# Bayesian Analysis Project:
# **Stroke Data**

Mahnaz Taheri & Doina Vasilev

May 2023

## 1 Introduction

The aim of our project is to showcase the main useful steps to carry out an analysis of the *Heart Stroke Dataset* following the Bayesian approach, in order to make inference on the parameters, and derive their posteriors.
In Section 2, a brief presentation of the dataset is given, and in the 3rd Section the main steps of the data cleaning process are shown. Section 4 contains a brief introduction on the main theoretical concepts included into the project, i.e., posterior distribution, logistic regression, and Metropolis- Hastings algorithm; and finally the resulting posteriors. In the end, we present an analysis on the relationship between the variables in the model, i.e., the mediation analysis.

## 2 The dataset

The dataset we're analyzing is the *Heart Stroke Dataset* we retrieved from Kaggle. The original dataset contains 5110 observations and 12 features, 6 of which are categorical, 5 numerical and 1 string variable (*index*).
We decided to drop some variables to reduce the parameter space, and simplify the computations of the posterior distributions, thus we obtained 5 independent variables and 1 outcome variable. The remaining features are the following:
**gender**: (*binary*) feature describing the sex of the patient. The two possible values are: **M** (male), and **F** (female).
**hypertension**: (*binary*) feature describing whether the patient suffers from hypertension (**1**), or not (**0**).
**heart_disease**: *binary* feature describing whether the patient suffers from a heart disease (**1**) or not (**0**).
**avg_glucose_level**: (*numeric*) feature the average glucose levels of each patient. The values fit within the interval [55.12, 271.74].
**bmi**: (*numeric*) feature describing the bmi index of each patient. The values fit within the interval [16.10, 44.90].
**stroke**: (*binary*) feature representing the output variable, i.e., what I wish to predict. It can take two possible values: **1**, if the patient has had (or is predicted to have) a stroke, or **0**, if patient has not had (or is not predicted to have) a stroke.

# 3 Data Cleaning

We've taken several steps in order to clean the data before making inference.

1. We replaced the **null values**: in the original dataset, the variable **bmi** contains several *NA* values.
   To replace them we used the **Random Forests** algorithm.
   The algorithm was first proposed by **L.Breiman (2001)**, and it is now implemented by the R package *randomForest*, as it was originally defined.
   The desirable characteristics of RF allow them to: [1] handle mixed types of missing data; [2] address interactions and nonlinearity; [3] scale to high-dimensions while avoiding overfitting; and [4] yield measures of variable importance useful for variable selection (**Tang. F & Ishwaran H.**; **2017**)
   The algorithm's approach for imputing missing values is the following: (1) pre-impute the data, (2) grow the forest, (3) update the original missing values using proximity of the data. (4) Iterate for improved results. (**Tang. F & Ishwaran H.**; **2017**). Thus, it's a reiterative process, where the data is repetitively imputed until it reaches best results.

2. Once we replaced the missing values, we decided to remove the outliers related to the **avg_glucose_level** and **bmi** variables.
   The resulting dataset contains 4276 observations.
   The boxplots representing the observations from the numerical variables, before and after the removal of the outliers are showcased in Figure 2.

3. Finally, we standardized the numerical variables with the aim of reducing the autocorrelations in the chains and thus improve the convergence of the Markov chain Monte Carlo algorithms.
   The distribution of the data, across the 5 variables is represented in Figure 1.

