Summer 2019

Regression Mixture Models: An Investigation of the Effects of Mixing Weights and Predictor Distributions

Phillip Sherlock

REGRESSION MIXTURE MODELS: AN INVESTIGATION OF THE EFFECTS OF MIXING WEIGHTS AND PREDICTOR DISTRIBUTIONS

by

Phillip Sherlock

Bachelor of Arts
University of Illinois at Chicago, 2011

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Educational Psychology and Research

College of Education

University of South Carolina

2019

Accepted by:

Christine DiStefano, Major Professor

Brian Habing, Major Professor

Xiaofeng S. Liu, Committee Member

Herman Knopf, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

DEDICATION

“Eternal God, in whom mercy is endless and the treasury of compassion inexhaustible, look kindly upon us and increase Your mercy in us, that in difficult moments we might not despair nor become despondent, but with great confidence submit ourselves to Your holy will, which is Love and Mercy itself.”

–St. Maria Faustina Kowalska (Diary, 949)

Dear Mom,

Thank you for your unending support. Thank you for encouraging me to find my passion. Thank you on behalf of all the students you inspire. Thank you for your constant witness to the power of education. Thank you for being patient and engaging me when I asked you ‘why’. Thank you for taking me to class with you as you finished your undergraduate studies. Thank you for teaching me that math is the universal language. When I was younger, first attempting to understand how math might be the language of things we can describe but not directly measure, I asked you, “How would one mathematically model love?” Know that while I don’t have a mathematical model for love, your motherhood has given me a comprehensive list of variables. Thank you for courageously choosing to bring me into this world—this would not have been possible without your loving sacrifice.

Love,

Phillip

ACKNOWLEDGEMENTS

"…To come to enjoy what you have not, you must go by a way in which you enjoy not.
To come to the knowledge you have not, you must go by a way in which you know not.
To come to the possession you have not, you must go by a way in which you possess not.
To come to be what you are not, you must go by a way in which you are not."

–St. John of the Cross (Ascent of Mount Carmel, Book 1, Chapter 13.11)

Thank you to my committee members, Drs. DiStefano, Habing, Liu, and Knopf. Working with each of you throughout my studies has truly been an honor and a blessing. Thank you all for your friendship and for giving me the opportunity to learn from you. Thank you for sharing the greatest gift of all—your time. The selflessness with which each of you live is a beautiful reminder that the only way to keep the gifts we have been given is to share them with others. Thank you for helping me to go the way in which I knew not. Thank you for your patience, despite my thinking many times I knew more than I did. The kindness each of you shared propelled me through the times I enjoyed not. Your gifts and talents are those that would stir a profound sense of pride in many people, but it is the humility each of you possess that inspires me to go the way in which I am not.

ABSTRACT

This study focused on understanding how several data characteristics associated with the investigation of effect heterogeneity (i.e., mixing weights, predictor distributions, and the inclusion of covariates) affected enumeration and parameter recovery with regression mixture models. The inclusion of *C on X* paths, where the latent class, C, is regressed on the predictor, X, allows predictor means to vary across classes, at two points in the model building process—during and after enumeration—was of interest. This main aim was accomplished by comparing the correct enumeration rates and parameter coverage rates with and without freely estimated predictor means across classes for models with two classes, considerable separation between groups, and a total sample size of 500. Findings from this study, in accordance with previous work, indicated that *C on X* paths, should only be included after enumeration (e.g., Nylund-Gibson & Maysen, 2014). Inclusion of *C on X* paths functionally frees the estimation of associated predictor means across classes. If these paths are included in the enumeration phase, over-extraction is typical when predictor variance differences are present. Results from this study supported findings from previous research that demonstrated the necessity of including the *C on X* path when predictor means vary across classes (Lamont, Vermunt, & Van Horn, 2016). Therefore, once the number of classes has been determined, *C on X* paths should be included in models just as researchers would freely estimate residual variances across classes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Y	Random dependent variable.
X	Random independent (i.e., predictor) variable.
Z	Random covariate.
c	Latent class.
J	Number of random Y variables.
K	Number of latent classes.
π	Mixing weight.
i	Index for i^{th} individual.
p	Index *for* p^{th} predictor variable.
N	Total sample size.
β_0	Y-intercept.
β_1	Slope.
ε	Error.
$N(,)$	Normally distributed random variable with specified (mean, variance)
μ	Population mean.
σ^2	Population variance.
α_k	Class-specific intercept

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
aBIC	Adjusted Bayesian Information Criterion
ANCOVA	Analysis of Covariance
BIC	Bayesian Information Criterion
ELL	English Language Learner
GLM	Generalized Linear Model
I.I.D.	Independently and Identically Distributed
RMM	Regression Mixture Mod

CHAPTER 1

INTRODUCTION

Intervention and prevention programs are commonly applied across many areas of the behavioral and social sciences in order to help individuals improve on an outcome of interest (e.g., increase students' academic achievement and social skills) or avoid some deleterious event (e.g., school dropout prevention). Social scientists have acknowledged that individuals have diverse experiences as members of different communities, schools, families, and peer groups (Bronfenbrenner, 2005; Elder, 1998; Patterson, DeBaryshe, & Ramsey, 1989; Sampson & Laub, 1993). Thus, in the context of the social sciences, an intervention or prevention program could have different effects for some respondents due to characteristics of a subgroup (e.g., gender, race, etc.) or some unobservable, previously unhypothesized dimension and not the (in)effectiveness of the treatment.

As an explanation for differential results, many developmental theories suggest heterogeneity in the effects of predictors on outcomes (Bauer, 2011). For example, ecological systems theory (Bronfenbrenner, 1977, 1989) infers that environmental influences on individuals' responses to an intervention give rise to differential effects, whereby individuals experience differences in the relationship between predictors and outcomes. The degree to which individuals respond positively or negatively to an intervention is influenced by their environmental responsivity; whereby, highly responsive individuals will benefit more from an intervention in the proper environment and diminished benefits in a less supportive environment compared to less responsive individuals (Blair, 2002; Klein Velderman, Bakersman-Kranenburg, Juffer, & van IJzendoorn, 2006). The Head Start model, which is based on Bronfenbrenner's ecological systems theory, is an early educational intervention system that highlights the multifaceted nature of child development (Bronfenbrenner & Morris, 1998). Head Start is designed with the intention of providing students from low-income families the necessary tools to enter school ready to learn. However, even though the treatment (i.e., Head Start) is uniformly applied to low-income children, not all children have the same home experiences regarding maternal support and environment. Thus, there is the potential for an interaction between a student's environment and the effectiveness of the Head Start intervention, which may manifest in differences in the observed results.

Recent research provides evidence for the existence of heterogeneity in outcomes specifically related to learning and development. Concerning children attending in Head Start, some children experienced long-term positive outcomes, while Head Start had little to no effect for other children (Cooper & Lanza, 2014). Results from Cooper and Lanza (2014) suggested that English language learner (ELL) children experienced an overall-positive effect from Head Start. However, this group can be further divided into two subgroups. Most ELL students in both groups had immigrant mothers without high school diplomas. However, the subgroups, and consequently differences in the children's academic development, were associated with the presence or lack thereof of the child's biological father. ELL children with a present biological father experienced greater, positive response to Head Start in terms of their reading and math scores than their peers with absent biological fathers. These results are in-line with what many behavioral theories suggest— that is, environmental influences on individuals' responses to an intervention give rise to differential effects, whereby individuals experience differences in the relationship between predictors and outcomes (Van Horn et al., 2015).

Traditionally, researchers have studied differential treatment effects with the inclusion of covariates (i.e., gender, race) as moderators in multiple regression models. This is commonly thought of as an interaction, where an individual's response to an intervention is a product of the average response to the intervention and characteristics of the individual (Aiken & West, 1991). For instance, in a multiple regression model that includes a slope for the effect of an intervention and a slope for the effect of gender, where male equals one, an interaction effect would be the product of multiplying the individual's values for the intervention exposure and gender by the value of slope coefficient for the interaction term.

Furthermore, attention to the presence of differential effects allows researchers to adequately address complex research questions involving interactions between behavior and environmental or social influences. Considering the Head Start example, environmental or social influences that lead to differential treatment effects, may be thought of as risk factors (Coie et al., 1993; Kellam, Koretz, & Moscicki, 1999), suggesting an underlying heterogeneity within populations of interest. Thus, research methodologies that can capture this underlying heterogeneity of individuals' experiences and account for its effect on the relationship between predictors and outcomes will lead to an increased ability to differentiate intervention efficacy. The ability to account for differential effects is imperative for accurately describing the generalizability of interventions, which requires researchers to disentangle the heterogeneous subgroups from seemingly homogeneous samples of intervention participants.

Although researchers in many fields have recognized the differential effect of environment on individuals' responses to interventions, most research designs do not explicitly include a mechanism for modeling population heterogeneity. In applied research, interaction terms within the generalized linear model (GLM) framework to examine group differences are most commonly employed (e.g., Analysis of Covariance, ANCOVA). For example, one might be interested in understanding how students' social skills (i.e., *Y*) develop as a function of the parenting styles with which the students are raised (i.e., *X*) across genders (i.e., *Z*). In this scenario, the researcher would estimate a GLM with an x by z interaction term to predict y; where the researcher is interested in understanding how the effect of parenting style on social skills varies across boys and girls. However, this approach does not distinguish between a model in which the effect of *X* on *Y* varies as a function of *Z* and a model in which the effect of *Z* on *Y* varies as a function of *X* (Kraemer, Kierman, Essex, & Kupfer, 2008). Furthermore, this approach is limited by the necessity of a priori identification of moderators and, in many situations, insufficient power to test multiple interaction terms (Boyce et al., 1998).

One alternative to using a GLM with interaction terms is the regression mixture model (Desarbo, Jedidi, & Sinha, 2001; Van Horn et al., 2009; Wedel & Desarbo, 1994. This is a type of finite mixture model that uncovers latent groups (i.e., classes) with similar characteristics which may have similar responses to a treatment or outcome. In this way, the regression mixture model explicitly models heterogeneity (i.e., differential effects) by allowing model parameters [e.g. intercepts, variances, and the effects of predictors on outcomes (i.e., slopes)] to vary across latent classes. The ability to empirically uncover qualitatively different groups of individuals with similar patterns on a set of response variables that differ between groups is an advantage not shared by regression interactions, as these models assume that the sample is drawn from a homogenous population with respect to the effects of the pre-specified predictors (including group memberships) on the dependent variables. Regression mixtures include a mechanism for uncovering and measuring subgroups of individuals that experience different responses to a particular intervention, whereby the association between a predictor (i.e., *X* variable) and an outcome (i.e., *Y* variable) differ across participants due to unobserved heterogeneity (i.e., presence of discrete subgroups).

Regression Mixture Models (RMMs) are best applied within the context of a theory-driven inquiry where there are a limited number of classes of individuals sharing similar relationships between predictors and outcomes (Van Horn et al., 2015). In the regression mixture framework, model parameters, including intercepts, slopes, and random errors can vary across discrete subgroups, referred to as latent classes. Latent class (henceforth, referred to simply as *class*) separation is measured by mean differences (i.e., intercepts) between classes and the effects of the predictors (i.e., slopes). Although RMMs do not assume equal variances of parameters (e.g., intercepts and slopes) across classes, non-normally distributed errors may bias parameter estimates (Van Horn et al., 2012).

Although GLMs may lead to similar conclusions as regression mixtures under certain conditions (such as when the predictors of differential effects are observed covariates such as gender and race), regression interactions are ill-equipped to detect unhypothesized heterogeneity (Van Horn et al, 2015). However, regression mixtures, are useful for building theories involving effect heterogeneity that would otherwise not be accounted for by observed variable interactions in the GLM framework. More "traditional" approaches to modeling heterogeneity are limited by the ability to include observed variables, whereas regression mixtures include latent classes that have the potential to identify differential effects beyond what can be attributed to observed variable interactions.

Van Horn and colleagues (2015) emphasize that the first step in regression mixture modeling should be justifying the theoretical existence of differential effects. For example, classical developmental theories suggest that poor parenting behaviors are associated with poor social adjustment in children (Campbell, Shaw, & Gilliom, 2000; Chorpita & Barlow, 1998; Rubin, Burgess, Dwyer, & Hastings, 2003). Evidence suggests that this relationship varies not only between children, but it may also depend on contextual and individual characteristics in both parents and children (Belsky, 2005). Results from regression mixtures provided evidence for the existence of effect heterogeneity in the relationship between parenting style and social skills beyond gender and ethnicity (Van Horn et al., 2015). Findings from this study revealed a subset of children who had higher than average social skills, which had a weak association with parenting style and was only partially explained by ethnicity (Van Horn et al., 2015). In other words, this means that the regression mixture uncovered an unexplainable source of variance in the relationship between the predictor and outcome through the inclusion of latent classes. However, this leads to a not-so-simple reality—that group *is* a source of heterogeneity that is modeled but must still be explained.

In many instances, social scientists unknowingly assume from the outset of a study that the effect of a predictor on an outcome is the same across a group of individuals. These same researchers, if there is reason to suspect that another variable (i.e., covariate) might influence the relationship between a predictor and an outcome, will include an interaction between the predictor of interest and a covariate, whereby the effect of the predictor of interest varies identically for the entire sample of individuals. Rather, in accordance with recommendations from Van Horn and colleagues (2015), researchers should instead assume the possibility of effect heterogeneity that would allow the relationships between the predictors and the outcome to vary across individuals, using latent classes in the regression mixture framework. Through empirical testing, researchers can determine whether the most parsimonious model features homogenous or heterogeneous relationships between predictors and an outcome. It is imperative that researchers regard differential effects as a possibility in a study not only for understanding the differences in how individuals respond to an intervention, beyond observed, hypothesized variables, but also for study planning—most importantly the number of people needed (i.e., sample size).

RMMs are an insightful tool for the applied researcher. When population heterogeneity is hypothesized, regression mixtures offer a succinct, powerful framework for testing nebulous relationships between predictors and outcomes that function differently for different individuals. However, the most important, yet unanswered question related to regression mixtures is the extent to which these models can retrieve population equations when the predictor distributions vary across classes and how this might be affected by differences in mixing weights (i.e., sample proportions) across classes. This is question is critical to applied researchers, because it often does not seem reasonable to assume from the outset of a study that individuals whose experiences are so substantively different as to warrant investigation of effect heterogeneity will also experience an exposure to a predictor that can be described by the same exact distribution—as described by parametric form, mean, and variance.

Previous studies have considered many basic parameterizations of regression mixture modeling, but more work is needed to understand the practicality of using these models in situations that resemble conditions found in applied studies. For example, researchers have investigated several total sample sizes for regression mixtures with two classes and two predictors in each class. However, existing literature has not addressed the ability of regression mixtures to retrieve population parameters when the class with the larger slopes have the smallest percentage of the overall sample. And although previous work has shown that researchers should pay attention to the possibility of different class predictor means, yet, no study has investigated the ability of regression mixtures to handle unequal predictor variances across classes.

