Application of MCMC to Estimate Parameters on Binary Response Data

Jingrui Mu

300130858

Course: **MAT5314**

School of Science, University of Ottawa

Professor: **Mayer Alvo**.

Content

1. Data Description

A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. I divided these data into two sets. The training set contains 532 subjects, and the testing set contains 200 subjects with missing values in the explanatory variables.

1.1 Format

These data frames contain the following columns:

Table1.1. Data Format

Variables	Description	Character
npreg	Number of Pregnancies	Integer
glu	Plasma Glucose Concentration in an Oral Glucose Tolerance Test	Integer
bp	Diastolic Blood Pressure (mm Hg)	Integer
skin	Triceps Skin Fold Thickness (mm)	Integer
bmi	Body Mass Index (Weight in kg/(height in m)\^2)	Float
ped	Diabetes Pedigree Function	Double
age	Age in Years	Integer
type	Yes or No, for diabetic according to WHO criteria	Logical

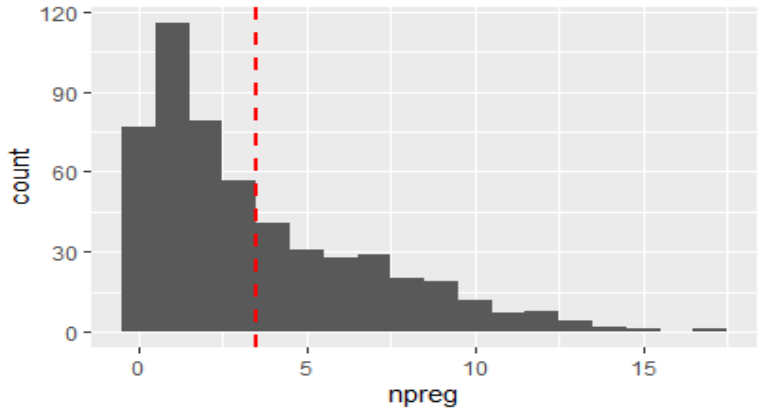
In the dataset, variable “type” is the response, Y=0 or Y =1. And the rest other 7 variables are assumed as explanatory variables.

Table 1.2 Response and Explanatory Variables

	Variables	Attribute
Y	type	Y=0, or Y=1
X1	npreg	Continuous
X2	glu	Continuous
X3	bp	Continuous
X4	skin	Continuous
X5	bmi	Continuous
X6	ped	Continuous
X7	age	Continuous

4.1 ***Description of Predictors***

To show relationship between these variables. Firstly, I did a summary of variable npreg and compares diabetes and non-diabetes in relation to npreg.



Firgure1.1 Histogram of npreg

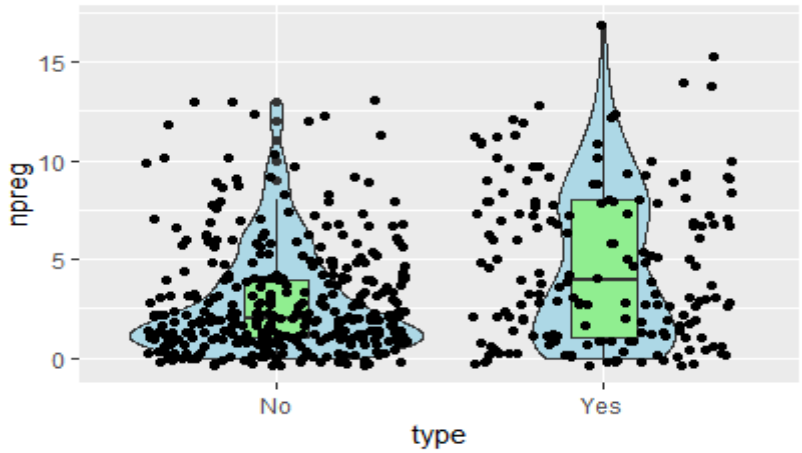
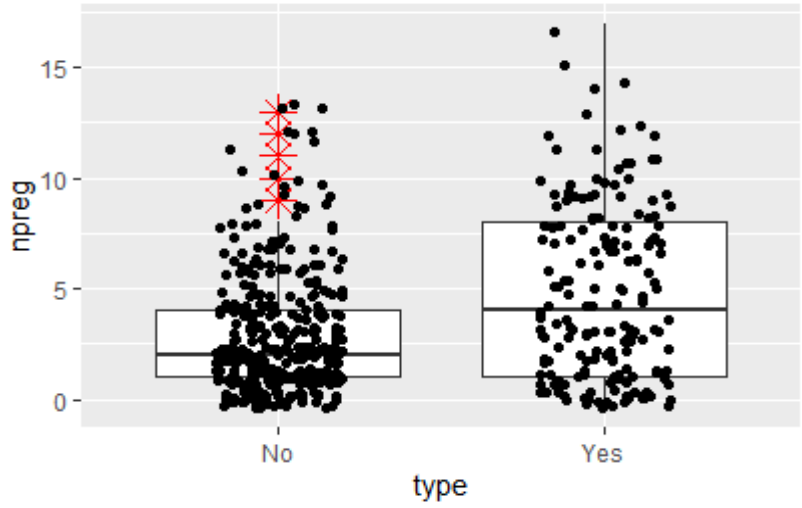


Figure1.2 Boxplot and Volin_plot of npreg



Firure1.3 Boxplot of npreg

From these two plots, which are combination of boxplot and volin_plot of npreg and the boxplot, respectively, we can see that the distributions of npreg on diabetic and

nondiabetic are totally different. Therefore, we can guess npreg is an important variable affecting a woman who is a diabetic or not.

Next, I used the same way to find the relationship between these variables.

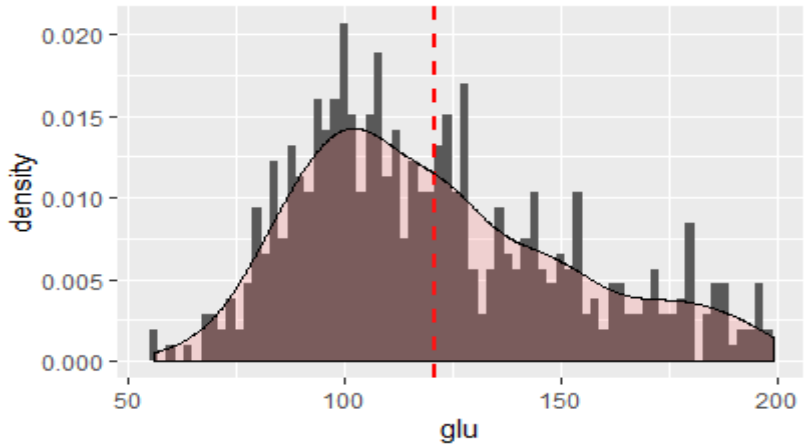


Figure 1.4 Combination of Histogram and density plot of glu

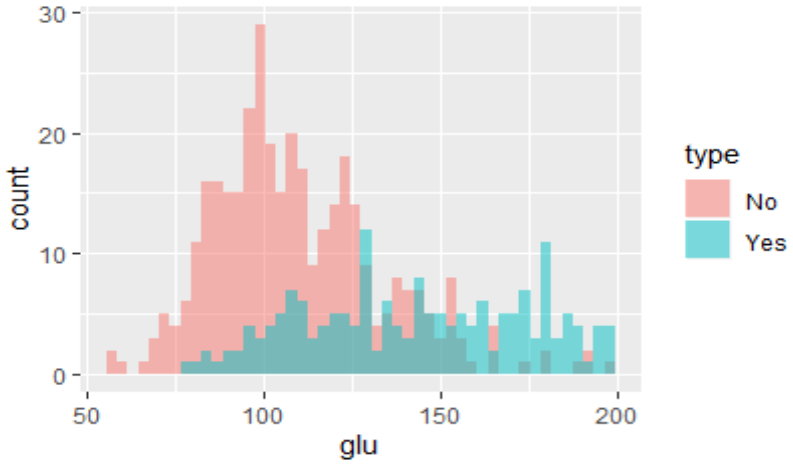


Figure 1.4 Histogram of glu on diabetic and nondiabetic

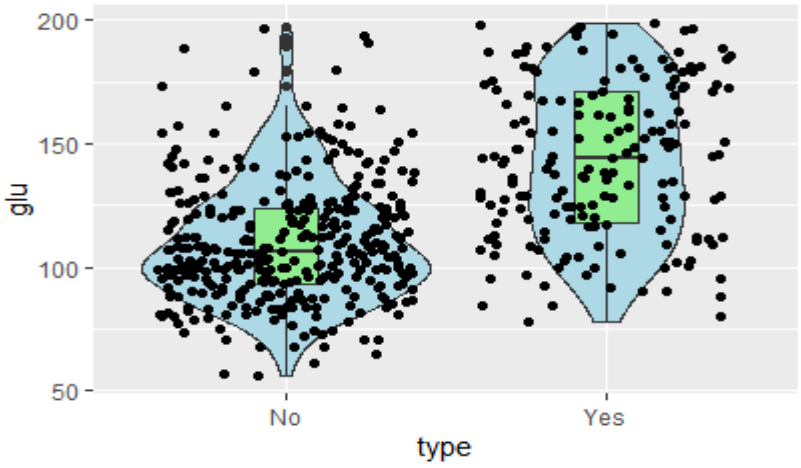


Figure 1.5 Boxplot and Volin_plot of glu

From these two plots (Figure 1.4 and Figure 1.5), we also can see that the distributions of glu o diabetic and nondiabetic are different. And in Chapter5, I will discuss the relationship between glu and probability of having diabetes.

