AN INTRODUCTION TO LONGITUDINAL DATA AND STATISTICAL SOFTWARE TO HANDLE LONGITUDINAL DATA

Dr. Kamal Kishore[*]

Abstract

Many research opportunities which are available with longitudinal data are unavailable in the world of Cross-sectional data. Longitudinal data are motivated by research studies used to assess the change of an outcome variable over a period of time. A very important characteristic of longitudinal data is that subjects are repeatedly measured over a period of time. The repeated nature of observations help the researcher to investigate how the variability of response varies in time with covariates. Moreover, longitudinal studies require fewer subjects to achieve high power as compared to cross-sectional studies. Repeated measurement from an individual is correlated in contrast to cross-sectional studies and is**a** positive aspect of longitudinal data.It is challenging to incorporate correlation into analysis but ignoring this correlation may produce biased results. There are many theoretical and practical issues which needs to be considered before analyzing longitudinal data. The objective of this article is to familiarize applied researchers with intricacies involved while dealing with longitudinal data.Second objective of this article is to give brief description about statistical software used to analyze longitudinal data.

Features of Longitudinal Data

Repeated measurement from subjects is defining feature of longitudinal data. This helps the researcher to investigate how the variability of response varies in time with covariates. Longitudinal data can be collected either from observational study or designed experiments (Fitzmaurice etal., 2004). Thus longitudinal data have some peculiar characteristics such as:

[*] Assistant Professor, Institute of Management, Christ University, Bengaluru

- Repeated measures from same subjects are positively correlated.
- Intra-individual correlation often decreases with increase in time, but rarely approaches value of zero.
- Longitudinal data set is mostly unbalanced.
- Most of the longitudinal data from studies consist of missing data.

Source of Correlation and Challenges in Longitudinal Data

There are potentially three sources of variation that leads to correlation among the repeated measures obtained from an individual. It is important to understand these sources of variability as advanced statistical techniques make different assumptions.

Between-Individual variability

Every individual is unique in the nature- nurture paradigm, so their response will vary within and between situations. In longitudinal studies, some individuals will respond highly, some on an average and some others consistently on a low than other individuals. This causes variation in response of individuals.

Within-Individual variability

The response of an individual on different occasions varies due to inherent biological variation. These biological variations are caused by a number of factors such as circadian rhythms or temperature, light, season, diet or infection etc., or in various combinations.

Measurement error

Variation in outcomes such as height and weight is almost negligible due to well-developed scales, whereas for many other outcomes such as depression, pain etc., the variability due to measurement error can be substantial. Measurement error is an integral part of all studies.

Challenges

Off late due to rapid development in methodology, computational power and integration of advanced techniques in standard software use of longitudinally studies have increased many folds. Moreover, one major reason for increase in longitudinal studies is belief that causality can be established with these types of studies. Although, this is one of the criteria among various criteria need to establish causality. These types of studies pose many challenges to researcher, some of which are given below.

- Repeated observations on same subjects are not independent, so we must account for correlation using sophisticated statistical techniques.
- It is difficult to deal with missing data as missing data patterns are much more sophisticated than cross-sectional data.
- Longitudinal data growth trajectories may be highly complicated and there may be large variation between subjects.
- Data format for application of advanced statistical techniques have to be converted to desired format.
- Different **types** of longitudinal studies poses different **types** of research questions and hence appropriate statistical method must be matched to theresearch question.
- Computationally intensive methods are required to account forheterogeneity among longitudinal measured units.

Characteristics of Longitudinal Data

Not every longitudinal study is amenable to analysis of change (Singer and Willett, 2003). These studies share three methodological features:

- Three or more waves of data.
- An outcome whose value changes systematically over time.
- A sensible metric for clocking time.

The study of change over time can be formulated in simple questions. Does the behavior of an individual change over time? Are there individual differences in intra-individual change? Cross-sectional data tell us nothing about the individual change. Two wave studies are also marginally better as it is linear due to restriction of two time point. It is not possible to comprehend whether the change is immediate, delayed or steady over the time interval. Moreover, it cannot separate measurement error from growth. More complex the trajectory, more the number of waves required. A minimum of three waves are required to measure change. The study should not only have sufficient number of waves of data but also long enough to record the change and transitions (Rogosa et al., 1982; Willett, 1988).

