intercept1	-4.5	-4.44575	0.16741	0.05425	<0.0001	-4.44841	0.16518	0.05159	<0.0001
	Intercept2	-4	-4.04900	0.20156	-0.04900	<0.0001	-4.09129	0.19894	-0.09129	<0.0001
	Linear2	1	1.01709	0.05782	0.01709	<0.0001	1.02911	0.05687	0.02911	<0.0001
	Intercept3	3.5	3.53672	0.27877	0.03672	<0.0001	3.45118	0.26781	-0.04882	<0.0001
	Linear3	-1	-1.01317	0.08134	-0.01317	<0.0001	-0.99661	0.07772	0.00339	<0.0001
	Intercept4	4	3.92532	0.37097	-0.07468	<0.0001	3.84505	0.32837	-0.15495	<0.0001
0.4	Intercept1	-4.5	-3.85544	0.10352	0.64456	<0.0001	-4.38056	0.14284	0.11944	<0.0001
	Intercept2	-4	-3.71962	0.20129	0.28038	<0.0001	-4.24662	0.19535	-0.24662	<0.0001
	Linear2	1	0.96062	0.05697	-0.03938	<0.0001	1.07920	0.05506	0.07920	<0.0001
	Intercept3	3.5	2.68244	0.24338	-0.81756	<0.0001	3.28533	0.24357	-0.21467	<0.0001
	Linear3	-1	-0.77011	0.07275	0.22989	<0.0001	-0.97131	0.07091	0.02869	<0.0001
	Intercept4	4	3.57913	0.34763	-0.42087	<0.0001	3.56211	0.24116	-0.43789	<0.0001
0.6	Intercept1	-4.5	-4.21744	0.11093	0.28256	<0.0001	-4.42373	0.13315	0.07627	<0.0001
	Intercept2	-4	-4.39089	0.19340	-0.39089	<0.0001	-4.39257	0.18775	-0.39257	<0.0001
	Linear2	1	1.13255	0.05393	0.13255	<0.0001	1.12288	0.05225	0.12288	<0.0001
	Intercept3	3.5	3.31580	0.23117	-0.18420	<0.0001	3.22811	0.22473	-0.27189	<0.0001
	Linear3	-1	-0.93409	0.07015	0.06591	<0.0001	-0.97784	0.06558	0.02216	<0.0001
	Intercept4	4	4.08380	0.53486	0.08380	<0.0001	3.37417	0.19821	-0.62583	<0.0001

[*] SE = Standard Error
[**] ρ = Correlation Level between Y_t and Z_t
[#] TPV = True Parameter Value
[$] GBDTM = Group-based dual trajectory modeling
[&] GBMTM = Group-based multi-trajectory modeling
Note: p-values are calculated based on the average mean and SE

- As the correlation level increased, the bias between the estimate and true parameter value increased. i.e. as the correlation level increased, Y_t were increasingly adjusted by Z_t.

Figures 5.4 showed the corresponding figures for Table 5.3 for Y_t in GBDTM and GBMTM with $\rho = 0.1$, 0.2, 0.4, and 0.6, respectively.

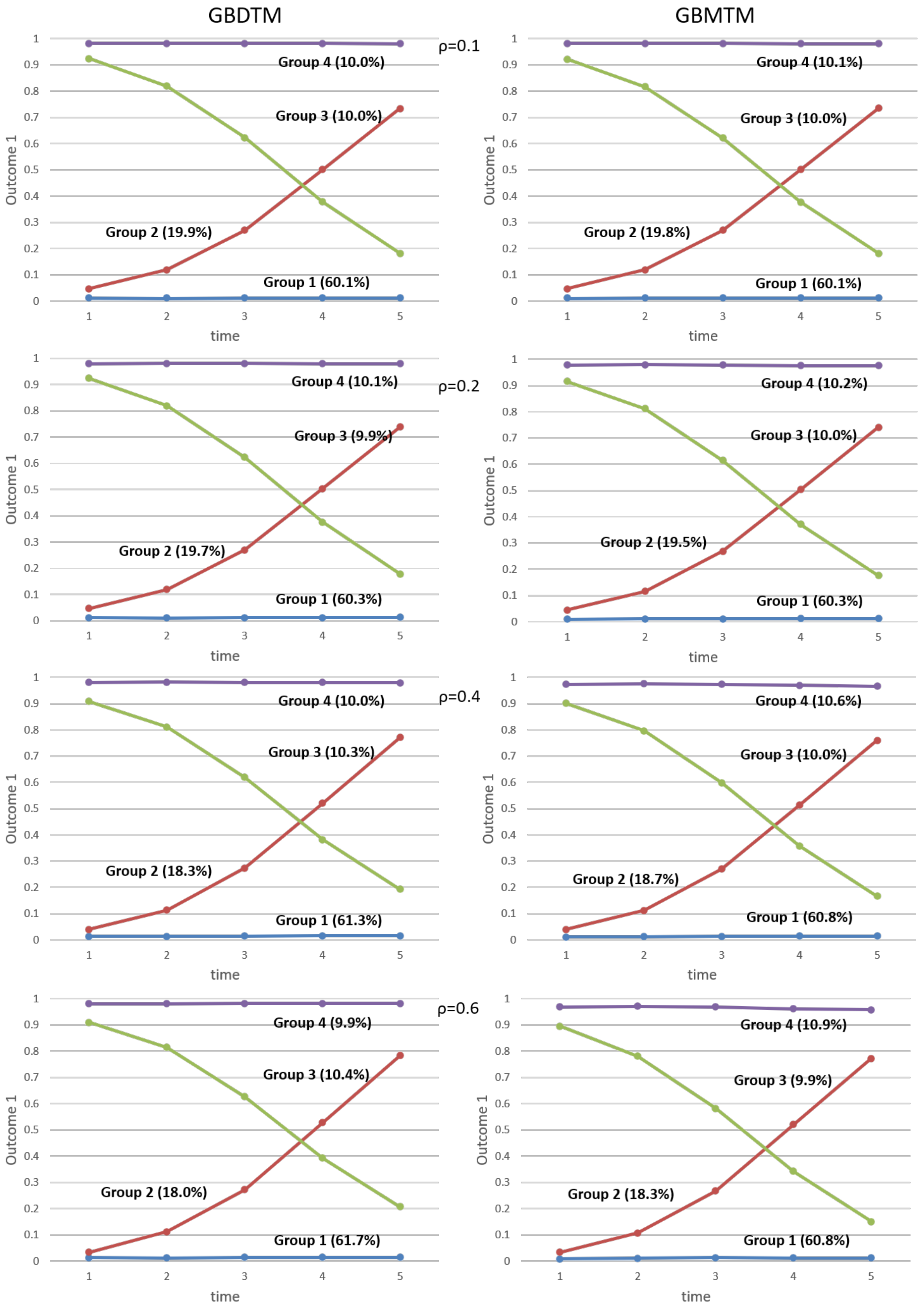


Figure 5.4: Simulation trajectory shapes of Outcome 1 in GBDTM and GBMTM with N = 4000

5.3.4 Estimates of parameters for Outcome 2 using GBTM, GBDTM and GBMTM

In the simulation, GBTM, GBDTM, and GBMTM were also developed for Outcome 2 (Z_t) using different correlation levels ($\rho = 0.1, 0.2, 0.4, 0.6$). Table 5.4 showed a comparison of parameter estimates in Z_t from GBTM, GBDTM, and GBMTM with N = 4000 (The results of N = 500 and 2000 can be seen in the Appendix A). In Table 5.4, the trajectory shapes for the models with different correlation levels were described as follows:

Table 5.4: Estimation of parameters for Outcome 2 on each polynomial trajectory in group-based trajectory modeling (GBTM), group-based dual trajectory modeling (GBDTM) and group-based multi-trajectory modeling (GBMTM) with sample size N = 4000 based on 500 simulated datasets

		GBTM			GBDTM			GBMTM		
ρ**	Parameter	Mean Estimates	Mean SE*	p-value	Mean Estimates	Mean SE*	p-value	Mean Estimates	Mean SE*	p-value
0.1	Intercept1	-9.01866	4.94689	0.068	-2.09866	0.08823	<0.0001	-2.01967	0.07186	<0.0001
	Linear1	-3.62862	1.04587	0.0005	-0.02449	0.02450	0.318	-0.05009	0.02223	0.024
	Intercept2	-0.69828	0.63490	0.272	-1.39155	0.13996	<0.0001	-2.25843	0.13285	<0.0001
	Linear2	-0.63494	0.24823	0.011	-0.04071	0.03828	0.288	0.12150	0.03808	0.0014
	Intercept3							-1.14361	0.15484	<0.0001
	Linear3							-0.20948	0.05159	<0.0001
	Intercept4							-1.41353	0.14154	<0.0001
	Linear4							-0.02817	0.04308	0.513
0.2	Intercept1	-8.45494	1.63698	<0.0001	-2.30075	0.07640	<0.0001	-2.17546	0.07526	<0.0001
	Linear1	1.17997	0.40459	0.004	0.02148	0.02368	0.364	-0.02809	0.02317	0.982
	Intercept2	-1.16258	0.53378	0.029	-1.16281	0.10609	<0.0001	-2.61595	0.13715	<0.0001
	Linear2	-0.03085	0.11648	0.792	0.00978	0.02998	0.744	0.29513	0.03740	<0.0001
	Intercept3							-0.61977	0.14099	<0.0001
	Linear3							-0.29408	0.04868	<0.0001
	Intercept4							-1.07252	0.12449	<0.0001
	Linear4							0.01604	0.03736	0.668
0.4	Intercept1	-0.86235	0.69347	0.214	-2.11837	0.07681	<0.0001	-2.19457	0.07619	<0.0001
	Linear1	-0.65634	0.38263	0.086	-0.04994	0.02428	0.04	-0.03252	0.02376	0.171
	Intercept2	-5.24975	1.41676	0.0002	-2.86950	0.15421	<0.0001	-2.92318	0.13771	<0.0001
	Linear2	0.98014	0.38436	0.011	0.53070	0.04128	<0.0001	0.52949	0.03687	<0.0001
	Intercept3	-0.16206	0.34009	0.634	0.20459	0.10674	0.055	0.49538	0.12861	0.0001
	Linear3	-0.08007	0.09012	0.375	-0.18278	0.03105	<0.0001	-0.47442	0.04517	<0.0001
	Intercept4							-0.18304	0.10772	0.089
	Linear4							0.01887	0.03259	0.563

		GBTM			GBDTM			GBMTM		
ρ**	Parameter	Mean Estimates	Mean SE*	p-value	Mean Estimates	Mean SE*	p-value	Mean Estimates	Mean SE*	p-value
0.6	Intercept1	-1.31898	0.27013	<0.0001	-1.92158	0.07173	<0.0001	-1.98290	0.07299	<0.0001
	Linear1	-0.34132	0.15871	0.032	-0.09727	0.02315	<0.0001	-0.08416	0.02345	0.0003
	Intercept2	-3.42048	0.55303	<0.0001	-2.97224	0.13622	<0.0001	-2.97177	0.13260	<0.0001
	Linear2	0.63677	0.16861	0.0002	0.68696	0.03746	<0.0001	0.68247	0.03653	<0.0001
	Intercept3	0.91688	0.17816	<0.0001	1.37535	0.08851	<0.0001	1.79018	0.14149	<0.0001
	Linear3	-0.19289	0.04298	<0.0001	-0.32841	0.02512	<0.0001	-0.70864	0.04691	<0.0001
	Intercept4							1.01726	0.11710	<0.0001
	Linear4							-0.05593	0.03532	0.113

[*] SE = Standard Error
[**] ρ = Correlation Level between Y_t and Z_t
[#] GBTM = Group-based trajectory modeling
[$] GBDTM = Group-based dual trajectory modeling
[&] GBMTM = Group-based multi-trajectory modeling
Note: p-values are calculated based on the average mean and SE

ρ **= 0.1:** Two linear polynomial trajectories were generated using GBTM for Z_t. The p-value of parameter estimates was significant for Linear 1 and Linear 2. In GBDTM, there were two constant trajectories generated because the parameters from Linear 1 and Linear 2 were not significant. In GBMTM, the first three groups followed a linear polynomial trajectory, and one consistent trajectory was found in Group 4. Figure 5.5

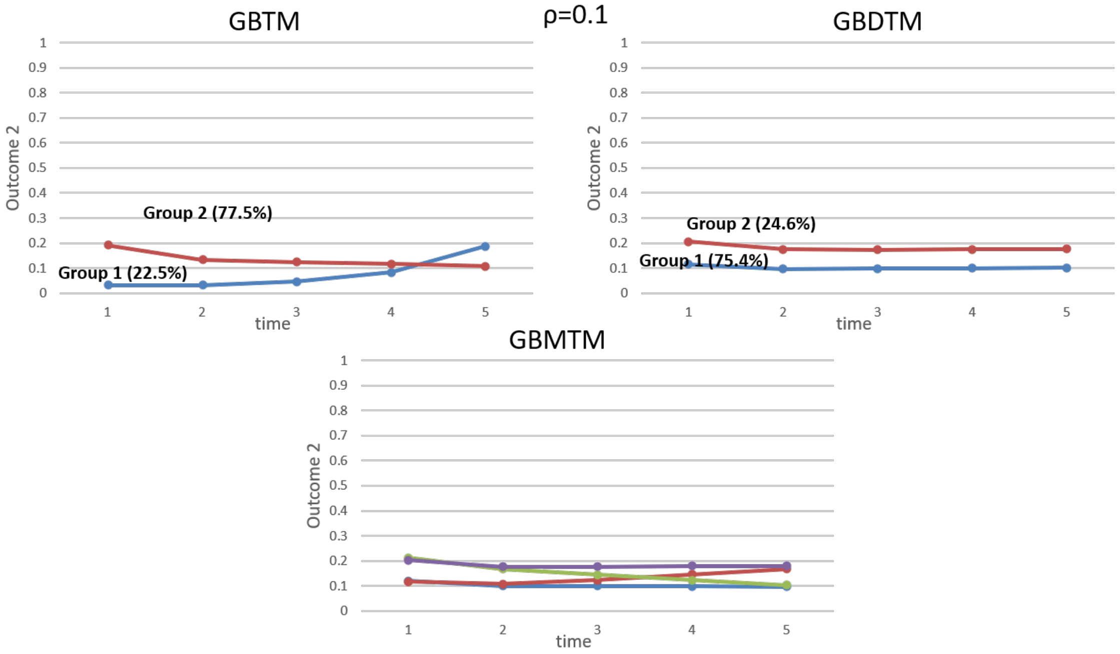


Figure 5.5: Simulation trajectory shapes of Outcome 2 in GBTM GBDTM and GBMTM with N = 4000 and $\rho = 0.1$

was the corresponding trajectories to Table 5.3 with $\rho = 0.1$.

ρ **= 0.2:** Two trajectory groups were found using GBTM and GBDTM. The trajectories from Group 1 and Group 2 were linear and constant in GBTM. Two constant trajectories were identified in GBDTM. In GBMTM, the paths from Group 1 & Group 4 were consistent, and the trajectories from Group 2 & Group 3 were linear. Figure 5.6 were the corresponding trajectories to Table 5.3 with $\rho = 0.2$.

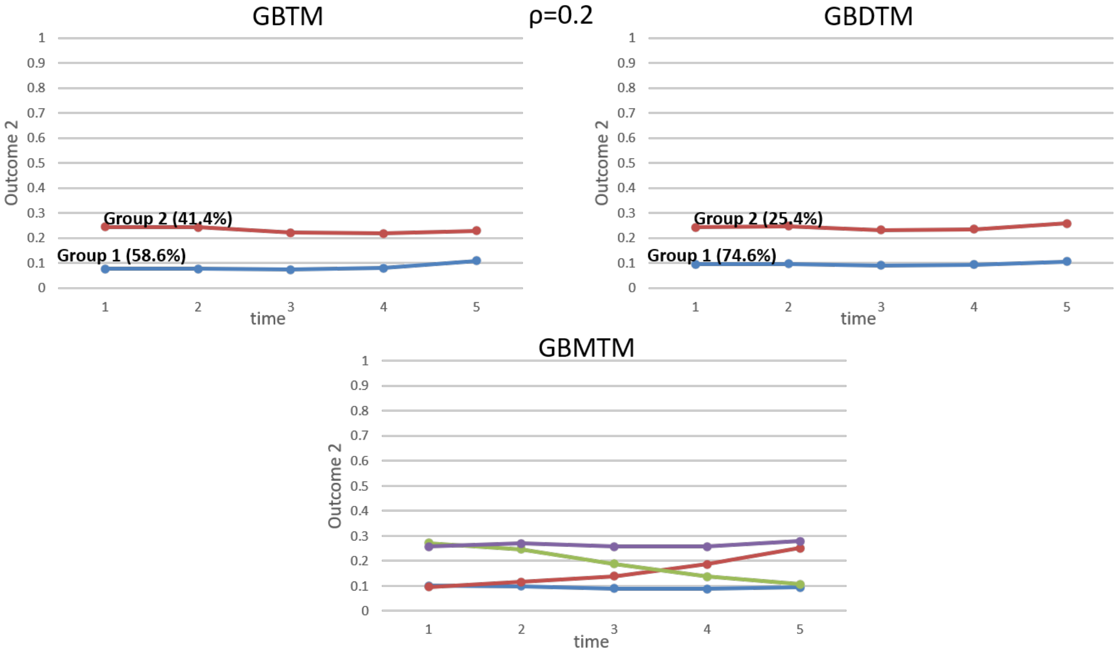


Figure 5.6: Simulation trajectory shapes of Outcome 2 in GBTM GBDTM and GBMTM with N = 4000 and $\rho = 0.2$

ρ **= 0.4:** Three trajectory groups were generated. The trajectory of Group 2 in GBTM was linear. Parameters of intercept and linear of the time variable were not significant in Group 1 and Group 3 from GBTM, which were defined as unknown polynomial trajectory shapes. In GBDTM, three linear trajectories were observed. In GBMTM, the path from Group 1 was constant; the trajectories from Group 2 & Group 3 were linear, and the trajectory from Group 4 had an unknown polynomial trajectory shape. Figure 5.7 were the corresponding trajectories to Table 5.3 with $\rho = 0.4$.

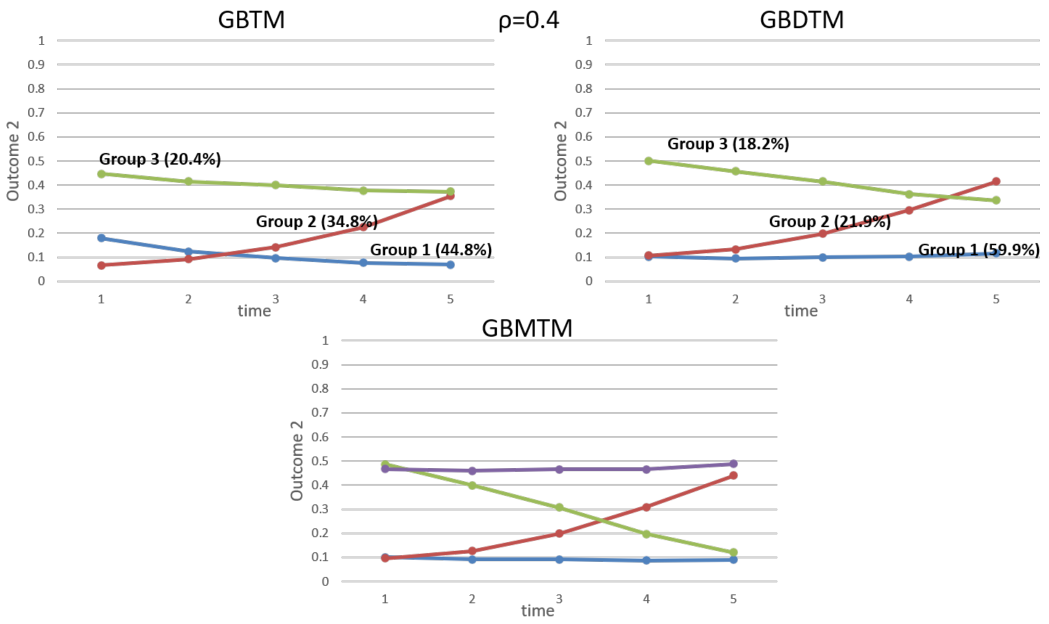


Figure 5.7: Simulation trajectory shapes of Outcome 2 in GBTM GBDTM and GBMTM with N = 4000 and $\rho = 0.4$

ρ = 0.6: Three linear paths were found in both GBTM and GBDTM. In GBMTM, trajectories in Group 1 to Group 3 were linear, and trajectory in Group 4 was constant. Figure 5.8 were the corresponding trajectories to Table 5.3 with $\rho = 0.6$.

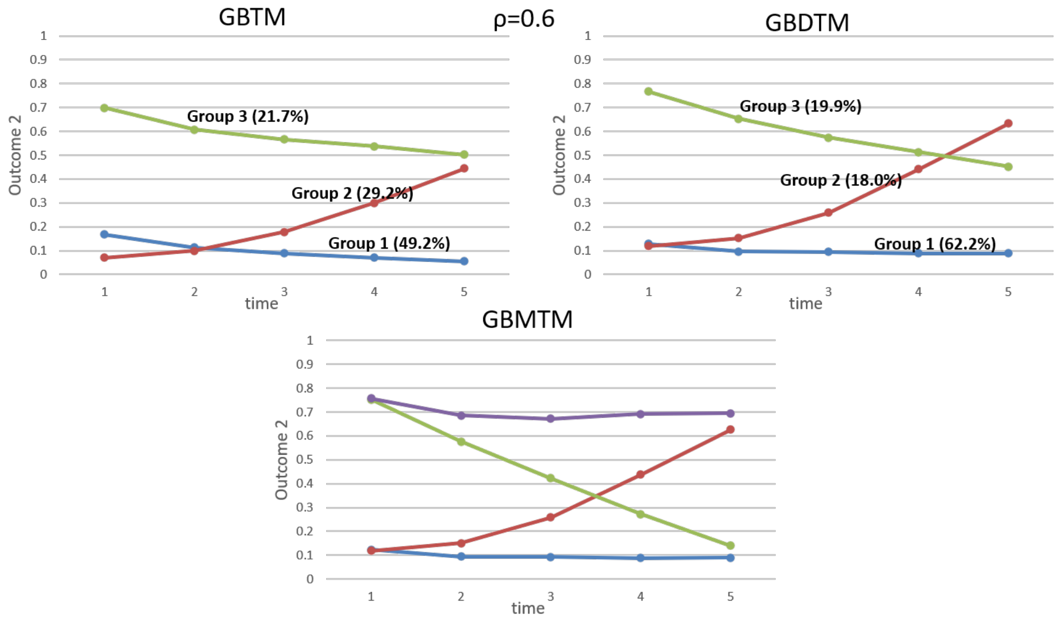


Figure 5.8: Simulation trajectory shapes of Outcome 2 in GBTM GBDTM and GBMTM with N = 4000 and $\rho = 0.6$

Another key finding was that we could compare the standard error variation for Z_t among different models or correlation levels. A comparison of the standard error from three trajectory models with the same correlation level revealed that GBTM had a larger standard error than GBDTM and GBMTM. Therefore, the standard error of trajectory parameter estimates was reduced in GBDTM and GBMTM in Z_t adjusted by Y_t. Moreover, the standard error of Z_t was declined with the correlation level (ρ) of the two outcomes getting larger in each kind of model.

5.3.5 Summary of simulation study based on trajectory models

Trajectories of GBTM, GBDTM, and GBMTM in two outcomes (Y_t, Z_t) with different correlation coefficients (ρ = 0.1, 0.2, 0.4, 0.6) were presented in Figures 5.9 - 5.12. Each Figure's A and D were related to Y_t and Z_t in GBTMs; each Figure's B and E were related to Y_t and Z_t in GBDTMs; each Figure's C and F were related to Y_t and Z_t in GBMTMs.

Outcome 1 Key Findings:

- Since Y_t was generated from the trajectory properties and the GBTM of Y_t was not influenced by Z_t, trajectory group membership and trajectory shapes were the same in each Figure's A.

- The trajectory shapes in each model with different levels of correlation were barely changed compared to the trajectories of Y_t in GBTM, GBDTM, and GBMTM. This was because the trajectory shape of Y_t was highly adjusted during data-generating, so the effect of Y_t from Z_t was limited in GBDTM and GBMTM.

- The variety of proportions from Y_t in the Figures increased as the correlation level increased.

Outcome 2 Key Findings:

- In GBTM and GBDTM, two trajectories were found with correlation level ρ =0.1

and 0.2, three trajectories were identified with correlation levels ρ =0.4 and 0.6. In GBMTM, four trajectories were found in each level of correlation.

- The distribution interval of probability for the mean paths increased as the correlation level increased. With correlation $\rho = 0.1$, the probability range of the average trajectories was around 0.1 to 0.2. However, when the correlation coefficient rising to 0.6, the range of probability was from 0.05 to 0.7.

- Compared to GBTM, trajectory shape and proportion had obviously changed in GBDTM. The trajectory shape and group members from Z_t were highly adjusted by Y_t in GBDTM.

- In GBMTM, the tendency of the trajectories in Z_t was similar to Y_t. For instance, the trajectory in Group 1 with a low probability in Y_t also had a relatively low probability in Z_t.

- The probability region was constringent as the correlation level decreased. With a correlation level of $\rho = 0.1$, the four trajectories were gathered in the tiny probability interval between 0.1 and 0.2.

Summary graphs of the simulation study result can be seen together and compared side by side in Figure 5.9 - 5.12.

