BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Categorical Data & Generalised Linear Model (CDA)

Semester 2, 2022

Prepared by:

Dr Thomas Fung
School of Mathematical and Physical Sciences,
Faculty of Science and Engineering
Macquarie University

Contents

Background 2
Unit summary 2
Workload requirements 3
Prerequisites 3
Co-requisites 3
Learning Outcomes 3
Unit content 3
Recommended approaches to study 4
Method of communication with coordinator(s) 4
Unit schedule 5
Assessment 6
Submission of assessments and academic honesty policy 6
Late submission of assessments and extension procedure 7
Learning resources 7
Software 7
Feedback 7
Required mathematical background 8
Changes to SVA since last delivery, including changes in response to student evaluation 8

Generalized Linear Models

Semester 2, 2022

Instructor contact details

Dr Thomas Fung	**Prof Benoit Liquet**
Room 626 12 Wally's Walk School of Mathematical & Physical Sciences Macquarie University	Room 630 12 Wally's Walk School of Mathematical & Physical Sciences Macquarie University
e-mail: thomas.fung@mq.edu.au	e-mail: benoit.liquet-weiland@mq.edu.au

Background

This unit, "Categorical Data Analysis and Generalized Linear Models" (CDA), is about statistical methods for analysing data when the response or outcome variable is non-normal.

Methods for contingency tables have a long history but are often somewhat ad hoc. Most methods for analysing categorical data, however, are special cases of Generalized Linear Models (GLMs). These include modelling count data (e.g., using Poisson regression); binary data (using logistic regression); data in more than two nominal categories (nominal or multinomial regression); or more than two ordered categories (ordinal logistic regression). GLMs provide a unifying framework that you will meet again in other units such as SVA and LCD.

There is an emphasis on the practical interpretation and communication of results to colleagues and clients who may not be statisticians.

Unit summary

This unit starts with the classical normal linear regression model. The family of generalized linear models is then introduced, and maximum likelihood estimators are derived. Models for counted responses, binary responses, continuous non-normal responses and categorical responses; and models for correlated responses, both normal and non-normal, and generalised additive models, are studied. All models and methods are illustrated using data sets from disciplines such as biology, actuarial studies and medicine.

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.

Prerequisites

Epidemiology (EPI), Mathematical Background for Biostatistics (MBB), Probability and Distribution Theory (PDT), Principles of Statistical Inference (PSI).

Co-requisites

Linear Models (LMR)

Learning Outcomes

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

1. Formulate a generalized linear model and derive its maximum likelihood estimators.
2. Answer research questions by exploring data graphically; selecting and applying appropriate modelling techniques; appraising underlying model assumptions and goodness of fit and modifying the analysis if required.
3. Perform model selection and test hypothesis.
4. Apply the generalized additive model to incorporate nonlinear forms of the predictors and use random effects or generalized estimating equations to model correlated data.
5. Use statistical software to create model output and interpret them.

Unit content

The unit is divided into 12 weeks of content, summarised in more detail below. Each week's content generally includes the following material:

1. "Lecture Notes" describe concepts and methods, and possibly including some examples.
2. Sometime, "Selected Readings" from published articles or textbooks will also be included.
3. One or more "Extended Examples" as part of the small group teaching activities exercises, illustrating the concepts/methods introduced in the notes and including more practically oriented exercises.

Study materials are downloadable from the eLearning unit site. Assignments and supplementary material, such as datasets will be posted to the unit site. Please note that we may not be able to post copies of copyright material (for example journal articles and book extracts)—for these you will have to rely on resources from 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 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.

Solutions to the exercises will be released at the middle of the following week (except those to be submitted for assessment, as described below). 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.

Method of communication with coordinator(s)

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

We strongly recommend that you post content-related questions to the Discussions tool in the eLearning site. In 2022 we are using the Learning Management system hosted by the University of Sydney and Macquarie University. You should already be familiar with the university student learning system from previous units.