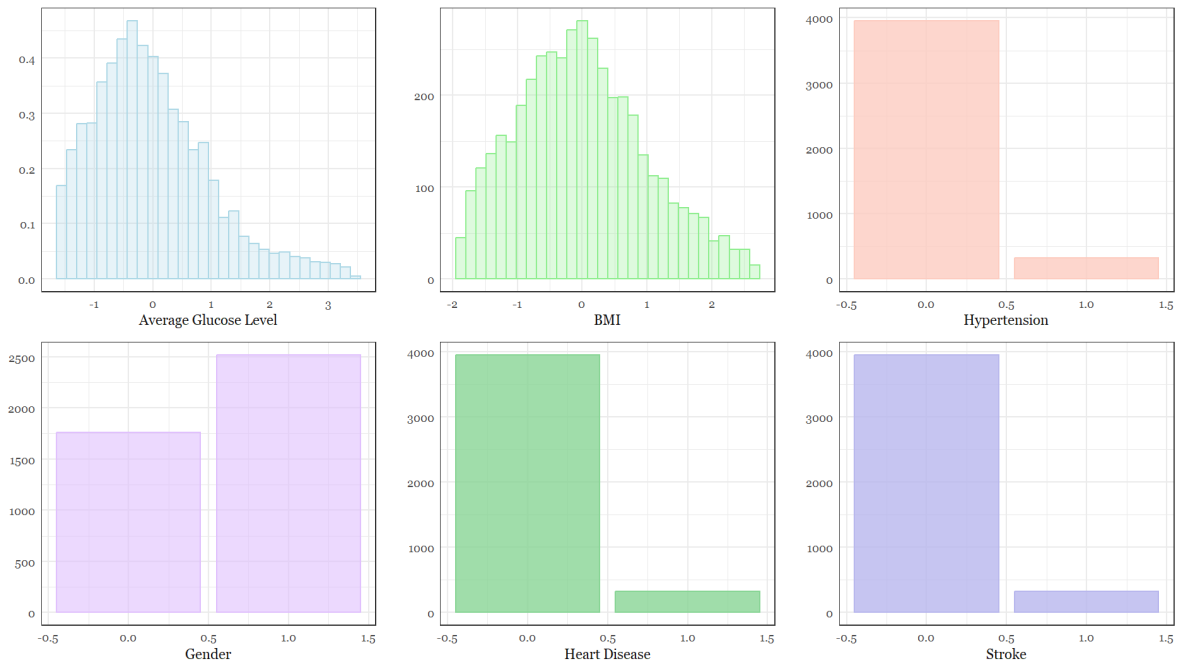


Figure 1: Distributions of the final dataset.

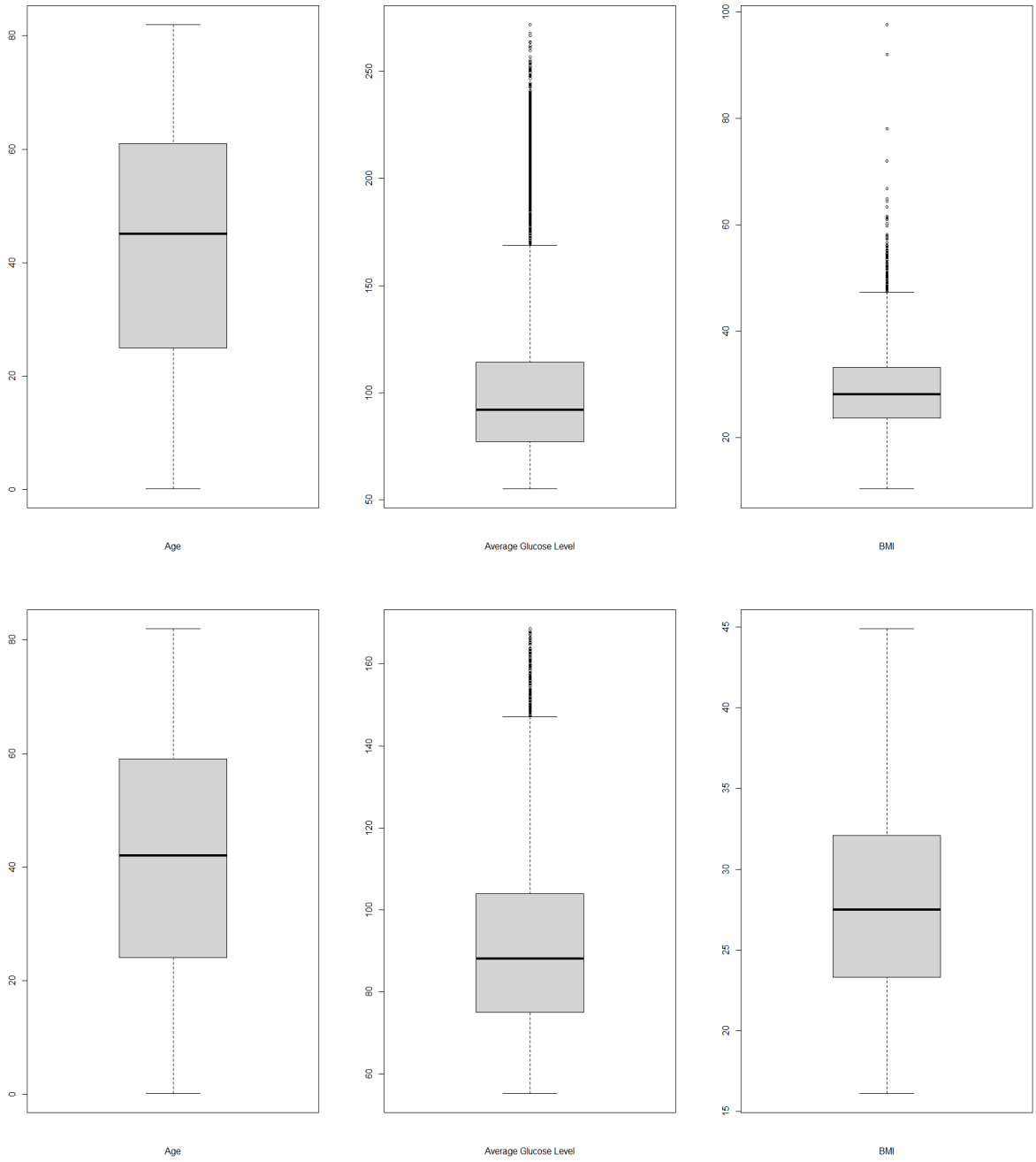


Figure 2: Boxplots representing the distribution of the data before and after the removal of the outliers.

