Practical Assessment, Research, and Evaluation

Volume 11 *Volume 11, 2006* Article 11

2006

Using Regression Mixture Analysis in Educational Research

Cody S. Ding

Recommended Citation

Ding, Cody S. (2006) "Using Regression Mixture Analysis in Educational Research," *Practical Assessment, Research, and Evaluation*: Vol. 11 , Article 11.
DOI: https://doi.org/10.7275/wgt2-b390
Available at: https://scholarworks.umass.edu/pare/vol11/iss1/11

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 11 Number 11, December 2006 ISSN 1531-7714

Using Regression Mixture Analysis in Educational Research

Cody S. Ding
University of Missouri-St. Louis

Conventional regression analysis is typically used in educational research. Usually such an analysis implicitly assumes that a common set of regression parameter estimates captures the population characteristics represented in the sample. In some situations, however, this implicit assumption may not be realistic, and the sample may contain several subpopulations such as high math achievers and low math achievers. In these cases, conventional regression models may provide biased estimates since the parameter estimates are constrained to be the same across subpopulations. This paper advocates the applications of regression mixture models, also known as latent class regression analysis, in educational research. Regression mixture analysis is more flexible than conventional regression analysis in that latent classes in the data can be identified and regression parameter estimates can vary within each latent class. An illustration of regression mixture analysis is provided based on a dataset of authentic data. The strengths and limitations of the regression mixture models are discussed in the context of educational research.

Regression models may be one of the most commonly used statistical analysis techniques in educational research. Typically, regression analysis is used to investigate the relationships between a dependent variable (either categorical or continuous) and a set of independent variables based on a sample from a particular population. Often the particular interest is placed on assessment of the effect of each independent variable on dependent variable, and such an effect is considered as the average effect value across all subjects in the sample. For example, if math achievement scores of 500 students are regressed on a measure of their motivation, the value for the slope or the regression coefficient quantifies the average change in math achievement across all 500 students for one unit change in motivation. What this implies is that these 500 students are treated as one homogenous group regarding motivation influences on math achievement, and the implicit assumption is that these students are from the same population with similar characteristics.

As will be described later in this paper, the basic assumption of a homogeneous group is often not realistic. This paper describes the use of regression mixture models as a tool to study the relationships between the dependent variable and a set of independent variables by taking into consideration unobserved population heterogeneity.

BACKGROUND

The general regression model found in any basic statistics test can be written as

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_n + \varepsilon_i \quad (1)$$

where β_0 is intercept; β_k is the regression slope or coefficient for a given independent variable *k*, and ε_i is error term for individual *i*. Equation 1 has one key feature. It assumes that all individuals are drawn from a single population with common population parameters.

