The 6th International Symposium on Emerging Information, Communication and Networks (EICN 2019) 
November 4-7, 2019, Coimbra, Portugal

Mixture Structural Equation Models for Classifying University Student Dropout in Latin America

Amelec Viloriaa*, Omar Bonerge Pineda Lezamab

*Universidad de la Costa, Cl. 58 # 55 – 66, Barranquilla 080001, Colombia 
*bUniversidad Tecnológica Centroamericana (UNITEC), San Pedro Sula 21101, Honduras

Abstract

This research seeks to develop a model that allows consider the different forms of heterogeneity of international dropout data and also classify students who continue studying and those who drop out. Specifically, through the use of Mixture Structural Equation Models (MSEM), the study seeks to develop a model for classifying dropout and applying it to an international database. The aim is then to determine the classification accuracy degree of the proposed model. The development and application of the model showed that the student’s health, the interpersonal relationships, and class attendance positively influence college adaptation, and in turn college adaptation positively influences college satisfaction. Additionally, the developed model can correctly classify 55.45% of continuing students and 61.68% of students who abandon their careers. These results suggest that the use of MSEM for international databases, characterized by heterogeneity, allows more robust and generalizable studies of dropouts in higher education.

© 2019 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the Conference Program Chairs.

Keywords: Mixture Structural Equations Models: Student Satisfaction: Student Dropout.

* Corresponding author. Tel.: +57-3046238313.
E-mail address: aviloria7@cuc.edu.co

1877-0509 © 2019 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2019.11.036
1. Introduction

Nowadays, Mixture Structural Equation Models are receiving greater attention in social sciences. This is because such models are appropriate when the data come from non-homogeneous populations, but rather from a mixture of different types of sampling units where the group of belonging is unknown. The structure of groups can be simple, as in the analysis of latent classes [1], or complex, as the mixture structural equation models (MSEM) [2]. In the case of MSEM, they are considerably attractive because they solve methodological challenges that other analysis methods are not able to solve. For example, [3] found that the maximum likelihood method for estimating structural equation models (SEM) may contain multiple solutions or degenerate solutions, where the identified groups have identical conditional probabilities. This paper presents an example of MSEM that overcomes the methodological challenges of SEM and allows the study of dropout in Latin American higher education. The objectives are: (a) develop a dropout classification model based on MSEM; (b) test the model using an international database (Alfa-Guia Project); and (c) determine the classification accuracy degree of the proposed model.

Currently, there is considerable debate about the appropriate method for measuring dropout in higher education at the international level. For example, according to [4], data related to higher education dropout provides a limited basis for comparative analysis at the international level because there is a very large difference between the existing systems of analysis. However, there are international studies that use standardized assessment instruments. For example, the study carried out by GUIA Project used standardized instruments applied to 21 higher education institutions: 5 from Europe and 16 from Latin America [5]. As a result of GUIA Project, several researches that seek to explain the phenomenon of abandonment using the obtained data were published [6] [7] [8]. These research works represent a significant advance in the study of abandonment in Latin America. However, the studies were conducted in specific countries and the findings were not generalized to the rest of the participating regions. Consequently, the Latin American literature on higher education dropout lacks studies with models of general application for different countries [9].

2. Estimation of MSEM for the study of dropout

Multiple linear regression models are one of the most widely used methods of statistical analysis in educational research [10]. Multiple linear regressions assume that all individuals are selected from a single group with common population parameters. However, in educational research such an assumption of homogeneous groups may be unrealistic. In particular, a sample of students may consist of students of different genders, courses, or enrollment status (students continuing or dropping out). The above suggests that multiple linear regression models should only be used when the effects of independent variables on the dependent variable do not differ among the groups under analysis. Because the source of population heterogeneity is based on the observed membership of each group, such as continuing or dropping out of school, data should be analyzed using methods that take into account the existence of multiple groups. In the case of the SEM methodology, according to [12], SEM models are characterized by two components.

