Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 114 (2017) 250–257

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS
October 30 – November 1, 2017, Chicago, Illinois, USA

Learning Curve Analysis Using Intensive Longitudinal and Cluster-Correlated Data

Xiao Zhong[a], Zeyi Sun[a*], Haoyi Xiong[a], Neil Heffernan[b], Md. Monirul Islam[a]

[a]Missouri University of Science and Technology, Rolla, MO 65409, USA
[b]Worcester Polytechnic Institute, Worcester, MA 01609, USA

Abstract

Intensive longitudinal and cluster-correlated data (ILCCD) can be generated in any situation where numerical or categorical characteristics of multiple individuals or study units are observed and measured at tens, hundreds, or thousands of occasions. The spacing of measurements in time for each individual can be regular or irregular, fixed or random, and the number of characteristics measured at each occasion may be few or many. Such data can also arise in situations involving continuous-time measurements of recurrent events. Generalized linear models (GLMs) are usually considered for the analysis of correlated non-normal data, while multivariate analysis of variance (MANOVA) is another option. In the paper, both GLMs and MANOVA are applied to ASSISTments online teaching system via the Wald test in order to estimate and predict the learning effects of the students after taking an Algebra course for three months in the system. Three case studies based on these two methodologies are presented in a mathematical and statistical manner.

Keywords: longitudinal data analysis; GLMs; MANOVA; Wald test; learning curve analysis; case study

1. Introduction

Online adaptive learning systems are increasingly being adopted in schools and families as supplements to traditional teaching or tutoring instructions and instruments. Among these systems, ASSISTments

* Corresponding author. Tel: +1-573-341-7745; fax: +1-573-341-6567.
Email address: sunze@mst.edu

10.1016/j.procs.2017.09.035

(www.assistments.org) is an online tutoring platform that provides "formative assessments that assist" [1]. For instance, on this platform, teachers may choose or add homework items and students can complete these items online. As students work on each item, they may receive real-time feedback on the correctness of their answers, or hints to improve their answers, or decomposition of multistep problems according to their own requests or automatic decisions made by the system. In addition, ASSISTments provides teachers with the reports on their students' homework performance so that the teachers can adjust their teaching or tutoring strategies, such as arranging more targeted homework reviews, assigning extra practice problems for certain topics, and so on. Prior researches have already established the promise of ASSISTments for improving the student performance in middle school mathematics through the homework support [1,2,3].

Furthermore, during the entire learning or tutoring period, say, in three months, students may be given a quiz or an exam regularly or irregularly by ASSISTments. For all the students taking the same tutoring course or program in the system, their performances are evaluated and recorded as the percentages of correctly answering the questions in each quiz or exam. These percentages actually consist of a group of intensive longitudinal and cluster-correlated data (ILCCD), because in general each student may correspond to more than one (learning) percentages. By analysing such ILCCD, researchers may discover some hidden patterns, learning trends, or new knowledge regarding online teaching, educational statistics, or behavioural psychology [4].

The significance of ILCCD analysis has been recognized under many situations where numerical or categorical characteristics of multiple individuals or study units are observed and measured at tens, hundreds, or thousands of occasions. Such data analysing procedures often have to consider the strong correlation among individuals, which is the major difficulty encountered in longitudinal data analysis. Many researches have been focused on the decomposition of the correlation or covariance matrix of the input data package. Generalized linear models (GLMs) and multivariate analysis of variance (MANOVA) are listed in the series. Moreover, as a generalization of statistical linear models, both methods gained a very important role in estimation and prediction for ILCCD with their flexibility and reliability when dealing with various types of input and output multivariate variables [5,6].

In this paper, we apply both GLMs and MANOVA via the Wald test to ASSISTments online teaching system to estimate and predict the learning effects of different students after taking an Algebra course for three months in the system. The analyzing procedures and results of three cases are presented with details. The remainder of the paper is organized as follows. Section 2 describes the natural data obtained from ASSISTments online teaching system. Section 3 introduces the statistical models briefly. Section 4 includes three case studies for balanced design and complete data based on these models. The conclusion and future work is given in the last section.

