Journal of Applied Sciences 5 (1): 187-191, 2005
ISSN 1607-8926

Bayesian Regression with Prior Non-sample Information on Mash Yield

[1]A. Ghafoor, [2]F. Muhammad and [2]I.A. Arshad
[1]Department of Statistics, Government College Farooka District, Sargodha, Pakistan
[2]Department of Mathematics and Statistics, Allama Iqbal Open University, Islamabad, Pakistan

Abstract: To increase the precision of estimated effect of a yield character "pod length" on mashbean grains yield, Bayesian regression technique with sample and prior non-sample information about pod length was applied on simple linear relation between mash grain yield and pod length. With the use of prior inequality information about regression coefficient on pod length, a reduction was observed in the estimated value of regression coefficient and its standard error. It was observed that prior inequality information about regression parameter is helpful to increase the precision of the regression estimates. Simulation procedure was developed to generate random residuals from Exponential (1) and Uniform (0, 1) distributions, to test the results. The results were compared with those based on original data set.

Key words: Bayesian, simulation, estimation, sampling, density function

INTRODUCTION

The effect of a fixed variable on response variable can be determined with the help of ordinary regression analysis. Sometimes, by introducing prior non-sample information about regression coefficient in regression analysis is very helpful to increase the precision of the estimated coefficients. In present study, prior inequality non-sample information about regression coefficient in simple linear relation between mash pod length (fixed variable) and mash grain yield (response variable), were introduced and checked the precision of estimated coefficient.

Faqir *et al.*[1] developed different prediction models for predicting mash grain yield and separated the traits that contributed positively and negatively towards the mash grain yield. It was observed that the effect of pod length is positively significant towards mash grain yield among other mash plant traits, such as plant height, days to flowering, days to first pod maturity, days to 90% maturity, branches per plant, pods per plant, seeds per pod, 100-seed weight, biological yield per plant. Geweke[2] and Griffith[3] discussed bayesian regression approach on consumption expenditure and income with both no prior information and prior non-sample inequality information about regression parameter with the assumption of known standard deviation and checked the influence of prior information on estimate and its standard error.

MATERIALS AND METHODS

Mash data was obtained form plant genetic resource institute at national agricultural research center Islamabad. The experimental material lasted for two years consisted of 37 mash genotypes arranged in Randomized Complete Block Design (RCBD) with three replications. Mash plant traits, such as plant height (X_1), days to flowering (X_2), days to first pod maturity (X_3), days to 90% maturity (X_4), branches per plant (X_5), pods per plant (X_6), pod length (X_7), seeds per pod (X_8), 100-seed weight (X_9), biological yield per plant (X_{10}) and grain yield per plant (Y) were measured. The study was initiated to increase the precision of the estimated effect of pod length on mash grain yield with the help of Bayesian regression technique including prior non-sample information. The simple linear regression model of mash grain yield (Y) on pod length (X_7) is given as:

$$y_i = \beta_0 + \beta_1 X_7 + \varepsilon_i$$

Before considering the prior inequality information about regression coefficient i.e. $c<\beta_1<d$,

$$b_1 \sim N\left\{ \beta_1, \sigma^2 \frac{1}{\sum (x_1 - \bar{x})^2} \right\} \quad (1)$$

It is useful to obtain the post-sample density function for β_1 with complete prior uncertainty under the assumption of known error variance σ^2. Also before a sample is taken,

Corresponding Author: Dr. I.A. Arshad, Department of Mathematics and Statistics, Allama Iqbal Open University, Islamabad, Pakistan E-mail: irshad_arshad@hotmail.com

the estimator of regression coefficient "b" has the probability density function:

To find the probability density for β_1, instead of treating b_1 as a random variable and β_1 as fixed, we treated b_1 as fixed and β_1 as random variable. The expression of uncertainty about β_1 after b_1 has been observed, can be obtained from equation (1) as:

$$\beta_1 = b_1 - Z * \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$

Where, b_1 and multiple of Z are constant and $Z \sim N(0, 1)$, so β_1 is normally distributed with mean and variance as given below:

$$\beta_1 \sim N\left\{ b_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \right\} \quad (2)$$

