Original Article

Identifying the Influencing Factors for the BMI by Bayesian and Frequentist Multiple Linear Regression Models: A Comparative Study

R. Vijayaragunathan, Kishore K. John[1], M. R. Srinivasan[2]

Department of Statistics, Indira Gandhi College of Arts and Science, Puducherry, [1]Department of Foreign Trade, Indira Gandhi College of Arts and Science, Puducherry, [2]Adjunct Professor, Chennai Mathematical Institute, Siruseri, Chennai, India

Abstract

Background: In this article, we attempt to demonstrate the superiority of the Bayesian approach over the frequentist approaches of the multiple linear regression model in identifying the influencing factors for the response variable. **Methods and Material:** A survey was conducted among the 310 respondents from the Kathirkamam area in Puducherry. We have considered the response variable, body mass index (BMI), and the predictors such as age, weight, gender, nature of the job, and marital status of individuals were collected with the personal interview method. Jeffreys's Amazing Statistics Program (JASP) software was used to analyze the dataset. In the conventional multiple linear regression model, the single value of regression coefficients is determined, while in the Bayesian linear regression model, the regression coefficient of each predictor follows a specific posterior distribution. Furthermore, it would be most useful to identify the best models from the list of possible models with posterior probability values. An inclusion probability for all the predictors will give a superior idea of whether the predictors are included in the model with probability. **Results and Conclusions:** The Bayesian framework offers a wide range of results for the regression coefficients instead of the single value of regression coefficients in the frequentist test. Such advantages of the Bayesian approach will catapult the quality of investigation outputs by giving more reliability to solutions of scientific problems.

Keywords: Bayesian regression model, BMI, JASP software, posterior distribution, variance inflation factor

Introduction

The multiple linear regression model is a widely used tool across all fields of research. The standard frequentist methods are used ubiquitously, but late Bayesian methods are used as an alternative to the frequentist methods for model comparisons. The concept of Bayesian linear regression on default Bayes factors for multiple regression designs was discussed by Liang *et al*. (2008).[1] Rouder and Morey (2012)[2] proposed the default Bayes factors for model selection in regression. Nevertheless, due to the lack of practical guidance and application of software for interpreting Bayes factors in multiple regression, these proposals did not gain much popularity among researchers earlier. The advancements in information technology in the last one decade have made it possible to apply Bayesian concepts more definitively. Yu *et al*. (2013)[3] developed Bayesian methods for variable selection, with a simple and efficient stochastic search variable selection (SSVS) algorithm proposed for posterior computation and demonstrated the same with simulated data. Kelter (2020)[4] who investigated the behavior of Bayesian indices of significance in medical research applying frequentist methods used Bayesian analysis on the simulation datasets to draw significant conclusions. Vijayaragunathan and Srinivasan (2020)[5] discussed Bayes factors for comparison of two-way analysis of variance models with illustration through the concept of the Bayesian multiple regression model.

Access this article online

Quick Response Code:

Website: www.ijcm.org.in

DOI: 10.4103/ijcm.ijcm_119_22

Address for correspondence: Dr. R. Vijayaragunathan, Department of Statistics, Indira Gandhi College of Arts and Science, Kathirkamam, Puducherry, India.
E-mail: rvijayaragunathan@gmail.com

How to cite this article: Vijayaragunatha R, John KK, Srinivasan MR. Identifying the influencing factors for the BMI by Bayesian and frequentist multiple linear regression models: A comparative study. Indian J Community Med 2023;48:659-65.

Received: 03-02-23, **Accepted:** 04-08-23, **Published:** 07-09-23

Ranjani *et al*. (2016)[6] analyzed the overweight and obesity rates in children and adolescents and found them increasing among lower and higher socioeconomic groups. Chu and Yuan (2018)[7] have proposed and compared the Bayesian hierarchical model approach to evaluate the treatment effect in basket trials. A study on body mass index (BMI) was conducted broadly based on parameters such as region, country, gender, and income based on data belonging to more than 30 years. In particular, the researchers focused on women from rural areas that belong to increasing obesity NCD Risk Factor Collaboration (NCD-RisC) (2017).[8] In a study by Sutin *et al*. (2021),[9] data were collected from three different time periods from the participants and analyzed against personal details such as weight, height, depression effect, and loneliness. It was also observed that the coronavirus pandemic increases the risk of incident depression and leads to a decline in well-being among the participants. Nikooseresht *et al*. (2015)[10] found that the application of the Bayesian technique for the treatment of any disease calls for integrating the results of multiple tests and is also based on available prior knowledge that needs to be analyzed before arriving at a decision.