2. Data description

Totally, there are 669 students involved in this ASSISTments project, among which, 621, 556 and 464 students were observed in each of three continuous months, respectively. Also, 279 students were observed in all three months. Therefore, the design of longitudinal study for all of the students is either balanced but not complete with *N* of 669 and *n* of 3, or unbalanced with *N* of 669 and *n* that is less than or equal to 3; while the design over the students who were observed in all three months is balanced and complete with *N* of 279 and *n* of 3. Five columns are included in the original data set: Student ID, Gender, Month1, Month2 and Month3. The measurements for Month1, Month2 and Month3 are the average percentages of each student correctly answering the exam questions in three continuous months, respectively. A sample of the raw data is given in Table 2.1.

Table 2.1. A Sample of the Raw Data

Student ID	Gender	Month1	Month2	Month3
136	f	29	56	32
137	f	38	58	64
139	f	48	29	57
140	f	18	52	69
141	f	25	0	
143	f	25	42	30

3. The statistical models and implementation

Let *n* be the number of repeated measures and *N* the number of subjects, the analysis of response profiles can be implemented in the general linear model for appropriate choices of X_i :

$$\underset{n\times 1}{Y_i} \mid \underset{n\times n}{X_i} = I \square \underset{n\times 1}{\beta} + \underset{n\times 1}{\varepsilon_i}\,, \qquad \varepsilon_i \square N(0,\Sigma_i),\quad i=1,2,...,N.$$

Or,

$$E(Y_i \mid X_i) = X_i\beta; \qquad i=1,2,...,N. \tag{1}$$

Especially, for balanced design and complete data, we assume that the covariance matrix of outcomes is unstructured and $\Sigma_i = \Sigma$; otherwise, we assume that the covariance matrix of outcomes has compound symmetry structure. The maximum likelihood estimations (MLEs) of β and Σ are given in Expressions (2) and (3), respectively. The restricted maximum likelihood (REML) estimations of β and Σ are given in Expressions (4) and (5), respectively.

Set $\hat{\Sigma}_{i_0} = I$, and $k \geq 0$, then

$$\hat{\beta}_{MLE_k} = (\sum_{i=1}^{N} X_i^T \hat{\Sigma}_{MLE_{i_k}}^{-1} X_i)^{-1} \sum_{i=1}^{N} (X_i^T \hat{\Sigma}_{MLE_{i_k}}^{-1} Y_i) \tag{2}$$

$$\hat{\Sigma}_{MLE_{i_{k+1}}} = \sum_{i=1}^{N} (Y_i - X_i\hat{\beta}_{MLE_k})(Y_i - X_i\hat{\beta}_{MLE_k})^T / N \tag{3}$$

$$\hat{\beta}_{REML_k} = (\sum_{i=1}^{N} X_i^T \hat{\Sigma}_{REML_{i_k}}^{-1} X_i)^{-1} \sum_{i=1}^{N} (X_i^T \hat{\Sigma}_{REML_{i_k}}^{-1} Y_i) \tag{4}$$

$$\hat{\Sigma}_{REML_{i_{k+1}}} = \sum_{i=1}^{N} [(Y_i - X_i\hat{\beta}_{REML_k})(Y_i - X_i\hat{\beta}_{REML_k})^T + X_i(\sum_{i=1}^{N} X_i^T \hat{\Sigma}_{REML_{i_k}}^{-1} X_i)^{-1} X_i^T] / N\,. \tag{5}$$

The sampling distribution of $\hat{\beta}_{MLE}$ or $\hat{\beta}_{REML}$ is asymptotically normal:

$$\hat{\beta}_{MLE} \overset{asymptotically}{\square} N(\beta, \hat{Cov}(\hat{\beta}_{MLE}))\,, \quad \text{where} \quad \hat{Cov}(\hat{\beta}_{MLE}) = (\sum_{i=1}^{N} X_i^T \hat{\Sigma}_{MLE_i}^{-1} X_i)^{-1}\,; \tag{6}$$

$$\hat{\beta}_{REML} \overset{asymptotically}{\square} N(\beta, \hat{Cov}(\hat{\beta}_{REML}))\,, \quad \text{where} \quad \hat{Cov}(\hat{\beta}_{REML}) = (\sum_{i=1}^{N} X_i^T \hat{\Sigma}_{REML_i}^{-1} X_i)^{-1}\,. \tag{7}$$

