Model Validation and Student Life Tables for Post-Secondary Enrollment

David Spencer, University of North Carolina Chapel Hill
Diana Gonzalez, California State Polytechnic University Pomona
John Lagergren, East Tennessee State University

Objectives

This report attempts to apply the methods of actuarial science to student progression in colleges and universities. It considers several possible models of progression expressed in matrix form, and fits these models to statewide undergraduate enrollment data obtained from the Integrated Postsecondary Education Data System. The best model as determined by the Akaike Information Criterion is used to create life tables.

Introduction

A seminal model for ecological population growth is known as the Leslie Model. In this model, females are grouped into age segments of equal size. The number of females in each age group i at iteration n is denoted $x_i(n)$, with the vector of these values denoted $x(n)$. Each group i is assumed to have a fixed probability that a member just entering the group will die before leaving the group, p_i, and a fixed average number of children per female while in that age group, b_i. These assumptions give the following model:

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \\ x_4(n) \end{bmatrix} = \begin{bmatrix} b_1 & b_2 & b_3 & b_4 \\ p_1 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 \\ 0 & 0 & p_3 & 0 \end{bmatrix} \begin{bmatrix} x_1(n-1) \\ x_2(n-1) \\ x_3(n-1) \\ x_4(n-1) \end{bmatrix} \qquad (1)$$

These parameters are subject to the constraints $0 \leq p_i \leq 1$ and $0 \leq b_i$. It turns out that, under the assumption that $\exists j \in \mathbb{N}$ with $b_j > 0$ and $b_{j+1} > 0$, the proportions of the total population in each age group approach constants as $n \to \infty$. This set of proportions at the limit is referred to as the stable age distribution.

The Leslie Model, though simple, manages to capture some of the basic features of ecological populations. We attempt to apply this model to student progression, grouping students by year of study rather than by age as in ecology, and having students drop out rather than having members of the population die. However, modeling student populations requires different assumptions, because there is no clear analogue of childbirth for undergraduate populations. One way to resolve this is to add a vector representing immigrants to the system (in other words, new students), which may or may not have nonzero entries after the first entry. The Leslie Model and the concept of an immigration vector led us to propose four models of student population dynamics, which we compare and analyze.

Data

Our data come from the Integrated Postsecondary Education Data System published by the National Center for Education Statistics. From this database, we were able to obtain statewide data by year on the number of first-year, second-year, third-year, and fourth-or-higher-year students at each school. We were also able to obtain the number of first-time freshmen at each school, which we denote $u(n)$ in the year n. We considered only Title IX compliant public schools. Grouping these schools by state (including the District of Columbia) gave us our dataset.

Models

Model 1

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \\ x_4(n) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ p_{12} & 0 & 0 & 0 \\ 0 & p_{23} & 0 & 0 \\ 0 & 0 & p_{34} & 0 \end{bmatrix} \begin{bmatrix} x_1(n-1) \\ x_2(n-1) \\ x_3(n-1) \\ x_4(n-1) \end{bmatrix} + \begin{bmatrix} \beta_1 n + \alpha_1 \\ \beta_2 n + \alpha_2 \\ \beta_3 n + \alpha_3 \\ \beta_4 n + \alpha_4 \end{bmatrix} \qquad (2)$$

p, α, and β are parameters. Model 1 closely resembles the Leslie model. However, instead of computing births from $x(n-1)$, new individuals enter each age group via an immigration vector with four entries that are linear in time. New individuals can enter in any group, rather than just the first group as in the Leslie model. Generally, we observed that populations approach a stable age distribution in Model 1, as in the Leslie Model.

Model 2

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \\ x_4(n) \end{bmatrix} = \begin{bmatrix} p_{11} & 0 & 0 & 0 \\ p_{12} & p_{22} & 0 & 0 \\ 0 & p_{23} & p_{33} & 0 \\ 0 & 0 & p_{34} & p_{44} \end{bmatrix} \begin{bmatrix} x_1(n-1) \\ x_2(n-1) \\ x_3(n-1) \\ x_4(n-1) \end{bmatrix} + \begin{bmatrix} u(n) \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad (3)$$

Model 2 is different from Model 1 in two ways:

1. Model 2 assumes that data on freshman enrollment are available. Many data sets do provide freshman enrollment, including those we studied.
2. Model 2 has nonzero entries on the main diagonal. This is to account for students who remain in the same classification for multiple years.

