Bayesian Modelling by Method of Normal Regression: A Case of Modelling Gluten Content in terms of Protein Content in a Variety of Wheat

Ram Prasad Khatiwada
Central Department of Statistics
Tribhuvan University, Kathmandu, Nepal
Email: ramktd@gmail.com

ABSTRACT

This article is about the Bayesian modelling of the parameters of a simple linear regression with normal errors. It studies the use of non-informative normal priors to the regression parameters. It has an application on modelling gluten content in terms of protein content of a variety of wheat. The exact estimations of credible sets of the regression parameters obtained from real and simulated data by using MCMC. The posterior estimates of the gluten content in terms of protein content are better in this regression model with normal non-informative prior.

Key words: Modelling, Bayesian regression, regression parameters, prior specification, MCMC.

INTRODUCTION

Models are the designed statements for predicting future events, capturing summarized trends and regularities in the observed data. A statistical model is a collection of probabilistic statements that describes and interprets present behaviour or predicts future performance. Statistical models are cheaply used to describe real life problems under uncertainty (Ntzoufras, 2009). It consists of three components: the response variable *Y*, the explanatory variables *X*, and a linking mechanism between the two sets of variables. The response variable *Y* is a stochastic part of the model because the outcome is uncertain before it is observed. In modelling procedure we put our interest to a certain outcomes of *Y* and to predict a future outcome of *Y*. *Y* is a stochastic variable, so

$Y| \ X_1, \ X_2, \ \dots \ \dots \ \dots, \ X_p \sim D(T),$ where $D(\theta)$ is a probability distribution with parameter θ.

The advantage of models is that they impose us to arrange and organize all information available in a logical way, which helps to define precisely the problem under study and facilitates exchange of knowledge. Models may be used for prediction when verified and validated, which may require data from both observation and experiments. To describe significant dependencies among variables, dependency modelling is used whereas, to describe the causal relations between determinant factors and performance measures causation models are used (Fayyad *et al.* 1996).

MODELS AND METHODS

Modelling in Bayesian Paradigm

If the underlying processes are not enough understood, models are designed based only on the observed data. Instead, models are constructed with existing expertise, by beginning with a flexible model specified by a set of parameters, and combined it with the statistical model of the generated data set. The former is the modelling technique in standard classical approach and the latter is the Bayesian modelling approach (O'Hagan, 1995). Bayesian modelling is the method of parametric modelling of data with prior information.

The strength of Bayesian approach is that they can make use of information that might not pertain exactly to the issue at hand. The information can be weighted according to relevance or quality, and sensitivity analysis can be used to assess the priority to be given for collecting more directly relevant data. Bayesian variants of Monte Carlo integration procedures have been devised to address these objections using Gaussian process models (Rasmussen & Ghahramani, 2003).

Let, *Y* be a random variable called response variable, which follows a probabilistic rule with density or probability function $f(y|\theta)$, where θ is the parameter vector. If, the independent and identically distributed sample of size 'n' of variable $y= [y_1, y_2, \dots\dots, y_n]^T$, then the joint distribution

$$f(y|\theta)=\prod_{i=1}^{n} f(y_i|\theta)$$

is called the likelihood of the model and contains the available information provided by the observed sample.

Models are constructed, usually, in order to asses or interpret causal relationship between the response variable Y and various characteristics expressed as a variable $X_j, j \in \upsilon$, called explanatory variables; *j* indicates a model term (or covariate) and υ , the set of all terms under consideration. The explanatory variable is linked with the response variable via a deterministic function and a part of the original parameter vectors is substituted by an alternative set of the parameters (denoted by β) that usually summarizes the effect of each covariate on the response variable.

In a Bayesian model selection, we calculate the posterior distribution over a set of models given a priori knowledge and some new observations (data). The knowledge is represented in the form of a prior over model structures *P(M)*, and their parameters *P(T/M)*, which define the probabilistic dependencies between the variables in the model (Beal, 2003).