Subjects and Methods

The concepts of frequentist and Bayesian approaches to multiple linear models will be discussed in this section.

Classical multiple linear regression model

The classical linear regression model may be written as follows:

$$y = X\beta + \epsilon \quad (1)$$

where $y = (y_1, y_2, ...y_n)^T$ is $n \times 1$ column vector, $X = (x_{i1}, x_{i2}, ..., x_{id})_{i=1}^{d}$ is $n \times d$ design matrix containing all variables, $\beta = (\beta_0, \beta_1, ..., \beta_{d-1})^T$ is the $d \times 1$ vector of parameters, and $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$ is the $n \times 1$ vector of errors and $\epsilon \sim N(0, \sigma^2 I)$.

In the classical context, the asymptotic distribution for $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$ and in the Bayesian context the distribution $\hat{\beta}$ of depend on the choice of the prior distribution, so it will not necessarily be a normal distribution.

Variance inflation factor (VIF)

When two variables are in near-perfect linear combinations with one another, collinearity will arise. When this involves more than two variables, it is known as multicollinearity. If the regression estimates are unstable and have high standard errors, it is an indication of the presence of multicollinearity. VIFs are a measure of the inflation in variances of estimates of the parameters due to collinearities that exist among the independent variables. The VIF for the j[th] explanatory variable is defined as follows:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (2)$$

where R_j^2 is the coefficient of determination obtained when X_j is regressed on the remaining $(k-1)$ variables. If VIF is 1, there is no correlation between predictor and remaining predictors, and then, the variance of the regression coefficient of a predictor is not inflated at all. If the VIF value is more than 5 or 10, a particular coefficient is estimated poorly or unstable because of near-nonlinear dependence among the regressor.

Condition index

This is like an alternate method to VIFs, which shows the degree of multicollinearity in a regression design matrix. Almost all the statistical packages used this index to find the collinearity between the variables. The condition indices are as follows:

$$C_j = \frac{\lambda_{max}}{\lambda_j} \quad (3)$$

where $\lambda_1, \lambda_2, ..., \lambda_k$ be the eigenvalues of $X'X$. However, if the eigenvalue is very small, a symptom of seriously ill-conditioned data and also the condition index are large, say more than 1000, which means that it is near-linearly dependent.

Bayesian multiple linear regression model

Bayesian linear regression is used when the data are distributed poorly or one has to deal with insufficient data. In the absence of data, the priors can take over, and this mechanism allows us to set prior based on the coefficients and the noise. In the Bayesian inference, the response variable y is not estimated as a single value; it is assumed to be drawn from a specific distribution. Rouder and Morey (2012)[2] discussed default Bayes factors for model selection in regression. Rockafellar *et al*. (2014)[11] presented a super-quantile regression with several numerical examples in the area of uncertainty quantification. Tsionas and Izzeldin (2018)[12] provided the Bayesian interpretation of the conditional value at risk; that is, super-quantile regression and computations are based on particle filtering using a special posterior distribution consistent with the super-quartile concept.

Consider a Bayesian linear model is $y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$, and then, the likelihood function is as follows:

$$f_y(X, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta)\right\} \quad (4)$$

We assume a conjugate prior distribution for $\beta \mid \sigma^2$ as follows: $N(m, \sigma^2 V)$. Thus, the data and prior information are involved in the interpretation of Bayesian inferences.

$$f(\sigma^2, m, V) = (2\pi\sigma^2)^{-\frac{n}{2}} |V|^{-\frac{d}{2}}$$

$$exp\left\{-\frac{1}{2\sigma^2}(\beta - m)^T V^{-1} (\beta - m)\right\} \quad (5)$$

The inverse-gamma distribution plays a conjugate prior distribution for σ^2 as follows:

$$f(a,b)=\frac{b^a}{\Gamma(a)}(\sigma^2)^{a-1}\,exp\,exp\left\{-\frac{b}{\sigma^2}\right\}$$

a >0, b>0 (6)

