BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Categorical Data Analysis and Generalized Linear Models (CDA)

Semester 2, 2020

Prepared by:
Prof Annette Dobson, Dr Mark Jones, Dr Michael Waller
School of Public Health
The University of Queensland

A/Prof Mark Chatfield (Queensland Institute of Medical Research) revised Stata codes and added do files and outputs for modules 4-6 in 2017

Contents

Categorical Data Analysis and Generalized Linear Models (CDA)

Semester 2, 2020

Instructor contact details

Instructor **Dr Michael Waller**

School of Public Health, Public Health Building
The University of Queensland
Herston Road, Herston, QLD, 4006

(07) 336 55116

m.waller@uq.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 categorical.

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.

Much of the material in CDA is similar to Annette Dobson's and Adrian Barnett's book "An Introduction to Generalized Linear Models" (third edition, Chapman Hall/CRC, 2008). The history of the relationship is that an early version of CDA was derived from an early version of the book but the material was changed over several years specifically for CDA. The revised (3rd edition) of the book was based on the CDA version but with changes. The CDA notes are designed specifically for distance delivery using the BCA model and are independent of and different from the book.

Unit summary

The aim of the unit is to enable you to use generalized linear models and other methods to analyse categorical data with proper attention to the underlying assumptions. There is an emphasis on the practical interpretation and communication of results to colleagues and clients who may not be statisticians.

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

The following BCA units are recommended pre-requisites

MBB: Mathematical Background for Biostatistics
EPI: Epidemiology
PDT: Probability and Distribution Theory
PSI: Principles of Statistical Inference
LMR: Linear Models

Modules 2 and 3 in particular build on material presented in Principles of Statistical Inference (PSI). Some students in previous years have commented that their PSI notes were useful in refreshing their memory on concepts such as Wald, Score and likelihood ratio tests. The statistical foundations that you develop in CDA will be invaluable to you in your career as a statistician and in subsequent BCA units such as Longitudinal and Correlated Data (LCD), Bayesian Statistical Methods (BAY) and Bioinformatics (BIF).

LMR is a recommended pre-requisite but due to timetabling constraints some students may be taking CDA and LMR concurrently. The extent to which this may be a problem depends on each student's prior knowledge and experience of statistical modelling, including multiple regression, analysis of variance and the use of diagnostics. For students who have done LMR you may want to refresh your knowledge of 'strategies for analysis' and the 'vagaries of model building'.

Co-requisites

LMR: Linear Models may be taken as a co-requisite

Learning Outcomes

On completion of this unit you should:

1. be able to explain and use standard methods for analysing data in contingency tables, including matched and stratified data;
2. understand the theory of GLMs and statistical inference based on GLMs for categorical data;
3. use correctly logistic regression models for binary, multinomial and ordinal categorical data
4. analyse correctly count data using Poisson regression.

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.

Study materials for all Modules are downloadable from the eLearning unit site. Assignments and supplementary material, such as datasets will be posted to the unit site.

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.

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 eLearning site), and yet not make you wait too long to see the sketch solutions.

Method of communication with coordinator(s)

We strongly recommend that you post content-related questions to the Discussions tool in the CDA area of BCA's eLearning site. In 2020 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" 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 "(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!).

At the start of semester we will send a welcome email and ask if you wish to receive a hard copy of the unit materials. If you respond positively to this question the unit materials will be posted to you, with your copy of this guide. The course notes are also available on the BCA eLearning site, along with the data sets for exercises and assignments. However the 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

Each module is scheduled to begin on a Monday and conclude on the Sunday of the following week.

Module 1. August 3 – August 16

Introduction to and revision of conventional methods for contingency tables especially in epidemiology: odds ratios and relative risks, chi-squared tests for independence, Mantel-Haenszel methods for stratified tables, and methods for paired data.

Module 2. August 17 – August 30

The exponential family of distributions; generalized linear models (GLMs), and parameter estimation for GLMs

Module 3. August 31 – September 13

Inference for GLMs – including the use of score, Wald and deviance statistics for confidence intervals and hypothesis tests, and residuals.

Module 4. September 14 – September 27

Binary variables and logistic regression models – including methods for assessing model adequacy.

Module 5. October 5 – October 18

Nominal and ordinal logistic regression for categorical response variables with more than two categories.

