FACULTY OF SCIENCE

SCHOOL OF MATHEMATICS AND STATISTICS

Semester 2, 2018

MATH3821

STATISTICAL MODELLING & COMPUTING

Cricos Provider Code: 00098G

School of Mathematics and Statistics, Faculty of Science
The Red Centre, Centre Wing - NSW 2052 Australia
Kensington campus

COURSE OUTLINE

MATH3821
Statistical Modelling & Computing
6 Units of Credit (6UOC)
(Semester 2, 2018)

Course authority and lecturer: Dr. Pierre Lafaye de Micheaux
Contact: Office: 2050
E-mail: lafaye@unsw.edu.au
Phone: (00.[+612]) 9385 7029
Web: http://web.maths.unsw.edu.au/~lafaye

Consultation hours: A suitable time slot will be chosen during the first lecture.
Please use e-mail to arrange an appointment outside this time.

1 Course Overview

The main purpose of this course is to give an introduction to flexible and modern approaches to regression and simulation methods using the statistical package R. In particular, various extensions of the linear models discussed in MATH2831 are considered: we consider regression models where we allow the mean response to be a quite general smooth function (nonparametric regression methods such as scatter plot smoothing, penalised splines, etc.) and regression models for data which are discrete or non-Gaussian (generalised linear models). The Bayesian linear model is also covered and simulation techniques such as Markov Chain Monte Carlo and classic Monte Carlo methods are used to fit these models. The lectures will be complemented with worked examples using the R data analysis and statistical programming software.

2 Pre-requisites, Exclusions

Pre-requisites: MATH2831 or MATH2931

3 Schedule and Format of the Course

There will be two (2) lectures per week (50 mn each), except for week 6.
There will be one (1) laboratory per week (110 mn), except for weeks 1 and 6.
Week 13 will be entirely dedicated to the oral presentations for the second assignment.

Activity	Day	From - To	Date	Room	Building
Lecture	Monday	09:05 – 09:55	July 23	1027	Macauley Theatre (E15)
	Wednesday	13:05 – 13:55	July 25		
	Monday	09:05 – 09:55	July 30		
	Wednesday	13:05 – 13:55	August 1		
	Monday	09:05 – 09:55	August 6		
	Wednesday	13:05 – 13:55	August 8		
	Monday	09:05 – 09:55	August 13		
	Wednesday	13:05 – 13:55	August 15		
	Monday	09:05 – 09:55	August 20		
	Wednesday	13:05 – 13:55	August 22		
	Monday	09:05 – 09:55	September 3		
	Wednesday	13:05 – 13:55	September 5		
	Monday	09:05 – 09:55	September 10		
	Wednesday	13:05 – 13:55	September 12		
	Monday	09:05 – 09:55	September 17		
	Wednesday	13:05 – 13:55	September 19		
	Wednesday	13:05 – 13:55	October 3		
	Monday	09:05 – 09:55	October 8		
	Wednesday	13:05 – 13:55	October 10		
	Monday	09:05 – 09:55	October 15		
	Wednesday	13:05 – 13:55	October 17		
	Monday	09:05 – 09:55	October 22		
	Wednesday	13:05 – 13:55	October 24		
Laboratory	Tuesday	10:05 – 11:55 (T10A only)	weeks 2–5,7-13	G12C	Red Centre Central Wing (H13)
		16:05 – 17:55 (T16A only)		M020	

Lectures: 23 hours over 12 weeks (weeks 1–5 and weeks 7–13)
Tutorials: 22 hours over 11 weeks (weeks 2–5 and weeks 7–13)
Recommended independent (out-of-class) study hours: 72 hours

Each class session will consist either of a lecture provided by the teacher, or of some laboratory time on a computer. Presence in class is **strongly encouraged** and the presence might be recorded.

4 Assessments

1. There will be two (2) group assignments (≤ 3 students) during the term of the session, each worth 10%. They will both have a mathematical and computational component. Each group will do an oral presentation (10 mn, including 1 mn for questions) of his work for the second assignment during week 13 (in the lab or lecture time). Consequently, the groups for the second assignment will be formed by the lecturer within each tutorial group. Students are expected to attend (and evaluate) the oral presentations of all the other groups in their tutorial group. Late assignments will *not* be accepted. You must submit your own work, or severe penalties will be incurred.