The purpose of this study is to support and add to the field's understanding of how to conduct research that assumes the possibility of differential effects. By simulating different conditions and parameterizations of regression mixtures, which are a method explicitly used for detecting differential effects, this study will add to the line of research aimed at understanding the requirements of implementing RMMs. Furthermore, this study will contribute to the applied researcher's understanding of the utility of RMMs that can be used to fine tune interventions and provide much needed targeted differentiation in education and related social services. Findings from this study can provide insight into the ability of mixture models to accurately detect subgroups and estimate differential effects among individuals when the following assumptions cannot be made: (a) equal-sized subgroups, (b) the largest subgroup had the largest slope, (c) the predictor means are equal across classes, and (d) the predictor variances are equal across classes.

CHAPTER 2

REVIEW OF LITERATURE

In this chapter, the regression mixture modeling framework and parameters to be estimated are described. In addition, recent methodological work investigating regression mixtures is detailed to provide both an introduction to regression mixtures and a review of relevant literature in order to ground and substantiate the need for the simulation study proposed in Chapter 3.

Mixture models are a flexible framework, with a seemingly endless amount of applications. The following chapters details work related to one specific type of mixture model—the univariate normal regression mixture with multiple continuous predictors. Although this investigation will not definitively dismiss situations in which the researcher after estimating several models finds that the predictor means and variances can be assumed equal, it is worth investigating whether regression mixtures can correctly enumerate and retrieve parameters when predictor distributions are indeed different as well as when mixing weights are equal and unequal.

2.1 REGRESSION MIXTURE FRAMEWORK

RMMs (Quandt, 1972) are a type of finite mixture model (Wedel & Desarbo, 1994). Other names used in the literature for finite mixture models include mixture models, latent class mixture models, latent class analysis, latent profile analysis, latent class regression models, RMMs, growth mixture models, hidden Markov models, hidden

Markov regression, hidden time series. Thus, finite mixture models refer to a broad class of models which assume that a sample of observations is drawn from a pre-specified number of *K* latent classes with pre-specified distributions but unknown mixing (i.e., sampling) proportions between the classes (Wedel & DeSarbo, 1994).

The purpose of using mixture models is to assign observations from a sample to the classes (i.e., mixing distribution) from which they were generated. In general, the sample decomposition approach of including mixture components (i.e., latent classes) from which individuals are drawn, has the benefit of detecting population heterogeneity. In cases where sample heterogeneity exists, two or more latent classes lead to a better fit and a more parsimonious model than assuming population homogeneity represented by one latent class. Aside from the ability to empirically uncover unhypothesized population heterogeneity, mixture models carry the added benefit of decreasing the error associated with the model by considering the differences between groups of individuals.

In general, finite mixture models, assume that *N* multivariate observations *Y* belong to a superpopulation, with *J* independently and identically distributed (I.I.D) random variables (i.e., *Y*) that are generated from a finite number of, *K*, groups, in proportions (i.e., mixing weights) $\pi_1, \ldots, \pi_k$. The mixing weights—prior probabilities used to assign observations to classes—are not known in advance, and generally fulfill the following:

$$\sum\nolimits_{k=1}^{K} \pi_s = 1,\ \ \pi_k > 0, k = 1, \ldots, K.$$

The mixture distributions (i.e., conditional densities) can belong to the same or different parametric families (i.e., normal, Poisson, gamma, binomial, etc.). It was assumed in this study that conditional densities belong to the same— that is, normal—parametric family (although this is not required).

RMMs can accommodate different types of predictors (e.g., dichotomous indicators, continuous predictors, etc.) and single or multivariate outcomes from one or multiple families. However, research to date has focused on models with conditional densities from the same parametric families. In general, the mixture modeling framework can be applied to any type of statistical model, as the procedure can simply be thought of as a method for decomposing any superpopulation into a mixture of distributions. For additional information, please refer to McLachlan and Peel (2000) for an in-depth review of finite mixture models.

Focusing on RMMs, these models have also been referred to in the literature as latent class regression models or cluster-wise regression models (Späth, 1979). Latent regression models were specifically introduced by Quandt (1972), as switching regression models. This special type of finite mixture model arises from a univariate (i.e., J = 1) mixture of normal distributions in which the dependent variable y is regressed on predictors differentially across latent classes. Traditional mixture models involved parsing individuals into latent classes based on means and variances for a set of outcomes; whereas, regression mixtures simultaneously cluster individuals into latent classes with separate regression equations. Specifically, if a heterogeneous population (i.e., superpopulation) is composed of two homogenous subgroups, a mixture model can be used to simultaneously detect the two clusters of individuals and estimate the two corresponding regression equations, rather than having only one inadequate regression equation. RMMs take the following form:

$$y_{\mathrm{ik}} = \beta_{0k} + \sum_{p=1}^{P} \beta_{pk} x_{ip} + \varepsilon_{ik},$$

where y_{ik} is the value for a continuous outcome variable, y for the i^{th} individual in the k^{th} class, β_{0k} is the intercept for the k^{th} class, P is the number of predictors, x_{ip} is the value for the p^{th} predictor variable, x for the i^{th} individual, and ε_{ik} is the random error for the i^{th} individual in class k with $k = 1, \ldots, K$, $i = 1, \ldots, n$, and $\varepsilon_{ik} \sim N(0, \sigma^2_k)$.

Furthermore, the probability that a sample individual is a member of a particular class can be expressed as a function of covariates (Muthén & Asparouhov, 2009; Wedel, 2002) specified by the following equation:

$$\Pr(c_i = k|z_i) = \frac{\exp(\alpha_k + \sum_{q=1}^{Q} \gamma_{qk} z_{iq})}{\sum_{s=1}^{K} \exp(\alpha_s \sum_{q=1}^{Q} \gamma_{qs} z_{iq})},$$

where c_i is the class-membership for the i^{th} individual in the k^{th} class, z_i is the observed value of the covariate z for membership in the k^{th} class for the i^{th} individual, α_k is the class-specific intercept, γ_k is the class-specific of effect of z. The predictors of class membership, z, can either be completely unique from the x predictors or they can partially or fully overlap (Muthén & Asparouhov, 2009; Wedel, 2002). In cases where there are no covariates predicting class-membership, the equation simplifies to the following:

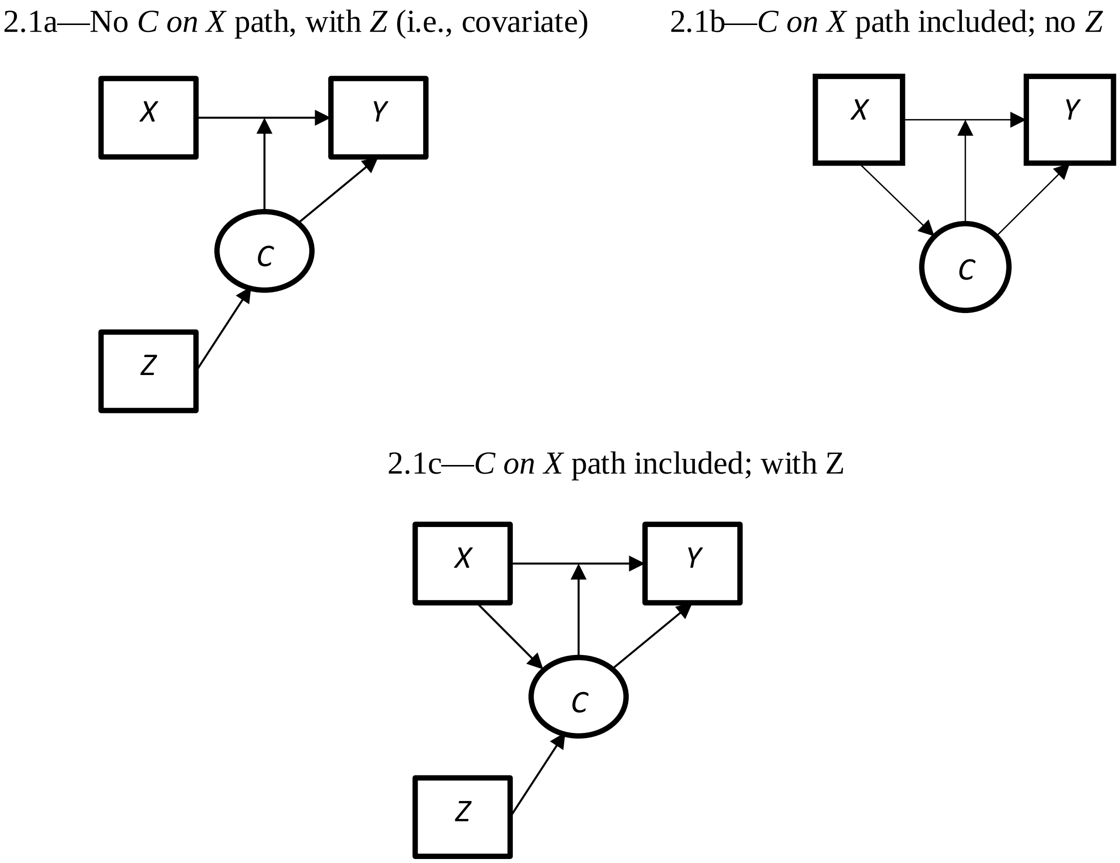
$$\Pr(c_i = k|z_i) = \frac{\exp(\alpha_k)}{\sum_{s=1}^{K} \exp(\alpha_s)}.$$

The use of latent variables is the most basic similarity between regression mixtures and related finite mixture models. Latent variables allow modeling of unobserved heterogeneity. Regression mixtures, like latent class analysis, use discrete latent variables referred to as latent classes. Practically, latent classes offer researchers the opportunity to separate a sample of individuals into subgroups based one or more variables. Specifically, regression mixtures explicitly model heterogeneity (i.e., differential effects) by allowing intercepts, variances, and the effects of predictors on outcomes (i.e., slopes) and class membership to vary across latent classes.

In cases where the *Z* and *X* variables either partially or fully overlap, meaning that *X* variables act as covariates (i.e., predictors of class membership), the assumption of equal means across classes is relaxed (Ingrassia, Minotti, & Vittadini, 2012). Thus, one can include a predictor, *X*, both as a predictor of *Y* and a predictor of class-membership. When *X* predicts *Y*, we refer to this as the *Y on X* path. When *X* predicts latent class membership, we refer to this as the *C on X* path, where *C* denotes the latent class variable.

In RMMs, one is able include the latent class on predictor (i.e., *C on X* path), which suggests that the independent variable, *X*, helps to predict class membership and functionally allows the mean of the predictors to vary across latent classes (Lamont, Vermunt, & Van Horn, 2016). Functionally, the *C on X* path is included in regression mixtures in order to freely estimate the mean of the *X* (i.e., predictor) across latent classes. Substantively, however, the *C on X* path, should not be interpreted unless it is theoretically meaningful (Lamont, Vermunt, & Van Horn, 2016).

The *C on X* path may be included for multiple predictors, distinctly, whereby the researcher assumes that the predictors each have separate, unconstrained means across classes. Figure 2.1(a) shows a model where *C* determines the relationship between *X* and *Y*, but membership in the k^{th} Class (i.e., *C*) is only determined by *an* individual's value of *Z*. Figure 2.1(b) shows a model where *C* determines the relationship between *X* and *Y*, and membership in the k^{th} Class (i.e., *C*) is determined by an individual's value of *X*. Figure 1(c) shows a model where *C* determines the relationship between *X* and *Y*, and membership in the k^{th} Class (i.e., *C*) is determined by an individual's values of both *X* and *Z*.



2.1a—No *C on X* path, with *Z* (i.e., covariate)

2.1b—*C on X* path included; no *Z*

2.1c—*C on X* path included; with Z

Figure 2.1: RMMs with and without an *X* predictor of class membership

2.2 REGRESSION MIXTURE ASSUMPTIONS

Regression mixtures, like other statistical models, follow assumptions. The most obvious assumption is that the individual observations emanate from smaller homogeneous subgroups (i.e., latent classes), which are unknown prior to estimation of the model. It follows that the assumptions which generally apply to the GLM also apply to the individual latent classes. In the GLM, it is assumed that the all individuals have the same relationship between the predictor and the outcome; whereas, in the regression mixture all individuals within a latent class have the same relationship between the predictor and the outcome. Unlike GLM interactions, however, which require homogeneous residual variance across the entire population, regression mixtures allow heterogeneity of residual variances to differ across classes, but they require homogeneity of residual variance within class. As I mentioned earlier, RMMs considered here include mixtures of normal distributions, requiring the assumption of normality of errors. Previous work has demonstrated that regression mixtures are sensitive to violating the assumption of homogeneity of within-class residual variance (Van Horn et al., 2012).

Although the number of classes, *K*, must be specified before estimating regression mixtures, researchers are able to freely estimate or constrain the class-specific intercepts, predictor slopes, predictor means, predictor variances, and residual variances. For example, suppose that a researcher specifies two classes before estimating a regression mixture with freely estimated intercepts, slopes, and residual variances while constraining the predictor means and variances. The results will feature two classes with separate linear equations, each differing based on their respective intercepts, slopes, and residual variances. However, the class equations will estimate the model assuming equal predictor means and variances. Taken together, class specific equations represent the mechanism by which regression mixtures measure differential effects or unobserved heterogeneity. When using regression mixtures, one assumes that a sample *N* is drawn from a population with *K* classes (i.e., sub-groups or sub-populations).

The process of choosing *K*—the number of classes—in a regression mixture model, is referred to as class enumeration. Class enumeration in RMMs is achieved through comparing some penalized information criterion across solutions with different values of *K*. Although the adjusted information criterion (AIC; Akaike, 1973) is commonly used for model selection, simulation studies have shown that AIC is upwardly biased with respect to class enumeration in mixture models (Nylund et al., 2007; Van Horn et al., 2009). Instead, researchers investigating class enumeration for regression mixture models have typically used the Bayesian Information Criterion (BIC; Schwarz, 1978) and sample-size adjusted BIC (aBIC; Sclove, 1987). The BIC and aBIC, do not rely on specific sampling distributions, but rather the observed data, considering the likelihood function, the sample size, and the number of parameters; wherein, the model with the smallest BIC and aBIC is typically chosen. The BIC is formally defined by the following formula:

$$\text{BIC} = \ln(n)\, q - 2\ln(\hat{L}),$$

where n is the sample size, q is the number of parameters estimated in the model, and $\hat{L}$ is the maximized value of the likelihood function. Both BIC and aBIC have been shown to be effective for class enumeration with mixture models (Van Horn et al., 2009).