Model 3

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \\ x_4(n) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ p_{12} & 0 & 0 & 0 \\ 0 & p_{23} & 0 & 0 \\ 0 & 0 & p_{34} & p_{44} \end{bmatrix} \begin{bmatrix} x_1(n-1) \\ x_2(n-1) \\ x_3(n-1) \\ x_4(n-1) \end{bmatrix} + \begin{bmatrix} u(n) \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad (4)$$

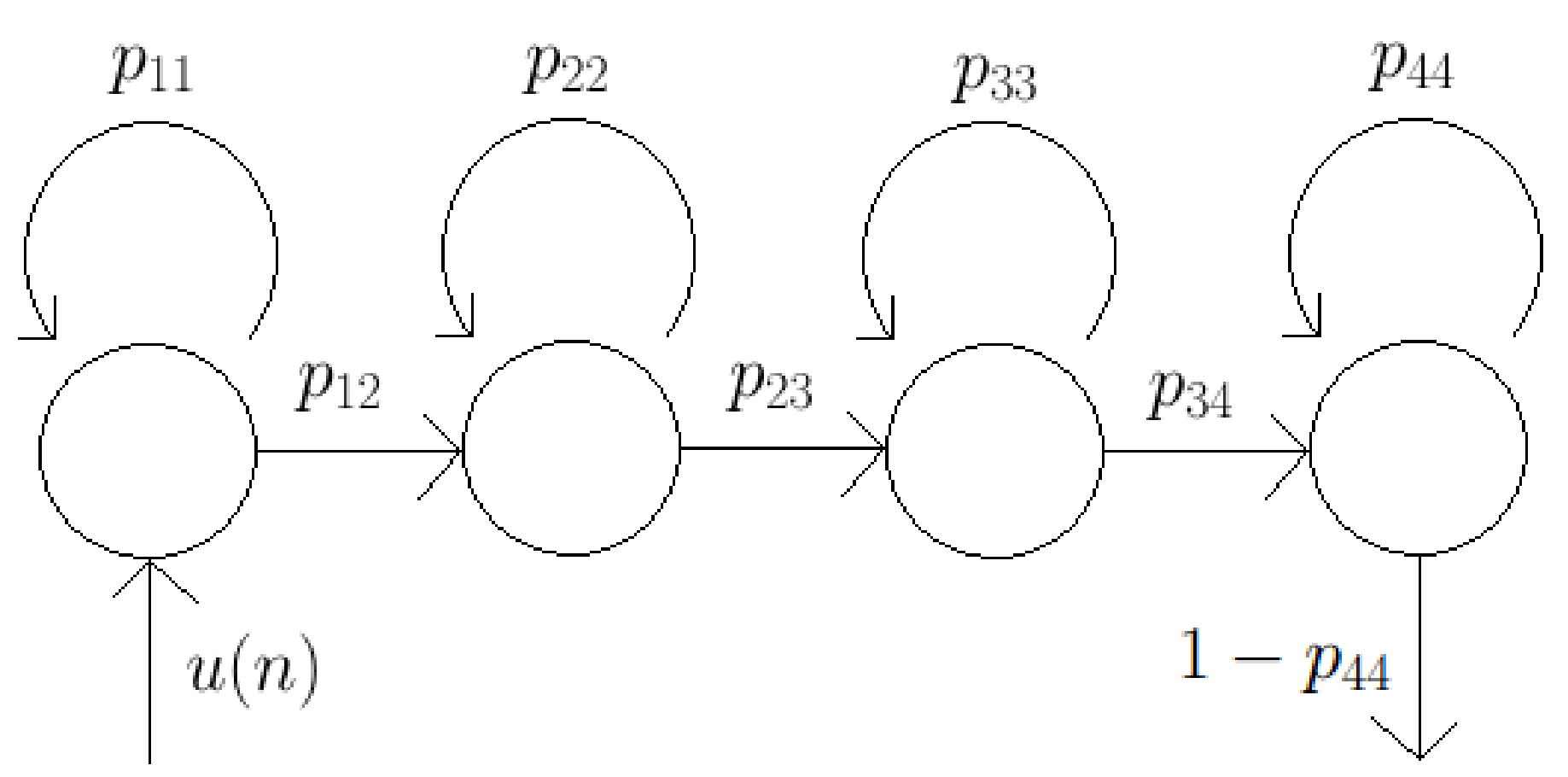
Model 3 is a modification of Model 2 in which we assume that the fourth year is the only year that can be repeated. The motivation for this is that the data sets used group together students who have been at school for four years or longer rather than providing figures for the number of students that have been in school for four, five, or six years.

