Kentaro Matsuura

Bayesian Statistical Modeling with Stan, R, and Python

Springer

Bayesian Statistical Modeling with Stan, R, and Python

Kentaro Matsuura

Bayesian Statistical Modeling with Stan, R, and Python

Springer

Kentaro Matsuura
HOXO-M Inc.
Tokyo, Japan

ISBN 978-981-19-4754-4 ISBN 978-981-19-4755-1 (eBook)
https://doi.org/10.1007/978-981-19-4755-1

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

About This Book

With the recent rise in popularity of Bayesian statistics, many books on this subject have been published. However, many of these books either only contain basic information, or contain too complicated formulas that are not easy to extend in practice. This book is, therefore, a departure from those books, and is intended to be a very practical book on Bayesian statistical modeling with real-world data analysis.

This book is about *Stan*, a software that conducts statistical modeling in the framework of Bayesian statistics. We also introduce how to use its R package *CmdStanR* and Python package *CmdStanPy*.

Statistical modeling is a process of fitting mathematical models with probabilistic distributions to the observed data, in order to understand phenomena and to further make predictions. In earlier days, such a method was not considered to be a practical method for the analysis of the real-world data, due to several reasons. For instance, it is not easy to collect a large amount of data and the computational time of the method was very long. In addition, fitting a complex model needs high mathematical skills, because the analytical solution is difficult and implementing its solution is even harder.

Yet nowadays, these are not the problems anymore. Today we can obtain a large amount of data relatively easily, and computers have become powerful and efficient compared to the earlier days. Also luckily, several programming languages that are specifically designed for statistical modeling have been developed. Therefore, statistical modeling has become one of the most efficient approaches for data analysis in the recent studies.

In this book, we use a new programing language, Stan, to conduct statistical modeling. Stan incorporates an excellent algorithm, and importantly, it is under active development by an awesome community. Its R and Python packages (CmdStanR and CmdStanPy) are released in parallel, which provide an easy starting point for the beginners. In Stan, many models can be written in about 30 lines of code. This includes not only the basic models such as multivariate regression and logistic regression, but also more advanced modeling such as hierarchical models, state-space models, Gaussian processes, and many others. Moreover, it enables the problem-specific model extensions based on the problem that every user is trying to solve for. We believe that the skillsets and the way of thinking about the real-world problem using Stan in this book will be helpful for every reader, even if the syntax of the Stan changes in the future, or one day another statistical modeling framework comes out and replaces Stan.

Chapter Structure

This book is structured in a way so that by reading from the beginning to the end, readers can obtain a systematical knowledge and skillsets for statistical modeling using Stan. The chapter structure in this book is organized as follows (see Fig. 1).

Roughly, the chapters can be divided into four large sections. Chapters 1 and 2 are the introduction, and we mainly focus on the theoretical background on statistical modeling and Bayesian inference. Chapters 3–5 are the guide of how to use Stan, specifically for the beginners. We introduce how to use Stan itself, as well as CmdStanR, and CmdStanPy through the example of simple regression analysis such as multiple linear regression and logistic regression. Chapters 6–10 are for more advanced users, and the important components and techniques to master statistical modeling are introduced. Chapters 6, 8, and 10 are the main chapters, which increase your statistical modeling approach repertoire. Chapters 7 and 9 introduce approaches for model improvement. Chapters 11–14 focus on advanced topics, which are applied and discussed quite widely when conducting the real data analysis.

We listed a summary of contents in each chapter: In Chap. 1, we give a brief introduction to statistical modeling and its features, as well as the recommended statistical modeling workflow. In Chap. 2, we concisely summarize the terminologies used in Bayesian inference and MCMC. These terminologies will be necessary to understand the later chapters. In Chap. 3, we introduce how to install Stan and describe its basic syntax. In Chap. 4, we show how to use Stan with its interface CmdStanR or CmdStanPy through an example of simple linear regression. Specifically, we discuss how to generate an MCMC sample, and obtain Bayesian confidence intervals and Bayesian prediction intervals from the MCMC sample. We also show how to explain

