Programing Languages for Statistics

STAT2005

Course Overview

Course Information

Instructor
Dr. WONG, Tat Wing
Office: LSB 112
Email: tw.wong@cuhk.edu.hk

Teaching Assistant
ZHANG, Zheng
Office: LSB G24
Email: 1155149267@link.cuhk.edu.hk

LIANG, Weichen
Office: LSB 141
Email: weichenliang@link.cuhk.edu.hk

Course Description

- This course aims at providing students with a basic knowledge of the high level programming language R and the statistical software SAS.
- Students will learn basic programming concepts in both software.
- For R, students will learn how to write functions and programs; perform simulations to solve statistical problems; produce graphics; understand matrices and vectors.
- For SAS, students will learn the principles of data management and exploratory data analysis.

Learning Outcomes

After finishing the course, students should be able to

1) understand the basic concepts in programming;
2) write commands to create, retrieve, and combine datasets;
3) produce suitable statistical graphics and descriptive statistics;
4) learn to apply computing software to assist data analysis.
5) write functions and programs;
6) perform simulations to solve statistical problem using programming;
7) prepare custom reports.

Why R?

- R is among the most popular open source language for data analysis.
- You can find R packages for a wide variety of statistical tasks
- R packages cover everything from Psychometrics, Genetics to Finance
- You can expect to use R in your future studies, e.g.,
  - STAT3001: Foundation of Financial and Managerial Statistics
  - STAT3008: Applied Regression Analysis
  - STAT4002: Applied Multivariate Analysis
  - STAT4005: Time Series
  - STAT4006: Categorical Data Analysis
  - RMSC4002: Financial Data Analytics with Machine Learning
  - RMSC4003: Statistical Modelling in Financial Markets

Why SAS?

- For your career prospects
- SAS is very popular for business analytics
  - Government bodies use SAS for data management and planning
  - Major banks use SAS to predict credit card fraud in real time
  - Businesses of all kinds use SAS to identify, reward and retain loyal customers
  - Pharmaceutical companies and researchers use SAS to mine and drive vital medical advances
- SAS has been around for almost 40 years, there are many tools built around it, many people trained in it that migration to another platform would be difficult
- SAS, unlike R is not open source and so if there are issues then there is someone accountable

Topic	Contents/ concepts
1. Introduction to R	Elements in statistical programming Basic operations of the R packages Data types: logical, numeric, character Data objects: vector, matrix, list, data frame factor Expressions and operators Built-in functions
2. Simulation and Exploratory Data Analysis	Simulation of random numbers from various distributions Monte Carlo simulation Exploratory data analysis
3. Basic Data Visualization	High level plots Histogram, pie chart, bar chart, dot chart etc. Low level graphics functions
4. Programming in R	Writing functions in R Loops and flow control Logical expressions If statements; for loops; while loops; repeat loops

5. Input, Output and Other Programming Issues	Error handling Formatted input and output Efficient programming Good programming habits
6. Scientific Computation in R	Basic definition of vector and matrix Matrix addition and multiplication Matrix inverse Solving system of linear equations Applications of matrices in statistics Finding roots of nonlinear equations Numerical optimization Numerical differentiation and integration
7. Installing Packages in R *(Optional)*	Pre-installed packages in the R library Installing packages from the Comprehensive R Archive Network (CRAN) Introduction to the *sas7bdat* package, the *rgl* package, the *quantmod* package, and the *tidyverse* package collection

8. Introduction to SAS	SAS program SAS data sets DATA and PROC steps SAS data libraries
9. Creating SAS Data Set	Creating SAS data set using Viewtable window Creating SAS data set using Import wizard Creating SAS data set using DATA step List input, column input and format input Retrieving an existing SAS data set
10. Data Manipulation in DATA Step	Basic SAS operators Assignment statements SAS built-in functions Conditional execution: IF statements Conditional execution: SELECT statements DO-Loops DO-WHILE Loops

11. Controlling Output Contents in DATA Step	Subsetting observations with OUTPUT and DELETE statements Subsetting variables with DROP and KEEP statements Creating more than one data set Conditional termination of DATA step with STOP statement Data set options
12. Combining SAS Data Sets *(Optional)*	Concatenation Interleaving One-to-One merge Matched merge Updating

Assessment Scheme

Type	Description	Weight
Assignments	4 sets of assignments will be given	20%
Mid-term Examination	October 25, 2023, cover Chapter 1-5	30%
Final Examination	To be arranged by the University, cover Chapter 1-6, 8-11	50%

Self-paced learning exercises

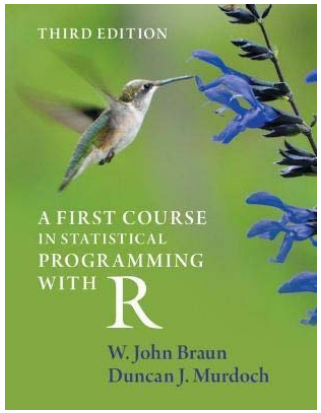
- There are self-paced exercise problems for you to practice what you have learnt in each chapter.
- They are optional, but you are welcome to discuss with us if you have any difficulty working on them.
- Solutions will also be provided for your references, but you are encouraged to try them before reading at the solutions.