By Bayes' rule, the posterior over models *M* observing data y is given by:

$$P(M/y) = \frac{P(M)\ P(y/M)}{P(y)}$$

The term $P(\,y \mid M\,)$ in the numerator is the *marginal likelihood* or *evidence* for a model *M*, which integrates over model parameters, and is the key quantity for Bayesian model selection. Also,

$$P(\,y \mid M\,) = \int P(\theta \mid M\,)\ P(\,y \mid \theta, M\,)\,d\theta$$

For model structure, we can compute the posterior distribution over parameters as

$$P(\theta \mid y, M\,) = \frac{P(\theta \mid M\,)P(\,y \mid \theta, M\,)}{P(\,y \mid M\,)}$$

The predictive density of a new response y′ given the responses *y = {y₁, y₂,……, yₙ}* obtained as

$$P(\,y' \mid y, M\,) = \int P(\theta \mid y, M\,)P(\,y' \mid \theta, y, M\,)d\theta\,,$$

or simply

$$P(\,y' \mid y, M\,) = \int P(\theta \mid y, M\,)P(\,y' \mid \theta, M\,)d\theta$$

If *y′* is conditionally independent of *y*|*T*, we can find posterior distribution of *x′* associated with the new response value *y′* as

$$P(\,x' \mid y', y, M\,) \propto \int P(\theta \mid y, M\,)P(\,x', y' \mid \theta, M\,)d\theta$$

The process of assembling information into a Bayesian model is a multi-stage one, using data and information of many types. It is important to note that, even though these models provide a structure into which the available data can be incorporated and use expert opinion; where there are no data, this does not mean that the models are a substitute for experimental data. The greatest advantage of Bayesian models is that they can be used to facilitate decision analysis despite inadequate data; this is especially important as some types of data are not likely to be readily collected at all (Beal, 2003). Bayesian analysis of the regression was first presented in the landmark paper by Lindley and Smith (1972).

Normal Regression Model

By regression we mean a statistical method used to model the relationship of one or more dependent (or response or outcome) variables to one or more independent (or explanatory) variables. The regression analysis is used with objectives of analyzing correlation, predict the values of dependent variable(s) given independent variable(s), infer cause and effect relationships and estimate systematic relationships and filter out noise.

In normal regression models the response variable *Y* is considered to be a continuous random variable distributed with the normal distribution with the parameters μ (mean) and σ^2 (variance). The normal regression model is summarized as:

$$Y \mid X_1, X_2, ..., X_p \sim N\Big(\mu(\,\beta, X_1, X_2, ..., X_p\,),\ \sigma^2\Big)$$

with, $\mu(\,\beta, X_1, X_2, ..., X_p\,) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$

where, $(\,\beta_0, \beta_1, ..., \beta_p\,)^T$ and σ^2 are the regression parameters.

An alternative formulation of the regression model is that representing response variable directly as a function of the explanatory variable plus a random normal error with mean *0* and variance σ^2.

$$Y = \beta_0 + \beta_1 X_1 + ...+ \beta_p X_p + \varepsilon\ ,$$

where $\varepsilon \sim N(\,0, \sigma_\varepsilon^{\ 2}\,)$

Likelihood Specification in Normal Regression Model

To simplify computational notation, we denote the response variable given explanatory variable

$Y \mid X_1, X_2, ..., X_p$ simply by *Y,*

and the expected value

E(Y|X₁, X₂, ……,Xₚ) by *E(Y)* or μ.

Let, $x_{i1}, x_{i2}, ..., x_{ip}$ be the values of the explanatory variable *X₁, X₂,…..,Xₚ* and with a sample size *n* corresponding to response values $y = (\,y_1, y_2, ..., y_n\,)^T$ for individuals *i=1,2,…,n* ; then the model is expressed as

$$Y_i \sim N\Big(\mu_i, \sigma^2\Big)$$

$\mu_i = \beta_0 + \beta_1 x_{i1} + ...+ \beta_p x_{ip}$ for *i = 1, 2, ……,n.*

Independent Prior Specification:

To make inferences about the regression coefficients, we obviously need to choose a prior distribution for β, σ^2. The basic way of assuming a priori regarding the parameters in the normal regression model is the use of independent distributions.