# 4 Posterior Distribution

As opposed to the frequentist approach, which bases the inference of parameters on the sampling distribution of the estimators, the Bayesian approach makes inference on the parameters integrating expert information (*prior distribution*) and experimental data.
The formal representation of Bayesian inference is condensed in the following formulas:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)} \tag{1}$$

In the multi-parameter space, Bayes' rule can be approximated to:

$$p(\beta, \sigma^2|y, X) \propto p(y|X, \beta, \sigma^2) \times p(\beta|\sigma^2) \times p(\sigma^2) \tag{2}$$

Where: $p(\beta, \sigma^2|y, X)$ is the posterior distribution, $p(y|X, \beta, \sigma^2)$ the likelihood function and $p(\beta|\sigma^2)p(\sigma^2)$ the priors.

The aim of our project is thus to compute the posterior distribution of the variable *stroke*, starting from defining the likelihoods and the priors, based on our knowledge of the problem.

## 4.1 Logistic Regression

To derive the posterior distribution of parameters inside *Logistic Regression* model, we first defined the prior distributions of the intercept, the coefficients, and the variance of the error term.

Based on the parameters inside the *Logistic Regression* model which is:

$$log\left(\frac{p(y=1|x)}{1-p(y=1|x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \tag{3}$$

Since we have *limited knowledge* of the priors, we assume that the intercept and the coefficients can take any value in real life. Thus, we set the priors on these parameters as normal distributions:

$$\begin{gathered}\beta_0 \sim \mathcal{N}(0, 10) \\ \beta_1 \sim \mathcal{N}(0, 10) \\ \beta_2 \sim \mathcal{N}(0, 10) \\ \beta_3 \sim \mathcal{N}(0, 10) \\ \beta_4 \sim \mathcal{N}(0, 10) \\ \beta_5 \sim \mathcal{N}(0, 10)\end{gathered}$$

These priors express our belief that the coefficients are likely to be close to zero, but can take on values within a range that we consider plausible. The choice of prior distribution and its parameters can have a strong influence on the posterior distribution and the results of the analysis, so it is important to choose priors that are appropriate for the data.
The error term is also distributed normally with $\mu = 0$ and variance unknown. So we define a prior distribution also for the variance which is the Gamma distribution. Since the variance can only take positive values, the appropriate distribution would be Gamma:

$$\sigma^2 \sim \mathcal{G}(0.6, 1) \tag{4}$$

Since $Y$ (*stroke*) is binary, the proper distribution in this case would be Bernoulli distribution.
Let us say $\pi$ is the probability of having a stroke for an individual $i$,
$Y_i|\pi_i = Bern \sim (\pi_i)$ .

Before building the Bayes rule, we should also define our likelihood function. We assume that the response variable $y$ follows a Bernoulli distribution with a probability of success given by the logistic function of the predictor variables:

$$p(y|x_1, x_2, x_3, x_4, x_5, \beta) = Bernoulli(y|p(x_1, x_2, x_3, x_4, x_5, \beta)) \tag{5}$$

Where:

1. $p(y|x_1, x_2, x_3, x_4, x_5, \beta)$ is the probability of the binary response variable $y$ taking the value 1 given the predictor variables $x_1, x_2, x_3, x_4, x_5$ and the model parameters $\beta$.

2. $Bernoulli(y|p(x_1, x_2, x_3, x_4, x_5, \beta))$ is the Bernoulli likelihood function, which gives the probability of observing the binary outcome $y$ given the success probability $p(x_1, x_2, x_3, x_4, x_5, \beta)$.

The success probability is given by the logistic function:

$$p(x_1, x_2, x_3, x_4, x_5, \beta) = \frac{1}{1+exp(-z)}$$

Where:

1. $z$ is the linear combination of the predictor variables and their corresponding coefficients, i.e., $z = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5$.

## 4.2 Metropolis-Hastings algorithm

Since we could not calculate the distribution of our parameters in a closed form, hence to obtain the posterior distribution of *stroke*, we exploited the **Metropolis-Hastings** algorithm. This is a Markov chain Monte Carlo algorithm for producing samples from distributions that are difficult to sample directly.
The algorithm includes the following steps: suppose we want to sample from a target distribution $\pi(\theta|x)$, and are provided with proposal distribution $q(\theta^*|\theta)$.

1. Initialize the iteration $j = 1$, and initialize the chain to $\theta^0$.

2. Generate a proposed value $\theta^*$, using the proposal distribution $q(\theta^*|\theta^{j-1})$.

3. Evaluate the acceptance probability $\alpha(\theta^{j-1}, \theta^*)$ of the proposed move, where

$$\alpha(\theta, \theta^*) = min\{1, \frac{\pi(\theta^*|x)q(\theta|\theta^*)}{\pi(\theta|x)q(\theta^*|\theta)}\} \tag{6}$$

4. Set $\theta^j = \theta^*$ with probability $\alpha(\theta^{j-1}, \theta^*)$, and set $\theta^j = \theta^{j-1}$ otherwise.

5. Change the counter from $j$ to $j + 1$ and return to Step 2. [1]

The strategy we're going to use is *Normal random walk*.

[1] Notes by Rossini L.

