Introduction to Software in Econometrics

Breast Cancer Bayesian Inference

Phung Dang Bao Ngoc i6290653
Mila Dimkovska i6286475

Group 6 Tutorial 1
EBS2072

School of Business and Economics
Maastricht University
January 2024

1 Introduction

Breast cancer is a significant global health concern, with researchers constantly seeking innovative approaches to understand the intricate interplay of risk factors associated with this disease. In this study, we employ Bayesian inference methods, specifically leveraging the power of the RStan package, to unravel the complexities of breast cancer. We start by giving an overview of the dataset we work with, and then we discuss our approach as well as our findings when implementing Bayesian Analysis.

In this paper, we use a statistical approach known as Bayesian inference to understand how different factors relate to the classification of breast cancer. Utilizing Bayesian methods provides a robust framework for integrating prior knowledge, handling uncertainties, and deriving posterior distributions that more accurately capture underlying statistical patterns in the data. We use visual tools such as trace plots and pair plots to show how our model behaves over time. Trace plots show us how our analysis explores different possibilities while pair plots focus on the relationships between different parameters, giving us insights into how breast cancer is categorized. This research helps us better grasp the complexities of breast cancer modeling and emphasizes the effectiveness of Bayesian methods in this field.

2 Data

We obtained our Comimbra Breast Cancer dataset via **Kaggle - Comimbra Breast Cancer Data**. The dataset under investigation features a binary dependent variable representing the presence or absence of breast cancer. The covariates include demographic and physiological factors such as Age, BMI (Body Mass Index), Glucose, Insulin, HOMA (Homeostatic Model Assessment), Leptin, Adiponectin, Resistin, and MCP.1 (Monocyte Chemoattractant Protein-1).

Firstly, we start with a summary table that captures key features of the data as well as we obtain some plots that can be useful for visually investigating some interesting properties of the sample, such as the correlation between the variables and possible extreme points(outliers). We decided to scale the data since it helps avoid numerical instabilities such as having very big and very small values simultaneously.

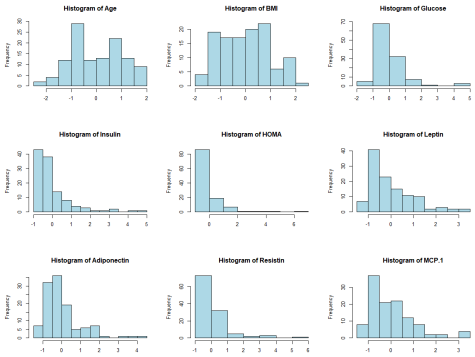
In the studied sample of 116 people, more than half were diagnosed with cancer (55%). The age of the sample ranges from 24 to 89, with a mean of 57. People's body weights exhibit a degree of variation, with an average BMI of 28. Some values related to metabolism, such as glucose and insulin, show diverse patterns. Moreover, fat-related indicators (Leptin and Adiponectin) and an inflammation marker (MCP.1) also differ among individuals. Overall, the summary statistics show that people with and without cancer in this sample have various characteristics, providing useful information for further study.

Summary Statistics

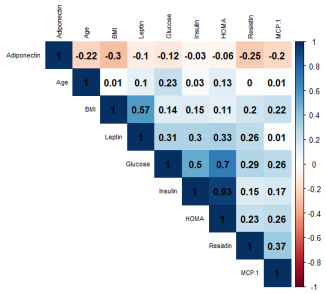
Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Classification	116	0.55	0.5	0	0	1	1
Age	116	57	16	24	45	71	89
BMI	116	28	5	18	23	31	39
Glucose	116	98	23	60	86	102	201
Insulin	116	10	10	2.4	4.4	11	58
HOMA	116	2.7	3.6	0.47	0.92	2.9	25
Leptin	116	27	19	4.3	12	37	90
Adiponectin	116	10	6.8	1.7	5.5	12	38
Resistin	116	15	12	3.2	6.9	18	82
MCP.1	116	535	346	46	270	700	1698

Furthermore, we visualized the distribution for each variable to identify any possible patterns and potential outliers in the data. From the histograms in Figure 1, we noticed that most of the parameters follow a right-skewed distribution such that the majority of the data points cluster to the left. Additionally, few parameters such as BMI seem to follow a normal distribution.

Investigating the correlation between parameters in Figure 2, we observed that both Insulin and Glucose levels are highly positively correlated with HOMA. If not addressed, this high correlation may lead to multicollinearity issue and affect the precision of our model's parameter estimates.