Model 4

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \\ x_4(n) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ p_{12} & 0 & 0 & 0 \\ 0 & p_{23} & 0 & 0 \\ 0 & 0 & p_{34} & p_{44} \end{bmatrix} \begin{bmatrix} x_1(n-1) \\ x_2(n-1) \\ x_3(n-1) \\ x_4(n-1) \end{bmatrix} + \begin{bmatrix} \beta_1 n + \alpha_1 \\ \beta_2 n + \alpha_2 \\ \beta_3 n + \alpha_3 \\ \beta_4 n + \alpha_4 \end{bmatrix} \qquad (5)$$

Model 4 is a hybrid of Model 1 and Model 3, which gives the model more control over what happens in the earlier years than Model 3.

Figure 1: A visual representation of student progression in Model 2.



Methods

We formulated four matrix models of student progression. Using the IPEDS data, we performed least-squares fits to each model, and compared the results using the Akaike Information Criterion. The Akaike Information Criterion is a metric that provides information about the relative suitability of models to data, though it makes no statement on the absolute suitability of any of the models. The AIC value for a model is

$$AIC = 2k - 2ln(L) \qquad (6)$$

where k is the number of parameters in the model and L is the least-squares sum for the best fit. Models with smaller AIC values are, by this metric, preferable to models with larger AIC values.

After computing the AIC values to determine the best-suited model, we performed the Durbin-Watson test to check for autocorrelation in the residuals. The Durbin-Watson statistic is the quantity

$$\rho = \frac{\sum_{j=2}^{N} r_j r_{j-1}}{\sum_{j=1}^{N} r_j^2} \qquad (7)$$

where r_j is the residual and N is the number of observations. Sufficiently small ρ is evidence of positive correlation between consecutive residuals; conversely, sufficiently large ρ is evidence of negative correlation between consecutive residuals. If neither is the case, there is insufficient evidence to disprove the null hypothesis.

Our final step was to compute means and standard errors using the bootstrapping methods. We will omit the details here, but bootstrapping consists of generating a sample of "similar" datasets based on the actual datasets. We performed least-squares fits to the best-suited model on the samples of similar datasets and found means and standard errors from these fits. To visualize the final results, we plotted clouds of best-fit curves given by the bootstrapped fits.

Results

Applying the AIC to the dataset for each state, we found that Model 2 (equation 3) is the most suitable for each of the 51 datasets, and that Model 3 (equation 4) is the second most suitable in terms of average AIC score across the 51 datasets. In addition to the clouds of best-fit curves, we also obtained point estimates for each state, given in the supplementary materials. Below are best-fit clouds for two states, one with a very tight cloud and one with a very widespread cloud.

Figure 2: Best-fit clouds for each student classification in Illinois and Indiana.