## 4.3 The model

The priors are defined as follows:

| **Parameter** | **Distribution** |
|---|---|
| $\beta_0$ | $\mathcal{N}(0, 10)$ |
| $\beta_1$ | $\mathcal{N}(0, 10)$ |
| $\beta_2$ | $\mathcal{N}(0, 10)$ |
| $\beta_3$ | $\mathcal{N}(0, 10)$ |
| $\beta_4$ | $\mathcal{N}(0, 10)$ |
| $\sigma^2$ | $\mathcal{G}(0.6, 1)$ |

Table 1: Prior distributions of the parameters in the dataset.

We chose the likelihood function based on the logistic regression, i.e., the *Bernoulli* distribution.
The model is defined as:

$$log\left(\frac{p(y=1|x)}{1-p(y=1|x)}\right) = \beta_0 + \beta_1\textbf{bmi} + \beta_2\textbf{gender} + \beta_3\textbf{hypertension} + \beta_4\textbf{avg\_glucose\_level} + \beta_5\textbf{heart\_disease}(7)$$

Once we have found the priors and the posteriors, we applied Metropolis-Hastings to obtain the posterior distributions.
The resulting posteriors are summarized in the following table:

| | mean | sd | hdi_3% | hdi_97% | mcse_mean | mcse_sd | ess_bulk | ess_tail | r_hat |
|---|---|---|---|---|---|---|---|---|---|
| **beta0** | -3.487 | 0.711 | -4.782 | -1.955 | 0.087 | 0.062 | 97.0 | 56.0 | 1.01 |
| **beta[0]** | -0.021 | 0.101 | -0.212 | 0.164 | 0.001 | 0.001 | 10793.0 | 11365.0 | 1.00 |
| **beta[1]** | 0.084 | 0.197 | -0.291 | 0.451 | 0.003 | 0.002 | 3741.0 | 6699.0 | 1.00 |
| **beta[2]** | 1.175 | 0.255 | 0.699 | 1.658 | 0.002 | 0.002 | 11469.0 | 15875.0 | 1.00 |
| **beta[3]** | -0.004 | 0.098 | -0.190 | 0.173 | 0.001 | 0.001 | 11619.0 | 12381.0 | 1.00 |
| **beta[4]** | 0.935 | 0.337 | 0.303 | 1.570 | 0.003 | 0.002 | 14726.0 | 20820.0 | 1.00 |
| **epsilon** | -0.081 | 0.697 | -1.541 | 1.226 | 0.085 | 0.061 | 98.0 | 57.0 | 1.01 |
| **sigma** | 0.535 | 0.678 | 0.000 | 1.771 | 0.049 | 0.034 | 214.0 | 379.0 | 1.00 |

Figure 3: Resulting posteriors.

Figure 4 showcases the path of the autocorrelations for each parameter. For most of the coefficients (from $\beta_1$ to $\beta_5$) the autocorrelations tend to fade as the number of iterations increases. The number of iterations we set is 50 000.
Increasing the number of iterations and *thinning* allows us to reduce the autocorrelations obtained through the Markov chain.
Unfortunately, even when exploiting such strategies the autocorrelations of the intercept and the sigma remain particularly high.

