Diabetes Prediction Using Bayesian Logistic Regression and Conventional Logistic Regression

Final Project for Bayesian Method for Data Science (DATS 6450)

Project Group 11: Li Du, Xiaochi Li, Yupeng Yang

Data Science, George Washington University

Introduction

What’s Diabetes?

- A number of diseases that involve problems with the hormone insulin and high blood glucose. As yet, there is no cure for diabetes.

- In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2015, diabetes was the direct cause of 1.6 million deaths.

- Early detection and treatments for potential patients are necessary to reduce the healthy risk of having diabetes.

Introduction

The Diabetes Dataset

- Pima Indians Diabetes Database from Kaggle.
- 768 observations, 8 variables as predictors and one target value.

1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1

Introduction

What did we do?

- Methods & Ideas:
  - Bayesian logistic regression model
  - Conventional logistic regression model

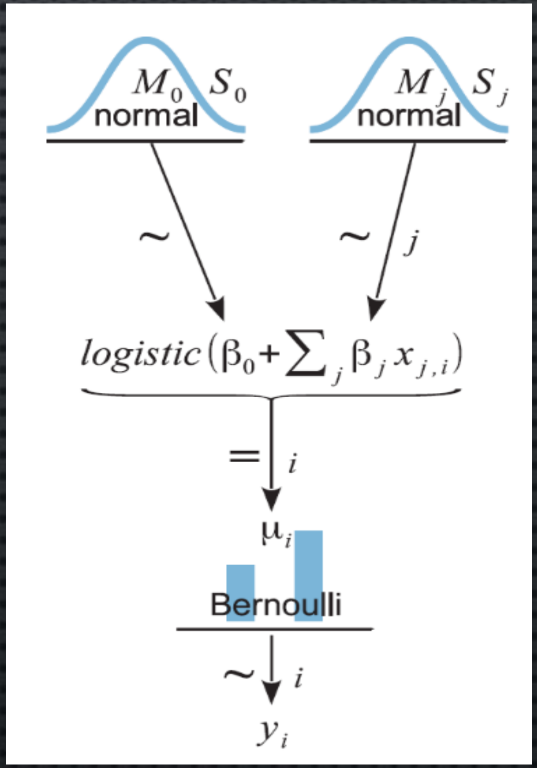
- Results & Analysis:
  - Importance of Data Preprocessing
  - Posterior distributions of coefficients and intercept
  - Effect of different Prior distributions

Methods & Ideas

- Bayesian General Linear Model
- $\mu = f(lin(x), [parameters])$ (1)
- $y \sim pdf(\mu, [parameters])$ (2)
- See Chapter 15

Lin(x) : logistic(x)
Pdf : Bernoulli

- Bayesian Logistic Model
- $\mu = logistic\left(\beta_0 + \sum \beta_j x_j\right)$ (3)
- $y \sim Bernoulli(\mu)$ (4)
- $logistic(x) = \frac{1}{1+e^{-x}}$ (5)
- See Chapter 21



Standardize:

$$z_j = \frac{x_j - \bar{x}_j}{s_{x\,j}}$$

*y[i] ~ dbern (ilogit (zbeta0 + sum (zbeta [1: Nx] * zx [i,1: Nx])))*

Methods & Ideas

Split

Train a conventional logistic model

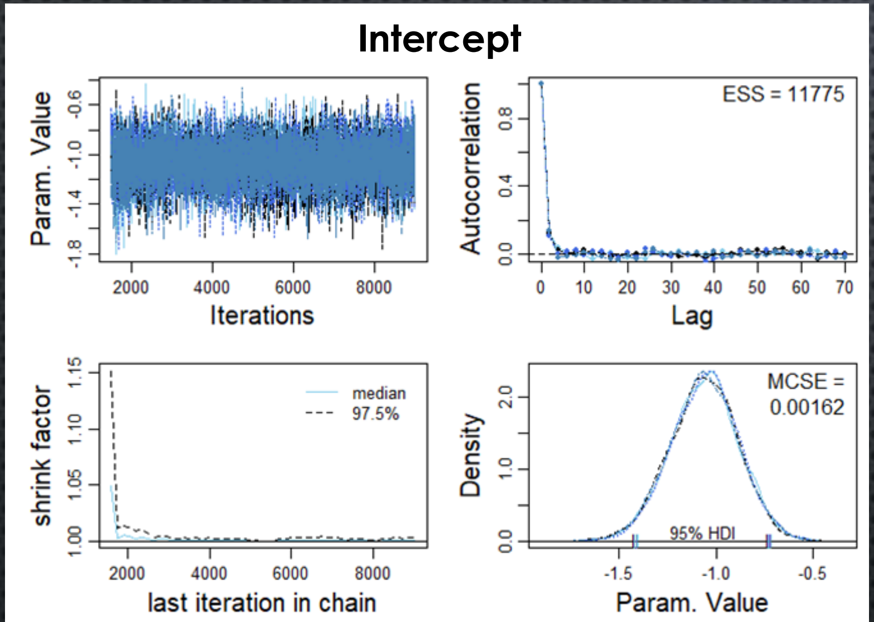
Train a Bayesian logistic model

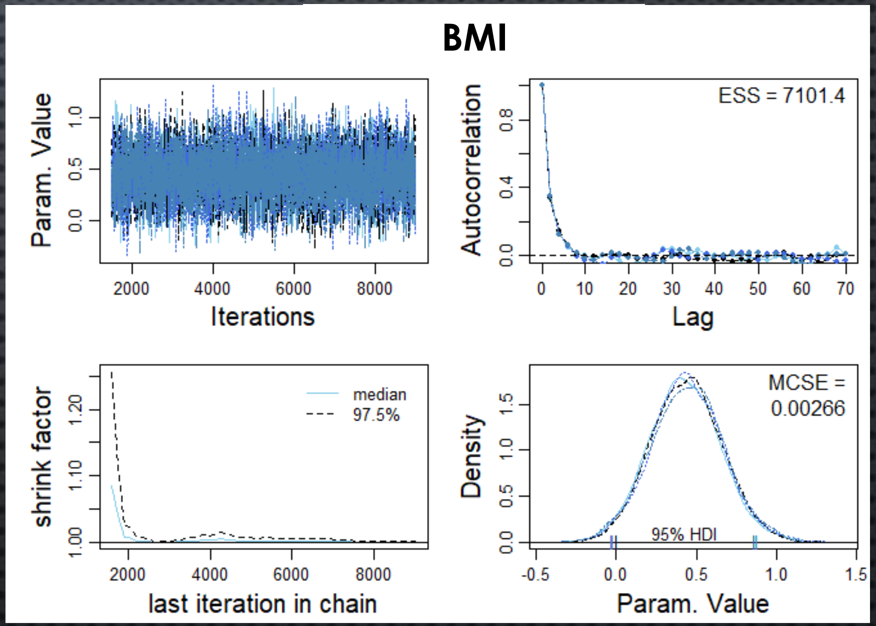
Use mode to build logistic function

Compare with AUC and Hit Rate

$$y = \mathrm{logistic}\left(mode(\beta_0) + \sum mode(\beta_j)x_j\right) \quad (7)$$

