COMPARISION OF DIFFERENT PREDICTION MODELS ON THE YIELD OF MASHBEAN (*Vagna mungo*)

MUHAMMAD F., GHAFOOR A.

Abstract

In classification of traits, it is observed that the plant traits such as pods per plant, pod length and biological yield per plant have positive contribution towards mash grain yield. Three regression procedures i.e. best subset regression, principal component regression and ridge regression were tried as in the first phase and for pod length, approximately same positive quantitative effect was observed towards mash grain yield for all three predictions models. Remaining mash plant traits contribute negatively towards mash grain yield. Also on the basis of criterion of goodness of fit, it was observed that best subset regression model is best and stable as compared to principal component and ridge regression prediction models, both on the basis of original data and simulation procedure.
A simulation procedure adopted to test the reliability of the results by generating random samples from normal (0,1) exponential (1) and uniform (0,1) distributions revealed that estimated effect for pod length for uniform (0,1) distribution tends to very close to original results as compared to other two distributions.

Key words: plant traits, yield, principal component regression, Bayesian regression, simulations.

INTRODUCTION

Achievement of self-sufficiency in agricultural production is an important step towards economic growth of Pakistan. Mashbean (*Vagna mungo*), is a rich source of quality plant proteins and an important crop of Pakistan.

Donachie and Haswell (1999), Wentzell and Andrews (1999) described multivariate calibration for the determination of trace metals in water matrices. Multivariate calibration models were constructed using multiple linear regression (MLR), principal components regression (PCR) and partial least squares (PLS) methods to identify which technique offers the better predictions for unknown sample. Finally, the calibration model constructed using PLS was found to provide the most accurate predictions for the unknown samples. Guler et al (2001) determined relationships between yield and yield components.

This study is initiated to provide comparison of different prediction models such as best subset regression, principal component regression and ridge regression based on original data and simulation data. Also to observe the effect of reference prior (based on experience of mash experts) on the estimated effects and its standard error of fixed trait on mash grain yield.

MATERIALS AND METHODS

Mash data were obtained form Plant Genetic Research Institute (PGRI) at National Agricultural Research Center (NARC) Islamabad. Data consisted of 37 pure lines of mash arranged in randomized complete block design (RCBD) with three replications. Eleven different traits including grain yield (Y) and morphological traits such as plant height (X_1), days to flowering (X_2), days to fist pod maturity (X_3), days to 90% maturity (X_4), branches per plant (X_5), pods per plant (X_6), pods length (X_7), seeds per pods (X_8), 100 - seed weight (X_9) and biological yield per plant (X_{10}), were measured in the data set.

The breeders are interested to obtain different prediction models for predicting mash grain yield (Y) on the basis of different fixed mash plant traits (X's) and the statisticians help in recommending models with respect to stability by simulation procedure. The three different regression procedures such as best subset, principal component and ridge regression are discussed as under.

Best Subset Regression

Whenever the pool of potential X variables is not very small, it is highly desirable that the investigator be able to concentrate on the limited number of regression models which are the "best" one according to a specified criterion. The limited might consist of the "best K=5 or 10 (say)" subsets according to the criterion employed so that the investigator can carefully choose the final model.

There now exist excellent computer algorithms for selecting best subsets of predictor variables in regression. A popular one is that given by Furnival and Wilson (1974), which computes only a fraction of possible regressions in determining the "best k" subsets. Three criteria such as R^2_p, MSE_p or R^2_{adj} and Mallows C_p may be applied for determining these "best K" subsets.

In addition to coefficient of multiple determination R^2_p and MSE_p or R^2_{adj}, Mallows Cp statistics has gained

popularity in recent years, Mallows (1973). This has the form:

$$C_P = \frac{SSE_P}{S^2} - (n - 2P) \qquad (2.1)$$

Where SSE_p is the Error sum of squares from the model containing p parameters (p is the number of parameter in the model including β_0). The "Best" is chosen after inspecting the C_p plot. We would look for a regression with a low C_p value about equal to P. When the choice is not clear-cut, it is a matter of personal judgment.

