Censored Distributions in Greta

Mlen-Too Wesley

ISYE 6420 - Fall 2024 - Final Project (EXTRA)

December 1, 2024

Introduction

Over the course, I have been using R as a programming language. It has been both challenging and rewarding. With that, I have been using a package called Greta. It is not as easy as PyMC in Python, but has helped me better learn the concepts of the course.

Two weeks ago, while submitting my final homework assignment, I extended Greta to support right censoring. This took a considerable amount of time. So, of course, this week, I decided to extend it to support more censoring distributions and publish on my GitHub.

And then, I decided to see if I could get it published to CRAN.

Unfortunately, this took up a lot of the time I should have been working on my actual project – but fortunately, I did manage to get it through all the CRAN checks. It was a great experience so I am submitting a short article I wrote on it with my final project.

Making Sense of Incomplete Data

Over the past few months, I've dived deeper into the world of math and stats behind data science and machine learning. Along the way, I discovered Bayesian statistics, and fell in love with its approach to understanding uncertainty.

What is Bayesian statistics?

At its core, Bayesian statistics is about using probabilities to make sense of uncertainty. It's a bit like how we naturally learn and update our beliefs in real life.

> "That which is probable depends upon the evidence at hand" — Thomas Bayes

For example, if you believe it's going to rain this time of year because it always rains, but see the sun shining today, you'd probably update that belief and go outside. Bayesian methods help us make sense of this process mathematically. Your initial assumptions (or "priors") and what you observe (the "data") are used to update what you know about the world (the "likelihood") – and this leads to a new belief (the "posterior").

Unlike the more common traditional or "frequentist" form of statistics, which focuses on exact conclusions drawn from measuring how frequent an event is likely to occur, Bayesian methods let you work more flexibly, incorporating uncertainty every step of the way.

Why is this important?

Bayesian statistics also lets us study data that's layered or grouped. For example, imagine you're studying the health of children in rural communities. Their health isn't just determined by factors such as access to medical care, but part of a larger web of influences.

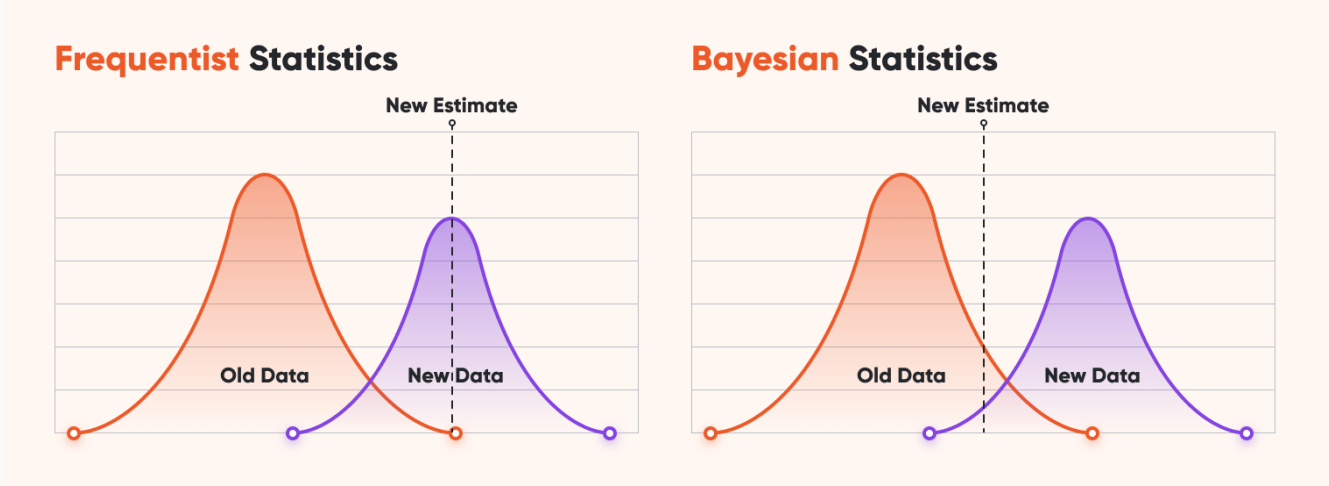
It might be tied to local agricultural production, which determines the availability of nutritious food. Agricultural production itself might also depends on broader economic policies, such as subsidies, trade regulations, or access to markets, which influence farming practices and food distribution. **Bayesian hierarchical models** help build models that reflect this layers of effects.

