BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Linear Models (LMR)

Semester 1, 2021

Prepared by:

Dr Timothy Schlub
Sydney School of Public Health
Faculty of Medicine and Health
University of Sydney

THE UNIVERSITY OF SYDNEY

Contents

Linear Models (LMR)

Semester 2, 2020

Instructor contact details

Dr. Timothy Schlub

Sydney School of Public Health
University of Sydney
(02) 9351 5992

tim.schlub@sydney.edu.au

Background

To be an effective practitioner of biostatistics it is vital to have a solid understanding of the theory and methods of linear models. The "R" in the subject codename LMR stands for "regression" analysis, which is another term for the methods of linear modelling. Although the term "linear" implies that we will be concerned with relationships that can be represented as straight lines, the methods actually cover a much broader range of relationships. This unit deals only with models for outcome variables that are continuously distributed. Such models are sometimes called "normal linear models" because statistical inference for them relies (to some extent) on normal distribution assumptions.

Unit summary

This unit will focus on developing, validating and interpreting multivariable linear models. We aim in this subject to provide a balance between theory and practice: mathematical proofs are not emphasised but sufficient mathematics is used to establish a solid grounding in the main concepts and to enable students to build on the basic material covered here. This subject provides core prerequisite knowledge in statistical modelling, which is built upon in other BCA modelling units such as Categorical Data Analysis (CDA) and Survival Analysis (SVA).

Many courses on regression and linear models emphasise the technical aspects of fitting and testing models. In practice, the hardest challenges facing an applied statistician relate to issues of how to construct and interpret appropriate models in such a way that you (the statistician) can help provide reasonable answers to empirical research questions. While technical material is generally easier to teach we place as much emphasis as possible on these less tangible issues, which are not any easier than the more technical material just because they involve less mathematics. Through the class discussions, we hope to reinforce the message that good applied statistical work requires a lot of judgment, and decisions that are not necessarily either right or wrong. After all, statistics is essentially the art of handling uncertainty!

Workload requirements

In 2017, LMR students reported that they spent a median time of 11 hours per week on this unit (min = 4hrs, interquartile range = 8hrs to 12hrs, max = 28hrs). The expected workload consists of guided readings, discussion posts, independent study, tutorial participation and completion of assessment tasks.

Prerequisites

Epidemiology (EPI), Mathematical Background for Biostatistics (MBB), Probability and Distribution Theory (PDT)

Co-requisites (must be taken before or concurrently with LMR)

Principals of Statistical Inference (PSI)

Learning Outcomes

At the completion of this unit the student will:

1. Have a sound understanding of the normal linear model including a theoretical grounding in the principles of least squares and likelihood-based estimation and related statistical inference, to the level of being able to manipulate equations required for deriving formulas for estimates and their standard errors for the standard models.
2. Understand the principles and practice of model checking and diagnostics, and the use of transformations, in particular the log transformation, to improve model fit; understand the appropriate use of analysis of covariance to adjust for confounding; have a good working knowledge of the theory and practice of multiple regression analysis; be familiar with the method of analysis of variance (up to 2 factor models) and its relationship to multiple regression; gain an introductory understanding of nonparametric smoothing for flexible regression modelling, and of the use of variance components and random effects models.
3. Have a strong grasp of practical issues involved in fitting linear models, including the ability to construct defensible models (use of dummy variables, choice of parameterisation, interaction and transformation of variables); demonstrate ability to fit models using modern statistical software and to interpret fitted models in terms that are useful to non-statisticians.

Unit content

The unit is divided into 6 modules, summarised in more detail below. Each module involves about 2 weeks of study (although this varies a little between modules) and generally includes the following materials:

1. Module notes describing concepts and methods, and including computational exercises and some exercises of a more “theoretical” nature.
2. A case study illustrating the concepts/methods introduced in the notes and including more practically oriented exercises.

Students should begin each module by reading through the module notes, 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. Outline solutions to these exercises will be provided online during the course of the allotted period allocated to each module. You should also work through all of the computing (Stata) examples in the notes for yourself on your own computer.

Case study should be worked through in parallel with its exercises. One or more exercises from some of the case studies will be required to be submitted by the due date (see assessment details and subject timetable).

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. Live online tutorials will be held for each module as per the timetable below. You are expected to prepare for these by reading the module and attempting the questions before the tutorial.

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) 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)

The best method to get in contact with Tim Schlub is by email (for administrative or personal enquiries) or by discussion board (for LMR content related enquiries). If you wish to discuss something over the phone, or in person, it is best to first arrange a mutually suitable time by email.