- Bp

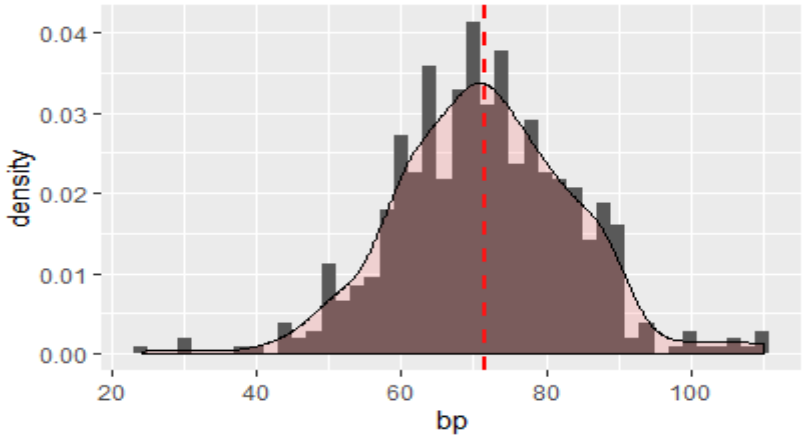


Figure 1.6 Histogram and Density Plot of Bp

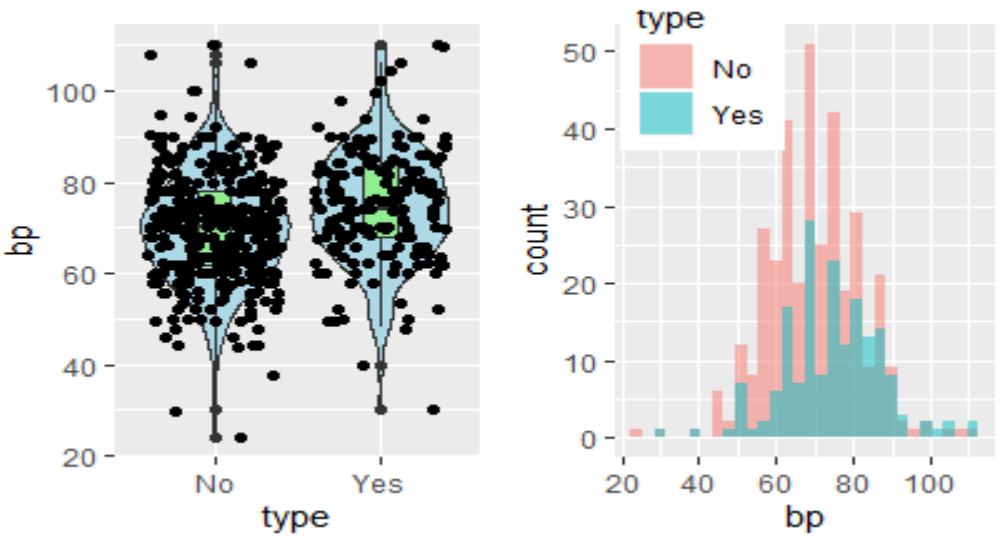


Figure 1.7 Histogram, Boxplot and Volinplot of Bp on diabetic and nondiabetic

About the BP, we can see that the Volinplots of it on diabetes and non-diabetes look similar. Maybe BP has litter influence on having diabetes. In Chapter 4, I will use Bayes Factors to exclude some unimportant variables.

- Skin

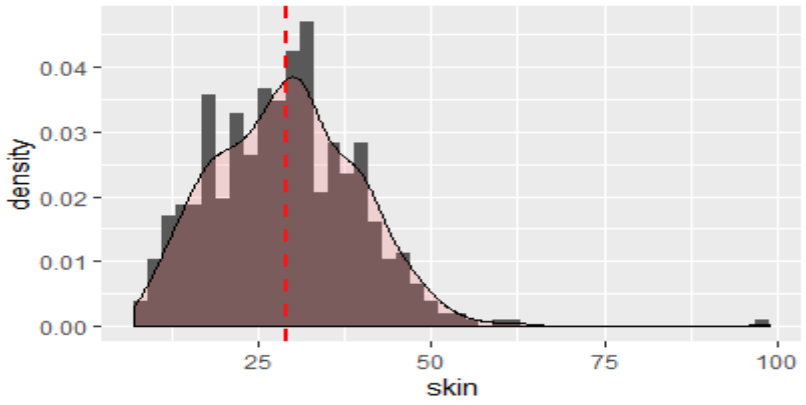


Figure 1.8 Histogram and Density Plot of Skin

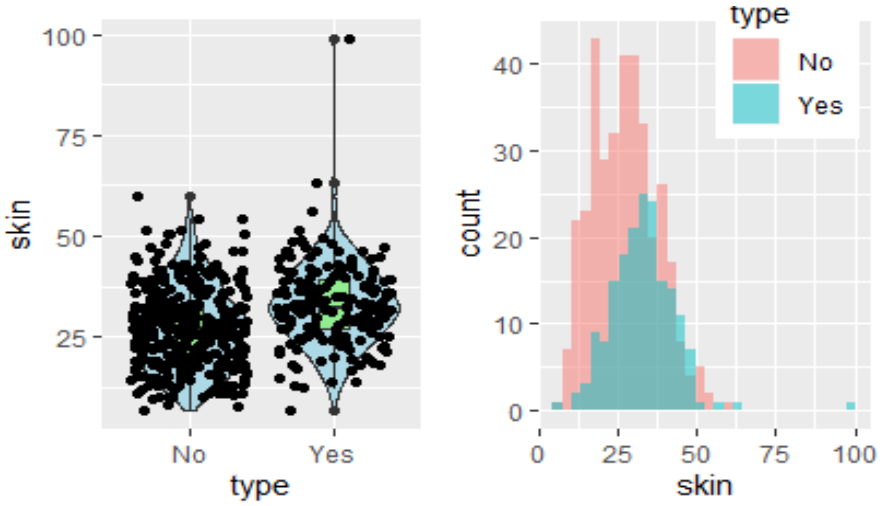


Figure 1.9 Histogram, Boxplot and Volinplot of Skin on Diabetic and Nondiabetic

From the volinplot of skin, we also can see that the different distributions of skin on diabetes and non-diabetes look similar. In Chapter4, I will use Bayes Factors to verify whether it is an important predictor.