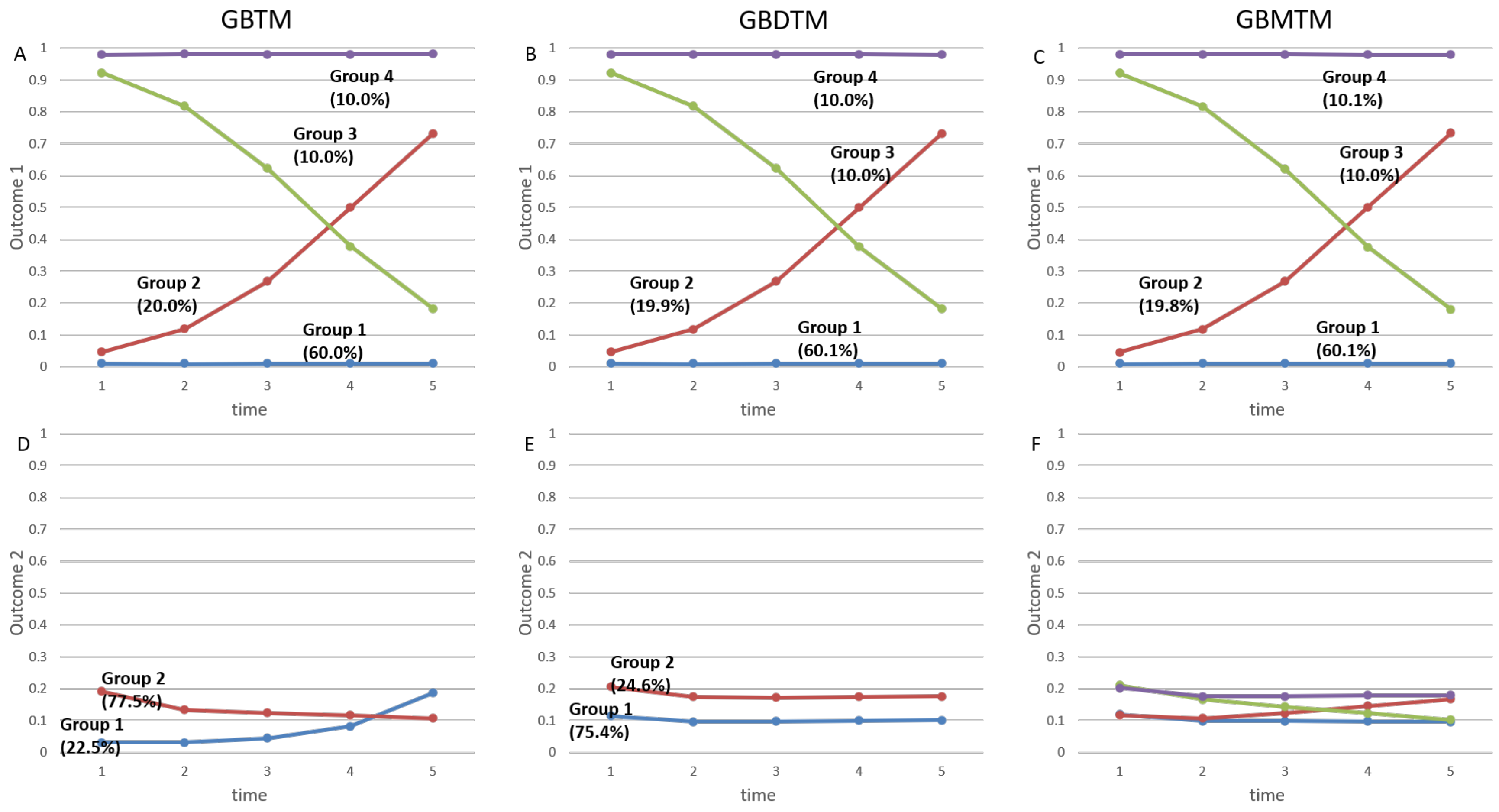


Figure 5.9: Average trajectories of three trajectory models from simulation study with sample size N = 4000 and correlation coefficient 0.1 based on 500 simulations

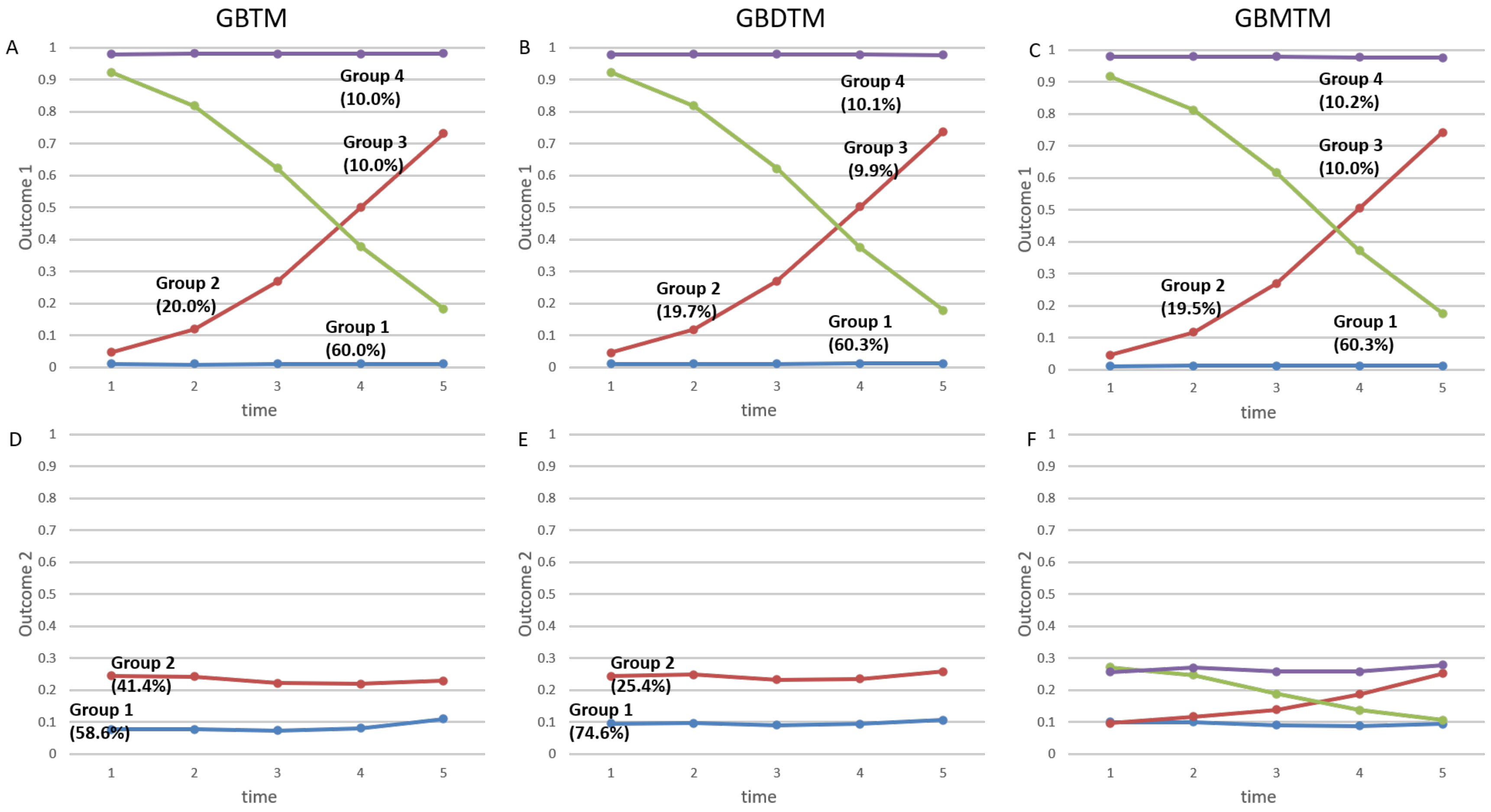


Figure 5.10: Average trajectories of three trajectory models from simulation study with sample size N = 4000 and correlation coefficient 0.2 based on 500 simulations

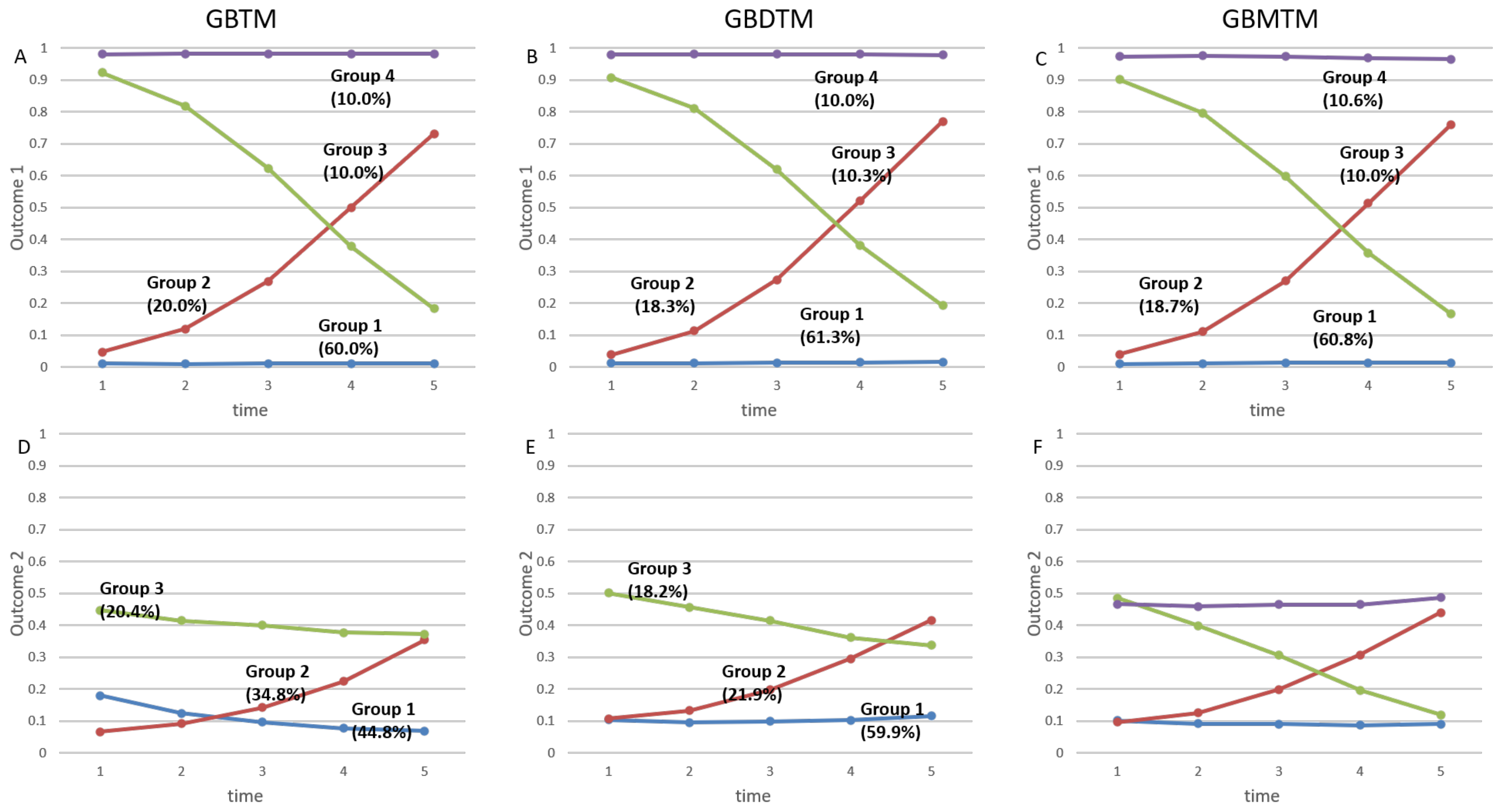


Figure 5.11: Average trajectories of three trajectory models from simulation study with sample size N = 4000 and correlation coefficient 0.4 based on 500 simulations

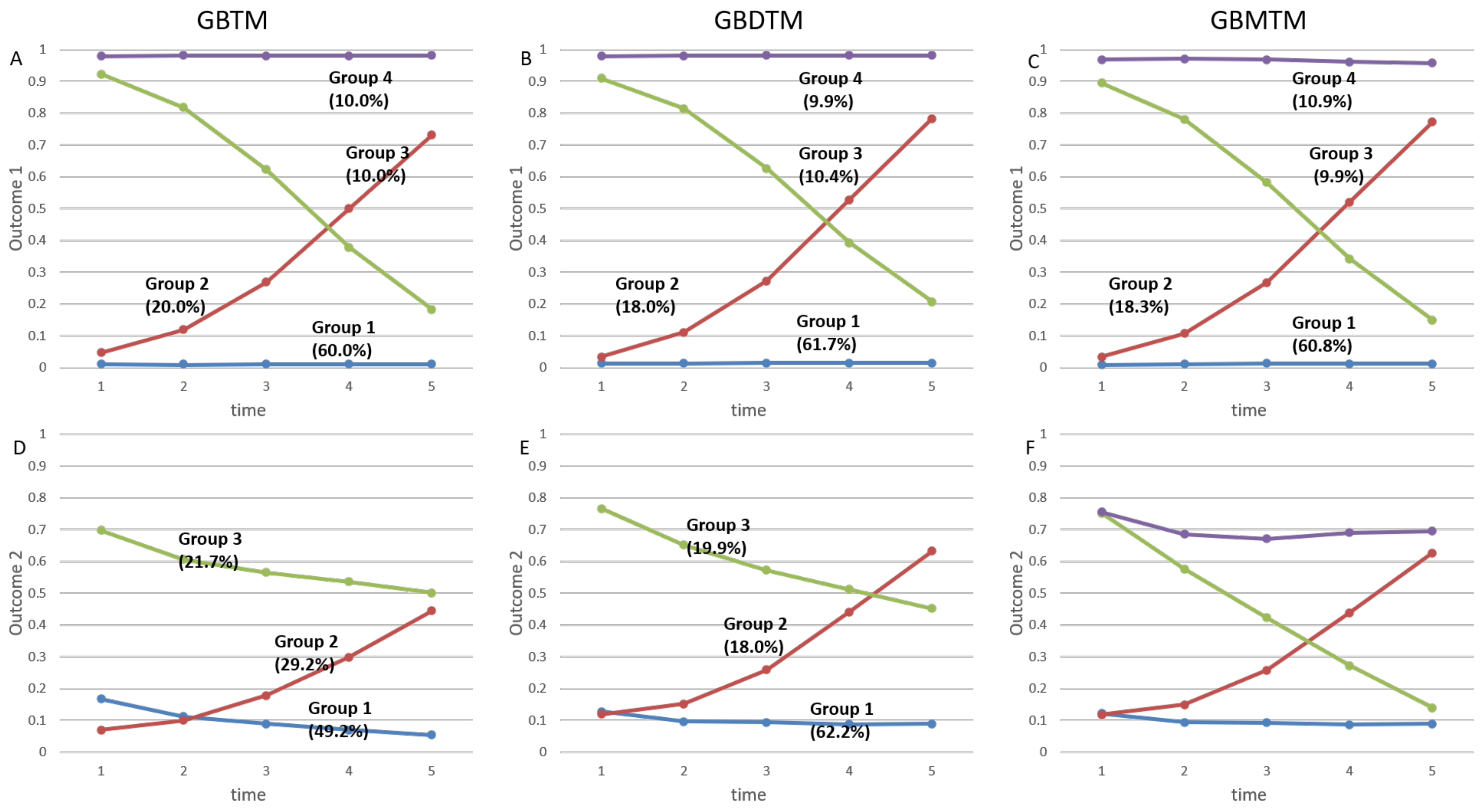


Figure 5.12: Average trajectories of three trajectory models from simulation study with sample size N = 4000 and correlation coefficient 0.6 based on 500 simulations

Chapter 6 DISCUSSION

6.1 Introduction

In this thesis, I applied group-based trajectory modeling (GBTM), group-based dual trajectory modeling (GBDTM), and group-based multi-trajectory modeling (GBMTM) to identify the trajectory trends over time with two associated longitudinal binary outcomes - depression and anxiety. Trajectory groups and the shape of the trajectories were selected based on BIC and AIC values and on posterior probability (Nagin, 2005). Risk factors for both outcomes were identified in both GBTMs and GBDTM. Moreover, I simulated two repeated measured outcomes with different association levels to further study the above trajectory models based on polynomial trajectory parameters, trajectory shapes, and trajectory tendencies.

6.2 Discussion of the application

Three trajectory models (GBTM, GBDTM, and GBMTM) were applied and developed with depression and anxiety outcomes from the KHPS dataset. In GBMTM, trajectory shapes varied significantly when compared to GBTM and GBDTM. This was because GBMTM must share the same proportion in each outcome, which restricted the flexibility of the model.

When trajectories were identified for both GBTM and GBDTM, the trends from different trajectory groups should involve variations (Nagin, 2005; Nagin & Tremblay, 2001). In this study, compared to GBTM, the shape of each depression and anxiety trajectory group remained similar to GBDTM. One reason for this finding was that the membership proportion variations from GBTM to GBDTM were small, but still not small enough to be ignored for both depression and anxiety. Another reason was that the portion of variation from GBTM to GBDTM might include many missing values that failed to influence the trajectory pathways. The study result showed that the membership proportion derived some changes from GBTM to GBDTM for both depression and anxiety. When the GBDTM depression trajectory groups were compared to that of GBTM, the members in the "low-flat", "low-to-high" and "high-curved" depression trajectory increased, but decreased in "low-to-middle" depression trajectory group. Meanwhile, comparing anxiety trajectories in GBDTM to GBTM, the members in the "low-flat" anxiety group decreased, but increased in the "low-to-middle" and "high-to-low" anxiety trajectory group; the "high-curve" anxiety trajectory groups remained unchanged.

Compared to the single outcome mixture models, the mixture model with multiple outcomes often had smaller standard errors for estimates when there were many missing values (Teixeira-Pinto et al., 2009). However, this was not observed in our analysis. The standard error of parameter estimates from GBTM to GBDTM in depression and anxiety trajectories was half decreased and half increased. The reason could be that missing outcomes depended on the subjects who failure to complete the survey. In our real dataset, if the individuals were missing in depression outcome measures, they would also be missing anxiety measures. The multi-outcome mixture model analysis involved more significant variables compared to the separate analysis (Teixeira-Pinto et al., 2009; Mayo-Wilson et al., 2017). In our application, the multivariate analysis for depression using GBDTM involved a total of six risk factors, one fewer than GBTM for depression. On the other hand, the multivariate analysis for anxiety using GBDTM involved a total of three risk factors, two more variables than GBTM for anxiety. De Oliveira showed that increased physical activity reduced anxiety probability in older people (de Oliveira, Souza, Rodrigues, Fett, & Piva, 2019). However, our study did not show this finding in GBTM and GBDTM with anxiety. Overall, the number of significant risk factors in GBDTM was more than separate GBTMs with depression and anxiety.

In GBDTM, among the four groups recognized as having different probabilities of being diagnosed with depression, a majority showed no depression and were generally unlikely to experience anxiety concomitantly. However, about 10% did experience depression during the follow-up period, most of whom showed a gradual increase in the probability of depression. Among individuals along this trajectory, 20% also experienced a moderate rise in anxiety risk over time. As for anxiety, diagnosis of which followed four different trajectories, the majority of respondents did not experience this condition and were also free of depression. However, 5% showed a slow increase trend for anxiety over time. This was accompanied by an increasing tendency to suffer from depression in just under half the cases. In general, being female, not involved in an income-generating activity in the older population, and membership in a trajectory suggesting risks for the alternate condition independently predicted a more vulnerable risk trajectory than the "low-flat" trajectory group for depression and anxiety.

Among the four depression trajectory groups, the constantly decreasing trajectory was not found for depression. The "high-curve" depression group was thought to have less likelihood of recovery among the older adults as they were more likely to experience reduced life satisfaction, income, quality of life, and poor health conditions (Dew et al., 2007; You et al., 2009; Jang, Small, & Haley, 2001). The "low-to-high" depression group had an intense increase in the occurrence of depression from 2009 to 2013 with a very small proportion (only 31 subjects). The markedly increased probability might have been precipitated by sudden and serious events, such as losing a spouse, physical incapacity, etc. However, among anxiety trajectories, a declining trajectory and a curved shape trajectory showed evidence of decreasing risk. A possible explanation for this observed decline is that individuals adapt or cope with their anxious feelings and no longer seek treatment. A second explanation might be that other, more pressing medical conditions emerge, eclipsing anxiety management. Thus, anxiety might still have been present but not identified (AAPG, 2019).

The association between depression and anxiety was identified from the trajectories' conditional probabilities and the logistic regression odds ratios. The study found that the "low-to-high" and "low-to-middle" depression groups also had a risk of being in the "low-to-middle" anxiety group. This suggested that older adults with an increasing likelihood of suffering from depression also have a greater chance of suffering from anxiety. This is consistent with other studies (Wetherell, Gatz, & Craske, 2003; Bassil, Ghandour, & Grossberg, 2011). Moreover, individuals in the "low-to-middle" depression group made up a high proportion of the "high-curve" anxiety trajectory group, suggesting that older patients who had severe anxiety may also suffer mild depression. The "high-curve" depression group members were more likely to have anxiety following the "high-to-low" and, less frequently, the "high-curved" anxiety trajectory; individuals in this particular overlap had severe mental health conditions and required more attention (Lenze, 2003). The inverse of these findings also supported the association between depression and anxiety; individuals in this study who did not have one of the study conditions (depression or anxiety) also tended not to have the other.

Our evaluation of demographic risk factors coincides to varying degrees with the literature. In the majority of depression and anxiety studies, sex did have an association with these conditions, suggesting older females generally were more at greater risk for depression and anxiety (McLean, Asnaani, Litz, & Hofmann, 2011; Girgus, Yang, & Ferri, 2017). Our study findings were consistent with results from other trajectory studies (Holmes et al., 2018; Montagnier et al., 2014; Kuchibhatla et al., 2012; El-Gilany, Elkhawaga, & Sarraf, 2018). However, some studies had found no sex-specific differences when investigating depression and anxiety (Spinhoven et al., 2017; Taylor & Lynch, 2004). This inconsistency might be related to different economic circumstances, social-cultural factors, psychosocial gender roles, or other population differences. In our study, age was a significant univariate influence for depression only, consistent with Holmes et al. (2018). Education level was not a significant predictor of either outcome. This result was consistent with some studies (Holmes et al., 2018; Norris & Murrell, 1988; Hong, Hasche, & Bowland, 2009), but not others (Spinhoven et al., 2017; Liang et al., 2011; Kuo et al., 2011; Byers et al., 2012; Hsu, 2012; Montagnier et al., 2014; Kuchibhatla et al., 2012; Andreescu et al., 2008). This lack of relationship our study revealed might be attributable to our study participants' relatively low education level overall.

Social factors were also known to influence various mental health problems, including depression and anxiety. Some studies suggested that older adults who lived alone or those without a partner, who live within an isolated social environment had a higher risk of depression or anxiety (El-Gilany et al., 2018; Kang et al., 2016; Chong et al., 2001; K. M. Mehta et al., 2003; Brown et al., 2002; Won & Choi, 2013). However, living alone and marital status did not relate to depression and anxiety in our study, which was consistent with other studies (Byers et al., 2012; Hsu, 2012; Montagnier et al., 2014; Rzewuska et al., 2015). Studies have also pointed to smoking or excessive drinking possibly also increasing the risk of depression and anxiety (Kuo et al., 2011; Byers et al., 2012; K. M. Mehta et al., 2003; Kirchner et al., 2007). Nevertheless, this association was not observed in our study, nor was it in others' work (Kuchibhatla et al., 2012; Kang et al., 2016). Some studies showed that homeownership reduces the risk of depression and anxiety (Kang et al., 2016; Chiao, Weng, & Botticello, 2011), but this association did not emerge in our multivariate analysis. However, income-generating activity was relevant in predicting both depression and anxiety, suggesting that people who still work and earn money later in life may have better mental health. Moreover, the older adults might have to work longer or delay their retirement to continue their financial circumstances (Lin, Dean, & Ensel, 2013; Newby & Moulds, 2011; Flint & Rifat, 1997a). Another possible reason is that the individuals at an older age are higher functioning overall in their ability to continue working (Hersen & Van Hasselt, 1992).

Chronic diseases (heart disease, stroke, diabetes, asthma, cancer, arthritis, osteoporosis, etc.) posed understandable challenges for older people and may impact mental health. In studying the relationship among depression, anxiety, and chronic disease, Clarke and Kay reviewed 159 papers published between 1995 and 2007 and found that depression was correlated with nearly all chronic diseases (Clarke & Currie, 2009). However, anxiety was only associated with heart disease, stroke, and diabetes mellitus. In our study, older adults with more than three chronic conditions, such as heart disease, stroke, cancer, and arthritis, etc., were more likely to develop depression. However, in anxiety trajectory groups, chronic disease was only significant in the "high-to-low" group from the univariate analysis, but not in the multivariate analysis. Studies had shown that in older adults, physical illness or disability is usually positively correlated with depression and anxiety (Kang et al., 2016; Knight, Nordhus, & Satre, 2003; Brenes et al., 2008; Hermans & Evenhuis, 2013). However, based on our multivariate analysis, physical/mental disability did not predict these outcomes.

6.3 Discussion of the simulations

In the simulations, I generated two relevant longitudinal outcomes and then developed and compared GBTM, GBDTM, and GBMTM with these two associated longitudinal outcomes. The data from repeated measurement Outcome 1 (Y_t) was generated with four clusters and defined trajectory pathways for each group (Haviland et al., 2011). The data from repeated measurement Outcome 2 (Z_t) was developed based on different levels of correlation coefficients with each measure of Y_t using logistic regression (Ocram, 2014; Touloumis, 2016). The correlation levels for each measure between two outcomes were $\rho = 0.1$, 0.2, 0.4, and 0.6. Each simulation was performed with sample size N = 500, 2000, and 4000 subjects. Five hundred replicated simulations were run in each scenario.

From the simulations, parameter estimates' bias with the trajectories from the models in Y_t could be large when the sample size was small (N = 500, 2000). Large bias was caused by the property of logistic regression to deal with the rare event or non-event (King & Zeng, 2001). As the sample size increased to N = 4000, parameter estimates moved closer to the real parameter value when used to generate Y_t. The parameters of Y_t were adjusted by Z_t in GBDTM and GBMTM. Studies showed that the outcomes could vary from one another in multi-outcome models with small standard errors compared to developments in the single-outcome models (Teixeira-Pinto et al., 2009; Mayo-Wilson et al., 2017). Our simulation study had the same findings in GBDTM and GBMTM compared to GBTM. In GBDTM and GBMTM, parameter estimates for Z_t had a smaller standard error than GBTM because of the adjustment by Y_t. Furthermore, as the two outcomes' correlation coefficient increased, the standard error decreased in Z_t in each of the three models. This was because the clusters of Z_t became much easier to be identified as the correlation increased with Y_t. Y_t was less influenced by Z_t in GBDTM and GBMTM because it was highly adjusted during data generation. On the other hand, Z_t was more influenced by Y_t in GBDTM and GBMTM, mostly when the two outcomes' correlation coefficient was low. Moreover, GBDTM included conditional probability when identifying the interrelationship between depression and anxiety directly. Therefore, if a significant association between two outcomes exists, and researchers are interested in studying their interrelationship, GBDTM should be preferred over GBTM.

In the simulation study, the proportion of the trajectory shape of Y_t in GBDTM and GBMTM was similar to GBTM. In our real data analysis, similar results were also discovered in both depression and anxiety outcomes. However, Z_t's average trajectories changed from GBTM to GBDTM due to Y_t's adjustment in the simulation study. In the GBMTM simulation results, four constructed groups were consistently identified for both outcomes. Although with low correlations, Z_t's four trajectories tended to diverge into two overall patterns with trajectories that overlapped within the patterns. As the correlation increased, the trajectory relationships became more distinct from one another and more similar between methods. Therefore, GBMTM and GBDTM are interchangeable with a high correlation (ρ = 0.8, 0.9).

Overall, different pathways for depression and anxiety were generated from the statistical approach, GBDTM. GBDTM included two outcomes simultaneously. Unlike GBTM, GBDTM involved the correlation between two outcomes when compared to GBTM and identified more risk factors. Therefore, GBDTM was better for modeling two correlated outcomes compared to GBTM. As Nagin explained, GBMTM was used to identify latent clusters of individuals following likely paths over time in multiple outcomes (Nagin et al., 2018). In this study, depression and anxiety do not share a similar trajectory shape in depression and anxiety outcome subgroups. Thus, GBMTM was not considered a better model than the other two models. Overall, GBDTM was more flexible in handling the different clusters of structures compared to GBMTM.

Strength and weakness

This study had limitations. First, the low prevalence of anxiety across the survey period limited predictor evaluation, particularly in the very small populated trajectories such as the "high-curved" anxiety trajectory group. Second, although the current study employed data from a sizeable subsample of the KHPS dataset, around 35% of the outcome measurements were missing, which might result in bias, even though I used two common methods to handle the missing data by maximum likelihood estimation (MLE) method and GBDTM. Third, the variables included in this study did not contain other potentially important health and psychosocial aspects that may be associated with depression and anxiety, such as stressful life events and social/family support information. Fourth, the specific cultural context of the Korean older adults in which this study was conducted may not be generalizable to other contexts.