Unit schedule

Semester 2, 2022 starts on Monday 25th July

Week	Week commencing	Topic	Assessment
1	25 July	The classical normal linear model	
2	1 August	Introduction to GLMs: The framework of generalized linear models is introduced, and the theory behind maximum likelihood estimation of the parameters started.	Assignment 1 available
3	8 August	Maximum likelihood estimation of the parameters; Poisson regression for count data	
4	15 August	Inference; comparison of models; deviance as a measure of fit; hypothesis testing	Assignment 1 due
5	22 August	Model checking: Definition of residuals in GLMs; checking for violation of model assumptions	
6	29 August	Model selection; overdispersion: Selection of models via AIC; the phenomenon of overdispersion; compound Poisson models to overcome it; the negative binomial model for counts	
7	5 September	Binary responses: logistic regression	Assignment 2 available
	12 - 25 September	Mid semester break	
8	25 Apr	Logistic regression contd; Zero-inflated models; Generalized additive models	
9	26	Regression models for ordinal and categorical responses	Assignment 2 due
10	9 May	Correlated data: Models for longitudinal data, and other data structures in which there is clustering or correlation between observations	Assignment 3 available
11	16 May	Correlated data	
12	23 May	Correlated data	
13	30 May	No Lecture	Assignment 3 due

Assessment

Assessment will be 3 written assignments worth 30%, 40% and 30% each, to be made available approximately at the time indicated in the Unit Schedule, and will be given at least 2 weeks to complete. 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
Assignment 1	Assignment	Weeks 1-2	1, 2, 3, 5	30%
Assignment 2	Assignment	Weeks 3-8	1, 2, 3, 5	40%
Assignment 3	Assignment	Weeks 8-12	1, 2, 3, 4, 5	30%

In general, you are required to submit your work typed in Word or similar (e.g. using Microsoft's Equation Editor for algebraic work) and part of the assessment is to make your work reproducible using RMarkdown. If extensive algebraic work is involved 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. This handwritten work should be scanned and collated into a single pdf file and submitted via the eLearning site. See the BCA Assessment Guide document for specific guidelines on acceptable standards for assessable work.

The instructors will generally avoid answering questions relating directly to the assessable material until after it has been submitted, but we encourage students to discuss the relevant parts of the notes among themselves, via eLearning. However **explicit solutions to assessable exercises should not be posted for others to use**, and 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

You should submit all your assessment material via eLearning unless otherwise advised. The use of Turnitin for submitting assessment items has been instigated within unit sites. For more detail please see pages 3-5 the BCA Student Assessment Guide.

The BCA pays great attention to academic honesty procedures. Please be sure to familiarise yourself with these procedures and policies at your university of enrolment. Links to these are available in the BCA Student Assessment Guide. Please also read carefully the Academic Honesty document in web page of this unit. When submitting assessments using Turnitin you will need to indicate your compliance with the plagiarism guidelines and policy at your university of enrolment before making the submission.

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 is not able to approve extensions beyond three days; for extensions beyond three days you need to apply to your home university, using their standard procedures.

Learning resources

There is no prescribed text for this unit. The following are useful references:

1. Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013). **Regression: Models, Methods and Applications**, Springer.
2. Faraway, J. J. (2016). **Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models**. CRC Press.
3. De Jong, P. and Heller, G.Z. (2008). **Generalized Linear Models for Insurance Data**, Cambridge University Press.
4. Wood, Simon N. (2017). **Generalized additive models: an introduction with R,** 2nd edition**.** CRC Press.
5. Stasinopoulos M. D., Rigby R. A., Heller G. Z., Voudouris V., De Bastiani F. (2017). **Flexible Regression and Smoothing: Using GAMLSS in R**. CRC Press.
6. Dobson, A. J. and Barnett, A. G. (2018). **An Introduction to Generalized Linear Models**, 4th edition, Chapman & Hall.
7. Lindsey, J.K. (1997). **Applying Generalized Linear Models**, Springer.
8. McCullagh, P. and Nelder, J.A. (1989). **Generalized Linear Models**, 2nd edition, Chapman & Hall.

Software

It is expected that you will be using R which is freely downloadable from the CRAN website. We recommend the use of the RStudio interface, also freely downloadable.

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 Blackboard

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.

Required mathematical background

Simple differentiation, integration and matrix algebra.

Changes to CDA since last delivery, including changes in response to student evaluation

CDA is now delivered by Macquarie University and this version is equivalent to their STAT8111.