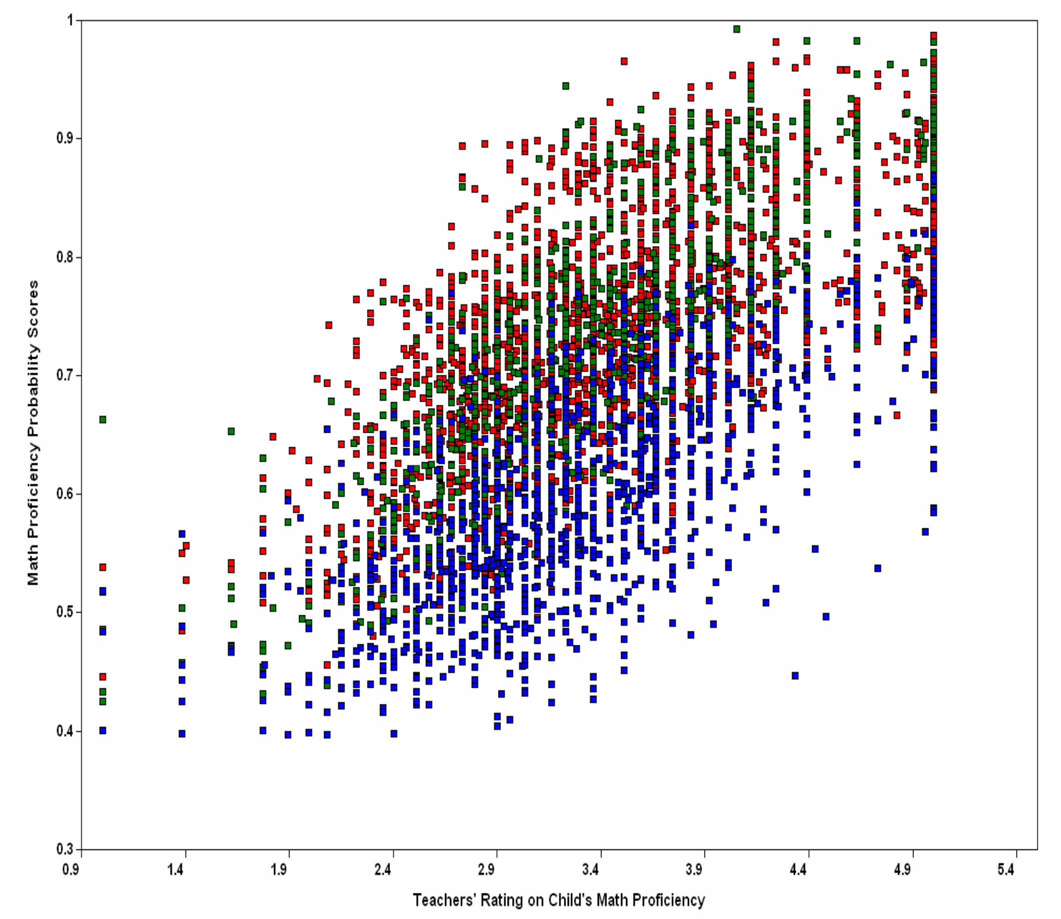
However, when a sample consists of various groups of individuals such as males and females, or different intervention groups, regression analysis can be performed to examine whether the effects of independent variables on a dependent variable differ across groups, either in terms of intercept or slope. These groups can be considered from different populations (e.g., male population or female population), and the population is considered heterogeneous in that these subpopulations may require different population parameters to adequately capture their characteristics. Since this source of population heterogeneity is based on observed group memberships such as gender, the data can be analyzed using regression models by taking into consideration multiple groups. In the methodology literature, subpopulations that can be identified beforehand are called groups (e.g., Lubke & Muthén, 2005; Muthen, 2001).

In this paper, nevertheless, special attention is devoted to the situations in which population heterogeneity is unobserved. In other words, group membership of individuals in the population is latent (McCutcheon, 1987; Waller & Meehl, 1998). For example, students may differ with respect to socioemotional development, and they may belong to either of two qualitatively different types, such as children with high math self-efficacy and children with low math self-efficacy. If we were to study the effect of socioemotional development on student math achievement using a regression model as represented in Equation 1, we would evaluate the average values of intercept and slope (or rate of change) across these two types of students, that is, there is one regression line that describes the relationships between student socioemotional development and math achievement. In this typical regression analysis, the investigator assumes that the sample is from a homogeneous population and that the common parameter estimates are adequate to depict the population characteristics represented in the sample. In other words, conventional regression model assumes that all individuals belong to a single population, and independent variables have the same influence on dependent variable for all individuals. For example, Figure 1 shows the association between teacher's rating on child's math proficiency level and his/her math test score based on a large dataset (the data will be discussed below). As can be seen in Figure 1, it is possible that there may be some distinct subgroups in the data, especially when national representative data are involved. If we ignore such heterogeneity in the data, regression model in Equation 1 may provide biased estimates for the data at hand. For instance, it is possible that children with a larger math gain may be more influenced by math self-efficacy with respect to his/her proficiency level than by school environment, while children with low math gain may be more influenced by both math self-efficacy and school environment. Thus, assumption of population homogeneity may not be realistic. As another example, the variation in reading development among poor readers may be affected more by family environment, whereas the variation in reading development for good readers may be more influenced by teaching methods or vice versa. Moreover, the variances of the residuals may also differ for these two groups of students, and such group differences in variance may contribute to the unequal variance across combinations of the levels of the independent variables.

