A Bayesian Logistic Regression Model to analyze the survival of breast cancer patients after the surgery

Hasini Gammune

University of Texas at Dallas

November 12, 2023

Section 1

Introduction

Introduction

- Logistic regression is a special case of regression analysis and is used when the dependent variable is nominally scaled or ordinally scaled.
- Logistic regression is the standard and the most reliable approach in the analysis of the binary and categorical outcome data.
- A Supervised machine learning method that is used to model the success probability of a certain class or event.
- A probabilistic model which automatically allows to compute the probability of success for a new data point.
- One of the most popular and powerfull ML models used in classifications.

Motivation

Example

Suppose a certain credit card company is using a logistic regression model to predict whether a credit card can be approved and suppose they train their developed model on very few negative data. Then under this circumstance, given a new data point, the developed model has a very low probability for the newly applied credit card being approved.

- Here the justifications relies totally on large sample arguments and training the model on fewer number of data gives unclear conclusions.
- A methodology is needed
  1. to capture the uncertainty about the model.
  2. in verifying whether the model parameters are meaningful.

A **Bayesian Logistic Regression Model** can be utilized to overcome this issue.

Research problem and Data

- Develop a Bayesian Logistic Regression model to classify the persons who are survived after the surgery and who are dead after the surgery from the given Haberman Cancer Survival data set.
- **Haberman Cancer Survival data set**: The data set contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.
- The response is the Survival status (class attribute)

```
1 = the patient survived 5 years or longer
0 = the patient died within 5 year
```

- The predictor variables are,
  1. Age of patient at time of operation (numerical)
  2. Number of positive axillary nodes detected (numerical)
  3. Tumour size (numerical)

Section 2

Model and Bayesian inference

Model

- Suppose Y_i, the survival status of the i^{th} individual follows a Bernoulli distribution with mean μ_i

$$Y_i|\mu_i \sim Bernoulli(\mu_i)$$

$$\text{Then} \qquad logit(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta^T \mathbf{x_i} = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

$$\text{and} \qquad \mu_i = \frac{exp(\beta^T \mathbf{x_i})}{1 + exp(\beta^T \mathbf{x_i})}$$

Bayesian inference

- Likelihood

$$
\begin{aligned}
f(\mathbf{y}|\beta) &= \prod_{i=1}^{n} \text{Bern}\left(y_i; \frac{\exp(\beta^T \mathbf{x_i})}{1+\exp(\beta^T \mathbf{x_i})}\right) \\
&= \prod_{i=1}^{n} \left(\frac{\exp(\beta^T \mathbf{x_i})}{1+\exp(\beta^T \mathbf{x_i})}\right)^{y_i} \left(\frac{1}{1+\exp(\beta^T \mathbf{x_i})}\right)^{1-y_i} \\
&= \exp\left(\sum_{i=1}^{n} (y_i(\beta^T \mathbf{x_i}) - \log(1+\exp(\beta^T \mathbf{x_i})))\right)
\end{aligned}
$$

- Prior

$$
\begin{aligned}
\beta &\sim \text{MN}(\mathbf{b}, \sigma_\beta^2 \mathbf{I}) \\
\pi(\beta) &= \frac{1}{\sqrt{(2\pi)^p |\sigma_\beta^2 \mathbf{I}|}} \exp\left(-\frac{1}{2\sigma_\beta^2}(\beta - \mathbf{b})^T(\beta - \mathbf{b})\right) \\
&\propto \exp\left(-\frac{1}{2\sigma_\beta^2}(\beta - \mathbf{b})^T(\beta - \mathbf{b})\right)
\end{aligned}
$$

Cont..

- Posterior

$$\pi(\beta|\mathbf{y}) \propto f(\mathbf{y}|\beta) \times \pi(\beta)$$
$$\propto \exp\left(\sum_{i=1}^{n}(y_i(\beta^T\mathbf{x_i}) - \log(1+\exp(\beta^T\mathbf{x_i})))\right)\exp\left(-\frac{1}{2\sigma_\beta^2}(\beta-\mathbf{b})^T(\beta-\mathbf{b})\right)$$

- The posterior distribution is **not in closed form**
- Gibbs Sampler method cannot be used.
- Here I used Metropolis-Hasting Algorithm to approximate the posterior distribution.

