Bayesian Data Analysis Workbook

Alex Liebscher

2020-07-17

Contents

Chapter 1

Introduction

Contained within this notebook is an approximate following of Bayesian Data Analysis (3rd edition) by Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin (2020).

My goal is to follow the book, with as few jumps as necessary, and copy their work, their writing, and their code. Through this, I hope to learn how to apply Bayesian principles to the everyday data analysis problem.

Most of the text used in this workbook is paraphrased directly from the original without shame. Their examples are sufficient and in many cases I would do a disservice altering their language. This workbook is simply an exercise to practice active reading, note-taking, and implementation. It is an opportunity to learn something about Bayesian inference from a collection of bright scholars to whom I trust to present a thorough, captivating read of the subject.

I anticipate this to be highly interactive, which is why I've opted to work directly in a `bookdown` format. Hopefully, some nice plots will be created along the way.

Some code might be compared to Aki Vehtari's. Vehtari also has published their own lecture notes, which could also be a helpful resource. Luckily too, it seems like most of the data is hosted on Gelman's site - this should come in handy.

1.1 Collaboration

I am looking for a reading group to incrementally work through the textbook and have discussions with. If you are looking to learn about Bayesian inference, I hope you won't be shy and we can struggle together. Email me at `alexliebscher0@gmail.com` and just say you're interested, I'll figure out the rest.

Moreover, feel free to clone this repo for a quick start.

Chapter 2

Probability and Inference

Bayesian data analysis may be broken down into three distinct steps:

1. Designing the full probability model
2. Conditioning on observed data
3. Evaluating the fit of the model

A primary motivation for adopting a Bayesian framework is that it allows us to interpret the statistical conclusions closer to our common-sense human intuition. The central feature to Bayesian analysis is the direct quantification of uncertainty.

In terms of notation, we define:

1. θ as the unobservable vector quantities or population parameters of interest.
    a. Example: the probabilities of survival under a control and a treatment for randomly chosen members of the population.
2. y as the observed data.
    b. Example: the known number of survivors and deaths in both the control and treatment groups.
3. $\tilde{y}$ as unknown, but potentially observable, quanitites.
    c. Example: the outcome (survival or death) of an unseen patient similar to those already in the experiment.

The y variables are called the "outcomes" and may be represented as, e.g. 1 if patient i survives and 0 is patient i dies, so that y takes the form of a vector. These values are considered *random* when making inferences because there is the possibility they could have been the opposite outcome due to the sampling process or the natural variation in the population.

For the time being, we consider the values of y to be independent and identically distributed (iid).

It also common to have explanatory variables or covariates. This may include age, previous health status, etc. X denotes this set of k variables across all n observations.

2.1 Bayesian inference

Conclusions in a Bayesian framework about our θ (remember, the unobservable quantities or population parameters), or our $\tilde{y}$ (our unknown, but possible, outcomes), stem from either $p(\theta|y)$ or $p(\tilde{y}|y)$. This is to say that our parameters, or our unknown outcomes, are identified through a probability statement, conditional on our observed data.

Knowing that this is what we're chasing, we can introduce Bayes' rule:

$$p(\theta, y) = p(\theta)p(y|\theta)$$

In this equation, $p(\theta)$ is called the *prior distribution*, and $p(y|\theta)$ is called the *sampling distribution.*

Note that we can also write this as $p(\theta, y) = p(y)p(\theta|y)$. Therefore, together with the last result,

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}$$

where the *prior predictive distribution* is $p(y) = \sum_\theta p(\theta)p(y|\theta)$ (or $p(y) = \int p(\theta)p(y|\theta)d\theta$ if θ is continuous).

This is called the *posterier density.* These formulas represent the core of Bayesian statistics.

The distribuion of $\tilde{y}$ is called the *posterior predictive distribution* and takes the form:

$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y)d\theta = \int p(\tilde{y}|\theta, y)p(\theta|y)d\theta$$

2.1.1 Inference about a genetic status

A first example, with setup found under Section 1.4 pg. 8.

First, we set up the prior distribution:

> Consider a woman who has an affected brother, which implies that her mother must be a carrier of the hemophilia gene with one 'good' and one 'bad' hemophilia gene. We are also told that her father is not affected; thus the woman herself has a fifty-fifty chance of having the gene. The unknown quantity of interest, the state of the woman, has just two values: the woman is either a carrier of the gene (θ $= 1$) or not ($\theta = 0$). Based on the information provided thus far, the prior distribution for the unknown θ can be expressed simply as $\mathrm{P}(\theta = 1) = \mathrm{P}(\theta = 0) = \frac{1}{2}$.

Second, we establish our data model and the likelihood formula:

There are two possible worlds here: first, the woman in question is affected; second, the women in question is *not* affected. Suppose she has two sons, neither of whom are affected. The status of the two sons is independent: one son's status does not affect the other's. They both, however, rely on the mother's (unknown) status; they are conditional upon her status. Thus, the two items of independent data generate the following likelihood functions:

$\mathrm{P}(\mathrm{son}_1 \text{ unaffected}, \mathrm{son}_2 \text{ unaffected } |\theta = 1) = (0.5)(0.5) = 0.25$

$\mathrm{P}(\mathrm{son}_1 \text{ unaffected}, \mathrm{son}_2 \text{ unaffected } |\theta = 0) = (1)(1) = 1$

Third, we establish the posterior distribution:

Using Bayes' rule, we can now combine the information we know from the data with our prior knowledge. If we let $y = (\mathrm{son}_1 \text{ status}, \mathrm{son}_2 \text{ status})$, then:

$$P(\theta = 1|y) = \frac{p(y|\theta = 1)P(\theta = 1)}{p(y|\theta = 1)P(\theta = 1) + p(y|\theta = 0)P(\theta = 0)} = \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = \frac{0.125}{0.625} = 0.2$$

A key aspect of Bayesian analysis is the ease at which we may add additional data to the mix. For example, suppose the mother has a third son. We don't need to recalculate the entire formula from where we started, instead we can substitute the newly found posterior distribution as the new data model ($y =$ $(\mathrm{son}_1 \text{ status}, \mathrm{son}_2 \text{ status}, \mathrm{son}_3 \text{ status})$):

$$P(\theta = 1|y) = \frac{p(y|\theta = 1)P(\theta = 1)}{p(y|\theta = 1)P(\theta = 1) + p(y|\theta = 0)P(\theta = 0)} = \frac{(0.2)(0.5)}{(0.2)(0.5) + (1.0)(0.8)} = \frac{0.1}{0.9} = 0.111$$

Point of confusion: On pg. 9 the authors say: "we use the previous posterior distribution as the new prior distribution". However, according to their definition of the prior distribution $p(\theta)$, it would seem as though they're actually replacing the *likelihood* function $p(y|\theta)$.

In any case, the new probability of the mother being a carrier is 11.1% given the status of her three sons.