The current study had many strengths. The KHPS dataset provided measures of outcomes annually for eight years, meaning that sufficient measurement time points could be used to develop depression and anxiety trajectories. Additionally, instead of self-reported signs, symptoms, or questionnaires, as in other studies, depression and anxiety outcomes in the KHPS dataset were collected from medical expenses, including prescription drug receipts or medical institutions/pharmacies, potentially leading to inadequate recognition of our sample outcomes. This is particularly true in the context of other chronic disease conditions (Manela et al., 1996). Therefore, the outcomes were more clinically valid. In this study, depression and anxiety were considered binary, which was different from most other studies using continuous scale outcomes. Furthermore, anxiety trajectories had barely been studied in older adults, so our research can be act as a guideline for future studies.

Chapter 7 CONCLUSION AND FUTURE RESEARCH

7.1 Conclusion

GBTM, GBDTM, and GBMTM were compared in both real data analysis and simulation. In the simulation study, GBDTM had less uncertainty in parameter estimation and so was always better than GBTM. A simulation study was conducted and showed that GBMTM could be instead of GBDTM when the correlation between two outcomes is high, or the data structure between two outcomes is similar. In this thesis, using the data from KHPS, GBTM, GBDTM, and GBMTM were applied to examine the tendency to suffer from depression and anxiety simultaneously. Since the correlation coefficients were between depression and anxiety significant but low, and group clusters for depression compared to anxiety were different, GBDTM was a better model than GBTM and GBMTM for the KHPS data.

Four trajectory groups of both depression and anxiety were generated for the KHPS dataset of older Koreans. The majority of older adults belong to the "low-flat" trajectory groups for depression and anxiety. This suggests that most older adults did not have depression and anxiety. Being female, having more than three chronic diseases, and not being involved in income-generating activities were significant predictors for the depression trajectory groups. Being female and not being involved in income-generating activities were significant predictors for the anxiety trajectory groups. Our findings were based on a large sample size, which guaranteed reliable differentiated trajectory groups and supported previous results found in the literature. Our findings can be used to assist health policy decision-makers in identifying individuals at risk for comorbid depression and anxiety and aid in devising supports for older individuals at risk of deteriorating mental health.

Main contributions

In this thesis, I applied GBTM, GBDTM, and GBMTM using the longitudinal binary depression and anxiety outcomes from the KHPS dataset, performed a simulation study, and compared the three trajectory models. GBDTM was selected as the best model. Conditional probabilities from GBDTM directly described the interrelationship between the depression and anxiety outcomes. Risk factors relevant to depression and anxiety outcomes were also identified with multivariate logistic regression analysis. South Korea is expected to become a "super-aged society" with over 20% of its population aged 65 years and older in 2026 and 38% in 2050. Our study used a general population sample, not enriched for a specific group. We think our findings may help health policy maker to develop appropriate depression and anxiety prevention programs.

We simulated GBTM, GBDTM, and GBMTM using two binary longitudinal outcomes with different correlation coefficient levels in each measurement. The characteristics of the three trajectory models were studied further in the simulation study. The simulation study also showed that GBDTM was always a better model than GBTM. GBMTM could be used instead of GBDTM if the correlation coefficient between two longitudinal outcomes was significantly high or with the similar data structures.

7.2 Future study

In this thesis, we studied GBTM, GBDTM, and GBMTM using two longitudinal binary outcomes. However, in some clinical studies, we might have longitudinal count outcomes. For instance, the trajectories from the patients with disabilities were measured by the number of basic activities they performed, called activities of daily living. In a mental health study, we might also be interested in the number of emergency visits or number of days an individual stayed in a hospital for depression and anxiety. Therefore, instead of the logistic polynomial function, we could consider using the zero-inflated Poisson (ZIP) model to identify the trajectories' paths with count data, including a lot of zeros. In the future, for GBDTM and GBMTM, we could also investigate the models' performance, including the mixed with continuous, binary and count outcomes.

We used co-current depression and anxiety outcomes in the older people from the KHPS dataset. However, GBDTM could also handle two linked effects that did not necessarily co-occur. The National Longitudinal Survey of Children and Youth (NLSCY) data was a longitudinal dataset from Statistics Canada (NLSCY, 2010). NLSCY data contained numerous factors correlated to a child's social, emotional, and behavioral development at multiple time measures. When the children were youth, their mental health was measured using an anxiety scale. However, as they grew into adolescence, mental health was measured using a depression scale. Therefore, GBDTM can be applied to find anxiety trajectories in youth and then depression trajectories as they became adolescents and young adults. Conditional probabilities could be used to study how anxiety in early childhood influences depression in adolescents and young adults.

GBDTM can only include two outcome variables at the same time. For more than two outcomes, another technique called the parallel process growth mixture model might be a suitable method to identify associations among multiple outcomes simultaneously (Wu et al., 2010).

In the application study, the risk factors were considered as time-independent covariates. However, GBTM, GBDTM, GBMTM can obtain time-dependent covariates as well. For example, an important event (such as the loss of a partner) may affect mental health during the measurement year. In these three trajectory models, risk factors mainly affect the proportion variation, but time-dependent variables could change trajectory shapes.

In our simulation study, Outcome 1 was simulated based on the trajectories' parameters. Outcome 2 was simulated based on the correlation coefficient from each measure of Outcome 1. Our study's simulation did not consider the missing data's influence, especially for data missing not at random. In the future, we would also study how the data missing not at random would influence GBDTM and GBMTM.

References

AAPG. (2019). *Anxiety and older adults: Overcoming worry and fear.* https://www.aagponline.org/index.php?src=gendocs&ref=anxiety.

Acierno, R., Brady, K., Gray, M., Kilpatrick, D. G., Resnick, H., & Best, C. L. (2002). Psychopathology following interpersonal violence: A comparison of risk factors in older and younger adults. *Journal of Clinical Geropsychology*, *8*(1), 13–23.

Allgulander, C., & Lavori, P. W. (1993). Causes of death among 936 elderly patients with 'pure'anxiety neurosis in stockholm county, sweden, and in patients with depressive neurosis or both diagnoses. *Comprehensive psychiatry*, *34*(5), 299–302.

Andreescu, C., Chang, C.-C. H., Mulsant, B. H., & Ganguli, M. (2008). Twelve-year depressive symptom trajectories and their predictors in a community sample of older adults. *International psychogeriatrics*, *20*(2), 221–236.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. *Advanced structural equation modeling: Issues and techniques*, *243*, 277.

Areán, P. A., & Cook, B. L. (2002). Psychotherapy and combined psychotherapy/pharmacotherapy for late life depression. *Biological psychiatry*, *52*(3), 293–303.

Baker, G., et al. (1954). Factor analysis of relative growth. *Growth*, *18*, 137–143.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (Vol. 904). John Wiley & Sons.

Bassil, N., Ghandour, A., & Grossberg, G. T. (2011). How anxiety presents differently in older adults. *Current Psychiatry*, *10*(3), 65–71.

Beekman, A. T., Bremmer, M. A., Deeg, D. J., Van Balkom, A. J., Smit, J. H., De Beurs, E., … Van Tilburg, W. (1998). Anxiety disorders in later life: a report from the longitudinal aging study amsterdam. *International journal of geriatric psychiatry*, *13*(10), 717–726.

Beekman, A. T., de Beurs, E., van Balkom, A. J., Deeg, D. J., van Dyck, R., & van Tilburg,

W. (2000). Anxiety and depression in later life: co-occurrence and communality of risk factors. *American Journal of psychiatry*, *157*(1), 89–95.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* springer.

Blazer, D., Burchett, B., Service, C., & George, L. K. (1991). The association of age and depression among the elderly: an epidemiologic exploration. *Journal of gerontology*, *46*(6), M210–M215.

Böhning, D. (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference*, *47*(1-2), 5–28.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol. 467). Hoboken, New Jersey: John Wiley & Sons.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization.* Cambridge,UK: Cambridge university press.

Brenes, G. A., Penninx, B. W., Judd, P. H., Rockwell, E., Sewell, D. D., & Wetherell, J. L. (2008). Anxiety, depression and disability across the lifespan. *Aging and Mental Health*, *12*(1), 158–163.

Brown, J. W., Liang, J., Krause, N., Akiyama, H., Sugisawa, H., & Fukaya, T. (2002). Transitions in living arrangements among elders in japan: does health make a difference? *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(4), S209–S220.

Bruce, M. L. (2001). Depression and disability in late life: directions for future research. *The American Journal of geriatric psychiatry*, *9*(2), 102–112.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological bulletin*, *101*(1), 147.

Bushway, S., & Weisburd, D. (2006). Acknowledging the centrality of quantitative criminology in criminology and criminal justice. *The Criminologist*, *31*(4), 1.

Butters, M. A., Bhalla, R. K., Andreescu, C., Wetherell, J. L., Mantella, R., Begley, A. E., & Lenze, E. J. (2011). Changes in neuropsychological functioning following treatment for late-life generalised anxiety disorder. *The British Journal of Psychiatry*, *199*(3), 211–218.

Byers, A. L., Vittinghoff, E., Lui, L.-Y., Hoang, T., Blazer, D. G., Covinsky, K. E., ... others (2012). Twenty-year depressive trajectories among older women. *Archives of general psychiatry*, *69*(10), 1073–1079.

Cambridge Dictionary, s. (2018). *Cambridge dictionary.* `https://dictionary.cambridge.org/dictionary/english/trajectory`.

Carroll, R., Wang, S., Simpson, D., Stromberg, A., & Ruppert, D. (1998). The sandwich (robust covariance matrix) estimator. *Unpublished manuscript*.

Caspi, A., & Roberts, B. W. (2001). Personality development across the life course: The argument for change and continuity. *Psychological Inquiry*, *12*(2), 49–66.

Cericola, V. (2015). *Quasi-newton methods.* `https://optimization.mccormick.northwestern.edu/index.php/Quasi-Newton_methods`.

Chiao, C., Weng, L.-J., & Botticello, A. L. (2011). Social participation reduces depressive symptoms among older adults: an 18-year longitudinal analysis in taiwan. *BMC public health*, *11*(1), 292.

Cho, M. J., Chang, S. M., Lee, Y. M., Bae, A., Ahn, J. H., Son, J., ... others (2010). Prevalence of dsm-iv major mental disorders among korean adults: a 2006 national epidemiologic survey (keca-r). *Asian journal of psychiatry*, *3*(1), 26–30.

Cho, M. J., & Lee, J. Y. (2005). Epidemiology of depressive disorders in korea. *Psychiatry Investig*, *2*(1), 22–7.

Cho, M. J., Lee, J. Y., Kim, B.-S., Lee, H. W., & Sohn, J. H. (2011). Prevalence of the major mental disorders among the korean elderly. *Journal of Korean medical science*, *26*(1), 1–10.

Cho, M. J., Seong, S. J., Park, J. E., Chung, I.-W., Lee, Y. M., Bae, A., ... others (2015). Prevalence and correlates of dsm-iv mental disorders in south korean adults: the korean epidemiologic catchment area study 2011. *Psychiatry investigation*, *12*(2), 164.

Chong, M.-Y., Chen, C.-C., Tsang, H.-Y., Yeh, T.-L., Chen, C.-S., Lee, Y.-H., … Lo, H.-Y. (2001). Community study of depression in old age in taiwan: prevalence, life events and socio-demographic correlates. *The British Journal of Psychiatry*, *178*(1), 29–35.

Chui, H., Gerstorf, D., Hoppmann, C. A., & Luszcz, M. A. (2015). Trajectories of depressive symptoms in old age: Integrating age-, pathology-, and mortality-related changes. *Psychology and aging*, *30*(4), 940.

Clarke, D. M., & Currie, K. C. (2009). Depression, anxiety and their relationship with chronic diseases: a review of the epidemiology, risk and treatment evidence. *Medical Journal of Australia*, *190*, S54–S60.

Cohen, J. (1978). Partialed products are interactions; partialed powers are curve components. *Psychological Bulletin*, *85*(4), 858.

Comstock, G. W., & Helsing, K. J. (1977). Symptoms of depression in two communities. *Psychological medicine*, *6*(4), 551–563.

Côté, S. M., Boivin, M., Liu, X., Nagin, D. S., Zoccolillo, M., & Tremblay, R. E. (2009). Depression and anxiety symptoms: onset, developmental course and risk factors during early childhood. *Journal of Child Psychology and Psychiatry*, *50*(10), 1201–1208.

Cudeck, R., & Harring, J. R. (2007). Analysis of nonlinear patterns of change with random coefficient models. *Annu. Rev. Psychol.*, *58*, 615–637.

Curran, P. J., Bauer, D. J., & Willoughby, M. T. (2004). Testing main effects and interactions in latent curve analysis. *Psychological Methods*, *9*(2), 220.

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: Combining autoregressive and latent curve models.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, *11*(2), 121–136.

De Beurs, E., Beekman, A., Geerlings, S., Deeg, D., Van Dyck, R., & Van Tilburg, W. (2001). On becoming depressed or anxious in late life: similar vulnerability factors but different effects of stressful life events. *The British Journal of Psychiatry*, *179*(5), 426–431.

Dekker, M. C., Ferdinand, R. F., Van Lang, N. D., Bongers, I. L., Van Der Ende, J., & Verhulst, F. C. (2007). Developmental trajectories of depressive symptoms from early childhood to late adolescence: gender differences and adult outcome. *Journal of Child Psychology and Psychiatry*, *48*(7), 657–666.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, *76*(374), 341–353.

de Oliveira, L. d. S. S. C. B., Souza, E. C., Rodrigues, R. A. S., Fett, C. A., & Piva, A. B. (2019). The effects of physical activity on anxiety, depression, and quality of life in elderly people living in the community. *Trends in psychiatry and psychotherapy*, *41*(1), 36–42.

Dew, M. A., Reynolds, C. F., Houck, P. R., Hall, M., Buysse, D. J., Frank, E., & Kupfer, D. J. (1997). Temporal profiles of the course of depression during treatment: predictors of pathways toward recovery in the elderly. *Archives of general psychiatry*, *54*(11), 1016–1024.

Dew, M. A., Whyte, E. M., Lenze, E. J., Houck, P. R., Mulsant, B. H., Pollock, B. G., ... Reynolds III, C. F., MD (2007). Recovery from major depression in older adults receiving augmentation of antidepressant pharmacotherapy. *American Journal of Psychiatry*, *164*(6), 892–899.

Diefenbach, G. J., & Goethe, J. (2006). Clinical interventions for late-life anxious depression. *Clinical Interventions in Aging*, *1*(1), 41.

Doraiswamy, P. M. (2001). Contemporary management of comorbid anxiety and depression in geriatric patients. *Journal of Clinical Psychiatry*, *62*(12), 30–35.

D'Unger, A., Land, K. M., & Nagin, P. (1998). How many latent classes of deliquent/criminal careers? results from mixed poisson regression analyses of the london, philadelphia, and racine cohorts studies. *American Journal of Sociology*, *103*, 1593–1630.

Ekström, J. (2011). A generalized definition of the polychoric correlation coefficient.

El-Gilany, A.-H., Elkhawaga, G. O., & Sarraf, B. B. (2018). Depression and its associated factors among elderly: A community-based study in egypt. *Archives of gerontology and geriatrics*, *77*, 103–107.

Elston, R., & Grizzle, J. E. (1962). Estimation of time-response curves and their confidence bands. *Biometrics*, *18*(2), 148–159.

Erdfelder, E. (1990). Deterministic developmental hypotheses, probabilistic rules of manifestation, and the analysis of finite mixture distributions. *Statistical methods in longitudinal research*, *2*, 471–509.

Everitt, B. (1984). Maximum likelihood estimation of the parameters in a mixture of two univariate normal distributions; a comparison of different algorithms. *The Statistician*, 205–215.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, *4*(3), 272.

Feng, X., Shaw, D. S., & Silk, J. S. (2008). Developmental trajectories of anxiety symptoms among boys across early and middle childhood. *Journal of abnormal psychology*, *117*(1), 32.

Fiske, A., Wetherell, J. L., & Gatz, M. (2009). Depression in older adults. *Annual review of clinical psychology*, *5*, 363–389.

Flint, A. J. (1994). Epidemiology and comorbidity of anxiety disorders in the elderly. *The American journal of psychiatry*.

Flint, A. J., & Rifat, S. L. (1997a). Anxious depression in elderly patients: response to antidepressant treatment. *The American Journal of Geriatric Psychiatry*, *5*(2), 107–

115.

Flint, A. J., & Rifat, S. L. (1997b). Two-year outcome of elderly patients with anxious depression. *Psychiatry research*, *66*(1), 23–31.

Forsell, Y. (2000). Predictors for depression, anxiety and psychotic symptoms in a very elderly population: data from a 3-year follow-up study. *Social psychiatry and psychiatric epidemiology*, *35*(6), 259–263.

Frankfurt, S., Frazier, P., Syed, M., & Jung, K. R. (2016). Using group-based trajectory and growth mixture modeling to identify classes of change trajectories. *The Counseling Psychologist*, *44*(5), 622–660.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York.

Frosio, I., Ferrigno, G., & Borghese, N. A. (2006). Enhancing digital cephalic radiography with mixture models and local gamma correction. *IEEE transactions on medical imaging*, *25*(1), 113–121.

Gardiner, J. C., Luo, Z., & Roman, L. A. (2009). Fixed effects, random effects and gee: what are the differences? *Statistics in medicine*, *28*(2), 221–239.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, *85*(410), 398–409.

Gelman, A., Carlin, J., Stern, H. S., & Rubin, D. (1995). Bayesian data analysis. 1995. *Chapman&Hall, London*.

Geman, S., & Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 401–414.

Geoffrey, M., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley-Interscience.

Ghosh, J. K., & Sen, P. K. (1984). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results.

Gilriches, Z. (1957). Hybrid corn: an exploration in the economics of technical change. *Econometrica*, *48*, 501–522.

Girgus, J. S., Yang, K., & Ferri, C. V. (2017). The gender difference in depression: are elderly women at greater risk for depression than elderly men? *Geriatrics*, *2*(4), 35.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, *73*(1), 43–56.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, 45–51.

Gompertz, B. (1833). A sketch of an analysis and notation applicable to the estimation of the value of life contingencies. In *Abstracts of the papers printed in the philosophical transactions of the royal society of london* (pp. 132–132).

Gould, C. E., O'Hara, R., Goldstein, M. K., & Beaudreau, S. A. (2016). Multimorbidity is associated with anxiety in older adults in the health and retirement study. *International journal of geriatric psychiatry*, *31*(10), 1105–1115.

Gower, R. M., & Richtárik, P. (2017). Randomized quasi-newton updates are linearly convergent matrix inversion algorithms. *SIAM Journal on Matrix Analysis and Applications*, *38*(4), 1380–1409.

Haviland, A. M., Jones, B. L., & Nagin, D. S. (2011). Group-based trajectory modeling extended to account for nonrandom participant attrition. *Sociological Methods & Research*, *40*(2), 367–390.

Henderson, A., Jorm, A., Korten, A., Jacomb, P., Christensen, H., & Rodgers, B. (1998). Symptoms of depression and anxiety during adult life: evidence for a decline in prevalence with age. *Psychological medicine*, *28*(6), 1321–1328.

Hermans, H., & Evenhuis, H. M. (2013). Factors associated with depression and anxiety in older adults with intellectual disabilities: results of the healthy ageing and intellectual disabilities study. *International journal of geriatric psychiatry*, *28*(7), 691–699.

Hersen, M., & Van Hasselt, V. B. (1992). Behavioral assessment and treatment of anxiety in the elderly. *Clinical Psychology Review*, *12*(6), 619–640.

Heun, R., Papassotiropoulos, A., & Ptok, U. (2000). Subthreshold depressive and anxiety disorders in the elderly. *European Psychiatry*, *15*(3), 173–182.

Himmelfarb, S., & Murrell, S. A. (1984). The prevalence and correlates of anxiety symptoms in older adults. *The Journal of psychology*, *116*(2), 159–167.

Holmes, S. E., Esterlis, I., Mazure, C. M., Lim, Y. Y., Ames, D., Rainey-Smith, S., ... others (2018). Trajectories of depressive and anxiety symptoms in older adults: a 6-year prospective cohort study. *International journal of geriatric psychiatry*, *33*(2), 405–413.

Hong, S.-I., Hasche, L., & Bowland, S. (2009). Structural relationships between social activities and longitudinal trajectories of depression among older adults. *The Gerontologist*, *49*(1), 1–11.

Horney, J., Osgood, D. W., & Marshall, I. H. (1995). Criminal careers in the short-term: Intra-individual variability in crime and its relation to local life circumstances. *American sociological review*, 655–673.

Hsu, H.-C. (2012). Group-based trajectories of depressive symptoms and the predictors in the older population. *International journal of geriatric psychiatry*, *27*(8), 854–862.

Huang, D. Y., Lanza, H. I., & Anglin, M. D. (2013). Association between adolescent substance use and obesity in young adulthood: a group-based dual trajectory analysis. *Addictive behaviors*, *38*(11), 2653–2660.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: relation to language input and gender. *Developmental psychology*, *27*(2), 236.

Hybels, C. F., Bennett, J. M., Landerman, L. R., Liang, J., Plassman, B. L., & Wu, B. (2016). Trajectories of depressive symptoms and oral health outcomes in a community sample of older adults. *International journal of geriatric psychiatry*, *31*(1), 83–91.

Isabella, S. (2017). *South korea is aging faster than any other developed country.* `https://qz.com/1066613/south-korea-demographic-time-bomb-its-aging-faster-than-any-other-developed-country-with-lowest-birth-rate-of-oecd-countries/`.

Isometsä, E. T., Henriksson, M. M., Aro, H. M., Heikkinen, M. E., Kuoppasalmi, K. I.,

& Lönnqvist, J. K. (1994). Suicide in major depression. *The American journal of psychiatry*.

Jang, Y., Small, B., & Haley, W. (2001). Cross-cultural comparability of the geriatric depression scale: comparison between older koreans and older americans. *Aging & Mental Health*, *5*(1), 31–37.

Jansen, R. (1993). Maximum likelihood in a generalized linear finite mixture model by using the em algorithm. *Biometrics*, 227–231.

Jones, B. L. (2020). *traj, group-based modeling of longitudinal data.* `https://www.andrew.cmu.edu/user/bjones/`.

Jones, B. L., & Nagin, D. S. (2007). Advances in group-based trajectory modeling and an sas procedure for estimating them. *Sociological methods & research*, *35*(4), 542–571.

Jones, B. L., & Nagin, D. S. (2011). A stata plugin for estimating group-based trajectory models.

Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A sas procedure based on mixture models for estimating developmental trajectories. *Sociological methods & research*, *29*(3), 374–393.

Juson, E. (2018). *11 signs and symptoms of anxiety disorders.* `https://www.healthline.com/nutrition/anxiety-disorder-symptoms`.

Kang, H., Bae, K. Y., Kim, S. W., Shin, H. Y., Shin, S., Yoon, J. S., & Kim, J.-M. (2017). Impact of anxiety and depression on physical health condition and disability in an elderly korean population. *Psychiatry investigation*, *14*(3), 240.

Kang, H., Bae, K. Y., Kim, S. W., Shin, S., Yoon, J. S., & Kim, J. M. (2016). Anxiety symptoms in korean elderly individuals: a two-year longitudinal community study. *International psychogeriatrics*, *28*(3), 423–433.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Kass, R. E., & Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, *90*(431), 928–934.

Kawachi, I., Sparrow, D., Vokonas, P. S., & Weiss, S. T. (1994). Symptoms of anxiety and risk of coronary heart disease. the normative aging study. *Circulation*, *90*(5), 2225–2229.

Kendall, P. C., & Watson, D. E. (1989). *Anxiety and depression: Distinctive and overlapping features.* Academic Press.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, 49–66.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., ... Kendler, K. S. (1994). Lifetime and 12-month prevalence of dsm-iii-r psychiatric disorders in the united states: results from the national comorbidity survey. *Archives of general psychiatry*, *51*(1), 8–19.

KHPS. (2020). *Korea health panel study.* `https://www.khp.re.kr:444/eng/main.do`.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, *9*(2), 137–163.

Kirchner, J. E., Zubritsky, C., Cody, M., Coakley, E., Chen, H., Ware, J. H., ... others (2007). Alcohol consumption among older adults in primary care. *Journal of General Internal Medicine*, *22*(1), 92–97.

Kirmayer, L. J., et al. (2001). Cultural variations in the clinical presentation of depression and anxiety: implications for diagnosis and treatment. *Journal of clinical psychiatry*, *62*, 22–30.

Knight, B. G., Nordhus, I. H., & Satre, D. D. (2003). Psychotherapy with older adults. *Handbook of Psychology*, 453–468.

Koo, S. K. (2018). Depression status in korea. *Osong public health and research perspectives*, *9*(4), 141.

Krishnan, K., Hays, J. C., & Blazer, D. G. (1997). Mri-defined vascular depression.

Kuchibhatla, M. N., Fillenbaum, G. G., Hybels, C. F., & Blazer, D. G. (2012). Trajectory classes of depressive symptoms in a community sample of older adults. *Acta Psychiatrica Scandinavica*, *125*(6), 492–501.

Kuo, S., Lin, K., Chen, C., Chuang, Y., & Chen, W. (2011). Depression trajectories and obesity among the elderly in taiwan. *Psychological medicine*, *41*(8), 1665.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14.

Lebowitz, B. D., Pearson, J. L., Schneider, L. S., Reynolds, C. F., Alexopoulos, G. S., Bruce, M. L., ... others (1997). Diagnosis and treatment of depression in late life: consensus statement update. *Jama*, *278*(14), 1186–1190.

le Cessie, S., & Van Houwelingen, J. (1994). Logistic regression for correlated binary data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *43*(1), 95–108.

Lenze, E. J. (2003). Comorbidity of depression and anxiety in the elderly. *Current Psychiatry Reports*, *5*(1), 62–67.

Lenze, E. J., Mulsant, B. H., Dew, M. A., Shear, M. K., Houck, P., Pollock, B. G., & Reynolds III, C. F. (2003). Good treatment outcomes in late-life depression with comorbid anxiety. *Journal of Affective Disorders*, *77*(3), 247–254.

Lenze, E. J., Mulsant, B. H., Shear, M. K., Alexopoulos, G. S., Frank, E., & Reynolds, C. F. (2001). Comorbidity of depression and anxiety disorders in later life. *Depression and Anxiety*, *14*(2), 86–93.

Lenze, E. J., Mulsant, B. H., Shear, M. K., Schulberg, H. C., Dew, M. A., Begley, A. E., ... Reynolds III, C. F. (2000). Comorbid anxiety disorders in depressed elderly patients. *American Journal of Psychiatry*, *157*(5), 722–728.

Lenze, E. J., & Wetherell, J. L. (2011). A lifespan view of anxiety disorders. *Dialogues in clinical neuroscience*, *13*(4), 381.

Liang, J., Xu, X., Quiñones, A. R., Bennett, J. M., & Ye, W. (2011). Multiple trajectories of depressive symptoms in middle and late life: racial/ethnic variations. *Psychology* *and aging*, *26*(4), 761.

Lim, H. J., Cheng, Y., Kabir, R., & Thorpe, L. (2020). Trajectories of depression and their predictors in a population-based study of korean older adults. *The International Journal of Aging and Human Development*, 0091415020944405.

Lin, N., Dean, A., & Ensel, W. M. (2013). *Social support, life events, and depression.* Academic Press.