Although the conventional regression model captures individual differences in intercept and slope, it is not always realistic to assume that a single-population model can account for all kinds of individual differences. Regression mixture models described here are a part of a general framework of finite mixture models (Lubke & Muthén, 2005; Muthen, 2001; Muthen & Muthen, 2000; Nagin & Tremblay, 2001; Vermunt & Magidson, 2002) and can be viewed as a combination of the conventional regression model and the classic latent class model (Lazarsfeld & Henry, 1968; McCutcheon, 1987). It should be noted that there are various types of regression

DOI: https://doi.org/10.7275/wgt2-b390

Figure 1. Scatter plot between children's math proficiency probability scores and teacher's rating of child's math self-concept of proficiency. Different color squares may suggest possible existence of different subpopulations or latent classes of children in the sample.



mixture models (e.g., Vermunt & Dijk, 2001), but this paper will only focus on the linear regression mixture model. The following sections will first describe some unique characteristics of the linear regression mixture model in comparison to the conventional linear regression model, including integration of covariates into the model. Second, a step-by-step regression mixture analysis of empirical data demonstrates how the linear regression mixture model may be used by incorporating population heterogeneity into the model.

Linear Regression Mixture Models

This paper focuses on applications of linear regression mixture models in the situations where population heterogeneity is unobserved (i.e., latent class) and observed group variables such as gender are incorporated in the analysis as covariates. Regression mixture models, also known as latent class regression analysis (Andersen, 2004; Bouwmeester, Sijtsma, & Vermunt, 2004; Vermunt & Magidson, 2005), are used to identify the relationships between the dependent variable and a set of independent variables along with the number of latent classes that best fit the data and to test potential predictors for a given latent class. Unlike conventional regression analysis, which assumes that the regression function in the sample arises from a single multivariate normal distribution, linear regression mixture model allows for heterogeneous regression functions by modeling a mixture of distinct multivariate normal distributions, each corresponding to a latent class. Individuals within each latent class share the same regression function.

Thus, regression mixture analysis relaxes the single population assumption to allow for parameter differences across unobserved subpopulations. This is accomplished by using latent classes, which implies that individuals vary around different regression functions. For example, in a study of the factors that may influence student math achievement, a researcher may include student self-efficacy, motivation, teaching methods, and classroom size as independent variables. The starting point of performing regression mixture analysis is first to identify the number of latent classes that best fit the data. Then the influences of independent variables on the dependent variable can be examined within each latent class. It may be possible that for a given latent class, only self-efficacy has any effect on math achievement, while for a second latent class, math achievement may be influenced by teaching methods and classroom size. Combined use of latent classes with regression models results in a very flexible analysis framework.

Since the linear regression mixture model is a part of finite mixture models (Muthen, 2001), multiple criteria are available to evaluate the number of latent classes for regression analysis because different indices provide information about different aspects of model fit. Comparisons between competing models assess relative fit to the data. For instance, likelihood ratio test (Lo, Mendell, & Rubin, 2001) can be used to compare regression mixture models with differing numbers of latent classes; a significant chi-square value (e.g., $p < .05$) indicates that the specified model is unlikely to be generated by a model with one less class. Also selection of a final model can be based on information criteria, such as Akaike information criterion (Akaike, 1973) or Bayesian information criterion (Schwarz, 1978). Lower observed criterion values are indicative of improved fit. Another index is Entropy (Ramaswamy, DeSarbo, Reibstein, & Robinson, 1993), which assesses the classification accuracy of placing people into classes based on their model-based probabilities. It ranges from 0.00 to 1.00, with higher values indicating better classification. It should be pointed out that although a number of model fit statistics can be used to evaluate a plausible model, the choice of a final model also depends on substantive considerations, previous research results, model parsimony, consistency with theory, and so on. It is difficult to identify the exact number of latent classes that represent true population heterogeneity (Bouwmeester et al., 2004).