Module 6. October 19 – November 1

Count data and Poisson regression

Unit schedule

Semester 2, 2020 starts on Monday 3 August

Week	Week Commencing	Module	Assessment
1	Monday 3rd August	1	
2	Monday 10th August	1	Assignment 1 Available Friday 14th August
3	Monday 17th August	2	
4	Monday 24th August	2	
5	Monday 31th August	3	**Assignment 1 Due Thursday 3rd September**
6	Monday 7th September	3	Assignment 2 Available Friday 11th September
7	Monday 14th September	4	
8	Monday 21st September	4	
			Assignment 2 Due Thursday 1st October
Monday 28th September to Sunday 4th October – Mid Semester Break			
9	Monday 5th October	5	
10	Monday 12th October	5	
11	Monday 19th October	6	Assignment 3 Available Wednesday 21st October
12	Monday 26th October	6	
			Assignment 3 Due Monday 9th November

Assessment

Assessment will include 3 written assignments worth 30%, 35% and 35% each, and to be completed within approximately 3 weeks. These assignments will be posted on the eLearning site together with an online Announcement broadcasting their availability.

- Assignment 1 will cover Module 1 and 2, and is worth 30% of the overall course mark. It is due before 2pm (EST) on Thursday 3rd September 2020.

- Assignment 2 will cover Module 3 and 4, and is worth 35% of the overall course mark. It is due before 2pm (EST) on Thursday 1st October 2020.
- Assignment 3 will cover all Modules, and is worth 35% of the overall course mark. It is due before 2pm (EST) on Monday 9th November 2020.

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. 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

Reference books:

Agresti A. “An Introduction to Categorical Data Analysis”, Wiley InterScience, 1996, ISBN 0-471-11338-7.

Agresti A. “Categorical Data Analysis” (second edition), Wiley, 2002, ISBN 0-471-36093-7

Agresti A. “Analysis of Ordinal Categorical Data”, Wiley, 1984

Dobson AJ and Barnett AG. “An Introduction to Generalized Linear Models” (third edition), published Chapman Hall / CRC in 2008, ISBN 978-1-58488-950-2.

Hilbe JM. “Logistic Regression Models”, Chapman & Hall/CRC Press, 2010

Kirkwood BR, Sterne JAC. "Essential Medical Statistics" (second edition) Blackwell, 2003, ISBN 0-86542-871-9.

Le CT. "Applied Categorical Data Analysis", Wiley, 1998.

Woodward M. "Epidemiology: Study Design and Data Analysis" (second edition), published Chapman Hall / CRC in 2005, ISBN 978-1-58488-415-6.

Hardin JW and Hilbe JM. "Generalized Linear Models and Extensions" (second edition), published Stata Press, 20 Feb 2007, ISBN 1597180149, 9781597180146.

Software

You will need to use statistical software for the exercises and assignments. Stata is the default for this unit.

Hilbe's book has detailed R commands corresponding to most of the Stata commands used in the book. Woodward's book and the supplementary materials on the web include examples using SAS and Stata. R code for many of the examples and exercises in Modules 2-6 is given in the book by Dobson and Barnett. Agresti's book includes an appendix about SAS (and some other software) commands for methods covered in CDA. For some exercises Excel may be a suitable tool. However, you may use whatever you like.

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.

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

The main issues for CDA that have been identified by previous students and the BCA peer review process are the jumps in concepts and methods between Module 1 and Modules 2-3 and again for Modules 4-6 – but at the end it does come together. BCA peer reviewers tend not to like the inclusion of Module 1 but students do like it and we have tried to connect it to the other modules whenever we can.

We did a major revision in 2012 where we changed the textbook used for module 1 so that module 1 would better connect with the material presented in module 2. We also edited module 3 to hopefully make it clearer and have removed non-essential material showing the derivation of the sampling distributions for various statistics presented. We created power-point slides with audio to enhance learning. You will get access to the data set to enable you to run your own analyses. Our plan was to create additional videos for the more mathematical material presented in modules 2 and 3. However on searching the internet we found many relevant online videos which provide good explanations of the concepts. Hence we have collated a list of recommended online videos for students to access to enhance their understanding.

In 2017 we made additional changes based on feedback from students the previous year. They included revising goodness-of-fit measures in Module 5 and providing more detailed exercise solutions and updating Stata codes in the examples.

In 2018 and 2019 online tutorials were also developed for study modules.