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

All content-related questions are to be posted to the Discussions tool in the LMR area of BCA's eLearning site. This year we are using the 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.

Module descriptions

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

Module 1

- formulate a simple linear regression model appropriate for a practical problem and interpret its parameters
- understand the least squares method of parameter estimation and its derivation
- use software to obtain estimates of regression model parameters, predicted values and residuals
- understand the basic elements of inference for the regression model parameters and fitted values
- formulate a linear model with a binary independent variable, select a parametrisation and recognise the relationship between the corresponding parameter estimates and tests and those of a two-group t-test
- understand how the outcome variance is partitioned in the analysis of variance table for simple regression

Module 2

- appreciate the role of residuals in diagnostic checking of regression models, including the use of appropriate graphical examinations of residuals
- understand the rationale for the standardisation of residuals, their properties and application
- have a general understanding of the use of nonparametric smoothing techniques to evaluate the shape of a regression function
- recognise when transformations of the response and/or covariate may be warranted and understand possible approaches to handle these, with particular emphasis on log transformations and the interpretation of resulting parameter estimates
- understand the key concepts of outliers and influence in regression models, and be able to implement diagnostic measures and displays to evaluate outlying and/or influential observations

Module 3

- understand and explain the effects of uncontrolled confounding, how it is detected, and the concept of its control by holding extraneous factors constant
- understand and describe the meaning of the effect of one covariate on an outcome variable adjusted for the effect of another covariate

- understand the application of linear regression methods to control for confounding of the effect of a binary risk factor by a binary confounder
- understand the construction, interpretation and checking of the analysis of covariance model for assessing the difference between group means with control of a continuous confounder
- understand and explain the concept of interaction, how it is assessed and how linear models containing interaction terms are interpreted

Module 4

- be familiar with the basic facts of matrix algebra and the way in which they are used in setting up and analysing regression models
- understand the multiple regression model as a generalisation of the simple regression model, and be able to interpret the coefficients of such a model
- understand the principal forms of statistical inference applied to the multiple regression model, and in particular how these relate to partitioning of the total sum of squares
- be familiar with the variety of ways in which multiple regression models are used and constructed in practice, including an understanding of the different purposes of modelling and their implications for model-building strategies
- understand and be able to use graphical tools for building and checking multiple regression models
- be familiar with popular techniques for variable selection in regression models (stepwise selection, all-subsets) and understand the dangers and limitations of using them
- understand the concept of collinearity in multiple regression and some strategies for dealing with it
- have extended their understanding of model diagnostic techniques in the context of multiple regression

Module 5

- be able to construct indicator variables to perform a regression with a categorical covariate and interpret parameter estimates
- extend understanding of the partitioning of the variability of an outcome into different sources in an ANOVA table
- be able to specify and estimate comparisons of interest, and understand the partitioning of the variability associated with a categorical predictor or factor into orthogonal components
- understand the rationale for methods used to control the over-interpretation of multiple comparisons in an analysis of variance, and appreciate some of the competing views on whether formal methods should be used for this purpose
- understand the simple one-way random effects or variance components model

- be familiar with the two-way ANOVA, including interaction effects and the problems of unbalanced data, and gain an introduction to higher-order ANOVA models

Module 6

- understand the concept of regression to the mean and the ways in which this can cause difficulty in the interpretation of data representing changes in an outcome measure
- understand the basic properties of the bivariate normal distribution
- understand the appropriate use of baseline values in the analysis of data from randomised trials, in particular the importance of conditioning on baselines using analysis of covariance, and appreciate some of the issues involved in handling baselines in non-randomised studies

Unit schedule

Week commencing		Module	Live zoom session	Case study 1	Assign. 1	Case study 2	Assign. 2
1	1-Mar	1. Simple linear regression: fundamentals	Introduction Wednesday 7:30pm				
2	8-Mar	1. Simple linear regression: fundamentals	Tutorial Wednesday 7:30pm (1hr)				
3	15-Mar	2. Simple linear regression: further topics	Q&A Monday 11:00am (30 mins)	Released			
4	22-Mar	2. Simple linear regression: further topics	Tutorial Wednesday 7:30pm (1hr)				
5	29-Mar	3. Regression with two predictors	Q&A Monday 11:00am (30 mins)	Due Monday	Released		
	5-Apr	Semester break					
6	12-Apr	3. Regression with two predictors	Tutorial Wednesday 7:30pm (1hr)				
7	19-Apr	4. Multiple regression	Q&A Monday 11:00am (30 mins)	Feedback			
8	26-Apr	4. Multiple regression	Q&A Monday 11:00am (30 mins)		Due Monday	Released	
9	3-May	4. Multiple regression	Tutorial Wednesday 7:30pm (1hr)		Feedback		
10	10-May	5. Analysis of variance	Q&A Monday 11:00am (30 mins)			Peer feedback	
11	17-May	5. Analysis of variance	Tutorial Wednesday 7:30pm (1hr)			Due Monday	Released
12	24-May	6. Analysis of change	Q&A Monday 11:00am (30 mins)				
13	31-May		Tutorial Wednesday 7:30pm (1hr)				
14	7-Jun						Due Monday