Model Fitting

- Random Walk Metropolis-Hastings

$$
\begin{aligned}
\mathbf{r}_{MH} &= \frac{\pi(\beta^*|\mathbf{y})}{\pi(\beta^{(t-1)}|\mathbf{y})} \frac{J(\beta^{(t-1)}|\beta^*)}{J(\beta^*|\beta^{(t-1)})} \\
&= \frac{\prod_{i=1}^{n} \text{Bern}\left(y_i; \frac{\exp(\beta^{*T}\mathbf{x_i})}{1+\exp(\beta^{*T}\mathbf{x_i})}\right)}{\prod_{i=1}^{n} \text{Bern}\left(y_i; \frac{\exp(\beta^{(t-1)T}\mathbf{x_i})}{1+\exp(\beta^{(t-1)T}\mathbf{x_i})}\right)} \frac{\text{MN}(\beta^*; \mathbf{b}, \sigma_\beta^2 \mathbf{I})}{\text{MN}(\beta^{(t-1)}; \mathbf{b}, \sigma_\beta^2 \mathbf{I})}
\end{aligned}
$$

- Proposal distribution

$$
J(\beta^*|\beta^{(t-1)}) \sim \text{MN}(\beta^{(t-1)}, k(\mathbf{X}^T\mathbf{X}))
$$

Section 3

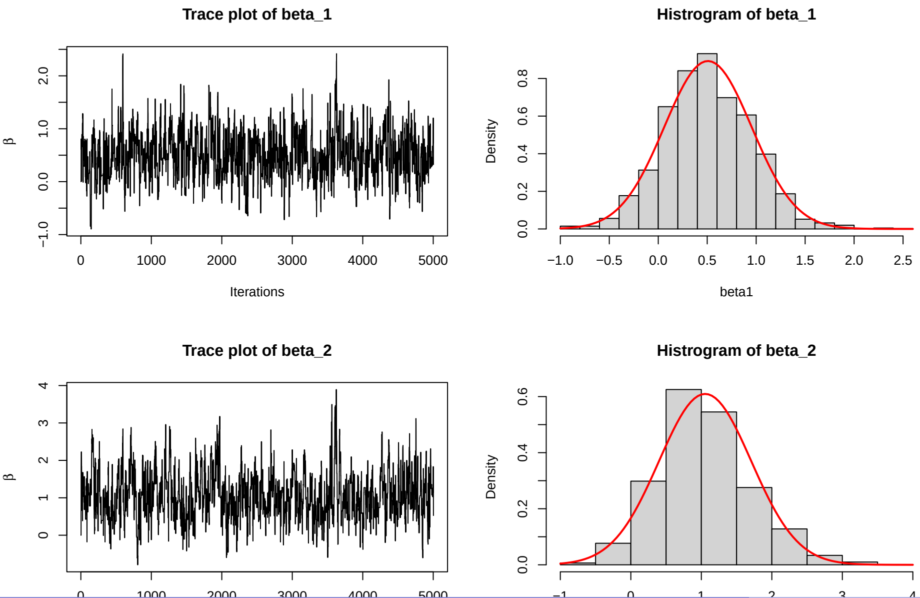
Simulated data analysis

Simulated data analysis

$$logit(\theta_i) = log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta^T \mathbf{x_i} + \epsilon = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i$$

```
##Simulating the data
set.seed(1111)
n<-50
X<-cbind(rep(1, n ),rnorm(n),rnorm(n))
beta<-c(5,0.5,1) #fix beta
epsilon<-rnorm(n,0,5) #generating error terms
theta<- exp((X %*% beta)+epsilon) /(1 + exp((X %*% beta) +epsilon))
y<-rbinom(n,1,theta)
```

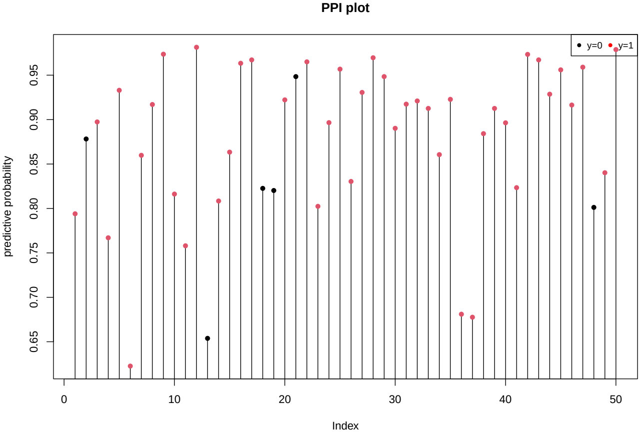
Validating the Algorithm



Posterior predictive checking

y=0	y=1
7	43

Table 1: Observed count of y_i



Section 4

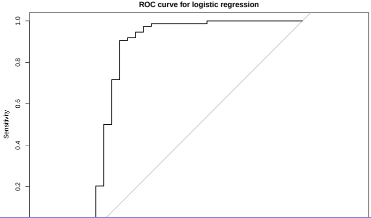
Real data analysis

Frequentist Approach

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	12.9329	4.3192	2.9943	0.0028
Age	-0.1906	0.0991	-1.9241	0.0543
Aux_nodes	-0.0424	0.0532	-0.7968	0.4256
tumour_size	-0.1354	0.0276	-4.9119	0.0000