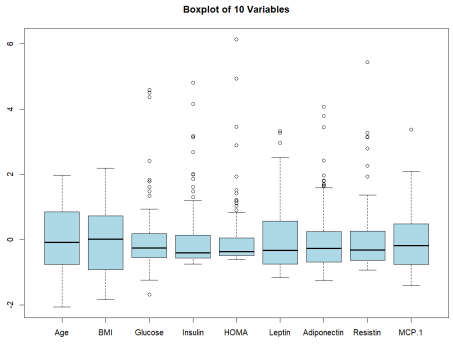
(a) Histograms for parameters (scaled)



(b) Correlation matrix of parameters

Figure 1: Histograms and correlation matrix

Finally, we constructed box plots to look for outliers that differ from the majority of the other data in a dataset. Noticeably, some variables such as HOMA, Resistin, etc. have extreme values that can be illustrated by points outside their respective boxes. Ignoring outliers or treating them inappropriately can lead to inaccurate conclusions. If the errors can be attributed to mistakes in data collection, it might be reasonable to delete them or decrease their magnitude, but if they are indeed some genuine extreme observations in the underlying sample we should not change them. In the corresponding dataset, we have decided not to change them and keep the initial form of the dataset.



3 Bayesian Model

In our research, we applied the Bayesian inference method, which is based on updating initial beliefs when given new information. We will investigate priors, likelihood, and posterior to do Bayesian analysis on the data introduced in Section 2.

3.1 Priors

The very first step was defining the priors that represent our initial beliefs before observing our data. We decided to conduct two different types of research, one built on uninformative and the other on informative prior.

Drawing on a decade of research exploring the relationship between age and breast cancer risk, it's observed that age plays a pivotal role in influencing a woman's susceptibility to breast cancer. As a woman's age increases, the risk of developing breast cancer rises significantly. We incorporate this understanding into our Bayesian model by assigning an informative prior for the coefficient associated with age. Since we scaled our data, the coefficients represent the change in standard deviations of the response variable for a one-standard-deviation change in the predictor.

Our prior belief is that the coefficient for age (θ_{Age}) follows a normal distribution with a mean of 2 and a standard deviation of 0.5. Similarly, some previous studies show a positive relation between body mass index and breast cancer risk. Higher BMI levels are often linked to increased Estrogen production in Adipose tissue, influencing the possibility of breast cancer. We express our belief that

the coefficient for BMI (θ_{BMI}) follows a normal distribution with a mean of 3 and a standard deviation of 0.5. Finally, our last informative prior regarding Glucose level is explained with a normal distribution with a mean of 2 and a standard deviation of 0.5. This reflects our expectation that glucose levels have a moderate impact on the log-odds of breast cancer risk. Regarding the other variables, since we didn't have any subjective belief or any expert opinion available, we assigned them to a standard normal distribution ($N(0, 1)$).

For an uninformative prior, we assumed a flat or non-informative distribution. This means we are not incorporating any specific prior knowledge or beliefs about the parameters, hence they are represented by experiencing a standard normal distribution.
Intuitively, while the uninformative prior has density spread over $(-2, 2)$, the informative priors with smaller standard deviation should have values concentrated around one region rather than spread out.

$$\text{Uninformative prior: } \theta[i] \sim N(0, 1)$$

$$\begin{aligned} \text{Informative prior:} \\ \theta[1] &\sim N(1, 0.5) \\ \theta[2] &\sim N(2, 0.5) \\ \theta[3] &\sim N(3, 0.5) \\ \theta[i] &\sim N(0, 1) \ \forall \ \text{i} = \{4 : 9\} \end{aligned}$$

3.2 Likelihood

The likelihood captures the distribution density that reflects from the actual data. We then constructed the likelihood function from a logistic regression, representing the probability of observing the given outcomes based on the model parameters. Specifically, the Bernoulli distribution is employed to model the likelihood of observing an individual with a breast cancer diagnosis.

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\theta_1 + \theta_2 X_2 + \theta 3 X_3 + ... + \theta_{10} X_{10})}}$$

$$y \sim \text{Bernoulli}(P)$$

When try plotting the shape of the likelihood function with respect to θ_2, or the θ_{Age}, we obtained the following shape in Figure 2. It can be seen that the likelihood has more concentrated values than the uninformative prior, which has spread-out values and the function shape has a lower peak. From the shapes of our priors and likelihood, it seems that the likelihood should have more impact over the posterior than the priors, since it has a more concentrated region of values.