A general linear regression mixture model can be formulated as follows:

$$y_{i(c)} = \beta_{0(c)} + \beta_{1(c)}x_1 + \beta_{2(c)}x_2 + \ldots + \beta_{k(c)}x_k + \varepsilon_{i(c)} \qquad (2)$$

Equation 2 has the appearance of a conventional regression model except for the subscript c ($c = 1, 2, \ldots C$). Subscript c in the equation indicates that the parameters may vary around different latent classes. In other words, individuals within each latent class c have the same parameter estimates, which, however, differ across latent classes. In words, Equation 2 says that a dependent variable can be predicted as a function of predictor variables, and a C-category latent class variable c is included, with each category representing a homogenous subpopulation having identical regression coefficients. As mentioned earlier, different types of regression mixture models exist. Depending on the scale type of the dependent variable, various regression mixture models can be estimated. For instance, if a dependent variable is continuous, the linear regression mixture model can be performed, as shown in Equation 2. On the other hand, if the dependent variable is dichotomous or nominal, binary or multinomial logistic regression mixture analysis can be formulated and performed, which would require a substantially different model. Moreover, for models containing $C > 1$ latent classes, covariates such as gender can be included in the model to improve classification of each case into the most likely class, that is, covariates can be used to predict the latent class membership.

Although there are a few software programs that can perform regression mixture analysis, the major computer programs for such an analysis are Mplus (Muthen & Muthen, 2001), GLLAMM (Skrondal & Rabe-Hesketh, 2004), or LatentGold (Vermunt & Magidson, 2005). In the following section, the LatentGold 4.0 program was used to demonstrate the linear regression mixture analysis based on a dataset of real data.

DOI: https://doi.org/10.7275/wgt2-b390

Illustration Of Regression Mixture Analysis

A concrete example in this section provides an illustration of how relationships between independent variables and a dependent variable in the potential presence of population heterogeneity may be investigated with the linear regression mixture model. The latent class variable *c* is used to model unknown heterogeneity, whereas observed group membership variables that are known to introduce heterogeneity are treated as covariates. In linear regression mixture analysis, one needs to specify the number of latent classes. In the model estimation process, the parameters of the model are estimated and the posterior probabilities with which each individual belongs to each of the classes are computed. The results include the model parameters such as within class regression coefficients, within class R^2, within class error variance, etc., and the posterior class probabilities for each individual.

RESEARCH QUESTIONS

To illustrate linear regression mixture analysis in comparison to conventional regression analysis, this example is framed around the following research questions:

1. What is the relationship between children's fifth grade math achievement, children's math self-concept, and teacher's rating on of children's math proficiency, approaches to learning, and self-control?

This research question addressed the issues of (a) whether self-reported math self-concept is predictive of children's math achievement; (b) how predictive teacher judgments of students' academic performance are; and (c) whether teachers' assessment of children's adaptive behaviors and approach to learning predicts children's math achievement.

Marsh, Relich, and Smith (1983) found that math self-concept was most highly correlated with math achievement (r = 0.55). In addition, it has been found that teacher judgment of children's academic competence has concurrent or predictive validity. For example, Hoge and Butcher (1984) found a regression coefficient of 0.71 between teacher's judgment and student's actual scores on standardized tests. In the studies they reviewed, Hoge and Coladarci (1989) indicated that judgment accuracy ranged from 0.28 to 0.92, with median correlation of 0.66. Thus, it would be interesting to replicate such a finding using a national representative sample of actual children.

Regarding teacher's rating of children's social competence, extensive research has taken place regarding the importance of social competence and the skills that contribute to that competence. Social competence has been found to be a significant predictor of academic achievement from K through sixth grade (Clark, Gresham, & Elliot, 1985). On a study of fifth-graders, Walker, Stieber, and Eisert (1991) have found teachers ratings of social skills to be the best predictor of future academic achievement, school adjustment, and delinquency in the next three year period. Therefore, teacher's ratings on approach to learning and self-control were used to see whether some of the findings could be replicated.

