Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 135 (2018) 671–677

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Computer Science and Computational Intelligence 2018

Linear regression model using bayesian approach for energy performance of residential building

Syarifah Diana Permai[a]*, Heruna Tanty[b]

[a]*Department of Statistics, School of Computer Science, Bina Nusantara University, Jakarta 11480*
[b]*Department of Mathematics, School of Computer Science, Bina Nusantara University, Jakarta 11480*

Abstract

In the statistics there are two types of points of view, Frequentist and Bayesian. The difference between Frequentist and Bayesian is the point of view in terms of looking at a parameter. Bayesian views a parameter as a random variable, it means the value is not a single value. The modeling method that most commonly used by researchers is linear regression model. The Frequentist methods that are often used in linear regression are Ordinary Least Square (OLS) and Maximum Likelihood Estimation (MLE). However, along with the Bayesian development, several studies have shown better modeling results than the Frequentist method. On the other hand, Bayesian approach is also used when assumptions in linear regression model using OLS are not met. Therefore, this research performs linear regression modeling with Bayesian approach. The analysis showed that linear regression model using OLS does not met all assumptions. It means the model is not good enough. Then, Bayesian approach can be used as an alternative for the model. The comparison of Bayesian and Frequentist modeling results using several criteria such as RMSE, MAPE and MAD. The results showed that the linear regression method using Bayesian approach is better than Frequentist method using OLS.

Keywords: Linear regression; bayesian; frequentist

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: syarifah.permai@binus.ac.id

10.1016/j.procs.2018.08.219

1. Introduction

Linear regression analysis is one of the most commonly used statistical methods for modeling cross section data. In regression modeling there are two kinds of variables, dependent variable (variables that are influenced or value depend on other variables) and independent variable (variable that is suspected to affect dependent variable). Several provisions of the dependent variable in the regression analysis can only be one variable and the data scale is interval / ratio so that data can only be in the form of numerical data. While the independent variable can be more than one variable and the data scale can be nominal / ordinal ie categorical data or interval / ratio of numerical data. If the number of independent variables used only one variable then called simple linear regression analysis, whereas if the number of variables used more than one is called by multiple linear regression analysis. On the scale of the data used for independent variables, if the data scale used is categorical data then called dummy regression analysis.

There are three goals in linear regression analysis, (1) form regression model to know the relation between dependent variable and independent variables, (2) to test whether there is influence of independent variables to dependent variable, and (3) to predict the value of variable dependent based on independent variables that have been determined. Formation of the regression model is done by estimating the parameters of the regression model. So that will yield regression coefficient for each independent variable. There are several methods that can be used to estimate regression model parameters. The most frequent method used by researchers is the Frequentist / Classic method by using OLS (Ordinary Least Square) or MLE (Maximum Likelihood Estimation) [1].

OLS method is also called the least squares method, done by minimizing the number of errors from the regression equation. Thus the regression model parameters are obtained from minimizing the error function of the equation. While MLE is done by maximizing the probability density function (probability density function) of a data. So by using both methods, there are classical assumptions that must be met from the results of regression modeling. Some assumptions that must be met are error independent, identic and normal distribution [2]. In addition to these two methods, there are other methods that can be used to estimate the regression model parameters that is Bayesian approach method. The difference between Frequentist and Bayesian methods is on the point of view of parameters. The Bayesian approach views parameters as a random variable so that the value is not a single value as in the Frequentist point of view [3].

Several studies have been carried out in applying Bayesian approaches to several models such as Chaturvedi doing Robust Bayesian analysis on linear regression modeling [4]. Parhusip analyzes susenas (revenue and expenditure) data by applying Bayesian regression modeling to estimate parameters and create confidence intervals. Azhar compared the Bayesian method with Likelihood in linear regression modeling applied to case studies of inflation, exchange rates and stock prices [2]. Iswari, Sumarjaya, Srinadi conducted a simple linear regression analysis and a confidence interval of regression parameters using simulation data where the prior distribution is unknown [3]. Kabir, et al. integrates Bayesian linear regression with Ordered Weighted Averaging (OWA) applied to predict a water distribution network in Calgary, Canada. Abidin used Bayesian approach to ridge regression modeling in cases of poverty in East Java [5]. Mutiarani, et al. performed Bayesian linear regression analysis using skew-symmetric error distribution and applied to survival analysis [6].