Lindley, D. V., & Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–41.

Lindsay, B. G., & Lesperance, M. L. (1995). A review of semiparametric mixture models. *Journal of statistical planning and inference*, *47*(1-2), 29–39.

Liu, H. (2007). Growth curve models for zero-inflated count data: An application to smoking behavior. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(2), 247–279.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, *74*(4), 817–827.

Lunn, D., Barrett, J., Sweeting, M., & Thompson, S. (2013). Fully bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62*(4), 551–572.

Manela, M., Katona, C., & Livingston, G. (1996). How common are the anxiety disorders in old age? *International Journal of Geriatric Psychiatry*, *11*(1), 65–70.

Mantella, R. C., Butters, M. A., Dew, M. A., Mulsant, B. H., Begley, A. E., Tracey, B., ... Lenze, E. J. (2007). Cognitive impairment in late-life generalized anxiety disorder. *The American Journal of Geriatric Psychiatry*, *15*(8), 673–679.

Mayo-Wilson, E., Fusco, N., Li, T., Hong, H., Canner, J. K., Dickersin, K., et al. (2017). Multiple outcomes and analyses in clinical trials create challenges for interpretation and research synthesis. *Journal of clinical epidemiology*, *86*, 39–50.

McArdle, J. J. (2014). A structural modeling experiment with multiple growth functions. In *Abilities, motivation and methodology* (pp. 93–140). Routledge.

McCoach, D. B. (2010). Hierarchical linear modeling. *The reviewer's guide to quantitative methods in the social sciences*, 123–140.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Vol. 37). Boca Raton, FL: CRC press.

McLachlan, G., & Krishnan, T. (2008). The em algorithm and extensions. whiley series in probability and statistics. *John Wiley & Sons, New York, USA, second edition. Moon, TK (1996). The expectation-maximization algorithm. IEEE Signal Processing Magazine*, *13*(6), 47–60.

McLaughlin, K. A., & King, K. (2015). Developmental trajectories of anxiety and depression in early adolescence. *Journal of abnormal child psychology*, *43*(2), 311–323.

McLean, C. P., Asnaani, A., Litz, B. T., & Hofmann, S. G. (2011). Gender differences in anxiety disorders: prevalence, course of illness, comorbidity and burden of illness. *Journal of psychiatric research*, *45*(8), 1027–1035.

Mehta, K. M., Simonsick, E. M., Penninx, B. W., Schulz, R., Rubin, S. M., Satterfield, S., & Yaffe, K. (2003). Prevalence and correlates of anxiety symptoms in well-functioning older adults: findings from the health aging and body composition study. *Journal of the American Geriatrics Society*, *51*(4), 499–504.

Mehta, P. D., Neale, M. C., & Flay, B. R. (2004). Squeezing interval change from ordinal panel data: Latent growth curves with ordinal outcomes. *Psychological methods*, *9*(3), 301.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1), 107–122.

Montagnier, D., Dartigues, J.-F., Rouillon, F., Pérès, K., Falissard, B., & Onen, F. (2014). Ageing and trajectories of depressive symptoms in community-dwelling men and women. *International journal of geriatric psychiatry*, *29*(7), 720–729.

Mustillo, S., Worthman, C., Erkanli, A., Keeler, G., Angold, A., & Costello, E. J. (2003). Obesity and psychiatric disorder: developmental trajectories. *Pediatrics*, *111*(4), 851–859.

Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological methods*, *4*(2), 139.

Nagin, D. S. (2005). *Group-based modeling of development.* Cambridge, Massaachusetts: Harvard University Press.

Nagin, D. S., Jones, B. L., Passos, V. L., & Tremblay, R. E. (2018). Group-based multi-trajectory modeling. *Statistical methods in medical research*, *27*(7), 2015–2023.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed poisson model. *Criminology*, *31*(3), 327–362.

Nagin, D. S., & Odgers, C. L. (2010). Group-based trajectory modeling (nearly) two decades later. *Journal of quantitative criminology*, *26*(4), 445–453.

Nagin, D. S., & Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: a group-based method. *Psychological methods*, *6*(1), 18.

Nelder, J. A., & Baker, R. J. (2004). Generalized linear models. *Encyclopedia of statistical sciences*, *4*.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Newby, J. M., & Moulds, M. L. (2011). Intrusive memories of negative events in depression: Is the centrality of the event important? *Journal of behavior therapy and experimental psychiatry*, *42*(3), 277–283.

NLSCY. (2010). *National longitudinal survey of children and youth.* `https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey\&SDDS=4450`.

Norris, F. H., & Murrell, S. A. (1988). Prior experience as a moderator of disaster impact on anxiety symptoms in older adults. *American Journal of Community Psychology*, *16*(5), 665–683.

Ocram. (2014). *Generate random correlated data between a binary and a continuous variable.* Cross Validated. Retrieved from `https://stats.stackexchange.com/q/12858`

(URL:https://stats.stackexchange.com/q/12858 (version: 2014-12-24))

OECD-data. (2019). *Suicide rates.* `https://data.oecd.org/healthstat/suicide-rates.htm`.

Olino, T. M., Klein, D. N., Lewinsohn, P. M., Rohde, P., & Seeley, J. R. (2010). Latent trajectory classes of depressive and anxiety disorders from adolescence to adulthood: descriptions of classes and associations with risk factors. *Comprehensive psychiatry*, *51*(3), 224–235.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460.

Orcutt, H. K., Erickson, D. J., & Wolfe, J. (2004). The course of ptsd symptoms among gulf war veterans: a growth mixture modeling approach. *Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies*, *17*(3), 195–202.

Paterniti, S., Alperovitch, A., Ducimetiere, P., Dealberto, M.-J., Lepine, J.-P., & Bisserbe, J.-C. (1999). Anxiety but not depression is associated with elevated blood pressure in a community group of french elderly. *Psychosomatic medicine*, *61*(1), 77–83.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, *185*, 71–110.

Pilla, R. S., & Lindsay, B. G. (2001). Alternative em methods for nonparametric finite mixture models. *Biometrika*, *88*(2), 535–550.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, *4*(1), 12–35.

Piquero, A. R. (2008). Taking stock of developmental trajectories of criminal activity over the life course. In *The long view of crime: A synthesis of longitudinal research* (pp. 23–78). Springer.

Pirlich, M., Schütz, T., Kemps, M., Luhman, N., Minko, N., Lübke, H. J., … Lochs, H.

(2005). Social risk factors for hospital malnutrition. *Nutrition*, *21*(3), 295–300.

Rabe-Hesketh, S., & Skrondal, A. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* New York, NY: Chapman and Hall/CRC.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111–163.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, California: Sage.

Raudenbush, S. W., Yang, M.-L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of computational and Graphical Statistics*, *9*(1), 141–157.

Robertson, T. B. (1908). On the normal rate of growth of an individual, and its biochemical significance. *Archiv für Entwicklungsmechanik der Organismen*, *25*(4), 581–614.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, *90*(429), 106–121.

Roeder, K., Lynch, K. G., & Nagin, D. S. (1999). Modeling uncertainty in latent class membership: A case study in criminology. *Journal of the American Statistical Association*, *94*(447), 766–776.

Russo, J., Vitaliano, P. P., Brewer, D. D., Katon, W., & Becker, J. (1995). Psychiatric disorders in spouse caregivers of care recipients with alzheimer's disease and matched controls: A diathesis-stress model of psychopathology. *Journal of abnormal psychology*, *104*(1), 197.

Rzewuska, M., Mallen, C. D., Strauss, V. Y., Belcher, J., & Peat, G. (2015). One-year trajectories of depression and anxiety symptoms in older patients presenting in general practice with musculoskeletal pain: A latent class growth analysis. *Journal of psychosomatic research*, *79*(3), 195–201.

Sammel, M. D., Ryan, L. M., & Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(3), 667–678.

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, *277*(5328), 918–924.

Sánchez, C. I., García, M., Mayo, A., López, M. I., & Hornero, R. (2009). Retinal image analysis based on mixture models to detect hard exudates. *Medical Image Analysis*, *13*(4), 650–658.

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality and Quantity*, *24*(4), 367–386.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, *8*(1), 3–15.

Schlattmann, P. (2009). *Medical applications of finite mixture models.* Verlag Berlin Heidelberg: Springer.

Schoevers, R., Beekman, A., Deeg, D., Jonker, C., & Tilburg, W. v. (2003). Comorbidity and risk-patterns of depression, generalised anxiety disorder and mixed anxiety-depression in later life: results from the amstel study. *International journal of geriatric psychiatry*, *18*(11), 994–1001.

Schoevers, R. A., Deeg, D., Van Tilburg, W., & Beekman, A. (2005). Depression and generalized anxiety disorder: co-occurrence and longitudinal patterns in elderly patients. *The American Journal of Geriatric Psychiatry*, *13*(1), 31–39.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Shen, Y.-C., Zhang, M.-Y., Huang, Y.-Q., He, Y.-L., Liu, Z.-R., Cheng, H., … Kessler, R. C. (2006). Twelve-month prevalence, severity, and unmet need for treatment of mental disorders in metropolitan china. *Psychological medicine*, *36*(2), 257–267.

Shevlin, M., & Miles, J. N. (1998). Effects of sample size, model specification and factor loadings on the gfi in confirmatory factor analysis. *Personality and Individual differences*, *25*(1), 85–90.

Singer, J. D., et al. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford university press.

Smith, K. (2018). *Anxiety vs. depression: How to tell the difference.* `https://www.psycom.net/anxiety-depression-difference`.

So, Y., & Kuhfeld, W. F. (1995). Multinomial logit models. In *Sugi 20 conference proceedings* (pp. 1227–1234).

Spinhoven, P., van der Veen, D., Voshaar, R. O., & Comijs, H. (2017). Worry and cognitive control predict course trajectories of anxiety in older adults with late-life depression. *European Psychiatry*, *44*, 134–140.

Stanley, M. A., Hopko, D. R., Diefenbach, G. J., Bourland, S. L., Rodriguez, H., & Wagener, P. (2003). Cognitive–behavior therapy for late-life generalized anxiety disorder in primary care: Preliminary findings. *The American Journal of Geriatric Psychiatry*, *11*(1), 92–96.

Strenio, J. F., Weisberg, H. I., & Bryk, A. S. (1983). Empirical bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics*, 71–86.

Sutin, A. R., Terracciano, A., Milaneschi, Y., An, Y., Ferrucci, L., & Zonderman, A. B. (2013). The trajectory of depressive symptoms across the adult life span. *JAMA psychiatry*, *70*(8), 803–811.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528–540.

Taylor, M. G., & Lynch, S. M. (2004). Trajectories of impairment, social support, and depressive symptoms in later life. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *59*(4), S238–S246.

Teixeira-Pinto, A., Siddique, J., Gibbons, R., & Normand, S. L. (2009). Statistical approaches to modeling multiple outcomes in psychiatric studies. *Psychiatric annals*, *39*(7), 729–735.

Titterington, D. M., Smith, A. F., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* Oakland, CA: Wiley.

Touloumis, A. (2016). Simulating correlated binary and multinomial responses under marginal model specification: The simcormultres package. *The R Journal*, *8*(2), 79-91. Retrieved from `https://journal.r-project.org/archive/2016/RJ-2016-034/index.html`

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10.

Vassiliadis, V., Spyroglou, I., Rigas, A., Rosenberg, J., & Lindsay, K. (2019). Dealing with the phenomenon of quasi-complete separation and a goodness of fit test in logistic regression models in the case of long data sets. *Statistics in Biosciences*, *11*(3), 567–596.

Verbeke, G., Molenberghs, G., & Rizopoulos, D. (2010). Random effects models for longitudinal data. In *Longitudinal research with latent variables* (pp. 37–96). Springer.

Victor, M. N. (2014). A mixture model for longitudinal trajectories. *International Journal of Statistics and Applications*, *4*(4), 181–191.

Vink, D., Aartsen, M. J., & Schoevers, R. A. (2008). Risk factors for anxiety and depression in the elderly: a review. *Journal of affective disorders*, *106*(1-2), 29–44.

Vlassis, N., & Likas, A. (2002). A greedy em algorithm for gaussian mixture learning. *Neural processing letters*, *15*(1), 77–87.

Wang, P. (1994). *Mixed regression models for discrete data* (Unpublished doctoral dissertation). University of British Columbia, Vancouver, BC.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of mathematical psychology*, *44*(1), 92–107.

Watkins, J. (2018). South korea's mental health problem — that koreans don't admit. *Around the world*.

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

Weisenbach, S. L., Boore, L. A., & Kales, H. C. (2012). Depression and cognitive impairment in older adults. *Current psychiatry reports*, *14*(4), 280–288.

Wetherell, J. L., Gatz, M., & Craske, M. G. (2003). Treatment of generalized anxiety disorder in older adults. *Journal of consulting and clinical psychology*, *71*(1), 31.

Wetherell, J. L., Sorrell, J. T., Thorp, S. R., & Patterson, T. L. (2005). Psychological interventions for late-life anxiety: a review and early lessons from the calm study. *Journal of Geriatric Psychiatry and Neurology*, *18*(2), 72–82.

WHO. (2017). *Depression and other common mental disorders: global health estimates* (Tech. Rep.). World Health Organization.

WHO. (2018). *Investing in treatment for depression and anxiety leads to fourfold return.* `https://www.who.int/news-room/headlines/13-04-2016-investing-in-treatment-for-depression-and-anxiety-leads-to-fourfold-return`.

WHO. (2020a). *Depression.* `https://www.who.int/news-room/fact-sheets/detail/depression`.

WHO. (2020b). *Meatal disorders.* `https://www.who.int/en/news-room/fact-sheets/detail/mental-disorders`.

Wicklin, R. (2013). *Simulating data with sas.* SAS Institute.

Wiesner, M., & Kim, H. K. (2006). Co-occurring delinquency and depressive symptoms of adolescent boys and girls: a dual trajectory modeling approach. *Developmental psychology*, *42*(6), 1220.

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

Wolfe, J. H. (1967). *Normix: Computational methods for estimating the parameters of multivariate normal mixtures of distributions.* (Tech. Rep.). NAVAL PERSONNEL RESEARCH ACTIVITY SAN DIEGO CALIF.

Wolitzky-Taylor, K. B., Castriotta, N., Lenze, E. J., Stanley, M. A., & Craske, M. G. (2010).

Anxiety disorders in older adults: a comprehensive review. *Depression and anxiety*, *27*(2), 190–211.

Won, M.-R., & Choi, Y.-J. (2013). Are koreans prepared for the rapid increase of the single-household elderly? life satisfaction and depression of the single-household elderly in korea. *The Scientific World Journal*, *2013*.

Wu, J., Witkiewitz, K., McMahon, R. J., Dodge, K. A., Group, C. P. P. R., et al. (2010). A parallel process growth mixture model of conduct problems and substance use with risky sexual behavior. *Drug and alcohol dependence*, *111*(3), 207–214.

Yen, Y.-C., Rebok, G. W., Gallo, J. J., Jones, R. N., & Tennstedt, S. L. (2011). Depressive symptoms impair everyday problem-solving ability through cognitive abilities in late life. *The American Journal of Geriatric Psychiatry*, *19*(2), 142–150.

Yoon, J. Y., Brown, R. L., Bowers, B. J., Sharkey, S. S., & Horn, S. D. (2015). Longitudinal psychological outcomes of the small-scale nursing home model: a latent growth curve zero-inflated poisson model. *International psychogeriatrics/IPA*, *27*(6), 1009.

You, K. S., Lee, H.-O., Fitzpatrick, J. J., Kim, S., Marui, E., Lee, J. S., & Cook, P. (2009). Spirituality, depression, living alone, and perceived health among korean older adults in the community. *Archives of Psychiatric Nursing*, *23*(4), 309–322.

Ziegel, E. R. (2004). Modelling binary data. *Technometrics*, *46*(1), 119–121.

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta psychiatrica scandinavica*, *67*(6), 361–370.

Zuckerman, H., Pan, Z., Park, C., Brietzke, E., Musial, N., Shariq, A. S., … others (2018). Recognition and treatment of cognitive dysfunction in major depressive disorder. *Frontiers in psychiatry*, *9*, 655.

Appendices

Appendix A SIMULATION RESULTS

A.1 Tables and figures of simulation results with two continuous longitudinal outcomes

Table A.1: Estimation in parameters of Outcome 1 on each polynomial trajectory in GBTM based on 500 simulated data sets

N	Parameters	True parameter value	Mean estimates	Mean SE*	Bias	P-value
500	Intercept1	5	5.000112	0.025847	0.000112	<0.0001
	Intercept2	20	20.00246	0.104992	0.002463	<0.0001
	linear2	-2	-2.00094	0.031656	-0.000943	<0.0001
	Intercept3	5	5.008564	0.148481	0.008564	<0.0001
	linear3	1.5	1.498366	0.044769	-0.001634	<0.0001
	Intercept4	28	28.00017	0.063312	0.000168	<0.0001
	Sigma	1	0.999254	0.014157	-0.000746	<0.0001
2000	Intercept1	5	5.000303	0.012916	0.000303	<0.0001
	Intercept2	20	19.99874	0.052464	-0.001257	<0.0001
	linear2	-2	-1.99967	0.015819	0.000334	<0.0001
	Intercept3	5	4.997734	0.074195	-0.002266	<0.0001
	linear3	1.5	1.500927	0.022371	0.000927	<0.0001
	Intercept4	28	27.9993	0.031637	-0.000702	<0.0001
	Sigma	1	0.999999	0.007074	-0.000001	<0.0001
4000	Intercept1	5	4.999974	0.009133	-0.000026	<0.0001
	Intercept2	20	19.99721	0.0371	-0.002788	<0.0001
	linear2	-2	-1.9994	0.011186	0.000604	<0.0001
	Intercept3	5	5.000508	0.052467	0.000508	<0.0001
	linear3	1.5	1.499828	0.015819	-0.000172	<0.0001
	Intercept4	28	28.00044	0.022372	0.000439	<0.0001
	Sigma	1	1.000276	0.005003	0.000276	<0.0001

* SE = Standard Error

Note: p-values are calculated based on the average mean and SE

Table A.2: Estimation of parameters for Outcome 1 on each polynomial trajectory in GBDTM and GBMTM with sample size N = 500 based on 500 simulated data sets

			GBDTM				GBMTM			
ρ**	Parameter	TPV#	Mean Estimates	Mean SE*	Bias	P-value	Mean Estimates	Mean SE*	Bias	P-value
0.1	Intercept1	5	5.0001	0.0258	5.0001	<0.0001	5.0001	0.0258	5.0001	<0.0001
	Intercept2	20	20.0025	0.1050	20.0025	<0.0001	20.0025	0.1050	20.0025	<0.0001
	Linear2	-2	-2.0009	0.0316	-2.0009	<0.0001	-2.0009	0.0316	-2.0009	<0.0001
	Intercept3	5	5.0086	0.1484	5.0086	<0.0001	5.0086	0.1485	5.0086	<0.0001
	Linear3	1.5	1.4984	0.0448	1.4984	<0.0001	1.4984	0.0448	1.4984	<0.0001
	Intercept4	28	28.0002	0.0633	28.0002	<0.0001	28.0002	0.0633	28.0002	<0.0001
	Sigma	1	0.9993	0.0142	0.9993	<0.0001	0.9993	0.0142	0.9993	<0.0001
0.2	Intercept1	5	5.0001	0.0258	5.0001	<0.0001	5.0001	0.0258	5.0001	<0.0001
	Intercept2	20	20.0025	0.1050	20.0025	<0.0001	20.0024	0.1050	20.0024	<0.0001
	Linear2	-2	-2.0009	0.0317	-2.0009	<0.0001	-2.0010	0.0317	-2.0010	<0.0001
	Intercept3	5	5.0086	0.1485	5.0086	<0.0001	5.0088	0.1485	5.0088	<0.0001
	Linear3	1.5	1.4984	0.0448	1.4984	<0.0001	1.4983	0.0448	1.4983	<0.0001
	Intercept4	28	28.0002	0.0633	28.0002	<0.0001	28.0001	0.0633	28.0001	<0.0001
	Sigma	1	0.9993	0.0140	0.9993	<0.0001	0.9993	0.0142	0.9993	<0.0001
0.4	Intercept1	5	5.0001	0.0258	5.0001	<0.0001	5.0000	0.0258	5.0000	<0.0001
	Intercept2	20	20.0025	0.1049	20.0025	<0.0001	20.0023	0.1050	20.0023	<0.0001
	Linear2	-2	-2.0009	0.0316	-2.0009	<0.0001	-2.0009	0.0317	-2.0009	<0.0001
	Intercept3	5	5.0086	0.1484	5.0086	<0.0001	5.0089	0.1485	5.0089	<0.0001
	Linear3	1.5	1.4984	0.0447	1.4984	<0.0001	1.4983	0.0448	1.4983	<0.0001
	Intercept4	28	28.0002	0.0633	28.0002	<0.0001	28.0002	0.0633	28.0002	<0.0001
	Sigma	1	0.9993	0.0141	0.9993	<0.0001	0.9992	0.0142	0.9992	<0.0001
0.6	Intercept1	5	5.0001	0.0259	5.0001	<0.0001	5.0001	0.0258	5.0001	<0.0001
	Intercept2	20	20.0025	0.1051	20.0025	<0.0001	20.0024	0.1050	20.0024	<0.0001
	Linear2	-2	-2.0009	0.0317	-2.0009	<0.0001	-2.0010	0.0317	-2.0010	<0.0001
	Intercept3	5	5.0086	0.1486	5.0086	<0.0001	5.0082	0.1485	5.0082	<0.0001
	Linear3	1.5	1.4984	0.0448	1.4984	<0.0001	1.4984	0.0448	1.4984	<0.0001
	Intercept4	28	28.0002	0.0634	28.0002	<0.0001	28.0001	0.0633	28.0001	<0.0001
	Sigma	1	0.9993	0.0138	0.9993	<0.0001	0.9993	0.0142	0.9993	<0.0001

* SE = Standard Error
** ρ = Correlation Level
# TPV = True Parameter Value
Note: p-values are calculated based on the average mean and SE

Table A.3: Estimation of parameters for Outcome 1 on each polynomial trajectory in GBDTM and GBMTM with sample size N = 2000 based on 500 simulated data sets

			GBDTM				GBMTM			
ρ**	Parameter	TPV#	Mean Estimates	Mean SE*	Bias	P-value	Mean Estimates	Mean SE*	Bias	P-value
0.1	Intercept1	5	5.0003	0.0129	0.0003	<0.0001	5.0003	0.0129	0.0003	<0.0001
	Intercept2	20	19.9987	0.0525	-0.0013	<0.0001	19.9987	0.0525	-0.0013	<0.0001
	Linear2	-2	-1.9997	0.0158	0.0003	<0.0001	-1.9997	0.0158	0.0003	<0.0001
	Intercept3	5	4.9977	0.0742	-0.0023	<0.0001	4.9977	0.0742	-0.0023	<0.0001
	Linear3	1.5	1.5009	0.0224	0.0009	<0.0001	1.5009	0.0224	0.0009	<0.0001
	Intercept4	28	27.9993	0.0316	-0.0007	<0.0001	27.9993	0.0316	-0.0007	<0.0001
	Sigma	1	1.0000	0.0071	0.0000	<0.0001	1.0000	0.0071	0.0000	<0.0001
0.2	Intercept1	5	5.0003	0.0129	0.0003	<0.0001	5.0003	0.0129	0.0003	<0.0001
	Intercept2	20	19.9987	0.0525	-0.0013	<0.0001	19.9988	0.0525	-0.0012	<0.0001
	Linear2	-2	-1.9997	0.0158	0.0003	<0.0001	-1.9997	0.0158	0.0003	<0.0001
	Intercept3	5	4.9977	0.0742	-0.0023	<0.0001	4.9986	0.0742	-0.0014	<0.0001
	Linear3	1.5	1.5009	0.0224	0.0009	<0.0001	1.5007	0.0224	0.0007	<0.0001
	Intercept4	28	27.9993	0.0316	-0.0007	<0.0001	27.9993	0.0316	-0.0007	<0.0001
	Sigma	1	1.0000	0.0071	0.0000	<0.0001	1.0000	0.0071	0.0000	<0.0001
0.4	Intercept1	5	5.0003	0.0129	0.0003	<0.0001	5.0004	0.0129	0.0004	<0.0001
	Intercept2	20	19.9987	0.0525	-0.0013	<0.0001	19.9987	0.0525	-0.0013	<0.0001
	Linear2	-2	-1.9997	0.0158	0.0003	<0.0001	-1.9997	0.0158	0.0003	<0.0001
	Intercept3	5	4.9977	0.0742	-0.0023	<0.0001	4.9978	0.0742	-0.0022	<0.0001
	Linear3	1.5	1.5009	0.0224	0.0009	<0.0001	1.5009	0.0224	0.0009	<0.0001
	Intercept4	28	27.9993	0.0316	-0.0007	<0.0001	27.9994	0.0316	-0.0006	<0.0001
	Sigma	1	1.0000	0.0071	0.0000	<0.0001	1.0000	0.0071	0.0000	<0.0001
0.6	Intercept1	5	5.0003	0.0129	0.0003	<0.0001	5.0003	0.0129	0.0003	<0.0001
	Intercept2	20	19.9987	0.0525	-0.0013	<0.0001	19.9988	0.0525	-0.0012	<0.0001
	Linear2	-2	-1.9997	0.0158	0.0003	<0.0001	-1.9997	0.0158	0.0003	<0.0001
	Intercept3	5	4.9977	0.0742	-0.0023	<0.0001	4.9981	0.0742	-0.0019	<0.0001
	Linear3	1.5	1.5009	0.0224	0.0009	<0.0001	1.5009	0.0224	0.0009	<0.0001
	Intercept4	28	27.9993	0.0316	-0.0007	<0.0001	27.9992	0.0316	-0.0008	<0.0001
	Sigma	1	1.0000	0.0071	0.0000	<0.0001	1.0000	0.0071	0.0000	<0.0001

* SE = Standard Error

** ρ = Correlation Level

# TPV = True Parameter Value

Note: p-values are calculated based on the average mean and SE

Table A.4: Estimation of parameters for Outcome 1 on each polynomial trajectory in GBDTM and GBMTM with sample size N = 4000 based on 500 simulated data sets