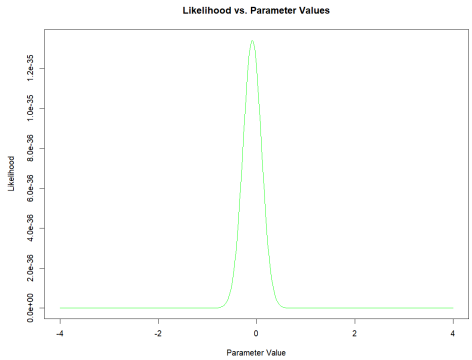


Figure 2: Likelihood function shape with respect to θ_{Age}

3.3 Simulation algorithm

With respect to our simulation, we implemented the NUTS (No-U-Turn Sampler) algorithm. NUTS is an extended version of Hamiltonian Monte Carlo (HMC). The reasons that drives our choice are the efficiency and adaptive tuning of NUTS. What this algorithm does is it starts at random values and takes steps in different directions to find the path that optimizes the most. To avoid going in circles (or in other words taking U-Turns), the algorithm stops exploring a path when it starts going back in the direction it came from.

Since the adaptive tuning of NUTS sampler takes place during the warm up process, we set our burn in draws to the first 500 iterations out of 1000 iterations we run. Although burn in draws discard a part of the sample and seem computationally effective, we maintain our burn in draws at 500 to help NUTS with adaptive tuning, so that NUTS can tune the size of steps it takes while exploring the paths, since too big or too small of the steps can deter the convergence of parameters. We also run 6 chains simultaneously so that it can help with mixing and convergent issues that the algorithm may encounter.

3.4 Posterior and STAN model results

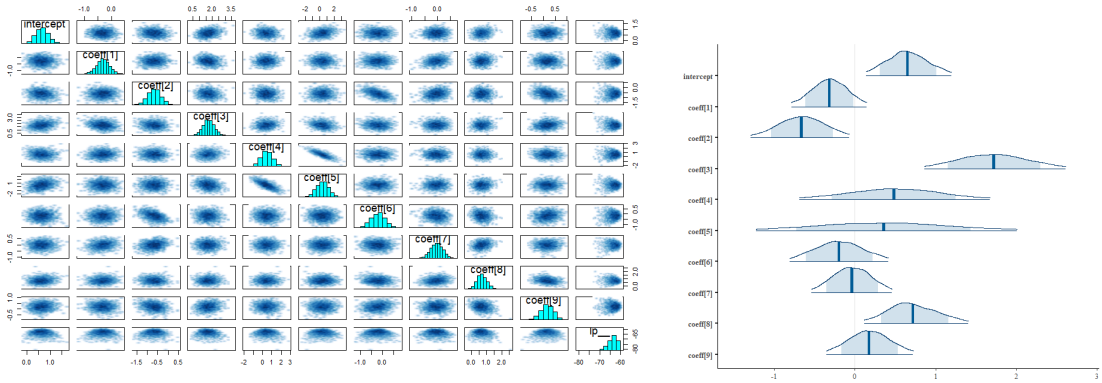
Running the corresponding model in RStan yields us certain results about the posterior distribution. Firstly, we examine the case where our prior is uninformative.

	Mean	SE Mean	SD	2.5%	25%	50%	75%	97.5%	n_{eff}	$\hat{R}$
Intercept	0.66	0.00	0.26	0.17	0.48	0.65	0.83	1.20	3453	1
coeff[1]	-0.32	0.00	0.24	-0.78	-0.48	-0.32	-0.16	0.16	3520	1
coeff[2]	-0.66	0.01	0.30	-1.27	-0.86	-0.66	-0.46	-0.09	3184	1
coeff[3]	1.71	0.01	0.44	0.90	1.42	1.70	1.99	2.63	3042	1
coeff[4]	0.45	0.01	0.59	-0.74	0.05	0.46	0.85	1.60	3007	1
coeff[5]	0.40	0.02	0.84	-1.21	-0.16	0.41	0.95	2.07	2760	1
coeff[6]	-0.20	0.01	0.30	-0.80	-0.41	-0.20	0.00	0.40	2833	1
coeff[7]	-0.03	0.00	0.25	-0.54	-0.19	-0.03	0.13	0.46	3487	1
coeff[8]	0.72	0.01	0.32	0.14	0.50	0.71	0.93	1.41	3158	1
coeff[9]	0.18	0.00	0.27	-0.35	0.00	0.17	0.35	0.71	3538	1
lp__	-63.73	0.07	2.36	-69.28	-65.04	-63.37	-61.98	-60.28	1222	1