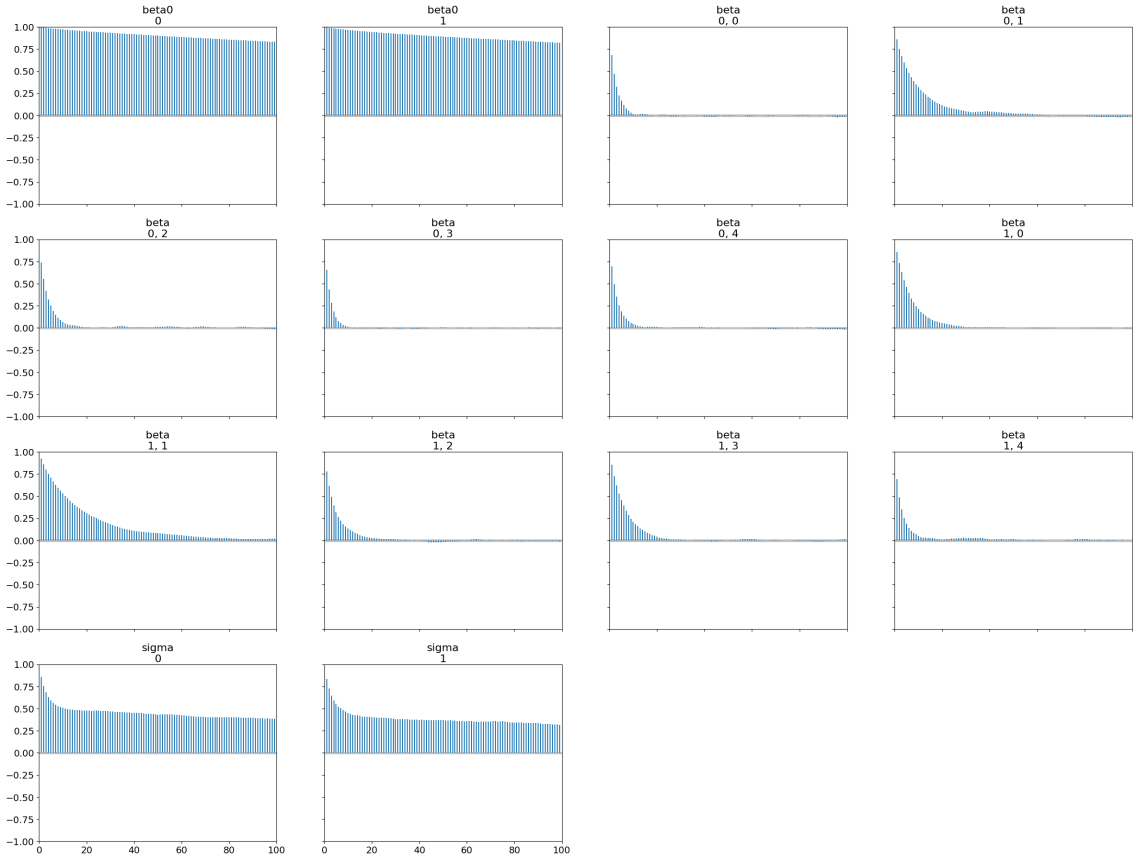


Figure 4: Autocorrelation plot of parameters

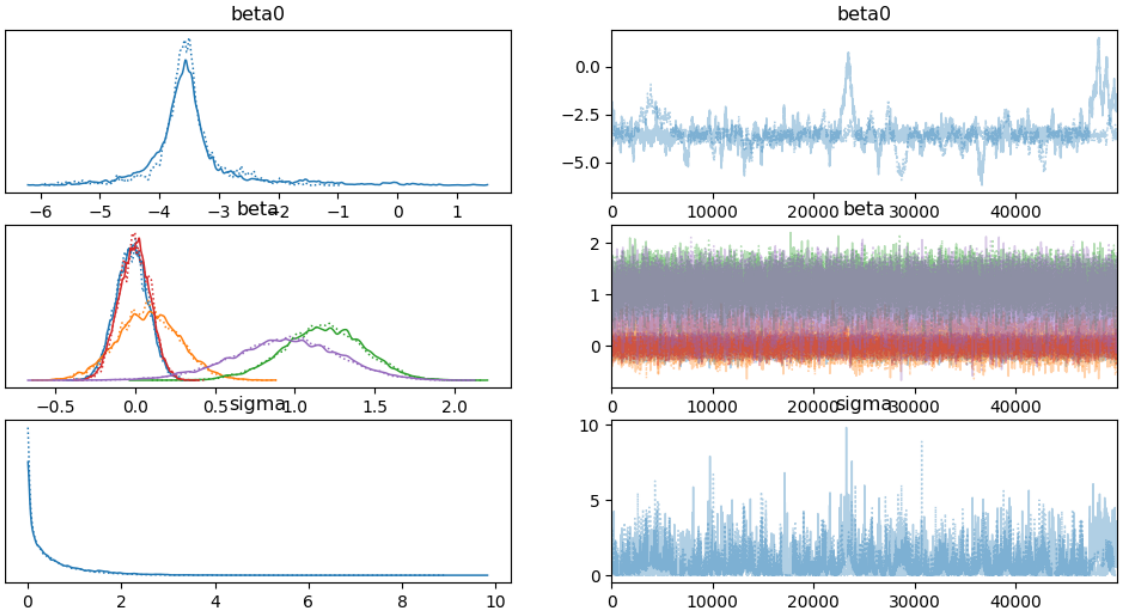


Figure 5: Trace plot of parameters

Moreover, in figure 5 one can observe:

1. On the left, one can observe the density plots for each posterior distribution.
   Density plots are one of the available diagnostic tools to assess the quality of the MCMC algorithm.
   They portray the posterior distributions after incorporating the priors, and showcase how the first adapts to the latter.
   In this case, the first plot showcases the intercept's posterior, with mean equal to -3.487; the second showcases the various distributions of the coefficients; and the last one portrays the distribution of the sigma.
   In the 2nd plot we can identify the posteriors of the coefficients and state the following:

   (a) The distributions of $\beta_1$ (bmi) and $\beta_4$ (avg_glucose_level) are centered around 0. This suggests that effects of such variables on the outcome are very small. Their distributions, however, are very concentrated, thus the estimates on the parameters do not show uncertainty.
   The coefficient $\beta_2$ (gender) is close to 0 as well, but its distribution is sparser.

   (b) On the other hand, the coefficients $\beta_3$ (hypertension) and $\beta_5$ (heart_disease) are well-distanced from the 0; however, their distributions showcase greater uncertainty with respect to the first 2.

2. On the right, the trace plots showcase the sequence of parameter values sampled by the algorithm as it explores the posterior distribution.
   Trace plots are diagnostic means, as well.
   We aim to a well-converged trace plots, because such plots will show the parameter values fluctuating around the posterior mean and not exhibiting any clear trend or pattern.
   In our case, the plots reflect once again what was observed using the autocorrelation plots: the coefficients from $\beta_1$ to $\beta_5$ are converging, as they are homogeneous and show no specific pattern.
   On the other hand, the plots on the intercept ($\beta_0$) and $\sigma^2$ show an upper trend around 25 000 iterations and toward the end as well.

