Journal of Statistical Software

September 2009, Volume 31, Book Review 2. *http://www.jstatsoft.org/*

Reviewer: Chiara Sabatti
University of California at Los Angeles

Reviewer: Christophe Lalanne
INSERM

Applied Statistical Genetics with R for Population-Based Association Studies

Andrea S. Foulkes
Springer-Verlag, New York, 2009.
ISBN 978-0-387-89553-6. 252 pp. USD 59.95 (P).
http://people.umass.edu/foulkes/asg.html

We have two separate and independent reviews for this book. They are both provided below.

Review by Chiara Sabatti

The book aims at filling a real gap in the literature. Statistical genetics topics are presented only in specialized texts that can often appear as unapproachable to graduate students in statistics, unless they have made a clear commitment to this branch of the discipline. In addition, while the statistical genetics community has been for a long time conscious of the importance of releasing software for anyone to use, and assuring the reproducibility of research, this software consists of a series of specialized packages, so that often the beginner is lost in the need of learning the syntax of many different programs. This specialization is largely unavoidable: statistical genetics deals with data with a complex dependence structure, determined by the familial relations of the subjects, the joint transmission of adjacent portions of the genome, etc.

Both explaining the implications of this dependence, and writing code to carry out data analysis taking family structure into account, require substantial time and the introduction of concepts, terminology, and notation unfamiliar to the average statistician.

In "population-based association studies"—the specific area investigated in this text, as the subtitle specifies—a number of these complications are avoided. In particular, subjects can, by and large, be considered as independent samples from a population. "Population-based association studies" are, then, in many ways, an ideal topic to introduce the student to the challenges of statistical genetics. Furthermore, relying on standard software, as R, becomes possible, and it helps to underscore the connections between the newly introduced problems in statistical genetics and more familiar statistical models.

Nevertheless, using R to illustrate statistical genetics computation, and data analysis, poses some challenges. By and large, most of the attention of geneticists is currently devoted to genome wide association studies (GWAS). These require genotyping hundreds of thousands of markers in thousands of individuals, generating datasets whose sizes make them difficult to manipulate in R. Possibly because of this, the author indicates that the primary focus of the book is on "candidate gene studies", which involve a much more limited number of markers. While this choice is understandable, it implies that readers are not exposed to what are the most current challenges of "applied statistical genetics".

One of the advantages of using R is the access to a large and ever-growing collection of specialized packages. *Applied Statistical Genetics with R* takes full advantage of this: just in Chapter 2 and 3, we see used **coin**, **genetics**, and **LDheatmap**, for example. Relying on packages, however, means that not much emphasis is put on actual coding of the different functions and methodologies presented, and that the reader/student has to deal with a syntax and data formats that are not uniform. This is perhaps made more extreme by the fact that the writer is not the author of the vast majority of the exemplified software. I found it disappointing that packages like **snpMatrix**, or **Rserve**, which aim at handling GWAS data, are simply mentioned without illustration.

After three introductory chapters on basic statistical and genetic concepts and association studies, the book deals with the problems of multiple comparison, unknown phase, and model building and prediction in high dimension: topic choices that I find relevant and stimulating.

Chapter 1 through 3 are interconnected and aim to outline the scientific questions behind genetic association studies, as well as introducing basic genetics and statistical concepts. I did not find them uniformly clear and accurate. The author seems to try to address a very broad (too broad) audience: one that needs to have explained the concept of "mean", and to be told how to read the symbol of factorial, and is as well ignorant of the basic elements of genetics. In my view, the natural audience for this text are statistics graduate students, not applied genetic researchers, who would not have any preference for the use of R and would be hardly persuaded to rely on software that cannot handle GWAS data. With this audience in mind, I would have avoided introducing really basic statistical concepts and provided a less concise view of the scientific questions tackled in genetics.

