BIO
STATISTICS
COLLABORATION
OF AUSTRALIA

Study Guide

Linear Models (LMR)

Semester 1, 2017

Prepared by:

Assoc/Professor Stephane Heritier

Department of Epidemiology and Preventive Medicine,
Monash University

Professor Andrew Forbes

Department of Epidemiology and Preventive Medicine,
Monash University

MONASH UNIVERSITY

POSTERA CRESCAM LAUDE
THE UNIVERSITY OF MELBOURNE

LINEAR MODELS (LMR)
Semester 1, 2017

Contents

Instructor contact details

Stephane Heritier

Department of Epidemiology and Preventive
Medicine, Alfred Hospital
Monash University

Ph: (03) 9903 0054
Email: Stephane.Heritier@monash.edu

Method of delivery and communication

Andrew Forbes from Monash University and John Carlin from University of Melbourne developed the material this subject (or “unit”) jointly. This semester, Stephane Heritier from Monash will be the primary coordinator, with Andrew functioning in a supportive role. The coordinators will be available by e-mail to answer questions related to the module notes and practical exercises, or to address any other issues that may arise, but cannot be expected to be available every day of the week.

We strongly recommend that you post content-related questions to the Discussion forum in the LMR area on the BCA’s 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. 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.

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 the case study for each module will usually 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%, by deadlines to be specified throughout the semester. For some modules a small fraction of the total marks will be allocated for the completion of brief online quizzes that we will set to help you check your understanding.

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%.

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 any and all matters between themselves, via eLearning, except that 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.

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.

In case you have not used Stata before, introductory instructional material is available for download from the eLearning website. 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 and 3 are of two weeks duration, Module 4 is of three weeks duration, Module 5 is of 10 days duration and Module 6 of 12 days duration.

Module 1: Simple linear regression: fundamentals

6 Mar – 19 Mar

No assessed exercises for this module

- scatterplot and least-squares regression line
- statistical inference for parameters of regression models
- binary covariates; connection with *t*-test

Module 2: Simple linear regression: further topics 20 Mar – 2 Apr

Assessed exercises from Module 2 due at 23.59 on 3 Apr

- model diagnostics: residuals, leverage, influence
- introduction to scatterplot smoothing & flexible curve fitting
- transformations

Module 3: Regression with two predictors 3 Apr – 23 Apr

No assessed exercises for this module

Note this module is allocated 3 weeks (due to mid-semester break)

- adjusting for a single covariate
- analysis of covariance
- confounding and interaction

Mid-semester break: 14-21 Apr

Note: tutors may be less available than normal during this period

Assignment 1: available Friday 14 April
due 23.59 Monday 8 May

Module 4: Multiple regression 24 Apr – 14 May

This module is allocated 3 weeks

Assessed exercises from Module 4 due at 23.59 on 15 May

- basic theory
- building explanatory and predictive models
- variable selection strategies
- collinearity

Module 5: Analysis of change 15 May – 24 May

Note that 10 days are allowed for this module

Assessed exercises from Module 5 due at 23.59 on 29 May

- regression to the mean
- using baseline data

Module 6: Analysis of variance 25 May – 6 June

Note that 12 days are allowed for this module

No assessed exercises for this module

- one-way, two-way and factorial ANOVA
- multiple comparison issues
- fixed and random effects

Assignment 2: available Thursday 1 June
due 23.59 Sunday 18 June

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

The most recent available student evaluation of LMR (Sem 2 2016) contained few specific suggestions for improvement and no substantial changes have been made. Exercise 1 assessed exercise has been removed to give students more time to start the semester. The size of the two major assignments has been slightly modified with respectively 30% and 35% weight. The 3 minor assignments are worth 10%, 20% and 5% for respectively module 2, 4 and 5.