Finally, figures 6 and 7 portray the 94% HDIs (High Density Intervals) for each parameter for the specified level of posterior probability, and assigns to each parameter a range of possible values.
The larger the intervals, the more uncertainty surrounds the parameters estimates.

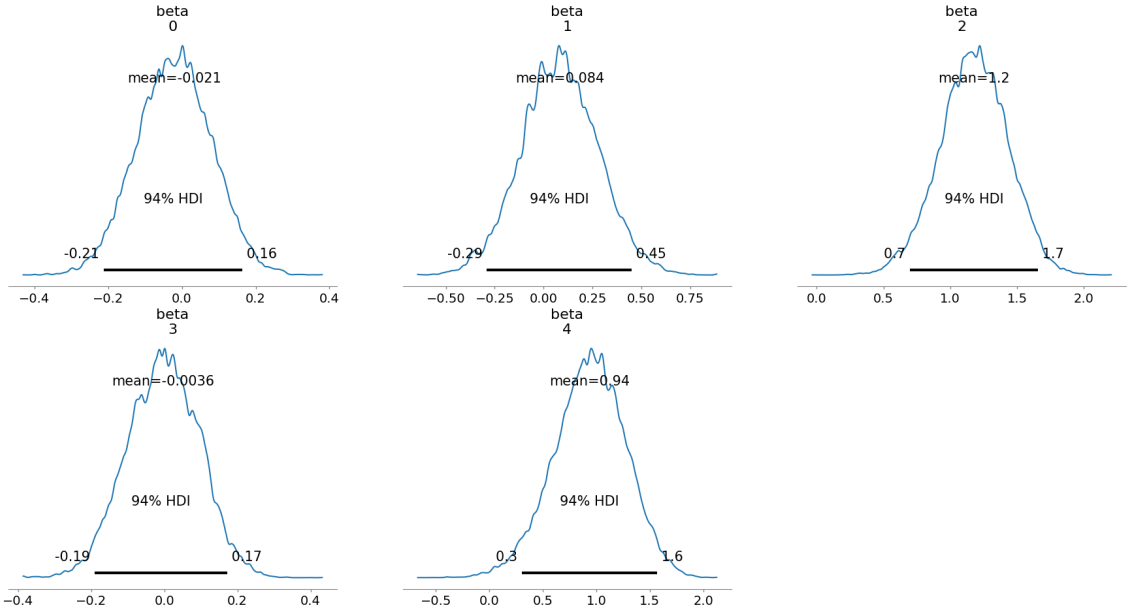


Figure 6: Posterior distributions of parameters

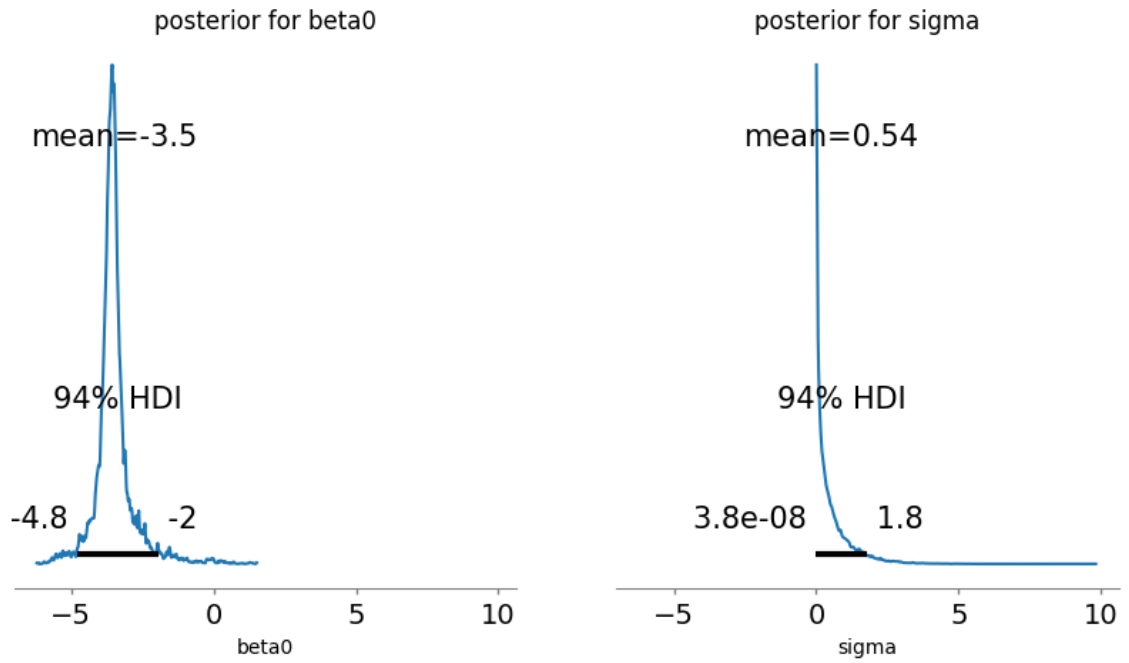


Figure 7: Posterior distributions of parameters

# 5 Mediation Analysis

Bayesian logistic regression mediation analysis is a statistical technique that can be used to explore the relationships between variables in a model, where one variable (the mediator) acts as an intermediary between two other variables (the independent and dependent variables).
This approach allows us to estimate the direct and indirect effects of the independent variable on the dependent variable, through the mediator. Overall, Bayesian logistic regression mediation analysis provides a powerful tool for investigating the mechanisms by which variables are related to each other.
In statistics, a mediator is a variable that explains the relationship between two other variables.
When a mediator exists, it means that the effect of one variable (the independent variable) on another variable (the dependent variable) is partially or fully explained by the mediator variable.