Principal Component Regression
In the development of model, the one of the difficulties that we sometimes encounter is the presence of highly inter-correlated predictor variables. A procedure, which analyzes the correlation structure in detail is principal component regression and is being extensively used John *et al.* (1987).
In the context of selection of best prediction model, we use the notation "centered" and scaled X matrix which is called "Z". Then applying principal component analysis on correlation matrix Z^TZ, the jth principal component is as:

$$W_j = \gamma_{1j}Z_1 + \gamma_{2j}Z_2 + \gamma_{3j}Z_3 + \cdots + \gamma_{(p-1)j}Z_{p-1} \qquad (2.2)$$

Thus, the procedure creates a set of artificial variables; W_j's via a linear transformation, as in (2.2) in such a way that the W vectors are orthogonal to each other. Finally, a prediction model for Y as a function of selected W_j's can be obtained by regressing dependent variable Y on some or all of the principal components scores as:

$$Y = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \ldots + \beta_{10} W_{10} \qquad (2.3)$$

Once the fitted equation is obtained in terms of the selected W_j's, it can be transformed back into a function of the original predictor variables.

Ridge Regression
Ordinary least squares (OLS) provides less precise estimates of the regression coefficients when applied to the non-orthogonal data. Ridge regression is one of the several methods that has been proposed by Hoerl *et al.* (1975) to remedy problem of non-orthogonality (multicolinearity), by modifying the method of least square to allow biased estimators of the regression coefficient but more precise (less standard error of the estimates) than unbiased estimates. The ridge regression coefficients b^R can be estimated as:

$$(r_{XX} + C * I) * b^R = r_{YX} \qquad (2.4)$$

Where r_{XX} is the (p-1) x (p-1) matrix of pair wise correlation between predictors, r_{YX} is a (p-1) x 1 vector of correlation between dependent and each predictor

and constant C reflects the amount of bias in the estimators usually varies from 0 and 1.

The Bayesian approach
To observe the effect of reference prior (non-informative prior) on the precision of the estimated effect of fixed trait on response variable, Bayesian regression approach with both sample and reference prior is used. The simple linear regression model of response (mash grain yield Y) on any more effective fixed trait (X_j) is given as:

$$Y_i = \beta_0 + \beta_j X_j + \varepsilon_i$$

Before considering reference prior i.e. $c < \beta_j < d$, it is useful to obtain the posterior distribution for β_j with complete uncertainty about reference prior under the assumption of known error variance σ^2 and it is obvious that:

$$b_j \sim N\{\beta_j, \sigma^2 \frac{1}{\Sigma(X_j - \bar{X})^2}\} \qquad (2.5)$$

Now instead of treating b_j as a random variable and β_j as fixed, we treat b_j as fixed and β_j as random variable. So the probability distribution for β_j that will express uncertainty about β_j after the sample has been observed is given as:

$$f(\beta_j \mid y) = [\frac{\Sigma(X_j - \bar{X})^2}{2\pi\sigma^2}]^{\frac{1}{2}} \exp\{-\frac{\Sigma(X_j - \bar{X})^2}{2\sigma^2}(\beta_j - b_j)^2\} \qquad (2.6)$$

So, $f(\beta_j|y)$ is an expression of uncertainty about β_j after the sample information has been observed and range of density $f(\beta_j|y)$ will remain same as $f(\beta_j)$.

Including reference prior about regression coefficient
The reference prior inequality about regression coefficient is of the form $c \leq \beta_j \leq d$, as used by Geweke (1986), Griffiths (1988) and Griffiths *et al.* (1988), where "c" and "d" are the limits that are specified by the expert prior to sample. As it is only an idea but we are uncertain where within interval (c, d), β_j might lie. Then in such case a probability density function that suggests that all values between c and d are equally likely is the uniform probability distribution given as:

$$f(\beta_j) = \begin{bmatrix} \dfrac{1}{d-c} & c \leq \beta_j \leq d \\ 0 & otherwise \end{bmatrix} \qquad (2.7)$$

The next question is how prior density function like (2.7) changes the posterior distribution for β_j. As our prior density function $f(\beta_j)$ attaches zero probability to the value of β_j outside the range (c, d), so the posterior distribution includes this information and the additional

information provided by the sample, must also attach zero probability to the value outside the range (c, d).