The posterior distribution of the parameters β and is σ^2 as follows:

$$f(y,X)\propto(\sigma^2)^{-\frac{d}{2}}\,exp\left\{-\frac{(\beta-\mu)^T\Lambda^1(\beta-\mu)}{2\sigma^2}\right\}$$

$$(\sigma^2)^{-\frac{n}{2}+a-1}\;e^{\left\{-\frac{m^TV^{-1}m-\mu^T\Lambda^1\mu+2b+y^Ty}{2\sigma^2}\right\}} \quad (7)$$

Thus,

$$f(y,X)\propto N(\mu,\sigma^2\Lambda)\times IG(a^*,b^*) \quad (8)$$

where $a^*=-\frac{n}{2}+a,$

$$b^*=\frac{m^TV^{-1}m-\mu^T\Lambda^1\mu+2b+y^Ty}{2}$$

Statistical Analysis and Results

We considered the example of BMI as a response variable with important independent variables such as age, weight, gender, job, and marital status of each individual for the linear regression model. To begin with this illustration, we applied the classical multiple linear regression model and identified the influencing factors of BMI. Additionally, the same illustration was applied in the Bayesian multiple linear regression model to identify the posterior distribution of independent variables.

Classical multiple linear regression model

The multiple regression equation of BMI on five predictors is as follows:

$$Y=24.141+0.025\,X_1+0.321\,X_2+2.595\,X_3-0.169\,X_4-0.467\,X_5 \quad (9)$$

where Y = BMI, X_1= age (in years), X_2= weight (in Kgs), X_3=gender (1=male, 2=female), X_4=job (5=agriculture, 4=hard work, 3 = office work, 2 = business, 1 = others), X_5= marital status (1 = married, 2 = unmarried).

From Table 1, the intercept is β_0= 24.141, and it shows that BMI would be expected if all the predictors are zero. The regression coefficient for the predictor "age" is 0.025, which indicates that for each year of age increase, the BMI increases by 0.025 if other predictors are the same. The regression coefficient for "weight" is 0.321, indicating that for every additional 10 kilograms of individual weight, BMI will increase by 3.21 if the other predictors are the same. Also, the predictor "gender" will add another 2.595 for males and 5.19 for females in BMI. However, the other two predictors such as "job" and "marital status" have negative values. According to the individual's nature of the job and marital status, the BMI will decrease by -0.169 and -0.467 (a married person's BMI to some extent more than the unmarried), respectively, but these do not significantly influence the response variable BMI.

The R-square value for the regression multiple linear models is 0.812; that is, 81.2% of total variations in the response variable BMI are explained by the predictors such as "age," "weight," and "gender" because their significant values are less than 0.05. Furthermore, in the ANOVA (Analysis of Variance) output, we observed that the F ratio is 247.901 and the corresponding *P* value is less than 0.05, which means all the regression coefficients are significantly different. In our results, the VIFs for all the predictors are all less than 10, which indicates that multicollinearity is not a serious concern.

From Table 2, there are five predictors so we have six dimensions and the eigenvalue for all the dimensions is less than 1 except the first dimension. It is an indication of multicollinearity as several eigenvalues are close to 0. The interpretation of the condition index is the square root of the ratio of the largest eigenvalue to the eigenvalue of the dimension. The condition index for dimensions 1 to 5 in Table 2 is less than 15; therefore, multicollinearity does not arise. However, the condition index for dimension 6 is 25.840, which is greater than 15, indicating the presence of multicollinearity. Furthermore, to find the variance proportions or variance decomposition proportions we used eigenvalues to calculate eigenvectors. For each condition index, the variance proportions are calculated for each predictor. The sum of the variance proportion for each predictor is 1. However, if the condition index is higher than 30 then the at least two predictor variance proportions are exceeded by 80% to 90%, which shows that multicollinearity is present between the respective

Table 1: Regression coefficients, their significant values, and variance inflation factors

