BIO
STATISTICS
COLLABORATION
OF AUSTRALIA

Study Guide

Linear Models (LMR)

Semester 2, 2017

Prepared by:

Dr Timothy Schlub
Sydney School of Public Health
Sydney Medical School
University of Sydney

THE UNIVERSITY OF
SYDNEY

LINEAR MODELS (LMR)
Semester 2, 2017

Contents

Instructor contact details

Dr. Timothy Schlub

Sydney School of Public Health
University of Sydney

Ph: (02) 9351 5992
Email: tim.schlub@sydney.edu.au

Method of delivery and communication

Blackboard discussion
It is expected that ***all academic questions related to the module notes and practical exercises will be posted on the discussion forum on the eLearning (Blackboard) site.*** This will enable other students to benefit from responses and indeed to respond themselves, and we will be encouraging as much interaction as possible within the class through this medium.

Elearning website
We will also use the eLearning site for posting course materials, although all of the core material (notes and readings) will also be sent out in paper form. You should have received instructions on how to access Blackboard from Erica in the BCA coordinating office. We expect that by this stage of your course most of you will have had experience with the Blackboard environment, but you are encouraged to refer to the “BCA eLearning Guide” that is available for download from the BCA website, at www.bca.edu.au/currentstudents.html.

Live online tutorials
Six live online tutorials will be held (one for each module) according to the timetable that follows. These tutorials will be held in blackboard collaborate and can be joined through by accessing the “Tutorials” folder on the eLearning website. A microphone and

speakers/headphones are required for the tutorials, with a webcam recommended but optional. A reliable internet connection with sufficient upload and download speeds for video conferencing is required for tutorial participation. Tutorials will be recorded and posted online for those who are unable to attend.

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.

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!

Unit Objectives

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. The coordinator will provide supplementary material explaining key issues by way of short videos and in response to Discussion postings.

The case study should then 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).

Assessment

Assessment will include two written assignments worth, respectively, 30% and 35%, to be completed in the middle and at the end of the semester (see dates below). In addition, students will be required to submit solutions to selected practical exercises from the Case Studies for Modules 2, 4 and 5, worth a total of 35% (10%, 20% and 5% respectively), by deadlines to be specified throughout the semester.

We require that you submit material for assessment by electronic means. Where the work involves algebraic derivations you may wish to scan (neatly!) handwritten work and submit the pdf version. In general we prefer that your work be typed in Microsoft Word (using Microsoft’s Equation Editor for algebraic work), because this facilitates electronic marking with comments.

Before submitting any assignment for assessment, please make sure you have read the Assessment Submission Instructions in the BCA Assessment Guide, which can be found here: www.bca.edu.au/currentstudents.html#assessmentguide. In particular, you should be familiar with the Academic Dishonesty and Plagiarism policy at the university in which you are enrolled.

Note that the Assessment Guide document also contains guidelines on standards that we expect (in terms of layout and expression, etc) for submitted work.

Where assignment work is submitted online using the eLearning Assessment tool, you will need to indicate your compliance with the plagiarism guidelines and policy before making the submission.

Please submit your assessment item on or before the due date. If you need an extension of time, contact the coordinator well before the due date. Although we endeavour to be flexible where possible, extension requests related to difficulties in managing your own time will not be viewed favourably. Many universities have formal policies that only allow extensions to be granted for specific reasons such as ill-health.

Late penalties: As per the standard BCA policy, where no extension has been granted, the mark obtained will be penalised by 5% of the total that would otherwise have been awarded per day late, up to a maximum of 50%. It is not the intention of this late penalty policy to cause a student to fail the unit when otherwise they would have passed. If deductions for late assignments result in the final unit mark for a student being less than 50, when otherwise it would have been 50 or greater, the student's final mark will be exactly 50.
NOTE: Submissions later than 14 days from the original due date will not be accepted and will receive a mark of zero, unless they have been approved by the unit coordinator, or special consideration has been granted.

Discussion of assessable material

The instructors will generally avoid answering questions relating directly to the assessable material until after it has been submitted. However, we encourage students to discuss the theory and content of the course via eLearning, noting that no explicit solutions to assessable exercises should not be posted for others to use, and also that students should not describe their solution to any specific assessment questions in any way.