SELECTED DIAGNOSTICS OF MCMC CHAINS





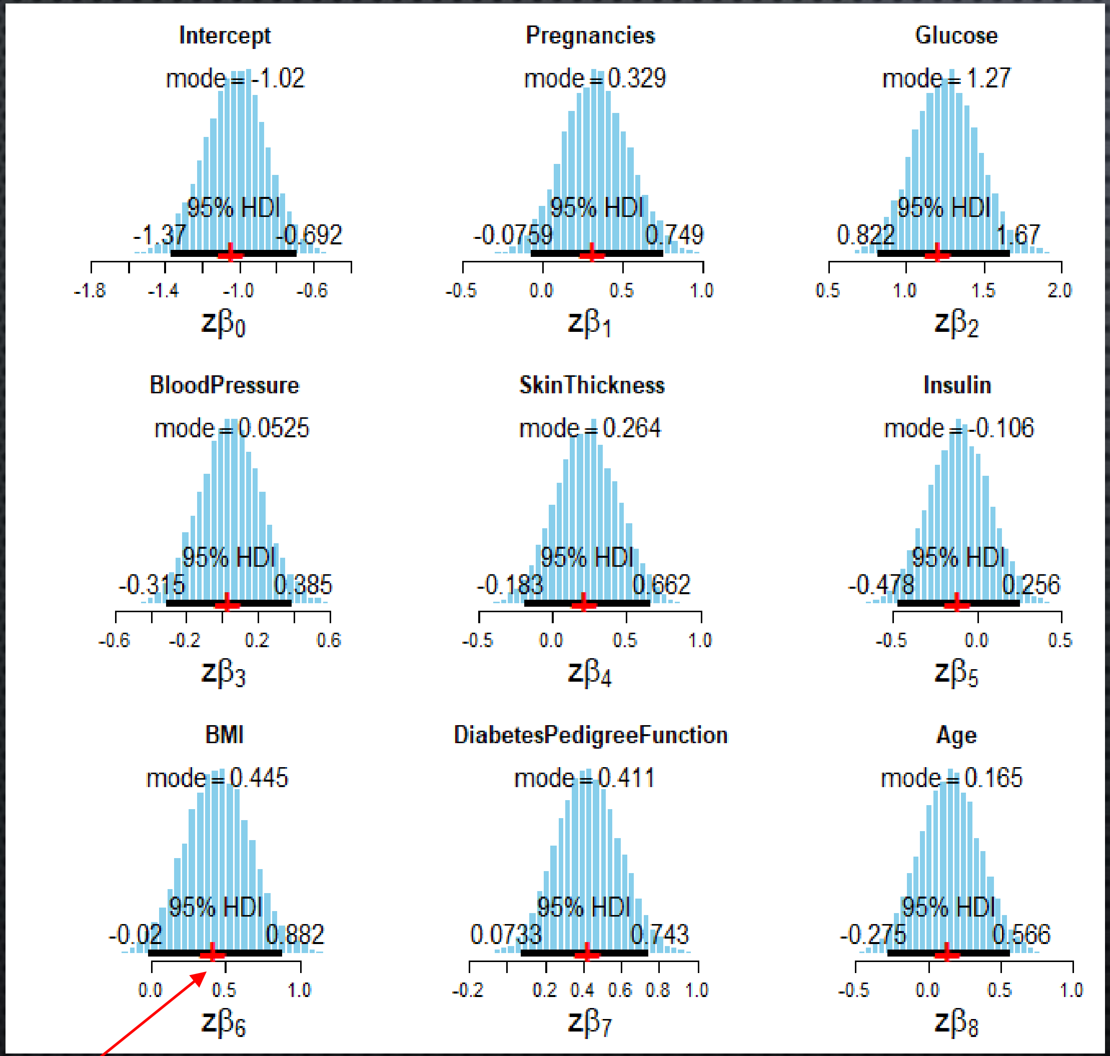
- Preprocessing : Remove missing values and kept 392 observation and 9 variable in the final dataset

Diagnostic information suggests good representativeness of the posterior distribution

Experimental Results & Analysis

Priors (β_0 - β_7) ~ dnorm (0, 0.25)

Note that in JAGS this indicates a normal distribution with mean as 0 and standard deviation of 2



coefficients from the traditional logistic regression

Logistic regression (hit rate=82%)			Bayesian logistic regression (hit rate=83%)		
	[Pred] 0	[Pred] 1		[Pred] 0	[Pred] 1
[Act] 0	74	17	**[Act] 0**	73	15
[Act] 1	4	22	**[Act] 1**	5	24