Longitudinal data helps in delineating cohort effect, age effect and period effect, which is not possible with cross-sectional data. The assessment of correlates or predictors of change **aremotivated** by research question such as "what kind of **persons** grow (learn) fastest?" (Raudenbush & Bryk, 2002; Singer & Willett, 2003). Individual time paths are the proper focus for analysis of change (Lenzenweger et al., 2004). Fundamental question in measuring change revolves around individual growth. Recently, due to surge in computational power and advanced methodology, researchers are able to find the answer about individual change(Singer & Willett, 2005).

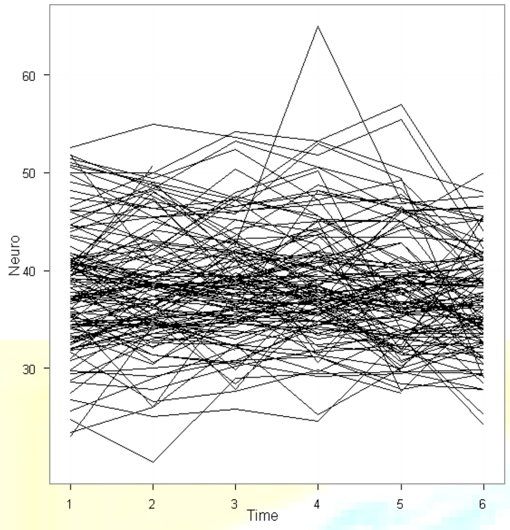
In many fields of biomedical, behavioral and organizations each individual is important along with the group. Do individual's start at same or different points? Why some individual's grow more than others? Are these changes related to some known factors? When change is being measured, every individual in the subject is being measured several times on successive occasions. Traditionally, a researcher tries to measure change with two waves of data, which is an inadequate strategy as it contains minimal information on individual change. Snapshots of before and after do not explore the various aspects of developmental trajectories. The past years have seen considerable progress in development of statistical methods for analysis of longitudinal data.

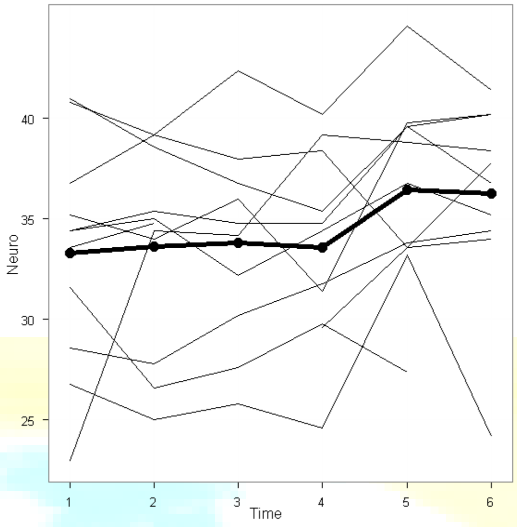
Visualizing longitudinal data

The questions regarding growth trajectory of individual and group discussed above can be visualized by plotting longitudinal data.

Figure A. *Spaghetti Plots of Observed Data on Each Subject* ***B.*** *Estimated Means (Solid Black Line) and Estimated Individual Trends for Randomly Selected Subset.*

Figure A. **Figure B.**





Visual displays of data are informative and should be preceded before any data analysis. Data display can be obtained from long as well as wide format. These plots serve as important tools to understand nature of longitudinal data which eventually helps in application of appropriate statistical technique. Spaghetti plot as displayed in figure A gives information on outcome variable of the subjects and individual differences in same during follow up period. This graph can help in identifying outliers but it is not very informative about mean response over time if sample size is more. In this type of situation, individual growth trajectory of a subset of group selected randomly can be plotted against group growth trajectory to assess relative performance and can be seen in figure B.

Data Structure

The data file in cross-sectional data used to have desired format required for application of suitable statistical techniques. It should be observed that organization of data itself in case of longitudinal data is not straightforward like cross-sectional data. Longitudinal data set therefore require explicit attention. In longitudinal dataset, the data can be arranged in two different ways: long format and wide format. Different statistical methods as well as different software require dataset in different **formats**. Due to this reason there may be a need to convert long format dataset to wide format and vice-versa. Both types of dataset are explained below.