There are some assumptions in linear regression using ordinary least square estimation. This regulations can cause some problems if the assumptions are not met. The existence of outliers can cause the linear regression model did not to meet the assumption of an independent residual. Bayesian linear regression can be used to accommodate outliers [7]. This research used energy efficiency dataset to determine the cooling equipment needed to maintain air conditions. Tsanas and Xifara performed energy performance of residential building using Iteratively Reweighted Least Square (IRLS) and Random Forest [8]. The main objectives of this study including: (1) to build a linear regression model using Ordinary Least Square (OLS) method, (2) to build a linear regression model using Bayesian approach, (3) to find out the best linear model.

2. Linear Regression Model

There are two linearity definitions that are linear to parameter and linear to variable. Modeling is called a linear model if the model is linear to the parameter. So even if the model is not linear to the variable or linear to the variable, as long as the linear model to the parameter, then the model is called linear model [9]. Multiple linear regression models can be used to evaluate the relationship between dependent variables with two or more independent variables. Before

doing linear regression modeling, it must first be ensured that there is a relationship or correlation between each independent variable to the dependent variable and the relationship between the dependent variable with all independent variables are linear. Testing whether or not the relationship between independent variables with variable dependent can use Pearson or Spearman correlation test [10]. There are several testing methods that can be used for the linearity test, that are White test, Terasvirta test and Reset test [11].

If there is one dependent variable (Y) of p independent variables (X_1, X_2, …,X_p), then multiple linear regression models can be written as follows [12]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{1}$$

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{2}$$

where
Y = dependent variable (n × 1)
X = the matrix of independent variable (n × (p +1))
β = vector of regression model parameters ((p +1) × 1)
ε = vector of error (n × 1)

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The β parameter is estimated to obtain the regression model, here is the formula to estimate the parameter:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y} \tag{3}$$

3. Bayesian Linear Regression

Bayesian linear regression is one of regression modeling with parameter estimation method using Bayesian approach. In the Bayesian approach there is a prior, likelihood distribution and posterior distribution. Parameter estimation by Bayesian approach is done by processing posterior distribution which multiplies the prior distribution with likelihood. In linear regression model using OLS estimation method, there is normal distributed error assumption that is $\varepsilon \sim N(0, \sigma^2)$. Since the error is normally distributed, the variables $(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)$ are also normally distributed. Thus the variables $(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) \sim N(\mathbf{X\beta}, \sigma^2)$ and probability density function (pdf) of these variables are as follows [13]

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta})\right] \tag{4}$$

Based on the probability density function (pdf) above can be defined likelihood function of these variables are as follows:

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta})\right] \tag{5}$$

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = (\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta})\right] \tag{6}$$

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-\frac{v}{2}} exp\left[-\frac{vs^2}{2\sigma^2}\right] \times (\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta})\right] \tag{7}$$

There are several prior distributions that can be used in the Bayesian approach of linear regression model, one of which is the distribution of prior conjugate [14]. Estimation of regression model parameters with Bayesian approach can

be done by iterating at marginal posterior. Posterior distribution is calculated by multiplying the prior distribution and likelihood function [13].

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \tag{8}$$

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}, \boldsymbol{X}) \propto p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)p(\sigma^2)p(\boldsymbol{\beta}|\sigma^2) \tag{9}$$

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}, \boldsymbol{X}) \propto (\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\boldsymbol{X\beta})\right] \times (\sigma^2)^{-\left(\frac{v}{2}+1\right)} exp\left[-\frac{vs^2}{2\sigma^2}\right] \times (\sigma^2)^{-k/2} exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{\beta}-\boldsymbol{\mu})\right] \tag{10}$$

The process of obtaining the estimation of regression model parameters with Bayesian approach can be done by using MCMC (Markov Chain Monte Carlo) algorithm. One of the commonly used algorithms in MCMC is Gibbs Sampling. Next is the iteration process to estimate the parameters until the burn-in conditions are met.

4. Methodology

This research uses secondary data which obtain from UCI Machine Learning Repository [15]. Data was collect from 12 different building shapes. The samples used are 768 samples. Based on energy efficiency cases, this research using several independent variables (X1, X2, X3, X4 and X5) and one variable dependent (Y). Here are the variables:

X1 = Relative Compactness
X2 = Surface Area
X3 = Wall Area
X4 = Overall Height
X5 = Glazing Area is the percentage of the floor area
Y = Cooling Load is a rate to remove the heat to maintain air conditions.

Furthermore, data is processed using linear regression modeling with Frequentist method. This model is estimate the parameter using Ordinary Least Square method. This data is also modeled by Bayesian method. The R package that can be used to Bayesian approach in linear regression model is MCMCpack [16 17]. Results from both models were compared to obtain the best model using RMSE, MAD and MSE criteria.

