Department of Psychology • University of Cape Town

Research Masters - Modern Statistical Methods with R

Convenor - Colin Tredoux

Lecturers – Colin Tredoux, and TBA

Welcome to this course. Statistical methods are important for psychologists and other researchers in the social sciences. Our questions are very complex, and analytic methods are needed that allow us to grasp, reduce, and explain this complexity. We will teach you some of the most important, cutting-edge methods available to researchers in the social sciences in this course, and we will at the same time introduce you to the R Statistical programming environment. R is open source software and is fast becoming a leading statistical platform. It is particularly good for working with large, complex data sets.

We look forward to introducing you to R, and to furthering your knowledge of methods of statistical analysis.

Prescribed books

Wickham, H. & Grolemund, G. (2017). *R for Data Science.* O'Reilly: Sebastapol, CA, USA. This book is available free online, in bookdown format, at r4ds.had.co.nz

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning.* New York: Springer (available for free electronic download at http://www- bcf.usc.edu/~gareth/ISL/). Note that this book is taught as part of a free MOOC, and you can download all video lectures and notes, solutions etc. from the same link).

Kline, R. (2010). *Principles and practice of structural equation modeling.* 2nd or 3rd or 4th edition. New York: Guilford Press. We suggest you buy this book, but it is possible that we will be able to get permission to use the two or three chapters we really need from it.

We will also prescribe some material for the sections on multilevel modeling, and deep learning, more about this closer to the time.

Recommended books

All of the Sage Quantitative Applications in the Social Sciences (QASS) books are available online through the UCT library, and are recommended reading

See also:

Chollet, F., & Allaire, J. J. (2018). *Deep Learning with R.* Manning Books.

Finch, W. H., Bolin, J. E., & Kelley, K. (2014). *Multilevel modeling using R.* CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevelhierarchical models* (Vol. 1). New York, NY, USA: Cambridge University Press.

Loehlin, J. C., & Beaujean, A. A. (2016). *Latent variable models: An introduction to factor, path, and structural equation analysis*. Taylor & Francis.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R.* New York: Springer.

Course pre-requisites

Successful completion of an honours Psychology degree, or at the discretion of the course convener. Note that the undergraduate and honours Psychology degrees at UCT emphasise quantitative methods - if you do not have this background, you should get in touch with the course convenor, who will prescribe material for you to study and prepare before you enter the course. Note that you need to bring a laptop to class for both lectures and tutorials. You should have R and RStudio installed on your laptop, ahead of time (go to r-project.org and rstudio.com respectively to download and install the software – it is free of charge). **Please ensure you do this ahead of time, and that your installations are working.**

Introductory R course: this is an intensive 15 hour introduction to R. It is scheduled for the week of 23 to 27 July 2018. We will provide details as soon as we can. You need to attend all the sessions.

Lectures: in Room 2A, Mondays at 14h00. Note that we start the lectures on 30 July, after an intensive introduction to R course in the week of 23 to 27 July.

Course convener, and lecturer: Colin Tredoux (colin.tredoux@uct.ac.za)

Tutor, and additional lecturer: TBA

Tutorials: in Room 2A, Thursdays from 09h00 to 11h00

Tutorial sessions will be held on a weekly basis. Completion of the weekly assignments will compose 25% of the final grade. These lab sessions are designed to give you hands-on experience running the analyses you will learn about during lectures.

Assessment

There will be weekly tutorial assignments which will make up 25% of the grade. There will be one midterm test (37.5%), and one final test (37.5%). Both tests will be practical in nature.

Course Schedule

WK	DATE	TOPIC	READING
1	23 to 27 July, 2018. This is an intensive 15 hour introduction to R	Introduction to R	Wickham & Grolemund, all chapters, except 22 to 25
2	30 July, 2 August	Multiple Regression	James, Chapters 1 to 3, Wicham and Grolemund, 22 to 25
3	6, 9 August	Multiple Regression	James, Chapter 3
4	13, 16 August	Classification	James, Chapter 4
5	20, 23 August	Resampling methods	James, Chapter 5
6	27, 30 August	Model selection and regularization	James, Chapter 6
7	3, 6 September	Nonlinear models	James, Chapter 7
	Spring vacation		
8	17, 24 September	Tree-based methods	James, Chapter 8
	Mid-semester test, September 21		
9	24, 27 September	Multilevel modeling	TBA
10	1, 4 October	Multilevel modeling	TBA
11	8, 11 October	Structural Equation Modeling	Kline, Chapters 5 & 6
12	15, 18 October	Structural Equation Modeling	Kline, Chapters 7 & 8
13	22, 25 October	Structural Equation Modeling	Kline, Chapters 9 & 10
14	29 October, 1 November	Deep learning / neural networks	TBA
15	8, 11 November	Deep learning / neural networks	TBA
16	15, 18 November	Revision	
	Final exam, 18 November		