2.3 EMPIRICAL EXAMPLE

To illustrate the information examined with a regression mixture model analysis, an applied example is provided. Consider a situation where researchers are interested in understanding how wages (dependent variable, *Y*) vary as a function of education and years of experience, and to see how these coefficients differ across gender. In a typical regression analysis and in a regression mixture where one assumes that the means of the predictors are equal across subgroups , education and years of experience would be included in the equation as predictor (i.e., *X*) variables and gender would be included as a covariate (i.e., *Z*) variable, whereby only gender would predict class membership (*C*). However, it is not always the case that predictor distributions are equal across subgroups. Furthermore, in this applied example, a researcher may have reason to believe that different subgroups vary not only with respect to the relationships between the predictors—education and years of experience—but also with respect to the distributions of those predictors across subgroups. Therefore, one can estimate a regression mixture wherein any combination of, or none of the predictors and covariates are used to predict class membership. However, due to the inability to determine a priori whether a variable should be included as a predictor of class membership, two models were estimated—the first where only the *X* variables are included in the model and the second in which the *X* variables and the *Z* variable (i.e., gender) are included as the *C on X* paths. Then after the final number of classes is determined, the estimates from the model with both *X* variables and the *Z* variable included as predictors of class membership will be interpreted. Using the publicly available, 1985 Current Population Survey—Determinants of Wages data set (N = 534), RMMs with the following equation:

$$Hourly\ Wage_{ik} = \beta_{0k} + \beta_{1k}{*}Education + \beta_{2k}{*}Experience + \beta_{3k}{*}Male + e_{ik},$$

In the following empirical example, the gender value for male = 1. For the entire sample, *hourly wage* ranges from 1 to 44.5 with a mean of 9.02. *Education* ranges from 2 to 18 with a mean of 13. *Years of experience* ranges from 0 to 55 with a mean of 17.8.

In order to determine the optimal number of classes, RMMs with one, two, and three classes (i.e., *k* = 1, 2, and 3) were estimated. First, enumeration, (i.e., the process through which the optimal classes solution is chosen), was determined based on evaluations of BIC across all *k* solutions. After selection of the optimal solution, the parameters from the final model were examined, and individual regression models for the number of classes were reported. As noted, one set of models included only predictors (*X*s and *Z*) measuring the dependent variable, wage (*Y*); the second set of models also included a model where gender (*Z*) is predicting the class(es) (*C*) and class determines the relationships between the predictors and the dependent variable (i.e., *C on X* paths). The final model contained class specific regression equations wherein education and years of experience predict hourly wage with, of course, a residual term

Table 2.1 shows the BIC values from the models estimated with and without the *C on X* paths for the empirical example. Based on BIC, a penalized-likelihood approach, the two-class model, which indicates two underlying latent subgroups, was chosen as optimal.

Table 2.1: BIC values for empirical example

	BIC	
Number of Classes	Without Covariates	With Covariates
1	3167.22	3138.14
2	3048.25	2992.26
3	Did not converge	Did not converge

At first glance, a researcher might assume that the two groups can be explained simply by the dichotomous gender covariate. However, a two-by-two cross tab with gender and most likely class membership based on posterior probability from the two-class model can be used to test how well gender functions as an observed covariate. If gender is a sufficient covariate, one would expect to see most women in one class and most men in the other. Table 2.2, the two-by-two cross tab for this example, does not show a clear separation based on gender.

Table 2.2: Two-by-two crosstab for empirical example

	Women	Men
Class 1	223	157
Class 2	22	132

The first class, which contains 71.2% of the overall sample—91% of the women and 54.3% of the men—is characterized by a lower intercept, with a smaller slope associated with education and a larger slope associated with experience, compared to Class 2. The second class, which contains 28.8% of the overall sample—9% of women and 45.7% of men—is characterized by a lower intercept, with a higher slope associated with education and a smaller slope associated with experience compared with Class 1. Table 2.3 includes the estimated parameters from the two-class model.

Table 2.3: Two-Class Model Estimates with C on X Paths Included

	Mean	SE
Class 1		
Intercept	.58	1.11
Slope 1	0.42	0.09
Slope 2	0.06	0.02
Residual	5.80	0.08
Education Mean	-0.48	0.11
Experience Mean	-0.09	0.02
Gender Mean	-2.40	0.48
Class 2		
Intercept	5.52	5.00
Slope 1	0.61	0.26
Slope 2	-0.02	0.08
Residual	32.65	8.43

Regarding the distributions of the predictors, sample participants assigned to Class 2 had both higher average education and experience. While the means of the predictor distributions are somewhat similar, the variance of experience for Class 1 is 156.15 while the variance for experience in Class 2 is 132.75. Individuals in Class 2, who have higher mean education and experience, with a smaller variance for years of experience show a stronger relationship between increased wages and education and years of experience.

Findings from this example could point to a latent psychological component that might be associated with wage differences. This hypothesis may also be supported by results from Risse, Farrell, and Fry (2018), which suggest that personality traits and psychological constructs are better predictors of wages than what can simply be captured by gender alone. Specifically related to work, men tend to be higher in hope for success, lower in fear of failure, and lower in agreeableness, which are associated with higher wages (Risse, Farrell, & Fry, 2018). And although men more often tend to have personality traits associated with higher wages, women can also have the disposition that is associated with higher wages, which explains, in part, the crossover between men and women between the two classes. Note, however, that other variables (e.g., occupation type, geographic area, and race/ethnicity) are not included; it is reiterated that the purpose of the analysis was to provide an illustration of how to interpret regression mixture modeling analyses and not to test a specific theory.

This example illustrates how an applied researcher might conduct an initial inquiry of effect heterogeneity. Prior to conducting a study that explicitly measured psychological constructs while recording observed wage, education, and experience data, a researcher who conducted the example analysis would have the empirical evidence to support an investigation of latent constructs associated with wage differences. Based on the results from the example above, a methodological researcher might be inclined to study the effects of differences in predictor distributions on the ability of regression mixtures to enumerate and accurately estimate class-specific regression parameters.

2.4 APPLIED STUDIES USING REGRESSION MIXTURES

Regression mixtures, although they are relatively new to the social sciences, have been used to study several types of differential effects. In their original application, Quandt (1972) used regression mixtures to study heterogeneity in housing construction. These models have also been applied to wage prediction (Quandt & Ramsey, 1978), and trade show performance (DeSarbo & Cron, 1988). Marketing researchers have also employed regression mixtures in order to better understand consumer behavior through population segmentation (Cleaver & Wedel, 2001; Desarbo, Jedidi, & Sinha, 2001.) More recently, Van Horn and colleagues (2015) used regression mixtures in order to investigate the existence of effect heterogeneity in the relationship between parenting style and social skills and found differential effects beyond gender and ethnicity. Related to education, regression mixtures have also been used to understand the heterogeneity in the effect of family resources on academic achievement (Van Horn et al., 2009; Lamont, Vermunt, & Van Horn, 2016; Jaki et al., 2019). Van Horn and colleagues (2015) detail regression mixtures and demonstrate their utility in testing specific hypotheses about differential effects and exploring heterogeneous effects of predictors. Although there are numerous differences between the applications of regression interactions and regression mixtures, regression mixtures do not require an observed predictor of differential effects (Van Horn et al., 2015).

2.5 SIMULATION STUDIES OF REGRESSION MIXTURES

When theory suggests that groups of individuals within a population are thought to have different relationships between a predictor and an outcome, effect heterogeneity can be investigated with GLM tests of interaction terms This traditional approach to studying effect heterogeneity relies on the inclusion of moderating variables (i.e., regression interactions). However, GLM interactions can fall short when effect heterogeneity exists beyond what is captured by a known variable *Z* (i.e., moderator) or the influential moderating variable has not been included in the model. In both cases, the larger heterogeneous population may contain multiple homogenous subgroups. If this is the case, the relationships between the predictors and outcome cannot be accurately modeled with one regression equation, because of an unmeasured latent dimension. Like the investigation of differential effects through regression interactions with a known covariate such as gender, which would involve two groups, regression mixtures first require specification of the number of homogeneous subgroups that emanate from the larger heterogeneous superpopulation. However, with regression mixtures, researchers do not presuppose that an observed variable, such as a gender covariate perfectly captures the heterogeneity that exists within a population. Rather, this process is initiated with theoretical justification and substantiated with empirical evidence. This process of determining the number of subgroups (i.e., classes) is known as enumeration. Furthermore, enumeration is a function of sample size, mixing weights, class separation, and predictors of class membership. Sample size refers to the size of the total sample, from which the classes are to be identified. Mixing weights refer to the proportions of the total sample that are contained within each of the classes. Class separation refers to the identifiability of the latent classes—how distinct they are from each other based on each of the parameters—slopes, covariates (i.e., predictors of class membership), and residuals.

Class membership is modeled in regression mixtures using covariates and *C on X* paths (i.e., the outcome, *Y*, is regressed on the predictor, *X*, and class, C, is also regressed on the predictor). Enumeration, which is the process of determining the *k* groups that represent the differential relationships between a set of predictors and an outcome across multiple latent classes in a regression mixture, is associated with the characteristics that define the subgroups within a superpopulation. Each of these characteristics contribute to the likelihood that a regression mixture can identify heterogeneity that exists within a sample and accurately describe their parameters.

In order to determine the reliability of regression mixtures to accurately uncover heterogeneity within populations, methodological researchers have begun to use Monte Carlo studies to systematically investigate how different parameterizations affect enumeration and parameter recovery. Findings from these studies, specifically regarding their effects on enumeration and parameter recovery, are detailed below.

Results from one such study investigating the effects of sample size on enumeration and parameter estimation in regression mixtures with one and two predictors found a direct relationship between class separation and the sample size needed to detect differential effects (Jaki et al, 2019). With little class separation, sample sizes of 3,000 were required for correct enumeration using BIC alone. This large sample size was needed regardless of whether the class sample sizes were balanced or unbalanced. However, Jaki and colleagues (2019) found that the smallest class in the three-class enumerated model often contained less than 10% of the individuals. Using 10% as the arbitrary criterion, they rejected the three-class model and accepted the two-class model when the smallest class contained less than 10% of individuals. Building on this concept, they recommended researchers be wary of the possibility of small classes representing a spurious finding and suggested that researchers consider both BIC and the proportion of the smallest class when enumerating (Jaki et al., 2019).

Jaki and colleagues (2019) were also interested in class-assignment of individuals. With balances samples greater than 1,000, models tended to over-assign individuals to the class with the larger effect size. However, with, smaller, unbalanced samples (i.e., N = 200 and N = 500) biased assignment was somewhat different compared to larger samples. When 75% of individuals truly belong to the class with the larger effect size, results yielded biased assignment with samples of 500 and 250, such that individuals tended to be over-assigned to the class with the smaller effect size (Jaki et al., 2019).

Concerning parameter estimation, parameter recovery for Class 1, containing the lower true slope value, had little bias. However, bias in all Class 2 (i.e., larger regression weight) parameters increased as the class separation decreased; wherein, the intercepts were upwardly biased (i.e., parameter estimates were higher than true values) while the regression weights and residual variances of were downwardly biased (i.e., lower than true values). It should also be noted that estimated standard errors of parameter estimates were too small, as evidenced by less than the target 95% coverage (i.e., estimate ± 1.96 SE) for the slope parameters, even when sample size was larger than 1,000. Continuing with the single predictor cases and large sample sizes (N ≥ 1,000), the distribution of the slope parameter was bimodal, with peaks around .2 and .7, (matching the population values—the simulated regression weights), with some outliers. However, cases with 500 and 200 individuals returned unimodal slope distributions, which indicates indistinguishability between the two classes.

Concerning the effect of class separation on enumeration, results indicated that increasing the intercept difference from 0 in both classes to 0 in one class and to 1 and 1.5 in the second class increased the percentage of correctly classified simulations to 70% and 95%, respectively. Returning to intercepts at zero for both classes and adding a second uncorrelated predictor with slope equal to the first predictor in each class (i.e., $y_{i|c=1} = 0 + .2x_{1\mathrm{i}} + .2x_{2\mathrm{i}} + \varepsilon_{i1}$; $y_{i|c=2} = 0 + .7x_{1i} + .7x_{2i} + \varepsilon_{i2}$) resulted in dramatic improvement in class enumeration. In conditions with equal (i.e., 50/50) and unequal (i.e., 75/25) sample proportions where class separation was due only to the differences in slopes between the two uncorrelated *X* predictors across classes (i.e., zero intercepts, equal residual variance, and equal predictor means and variances), the BIC correctly retrieved the two-class solution 97% of the time, even with samples as small as 500. Parameter coverage rates were only slightly less than 0.95 in the two predictor conditions with balanced and unbalanced samples of 500. (Jaki et al., 2019). Furthermore, when errors are normally distributed, there was an interaction between class separation and sample size on parameter recovery and class enumeration (Jaki et al., 2019).

Related to the inclusion of *X* variables as predictors of class membership, results from a study by Lamont, Vermunt, and Van Horn (2016) indicated that violating the equal predictor (i.e., *X*) means assumption and not including the *C on X* path was associated with an increased probability of selecting additional latent classes and biased class proportions. In the same study, however, results suggested that incorrectly constraining the class predictor means (by not including the *C on X* path) rarely led to a substantively different interpretation of the solution (Lamont et al, 2016). Results from the applied portion of the study by Lamont and colleagues (2016) showed that parameter estimates tended to be generally similar with and without the *C on X* path, but standard errors were higher when the path was included. The work by Lamont and colleagues (2016) shed some light on predictor mean differences across latent classes, but their study did not change both mean and variance values across classes.

In a simulation study investigating the effect of constraining the variances of normally distributed class-specific residuals to be equal, results indicated that class enumeration with constrained residual variances was affected only as the difference in residual variances across classes increased (Kim et al., 2016). Moreover, when freely estimating residual variances with two uncorrelated predictors, selecting the correct number of classes was related to class separation. Specifically, greater class separation as indicated by larger differences in residual variances (i.e., equal to 1) led to correct enumeration more often across each of the three intercept difference conditions (i.e., 0, 0.5, and 1) (Kim et al., 2016). However, in the study conducted by Kim and colleagues (2016), the moderate residual variance difference (i.e., 0.5) condition resulted in correct enumeration less often across all intercept differences compared to the conditions with no variance differences across classes. In terms of parameter estimation, results indicated that sample proportions and regression weights illustrated bias when unequal residual variances were constrained. Based on findings from their simulation study, Kim and colleagues (2016) recommended that researchers freely estimate residual variances across classes, and only impose the residual variance constraint if the models with and without the constraint have similar fit and substantive interpretation. If models with and without the residual variance equality constraint have similar fit, but different substantive interpretations, researchers should proceed with caution.

2.6 PURPOSE OF STUDY

While regression mixture modeling may be a novel approach to dealing with effect heterogeneity, there has been limited study of these designs. Simulation studies involving RMMs in the literature have included one *X* variable (Van Horn et al., 2015; Lamont, Vermunt, & Van Horn, 2016; Jaki et al., 2019) or multiple *X* variables (i.e., 2, Kim et al., 2016; Jaki et al., 2019) in the regression equations. However, in studies involving multiple *X* variables, balanced samples (Kim et al, 2016; Jaki et al., 2019) and conditions of unbalanced samples where the greatest number of participants were generated from the equation with the larger slope (Jaki et al., 2019) are often used.