IJMRA

Long Format

Long format is also known as person-period data set. Repeated measures of subjects appear vertically one below another (Long, 2012) and therefore, also known as univariate format. Static or fixed variables appear in an additional column and their values remain fixed for different time. All person-period data set contain four types of variables (Singer & Willett, 2003):

- A subject identifier
- A time indicator
- Outcome variable(s)
- Predictor variable(s)

The person-period data set can easily accommodate any data collection schedule consisting of multiple numbers of outcomes with different combinations of time-varying and time-invariant covariates. The data set for subjects that measured six times in long format can be depicted as follows:

Table 1. *Long format*

Subject	Gender	Family	Edu	Time	Neuro
1	1	0	1	1	36.2
1	1	0	1	2	39.4
1	1	0	1	3	42
1	1	0	1	4	45.3
1	1	0	1	5	39
1	1	0	1	6	46
2	0	1	0	1	27.5
2	0	1	0	2	29.4
2	0	1	0	3	30.6
2	0	1	0	4	33.2
2	0	1	0	5	37.9
2	0	1	0	6	36.2

Wide Format

It is also known as person-period data set. In this data set, data collected at different time points appear in multiple columns (Long, 2012). The primary advantage of wide format data set is the ease with which each person's growth record can be visualized in a temporal sequence. Repeated measures are distinguished from static variable by multiple column labels. It is also known as multivariate format and is popular for cross-sectional data format. In this set-up each subject has only one row of data to record time dependent and time independent variables. Total number of rows in this set-up is equal to number of subjects used in the study. According to Singer and Willett (2003), despite the ease with which each person's empirical growth record can be visualized, the person-period data set has four disadvantages that make it a poor choice for most longitudinal analyses:

- It leads to non-informative summaries
- It omits an explicit "time" variable
- It is inefficient when the number and spacing of follow-up varies accordingto individuals
- It cannot easily handle the presence of time-varying covariates.

The data set in wide or person-period set-up for subjects measured six times can be represented as:

Table 2.*Wide format*

Sub	Gender	Family	Edu	Time1	Time2	Time3	Time4	Time5	Time6
1	1	0	1	36.2	39.4	42	45.3	39	46
2	0	1	0	28.6	27.5	29.4	30.6	33.2	36.2

Missing Data

Attrition, omission and planned missingness have limited the ability of researchers to conduct most appropriate analyses (Duncan et al., 2006). The missing data problem is critical in longitudinal data analysis as structure of missing mechanism is complicated due to involvement of both intermittent missing pattern and dropout. In, longitudinal studies subjects are measured repeatedly and despite taking all the precautions some subjects miss their schedule. Intermittent missing pattern arise as participants occasionally miss their schedules but remain part of the study. In case of dropout, participants prematurely withdraw from the study. This leads to sample **consisting** of subjects with incomplete observations. Identification of assumptions regarding missing data are critical to obtain unbiased parameter estimates. Missing data can be broadly divided into three categories as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Little & Rubin, 2002; Rubin, 1976).

MCAR

Data is said to be at MCAR when probability of missing response does not depend on the observed or missing values. In MCAR mechanism, observed values can be considered as random sub sample of the total values. The MCAR assumption is the most stringent of the missing data mechanisms, and can be checked using Little's multivariate test for MCAR. Presence or absence of an observation is completely unrelated to the value that might have been observed. Such data are easy to handle as cases with missing data can be excluded. Even though this may not be efficient such an approach will not introduce a systematic bias.

MAR

This means that '***missingness'***depends on the observed values but not on missing value that may have been obtained. In this case the probability of missing data on a particular variable *W* can depend on other observed variables but not on *W*itself. This cannot be checked from sample values. MCAR and MAR are known as ignorable mechanisms. If a suitable statistical model can be developed to describe the observed data, then valid inference can be made using ML methods applied to observed incomplete data.

MNAR

MNAR means that '***missingness***'depends on both the observed and missing values. Since probability of missing data is related to some elements of missing values, it is also known as informative or non-ignorable mechanism. The term non-ignorable refers to fact that missing data mechanism cannot be ignored. When missingness is non-ignorable, it means that future unobserved response cannot be predicted, conditional on past response; instead a model to incorporate missing data mechanism is required (Nakai & Ke, 2011). Imputation methods, including multiple imputation, rely on the assumption that data are missing at random (MAR)and are not applicable for MNAR mechanism.

Traditionally, subjects with missing data are deleted from analysis but this approach should not be adopted due to two major reasons. First, longitudinal studies are costly and time consuming and deletion technique will reduce the sample size. Second, the subject with missing data may follow altogether different pattern and thus analysis with complete data may **lead**to false conclusion.

