BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Regression Modelling for Biostatistics 2 (RM2)

Semester 1, 2022

Prepared by:

Dr Michael Waller
School of Public Health
University of Queensland

Prof Gillian Heller
NHMRC Clinical Trials Centre
University of Sydney

THE UNIVERSITY OF QUEENSLAND

THE UNIVERSITY OF SYDNEY

Contents

Regression Modelling for Biostatistics 2 (RM2)

Semester 1, 2022

Contact details

Dr Michael Waller

School of Public Health,
University of Queensland

(07) 3365 5116

m.waller@uq.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 via email to 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 categorical 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.

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.

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)).

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, Zero-Inflated models, 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 attack the exercises independently (or via the eLearning site), 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 coordinator(s)

We strongly recommend that you post content-related questions to the Discussions tool in the RM2 area of BCA's eLearning site. In 2022 we are using the Canvas Learning Management system 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 "Getting Started in Canvas" document available on the Student Resources page of the BCA website.

Questions about administrative aspects or course content can be emailed to the coordinator, and when doing so please use "(RM2):" 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!).

At the start of semester we will send a welcome email. The course notes are available on the BCA eLearning site, along with the data sets 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 try and 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. February 28 – March 13

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. March 14 – April 3

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

Module 3. April 4 – May 1

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. May 2 – May 15

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. May 16 – May 29

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

Semester 1, 2022 starts on Monday 28 Feb

Week	Week Commencing	Module	Assessment
1	Monday 28th February	1	
2	Monday 7th March	1	
3	Monday 14th March	2	
4	Monday 21th March	2	Assignment 1 Available Friday 25th March
5	Monday 28th March	2	
6	Monday 4th April	3	
7	Monday 11th April	3	**Assignment 1 Due Monday 11th April**
Monday 18th April to Sunday 24th April – Mid Semester Break			
			Assignment 2 Available Friday 22nd April
8	Monday 25th April	3	
9	Monday 2nd May	4	
10	Monday 9th May	4	**Assignment 2 Due Monday 9th May**
11	Monday 16th May	5	Assignment 3 Available Friday 20th May
12	Monday 23rd May	5	
			Assignment 3 Due Friday 10th June

Assessment

Assessment will include 2 written assignments worth 30% each, and one presentation assessment worth 40%. Each assessment 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	Module 1 and 2	1,2,3,6	30%
Assessment 2	Assignment	Module 3	1,2,3,4,5,	30%
Assessment 3	Assignment and presentation	Module 1, 2, 3, 4, 5	1,2,3,4,5,6,7	40%

Assessment 3 will include a presentation delivery. The planned dates for these presentation deliveies are between 6th June and 10th June.

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; alternatively if you are an R user we encourage the use of RMarkdown. 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 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.

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 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 text book 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 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 13-17. 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.

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 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

This the first delivery of RM2.

Acknowledgments

Prof Gillian Heller (USYD CTC) is a co-coordinator of this unit. This version of RM2 was developed by Dr Michael Waller (UQ) and Prof Gillian Heller (USYD CTC).

We would like to acknowledge Prof Gillian Heller (USYD CTC, MACQ), Dr Ken Beath (MACQ), 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.