In the context of Bayesian logistic regression, the mediation test gives us estimates of the **direct**, **indirect**, and **total** effects of the independent variable on the dependent variable:

1. The direct effect is the effect of the independent variable on the dependent variable that is not explained by the mediator.

2. The indirect effect is the effect of the independent variable on the dependent variable that is explained by the mediator.

3. The total effect is the sum of the direct and indirect effects.

First we have to define our causal model.
In our case the causal model will have two paths: one from the independent variable (*smoking status*) to the outcome (*stroke*) and another path through mediator variable (*heart_disease*) to the outcome.
We then specified the distributions for the model parameters which are Normally distributed based on our assumptions.
The logistic regression is once built with smoking status as an independent variable and the heart disease as the response. Another regression model is built this time with the heart disease and smoking status as predictors and the stroke as dependent variable.
In the next step we want to estimate the indirect and direct effect of the independent variables on the outcome through the mediator, using this formula:
indirect effect = mediator effect * coefficient of the mediator in the logistic regression model.
The estimation of the total effect is done by adding the direct effect to the indirect effect. The direct effect in this case is the coefficient of smoking status.
By calculating the posterior distributions, we obtain these plots (Figure 9) in which can be seen they are all converging.
Finally, in Figure 10 we can see the posterior distributions of the direct, indirect, and total effect. As it is shown, the direct and indirect effect have different means.

The posterior mean of the direct effect (x to y) is 0.36 which means if the probability of the smoking status being one increases, the probability of the person having a stroke increases by 43%.

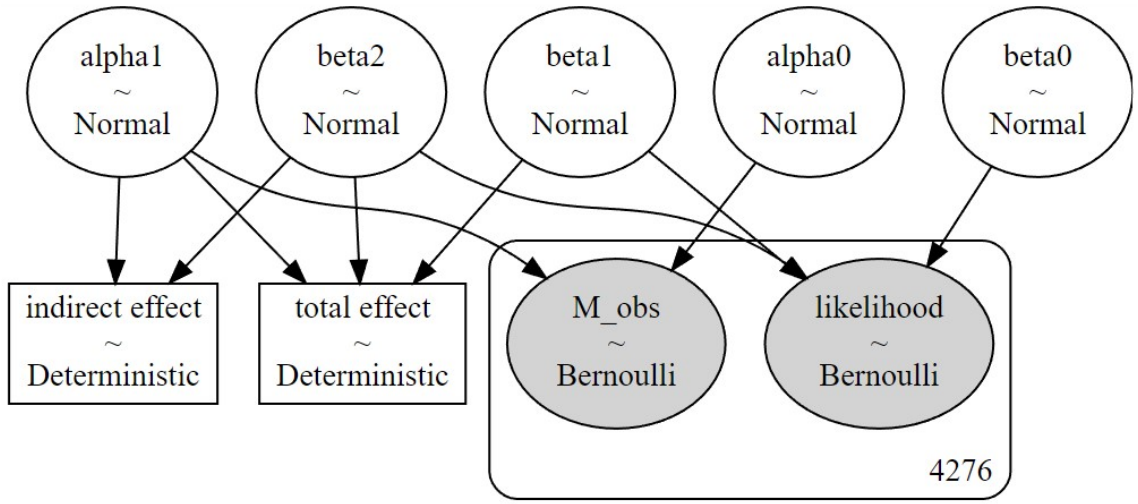


Figure 8: This graph depicts the graph of the model.

The posterior mean of the indirect effect (x to m to y) is 0.43, meaning if the probability of the smoking_status=one increases, the probability of the person having a stroke increases by 53%. The probability that the indirect effect is zero is very small.

The posterior mean of the total effect is 0.79. This also means that if the probability of the smoking status being one increases, the probability of the person having a stroke increases by 120%!.

If the indirect effect is more than the direct effect, it means that the mediator variable is playing a more significant role in explaining the relationship between the independent and dependent variables. In other words, this could indicate that the mediator variable is an important factor to consider when trying to understand the relationship between the independent and dependent variables. It could also suggest that the direct effect may be partially mediated by the mediator variable.