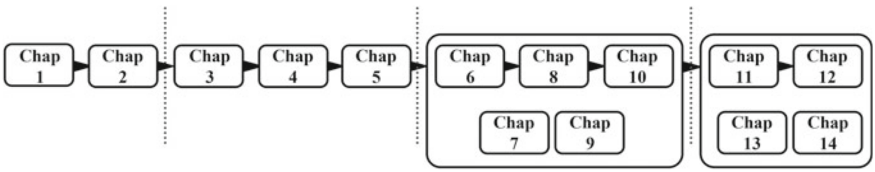

Fig. 1 A flow diagram showing the connections between chapters

the obtained inference results. In Chap. 5, we follow the statistical modeling workflow introduced in Chap. 1. The examples of multiple linear regression, logistic regression, and Poisson regression are also given. We also show examples of how to use plots to check the used models. In Chap. 6, we deviate slightly from the Stan itself and introduce basic probabilistic distributions. We mainly focus on the distributions that are used frequently in analytical practice. In Chap. 7, we summarize some potential issues and troubles when applying regression models to realistic problems. This chapter also includes the introductions on non-linear models, censored data, and data with outliers. In Chap. 8, we introduce hierarchical models (multilevel models). This type of model is used to account for group differences and individual differences. It will be one of the most widely used model types in the future. In Chap. 9, we introduce the troubleshooting approach when the MCMC does not converge, with the focus on setting the weakly informative priors. In Chap. 10, we introduce how to address discrete parameters. The current version of Stan has the limitation that is not able to use the discrete parameters, and we introduce how to solve such issues. In Chap. 11, we use state-space models for time series data. In Chap. 12, we use Markov Random Field models and Gaussian Processes for spatial data. In Chap. 13, we introduce how to efficiently use the MCMC sample from a posterior distribution (or a predictive distribution). In Chap. 14, we discuss some advanced topics, specifically we choose some examples from survival time analysis, matrix factorization, and model selection using information criteria.

Prerequired Background Knowledge

We do not require the readers to have the previous experience on statistical modeling. However, since building models and checking the models are time-consuming and labor-intensive processes, we hope the readers have the mindset and enthusiasm to enjoy such processes.

To conduct statistical modeling, it requires a wide range of knowledge and skills such as the statistics knowledge (especially probability distributions), skill to imagine the mechanisms of the observed phenomenon and to express them using mathematical formulas as well as programming languages. One of the aims of this book is to extend and improve these types of skillsets. That being said, in this book, we assume that the readers have the following basic knowledge and skillsets already as the requirement.

- The fundamental knowledge of probability and statistics: probability, probability distributions, probability density functions, conditional probability, joint probability, marginal probability, correlation coefficients. The knowledge of statistical tests is not required. The next section introduces each of these terminologies briefly, but it would also be helpful to look up some web resources including Wikipedia for those who are less familiar.
- Programming skills: You should be familiar with either R or Python. Specifically, we require the readers to be able to conduct fundamental data processing and data visualizations.

The Terminologies and Symbols Used in This Book

Sum and Products

Σ represents the summation:

$$\sum_{k=1}^{K} a_k = a_1 + a_2 + ... + a_K,$$

and Π represents the product:

$$\prod_{k=1}^{K} a_k = a_1 \times a_2 \times ... \times a_K.$$

Probability Distribution

Probability distribution is the expression of how likely a random variable takes each of the possible values. It also can be simply called as distributions. The probability distribution of random variable a is written as $p(a)$.

Probability Mass Function (PMF)

When the random variable is a discrete variable, the probability distribution is called a *probability mass function.*

Probability Density Function (PDF)

When the random variable is a continuous variable, the probability distribution is called a *probability density function.* It is also called a *density*. Note that if the probability density function is denoted as $p(x)$, the value at $p(x)$ (for instance, $p(x = 0)$) is not a probability. Rather, the integral of $p(x)$ is a probability. For instance, the probability that $0 \leq x \leq 0.2$ can be expressed as follows:

$$\int_{0}^{0.2} p(x)dx$$

Joint Distribution

When there are multiple random variables, a *joint distribution* expresses the probability that how likely each of these random variables takes the possible values for them. It is a type of probability distribution. When there are two random variables a and b, we write their joint distribution as $p(a, b)$, and when we have total of K random variables $\theta_1, \theta_2, \ldots, \theta_K$, we write their joint distribution as $p(\theta_1, \theta_2, \ldots, \theta_K)$.