Model	Coefficient	Unstandardized	S.E.	Standardized	*t*	*P*	Collinearity	
							Tolerance	VIF
H0	Intercept	24.141	0.267		90.352	<0.001		
H1	Intercept	0.626	1.147		0.545	0.586		
	Age	0.025	0.009	0.083	2.822	0.005	0.752	1.331
	Weight	0.321	0.009	0.918	34.468	<0.001	0.924	1.082
	Gender	2.595	0.255	0.279	10.158	<0.001	0.865	1.156
	Job	-0.169	0.112	-0.043	-1.502	0.134	0.804	1.244
	Marital	-0.467	0.330	-0.042	-1.417	0.158	0.743	1.345

Table 2: Collinearity diagnostics for the multiple regression model

Dimension	Eigenvalue	Condition index	Variance proportions					
			Intercept	age	Weight	gender	Job	Marital
1	5.521	1.000	0.000	0.003	0.001	0.002	0.004	0.002
2	0.227	4.935	0.000	0.189	0.001	0.009	0.237	0.041
3	0.103	7.330	0.000	0.161	0.007	0.539	0.008	0.052
4	0.085	8.045	0.001	0.082	0.002	0.001	0.670	0.441
5	0.056	9.916	0.002	0.205	0.520	0.034	0.044	0.159
6	0.008	25.840	0.996	0.359	0.470	0.414	0.038	0.305

predictors, but in our example, all the variance proportions are below 80%, which means that multicollinearity does not occur.

The Q–Q plot (quantile–quantile plot) is a graphical tool to help us assess a set of data drawn from the specified theoretical distribution. Here, we used a Q–Q plot to check the assumption that the dependent variable is normally distributed. From Figure 1, the dependent variable is normally distributed since most of the points are on the line. The difference between the observed value and the predicted value is residuals that are equal across the values of the dependent variable, which is known as homoscedastic. In this case, the points are randomly scattered around the x-axis, and therefore, the data are homoscedastic.

Bayesian multiple linear regression model

The aim of using Bayesian linear regression is to determine the posterior distribution for the model parameters instead of finding the single "best" value of the model parameters in the classical linear regression model. In the Bayesian multiple linear regression model, we used a uniform distribution as a prior probability of the model, and it produces a probability value of 0.077; furthermore, we compare 15 possible models to identify the best model as shown in Table 3.

The posterior probability of the model for the top three models is 0.942, and the remaining models have the least probability. The prior probabilities will be redistributed to the posterior probabilities for these models according to the strength of independent variables. The best model consists of the three predictors such as "age," "weight," and "gender," and its posterior distribution probability value is 0.614, which is among all possible models, and this model result explains 61.4% of the response variable compared with all other models. Additionally, the second-best model consists of four predictors such as "age," "weight," "gender," and "job" of the individuals and will provide 17.4% posterior probability compared with other models. Similarly, the third-best model also provides almost the same result of 15.4% when the "job" is replaced by "marital" in the model. However, the posterior distribution probabilities for the other remaining models are 2.6%, 1.3%, 1.0%, and less than 1% compared with other models. Interestingly, the multiple linear regression model (full model) consisting of all the predictors provides only 2.6% of the posterior probability compared with all other possible models.