$$f\left(\beta,\sigma^2\right)=\prod_{j=0}^{p} f\left(\beta_j\right)\cdot f\left(\sigma^2\right)$$

$$\beta_j \sim N\left(\mu_{\beta_j},\varsigma_j^{\,2}\right) \qquad \text{for, } j=0,1,\ldots\ldots,p$$

$$\sigma^2 \sim inv \text{ gamma } (a,\ b)$$

The computational software, WinBUGS, for Bayesian analysis prefer to use precision (τ) instead of variance σ^2. So, the specification is expressed as

$$f(\beta,\tau)=\prod_{j=0}^{p} f\left(\beta_j\right)\cdot f(\tau) \quad \text{and } \tau \sim gamma(a,b)$$

The prior mean and variance of precision parameter τ are

$$E(\tau)=\frac{a}{b} \text{ and } Var(\tau)=\frac{a}{b^2} \text{ respectively.}$$

Conjugate Prior Specification:

The normal distribution is assigned as conjugate prior for the β|σ^2 and an inverse gamma distribution for σ^2 for the normal regression model. The priori for the joint distribution of $[\,\beta,\sigma^2\,]$ follows *normal-inverse gamma* distribution. We symbolize it as

$$\beta \mid \sigma^2 \sim N_p\left(\mu_\beta,\ c^2 V\sigma^2\right) \text{ and } \sigma^2 \sim IG\left(a,b\right)$$

where, $V=(X^T X)^{-1}$ and c^2 is a parameter controlling overall magnitude of the prior variance (Zellner, 1986); the default choice of $c^2 = n$ (Kass & Washerman, 1995).

Posterior Updating

The object of statistical inference is the posterior distribution of the parameters β$_0$,…,β$_k$ and σ^2. By Bayes' Rule, this is simply

$$f(\beta_0,\beta_1,\ldots\ldots,\beta_k,\ \sigma^2 \mid Y,X) \propto f(\ \beta_0,\beta_1,\ldots\ldots,\beta_k,\ \sigma^2\ )\times \prod_i f(y_i \mid \mu_i,\sigma^2\ )$$

In case of simple linear regression

$$f(\beta_0,\beta_1,\sigma^2 \mid Y,X) \propto f(\ \beta_0,\beta_1,\sigma^2\ )\times \prod_i f(y_i \mid \mu_i,\sigma^2\ )$$

Simple Linear Regression with Normal Errors

Simple linear regression is the statistical method used to model the relationship of one dependent (or response or outcome) variable to one independent (or explanatory) variable. In simple linear regression, we assume mean of dependent variable Y is linearly related to independent variable X.

$$E[Y|X]=\beta_0+\beta_1 X$$

$$Y=\beta_0+\beta_1 X+\varepsilon \quad \text{where } \varepsilon \sim N(0,\sigma_\varepsilon^{\,2}) \text{ and } E(\varepsilon)=0$$

For simplicity this model is written as $E[Y|X]=\alpha+\beta X$ or $Y=\alpha+\beta X+\varepsilon$ with common notation $\beta_0=\alpha$ and $\beta_1=\beta$.

For simple linear regression, data consist of a sample of (Y_i,X_i) pairs and we infer whether or not distribution of Y depends on X, estimate coefficients of relationship between *Y* and *X*, find credible intervals for coefficients of the slope and intercept, evaluate how much of the variability in *Y* is explained by *X*, predict not-yet-observed *Y* for a given *X* value and evaluate adequacy of model.

The likelihood function is given by

$$f(\,y \mid \alpha,\eta,\sigma\,)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\,y_i-\alpha-\beta\,x_i\,)^2\right\}$$

$$=\frac{1}{(\,2\pi\,)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\,y_i-\alpha-\beta\,x_i\,)^2\right\}$$

The standard estimates for α, β are

$$b=\frac{S_{xy}}{S_{xx}} \text{ and } a=\bar{y}-b\,\bar{\text{x}} \text{ respectively.}$$

where, $S_{xy}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)$, $S_{xx}=\sum_{i=1}^{n}(\,x_i-\bar{x}\,)^2$, $\bar{y}=\sum_i y_i/n$ and $\bar{x}=\sum_i x_i/n$

The estimates *a* and *b* are often called ordinary least square (OLS) estimates because they minimize the sum of squared deviations from the regression line.

$$(\,a,b\,)=\arg\min\{See\}, \qquad \text{where,}$$

$$See=\sum_{i=1}^{n}(y_i-a-bx_i)^2$$

a and *b* are the maximum likelihood estimates of α andβ, if the error term is normally distributed.