Marginalization and Marginal Distribution

Marginalization is the elimination of variables by summing up or integrating a joint distribution with respect to random variables in the joint distribution over their possible values. The probability distribution obtained by marginalization is called a *marginal distribution*. For instance, by summing up the joint probability with respect to a discrete random variable a, we can obtain the marginal distribution $p(b)$ from $p(a, b)$:

$$p(b) = \sum_a p(a, b)$$

Similarly, when a is a continuous random variable, by integrating the joint distribution with respect to a, we can obtain the marginal distribution $p(b)$ from $p(a, b)$:

$$p(b) = \int p(a, b) da$$

Figure 2 shows the relationship between joint distribution and marginal distribution. The example where both a and b are discrete variables is shown in the left table of Fig. 2. The table shows the distribution of the species and sexes of animals in a pet store. The region surrounded by the solid gray line is the joint distribution $p(Animal, Sex)$. The region surrounded by the dotted gray line on the right side is the marginal distribution $p(Sex)$, which is obtained by summing up the probability with respect to *Animal*. Similarly, the region surrounded by the dashed gray line on the bottom side is the marginal probability *p*(*Animal*), which is obtained by summing up the probability with respect to Sex. It is called a marginal distribution because we commonly show the summed probability on the margins of a table as shown here. The right plot on Fig. 2 shows when both a and b are continuous variables. The contour lines represent the joint distribution $p(a, b)$. The density plot on the right side of the plot is $p(b)$ which is obtained by integrating $p(a, b)$ with respect to a. The density distribution on the top is $p(a)$, which is obtained by integrating $p(a, b)$ with respect to b.

Conditional Probability Distribution

We consider a joint distribution $p(a, b)$. The distribution of the random variable a, when a value b_0 is given to the random variable b, is called the *conditional probability distribution*, and is written as $p(a|b_0)$. The following equation holds true:

$$p(a|b_0) = \frac{p(a, b_0)}{p(b_0)}$$

Figure 3 shows the conditional probability distribution using the data from Fig. 2. The region surrounded by the solid gray line on the left table on Fig. 3 is the conditional probability distribution $p(a|b_0)$ when the sex b_0 = Female is given. The density

b \ a	Cats	Dogs	Birds	Total
Male	0.19	0.23	0.12	0.54
Female	0.22	0.16	0.08	0.46
Total	0.41	0.39	0.20	1.00

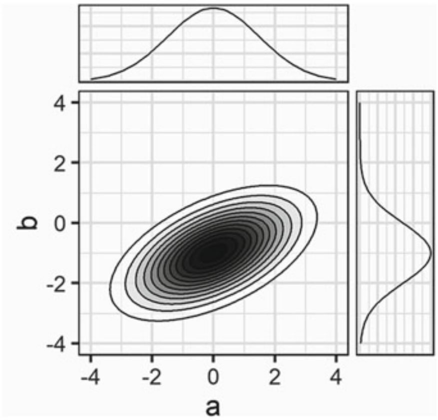

Fig. 2 Examples of joint distribution and marginal distribution

b \ a	Cats	Dogs	Birds	Total
Male				
Female	$\frac{0.22}{0.46}$	$\frac{0.16}{0.46}$	$\frac{0.08}{0.46}$	$\frac{0.46}{0.46}$

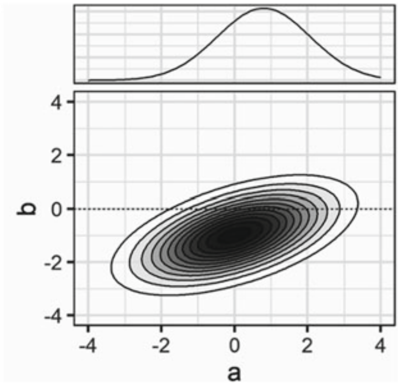

Fig. 3 Examples of conditional probability distribution

function on the top right of Fig. 3 is the conditional probability distribution $p(a|b_0)$ when $b_0 = 0$ is given.

In addition, when the location and the shape of the probability distribution of a random variable y is determined by parameter θ, we write it as $p(y|\theta)$.