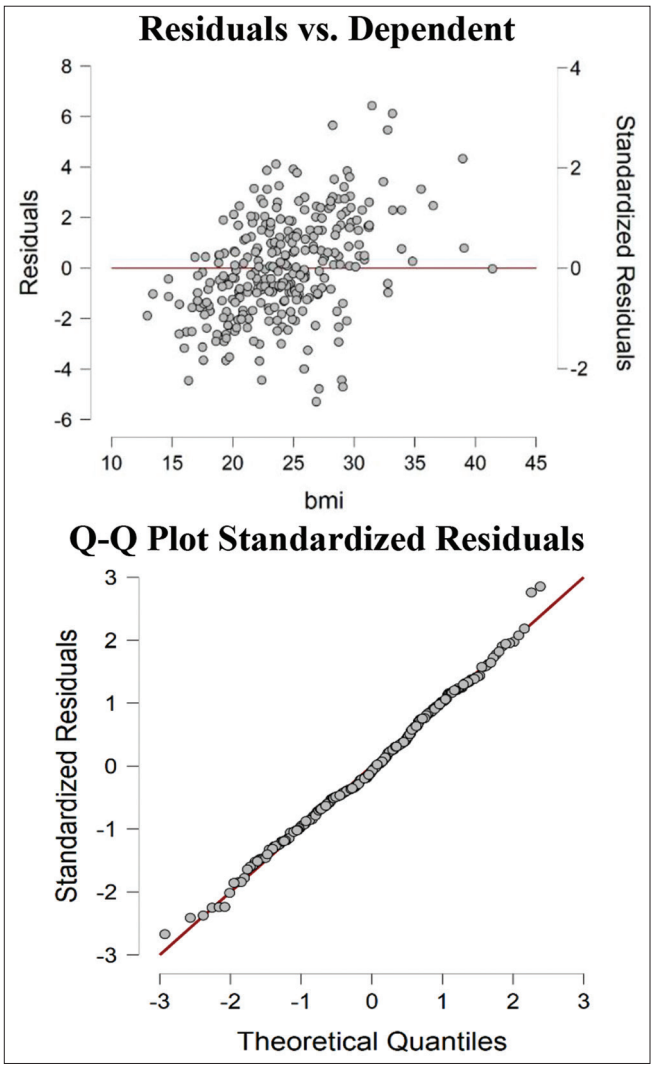

Figure 1: Scatter plot for residual against BMI and Q–Q plot for standardized residuals

The R-square value for the multiple linear regression equation with all predictors is 0.812, and the R-square value for the best model, which consists of three predictors such as "age," "weight," and "gender," is 0.808. These two models provide almost the same results in model fitting. The inclusion probabilities for the predictors such as "weight," "gender," and "age" are 0.846, 0.615, and 0.462, respectively. Also, the probability of inclusion given the data for "intercept," "weight,"

"gender," and "age" has the highest probability, which indicates that these predictors are compulsorily included in the model. However, the other predictors such as "job" and "marital status" have small probabilities to include in the model. The mean and SD (Standard Deviation) of regression coefficients are given in Table 4, and the mean of intercept, age, weight, gender, job, and marital status is 24.141, 0.031, 0.321, 2.688, -0.046, and -0.121; their standard deviations are 0.117, 0.010, 0.009, 0.256, 0.102, and 0.287, respectively. Interestingly, the regression coefficients in the frequentist and Bayesian multiple linear regression model are reasonably different.

The main advantage of Bayesian approaches to multiple linear regression is that each regression coefficient follows a specific distribution based on the prior. Here, we used Jeffreys–Zellner–Siow prior to find the Bayes factors. Similar to the ordinary least-squares procedure, the posterior means and standard deviations of the coefficients are extracted from the posterior distributions. The mean of regression coefficients is completely different from the frequentist multiple regression model. The posterior distributions of the regression coefficients with the spread of the distribution related to the standard errors are visualized in Figure 2. We may give a clearer and more useful summary regression model to the five predictors as in the upper and lower bounds of 95% credible intervals of all coefficients. The 95% credible interval of predictors for "age" is (0.014,0.051), for "weight" is (0.304,0.339), for "gender" is (2.199,3.152), for "job" is (−0.280,0.001), and for "marital status" is (−0.820,0.004).

The weightage of inclusion of independent variables in the model "weight," "gender," and "age" is 3.619×10^{98}, 3.1×10^{20}, and 34.78, respectively, but the inclusion of the other two independent variables such as "job" and "marital status" is negligible.

Discussion and Conclusions

In contemporary research, the application of Bayesian concepts has become a trend in model-building problems. In this study, we have employed JASP software to do statistical analysis for finding frequentist and Bayesian multiple linear regression models. We built both the frequentist and Bayesian multiple linear regression models using obesity data collected from a section of people in Puducherry State, India. As the dataset is not big, applying the Bayesian concept is more appropriate. In the frequentist method, the percentage of fitting of model, significance of levels of the regression coefficients, and for finding multicollinearity, we used the measures of VIF and condition index. Particularly, the BMI of the individual is influenced by the three predictors such as "age," "weight," and "gender." Interestingly, the Bayesian multiple regression model provides the "best" model among the list of models displaying how much data support the model. Furthermore, the independent variables form their own posterior distribution with inclusion probability for the model and also provide credible intervals for the predictors. From our study, the "weight" of the individual has a high probability to be included in the model and its posterior distribution shows that there is