			GBDTM				GBMTM			
ρ**	Parameter	TPV#	Mean Estimates	Mean SE*	Bias	P-value	Mean Estimates	Mean SE*	Bias	P-value
0.1	Intercept1	5	5.0000	0.0091	0.0000	<0.0001	5.0000	0.0091	0.0000	<0.0001
	Intercept2	20	19.9972	0.0371	-0.0028	<0.0001	19.9972	0.0371	-0.0028	<0.0001
	Linear2	-2	-1.9994	0.0112	0.0006	<0.0001	-1.9994	0.0112	0.0006	<0.0001
	Intercept3	5	5.0005	0.0525	0.0005	<0.0001	5.0005	0.0525	0.0005	<0.0001
	Linear3	1.5	1.4998	0.0158	-0.0002	<0.0001	1.4998	0.0158	-0.0002	<0.0001
	Intercept4	28	28.0004	0.0224	0.0004	<0.0001	28.0004	0.0224	0.0004	<0.0001
	Sigma	1	1.0003	0.0050	0.0003	<0.0001	1.0003	0.0050	0.0003	<0.0001
0.2	Intercept1	5	5.0000	0.0091	0.0000	<0.0001	5.0000	0.0091	0.0000	<0.0001
	Intercept2	20	19.9972	0.0371	-0.0028	<0.0001	19.9972	0.0371	-0.0028	<0.0001
	Linear2	-2	-1.9994	0.0112	0.0006	<0.0001	-1.9994	0.0112	0.0006	<0.0001
	Intercept3	5	5.0005	0.0525	0.0005	<0.0001	5.0005	0.0525	0.0005	<0.0001
	Linear3	1.5	1.4998	0.0158	-0.0002	<0.0001	1.4998	0.0158	-0.0002	<0.0001
	Intercept4	28	28.0004	0.0224	0.0004	<0.0001	28.0004	0.0224	0.0004	<0.0001
	Sigma	1	1.0003	0.0050	0.0003	<0.0001	1.0003	0.0050	0.0003	<0.0001
0.4	Intercept1	5	5.0000	0.0091	0.0000	<0.0001	5.0000	0.0091	0.0000	<0.0001
	Intercept2	20	19.9972	0.0371	-0.0028	<0.0001	19.9970	0.0371	-0.0030	<0.0001
	Linear2	-2	-1.9994	0.0112	0.0006	<0.0001	-1.9994	0.0112	0.0006	<0.0001
	Intercept3	5	5.0005	0.0525	0.0005	<0.0001	5.0002	0.0525	0.0002	<0.0001
	Linear3	1.5	1.4998	0.0158	-0.0002	<0.0001	1.4999	0.0158	-0.0001	<0.0001
	Intercept4	28	28.0004	0.0224	0.0004	<0.0001	28.0005	0.0224	0.0005	<0.0001
	Sigma	1	1.0003	0.0050	0.0003	<0.0001	1.0002	0.0050	0.0002	<0.0001
0.6	Intercept1	5	5.0000	0.0091	0.0000	<0.0001	4.9999	0.0091	-0.0001	<0.0001
	Intercept2	20	19.9972	0.0371	-0.0028	<0.0001	19.9967	0.0371	-0.0033	<0.0001
	Linear2	-2	-1.9994	0.0112	0.0006	<0.0001	-1.9993	0.0112	0.0007	<0.0001
	Intercept3	5	5.0005	0.0525	0.0005	<0.0001	5.0005	0.0525	0.0005	<0.0001
	Linear3	1.5	1.4998	0.0158	-0.0002	<0.0001	1.4998	0.0158	-0.0002	<0.0001
	Intercept4	28	28.0004	0.0224	0.0004	<0.0001	28.0003	0.0224	0.0003	<0.0001
	Sigma	1	1.0003	0.0050	0.0003	<0.0001	1.0002	0.0050	0.0002	<0.0001

[*] SE = Standard Error

[**] ρ = Correlation Level

[#] TPV = True Parameter Value

Note: p-values are calculated based on the average mean and SE

Table A.5: Estimation of parameters for Outcome 2 on each polynomial trajectory in GBTM, GBDTM and GBMTM with sample size N = 500 based on 500 simulated data sets

		GBTM			GBDTM			GBMTM		
ρ**	Parameter	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value
0.1	Intercept1	11.14243	0.24347	<0.0001	10.94462	0.25599	<0.0001	12.15902	0.38271	<0.0001
	Linear1							-0.30787	0.11539	0.008
	Intercept2	19.32823	2.37897	<0.0001	17.32663	1.65447	<0.0001	13.30223	0.66287	<0.0001
	Linear2	-1.88081	0.62391	0.002	-1.27227	0.42461	0.003	-0.46967	0.19986	0.019
	Intercept3							11.55518	0.39973	<0.0001
	Linear3									
	Intercept4							12.99391	0.39973	<0.0001
	Linear4									
	Sigma	6.20376	0.09513	<0.0001	6.20886	0.09464	<0.0001	6.31010	0.08938	<0.0001
0.2	Intercept1	12.17200	0.56509	<0.0001	12.32673	0.39984	<0.0001	12.36543	0.38530	<0.0001
	Linear1	-0.16283	0.13988	0.244	-0.28154	0.11024	0.011	-0.30739	0.11617	0.008
	Intercept2	20.41423	2.57990	<0.0001	16.91359	0.96551	<0.0001	14.69860	0.66735	<0.0001
	Linear2	-1.49169	0.67564	0.027	-0.56576	0.27440	0.039	-0.61272	0.20121	0.002
	Intercept3							12.27933	0.94378	<0.0001
	Linear3							-0.05431	0.28456	0.845
	Intercept4							16.16347	0.94378	<0.0001
	Linear4							-0.28768	0.28456	0.312
	Sigma	6.30194	0.09766	<0.0001	6.30149	0.09375	<0.0001	6.35149	0.08999	<0.0001
0.4	Intercept1	12.13528	0.65973	<0.0001	12.03961	0.44114	<0.0001	12.38994	0.37358	<0.0001
	Linear1	-0.11481	0.16237	0.48	-0.17428	0.12052	0.148	-0.28427	0.11264	0.012
	Intercept2	21.51765	2.57710	<0.0001	17.69697	0.91788	<0.0001	17.57014	0.64706	<0.0001
	Linear2	-1.86187	0.66558	0.005	-0.97365	0.24322	<0.0001	-0.97685	0.19509	<0.0001
	Intercept3	20.64094	1.83014	<0.0001	20.74687	0.93524	<0.0001	12.28144	0.91508	<0.0001
	Linear3	-0.02337	0.55957	0.967	-0.21692	0.28162	0.441	0.25704	0.27591	0.352
	Intercept4							20.66413	0.91508	<0.0001
	Linear4							-0.22318	0.27591	0.419
	Sigma	6.09891	0.09650	<0.0001	6.11187	0.09170	<0.0001	6.15833	0.08725	<0.0001
0.6	Intercept1	11.55588	0.45808	<0.0001	11.59849	0.33427	<0.0001	11.82820	0.33677	<0.0001
	Linear1	-0.06943	0.11638	0.551	-0.12069	0.09931	0.224	-0.23851	0.10154	0.019
	Intercept2	20.55217	1.13663	<0.0001	19.61730	0.64650	<0.0001	20.19467	0.58331	<0.0001
	Linear2	-1.43899	0.27719	<0.0001	-1.18401	0.18571	<0.0001	-1.36438	0.17587	<0.0001
	Intercept3	25.22917	0.91249	<0.0001	25.08799	0.82094	<0.0001	11.73829	0.82492	<0.0001
	Linear3	-0.22610	0.27274	0.407	-0.19719	0.24614	0.423	0.61195	0.24872	0.014
	Intercept4							25.06169	0.82492	<0.0001
	Linear4							-0.19545	0.24872	0.432
	Sigma	5.55654	0.08497	<0.0001	5.55254	0.08064	<0.0001	5.55161	0.07865	<0.0001

*SE = Standard Error

** ρ = Correlation Level

Note: p-values are calculated based on the average mean and SE

Table A.6: Estimation of parameters for Outcome 2 on each polynomial trajectory in GBTM, GBDTM and GBMTM with sample size N = 2000 based on 500 simulated data sets

		GBTM			GBDTM			GBMTM		
ρ**	Parameter	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value
0.1	Intercept1	11.17759	0.10708	<0.0001	10.88372	0.14262	<0.0001	12.16308	0.19119	<0.0001
	Linear1							-0.31206	0.05765	<0.0001
	Intercept2	18.99432	1.16338	<0.0001	16.34449	0.91462	<0.0001	13.25117	0.33116	<0.0001
	Linear2	-1.80098	0.30809	<0.0001	-1.01763	0.21836	<0.0001	-0.45190	0.09985	<0.0001
	Intercept3							12.22020	0.46833	<0.0001
	Linear3							-0.21008	0.14121	0.137
	Intercept4							13.91059	0.46833	<0.0001
	Linear4							-0.29711	0.14121	0.035
	Sigma	6.20861	0.04744	<0.0001	6.21970	0.04809	<0.0001	6.31207	0.04465	<0.0001
0.2	Intercept1	12.18309	0.28510	<0.0001	12.38086	0.19466	<0.0001	12.35084	0.19257	<0.0001
	Linear1	-0.16546	0.06970	0.018	-0.29234	0.05379	<0.0001	-0.30426	0.05806	<0.0001
	Intercept2	20.29445	1.19166	<0.0001	16.74554	0.46051	<0.0001	14.71712	0.33354	<0.0001
	Linear2	-1.55305	0.29962	<0.0001	-0.54493	0.13075	<0.0001	-0.61858	0.10057	<0.0001
	Intercept3							12.35811	0.47170	<0.0001
	Linear3							-0.07451	0.14222	0.6
	Intercept4							16.11928	0.47170	<0.0001
	Linear4							-0.27637	0.14222	0.052
	Sigma	6.30714	0.04900	<0.0001	6.31508	0.04670	<0.0001	6.35749	0.04497	<0.0001
0.4	Intercept1	12.23223	0.32197	<0.0001	12.10871	0.22043	<0.0001	12.38289	0.18670	<0.0001
	Linear1	-0.12883	0.07847	0.101	-0.18764	0.05937	0.002	-0.28311	0.05629	<0.0001
	Intercept2	21.35967	1.43695	<0.0001	17.56298	0.45888	<0.0001	17.56650	0.32337	<0.0001
	Linear2	-1.87544	0.36436	<0.0001	-0.93231	0.11794	<0.0001	-0.97791	0.09750	<0.0001
	Intercept3	20.64149	0.81366	<0.0001	20.78496	0.46653	<0.0001	12.32452	0.45732	<0.0001
	Linear3	-0.12518	0.26098	0.631	-0.23803	0.13963	0.088	0.24023	0.13789	0.081
	Intercept4							20.70349	0.45732	<0.0001
	Linear4							-0.23676	0.13789	0.086
	Sigma	6.11275	0.04847	<0.0001	6.13197	0.04617	<0.0001	6.16366	0.04360	<0.0001
0.6	Intercept1	11.58802	0.22373	<0.0001	11.60353	0.16577	<0.0001	11.81382	0.16825	<0.0001
	Linear1	-0.07306	0.05699	0.2	-0.11946	0.04935	0.015	-0.23565	0.05073	<0.0001
	Intercept2	20.62778	0.56167	<0.0001	19.61711	0.32462	<0.0001	20.20968	0.29141	<0.0001
	Linear2	-1.45285	0.13690	<0.0001	-1.18474	0.09293	<0.0001	-1.36882	0.08786	<0.0001
	Intercept3	25.19402	0.45427	<0.0001	25.05570	0.41244	<0.0001	11.73262	0.41212	<0.0001
	Linear3	-0.22736	0.13579	0.094	-0.20294	0.12412	0.102	0.60953	0.12426	<0.0001
	Intercept4							25.04092	0.41212	<0.0001
	Linear4							-0.20163	0.12426	0.105
	Sigma	5.55925	0.04243	<0.0001	5.56059	0.04030	<0.0001	5.55454	0.03929	<0.0001

[*] SE = Standard Error

[**] ρ = Correlation Level

Note: p-values are calculated based on the average mean and SE

Table A.7: Estimation of parameters for Outcome 2 on each polynomial trajectory in GBTM, GBDTM and GBMTM with sample size N = 4000 based on 500 simulated data sets

		GBTM			GBDTM			GBMTM		
ρ**	Parameter	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value
0.1	Intercept1	11.17556	0.07371	<0.0001	10.89472	0.10117	<0.0001	12.15828	0.13523	<0.0001
	Linear1							-0.31162	0.04077	<0.0001
	Intercept2	19.10711	0.81196	<0.0001	16.41428	0.71249	<0.0001	13.26844	0.23422	<0.0001
	Linear2	-1.82617	0.21558	<0.0001	-1.03054	0.16800	<0.0001	-0.45440	0.07062	<0.0001
	Intercept3							12.16496	0.33124	<0.0001
	Linear3							-0.20393	0.09987	0.042
	Intercept4							13.94418	0.33124	<0.0001
	Linear4							-0.30954	0.09987	0.002
	Sigma	6.20839	0.03341	<0.0001	6.21986	0.03429	<0.0001	6.31507	0.03158	<0.0001
0.2	Intercept1	12.18907	0.19897	<0.0001	12.38891	0.13609	<0.0001	12.33911	0.13609	<0.0001
	Linear1	-0.16744	0.04856	0.001	-0.29394	0.03767	<0.0001	-0.30224	0.04103	<0.0001
	Intercept2	20.22717	0.83359	<0.0001	16.68416	0.31771	<0.0001	14.72557	0.23572	<0.0001
	Linear2	-1.53699	0.20708	<0.0001	-0.53351	0.09098	<0.0001	-0.62218	0.07107	<0.0001
	Intercept3							12.32339	0.33336	<0.0001
	Linear3							-0.06129	0.10051	0.542
	Intercept4							16.12047	0.33336	<0.0001
	Linear4							-0.27343	0.10051	0.007
	Sigma	6.30650	0.03463	<0.0001	6.31608	0.03299	<0.0001	6.35549	0.03178	<0.0001
0.4	Intercept1	12.23950	0.22218	<0.0001	12.08674	0.16921	<0.0001	12.36293	0.13209	<0.0001
	Linear1	-0.12945	0.05407	0.017	-0.18110	0.04391	0.0003	-0.27808	0.03983	<0.0001
	Intercept2	21.37750	1.00479	<0.0001	17.60055	0.34655	<0.0001	17.58927	0.22878	<0.0001
	Linear2	-1.88774	0.24978	<0.0001	-0.94148	0.08706	<0.0001	-0.98733	0.06898	<0.0001
	Intercept3	20.61455	0.54429	<0.0001	20.71627	0.33119	<0.0001	12.29097	0.32355	<0.0001
	Linear3	-0.11541	0.17886	0.519	-0.22189	0.09905	0.025	0.25117	0.09755	0.01
	Intercept4							20.64587	0.32355	<0.0001
	Linear4							-0.22366	0.09755	0.022
	Sigma	6.11703	0.03430	<0.0001	6.13767	0.03405	<0.0001	6.16844	0.03085	<0.0001
0.6	Intercept1	11.59743	0.15800	<0.0001	11.61647	0.11702	<0.0001	11.81541	0.11898	<0.0001
	Linear1	-0.07681	0.04024	0.056	-0.12282	0.03483	0.0004	-0.23672	0.03587	<0.0001
	Intercept2	20.64707	0.39625	<0.0001	19.65308	0.23037	<0.0001	20.24129	0.20608	<0.0001
	Linear2	-1.46007	0.09631	<0.0001	-1.19587	0.06580	<0.0001	-1.37900	0.06213	<0.0001
	Intercept3	25.19054	0.32071	<0.0001	25.04292	0.29247	<0.0001	11.72714	0.29144	<0.0001
	Linear3	-0.22606	0.09600	0.019	-0.19856	0.08813	0.024	0.60669	0.08787	<0.0001
	Intercept4							25.02482	0.29144	<0.0001
	Linear4							-0.19842	0.08787	0.024
	Sigma	5.56182	0.03001	<0.0001	5.56479	0.02851	<0.0001	5.55621	0.02779	<0.0001

*SE = Standard Error
** ρ = Correlation Level
Note: p-values are calculated based on the average mean and SE

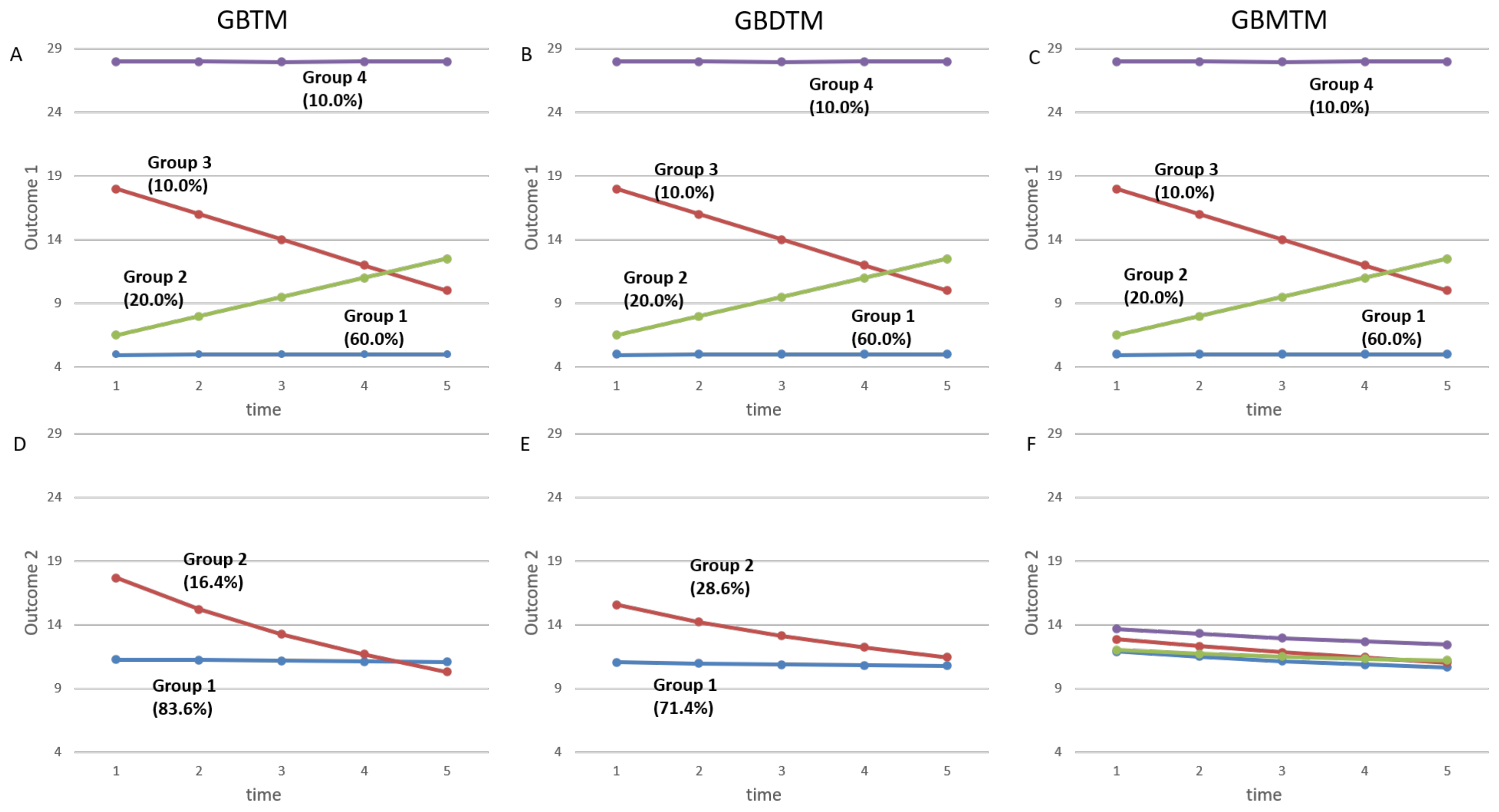


Figure A.1: Average trajectories of three trajectory models from simulation study with sample size N = 4000 and correlation coefficient 0.1 based on 500 simulations

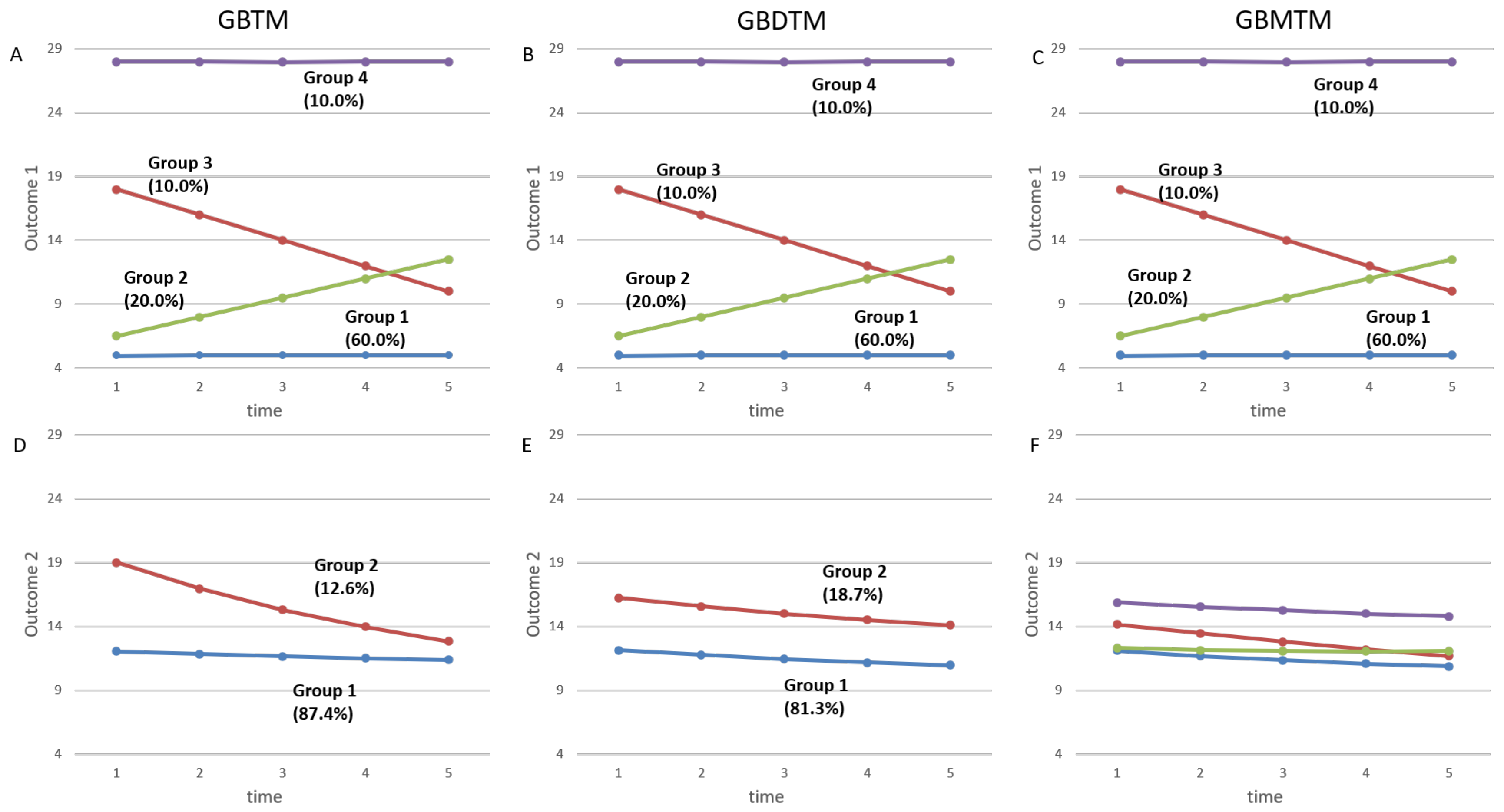


Figure A.2: Average trajectories of three trajectory models from simulation study with sample size N = 4000 and correlation coefficient 0.2 based on 500 simulations

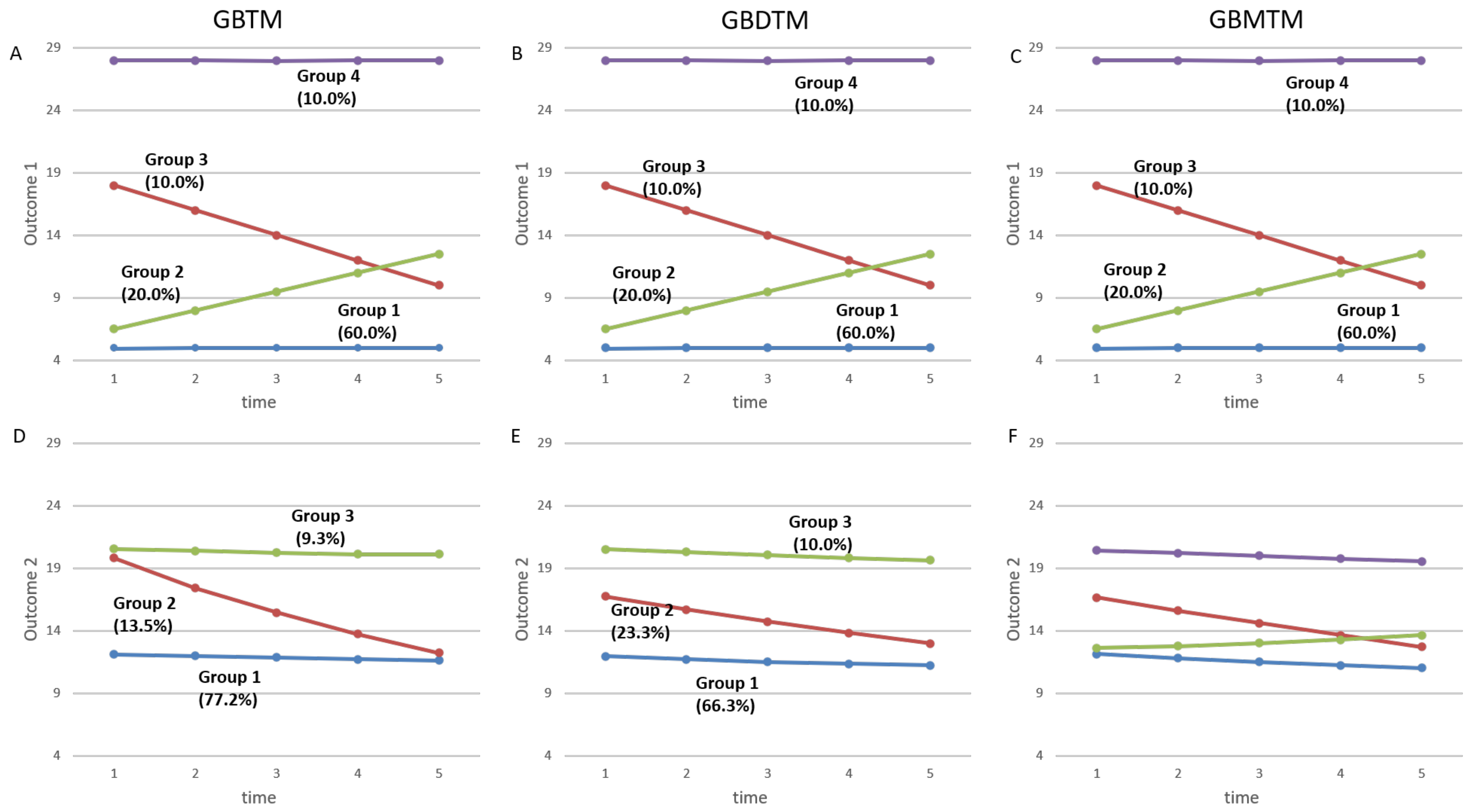


Figure A.3: Average trajectories of three trajectory models from simulation study with sample size N = 4000 and correlation coefficient 0.4 based on 500 simulations