Point estimation of β_j

$$E\,[\,L\,(\,\beta_{\,j}\,,\,b_{\,j}\,)\,]\;=\;E\,[\,c\,(\,\beta_{\,j}\,-\,b_{\,j}\,)^{\,2}\,]\;=\;\int c\,(\,\beta_{\,j}\,-\,b_{\,j}\,)^{\,2}\,f\,(\,\beta_{\,j}\,\mid\,y\,)\,d\beta_{\,j}$$

Here β_j is as random variable and $f\,(\beta_j\mid y)$ as its probability density function. When the loss function is quadratic, the point estimate for an unknown parameter would be the mean of the posterior distribution, which also minimizes the expected loss.

$$E_{\,N}\;(\,\beta_{\,j}\,)\;=\;\int_{c}^{d}\beta_{\,j}\,f\,(\,\beta_{\,j}\,\mid\,y\,)\,d\beta_{\,j}$$

Simulation procedure

In any kind of research work, it is good practice to test the reliability of recommended results by simulation procedure. Here simulation of prediction model and Bayesian approach with reference prior is made by Minitab statistical package. To obtain random samples of response variable, the steps of simulation procedure are

- In the first step we generate samples of random vectors of residuals each consists of 37 observations from normal distribution with mean zero and unit variance.

To choose a single point estimate for β_j we have to minimize the expected symmetric quadratic loss function.

- With the assumption of known variance σ^2, the random residuals are converted so that $\varepsilon_i \sim N\,(0, \sigma^2)$.
- The random mash grain yield can be obtained by adding fitted mash yield on original data and random residuals with zero mean and σ^2 variance, obtained in preceding step.

All the methods discussed under section 2 are applied on the randomly generated data and results are discussed below.

RESULTS AND DISCUSSION

Best subset, ridge and principal component regression models

Three prediction model, based on best subset, principal component and ridge regression, are developed for predicting mash grain yield on the basis of different physical and yield related traits which are as under:

Best subset regression based on three different criteria such as C_P Mallow's statistics, R^2_P and MSE_P is given as

$$\text{Grain yield} = -9.63 - 0.0436\ X_1 - 0.0345\ X_4 - 0.0858\ X_5 + 0.118 X_6$$
$$+ 3.27\ X_7 + 0.122\ X_{10} \qquad\qquad (3.1)$$

With coefficient of multiple determination, $R^2 = 94.10\%$ and Adjusted coefficient of multiple determination $R^2_{adj} = 92.90\%$. Also the unexplained variation becomes 35.285.

Principal component regression model based on first 5 principal components contributing 87% of total variation present in the data is given as:

$$\text{Grain yield} = -12.9753 - 0.0215\ X_1 - 0.0216\ X_3 - 0.0252\ X_4 + 0.1746\ X_5 + 0.0504\ X_6$$
$$+ \ .... + 1.6216\ X_7 + 1.4418\ X_8 + 0.2230 X_9 + 0.1048\ X_{10} \qquad (3.2)$$

Which is the final Principal component regression model for predicting mash grain yield, free from multicollinearity effect and any extra ordinary small estimate, with coefficient of multiple determination $R^2 = 86.84\%$ and adjusted coefficient of multiple determination $R^2_{adj} = 81.79\%$.

Ridge regression model observed from ridge trace and method due to Hoerl *et al.* (1975) with biasing constant C=0.024 is as:

$$\text{Grain yield} = -12.01 - 0.0424\ X_1 + 0.0548\ X_2 - 0.0614\ X_4 - 0.0634\ X_5 + 0.1114\ X_6$$
$$+ \ ..... + 3.2392\ X_7 + 0.1217 X_8 + 0.3337 X_9 + 0.1170\ X_{10} \qquad (3.3)$$

Which is the final Ridge regression model for predicting mash grain yield, free from multicollinearity effect and any extra ordinary small estimate, with coefficient of multiple determination R^2 = 94.38% and adjusted coefficient of multiple determination R^2_{adj} = 92.22%. It is clear from above measures of goodness of fit that there is a close agreement between statistics of goodness of fit for ridge regression and subset regression and better than statistics obtained from principal component regression. But we prefer the best subset regression model for predicting mash grain yield because of having fewer numbers of regressors. The stability of prediction model with respect to future prediction was also tested and best subset regression model was found most stable and better for future prediction. In this study it was also observed that the outlying observation w.r.t. response must be used with great care because the stability of the model w.r.t. future prediction is seriously damaged by outlying observation. It is clear from the above model all the plant traits except pods/plant (X_6), pod length (X_7) and biological yield/plant (X_{10}) have negative contribution towards mash yield. Positive and approximately same quantitative contribution of pod length (X_7) is observed towards mash yield for all above three prediction models. So we can safely recommend the mash breeders that any mash genotype having more pod length as compared to the others genotypes contributes more toward mash yield. Above three prediction models not only provide the breeders a way for predicting yield of any variety on the basis of physical and yield traits of mash plant also provide an idea about the traits which are contributing positively or negatively towards mash yield.