- Bmi

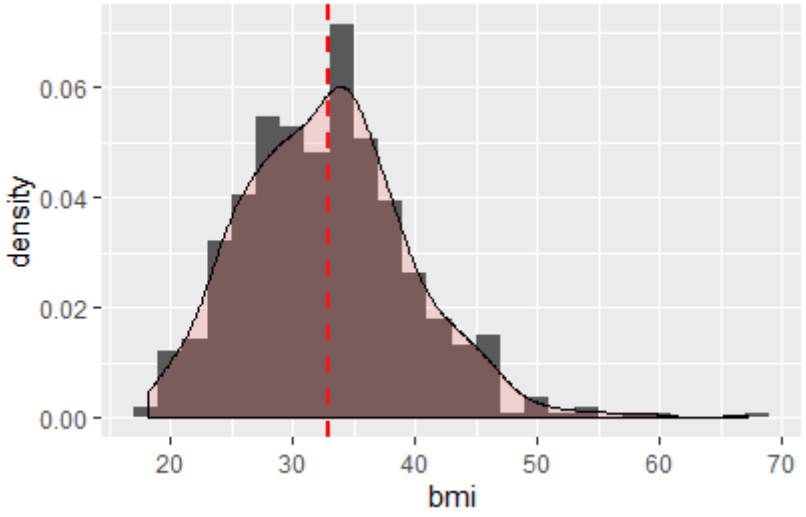


Figure 1.10 Histogram and Density Plot of Bmi

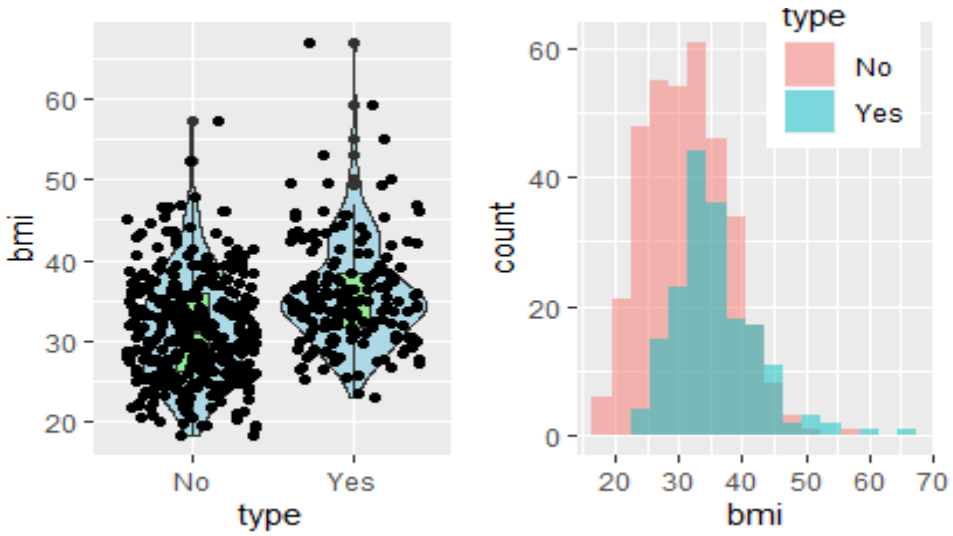


Figure 1.11 Histogram, Boxplot and Volinplot of bmi on Diabetic and Nondiabetic

The Volinplot of BMI shows that the distributions of bmi on diabetes and non-diabetes is totally different. For diabetes group, the median of BMI is much higher than non-diabetes group. In Chapter5, I will also discuss the relationship between BMI and probability of having diabetes.

- Ped

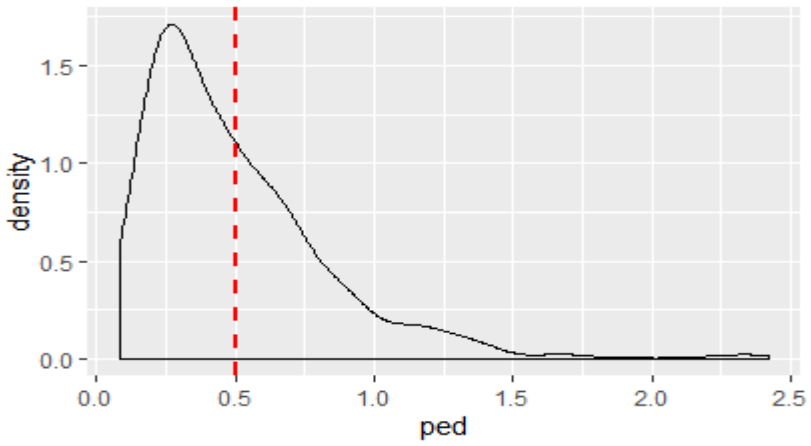


Figure 1.12 Density Plot of ped

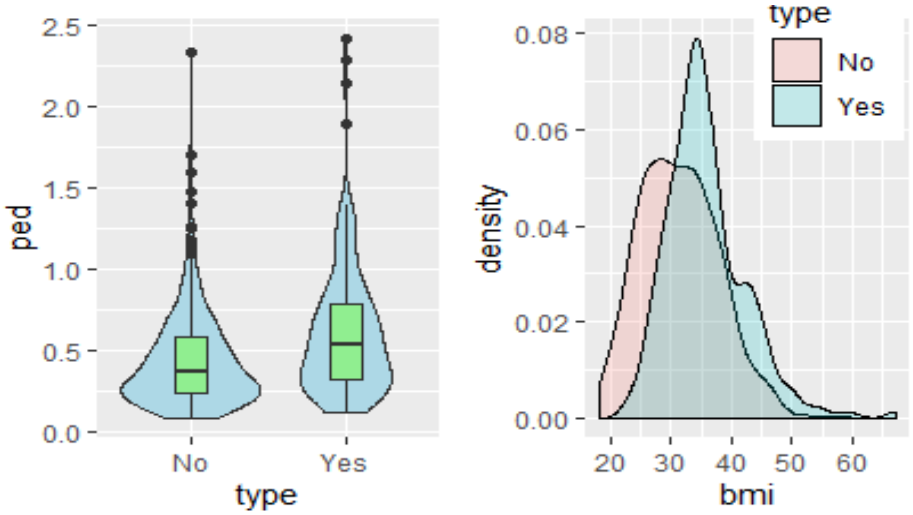


Figure 1.13 Histogram, Boxplot and Volinplot of Ped on Diabetic and Nondiabetic

- Age

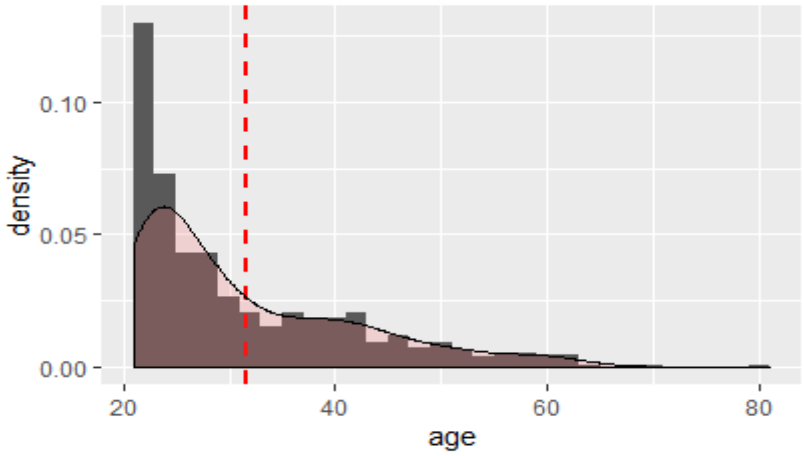


Figure 1.14 Histogram and Density Plot of age

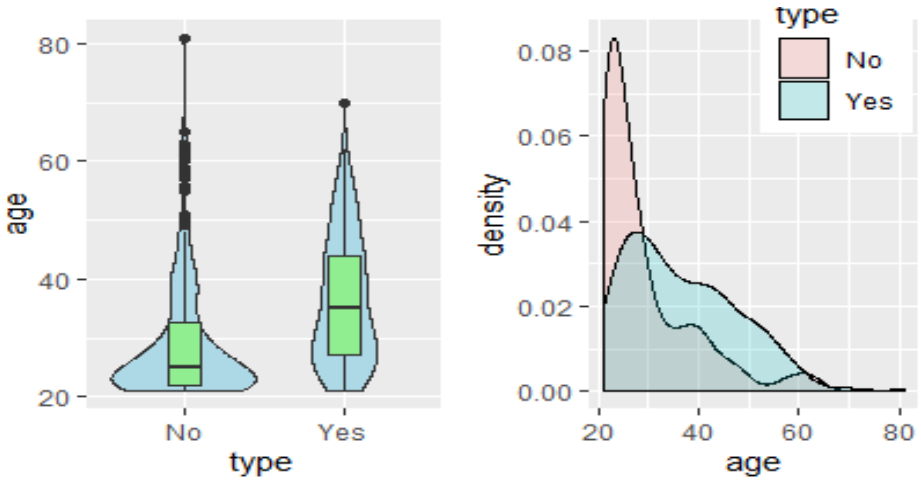


Figure 1.15 Histogram, Boxplot and Volinplot of Ped on Diabetic and Nondiabetic

Figure 1.15 displays the distribution of age on diabetes and non-diabetes. For diabetic group, the median of age is much larger than nondiabetic group. The median of age in diabetic group is about 35 years old, but in nondiabetic group, the median of age is close to 28 years old.

Note: all of the code about plots is in the Appendix-1. And these plots above all are plotted by using package "ggplot2"

4.1 ***Objectives***

1) Using Bayesian Probit Model with different priors on β and Bayesian Logit Model to estimate the posterior distribution of parameters of interest. In Chapter2, I use these two different models to predict responses with new dataset X. For predicting responses, I also use Support Vector Classifier (SVC) model.
2) For testing the prediction accuracy of Bayesian models, ROC curve is used to compare Bayesian Models with SVC Model in Chapter3.
3) In Chapter4, for promoting models to get a more accurate result, I use Bayes Factors to exclude two unimportant predictors.
4) In Chapter5, I am interested the relationship between glu and probability of having diabetes, BMI and chance of having DMII, and the relation of age and having diabetes, so I discuss them based on previous Bayesian Probit Model, which has a better performance. With these discussions, I draw some conclusions about how to keep away from diabetes.