2. Do children in different latent classes vary in terms of children's gender and race?

It is important to note that many of the variables may be related to children's math achievement, and they are not explored in this investigation. The variables examined here were just a few of the variables that can/should be examined in the data and were selected to demonstrate the range of information that may be obtained from the linear regression mixture analysis and may help shape the design for the future studies. Readers, however, are cautioned not to draw definitive causal inferences based on the results presented in this example, but rather focus on the proposed analysis paradigm.

METHODOLOGY

Data

The data used in this illustrative analysis were from the Early Childhood Longitudinal Study (ECLS), an ongoing study by the U.S. Department of Education, National Center for Education Statistics that focuses on children's early school experiences beginning with kindergarten (Tourangeau, Nord, Lê, Pollack, & Atkins-Burnett, 2006). The study follows a nationally representative sample of children from kindergarten through fifth grade. The sample reflected all children from various racial and language background. Sampling for the ECLS was based on a dual frame, multi-stage sampling design, with 100 primary sampling units (PSU). For simplicity, only the data collected during 2004 from the fifth graders was in this paper. The sample size in the current analysis was 1,342 children, which included 650 males and 692 females. Among the total analysis sample of children, 797 were White, 126 were Black, 230 were Hispanic, 141 were Asian, and 48 were multiracial.

Measures

In the present analysis, four measures are used as independent variables. They are:

Self-Description Questionnaire—Math Self-Concept (Marsh, 1990). This measure assesses how children think and feel about themselves in terms of math competence. This scale includes eight items on math grades, the difficulty of math work, and interest in and enjoyment of math, with the score scale ranged from 1 to 4. The analysis used the average score of each participant.

Academic Rating Scale-Math. This is the teacher's rating of children's academic performance in math. Teachers were asked to rate each child's proficiency in the following areas: number concepts, measurement, operation, geometry, math strategies, and beginning algebraic thinking, with the score scale ranged from 1 to 5. The analysis used the average score of each participant.

Social Rating Scale-Approach to Learning. This is the teacher's judgment of children's social competence. The approach to learning scale measures behaviors that affect the ease with which children can benefit from the learning environment. It includes six items that rate the child's attentiveness, task persistence, eagerness to learn, learning independence, flexibility, organization, and following classroom rules, with the score scale ranged from 1 to 4. The analysis used the average score of each participant.

Social Rating Scale-Self-Control. It has four items that rate the child's ability to control behavior by respecting the property rights of others, controlling temper, accepting peer ideas for group activities, and responding appropriately to peer pressure, with the score scale ranged from 1 to 4. The analysis used the average score of each participant.

In all above measures, the scores were coded positively, with high scores indicating higher self-concept, and higher teacher rating on academic and social competence. The reported reliability for these independent variables ranged from .79 to .92 (Tourangeau et al., 2006).

Analysis

The dependent variable used is a composite math proficiency probability score that was computed as an average across nine math skill levels: count/number, relative size, ordinality/ sequence, add/subtract, multiple/divide, place value, rate and measurement, fractions, and area/volume. The probability scores were from 0.00 to 1.00, with a larger probability score indicating an overall higher achievement across these math skill levels.

In addition, children's gender and race are included as covariates. They are used to increase the classification accuracy of individuals into each latent class. In this paper, children's race is represented in five categories: White, Black, Hispanic, Asian (which includes Pacific Islanders and American Indians), and multiracial.

Since the scores of dependent variable used are continuous, the appropriate regression mixture model is a linear analysis. The analysis is exploratory with respect to the sources of latent population heterogeneity. Commonly, a key interest in an exploration of population heterogeneity is to determine the number of latent classes that best fit the data.

Therefore, regression mixture models ranging from a 1-class latent model to a 4-class mixture model are tested. The analysis was performed using LatentGold 4.0. In all of these models, the dependent variable is math proficiency probability scores, and the same set of independent variables is used, with child's gender and race as covariates. Among these four models, we sought a model with smallest AIC information criterion values. After 4

DOI: https://doi.org/10.7275/wgt2-b390

classes were extracted, a 3-class regression model performed somewhat better than other models, with AIC being smallest. It was interesting to notice that the model with 1-class had the largest AIC in comparison with other models, which suggested population homogeneity was not likely to be a realistic assumption in the sample. Based on empirical and substantive consideration, the 3-class linear regression model was selected as optimal. In this 3-class model, regression coefficients and error variance were class dependent, that is, they were freely estimated without any equality constraints.