Model comparison

- Use the mode values of the posteriors from Bayesian logistic regression
- Test both models using the testing set

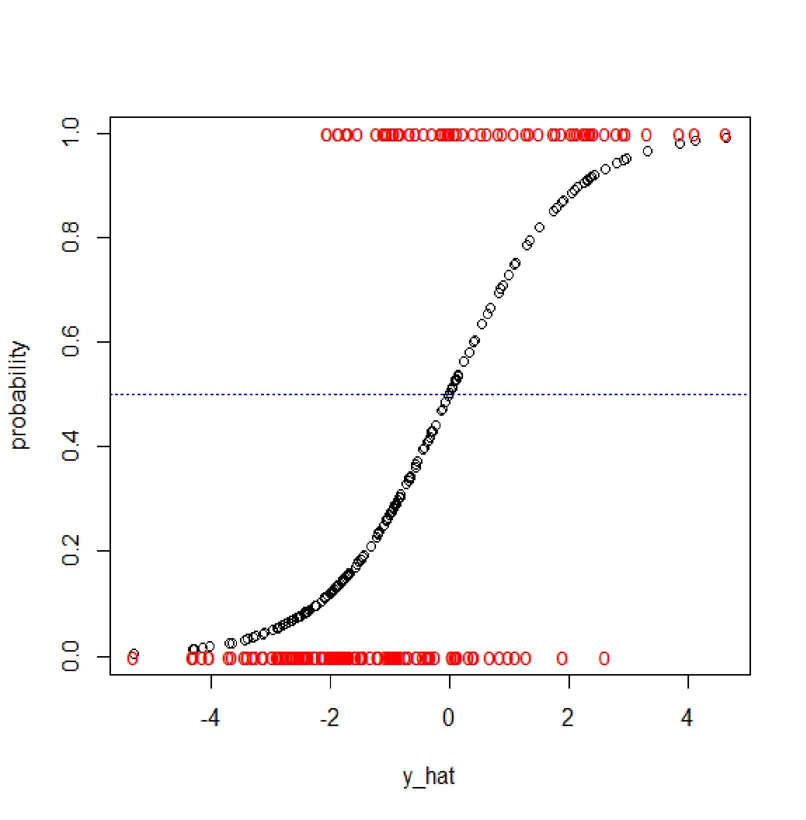
The two regression models agreed very well

Experimental Results & Analysis

Sensitivity analysis I : Unbalanced classes

- The preprocessed dataset has unequal number of records in the 1 and 0 class.
- Parameter estimates become more uncertain (i.e. larger HDIs) when records are unbalanced.

