
Preface

As an area of statistical application, environmental epidemiology and more specifically, the estimation of health risk associated with the exposure to environmental agents, has led to the development of several statistical methods and software that can then be applied to other scientific areas. The statistical analyses aimed at addressing questions in environmental epidemiology have the following characteristics. Often the signal-to-noise ratio in the data is low and the targets of inference are inherently small risks. These constraints typically lead to the development and use of more sophisticated (and potentially less transparent) statistical models and the integration of large high-dimensional databases. New technologies and the widespread availability of powerful computing are also adding to the complexities of scientific investigation by allowing researchers to fit large numbers of models and search over many sets of variables. As the number of variables measured increases, so do the degrees of freedom for influencing the association between a risk factor and an outcome of interest.

We have written this book, in part, to describe our experiences developing and applying statistical methods for the estimation for air pollution health effects. Our experience has convinced us that the application of modern statistical methodology in a reproducible manner can bring to bear substantial benefits to policy-makers and scientists in this area. We believe that the methods described in this book are applicable to other areas of environmental epidemiology, particularly those areas involving spatial-temporal exposures.

In this book, we use the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) and Medicare Air Pollution Study (MCAPS) datasets and describe the R packages for accessing the data. Chapters 4, 5, 6, and 7 describe the features of the data, the statistical concepts involved, and many of the methods used to analyze the data. Chapter 8 then shows how to bring all of the methods together to conduct a multi-site analysis of seasonally varying effects of PM_{10} on mortality.

A principal goal of this book is to disseminate R software and promote reproducible research in epidemiological studies and statistical research. As

a case study we use data and methods relevant to investigating the health effects of ambient air pollution. Researching the health effects of air pollution presents an excellent example of the critical need for reproducible research because it involves all of the features already mentioned above: inherently small risks, significant policy implications, sophisticated statistical methodology, and very large databases linked from multiple sources. The complexity of the analyses involved and the policy relevance of the targets of inference demand transparency and reproducibility.

Throughout the book, we show how R can be used to make analyses reproducible and to structure the analytic process in a modular fashion. We find R to be a very natural tool for achieving this goal. In particular, for the production of this book, we have made use of the tools described in Chapter 3.

All of the data described in the book are provided in the **NMMAPS**lite and **MCAPS** R packages that can be downloaded from CRAN.¹ We have developed R packages for implementing the statistical methodology as well as for handling the databases. Packages that are not available from CRAN can be downloaded from the book's website.²

We would like to express our deepest appreciation to the many collaborators and students who have worked with us on various projects, short courses, and workshops that we have developed over the years. In particular, Aidan McDermott, Scott Zeger, Luu Pham, Jon Samet, Tom Louis, Leah Welty, Michelle Bell, and Sandy Eckel were all central to the development of the software, databases, exercises, and analyses presented in this book. Several anonymous reviewers provided helpful comments that improved the presentation of the material in the book. In addition, we would like to thank Duncan Thomas for many useful suggestions regarding an early draft of the manuscript. Finally, this work was supported in part by grant ES012054-03 from the National Institute of Environmental Health Sciences.

Baltimore, Maryland,
April 2008

Roger Peng
Francesca Dominici

¹ <http://cran.r-project.org/>

² <http://www.biostat.jhsph.edu/~rpeng/useRbook/>

Statistical Methods for Environmental Epidemiology with
R

A Case Study in Air Pollution and Health

Peng, R.D.; Dominici, F.

2008, X, 144 p., Softcover

ISBN: 978-0-387-78166-2