2. There will be one (1) mid-session test using the R software. It will occur in week 11 in your lab room and will account for 20% of the final mark.

3. The final examination (2 hours) will have a purely written form and will account for 60% of the total mark. It will occur in November. It will cover the whole course content.

Task	Weighting	Duration	Date Due
Assignment 1	10%	2 weeks	end of week 5
Assignment 2	10%	2 weeks	end of week 12
Mid-session test	20%	1.5 hours	week 11
End-of-session Examination	60%	2 hours	TBA

Knowledge and Abilities Assessed

- The Assignments will assess your ability to write short R codes, as well as your ability to demonstrate theoretical results. They will also assess your ability to communicate statistics orally and in writing.
- The mid-session test will assess your ability to use R packages and functions appropriately, as well as your theoretical understanding of the assumptions and limitations behind several statistical techniques.
- The Final Examination will mainly test your knowledge of the course content (no document allowed), your ability to prove mathematical results as seen in class, your ability to interpret correctly the output of statistical software when used to analyse a real data set, and how you can relate this output to the theoretical formulas presented in the lectures.

Note: Supplementary Examinations for S2 2018, for both final examination special consideration applications and those students who achieved a final mark from 45–49, will be scheduled between Saturday 8 December to Saturday 15 December 2018. See the section 'Administrative Matters' thereafter for more details.

5 Course Aims

The aim of MATH3821 is that at the end of session you should understand the theory, concepts and techniques involved in the syllabus and be able to apply those concepts and techniques to the solution of appropriate problems. The R packages will allow you to solve problems computationally.
The activities and assessment for the course will contribute to the core science graduate attributes of 'Research, inquiry and analytical thinking abilities', 'Capability and motivation for intellectual development' and 'Communication'. New ideas, skills and methods are introduced, discussed and demonstrated in lectures. Then students develop these skills by applying them to specific tasks in tutorial-like periods and in assessments. Active student participation in tutorial-like periods is expected.

6 Course Learning Outcomes

Upon successful completion of the requirements of this course, you should have the knowledge and skills to:

✓ correctly state definitions introduced in the lecture notes;

✓ have working knowledge of appropriate theorems;

✓ apply the concepts and techniques of MATH3821 to solve appropriate problems;

✓ have the ability to use specific and general results given specified assumptions;

✓ use terminology and reporting styles appropriately and successfully to communicate information and understanding;

✓ choose the appropriate R package to apply effectively a statistical or computing technique;

✓ describe the output of statistical software and interpret the results;

✓ relate the output of statistical software to the theoretical formulas presented in the lectures;

✓ clarify the models, hypotheses, intuitions, strengths and weaknesses of the various approaches;

✓ formulate and discuss effectively the results, in written and oral form.

The above outcomes are related to the development of the Science Faculty Graduate Attributes, in particular: 1. **Research, inquiry and analytical thinking abilities**, 4. **Communication**, 6. **Information literacy**.

7 Resources

Lecture notes (digital slides created with RMarkdown/LaTeX) provide the main reference source for this course.

The slides used in class will be posted on the course web site in advance:
UNSW Sydney's Moodle web page https://moodle.telt.unsw.edu.au
and/or my personal web page http://web.maths.unsw.edu.au/~lafaye
It is strongly advised to download and/or print these slides, and to bring them in class for annotation purpose.

There is no single mandatory textbook as the content of the course will be defined by the lectures. But some useful books are listed in the References section at the end of this document.

You should check regularly Moodle for new materials, as well as for announcements about assessment tasks.

8 Teaching Strategies Underpinning the Course

New ideas and skills are first introduced and demonstrated in lectures, then students develop these skills by applying them to specific tasks. Computing skills are developed and practiced.

We believe that effective learning is best supported by a climate of inquiry, in which students are actively engaged in the learning process. Hence this course is structured with a strong emphasis on problem-solving tasks and students are expected to devote the majority of their class and study time to the solving of such tasks. Effective learning is achieved when students attend all classes, have prepared effectively for classes by reading through previous lecture notes and by having made a serious attempt at doing for themselves the problems. Furthermore, lectures should be viewed by students as an opportunity to learn, rather than just copy down or skim over lecture notes.

