MATEMATIKA: MJIAM, 2019, Volume 35, Number 2, 229–235

Multiple Linear Regression Model of Rice Production using Conjugate Gradient Methods

Nur Idalisa*, Mohd Rivaie, Nurul Hafawati Fadhilah, Nur Atikah, Anis Shahida and Nur Hidayah Mohd Noh

Jabatan Matematik dan Statistik, Fakulti Sains Komputer dan Matematik
Univesiti Teknologi Mara (UiTM) Cawangan Terengganu
Kampus Kuala Terengganu, Malaysia

*Corresponding author: nuridalisa@tganu.uitm.edu.my

Article history
Received: 2 October 2018
Received in revised form: 10 April 2019
Accepted: 7 May 2019
Published online: 1 August 2019

Abstract Regression is one of the basic relationship models in statistics. This paper focuses on the formation of regression models for the rice production in Malaysia by analysing the effects of paddy population, planted area, human population and domestic consumption. In this study, the data were collected from the year 1980 until 2014 from the website of the Department of Statistics Malaysia and Index Mundi. It is well known that the regression model can be solved using the least square method. Since least square problem is an unconstrained optimisation, the Conjugate Gradient (CG) was chosen to generate a solution for regression model and hence to obtain the coefficient value of independent variables. Results show that the CG methods could produce a good regression equation with acceptable Root Mean-Square Error (RMSE) value.

Keywords Conjugate Gradient; regression; rice production

Mathematics Subject Classification 62P12, 65F10

1 Introduction

The rapid growth in technology has increased the demands and expectations of powerful solvers in terms of time and memory. The main purpose of this paper is to solve least square problem particularly multiple linear regression equation. This paper focuses on rice as a case study because it is the largest and most widely cultivated food crop in Malaysia. This study is twofold. The objectives of this study are:

a) To examine the production of rice in Malaysia using least square method with four independent variables,

b) Ascertain the performance of CG with regards to its reliability and consistency as compared to the direct method to solve the least square problem.

The paper is organized as follows: the next section will provide a review of the related literature to highlight the need of an examination of rice production and also the use of CG. This is followed by a description of the method utilised to collect and analyse the data. In subsequent sections, the paper presents its argument on the performance consistency of CG and makes recommendations based on the conclusions of this study.

2 Conjugate Gradient Method

The Conjugate Gradient (CG) method was originally proposed by R. Hestenes and Eduard Stiefel in 1952 [1]. It is a method to solve a Symmetric-Positive-Definite (SPD) system of equations. The name 'Conjugate gradient' is derived from the fact that the successive search directions are conjugated with respect to the coefficient matrix of the symmetric positive definite system of equations. Conjugate Gradient (CG) method is famous due to its simplicity of formula and low memory requirement. The search direction d_k of CG method can be defined as

$$d_k = \begin{cases} -g_k & \text{if } k = 0, \\ -g_k + \beta_k d_{k-1} & \text{if } k \geq 1, \end{cases}$$

where g_k is the gradient of $f(x)$ and β_k is CG coefficients. Andrei [2] classified CG into five categories and the chosen β_k will affect the performance of method. Hence, in this paper, the well-known formula β_k chosen are Fletcher and Reeves (FR) (1964) [3] and Polak, Ribiere and Polyak (PRP) (1969) [4]. The FR method is the first CG algorithm for nonlinear functions while PRP method is one of the most efficient nonlinear CG methods. The formula of these methods as follows,

$$\beta_k^{FR} = \frac{g_k^T g_k}{\|g_{k-1}\|^2} \tag{1}$$

$$\beta_k^{PRP} = \frac{g_k^T (g_k - g_{k-1})}{\|g_{k-1}\|^2} \tag{2}$$

There are many versions of CG algorithm depending on what type of modifications or improvements were done by the researchers. The following is the CG algorithm that was used throughout this paper.

Algorithm 1 (Conjugate gradient method)

Step 1. Initialisation:

a. Input matrices A, b, and X. Choose the initial vector matrix X_0 (as an estimate of the solution), k=0.

b. Set $r = b$, $d = r$

c. Specify the convergence tolerance ε .

Step 2. If $\|r\| < \varepsilon$, stop. Otherwise go to Step 3.

