BIO STATISTICS COLLABORATION OF AUSTRALIA

Unit of Study Notes

Categorical Data Analysis and Generalized Linear Models (CDA)

Semester 2, 2017

Prepared by,
Prof Annette Dobson and Dr Mark Jones
School of Population Health
The University of Queensland

THE UNIVERSITY OF QUEENSLAND

Contact details

Welcome to CDA. The unit co-ordinator and UQ program co-ordinator is **Mark Jones**. Mark Jones will be communicating with you throughout the unit and will be marking the assessments.

Our contact details are as follows:

Dr Mark Jones

School of Population Health, Public Health Building
University of Queensland
Herston Road, Herston, QLD 4006
E-mail: m.jones@sph.uq.edu.au
Office phone: 07-3365 5116; Fax: 07-3365 5540

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.

Aim of the Unit

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.

Objectives

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.

Assumed knowledge

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'.

Overview

The unit is organised into six Modules each taking 2 weeks. The modules are in three distinct groups.

Module 1 is a refresher of the disparate methods for analysing categorical data that you have encountered previously in introductory statistics and epidemiology units. We think it is important to revise this material so you are better able to link it to the approach presented in CDA. As there are many excellent textbooks on these topics we have used excerpts from one of these for this module: Agresti's book "*Categorical Data Analysis*" (second edition, John Wiley & Sons, Inc. 2002).

Modules 2 and 3 are very different. They establish the statistical foundations for a unified approach to modelling categorical (and other forms) of data, namely generalized linear models (GLMs). These modules rely heavily on PSI (and hence MBB and PDT).

Beware: Some students experience shock at the gear change between Module 1 and Modules 2 and 3 – do not panic!

Modules 4-6 bring it all together. They work through specific GLMs needed for the types of problems introduced in Module 1 but in a unified way that also links closely to LMR.

Contents

Module 1. July 31 – August 13

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 14 – August 27

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

Module 3. August 28 – September 10

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

Module 4. September 11 – September 24

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

Module 5. October 2 – October 15

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

Module 6. October 16 – October 29

Count data and Poisson regression

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.

Timetable

Week beginning Monday	Module	Co-ordinator	Assignment to be submitted Monday
31 July	Module_1	Mark Jones	
7 August	"		
14 August	Module_2	"	
21 August	"		
29 August	Module_3	"	
4 September	"		
11 September	Module_4	"	Assignment 1
18 September	"		
25 September	Break		
2 October	Module_5	"	Assignment 2
9 October			
16 October	Module_6	"	
23 October	"		
30 October	Study		Assignment 3

Method of Delivery & Communication

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. You can add your own information to this forum, if you wish, so that others in the course can contact you. For example:

Jonathan Bloggs

j.bloggs@ctc.edu.au

ph: 02-9999-9999

NHMRC Clinical Trials Centre, Sydney

Jonathan is a trainee biostatistician at the Clinical Trials Centre. He is currently working with trials of new medications for diabetes and heart disease.

This is entirely optional. If you would like to be part of the forum, but without your contact details, that will be fine as well.

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. Some students in the past have said they haven't used the discussion board as much as they would have liked, as they didn't want to be seen to be colluding in the preparation of assignments. We encourage discussion about the course material, and assignments, as long as worked answers are not given.

Assessment

The assessment is based entirely on assignments. There is no examination and no marks awarded for online discussions. There are two assignments each worth 35% of the marks and one assignment worth 30%. These will involve analysing real data sets. They will give you scope to demonstrate insight and flair!

The due dates for the assessment items are shown in the Timetable on page 7. They are due in on Mondays.

Assignments will be posted on the eLearning site.

- Write your assignment in any style (e.g. journal, technical report) but make sure that the layout is clear and that all the questions are answered.
- Marks will be allocated for presentation.
- Do not be afraid to use long, but descriptive and specific, headings or sub-headings (e.g. “Methods for assessing statistical interactions” instead of “Methods”).
- Remember to define any acronyms you use, and briefly explain any new terminology or assumptions.
- Marks will be lost (from the style section) for assignments that are too long or include irrelevant material that indicates that you did not understand the question.
- Raw computer output is not acceptable.

The following two documents available on the BCA website as resources for current students may be helpful:

Guide for Reporting Statistical Results
Referencing Style Guide
They are available at www.bca.edu.au/currentstudents.html

Before commencing the course, you should read the BCA assessment guide (Appendix), and the information about the plagiarism policy of your home university.

Assessment deadlines are important.

Extensions or late submissions policy
Requests for an extension an assignment must be made in advance of the due date. Requests must be made directly to the module coordinator by email. The module coordinator will reply with the decision as to whether an extension has been granted and the new due date.

Extensions can cause delays in feedback for other students who submitted on time. Also due to prerequisites, late results may preclude you from studying subsequent units. Different universities have different result submission deadlines. BCA results have to be transmitted between universities, which shortens the available time.

Feedback

Your Assignments will be returned to you via the eLearning site.

Outline answers for exercises in each Module will be posted on the eLearning site after students have had a chance to attempt the exercises without access to the solutions.

Model answers for Assignments are not really appropriate as there is hardly ever a unique best solution. With permission of the students concerned I would like to adopt a system used in CDA in previous years. This involves posting on the eLearning site for each Assignment the work of two student assignments who received high marks.

Complaints policy

Please see the BCA complaints policy in the Assessment Guide and in online assessment submission pages.

Summary of changes to materials and/or procedures since last delivery

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.

Last year we made additional changes based on feedback from students the previous year. They include moving some of the more mathematical material in module 2 to the appendices as well as revising and adding another example to module 3. We plan to introduce new assignments for 2017, revise goodness-of-fit measures in Module 5, and provide more detailed exercise solutions and update Stata codes in the examples.

Appendix – BCA Assessment Guide

Can be downloaded from:
http://www.bca.edu.au/linked%20docs/Student%20resources/BCA_assessment_guide_student.pdf

Please note that any previous instructions for the 'own work' declarations are now redundant as assignments will be submitted via Turnitin.