Growth Modeling with LMS Data

Data Preparation, Plotting, and Screening

Jemma Bae Kwon
Michigan Virtual Learning Research Institute

September 2017

MICHIGAN VIRTUAL LEARNING®
RESEARCH INSTITUTE

About Michigan Virtual Learning Research Institute

In 2012, the Governor and Michigan Legislature passed legislation requiring *Michigan Virtual*™, formally *Michigan Virtual University*®, to establish a research center for online learning and innovation. Known as *Michigan Virtual Learning Research Institute*® (*MVLRI*®), this center is a natural extension of the work of *Michigan Virtual*. Established in 1998, *Michigan Virtual's* mission is to advance K-12 digital learning and teaching through research, practice, and partnerships. Toward that end, the core strategies of *MVLRI* are:

- Research — Expand the K-12 online and blended learning knowledge base through high-quality, high impact research;
- Policy — Inform local, state, and national public education policy strategies that reinforce and support online and blended learning opportunities for the K-12 community;
- Innovation — Experiment with new technologies and online learning models to foster expanded learning opportunities for K-12 students; and
- Networks — Develop human and web-based applications and infrastructures for sharing information and implementing K-12 online and blended learning best practices.

Michigan Virtual dedicates a small number of staff members to *MVLRI* projects as well as augments its capacity through a fellows program drawing from state and national experts in K-12 online learning from K-12 schooling, higher education, and private industry. These experts work alongside *Michigan Virtual* staff to provide research, evaluation, and development expertise and support.

Suggested Citation: Kwon, J. B. (2017). *Growth modeling with learning management system data: Data preparation, plotting, and screening*. Lansing, MI: Michigan Virtual University. Retrieved from https://mvlri.org/research/publications/growth-modeling-with-lms-data/

Introduction

Several years ago, Ferdig and Cavanaugh (2011) lamented K-12 online and blended programs' state of data collection and analysis to inform and enhance their practices. Since then, the field has improved through increased research and evaluation efforts using various statistical modeling techniques with educational data that were stored in database systems. Those efforts provided stakeholders with opportunities to take advantage of information and knowledge through extracting, analyzing, and interpreting data to better track students' learning activities, understand learners in online courses, and identify their needs.

Taking the example of the *Michigan Virtual Learning Research Institute*® (*MVLRI*®)'s quantitative research project, Lin and colleagues (under review) used an entire school year's data from a learning management system (LMS) and explored the association between class size and students' final grades through fractional polynomial analysis with multilevel modeling. To our best knowledge, it was the first study to answer such questions as "what is the optimal class size?" in the K-12 online context. DeBruler and Bae (2016) examined the relationship of locale codes with subject areas and students' gender, race, and course completion status using descriptive analysis and logistic regression. The study enhanced our understanding about learners in virtual courses by depicting how virtual schools are serving students in each locale.

More recently, Kwon (2017b) used cross-classified, multilevel modeling to model data at the enrollment-, student-, and instructor-level simultaneously while addressing the unique data structure — course enrollments nested in students and teachers — separately. The study provided empirical evidence of poor performance of those students who took the virtual course in order to recover credits.

In placing more fine-grained variables at its analytic center, Lowes and Lin (2017) focused timestamped data and records from the grade-book for Algebra1A courses. For the purpose of investigating how students paced themselves throughout the semester, sequential pattern mining technique, cluster analysis, and survival analyses were used. Kwon (2017a, 2017c) also explored more fine-grained variables from time-stamp and grade-book data by the use of time series cluster analysis with hierarchical clustering as well as with flat partition clustering, in an attempt to capture meaningful learning profiles. These studies provided insights into student course behavior.