Software Packages for Longitudinal Data Analysis

Longitudinal research is time consuming and hence investigators try to avoid this design. However, it is necessary to take up longitudinal studies to understand how the cohort behaves over a period of time. The advances in methodology and computation are also a result of this. The main reason that can be attributed to this increased use is availability of efficient statistical techniques and user friendly software packages to analyze this type of data. Longitudinal data can have continuous or categorical outcome variable. GEE, LGC using SEM, mixed models (linear mixed models and generalized linear mixed models) are the advanced statistical techniques which provide flexibility and power to handle longitudinal data. An attempt is made here to give a brief summary of different software packages available for applying the above mentioned techniques in analyzing longitudinal data.

EQS

EQS is a software for LGC using SEM which uses simple and straight forward specification language to describe the model to be analyzed. It is equipped with extensive syntax error-checking, to make use of the program relatively easy and error free (Duncan et al., 2006). The output in EQS is produced in the form of text file and can be exported to get path diagram with parameter estimates. The current release of EQS is 6.2 and more information regarding this software can be obtained from official EQS manual by Bentler(2005).

HLM

HLM (Hierarchical linear models) is a dedicated software package which can handle multilevel or mixed effect models. The HLM program allows for continuous, count, nominal and ordinal outcome variables and assumes a functional relationship between the expectation of outcome and linear combination of a set of explanatory variable. HLM 7.01 for windows is the currently available commercial software. Introduction and advances of HLM 7.01 can be obtained from book written by Garson (2013).

LISREL

LISREL (Linear Structural Relationships) is one of the most widely used and oldest dedicated software for SEM. It accepts LISREL and SIMPLIS input as different command languages in input file. LISREL input is written in matrix notation whereas all the matrix notations can be avoided in SIMPLIS. The latest release is LISREL 9.2. More information regarding LISREL can be obtained from scientific software international (SSI).

MLwiN

MLwiNhas been created by center of multilevel modeling team based at **Universityof** Bristol with colleagues in other centers. One of the prominent features of MLwinN is to allow model setup and results display using standard model equations. The current version of MLwiN has made easy to use provision where wide variety of multilevel models, including generalized multilevel models for non-normal response variable can be fitted. The current version of the software is MLwiN 2.32. The detailed information regarding MLwiN can be obtained from university of Bristol web page.

Mplus

Mplus is a statistical modeling program that provides a flexible tool to variety of SEM models, estimators and algorithms for analyzing data. It consists of set of eleven commands, which offer several options for different type of data. Mplus is a versatile program which have many features that are not offered by other programs (Duncan et al., 2006).Mplus Version 7.3 is now commercially available.Mplus has extensive documentation for learning and advanced analysis on site.

R

Ris a free software environment for statistical computing and graphics (Team, 2005). R is widely popular among researchers for data analysis and**used** to develop new statistical software's. R **comes** with few built in base packages, but many are available as add-ons in the form of dedicated packages. The packages namely, 'gee' (Carey, 2011) and 'geepack' (Halekoh, Højsgaard, & Yan, 2006) can be used for GEE procedures. The QIC index for model selection can be calculated with MESS package (Ekstrom, 2013). The mixed models in R are available through 'nlme' (Pinheiro, Bates, DebRoy, & Sarkar, 2010) and 'lme4' (Bates, Maechler, Bolker, & Walker, 2013) packages. These are very mature and widely used packages for dealing with longitudinal/clustered data. R also offers a variety of dedicated software's for structural equation modeling. 'OpenMx', 'lavaan' and 'sem' are widely used programmers in R platform. OpenMx is free, full featured, open source structural equation modeling software (Boker et al., 2011). The acronym lavaan stand for latent variable analysis and its long term goal is to provide a collection of tools that can be used to **explore,** estimate, and understand a wide family of models (Rosseel, 2012). The package sem(Fox, Kramer, & Friendly, 2010) contains function for fitting general linear structural equation models. Both, lavaan and sem are not well developed as compared to OpenMx for categorical outcome variables.

SAS

SAS is comprehensive statistical software used widely by applied researchers. SAS 9.4 is most recent release of SAS institute. GEE in SAS is carried out with 'PROC GENMOD' command by specifying a repeated statement in which clustering information and working correlation matrix is provided by the user. GENMOD procedures have option for different models and working correlation structures. The LMM, GLMM and NLMM in SAS can be analyzed with option like, 'PROC MIXED', 'PROC GLIMMIX' and 'PROC NLMIXED'. LGC modeling technique using SEM has become popular statistical technique for modeling longitudinal data. The 'CALIS' procedure in SAS can be used to fit LGC models. SAS reports a broad range of fit indices.