Only one study has examined the effect of allowing the mean of the predictor to vary across latent classes, and each class equation only included one predictor (Lamont, Vermunt, & Van Horn, 2016). Although it is widely accepted that researchers should freely estimate the means of the *X* variables across classes in the model building process (Lamont, Vermunt, & Van Horn, 2016), no studies have examined the efficacy of RMMs when the variances of the *X* variables differ across classes. Jaki and colleagues (2019) only included unbalanced designs where the larger regression weights coincided with the larger sample proportion.

To address gaps in the current literature, this study will investigate parameterization issues in regression mixtures that will contribute to the field's understanding of models that can be used to fine tune interventions and provide much needed improvements in approaches to identifying effect heterogeneity. Findings from this fully crossed simulation study will provide insight into the ability of mixture models to accurately detect subgroups and estimate differential effects among individuals when the following assumptions cannot be made: (a) equal-sized subgroups, (b) the largest subgroup had the largest slope, (c) the predictor means are equal across classes, and (d) the predictor variances are equal across classes. Results from this study will offer much needed insights into class enumeration and parameter recovery in regression mixtures when the means and variances of predictors are both equal and unequal, the mixing weights for the smaller and larger effect sizes are equal and unequal (i.e., 50/50; 25/75; 75/25), and the *C on X* paths for two predictors are omitted and included.

This study is important as it can contribute to the field's understanding of how well regression mixtures recover classes and detect effect heterogeneity while accurately recovering the associated parameters. Testing the ability of RMMs to correctly enumerate and recover parameters associated with classes that vary in the predictor means and variances, when it is unknown whether the group with the greater response will represent the majority of the sample, is essential to understanding the limitations of these models. If the regression mixtures, based on the conditions in this study, show promise in retrieving and accurately describing subgroups that have different predictor distributions in addition to the differential effects, this will be a step forward in giving applied researchers the confidence to apply these RMMs.

CHAPTER 3

METHODS

This study investigated the accuracy of regression mixtures relative to the enumeration of classes (i.e., selecting the correct number of classes) and estimation of parameters across a variety of situations with the aim of understanding the effects of sample sizes, predictor means, and predictor variances. A simulation study was used for the investigations. As the purpose of this study was to ascertain the utility of regression mixtures in applied settings, these models were appraised, under the outlined conditions, based on how well they first enumerated the classes (i.e., correctly choose the number of classes from which the data originated) and then how well they retrieved the parameters associated with those population subgroups (i.e., slopes, residuals, etc.). This chapter details the population model of the regression mixture investigated, a description of the features manipulated in this simulation study, and a description of how the findings were evaluated.

3.1 POPULATION MODEL

The population (i.e., true) model that served as the basis for this simulation study is a two-class, two-predictor model. The parameters in each of the models include the Y-intercept (i.e., β_0), the slope for X_1 (i.e., β_1), the slope for X_2 (i.e., β_2), and the error term (i.e., ε). The general model with two classes and two predictors and equal residual variances can be written as

Class 1: $y_{i|c=1} = \beta_{01} + \beta_{11}X_{1i} + \beta_{21}X_{2i} + \varepsilon_{i1}$, $\varepsilon_{i1} \sim N(0, \sigma^2)$

Class 2: $y_{i|c=2} = \beta_{02} + \beta_{12}X_{1i} + \beta_{22}X_{2i} + \varepsilon_{i2}$, $\varepsilon_{i2} \sim N(0, \sigma^2)$.

As a simulation study contains an infinite number of conditions that can be manipulated, the following limitations were placed on the population model: the variance of *Y* equaled 1 for every class in every condition and the difference between the intercepts was always 1, with the second latent class always having the larger intercept. The differences in the residual variances and the intercepts were such that each class in each condition had Var(Y) = 1 to pinpoint the effects of manipulated features of interest—mixing weights, predictor distributions, and the omission and inclusion of *C on X* paths.

In this study, there was no correlation between predictors within class (i.e., Pearson correlation coefficient r = 0), indicating the complete absence of multicollinearity. The outcome *Y*, with zero correlation between the predictors and (large) differences in the intercepts and residual variance across classes, was generated according to the following equations:

Class 1: $y_{i|c=1} = 0 + .2x_{1i} + .2x_{2i} + \varepsilon_{i1}$, $\varepsilon_{1} \sim N(\mu, \sigma^2)$;

Class 2: $y_{i|c=2} = 1 + .7x_{1i} + .7x_{2i} + \varepsilon_{i2}$, $\varepsilon_{2} \sim N(0, .02)$.

Furthermore, each condition has Var(*Y*) = 1.

Using Var(*Y*) = Var(*X*) + Var(*E*) with zero correlation between the predictors, we had two situations observed within the simulation. In conditions where the variances of all the *X*'s equal 1, the residual variances for Class 1 = 0.92 and Class 2 = 0.02. This is the result of the equation above, where Var(*Y*) = 1 = $.2^2 + .2^2 + .92$. In conditions where Class 1|X_1 had a variance of 2, the residual variance for class 1 = 0.84 and Class 2 = 0.02. This is the result of the equation Var(Y) = 1 = $2(.2^2) + 2(.2^2) + 0.84$.

3.2 SIMULATION CONDITIONS

Each of the simulation conditions contained a constant total sample size equal to 500, an intercept difference equal to 1 across classes, and Var(*Y*) equal to 1 for each of the classes. The effects of the following characteristics on enumeration and parameter recovery were investigated: mixing weights, class separation, and predictors of class membership. Mixing weights refer to the proportions of the total sample that are contained within the classes. Class separation refers to the identifiability of the latent classes—how distinct they are from each other based on each of the parameters—slopes, covariates (i.e., predictors of class membership), and residuals. Class membership is modeled in regression mixtures using covariates and *C on X* paths (i.e., the class, C, is regressed on the outcome, *Y*, while *Y* is also regressed on *X*).

3.3 DATA GENERATION AND SIMULATION

Data was generated using the software package R (see Appendix A) and all RMMs will be fit using Mplus. The Mplus software package (version 7.4; Muthén & Muthén, 1998-2015) was used for all analyses; wherein, model parameters were estimated using maximum likelihood with robust standard errors, as this is the default estimator for mixture models in Mplus. For each condition, 500 replications were computed and analyzed. Non-converging replications were recorded and removed from subsequent analyses.

3.4 SAMPLE PROPORTIONS

The first manipulated feature in the simulation was the proportion of the sample generated from each of the classes. This study featured three scenarios for sample size differences with the following proportions (i.e., Class1 percentage/Class 2 percentage) —

(1) 50/50; (2) 75/25; (3) 25/75. These unbalanced conditions were a direct extension of work by Jaki and colleagues (2019), who only used the 50/50 scenario and the 25/75 split, where most of the individuals were assigned to the class with the larger regression weights. Building on the work by Jaki et al (2019), this study included conditions in which the smaller percentage of individuals are generated from the class with the larger regression weights.

3.5 PREDICTOR MEAN DIFFERENCE

The second manipulated feature in the study was the mean difference of the predictors. This study featured two scenarios for differences in the means of the X variables. The first predictor means difference scenario featured no mean difference across the latent classes for all X variables, whereby the means of both predictors in each of the classes was distributed as standard normal variates. The second predictor mean difference scenario featured a mean difference of 1 for X_1 in Class 1 relative to its counterpart in Class 2. In the predictor mean discrepant condition, the X_1 variable in Class 1 was distributed as N(1, variance depends on condition); whereas, the X_1 variable in Class 2 was distributed as N(0, variance depends on condition).

3.6 PREDICTOR VARIANCE DIFFERENCE

The third manipulated feature in the proposed simulation was the difference in the variances of the predictors across classes. The first predictor variance scenario featured equal predictor variances across classes, wherein each predictor followed a standard normal distribution. The second predictor variance scenario featured a greater X_1 variance in Class 1 relative to the X_1 variable in Class 2. In the predictor variance discrepant condition, the X_1 variable in Class 1 was distributed as N(mean depends on condition, 2); whereas, the X_1 variable in Class 2 was distributed as N(mean depends on condition, 1).

3.7 TESTED MODELS

RMMs with three sample proportion conditions, two predictor mean conditions, and two predictor variance conditions were analyzed under two possible conditions—(1) without *C on X* paths, where the *X* variables are not included as *Z* variables (i.e., covariates predicting class membership); and (2) with *C on X* paths, where the *X* variables also predict class membership (i.e., act as covariates). Correctly specified models, when predictor mean parameter estimation is congruent with the simulated data—(a) differing predictor mean parameters across classes are not constrained to be equal; and (b) equal predictor mean parameters across classes are constrained to be equal were specified in order to determine the utility of the *C on X* paths across several conditions. These comparisons will be essential for determining the effect of equality constraints on predictor means. With a total number of 24 conditions, the simulation study is fully crossed with respect to mixing weights (3 conditions), predictor mean differences (2 conditions), and predictor variance differences (2 conditions). Table 3.1 includes a summary of the conditions manipulated in the simulation. The simulation code for Condition 1 is given in Appendix A.

Table 3.1: Simulation conditions

Condition	Description	Mixing Weights
No *C on X* Paths		
1	All *X*s ~ N(0, 1); $\varepsilon_{class|1}$ ~ N(0, .92); $\varepsilon_{class|2}$ ~ N(0, .02)	50/50
2		75/25
3		25/75

4	Class 1|X_1 ~ N(0, 2); $\varepsilon_{class\|1}$ ~ N(0, .84); $\varepsilon_{class\|2}$ ~ N(0, .02)	50/50
5		75/25
6		25/75
7	Class 1|X_1 ~ N(1, 1); $\varepsilon_{class\|1}$ ~ N(0, .92); $\varepsilon_{class\|2}$ ~ N(0, .02)	50/50
8		75/25
9		25/75
10	Class 1|X_1 ~ (1, 2);	50/50
11	$\varepsilon_{class\|1}$ ~ N(0, .84); $\varepsilon_{class\|2}$ ~ N(0, .02)	75/25
12		25/75
C on X Paths Included		
13	All *X*s ~ N(0, 1); $\varepsilon_{class\|1}$ ~ N(0, .92); $\varepsilon_{class\|2}$ ~ N(0, .02)	50/50
14		75/25
15		25/75
16	Class 1| X_1 ~ N(0, 2); $\varepsilon_{class\|1}$ ~ N(0, .84); $\varepsilon_{class\|2}$ ~ N(0, .02)	50/50
17		75/25
18		25/75
19	Class 1| X_1 ~ N(1, 1); $\varepsilon_{class\|1}$ ~ N(0, .92); $\varepsilon_{class\|2}$ ~ N(0, .02)	50/50
20		75/25
21		25/75
22	Class 1| X_1 ~ N(1, 2); $\varepsilon_{class\|1}$ ~ N(0, .84); $\varepsilon_{class\|2}$ ~ N(0, .02)	50/50
23		75/25
24		25/75

3.8 OUTCOMES

Regression mixtures for each of the conditions were estimated with k = 1, 2, and 3. Then, for each of the conditions, the percentage of replications in which the BIC indicates the true two-class solution to have better relative fit over the one- and three-class solutions were reported. For each of the replications in which the BIC correctly identifies the two-class solution as having the best relative fit, the mean parameter values, median standard errors, and 95% coverage rates will be reported for each of the parameters in each of the conditions. In order to determine whether the inclusion of the *C on X* paths lead to be better parameter coverage (i.e., estimate ± 1.96 SE), proportion tests for the 95% coverages will be reported for each of the parameterizations with and without the *C on X* paths.

Each condition will attempt to recover data generated from two latent classes, where each class included two predictors. In each condition, Class 1 had an intercept of 0 and contain two predictors having slopes equal to 0.2; whereas, Class 2 will always have an intercept of 1 and two predictors with both variables having slopes equal to 0.7. Previous work by Jaki and colleagues (2019) found that when class separation was due only to the differences in slopes between the two uncorrelated *X* predictors across classes (i.e., zero intercepts, equal residual variance, and equal predictor means and variances), regression mixtures were correctly enumerated in 97% of replications using BIC alone, while obtaining 0.95 parameter coverage with the same sample size—in both balanced and unbalanced conditions. This proposed study will extend the work by Jaki and colleagues (2019) by varying predictor means, predictor variances, and sample proportions (i.e., smaller proportion associated with large and small effects classes).

In order to determine the impacts of the various model specifications for the crossed simulation conditions of the simulation conditions, this study examined two outcomes: a) class enumeration based on BIC, and b) recovery of parameter estimates (i.e., intercepts, slopes, residual variance, and predictor means).

3.9 CLASS ENUMERATION

In order to examine class enumeration, BIC values for the one-class solution vs. the two-class solution and the two-class solution vs. the three-class solution were analyzed for every simulation condition. Therefore, the outcome of interest for the class enumeration portion of the study was the percentage of replications that chose the two-class solution (i.e., correctly enumerated) based on BIC values. The accuracy of the RMMs in recovering parameters estimates for correctly enumerated models (i.e., BIC chose the two-class model over the one- and three-class models) were examined for 95% coverage when the models were estimated with and without the *C on X* paths.

3.10 PARAMETER RECOVERY

Comparing recovered parameters between the crossed conditions provided the opportunity to pinpoint how various parameterizations affect the efficacy of regression mixtures in capturing population heterogeneity. Estimated parameters for models in which the BIC selects the two-class solution were compared to the true parameter values. In order to analyze difference in 95% coverage rates between conditions with and without the *C on X*, twelve two-sample proportion tests were conducted for each parameter. The twelve two-sample tests were the result of comparing the two types of models (i.e.., with and without *C on X* paths) across all combinations of the following factors: (1) mixing weight for the smaller effect size class (i.e., three levels—.25, .50, and .75); (2) mean discrepancy in X_1 across classes (i.e., two levels—0 and 1); (3) variance discrepancy in X_1 across classes (i.e., two levels—1 and 2).

3.11 SUMMARY

The focus of this study was to determine the ability of regression mixtures to correctly enumerate classes and recover parameters representing distinct subgroups from heterogeneous populations using models with balanced and unbalanced sample proportions each having multiple predictors with equal and unequal means and variances.

This study included several conditions often encountered in practice by including the following characteristics: (a) multiple predictors, (b) unbalanced designs (in terms of sample size and parameter size), (c) differing predictor means across latent classes, and (d) differing predictor variances across latent classes.