Notation of $y \sim p(y)$

The expression of $y \sim p(y)$ represents that the values of a random variable y is generated probabilistically from a probability distribution $p(y)$. In other words, a random variable y follows a probability distribution $p(y)$. For instance, if we write a Poisson distribution with parameter λ as Poisson$(y|\lambda)$, then $y \sim$ Poisson$(y|\lambda)$ represents that the values of y is probabilistically generated from Poisson$(y|\lambda)$, or y follows Poisson$(y|\lambda)$. We can also write it by omitting y, like $y \sim$ Poisson(λ).

Independence

When we say that two random variables a and b are mutually *independent*, we mean that this equation holds true: $p(a, b) = p(a)p(b)$. Further, from the definition of conditional probability distribution, this is equivalent to $p(a|b) = p(a)$.

Normalization

In this book, we use the term *normalization* to refer multiplying a constant to a function (or dividing by a constant), for the function to satisfy the condition of the probability density function. The constant that is used for normalization is called the normalization constant.

Partial Derivative

Consider a function of two variables $f(\theta_1, \theta_2)$. We consider the derivative by fixing θ_2 as a constant and changing θ_1

$$\lim_{\Delta\theta_1 \to 0} \frac{f(\theta_1 + \Delta\theta_1, \theta_2) - f(\theta_1, \theta_2)}{\Delta\theta_1}$$

and this is called *partial derivative* of $f(\theta_1, \theta_2)$ with respect to θ_1. The same would apply for multivariate function $f(\theta_1, \theta_2, \dots, \theta_K)$, and when taking its derivative, we only need to consider all variables as constants, except for θ_1.

Uppercase Letters

In this book, all the uppercase letters such as Y and Age refer to the given data, unless otherwise specified.

Vectors and Matrices

In this book, vectors are denoted with arrows on top of each letter ($\overrightarrow{x}$), whereas matrices are denoted with uppercase letters with bold font (**X**). When it is only said to be a vector, its default is a column vector. Sometimes we use transpose sign T with a row vector to express a column vector. The main reason of doing this is simply to save the paper space.

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = (x_1\ x_2\ x_3)^T$$

Subscripts and "[]"

There are two types of subscripts. One is to represent the contents of the variable like σ_y (which represents σ, the standard deviation, of y). The other type is to represent the elements in a vector or a matrix. For instance, when we want to specify the k-th element in a vector $\overrightarrow{\theta}$, we write it as θ_k. In the latter case, sometimes we use "[]" to represent θ_k as of $\theta[k]$. Similarly, we write $\theta[n, k]$ to represent $\theta_{n,k}$ as well. Additinoally, $\overrightarrow{Y[3]}$ represents that the third element of Y is a vector.

Others

In the field of statistical modeling, terms such as mixed models, mixed effect models, fixed effects, and random effects are frequently used. This book also deals with the models that are related to these terms but does not use these. This is because we consider it important to understand model formulas rather than the terms themselves. In addition, since multilevel models and hierarchical models have the same mathematical formula, they are both called hierarchical models in this book.

The Source Code Used in the Book

All the source code and data in each chapter, including the visualization part, are freely available on this book's GitHub repository: https://github.com/MatsuuraKentaro/Bayesian_Statistical_Modeling_with_Stan_R_and_Python.

Also, in some of the chapters, we added some practice examples where the readers can code by themselves and deepen the understanding. The answers to those practical examples are also released on the same GitHub repository. Please use it as a reference. Lastly, in this book, we kept the following formats consistent:

- The set of equations to represent a model (Model Formula X.Y)
- Stan code (`modelX-Y.stan`)
- R and Python code (`run-modelX-Y.R, run-modelX-Y.py`)

The computing environment under which this book was written was Windows 11 (64bit), R 4.1.3, Stan 2.29.2, CmdStanR 0.5.0, CmdStanPy 1.0.1.

Tokyo, Japan Kentaro Matsuura

Acknowledgements First of all, I am particularly grateful to Shuonan Chen for her great contribution on writing this book in English. I would also like to thank Masaki Tsuda for offering to review this book. Next, I wish to express my gratitude to Prof. Motohiro Ishida and Yutaka Hirachi in Springer for giving me the opportunity to write this book. Finally, I would like to thank my family, including my cats.

Contents

Part IV Advanced Topics for Real-World Data Analysis