Table 3: Bayesian multiple linear model comparisons for the response variable "BMI"

Models	P (M)	P (M|data)	BF_M	BF_{10}	R^2
Age + weight + gender	0.077	0.614	19.086	1.000	0.808
Age + weight + gender + job	0.077	0.174	2.524	0.283	0.811
Age + weight + gender + marital	0.077	0.154	2.178	0.250	0.811
Age + weight + gender + job + marital	0.077	0.026	0.323	0.043	0.812
Weight + gender + marital	0.077	0.013	0.153	0.021	0.803
Weight + gender + job	0.077	0.010	0.117	0.016	0.803
Weight + gender + job + marital	0.077	0.009	0.113	0.015	0.807
Weight + gender	0.077	8.10e-4	0.010	0.001	0.795
Age + weight + job	0.077	8.14e-22	9.77e-21	1.32e-21	0.732
Age + weight	0.077	7.93e-22	9.52e-21	1.29e-21	0.727
Weight	0.077	4.07e-22	4.88e-21	6.63e-22	0.719
Marital	0.077	4.09e-100	4.91e-99	6.66e-100	0.030
Null model	0.077	9.30e-101	1.11e-99	1.51e-100	0.000

Table 4: Posterior summaries of regression coefficients

Coefficient	P (incl)	P (excl)	P (incl|data)	P (excl|data)	BF inclusion	Mean	SD	95% credible interval	
								Lower	Upper
Intercept	1.000	0.000	1.000	0.000	1.000	24.141	0.117	23.914	24.359
Age	0.462	0.538	0.968	0.032	34.780	0.031	0.010	0.014	0.051
Weight	0.846	0.154	1.000	0.000	3.629e+98	0.321	0.009	0.304	0.339
Gender	0.615	0.385	1.000	0.000	3.100e+20	2.688	0.256	2.199	3.152
Job	0.385	0.615	0.219	0.781	0.449	-0.046	0.102	-0.280	0.001
Marital status	0.385	0.615	0.202	0.798	0.404	-0.121	0.287	-0.820	0.004