RESULTS

Table 1 provides the regression coefficients for each of the three latent classes, along with the estimated class proportions and the mean math probability scores. Table 2 shows the classification profile information. It can be seen that for Class 1, which consisted of 57% of the sample, math achievement was significantly associated with only teacher's rating on math competence. This variable only accounted for about 49% of the variance in math achievement. What this implied was that for individuals within this class teacher judgment of these children's math competence was statistically accurate in predicting their actual achievement. Other information provided in the analysis, as shown under *Covariates* in Table 1, was that male children were more likely to be members of Class 1 than female children, and White children were also likely to be members of Class 1 than children of other ethnical background. The class proportion size for Class 1 suggested (as shown in Table 2) 57% were male children and 43% were female children. White children consisted of 82% of Class 1 individuals.

For individuals in latent Class 2, their math achievement was significantly associated with teacher rating on math competence and on approach to learning. These two variables accounted for about 63% of variances in math achievement. Thus, it seemed that children with higher math achievement had a higher teacher rating on math competence and approach to learning. This class consisted of 39% of the total sample, of which 61% were female children and 39% were male children. Class 2 also had 33% White children, 18% Black children, 29% Hispanic children, 16% Asian children, and 4% of multiracial children (see Table 2).

For Class 3, children's math achievement was significantly associated with children's math self-concept, teacher rating of math competence and of self-control. Children with high math score, thus, tended to report a higher math self-concept and had a higher teacher rating for math competence and self-control. There was some information about children in Class 3 that was interesting to note: (1) about 95% of the variance in math achievement was accounted for by these three variables; (2) this class consisted of about 4% of the total sample, of which 75% were female children; (3) among these 4% children, 62% were Asian, 15% were Hispanic, 14% were multiracial, 8% Black, and about 1% were White (see Table 2); and (4) White children were less likely to be members of this class (□ = -2.73, *p* < .05).

To contrast the linear regression mixture model with the conventional regression analysis, a conventional regression analysis was performed with the same dependent variable and independent variables, while controlling for gender and race. The results are shown in the last column of Table 1. It can be seen that children's math achievement was significantly related to child's math self-concept, teacher rating on math competence, and teacher rating of approach to learning. Teacher rating of child's self-control was not significantly related to math achievement. Thus, the conclusion could be that on average children who had high math scores tended to report high self-concept in math and had higher teacher ratings of math competence and approach to learning.

Table 1.Parameter Estimates and Model-Based Class Size

	Class 1	Class 2	Class 3	
Class proportion size	57%	39%	4%	
Mean math prob. scores	.75	.63	.66	
	Regression Coefficients			β[b]
Math Self-concept	0.006 (0.005)	0.005 (0.005)	0.034* (0.008)	0.006* (0.003)
ARS-Math	0.084** (0.006)	0.097** (0.005)	0.065* (0.017)	0.089** (0.004)
SRS—Learning	0.015 (0.008)	0.026* (0.006)	0.028 (0.078)	0.023** (0.006)
SRS-Self-Control	0.005 (0.007)	-0.016 (0.015)	0.22** (0.033)	-0.0003 (0.006)
Error Variance	0.005**	0.003**	0.001	0.065
R^2	0.49	0.63	0.94	0.44
Covariates				
Gender				
Male	0.519[a]	-0.044	-0.475	
Female	-0.519	0.044	0.475	
Race				
White	2.320[a]	0.409	-2.730[a]	
Black	-0.525	0.401	0.123	
Hispanic	-0.196	0.168	0.027	
Asian	-1.234	-0.351	1.586	
Multiracial	-0.363	-0.628	0.992	

Note. Standard errors are in parentheses. [a] indicates regression coefficients significantly differ from zero at $p < .05$. [b] indicates regression coefficients from conventional regression analysis. * $p < .05$.