Assuming the prior distribution for (α,β and ϕ) as

$$g(\alpha,\beta,\phi)=\frac{1}{\phi} \text{ , where } \phi=\sigma^2$$

then, *a* and *b* are the posterior expected values of α and β if *n > 2.*

Bayesian Estimation of the Parameters in Simple Linear Regression

Let us re-parameterize $Y=\alpha+\beta X+\varepsilon$ as

$\eta=\alpha+\beta\overline{X}$ where, $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ is the sample mean,

then $E(Y\mid X)=\eta+\beta(X-\overline{X})$

β is called the slope of the regression line (simply, regression coefficient) and η is sometimes called the intercept, although this term is usually used for α.

The likelihood function for the re-parameterized form is

$$f(y\mid x,\eta,\beta,\sigma)=\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\{y_i-\eta-\beta(x_i-\bar{x})\}^2\right]$$

The regression estimates in the re-parameterized model are

$$b=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})/\sum_{i=1}^{n}(x_i-\bar{x})^2 \text{ and } h=\bar{y}$$

Replacing σ^2 by ϕ, the likelihood function is

$$f(y\mid x,\eta,\beta,\phi)=(2\pi\phi)^{-n/2}\exp\left[-\frac{1}{2\phi}\sum_i\{y_i-\eta-\beta(x_i-\bar{x})\}^2\right]$$

By assigning the reference prior, $g(\alpha,\beta,\phi)=1/\phi$, the posterior is obtained as prior times likelihood:

$$Posterior \propto Prior \times Likelihood$$

$$f(\underline{y}\mid\underline{x},\eta,\beta,\phi)\propto\frac{1}{\phi}\times(2\pi\phi)^{-n/2}\exp\left[-\frac{1}{2\phi}\sum_i\{y_i-\eta-\beta(x_i-\bar{x})\}^2\right]$$

$$=(2\pi)^{-n/2}\phi^{-(n+2)/2}\exp\left[-\frac{1}{2\phi}\sum_i\{y_i-\eta-\beta(x_i-\bar{x})\}^2\right]$$

Algebraic manipulation of squared deviations

$$\sum_i\{y_i-\eta-\beta(x_i-\bar{x})\}^2=See+n(\eta-h)^2+Sxx(\beta-b)^2$$

given that,

$$Sxx=\sum_i(x_i-\bar{x})^2 \text{ and } See=\sum_i\{y_i-a-b(x_i-\bar{x})\}^2$$

Thus, the posterior density

$$f(\underline{y}\mid\underline{x},\eta,\beta,\tau)\propto(\phi)^{-(n+2)/2}\exp\left[-\frac{1}{2\phi}\sum_i\left\{See-n(\eta-h)^2+Sxx(\beta-b)^2\right\}\right]$$

The first term in the exponential does not involve η and β, and the second part is proportional to a bivariate normal density.

Conditional on the variance, ϕ, the posterior distribution for η and β is bivariate normal. The mean of η is *h* and its variance *is* ϕ/n and mean of β is *b* and its variance is ϕ/Sxx. Conditional on ϕ, η and β are independent but untransformed intercept α is not independent of β.

On the subject of posterior distribution of ϕ, we find the posterior distribution for ϕ/See, which is

$\phi/See\sim$ inv. chi-square with *n-2* df.

Posterior distribution of precision parameter

$\tau=1/\phi\sim$ Gamma*(c, d)* with $c=(n-2)/2$ and $d=2/See$

Marginal posterior distribution for η and β are given as

$$\eta=\sqrt{n}\frac{(\eta-h)}{See/(n-2)}\sim t_{n-2} \text{ and } \beta=\sqrt{Sxx}\frac{(\beta-b)}{See/(n-2)}\sim t_{n-2}$$

If we marginalize out ϕ, η and β are not independent. The theory behind these distributions can be found in Draper and Smith (1981) and Lee (1997).