Table 1: Summary RSTAN (iterations = 1000, chains = 6) (uninformative prior)

As reported in Table 1, sufficiently large n_{eff} and ($\hat{R} = 1$) indicate that the chains have no mixing issues and converge to the same area of values. The 95% intervals are regions with starting value reported as 2.5% quantile and ends at 97.5% in Table 1 and graphed in Figure 3b, which we will compare in later part of this report with the other model where an informative prior is used. The convergence of chains is also depicted in the trace plot in Figure 5.

The distribution of parameters illustrated in the pair plot (Figure 3.1) suggests symmetric distribution for all coefficients; therefore, it is intuitive to report the mean as our findings for the coefficient estimates. However, some observed estimated coefficients do not align with intuition. For example, the age and the dependent variable cancer risk are reported to share a negative relationship, which is in contrast to logical reasoning in real life. The log posterior density (reported as lp__ in Table 1) that has low n_{eff} also suggests that this model is not fitting the data very well. We run the model with informative prior to compare its performance with the results we obtained.



(a) Pair plot between parameters

(b) Posterior density plot of parameters

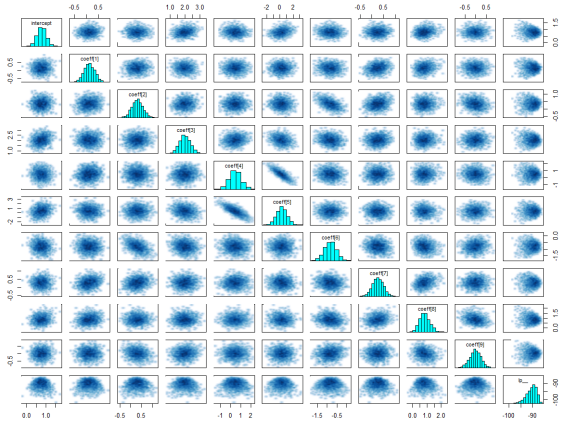
Figure 3: Statistics for posterior analysis with **uninformative prior**

In our analysis with informative prior, we had a different belief about our first four parameters including the intercept. We set up our priors for the first three coefficients and intercept to follow normal distribution with different means comparing to the first model and smaller standard deviation. In other words, we believe some of the parameter estimates are more concentrated in certain regions rather than spread out with $\sim N(0, 1)$. That is the reason why the 95% region for this model are narrower than the previous model as reported in Table 1 and Figure 4. The $\hat{R} = 1$ and n_{eff} indicate that there were no mixing and converging issues. We believe that for lp__, the increase in n_{eff} comparing to previous model implies that this model may have a better fit for the data.

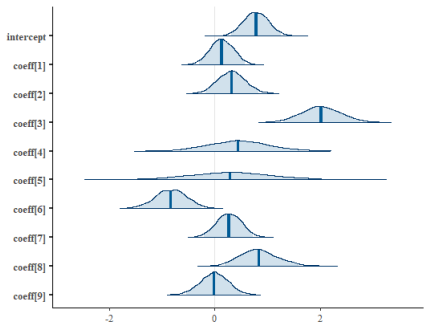
	Mean	SE Mean	SD	2.5%	25%	50%	75%	97.5%	n_{eff}	$\hat{R}$
Intercept	0.78	0.00	0.24	0.31	0.61	0.77	0.94	1.26	2870	1
Coeff[1]	0.12	0.00	0.22	-0.30	-0.03	0.12	0.27	0.55	3230	1
Coeff[2]	0.31	0.00	0.25	-0.17	0.14	0.31	0.49	0.83	3266	1
Coeff[3]	2.01	0.01	0.36	1.34	1.77	2.01	2.26	2.71	3505	1
Coeff[4]	0.43	0.01	0.59	-0.71	0.03	0.43	0.81	1.61	2537	1
Coeff[5]	0.31	0.02	0.83	-1.30	-0.26	0.33	0.88	1.94	2706	1
Coeff[6]	-0.84	0.01	0.30	-1.45	-1.03	-0.83	-0.63	-0.28	2648	1
Coeff[7]	0.27	0.00	0.24	-0.22	0.10	0.27	0.43	0.74	3002	1
Coeff[8]	0.85	0.01	0.35	0.23	0.59	0.82	1.07	1.60	2592	1
Coeff[9]	-0.02	0.00	0.26	-0.55	-0.20	-0.02	0.15	0.49	3724	1
lp__	-90.53	0.06	2.22	-95.86	-91.83	-90.23	-88.90	-87.14	1496	1