Step 3. Compute

a. The CG coefficient β_k based on Equation (1) and Equation (2).

b. Set $d_k = r_k + \beta_k d_{k-1}$

c. Set $\alpha_k = \frac{r_k^T r}{d_k^T A d_k}$

Step 4. Updates

a. $X_{k+1} = X_k + \alpha_k d_k$

b. $r_{k+1} = r_k - \alpha_k A d_k$

Step 4. Stopping criteria

a. Check convergence and stopping criteria.

b. Go to Step 2 with $k = k + 1$ if necessary.

3 Conceptual Framework

Rice is the second most important crop in the world after wheat [5, 6], with Asia being the largest producer and consumer [7]. It constitutes one of the most important staple foods of over half of the world's population. More than half of the world's population depends on rice for food calories and protein, especially in developing countries. The world is currently witnessing rapid growth in rice consumption where the world will need about 760 million tons of paddy by the year 2025. The estimation is based on 35 percent more than the rice production in the year 1996 [8]. Another study suggest that the world's demand will be increasing up to 40% more rice by the year 2030 [9]. These estimates are based in view of continued population growth, a drastic reduction in growth of rice production during 1990s [8, 9] and economic prosperity. Its increase in demand will have to be met with less land, less water, less labour and fewer chemicals.

This paper aims to investigate the use of classical CG methods in rice production in Malaysia and to determine the performance consistency of CG when applied in a context other than its original intent. A data set was created with four independent variables namely paddy production, planted area, human population and domestic consumption that could be used as a factor that contributes to rice production in Malaysia. The data consists of thirty-five annual time series data sets from 1980 to 2014. The conceptual framework is as shown in Figure 1.

4 Material and Methods

Data Collection

The 35 annual data set was collected from Department of Statistics Malaysia and Index Mundi website from year 1980 to 2014 consisting of one dependent variable (rice production) and four independent variables (paddy production, planted area, human population and domestic consumption).

To ensure that the four independent variables that were chosen can predict the dependent variables, each independent variable is evaluated by multiple regression analysis in SPSS.

From Table 1, the significance value for paddy production, planted area, human population and domestic consumption are less than 0.5 which means that these four independent variables play an important role in determining rice production.

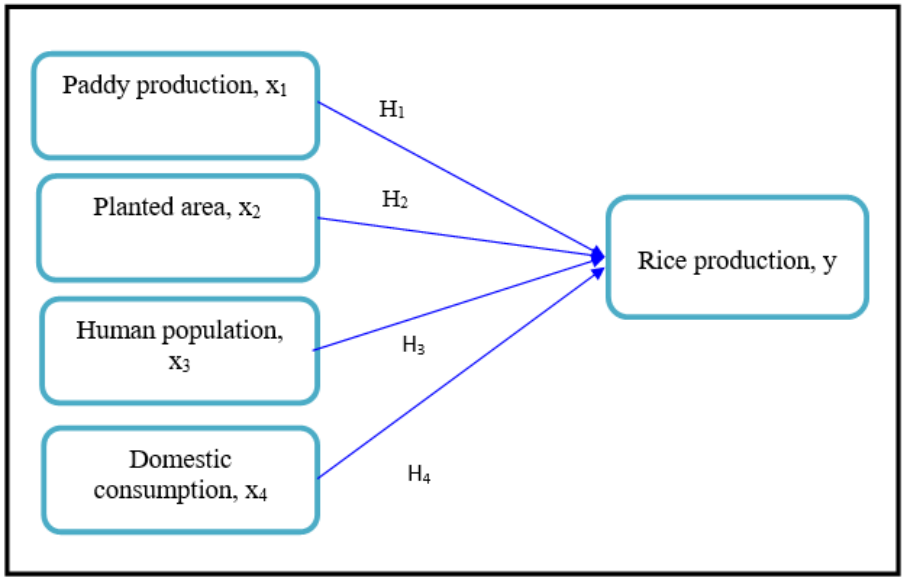

Figure 1: Conceptual Framework

Table 1: Beta Coefficients for Independent Variables

Independent Variable	Beta Coefficient	Significance Level
Paddy Production	0.994	0.000
Planted Area	0.005	0.039
Human Population	−0.016	0.016
Domestic Consumption	0.019	0.022