Although previous individual studies have had limitations and posed challenges when LMS data was used for research, the methods and findings of those studies contribute to the evolution of this field of study. In this context, we seek to apply other analytic approaches to the LMS data — specifically, growth modeling. Growth modeling was selected because the timestamped data are readily re-structured in the way of longitudinal, repeated measures. Above all, growth modeling can be viewed as an answer to the call — a rigorous methodology that enables us to address research questions on intra-individual changes as well as inter-individual differences within the K-12 online learning context. In this report, we describe key preliminary steps prior to fitting growth models.

Growth Modeling

Growth modeling is one of the analytic approaches in panel study that is defined as research that observes the same individuals at different points in time. By collecting information over time, the research could depict developmental changes or patterns of behavioral changes. Timestamped data in an LMS can be transformed easily to the longitudinal data that enable us to investigate changes in student behaviors in the course over time. In doing so, the first step is to define the timing metric variable, for example, measurement occasion on a weekly, monthly, or quarterly basis. After extracting and structuring the longitudinal data, practical preliminary steps to growth modeling, including data plotting and data screening, are extremely important. Those processes are presented in the following sections.

Data Preparation

For the study focus, mathematics courses in the 2015-16 academic year were chosen. The courses included Algebra 1, Algebra 1 (A), Algebra 1 (B), Algebra 2 (A), Algebra 2 (B), AP Calculus AB (A), AP Calculus AB (B), AP Calculus BC (A), AP Calculus BC (B), AP Statistics (A), AP Statistics (B), Calculus (A), Calculus (B), Geometry (A), Geometry (B), Math Tracks, Mathematics 6 (A), Mathematics 6 (B), Mathematics 7 (A), Mathematics 7 (B), Mathematics 8 (A), Mathematics 8 (B), Mathematics of Baseball, Personal Finance (A), Personal Finance (B), Pre-Algebra (A), Pre-Algebra (B), Pre-Calculus (A), Pre-Calculus (B), Probability and Statistics (A), Probability and Statistics (B), and Trigonometry. From the study, we excluded courses that were specified as at the elementary/middle school-level or whose gradebook data were not stored in the database. The study included gradebook data from 2,849 enrollments in the three semesters (fall, spring, and summer) as well as trimesters.

The dependent variable was student earned scores on a time series basis. As each course had a different grading system, we transformed those scores into the ratio of earned scores until a particular time to the possible course score. The database structured the raw data in the long format, where each row contained data for each timestamp in seconds. To reshape those data, we first defined the timing metric variable to be studied.

We decided to focus on the monthly greatest earned score. Using the LEFT function in Microsoft Excel, the cells for timestamps were cleaned up by removing entries for day, hour, minute, and second from each timestamp cell. This step created data containing multiple entries of the dependent variable at the same measurement occasion (i.e., month). Using the GSORT function and commands to deal with duplicates in Stata, we restructured the data set to contain the greatest earned score per month; thus, the final data contained individual students' greatest earned score, month by month for various mathematics courses.

Data Plotting

In order to take a close look at the data, we first produced visualizations. This process helped us glean information about potential models, possible time metrics, and outliers. Using the PROFILEPLOT command in Stata, the trajectory of proportions of earned scores out of the possible course points against time were plotted, as shown in Figures 1 through 3. The vertical axis represents the dependent variable, and the horizontal axis displays the measurement occasions.

Figure 1. Longitudinal Plot of Earned Scores for Regular Semesters (5 measurement occasions)