2 Models

2.1 Bayesian Probit Model

Albert and Chib (1993) developed a Gibbs sampling method by regarding this problem as a missing-data problem for simulating from the posterior distribution.In the dataset, there are observed binary responses $y_1, \cdots, y_n$. Associated with the ith response, one observes the values of k explanatory variables $x_{i1,}x_{i2}, \cdots, x_{ik}$. In the probit regression model, the probability of $y_i = 1$, p_i=1, is written as:

$$\theta_i = P(y_i = 1) = F(\beta_0 + x_{i1}\beta_1 + \cdots + x_{ik}\beta_k), \tag{2.1}$$

Where $F(\cdot)$ is the cdf of a standard normal distribution.

1) Uninformative Uniform Prior on $\boldsymbol{\beta}$

I place a uniform prior on β, then the posterior density is given by,

$$g(\beta|y) \sim \prod_{i=1}^{n} \theta_i^{y_i} (1-\theta_i)^{y_i}, y_i = 0 \; or \; 1 \tag{2.2}$$

In the dataset, the binary response y_i is an indicator of whether a woman is a diabetic

or nondiabetic, where y_i=1 indicates the person is a nondiabetic, and $y_i = 0$ indicates the person is a diabetic. Define there is a continuous variable than can measure health:

$$Z_i = x_i^T \beta + \varepsilon_i, Where\ \varepsilon_i \sim N(0, \sigma^2) \tag{2.3}$$

So We can regard this problem as modeling a normal regression model on latent variables $Z_1, Z_2, \cdots, Z_n$ and we can only observe whether $Z_i > 0$ $(y_i = 1)$ or not.

Next, I use the Gibbs sampling to estimate parameters. In the algorithm, we will add the latent variables $Z = (Z_1, Z_2, \cdots, Z_n)$ to the estimated parameters and sampling from the joint posterior distribution of Z and β.

Based on the dataset and Probit Regression Model, I define Z_i as follows:

$$Z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \varepsilon_i \tag{2.4}$$

And the probability of diabetic is given by:

$$\begin{gathered} \theta_i = F(x_i^T \beta) \\ y_i | \theta_i \sim Bernoulli(\theta_i) \end{gathered} \tag{2.5}$$

As it mentioned above, I place an uninformative uniform prior on β, and assume ε_i represents identical independent distribution sample from standard normal distribution.

$$\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \sim iidN(0,1)$$

Referring to "Bayesian Computation with R" by Jim Albert and lecture notes, I rewrite the joint density of (Z, β) to get the full conditional distribution of each parameter given the rest:

$$\begin{gathered} g(\beta | Z, data) \sim N_8((X^T X)^{-1} Z, (X^T X)^{-1}) \\ g(Z_i | \beta, data) \sim N(x_i^T \beta, 1) I(Z_i > 0), y_i = 1 \\ \text{Where } x_i^T = (1, x_{i1}, x_{i2}, \cdots, x_{i7}) \end{gathered} \tag{2.6}$$

After writing all conditional distributions, we can generate values of parameters of interest, $\beta_0, \beta_1, \cdots, \beta_7,\ based\ on\ Gibbs\ Sampling$. The following table is results of the simulation:

Table 2.1 Estimated Parameters of Bayesian Probit Model

parameters	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
mean	-5.588	0.071	0.021	-0.005	0.004	0.049	0.660	0.016
sd	0.538	0.024	0.002	0.006	0.008	0.013	0.195	0.008

Note: the full code about Bayesian Probit Model is on the Appendix-2. I use the function *bayes.probit* form LearnBayes Package.

2) Informative Prior on β

Referring to the Section 10.3.2 from "Bayesian Computation with R", I place an informative prior on β to get values of parameters of interest. Suppose β is assigned to a multivariate normal prior with mean vector β_0 and variance-covariance matrix V_0. Based on the lecture notes, I rewrite the joint density and get the conditional distribution of β:

$$g(\beta|Z, data) \sim N_8(\beta^1, V_1)$$
$$\beta^1 = (X^T X + V_0^{-1})^{-1}(X^T Z + V_0^{-1}\beta^0), V_1 \quad (2.7)$$
$$= (X^T X + V_0^{-1})^{-1}$$

A convenient choice for the prior is the Zellner g prior introduced in Section 9.3 of "Bayesian Computation with R". Let β have a normal distribution with mean vector 0 and variance-covariance matrix $c(X^T X)^{-1}$, where c is a large value, c=100. And then in the next chapter, I still assign β^P (β^P represents a subset of predictors) a prior of the same functional form with the same value c when choosing appropriate predictors. The following table is the results of Gibbs Sampling with informative prior on β:

Table 2.2 Estimated Parameters of Bayesian Probit Model with Informative Prior

parameters	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
mean	-5.364	0.069	0.020	-0.004	0.004	0.046	0.632	0.015
sd	0.515	0.024	0.002	0.006	0.008	0.013	0.190	0.008

Note: the full code about Bayesian Probit Model is on the Appendix-2.

Based on these tables, we can see that the two Bayesian Probit Models with different

priors have the similar results.

2.2 ***Bayesian Logit Model***

In Logistic Regression, we often regress the binary response Y_i onto the covariate X_i as follows:

$$logit(P(Y_i = 1)) = X_i^T \beta \tag{2.8}$$

Where X_i is a column of the design matrix. Equivalently,

$$P(Y_i = 1) = exp(X_i^T \beta)/1 + exp(X_i^T \beta) \tag{2.9}$$

Similarly, we can write the joint density of $(\beta, data)$,

$$g(\beta|y) \sim (\prod_{i=1}^{n} \theta_i^{y_i} (1-\theta_i)^{y_i}) g(\beta), y_i = 0 \; or \; 1 \tag{2.10}$$

$\theta_i = P(Y_i = 1)$, $g(\beta)$ is the prior on β

And I place a multivariate normal prior on β:

$$\beta \sim \mathrm{N}_8(\beta_0, \Sigma_0)$$

$$\beta_0 = (-6,0,0,0,0,0,0,0), \Sigma_0 = 10 \times I_8$$

The following table is the values of posterior of β:

Table 2.3 Estimated Parameters of Bayesian Logit Model

parameters	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
mean	-7.195	0.133	0.038	-0.015	0.009	0.045	0.717	0.008
sd	0.522	0.044	0.009	0.006	0.014	0.021	0.451	0.019

Note: I define a new function based on MH sampling, which is *Bayes.logistic*, to simulate the posterior of interested parameters. The full code about Bayesian Logit Model is on the Appendix-3.

Form these tables above, we can see that the results of Bayesian Logit Model and Probit Model are a little different. The standard deviation of Logit Model is larger than Probit Model. In next chapter, I will discuss the accuracy of these two models based on ROC.

2.3 Support Vector Classifier

The original SVM algorithm was invented by Vladimir N. Vapnik and Alexey Ya,

Chervonenkis in 1963. And now SVM has been widely used when doing classifications. Referring to Section 9.3 from "An Introduction of Statistical Learning with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, we can know that the support vector classifier classifies an observation depending on which side of hyperplane it lies. The hyperplane can correctly separate most of the observations into two classes, but may still misclassify a few observations. The solution to the optimization problem is as follows:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n} mize \; M, \sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1-\varepsilon_i)$$

$$\varepsilon_i \geq 0, \sum_{i=1}^{n} \varepsilon_i \geq C$$

Where C is an nonnegative turning parameter and M is the width of the Margin. We seed to make this margin as large as possible. $\varepsilon_1, I, \varepsilon_n$ are slack variables that allow individual observations to be on the wrong side of the hyperplane. Based on these above, the loss function of SVC can be written as follows:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} imize \left\{ \sum_{i=1}^{n} \max\left[0, 1-y_i f(x_i)\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2.4}$$

It's very similar to the loss function of Logistic Regression. I used SVC to predict response Y on dataset X. In next chapter, I will compare these models based on ROC curves.

3 Comparing Models

3.1 Comparing between Bayesian Probit Models

1) Bayesian Probit Model with noninformative prior on $\boldsymbol{\beta}$.

Based on Section 2.1, I generated the values of parameters of interest, $\beta_0, \beta_1, I, \beta_7$. Now I am interested in predicting response Y's with testing dataset X. Based on the results, ROC curve is used to test the goodness of the model.