Simulation results
It is very important to test the stability and reliability of the previously developed and recommended models (given in section 3.1) by simulation. About 3000 random samples of mash grain yield each consists of 37 observations (genotypes) were generated from normal (0, 1) and exponential (1) distributions and then transformed into residuals with zero mean with 1.20 known variance. The fitted values for each model, obtained by substituting original fixed traits, plus randomly generated residuals provided us random mash grain yield.
To compare the stability and precision of regression estimates obtained from three-prediction models, about 3000 times' regression procedures were performed and vectors of estimates were stored. All the procedure including random sample generation, performing regression analysis and storing their estimates were done through a program written in Minitab 11 (statistical package). Estimated parameter by three different regression methods on original data and

randomly generated data by both normal (0, 1) and exponential (1) distribution are presented in Table-3.1. The estimated coefficients and their standard errors under randomly generated normal and exponential data were obtained by averaging all vectors of estimates and square root of diagonal elements of variance-covariance matrix of values for each estimate respectively. In Table-3.1, BSR, RR and PCR stand for best subset regression, ridge regression and principal component regression respectively. In the Table-3.1 it is quite clear that the estimated parameters by best subset regression (BSR), ridge regression (RR) and principal component regression (PCR) from original data are very close to the estimates obtained from randomly generated data from normal and exponential distribution. However the standard errors of the estimates from best subset regression (BSR) are less as compared to the ridge regression and principal component regression in both data sets under normal (0,1) and exponential (1) distributions. It means that the estimates obtained from best subset regression (BSR) are more precise and stable than estimates from ridge regression (RR) and principal component regression (PCR). So we conclude that best subset regression model is more stable and reliable for predicting mash grain yield. Our recommendation about best subset regression in simulation is same as recommendation made previously on the original data.

Bayesian approach
As we observed that only the mash plant trait pod length (X_7) contributed more effectively and positively towards mash grain yield (Y). Now using Bayesian regression with reference prior (prior non-sample information). Before starting Bayesian approach it is important to discuss the following two features.

1. Reference prior (non-informative prior) information about regression parameter i.e. $0 < \beta_7 < 4.5.$ is based on previous data and suggested by mash breeders.

2. We assumed that σ^2 is known, and value of σ^2 =1.20, is obtained by pooling mean square error from previous data.

The residuals were tested and found normally distributed as done in previous section and all other regression assumptions were also tested and found desirable.
From equation 2.5, after treating b_7 as fixed and β_7 as random it can be derived that,

$$\beta_7 \sim N(3.17, \quad 1.120)$$

So the posterior distribution $f(\beta_7|y)$ for β_7 with no reference prior, about regression coefficient β_7, included after the sample has been observed according to the equation (2.6) is given as:

$$f(\beta_7 \mid y) = \frac{1}{\sqrt{2(1.120)\pi}} \exp\left\{-\frac{1}{2(1.120)}(\beta_7 - 3.17)^2\right\}$$

Including reference prior about regression coefficient β_7

The reference prior inequality $0 \le \beta_7 \le 4.5$ can be expressed in term of prior uniform density function as in equation (2.7) is given as:

$$f(\beta_7) = \begin{bmatrix} \dfrac{1}{4.5} & 0 \le \beta_7 \le 4.5 \\ 0 & elsewhere \end{bmatrix}$$

Then the posterior distribution that includes this information and the information provided by the sample must also attach zero probability to the value of β_7 outside the range (0, 4.5) and is given as:

$$f_N(\beta_7 \mid y) = \frac{1}{\sqrt{2(1.120)\pi}} \exp\left\{-\frac{1}{2(1.120)}(\beta_7 - 3.17)^2\right\}$$

$$0 \le \beta_7 \le 4.5 \qquad (3.1)$$

Here "N" is refers for normal distribution. But it was also observed that probability of β_7 lying outside the range (0, 4.5) also exists which can be calculated as:

$$P(\beta_7 > 4.5) = P(\frac{\beta_7 - b_7}{\sigma_{\beta_7}} = \frac{4.5 - 3.17}{1.058}) = 0.1056$$