5. Results and Discussions

Before do the analysis using regression model, the first step is correlation test to determine whether there is a relationship between independent variable and independent variable. Furthermore is linear regression analysis. The parameter estimation methods of regression model using Ordinary Least Square and Bayesian. Here is the result of correlation analysis between independent variables with variable dependent.

H_0 : $\rho = 0$ (There is no correlation between two variables)
H_1 : $\rho \neq 0$ (There is correlation between two variables)

Table 1. Correlation test between dependent variable and independent variables

Variable	Correlation coefficient	p-value
X_1	0.6343391	$< 2.2 \times 10^{-16}$
X_2	-0.6729989	$< 2.2 \times 10^{-16}$
X_3	0.427117	$< 2.2 \times 10^{-16}$
X_4	0.8957852	$< 2.2 \times 10^{-16}$
X_5	0.207505	6.457×10^{-9}

Based on Tabel 1 above the correlation test showed that the independent variables (Relative Compactness, Surface Area, Wall Area, Overall Height and Glazing Area) are correlated with the dependent variable. It is shown from all p-value is less than α, with α value is 0.05. Therefore concluded the correlation value is not equal to zero, in other words the variables are correlated each other, so it can be concluded that there is a relationship between independent variables with variable dependent. Because there is a relationship between the independent variables to the dependent variable, then the data can be continued on regression modeling. Table 2 is the output of multiple linear regression modeling between Cooling Load variable with Relative Compactness, Surface Area, Wall Area, Overall Height and Glazing Area using OLS method.

Table 2. Linear regression model using OLS method

	Estimate	T value	p-value
Intercept	97.76185	4.71	2.94×10^{-06}
X1	-70.7877	-6.307	4.80×10^{-10}
X2	-0.08825	-4.738	2.57×10^{-06}
X3	0.044682	6.162	1.16×10^{-09}
X4	4.283843	11.62	$< 2 \times 10^{-16}$
X5	14.81797	17.082	$< 2 \times 10^{-16}$

Based on the regression model below, 3 variables give a positive influence and 2 variables give negative influence. This is shown from the estimate value of the parameter coefficients. Surface area variable and wall area variable can be interpreted if the area is wider, then the cooling load is low. Otherwise, relative compactness variable, overall height variable and glazing area variable can be interpreted if that variables are increases then cooling load will also increase. Further testing of significance parameters. This is done to determine the influence of each independent variable to the dependent variable. The result of simultaneously and partially test indicate that all variables affect the cooling load.

$$Y = 97.76185 - 70.7877\, X_1 - 0.08825\, X_2 + 0.044682\, X_3 + 4.283843\, X_4 + 14.81797\, X_5$$

The results of assumptions test on the regression model showed that residual is not normally distributed, residual is not independent and residual is not identical. This indicates that the IIDN assumptions on the OLS regression model is not met. The VIF test for the model show that multicollinearity has occurred. Based on this results, the multiple linear regression using the OLS parameter estimation method does not match the result because the assumptions of linear regression using OLS are not met. Therefore, another parameter estimation method is used to form the regression model. One of method for estimating the parameters that can be used is the Bayesian approach.

In this research, multiple linear regression modeling analysis with Bayesian parameter estimation is used. The prior distribution used in this study is the Normal distribution for the β parameter and the Gamma inverse distribution for the parameter σ^2. The algorithm used is Gibbs Sampler using Markov Chain Monte Carlo (MCMC) method. Iteration used as many as 10000 with Burn in at 500 and thin of 1.

Table 3. Linear regression model using Bayesian approach

	Estimate	Quantile (2.5%)	Quantile (97.5%)
Intercept	97.7747	57.41889	137.8781
X1	-70.8027	-92.5945	-49.0122
X2	-0.08824	-0.12443	-0.05174
X3	0.04466	0.03031	0.05878
X4	4.28452	3.57349	5.01809
X5	14.81861	13.14172	16.54031

The model of linear regression analysis with Bayesian approach is

$$Y = 97.7747 - 70.8027\ X_1 - 0.08824\ X_2 + 0.04466\ X_3 + 4.28452\ X_4 + 14.81861\ X_4$$

The parameters of linear regression model with Bayesian approach above is not too different from OLS method. Then, to determine the best method using several criteria by comparing the values of RMSE, MAD and MAPE.

