P08801 Data Science Foundations

Course Description

The broad aim of the module is to present an overview of core data science concepts and tools, focusing on real-life data science research questions with practical exposure to R and/ or Python programming as an integral part of the course.

Relationship with other Modules

Level 7 x.5 module

- Compulsory for
  - MSc Data Analytics for Government (TE62A)
  - PG Diploma in Data Analytics for Government (TE62B)
  - PG Certificate in Data Analytics for Government (TE62C)
  - MSc Data Analytics (TE63A)
  - PG Diploma in Data Analytics (TE63B)
  - PG Certificate in Data Analytics (TE63C)

Learning Outcomes

The Brookes attributes to which the learning outcomes are linked are as follows:

a. Academic literacy
b. Research literacy
c. Critical self-awareness and personal literacy
d. Digital and information literacy
e. Active citizenship

On successful completion of this module, students will be able to:

	Brookes Attribute Developed	Other BAs Developed
1. Demonstrate the ability to identify and integrate data of various types from traditional and alternative sources, and make informed judgements about their use in data science research	a	c
2. Critically evaluate the methodologies applied in data collection, data processing, data analysis & dissemination of research findings	a	b
3. Critically assess methods and data strengths and limitations combined to application of R and/or Python	d	

Outline Syllabus

- Core data science concepts
- Different types of data sources available (admin data, survey data, open data, big data, etc)
- Data collection, including innovative data collection methods
- Challenges associated with unstructured data
- Treatment of different data types
- Basic data analysis (structured and unstructured data)
- Presentation of data through basic data visualisations

Teaching Learning and Assessment Strategy

The teaching of the theoretical aspects of data science will largely be done in the lectures, crucially contributing to the students' Academic Literacy.

The lectures will guide students to the literature where they will evaluate the relative strengths and weaknesses of various approaches to data science, developing their Research Literacy skills. The lectures will also teach students how to analyse and critically evaluate different approaches, so enhancing their Academic Literacy. All of this material will be assessed through the coursework.

Developmental and experimental material will largely be taught in the practical sessions, and formative exercises with fully worked solutions to a selection of problems will be provided. The practical sessions will include the use of software tools. This material will be assessed in the practical assignment in which the student will utilise their knowledge of data science to solve particular problems. This will involve recourse to technical documentation, thus enhancing the students Digital Information Literacy

Academic, Research and Digital Information Literacy will be assessed through the coursework components.

Learning Hours

Scheduled learning and teaching activities

Lectures	12 hours
Practicals/workshops	12 hours

Guided independent study

Directed/independent study	46 hours
Assessment prep	30 hours

Assessment Tasks

Summative Assignments	**Word Count/ Exam Length**	**LOs Assessed**	**Weighting %**
Coursework: (Indicative assignment)			
Short written report detailing the application of data science to the solution of selected problems. Each problem will necessarily involve the use of programming in R and/or Python	4 problems	1-3	100%

Opportunities for formative assessment and feedback

Students are expected to use lecture content to guide their own study of the module content. Verbal feedback will be given at all practical classes, and in-class discussion of practical exercises will provide further formative feedback.

Indicative Reading List

Author	**Title**	**Publisher**	**Date**
Igual L & Segui S	*Introduction to Data Science*	Springer	2017
Mahmood Z	*Data Science and Big Data Computing: Frameworks and Methodologies*	Springer	2016

Validation History

Date module first approved: June 2017

P08801 Data Science Foundations
Proposed Timetable

Time	Monday	Tuesday	Wednesday	Thursday	Friday
10:00-12:00	**Session 1** Intro to Data Science	**Session 4** Data Pre-Processing	**Session 7** Advanced DBs	**Session 10** Semi-Structured Data: XML namespaces and DTDs	**Session 11** Data Visualisation Theory
12:00-12:45	**LUNCH**				
12:45-14:45	**Session 2** Data sources	**Session 5** Database Theory	**Session 8** Optimising the DB for Data Analytics	Free slot	**Session 12** Data Visualisation for Data Science
14:45-15:00	**BREAK**				
15:00-16:30	**Session 3** Sampling Distribution and Bootstrap	**Session 6** MySQL	**Session 9** Semi-Structured Data: an intro to XML	Free slot	**Session 13** Recap

Upcoming runs of this module

15th – 19th January 2018, Newport, £420
Assessment deadline 26th February 2018
Applications close 22nd December 2017

21st – 25th May 2018, Titchfield, £420
Assessment deadline 2nd July 2018
Applications close 4th May 2018

11th – 15th June 2018, Edinburgh, £550
Assessment deadline 23rd July 2018
Applications close 12th March 2018

P08803 Statistical Programming

Course Description

The aim of the module is to introduce core programming techniques in R essential for performing data manipulation, data processing and data analyses of traditional and alternative data sources through practical sessions.

