BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Longitudinal and Correlated Data (LCD)

Semester 1, 2023

Prepared by:

Andrew Forbes and Jessica Kasza
Department of Epidemiology and Preventive Medicine,
Monash University

John Carlin, Lyle Gurrin and John Holmes
School of Population and Global Health, University of Melbourne,
(*John Holmes only*) School of Mathematics and Statistics, University of Melbourne and (*John Carlin only*) Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute

MONASH University

THE UNIVERSITY OF MELBOURNE

Contents

Longitudinal and Correlated Data (LCD)
Semester 1, 2023

Contact details

Lyle C. Gurrin
BSc(Hons) PhD GradCertUniTeach AStat FRSS

Professor of Biostatistics
Centre for Epidemiology and Biostatistics
Melbourne School of Population and Global Health
The University of Melbourne

tel1: +61 3 8344 0731
tel2: +61 4 0699 6731
email: lgurrin@unimelb.edu.au

Background

Longitudinal and correlated data arise in many settings in health and medical research. Common examples include studies involving repeated measurements of individuals over time, in clinical trials and cohort studies, and cluster-randomised trials where participants are clustered within natural units such as schools or medical practices. The common characteristic of these data structures is that of correlated measurements either within an individual or within a cluster of individuals. Standard methods of statistical analysis assume independent observations and therefore do not accommodate this correlation, and more sophisticated methods need to be considered. There have been significant developments in these methods and their availability in statistical software packages in recent decades.

Context within the program

This unit builds on the knowledge and skills that students gained in the unit Regression Modelling for Biostatistics 1 (RM1), particularly linear regression for continuously-valued outcomes and logistic regression for binary outcomes. To accommodate the correlated data structures that typically result from longitudinal and cluster studies requires extending both the statistical models (that provide an idealised generative process for the data) and the estimation methods (that are applied to data to estimate the model parameters).

Prerequisites

Epidemiology (EPI), Mathematical Foundations for Biostatistics (MFB), Principles of Statistical Inference (PSI), Regression Modelling for Biostatistics 1 (RM1).

For University of Melbourne students the corresponding units are Epidemiology 1 (POPH90014 = EPI), Probability and Inference in Biostatistics (MAST90100 = PSI),

Foundations of Regression (MAST90102 = RM1) and Advanced Regression (MAST90099 = RM2).

Co-requisites

Nil

Unit summary

This unit covers statistical models for longitudinal and correlated data in medical research. The concept of hierarchical data structures is developed, together with simple numerical and analytical demonstrations of the inadequacy of standard statistical methods. Beginning with models based on normal distributions, appropriate statistical methods involving generalised estimating equations and mixed linear models are developed and explored using the R and Stata statistical software packages. The limitations of traditional repeated measures analysis of variance are briefly discussed. Extensions to non-normal outcomes are developed and using a set of case studies, approaches based on generalised estimating equations (GEE) and generalised linear mixed models (GLMM) are developed and contrasted. Throughout, emphasis is placed on interpretation issues focussing on the underlying clinical or public health research question.

Workload requirements

The expected workload for this unit is 10-12 hours per week on average. This comprises of guided readings, discussion posts, independent study and the completion of assessment tasks.

Learning Outcomes

At the completion of this unit students should be able to:

1. Recognise the existence of correlated or hierarchical data structures, and describe the limitations of standard methods in these settings
2. Develop and analytically describe appropriate models for longitudinal and correlated data based on subject matter considerations
3. Be proficient at using statistical software packages (Stata and R) to fit models and perform computations for longitudinal data analyses, and to correctly interpret results
4. Express the results of statistical analyses of longitudinal data in language suitable for communication to medical investigators or publication in biomedical or epidemiological journal articles

Unit content

The unit is divided into 6 modules, summarised in more detail below. Each module will involve either 2 or 3 weeks of study and generally includes the following material:

1. Module notes describing concepts and methods, including some exercises of a more “theoretical” nature.

2. Selected readings from published articles or textbooks.
3. One or more extended examples illustrating the concepts/methods introduced in the notes and including more practically oriented exercises.