Also the probability distribution for β_1 that will express uncertainty about β_1 after the sample has been observed is:

$$f(\beta_1|y) = \left[\frac{\sum (X_1 - \bar{X})^2}{2\pi\sigma^2} \right]^{\frac{1}{2}} \exp\left\{ \frac{\sum (X_i - \bar{X})^2}{2\sigma^2} (\beta_1 - b_1)^2 \right\} \quad (3)$$

The notation $f(\beta_1| y)$ is used rather than just $f(\beta_1)$ to denote that y is given. So, $f(\beta_1| y)$ is an expression of uncertainty about β_1 after the sample information has been observed and range of density $f(\beta_1| y)$ will remain same as $f(\beta_1)$.

Including prior information about regression coefficient: The prior inequality information on regression coefficient is of the form $c \le \beta_1 \le d$, Where, "c" and "d" are the limits that are specified by the expert prior to sample. This prior inequality information can be expressed in term of prior density function. As it is only an idea from prior inequality information that β_1 lies within "c" and "d", but we have no idea where within interval (c, d), β_1 might lie. Then in such case a probability density function that suggests that all values between c and d are equally likely is only the uniform probability distribution, given as:

$$f(\beta_1) = \begin{bmatrix} 1 & c \le \beta_1 \le d \\ 0 & otherwise \end{bmatrix} \quad (4)$$

The next question is how prior density function like (4) changes the post sample density function for β_1. As our prior density function $f(\beta_1)$ attaches zero probability to the value of β_1 outside the range (c, d). So the post sample density function includes this information and the additional information provided by the sample, must also attach zero probability to the value outside the range (c, d), which is given same as in equation (3).

Point estimation of β_1: To choose a single point estimate for β_1 so as to minimize the losses that occur from over and underestimation, the function that describes the losses happen in regression coefficient β_1 is called loss function and is denoted by $L(\beta_1, b_1)$. In other words we can say that loss function $L(\beta_1, b_1)$ is the function of error $(\beta_1 - b_1)$, which may be positive or negative. Symmetric quadratic loss function is considered here for point estimation.

$$L(b, \beta_1) = c(\beta_1 - b_1)^2$$

Where, c is a constant. This function is symmetrical in the sense that losses from overestimation are identical to the losses from underestimation and is quadratic because quadratic in estimation error $|\beta_1 - b_1|$. The point estimate for β_1 can be obtained by minimizing the average loss function. The expression of average loss is given as:

$$E\left[L(\beta_1, b_1)\right] = E\left[c(\beta_1 - b_1)^2\right] = \int c(\beta_1 - b_1)^2 f(\beta_1|y) d\beta_1$$

Here β_1 is as random variable and $f(\beta| y)$ as its probability density function. When the loss function is quadratic, the point estimate for an unknown parameter would be the mean of the post sample density function, which minimizes the expected loss.

$$E_{1N}(\beta_1) = \int_c^d \beta_1 f(\beta_1|y) d\beta_1$$

Which is weighted average of all values of β_1 with probabilities as weights.

Simulation procedure: To test the reliability of recommended results, following simulation procedure was adopted.

- 500 samples of random vectors of residuals each consists of 37 observations from normal distribution with mean zero and unit variance were generated.
- With the assumption of known variance σ^2. The random residuals are converted so that we obtain $\varepsilon_i \sim N(0, \sigma^2)$.
- The random mash grain yield was obtained by adding fitted mash yield based on original data and random residuals with zero mean and σ^2 variance.