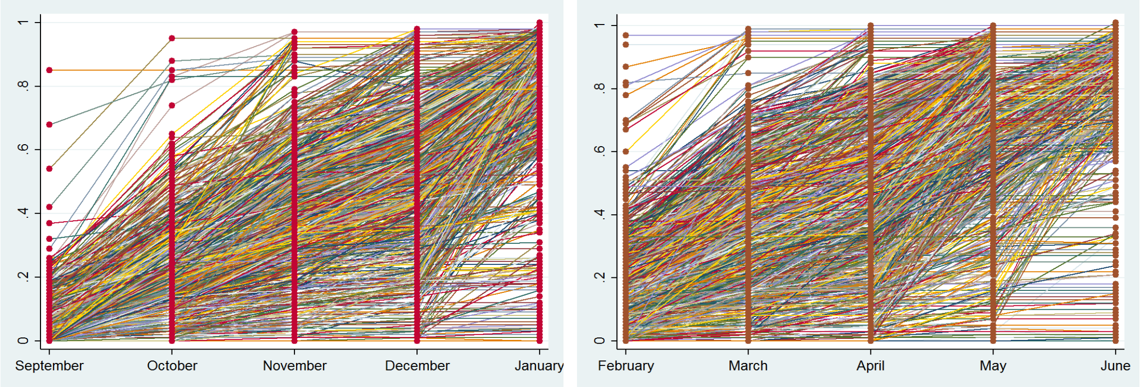


Figure 2. Longitudinal Plot of Earned Scores for Summer Semester (3 measurement occasions)

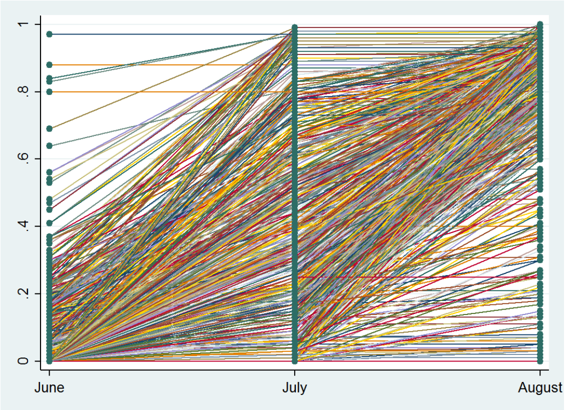
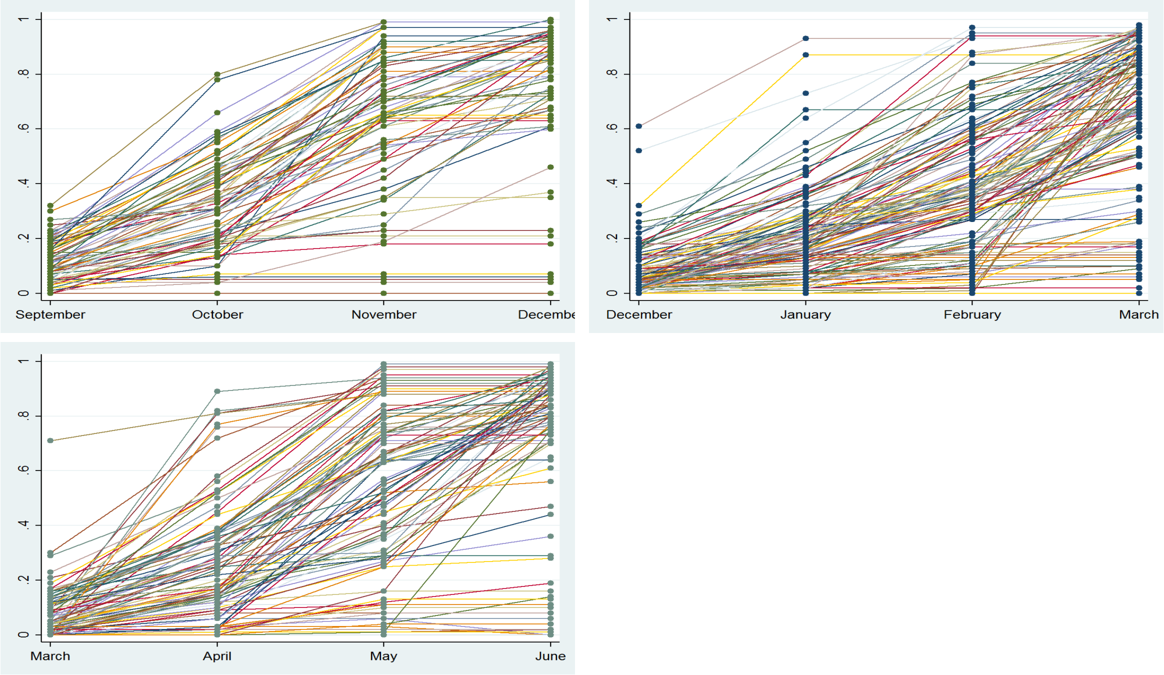


