BIO
STATISTICS
COLLABORATION
OF AUSTRALIA

# Study Guide

## Regression Modelling for Biostatistics 2 (RM2)

Semester 2, 2024

Prepared by:

Prof Gillian Heller and Dr Ken Beath
NHMRC Clinical Trials Centre
University of Sydney

Dr Michael Waller
School of Public Health
University of Queensland

## Contents

# Regression Modelling for Biostatistics 2 (RM2)

**Semester 2, 2024**

## Contact details

| Professor Gillian Heller | Dr Ken Beath |
|---|---|
| NHMRC Clinical Trials Centre<br>University of Sydney | NHMRC Clinical Trials Centre<br>University of Sydney |
| (02) 8036 5250 | |
| gillian.heller@sydney.edu.au | ken.beath@sydney.edu.au |

If you have any general BCA queries, please contact: Jaqe Vaughan at the BCA Coordinating Office on 02 9562 5076/54 or via email bca@sydney.edu.au

## Background

This unit builds on the material taught in Regression Modelling for Biostatistics 1 and covers generalized linear models and survival analysis techniques.

The aim of this unit is to enable students to implement generalized linear models (GLMs) for analysis of non-normal data, and survival analysis methods for time-to-event data, with proper attention to the underlying assumptions. A major focus is on selection of appropriate methods, assessing the model fit and diagnostics of GLMs and survival models, and the practical interpretation and communication of model results.

This unit is the final core taught subject in the BCA Masters program.

## Context within the program

This unit is the final core taught subject in the BCA Masters program.

## Prerequisites

Epidemiology (EPI), Mathematical Foundations for Biostatistics (MFB) (or Mathematical Background for Biostatistics (MBB) and Probability and Distribution Theory (PDT)), Principles of Statistical Inference, Regression Modelling for Biostatistics 1 (RM1) (or Linear Models (LMR)).

## Unit summary

This unit presents the theory and application of generalized linear models (GLMs) and survival analysis. The unit covers the implementation of GLMs to analyse count data using Poisson and negative binomial regression; how logistic regression models can be applied to binary, multinomial, and ordinal data; and the use of GLMs with continuous data. The unit presents methods to analyse time to event survival data including the Kaplan Meier curve and the Cox proportional hazards regression model.

## Workload requirements

The expected workload for this unit is 10-12 hours per week on average, consisting of guided readings, discussion posts, independent study, module exercises, videos, tutorials and completion of assessment tasks.

## Learning Outcomes

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

1. Explain the theory of GLMs and statistical inference based on GLMs
2. Analyse data using logistic regression models for binary, multinomial and ordinal categorical data
3. Analyse count and rate data using Poisson regression, Negative Binomial, and continuous data using GLMs
4. Explain the nature of survival data and summarise and display survival data using nonparametric methods, including the Kaplan-Meier curve
5. Analyse survival data using the Cox proportional hazards model, including time dependent covariates and the stratified Cox model
6. To assess and evaluate the model fit and diagnostics of GLMs and survival models
7. Synthesise results of analyses to present and communicate findings

## Unit content

The unit is divided into 5 modules, summarised in more detail below. Each module will involve approximately 2 to 3 weeks of study and includes the following material:

1. Module notes describing concepts and methods, and including examples and exercises involving data analysis.
2. Selected readings from published articles or textbooks.
3. Videos of analysis examples and tutorials where methods are demonstrated in statistical software.

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 will be posted online at the midway point of the allocated time for the module. This is intended to encourage you to attempt the exercises independently, and yet not make you wait too long to see the sketch solutions.

Videos of the live Tutorials will also be posted on the eLearning site after each Tutorial.

## Method of communication with coordinators

We strongly recommend that you post content-related questions to the Discussions tool in the RM2 area of BCA's eLearning site, which is the Canvas Learning Management system hosted by the University of Sydney. You will 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.

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!).

The course notes are available on the BCA eLearning site, in both html and pdf formats, along with the datasets for exercises and assignments. However some readings may not be available on the BCA eLearning site hence these may be emailed to you.

We would like to encourage the use of the discussion board facilities on the eLearning site, in order to reduce the isolation of studying by distance. Firstly, you will see a 'Student Introductions' forum on the discussion board.

When you log in to the eLearning site, you will see under 'Discussions' various forum headings. We will include some general discussion points in each module to encourage discussion amongst the group, but would like you to discuss matters and help each other as much as you can. We encourage discussion about the course material, as long as assignment answers are not given.

## Module descriptions

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

**Module 1** 29 July – 11 Aug

The Exponential Family of Distribution and Generalized Linear Models. Maximum Likelihood Estimation for GLMs. Inference for GLMs, including Likelihood Ratio test, Wald statistic and the Deviance. Checking the model assumptions and assessing the goodness of fit of GLMs. Selection of distribution and choice of link function for a GLM.