Study materials for all Modules are downloadable from the corresponding University of Sydney *Canvas* website. Assignments and supplementary material, such as datasets will be posted to the unit website.

Recommended approaches to study

Students should work through each module systematically, following the module notes and any readings referred to, and working through the accompanying exercises. You will learn more efficiently if you tackle the exercises systematically while working through the notes. You are encouraged to post any content-related questions to eLearning, whether they relate directly to a given exercise, a request for clarification or further explanation of an area in the notes. You should also work through all the computational examples in the notes for yourself on your own computer.

Outline solutions to the exercises in each module (except those to be submitted for assessment, as described below) will be posted online at the midway point of the allocated time period for the module. This is intended to encourage you to attack the exercises independently (or via the Canvas site), and yet not make you wait too long to see the sketch solutions.

Method of communication with coordinator

Questions about administrative aspects or course content can be emailed to the coordinator(s), and when doing so please use "LCD:" in the Subject line of your email to assist in keeping track of our email messages. Coordinator/s will be available to answer questions related to the module notes and practical exercises, and to address any other issues that require clarification. However, please note that instructors are not necessarily available every day of the week and you should expect that it may take a day or so to respond to questions (possibly longer over weekends and during breaks!).

We strongly recommend that you post content-related questions to the Discussions tool (accessible from the Modules link on the front page) in the LCD area of BCA's Canvas website, hosted by the University of Sydney. You may be familiar with the system from previous BCA units and will receive any specific instructions on using the eLearning site this semester from the BCA Coordinating Office. There is also a "Let's get started" document available on the Student Resources page of the BCA website.

Relying on Canvas for content-related communication and problem-solving will enable other students to benefit from responses and indeed to respond themselves, and we try to encourage as much interaction as possible within the class through this medium. We will also use Canvas for posting all course materials although some of the core material (particularly selected readings, whose reproduction is subject to copyright considerations) is also sent out in paper form.

Module descriptions

Below is an outline of the study modules, followed by a timetable and assessment description table.

Each module is scheduled to begin on a Monday and conclude on the Sunday of the following week. The due date for submission of the short assessment for each of Modules 1 to 5 is 11:59pm on the Monday immediately after the completion of the module, as indicated in the assessment table below. Assignments No. 1 and No. 2 are due at least one week after the conclusion of Modules 3 and Module 6 respectively. Note that no new material will be introduced in the week before an Assignment is due.

Module 1: Introduction to correlated data using paired data and simple clustered data.

- Paired data: the simplest correlated data structure
- Advantage of modelling approach e.g. with missing data, to enable use of both within- and between-subject information where possible, leading to simple random effects model.
- Extension to exchangeable clustered data with varying numbers of individuals per cluster, and consideration of between-cluster effects
- Introduction to generalised estimating equations (GEE)

Module 2: Overview of different correlated and longitudinal data structures and related research questions

- Examples of two major types of problem: cluster-randomised trials and repeated-measures longitudinal studies.
- Simple approaches to analysis: graphical display (trajectory plots, pairwise correlations), and summary measures approach to analysis
- Cluster-randomised trials: design effect and simple approaches to analysis.

Module 3: Methods for continuous outcome measures based on generalised estimating equations (GEE)

- The marginal model approach to handling correlation within clusters or individuals (by generalising the standard regression model to allow correlated error terms)
- Robust (information-sandwich) standard errors.
- Random effects specifications, i.e. conditional/ multilevel/ hierarchical structure and relationship to marginally specified models

Module 4: Methods for continuous outcome measures based on normal mixed models, with likelihood-based estimation.

- Alternative approaches to estimation: weighted/generalised least squares, maximum likelihood and REML.
- Separating between- and within-individual (or group) effects
- Classical repeated measures ANOVA and relationship to modern modelling approaches.
- Missing data: importance of assumptions about mechanism for missingness, and implications for GEE and likelihood-based estimation.