Figure 3. Longitudinal Plot of Earned Scores for Trimesters (4 measurement occasions)



These plots are often called “spaghetti” plots and may not be readily interpretable in terms of overall trends over time. From the graphical exploration of data, we could, however, gain an insight into data and potential analytic approaches as follows.

- Outliers at the bottom region of the plot, from the first time-occasion to the last occasion, need to be observed more closely. If these cases are related to the withdrawn enrollment record in the midst of academic term, they should be removed from the study sample.
- We found various patterns of trajectories, including the linear growth as well as nonlinearity. Furthermore, various patterns in trajectories made it possible to infer a subset of students whose growth patterns are more similar to each other than those in others. As such, growth mixture modeling (GMM) should be considered as an analytic approach. GMM will be discussed in the last section of the report.
- We would combine three trimester data sets for the final modeling with the time metrics of the order of months to obtain sufficient sample size for the trimester data in future research. Given that overall trends are different among the three data sets, results from GMM need to be re-examined by the three different terms, which are Trimester 1 from September to December, Trimester 2 from December to March, and Trimester 3 from March to June.

Data Screening

To obtain fundamental information about the data, descriptive statistics on the chosen time metric were explored. Table 1 presents the result.

Table 1. Data Summary and Univariate Descriptive Statistics of Dependent Variable

	n	Pass %	1st Month %		2nd Month %		3rd Month %		4th Month %		5th Month %	
			M	SD	M	SD	M	SD	M	SD	M	SD
Fall	935	85	8	7	23	14	40	21	52	25	76	25
Spring	893	87	17	14	33	22	45	26	67	27	77	25
Summer	700	81	7	13	35	29	73	28	--	--	--	--
Trimester	321	70	--	--	--	--	--	--	--	--	--	--
Trimester1	--	--	11	8	31	18	61	28	71	29	--	--
Trimester2	--	--	8	9	19	17	36	26	60	32		
Trimester3	--	--	6	9	24	22	49	32	64	35	--	--
Total	**2,849**	**83**	--	--	--	--	--	--	--	--	--	--

During the fall semester, students appear to earn 8% of possible course points in September and continue upturns by earning 23% in October, 40% in November, 52% in December, and 76% in January, on average. Among those in the fall sample, students who withdrew from the course showed a substantial gap in growth patterns with averages of 3%, 7%, 9%, 9%, and 12% and standard deviations of 5%, 10%, 11%, 11%, and 19%, at the five time segments.

The growth pattern of students in the spring semester shows an upward trajectory with earned-scores of 17% of course points in February, 33% in March, 45% in April, 67% in May, 77% in June, with respective standard deviations of 14%, 22%, 26%, 27%, and 25%. The withdrawn cases indicate a discrepancy in growth patterns with averages 9%, 10%, 10%, 13%, and 13% and with standard deviations of 12%, 13%, 14%, 17%, and 17% at the respective time segments.

While having a slightly low passing rate in comparison to regular semesters, the summer semester shows the trend connecting points at 7% of course points in June, 35% in July, and 73% in August with respective standard deviations of 13%, 29%, and 28%. A discrepancy between the semester data and the subset of withdrawn cases in univariate descriptive statistics was also found (Mean: 3%, 5%, and 7%; SD: 5%, 6%, and 9%).

