



Study Guide

Data Management and Statistical Computing (DMC)

Semester 1, 2023

Prepared by:

Dr Gabriella Lincoln School of Public Health
The University of Adelaide

Copyright © The University of Adelaide, The University of Queensland



Contents

Contact details	2
Background	2
Context within the program.....	2
Prerequisites	2
Co-requisites	3
Unit summary.....	3
Workload requirements.....	4
Learning Outcomes	4
Unit content	4
Recommended approaches to study	6
Method of communication with coordinator(s)	6
Module descriptions	7
Unit schedule	8
Assessment	8
Submission and academic honesty policy	9
Late submission and extension procedure.....	9
Learning resources	10
Software requirements and assumed knowledge	10
Required mathematical background.....	11
Feedback	11
Unit changes, including response to recent student evaluation.....	11
Acknowledgments.....	11

Data Management and Statistical Computing (DMC)

Semester 1, 2023

Contact details

Instructor

Dr Gabriella Lincoln

School of Public Health Faculty of Health
and Medical Sciences
The University of Adelaide
North Terrace Campus
Floor 4, Rundle Mall Plaza
SA 5000

gabriella.lincoln@adelaide.edu.au

If you have any general BCA queries, please contact: Karolina Kulczynska-Le Breton or Emily Higginson at the BCA Coordinating Office on 02 9562 5076/54 or email bca@sydney.edu.au

Background

The aim of this unit is to provide students with the knowledge and skills required to undertake moderate to high-level data manipulation and management in preparation for statistical analysis of data typically arising in health and medical research.

Context within the program

DMC provides the foundational knowledge on data management and statistical computing required to undertake the coursework for all units in the BCA program and represents the basic level of computing proficiency expected of a practising biostatistician.

Competent usage of more than one statistical package to perform common tasks is essential, as in recent years new statistical methods emerge on different software platforms with increasing frequency. This aspect is also reflected in the delivery of the BCA units that follow DMC, where the methods covered therein have been developed in R but not in Stata, or vice versa.

Prerequisites

None.

Co-requisites

None.

Unit summary

This unit introduces the software packages Stata and R, with the aim of providing a foundation to build upon in further studies and biostatistical career.

The unit is delivered through the Canvas eLearning platform at the University of Sydney.

Unit content will be uploaded to the Canvas e-learning platform in PDF format, including course modules notes, exercises, assignments, coursework solutions and supplementary material (except readings, which for this unit will be mostly found on the recommended textbooks or in publications available through University libraries).

Regular discussions on unit material will take place on Canvas' Discussion Board. Tutorial sessions will be held during the unit, bar during the mid-semester break and on the last week of the unit. Participation to the tutorials is encouraged and recommended, and details on schedule and content will be posted on Canvas.

Students are expected to engage with the Canvas eLearning platform often to ensure that they keep abreast of any notifications relating to the unit.

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. As in any unit, please note that the actual workload may vary for different students.

Learning Outcomes

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

1. Gain experience in data manipulation and management using two major statistical software packages (Stata and R)
2. Learn how to display and summarise data using statistical software
3. Become familiar with the checking and cleaning of data
4. Learn how to link files through use of unique and non-unique identifiers
5. Acquire fundamental programming skills for efficient use of software packages
6. Learn key principles regarding confidentiality and privacy in data storage, management and analysis

Unit content

The unit is divided into 3 modules, summarised in more detail below. Each module will involve approximately 4 weeks of study and generally includes the following material:

Module 1: Importing and exporting data; recoding and formatting data; labelling variables and values; use of date data, displaying and summarising data. Construction of suitable programming scripts to reproduce results.

Module 2: Graphs, Data management and Statistical Quality Assurance Methods. Includes advanced graphics for production of publication-quality graphs.

