Faculty of Medicine
Centre for Big Data Research in Health

HDAT9600
Health Data Analytics: Statistical Modelling I

COURSE OUTLINE

Term 2 2019

TABLE OF CONTENTS PAGE

COURSE INFORMATION

This course provides a sound grounding in the theory and practice of fitting statistical regression models, including linear models; generalised linear models (GLMs) for outcomes that are non-linear, binary or count; and survival analysis and time to event models. A major theme of the course is best practice in model fitting, including thorough exploratory data analysis, model assumption checking, data preparation and transformation, including the use of imputation, and careful attention to model adequacy and diagnostics. The presentation and visualisation of statistical models is considered, with emphasis on the explanatory insights that can be gained from well-constructed models.

COURSE LEARNING OUTCOMES

On completion of this course students should be able to:

- construct GLMs with appropriate covariate sets in health research scenarios.
- appraise model fit using a variety of model diagnostics.
- compose narratives of GLM interpretation within the framework of statistical inference.
- visualise 'standard' statistical techniques as special simplified cases of the GLM.

COURSE CONVENOR

Andrew Blance a.blance@unsw.edu.au OL tag @andb

RECOMMENDED TEXTBOOKS

All required course material is provided online, based on these textbooks:

1. Faraway, Julian. *Linear Models with R.* 2nd Edition (2016), Chapman & Hall, CRC Press
2. Faraway, Julian. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* 2nd Edition (2016), Chapman & Hall, CRC Press.
3. Harrel, Frank. *Regression Modelling Strategies with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd Edition (2015), Springer International Publishing.
4. Broström, Göran. *Event History Analysis with R*. 1st Edition (2012), Hoboken: CRC Press.

Interested students are encouraged to read further about the covered material. Note that all essential content necessary for assessment purposes is included within the course materials.

COURSE STRUCTURE and TEACHING STRATEGIES

This is an online learning course comprising of:

- 10 chapters hosted on www.openlearning.com
- Interactive learnr documents in RStudio

This course is equivalent to a UNSW standard 6 units-of-credit postgraduate course, which should involve about 150 hours of study and learning activities. The formal learning activities are approximately 100 hours throughout the session and students are expected (and strongly recommended) to do additional hours of self-study.

The course is accessed via www.openlearning.com. Core material will be delivered as learnr documents opened in R Studio with interactive coding activities and explanatory text.

Assumed knowledge: This course is taught using basic R programming. If you are new to R or need a refresher, the first chapter includes an introduction to R module. Statistical Modelling builds on the concepts taught in Statistical Foundations for Health Data Science. If you have not completed this prerequisite course, it is expected students to have an intermediate level of statistics knowledge.

COURSE RESOURCES

1. Course materials

- Written notes (in learnR format)
- Course videos
- Course exercises

2. R & RStudio

- **Install R** on your computer: **https://www.r-project.org/**
- **Install RStudio** on your computer: **https://www.rstudio.com/**

COURSE EVALUATION AND DEVELOPMENT

For course evaluation, feedback will be gathered towards the end of the course via MyExperience and Survey Monkey for blended students and webstream students, respectively. Information will be provided in OpenLearning how to complete this survey. Student feedback is taken seriously, and continual improvements will be made to the course based, in part, on such feedback.

ASSESSMENT PROCEDURES

The assessment for this course comprises two major components. First, (three sets of) individual in-course assignments totalling 70%. Second, a final project accounting for the remaining 30%.

1. Three sets of individual exercises: 70%

   The in-course assignment component is further broken down into three assignments.

   - Linear models (chapters 2-4) - 30%
   - Generalised liner models (chapters 5-7) - 20%
   - Survival analysis (chapters 8-9) - 20%

   The assignments consist of exercises based on content for that section and/or previous sections. They consist of a number of programming exercises in the R language used throughout the course, and some short-answer questions.
   These exercises will re-iterate the concepts and best practice covered in each chapter's material. No new concepts or materials will be introduced in or required for the exercises.

2. A group project: 30%

   Each group will undertake an assignment which requires:

   - Exploratory data analysis of a provided data set (each group must use a different data set).
   - Propose two appropriate regression models to address contextual questions (suggested questions will be provided, although you may pose your own), and propose one or more underlying causal models which will be used to inform model construction.
   - Fit the proposed models and evaluate their technical attributes (goodness-of-fit, model diagnostics, predictive performance if applicable, etc.).
   - Report all the above in the form of a "literate programming "document which interleaves narrative text, program code and statistical outputs as an R knitr document (very similar to an R or Jupyter notebook).
   - For this group project, an individual's mark will be composed of an individual component (which may vary amongst team members) and a group component (which will be the same for all team members). A rubric developed specifically for the purposes of the assessment will be used.

Late submissions

If you submit assessments late without special consideration, a 5% penalty deduction for every day late will be applied. For example, if you submit an assessment 3 days late, then 15% (5% x 3 days) will be deducted from the assessment mark. Thus, if your assessment was marked as 80% but was submitted 3 days late, then your final mark will be 65% only. Assessments will not be marked if submitted 14 or more days after the assessment due date and will receive a score of 0.

COURSE TIMETABLE

All times 17:00, unless indicated.

Week	Released	Chapter	F2F	Assessment due
1	9am 3 June	Intro to R, RStudio, knitr & rmarkdown Exploratory data analysis	4 June	
2	7 June	Linear modelling I	11 June	
3	14 June	Linear modelling II	18 June	
4	21 June	Linear modelling III	25 June	28 June
5	28 June	Generalised linear modelling I	2 July	
6	5 July	Generalised linear modelling II	9 July	
7	12 July	Generalised linear modelling II	16 July	19 July
8	19 July	Survival analysis I	23 July	
9	26 July	Survival analysis II	30 July	2 Aug
10	2 Aug	Review Final project	6 Aug	 23 Aug

GENERAL INFORMATION

Special Consideration

If you find you are going to miss an assessment due to illness, misadventure or circumstances beyond your control, you need to apply for special consideration. To do this:

1. Tell the course convenors as soon as possible via email. Applications for special consideration will not normally be received more than 3 days after the assessment due date.
2. Submit supporting evidence. This may include a medical certificate or a supporting email from your supervisor for extenuating work commitments.

If your application for special consideration is approved, your course convenors will discuss with you how you can complete your assessment.

Academic Integrity and Plagiarism

The UNSW Student Code outlines the standard of conduct expected of students with respect to their academic integrity and plagiarism.

More details of what constitutes plagiarism can be found here.