Table 2: Summary STAN (iteration = 1000, chains = 6) (informative prior)

According to the pair plot, the coefficients have symmetric distribution so we reported the mean of the coefficient estimates as our results. The coefficient estimates we obtained are different comparing the ones with an uninformative prior. For example, now we obtain positive means for coeff[1], coeff[2], and coeff[7], while the coeff[9] now has negative estimates. Since we scaled our data, this suggests that according to our model, one increase in standard deviation of the scaled coeff[1] leads to the predicted cancer risk increases by 0.12 standard deviation.

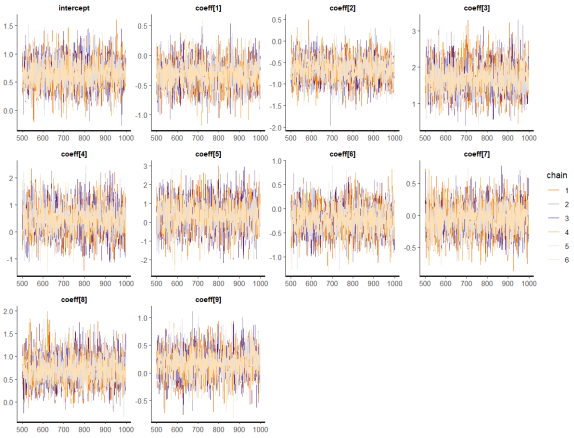


(a) Pair plot between parameters

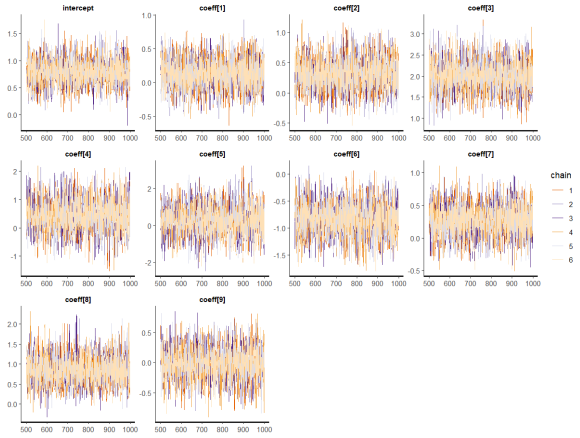


(b) Posterior density plot of parameters

Figure 4: Statistics for posterior analysis with **informative prior**



(a) Trace plot (uninformative prior)



(b) Trace plot (informative prior)

Figure 5: Trace plots of two priors

4 Conclusion

In summary, we used Bayesian inference method to explore our data of Breast Cancer Diagnosis, by implementing NUTS sampler as simulation algorithm to investigate the relationship between the priors, the likelihood, and the posteriors. We mainly focused on two cases of priors: one uninformative and one informative, and explored how given the same inputs, the different priors can yield different posteriors and interpret these results. The intuition behind our findings is that when we used a more constrained prior, the posterior will have a more concentrated density comparing to a flat one. This can be illustrated by posterior plots and ranges of intervals reported in our research.

Surprisingly, the results came from STAN simulation with an uninformative prior deviate from logical expectations, as for example the Age parameter was reported to have a negative correlation with the Breast Cancer risk, which was contradictory to the popular expectation that age should be one of the main factors driving this risk. However, when introducing an informative prior, the coefficient estimates change drastically and seemingly fits the data slightly better.

This report is still limited in certain aspects. When comparing priors, we have only addressed the difference between uninformative and weakly informative prior. Our prior was intentionally chosen to be less specific since we wanted to avoid running into huge amount of divergent transitions that arise from overly specific priors that were not reflected similarly in the data. We also have yet to tackle the correlation between parameters when running our estimation; therefore, our model's precision may have been reduced. Another extension that would be a meaningful addition to the report would be the implementation of different sampling algorithms that explore how different algorithms would affect the obtained results.

References

[1] Silverman, M. P., & Lipscombe, T. C. (2022). *Exact Statistical Distribution of the Body Mass Index (BMI): Analysis and Experimental Confirmation. Open Journal of Statistics, 12*(3), [Page Range if available]. `https://doi.org/10.4236/ojs.2022.123022`.