BIO
STATISTICS
COLLABORATION
OF AUSTRALIA

Study Guide

LONGITUDINAL & CORRELATED DATA (LCD)

Semester 1, 2017

Prepared by:

Andrew Forbes
Department of Epidemiology and Preventive Medicine,
Monash University

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

MONASH University

THE UNIVERSITY OF
MELBOURNE

LONGITUDINAL & CORRELATED DATA (LCD)
Semester 1, 2017

Instructor contact details

A/Prof Lyle Gurrin

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

Tel: (03) 8344 0731

Email:

lgurrin@unimelb.edu.au

Dr Elasma Milanzi

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

Tel: (03) 8344 1788

Email:

elasma.milanzi@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.

Unit summary

This subject 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 SAS 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.

Objectives

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 SAS) 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

Method of delivery and communication

Two instructors (Andrew Forbes, John Carlin) were jointly responsible for the development of the material for this subject/unit and have in the past typically alternated taking the major role in coordination of the unit. This semester Lyle Gurrin will be taking primary responsibility as unit coordinator. Professors Carlin and Forbes will, however, keep an eye on the action! Dr Elasma Milanzi, a Lecturer in Biostatistics at the Melbourne School of Population and Global Health will also play a substantial role in the delivery of the unit. Her details and those of A/Prof Gurrin are above.

Questions about administrative aspects or course content can be emailed to the coordinator, and when doing so please use “LCD:” in the Subject line of your email to assist in keeping track of our email messages. Lyle 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 in the LCD area of BCA’s eLearning site. In 2017 we will continue to us the Blackboard system hosted by the University of Sydney. You should 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 “Getting Started” document available on the Student Resources page of the BCA website.

Relying on Blackboard 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 Blackboard 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.

Unit content

The unit is divided into 6 modules, summarised in more detail below. Each module will involve approximately 2 weeks of study and generally 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.

3. One or more extended examples illustrating the concepts/methods introduced in the notes and including more practically oriented exercises.

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 are encouraged to post any content-related questions to eLearning, whether they relate directly to a given exercise, or are a request for clarification or further explanation of an area in the notes. You should also work through all of 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 Blackboard site), and yet not make you wait too long to see the sketch solutions.

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 below).

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 Blackboard, 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.

You should submit material for assessment using the Assignments tool in eLearning. 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. In general, we prefer that your work be typed in Word or similar and recommend the use of Microsoft's Equation Editor for algebraic work which is now much easier to use after considerable evolution in new releases of MicroSoft Office. See the separate BCA document for specific guidelines on acceptable standards for assessable work.

Note that where assignment work is submitted online using the Blackboard Assessment tool, you will need to indicate your compliance with the plagiarism guidelines and policy before making the submission.

Please submit your assessment items on or before the due date. If you need an extension of time, for a **legitimate** reason such as a health problem, contact the coordinator, preferably well before the due date.

Late penalties: Where no extension has been granted, the mark obtained will be penalized by 5% of the total that you would have received per day late, up to a maximum of 50%, according to standard BCA policies.

In addition, please pay careful attention to the following documents that are included in the hardcopy package: BCA Assessment Policies and Procedures (including Universities' Plagiarism Policies), and the Assignment Cover Sheet.

Reference Books

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.]

Diggle PJ, Heagerty P, Liang K-Y, Zeger SL: *Analysis of Longitudinal Data* 2nd Edn. Oxford UP, 2002.

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

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

Brown H, Prescott R. *Applied Mixed Models in Medicine,* Wiley 1999.

Software

For this subject you will need to have access to, and a working familiarity with, either Stata or SAS and preferably both. Some of the course was originally developed with a dependence on SAS but the difficulties some students face in getting access to SAS, as well as the greater ease of use of Stata (and its much improved capacity for fitting mixed models) mean that SAS will not be an absolute requirement.

Stata 12 was released in July 2011 and we assume you are using at least this version. However, we expect most of you would be using Stata 13 or 14, the latter of which was released in the second half of 2015. We are not aware of any major differences between Stata versions that affect the material, but minor issues will be pointed out in Blackboard postings. If you only have an earlier version of Stata (v11 or earlier) then please email me (Lyle) and we can discuss your options. 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 SAS, the notes assume you have version 9.4, although slightly earlier versions should not have any important differences.

Timetable for modules of study and assessment tasks

Below (over the page) is an outline of the study modules and assessment tasks for this unit together with a timetable. Note that BCA Semester 1 starts Monday 6th March regardless of the official at the university at which you are enrolled.

Study materials for all Modules are contained in your mail-out package. Supplementary material, such as datasets, and Assignments, will be posted to Blackboard. We will also post the lecture notes on Blackboard, but 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 the hard copy mail-out or on your home university's library resources.

It is intended that students will work through the material for each module, including completion of practice exercises, by the end-date of the module. As stated above, we encourage online discussion of topics and exercises, which makes it important to work at a consistent pace with the rest of the class, as far as possible.

We have scheduled each module 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.**

Module 1: Introduction to correlated data using paired data and simple clustered data.
Mon 6th March – Sun 19th March

- 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 1 exercise: Due 11:59pm Mon 20th March

Module 2: Overview of different correlated and longitudinal data structures and related research questions
Mon 20th March – Sun 2nd April

- 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 2 exercise: Due 11:59pm Mon 3rd April

Module 3: Methods for continuous outcome measures based on generalised estimating equations (GEE)
Mon 3rd April – Sun 16th April

- 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 3 exercise: Due 11:59pm Tue 18th April
(Easter is Fri 14th April – Mon 17th April)

Assignment 1 due: 11:59pm Mon 24th April
Note the time allowed for completion of the assignment.
There is one week with no new material Mon 17th April to Sun 23rd April.

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

Mon 24th April – Sun 7th May

- 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 4 exercise: Due 11:59pm Mon 8th May

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

Mon 8th May – Sun 21st May

- 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, in particular interpretation of "subject-specific" and "population-average" parameters.

Module 5 exercise: Due 11:59pm Mon 22nd May

Module 6: Methods for count data; transitional models

Mon 22nd May – Sun 4th June

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

No Module 6 exercise is required for submission.

Assignment 2 due: 11:59pm, Mon 12th June