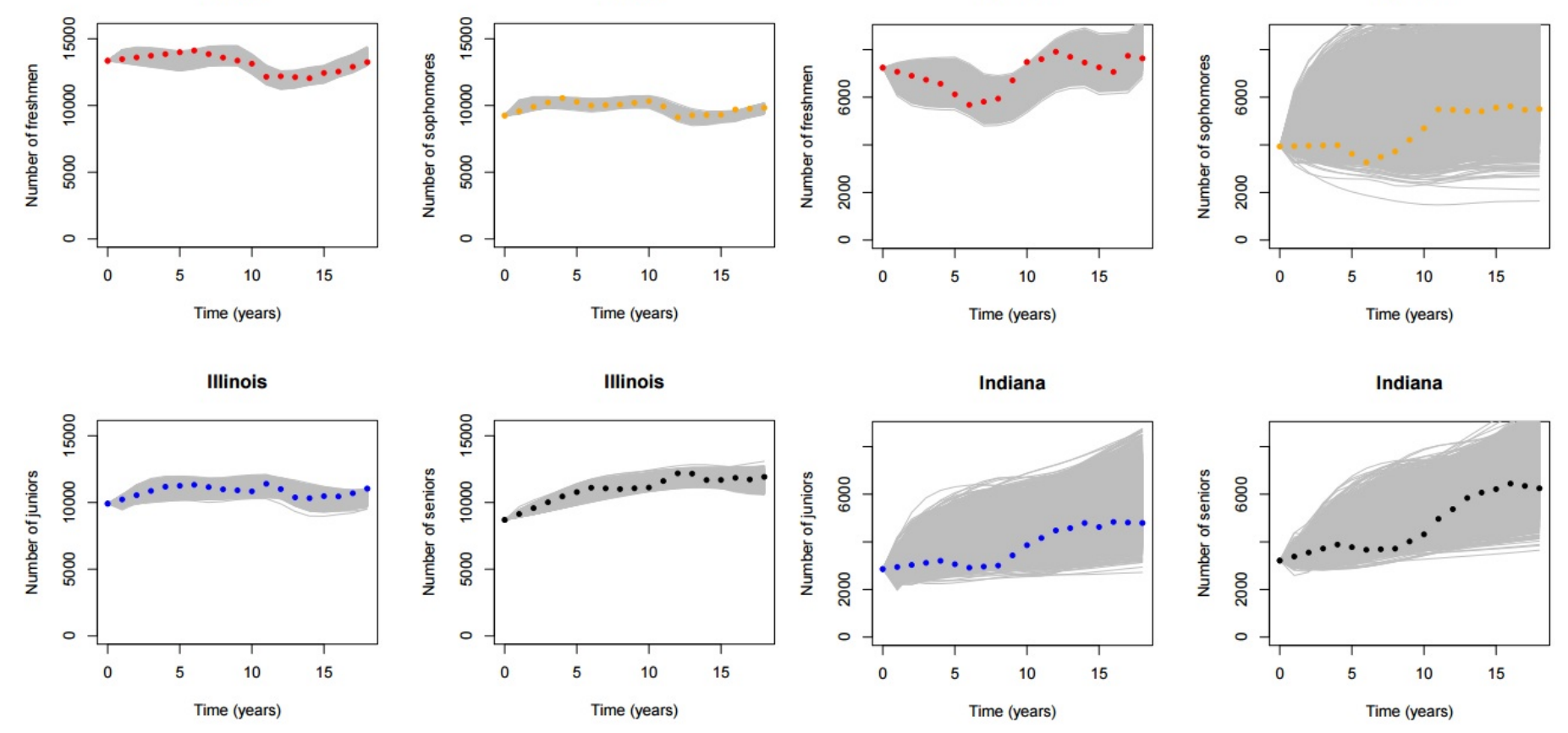
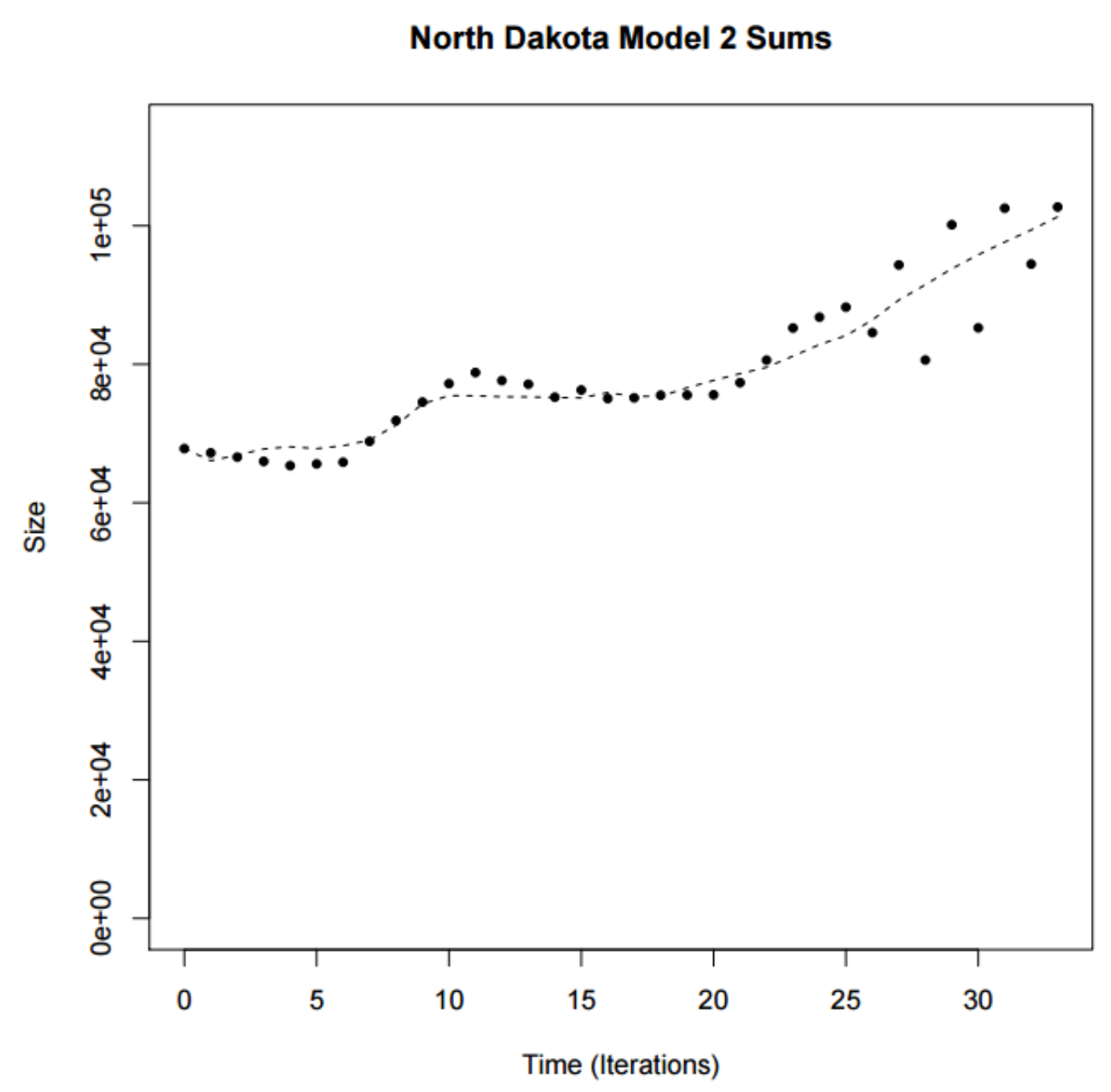