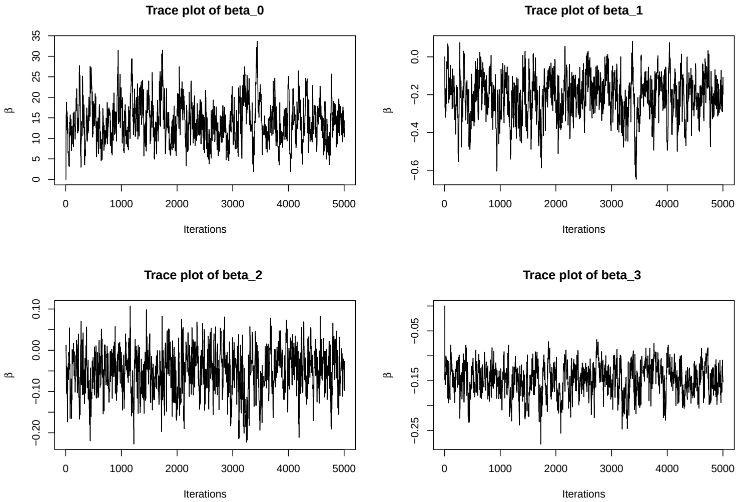
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	12.9329	4.3192	2.9943	0.0028
Age	-0.1906	0.0991	-1.9241	0.0543
Aux_nodes	-0.0424	0.0532	-0.7968	0.4256
tumour_size	-0.1354	0.0276	-4.9119	0.0000

Table 2: Beta coefficients from GLM output

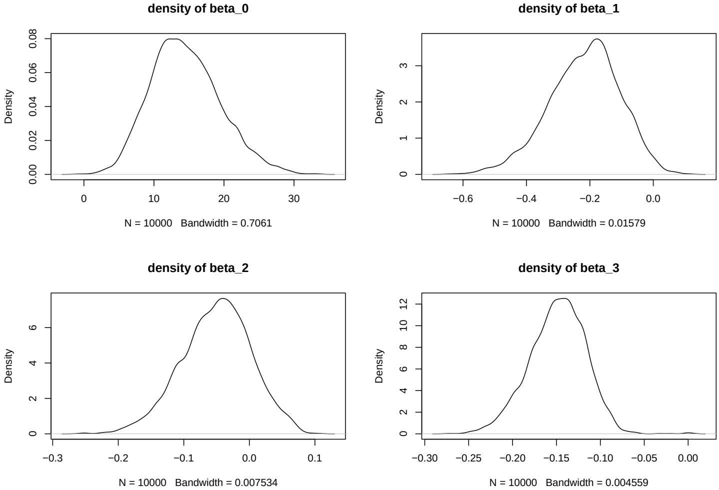


Bayesian Approach

- The proposal distribution is adjusted such that $k = 10$ and so the acceptance rate is 0.4612.



Posterior densities of βs

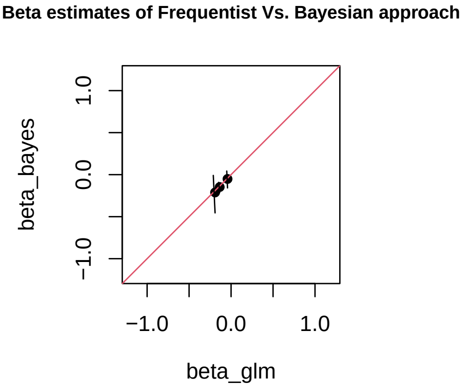


- Shape wise the β distributions look more like symmetric.

β estimates

	Frequentist approach	Bayesian approach
β_0	12.93288915	15.32778623
β_1	-0.19061434	-0.23602491
β_2	-0.04237426	-0.05466828
β_3	-0.13535657	-0.14782607