Table 2. Covariates Associated With Latent Class Membership

	Class1	Class2	Class3
Covariates			
Gender			
Male	56.63%	38.81%	25.36%
Female	43.37%	61.19%	74.64%
Race			
White	82.21%	32.62%	0.81%
Black	3.43%	18.47%	7.96%
Hispanic	8.99%	28.87%	14.84%
Asian	3.05%	15.63%	62.32%
Multiracial	2.32%	4.40%	14.07%

CONCLUSIONS

To address the research question regarding relationship between children's math achievement with math self-concept and teacher judgment of math competence and of social competence, the findings indicated that teacher judgment of math competence was statistically accurate in predicting children's math performance across all three latent classes. This was a quite robust finding and replicated the previous findings about accuracy of the teacher judgment (e.g., Hoge & Butcher, 1984). However, child's math self-concept and teacher ratings of their approach to learning and self-control were statistically significantly associated with math achievement only for distinct subgroups of children. That is, this relationship depended on types of children in the population. Thus, the previous findings concerning this association were replicated only for some children, particularly children of specific ethnic groups. For instance, teacher rating of self-control was found to be statistically significantly related to math performance for children who consisted of only 4% of the sample, and 62% of whom were Asian children, and 75% of whom were female children. It was interesting to note that if the conclusions were based on the results from conventional regression analysis, then the previous findings would be replicated in that child's math self-concept would be a strong predicator of actual math performance (Marsh et al., 1983), of social competence, and of approach to learning; however, self-control would NOT be predicative of math performance (e.g., Clark et al., 1985) for "average" children. Population heterogeneity in the sample, therefore, would be completely overlooked and valuable information regarding differential subgroup performance would be lost in explaining mathematics achievement.

DISCUSSION

Regression mixture models are a tool to investigate population heterogeneity. As anticipated, this application of regression mixture modeling to an actual data set indicated that multiple latent classes might be embedded with the single regression functional form. Compared to conventional regression analysis that assumes one equation would fit all individuals, a regression mixture analysis can provide a detailed description of subpopulations of individuals within a sample. In the illustration, the conventional regression analysis revealed only average results across all children, the error variance was quite large, and R^2 was quite small in comparison to the results of linear regression mixture analysis. For instance, the error variance was close to zero and R^2 was 0.94 for Class 3, indicating a good fit between the model and the data from these individuals. In contrast, the conventional regression model had a inferior model-data fit. Thus, regression mixture models may improve predictability because the individual differences are systematically classified to form homogeneous groups. The regression mixture analysis resulted in subpopulations with specific patterns of regression function, and with differing proportions of female and ethnical children.

It should be pointed out that regression mixture modeling is a different analytical technique for studying population heterogeneity than multiple group modeling. The purpose of regression mixture analysis is to identify differing regression functions across latent classes, and such an approach is appropriate if the interest is in detecting and characterizing the relationships among variables according to subpopulations of individuals. The observed grouping variables such as gender may be used as covariates to help predict the latent class membership. For instance, in the illustration, Class 1 is predicted by gender and race, while Class 2 is not predicted by either grouping variables. Thus, the latent class has a different interpretation, and it is used to describe a different kind of heterogeneity in the sample. But one should realize that classification of individual into latent classes is model dependent and it is not intrinsic to the individuals in the sample (Lubke & Muthén, 2005). On the other hand, the purpose of multiple group regression analysis is to compare these groups with respect to their regression functions, and the observed group membership is an intrinsic characteristic of the individual (such as individuals are either male or female).

Regression mixture analysis is not without its limitations. First is the determination of the proper number of latent classes in the data. As Bauer and

Curran (2003) suggested,, mixture modeling can detect population heterogeneity as well as distribution skewness. If there exists non-normality within class, non-normality of observed variables, or non-linearity, the latent class may simply describe the skewness, and may not reflect latent classes of individuals in the sample. Thus, in addition to ensuring the normality and linearity assumptions, one should also consider at a conceptual level whether an additional class is providing meaningful information about the heterogeneity.