STATA

STATA (2011) is a general purpose statistical analysis package that can be used to perform different statistical tests. STATA is mainly menu driven software but can be used for programming also. The GEE model in STATA can be fitted using ‘xtgee’ command, which is a part of ‘xt’ cross-sectional time series analysis tools (Horton & Lipsitz, 1999). The QIC module (Cui, 2007), which is equivalent of AIC criterion for maximum likelihood estimation based model can be used as a fit index for GEE. The different categories of mixed models in STATA such as ‘xtreg’ (continuous response), ‘xtlogit’ (binary response), ‘xtpois’ (count response variable) and many more can be fitted in STATA. The detailed information about ‘sem’ command can be obtained from STATA SEM manual. SEM encompasses a broad array of models ranging from linear regression to measurement models to simultaneous equations which include latent growth curve models (LGC) also. STATA 13 is the latest release available to users.

SPSS

A statistical package for social sciences (SPSS) is comprehensive statistical software where most commands are available through graphical user interface or through command syntax. The GEE procedure in SPSS is implemented using ‘GENLIN’ command. It is available in the Advanced Statistics option. The GEE procedure is available as sub-option in the ‘Generalized Linear Models’ option from menu ‘Analyze’. The between-subject and within-subject variables along with working correlation structure are to be selected form available options (Subba, Thennarasu, & Subbakrishna, 2012). SPSS is capable of performing various types of popular mixed models using MIXED model options. The mixed models procedures in SPSS are a part of the advanced models module that can be used in combination with base package. LGC models in SPSS can be analyzed with Amos package. Model in Amos can be fitted with standard commands and path diagrams using Graphics. SPSS Amos enable researchers to specify, estimates, assess and present models to show hypothesized relationship among variables. SPSS 22 is latest commercial release.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. *R Package Version*, 0–1.

Bentler, P. M. (2005). EQS 6: Structural Equation Program Manual (Multivariate Software, Encino, CA).

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., … Bates, T. (2011). OpenMx: an open source extended structural equation modeling framework. *Psychometrika*, *76*(2), 306–317.

Carey, V. J. (2011). gee: Generalized Estimation Equation solver, R package version 4.13-17, http://CRAN.R-project.org/package=gee.

Cui, J. (2007). QIC program and model selection in GEE analyses. *Stata Journal*, *7*(2), 209–220.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Routledge.

Ekstrom, C. (2013). MESS: Miscellaneous esoteric statistical scripts, R package version 0.1-6, http://CRAN.R-project.org/package=MESS.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis* (Vol. 998). Wiley.

Fox, J., Kramer, A., & Friendly, M. (2010). sem: structural equations models. R package version 0.9-20. htt p. *CRAN. R-Project. Org/package= Sem*.

Garson, G. D. (2013). Introductory guide to HLM with HLM 7 software. *Hierarchical Linear Modeling: Guide and Applications. Sage Publications, Los Angeles et Al*.

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, *15*(2), 1–11.

Horton, N. J., & Lipsitz, S. R. (1999). Review of software to fit generalized estimating equation regression models. *The American Statistician*, *53*(2), 160–169.

Lenzenweger, M. F., Johnson, M. D., & Willett, J. B. (2004). Individual growth curve analysis illuminates stability and change in personality disorder features: the longitudinal study of personality disorders. *Archives of General Psychiatry*, *61*(10), 1015.

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data.

Long, J. D. (2012). *Longitudinal data analysis for the behavioral sciences using R*. SAGE Publications, Incorporated.

Nakai, M., & Ke, W. (2011). Review of the Methods for Handling Missing Data in Longitudinal Data Analysis, *5*(1), 1–13.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2010). the R Core team (2009) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-96. *R Foundation for Statistical Computing, Vienna*.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). SAGE Publications, Incorporated.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*(3), 726.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical*.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis* (First). New York, NY: Oxford University Press.

Singer, J. D., & Willett, J. B. (2005). Individual Growth Modeling□ :Modern Methods for Studying Change Individual Growth Modeling□ : Modern Methods for Studying Change, (May).

Team, Rd. C. (2005). *R: A language and environment for statistical computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. url: http://www. R-project. org.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422.

IJMRA