BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Longitudinal and Correlated Data

Semester 1, 2024

Prepared by:

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

John Carlin, Lyle Gurrin, John Holmes and Koen Simons
School of Population and Global Health, University of Melbourne,
(*John Holmes only*) School of Mathematics and Statistics,
Canterbury University, New Zealand
(*John Carlin only*) Clinical Epidemiology and Biostatistics Unit,
Murdoch Children's Research Institute
(*Koen Simons only*) School of Public Health and Community
Medicine, University of Gothenburg University, Sweden

MONASH University

THE UNIVERSITY OF MELBOURNE

Contents

Longitudinal and Correlated Data (LCD)

Semester 1, 2024

Contact details of unit co-ordinator

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**

If you have any general BCA queries, please contact: Jacqueline (Jaqē) Vaughan or Emily Higginson at the BCA Coordinating Office on **02 9562 5076/54** or email **bca@sydney.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, consisting of guided readings, discussion posts, independent study, and 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 approximately 2 or 3 weeks of study and includes the following material:

1. Module notes describing concepts and methods, and including some exercises of a more "theoretical" nature;
2. Selected readings from published articles or textbooks; and
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 accessed from the eLearning unit site. Assignments and supplementary material such as datasets will be available within each Assignment item. Please note that we are not able to post copies of copyright material (journal articles and book extracts)—for these you will have to rely on your home university's library.

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 a lot more efficiently if you tackle the exercises systematically as you work through 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 for the module. This is intended to encourage you to attack the exercises independently (or via the eLearning site), and yet not make you wait too long to see the sketch solutions.

Make the most of this unit by engaging with coordinators and fellow students on the Discussion Board and in Tutorials. These are safe spaces to discuss the course material and related ideas and students are encouraged to make the most of them by engaging in respectful discussion.

Questions about Assignments should be directed to the coordinator in the first instance to avoid any Academic Honestly issues.

Method of communication with coordinator(s)

Questions about administrative aspects or course content can be emailed to the coordinator. 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.

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 Discussion Board in the unit site.

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 required exercises from each module is 11:59pm on the day immediately following the completion of the module, as indicated 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; and
- 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; and
- 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; and
- 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; and
- 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; and
- Transitional or Markov models: application to modelling change or incidence.

Unit schedule

Semester 2, 2024 starts on Monday 26th February 2024

Week	Week commencing	Module	Topic	Assessment
1	Feb 26	1	Introduction to correlated data using paired data and simple clustered data	
2	Mar 4	1		
3	Mar 11	2	Overview of different correlated and longitudinal data structures and related research questions	
4	Mar 18	2		Module 1 – 2 assessment due **March 25**
5	Mar 25	3	Methods for continuous outcome measures based on generalised estimating equations (GEE)	
6	Apr 1	3	Mid Semester break including Easter break starting Friday March 29	
7	Apr 8	3	Questions about **Assignment No. 1** (no new material this week)	Module 1 – 3 assessment due **April 15**
8	Apr 15	4	Methods for continuous outcome measures based on normal mixed models, with likelihood-based estimation	
9	Apr 22	4		
10	Apr 29	5	Methods for discrete data: GEE and generalized linear mixed models (GLMM)	
11	May 6	5		Module 4 – 5 assessment due **May 13**
12	May 13	6	Count data, transition models	
13	May 20	6		
14	May 29	6	Questions about **Assignment No. 2** (no new material this week)	Module 4 – 6 assessment due **June 5**

Assessment

- Two sets of modules exercises worth 20% each, to be made available after Module 2 and Module 5, and to be completed within approximately two weeks; and
- Two written reports worth 30% each, to be made available after Module 3 and Module 6, and to be completed within approximately three weeks.

In addition, students will be required to submit solutions to selected practical exercises (one from each module except Module X), worth a total of XX%, by deadlines specified throughout the semester (see table below).

Assessments are due by 11:59pm on the stated day.