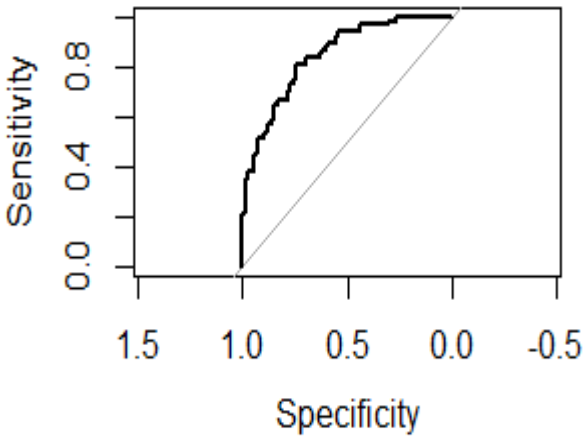


Figure 3.1 ROC Curve of Bayesian Probit Model with noninformative prior

The ROC curve has an AUC of 0.845 suggesting a good classification accuracy.

2) Bayesian Probit Model with Informative prior on β

Similarly, with Bayesian Probit Model having Informative prior on β, I got the posterior distribution of parameters, $\beta_0, \beta_1, I, \beta_7$. ROC curve of this Model is as follows:

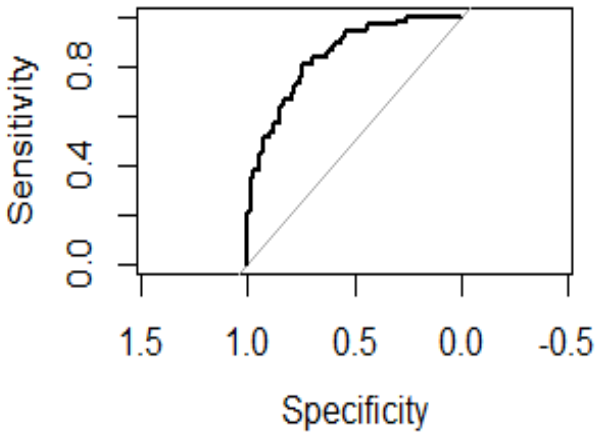


Figure 3.2 ROC curve of Bayesian Probit Model with Informative prior on β

The ROC curve has an AUC of 0.8454 also suggesting a good classification accuracy. These two Bayesian Probit Models both has a good performance. Therefore, different priors have little influence on accuracy of predicting.

3.2 Bayesian Logit Model

As it mentioned above, in the Bayesian Logit Model, the multivariate normal distribution was chosen as the prior and I generated the values of parameters $\beta_0, \beta_1, I, \beta_7$ with MH sampling. Also, the same dataset X is used to predict response Y's. The ROC curve is as follows:

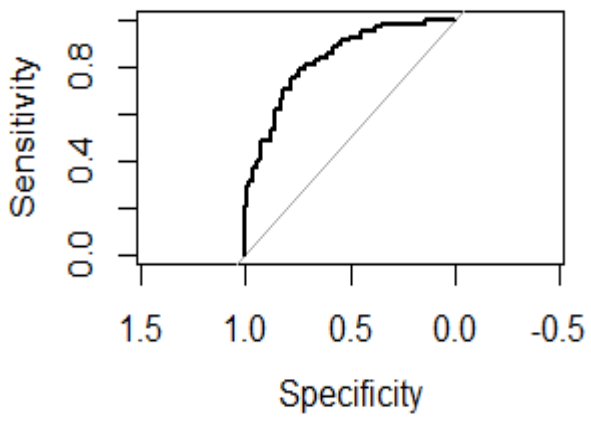


Figure 3.3 ROC curve of Bayesian Logit Model

The ROC curve has an AUC of 0.8363 also suggesting a good classification accuracy. But obviously, Bayesian Probit Models have a better prediction accuracy.

3.3 Support Vector Classifier

SVC or SVM (Support Vector Machine) has been used widely to do classifications. In this section, I compare SVC with Bayesian Models (Probit and Logit) to verify Bayesian Probit or Logit Regression Models also have a pretty good prediction accuracy. As it mentioned above, we solve the optimization problem to obtain the predictions.

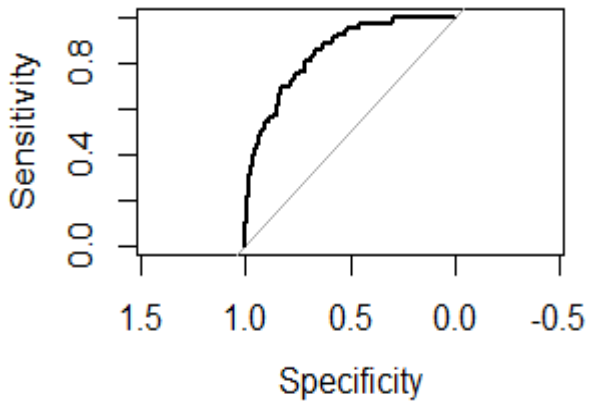


Figure 3.4 ROC curve of SVC model

The SVC ROC curve has an AUC of 0.8485, which is larger than Bayesian Models'. But Bayesian Probit Models' AUC is very close to it. Therefore, we can say Bayesian Probit Models are credible when predicting responses on new testing dataset.

Note: all of the code in chapter3 is attached in Appendix-4.

4 Models Promotion

Based on the plots of chapter1, I find some distributions of explanatory variables on different response Y look similar, which means I may put some unimportant predictors into the previous models. Referring to section10.3 and Chapter 8 from "Bayesian Computation with R", I use Bayes Factors to choose the better model with a subset of predictions.

From the plots of Chapter1, we can see that distributions of predictors, bp and skin, are similar on responses Y (0 or 1). So I compare the "Full" model with "excluded bp and skin" model. The log marginal densities of these two models are given respectively by -251.8974, and -248.487, so the Bayes Factor in support of the full model containing both variables is as follows:

$$BF = exp(-251.8974)/exp(-248.487) = 0.033$$

It indicates that there is support for excluding predictors, bp and skin, in the model. Table 4.1 displays the Bayes Factors comparing full model, excluding bp and skin model, only excluding bp model, and only excluding skin model. From the table, it is

clear that bp and skin are not so important predictors.

Table 4.1 Bayes Factors comparing models mentioned above

	Model 2			
Model 1	Full Model	Bp Excluded	Skin Excluded	Bp, Skin excluded both
Full Model	1	0.204	0.179	0.033
Bp Excluded	4.911	1	0.8802	0.162
Skin Excluded	5.579	1.136	1	0.184
Bp, Skin excluded both	30.276	6.165	5.427	1

Note: Each number represents the Bayes factor in support of Model1 over Model 2. All of code in the part is in Appendix-5.

Next, I only put five predictors, excluded bp and skin in the Bayesian Probit Model. From the ROC Curve, which has an AUC of 0.8466, we can see that the reduced model is better than full model.

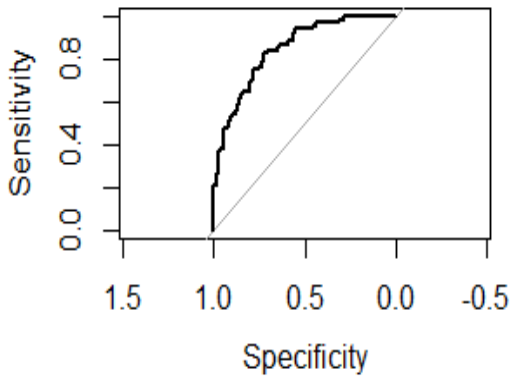


Figure 4.1 ROC Curve of reduced Bayesian Probit Model

Finally, the reduced model is as follows:

$$p_i = P(Y_i = 1) = F(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_6 x_{i6} + \beta_7 x_{i7})$$
$$y_i \mid p_i \sim Bernoulli(p_i) \tag{4.1}$$

5. Conclusion

In this chapter, I find the relationship between different explanatory variables and responses based on previous model on chapter4. Firstly, I estimate the probability of woman having diabetic for glu (Plasma Glucose Concentration) from 65 to 200 since glu is an important indicator for diabetics in medical issues.