Reference Books

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 required readings from selected texts will be provided (in photocopy form) to students. Of the books listed, the one that we make the most use of is the first one (Kutner *et al*), and we recommend this book as a comprehensive (if rather tedious!) text that covers an enormous range of material at a relatively inexpensive price. Unfortunately we understand that the book has recently gone out of print, so you may be limited to tracking down second-hand copies if you wish to buy one.

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

> N.B. The 5th edition of this textbook contained very minor changes (relevant to LMR) from the previous 4th edition, although the authorship changed. The previous edition is Neter J, Kutner M, Nachtsheim C, Wasserman W. *Applied Linear Statistical Models* , 4th Edition, Irwin, Chicago, 2005. Either edition is fine for this subject. The readings provided in your notes are from the 5th edition, but a table indicating the corresponding pages from the 4th edition is available on request from the coordinator.

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.

Weisberg S. *Applied Linear Regression* (2nd ed) New York: Wiley, 1985

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 we require use of the Stata statistical package. 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.

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).

Timetable for modules of study and assessment tasks

Below is an outline of the study modules and assessment tasks with a timetable.

All module notes will be available online and module readings will be contained in the mail-out package you have received (or will soon be receiving). Case studies with required exercises and Assignments will be posted to eLearning.

It is intended that students will work through the material for each module, including completion of practice exercises, by the end date of the module. We encourage online discussion of topics and exercises, which makes it important to work at a consistent pace with the rest of the class, as far as possible.

There are six modules in this unit and their scheduled dates are indicated below, along with the dates that assignments and assessed exercises are due.

Note that Modules 1, 2, 3 and 6 are of two weeks duration, Module 4 is of three weeks duration, and Module 5 is 1 week duration.

Timetable:

				Assessments				
Week number and start date		Module	Live online tutorial	Case study mod. 2	Assign. 1	Case study mod. 4	Case study mod. 5	Assign. 2
1	31-Jul	1. Simple linear regression: fundamentals						
2	7-Aug	1. Simple linear regression: fundamentals	Thurs. 7:30pm					
3	14-Aug	2. Simple linear regression: further topics		Released				
4	21-Aug	2. Simple linear regression: further topics	Thurs. 7:30pm					
5	28-Aug	3. Regression with two predictors		**Due Monday**	Released			
6	4-Sep	3. Regression with two predictors	Thurs. 7:30pm					
7	11-Sep	4. Multiple regression		Feedback		Released		
8	18-Sep	4. Multiple regression	Thurs. 7:30pm		**Due Monday**			
	25-Sep	Mid semester break						
9	2-Oct	4. Multiple regression			Feedback			
10	9-Oct	5. Analysis of change	Thurs. 7:30pm			**Due Monday**	Released	
11	16-Oct	6. Analysis of variance						
12	23-Oct	6. Analysis of variance	Thurs. 7:30pm			Feedback	**Due Monday**	Released
13	30-Oct							
14	6-Nov						Feedback	
15	13-Nov							**Due Monday**

Note: All assessment tasks are due on Monday at 11:59pm. Live online tutorials are held on a Thursday starting at 7:30pm.

Module Objectives

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

- 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

Module 6

- 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

Changes to LMR since last delivery

Feedback from LMR in 2016 indicated that the assignment burden was too high, particularly at the start of the semester. Additionally, students requested further guidance on the process of building a multiple regression model. To address this feedback, the assessment component of the module 1 case study that was previously worth 10% is now just part of the regular tutorial (and not worth anything). The module 4 case study that was previously worth 10% is now worth 20% and has been expanded to include some model building strategy guidance that will help students prepare for Assignment 2. These changes are intended to shift some of the assessment burden from the start of the semester, into the middle of the semester where there was previously little to do for students. It is also an overall reduction in assessment load for students.

Feedback from 2016 also indicated that students would like the opportunity to engage more with each other and with their teachers. To address this feedback, live online tutorials are being introduced for the first time this year.