Relationship with other Modules

Level 7 x.5 module

- Compulsory for
  - MSc Data Analytics for Government (TE62A)
  - PG Diploma in Data Analytics for Government (TE62B)
  - PG Certificate in Data Analytics for Government (TE62C)
  - MSc Data Analytics (TE63A)
  - PG Diploma in Data Analytics (TE63B)
  - PG Certificate in Data Analytics (TE63C)

Learning Outcomes

The Brookes attributes to which the learning outcomes are linked are as follows:

a. Academic literacy
b. Research literacy
c. Critical self-awareness and personal literacy
d. Digital and information literacy
e. Active citizenship

On successful completion of this module, students will be able to:

	Brookes Attribute Developed	Other BAs Developed
1. Undertake complex data analysis tasks by performing data entry, data manipulation and statistical procedures in R with application to survey data, administrative data, census data and big data	d	
2. Identify and select appropriate built-in functions, and implement technical and analytical processes not readily available in the software by using iterative methods and simulations to solve complex problems	c	d
3. Identify and evaluate the theoretical aspects of statistical programming	a	

Outline Syllabus

- Introduction to R
- Data types and structures
- Importing and exporting data
- Basic statistical methods
- Basic graphics
- Advanced graphics
- Further statistical methods
- Basic programming
- Further programming
- Packages
-

Teaching Learning and Assessment Strategy

Students completing module P08803 will have been given the opportunity to:

- Attend lectures which will teach the theoretic aspects of statistical programming;
- Participate in practical classes using R.

Module P08803 is assessed by two programming assignments.

All the teaching and learning activities are targeted to development and training of specific abilities, as listed in the learning outcomes and to enable the successful completion of all the above assessed tasks.

The teaching and learning activities for module P08803 include lectures, practical classes, review and formative feedback sessions.

Attendance and active participation to the theoretical lectures and practical classes, along with the necessary independent study and students' effort, will allow the development of an adequate knowledge-base of the principles of statistical programming.

Learning Hours

Scheduled learning and teaching activities

Lectures	12 hours
Practicals/workshops	12 hours

Guided independent study

Directed/independent study	46 hours
Assessment prep	30 hours

Assessment Tasks

Summative Assignments	**Word Count/ Exam Length**	**LOs Assessed**	**Weighting %**
Coursework: (Indicative assignment)			
1. Use built-in functions of R to solve given problems	2 problems	1-2	50%
2. Develop a software tool using R to analyse a given data set, and report on the findings		1-3	50%

Opportunities for formative assessment and feedback

Students are expected to use lecture content to guide their own study of the module content. Verbal feedback will be given at all practical classes.

Fully worked solutions will be given for the weekly exercises to enhance students' assessment literacy.

Indicative Reading List

Author	**Title**	**Publisher**	**Date**
Dalgaard P	*Introductory Statistics with R*	Springer	2008
Zuur A et al	*A Beginner's Guide to R*	Springer	2009

Validation History

Date module first approved: June 2017

P08803 Statistical Programming
Proposed Timetable

Time	Monday	Tuesday	Wednesday	Thursday	Friday
10:00-12:00	**Session 1** Introduction to statistical programming with R	**Session 3** Importing data	**Session 5** Basic graphics	**Session 7** Further statistical methods	**Session 9** Further programming
12:00-13:00	**LUNCH**				
13:00-14:45	**Session 2** Data types and structures	**Session 4** Basic statistical methods	**Session 6** Advanced graphics	**Session 8** Basic programming	**Session 10** Creating packages
14:45-15:00	**BREAK**				
15:00-16:30	**Practical session**	**Practical session**	**Practical session**	**Practical session**	Recap, gaining credit, and **practical session**

Upcoming runs of this module

21st – 25th May 2018, Newport, £420
Assessment deadline 2nd July 2018
Applications close 4th May 2018

P08811 Regression Modelling

Course Description

This module will introduce the basic regression model - residual analysis, model building and selection, and the handling of categorical variables. Also, Logistic regression (binary response regression) will be introduced, assessing the model fit and model building and selection. Finally, Multiple regression and Multivariate regression modelling will be introduced.