Chapter 4 (on multiple comparison), 6 (classification and regression trees), and 7 (further topics in high dimensional data analysis) give a nice introduction of these topics. Genotype data is used in the examples, but the overall reference to statistical genetics is somewhat superficial. For example, the dependency between association tests generated by association between neighboring markers is mentioned and few approaches to account for this are summarized, but only at the end of Chapter 4 and without data examples: still, this is one of the most distinguishing features of multiple comparisons in gene mapping problems. The same dependence is ignored when describing methods of imputation for missing genotypes in the following chapters.

Chapter 5 deals with a genuinely genetics problem: the reconstruction of unknown phase from multilocus genotype data. There are not many textbooks that present this material, and hence I found this chapter one of the most valuable in the book. Unfortunately, some of the most interesting algorithms are not coded in R and hence the data analysis examples are limited.

One interesting aspect of the book is that it contains a number of references to recent literature

dealing with the scientific problems presented, even when the content of this literature is not explicitly discussed. The textbook, then, serves as a starting point for further reading and this is a great way of introducing statistical genetics problems to a general audience. From this point of view, and in many other ways, the book feels like the transcription of lecture notes of an introductory class. This is certainly the way in which many great texts were developed. It is this reader's impression that this book has room to grow in a more mature text in following editions.

Review by Christophe Lalanne

Andrea S. Foulkes is an Associate Professor of Biostatistics at the University of Massachusetts School of Public Health and Health Sciences, and she authored the **mirf** package (for multiple imputation and random forests).

This book provides a gentle introduction to genome-wide association studies (GWAS) within both a theoretical and methodological perspective. It will especially be a useful ressource to those interested in the ever growing interdisciplinary approach to "genetic epidemiology", which according to Thomas (2004) involves the "study of the joint action of genes and environmental factors in causing disease in human populations and their patterns of inheritance in families". Since this book focuses on association studies and gene variants, or single nucleotide polymorphisms (SNPs), it will also nicely complement handbook by Gentleman, Carey, Huber, Irizarry, and Dudoit (2005) on statistics in genetics within the **Bioconductor** project (`http://www.Bioconductor.org/`) as the latter is rather centred on microarrays technology and proteomics. Data sets used throughout the book and associated R scripts can be found on the companion website.

GWAS focus on the relationships between the genetic sequence information (loosely refered to as the genotype) and a trait or phenotype (e.g., cholesterol level or related disease) measured in vivo or in vitro in unrelated individuals. Whereas linkage or candidate gene studies rely on gene expression product, here single base pair changes occuring in at least 1% of the population are used as a proxy to reflect spatial loci of variability on the whole genome. What are the challenges raised by GWAS? On the one hand, the ever growing mass of data collected by various centers calls for robust testing procedures, since high-throughput SNP-chip technology can now deliver 10^6 probes at once and that human genotype might be tested against several phenotypes of interest. On the other hand, the end user also needs computationally efficient reduction techniques, such that meaningful associations between markers can be uncovered and interpreted from these high-dimensional data sets.

The book is divided into seven parts organized so as to offer a progressive introduction into relevant methodological aspects of the statistical analysis of genotype data, with an emphasis on the use of association or population-based studies in health-related applications. The first two chapters review genetic theory and the underlying statistical principles aiming at discovering association between candidate genes or single polymorphisms and a disease of interest, here for example HIV infection. It also introduces the reader to the challenges relevant to this kind of multivariate data as well as the need to consider the natural stratification that arises as a consequence of multi-centric enrollment involving people of differing race and ethnies.

Chapter 3 provides a deeper introduction to SNPs data and concepts related to genotype such as linkage disequilibrium (LD), which refers to the association between the alleles present at

each of two sites on a genome, and the Hardy-Weinberg equilibrium (HWE) theorem allowing to test for the independence of alleles at a given locus between two homologous chromosomes. LD is of particular importance because a set of SNPs may not directly explain the variations observed in the trait under consideration but they may be correlated with a true disease causing variant or a known biomarker instead. Application of HWE is useful when trying to uncover population substructures or genotyping errors. Missing data (arising as consequence of low or false genotyping rate), are considered in a separate chapter although their analysis as well as imputation methods are part of data preprocessing steps.