When it came to the three trimester terms, the first term's trajectory had connecting points at 11% of course points in September, 31% in October, 61% in November, and 71% in December. The second term indicates that an overall growth slow down with the averages of 8% of course points in December, 19% in January, 36% in February, and 60% in March. The last trimester also showed a relatively slower upward trend in comparison to Trimester 1 with averages of 6% of course points in March, 24% in April, 49% in May, and 64% in June. From their univariate descriptive statistics, we found withdrawn cases to be outliers (Trimester 1: M=3%; SD= 4% for all four time segments / Trimester 2: M=5%, 10%, 12%, 15%; SD=5%, 5%, 6%, 12% / Trimester 3: M= 1%, 3%, 3%, 3%; SD=2%, 2%, 3%, 3%).

Discussion

From data plotting and screening, we found that the majority of students showed an upward trajectory in the middle region of the plot. Some outliers on the bottom area (i.e., those that remained unchanged in cumulative earned-scores throughout the semester/trimester) could be students who withdrew from the course. There were also a few extremes on the upper region of the plot, which featured a sharp growth in the early part of the term that then remained unchanged until the semester ended (i.e., exceeded the course passing mark earlier than the official course schedule). Overall, high standard deviations indicate that data points were spread out, and also growth patterns on plots showed a great deal of variation. Accordingly, it is worth considering removal of withdrawn cases from the study sample when fitting growth models and that modeling should address the potential heterogeneity of growth trajectories under the GMM framework.

Conventional growth modeling assumes individuals from a homogeneous population, their growth equivalently affected by covariates, and a single average trajectory to describe an entire population. However, heterogeneity of growth trajectories can often be speculated (Grimm, Ram, & Estabrook, 2017). To take an example in the context of the current study, patterns of course engagement as well as content mastery over time are often different between students who took the virtual course because their local schools did not offer particular AP courses and those who did so to recover the credit that they failed to obtain from taking the course previously. GMM is considered a robust approach against this assumption on the homogeneous population. Specifically, latent class growth analysis that is a special type of GMM should address this matter by identifying distinct clusters prior to performing GMM (Jung & Wickrama, 2008). As the next step, we will conduct latent class growth analysis to better understand how student course behaviors unfold over time in virtual mathematics courses.

References

DeBruler, K. & Bae, J. (2016). *Educating students across locales: Understanding enrollment and performance across virtual schools*. Lansing, MI: Michigan Virtual University. Retrieved from http://media.mivu.org/institute/pdf/locale.pdf

Grimm, K. J., Ram, N., & Estabrook, R. (2017). *Growth Modeling: Structural Equation and Multilevel Modeling Approaches*. Guilford Press. NY.

Jung, T. & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, 2, 302-317.

Kwon, J. B. (2017 a). *Exploring patterns of time investment in courses using time-series clustering analysis of time spent in the course.* Lansing, MI: Michigan Virtual University. Retrieved from https://mvlri.org/research/publications/exploring-patterns-of-time-investment-in-courses/

Kwon, J. B. (2017 b). Examining *credit recovery learning profile from time-series clustering analysis of attempted scores*. Lansing, MI: Michigan Virtual University. Retrieved from http:// http://media.mivu.org/institute/pdf/creditrec2.pdf

Kwon, J. B. (2017 c). *Examining credit recovery experience at a State Virtual School*. Lansing, MI: Michigan Virtual University. Retrieved from http://media.mivu.org/institute/pdf/creditrec.pdf

Lin, C.-H., & Bae, J. (under review). *The effect of class size in online K-12 courses*. Manuscript submitted to a peer-reviewed journal.

Lowes. S. & Lin, P. (2017). *Student pathways through online algebra 1 courses.* Lansing, MI: Michigan Virtual University. Retrieved from http://media.mivu.org/institute/pdf/algebrapath.pdf

3101 TECHNOLOGY BLVD., SUITE G • LANSING, MI 48910 • MICHIGANVIRTUAL.ORG • 888.532.5806

MICHIGAN™ VIRTUAL

MICHIGAN VIRTUAL LEARNING® RESEARCH INSTITUTE

MICHIGANVIRTUAL.ORG

MVLRI.ORG