Assessment

Assessment will include 2 written assignments worth 30% and 40% respectively, to be made available in the middle and at the end of the semester, 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. In addition, students will be required to submit solutions to two case studies, worth 10% and 20% respectively by deadlines specified throughout the semester (see table below).

Assessment name	Assessment type	Coverage	Learning objectives	Weight	Due on Monday 11:59pm
Case study 1	Assignment	Module 1 - 2	1,2,3	10%	Week 5
Assignment 1	Assignment	Module 1 - 3	1,2,3	30%	Week 8
Case study 2	Oral presentation	Module 1 – 4	1,2,3	20%	Weeks 10 & 11
Assignment 2	Assignment	Modules 1 - 6	1,2,3	40%	Week 14

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 we strongly recommend that you become familiar with equation typesetting software such as this. 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.
This guide will also be included in hardcopy in your package of notes.

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

We adhere to standard BCA policy for late penalties for submitted work, i.e. a 5% deduction from the earned mark for each day the assessment is late, up to a maximum of 50%. 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 the unit, but a number of reference books are recommended as background material (see list below). The module notes and case studies form the primary material for this unit, and readings from selected texts will be referred to.

Weisberg S. *Applied Linear Regression* (3rd ed) New York: Wiley, 2005
This book is available online from most university libraries. If it is not available at your library, you can also obtain a free copy from https://www.academia.edu/.

Kutner M, Nachtsheim C, Neter J, Li W. *Applied Linear Statistical Models*, 5th Edition, McGraw-Hill/Irwin, New York, 2005.

Hamilton LC, *Regression with Graphics: A Second Course in Applied Statistics,* Duxbury Press, 1992.

Draper N, Smith H. *Applied Regression Analysis*, 3rd Edition, Wiley, 1998.

Selvin S. *Practical Biostatistical Methods*, Duxbury Press, 1995.

Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. *Regression Methods in Biostatistics*, 2nd ed., Springer, 2012.

Students in this course are sometimes interested in understanding more of the theoretical details underlying the methods of statistical inference for linear models. The following textbooks may be useful for this purpose (but be aware that these are not for the mathematically faint-hearted!):

Bain LJ, Engelhardt M. *Introduction to Probability and Mathematical Statistics* (2nd ed) Duxbury Press, 2000.

Casella G, Berger RL. *Statistical Inference* (2nd ed) Duxbury Press, 2001 (or 1st ed, 1990).

Software

For this subject you can use either the Stata statistical package or R. If using Stata, the notes assume the use of release 12 or later of Stata. Most of the commands we use should work fine in older versions (as long as they are not too old!), although there was an important change relevant to LMR with the introduction of “factor variables” in Stata 12.

We will be presenting all examples in the notes using Stata, as well as providing sample Stata code in appropriate places and datasets in Stata format. Corresponding R code is provided separately.

NOTE: There are three flavours of Stata – Small, "Intercooled" and Special Edition. We recommend use of Intercooled Stata, although Small Stata should actually be adequate for the examples covered in this course. Special Edition is for enormous datasets (i.e. up to 32,000 variables).

Feedback

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 Blackboard
- Interactive discussion of topics during the live online tutorials
- Interactive discussion during live Q&A online video conferences

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. Informal feedback direct to the teachers or discussion board is also encouraged and appreciated throughout the whole semester.

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

Feedback for LMR in 2020 indicated positive experiences to the few pre-recorded lectures that was available for half of the modules. So these will be continued and expanded on to include a lecture for all modules. Additionally, the some readings were not easily available. The course notes have therefore been removed so that all required readings are freely available online (though a few optional ones are still only available through physical copies at a library). This year we are also introducing live Q&A sessions on weeks when there is no tutorial, to have a chance to catch up and ask any quick questions you might have.