Introduction to R - Schedule

Session	DATE and TIME	TOPIC	READING
1	23 July 2018 09h00 – 12h00	Using R Studio Base package operations Descriptive statistics	Wickham & Grolemund W & G), Ch 1
2	24 July 2018 09h00 – 12h00	Writing scripts Visualizing data with ggplot Data transformation	W & G : 2 - 6
3	25 July 2018 09h00 – 12h00	Exploratory data analysis R Studio projects R Markdown reports, notebooks Importing data	W & G : 7, 8, 27, 10, 11
4	26 July 2018 09h00 – 12h00	Tidy data	W & G : 12
5	27 July 2018 09h00 – 12h00	Relational data Strings, factors, dates, times Programming	W & G : 13 – 16, 17 – 18, 19 – 21

All sessions will be in the Bessie Head laboratory, Beattie Building

Assessment - the first assignment for the course is due on Wednesday 1 August at 23h55, and assesses material taught in the Introduction to R course

Department of Psychology • University of Cape Town
Research Masters - Modern Statistical Methods with R

Assignment 1 – Introduction to R

Due Wednesday 1 August at 23h55

Note: all assignments for this course must be submitted (uploaded to Vula, in the correct assignment tab) as either R Markdown documents, or as R Notebooks. In either case they should be compilable (i.e. we must be able to compile them from within R Studio). You are also required to compile the R Markdown document, or R Notebook, as a pdf or html report and to submit it. The report should be in a suitable form and quality for reading by an audience reasonably well informed about statistical methods i.e. it should not be a collection of notes to yourself. For this first assignment, however, only question 7 is amenable to this treatment, but you should answer the other questions in the R Markdown or Notebook document, nonetheless.

Question. You will find the dataset named younglives_peru.xlsx on Vula, in the assignment tab. This dataset contains longitudinal data about the life circumstances of young children in Peru, as measured in 2002, 2005, 2009, 2012, and 2015 (see http://www.younglives.org.uk/). The first ten rows of the data are shown below, for a selection of the variables contained in over 250 data files. A second data file, named younglives_peru_wealth.sav, is also in the assignment tab, and contains data on the so-called 'wealth index', which is a composite measure of the wealth of the family the index child comes from.

1	Import the two datasets into R, and join them, using childid as the key (matching) variable.	**(10)**
2	Select only the variables measuring wealth index, childid, child's language, and receptive vocabulary (ppvtraw). Retain the dataset.	**(3)**
3	Keep only the cases where the child's language was Spanish. Retain the dataset.	**(3)**
4	Create a wide format data file so that the variables wealth index and receptive vocabulary appear for each of rounds 2 to 5, along with childid.	**(10)**
5	Create a new variable that averages the wealth per child across rounds.	**(2)**
6	Rename all your variables so that they are lower case, in a single line of code.	**(2)**
7	What is the relationship between wealth and receptive vocabulary over time? Is it consistent across individual children, or do there appear to be individual differences in the overall level, and rate of change over time? Create one or more visual representations of this relationship with ggplot that will tell your story effectively, but do not report more than four figures. Report descriptive statistics as you think appropriate. Write no more than 2 pages in total for this question.	**(30)**
8	Write a function that computes the 'trimmed mean'. The function should take a vector as an input argument, and a percentage to trim from the top and bottom of the vector as an additional argument. The function should return the trimmed mean as a value. You can check your answer with the R function mean(), but do not include a call to it in your function itself.	**(15)**
9	In answering 1 to 6 above, use the packages dplyr, tidyr, psych, and ggplot2, especially. You should answer each question as a separate section in your report. You should apply what you have learnt from the text *R for Data Science*. Marks will be given for how effectively you do this, the readability of your report, the formatting and style of your R code, and whether it compiles successfully, among other things.	**(25)**