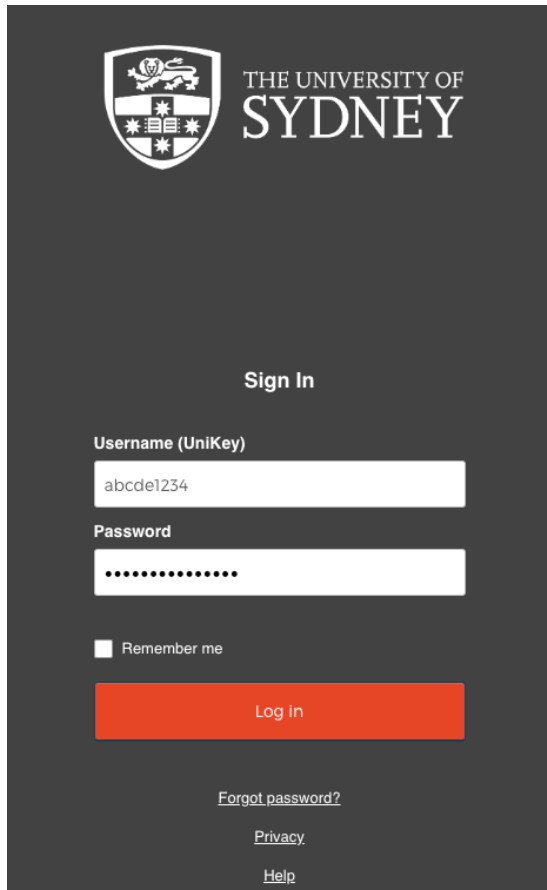
Module 3: More Advanced Statistical Computing: Using functions to generate new variables, appending, merging and transposing data; programming skills including macros, loops, user-defined functions and programs.

Study materials for all Modules are downloadable from the Canvas unit site. Canvas is administered by the BCA Office at the University of Sydney.

A Unikey will be provided for you to access Canvas (see image below). If you are a continuing student, please use the login details you used previously. For any issues with access, please contact the BCA office at bca@sydney.edu.au.

[The University of Sydney - Sign In](#)

Canvas Student App



The image shows a dark-themed sign-in page for The University of Sydney. At the top left is the university's crest, and to its right is the text 'THE UNIVERSITY OF SYDNEY'. Below this, the text 'Sign In' is centered. There are two input fields: 'Username (UniKey)' containing 'abcde1234' and 'Password' with masked characters. A 'Remember me' checkbox is present below the password field. A red 'Log in' button is at the bottom. At the very bottom, there are links for 'Forgot password?', 'Privacy', and 'Help'.



Assignments and supplementary material, such as datasets will be posted to the unit site.

Please note that we will not automatically post out copies of the course notes to all students. If you do wish to have a hard copy of the course notes, please request these from the Course Coordinator as soon as possible. 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.

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.

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 much more efficiently if you tackle the exercises systematically as you work through the notes.* You will find it easier to learn computing language syntax by recreating all the computational examples in the notes for yourself on your own computer.

As a grace note: we advise those new to statistical computing to practice coding in shorter but frequent sessions. Twenty minutes daily during the working week will help to assimilate the content much more effectively than coding for ninety minutes in one sitting at the weekend.

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 unit coordinator in the first instance to avoid any Academic Honesty issues.

Method of communication with coordinator(s)

Questions about administrative aspects or unit content can be emailed to the coordinator. Please use "(DMC: BCA):" 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.

Module descriptions

Each of the three modules for this unit are divided into part A and B (apart from a brief addendum for Module 2 referred to as “2C”). Modules run for a total of 4 weeks each and are scheduled to begin on a Monday. **The submission of the assessable exercises and assignments from each module is 2:00 PM ACST on the due date (please note this is the local time in Adelaide, as this is the delivering institution this semester).**

Below is the outline of the topics covered in each module.