APPLICATION OF MODEL

Sample and Data

Independent samples were collected for a variety of wheat to study the relationship between the percentage of protein and gluten content (Khatiwada, 2011). Protein content is vital for baking quality of wheat flour whereas gluten is the main structure for forming protein. Sahin and Sumnu (2006) explain that 'proteins are surface active compounds, comparable with low molecular weight emulsifiers (surfactants), result in lowering of interfacial tension of fluid interfaces, emulsify an oil phase in water and stabilize the emulsion'. It helps in increasing flavor, self- life of the product and helps to make the product soft.

The summary statistics regarding percentage of protein content and gluten content obtained from 20 samples is given in the Table 1.

Table 1. Summary statistics of the percentage of protein and gluten content

Content	Mean	S.D.	SE of mean	Variance	min	max	range	95% CI
Protein (x)	13.32	2.28	0.509	5.188	9.11	16.67	7.56	12.26 – 14.39
Gluten (y)	5.36	0.93	0.207	0.858	3.44	7.21	3.77	4.93 – 5.79

RESULTS

Summaries of the OLS Method

The correlation coefficient between the percentage of protein and the gluten contents (0.935) is significant (p <0.000) with SE=0.337. By ordinary least square (OLS) method the regression coefficients are obtained as α =0.29 and β = 0.38 with SE(α)=0.458 and SE(β)=0.034 respectively. The values of R-square and adjusted R-square are 0.875 and 0.868 respectively. The classical simple regression line of the percentage of gluten content on protein contained by OLS method is given by *Y=0.29+0.38X.*

Bayesian Regression Summaries from Real Data

The following summaries were obtained on applying Bayesian method of regression on the real data of the percentage of protein and gluten content.

Posterior distributions of the parameters (α and β), conditional on ϕ :

- Posterior mean of $\eta = 5.36$, thus, the intercept (α) has a normal distribution with mean 5.36 and variance ϕ/20
- Posterior mean of $\beta = b = Sxy / Sxx = 0.381$, thus, the slope (β) has a normal distribution with mean 0.381 and variance, $\phi / Sxx = \phi / 98.56$.
- The posterior mean of the slope and intercept are the same as that of the (OLS) estimates.

Posterior distributions of the parameters (α and β), for un-conditionality on ϕ:

- Posterior α has a t distribution with 18 df, center 5.36 and standard error $= S / \sqrt{n}$ =0.08
- Posterior β has a t distribution with 18 df , center 0.381 and standard error $= S / \sqrt{Sxx}$ = 0.034
- Posterior precision $\tau = 1/\phi$ has a gamma (*c, d*) distribution, where $c = (n-2)/2 = 9$ and $d = 2/See = 0.98$

The 95% credible interval (*CI*) and 50% highest density region (*HDR*) of the parameters α and β obtained from the real data, updated using a non-informative prior, are given in the Table 2.

Table 2. Posterior summaries of the HDR and CI of the regression parameters

Parameters	Mean	Se	50% HDR (given, $t_{0.75,18}$=0.688)	95% CI (given, $t_{0.95,18}$=2.101)
α	5.360	0.080	5.31—5.41	5.19—5.53
β	0.381	0.034	0.357—0.404	0.31—0.45

The posterior summaries of the variance (ϕ) are obtained from the inverted chi-square with *n-2* degrees of freedom. The 50% HDR for the variance (ϕ)

$$= \left(\frac{See}{\ddot{\chi}^2_{(0.75,18df)}}, \frac{See}{\underline{\chi}^2_{(0.75,18df)}} \right)$$

$$= \left(\frac{2.05}{22.399}, \frac{2.05}{14.218} \right) = (0.092,\ 0.144)$$

The 95% credible set for variance (ϕ)

$$= \left(\frac{See}{\ddot{\chi}^2_{(0.95,18df)}}, \frac{See}{\underline{\chi}^2_{(0.95,18df)}} \right)$$

$$= \left(\frac{2.05}{32.608}, \frac{2.05}{8.548} \right) = (0.063,\ 0.239)$$

Results from the Simulated Data

For the model assessment posterior estimates were generated by using MCMC via Win BUGS. MCMC offers a way of using numerical methods to sum over the uncertainty about the parameters in the model in order to summarize the marginal distributions even in the absence of an accessible analytic solution. The non-informative normal priors were used for the regression parameters and a gamma prior for the precision parameter to update the normal regression model for the protein and gluten content data. 5000 iterations were performed to look at the convergence of the model and the results were taken discarding initial preliminary 500 iterations. The likelihood of the percentage of gluten content in a wheat variety is, then obtained as:

$Yi \sim N(\mu_i, \sigma^2)$, where, $\mu i = \alpha + \beta x_i$ (x_i is the proportion or percentage of protein content)

The non informative priors were taken *as*

$\alpha \sim N(0, 1000)$, $\beta \sim N(0, 1000)$ and $\tau \sim gamma(0.1, 0.1)$

The summary of the posterior densities of the parameters [intercept (α), regression coefficient (β), precision (τ) and the predicted values of the means (μ) of the response variable *Y*], obtained from simulated data, using MCMC via WinBUGS, are given in the Table 3.

Based on the posterior distribution of the values of the parameters, the fitted model is obtained as

$$\hat{Y} = -0.9193 + 0.4358X \ .$$

The density plots and trace plots of alpha and beta are drawn using software WinBUGS and presented. in Figure 2. The density plots (Figure 1) show that the posterior values of the alpha and beta are better fitted to the normal distribution. The trace plots (Figure 2) show that the convergence of the model is satisfactory. Figures 3 depicts the box plots of the average predicted values (μ_i) of response variable y_i and the scatter plot with the fitted line for μ_i is given in the Figure 4.

The Bayesian version of MSE and R-square were obtained 0.3419 and 0.903 respectively for the simulated data. Simulated data has large R^2 value and less mean square errors of the estimate. This model gives the predicted values very close to those values obtained in classical regression, because of the use of a non-informative normal prior with large variance. However, this model has heavy tail distribution with the large values of standard deviations of the estimates.

Table 3. Summary of the posterior density and the predicted values of means

node	mean	se	MC error	median	2.5%	97.5%
alpha	-0.9193	31.77	0.4556	-0.9823	-63.07	60.66
beta	0.4358	31.73	0.5518	0.1735	-60.56	63.80
tau	0.9889	3.097	0.0418	0.0049	2.39E-15	9.57
mu[1]	3.051	291.40	5.0710	1.171	-558.90	587.60
mu[2]	4.554	400.40	6.9690	0.966	-772.80	808.70
mu[3]	3.922	354.60	6.1710	0.316	-681.10	717.00
mu[4]	5.352	458.30	7.9770	1.401	-883.50	925.20
mu[5]	6.010	506.10	8.8090	1.282	-973.40	1021.00
mu[6]	4.127	369.40	6.4300	0.334	-711.70	746.80
mu[7]	5.914	499.20	8.6880	1.580	-959.80	1008.00
mu[8]	5.831	493.20	8.5830	1.516	-948.10	995.50
mu[9]	5.339	457.40	7.9610	1.476	-881.80	923.40
mu[10]	3.500	324.00	5.6370	0.930	-620.70	654.80
mu[11]	6.345	530.50	9.2340	1.071	-1019.00	1069.00
mu[12]	4.036	362.80	6.3140	0.125	-698.00	733.50
mu[13]	4.284	380.80	6.6280	0.818	-734.20	769.50
mu[14]	5.487	468.20	8.1480	1.344	-900.90	944.30
mu[15]	4.249	378.30	6.5840	0.711	-729.30	764.50
mu[16]	4.219	376.10	6.5450	0.576	-724.90	760.00
mu[17]	5.988	504.60	8.7820	1.401	-970.30	1018.00
mu[18]	4.027	362.20	6.3030	0.194	-696.70	732.20
mu[19]	6.154	516.60	8.9910	2.019	-993.60	1042.00
mu[20]	5.326	456.40	7.9440	1.550	-880.10	921.50