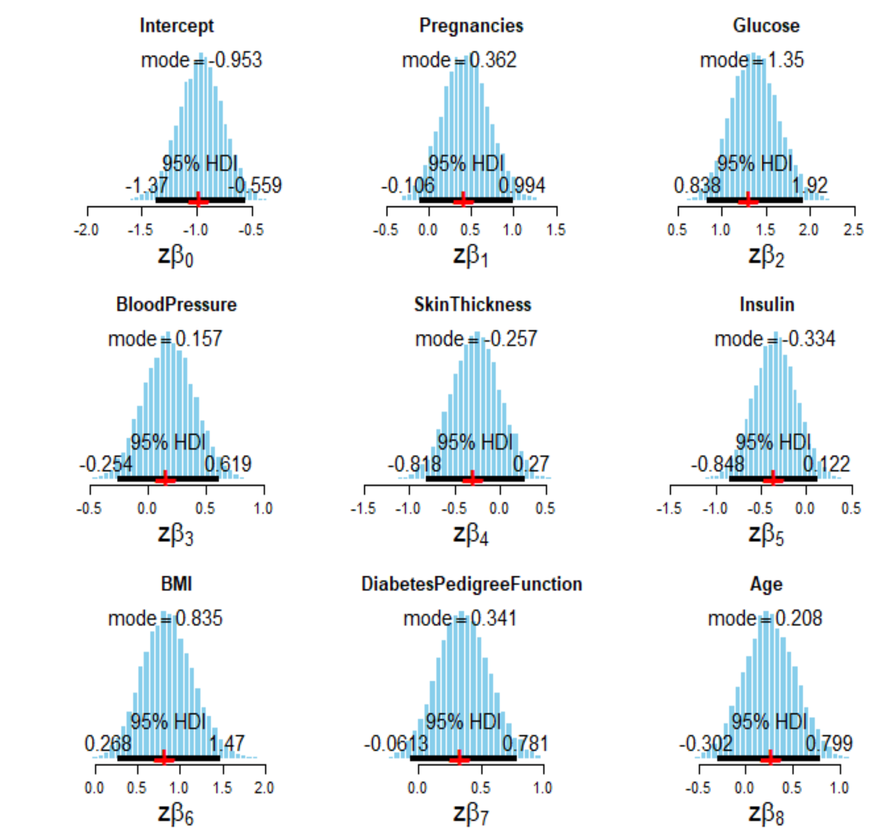
Data	Label 0	Label 1
Original	262	130
Balanced	130	130