Let *L* denote a vector or matrix of known weights. We may consider a hypothesis test with the null and alternative hypotheses defined as

$$\begin{aligned} H_0 &: L\beta = 0 \\ H_a &: L\beta \neq 0 \end{aligned} \tag{8}$$

Then, an approximate 95% confidence interval for $L\beta$ can be obtained by the following formula if $L\beta$ is a one-dimension vector:

$$L\hat{\beta} \pm 1.96\sqrt{L\hat{Cov}(\hat{\beta})L^T}\,. \tag{9}$$

We can also perform a Wald test or likelihood ratio test after defining the constant vector L corresponding to our scientific interest. Specifically, to draw a statistically meaningful conclusion, the Wald test statistic can be calculated via the following formula and then compared to a χ^2 distribution with degrees of freedom equal to the number of rows of L:

$$W^2 = (L\hat{\beta})^T \{LC\hat{o}v(\hat{\beta})L^T\}^{-1}(L\hat{\beta}). \tag{10}$$

As of the likelihood ratio test, the maximized log-likelihood for constrained model with null hypothesis and unconstrained model with alternative hypothesis are required to make a comparison to draw the conclusion [5].

4. Case studies for balanced design and complete data

In this section, by performing three different but correlated hypothesis testing procedures based on the models explained in Section 3 over the 279 students who were observed and measured in all three months when taking the Algebra course on ASSISTments, we aim to answer three types of scientific questions relevant to the learning performances of the students. These questions are arranged in a hierarchical order– are all the 279 students performing similarly in their three-month learning period? If not, are the female and male students among the 279 students performing significantly differently in these three months? If not, are the learning performances of these two groups following significantly different trends or growth curves?

4.1. Case 1: One sample repeated measures

The 279 students' performances are measured repeatedly in three continuous months. The goal is to quantify the differences of their performances in these three months.

Let $Y_i = (Y_{i1}, Y_{i2}, Y_{i3})^T$, $\beta = (\beta_0, \beta_1, \beta_2)^T$,

$$X_i = \begin{pmatrix} 100 \\ 110 \\ 101 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 \sigma_{12} \sigma_{13} \\ \sigma_{21} \sigma_2^2 \sigma_{23} \\ \sigma_{31} \sigma_{32} \sigma_3^2 \end{pmatrix}, \quad N = 279. \tag{11}$$

Then β_0, $\beta_0 + \beta_1$ and $\beta_0 + \beta_2$ represent the mean performances of all students in the 1st month, 2nd month and 3rd month, respectively. To reach the goal set above, we perform the following steps with our own code in R:

a) Using the formulas given in Section 3, we obtain the MLEs and REML estimates of β and Σ as

$$\hat{\beta}_{MLE} = \hat{\beta}_{REML} = \begin{pmatrix} 0.36 \\ 0.0790681 \\ 0.1383513 \end{pmatrix}$$

$$\hat{\Sigma}_{MLE} = \begin{pmatrix} 0.051375 & 0.019766 & 0.023657 \\ 0.019766 & 0.052107 & 0.020452 \\ 0.023657 & 0.020452 & 0.058248 \end{pmatrix} \qquad \hat{\Sigma}_{REML} = \begin{pmatrix} 0.05156 & 0.019837 & 0.023742 \\ 0.019837 & 0.052294 & 0.020525 \\ 0.023742 & 0.020525 & 0.058458 \end{pmatrix}$$

$$C\hat{o}v\hat{\beta} = \begin{pmatrix} 1.85E-04 & -0.00011 & -9.97E-05 \\ -1.14E-04 & 0.00023 & 1.02E-04 \\ -9.97E-05 & 0.000102 & 2.24E-04 \end{pmatrix}$$

b) Let $L = \begin{pmatrix} 010 \\ 001 \end{pmatrix}$, we calculate the Wald test statistic and p-value:

$$W^2 = (L\hat{\beta})^T \{LC\hat{o}v(\hat{\beta})L^T\}^{-1}(L\hat{\beta}) = 86.79573 > \chi^2_{2,0.025} = 7.38$$

$$p - value = 0$$

Since p-value is 0, we reject the null hypothesis at the significance level of 0.05 and conclude that the mean performances or the mean scores of 279 students are different in the three-month learning period at the significance level of 0.05.