Module 5: Methods for discrete data: GEE and generalized linear mixed models (GLMM)

- Binary outcomes and logistic regression models: generalising to correlated data. Methods focussing on the marginal mean structure: estimating equations in general and GEE. Linear marginal model no longer corresponds to a linear conditional model.
- Methods using a full (multilevel) model specification.
- Advantages and disadvantages of each approach, particularly in the interpretation of “subject-specific” and “population-average” parameters.

Module 6: Methods for count data; transition models

- Poisson regression model, using GEE and GLMM approaches.
- Negative-binomial model.
- Transitional or Markov models: application to modelling change or incidence.

Unit schedule 2023

Week	Week starting	Module	Topic	Assessment
1	February 27	1	Introduction to correlated data using paired data and simple clustered data	
2	March 6	1		Mod 1 Assess due March 13
3	March 13	2	Overview of different correlated and longitudinal data structures and related research questions	
4	March 20	2		Mod 2 Assess due March 27
5	March 27	3	Methods for continuous outcome measures based on generalised estimating equations (GEE)	
6	April 3	3		
7	April 10	3	Mid Semester break including Easter break starting Friday 7 April	Mod 3 Assess due April 17
8	April 17		Questions about **Assignment No. 1** (no new material this week)	Assignment No. 1 due April 24
9	April 24	4	Methods for continuous outcome measures based on normal mixed models, with likelihood-based estimation	
10	May 1	4		Mod 4 Assess due May 8
11	May 8	5	Methods for discrete data: GEE and generalized linear mixed models (GLMM)	
12	May 15	5		Mod 5 Assess due May 22
13	May 22	6	Count data, transition models	
14	May 29	6		**NO** Mod 6 Assess
15	June 5		Questions about **Assignment No. 2**	Assign No. 2 due June 12
16	June 12			

Assessment

Assessment will include two written assignments worth 30% each, to be made available in the middle and at the end of the semester, and to be completed within approximately two weeks. In addition, students will be required to submit solutions to selected practical exercises (one from each module **except** Module 6), worth a total of 40%, by deadlines specified throughout the semester (see table below).

Assessment name	Assessment type	Coverage	Learning objectives	Weight
Module 1	Short Assessment	Module 1	1,2	8%
Module 2	Short Assessment	Module 2	1,2,3,4	8%
Module 3	Short Assessment	Module 3	1,2,3,4	8%
Assignment No. 1	Assignment	Modules 1-3	1,2,3,4	30%
Module 4 exercises	Short Assessment	Module 4	1,2,3	8%
Module 5 exercises	Short Assessment	Module 5	1,2,3	8%
Assignment No. 2	Assignment	Modules 1-6	1,2,3,4	30%

You should submit material for assessment using the Assignments tool in *Canvas*. Where the work involves algebraic derivations that you find easier to complete by hand then you should scan your work to electronic form for submission. This handwritten work should be scanned and collated into a single pdf file and submitted via the *Canvas* site. In general, we prefer that your work be typed in Microsoft Word or similar and recommend the use of either LaTeX or Microsoft's Equation Editor for algebraic work which is now much easier to use than previous versions. See the BCA Assessment Guide document for specific guidelines on acceptable standards for assessable work.

Please note that the instructors will not answer questions online relating directly to the assessable material until after it has been submitted. However, with respect to the five module-based assessments, we encourage students to discuss any related material between themselves, via Canvas, *as long as explicit solutions to the exercises are not posted for others to use,* and each student's submitted work is clearly their own, with anything derived from other students' discussion contributions clearly attributed to the source. Note that in contrast, the two "major" assignments require completely independent work by students.

Submission and academic honesty policy

You should submit all your assessment material via the Canvas site unless otherwise advised. The use of Turnitin for submitting assessment items has been instigated within unit sites. For more detail please see pages 3-5 the BCA Student Assessment Guide. This guide will also be included in hardcopy in your package of notes.