The first component is a confirmatory model that relates latent variables with all their corresponding manifest variables (indicators) and takes errors into account. This component can be considered as the regression model that relates manifest variables to a small number of latent variables. The second component is again a type of structural regression that relates endogenous (dependent) latent variables to the linear terms of some endogenous and exogenous (independent) latent variables. Since latent variables are random, they cannot be directly analyzed by conventional techniques such as multiple linear regression models that are based on direct observations. For SEM models, the random manifest vector covariance matrix contains all the unknown parameters of the model. Thus, the classical methods of SEM analysis focus on the covariance matrix sample and not on the direct measurements of the random vectors. When the distribution of the random vector is normal multivariate and the sample size is large enough, the asymptotic distribution of the covariance matrix approximates exactly the normal multivariate distribution, and the result can be considered correct [13] [14] [15].
However, in more complex situations, which are common in educational research, the analysis of the covariance structure may encounter theoretical and computational problems [16]. For example, for dichotomous, categorical, or ordinal measurements, the covariance matrix sample cannot be used. Additionally, the existence of subgroups in the data can cause observations to be correlated, leading to problems in the covariance structure with the covariance matrix sample. Thus, the use of international data can be challenging due to the existence of subgroups that cause problems in the covariance structure. Consequently, if heterogeneity in internationally sourced data is ignored, multiple linear regression models and endogenous and exogenous relationships of SEM models can give biased estimates. This leads to the conclusion that the use of international databases should take into account the existence of subgroups that break homogeneous, normal multivariate population assumptions, with randomly distributed missing data and that would lead to the calculation of biased estimates if SEM is used. MSEMs are a hybrid approach between SEM and latent class analysis and were proposed by [17].

MSEMs are considered as an extension of the conventional SEM model, similar to multi-group analysis. However, an important difference between MSEM and standard multigroup analyses is that MSEM group membership is latent. By integrating latent classes into the SEM model, different forms of unobserved heterogeneity can be detected. Methodologically speaking, the subpopulations that can be identified a priori are called groups [18]. However, there are situations in which group membership is not clearly defined. This research pays special attention to situations in which the heterogeneity of the population is unknown, such as dropout phenomena in higher education. In other words, our research considers that the group membership of students in the population can be considered as latent [3] and could be determined according to methods that take into account the heterogeneity of the data. Consequently, our research seeks to develop a model that allows to consider the different forms of heterogeneity of international dropout data and to classify students who continue and drop out of their studies.

3. The Mixture Structural Equation Model

According to [14], MSEMs should not be used unless the membership of each independent observation is known and can be specified precisely. In general, MSEMs arise from a mixed population of K components with associated probability densities and mixed proportions. These types of models are used for heterogeneity modeling, extreme data management, and density estimation. Techniques such as Bayesian methods combined with Markov Chain Monte Carlo (MCMC) methods are used for this purpose. The mathematical details of MSEM estimation with Bayesian methods combined with MCMC methods used for this study are developed in detail in [13] [14] [17]. In general, these types of methods consider that the distribution of latent variables can be estimated from the data, rather than being assumed in some parametric way. According to [5], Bayesian methods combined with MCMC methods are the art of recovering hidden groups of observed data based on a statistical model. Specifically, from the point of view of higher education dropout, it is assumed that the institution does not know which students are continuing and which students dropped out. Consequently, it is sought that the model can correctly "guess" whether the student drops out or continues to study. In general, the algorithm "guesses" the parameters of the group distributions. It calculates the subsequent distribution of the groups. Then, it updates the parameters after obtaining a convergent result where the parameters of the group distributions are stable.

4. Method

The data used for this illustrative analysis was extracted from the GUIA Project repository [17]. The sample obtained from the repository consists of 12548 respondents from 14 countries in 19 institutions. Specifically, 48% of the respondents were under 19 years of age and 52% of the respondents were 20 years of age or older. Moreover, 50.6% of respondents were men and 49.4% were women.

Based on the GUIDE Project repository, the following variables were used for the analysis: (a) student’s health, with a 5-point ordinal scale that measures the student's health status during his or her stay at the university; (b) class attendance, a 5-point ordinal variable that measures whether the student regularly attended the regular courses; (c) interpersonal relationships, which is a latent variable that consists of two 5-point ordinal measures that assess the student's relationship with his or her teachers and classmates during his or her stay at the university; (d) college
adjustment, which is a latent variable consisting of two 5-point ordinal measures and seeks to measure the student's academic and social adjustment to college life; (e) student satisfaction, a latent variable that consists of three 5-point ordinal measures and that seeks to measure the degree of satisfaction with the orientation of the plan and study programs, the coordination between subjects and the content of the subjects; and (f) student abandonment, which is a binomial variable that measures whether the student is enrolled in the same career in which he was enrolled in the period 2016-2018. The reliability of the three latent variables was then calculated using Cronbach's Alpha, which was greater than 0.68 for all variables.

Since the total sample of 12548 respondents was characterized by: (a) missing data that did not meet the axiom random distribution (MAR); and (b) unbalanced groups for students who continued or dropped out of their careers; a sample of 6158 data was randomly obtained. This statistical manipulation was performed because, according to [14], for the correct application of mixture structural equation models, the data must meet the MAR criterion and have balanced samples for the variables under analysis. Next, a learning set was randomly generated with 50% of the sample and the proposed model was implemented using AMOS 22 through the module for mixed models with training data.