4.2. Case 2: One-way multivariate ANOVA (MANOVA)

The performances of 279 students (167 male students and 112 female students) are measured repeatedly in three months. The goal is to test if the mean performances of male students and female students are the same in these three months. That is, in this case, we have $G = 2$, $N_g = (N_{male}, N_{female}) = (167,112)$, $n = 3$. Also, with two groups measured at three months, there are $2 \times 3 = 6$ mean parameters to consider.

For the first group (female students), let the design matrix

$$X_i = \begin{pmatrix} 100000 \\ 101000 \\ 100100 \end{pmatrix}. \tag{12}$$

For the second group (male students), let the design matrix

$$X_i = \begin{pmatrix} 110000 \\ 111010 \\ 110101 \end{pmatrix}. \tag{13}$$

And $$\Sigma = \begin{pmatrix} \sigma_1^2 \sigma_{12} \sigma_{13} \\ \sigma_{21} \sigma_2^2 \sigma_{23} \\ \sigma_{31} \sigma_{32} \sigma_3^2 \end{pmatrix}, Y_i = (Y_{i1}, Y_{i2}, Y_{i3})^T, \beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^T, N = 279. \tag{14}$$

Then β_0, $\beta_0 + \beta_2$ and $\beta_0 + \beta_3$ represent the mean performances of all female students in the 1st month, 2nd month and 3rd month, respectively; $\beta_0 + \beta_1, \beta_0 + \beta_1 + \beta_2 + \beta_4$ and $\beta_0 + \beta_1 + \beta_3 + \beta_5$ represent the mean performances of all male students in the 1st month, 2nd month and 3rd month, respectively. Following the process similar to Case 1, we can test whether the mean performances of male students and female students are different or not in these three continuous months:

a) From the formulas in Section 3, we obtain the MLEs and REML estimates of β and Σ as

$$\hat{\beta}_{MLE} = \hat{\beta}_{REML} = \begin{pmatrix} 0.38446429 \\ -0.04087147 \\ 0.0625 \\ 0.10928571 \\ 0.02767964 \\ 0.0485586 \end{pmatrix}$$

$$\hat{\Sigma}_{MLE} = \begin{pmatrix} 0.050974 & 0.01963 & 0.023732 \\ 0.01963 & 0.052065 & 0.020476 \\ 0.023732 & 0.020476 & 0.058234 \end{pmatrix} \qquad \hat{\Sigma}_{REML} = \begin{pmatrix} 0.051342 & 0.019778 & 0.023904 \\ 0.019778 & 0.052441 & 0.020624 \\ 0.023904 & 0.020624 & 0.058654 \end{pmatrix}$$

$$C\hat{o}v\hat{\beta} = \begin{pmatrix} 0.000458 & -0.00046 & -0.00028 & -0.00024 & 0.000282 & 0.000245 \\ -0.00046 & 0.000766 & 0.000282 & 0.000245 & -0.00047 & -0.00041 \\ -0.00028 & 0.000282 & 0.000573 & 0.000253 & -0.00057 & -0.00025 \\ -0.00024 & 0.000245 & 0.000253 & 0.000555 & -0.00025 & -0.00056 \\ 0.000282 & -0.00047 & -0.00057 & -0.00025 & 0.000958 & 0.000422 \\ 0.000245 & -0.00041 & -0.00025 & -0.00056 & 0.000422 & 0.000928 \end{pmatrix}$$

b) Let $L = \begin{pmatrix} 010000 \\ 000010 \\ 000001 \end{pmatrix}$, *the Wald test statistic and p-value are calculated as*

$$W^2 = (L\hat{\beta})^T \{LC\hat{o}v(\hat{\beta})L^T\}^{-1}(L\hat{\beta}) = 3.211857 < 9.35 = \chi^2_{3,0.025}$$

$$p - value = 0.3974529$$

Since the p-value is greater than 0.05, we cannot reject the null hypothesis at the significance level of 0.05. Therefore, we conclude that the mean performances of male and female students are not different in three-month tutoring period at the significance level of 0.05.