5 Result and discussion

The multiple linear regression analysis was carried out by using paddy productions (x_1), planted area (x_2), human population (x_3) and domestic consumption (x_4) data as independent variables and rice yield (y) data as the dependent variable. The multiple linear regression model is shown in Equation 3. These independent variables are hypothesised to influence the rice production.

$$y_i = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 \tag{3}$$

The algorithm to estimate the equation of the MLR line is called the ordinary least squares (OLS) estimation. OLS regression analysis assumes the relationships between independent and dependent variables are constant throughout the dataset and quantify this relationship with a coefficient [10]. In the least square method, the estimators are the values of parameters $X = (a_0, a_1, a_2, a_3, a_4)^T$ which minimise the objective function as follows:

$$S = \min \sum_{i=1}^{N} E_i^2 = \sum_{i=1}^{N} ((a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4))^2 \tag{4}$$

where N is the number of observed rice yield data. The concept from Calculus can be used by differentiating Equation 4 with respect to all the parameters involved to minimize the sum

of squares, S. At the minimum all the partial derivatives $\frac{\partial S}{\partial a_0}, \frac{\partial S}{\partial a_1}, \ldots, \frac{\partial S}{\partial a_4}$vanish. Writing the equations in matrix form will lead to the following system of linear equation.

$$\begin{bmatrix} N & \sum x_1 & \sum x_2 & \sum x_3 & \sum x_4 \\ \sum x_1 & \sum x_1^2 & \sum x_1x_2 & \sum x_1x_3 & \sum x_1x_4 \\ \sum x_2 & \sum x_2x_1 & \sum x_2^2 & \sum x_2x_3 & \sum x_2x_4 \\ \sum x_3 & \sum x_3x_1 & \sum x_3x_2 & \sum x_3^2 & \sum x_3x_4 \\ \sum x_4 & \sum x_4x_1 & \sum x_4x_2 & \sum x_4x_3 & \sum x_4^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum x_1y \\ \sum x_2y \\ \sum x_3y \\ \sum x_4y \end{bmatrix} \tag{5}$$

For ease of explanation, the coefficient matrix for the linear Equation 5 can be denoted as A and the constant vector as the letter b. Following are the matrices obtained using the data selected in this study:

$$A = \begin{bmatrix} 35 & 74333633 & 23704424 & 717527606 & 67323000 \\ 74333633 & 1.60635E+13 & 5.03944E+13 & 1.57473E+15 & 1.47159E+14 \\ 23704424 & 5.03944E+13 & 1.60635E+13 & 4.85852E+14 & 4.56017E+13 \\ 717527606 & 1.57473E+15 & 4.85852E+14 & 1.58705E+16 & 1.46556E+15 \\ 67323000 & 1.47159E+14 & 4.56017E+13 & 1.46556E+15 & 136349629 \end{bmatrix}$$

$$b = \begin{bmatrix} 47897000 \\ 1.03739E+14 \\ 3.24722E+13 \\ 1.01462E+15 \\ 9.4822E+13 \end{bmatrix}$$

By using the inversion of matrix method, the following solution is obtained

$$y = -1875676.356 - 0.002542584x_1 + 3.932028222x_2 + 0.028770676x_3 - 0.001716303x_4.$$

The CG method has been used as comparison with the aforementioned inversion of matrix method to solve ***least*** squares problems of Equation (5). The CG algorithm is explained in Algorithm 1. Two classical CG methods have been chosen for this study which are FR and PRP method. The MATLAB2017a software has been used to perform all calculations and data analyses in Intel®Core™i3-5005U CPU @ 2.00GHz RAM 6 GB PC environment.

In order to measure the model performance, the root mean square error (RMSE) and R-square (R^2) have been calculated as standard metric for model errors [11] [12]. The RMSE is computed by

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{\bmod l} - y_{obs}\right)^2}$$

where $y_{\bmod l}$ refers to the predicted model value and y_{obs} refers to observed rice yield value. While, R^2 is computed by

$$R^2 = \frac{\sum_{i=1}^{N}\left(y_{\bmod l} - \bar{y}\right)^2}{\sum_{i=1}^{N}\left(y_{obs} - \bar{y}\right)^2}$$

where $\bar{y}$ is mean of observed rice yield data. R^2 is a statistical measure of how close the data are to the fitted regression line. The R^2 is always between 0 and 1 which the higher the R^2, the better the model fits the observed data [13] [14].