RESULTS AND DISCUSSION

Faqir *et al.*[1] estimated regression coefficients of the relationship of fixed mash plant traits (X's) on mash grain

yield (Y) so that best prediction model for predicting mash yield can be developed. It was observed that only the trait pod length (X_7) contributed positively and more effectively towards the mash grain yield (Y). Now we will discuss how the estimated regression coefficient of pod length and its standard error changes with the introduction of prior non-sample inequality information about regression coefficient of pod length by Bayesian approach.

Prior non-sample inequality information about regression parameter i.e. $0<\beta_1<4.5$.is based on previous data. It was assumed that σ^2 is known and value of $\sigma^2=1.20$, is obtained by pooling mean square error from previous data. The residuals were tested and found normally distributed and all other regression assumptions such as autocorrelation, multicollinearity and hetroscedasticity were tested and found desirable.

Summary of estimates based on sampling theory procedure: The estimated simple linear regression model of grain yield (Y) on pod length (X_7) is given as:

$$\hat{y} = -12 + 3.17\ X_7$$
$$\text{S.E.}\quad (4.471)\quad (1.069)$$

Here b_1 is the estimate of the regression coefficient of pod length (X_7) with standard error of estimate i.e. SE (b_1) = 1.069. A 95% confidence intervals for regression coefficient is given as:

$$0.97 \le \beta_1 \le 5.37$$

The interval suggests that the effect of pod length (X_7) on grain yield (Y) lies between 0.97 and 5.37. But according to the prior inequality information that effect of pod length (X_7) on grain yield (Y) should lies between 0 and 4.5. A little difference between calculated confidence interval and prior inequality information could be ignored because of having same width for both intervals.

Expressing uncertainty about regression parameter with no prior information: Before considering the prior inequality information $0<\beta_1<4.5$, from equation (2) it follows that β_1 is normally distributed with mean and variance given as:

$$\beta_1 \sim N(3.17, 1.120)$$

So the post sample density function for β_1 with no prior inequality information about regression coefficient β_1, after the sample has been observed is given according to the equation (3).

$$f(\beta_1 | y) = \frac{1}{\sqrt{2(1.120)\pi}} \exp\left\{-\frac{1}{2(1.120)}(\beta_1 - 3.17)\right\}$$

Including prior information about regression coefficient β_1: The prior inequality information $0 \le \beta_1 \le 4.5$ can be expressed in term of prior uniform density function as in equation (4) is given as:

$$f(\beta_1) = \begin{bmatrix} 1 & 0 \le \beta_1 \le 4.5 \\ 0 & \text{elsewhere} \end{bmatrix}$$

As prior density function f (β_1) attaches zero probability to the value β_1 of outside the range (0, 4.5). Then the post sample density function that includes this information and the information provided by the sample, must also attach zero probability to the value of β_1 outside the range (0. 4.5) and is given as:

$$f_N(\beta_1 | y) = \frac{1}{\sqrt{2(1.120)\pi}} \exp\left\{-\frac{1}{2(1.120)}(\beta_1 - 3.17)^2\right\} \qquad (5)$$
$$0 \le \beta_1 \le 4.5$$

Here "N" is used as a subscript of $f(\beta_1|y)$ to refer to the normal distribution that express our uncertainty about β_1 after sample has been observed. But from equation (2), the probability of β_1 lying outside the range (0, 4.5) is given as:

$$P(\beta_1 > 4.5) = P\left(\frac{\beta_1 - b_1}{\sigma_{\beta_1}} = \frac{4.5 - 3.17}{1.058}\right) = 0.1056$$

Which cannot be ignored and need to truncate the post-sample density function with prior inequality information included about β_1` given in equation (3.1). So in such situation it is necessary to modify the density $f_N$ `(β_1| y) so that P (β >$_1$4.5) = 0. Then such modified post sample density function is called truncated post sample density function for β_1. 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 post sample density function. Truncated post sample normal density function is obtained by dividing the density $f_N(\beta_1|y)$ to value (1-0.1056) and denoted by $f_{TN}(\beta_1|y)$.