Chapter 4 tackles the problem of multiple comparisons, and provides a concise and elegant summary of the various procedures derived so far. It discusses the notion of type I error (i.e., falsely rejecting the null) and type II error (false negatives), and the control of family-wise error or false discovery rate (FDR) using single-step and step-down adjustment methods. Examples include the conservative Bonferroni correction or more elaborated techniques like conditional FDR. Resampling-based methods are also discussed before concluding the chapter. Suggested additional readings might include Dudoit and van der Laan (2008), for example.

Chapter 5 deals with the problem of haplotype block reconstruction with EM algorithm and the study of haplotype-trait association. The latter is usually framed in a two-stage (likelihood-based) regression modeling approach, involving (1) the estimation of posterior haplotype probabilities based on genotype information alone and (2) the estimation of inferred haplotype-trait association. The last two chapters are concerned with the multivariate modeling of so-called high-dimensional data both at the individual and gene level. Included modeling approaches are: Classification and regression trees, random forests, logic regression, multivariable adaptive regression splines, and bayesian variable selection.

The text is regularly interleaved with numerical illustrations, which gives the reader the opportunity to assimilate part of the theoretical groundings. Since R snippets are kept deliberately oversimplified, even the non-experienced R user may be able to go beyond illustrative applications on his/her own. Useful R packages for statistical genetics are listed in the appendix but the reader is encouraged to also take a look at the Comphrehensive R Archive Network Task View on Genetics (`http://CRAN.R-project.org/view=Genetics`). Actually, both **GenABEL** and **SNPassoc** allow to test for various genetic models (dominant, co-dominant, recessive, etc.) on whole-genome scan.

This new book in the Springer *Use R!* series certainly fills the lacking R ressources on this rapidly evolving field in statistical genetics, although it would have benefited from more illustrations. Indeed, genetic data not only challenge the computational efficiency of modern regression techniques but also call for high-dimensional visualization techniques. Given the number of cofactors that might be considered at the same time, such graphical devices should be interactive and provide dynamic facilities, as for example what is proposed in **GGobi** (`http://www.ggobi.org/`) although this particular software may suffer from the huge volume of data to be displayed. Meta-analysis on previous GWAS is also becoming an important issue before making robust inference on gene-disease association (e.g., Ioannidis, Thomas, and Daly 2009). Likewise, a dedicated chapter on data storage and computing issues would have been very interesting as GWAS call for very fine-tuned optimization and robust data structures under R, especially when one is interested in testing a large number of SNPs against several phenotypes. In this regard, the **snpMatrix** package David C. Clayton and the **GGbase** package by Vince J. Carey offer efficient solutions for the importation and manipulation of SNPs data. Needless to say, R proves to be a viable alternative to widely used softwares like

PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/), **MERLIN** (http://www.sph.umich.edu/csg/abecasis/merlin/) or Stata (http://www.stata.com/). In the near future, it would be interesting to see how statistical genetics with R could benefit from the use of MPI or GPU architectures.

References

Dudoit S, van der Laan MJ (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer-Verlag, New York.

Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S (2005). *Bioinformatics and Computational Biology Solutions Using R and **Bioconductor***. Springer-Verlag, New York.

Ioannidis JPA, Thomas G, Daly MJ (2009). "Validating, Augmenting and Refining Genome-Wide Association Signals." *Nature Reviews Genetics*, **10**, 318–329.

Thomas DC (2004). *Statistical Methods in Genetic Epidemiology*. Oxford University Press, Oxford.

Reviewer:

Chiara Sabatti
Departments of Statistics and Human Genetics
University of California Los Angeles
E-mail: csabatti@ucla.edu
URL: http://www.genetics.ucla.edu/labs/sabatti/home/

Christophe Lalanne
INSERM U669 and Department of Clinical Research
Hôpital Saint-Louis (Paris)
E-mail: ch.lalanne@gmail.com
URL: http://www.aliquote.org/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 31, Book Review 2 *Published:* 2009-09-19
September 2009