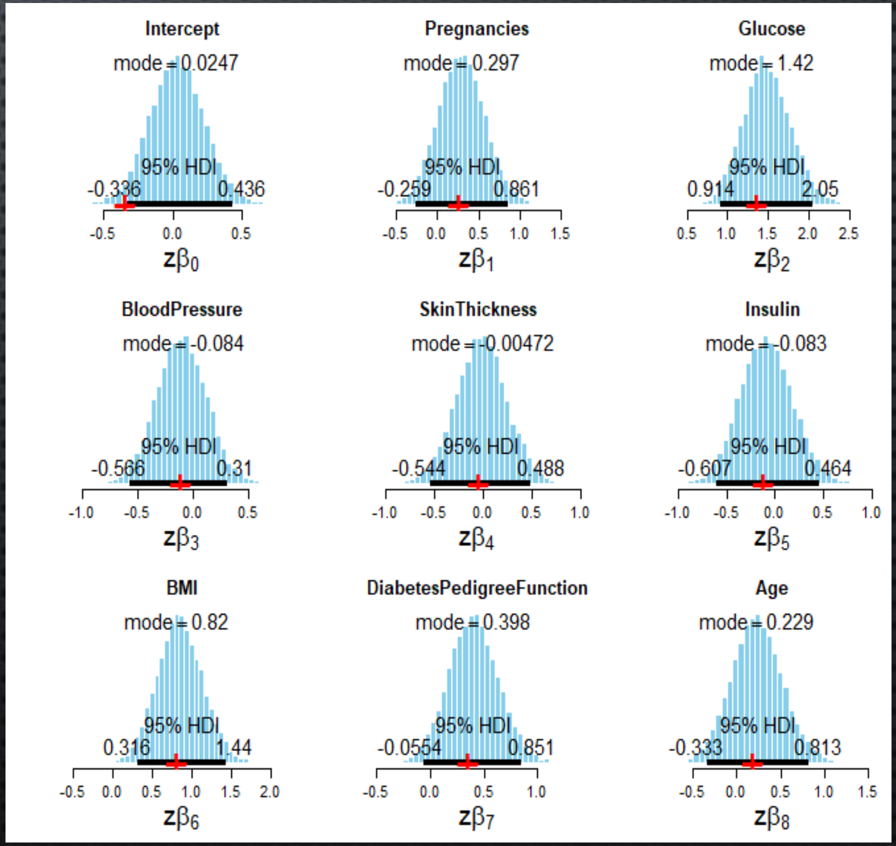
Experimental Results & Analysis

Sensitivity analysis I : Unbalanced classes

- Posterior without balanced data



- Posterior with balanced data

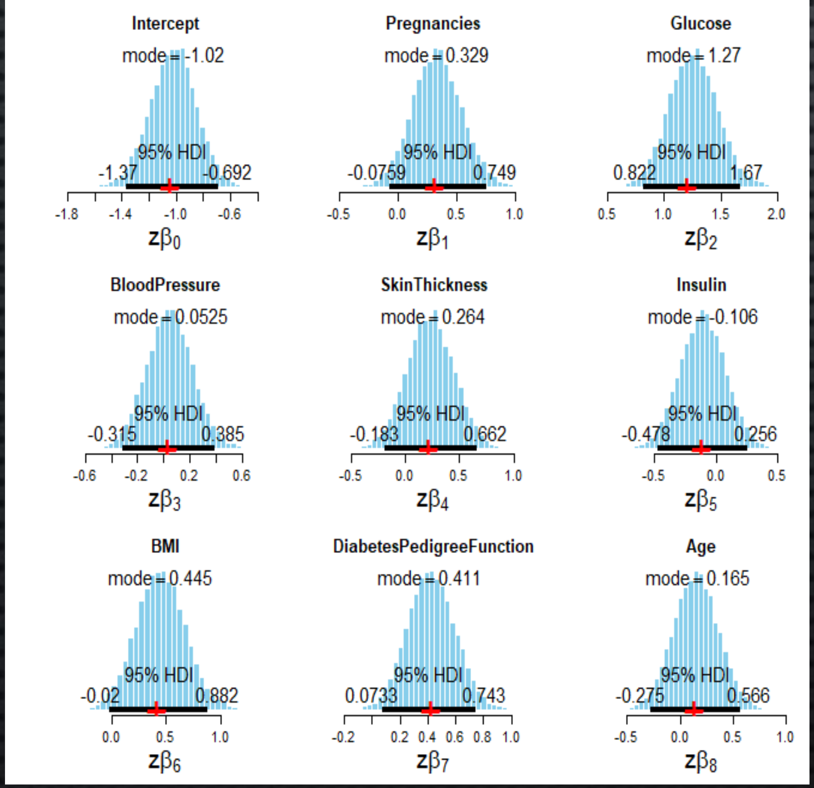


Experimental Results & Analysis

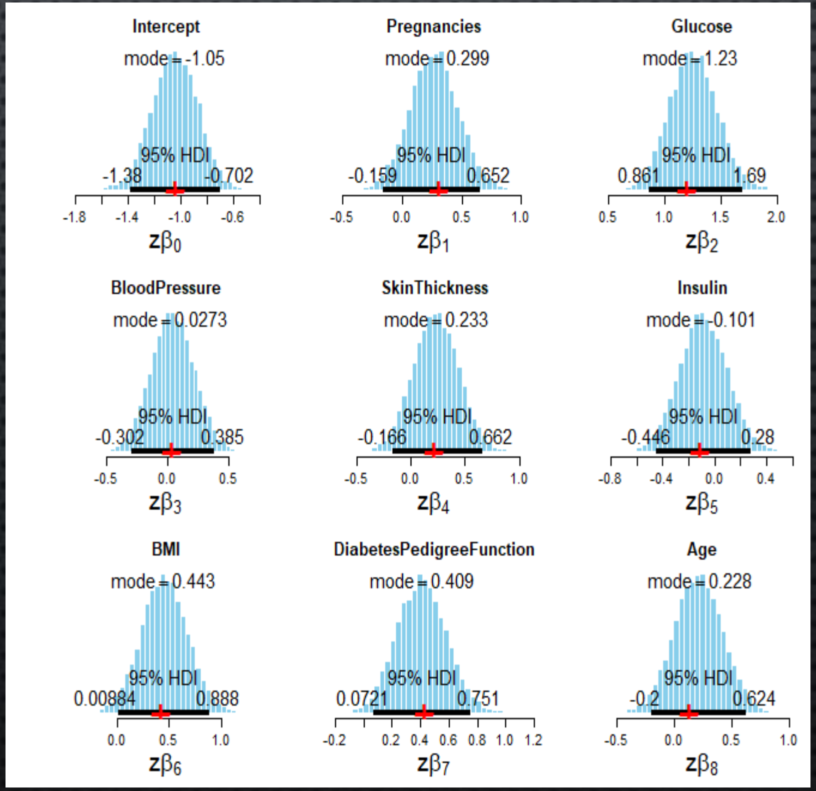
Sensitivity analysis II: different priors