MATH3821 is taught through carefully planned lectures that logically develop the concepts and techniques specified in the course. Examples are emphasised as they provide the underlying motivation for the course, and because students best understand the general theory when it is developed from simple, and then more complex examples. Small group tutorials allow students to apply the material introduced in the lectures. These tutorials provide the opportunity for individual assistance. Students are expected to print out and bring their tutorial exercises and are expected to work conscientiously at the exercises in the tutorials.

Students are encouraged to give constructive feedback during the teaching session. They are encouraged to work collaboratively with other students to develop their understanding and their problem solving skills.

9 Topics Covered

If time permits, it is intended that the following topics will be covered in the given order, through a mix of lectures and practical tutorials using the **R** software. Any variation from this will be indicated by the lecturer. The use of **R** functions and packages will be illustrated with examples of real and/or simulated data sets. Mastering the theory is important for a deep understanding of how to adequately use statistical methods, thus parts of the lectures will be devoted to this. The lecture notes are the most direct way to learn about the topics. To go further, you can consult the selected references.

1. **Linear models:** introduction to the **R** software, the simple linear regression (SLR) model, model selection and validation, residuals, categorical variables, ANOVA, relationships to t-tests; (week 1)

2. **Binomial regression:** logistic regression, Wald tests, deviance, sequential tests, residuals, choosing models and transformations, multiple predictors, probit regression, classification; (week 2)

3. **Generalized linear models (GLM):** exponential family, link functions, iteratively weighted least squares, weighted least squares regression, scaled deviance, Poisson regression, Poisson distribution, maximum likelihood estimation, hypothesis testing, Poisson approximation to the binomial; (week 3)
4. **Smooth and non-parametric regression:** polynomial regression, scatterplot smoothing, idea of a basis, splines, smoothing splines, regression splines, selection of the smoothing parameters, bias/variance trade-off, expressions for MSE and PSE, cross-validation, generalised cross-validation, degrees of freedom of a smoother, confidence bands for smoothers, smoothing with multiple predictors, the curse of dimensionality; (week 4)
5. **Kernel smoothing and density estimation:** link between weighted least squares regression and local regression, running mean smoother, running line smoother, loess algorithm, kernel smoothing, local likelihood, histograms, bias and variance for histograms, choice of bin width, nonparametric density estimation, empirical distribution function, smooth density estimation, MISE for kernel estimators, variability plots, bivariate kernel density estimation, multivariate kernel density estimation; (week 5)
6. No teaching during week 6.
7. **Generalized additive models (GAM):** additive models, interactions in additive models, backfitting algorithm, projection pursuit regression; (week 7)
8. **Bayesian Inference:** the Bayesian philosophy, the Bayesian method, functions of parameters, large sample properties of Bayes' procedures, flat priors, improper priors and noninformative priors, multiparameter problems, bayesian testing, strengths and weaknesses of Bayesian inference; (week 8)
9. **The Bayesian linear model:** credible intervals for coefficients, prediction, residuals and diagnostics; (week 9)
10. **Simulation methods:** approximating integrals, generating discrete random variables, the rejection method, rejection sampling, the inverse transform method; (week 10)
11. **Markov Chain Monte Carlo:** Markov chain theory, the Gibbs Sampler, the Metropolis-Hastings algorithm, the random-walk Metropolis sampler, the independence sampler, hybrid chains, hybrid MCMC, the reversible jump MCMC; (week 11)
12. **The Bootstrap:** generalities, bootstrap confidence intervals; (week 12)
13. No teaching during week 13 which is used for the oral presentations.

10 Course Evaluation and Development

The School of Mathematics and Statistics evaluates each course each time it is run. We carefully consider the student responses and their implications for course development. It is common practice to discuss informally with students how the course and their mastery of it are progressing. Thank you in advance for your contribution!

11 Administrative Matters

It is the student's responsibility to be familiar with UNSW and School of Mathematics and Statistics policies.