Table 3: Beta coefficients from both methods

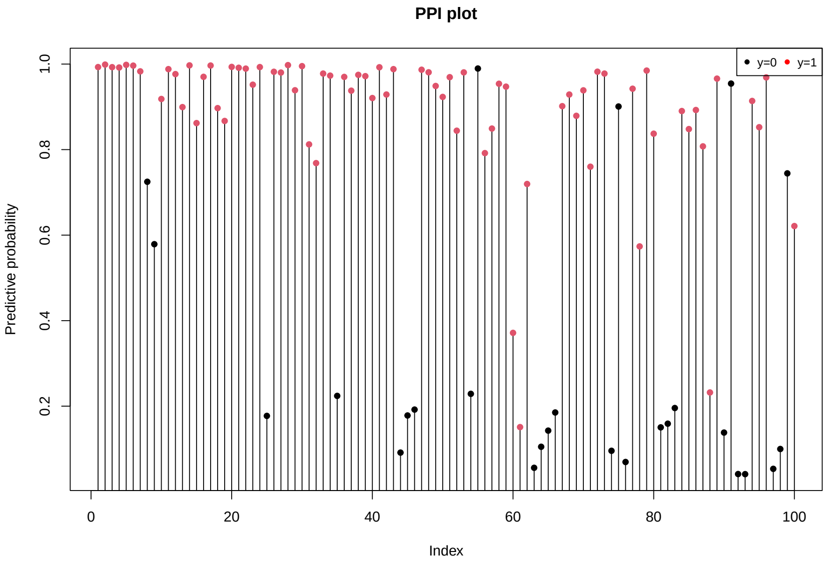


Credible Intervals for β

	Frequentist Approach		Bayesian Approach	
	2.5 %	97.5 %	2.5 %	97.5 %
β_0	4.4675	21.3983	6.4539	26.4662
β_1	-0.3848	0.0036	-0.4819	-0.0371
β_2	-0.1466	0.0619	-0.1572	0.0395
β_3	-0.1894	-0.0813	-0.2107	-0.0943

Table 4: 95% CI from Frequentist Approach and Bayesian Approach

Posterior predictive checking



Cont..

y=0	y=1
26	74

Table 5: Observed count of the survival status

- The observed data for $y = 1$ (the patient survived 5 years or longer) have a higher probability to be sampled in the predictive distribution.
- $y = 1$ **(the patient survived 5 years or longer) have a higher posterior predictive distribution of inclusion.**
- The model fits the data very well.

Regularization

- Spike and Slab Prior

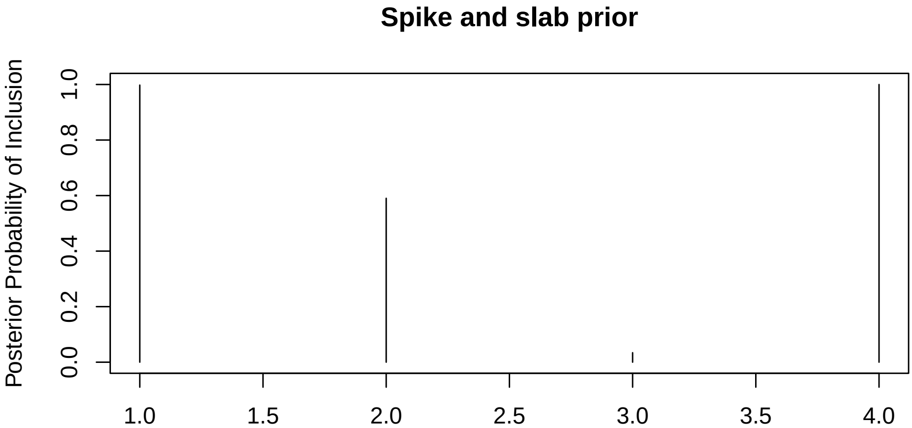
$$\beta_j|\sigma^2, \gamma_j \sim (1-\gamma_j)\mathbf{I}_0(\beta_j) + \gamma_j \mathrm{N}(0, h\sigma^2)$$
$$\sigma^2 \sim \mathrm{IG}(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$$
$$\gamma_j \sim \mathrm{Bern}(\omega)$$
$$\text{Hyperprior} \quad \omega \sim \mathrm{Beta}(a, b)$$

- Prior Setting : $\omega = 0.6$ h=1
- Add-Delete algorithm was used to update γ_j.

Prediction evaluation

	MVN Prior	Regularization
1	0.38	0.26

Table 6: MSE comparison



Section 5

Summary

Summary

- The β estimates from the two methods closely align.
- The Bayesian model captures the uncertainty which is not covered by the frequentist approach.
- A sensitivity analysis can be performed by varying the prior settings.
- The posterior distribution can be approximated using Grid approximation and Acceptance-rejection sampling also.
- Bayesian methodology can be used to overcome the small sample issue through a regularization methodology.

Section 6

References

References

- DuMouchel, W. (2012). Multivariate Bayesian logistic regression for analysis of clinical study safety issues. Statistical Science, 27(3), 319-339.
- Hoff, P. D. (2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer.
- O'brien, S. M., & Dunson, D. B. (2004). Bayesian multivariate logistic regression. Biometrics, 60(3), 739-746.
- O'Malley, A. J., & Zou, K. H. (2006). Bayesian multivariate hierarchical transformation models for ROC analysis. Statistics in Medicine, 25(3), 459-479.
- Chen, M. H., & Dey, D. K. (2003). Variable selection for multivariate logistic regression models. Journal of Statistical Planning and Inference, 111(1-2), 37-55.
- Cawley, G. C., & Talbot, N. L. (2006). Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. Bioinformatics, 22(19), 2348-2355.

Any Questions?

Thank You!