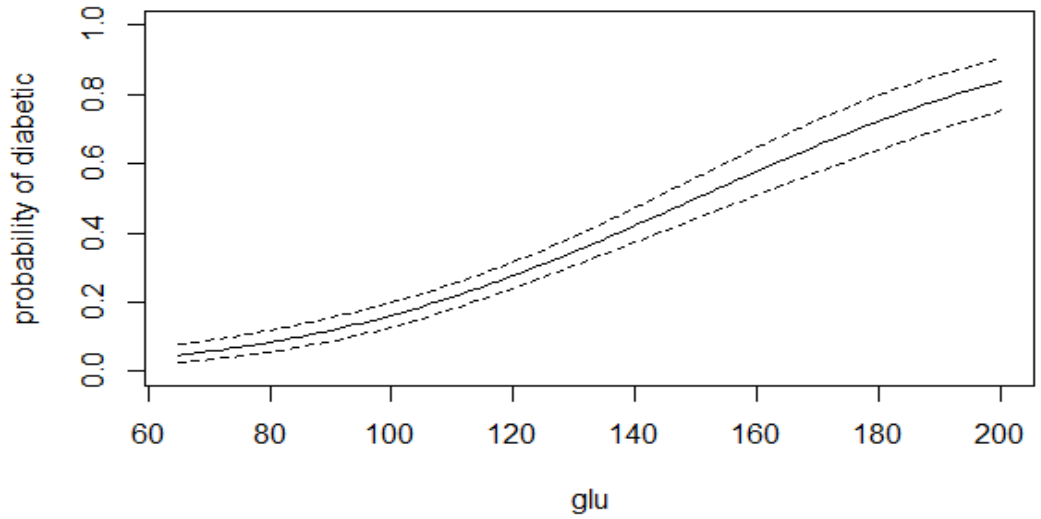


Figure 5.1 Posterior distribution of probability of diabetic for glu from 65 to 200. For each glu, the 5^{th}, 50^{th}, 95^{th} percentiles of the posterior are plotted.

For each glu, the solid line is the location of the median of having diabetic probability for a woman and the interval between the dashed lines corresponds to a 90% interval estimate for this probability. It clearly shows the relationship between glu predictors and responses. We can say that if a woman's glu is higher, the probability of having diabetic for her will be higher.

Similarly, I compute the probability of having diabetic for woman at different ages.

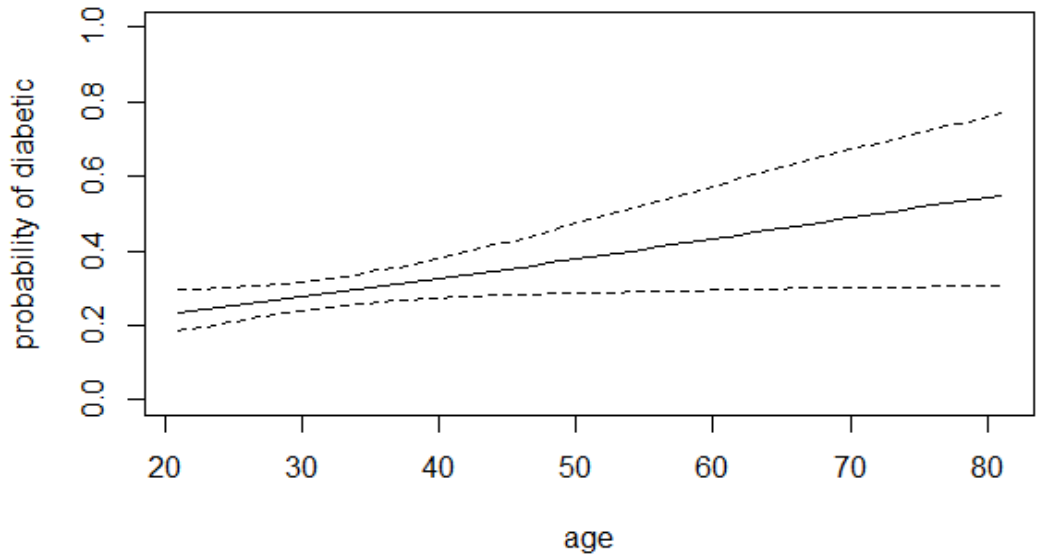


Figure 5.2 Posterior distribution of probability of having diabetic for women of different ages.

This plot clearly shows the relationship between age and the probability of having diabetic. For elder women, they are more likely to have diabetic than younger people. However, for women from 20 years old to 40 years old, we don't have significant evidence to say elder women are more easily to have diabetic. In other words, if we are beyond 40 years old, we need to be careful since the probability of having diabetic increases quickly for each increment unit in age.

Finally, for women with different BMIs, I compute the probability of having diabetic since BMI is an important index to measure the degree of obesity or healthiness for a person. Actually, most of diabetic patients also have obesity problem.

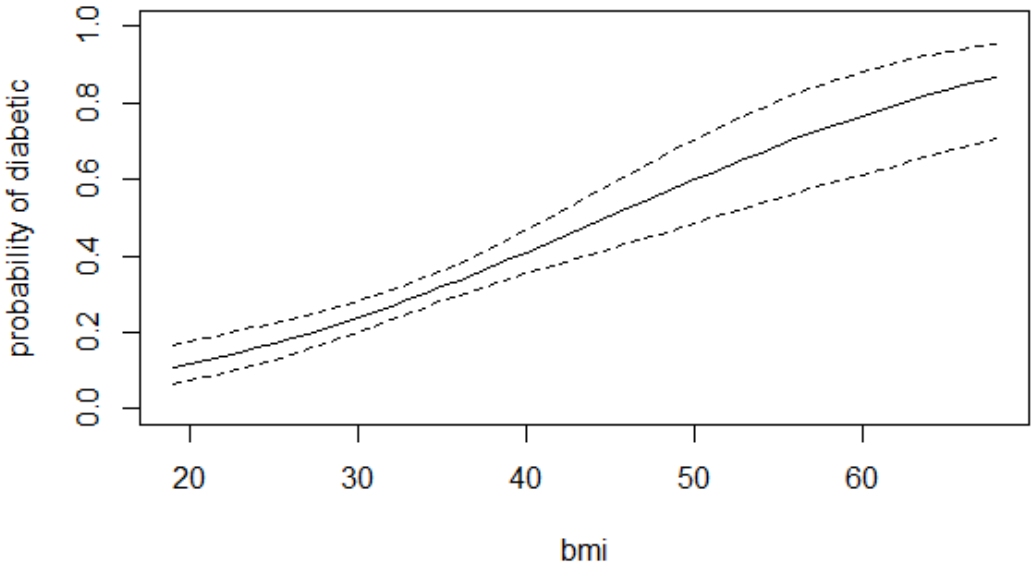


Figure 5.2 Posterior distribution of probability of having diabetic for women with different BMIs.

From this plot, we can say that BMI is an important indicator for diabetics and there is significant evidence to show that higher BMI, higher the probability of having diabetics. Therefore, exercising is a good way for us to keep away from diabetics.

Conclusions: Based on previous chapters, we can see that Bayesian Probit or Logit Models also have a good prediction accuracy. But in this example, Bayesian Probit Model has a better performance than Logit Model's. And after discussing the three relationships between glu and probability of having diabetic, age and probability of it, BMI and probability of it respectively, I find higher BMI, the probability of having diabetic will also be higher. And glu also has the same relationship as BMI's.

6 References

[1] Pr. Mayer Alvo. University of Ottawa, Course of Lectures, 2019

[2] Jim Albert, “Bayesian Computations with R”, 2009

[3] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, “An Introduction to Statistical Learning with Applications in R”, 2013

[4] Nicholas G. Polson, James G. Scott and Jesse Windle (2013), *“Bayesian Inference for Logistic Models using Polya-Gamma Latent Variables”, Journal of the American Statistical Association,* 1339-1349.

[5] Zellner, A. (1986), “On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions”, in P.K.Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Amsterdam: North-Holland.