Figure 3: Best-fit curve for total students in North Dakota



Conclusions

In most cases, the clouds of best-fit curves are reasonably tight. Also, as seen in the above example, the total number of students matches up very well with the model's totals (this holds true in general, not just for North Dakota). However, looking at the actual values, many parameter estimates almost certainly do not reflect reality - for example, the results estimate that 53% of Georgia students repeat their first year. It seems likely that we either made an incorrect assumption or investigated no suitable models.

One possible issue in our investigation is the way in which we applied the AIC. Models 2 and 3 make use of freshman enrollment data while Models 1 and 4 do not, so Models 2 and 3 have an inherent advantage in the AIC comparison. After all, Models 1 and 4 spend eight variables predicting values that Models 2 and 3 get for free. Models 2 and 3 do in fact have the best average AIC scores. The fact that our model requires this data also means that we cannot forecast overall enrollment without first forecasting freshman enrollment. This disadvantage of Models 2 and 3 is not represented in their AIC scores.

Acknowledgements

Funding provided by the National Science Foundation, award DMS-1263009. This research was conducted at East Tennessee State University as part of a Research Experience for Undergraduates in June 1 - August 1, 2015. Professor Ariel Cintron-Arias of East Tennessee State University was the faculty supervisor.

Contact Information

- Email: djspence@live.unc.edu
- Phone: (704) 451 3564