> "Everything is connected... no one thing can change by itself" – Paul Hawken

Now, let's talk about **censored data** – a challenge that often comes up in real-world studies. Sometimes, when working with data, you don't get the full picture:

- Someone might leave the study early or start late (right and left censoring)
- Measuring equipment may suffer a temporary malfunction (interval-censored)

Accounting for this type of uncertainty is critical for drawing meaningful conclusions. That's where **censored distributions** shine – they allow us to model incomplete data while still capturing its underlying structure.



R and Greta

The **R programming language** is loved by statisticians and data scientists for its flexibility, ease of use, and open source approach. Greta, also written for R, is a probabilistic programming language that makes Bayesian modeling more accessible, automating much of the complex math while giving you the freedom to design custom models.

> Simple and scalable statistical modelling in R

If you're familiar with Python, you might have heard of **PyMC**, a similar tool used for Bayesian modeling. **Greta** is like it's R-rated sister from another mister, built on **TensorFlow**, a very popular library from **Google** that is often used in machine learning workflows for its scalable and efficient computation.

Censored distributions for Greta

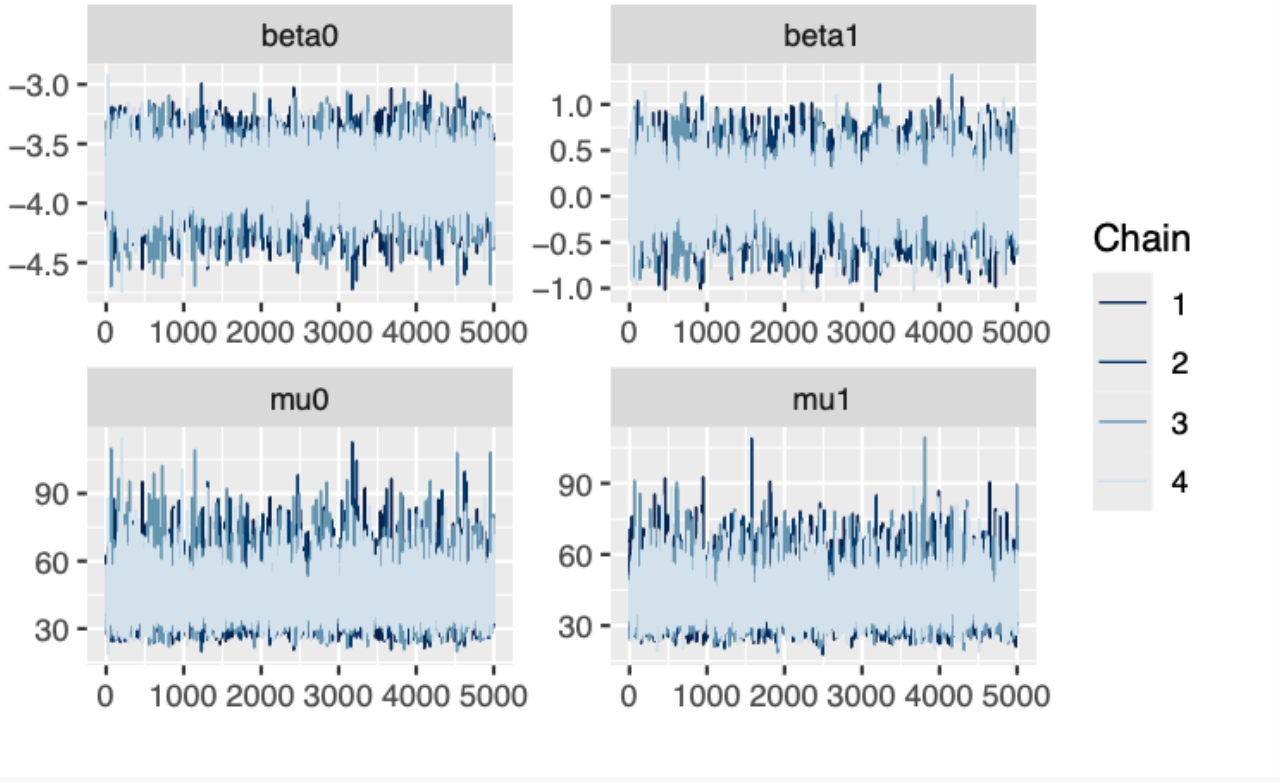
Last month, I noticed that Greta lacked direct support for censored distributions and decided to build a solution. This was both challenging and rewarding:

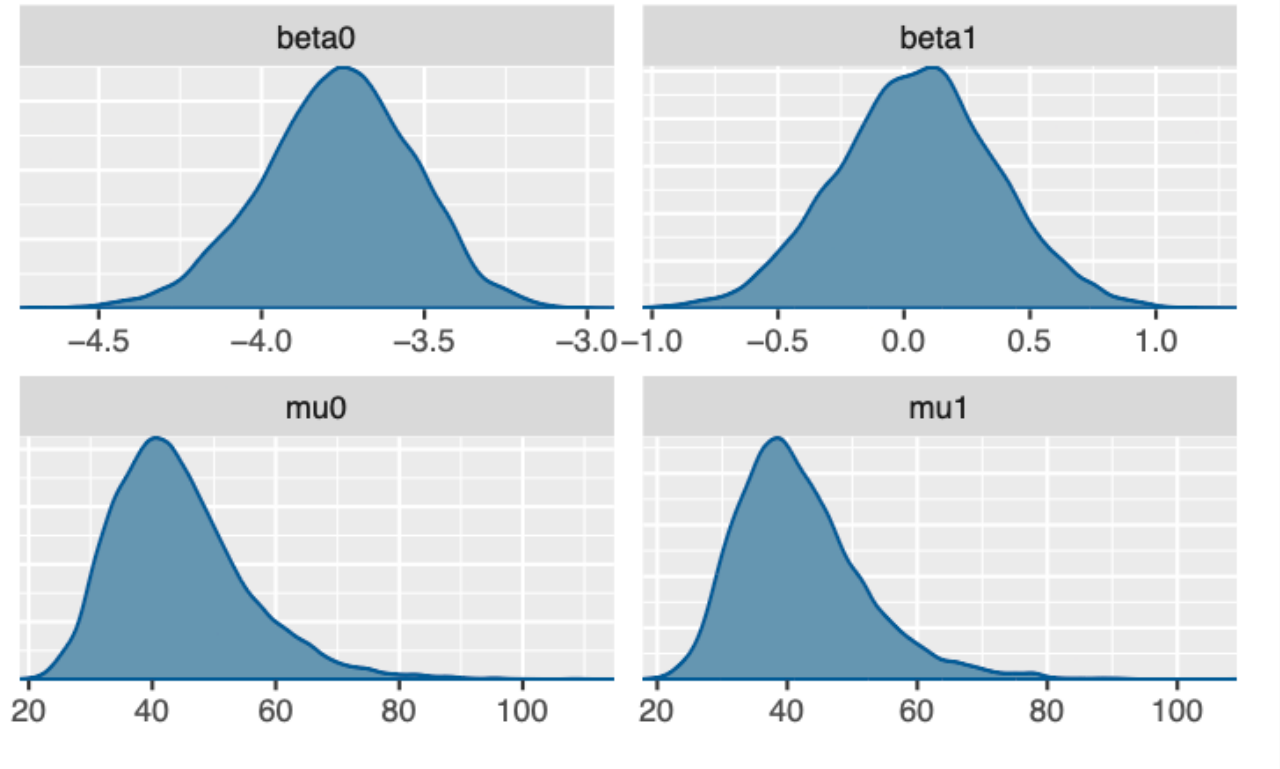
1. Extending Greta's capabilities to support censored versions of distributions like Normal, Exponential, Beta, Weibull, and more
2. Wrangling with unit tests to ensure every edge case was handled
3. Configuring GitHub workflows and actions to automate builds, tests, and checks
4. Navigating the rigorous CRAN submission process to meet its high standards for quality and reproducibility

Today, I'm thrilled to announce that my project **is now live on CRAN**! 🎉

Here's some examples of its use:

```
library(readr)
library(greta)
library(greta.censored)
library(bayesplot)

data <- read.csv("data.csv")

time <- as_data(data$time)
observed <- as_data(data$observed)
group <- as_data(data$group)

beta0  <- normal(0, 100)
beta1  <- normal(0, 100)
lambda <- exp(beta0 + beta1 * group)

mu0 <- exp(-beta0)
mu1 <- exp(-beta0 - beta1)

distribution(time) <- exponential_censored(lambda, observed, dim = length(time))
cancer_model <- model(beta0, beta1, mu0, mu1)

draws <- mcmc(cancer_model, warmup = 1000, n_samples = 5000, chains = 4)

mcmc_trace(draws)
mcmc_dens(draws)
```


beta0
beta1
mu0
mu1
−3.0
−3.5
−4.0
−4.5
1.0
0.5
0.0
−0.5
−1.0
90
60
30
0
1000
2000
3000
4000
5000
Chain
1
2
3
4


beta0
beta1
mu0
mu1
−4.5
−4.0
−3.5
−3.0
−1.0
−0.5
0.0
0.5
1.0
20
40
60
80
100

It's my first open-source package and I couldn't be more excited to share it with the world. Whether you're a researcher, data scientist, or student, I hope this package proves useful for your work.

- CRAN: https://cran.r-project.org/web/packages/greta.censored
- GitHub: https://github.com/mtwesley/greta.censored
- Documentation: https://mtwesley.github.io/greta.censored