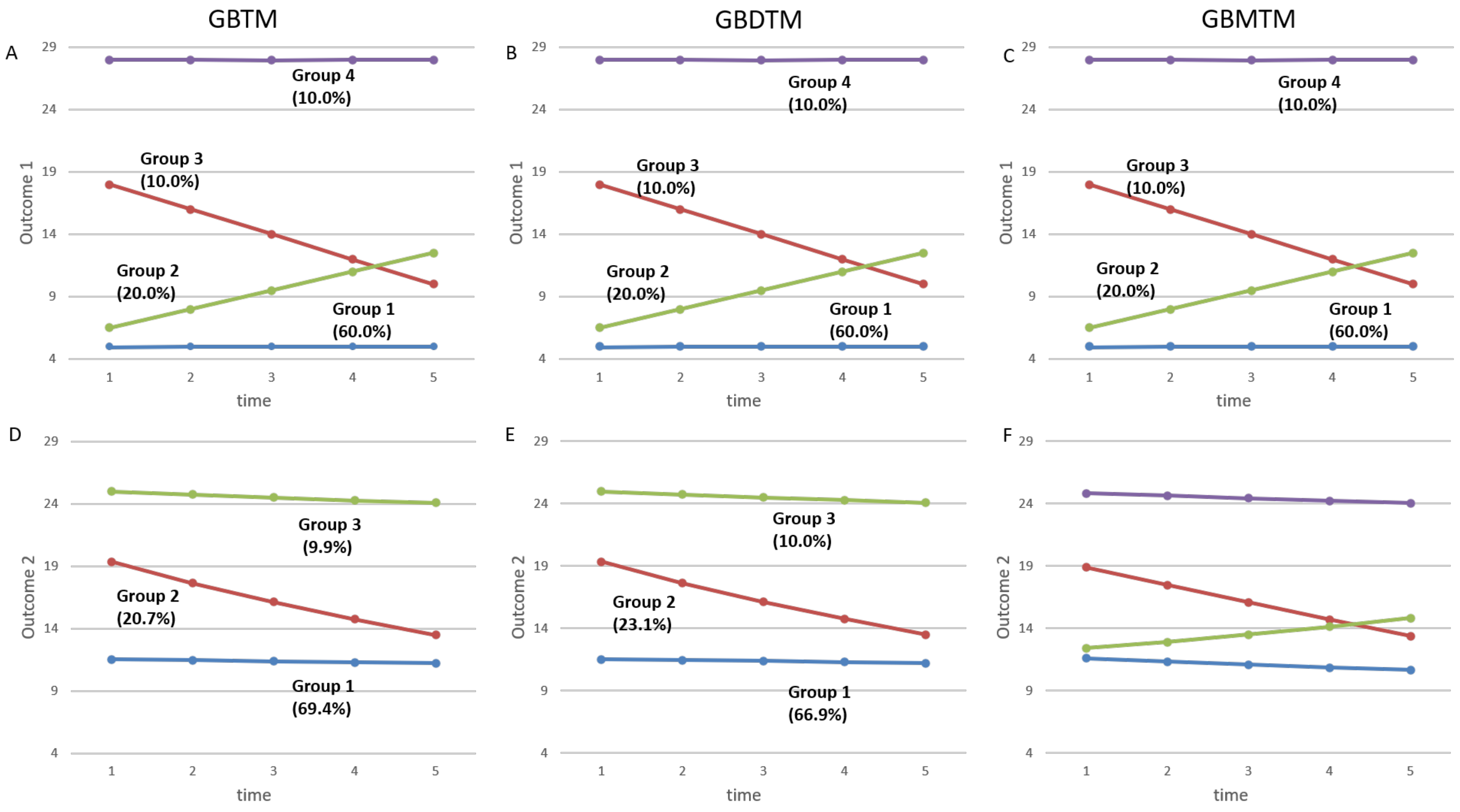


Figure A.4: Average trajectories of three trajectory models from simulation study with sample size N = 4000 and correlation coefficient 0.6 based on 500 simulations

A.2 Tables and figures of simulation results with one continuous and one binary longitudinal outcome

Table A.8: Estimation of parameters of Outcome 1 on each polynomial trajectory in GBTM based on 500 simulated data sets

N	Parameters	True parameter value	Mean estimates	Mean SE*	Bias	P-value
500	Intercept1	5	5.000112	0.025847	0.000112	<0.0001
	Intercept2	20	20.00246	0.104992	0.002463	<0.0001
	linear2	-2	-2.00094	0.031656	-0.000943	<0.0001
	Intercept3	5	5.008564	0.148481	0.008564	<0.0001
	linear3	1.5	1.498366	0.044769	-0.001634	<0.0001
	Intercept4	28	28.00017	0.063312	0.000168	<0.0001
	Sigma	1	0.999254	0.014157	-0.000746	<0.0001
2000	Intercept1	5	5.000303	0.012916	0.000303	<0.0001
	Intercept2	20	19.99874	0.052464	-0.001257	<0.0001
	linear2	-2	-1.99967	0.015819	0.000334	<0.0001
	Intercept3	5	4.997734	0.074195	-0.002266	<0.0001
	linear3	1.5	1.500927	0.022371	0.000927	<0.0001
	Intercept4	28	27.9993	0.031637	-0.000702	<0.0001
	Sigma	1	0.999999	0.007074	-0.000001	<0.0001
4000	Intercept1	5	4.999974	0.009133	-0.000026	<0.0001
	Intercept2	20	19.99721	0.0371	-0.002788	<0.0001
	linear2	-2	-1.9994	0.011186	0.000604	<0.0001
	Intercept3	5	5.000508	0.052467	0.000508	<0.0001
	linear3	1.5	1.499828	0.015819	-0.000172	<0.0001
	Intercept4	28	28.00044	0.022372	0.000439	<0.0001
	Sigma	1	1.000276	0.005003	0.000276	<0.0001

* SE = Standard Error
Note: p-values are calculated based on the average mean and SE

Table A.9: Estimation of parameters for Outcome 1 on each polynomial trajectory in GBDTM and GBMTM with sample size N = 500 based on 500 simulated data sets

			GBDTM				GBMTM			
ρ**	Parameter	TPV#	Mean Estimates	Mean SE*	Bias	P-value	Mean Estimates	Mean SE*	Bias	P-value
0.1	Intercept1	5	5.0001	0.0258	0.0001	<0.0001	5.0001	0.0258	0.0001	<0.0001
	Intercept2	20	20.0025	0.1050	0.0025	<0.0001	20.0025	0.1050	0.0025	<0.0001
	Linear2	-2	-2.0009	0.0317	-0.0009	<0.0001	-2.0009	0.0317	-0.0009	<0.0001
	Intercept3	5	5.0086	0.1485	0.0086	<0.0001	5.0086	0.1485	0.0086	<0.0001
	Linear3	1.5	1.4984	0.0448	-0.0016	<0.0001	1.4984	0.0448	-0.0016	<0.0001
	Intercept4	28	28.0002	0.0633	0.0002	<0.0001	28.0002	0.0633	0.0002	<0.0001
	Sigma	1	0.9993	0.0141	-0.0007	<0.0001	0.9993	0.0142	-0.0007	<0.0001
0.2	Intercept1	5	5.0001	0.0258	0.0001	<0.0001	5.0001	0.0258	0.0001	<0.0001
	Intercept2	20	20.0025	0.1050	0.0025	<0.0001	20.0025	0.1050	0.0025	<0.0001
	Linear2	-2	-2.0009	0.0317	-0.0009	<0.0001	-2.0009	0.0317	-0.0009	<0.0001
	Intercept3	5	5.0086	0.1485	0.0086	<0.0001	5.0086	0.1485	0.0086	<0.0001
	Linear3	1.5	1.4984	0.0448	-0.0016	<0.0001	1.4984	0.0448	-0.0016	<0.0001
	Intercept4	28	28.0002	0.0633	0.0002	<0.0001	28.0002	0.0633	0.0002	<0.0001
	Sigma	1	0.9993	0.0142	-0.0007	<0.0001	0.9993	0.0142	-0.0007	<0.0001
0.4	Intercept1	5	5.0001	0.0259	0.0001	<0.0001	5.0001	0.0258	0.0001	<0.0001
	Intercept2	20	20.0025	0.1050	0.0025	<0.0001	20.0025	0.1050	0.0025	<0.0001
	Linear2	-2	-2.0009	0.0317	-0.0009	<0.0001	-2.0009	0.0317	-0.0009	<0.0001
	Intercept3	5	5.0086	0.1486	0.0086	<0.0001	5.0086	0.1485	0.0086	<0.0001
	Linear3	1.5	1.4984	0.0448	-0.0016	<0.0001	1.4984	0.0448	-0.0016	<0.0001
	Intercept4	28	28.0002	0.0633	0.0002	<0.0001	28.0002	0.0633	0.0002	<0.0001
	Sigma	1	0.9993	0.0140	-0.0007	<0.0001	0.9993	0.0142	-0.0007	<0.0001
0.6	Intercept1	5	5.0001	0.0259	0.0001	<0.0001	5.0001	0.0258	0.0001	<0.0001
	Intercept2	20	20.0025	0.1051	0.0025	<0.0001	20.0025	0.1050	0.0025	<0.0001
	Linear2	-2	-2.0009	0.0317	-0.0009	<0.0001	-2.0009	0.0317	-0.0009	<0.0001
	Intercept3	5	5.0086	0.1486	0.0086	<0.0001	5.0086	0.1485	0.0086	<0.0001
	Linear3	1.5	1.4984	0.0448	-0.0016	<0.0001	1.4984	0.0448	-0.0016	<0.0001
	Intercept4	28	28.0002	0.0634	0.0002	<0.0001	28.0002	0.0633	0.0002	<0.0001
	Sigma	1	0.9993	0.0139	-0.0007	<0.0001	0.9993	0.0142	-0.0007	<0.0001

* SE = Standard Error

** ρ = Correlation Level

# TPV = True Parameter Value

Note: p-values are calculated based on the average mean and SE

Table A.10: Estimation of parameters for Outcome 1 on each polynomial trajectory in GBDTM and GBMTM with sample size N = 2000 based on 500 simulated data sets

			GBDTM				GBMTM			
ρ**	Parameter	TPV#	Mean Estimates	Mean SE*	Bias	P-value	Mean Estimates	Mean SE*	Bias	P-value
0.1	Intercept1	5	5.0003	0.0129	0.0003	<0.0001	5.0003	0.0129	0.0003	<0.0001
	Intercept2	20	19.9987	0.0525	-0.0013	<0.0001	19.9987	0.0525	-0.0013	<0.0001
	Linear2	-2	-1.9997	0.0158	0.0003	<0.0001	-1.9997	0.0158	0.0003	<0.0001
	Intercept3	5	4.9977	0.0742	-0.0023	<0.0001	4.9977	0.0742	-0.0023	<0.0001
	Linear3	1.5	1.5009	0.0224	0.0009	<0.0001	1.5009	0.0224	0.0009	<0.0001
	Intercept4	28	27.9993	0.0316	-0.0007	<0.0001	27.9993	0.0316	-0.0007	<0.0001
	Sigma	1	1.0000	0.0071	0.0000	<0.0001	1.0000	0.0071	0.0000	<0.0001
0.2	Intercept1	5	5.0003	0.0129	0.0003	<0.0001	5.0003	0.0129	0.0003	<0.0001
	Intercept2	20	19.9987	0.0525	-0.0013	<0.0001	19.9987	0.0525	-0.0013	<0.0001
	Linear2	-2	-1.9997	0.0158	0.0003	<0.0001	-1.9997	0.0158	0.0003	<0.0001
	Intercept3	5	4.9977	0.0742	-0.0023	<0.0001	4.9977	0.0742	-0.0023	<0.0001
	Linear3	1.5	1.5009	0.0224	0.0009	<0.0001	1.5009	0.0224	0.0009	<0.0001
	Intercept4	28	27.9993	0.0316	-0.0007	<0.0001	27.9993	0.0316	-0.0007	<0.0001
	Sigma	1	1.0000	0.0071	0.0000	<0.0001	1.0000	0.0071	0.0000	<0.0001
0.4	Intercept1	5	5.0003	0.0129	0.0003	<0.0001	5.0003	0.0129	0.0003	<0.0001
	Intercept2	20	19.9987	0.0525	-0.0013	<0.0001	19.9987	0.0525	-0.0013	<0.0001
	Linear2	-2	-1.9997	0.0158	0.0003	<0.0001	-1.9997	0.0158	0.0003	<0.0001
	Intercept3	5	4.9977	0.0742	-0.0023	<0.0001	4.9977	0.0742	-0.0023	<0.0001
	Linear3	1.5	1.5009	0.0224	0.0009	<0.0001	1.5009	0.0224	0.0009	<0.0001
	Intercept4	28	27.9993	0.0316	-0.0007	<0.0001	27.9993	0.0316	-0.0007	<0.0001
	Sigma	1	1.0000	0.0070	0.0000	<0.0001	1.0000	0.0071	0.0000	<0.0001
0.6	Intercept1	5	5.0003	0.0129	0.0003	<0.0001	5.0003	0.0129	0.0003	<0.0001
	Intercept2	20	19.9987	0.0525	-0.0013	<0.0001	19.9987	0.0525	-0.0013	<0.0001
	Linear2	-2	-1.9997	0.0158	0.0003	<0.0001	-1.9997	0.0158	0.0003	<0.0001
	Intercept3	5	4.9977	0.0742	-0.0023	<0.0001	4.9977	0.0742	-0.0023	<0.0001
	Linear3	1.5	1.5009	0.0224	0.0009	<0.0001	1.5009	0.0224	0.0009	<0.0001
	Intercept4	28	27.9993	0.0316	-0.0007	<0.0001	27.9993	0.0316	-0.0007	<0.0001
	Sigma	1	1.0000	0.0070	0.0000	<0.0001	1.0000	0.0071	0.0000	<0.0001

*SE = Standard Error
** ρ = Correlation Level
$^{\#}$ TPV = True Parameter Value
Note: p-values are calculated based on the average mean and SE

Table A.11: Estimation of parameters for Outcome 1 on each polynomial trajectory in GBDTM and GBMTM with sample size N = 4000 based on 500 simulated data sets

			GBDTM				GBMTM			
ρ**	Parameter	TPV#	Mean Estimates	Mean SE*	Bias	P-value	Mean Estimates	Mean SE*	Bias	P-value
0.1	Intercept1	5	5.0000	0.0091	0.0000	<0.0001	5.0000	0.0091	0.0000	<0.0001
	Intercept2	20	19.9971	0.0371	-0.0029	<0.0001	19.9971	0.0371	-0.0029	<0.0001
	Linear2	-2	-1.9994	0.0112	0.0006	<0.0001	-1.9994	0.0112	0.0006	<0.0001
	Intercept3	5	5.0002	0.0525	0.0002	<0.0001	5.0002	0.0525	0.0002	<0.0001
	Linear3	1.5	1.4999	0.0158	-0.0001	<0.0001	1.4999	0.0158	-0.0001	<0.0001
	Intercept4	28	28.0004	0.0224	0.0004	<0.0001	28.0004	0.0224	0.0004	<0.0001
	Sigma	1	1.0003	0.0050	0.0003	<0.0001	1.0003	0.0050	0.0003	<0.0001
0.2	Intercept1	5	5.0000	0.0091	0.0000	<0.0001	5.0000	0.0091	0.0000	<0.0001
	Intercept2	20	19.9971	0.0371	-0.0029	<0.0001	19.9971	0.0371	-0.0029	<0.0001
	Linear2	-2	-1.9994	0.0112	0.0006	<0.0001	-1.9994	0.0112	0.0006	<0.0001
	Intercept3	5	5.0002	0.0525	0.0002	<0.0001	5.0002	0.0525	0.0002	<0.0001
	Linear3	1.5	1.4999	0.0158	-0.0001	<0.0001	1.4999	0.0158	-0.0001	<0.0001
	Intercept4	28	28.0004	0.0224	0.0004	<0.0001	28.0004	0.0224	0.0004	<0.0001
	Sigma	1	1.0003	0.0050	0.0003	<0.0001	1.0003	0.0050	0.0003	<0.0001
0.4	Intercept1	5	5.0000	0.0091	0.0000	<0.0001	5.0000	0.0091	0.0000	<0.0001
	Intercept2	20	19.9971	0.0371	-0.0029	<0.0001	19.9971	0.0371	-0.0029	<0.0001
	Linear2	-2	-1.9994	0.0112	0.0006	<0.0001	-1.9994	0.0112	0.0006	<0.0001
	Intercept3	5	5.0002	0.0525	0.0002	<0.0001	5.0002	0.0525	0.0002	<0.0001
	Linear3	1.5	1.4999	0.0158	-0.0001	<0.0001	1.4999	0.0158	-0.0001	<0.0001
	Intercept4	28	28.0004	0.0224	0.0004	<0.0001	28.0004	0.0224	0.0004	<0.0001
	Sigma	1	1.0003	0.0050	0.0003	<0.0001	1.0003	0.0050	0.0003	<0.0001
0.6	Intercept1	5	5.0000	0.0091	0.0000	<0.0001	5.0000	0.0091	0.0000	<0.0001
	Intercept2	20	19.9971	0.0371	-0.0029	<0.0001	19.9971	0.0371	-0.0029	<0.0001
	Linear2	-2	-1.9994	0.0112	0.0006	<0.0001	-1.9994	0.0112	0.0006	<0.0001
	Intercept3	5	5.0002	0.0525	0.0002	<0.0001	5.0002	0.0525	0.0002	<0.0001
	Linear3	1.5	1.4999	0.0158	-0.0001	<0.0001	1.4999	0.0158	-0.0001	<0.0001
	Intercept4	28	28.0004	0.0224	0.0004	<0.0001	28.0004	0.0224	0.0004	<0.0001
	Sigma	1	1.0003	0.0050	0.0003	<0.0001	1.0003	0.0050	0.0003	<0.0001

[*] SE = Standard Error
[**] ρ = Correlation Level
[#] TPV = True Parameter Value
Note: p-values are calculated based on the average mean and SE

Table A.12: Estimation of parameters for Outcome 2 on each polynomial trajectory in GBTM, GBDTM and GBMTM with sample size N = 500 based on 500 simulated data sets

		GBTM			GBDTM			GBMTM		
ρ**	Parameter	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value
0.1	Intercept1	-4.29708	2.113521	0.042	-3.20394	0.361038	<0.0001	-3.20971	0.313708	<0.0001
	Linear1	-3.92716	0.768494	<0.0001	-0.011	0.106448	0.918	0.005179	0.094404	0.956
	Intercept2	-1.22538	1.737032	0.481	-1.74845	0.45284	0.0001	-2.32647	0.408548	<0.0001
	Linear2	-0.14547	0.469996	0.757	-0.00619	0.128783	0.962	-0.11798	0.130753	0.367
	Intercept3							-3.3652	0.762096	<0.0001
	Linear3							0.121227	0.217368	0.577
	Intercept4							-1.86921	0.432027	<0.0001
	Linear4							0.027641	0.12896	0.83
0.2	Intercept1	-3.57016	0.542473	<0.0001	-2.84622	0.262878	<0.0001	-3.04972	0.289312	<0.0001
	Linear1	0.053929	0.160535	0.737	-0.01515	0.076132	0.842	0.013769	0.086678	0.874
	Intercept2	-0.85997	0.626733	0.17	-0.7736	0.298393	0.01	-1.57742	0.317184	<0.0001
	Linear2	0.022894	0.156363	0.884	-0.0081	0.089912	0.928	-0.182	0.103928	0.08
	Intercept3							-3.10435	0.645264	<0.0001
	Linear3							0.16536	0.181349	0.362
	Intercept4							-0.834	0.320036	0.009
	Linear4							0.0457	0.095742	0.633
0.4	Intercept1	-9.88555	1.865907	<0.0001	-2.9248	0.291386	<0.0001	-2.76569	0.250454	<0.0001
	Linear1	1.139762	0.434609	0.009	0.090037	0.0807	0.265	0.033371	0.074374	0.654
	Intercept2	2.676267	1.51892	0.078	-0.63955	0.289161	0.027	-0.51842	0.237108	0.028
	Linear2	-1.53156	0.356897	<0.0001	-0.17742	0.083817	0.034	-0.22658	0.076645	0.003
	Intercept3	0.598915	0.437362	0.012	0.548784	0.337531	0.104	-2.92182	0.521612	-5.60152
	Linear3	0.684539	0.178391	0.049	0.204972	0.108391	0.059	0.328139	0.140356	0.019
	Intercept4							0.550767	0.334316	0.099
	Linear4							0.19911	0.106895	0.063
0.6	Intercept1	-2.89242	0.403049	<0.0001	-2.53397	0.193022	<0.0001	-2.51474	0.189529	<0.0001
	Linear1	0.30388	0.143186	0.034	0.263474	0.054276	<0.0001	0.258721	0.052379	<0.0001
	Intercept2	-1.48834	0.966516	0.124	-1.97204	0.46949	<0.0001	1.177734	0.246615	<0.0001
	Linear2	0.579334	0.272048	0.033	0.716815	0.13665	<0.0001	-0.01785	0.073857	0.81
	Intercept3	8.105447	2.875524	0.005	1.471718	0.678589	0.03	-3.04427	0.417321	<0.0001
	Linear3	-0.82104	0.654997	0.21	2.468193	0.204369	<0.0001	0.97128	0.126296	<0.0001
	Intercept4							-3.17667	1.425478	0.026
	Linear4							8.654814	0.73231	<0.0001

*SE = Standard Error
** ρ = Correlation Level
Note: p-values are calculated based on the average mean and SE

Table A.13: Estimation of parameters for Outcome 2 on each polynomial trajectory in GBTM, GBDTM and GBMTM with sample size N = 2000 based on 500 simulated data sets

		GBTM			GBDTM			GBMTM		
ρ**	Parameter	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value
0.1	Intercept1	-6.35872	1.521438	<0.0001	-3.16539	0.161227	<0.0001	-3.1972	0.155613	<0.0001
	Linear1	0.285748	0.399375	0.474	-0.00458	0.044767	0.919	0.003334	0.04685	0.943
	Intercept2	-0.30919	1.095129	0.778	-1.75941	0.200488	<0.0001	-2.29535	0.198984	<0.0001
	Linear2	-0.36412	0.28499	0.202	-0.01303	0.057902	0.823	-0.11045	0.063358	0.081
	Intercept3							-3.23433	0.356033	<0.0001
	Linear3							0.10416	0.102289	0.309
	Intercept4							-1.85975	0.213517	<0.0001
	Linear4							0.028402	0.063689	0.656
0.2	Intercept1	-2.70449	0.191085	<0.0001	-2.82723	0.12692	<0.0001	-3.02601	0.143146	<0.0001
	Linear1	-0.02971	0.050353	0.555	-0.0147	0.037017	0.691	0.009241	0.042971	0.83
	Intercept2	-0.84071	0.295684	0.004	-0.77269	0.145003	<0.0001	-1.58353	0.157072	<0.0001
	Linear2	0.010502	0.070093	0.881	-0.0142	0.04414	0.748	-0.17163	0.051159	0.0007
	Intercept3							-3.00865	0.30814	<0.0001
	Linear3							0.154322	0.086902	0.076
	Intercept4							-0.82309	0.159039	<0.0001
	Linear4							0.041682	0.047594	0.381
0.4	Intercept1	-5.98202	1.486543	<0.0001	-2.8928	0.134273	<0.0001	-2.76305	0.124607	<0.0001
	Linear1	0.658811	0.270541	0.015	0.086527	0.037885	0.022	0.034485	0.036973	0.351
	Intercept2	0.226995	0.902845	0.802	-0.70499	0.137948	<0.0001	-0.52613	0.118152	<0.0001
	Linear2	-0.44148	0.177864	0.013	-0.15965	0.039859	<0.0001	-0.22328	0.038136	<0.0001
	Intercept3	0.527904	0.254678	0.019	0.554893	0.166777	0.0009	-2.86492	0.25401	<0.0001
	Linear3	0.338696	0.113077	0.003	0.198548	0.053431	0.0002	0.321743	0.068513	<0.0001
	Intercept4							0.552484	0.166034	0.0009
	Linear4							0.196522	0.053035	0.0002
0.6	Intercept1	-2.56617	0.164474	<0.0001	-2.50713	0.095106	<0.0001	-2.49661	0.094223	<0.0001
	Linear1	0.260756	0.05453	<0.0001	0.25838	0.026625	<0.0001	0.254886	0.026045	<0.0001
	Intercept2	-1.21532	0.49865	0.015	-2.18284	0.239563	<0.0001	1.172443	0.122332	<0.0001
	Linear2	0.466354	0.112052	<0.0001	0.758452	0.068612	<0.0001	-0.01953	0.036699	0.595
	Intercept3	5.510738	1.654844	0.045	1.778876	0.360902	<0.0001	-3.02492	0.207055	<0.0001
	Linear3	-0.66665	0.332218	0.0008	0.728153	0.164096	<0.0001	0.966528	0.062663	<0.0001
	Intercept4							0.530964	0.840247	0.527
	Linear4							2.980181	0.561091	<0.0001

*SE = Standard Error
** ρ = Correlation Level
Note: p-values are calculated based on the average mean and SE

Table A.14: Estimation of parameters for Outcome 2 on each polynomial trajectory in GBTM, GBDTM and GBMTM with sample size N = 4000 based on 500 simulated data sets

		GBTM			GBDTM			GBMTM		
ρ**	Parameter	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value
0.1	Intercept1	-4.42717	0.882989	<0.0001	-3.1638	0.107505	<0.0001	-3.21314	0.110382	<0.0001
	Linear1	-0.1564	0.298238	0.6	-0.00274	0.03078	0.93	0.007316	0.033163	0.825
	Intercept2	-1.72545	0.749014	0.021	-1.7685	0.140671	<0.0001	-2.31079	0.14145	<0.0001
	Linear2	-0.00631	0.17766	0.972	-0.00934	0.040889	0.82	-0.11004	0.045029	0.015
	Intercept3							-3.22232	0.249872	<0.0001
	Linear3							0.104184	0.071769	0.154
	Intercept4							-1.85897	0.150393	<0.0001
	Linear4							0.031868	0.044804	0.477
0.2	Intercept1	-2.68062	0.123949	<0.0001	-2.83189	0.089673	<0.0001	-3.02735	0.101157	<0.0001
	Linear1	-0.02878	0.032319	0.373	-0.01262	0.026144	0.63	0.009993	0.030353	0.742
	Intercept2	-0.83255	0.208617	<0.0001	-0.76875	0.102329	<0.0001	-1.58111	0.110993	<0.0001
	Linear2	0.011529	0.049597	0.816	-0.01668	0.031246	0.594	-0.17307	0.036171	<0.0001
	Intercept3							-3.04418	0.218684	<0.0001
	Linear3							0.166624	0.061357	0.007
	Intercept4							-0.82013	0.112363	<0.0001
	Linear4							0.039827	0.033636	0.236
0.4	Intercept1	-4.75503	0.95872	<0.0001	-2.89306	0.092444	<0.0001	-2.77155	0.088166	<0.0001
	Linear1	0.443051	0.172042	0.01	0.088169	0.026422	0.0008	0.037575	0.026122	0.15
	Intercept2	-0.52926	0.607479	0.384	-0.70253	0.094336	<0.0001	-0.51946	0.083437	<0.0001
	Linear2	-0.17434	0.09998	0.081	-0.16082	0.027974	<0.0001	-0.22461	0.026938	<0.0001
	Intercept3	0.483094	0.167174	0.004	0.553322	0.11763	<0.0001	-2.83942	0.178539	<0.0001
	Linear3	0.270096	0.073647	0.0002	0.198047	0.037699	<0.0001	0.314218	0.048261	<0.0001
	Intercept4							0.551394	0.117269	<0.0001
	Linear4							0.19683	0.037461	<0.0001
0.6	Intercept1	-2.5696	0.11151	<0.0001	-2.4982	0.066901	<0.0001	-2.49244	0.066515	<0.0001
	Linear1	0.268259	0.034174	<0.0001	0.25607	0.01867	<0.0001	0.253932	0.018393	<0.0001
	Intercept2	-1.139	0.368317	0.002	-2.34803	0.175385	<0.0001	1.172436	0.086501	<0.0001
	Linear2	0.448771	0.076797	<0.0001	0.79325	0.050219	<0.0001	-0.01935	0.025944	0.456
	Intercept3	3.987474	1.199847	0.0009	2.08077	0.249412	<0.0001	-3.00811	0.145744	<0.0001
	Linear3	-0.38157	0.23722	0.108	0.154405	0.119836	0.198	0.960574	0.04409	<0.0001
	Intercept4							2.032224	0.605113	0.0008
	Linear4							1.499547	0.40833	0.0002