Relationship with other Modules

Level 7 x.5 module

- Acceptable for
  MSc Data Analytics for Government (TE62A)
  PG Diploma in Data Analytics for Government (TE62B)
  PG Certificate in Data Analytics for Government (TE62C)
  MSc Data Analytics (TE63A)
  PG Diploma in Data Analytics (TE63B)
  PG Certificate in Data Analytics (TE63C)

Learning Outcomes

The Brookes attributes to which the learning outcomes are linked are as follows:

a. Academic literacy
b. Research literacy
c. Critical self-awareness and personal literacy
d. Digital and information literacy
e. Active citizenship

On successful completion of this module, students will be able to:

	Brookes Attribute Developed	Other BAs Developed
1. Critically evaluate the techniques involved in fitting regression models, in order to select and use those most appropriate	a	
2. Evaluate and synthesise findings obtained from applying linear/nonlinear regression models	a	
3. Use a statistical package to implement different regression analysis techniques and interpret and critically assess the outputs	a	

Outline Syllabus

- Simple linear regression models; estimation, prediction and inference
- Linear regression model assumptions; diagnostics and remedies
- Multiple linear regression models
- Regression diagnostics; influential observations
- Multicollinearity; detection and remedial measures
- Categorical independent variables
- Stepwise procedures and model selection criteria
- Logistic regression
- Introduction to multivariate regression models

Teaching Learning and Assessment Strategy

Students completing module P08811 will have been given the opportunity to:

- Attend lectures which will teach the theory and application of multiple linear regression and logistic regression models to sets of data, show the application of multiple linear regression and logistic regression models in a variety of contexts and discuss the limitations of linear regression
- Participate in supervised practical classes and use statistical packages SPSS in supervised practical sessions to implement regression analysis

Assessment comprises of two assessed tasks, a class test (40%) and a coursework (60%).

All the teaching and learning activities are targeted to development and training of specific abilities, as listed in the learning outcomes and to enable the successful completion of all the above assessed tasks.

The teaching and learning activities for module P08811 include lectures, problem-solving/tutorial sessions, practical labs, review and formative feedback sessions.

Attendance and active participation to the theoretical lectures, along with the necessary independent study and students' effort, will allow the development of an adequate knowledge-base of principles of regression models and regression analysis.

Data analysis capabilities, the ability to select and apply relevant statistical regression models to real data are developed through the practical sessions, where students' active participation and independent classroom-work is continuously promoted and expected.

The coursework is designed to assess the student's knowledge and understanding of the course content and will require skills in explaining the purpose of linear/nonlinear regression, choosing, right models, fitting models to statistical data, testing and evaluating the fit, interpreting the results and communicating findings through written reports.

Learning Hours

Scheduled learning and teaching activities
Lectures 12 hours
Practicals/workshops 12 hours

Guided independent study
Directed/independent study 46 hours
Assessment prep 30 hours

Assessment Tasks

Summative Assignments	**Word Count/ Exam Length**	**LOs Assessed**	**Weight %**
Coursework: (Indicative assignment)			
Class Test Topics assessed: Matrix algebra representation of linear regression, calculation and interpretation of model parameters; Diagnostics for curvilinear relationship and influential point; Regression model with one binary and one continuous independent variable, testing understanding of adjusted means and interactions	1 hour	1-2	40%
Assignment Topics will be assessed through the analysis of individualised data sets: Regression of continuous and categorical independent variables; Ability to code dummy variables and interactions, to understand and interpret parameters and to interpret and apply results correctly; Logistic regression	3000 words	1-3	60%

Opportunities for formative assessment and feedback

Students are expected to use lecture content to guide their own study of the module content. Verbal feedback will be given at all practical classes, and students will have the opportunity to discuss the written feedback on their assessed coursework. Students will be given advice following all coursework submissions to help them improve their performance in subsequent coursework.

Indicative Reading List

Author	**Title**	**Publisher**	**Date**
Freund R et al	*Regression Analysis* (2nd ed)	Academic Press	2006
Keith T	*Multiple Regression and Beyond*	Allyn & Bacon	2006
Montgomery D C et al	*Introduction to Linear Regression Analysis* (4th ed)	Wiley	2006

Validation History

Date module first approved: June 2017

P08811 Regression Modelling
Proposed Timetable

Time	Monday	Tuesday	Wednesday	Thursday	Friday
10:00-12:00	**Session 1** Simple linear regression	**Session 4** Residuals analysis: influential observation	**Session 6** Categorical independent variable	**Session 8** Introduction to logistic regression	**Session 10** Generalised linear models 1
12:00-13:00	**LUNCH**				
13:00-14:45	**Session 2** Linear regression model assumptions	**Session 5** Multi-collinearity	**Session 7** Model selection	**Session 9** Further logistic regression	**Session 11** Generalised linear models 2
14:45-15:00	**BREAK**				
15:00-16:30	**Session 3** Multiple linear regression	**Practical session**	**Practical session**	**Practical session**	**Practical session**

Upcoming runs of this module

11th – 15th June 2018, Newport, £420
Assessment deadline 23rd July 2018
Applications close 4th May 2018