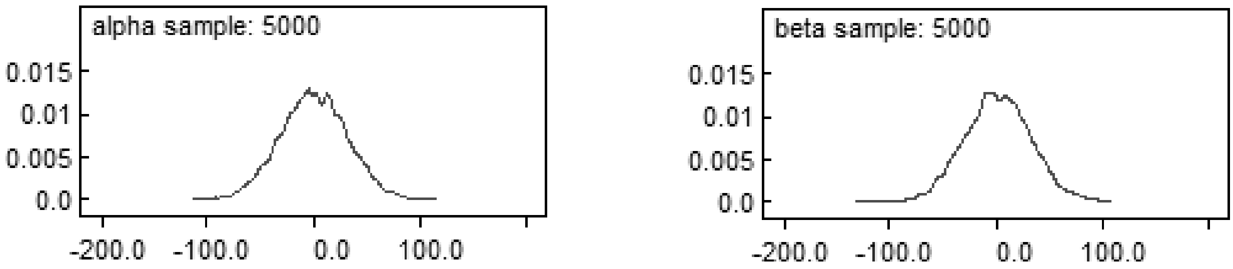

Fig. 1. Posterior density plots of alpha and beta from the simulated data using MCMC

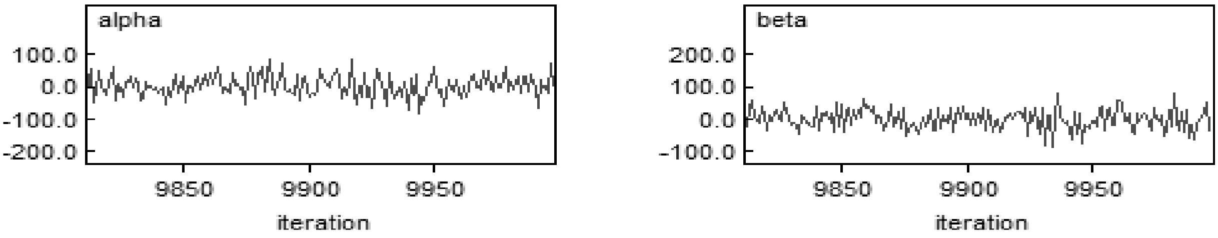

Fig. 2. Trace plots of alpha and beta for last 200 iterations from simulated data using MCMC