* SE = Standard Error
** ρ = Correlation Level
Note: p-values are calculated based on the average mean and SE

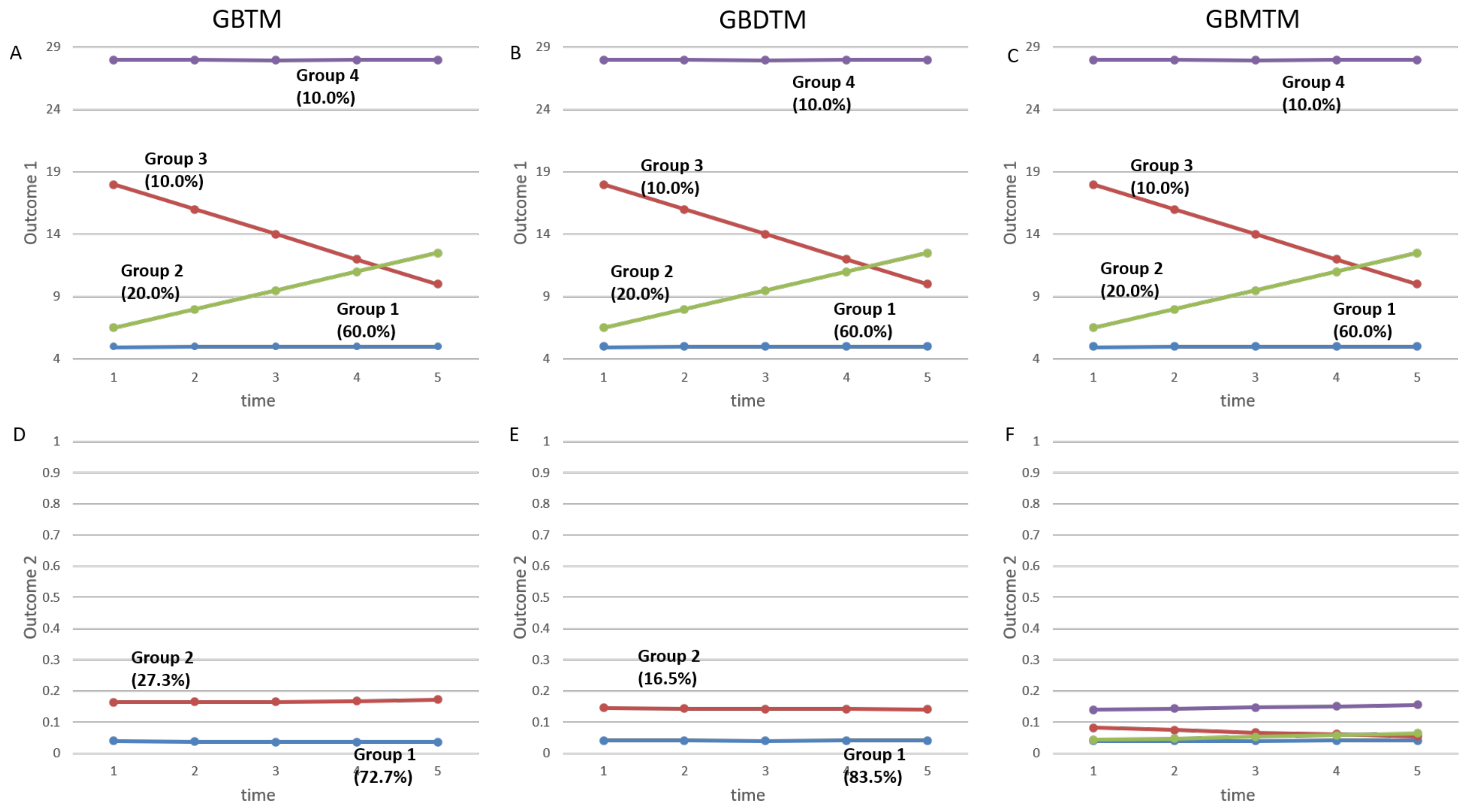


Figure A.5: Average trajectories of three trajectory models from simulation study with sample size N = 4000 and correlation coefficient 0.1 based on 500 simulations

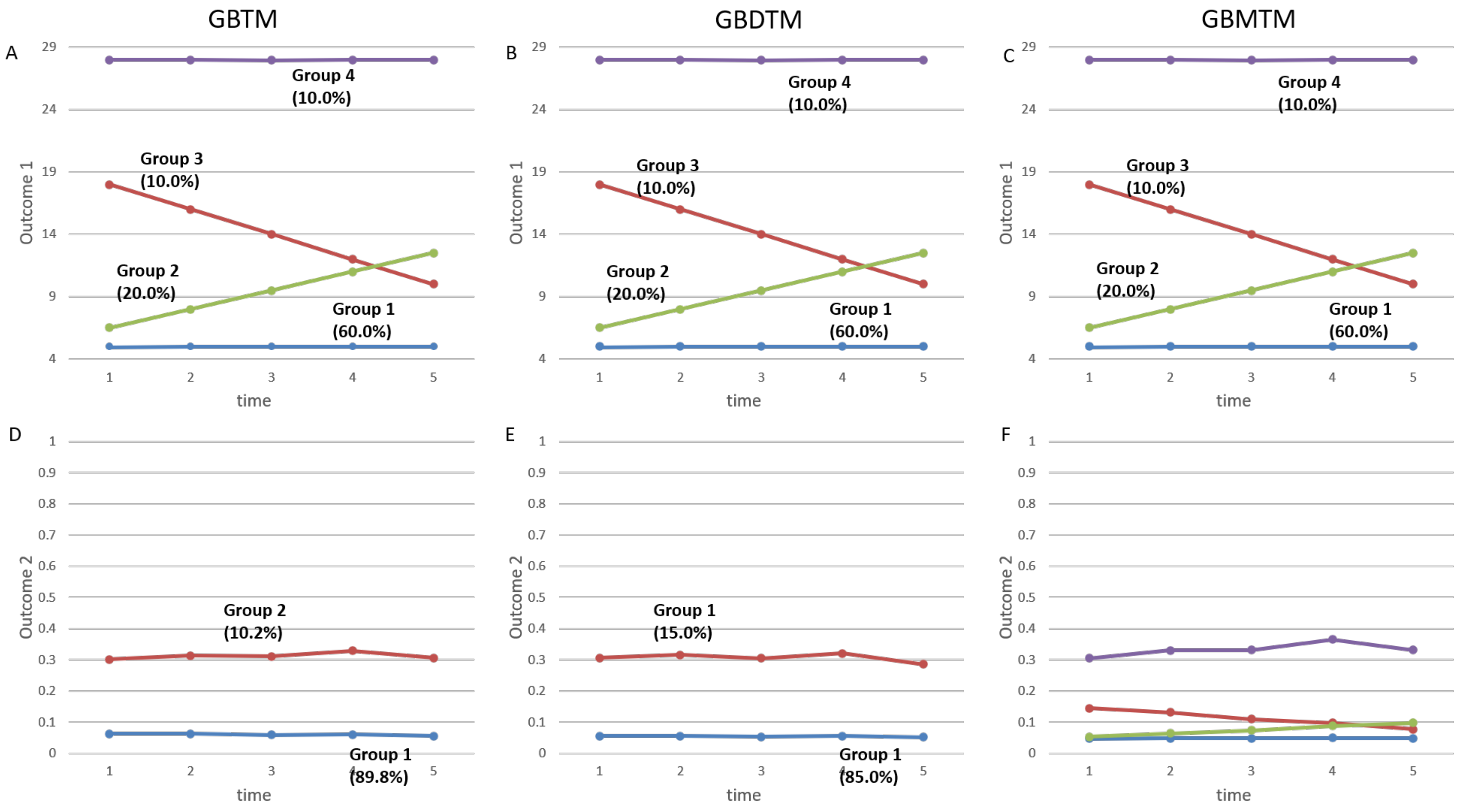


Figure A.6: Average trajectories of three trajectory models from simulation study with sample size N = 4000 and correlation coefficient 0.2 based on 500 simulations

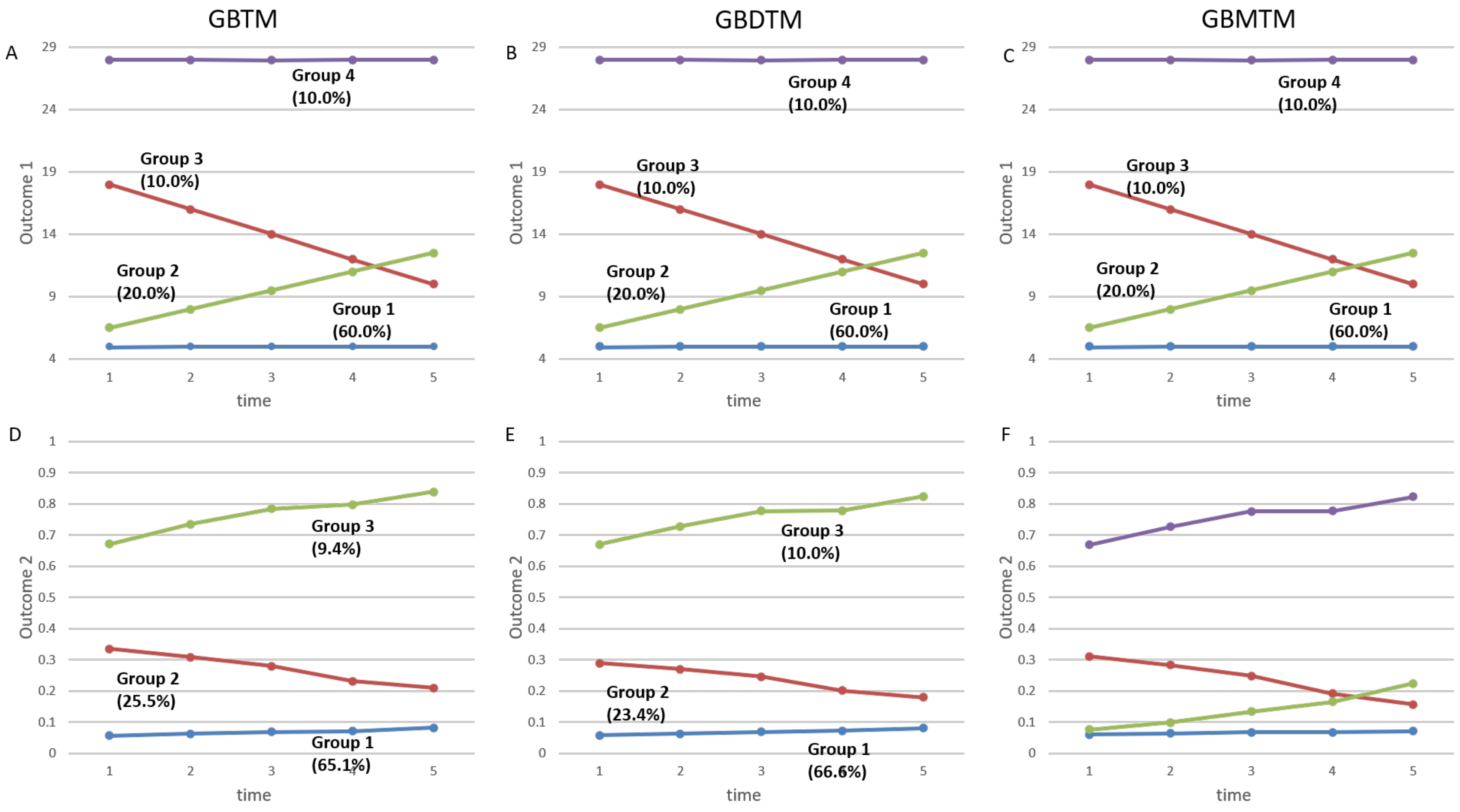


Figure A.7: Average trajectories of three trajectory models from simulation study with sample size N = 4000 and correlation coefficient 0.4 based on 500 simulations

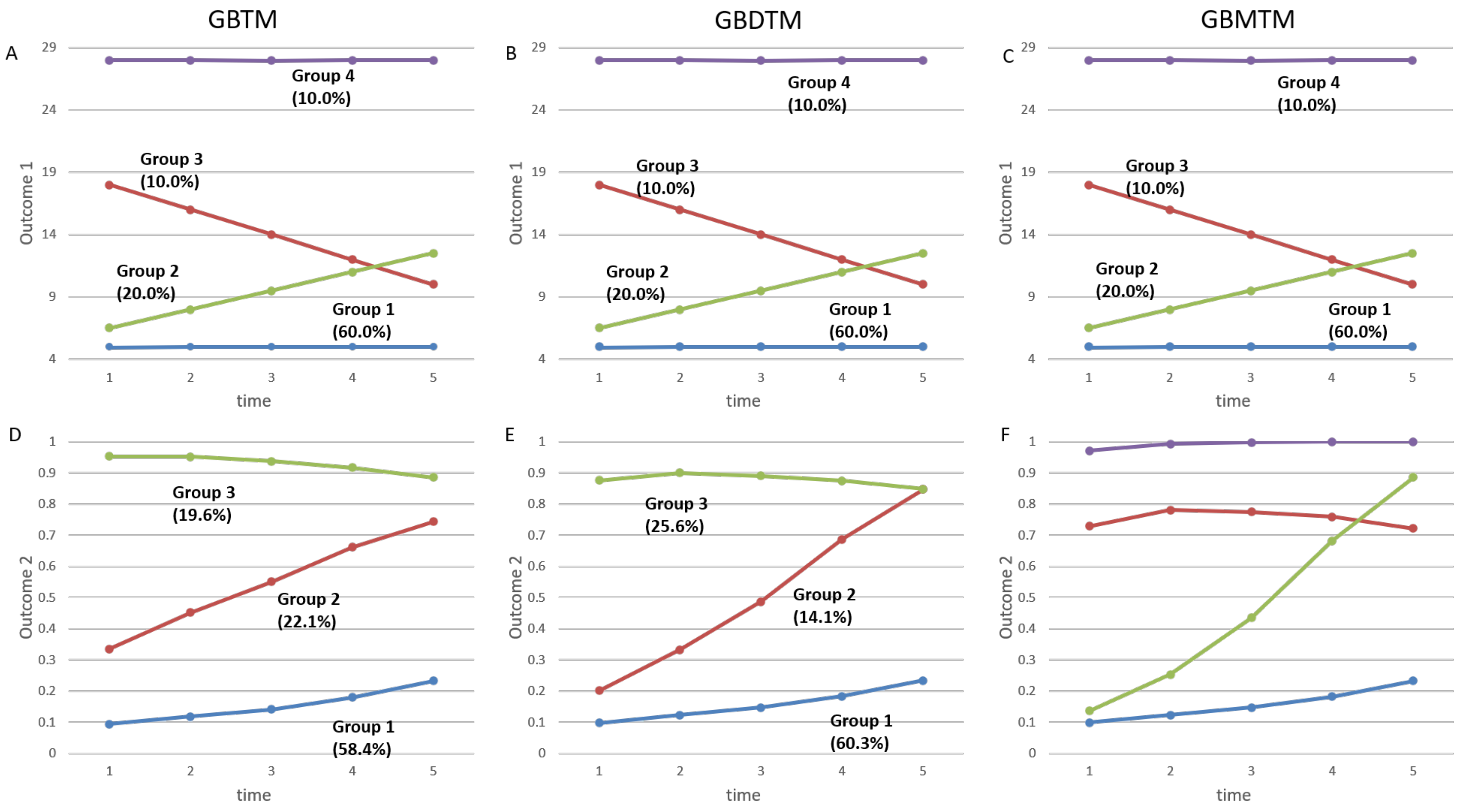


Figure A.8: Average trajectories of three trajectory models from simulation study with sample size N = 4000 and correlation coefficient 0.6 based on 500 simulations

A.3 Tables and figures of simulation results with two binary longitudinal outcomes

Table A.15: Estimation of parameters for Outcome 1 on each polynomial trajectory in GBDTM and GBMTM with sample size N = 500 based on 500 simulated data sets

			GBDTM				GBMTM			
ρ**	Parameter	TPV#	Mean Estimates	Mean SE*	Bias	P-value	Mean Estimates	Mean SE*	Bias	P-value
0.1	Intercept1	-4.5	-5.2566	0.9204	-0.7566	<0.0001	-5.1338	0.8048	-0.6338	<0.0001
	Intercept2	-4	-4.1256	1.1736	-0.1256	<0.0001	-4.1352	0.5934	-0.1352	<0.0001
	Linear2	1	1.0303	0.4626	0.0303	<0.0001	1.0332	0.1697	0.0332	<0.0001
	Intercept3	3.5	3.7571	1.3843	0.2571	<0.0001	3.7101	0.8736	0.2101	<0.0001
	Linear3	-1	-1.0821	0.4898	-0.0821	<0.0001	-1.0686	0.2567	-0.0686	<0.0001
	Intercept4	4	10.2503	2.5578	6.2503	<0.0001	7.3520	1.3738	3.3520	<0.0001
0.2	Intercept1	-4.5	-4.9872	0.7195	-0.4872	<0.0001	-4.9303	0.6951	-0.4303	<0.0001
	Intercept2	-4	-4.1742	0.6859	-0.1742	<0.0001	-4.2029	0.5915	-0.2029	<0.0001
	Linear2	1	1.0465	0.2128	0.0465	<0.0001	1.0548	0.1683	0.0548	<0.0001
	Intercept3	3.5	3.7938	0.9586	0.2938	<0.0001	3.6475	0.8442	0.1475	<0.0001
	Linear3	-1	-1.0965	0.2909	-0.0965	<0.0001	-1.0593	0.2479	-0.0593	<0.0001
	Intercept4	4	6.9596	1.3198	2.9596	<0.0001	6.3150	1.3231	2.3150	<0.0001
0.4	Intercept1	-4.5	-4.4290	0.4394	0.0710	<0.0001	-4.6201	0.4870	-0.1201	<0.0001
	Intercept2	-4	-4.3750	0.6188	-0.3750	<0.0001	-4.3631	0.5788	-0.3631	<0.0001
	Linear2	1	1.1206	0.1764	0.1206	<0.0001	1.1073	0.1627	0.1073	<0.0001
	Intercept3	3.5	3.5637	0.8100	0.0637	<0.0001	3.4758	0.7524	-0.0242	<0.0001
	Linear3	-1	-1.0310	0.2430	-0.0310	<0.0001	-1.0310	0.2208	-0.0310	<0.0001
	Intercept4	4	7.5713	1.2930	3.5713	<0.0001	4.7134	0.9501	0.7134	<0.0001
0.6	Intercept1	-4.5	-4.2881	0.3512	0.2119	<0.0001	-4.6947	0.4822	-0.1947	<0.0001
	Intercept2	-4	-4.4980	0.5696	-0.4980	<0.0001	-4.5025	0.5542	-0.5025	<0.0001
	Linear2	1	1.1581	0.1584	0.1581	<0.0001	1.1481	0.1538	0.1481	<0.0001
	Intercept3	3.5	3.5399	0.7454	0.0399	<0.0001	3.3953	0.6858	-0.1047	<0.0001
	Linear3	-1	-1.0103	0.2287	-0.0103	<0.0001	-1.0321	0.2027	-0.0321	<0.0001
	Intercept4	4	7.9758	1.4463	3.9758	<0.0001	4.1803	0.8042	0.1803	<0.0001

* SE = Standard Error

** ρ = Correlation Level

# TPV = True Parameter Value

Note: p-values are calculated based on the average mean and SE

Table A.16: Estimation of parameters for Outcome 1 on each polynomial trajectory in GBDTM and GBMTM with sample size N = 2000 based on 500 simulated data sets

			GBDTM				GBMTM			
ρ**	Parameter	TPV#	Mean Estimates	Mean SE*	Bias	P-value	Mean Estimates	Mean SE*	Bias	P-value
0.1	Intercept1	-4.5	-4.5020	0.2626	-0.0020	<0.0001	-4.4961	0.2584	0.0039	<0.0001
	Intercept2	-4	-4.0442	0.2882	-0.0442	<0.0001	-4.0549	0.2867	-0.0549	<0.0001
	Linear2	1	1.0137	0.0829	0.0137	<0.0001	1.0171	0.0824	0.0171	<0.0001
	Intercept3	3.5	3.5157	0.3963	0.0157	<0.0001	3.4992	0.3927	-0.0008	<0.0001
	Linear3	-1	-1.0040	0.1154	-0.0040	<0.0001	-1.0013	0.1142	-0.0013	<0.0001
	Intercept4	4	4.8159	0.7128	0.8159	<0.0001	4.5589	0.6476	0.5589	<0.0001
0.2	Intercept1	-4.5	-4.4452	0.2412	0.0548	<0.0001	-4.4462	0.2377	0.0538	<0.0001
	Intercept2	-4	-4.0713	0.2891	-0.0713	<0.0001	-4.1106	0.2848	-0.1106	<0.0001
	Linear2	1	1.0245	0.0829	0.0245	<0.0001	1.0358	0.0814	0.0358	<0.0001
	Intercept3	3.5	3.5263	0.3959	0.0263	<0.0001	3.4466	0.3811	-0.0534	<0.0001
	Linear3	-1	-1.0087	0.1155	-0.0087	<0.0001	-0.9934	0.1106	0.0066	<0.0001
	Intercept4	4	4.2254	0.6926	0.2254	<0.0001	4.1051	0.5687	0.1051	<0.0001
0.4	Intercept1	-4.5	-4.2579	0.1814	0.2421	<0.0001	-4.3855	0.2049	0.1145	<0.0001
	Intercept2	-4	-4.2965	0.2874	-0.2965	<0.0001	-4.2707	0.2798	-0.2707	<0.0001
	Linear2	1	1.1030	0.0808	0.1030	<0.0001	1.0867	0.0788	0.0867	<0.0001
	Intercept3	3.5	3.3116	0.3535	-0.1884	<0.0001	3.2758	0.3457	-0.2242	<0.0001
	Linear3	-1	-0.9471	0.1047	0.0529	<0.0001	-0.9701	0.1006	0.0299	<0.0001
	Intercept4	4	4.6069	0.6675	0.6069	<0.0001	3.6409	2.5987	-0.3591	<0.0001
0.6	Intercept1	-4.5	-4.2152	0.1577	0.2848	<0.0001	-4.4271	0.1911	0.0729	<0.0001
	Intercept2	-4	-4.3988	0.2749	-0.3988	<0.0001	-4.4052	0.2673	-0.4052	<0.0001
	Linear2	1	1.1359	0.0767	0.1359	<0.0001	1.1271	0.0744	0.1271	<0.0001
	Intercept3	3.5	3.3339	0.3290	-0.1661	<0.0001	3.2373	0.3188	-0.2627	<0.0001
	Linear3	-1	-0.9391	0.0999	0.0609	<0.0001	-0.9815	0.0931	0.0185	<0.0001
	Intercept4	4	4.7474	0.7163	0.7474	<0.0001	3.4082	0.2919	-0.5918	<0.0001

* SE = Standard Error

** ρ = Correlation Level

# TPV = True Parameter Value

Note: p-values are calculated based on the average mean and SE

Table A.17: Estimation of parameters for Outcome 2 on each polynomial trajectory in GBTM, GBDTM and GBMTM with sample size N = 500 based on 500 simulated data sets

		GBTM			GBDTM			GBMTM		
ρ**	Parameter	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value
0.1	Intercept1	-34.1256	28.44168	0.231	-2.53802	1.613088	0.116	-2.02628	0.206392	<0.0001
	Linear1	6.218871	11.39287	0.585	0.024373	0.793961	0.976	-0.05092	0.063944	0.426
	Intercept2	-0.34518	1.184791	0.771	-0.78154	1.206135	0.518	-2.28989	0.386017	<0.0001
	Linear2	-0.56958	0.39233	0.147	-0.41185	0.595776	0.49	0.125616	0.110335	0.255
	Intercept3							-1.11686	0.461797	0.016
	Linear3							-0.22603	0.157603	0.152
	Intercept4							-1.41532	0.406314	0.0005
	Linear4							-0.03688	0.124425	0.767
0.2	Intercept1	-13.4285	5.389076	0.013	-2.317	0.43628	<0.0001	-2.18658	0.215832	<0.0001
	Linear1	2.021792	1.463652	0.167	0.022154	0.162939	0.892	-0.02719	0.066499	0.683
	Intercept2	2.236319	1.453343	0.124	-1.11195	0.399206	0.005	-2.65166	0.397411	<0.0001
	Linear2	-0.91551	0.458455	0.046	-0.00425	0.148924	0.978	0.299852	0.108131	0.006
	Intercept3							-0.60683	0.417478	0.146
	Linear3							-0.30793	0.147243	0.037
	Intercept4							-1.07287	0.355315	0.003
	Linear4							0.013517	0.106762	0.899
0.4	Intercept1	6.034007	2.038703	0.003	-2.6046	0.376824	<0.0001	-2.19273	0.217416	<0.0001
	Linear1	-4.64712	1.130331	<0.0001	-0.21114	0.108102	0.051	-0.03394	0.068063	0.618
	Intercept2	-20.0857	3.977221	<0.0001	-2.19077	0.418463	<0.0001	-2.9489	0.398835	<0.0001
	Linear2	4.29121	1.02782	<0.0001	0.193341	0.135639	0.154	0.536748	0.106603	<0.0001
	Intercept3	3.851973	1.451661	0.008	0.056418	0.311699	0.856	0.487842	0.3735	0.192
	Linear3	-0.84414	0.428097	0.049	-0.14118	0.103	0.171	-0.4811	0.131897	0.0003
	Intercept4							-0.14836	0.3071	0.629
	Linear4							0.010082	0.093104	0.914
0.6	Intercept1	0.187526	1.349595	0.89	-1.91394	0.211279	<0.0001	-1.9867	0.20871	<0.0001
	Linear1	-1.37738	0.758076	0.069	-0.10021	0.069685	0.15	-0.0842	0.067215	0.211
	Intercept2	-7.34986	2.453803	0.003	-3.03934	0.413561	<0.0001	-3.00988	0.38631	<0.0001
	Linear2	1.730669	0.740785	0.019	0.699516	0.112775	<0.0001	0.689345	0.106206	<0.0001
	Intercept3	1.140002	0.583805	0.051	1.42061	0.261713	<0.0001	1.832053	0.415386	<0.0001
	Linear3	-0.22886	0.14891	0.125	-0.33854	0.074538	<0.0001	-0.73079	0.139948	<0.0001
	Intercept4							1.023941	0.334793	0.002
	Linear4							-0.05683	0.101085	0.574

[*] SE = Standard Error
[**] ρ = Correlation Level
Note: p-values are calculated based on the average mean and SE

Table A.18: Estimation of parameters for Outcome 2 on each polynomial trajectory in GBTM, GBDTM and GBMTM with sample size N = 2000 based on 500 simulated data sets