Table 4. Comparison between OLS and bayesian in linear regression model

Estimation Method	RMSE	MAD	MAPE
Ordinary Least Square	3.187969	2.247922	8.925759%
Bayesian	3.187969	2.247842	8.925359%

Tabel 4 the comparison of criteria showed that the RMSE (Root Mean Square Error) value of OLS method is smaller than the linear regression model using Bayesian. While based on the criteria of MAD (Mean Absolute Deviance) and MAPE (Mean Absolute Percentage Error) showed that linear regression model using Bayesian is smaller than OLS. The smaller value of RMSE, MAD and MAPE indicate better model. Because based on 3 criteria there are two criteria showed that the values of linear regression model with Bayesian approach are smaller than value of OLS method, then in this case better to use linear regression model with Bayesian approach. In addition, because the linear regression model with OLS does not meet the classical assumptions, it also makes the model not good enough.

6. Conclusion

The results of this research indicate that multiple linear regression modeling using OLS showed that the assumption does not occur multicollinearity are not met. While other assumptions related to the classical assumptions of IIDN (residual normally distributed, residual independent and residual identic) are not met. Although all parameters of the independent variables are significant in the model, the model is not good enough because the linear regression assumptions are not met. RMSE criteria of OLS is smaller than Bayesian, but since assumption is not fulfilled, it is better to apply Bayesian approach to regression model. Based on other criteria, MAD and MAPE showed that the criteria value of Bayesian regression model is less than OLS. It can be concluded that the linear regression model using Bayesian approach is better than OLS method.

References

1. Anderson DR, Sweeney DJ, Williams TA. Statistics for Business and Economics. 10th ed. USA: Thomson South-Western; 2008.
2. Azhar JA. Perbandingan Metode Bayes dan Metode Likelihood Dalam Mengestimasi Parameter Model Regresi Linier. Laporan Skripsi. Yogyakarta: UIN Sunan Kalijaga, Program Studi Matematika, Fakultas Sains dan Teknologi; 2012.
3. Iswari AAIA, Sumarjaya IW, Srinadi IGAM. Analisis Regresi Bayes Linier Sederhana dengan Prior Noninformatif. E-Jurnal Matematika. 2014; 3(2).
4. Chaturvedi A. Robust Bayesian Analysis of The Linear Regression Model. Journal of Statistical Planning and Inference. 1996; 50.
5. Abidin FP. Penanganan Multikolinieritas dengan Bayesian Ridge Regression pada Kasus Kemiskinan Kabupaten/Kota di JawaTimur. Jurnal Mahasiswa Statistika, Universitas Brawijaya. 2015; 3(1).
6. Mutiarani V, Setiawan A, Parhusip HA. Penerapan Model Regresi Linier Bayesian Untuk Mengestimasi Parameter dan Interval Kredibel. In Seminar Nasional Matematika dan Pendidikan Matematika FMIPA UNY; 2012; Yogyakarta.

7. West M. Outlier Models and Prior Distributions in Bayesian Linear Regression. Journal of the Royal Statistical Society Series B (Methodology). 1994; 46(3).
8. Tsanas A, Xifara A. Accurate Quantitative Estimation of Energy Performance of Residential Buildings Using Statistical Machine Learning Tools. Energy and Buildings. 2012; 49.
9. Gujarati DD, Porter DC. Basic Econometrics. 5th ed. New York: McGraw-Hill/Irwin; 2009.
10. Hauke J, Kossowski T. Comparison of Value of Pearson's and Spearman's Correlation Coefficient on The Sama Sets of Data. Quaestiones Geographicae. 2011; 30(2).
11. Terasvirta T, Lin CF, Granger CWJ. Power of The Neural Network Linearity Test. Journal of Time Series Analysis. 1993; 14(2).
12. Draper NRD, Smith H. Applied Regression Analysis. 3rd ed. Canada: John Willey & Sons, Inc; 1998.
13. Mutiarani V. Estimasi Parameter dan Interval Kredibel Dalam Model Regresi Linier Bayesian. Laporan Tugas Akhir. Salatiga: Universitas Kristen Satya Wacana, Program Studi Matematika, Fakultas Sains dan Matematika; 2013.
14. Rubio FJ, Genton MG. Bayesian Linear Regression With Skew-Symmetric Error Distributions With Applications to Survival Analysis. Statistics in Medicine. 2016; 35.
15. Xifara A. UCI Machine Learning Repository. [Online].; 2012 [cited 2018 June 6. Available from: https://archive.ics.uci.edu/ml/datasets/Energy+efficiency.
16. Martin AD, Quinn KM. Applied Bayesian Inference in R Using MCMCpack. R News. 2006; 6(1).
17. Martin AD, Quinn KM. MCMCpack: Markov Chain Monte Carlo in R. Journal of Statistical Software. 2011; 42(9).