5. Results

Table 1 presents the standardized regression coefficients for students who continue to study and those who dropped out of their careers, the classification profile of the data obtained using Bayesian methods combined with MCMC methods, and for students who continue to study their career (51.33% of the sample), all hypotheses were validated because the coefficients are positive and greater than zero with 95% confidence. Similarly, for students who dropped out of their careers (48.15% of the sample), all hypotheses were also validated with positive coefficients and 95% confidence that their values are different from zero.

Consequently, the study presents evidence not to reject any of the proposed hypotheses (H1-H5) both for students who continue to study and those who abandoned their careers, as well as the proportions of cases correctly classified by Bayesian methods combined with MCMC methods. Specifically, the developed model was able to correctly classify 55.45% of the students who continue to study their careers, and 61.68% of the students who dropped out of their careers. These results are promising since they show that the proposed model performs a correct classification of cases better than randomly. Moreover, Table 1 also shows the AUC for the ROC curve that is used to determine the sensitivity of binary classifiers. The AUC value obtained for the mixture model was 60.75% and, according to [3], can be considered within the acceptable range (60 to 80%). Consequently, the mixture structural equation model performs an acceptable classification of students who continue or dropped out of their careers.

Table 1. Regression Standardized Coefficients and data classification profile

<table>
<thead>
<tr>
<th>Relationships</th>
<th>Hypothesis</th>
<th>Students who continue their career</th>
<th>Students who dropped out of their career</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Regression Coefficient</td>
<td>Confidence Interval at 95%</td>
</tr>
<tr>
<td>Student health → College adjustment</td>
<td>H1</td>
<td>0.17</td>
<td>0.10, 0.17</td>
</tr>
<tr>
<td>Interpersonal Relationships → College adjustment</td>
<td>H2</td>
<td>0.45</td>
<td>0.48, 0.69</td>
</tr>
<tr>
<td>Attendance to classes → College Adjustment</td>
<td>H3</td>
<td>0.085</td>
<td>0.069, 0.178</td>
</tr>
<tr>
<td>Interpersonal Relationships → Student satisfaction</td>
<td>H4</td>
<td>0.19</td>
<td>0.154, 0.28</td>
</tr>
<tr>
<td>College adjustment → Student satisfaction</td>
<td>H5</td>
<td>0.052</td>
<td>0.035, 0.27</td>
</tr>
<tr>
<td>-Proportion -Percentage of correctly classified cases.</td>
<td></td>
<td>51.33%</td>
<td>48.15%</td>
</tr>
</tbody>
</table>
6. Conclusion

The results presented in Table 1 show that student health, interpersonal relationships, and class attendance contribute positively to college adjustment. Then, adjustment to college positively influences satisfaction with college. These findings were robust for both groups under analysis with a 95% confidence level that they are positive and different from zero. The results also suggest that, compared to student health and class attendance, interpersonal relationships have the greatest influence on college adjustment. On the other hand, the results suggest that, for students continuing their careers, interpersonal relationships have the greatest influence on student satisfaction. For students who dropped out, college adjustment has the greatest influence on student satisfaction. Together, these results suggest that the factors influencing their adjustment to college for both groups are similar. However, the factors that determine their satisfaction with college differ for both groups. This difference is due to the fact that students who dropped out of college show that the importance of adapting to college is greater, because their levels of adaptation to college were lower. In the case of the second research question, the aim was to determine whether the proposed causal relationships allowed the correct classification of those students who continued or abandoned their careers.

The proposed model allowed classify 55.45% of students who continue their careers and 61.68% of students who dropped out of their careers. By applying the ROC curve, an acceptable AUC value (60.75%) was obtained. As a result, the proposed model made it possible to classify students who continued or dropped out of their careers at a better than randomly. These results suggest that this model is applicable to the different contexts of Latin America since it correctly "guessed" the distribution parameters of the groups of students until converging to a result of stable parameters for the countries participating in the sample. However, the AUC could be improved by adding other variables that influence student satisfaction. The proposed research tool makes it possible to investigate data with heterogeneous data distributions. As anticipated, the application of MSEM in real data indicates the existence of groups (continuing students vs. academic dropouts) with non-normal distributions that cannot be analyzed by common analysis methods, such as multiple linear regression or SEM. The proposed method does not assume that the results are applicable to all individuals in the sample; rather, it gives a detailed description of the subpopulations of individuals within the sample. In this sense, the proposed method allows the identification of different parameters for the characterization of relationships between variables according to student subpopulations.

References