Second, a model identification index such as AIC may not provide sufficient evidence for models of heterogeneity (Bauer & Curran, 2004). There is no consensus, so far, regarding which model identification index can be used to select "best" models. Therefore, ambiguity in model selection will continue. In this paper, linear regression mixture analysis is used as one possible way of exploring the data; such an approach is similar to conventional exploratory regression analysis and results should be regarded as preliminary. Independent replication of the study would be essential for generalizing the results.

Readers should keep these limitations in mind when applying regression mixture models. But it seems that regression mixture models are a useful tool and can be used to model heterogeneity in regression function, thus leading to improved regression solutions. In a sense, conventional regression models are a special case of regression mixture models where only one class is assumed and aggregate regression function is concerned. However, it would be necessary to investigate this constraint that a set of common parameter estimates is sufficient to capture the population characteristics. Regression mixture models, on the other hand, places the regression structure in a much more flexible way.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267-281). Akademiai Kiado, Budapest.

Andersen, E. B. (2004). Latent regression analysis based on the rating scale model. *Psychology Science, 46*(2), 209-226.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods, 9*, 3-29.

Bauer, J. D., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods, 8*, 338-363.

Bouwmeester, S., Sijtsma, K., & Vermunt, J. K. (2004). Latent class regression analysis for describing cognitive developmental phenomena: An application to transitive reasoning. *European Journal of Developmental Psychology, 1*(1), 67-86.

Clark, L., Gresham, F. M., & Elliot, S. N. (1985). Development and validation of a social skills assessment measure: The RROSS-C. *Journal of Psychoeducational Assessment, 4*, 347-356.

Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement level. *Journal of Educational Psychology, 76*, 777-781.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research, 59*, 297-313.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin Company.

Lo, Y., Mendell, N., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767-778.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*(1), 21-39.

Marsh, H. W. (1990). *Self-Description Questionnaire Manual.* Campbelltown N. S. W., Australia: University of Western Sydney, Macarthur.

Marsh, H. W., Relich, J., & Smith, I. D. (1983). Self-concept: The construct validity of interpretations based upon the SDQ. *Journal of Personality and Social Psychology, 45*, 173-187.

McCutcheon, A. L. (1987). *Latent class analysis.* Thousand Oaks, CA: Sage Publications, Inc.

Muthen, B. O. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In L. M. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 291-322). Washington, DC: APA.

Muthen, B. O., & Muthen, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24*, 882-891.

Muthen, L. K., & Muthen, B. O. (2001). Mplus User's Guide. Los Angeles, CA: Muthen & Muthen.

Nagin, D., & Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods, 6*, 18-34.

Ramaswamy, V., DeSarbo, W., Reibstein, D., & Robinson, W. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science, 12*, 103-124.

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

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

Tourangeau, K., Nord, C., Lê, T., Pollack, J. M., & Atkins-Burnett, S. (2006). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Combined User's Manual for the ECLS-K Fifth-Grade Data Files and Electronic Codebooks (NCES 2006–032).* Washington, DC: U.S. Department of Education: National Center for Education Statistics.

Vermunt, J. K., & Dijk, L. V. (2001). A nonparametric random-coefficients approach: The latent class regression model. *Multilevel Modeling Newsletter, 13*(6-13).

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & L. M. Allan (Eds.), *Applied latent class analysis.* Cambridge: Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2005). *Latent Gold 4.0 user's guide.* Belmont, MA: Statistical Innovations Inc.

Walker, H. M., Stieber, S., & Eisert, D. (1991). Teacher ratings of adolescent social skills: Psychometric characteristics and factorial replicability across age-grade ranges. *School Psychology Reivew, 20*(2), 301-314.

Waller, N., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks, CA: Sage Publications, Inc.

Citation

Ding, Cody (2006). Using Regression Mixture Analysis in Educational Research. *Practical Assessment Research & Evaluation*, 11(11). Available online: http://pareonline.net/getvn.asp?v=11&n=11

Authors

Correspondence concerning this paper should be addressed to

Cody Ding
404 Marillac Hall
Division of Educational Psychology
University of Missouri-St. Louis
St. Louis, MO 314-516-6562

Email: dinghc [at] umsl.edu