The BCA pays great attention to academic honesty procedures. Please be sure to familiarise yourself with these procedures and policies at your university of enrolment.

Links to these are available in the BCA Student Assessment Guide. When submitting assessments using Turnitin you will need to indicate your compliance with the plagiarism guidelines and policy at your university of enrolment before making the submission.

A special note regarding "contract cheating" sites: Unfortunately there have been instances in the past of students using such websites to post assignment questions and receive solutions (usually for a fee). We have arrangements with these sites to identify the students posting questions or accessing the solutions, and such students will be referred to and face disciplinary processes at their home university.

Late submission and extension procedure

We adhere to standard BCA policy for late penalties for submitted work, i.e. a 5% deduction from the earned mark for each day the assessment is late, up to a maximum of 50%. Extensions are possible, but these need to be applied for (by email) as early as possible. The Unit Coordinator is not able to approve extensions beyond three days; for extensions beyond three days you need to apply to your home university, using their standard procedures.

Learning resources

There is no single prescribed text for the subject, but a number of reference books are recommended as background material (list below). The first book in the list is the one that we find closest to our approach in LCD (although it appeared after the first draft of the course was written), so if you were to obtain one book this would be our recommendation. The module notes and case studies form the primary material for this subject, and required readings from selected texts, are provided in the mailout package.

Fitzmaurice G, Laird N, Ware J. *Applied Longitudinal Analysis*, John Wiley and Sons, 2004. [Note that a 2nd edition appeared in 2012. All readings for this semester are taken from the 2004 edition – they differ minimally from the 2012 edition]

Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*, 2nd Edition, Oxford University Press, 2002.

Singer JD, Willett JB. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*, Oxford University Press, 2003.

Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*, Springer, 2000.

Brown H, Prescott R. *Applied Mixed Models in Medicine*, 3rd Edition, Wiley, 2015.

Software requirements and assumed knowledge

For this subject you will need to have access to, and a working familiarity with, either Stata or R. While some of the course was originally developed with a dependence on SAS, the difficulties some students faced in getting access to SAS, as well as the greater ease of use of Stata and availability of R (and its much improved capacity for fitting mixed models) means that SAS is no longer used in LCD. All methods in this unit can be

conducted using Stata alone or R alone, however, the set of examples given with Stata code is somewhat more complete than the set of examples with R code.

Students using Stata will need at least version 13 that was released in July 2013. The current version is Stata 17 released in April 2021. We are not aware of any major differences between Stata versions that affect the material, but minor issues will be pointed out in *Canvas* postings. Importantly, whichever version you are using, please ensure that you have performed the online update to the latest update of that version. (Use the command **`update query`**).

For R, the notes assume you are working with the latest version, although slightly earlier versions should not have any important differences. The latest version of R is R 4.2.2 "Innocent and Trusting" released in October 2022.

Required mathematical background

No additional mathematical background is required beyond what is covered in Mathematical Foundations for Biostatistics (MFB), Principles of Statistical Inference (PSI), Regression Modelling for Biostatistics 1 (RM1).

Feedback

Our feedback to you

The types of feedback you can expect to receive in this unit are:

- Formal individual feedback on submitted exercises assignments
- Responses to questions posted on Canvas

Your feedback to us

One of the formal ways students have to provide feedback on teaching and their learning experience is through the BCA student evaluations at the end of each unit. The feedback is anonymous and provides the BCA with evidence of aspects that students are satisfied with and areas for improvement.

Unit changes, including response to recent student evaluation

LCD was last delivered in Semester 1 2022. There have been only minor changes since that delivery in the form of typos and minor edits for greater clarification of the text, and an extension of coding examples done in R to complement existing Stata code.

Acknowledgments

Two instructors, Andrew Forbes (Monash University) and John Carlin (MCRI and University of Melbourne) were jointly responsible for the development of the material for this unit.