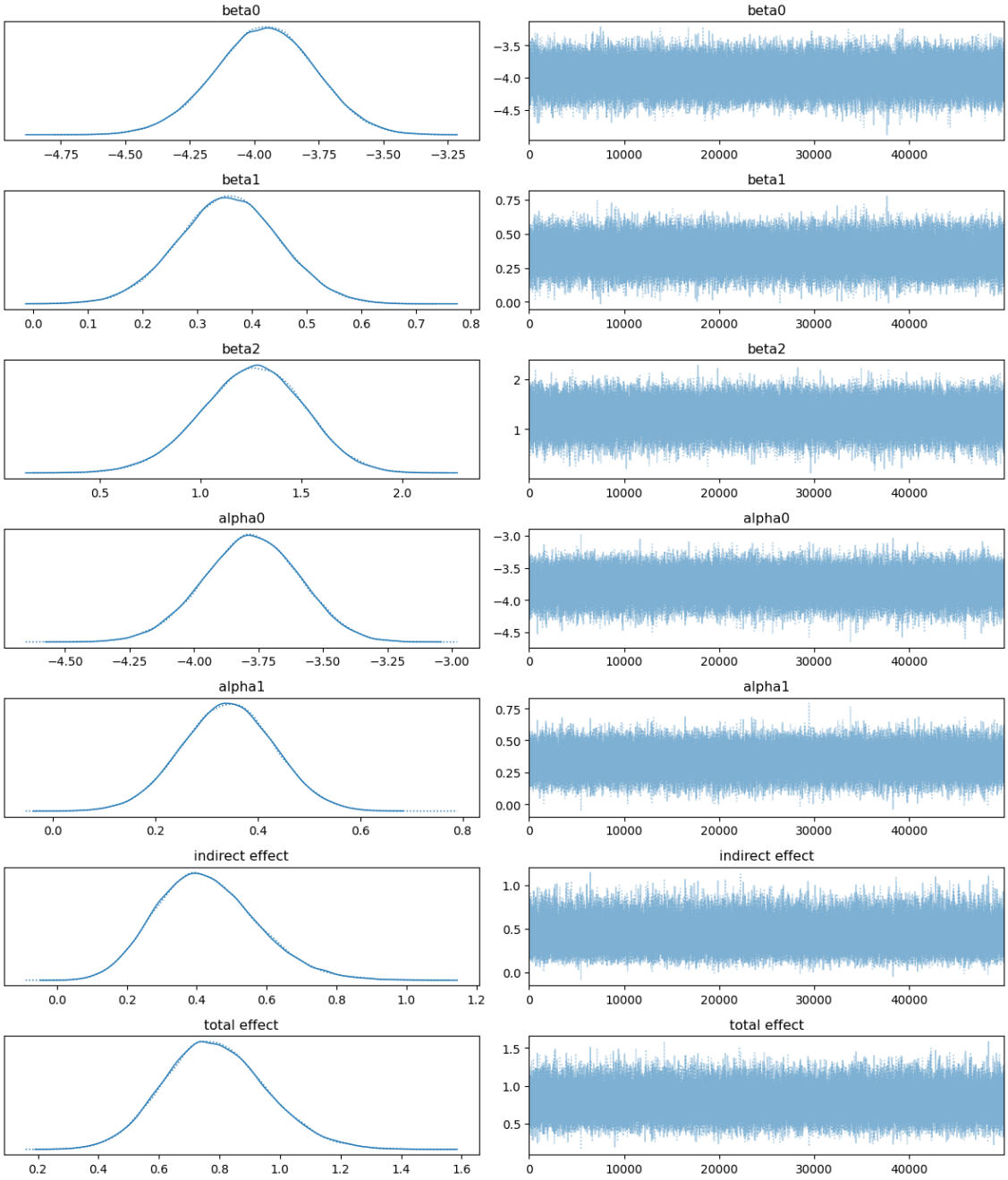


Figure 9: Trace plot of parameters

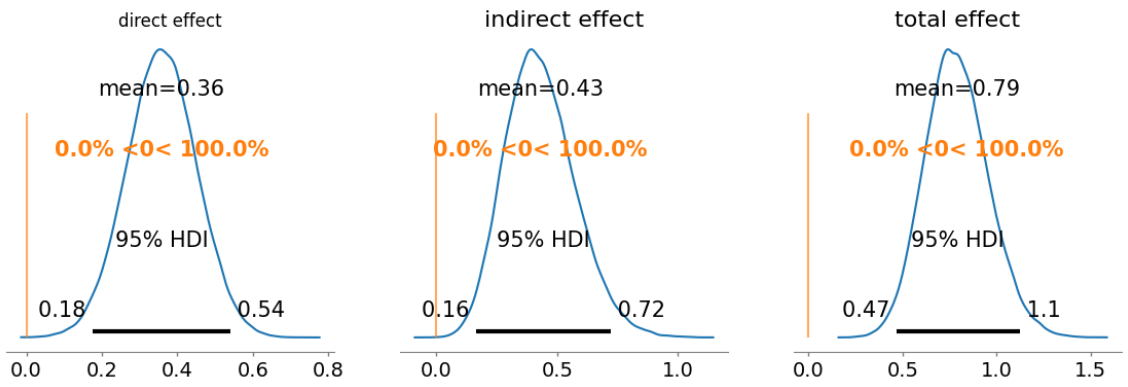


Figure 10: The figure illustrates the posterior distributions of direct, indirect and total effect