This cannot be ignored and needed to truncate the posterior distribution given in equation (3.1) with reference prior information included about β7. Here truncation means shifting the probability (area) greater than "4.5" proportionally over the remainder of the density function, then the resulting distribution is called truncated posterior distribution which is given as:

$$f_{TN}(\beta_7 \mid y) = \frac{(0.8954)^{-1}}{\sqrt{2(1.120)\pi}} \exp\left\{-\frac{1}{2(1.120)}(\beta_7 - 3.17)^2\right\} \qquad 0 \le \beta_7 \le 4.5$$

Here "TN" refers to truncated normal distribution. Both normal and truncated normal posterior distribution is shown in figure 3.1.
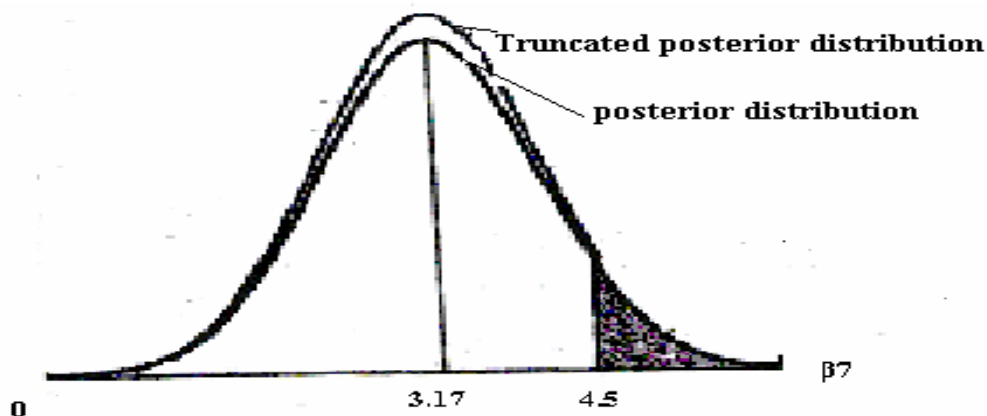
Fig. 3.1   posterior and truncated posterior distributions

As indicated in section-2 that the mean of the truncated posterior distribution is the parameter estimate that also minimizes expected loss, which is given as:

$$E_{TN}(\beta_7) = \int_0^{4.5} \beta_7 f_{TN}(\beta_7 \mid y) d\beta_7$$

It is very difficult to solve this integral, so we generate 5,000 observations from posterior distribution $f_N(\beta_7 | y)$, by using $\beta_7 \sim N(3.17, 1.120)$. Observations greater than 4.5 were discarded to obtain a random sample from truncated posterior distribution $f_{TN}(\beta_7|y)$. Now to obtain point estimate of β_1, we simply take the mean of retained observations from the sample of 5,000 observations, which becomes i.e. $\hat{E}_{TN}[\beta_7] = 2.9505$. This estimate of β_7 is lower than the estimate 3.17 obtained from approach that does not take into account the prior information. Also the sample variance from the retained observations on β_7 which comes out to be $\hat{Var}_{TN}[\beta_7] = 0.8973$ is less than 1.069 obtained from regression approach that does not take into account the prior information. So the introduction of reference prior reflects the reduction in dispersion of estimate. In other words we can say that precision of the regression estimates increases with the introduction of reference prior.

Simulation results

The results obtained using original data are verified by generating random samples 5000 each from exponential (1) and uniform (0, 1) distribution by using procedure given in section 2.5. In Table 3.2 it is quite clear that estimated effect of pod length from randomly generated data from Uniform (0, 1) distribution tends to be very close to the original estimated effect as compared to the normal (0, 1) and exponential (1). Also standard error of the estimate of effect and truncated probability for random sample from uniform (0, 1) decreased as compared to the other two distributions.

REFERENCES

DONACHIE A. AND HASWELL SJ. (1999) : Application and comparisons of chemometric techniques for calibration modeling using electrochemical/ICP-MS data for trace elements in UHQ water and humic acid matrices. *Analytica Chimica Acta*. 378(1-3), 235-243.

FURNIVAL G.M. AND WILSON R.W. (1974) : Regression by leaps and bounds. *Technometrics*. 16, 499-511.

GEWEKE J. (1986): Exact inference in the inequality constrained normal linear regression model. *Journal of Applied Econometrics*, 1, 27-141.

