Chapman & Hall / CRT Statistics Books: Proposal

## Proposed title

**Visualizing Multivariate Data and Models in R**

## Author

Michael Friendly, York University, [friendly@yorku.ca](mailto:friendly@yorku.ca), with assistance from Udi Alter, York University, [udialter@yorku.ca](mailto:udialter@yorku.ca)

## Contents

An in-progress outline with some description is attached. Reviewers who wish to see the current state can visit the GitHub repository, <https://github.com/friendly/Vis-MLM-quarto>. At some point soon, the rendered current version of the book will be made available online.

## Subject

This book is about graphical methods developed recently for multivariate data, and their uses in understanding relationships when there are several aspects to be considered together. Data visualization methods for statistical analysis are well-developed for simple linear models with a single outcome variable. However, with applied research in the social and behavioral sciences, it is often the case that the phenomena of interest (e.g., depression, job satisfaction, academic achievement, childhood ADHD disorders, etc.) can be measured in several different ways or related aspects.

For example, if academic achievement can be measured for adolescents by reading, mathematics, science and history scores, how do predictors such as parental encouragement, school environment and socioeconomic status affect all these outcomes? In a similar way? In different ways? In such cases, much more can be understood from a multivariate approach that considers the correlations among the outcomes. Yet, sadly, researchers typically examine the outcomes one by one which often only tells part of the data story.

The statistical leap from univariate to multivariate methods is quite simple because the classical models for ANOVA and regression:

* Generalize directly to analogous models for MANOVA and multivariate regression,
* Statistical multivariate tests are straightforward extensions of standard *F*- and *t*-tests but give richer detail about the relationships among response variables.
* Univariate intervals become ellipses (and ellipsoids!), showing location, covariation, and precision.

This book brings together a collection of novel techniques I, and others, have developed over the past 15 years and implemented in mature R packages. Some key substantive features of the book are:

* Statistical data visualization is cast in a general framework by goal (see the data, visualize a model, diagnose problems), rather than a categorization by graphic types.
* Data visualization is seen as a combination of exposure—plotting the raw data--- and summarization--- plotting statistical summaries--- to highlight what should be noticed. For example, data ellipses and confidence ellipses are widely used as simple, effective summaries of data and fitted model parameters. When the data is complex, the idea of visual thinning can be used to balance the tradeoff.
* These graphical tools can be used to understand or explain a wide variety of statistical concepts, phenomena, and paradoxes such as Simpson’s paradox, effects of measurement error, and so forth.
* The HE (“hypothesis – error”) plot framework provides a simple way to understand the results of statistical tests and the relations among response outcomes in the multivariate linear model.
* Dimension reduction techniques such as PCA and discriminant analysis are presented as “multivariate juicers,” able to squeeze the important information in high-dimensional data into informative two-dimensional views.

Some key pedagogical features are:

* Conceptual diagrams are used throughout the book to illustrate the main ideas behind statistical and graphical methods.
* Infoboxes or “callouts” are used to explain related ideas to interfere little with the main flow of the narrative.
* All examples are “live,” in the sense that they are directly produced by the R code in the book. The online version will allow these to be either shown directly in the text or toggled on/off by the reader.

## Related books

There are several textbooks and monographs on multivariate data analysis, but none focus mainly on data visualization and graphical methods. Everitt & Hothorn (2011) do use R, but not with any incisive graphics.

A few texts on multivariate methods are:

* Timm, N. H. (1975). *Multivariate Analysis with Applications in Education and Psychology*. Wadsworth (Brooks/Cole)
* Everitt, B., & Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer New York
* Tabachnick, B. G., & Fidell, L. S. (2019). *Using Multivariate Statistics* (7 ed.). Pearson.

## Audience

This book is aimed at advanced undergraduates, graduate students and researchers wishing to learn and apply multivariate techniques.

## Production

At this stage of writing, it is difficult to predict the number of pages, but I expect it to be in the range 300-400, with perhaps 100 figures. Because this is a book on data graphics, color will be essential throughout.

The book is being written using RStudio, Rmarkdown and Quarto. In writing, I’m using primarily HTML format and would like to make pre-publication drafts publicly available for comment and suggestions by colleagues. I expect to be able to have this translated to LaTeX, using the standard CRC krantz.cls, though some assistance with this might be necessary.

I expect to complete a first draft of the book in the summer of 2024.

## Reviews

John Fox, [jfox@mcmaster.ca](mailto:jfox@mcmaster.ca)

Georges Monette, [georges@yorku.ca](mailto:georges@yorku.ca)

Dianne Cook, [dicook@monash.edu](mailto:dicook@monash.edu)

Wayne Oldford, [rwoldford@uwaterloo.ca](mailto:rwoldford@uwaterloo.ca)

## Marketing

Key features:

Keywords: data visualization; multivariate analysis; data analysis; statistics; R

Markets:

* STA07A-Statistics-Statistical Theory & Methods
* STA12A-Statistics-Statistics for the Social and Behavioral Sciences
* CMS08-Computer Science & Engineering-Visualization