**Module 2** 12 Aug – 1 Sep

AIC and BIC statistics. GLMs for continuous outcome data. Analysis of count and rate data using Poisson regression and Negative Binomial models. Logistic regression models for binary, multinomial and ordinal categorical data

**Module 3** 2 – 22 Sep

Life tables. The nature of survival data, including censoring; the survival function: definition and estimation via the Kaplan-Meier curve; Kaplan-Meier estimate of the survival function: confidence intervals and hypothesis testing. the stset command in Stata; Surv function in R; The density, survival, hazard and cumulative hazard functions; the Nelson-Aalen estimate of the cumulative hazard function; Definition of the proportional hazards model; construction of the partial likelihood for the Cox model.

**Module 4** 30 Sep – 13 Oct

Hypothesis testing on the coefficients of the Cox model; estimation of the baseline functions S0(t) and H0(t), and their adjustment for covariate values; the effect of a change in scale and origin of units of measurement of covariates. Model diagnostics for the Cox PH model;

**Module 5** 14– 27 Oct

Time-dependent covariates in the Cox model; Stratified Cox Model. Parametric survival time models; discrete-time logistic model. Sample size for survival.

## Unit schedule

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.

Semester 2, 2024 starts on Monday 29th July.

| Week | Week Commencing | Module | Assessment |
|---|---|---|---|
| 1 | Monday 29th July | 1 | |
| 2 | Monday 5th August | 1 | |
| 3 | Monday 12th August | 2 | |
| 4 | Monday 19th August | 2 | Assignment 1 available Friday 23rd August |
| 5 | Monday 26th August | 2 | |
| 6 | Monday 2nd September | 3 | |
| 7 | Monday 9th September | 3 | **Assignment 1 due Monday 9th September**<br>Assignment 2 available Friday 13th September |
| 8 | Monday 16th September | 3 | |
| Monday 23rd – Sunday 29th September – Mid Semester Break | | | |
| 9 | Monday 30th September | 4 | |
| 10 | Monday 7th October | 4 | **Assignment 2 due Tuesday 8th October** |
| 11 | Monday 14th October | 5 | Assignment 3 available Friday 18th October |
| 12 | Monday 21st October | 5 | |
| | | | **Assignment 3 due Friday 8th November** |

## Assessment

Assessment will consist of two written assignments worth 30% each, and one written and presentation assignment worth 40%. Each assignment is to be completed within approximately 2.5 weeks. These assignments will be posted on the eLearning site together with an online Announcement broadcasting their availability.

| Assessment name | Assessment type | Coverage | Learning objectives | Weight |
|---|---|---|---|---|
| Assessment 1 | Assignment | Modules 1 and 2 | 1,2,3,6 | 30% |
| Assessment 2 | Assignment | Module 3 | 1,2,3,4,5, | 30% |
| Assessment 3 | Assignment and presentation | Modules 1, 2, 3, 4, 5 | 1,2,3,4,5,6,7 | 40% |

In general, you are required to submit work typed in Word or similar. We recommend you become familiar with equation typesetting software such as Microsoft's Equation Editor for algebraic work; if you are an R user we strongly encourage the use of RMarkdown, or alternatively knitr if you are also a LaTex user. 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 their own, with anything derived from other students' discussion contributions clearly attributed to the source.

## Submission of assessments 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.

## Use of ChatGPT and other generative AI tools in assessment tasks

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 AI tools (e.g., ChatGPT), successful completion of these components will require students to critically engage in specific contexts and tasks for which artificial intelligence 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 of assessments 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

The primary textbook for the unit is.

Vittinghoff E, Glidden D, Shiboski S, McCulloch C. *Regression Methods in Biostatistics: Linear, logistic, survival and repeated measures models*. 2nd Edition. Springer Verlag 2012

Other recommended resources will be provided throughout the course.

## Software requirements and assumed knowledge

For this subject you will need to have access to R or Stata. Code and output in the module notes are given in both R and Stata, and students may choose to work in either software language.

Stata 17 was released in April 2021, and we assume you are using this version. However, some of you may be using Stata 15 or Stata 16. We are not aware of any major differences between Stata versions that affect the material, but minor issues will be pointed out in eLearning postings.

The most recent versions of R and RStudio are available to download from
https://www.r-project.org/
https://www.rstudio.com/products/rstudio/download/

For help with R, please see Learning R in the Student Resources site.

For help with Stata please use the help functions within Stata.
Please post on the Discussion board with questions regarding R and Stata.

## Required mathematical background

Students who have undertaken the pre-requisites will have the required mathematical background for the course. This unit is practical in nature and is focused on the application of GLM and survival models on datasets using statistical software, with proper attention to the underlying assumptions.

## Feedback

*Our feedback to you:*

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

- Formal individual feedback on submitted 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

The only change this semester is the method for time-dependent covariates in R has been changed to make it closer to the Stata method, and to allow more complicated time-dependent covariates.

## Acknowledgments

Prof Gillian Heller (USYD CTC) and Dr Ken Beath are co-ordinators of this unit in this semester. This version of RM2 was developed by Dr Michael Waller and Prof Gillian Heller.

Michael Waller and Gillian Heller would like to acknowledge Dr Ken Beath (University of Sydney, prev. Macquarie University), A/Prof Mark Jones and Prof Annette Dobson (both UQ) who have developed earlier BCA units and other materials which have been included as part of this delivery.