Assessment Policies

The School of Mathematics and Statistics has a set of assessment policies that you can consult at
http://www.maths.unsw.edu.au/currentstudents/assessment-policies

If you are absent (e.g., ill) from the final examination, you must apply for special consideration using the UNSW Special Consideration online service. Information regarding additional assessments are available at http://www.maths.unsw.edu.au/currentstudents/additional-assessment

If your final mark is in the range 45–49, you are automatically eligible for a deferred examination, but your final mark, if you pass the examination, will be capped at 50. This capping will not apply if you were ill for the examination and have applied on-line in the usual way. If you are ill on the day of the examination, then you should not sit the examination, but should apply as above. If you are ill and your mark during the semester is less than 40% you are unlikely to be granted a deferred examination.

The Use of Calculators in the Examination

The University does **not** supply calculators in the final examination. You should look at the web page https://www.maths.unsw.edu.au/currentstudents/exam-information-and-timetables

School Rules and Regulations

Fuller details of the general rules regarding attendance, release of marks, special consideration, etc., are available at
http://www.maths.unsw.edu.au/currentstudents/student-services

Plagiarism and Academic Honesty

Plagiarism is presenting another person's work or ideas as your own. Plagiarism is a serious breach of ethics and is not taken lightly. It undermines academic integrity and it will not be tolerated.

UNSW Sydney has a set of rules and regulations for academic conduct, honesty and plagiarism. See
http://www.lc.unsw.edu.au/academic-integrity-plagiarism
and
https://www.gs.unsw.edu.au/policy/documents/studentcodepolicy.pdf

12 References to Complement the Lecture Notes

1. P. Lafaye de Micheaux, R. Drouilhet and B. Liquet (2013). *The* R *Software : Fundamentals of Programming and Statistical Analysis*. Statistics and Computing. New York: Springer.
   http://link.springer.com/book/10.1007/978-1-4614-9020-3
   Translations exist in French, Chinese and Indonesian.
   http://biostatisticien.eu/springeR

2. G. James, D. Witten, T. Hastie and R. Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R*, Springer.

3. T. Hastie, R. Tibshirani and J. Friedman (2008). *The Elements of Statistical Learning: Data Mining, Inference and Predictions*, Second Edition, Springer.

4. T. Hastie and R. Tibshirani (1990). *Generalized Additive Models*, Chapman and Hall.

5. J. Scott Long (1997). *Regression Models for Categorical and Limited Dependent Variables*, Sage publications.

6. P. J. Green and B. W. Silverman (1994). *Nonparametric Regression and Generalised Linear Models*, Chapman and Hall.

7. A. J. Dobson (2002). *An introduction to Generalised linear models*, Second Edition, Chapman and Hall.

8. J. Rawlings, S. Pantula and D. Dickey (1998). *Applied Regression Analysis: A Research Tool*, Second Edition, Springer.

9. C. Bishop (2006). *Pattern Recognition and Machine Learning*, Springer.

10. D. Hosmer and S. Lemeshow (2000). *Applied Logistic Regression*, Second Edition, Wiley.

11. B. Efron and T. Hastie (2016). *Computer Age Statistical Inference Algorithms, Evidence, and Data Science*, Cambridge University Press.

12. P. McCullagh and J. Nelder (1989). *Generalized Linear Models*, Second Edition, Chapman and Hall.

13. J. Pinheiro and D. Bates (2000). *Mixed-Effect Models in S and S-PLUS*, Springer.

14. S. Weisberg (2005). *Applied Linear Regression*, Third Edition, Wiley.

15. S. Sheather (2009). *A Modern Approach to Regression with R*, Springer.

16. W. Venables and B. Ripley (2002). *Modern Applied Statistics with S*, Fourth Edition, Springer.

17. G. Wahba (1990). *Spline Models for Observational Data*, SIAM: Society for Industrial and Applied Mathematics.

18. S. Wood (2006). *Generalized Additive Models: an introduction with R*, Chapman and Hall.

19. D. Ruppert, M. P. Wand and R. J. Carroll (2003), *Semiparametric Regression*, Cambridge University Press.

20. C. J. Lloyd (1999), *Statistical Analysis of Categorical Data*, Wiley.

21. A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin (2004), *Bayesian Data Analysis*, Chapman and Hall.

Some of the above books might be available at your local library: https://primoa.library.unsw.edu.au/primo-explore/search?vid=UNSWS

Last update: July 15, 2018.