In summary, the simulation study consisted of a fully crossed design with 24 cells comprised of 3 sample proportions (i.e., 50/50, 75/25, 25/75) x 2 predictor mean conditions x 2 predictor variance conditions x 2 tested models (i.e., with and without *C on X* paths). Outcomes included enumeration (i.e., percentage of replications where the BIC correctly chooses two classes), parameter recovery (i.e., mean estimate, median estimate, median standard error, 95% coverage rate for each parameter, and proportion tests to distinguish statistical significance in 95% coverage across models with and without *C on X* paths.

CHAPTER 4

RESULTS

Results from the simulation, including convergence rates, enumeration, and parameter recovery, are detailed in this chapter. Overall, the results point to the need for researchers to enumerate without *C on X* paths, as evidenced by the overall correct enumeration rates for models without *C on X* paths across all conditions. This result was especially apparent when the true variances of predictors vary across classes.

4.1 CONVERGENCE RATES

Overall, the convergence rates for all conditions were equal to or exceeded 99%. Convergence rates for each of the conditions without the *C on X* paths were greater than .99 for all values of *K*. Convergence rates for models with the *C on X* paths were 1 for all conditions when *k* =1 and *k*=2. The minimum convergence rate for k=3 when the *C on X* paths were included was .99. Table 4.1 shows the specific convergence rates by condition.

4.2 CLASS ENUMERATION

The first outcome of interest in this simulation study was class enumeration. Class enumeration, which is related to determining the number of underlying subpopulations within the sample, was determined using a penalized likelihood criterion—namely, BIC. Table 4.2 contains (1) the percentages of replications wherein the BIC chose the two-class solution over the one- and three-class solutions; and (2) percentages of replications wherein the BIC chose the three-class solution over the two-class solution.

Table 4.1: Model converge rates by condition

Condition	One-Class	Two-Class	Three-Class
1	1	1	0.998
2	1	1	0.994
3	1	1	1
4	1	1	0.996
5	1	1	0.998
6	1	1	1
7	1	1	0.998
8	1	1	0.996
9	1	1	0.998
10	1	1	0.998
11	1	1	1
12	1	1	0.998
13	1	1	0.998
14	1	1	0.998
15	1	1	1
16	1	1	0.998
17	1	1	0.99
18	1	1	1
19	1	1	0.998
20	1	1	0.994
21	1	1	1
22	1	1	0.998
23	1	1	0.996
24	1	1	0.998

Overall, conditions 1-12, which did not include the *C on X* paths correctly enumerated with BIC more often than the models including the *C on X* paths. However, when there was not a predictor variance discrepancy—either with or without a predictor mean difference—the models with the *C on X* paths resulted in correct enumeration slightly more often. The lowest correct enumeration rate was 95% across conditions 1-12 (i.e., No

C on X paths), while the lowest correct enumeration for the conditions 13-24 (i.e., including the *C on X* paths) was 60.6%.

This difference in enumeration is more apparent in the predictor variance discrepant conditions. Specifically, conditions 4-6 and 10-12 yielded correct enumeration equal to or above 95% of the replications as compared to conditions 16-18 and 22-24 (which included the *C on X* paths) where the two-class solution was optimal between 60.6% and 68.4% of the time. Even though there were discrepant variances across the X_1 predictor in conditions 4-6 and discrepant X_1 means and variances in conditions 10-12, not including the *C on X* paths led to correct enumeration more often than in conditions 16-18 and 22-24 when the predictors were included as covariates. The advantage of not including the *X* variables as covariates predicting class membership was highlighted when the mixing weights were 50/50 and 25/75. Therefore, the difficulty of correctly enumerating regression mixtures when including covariates is compounded when the mixing weight associated with the smaller effect size class is either equal to or smaller than the mixing weight associated with larger effect size class.

Table 4.2: Enumeration using BIC for all conditions

Cond.	*X* values	Mix	BIC chose 2 over 1 and 3	BIC chose 3 over 2
No C on X Paths				
1	All *X*s ~ *N*(0, 1)	50/50	0.972	0.028
2		75/25	0.962	0.038
3		25/75	0.988	0.012
4	Class 1| X_1 ~ *N*(0, 2)	50/50	0.976	0.024
5		75/25	0.972	0.028
6		25/75	0.964	0.036
7	Class 1| X_1 ~ *N*(1, 1)	50/50	0.97	0.03
8		75/25	0.974	0.026

9		25/75	0.97	0.03
10	Class 1| $X_1 \sim N(1, 2)$	50/50	0.974	0.026
11		75/25	0.962	0.038
12		25/75	0.95	0.05
C on X Paths Included				
13	All *X*s ~ (0, 1)	50/50	0.996	0.004
14		75/25	0.998	0.002
15		25/75	0.994	0.006
16	Class 1| $X_1 \sim N(0, 2)$	50/50	0.606	0.394
17		75/25	0.684	0.316
18		25/75	0.624	0.376
19	Class 1| $X_1 \sim N(1, 1)$	50/50	0.99	0.01
20		75/25	0.998	0.002
21		25/75	0.998	0.002
22	Class 1| $X_1 \sim N(1, 2)$	50/50	0.636	0.364
23		75/25	0.682	0.318
24		25/75	0.664	0.336

The second outcome of interest in this simulation study was parameter recovery when the BIC chose the two-class solution. Table 4.3 contains the recovered parameters for conditions 1-12 (i.e., conditions which did not include the C on X paths) for replications where the BIC chose the two-class solution.

Intercept coverage values were high (i.e., above .90) when there was no predictor mean difference. However, the coverage rates for the Class 1 intercept suffered when there was a predictor mean difference (95% coverage ranging from .83 to .87). In general, coverage rates were higher for recovering the Class 2 intercept values (95% coverage ranging from .94 to .96), where there was never a predictor mean difference that would necessitate inclusion of the *C on X* paths. The lowest coverage was observed with Class 1 conditions including a mean difference in the X_1 parameter. In these situations, the true population value was captured within the 95% coverage interval only 76% of the time when no X_1 predictor variance difference was present and 83% of the time with an X_1 predictor variance difference. Residual variance coverages were higher across both classes with a correctly specified predictor mean equality (95% coverage ranging from .88 to .95), compared to conditions with an incorrectly specified predictor mean equality (95% coverage ranging from .71 to .90). All the median parameter estimates were close to the mean parameter estimates, indicating an approximate normal distribution for the estimated parameters.

Table 4.3: Parameter estimates for conditions 1-12 (Predictors *not included* as covariates)

		50/50			75/25			25/75		
		Mean		95	Mean		95	Mean		95
	True	Med	SE	Cov	Med	SE	Cov	Med	SE	Cov
All *X* Predictor Distributions Equal *X* ~ *N*(0,1); No *C on X* Paths										
Class 1										
Int	0.00	0.00			0.00			-0.01		
		0.00	0.07	0.94	0.00	0.05	0.94	-0.01	0.09	0.95
Slope 1	0.20	0.20			0.20			0.20		
		0.20	0.06	0.95	0.20	0.05	0.92	0.21	0.09	0.93
Slope 2	0.20	0.20			0.20			0.20		
		0.20	0.06	0.94	0.20	0.05	0.95	0.20	0.09	0.93
Resid	0.92	0.91			0.91			0.88		
		0.91	0.08	0.93	0.91	0.07	0.95	0.88	0.11	0.89
Class 2										
Int	1.00	1.00			1.00			1.00		
		1.00	0.01	0.96	1.00	0.02	0.95	1.00	0.01	0.94
Slope 1	0.70	0.70			0.70			0.7		
		0.70	0.01	0.95	0.70	0.02	0.95	0.7	0.01	0.94
Slope 2	0.70	0.70			0.70			0.70		
		0.70	0.01	0.93	0.70	0.02	0.92	0.70	0.01	0.95
Resid	0.02	0.02			0.02			0.02		
		0.02	0	0.93	0.02	0.00	0.89	0.02	0.00	0.96

Class 1| X_1 Variance Increase to X ~ N(0, 2); All Equal X Means; No C *on* X Paths										
Class 1										
Int	0.00	0.02			0.01			0.03		
		0.01	0.06	0.91	0.01	0.05	0.93	0.03	0.09	0.92
Slope 1	0.20	0.19			0.19			0.19		
		0.19	0.04	0.93	0.19	0.03	0.93	0.19	0.06	0.94
Slope 2	0.20	0.20			0.20			0.20		
		0.20	0.06	0.94	0.20	0.05	0.95	0.20	0.09	0.92
Resid	0.84	0.84			0.84			0.82		
		0.84	0.08	0.94	0.83	0.06	0.94	0.81	0.10	0.92
Class 2										
Int	1.00	1.00			1.00			1.00		
		1.00	0.01	0.94	1.00	0.02	0.94	1.00	0.01	0.95
Slope 1	0.70	0.70			0.70			0.70		
		0.70	0.01	0.95	0.70	0.02	0.93	0.70	0.01	0.94
Slope 2	0.70	0.70			0.70			0.70		
		0.70	0.01	0.94	0.70	0.02	0.95	0.70	0.01	0.95
Resid	0.02	0.02			0.02			0.02		
		0.02	0.00	0.92	0.02	0.00	0.89	0.02	0.00	0.95
Class 1 | X_1 Mean Increase to X ~ N(1, 1); All Equal X Variances; No C *on* X Paths										
Class 1										
Int	0.00	0.07			0.06			0.09		
		0.07	0.08	0.84	0.06	0.07	0.83	0.09	0.10	0.84
Slope 1	0.20	0.14			0.15			0.12		
		0.14	0.05	0.76	0.15	0.04	0.75	0.12	0.07	0.78
Slope 2	0.20	0.24			0.23			0.25		
		0.24	0.06	0.91	0.22	0.05	0.88	0.24	0.09	0.92
Resid	0.92	0.85			0.87			0.83		
		0.86	0.08	0.79	0.87	0.07	0.80	0.82	0.11	0.81
Class 2										
Int	1.00	1.00			1.00			1.00		
		1.00	0.01	0.94	1.00	0.02	0.95	1.00	0.01	0.93
Slope 1	0.70	0.70			0.70			0.70		
		0.70	0.01	0.94	0.70	0.02	0.94	0.70	0.01	0.93
Slope 2	0.70	0.70			0.70			0.70		
		0.70	0.01	0.96	0.70	0.02	0.95	0.70	0.01	0.93
Resid	0.02	0.02			0.02			0.02		
		0.02	0.00	0.85	0.02	0.00	0.79	0.02	0.00	0.84
Class 1 | X_1 Mean and Variance Increase to X ~ N(1, 2); No C *on* X Paths										
Class 1										
Int	0.00	0.06			0.05			0.07		
		0.06	0.07	0.87	0.05	0.06	0.86	0.07	0.10	0.87
Slope 1	0.20	0.16			0.17			0.16		
		0.16	0.04	0.83	0.17	0.03	0.85	0.16	0.05	0.85
Slope 2	0.20	0.23	0.06	0.92	0.22	0.05	0.90	0.24	0.08	0.91

		0.23			0.22			0.24		
Resid	0.84	0.79			0.81			0.77		
		0.79	0.07	0.86	0.81	0.06	0.90	0.77	0.10	0.81
Class 2										
Int	1.00	1.00			1.00			1.00		
		1.00	0.01	0.96	1.00	0.02	0.96	1.00	0.01	0.94
Slope 1	0.70	0.70			0.69			0.70		
		0.70	0.01	0.95	0.69	0.02	0.90	0.70	0.01	0.95
Slope 2	0.70	0.70			0.70			0.70		
		0.70	0.01	0.95	0.70	0.02	0.92	0.70	0.01	0.93
Resid	0.02	0.02			0.02			0.02		
		0.02	0	0.79	0.02	0.00	0.71	0.02	0.00	0.85

Table 4.4 contains the recovered parameters for conditions 13-24 (i.e., included *C on X* paths) for replications where the BIC chose the two-class solution. Overall, the models with the *C on X* paths yielded higher 95% coverage rates for the parameters in common with the models without the *C on X* paths. Again, all median parameter estimates were close to the mean parameter estimates, providing no blatant evidence against normality.

Table 4.4: Parameter estimates for conditions 13-24 (Predictors *included* as covariates)

		50/50			75/25			25/75		
		Mean		95	Mean		95	Mean		95
	True	Med	SE	Cov	Med	SE	Cov	Med	SE	Cov
All *X* Predictor Distributions Equal *X* ~ *N*(0,1); *C on X* Paths										
Class 1										
Int	0.00	0.00	0.07	0.94	0.00	0.05	0.97	-0.01	0.10	0.93
		-0.01			0.00			0.00		
Slope 1	0.20	0.20	0.06	0.94	0.20	0.05	0.95	0.20	0.09	0.96
		0.20			0.20			0.20		
Slope 2	0.20	0.20	0.06	0.94	0.20	0.05	0.94	0.20	0.09	0.94
		0.20			0.20			0.20		
Resid	0.92	0.91	0.08	0.91	0.91	0.07	0.92	0.90	0.12	0.91
		0.90			0.91			0.90		
X_1	0.00	-0.01	0.11	0.97	0.00	0.14	0.95	0.00	0.12	0.96
		0.00			0.00			0.00		
X_2	0.00	0.01	0.11	0.95	0.01	0.14	0.94	-0.01	0.12	0.96

		0.02			0.01			-0.01		
Class 2										
Int	1.00	1.00 1.00	0.01	0.92	1.00 1.00	0.02	0.96	1.00 1.00	0.01	0.96
Slope 1	0.70	0.70 0.70	0.01	0.93	0.70 0.70	0.02	0.95	0.70 0.70	0.01	0.95
Slope 2	0.70	0.70 0.70	0.01	0.94	0.70 0.70	0.02	0.92	0.70 0.70	0.01	0.96
Resid	0.02	0.02 0.02	0.00	0.92	0.02 0.02	0.00	0.93	0.02 0.02	0.00	0.94
Class 1| *X*$_1$ Variance Increase to *X* ~ *N*(0, 2) with All Equal *X* Means; *C on X* Paths										
Class 1										
Int	0.00	0.02 0.02	0.06	0.95	0.01 0.01	0.05	0.95	0.02 0.03	0.09	0.94
Slope 1	0.20	0.20 0.19	0.04	0.96	0.20 0.19	0.04	0.94	0.19 0.19	0.06	0.92
Slope 2	0.20	0.19 0.20	0.06	0.92	0.20 0.20	0.05	0.94	0.20 0.20	0.09	0.95
Resid	0.84	0.84 0.83	0.08	0.94	0.84 0.83	0.06	0.94	0.81 0.80	0.10	0.89
X$_1$	0.00	0.02 0.02	0.08	0.92	0.03 0.03	0.08	0.89	0.01 0.01	0.12	0.92
X$_2$	0.00	-0.01 -0.02	0.11	0.96	-0.01 -0.01	0.14	0.96	0.00 0.00	0.12	0.98
Class 2										
Int	1.00	1.00 1.00	0.01	0.95	1.00 1.00	0.02	0.96	1.00 1.00	0.01	0.94
Slope 1	0.70	0.70 0.70	0.01	0.94	0.70 0.70	0.02	0.92	0.70 0.70	0.01	0.94
Slope 2	0.70	0.70 0.70	0.01	0.92	0.70 0.70	0.02	0.95	0.70 0.70	0.01	0.95
Resid	0.02	0.02 0.02	0.00	0.89	0.02 0.02	0.00	0.89	0.02 0.02	0.00	0.93
Class 1| *X*$_1$ Mean Increase to *X* ~ *N*(1, 1); All Equal *X* Variances; *C on X* Paths										
Class 1										
Int	0.00	0.01 0.01	0.09	0.95	0.00 0.00	0.08	0.95	-0.02 -0.02	0.13	0.92
Slope 1	0.20	0.20 0.19	0.06	0.97	0.20 0.20	0.05	0.95	0.20 0.20	0.09	0.93
Slope 2	0.20	0.20 0.20	0.06	0.96	0.20 0.20	0.05	0.94	0.20 0.20	0.09	0.95