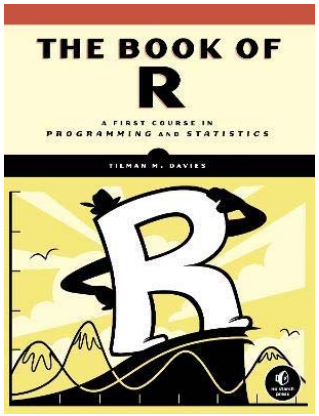
Academic honesty and plagiarism

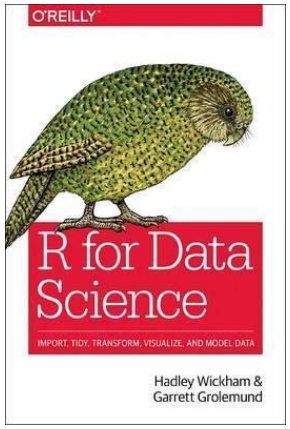
- The Chinese University of Hong Kong places very high importance on honesty in academic work submitted by students, and adopts a policy of **zero** tolerance on cheating and plagiarism.
- Any related offence will lead to disciplinary action including termination of studies at the University. Relevant information can be allocated via: http://www.cuhk.edu.hk/policy/academichonesty/

References for R

1) Braun, W. J., & Murdoch, D. J. (2021). A first course in statistical programming with R. Cambridge University Press.
2) Davies, T.M. (2016). The Book of R: A First Course in Programming and Statistics. No Starch Press.
3) Venables, W. N., Smith, D. M., & R Development Core Team. (2009). An introduction to R.
4) Wickham, H., & Grolemund, G. (2016). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media

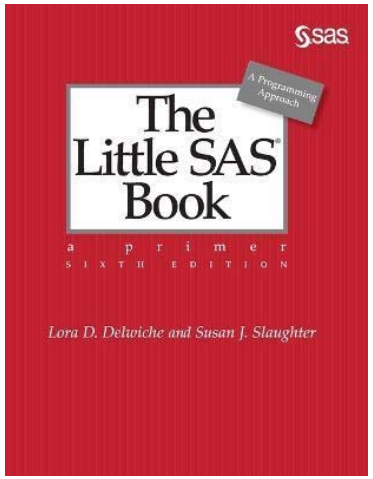


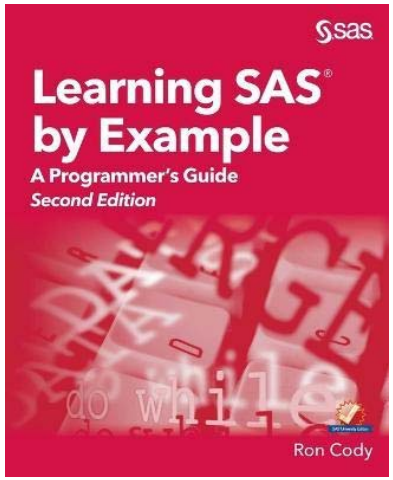




References for SAS

1) Delwiche, L. D., & Slaughter, S. J. (2019). The little SAS book: a primer. SAS Institute.
2) Cody, R. (2018). Learning SAS by Example: A Programmer's Guide. SAS Institute.
3) SAS Institute. (2019) SAS Certified Specialist Prep Guide: Base Programming Using SAS 9.4. SAS Institute.





Computer Software

You should have R and SAS available in your computer

- R is available for free from the official website:

  https://cran.r-project.org/

- SAS is a commercial software. You are suggested to use it online via SAS OnDemand for Academics

  https://www.sas.com/en_us/software/on-demand-for-academics.html

  You need to create an account with SAS.

  You could also get an offline version with SAS Educational License (MS Windows only) from the Department of Statistics if

  i. You are student from the Department of Statistics (STAT, RMSC, QFRM, CDAS major), or
  ii. You have enrolled in STAT2005 after the add/drop period.

  Please contact Mr. Julian Wong (LSB G23, julianwong@cuhk.edu.hk) for installation of the software.

Purpose of this course

- To reach a point where these software become another useful tool for you to use when “doing” statistics
- We hope that this is the beginning of your regular use of statistical software for the rest of your degree
- Without regular use or practice, languages tend to be forgotten
- So please continue to use R and SAS well after the final exam for this course. Then you will be able to say on your CVs that you are “fluent” in them.