$$= \frac{f_N(\beta_1 | y)}{0.8954}$$

$$f_{TN}(\beta_1 | y) = \frac{f_N(\beta_1 | y)}{[1 - P(\beta_1 > 4.5)]}$$

Finally the truncated post-sample normal density function is given as:

$$f_{TN}(\beta_1 | y) = \frac{(0.8954)^{-1}}{\sqrt{2(1.120)\pi}} \exp\left\{-\frac{1}{2(1.120)}(\beta_1 - 3.17)^2\right\}$$

$$0 \le \beta_1 \le 4.5$$

Point estimation of β_1: As indicated in the section material and methods that the mean of the post-sample density function (which is now truncated post sample density function) is the parameter estimate that also minimizes expected loss. The point estimate of β_1, is as follows:

$$E_{TN}(\beta_1) = \int_0^{4.5} \beta_1 f_{TN}(\beta_1 | y)\, d\beta_1$$

It is very difficult to solve this integral, so we generated 5,000 observations from post sample density function $f_N(\beta_1|y)$, by using $\beta_1 \sim N(3.17, 1.20)$. Observations greater than 4.5 were discarded to obtain a random sample from truncated post-sample density function $f_{TN}(\beta_1|y)$. The sample mean of remaining observations is an estimate of the mean of $f_{TN}(\beta_1|y)$. The summery of artificially generated random sample is given in Table 1

Now to obtain point estimate of β_1, we simply take the mean of retained observations from the sample of 5,000 observations, which is given as under

$$\hat{E}_{TN}[\beta_1] = 2.9505$$

This estimate of β_1 is lower than the estimate 3.17 obtained from approach that does not take into account the prior information. The sample variance from the retained observations on β_1 is an estimate of the variance of $f_{TN}(\beta_1|y)$ and given as:

$$\hat{V}ar_{TN}[\beta_1] = 0.8052$$

Table1: Observations and Estimated Probabilities from artificially generated sample

	Number of observations
Total	5000
Greater than 4.5	538
(Discarded observations)	
Observation from truncated density	4462
Estimated probability P (β_1>4.5)	0.1076
True probability P (β_1>4.5)	0.1056

Table 2: Mean and standard deviation of post sample density function

Statistics	Uncertain prior information	Inequality prior information $0<\beta_1<4.5$
Mean	3.17	2.9505
Standard deviation	1.058	0.89737

The standard deviation of the truncated distribution, 0.89737, is less than that of 1.069 obtained from regression approach that does not take into account the prior information, reflecting the reduction in dispersion when prior information are used (Table 2).

Simulation results: The results obtained by using original data is verified by generating random samples from exponential (1) and uniform (0, 1) distribution. In Table 3, it is clear that estimated effect of pod length from randomly generated data from Uniform (0, 1) distribution tends to 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.

CONCLUSION

The study was initiated to test the effect of including prior non-sample inequality information about regression parameter on estimated effect and its standard error. It was observed that when $0<\beta_1<4.5$ introduced as prior non-sample information, the estimated effect of pod length on grain yield and its standard error was decreased. A simulation procedure was also adopted to test the reliability of the results by generating random samples from normal (0,1) exponential (1) and uniform (0, 1) distributions and it was concluded that the estimated effect for pod length for uniform (0,1)

Table 3: Summery of results from randomly generated data

Statistics	Data generated from normal distribution	Data generated from exponential (1) distribution	Data generated from 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

distribution tends to very close to original results as compared to other two distributions.

REFERENCES

1. Faqir, M., I.A. Arshad and A. Ghafoor, 2004. Development of various prediction models for mash yield with comparison. (Submitted).
2. Geweke, J., 1986, Exact inference in the inequality constrained normal linear regression model. J. Applied Econometrics, 1: 127-141.
3. Griffith, W.E., 1988. Bayesian Econometrics and how to get rid of those wrong signs. Review of Marketing and Agricultural Economics, 56: 36-56.