[6] Congdon P. Bayesian model choice based on Monte Carlo estimates of posterior model probabilities[J]. Computational Statistics & Data Analysis, 2004

[7] Altaleb A, Chauveau D. Bayesian analysis of the Logit model and comparison of two Metropolis-Hastings strategies[J]. Computational Statistics & Data Analysis, 2001, 39: 137-152

[8] Lahiri K and Gao J. Bayesian analysis of nested logit model by Markov chain Monte Carlo[J]. Journal of Econometrics, 2002, 111:103-133

[9] Hastings W K. Monte Carlo sampling methods using Markov chains and their application[J]. Biometrika, 1970, 57:97-109

[10] Gibbons R D and Wilcox-Gok V. Health Service Utilization and Insurance Coverage: A Multivariate Probit Analysis[J]. Journal of the American Statistical Association, 1998, 441:63-72

[11] Dr. Hari M. Koduvely, “Learning Bayesian Models with R”, 2015

Appendix-1

```
getwd()
```

```
## [1] "F:/Bayesian_Project"
```

```
setwd("F:/Bayesian_project")
getwd()
```

```
## [1] "F:/Bayesian_project"
```

```
Pima.data<-read.csv("F:/Bayesian_project/Pima.csv",sep = ",",header=
TRUE)
library(ggplot2)
#npreg
dev.off()
```

```
## null device
##           1
```

```
plot_npreg<-ggplot(Pima.data,aes(x=npreg))+geom_histogram(binwidth =
 1)+geom_vline(aes(xintercept=mean(npreg, na.rm=T)),color="red",line
type="dashed",size=1)
# plot_npreg

boxplot_npreg<-ggplot(Pima.data,aes(x=type,y=npreg))+geom_boxplot(ou
tlier.colour="red",outlier.shape=8,outlier.size=4)+geom_jitter(shape
=16,position=position_jitter(0.2))
# boxplot_npreg
volin_npreg<-ggplot(Pima.data,aes(x=type,y=npreg))+geom_violin(alpha
=0.999,width=1)+geom_jitter(colour='black')
# volin_npreg
mix_npreg<-ggplot(Pima.data, aes(x=type, y=npreg))+geom_violin(fill=
"lightblue")+geom_boxplot(fill="lightgreen", width=.2)+geom_jitter(c
olour='black')
# mix_npreg

#glu
plot_glu<-ggplot(Pima.data,aes(x=glu))+geom_histogram(aes(y=..densit
y..),binwidth = 2)+geom_vline(aes(xintercept=mean(glu, na.rm=T)),col
or="red",linetype="dashed",size=1)+geom_density(alpha=.2, fill="#FF6
666")
# plot_glu
glu_histogram<-ggplot(Pima.data, aes(x=glu, fill=type))+geom_histogr
am(binwidth=3, alpha=.5, position="identity")
# glu_histogram
mix_glu<-ggplot(Pima.data, aes(x=type, y=glu))+geom_violin(fill="lig
```

```
htblue")+geom_boxplot(fill="lightgreen", width=.2)+geom_jitter(colou
r='black')
# mix_glu

# bp
plot_bp<-ggplot(Pima.data,aes(x=bp))+geom_histogram(aes(y=..densit
y..),binwidth = 2)+geom_vline(aes(xintercept=mean(bp, na.rm=T)),colo
r="red",linetype="dashed",size=1)+geom_density(alpha=.2, fill="#FF66
66")
# plot_bp
# par(mfrow=c(1,2))
# Define a function to display more than one plots in one picture
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL)
{
  library(grid)
  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)
  numPlots = length(plots)
  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols))
  }
  if (numPlots==1) {
    print(plots[[1]])
  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(lay
out))))
    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this
subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}
```

```
bp_plot<-ggplot(Pima.data, aes(x=type, y=bp))+geom_violin(fill="ligh
tblue")+geom_boxplot(fill="lightgreen", width=.2)+geom_jitter(colour
='black')
bp_histogram<-ggplot(Pima.data, aes(x=bp, fill=type))+geom_histogram
(binwidth=3, alpha=.5, position="identity")+theme(legend.position = c
(.23,.9))
# multiplot(bp_plot,bp_histogram,cols=2)

# skin
plot_skin<-ggplot(Pima.data,aes(x=skin))+geom_histogram(aes(y=..dens
ity..),binwidth = 2)+geom_vline(aes(xintercept=mean(skin, na.rm=T)),
color="red",linetype="dashed",size=1)+geom_density(alpha=.2, fill="#
FF6666")
# plot_skin
skin_plot<-ggplot(Pima.data, aes(x=type, y=skin))+geom_violin(fill="
lightblue")+geom_boxplot(fill="lightgreen", width=.2)+geom_jitter(co
lour='black')
skin_histogram<-ggplot(Pima.data, aes(x=skin, fill=type))+geom_histo
gram(binwidth=3, alpha=.5, position="identity")+theme(legend.positio
n = c(.75,.9))
# multiplot(skin_plot,skin_histogram,cols=2)

# bmi
plot_bmi<-ggplot(Pima.data,aes(x=bmi))+geom_histogram(aes(y=..densit
y..),binwidth = 2)+geom_vline(aes(xintercept=mean(bmi, na.rm=T)),col
or="red",linetype="dashed",size=1)+geom_density(alpha=.2, fill="#FF6
666")
# plot_bmi
bmi_plot<-ggplot(Pima.data, aes(x=type, y=bmi))+geom_violin(fill="li
ghtblue")+geom_boxplot(fill="lightgreen", width=.2)+geom_jitter(colo
ur='black')
bmi_histogram<-ggplot(Pima.data, aes(x=bmi, fill=type))+geom_histogr
am(binwidth=3, alpha=.5, position="identity")+theme(legend.position
= c(.75,.9))
# multiplot(bmi_plot,bmi_histogram,cols=2)

# ped
plot_ped<-ggplot(Pima.data,aes(x=ped))+geom_histogram(aes(y=..densit
y..),binwidth = 2)+geom_vline(aes(xintercept=mean(ped, na.rm=T)),col
or="red",linetype="dashed",size=1)+geom_density(alpha=.2, fill="#FF6
666")
# plot_ped
ped_plot<-ggplot(Pima.data, aes(x=type, y=ped))+geom_violin(fill="li
```

```
ghtblue")+geom_boxplot(fill="lightgreen", width=.2)
ped_histogram<-ggplot(Pima.data, aes(x=ped, fill=type))+geom_density
(alpha=.2, position="identity")+theme(legend.position = c(.75,.9))
# multiplot(ped_plot,ped_histogram,cols=2)

# age
plot_age<-ggplot(Pima.data,aes(x=age))+geom_histogram(aes(y=..densit
y..),binwidth = 2)+geom_vline(aes(xintercept=mean(age, na.rm=T)),col
or="red",linetype="dashed",size=1)+geom_density(alpha=.2, fill="#FF6
666")
# plot_age
age_plot<-ggplot(Pima.data, aes(x=type, y=age))+geom_violin(fill="li
ghtblue")+geom_boxplot(fill="lightgreen", width=.2)
age_histogram<-ggplot(Pima.data, aes(x=age, fill=type))+geom_density
(alpha=.2, position="identity")+theme(legend.position = c(.75,.9))
# multiplot(age_plot,age_histogram,cols=2)
```

Appendix-2

```
library(LearnBayes)
Pima.data<-read.csv("F:/Bayesian_project/Pima.csv",sep=",",header=TR
UE)
fit1=glm(type~npreg+glu+bp+skin+bmi+ped+age,family=binomial(link=pro
bit),data=Pima.data)
# summary(fit1)
m=10000
X=cbind(1,Pima.data$npreg,Pima.data$glu,Pima.data$bp,Pima.data$skin,
Pima.data$bmi,Pima.data$ped,Pima.data$age)
type<-c(Pima.data$type)

#Bayesian Probit Model with uninformative prior
fit2=bayes.probit(type,X,m)
# apply(fit2$beta,2,mean)
# fit2$log.marg

#Bayesian Probit Model with informative prior
Y<-type
beta0=c(0,0,0,0,0,0,0,0); c0=100
P0=t(X)%*%X/c0
fit3<-bayes.probit(Y,X,m,list(beta=beta0,P=P0))
#fit3$log.marg
```

Appendix-3

```
logit<-function(x){log(x/(1-x))}
expit<-function(x){exp(x)/(1+exp(x))}

## computes the joint disribution
## computes the joint disribution
log_post<-function(Y,X,beta,n){
  #prob<-rep(0,n)
  #like<-rep(0,n)
  X1<-X[,1]
  X2<-X[,2]
  X3<-X[,3]
  X4<-X[,4]
  X5<-X[,5]
  X6<-X[,6]
  X7<-X[,7]
  X8<-X[,8]
  prob1<-expit(beta[1]*X1+beta[2]*X2+beta[3]*X3+beta[4]*X4+beta[5]*X
5+beta[6]*X6+beta[7]*X7+beta[8]*X8)
  like<- sum(dbinom(Y,1,prob1,log=TRUE))
  prior<-sum(dnorm(beta,0,10,log=TRUE))
  val=like+prior
  return(val)
}
Bayes.logistic<-function(Y,X,n.samples,can.sd,n){
  #Initial values:
  beta <-c(-6,0,0,0,0,0,0,0)
  # Keep track of the samples
  keep.beta<- matrix(0,n.samples,8)
  keep.beta[1,] <- beta
  acc <- att <- rep(0,8)
  curlp<-log_post(Y,X,beta,n) # log posterior at current beta
  for(i in 2:n.samples){
    #Update beta using MH sampling:
    for(j in 1:8){
      att[j] <- att[j] + 1
      # Draw candidate:
      canbeta    <- beta
      canbeta[j] <- rnorm(1,beta[j],can.sd)
      canlp      <- log_post(Y,X,canbeta,n)
      # Compute acceptance ratio:
      R <- exp(canlp-curlp)
      U <- runif(1)
      if(U<R){
```

```
        beta<-canbeta
        curlp<-canlp
        acc[j]<-acc[j]+1
      }
    }
    keep.beta[i,]<-beta
  }
  # Return the posterior samples of beta and
  # the Metropolis acceptance rates
  list(beta=keep.beta,acc.rate=acc/att)
}
burn<-200
n.samples<-1000
Pima.data<-read.csv("F:/Bayesian_project/Pima.csv",sep=",",header=TR
UE)
data<-Pima.data[,2:9]
X<-cbind(1,data[,1:7])
Y<-data[,8]
n<-length(Y)
logit_fit<-Bayes.logistic(Y,X,n.samples=n.samples,can.sd=0.03,n)
# apply(logit_fit$beta,2,mean)
# apply(logit_fit$beta,2,sd)
```