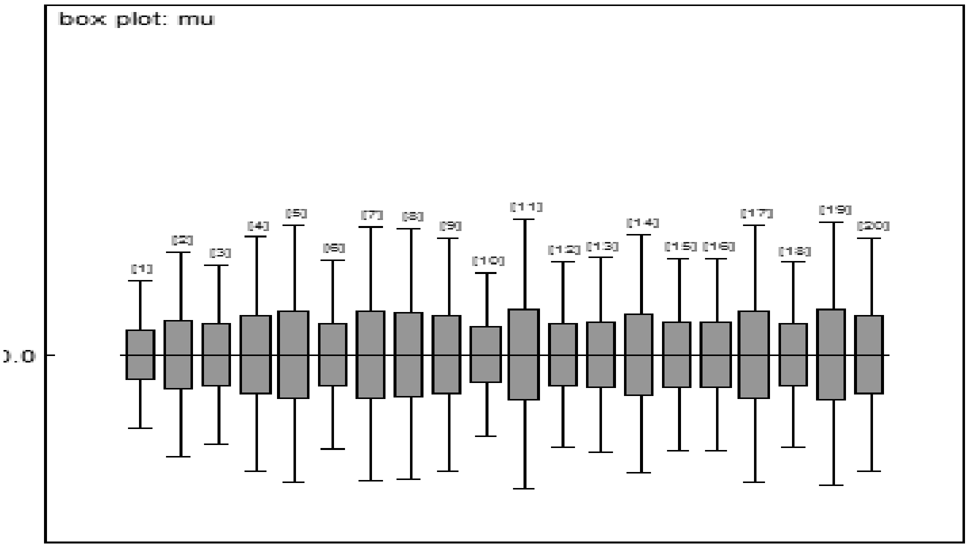

Fig. 3. Box plots of the Predicted mean values from the simulated data

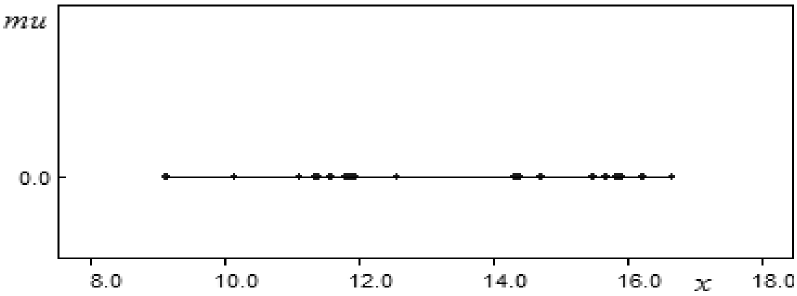

Fig. 4. Scatter plot and model fit of the predicted mean values

CONCLUSION

In this study, the posterior densities for the parameters of the normal regression were obtained under consideration of a non-informative prior distribution. The credible intervals and HDR for the parameters (intercept, regression coefficient and precision) were obtained combining the real data to the specified prior. The point posterior estimates obtained for the intercept and slope were adequately akin to classical estimates. The study was completed using simulations by MCMC, which helped to sum over the uncertainty about the parameters in the model and to generate posterior densities of the parameters of interest. The technique for reaching an eventually distribution fitting of the posterior estimates, known as a convergence test, was used by monitoring one chain for a long time in MCMC iterations. An sensitive and easily implemented diagnostic tool for the model, known as a trace-plot, was used to plot the parameter values at time (t) against the iteration number and studied the shape of the plot and model found to be fitted satisfactorily. Runs of the means were obtained to test whether the posterior distributions of the parameters influenced or not. The kernel density plots were plotted to summarize the posterior distribution of the parameters. From the results of the study, it is found that the modelling of the percentage of gluten content in terms of protein content is better in normal regression model with normal non-informative prior.

REFERENCES

Beal, M. J. 2003. *Variational Algorithm for Approximate Bayesian Inference*. PhD thesis, Neuroscience unit, university of London. www.cse.buffalo.edu

Draper, N.R., and Smith, H. 1981. *Applied Regression Analysis*, (2nd Ed). Wiley, New York.

Fayyad, U. M., Piatesky-Shapiro, G. and Smyth P. 1996. From Data Mining to knowledge Discovery: An Overview. In: *Advances in Knowledge Discovery and Data Mining.* (eds.) Fayyad, U. M., Piatesky-Shapiro. G, Smyth P. and Uthurusamy R. AAAI-Press. Menlo Park, Canada**.** 1-37p.

Kass, R. and Washerman, L. 1995. A reference Bayesian test for nested hypotheses and its relationship to the Shewarz criterion. *Journal of the American Statistical Association* **90**: 928-934.

Khatiwada, R. P. 2011. *Bayesian Modelling Approaches on some Issues of Agro-Food Production and Quality Control*. PhD Thesis, Central Department of Statistics, Institute of Science and Technology, Tribhuvan University, Kathmandu, Nepal.

Lee, P. 1997. *Bayesian Statistics: An Introduction,* (2nd ed). Arnold, London.

Lindley, D. V. and Smith, A. M. 1972. Bayes estimate for the linear model(with Discussion). *Journal of the Royal statistical society,* Series B **34:** 1-41

Ntzoufras, I. 2009. *Bayesian Modelling using WinBUGS.* John Wiley and Sons, New Jersey, Canada.

O'Hagan, A. 1995. Fractional Bayes factors for model Comparison. *Journal of the Royal Statistical Society*, Series B **57**: 57-69.

Rasmussen, C. E. and Ghahramani, Z. 2003. *Bayesian Monte Carlo*. Gatsby Computational Neuroscience Unit, University College London. http://www.gatsby.ucl.ac.uk

Sahin S. and Sumnu, S. G. 2006. *Physical Properties of Foods.* Springer, US.

Zellner, A. 1986. On assessing prior distribution and Bayesian regression Analysis using g-prior distribution, In: *Bayesian inference and Decision Techniques: Essays in Honour of Bruno de Finetti, (*eds.) P Goel and A Zellner*,* North Holand, Amsterdam. 233-243p.