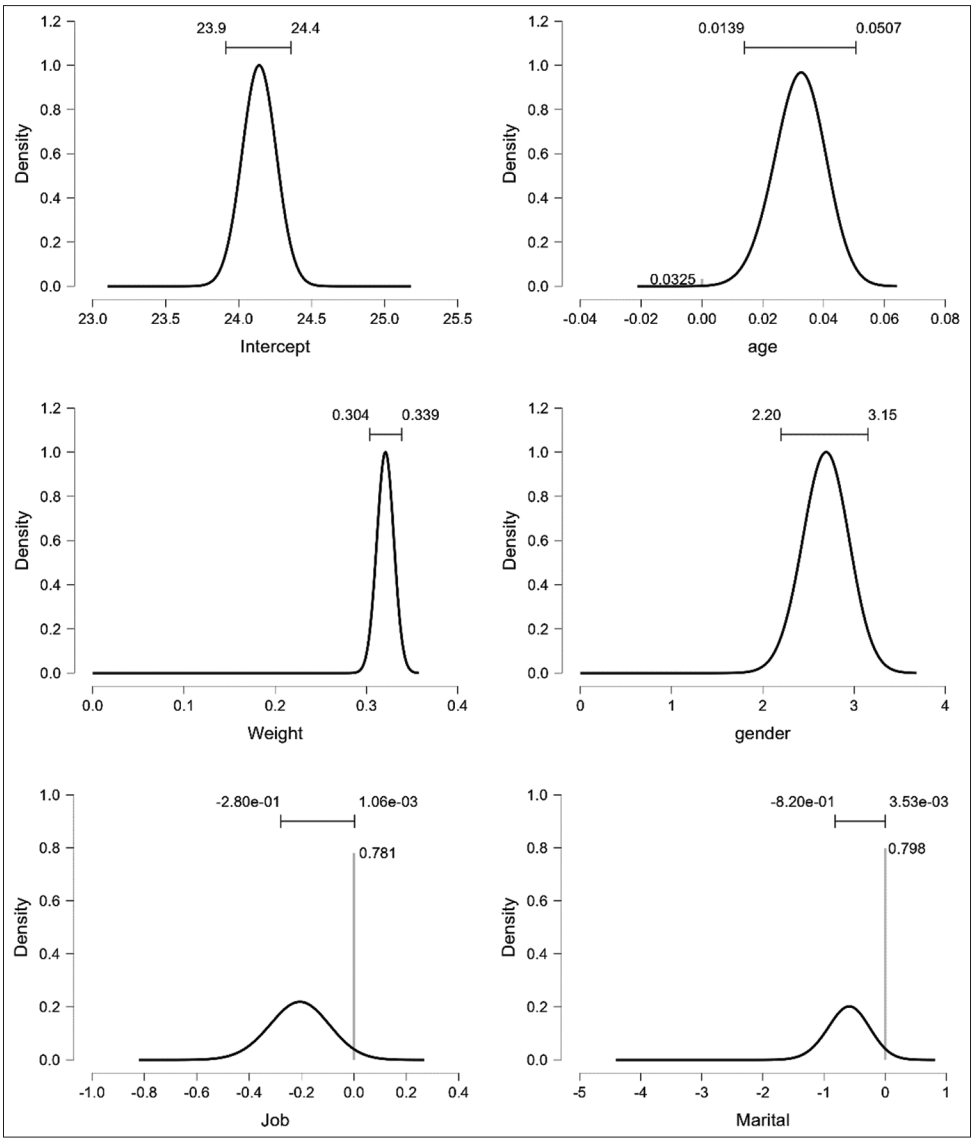

Figure 2: Posterior distributions for intercept and five predictors

less variability. Similarly, the "age," "gender," and "intercept" have high inclusion probabilities. These results may be helpful to doctors or nutritionists to reconstruct the new structure of the BMI for region-wise or country-wise groups of people and provide pertinent treatment to the patients. To conclude, the Bayesian framework offers a wide range of results for the regression coefficients instead of the single value of regression coefficients in the frequentist test. It may be useful to the researcher to draw expressive conclusions for a better understanding of variables in the data.

Acknowledgments

The authors would like to thank the editor and the reviewers for their constructive comments and suggestions, which highly improved the paper.

Author's contributions

All the authors contributed and approved the final manuscript.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g priors for Bayesian variable selection. J Am Stat Assoc 2008;103:410-23.
2. Rouder JN, Morey RD. Default Bayes factors for model selection in regression. Multivariate Behav Res 2012;47:877-903.
3. Yu T, Xiang L, Wang HJ. Quantile regression for survival data with covariates subject to detection limits. Biometrics 2021;77:610-21.
4. Kelter R. Simulation data for the analysis of Bayesian posterior significance and effect size indices for the two-sample *t*-test to support reproducible medical research. BMC Res Notes 2020;13:1-3.
5. Vijayaragunathan R, Srinivasan MR. Bayes factors for comparison of two-way ANOVA models. J Stat Theory Appl 2020;19:540-6.
6. Chu Y, Yuan Y. A Bayesian basket trial design using a calibrated Bayesian hierarchical model. Clin Trials 2018;15:149-58.

7. NCD Risk Factor Collaboration (NCD-RisC). Rising rural body-mass index is the main driver of the global obesity epidemic in adults. Nature 2019;569:260-4.
8. Sutin AR, Stephan Y, Luchetti M, Aschwanden D, Strickhouser JE, Lee JH, *et al*. BMI, weight discrimination, and the trajectory of distress and well-being across the coronavirus pandemic. Obesity (Silver Spring) 2021;29:38-45.
9. Nikooseresht Z, Rimaz S, Asadi-Lari M, Nedjat S, Merghati-Khoee E, Saiepour N. Reliability and validity of the Iranian version of the human immunodeficiency virus specific World Health Organization quality of life BREF questionnaire. J Biostat Epidemiol 2015;1:37-44.
10. Rockafellar RT, Royset JO, Miranda SI. Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. Eur J Oper Res 2014;234:140-54.
11. Ranjani H, Mehreen TS, Pradeepa R, Anjana RM, Garg R, Anand K, *et al*. Epidemiology of childhood overweight and obesity in India: A systematic review. Indian J Med Res 2016;143:160-74.
12. Tsionas MG, Izzeldin M. Bayesian CV@R/super-quantile regression. J Appl Stat 2018;45:2943-57.