		GBTM			GBDTM			GBMTM		
ρ**	Parameter	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value	Mean Estimates	Mean SE*	P-value
0.1	Intercept1	-17.0851	10.07553	0.09	-2.11554	0.146425	<0.0001	-2.01117	0.101588	<0.0001
	Linear1	2.299	2.390585	0.336	-0.02365	0.042104	0.575	-0.05307	0.031471	0.092
	Intercept2	-0.60751	0.618343	0.316	-1.32246	0.207194	<0.0001	-2.27133	0.189522	<0.0001
	Linear2	-0.6485	0.197488	0.001	-0.10353	0.056386	0.066	0.122601	0.054287	0.024
	Intercept3							-1.14558	0.220057	<0.0001
	Linear3							-0.20752	0.07335	0.005
	Intercept4							-1.42302	0.200936	<0.0001
	Linear4							-0.02559	0.061074	0.675
0.2	Intercept1	-7.18007	2.203965	0.001	-2.30151	0.110442	<0.0001	-2.17754	0.106524	<0.0001
	Linear1	0.794501	0.564298	0.159	0.021243	0.033574	0.527	-0.02729	0.032785	0.405
	Intercept2	-1.13471	0.712412	0.111	-1.14955	0.149867	<0.0001	-2.59837	0.194192	<0.0001
	Linear2	-0.05126	0.178136	0.774	0.004668	0.042358	0.912	0.291026	0.053019	<0.0001
	Intercept3							-0.61161	0.199523	0.002
	Linear3							-0.2966	0.069029	<0.0001
	Intercept4							-1.07176	0.176837	<0.0001
	Linear4							0.01504	0.053082	0.777
0.4	Intercept1	0.491399	0.901457	0.586	-2.30243	0.130601	<0.0001	-2.19111	0.107742	<0.0001
	Linear1	-1.41502	0.454334	0.002	0.057611	0.040061	0.15	-0.03276	0.033606	0.33
	Intercept2	-7.16872	1.663208	<0.0001	-2.56422	0.196631	<0.0001	-2.93579	0.196085	<0.0001
	Linear2	1.41504	0.444725	0.001	0.387754	0.054878	<0.0001	0.536	0.052451	<0.0001
	Intercept3	0.319902	0.604378	0.597	0.020349	0.152605	0.894	0.475685	0.182171	0.009
	Linear3	-0.18952	0.176606	0.283	-0.13821	0.044456	0.002	-0.46765	0.063938	<0.0001
	Intercept4							-0.1963	0.152703	0.199
	Linear4							0.022105	0.04617	0.632
0.6	Intercept1	-1.09271	0.418998	0.009	-1.92381	0.101775	<0.0001	-1.98664	0.103445	<0.0001
	Linear1	-0.50444	0.249112	0.043	-0.09686	0.032863	0.003	-0.08334	0.03321	0.012
	Intercept2	-3.6926	0.934042	<0.0001	-2.97658	0.19374	<0.0001	-2.97406	0.187952	<0.0001
	Linear2	0.720922	0.279633	0.01	0.687866	0.053226	<0.0001	0.682656	0.051773	<0.0001
	Intercept3	0.922548	0.250013	0.0002	1.384205	0.126314	<0.0001	1.801705	0.201082	<0.0001
	Linear3	-0.19154	0.060846	0.002	-0.32995	0.035844	<0.0001	-0.71306	0.066737	<0.0001
	Intercept4							1.017253	0.165928	<0.0001
	Linear4							-0.05496	0.050065	0.273

[*] SE = Standard Error
[**] ρ = Correlation Level
Note: p-values are calculated based on the average mean and SE

Appendix B SIMULATION CODE EXAMPLE

B.1 Simulation of two binary outcomes with correlation level 0.6

```
let NumSamples = 500;/* Specify number of simulation */

%let NumPre = 5;    /* Specify number of preliminary simulation */

%let N = 4000;        /* Specify sample size */

%let nCont = 5; /*specify number of measures*/


/*Generate 5 preliminary Outcome 1*/

data PreSim1;

call streaminit(89036);

array y[&nCont];

array z[&nCont];

array t[&nCont] (1 2 3 4 5);

array eta1[&nCont];

array mu[&nCont];

do SampleID = 1 to &NumPre;

do i = 1 to &N;

do j = 1 to dim(y);

type=i/&N;

if type=<0.6 then do;

eta1[j] = -4.5;

mu[j] = logistic(eta1[j]);

y[j] = rand("Bernoulli", mu[j]);

sp=1;

end;

if 0.6<type=<0.8 then do;

eta1[j] = -4+1*j;
```

```
mu[j] = logistic(eta1[j]);
y[j] = rand("Bernoulli", mu[j]);
sp=2;
end;
if 0.8<type=<0.9 then do;
eta1[j] = 3.5-1*j;
mu[j] = logistic(eta1[j]);
y[j] = rand("Bernoulli", mu[j]);
sp=3;
end;
if 0.9<type then do;
eta1[j] = 4;
mu[j] = logistic(eta1[j]);
y[j] = rand("Bernoulli", mu[j]);
sp=4;
end;
end;
output;
end;
end;
run;

/*Generate 5 preliminary Outcome 2*/
data PreSim1;
set PreSim1;
array eta[&nCont];
array mu[&nCont];
array y[&nCont];
array z[&nCont];
```

```
eta1 =-2+3.2*y1;

mu1 = logistic(eta1);

z1 = rand("Bernoulli", mu1);

eta2 =-2.3+3.1*y2;

mu2 = logistic(eta2);

z2 = rand("Bernoulli", mu2);

eta3 =-2.3+3*y3;

mu3 = logistic(eta3);

z3 = rand("Bernoulli", mu3);

eta4 =-2.3+3.1*y4;

mu4 = logistic(eta4);

z4 = rand("Bernoulli", mu4);

eta5 =-2.2+3*y5;

mu5 = logistic(eta5);

z5 = rand("Bernoulli", mu5);

output;

run;


/*Output correlation between preliminary Outcome 1 and 2 in each measurement

to make sure they have the correlation coefficient around 0.6*/

proc corr data = SIMREG3 spearman;

var y1-y5 z1-z5;

by sampleID;

ods output spearmanCorr = output.CorrC1_4000_06;

run;


/*Detemine number of trajectory groups and initial value of Outcome 2 in GBTM and GBDTM*/
```

```
%macro trajpre;

%do NumPre=1 %to 5;

data PreSim1;

set PreSim1;

if SampleID~=&NumPre then delete;

run;

PROC TRAJ DATA=SIMREG4 OUTPLOT=OP1_&NumSamples OUTSTAT=OS1_&NumSamples OUT=OF1_&NumSamples OUT

     ID i; VAR z1-z5; INDEP T1-T5;

     MODEL logit; NGROUPS 3; ORDER 1 1 1;

;

RUN;

/*%TRAJPLOT(Op1_&NumSamples,OS1_&NumSamples,'Variable2 vs. Age','Cnorm Model','Variable2','Age

%end;

%mend trajpre;

%trajpre


/*Generate Outcome 1 in full simulation*/

data SimReg1;

call streaminit(89025);

array y[&nCont];

array z[&nCont];

array t[&nCont] (1 2 3 4 5);

array eta1[&nCont];

array mu[&nCont];

do SampleID = 1 to &NumSamples;

do i = 1 to &N;

do j = 1 to dim(y);

type=i/&N;

if type=<0.6 then do;
```

```
eta1[j] = -4.5;

mu[j] = logistic(eta1[j]);

y[j] = rand("Bernoulli", mu[j]);

sp=1;

end;

if 0.6<type=<0.8 then do;

eta1[j] = -4+1*j;

mu[j] = logistic(eta1[j]);

y[j] = rand("Bernoulli", mu[j]);

sp=2;

end;

if 0.8<type=<0.9 then do;

eta1[j] = 3.5-1*j;

mu[j] = logistic(eta1[j]);

y[j] = rand("Bernoulli", mu[j]);

sp=3;

end;

if 0.9<type then do;

eta1[j] = 4;

mu[j] = logistic(eta1[j]);

y[j] = rand("Bernoulli", mu[j]);

sp=4;

end;

end;

output;

end;

end;

run;
```

```
/*GBTM for Outcome 1*/

%macro traj1;

%do NumSamples=1 %to 500;

data SIMREG2;

set SIMREG1;

if SampleID~=&NumSamples then delete;

run;

PROC TRAJ DATA=SIMREG2 OUTPLOT=OP&NumSamples OUTSTAT=OS&NumSamples OUT=OF&NumSamples OUTEST=OE&

     ID i; VAR y1-y5; INDEP T1-T5;

     MODEL logit; NGROUPS 4; ORDER 0 1 1 0;

         /*starting value outcome 1*/

         start  -4.5

                        -4   1

                        3.5   -1

                        4

                        60 20 10 10;

RUN;

/*%TRAJPLOT(Op&NumSamples,OS&NumSamples,’Variable vs. Age’,’Cnorm Model’,’Variable’,’Age’)*/

%end;

%mend traj1;

%traj1


/*result output Outcome 1 in GBTM*/

DATA output.OEC1_4000_06;

SET OE1-OE500;

RUN;

DATA output.OFC1_4000_06;

SET OF1-OF500;

RUN;
```

```
DATA output.OPC1_4000_06;
SET OP1-OP500;
RUN;
DATA output.OSC1_4000_06;
SET OS1-OS500;
RUN;

/*generate Outcome 2 in full simulation*/
data simreg3;
set simreg1;
array eta[&nCont];
array mu[&nCont];
array y[&nCont];
array z[&nCont];

eta1 =-2+3.2*y1;
mu1 = logistic(eta1);
z1 = rand("Bernoulli", mu1);
eta2 =-2.3+3.1*y2;
mu2 = logistic(eta2);
z2 = rand("Bernoulli", mu2);
eta3 =-2.3+3*y3;
mu3 = logistic(eta3);
z3 = rand("Bernoulli", mu3);
eta4 =-2.3+3.1*y4;
mu4 = logistic(eta4);
z4 = rand("Bernoulli", mu4);
eta5 =-2.2+3*y5;
mu5 = logistic(eta5);
```

```
z5 = rand("Bernoulli", mu5);
output;
run;

/*Output correlation between Outcome 1 and Outcome 2 in each measurement*/
proc corr data = SIMREG3 spearman;
var y1-y5 z1-z5;
by sampleID;
ods output spearmanCorr = output.CorrC1_4000_06;
run;

/*Build GBTM for Outcome 2*/
%macro traj2;
%do NumSamples=1 %to 500;
data SIMREG4;
set SIMREG3;
if SampleID~=&NumSamples then delete;
run;
PROC TRAJ DATA=SIMREG4 OUTPLOT=OP1_&NumSamples OUTSTAT=OS1_&NumSamples OUT=OF1_&NumSamples OUT
     ID i; VAR z1-z5; INDEP T1-T5;
     MODEL logit; NGROUPS 3; ORDER 1 1 1;
/*starting value outcome 2 based on mean of 5 simulation with sample size 4000*/
         start              -1.469565    -0.245484
                                                -3.581290     0.70505
                                                0.859367    -0.177431
                   54.581498     23.386612    22.031890

;
RUN;
```

```
/*%TRAJPLOT(Op1_&NumSamples,OS1_&NumSamples,'Variable2 vs. Age','Cnorm Model','Variable2','Age
%end;
%mend traj2;
%traj2

/*result output Outcome 2 in GBTM*/
DATA output.OEC2_4000_06;
SET OE1_1-OE1_500;
RUN;
DATA output.OFC2_4000_06;
SET OF1_1-OF1_500;
RUN;
DATA output.OPC2_4000_06;
SET OP1_1-OP1_500;
RUN;
DATA output.OSC2_4000_06;
SET OS1_1-OS1_500;
RUN;

/*Build GBDTM for Outcome 1 and 2*/
%macro dual1;
%do NumSamples=1 %to 500;
data SIMREG4;
set SIMREG3;
if SampleID~=&NumSamples then delete;
run;
PROC TRAJ DATA=simreg4 OUTPLOT=OP2_&NumSamples OUTSTAT=OS2_&NumSamples OUT=OF2_&NumSamples OUTE
     ID i;
     VAR y1-y5; INDEP T1-T5; MODEL logit; NGROUPS 4; ORDER 0 1 1 0;
     VAR2 z1-z5;  INDEP2 T1-T5; MODEL2 logit;  NGROUPS2 3; ORDER2 1 1 1;
         /*same start value used with GBTM*/
         start  -4.5
                        -4   1
                        3.5   -1
                        4
                        60 20 10 10
                                                -1.469565    -0.245484
                                                -3.581290     0.70505
                                                0.859367    -0.177431
                   54.581498     23.386612    22.031890
                                   54.581498     23.386612    22.031890
                                   54.581498     23.386612    22.031890
                                   54.581498     23.386612    22.031890
                        ;
RUN;
/*%TRAJPLOT(OP2_&NumSamples,Os2_&NumSamples,'Opposition vs. Age','Cnorm Model','Opposition','S
%TRAJPLOT(OP3_&NumSamples,Os3_&NumSamples,'Opposition vs. Age','Cnorm Model','Opposition','Scal
%end;
%mend dual1;
%dual1;

/*Result output in GBDTM for Outcome 1 and Outcome 2*/
DATA output.OEC3_4000_06;
SET OE2_1-OE2_500;
RUN;
DATA output.OFC3_4000_06;
SET OF2_1-OF2_500;
RUN;
```

```
DATA output.OPC3_4000_06;
SET OP2_1-OP2_500;
RUN;
DATA output.OSC3_4000_06;
SET OS2_1-OS2_500;
RUN;
DATA output.OPC4_4000_06;
SET OP3_1-OP3_500;
RUN;
DATA output.OSC4_4000_06;
SET OS3_1-OS3_500;
RUN;

/*Build GBMTM for Outcome 1 and Outcome 2*/
%macro MULT1;
%do NumSamples=1 %to 500;
data SIMREG4;
set SIMREG3;
if SampleID~=&NumSamples then delete;
run;
PROC TRAJ DATA=simreg4 OUTPLOT=OP4_&NumSamples OUTSTAT=OS4_&NumSamples OUT=OF4_&NumSamples OUT
     ID i;
     VAR y1-y5; INDEP T1-T5; MODEL logit; ORDER 0 1 1 0;
     VAR2 z1-z5;  INDEP2 T1-T5; MODEL2 logit; ORDER2 1 1 1 1;
        MULTGROUPS 4;
         start   -4.5
                        -4   0.81
                        3.5   -1
                        4
```

```
 0 0 0 0 0 0 0 0

                        60 20 10 10
                        ;
RUN;
/*%TRAJPLOT(OP4_&NumSamples,Os4_&NumSamples,’Opposition vs. Age’,’Cnorm Model’,’Opposition’,’S
%TRAJPLOT(OP5_&NumSamples,Os5_&NumSamples,’Opposition vs. Age’,’Cnorm Model’,’Opposition’,’Scal
%end;
%mend MULT1;
%MULT1

/*Result output GBMTM for Outcome 1 and Outcome 2*/
DATA output.OEC5_4000_06;
SET OE4_1-OE4_500;
RUN;
DATA output.OFC5_4000_06;
SET OF4_1-OF4_500;
RUN;
DATA output.OPC5_4000_06;
SET OP4_1-OP4_500;
RUN;
DATA output.OSC5_4000_06;
SET OS4_1-OS4_500;
RUN;
DATA output.OPC6_4000_06;
SET OP5_1-OP5_500;
RUN;
DATA output.OSC6_4000_06;
SET OS5_1-OS5_500;
```

```
RUN;

/*Output average of parameter estimates and standard error from GBTM GBDTM and GBMTM model*/
proc means data =output.Oec1_4000_06(WHERE=(_TYPE_="PARMS")) mean;
var INTERC01 -- _AIC_;
run;
proc means data =output.Oec1_4000_06(WHERE=(_TYPE_="STDERR")) mean;
var INTERC01-- _AIC_;
run;
proc means data =output.Oec2_4000_06(WHERE=(_TYPE_="PARMS")) mean;
var INTERC01 -- _AIC_;
run;
proc means data =output.Oec2_4000_06(WHERE=(_TYPE_="STDERR")) mean;
var INTERC01 -- _AIC_;
run;
proc means data =output.Oec3_4000_06(WHERE=(_TYPE_="PARMS")) mean ;
var INTERC01 -- _AIC_;
run;
proc means data =output.Oec3_4000_06(WHERE=(_TYPE_="STDERR")) mean;
var INTERC01 -- _AIC_;
run;
proc means data =output.Oec5_4000_06(WHERE=(_TYPE_="PARMS")) mean;
var INTERC11--_AIC_;
run;
proc means data =output.Oec5_4000_06(WHERE=(_TYPE_="STDERR")) mean;
var INTERC11--_AIC_;
run;
```

```
/*Output trajectory average mean each measurement in every model*/
proc means data =output.Opc1_4000_06(WHERE=(T=1)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc1_4000_06(WHERE=(T=2)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc1_4000_06(WHERE=(T=3)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc1_4000_06(WHERE=(T=4)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc1_4000_06(WHERE=(T=5)) mean;
var AVG1--U95M4;
run;

proc means data =output.Opc2_4000_06(WHERE=(T=1)) mean;
var AVG1--U95M3;
run;
proc means data =output.Opc2_4000_06(WHERE=(T=2)) mean;
var AVG1--U95M3;
run;
proc means data =output.Opc2_4000_06(WHERE=(T=3)) mean;
var AVG1--U95M3;
run;
proc means data =output.Opc2_4000_06(WHERE=(T=4)) mean;
var AVG1--U95M3;
run;
```

```
proc means data =output.Opc2_4000_06(WHERE=(T=5)) mean;
var AVG1--U95M3;
run;

proc means data =output.Opc3_4000_06(WHERE=(T=1)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc3_4000_06(WHERE=(T=2)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc3_4000_06(WHERE=(T=3)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc3_4000_06(WHERE=(T=4)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc3_4000_06(WHERE=(T=5)) mean;
var AVG1--U95M4;
run;


proc means data =output.Opc4_4000_06(WHERE=(T=1)) mean;
var AVG1--U95M3;
run;
proc means data =output.Opc4_4000_06(WHERE=(T=2)) mean;
var AVG1--U95M3;
run;
proc means data =output.Opc4_4000_06(WHERE=(T=3)) mean;
var AVG1--U95M3;
```

```
run;
proc means data =output.Opc4_4000_06(WHERE=(T=4)) mean;
var AVG1--U95M3;
run;
proc means data =output.Opc4_4000_06(WHERE=(T=5)) mean;
var AVG1--U95M3;
run;

proc means data =output.Opc5_4000_06(WHERE=(T=1)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc5_4000_06(WHERE=(T=2)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc5_4000_06(WHERE=(T=3)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc5_4000_06(WHERE=(T=4)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc5_4000_06(WHERE=(T=5)) mean;
var AVG1--U95M4;
run;

proc means data =output.Opc6_4000_06(WHERE=(T=1)) mean;
var AVG1--U95M4;
run;
proc means data =output.Opc6_4000_06(WHERE=(T=2)) mean;
var AVG1--U95M4;
```

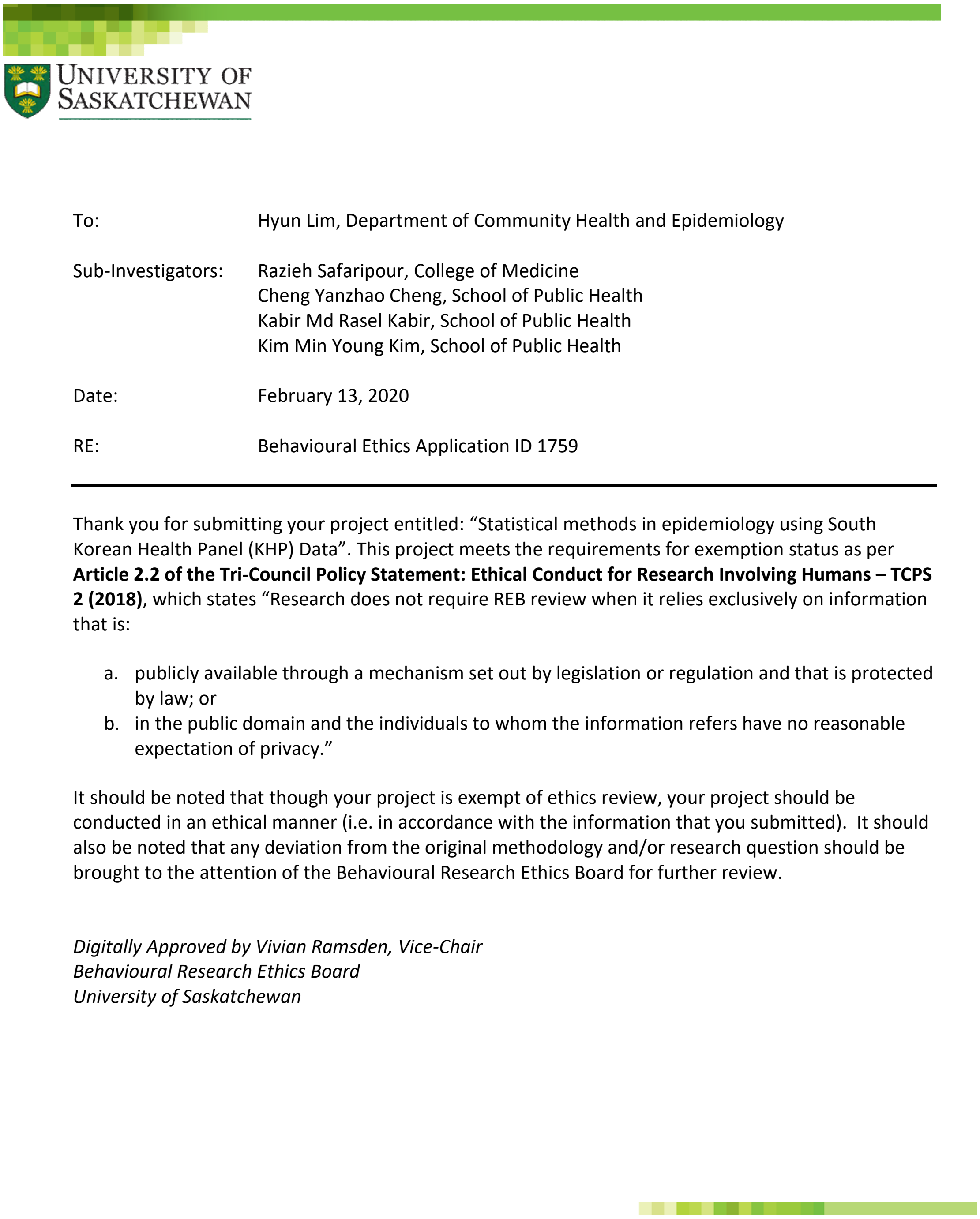
```
run;

proc means data =output.Opc6_4000_06(WHERE=(T=3)) mean;

var AVG1--U95M4;

run;

proc means data =output.Opc6_4000_06(WHERE=(T=4)) mean;

var AVG1--U95M4;

run;

proc means data =output.Opc6_4000_06(WHERE=(T=5)) mean;

var AVG1--U95M4;

run;


/*Output the average proportion of trajectory groups in each model*/

data output.Osc1_4000_06;

   set output.Osc1_4000_06;

   if list>=4 then list=0;

   list+1;

run;

proc means data =output.Osc1_4000_06 (WHERE=(list=1)) mean;

var PI;

run;

proc means data =output.Osc1_4000_06 (WHERE=(list=2)) mean;

var PI;

run;

proc means data =output.Osc1_4000_06 (WHERE=(list=3)) mean;

var PI;

run;

proc means data =output.Osc1_4000_06 (WHERE=(list=4)) mean;

var PI;

run;
```

```
data output.Osc2_4000_06;
   set output.Osc2_4000_06;
   if list>=3 then list=0;
   list+1;
run;
proc means data =output.Osc2_4000_06 (WHERE=(list=1)) mean;
var PI;
run;
proc means data =output.Osc2_4000_06 (WHERE=(list=2)) mean;
var PI;
run;
proc means data =output.Osc2_4000_06 (WHERE=(list=3)) mean;
var PI;
run;

data output.Osc3_4000_06;
   set output.Osc3_4000_06;
   if list>=4 then list=0;
   list+1;
run;
proc means data =output.Osc3_4000_06 (WHERE=(list=1)) mean;
var PI;
run;
proc means data =output.Osc3_4000_06 (WHERE=(list=2)) mean;
var PI;
run;
proc means data =output.Osc3_4000_06 (WHERE=(list=3)) mean;
var PI;
```

```
run;

proc means data =output.Osc3_4000_06 (WHERE=(list=4)) mean;

var PI;

run;



data output.Osc4_4000_06;

   set output.Osc4_4000_06;

   if list>=3 then list=0;

   list+1;

run;

proc means data =output.Osc4_4000_06 (WHERE=(list=1)) mean;

var PI;

run;

proc means data =output.Osc4_4000_06 (WHERE=(list=2)) mean;

var PI;

run;

proc means data =output.Osc4_4000_06 (WHERE=(list=3)) mean;

var PI;

run;




data output.Osc5_4000_06;

   set output.Osc5_4000_06;

   if list>=4 then list=0;

   list+1;

run;

proc means data =output.Osc5_4000_06 (WHERE=(list=1)) mean;
```

```
var PI;

run;

proc means data =output.Osc5_4000_06 (WHERE=(list=2)) mean;

var PI;

run;

proc means data =output.Osc5_4000_06 (WHERE=(list=3)) mean;

var PI;

run;

proc means data =output.Osc5_4000_06 (WHERE=(list=4)) mean;

var PI;

run;


/*Output the average correlation between each measure of Outcome 1 and Outcome 2*/

proc means data =output.corrc1_4000_06 (WHERE=(variable="y1")) mean;

var z1;

run;

proc means data =output.corrc1_4000_06 (WHERE=(variable="y2")) mean;

var z2;

run;

proc means data =output.corrc1_4000_06 (WHERE=(variable="y3")) mean;

var z3;

run;

proc means data =output.corrc1_4000_06 (WHERE=(variable="y4")) mean;

var z4;

run;

proc means data =output.corrc1_4000_06 (WHERE=(variable="y5")) mean;

var z5;

run;
```

B.2 Ethics Approval Letter for KHPS data



UNIVERSITY OF SASKATCHEWAN

To: Hyun Lim, Department of Community Health and Epidemiology

Sub-Investigators: Razieh Safaripour, College of Medicine
Cheng Yanzhao Cheng, School of Public Health
Kabir Md Rasel Kabir, School of Public Health
Kim Min Young Kim, School of Public Health

Date: February 13, 2020

RE: Behavioural Ethics Application ID 1759

Thank you for submitting your project entitled: “Statistical methods in epidemiology using South Korean Health Panel (KHP) Data”. This project meets the requirements for exemption status as per **Article 2.2 of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2018)**, which states “Research does not require REB review when it relies exclusively on information that is:

a. publicly available through a mechanism set out by legislation or regulation and that is protected by law; or
b. in the public domain and the individuals to whom the information refers have no reasonable expectation of privacy.”

It should be noted that though your project is exempt of ethics review, your project should be conducted in an ethical manner (i.e. in accordance with the information that you submitted). It should also be noted that any deviation from the original methodology and/or research question should be brought to the attention of the Behavioural Research Ethics Board for further review.

Digitally Approved by Vivian Ramsden, Vice-Chair
Behavioural Research Ethics Board
University of Saskatchewan

Figure B.1: Ethics Approval Letter for KHPS data