4.3. Case 3: Growth curve with covariates

Same as Case 2, the performances of 279 students from two groups (male students/female students) are observed at the same occasions of the 1st month, 2nd month and 3rd month. In this case, the goal is to see if the mean responses of the student performance in each of the two groups have the same growth curves or if the mean responses of the student performance in each group have the same trend over time t_j, $j = 1,2,3$.

Suppose that each student follows his or her own growth curve indexed by a $q \times 1$ parameter vector β_i, which is a deterministic function of a $q \times p$ time-invariant covariate matrix A_i, that is, $\beta_i = A_i\beta$; here, β is a $p \times 1$ unknown parameter vector. Thus, the model used for this case can be defined as

$$\underset{n\times 1}{Y_i} = \underset{n\times q}{Z}\ \underset{q\times p}{A_i}\ \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon_i} = \underset{n\times p}{X_i}\ \underset{p\times 1}{\beta} + \varepsilon_i, \qquad \varepsilon_i \,\square\, N(0,\Sigma),\ i = 1,2,...,N. \tag{15}$$

We also assume that the covariance matrix of outcomes $\Sigma_i = \Sigma$ and unstructured. For this specific problem, we have

$$Y_i = (Y_{i1}, Y_{i2}, Y_{i3})^T, \beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^T, N = 279, n = 3, \tag{16}$$

the design matrix for the first group (female students) is

$$X_i = \begin{pmatrix} 10t_1t_1^2 00 \\ 10t_2t_2^2 00 \\ 10t_3t_3^2 00 \end{pmatrix} = \begin{pmatrix} 101100 \\ 102400 \\ 103900 \end{pmatrix}, \tag{17}$$

and the design matrix for the second group (male students) is

$$X_i = \begin{pmatrix} 11t_1t_1^2t_1t_1^2 \\ 11t_2t_2^2t_2t_2^2 \\ 11t_3t_3^2t_3t_3^2 \end{pmatrix} = \begin{pmatrix} 111111 \\ 112424 \\ 113939 \end{pmatrix}. \tag{18}$$

Then, β_0 represents the mean performance of all female students at the starting point; the sign and magnitude of β_2 and β_3 determine whether the mean performance of all female students is increasing or decreasing over time (three months) and how the rate of change depends on time, respectively. Also, $\beta_0 + \beta_1, \beta_2 + \beta_4, \beta_3 + \beta_5$ have the similar meanings for male students. The following process can lead us to draw a conclusion with respect to the goal set above:

a) Calculate the MLEs and REML estimates of β and Σ as

$$\hat{\beta}_{MLE} = \hat{\beta}_{REML} = \begin{pmatrix} 0.30625 \\ -0.07535 \\ 0.086071 \\ -0.00786 \\ 0.037881 \\ -0.0034 \end{pmatrix}$$

$$\hat{\Sigma}_{MLE} = \begin{pmatrix} 0.050974 & 0.019636 & 0.023732 \\ 0.019636 & 0.052065 & 0.020476 \\ 0.023732 & 0.020476 & 0.058234 \end{pmatrix} \qquad \hat{\Sigma}_{REML} = \begin{pmatrix} 0.051342 & 0.019778 & 0.023904 \\ 0.019778 & 0.052441 & 0.020624 \\ 0.023904 & 0.020624 & 0.058654 \end{pmatrix}$$