Resid	0.92	0.91 0.91	0.09	0.93	0.92 0.91	0.07	0.94	0.89 0.89	0.12	0.90
X_1	1.00	0.32 0.91	0.13	0.62	0.31 0.89	0.16	0.62	0.18 0.89	0.14	0.57
X_2	0.00	0.00 -0.01	0.12	0.95	0.00 0.01	0.15	0.94	0.01 0.01	0.13	0.95
Class 2										
Int	1.00	1.00 1.00	0.01	0.94	1.00 1.00	0.02	0.95	1.00 1.00	0.01	0.95
Slope 1	0.70	0.70 0.70	0.01	0.94	0.70 0.70	0.02	0.94	0.70 0.70	0.01	0.95
Slope 2	0.70	0.70 0.70	0.01	0.95	0.70 0.70	0.02	0.95	0.70 0.70	0.01	0.95
Resid	0.02	0.02 0.02	0.00	0.92	0.02 0.02	0.00	0.92	0.02 0.02	0.00	0.93
Class 1 | X_1 Mean and Variance Increase to $X \sim N(1, 2)$; *C on X* Paths										
Class 1										
Int	0.00	0.01 0.01	0.08	0.93	0.02 0.02	0.06	0.94	0.01 0.01	0.12	0.97
Slope 1	0.20	0.20 0.20	0.04	0.92	0.19 0.19	0.04	0.95	0.2 0.2	0.06	0.95
Slope 2	0.20	0.20 0.20	0.06	0.95	0.20 0.20	0.05	0.94	0.19 0.19	0.09	0.94
Resid	0.84	0.83 0.83	0.08	0.91	0.84 0.83	0.06	0.96	0.82 0.81	0.11	0.91
X_1	1.00	0.16 0.61	0.10	0.15	0.25 0.56	0.10	0.06	0.04 0.61	0.13	0.36
X_2	0.00	0.00 0.01	0.11	0.95	-0.01 -0.01	0.14	0.97	0.01 0.02	0.13	0.96
Class 2										
Int	1.00	1.00 1.00	0.01	0.95	1.00 1.00	0.02	0.94	1.00 1.00	0.01	0.95
Slope 1	0.70	0.70 0.70	0.01	0.94	0.70 0.70	0.02	0.93	0.70 0.70	0.01	0.95
Slope 2	0.70	0.70 0.70	0.01	0.95	0.70 0.70	0.02	0.96	0.70 0.70	0.01	0.92
Resid	0.02	0.02 0.02	0.00	0.92	0.02 0.02	0.00	0.92	0.02 0.02	0.00	0.95

4.3 PROPORTION TESTS OF COVERAGE RATES

Two models were estimated in this study—models with and without *C on X* paths for the predictors. Overall, models including the additional covariate path appeared to have better coverage across most conditions, proportion tests were used to better pinpoint potential differences in the 95% coverage rates between the models across conditions. For conditions with and without the *C on X*, 12 two-sample proportion tests were conducted for each parameter. These tests serve to infer

The twelve two-sample tests were the result of comparing the two types of models (i.e.., with and without *C on X* paths) across all combinations of the following factors: (1) mixing weight for the smaller effect size class (i.e., three levels—.25, .50, and .75); (2) mean discrepancy in X_1 across classes (i.e., two levels—0 and 1); (3) variance discrepancy in X_1 across classes (i.e., two levels—1 and 2).

Table 4.5 contains the 95% confidence intervals for the differences in the proportions between the 95% coverage rates for the models with and without the *C on X* paths. A negative difference value indicates that the coverage rate from the model with the *C on X* paths had a higher coverage rate, while a positive value indicates that the model without *C on X* paths had a higher coverage rate.

The differences in parameter coverage, especially for Class 1 intercepts, Class 1|X_1 slopes, Class 1|X_2 slopes, Class 1 residual variances, and Class 2 residual variances are more evident when there was an X_1 mean difference. The model with the *C on X* paths resulted in significantly better coverage rates for 6/6 Class 1 intercepts, 6/6 X_1 slope parameters, 2/3 of the X_2 slope parameters, 5/6 of the Class 1 residual variances, and 6/6 Class 2 residual variances when the population models contained predictor mean differences. Of all the significant differences, there was only one in which the model without the *C on X* paths resulted in better coverage—class 2 intercept in condition 1 vs. condition 13. This finding makes sense because there was no mean difference between the predictors across classes in this condition. However, there were many more instances in which the models including the *C on X* paths led to better parameter recovery, especially in conditions where the predictors associated with β_1 had different means across classes. Between the models with and without the *C on X* paths, the confidence intervals for the differences in the 95% coverage rates for the slope of X_1 ranged in magnitude from 4% for the lower bound to 25% for the upper bound.

Table 4.5: Confidence intervals from proportion tests for coverage rates

Class 1

							Resid	Resid
Condition	Int L	Int U	X_1 L	X_1 U	X_2 L	X_2 U	L	U
1 v 13	-0.02	0.04	-0.02	0.04	-0.03	0.04	-0.01	0.06
2 v 14	-0.05	0.00	-0.07	0.00	-0.02	0.04	-0.01	0.06
3 v 15	-0.01	0.05	**-0.07**	**-0.01**	-0.04	0.03	-0.06	0.02
4 v 16	-0.07	0.00	-0.06	0.00	-0.02	0.06	-0.04	0.03
5 v 17	-0.05	0.02	-0.04	0.03	-0.03	0.04	-0.03	0.04
6 v 18	-0.06	0.02	-0.02	0.06	-0.06	0.01	-0.01	0.07
7 v 19	**-0.15**	**-0.07**	**-0.24**	**-0.16**	**-0.08**	**-0.02**	**-0.18**	**-0.09**
8 v 20	**-0.16**	**-0.08**	**-0.25**	**-0.16**	**-0.09**	**-0.02**	**-0.18**	**-0.09**
9 v 21	**-0.12**	**-0.04**	**-0.20**	**-0.11**	-0.06	0.00	**-0.14**	**-0.05**
10 v 22	**-0.10**	**-0.02**	**-0.14**	**-0.04**	-0.07	0.00	-0.10	0.00
11 v 23	**-0.10**	**-0.01**	**-0.14**	**-0.06**	**-0.09**	**-0.01**	**-0.09**	**-0.02**
12 v 24	**-0.14**	**-0.06**	**-0.14**	**-0.06**	**-0.07**	**0.01**	**-0.14**	**-0.05**

Class 2

							Resid	Resid
Condition	Int L	Int U	X_1 L	X_1 U	X_2 L	X_2 U	L	U
1 v 13	**0.01**	**0.07**	-0.01	0.05	-0.03	0.03	-0.02	0.05
2 v 14	-0.04	0.01	-0.02	0.04	-0.03	0.03	-0.07	0.00
3 v 15	-0.05	0.01	-0.04	0.02	-0.04	0.02	-0.01	0.04
4 v 16	-0.04	0.03	-0.03	0.04	-0.02	0.06	-0.01	0.08

5 v 17	-0.05	0.02	-0.03	0.05	-0.03	0.03	-0.03	0.03
6 v 18	-0.02	0.05	-0.03	0.04	-0.04	0.03	-0.01	0.06
7 v 19	-0.03	0.03	-0.03	0.03	-0.02	0.03	**-0.11**	**-0.03**
8 v 20	-0.03	0.03	-0.03	0.03	-0.02	0.03	**-0.18**	**-0.09**
9 v 21	-0.05	0.01	-0.05	0.01	-0.06	0.01	**-0.13**	**-0.05**
10 v 22	-0.02	0.04	-0.03	0.04	-0.03	0.03	**-0.18**	**-0.09**
11 v 23	-0.01	0.05	-0.07	0.02	-0.07	0.00	**-0.27**	**-0.16**
12 v 24	-0.05	0.02	-0.04	0.03	-0.03	0.05	**-0.14**	**-0.05**

Note. Bold values indicate significant differences at α = .05.

Convergence rates for all conditions were equal to or greater than .99, which suggests that the conditions supported subsequent analyses. Overall, conditions 1-12, which did not include the *C on X* paths correctly enumerated with BIC more often than the models including the *C on X* paths. This difference in enumeration is more apparent in the predictor-variance discrepant conditions. Specifically, conditions 4-6 and 10-12 all had correct enumeration greater than or equal to 98% as compared to conditions 16-18 and 22-24 (which included the *C on X* paths) where the two-class solution was identified between 60.6% and 68.4% of the time. The advantage of not including the *X* variables as covariates predicting class membership was highlighted when the mixing weights were 50/50 and 25/75. Therefore, the difficulty of correctly enumerating regression mixtures when including covariates is compounded when the mixing weight associated with the smaller effect size class is either equal to or smaller than the mixing weight associated with larger effect size class. Overall, the models with the *C on X* paths had better 95% coverage rates for the parameters in common with the models without the *C on X* paths. The differences in parameter coverage, especially for class 1 intercepts, Class 1| X_1 slopes, Class 1| X_2 slopes, Class 1 residual variances, and Class 2 residual variances are more evident when there is an X_1 mean difference. There was an especially important finding related to the ability of models with *C on X* paths led to recover slopes from predictors with mean shifts across classes. The confidence intervals for the differences in the 95% coverage rates for the slope of the *Xs*, when the associated X_1 predictors had different variances across classes, ranged in magnitude from 4% for the lower bound to 25% for the upper bound.

4.4 INCLUDING *C ON X* PATHS AFTER ENUMERATION

As a follow-up to investigating the feasibility of including *C on X* paths after enumeration without those paths, models including *C on X* paths for both predictors were estimated for the same data sets generated for cases 1-12. This examination serves as another test aimed at ruling out the possibility that the overall improved parameter coverages when including the *C on X* paths after enumeration was an anomalous finding (i.e., simply a function of the data sets generated for cases 1-12). Table 4.6 shows the results from the two-class models with *C on X* paths that were estimated for the data originally generated for conditions 1-12. The models with *C on X* paths do not appear to be any worse at parameter recovery, as evidenced by the 95% coverage rates, than the models not including the *C on X* paths.

Table 4.6: Two-class models with C on X paths using data from cases 1-12

		50/50			75/25			25/75		
	True	Mean Med	SE	95 Cov	Mean Med	SE	95 Cov	Mean Med	SE	95 Cov
All *X* Predictor Distributions Equal *X* ~ *N*(0,1); *C on X* Paths										
Class 1										
Int	0.00	0.00 0.00	0.07	0.94	0.00 0.00	0.05	0.94	-0.01 -0.01	0.09	0.95
Slope 1	0.20	0.20 0.20	0.06	0.95	0.20 0.20	0.05	0.92	0.20 0.21	0.09	0.92
Slope 2	0.20	0.20 0.20	0.06	0.93	0.20 0.20	0.05	0.95	0.20 0.20	0.09	0.93

Resid	0.92	0.91 0.91	0.08	0.93	0.91 0.91	0.07	0.94	0.88 0.88	0.12	0.90
X_1	0.00	0.01 0.01	0.11	0.95	0.01 0.01	0.14	0.95	-0.01 0.00	0.12	0.94
X_2	0.00	-0.01 -0.01	0.11	0.96	0.00 0.00	0.14	0.96	0.00 0.00	0.12	0.95
Class 2										
Int	1.00	1.00 1.00	0.01	0.96	1.00 1.00	0.02	0.95	1.00 1.00	0.01	0.94
Slope 1	0.70	0.70 0.70	0.01	0.95	0.70 0.70	0.02	0.95	0.70 0.70	0.01	0.94
Slope 2	0.70	0.70 0.70	0.01	0.94	0.70 0.70	0.02	0.93	0.70 0.70	0.01	0.95
Resid	0.02	0.02 0.02	0.00	0.93	0.02 0.02	0.00	0.89	0.02 0.02	0.00	0.95
Class 1| X_1 Variance Increase to $X \sim N(0, 2)$ with All Equal X Means; *C on X* Paths										
Class 1										
Int	0.00	0.01 0.01	0.06	0.92	0.01 0.01	0.05	0.93	0.03 0.03	0.09	0.92
Slope 1	0.20	0.19 0.19	0.04	0.93	0.2 0.2	0.04	0.93	0.19 0.19	0.06	0.94
Slope 2	0.20	0.20 0.20	0.06	0.94	0.2 0.2	0.05	0.95	0.20 0.20	0.09	0.93
Resid	0.84	0.84 0.84	0.08	0.94	0.84 0.83	0.06	0.94	0.82 0.81	0.10	0.93
X_1	0.00	0.02 0.03	0.08	0.93	0.03 0.03	0.08	0.90	0.01 0.00	0.12	0.94
X_2	0.00	-0.01 -0.01	0.11	0.95	0.00 -0.01	0.14	0.94	-0.01 0.00	0.12	0.96
Class 2										
Int	1.00	1.00 1.00	0.01	0.94	1.00 1.00	0.02	0.94	1.00 1.00	0.01	0.95
Slope 1	0.70	0.70 0.70	0.01	0.95	0.70 0.70	0.02	0.92	0.70 0.70	0.01	0.94
Slope 2	0.70	0.70 0.70	0.01	0.94	0.70 0.70	0.02	0.95	0.70 0.70	0.01	0.95
Resid	0.02	0.02 0.02	0.00	0.92	0.02 0.02	0.00	0.89	0.02 0.02	0.00	0.95

Class 1| X_1 Mean Increase to $X \sim N(1, 1)$ with All Equal X Variances; *C on X* Paths Included