GRIFFITH E. (1988) : Bayesian econometrics and how to get rid of those wrong signs. *Review of Marketing and Agricultural Economic*. 56, 36-56.

GRIFFITH W. E., HILL R. C. AND JUDGE G. G. (1988) : Learning and practicing Econometrics. *2nd Edition, John Wiely and Sons, Inc (New York)*.

GULER, M., ADAK, MS, AND ULUKAN, H. (2001). Determining relationships among yield and some yield components using path coefficient analysis in chickpea. European Journal of Agronomy .14(2), 161-166.

HOREL A.E., KENNARED R.W. AND BALDWIN K.F. (1975) : Ridge regression: some simulations. *Communication in statistics*. 4, 105-123.

JOHN N., WILLIAM W. AND MICHAL H. K. (1987) : Applied linear statistical model. *2nd Edition Toppan Comp*. Tokyo (Japan).

MALLOWS C. L. (1973) : Some comments on C_p. *Technometrics*. 15 661-675.

WENTZELL P. D. AND ANDREWS D. T. (1999) : Estimation of hydrocarbon types in light gas oils and diesel fuels by ultraviolet absorption spectroscopy and multivariate calibration. *Canadian Journal of Chemistry*. 77(3), 391-400.

Corresponding author:

Prof. Muhammad F.
Department of Mathematics and Statistics, Allama Iqbal Open University, Islamabad, Pakistan

Tab. 3.1 :   **Comparison of different models on randomly generated data**

Estmd. Coeffs.	No of Obsn	Results of Original Data			Results of Randomly Generated data from Normal Distribution (0, 1)						Results of Randomly Generated data from Exponential Distribution (1)					
		B.S.R.	R. R.	P.C.R.	S.S.R	St.Dev. (BSR)	R.R.	St.Dev. (RR)	P.C.R.	St.Dev. (PCR)	B.S.R.	St.Dev. (BSR)	R. R.	St.Dev. (RR)	P.C.R.	St.Dev. (PCR)
b0	1000	-9.63	-11.94	-8.89	-9.47	3.60	-11.75	4.75	-12.90	4.83	-9.78	3.54	-11.93	4.94	-12.77	4.68
b1	1000	-0.04	-0.04	-0.06	-0.04	0.02	-0.04	0.02	-0.02	0.02	-0.04	0.02	-0.04	0.02	-0.02	0.02
b2	1000	*	0.05	0.00	*	*	0.05	0.07	*	*	*	*	0.06	0.08	*	*
b3	1000	*	0.00	-0.02	*	*	*	*	-0.02	0.06	*	*	*	*	-0.02	0.06
b4	1000	-0.03	-0.06	-0.01	-0.03	0.02	-0.06	0.04	-0.03	0.06	-0.03	0.02	-0.06	0.04	-0.03	0.06
b5	1000	-0.09	-0.07	-0.08	-0.09	0.06	-0.06	0.06	0.18	0.06	-0.09	0.06	-0.07	0.06	0.17	0.06
b6	1000	0.12	0.11	0.09	0.12	0.02	0.11	0.03	0.05	0.03	0.12	0.02	0.11	0.03	0.05	0.03
b7	1000	3.27	3.22	3.41	3.22	0.91	3.23	1.11	1.66	1.13	3.31	0.90	3.23	1.12	1.61	1.08
b8	1000	*	0.12	-0.17	*	*	0.10	0.63	1.42	0.64	*	*	0.12	0.64	1.45	0.62
b9	1000	*	0.34	-0.06	*	*	0.32	0.58	0.21	0.59	*	*	0.33	0.57	0.19	0.59
b10	1000	0.12	0.12	0.17	0.12	0.05	0.12	0.05	0.10	0.06	0.12	0.05	0.12	0.06	0.11	0.06

Tab. 3.2 : Summary of results from randomly generated data

Statistics	Normal distribution	Exponential(1) distribution	Uniform (0, 1) distribution
Estimate of β_1 and Standard error including prior information.	2.9505 (0.89737)	3.0516 (0.8289)	3.13 (0.3079)
Total generated observations	5,000	5,000	5,000
Observation from Truncated density	4462	4836	4960
Discarded Observations	538	164	40
Estimated Probability P(β_1> 4.5)	0.1076	0.0328	0.008

It is also clear from Table 3.2 that standard error of estimate with exponential (1) distribution and normal (0, 1) distribution data is smaller than Uniform (0, 1) and original data