$$C\hat{o}v\hat{\beta} = \begin{pmatrix} 0.00586 & -0.00586 & -0.00633 & 0.001526 & 0.006328 & -0.00153 \\ -0.00586 & 0.009791 & 0.006328 & -0.00153 & -0.01057 & 0.002549 \\ -0.00633 & 0.006328 & 0.007394 & -0.00183 & -0.00739 & 0.001826 \\ 0.001526 & -0.00153 & -0.00183 & 0.00046 & 0.001826 & -0.00046 \\ 0.006328 & -0.01057 & -0.00739 & 0.001826 & 0.012353 & -0.00305 \\ -0.00153 & 0.002549 & 0.001826 & -0.00046 & -0.00305 & 0.000768 \end{pmatrix}$$

b) Let $L = \begin{pmatrix} 010000 \\ 000010 \\ 000001 \end{pmatrix}$, we obtain the Wald test statistic and the p-value as

$$W^2 = (L\hat{\beta})^T \{LC\hat{o}v(\hat{\beta})L^T\}^{-1}(L\hat{\beta}) = 3.211857 < 9.35 = \chi^2_{3,0.025}$$

$$p - value = 0.3974529$$

Thus, we cannot reject the null hypothesis at the significance level of 0.05, and we conclude that the mean performances of male students and female students among (at least) these 279 students don't have different growth curves during three-month study at the significance level of 0.05.

5. Conclusion and future work

In the paper, three case studies, different but correlated, regarding the ILCCD analysis are explored in a hierarchical order over a specific data set that are automatically collected by ASSISTments online teaching system. Based on the Wald test, both GLMs and MANOVA are implemented in each case with details to get some initial understanding of the outcomes from the online teaching system. The conclusions drawn from the testing results are consistent with the discovery of educational researchers [7]. Usually, the precision of model-based statistical approaches relies on the assumptions made for the input data, which is not quite applicable in certain environments. Therefore, the data-driven and assumption-free techniques, such as support vector machines (SVMs), artificial neural networks (ANNs), deep neural networks (DNNs), and their variants in machine learning and artificial intelligence fields, may be an alternative to achieve higher efficiency in estimation and prediction for certain kind of intensive longitudinal and cluster-correlated data [8,9,10].

For future work, reasonable sampling strategies may be considered in order to grab and analyze the most representative data sets among large amounts of longitudinal data sources available on ASSISTments and other online teaching systems.

References

[1] Feng, M. (2014). Towards Uncovering the Mysterious World of Math Homework. *Proceedings of the 7th International Conference on Educational Data Mining*. EDM 2014. pp 425-426.

[2] Singh, R., Saleem, M., Pradhan, P., Heffernan, C., Heffernan, N., Razzaq, L., & Dailey, M. (2011). Improving K-12 homework with computers. In *Proceedings of the Artificial Intelligence in Education Conference*, Auckland, New Zealand.

[3] Kelly, Y., Heffernan, N., Heffernan, C., Goldman, S., Pellegrino, G. & Soffer, D. (2013). Estimating the effect of web-based homework. In Lane, Yacef, Motow & Pavlik (Eds) *The Artificial Intelligence in Education Conference* (pp. 824-827). Springer-Verlag.

[4] Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics.

[5] Laird, N. (2004). *Analysis of Longitudinal and Cluster-Correlated Data*. NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 8. The Institute of Mathematical Statistics and the American Statistical Association.

[6] Walls, T. A., & Schafer, J. L. (2006). *Models for Intensive Longitudinal Data*. Oxford University Press.

[7] Cooper, H., Robinson, J., & Patall, E. (2006). Does homework improve academic achievement? A synthesis of research, 1987-2003. *Review of Educational Research*, 76: 1-62

[8] Zhong, X. (2000). *Ph.D. Dissertation: Classification and Cluster Mining (unpublished).* Zhejiang University, Hangzhou, China.

[9] Zhong, X. (2004). *Master Thesis: A Study of Several Statistical Classification Methods with Application on Microbial Source Tracking.* WPI, Worcester, MA. Available: http://www.wpi.edu/Pubs/ETD/Available/etd-0430104-155106

[10] Zhong, X., & Enke, D. (2017). Forecasting Daily Stock Market Return Using Dimensionality Reduction. *Expert Systems with Applications*, 67: 126-139.