Class 1										
Int	0.00	0.00 0.00	0.09	0.95	0.00 0.00	0.08	0.95	0.00 0.00	0.14	0.93
Slope 1	0.20	0.20 0.20	0.06	0.95	0.20 0.20	0.05	0.93	0.20 0.20	0.09	0.95
Slope 2	0.20	0.20 0.20	0.06	0.96	0.20 0.20	0.05	0.93	0.20 0.20	0.09	0.95
Resid	0.92	0.91 0.91	0.08	0.93	0.91 0.91	0.07	0.92	0.89 0.89	0.12	0.92
X_1	1.00	0.33 0.91	0.13	0.62	0.45 0.93	0.16	0.69	0.30 0.91	0.14	0.61
X_2	0.00	-0.01 -0.01	0.12	0.96	0.01 0.01	0.14	0.96	-0.01 -0.01	0.13	0.96
Class 2										
Int	1.00	1.00 1.00	0.01	0.94	1.00 1.00	0.02	0.95	1.00 1.00	0.01	0.92
Slope 1	0.70	0.70 0.70	0.01	0.95	0.70 0.70	0.02	0.96	0.70 0.70	0.01	0.94
Slope 2	0.70	0.70 0.70	0.01	0.96	0.70 0.70	0.02	0.96	0.70 0.70	0.01	0.93
Resid	0.02	0.02 0.02	0.00	0.95	0.02 0.02	0.00	0.92	0.02 0.02	0.00	0.92

Class 1 | X_1 Mean and Variance Increase to $X \sim N(1, 2)$; *C on X* Paths

Class 1										
Int	0.00	0.02 0.02	0.08	0.96	0.02 0.01	0.06	0.94	0.01 0.01	0.12	0.92
Slope 1	0.20	0.19 0.19	0.04	0.95	0.19 0.19	0.03	0.93	0.20 0.19	0.06	0.94
Slope 2	0.20	0.20 0.20	0.06	0.94	0.20 0.20	0.05	0.93	0.20 0.20	0.09	0.93
Resid	0.84	0.83 0.82	0.08	0.92	0.83 0.83	0.06	0.94	0.82 0.81	0.11	0.90
X_1	1.00	0.18 0.63	0.10	0.14	0.21 0.55	0.10	0.07	0.13 0.68	0.13	0.40
X_2	0.00	0.00 0.00	0.11	0.96	-0.01 -0.01	0.14	0.95	-0.01 -0.01	0.13	0.96
Class 2										
Int	1.00	1.00 1.00	0.01	0.96	1.00 1.00	0.02	0.96	1.00 1.00	0.01	0.94

Slope 1	0.70	0.70 0.70	0.01	0.95	0.70 0.70	0.02	0.92	0.70 0.70	0.01	0.95
Slope 2	0.70	0.70 0.70	0.01	0.96	0.70 0.70	0.02	0.92	0.70 0.70	0.01	0.93
Resid	0.02	0.02 0.02	0.00	0.93	0.02 0.02	0.00	0.88	0.02 0.02	0.00	0.93

Proportion tests were conducted for each of the parameters, comparing the results from conditions 1-12 (i.e., the models that did not include the *C on X* paths), to the models including the paths using the same data. Table 4.7 presents the results from those proportion tests. As was the case with the data originally generated for cases 13-24, estimating models with the *C on X* paths for the data used in cases 1-12 shows no systematic difference in parameter recovery for cases 1-6—those data did not include a mean difference in the X_1 predictors. However, the difference in parameter recovery becomes apparent and statistically significant when there are X_1 predictor mean differences across classes. Concerning the 95% coverage rates, the models including the *C on X* paths outperformed the models with respect to 95% coverage for the Class 1 intercept, Class 1|X_1 slope, and the residual variances from both classes.

Table 4.7: Confidence intervals without/with *C on X* paths– Cases 1-12

Class 1

Cases	Int L	Int U	X_1 L	X_1 U	X_2 L	X_2 U	Resid L	Resid U
1 v 1B	-0.03	0.03	-0.03	0.03	-0.02	0.04	-0.03	0.04
2 v 2B	-0.03	0.03	-0.04	0.03	-0.03	0.03	-0.03	0.04
3 v 3B	-0.03	0.03	-0.03	0.04	-0.04	0.03	-0.05	0.03
4 v 4B	-0.04	0.03	-0.04	0.03	-0.03	0.04	-0.04	0.03
5 v 5B	-0.03	0.04	-0.03	0.04	-0.03	0.03	-0.03	0.03
6 v 6B	-0.03	0.04	-0.04	0.03	-0.04	0.03	-0.04	0.03
7 v 7B	**-0.15**	**-0.07**	**-0.23**	**-0.14**	**-0.09**	**-0.02**	**-0.18**	**-0.09**
8 v 8B	**-0.16**	**-0.08**	**-0.23**	**-0.14**	-0.08	0.00	**-0.16**	**-0.07**

9 v 9B	**-0.14**	**-0.05**	**-0.21**	**-0.12**	-0.06	0.01	**-0.15**	**-0.06**
10 v 10B	**-0.12**	**-0.05**	**-0.17**	**-0.08**	-0.05	0.02	**-0.10**	**-0.02**
11 v 11B	**-0.12**	**-0.04**	**-0.12**	**-0.04**	-0.06	0.01	-0.07	0.00
12 v 12B	**-0.10**	**-0.02**	**-0.13**	**-0.05**	-0.06	0.01	**-0.13**	**-0.04**
Class 2								
							Resid	Resid
Cases	Int L	Int U	*X_1* L	*X_1* U	*X_2* L	*X_2* U	L	U
1 v 1B	-0.03	0.03	-0.03	0.03	-0.04	0.03	-0.04	0.03
2 v 2B	-0.03	0.03	-0.02	0.04	-0.04	0.03	-0.04	0.04
3 v 3B	-0.03	0.03	-0.03	0.03	-0.03	0.03	-0.02	0.03
4 v 4B	-0.03	0.03	-0.03	0.03	-0.03	0.03	-0.03	0.04
5 v 5B	-0.03	0.03	-0.03	0.04	-0.03	0.03	-0.03	0.03
6 v 6B	-0.03	0.03	-0.03	0.03	-0.03	0.03	-0.03	0.03
7 v 7B	-0.03	0.03	-0.04	0.02	-0.02	0.03	**-0.15**	**-0.07**
8 v 8B	-0.03	0.03	-0.05	0.01	-0.04	0.02	**-0.17**	**-0.08**
9 v 9B	-0.03	0.04	-0.05	0.02	-0.04	0.03	**-0.12**	**-0.03**
10 v 10B	-0.03	0.03	-0.03	0.03	-0.04	0.02	**-0.19**	**-0.10**
11 v 11B	-0.02	0.03	-0.06	0.02	-0.03	0.04	**-0.22**	**-0.11**
12 v 12B	-0.03	0.03	-0.03	0.03	-0.03	0.03	**-0.13**	**-0.04**

Note. Bold values indicate significant differences at α = .05.

CHAPTER 5

DISCUSSION

The focus of this study was to determine the ability of regression mixtures to correctly enumerate classes and recover parameters from superpopulations using models with balanced and unbalanced sample proportions each having multiple predictors with equal and unequal means and variances. This study included several conditions encountered in practice, such as: (a) multiple predictors, (b) unbalanced designs (in terms of sample size and parameter size), (c) differing predictor means across latent classes, and (d) differing predictor variances across latent classes. Although previous simulation studies of RMMs have investigated the effects of constraining residual variances and predictor means, in the presence and absence of differences in these terms across classes, this study fills in a gap in the RMM literature related to how well these models handle differences in predictor variances across classes. This was the first study to shed light on the effect of predictor variance differences across classes with respect to enumeration and parameter recovery.

A simulation study was conducted to examine the effects of mixing weights, differences in predictor distributions across classes, and the omission vs. inclusion of *C on X* (i.e., functionally frees estimation of *X* means) paths on enumeration and parameter recovery with regression mixtures. The simulation study consisted of a fully crossed design with 24 cells comprised of 3 sample proportions (i.e., 50/50, 75/25, 25/75) x 2 predictor mean conditions (i.e., Class 1— Normally distributed X_1 with either variance equal to 0 or 1) x 2 predictor variance conditions (i.e., Class 1—Normally distributed X_1 with either variance equal to 1 or 2) x 2 tested models (i.e., with and without *C on X* paths).

The first outcome of interest from the simulation study—class enumeration—indicated a strong preference for initially estimating models, for the purpose of enumeration, without the *C on X* paths or the inclusion of covariates. Class enumeration, which is related to determining the number of underlying subpopulations within the sample, was determined using a penalized likelihood criterion—namely, BIC. The percentages of replications wherein the BIC chose the two-class solution over the one- and three-class solutions and the percentages of replications wherein the BIC chose the three-class solution over the two-class solution were used to evaluate enumeration across conditions. Overall, conditions (i.e., Table 4.2 – cells 1-12), which did not include the *C on X* paths correctly enumerated with BIC more often than the models including the *C on X* paths. However, when there was not a predictor variance discrepancy—either with or without a predictor mean difference—the models with the *C on X* paths resulted in correct enumeration slightly more often. The lowest correct enumeration rate across conditions without the *C on X* paths, where the predictor means are constrained to be equal, was 95%. This indicates that enumerating without the *C on X* paths appears to be robust across conditions with and without predictor mean and variance differences when the mixing weights are balanced and unbalanced. The lowest correct enumeration for the conditions 13-24 (i.e., including the C on X paths) was 52%. This difference in enumeration is more apparent in the predictor variance discrepant conditions.

Specifically, conditions 4-6 and 10-12 all had correct enumeration more often than 98% as compared to conditions 16-18 and 22-24 (which included the C on X paths) where the two-class solution was recovered between 52% and 87% of the time.

Even when there were discrepant variances for the X_1 predictor and discrepant X_1 means and variances, meaning that the true distributions for the X_1 predictors were different across classes, not including the *C on X* paths led to correct enumeration more often than in conditions when the predictors were included as covariates (i.e., *C on X* paths). The advantage of not including the *X* variables as covariates predicting class membership was highlighted when the mixing weights were 50/50 and 25/75. Therefore, the difficulty of correctly enumerating regression mixtures when covariates are included in the model is compounded when the mixing weight associated with the smaller effect size class is either equal to or smaller than the mixing weight associated with larger effect size class.

Results of the current study supported findings from Nylund-Gibson & Masyn (2016), who found that including covariates that predict latent class membership (e.g., *Z* variables and *C on X* paths) led to over-extraction of classes in the enumeration phase. In order to avoid incorrect enumeration, it was suggested that the number of classes should be chosen based on comparisons between models not including covariates (Nylund-Gibson & Masyn, 2016). Then, once the number of classes has been determined, covariates predicting class membership can be added to models. Furthermore, results support findings from Lamont, Vermunt, and Van Horn (2016), who showed that failing to include *C on X* paths to account for predictor mean differences can result in over-extraction of classes. Results from this study revealed that overextraction encountered by not including *C on X* paths in the presence of a predictor mean difference is present to a greater extent when there is a predictor variance difference and the *C on X* paths are included. This means that overextraction is a problem during enumeration when the predictor means are freely estimated, and the true predictor distributions have unequal variances. The percentage of times the model with constrained predictor means resulted in an incorrect extraction of three classes when the BIC should have chosen only two classes never exceeded 2%-- even when there was a predictor mean difference. This provides evidence that enumeration is more sensitive to differences in predictor variances than predictor means. In contrast, with freely estimated predictor means in the presence of a predictor variance difference, resulted in incorrect enumeration ranging between 13% and 48% of the time. Therefore, enumeration should always be conducted using models that do not have freely estimated predictor means (i.e., constrained predictor means – no *C on X* paths).

In line with the results from by Lamont, Vermunt, and Van Horn (2016), this study showed that when the class separation is large, (e.g., intercept difference equal to 1), using BIC for enumeration does not appear to be affected by the inclusion of *C on X*. This result was observed whether there was a predictor mean difference across classes or not (in the absence of predictor variance differences across classes). For the applied researcher, this means that C on X paths should not be included during enumeration, or the process of selecting the number of classes, with regression mixture models. This study helps to better understand enumeration as a variance difference in predictor distributions across classes (with and without predictor mean differences)—*C on X* paths were found to negatively impact enumeration.

Findings from this study suggest that not including covariates, especially in the form of *C on X* paths during enumeration to be especially apparent with regards to regression mixtures when the variances of the predictors are not equal across classes. Therefore, applied researchers using mixture models, should first estimate and enumerate without covariates (i.e., *Z* variables)—including the use of *X* variables with a *C on X* path. Then, after the optimal number of classes has been chosen based on BIC, the model with *k* classes including covariates and *C on X* paths should be estimated to obtain final parameter estimates. Although it is not always the case that predictor means will vary across classes, this study demonstrates that including the *C on X* paths leads to better parameter recovery when the predictor means are indeed different across classes.

Although using the BIC index to identify the optimal number of classes led to correct enumeration more often in models without the *C on X* paths, the models including the predictors as covariates led to better parameter recovery, especially when a predictor mean difference was present across classes, regardless of differences in predictor variances. Differences in 95% coverage rates between conditions with and without the *C on X* paths were analyzed using two-sample proportion tests for each of the common parameters. The two-sample tests of proportions arose from comparing the two types of models (i.e., with and without *C on X* paths) across all combinations of the following factors: (1) mixing weight for the smaller effect size class (i.e., three levels—.25, .50, and .75); (2) mean discrepancy in X_1 across classes (i.e., two levels—0 and 1); (3) variance discrepancy in X_1 across classes (i.e., two levels—1 and 2). The 95% confidence intervals for the differences in the proportions between the 95% coverage rates for the models with and without the *C on X* paths indicated the benefit of including the *C on X* paths after first enumerating without covariates. The difference in coverage rates was most apparent for the Class 1 intercept an X_1 slope parameter, when a predictor mean difference across latent classes was present, proportion tests showed significant differences between coverage rates between the models with and without the *C on X paths*. For each of the conditions in which the X_1 predictor in Class 1 had a greater mean than the X_1 predictor in Class 2, the models with the *C on X* paths obtained greater coverage rates for the X_1 slope parameter in class 1. For the same comparisons, involving cases with mean differences in the X_1 predictor, models including *C on X* paths also obtained significantly greater coverage rates for the Class 1 intercept values and the residual variances for Class 2. These results here are in line with Lamont et al (2016) who also found that failing to include the *C on X* path resulted in reduced parameter coverage rates. Similarly, Kim et al. (2019), illustrated issues related to improperly constraining discrepant residual variances across classes. These findings, when considered alongside findings point to the intuitive understanding that parameter recovery with regression mixtures will be compromised when parameters that are indeed different across classes are held constant in estimation.

Based on results from this study, applied researchers should enumerate RMMs, much like what is done with latent class models –that is, without covariates and *C on X* paths. After the optimal *k* number of classes has been determined, covariates and *C on X* paths aligned with theory should be added in order to avoid biased parameter estimates that would result from unnecessarily constraining predictor means.