Vague prior

- Priors (β_0 - β_7) ~ Norm(0, 2)

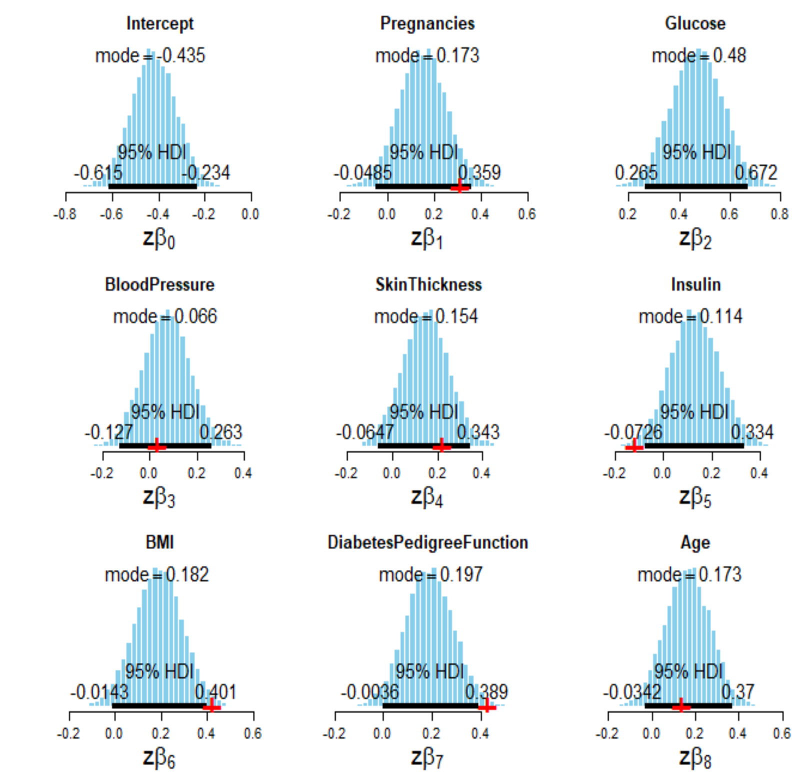


- Priors come from the posteriors after running 50 random selected samples



Relatively informative prior

- Priors (β_0 - β_7) ~ Norm(0, 0.14)



Conclusion

- Bayesian logistic regression and conventional logistic regression only display marginally differences in their performance gauged by hit rates.
- Unbalanced number of records in the two classes is not a dictating factor to the posterior distributions of the coefficients, whereas samples size and selection of variables play a more significant role.
- The evaluation of various priors also suggests that the posterior is synergy of prior and likelihood that is determined by the dataset.
- Altering priors had a significant effect on the posterior distribution of the coefficients.

REFERENCE

- Diabetes Health Center, https://www.webmd.com/diabetes/default.htm
- Diabetes World Health Organization, http://www.who.int/mediacentre/factsheets/fs312/en/
- Diabetes Database, https://www.kaggle.com/uciml/pima-indians-diabetes-database/data
- Case Study on Pima Indian Diabetes. Available at https://www.kaggle.com/hinchou/case-study-on-pima-indian-diabetes
- Kruschke, J.. Doing Bayesian data analysis, 2nd Edition: A Tutorial with R, JAGS and Stan. 2015 Academic Press/Elsevier.