UNSW SYDNEY

FACULTY OF SCIENCE
SCHOOL OF MATHEMATICS AND STATISTICS

MATH5806

APPLIED REGRESSION ANALYSIS

Semester 2, 2017

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

COURSE OUTLINE

MATH5806
Applied Regression Analysis
6 Units of Credit (6UOC)
(Semester 2, 2017)

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: We will decide on suitable time slots during the first lecture.

1 Course Overview

The topics covered in this course include (and are subject to change):

linear regression; weighted least squares; generalised linear models; fitting GLMs and diagnostics; poisson, binomial regression; analysis of variance; penalised regression methods; splines; penalised splines; thin plate splines; variable selection; generalised cross-validation; local likelihood; kernel smoothing; generalised additive models; multinomial logit analysis; ordinal logistic regression.

The lectures will be complemented with worked examples using R.

2 Prerequisites, Exclusions

Prerequisite: 24 units of level III mathematics or a degree in a numerate discipline or permission of the Head of Department. Basic knowledge in the R statistical software is recommended.

3 Schedule and Format of the Course

Activity	Day	From - To	Date	# sessions	Room	Building
Lectures	Friday	17:00 – 20:00	August 4 August 11 August 18 August 25 September 1 September 8 September 15 September 22 October 6 October 13 October 20 October 27	1	G32	OMB

Lectures: 36 hours
Independent study hours: 72 hours

Each class session will consist of a lecture provided by the teacher, followed by some practicals to do in class.

4 Teaching Strategies Underpinning the Course and Access to the Web Site of the Course

Lecture notes provide a brief reference source for this course. New ideas and skills are first introduced and demonstrated in lectures, then students develop these skills by applying them to specific tasks in tutorials. 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 in tutorials, 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, in the case of lectures, and, in the case of tutorials, by having made a serious attempt at doing for themselves the tutorial problems prior to the tutorials. Furthermore, lectures should be viewed by students as an opportunity to learn, rather than just copy down or skim over lecture notes.

The slides used in class will be posted on the course web site[1] in advance.

5 Assessments

UNSW assesses students under a standards based assessment policy. For how this policy is applied in the School of Mathematics and Statistics see
http://www.maths.unsw.edu.au/currentstudents/assessment-policies

- There will be two assignments during the term of the session. These will have a mathematical and computational component. Each assignment will contribute at least 10% to the final mark.
- Depending on students numbers, there will be individual or group projects. This will be a thorough statistical analysis, with a substantial amount of computational work. A professional report will be handed in, which details the data sets, methods and assumptions used, as well as inference and conclusions. Based on this report, a student's ability to communicate statistics in writing will be assessed. If time allows, 10–15 minute in-class presentation will be given for each project, to also assess the ability to communicate statistics orally. This project might account for up to 30% of the final mark.
- The final exam will have a purely written form and will account for 50% of the total mark.

6 Important Administrative information

The school has strict rules for academic conduct and plagarism.

- Presence in class is **strongly encouraged** and the presence might be recorded.
- 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.

Further rules and regulations, particularly on plagiarism and academic honesty, can be found at
https://www.maths.unsw.edu.au/currentstudents/assessment-policies.

Note: The information contained herein is for general guidance of students and is as accurate as possible at the date of issue. You will be informed of any changes.

[1] UNSW Moodle and/or http://web.maths.unsw.edu.au/~lafaye

7 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!

8 Course Aims

The aim of this course is to introduce students to modern regression models and to provide hands-on experience with computing methods needed for applications to real data. 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 tutorials and assessments. Active student participation in tutorials is expected.

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

- ✓ choose the appropriate regression technique to analyse a given data set;
- ✓ choose the appropriate R package to apply effectively this technique;
- ✓ interpret the output and the results;
- ✓ understand the theoretical fundations on which these techniques rely;
- ✓ understand the models, hypotheses, intuitions, and strengths and weaknesses of the various approaches;
- ✓ communicate effectively the results, in written or oral form.

9 Course Content

Regression is a set of statistical techniques widely used to analyse relationships between several variables. We will cover the following topics, if time permits, through a mix of lectures and practical tutorials using the R software:

1. **Introduction:** concepts such as prediction, inference, estimation methods, maximum likelihood estimation, least squares estimation, simple linear regression, exponential family of distributions, introduction to R;
2. **Generalized Linear Models (GLMs):** flexible general framework including linear regression, logistic regression and Poisson regression as special cases, methods of estimation, model fitting and statistical inference are discussed;
3. **General Linear Models and analysis of residuals:** multiple regression, ANOVA (Analysis of Variance) is used for a continuous response variable and categorical explanatory variables (factors), ANCOVA (Analysis of Covariance) is used when at least one of the explanatory variables is continuous;
4. **Assessing Model Accuracy:** definition of error, bias-variance trade-off, variable selection, cross validation;
5. **Shrinkage Methods:** by shrinking the coefficient estimates towards zero, their variance is being reduced. Ridge regression and Lasso are two best-known examples of shrinkage methods;

6. **Nonlinear Regression:** polynomial regression, step functions, regression splines, smoothing splines, local regression and generalized additive models; Nonlinear models are more complex in terms of interpretation and inference. However, their advantage lies in their predictive power. Polynomial regression extends the linear model by raising the original predictors to a power. For instance, the cubic regression will have three variables: X_1, X_2 and X_3 as predictors. Step functions cuts the range of a variable into K distinct regions, which produces a piecewise constant fit. Regression splines are extension of polynomials and step functions and involve dividing the range of X into K distinct regions. Within each region a polynomial function is fit to data. Smoothing splines are similar to regression splines and result from minimizing the residual sum of squares subject to smoothing penalty. In Local regression the regions are allowed to overlap. Generalized additive models extend the methods above to multiple predictors;

7. **The Bootstrap:** The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.

8. **Dimension Reduction Methods:** Principal components analysis (PCA) is a popular approach for deriving a low- dimensional set of features from a large set of variables. Partial least squares (PLS) is a supervised alternative to Principal Component Regression.

The use of R packages will be illustrated with examples of real or simulated data sets.

10 References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

18. 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
http://biostatisticien.eu/springeR
Note: versions exist in French, Chinese and Indonesian.

Some of the above books might be available at your local library: http://primoa.library.unsw.edu.au/primo_library/libweb/action/search.do?&vid=UNSWS&reset_config=true

Last update: July 17, 2017.