5.1 LIMITATIONS AND FUTURE STUDY

Although this study found promising findings related to the utility of regression mixtures run as a two-step procedure, this simulation study used a limited set of conditions with only one total sample size, large class separation, uncorrelated predictors, one discrepant predictor mean condition, one discrepant variance condition, and only two effect sizes across two latent classes. It will be important for researchers to extend this work in order to determine the extent to which two-step regression mixtures are able to correctly enumerate and recover parameters when predictors have multicollinearity. There will also be a need to determine how well the two-step regression mixture procedure is able to operate with different sample sizes, larger numbers of underlying latent classes (i.e., three, four, etc.), and different predictor distribution conditions. Most importantly, it will be necessary to investigate the performance of other multi-step estimation procedures. This study points to the strength of first estimating regression mixtures without covariates, and then after correct enumeration including the *C on X* paths for final parameter estimation. However, a three-step procedure, wherein clusters are first determined using only the predictors should also be compared to the two-step procedure described in this study.

Future research, in addition to testing the claims put forth in this study, should be conducted with regression mixtures involving more predictors belonging to different parametric families. This will be paramount in establishing and understanding the practical applications for these models. This study explained the importance of excluding normally distributed covariates during enumeration, which becomes especially problematic when the variances of the predictors vary across classes. Building on this, interested researchers must investigate regression mixtures involving various types of predictors (i.e., categorical, count, etc.)

5.2 SUMMARY

This study focused on understanding how several data characteristics associated with the investigation of effect heterogeneity (i.e., mixing weights, predictor distributions, and the inclusion of covariates) affected enumeration and parameter recovery with RMMs. The inclusion of *C on X* paths, which allow predictor means to vary across classes, at two points in the model building process—during and after enumeration—was of interest. This main aim was accomplished by comparing the enumeration rates and parameter coverages with and without freely estimated predictor means across classes for models with two classes, considerable separation between groups, and a total sample size of 500. Findings indicated that *C on X* paths, should only be included after enumeration (e.g., Nylund-Gibson & Maysen, 2014). Inclusion of *C on X* paths functionally free the estimation of associated predictor means across classes. If these paths are included in the enumeration phase, over-extraction is typical when predictor variance differences are present. Results from this study supported findings from previous research that demonstrated the necessity of including the *C on X* path when predictor means vary across classes (Lamont, Vermunt, & Van Horn, 2016). Therefore, once the number of classes has been determined, *C on X* paths should be included in models just as researchers would freely estimate residual variances across classes.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Akaike, H. (1973). *Information theory and extension of the maximum likelihood principle*. In B. N. Petrov & F. Csaki (Eds.), Proceedings of the Second International Symposium on Information Theory, Budapest, Hungary (pp. 267-281). Berlin, Germany: Springer.

Bauer, D. J. (2011). Evaluating individual differences in psychological processes. *Current Directions in Psychological Science*, *20*, 115-118.

Belsky, J. (2005). Differential susceptibility to rearing influence: An evolutionary hypothesis and some evidence. In B. Ellis & D. Bjorklund (Eds.), *Origins of the social mind: Evolutionary psychology and child development* (pp. 139-163). New York, NY: Guilford.

Blair, C. (2002). Early intervention for low birth weight preterm infants: The role of negative emotionality in the specification of effects. *Development and Psychopathology*, *14*, 311-332.

Boyce, W. T., Frank, E., Jensen, P. S., Kessler, R. C., Nelson, C. A., Steinberg, L., & Mac Arthur Foundation Research Network on Psychopathology and Development. (1998). Social context in developmental psychopathology: Recommendations for future research from the MacArthur Network on

Psychopathology and Development. *Development and Psychopathology*, *10*, 143-164.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist, 32*, 513-531.

Bronfenbrenner, U. (1989). Ecological systems theory. *Annals of Child Development*, *6*, 187-249.

Bronfenbrenner, U. (Ed.). (2005). *Making human beings human: Bioecological perspectives on human development*. Thousand Oaks, CA: Sage.

Bronfenbrenner, U., & Morris, PA. (1998). The ecology of developmental processes. In R. M. Lerner (Ed.), *Handbook of child psychology*: Vol. 1. *Theoretical models of human development* (5th ed.) (pp. 993 – 1028). Hoboken, NJ: Wiley & Sons, Inc.

Campbell, S. B., Shaw, D. S., & Gilliom, M. (2000). Early externalizing behavior problems: Toddlers and preschoolers at risk for later maladjustment. *Development and Psychopathology*, *12*, 467-488.

Chorpita, B. H., & Barlow, D. H. (1998). The development of anxiety: The role of control in the early environment. *Psychological Bulletin*, *124*, 3-21.

Cleaver, G., & Wedel, M. (2001). Identifying random-scoring respondents in sensory research using finite mixture regression models. *Food Quality and Preference*, *12*, 373-384.

Coie, J. D., Watt, N. F., West, S. G., Hawkins, J. D., Asarnow, J. R., Markman, H. J., Ramey, S. L., Shure, M. B., & Long, B. (1993). The science of prevention: A conceptual framework and some directions for a national research program. *American Psychologist*, *48,* 1013–1022.

Cooper, B. R., & Lanza, S. T. (2014). Who Benefits Most from Head Start? Using Latent Class Moderation to Examine Differential Treatment Effects. *Child Development*, *85*(6), 2317–2338.

Desarbo, W. S., & Cron, W. L. (1988) A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, *5*, 249-282.

Desarbo, W. S., Jedidi, K., & Sinha, I. (2001). Customer value analysis in a heterogeneous market. *Strategic Management Journal*, *22*, 845-857.

Elder, G. H. (1998). The life course developmental theory. *Child Development*, *69*, 1–12.

Ingrassia, S., Minotti, S. C., & Vittadini, G. (2012). Local statistical modeling via a clusterweighted approach with elliptical distributions. *Journal of Classification*, *29*, 363-401.

Jaki, T., Kim, M., Lamont, A., George, M., Chang, C., Feaster, D., & Van Horn, M. L. (2019). The Effects of Sample Size on the Estimation of Regression Mixture Models. *Educational and Psychological Measurement*, *79*(2), 358–384.

Kellam, S. G., Koretz, D., & Moscicki, E. K. (1999). Core elements of developmental epidemiologically based prevention research. *American Journal of Community Psychology*, 27, 463–482.

Kim, M., Lamont, A. E., Jaki, T., Feaster, D., Howe, G., & Van Horn, M. L. (2016). Impact of an equality constraint on the class-specific residual variances in regression mixtures: A Monte Carlo simulation study. *Behavior Research Methods*, 48(2), 813-826.

Klein Velderman, M., Bakersman-Kranenburg, M. J., Juffer, F., & van IJzendoorn, M. H.

(2006). Effects of attachment-based interventions on maternal sensitivity and infant attachment: Differential susceptibility of highly reactive infants. *Journal of Family Psychology*, *20*, 266-274.

Kraemer, H. C., Kierman, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, 27, S101-S108.

Lamont, A. E., Vermunt, J. K., & Van Horn, M. L. (2016) Regression Mixture Models: Does Modeling the Covariance Between Independent Variables and Latent Classes Improve the Results?, *Multivariate Behavioral Research*, *51*(1), 35-52.

Nylund-Gibson, K. & Masyn, K. E. (2016) Covariates and Mixture Modeling: Results of a Simulation Study Exploring the Impact of Misspecified Effects on Class Enumeration, *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 782-797.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

Muthén, B. O., & Asparouhov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society*, Series A, *172*, 639-657.

Muthén, L. K. & Muthén, B. O. Mplus (Version 7.4). Los Angeles, CA: Muthén & Muthén; 1998–2015.

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55, 463-469.

Nylund, K. L., Asparauhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 535-569.

Patterson, G. R., DeBaryshe, B. D., & Ramsey, E. (1989). A developmental perspective on antisocial behavior. *American Psychologist*, *44*, 329–335.

Quandt, R. E. (1972). A new approach to estimating switching regression. Journal of the *American Statistical Association*, *67*, 306-310.

Quandt, R. E., & Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, *73*, 730-738.

R Development Core Team. (2018). R: A language and environment for statistical computing (Version 3.5). Vienna, Austria: R Foundation for Statistical Computing.

Risse, L., Farrell, L., & Fry, T. R. L. 2018. "Personality and pay: do gender gaps in confidence explain gender gaps in wages?" *Oxford Economic Papers*, *70*(4), 919-949.

Rubin, K. H., Burgess, K. B., Dwyer, K. M., & Hastings, P. D. (2003). Predicting preschoolers' externalizing behaviors from toddler temperament, conflict, and maternal negativity. *Developmental Psychology*, *39*, 164-176.

Sampson, R. J., & Laub, J. H. (1993). *Crime in the making: Pathways and turning points through life*. Cambridge, MA: Harvard University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Sclove, L. S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333–343.

Späth, H. (1979). Algorithm 39: Clusterwise linear regression. *Computing*, *22*, 367-373.

Van Horn, M. L., Jaki, T., Masyn, K., Howe, G., Feaster, D. J., Lamont, A. E., . . . Kim, M. (2015). Evaluating differential effects using regression interactions and regression mixture models. *Educational and Psychological Measurement*, *75*, 677-714.

Van Horn, M. L., Jaki, T., Masyn, K., Ramey, S. L., Antaramian, S., & Lemanski, A. (2009). Assessing differential effects: Applying regression mixture models to identify variations in the influence of family resources on academic achievement. *Developmental Psychology*, *45*, 1298-1313.

Van Horn, M. L., Smith, J., Fagan, A. A., Jaki, T., Feaster, D. J., Masyn, K., . . . Howe, G. (2012). Not quite normal: Consequences of violating the assumption of normality in regression mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, 19, 227-249.

Wedel, M. (2002). Concomitant variables in finite mixture models. *Statistica Neerlandica*, *56*, 362–375. doi:10.1111/1467-9574.t01-1-00072

Wedel, M., & Desarbo, W. S. (1994). A review of recent developments in latent class regression models. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 352-388). Cambridge, MA: Blackwell Business.

APPENDIX A

R CODE FOR CONDITION 1 WITH K = 1:2

```
library(stringr)
Datagen=function(num,reps,flnm,n.c1,n.c2,mean.c1.x1,sd.c1.x1,mean.c2.x1,sd.c2.x1,mean.c1.x
2,sd.c1.x2,mean.c2.x2,sd.c2.x2)
{
dat=matrix(NA,ncol=4,nrow=num)

# generate class membership and save to col.3 in datemp
class=c(rep(1,n.c1),rep(0,n.c2))     #class[class==0]=2
dat[,4]=class

# generate covariate 1 and save it to col.1
dat[,1][class==1]= rnorm (sum(class==1),mean=mean.c1.x1,sd=sqrt(sd.c1.x1))
dat[,1][class==0]= rnorm (sum(class==0),mean=mean.c2.x1,sd=sqrt(sd.c2.x1))

# generate covariate 2 and save it to col.2
dat[,2][class==1]= rnorm (sum(class==1), mean.c1.x2,sd=sqrt(sd.c1.x2))
dat[,2][class==0]= rnorm (sum(class==0), mean.c2.x2,sd=sqrt(sd.c2.x2))
# generate error terms
rande1=rnorm(sum(class==1),mean=0,sd=sqrt(.92))
rande2=rnorm(sum(class==0),mean=0,sd=sqrt(.02))

# generate response and save it to col.2
dat[,3][class==1]=dat[,1][class==1]*0.2 + dat[,2][class==1]*0.2 + rande1
dat[,3][class==0]= 1 + dat[,1][class==0]*0.7 + dat[,2][class==0]*0.7 + rande2

file.str=paste(flnm,reps,".txt",sep="")
write.table(dat,file.str,row.names=F,col.names=F)
}

for(i in 1:500) {
Datagen(500,i,"C:/Users/philr/Documents/Reg_Mix/data/case_1/case_1_",250,250,0,1,0,1,0,1,
0,1)
}

###
# This code generates the mplus input file to run a regression mixture for
# one and two classes
reps=i
flnm="C:/Users/philr/Documents/Reg_Mix_2/data/case_1/case_1_"
file.str=paste(flnm,reps,".txt",sep="")
# infile is the data file to be analyzed
# reps is the replication number
# saveloc is the file location to which the estimates will be written
# mpinput is the file location to which the mplus input file will be written

mplusin=function(infile, reps, saveloc,saveloc2, mpinput){
mpmat<-'title:  a latent class model assuming cross-sectional data;'
mpmat<-rbind(mpmat, paste('data: file is ', infile, ';', sep=''))
mpmat<-rbind(mpmat,'variable:')
mpmat<-rbind(mpmat,'')
mpmat<-rbind(mpmat,'names are  x1 x2 y cl; ')
mpmat<-rbind(mpmat,'')
mpmat<-rbind(mpmat,'usevariables are x1 x2 y;')
mpmat<-rbind(mpmat,'classes=c(',k,');')
mpmat<-rbind(mpmat,'')
mpmat<-rbind(mpmat,'analysis:')
mpmat<-rbind(mpmat,'type=mixture;')
mpmat<-rbind(mpmat,paste('STSEED=', sample(1:1000000, 1, replace=FALSE), ';'))
mpmat<-rbind(mpmat,'model:')
mpmat<-rbind(mpmat,'%overall%')
mpmat<-rbind(mpmat,'y on x1 x2;')
mpmat<-rbind(mpmat,'y;')
mpmat<-rbind(mpmat,'%c#1%')
mpmat<-rbind(mpmat,'y on x1 x2;')
mpmat<-rbind(mpmat,'y;')
mpmat<-rbind(mpmat,paste('Savedata: results are ',saveloc, ';', sep=''))

mpmat<-rbind(mpmat,paste('file is ',saveloc2, ';', sep=''))

mpmat<-rbind(mpmat,'save is cprob;')

write(mpmat,mpinput)

}


flnm="C:/Users/philr/Documents/Reg_Mix_2/data/case_1/case_1_"

svname="C:/Users/philr/Documents/Reg_Mix_2/results/case_1/case_1_"

svname2="C:/Users/philr/Documents/Reg_Mix_2/cprob/case_1/case_1_"

inname="C:/Users/philr/Documents/Reg_Mix_2/inputs/case_1/case_1_"


for(k in 1:2) {

for(i in 1:500) {

file.str=paste(flnm,i,".txt",sep="")

sv.str=paste(svname,i,"_",k,".txt",sep="")

sv.str2=paste(svname2,i,"_",k,".txt",sep="")

in.str=paste(inname,i,"_",k,".txt",sep="")

mplusin(file.str,i,sv.str,sv.str2,in.str)

}

}


# setwd sets the work directory, usually this is the location of mplus.exe

setwd("C:/Program Files/Mplus")
```

```
for(k in 1:2) { for (i in 1:500) {

inmat=paste(inname,i,"_",k,".txt",sep="")

inmat=rbind(inmat,"C:/Users/philr/Documents/Reg_Mix_2/results/case_1/temp.txt")

write(inmat,"C:/Users/philr/Documents/Reg_Mix_2/infiles/condition_1.txt")

shell("Mplus < C:/Users/philr/Documents/Reg_Mix_2/infiles/condition_1.txt")

}

}
```