Assessment name	Assessment type	Coverage	Learning objectives	Weight
Module 1-2 exercises	Short-answer questions	Module 1-2	1-2	20%
Module 1-3 assignment	Report	Module 1-3	1-3	30%
Module 4-5 exercises	Short-answer questions	Module 4-5	2-4	20%
Module 4-6 assignment	Report	Module 4-6	1-4	30%
Online quizzes	Non-assessed	Various	Various	0%

In general, you are required to submit your work as a document prepared in Microsoft Word or something similar. We strongly recommend you become familiar with equation typesetting software such as Microsoft's Equation Editor (or use LaTeX) for algebraic work. You may submit neatly handwritten work, however, please note that marks will potentially be lost if the solution cannot be understood by the markers due to unclear or illegible writing. Handwritten work should be scanned and collated into a single pdf file and submitted via the eLearning site. See the BCA Assessment Guide for guidelines on acceptable standards for assessable work.

Students are encouraged to discuss relevant topics in the Discussion Board. Please avoid posting questions relating directly to assessable material. These should be emailed to the Unit Coordinator in the first instance.

Explicit solutions to assessable exercises should not be posted for others to use. Each student's submitted work must be clearly their own, with anything derived from other students' discussion contributions clearly attributed to the source.

Submission and academic honesty policy

All assessment material should be submitted via the relevant Assessment module in Canvas unless otherwise advised. Turnitin plagiarism detection is applied to all

submissions. For detailed information, please see the BCA Assessment Guide, which includes links to the Academic Honesty policies at member universities. Please familiarise yourself with the procedures and policies at your home university. You will need to indicate your compliance with the plagiarism guidelines and policy at your home university.

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 student posting questions or accessing the solutions, and such students will be referred to and face disciplinary processes at their home university.

Use of ChatGPT or other generative AI tools

The assessment tasks in this Unit have been designed to be challenging, authentic and complex. Although individual assessment components may provide specific guidance regarding the use of generative artificial intelligence (AI) tools (e.g., ChatGPT), successful completion of these components will require students to critically engage in specific contexts and tasks for which AI will provide only limited support and guidance. In all cases, a failure to reference the use of generative AI may constitute student misconduct under the Student Code of Conduct of your University of enrolment. To successfully complete assessment tasks, students will be required to demonstrate detailed comprehension of their written submission independent of AI tools.

Late submission and extension procedure

The standard BCA policy for late penalties for submitted work is a 5% deduction from the earned mark for each day the assessment is late, up to a maximum of 10 days (including weekends and public holidays). Extensions are possible, but these need to be applied for (by email) as early as possible. The Unit Coordinator can approve extensions up to three days; for extensions beyond three days, you must apply to your home university, using their standard procedures.

Learning resources

There is no single prescribed text for the subject, but a few 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.

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 18 released on April 25th, 2023. 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 4.3.2 "Eye Holes" released for Hallowe'en on October 31st 2023. For help with R, please see Learning R in the Student Resources site.

If you have not yet organised access to these packages, you should do so as soon as possible. This is a practical course which requires regular use of the relevant software; delays in gaining access to these packages may impact your ability to complete the course. Information on how to download R and RStudio, and access Stata can be found in the BCA Textbook and Software Guide.

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
- Feedback from non-assessed online quizzes
- Responses to questions posted on *Canvas*

Your feedback to us

One of the formal ways students provide feedback on teaching and their learning experience is through the BCA student evaluation survey at the end of each semester. 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 2023. Apart from new assessment tasks, there have been only minor changes since that delivery in the form of correcting typographical errors and minor edits for greater clarification of the text. An extension of coding examples done in R to complement existing Stata code was completed in 2022.

Acknowledgments

Professor Andrew Forbes (Monash University) and Professor John Carlin (Murdoch Children's Research Institute (MCRI) and University of Melbourne) were jointly responsible for the development of the material for this unit. Dr John Holmes (University of Canterbury, New Zealand) added the data analysis examples using *R* in 2022, and Dr Koen Simons overhauled the assessment tasks in 2023.