Module 1:

- Reading in and importing datasets in various file formats into R and Stata
- Exporting datasets in Stata, R and other formats
- Exploring datasets descriptively
- Investigating potential data management issues
- Documentation and reproducibility of data manipulations
- Working with dates

Module 2:

- Planning data collection and database design
- Data cleaning and monitoring
- Preparing datasets for analysis
- Investigating relationships between variables
- Missing data basics
- Graphical displays for descriptive analysis

Module 3:

- Data manipulations and wrangling with numerical and string functions
- Merging, appending and reshaping datasets
- Local and global macros, scalars
- Loops
- Creating custom programs
- Saving estimation results for later use

Please ensure that you allow sufficient time for Module 3 as there is a leap in the level of difficulty moving from Module 1 and 2.

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

Unit schedule

Semester 1, 2023 starts on Monday **27 February, 2023**.

Live tutorials will be held mostly every other week and will be mostly guided by the topics raised from the questions posted by students on the Discussion Board.

Please ensure you check on Canvas (website or app) on a regular basis to stay updated on the upcoming tutorial dates.

Week	Week commencing	Module	Assessment
1	27 February	1A	
2	6 March	1A	Ass1 released on 6 March
3	13 March	1B	
4	20 March	1B	
5	27 March	2A	Ass1 due: 27 March
6	3 April	2A	Ass2 released 3 April
Mid-semester break 10 April			
7	17 April	2B	
8	24 April	2B + 2C	
9	1 May	3A	Ass2 due: 1 May
10	8 May	3A	
11	15 May	3B	Ass3 released on 15 May
12	22 May	3B	
13	29 May		
	5 June		Ass3 due: 5 June

Assessment

Assessment includes 3 written assignments worth 30 or 35% each as per table below, and will be made available as per timetable above.

Assessment name	Assessment type	Coverage	Learning Outcomes	Weight
Assignment 1	Assignment	Module 1	1, 3, 5	30%
Assignment 2	Assignment	Module 2	1, 2, 3, 5	35%
Assignment 3	Assignment	Module 1, 2 & 3	1,2,3,4,5,6	35%

Assignments should be submitted via the assignment submission tool on Canvas; if you experience difficulties with this submission method, assignments can be submitted via email.

In general, you are required to submit work typed in Word or similar. We strongly recommend you become familiar with equation typesetting software such as Microsoft's Equation Editor 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. However, 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.

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

Access to the following textbooks is recommended, as these texts contain the further reading material referenced in the Notes:



Stata

Juul S, Frydenberg M. An Introduction to Stata for Health Researchers, 5th ed. Stata Press, 2021. To purchase:
<https://www.stata.com/bookstore/introduction-stata-health-researchers/>

R

Wickham H, Grolemund G. R for Data Science. O'Reilly 2017 (freely available online at <https://r4ds.had.co.nz/>)

Software requirements and assumed knowledge

No previous computing or programming knowledge is assumed for this course.

However, as pointed out previously, access to the following software packages should be organised ahead of the start of the course:

- Stata: version 14 or later. The University of Adelaide provides free licences to their students as do some other Universities. Please check with your Program Coordinator. Should you require to purchase a licence, [SDAS - Australia sales@surveydesign.com.au](mailto:sales@surveydesign.com.au)
- R : [The Comprehensive R Archive Network \(r-project.org\)](http://TheComprehensiveRArchiveNetwork.com)
- RStudio IDE: [RStudio Desktop - Posit](https://posit.co/)

This is a practical unit designed to develop computing and programming skills in Stata and R; delays in gaining access to the software may impact your ability to complete the unit.

For help with R, also please see Learning R in the Student Resources site on Canvas. If you have not yet organised access to these packages, you should do so as soon as possible. This unit 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

Mathematical proficiency at pre-uni level (basic algebra and statistics, such as familiarity with percentiles, IQR, standard deviation, mean and median).

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 Blackboard

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

DMC was last delivered in Semester 2 2022. References for the sections in the new edition of the Stata textbook were updated. Otherwise, no major unit changes were implemented.

Acknowledgments

In Semester 1, 2020, the R notes were redeveloped by Dr Jennie Louise at The University of Adelaide, and further changes were implemented including incorporation of online tutorials and video content. Further changes were implemented in Semester 1, 2022, including additional modification of the course notes.