Appendix-4

```
library(LearnBayes)
Pima.train<-read.csv("F:/Bayesian_project/pima.train.csv",sep=",",he
ader=TRUE)
Xtest=Pima.train[,2:8]
Ytest=Pima.train[,9]
Xtest<-cbind(1,Xtest)
Pima.data<-read.csv("F:/Bayesian_project/Pima.csv",sep=",",header=TR
UE)
m=10000
X=cbind(1,Pima.data$npreg,Pima.data$glu,Pima.data$bp,Pima.data$skin,
Pima.data$bmi,Pima.data$ped,Pima.data$age)
type<-c(Pima.data$type);Y<-type
fit2=bayes.probit(type,X,m)
# ROC for Bayesian Probit Model with Uninformative prior
pi61<-fit2$beta%*%t(Xtest)
p61<-exp(pi61)/(1+exp(pi61))
ypred.probit61<-colMeans(p61)
table(ypred.probit61,Ytest)
library(pROC)

roc(Ytest,ypred.probit61,plot=T)
# ROC for Bayesian Probit Model with Inforamative prior
beta0=c(0,0,0,0,0,0,0,0); c0=100
P0=t(X)%*%X/c0
fit3<-bayes.probit(Y,X,m,list(beta=beta0,P=P0))
# fit3$log.marg
pi51<-fit3$beta%*%t(Xtest)
p51<-exp(pi51)/(1+exp(pi51))
ypred.probit51<-colMeans(p51)
table(ypred.probit51,Ytest)
library(pROC)
roc(Ytest,ypred.probit51,plot=T)

#ROC for Bayesian Logit Model
pi31<-logit_fit$beta%*%t(Xtest)
p31<-exp(pi31)/(1+exp(pi31))
ypred.logit1<-colMeans(p31)
# table(ypred.logit1,Ytest)
# roc(Ytest,ypred.logit1,plot=T)

# SVC
Pima.data<-read.csv("F:/Bayesian_project/Pima.csv",sep=",",header=TR
UE)
```

```r
X_SVM=Pima.data[,2:8]
Y_SVM=Pima.data[,9]
dat=data.frame(x=X_SVM,y=as.factor(Y_SVM))
library(e1071)
svmfit=svm(y~., data=dat,kernel="linear",cost=10)
# svmfit$index
# summary(svmfit)
set.seed(1)
tune.out<-tune(svm,y~.,data=dat,kernel="linear",ranges=list(cost=c
(0.001,0.01,0.1,1,5,10,100)))
# summary(tune.out)
bestmod=tune.out$best.model
Pima.train<-read.csv("F:/Bayesian_project/pima.train.csv",sep=",",he
ader=TRUE)
Xtest=Pima.train[,2:8]
Ytest=Pima.train[,9]
testdat=data.frame(x=Xtest,y=as.factor(Ytest))
ypred=predict(bestmod,testdat)
# table(predict=ypred,truth=testdat$y)
svmfit.opt<-svm(y~.,data=dat,kernel="linear",cost=0.01,decision.valu
es=T)
fitted<-attributes(predict(svmfit.opt,testdat,decision.values=TRUE))
$decision.values
```

```r
roc(Ytest,fitted,plot=T)
Ytest<-as.numeric(Ytest)
# class(fitted)
fit_pred<-predict(svmfit.opt,newx=testdat[-y,])
```

Appendix_5

```
library(LearnBayes)
Pima.data<-
read.csv("F:/Bayesian_project/Pima.csv",sep=",",header=TRUE)
fit1=glm(type~npreg+glu+bp+skin+bmi+ped+age,family=binomial(link=pro
bit),data=Pima.data)
m=10000
X=cbind(1,Pima.data$npreg,Pima.data$glu,Pima.data$bp,Pima.data$skin,
Pima.data$bmi,Pima.data$ped,Pima.data$age)
type<-c(Pima.data$type);Y<-type

beta0=c(0,0,0,0,0,0,0,0); c0=100
P0=t(X)%*%X/c0
fit3<-bayes.probit(Y,X,m,list(beta=beta0,P=P0))

fit4<-bayes.probit(Y,X[,-c(4,5)],m,list(beta=beta0[-c(4,5)],P=P0[-
c(4,5),-c(4,5)]))

BF<-exp(fit3$log.marg)/exp(fit4$log.marg)
BF1<-exp(fit4$log.marg)/exp(fit3$log.marg)

fit5<-bayes.probit(Y,X[,-4],m,list(beta=beta0[-4],P=P0[-4,-4]))
BF2<-exp(fit5$log.marg)/exp(fit3$log.marg)

fit6<-bayes.probit(Y,X[,-5],m,list(beta=beta0[-5],P=P0[-5,-5]))
BF3<-exp(fit6$log.marg)/exp(fit3$log.marg)
BF4<-exp(fit3$log.marg)/exp(fit5$log.marg)
BF5<-exp(fit3$log.marg)/exp(fit6$log.marg)
BF6<-exp(fit5$log.marg)/exp(fit6$log.marg)
BF7<-exp(fit5$log.marg)/exp(fit4$log.marg)
BF8<-exp(fit6$log.marg)/exp(fit4$log.marg)
BF9<-exp(fit6$log.marg)/exp(fit5$log.marg)
BF10<-exp(fit4$log.marg)/exp(fit5$log.marg)
BF11<-exp(fit4$log.marg)/exp(fit6$log.marg)
```

Appendix 6

```
     library(LearnBayes)
Pima.data<-read.csv("F:/Bayesian_project/Pima.csv",sep=",",header=TR
UE)
m=10000
X=cbind(1,Pima.data$npreg,Pima.data$glu,Pima.data$bp,Pima.data$skin,
Pima.data$bmi,Pima.data$ped,Pima.data$age)
type<-c(Pima.data$type);Y<-Pima.data$type
beta0=c(0,0,0,0,0,0,0,0); c0=100
P0=t(X)%*%X/c0
fit4<-bayes.probit(Y,X[,-c(4,5)],m,list(beta=beta0[-c(4,5)],P=P0[-c
(4,5),-c(4,5)]))
glu<-seq(65,200)
nglu<-length(glu)
X1<-cbind(1,rep(mean(X[,2]),nglu),glu,rep(mean(X[,6]),nglu),rep(mean
(X[,7]),nglu),rep(mean(X[,8]),nglu))
p.glu=bprobit.probs(X1,fit4$beta)
plot(glu,apply(p.glu,2,quantile,0.5),type="l",ylim=c(0,1),xlab="glu
",ylab="probability of diabetic")
lines(glu,apply(p.glu,2,quantile,0.05),lty=2)
lines(glu,apply(p.glu,2,quantile,0.95),lty=2)

age<-seq(21,81)
nage<-length(age)
X2<-cbind(1,rep(mean(X[,2]),nage),rep(mean(X[,3]),nage),rep(mean(X[,
6]),nage),rep(mean(X[,7]),nage),age)
p.age=bprobit.probs(X2,fit4$beta)
plot(age,apply(p.age,2,quantile,0.5),type="l",ylim=c(0,1),xlab="age
",ylab="probability of diabetic")
lines(age,apply(p.age,2,quantile,0.05),lty=2)
lines(age,apply(p.age,2,quantile,0.95),lty=2)

bmi<-seq(19,68)
nbmi<-length(bmi)
X3<-cbind(1,rep(mean(X[,2]),nbmi),rep(mean(X[,3]),nbmi),bmi,rep(mean
(X[,7]),nbmi),rep(mean(X[,8]),nbmi))
p.bmi=bprobit.probs(X3,fit4$beta)
plot(bmi,apply(p.bmi,2,quantile,0.5),type="l",ylim=c(0,1),xlab="bmi
",ylab="probability of diabetic")
lines(bmi,apply(p.bmi,2,quantile,0.05),lty=2)
lines(bmi,apply(p.bmi,2,quantile,0.95),lty=2)
```