Table 2 summarises the estimation of the model coefficients for OLS Regression using the inverse method and CG methods with RMSE values and R^2. The CPU time and number of iterations are also obtained for CG methods.

Table 2: Summary of Result

Method		Inverse method	Conjugate Gradient	
			FR	PRP
CPU Time (s)		0.004628	0.004527309	0.006152562
Number of Iteration		-	13	14
Parameters	a0	−1875676.356	−1658276.703	−1658276.702
	a1	−0.002542584	−0.001498206	−0.001498206
	a2	3.932028222	3.291499856	3.291499855
	a3	0.028770676	−0.00257371	−0.00257371
	a4	−0.001716303	−0.088784039	−0.08878403
RMSE ($\times 10^4$)		8.274928	10.508817	10.508817
R^2		0.182345293	0.845910483	0.845938056

CPU time is used to compare running time speed for each method. The iterations number refers to how many iterations is required to meet the stopping criteria. The stopping criteria for this paper are $\varepsilon = 10^{-12}$ as the convergence tolerance. For this research, CG has run 100 randomly initial input and the RMSE was recorded as well as the CPU time, number of iteration and the OLS parameter. The performance of the method can be observed through the influence of the least RMSE and CPU time, with the R square close to 1. From Table 2, it can be observed that CG methods are comparable with the direct method. The CG—PRP method is the best method that produced the closest R^2 value to 1. For all initial points, FR required 13 iterations and PRP required 14 iterations. Therefore, it can be concluded that the initial vector for CG does not affect the number of iterations.

6 Conclusion

The main result of this paper is that the conjugate gradient method converges to the solution of the least square problem to solve the multiple linear regression of rice production. From the analysis in this paper, the four constructs were found to be significant to the rice production in Malaysia. The reliability of the CG methods was found to be consistent with the result obtained from the direct method. This finding has implications on the theoretical foundations of CG and support the notion of a unified view of examining iterative technique adoption. Since there are variants of CG method, it is suggested to apply other CG method to improve the results for this study.

References

[1] Hestenes, M. R. and Stiefel, E. Methods of Conjugate Gradients for Solving Linear Systems. *J. Res. Nat. Bur. Stand.* 1952.49: 409-436.

[2] Andrei. N. Open Problems in Nonlinear Conjugate Gradient Algorithms for Unconstrained Optimization. *Bull. Malays. Math. Sci. Soc.* 2011. 34(2):319-330.

[3] Fletcher, R., & Reeves, C. M. Function minimization by conjugate gradients. *The Computer Journal*. 1964. 7(2): 149-154.

[4] Polak,E., Ribiere,G. and Polyak. Note sur la convergence de direction conjugees. *Rev. Francaise Inform. RecherceOperationelle*. 1969. 3: 35-43.

[5] Matthews, R. B., Kropff, M. J., Bachelet, D., & Van Laar, H. H.(Eds.). Modeling the impact of climate change on rice production in Asia. *Int. Rice Res. Inst.* 1995.

[6] Banik, M. Cold injury problems in Boro rice. In *Workshop on Modern Rice Cultivation in Bangladesh*. Bangladesh Rice Res. Inst. Joydebpur, Gazipur, Bangladesh. 1999, February. pp. 14-16.

[7] Gumma, M. K., Nelson, A., Thenkabail, P. S., & Singh, A. N. Mapping rice areas of South Asia using MODIS multitemporal data. *Journal of applied remote sensing*. 2011. 5(1): 053547.

[8] Brown, L. R. *Tough choices: facing the challenge of food scarcity*. WW Norton & Company. 1996.

[9] Brown, L. R., & Mitchell, J. D. *The Agricultural Link: How Environmental Deterioration Could Disrupt Economic Progress*. Washington, DC: Worldwatch Institute. 1997.pp 73.

[10] Deerfield A. Quantile Regression Analysis of Cooperative Learning Effects. *International Review of Economics Education*. 2018.

[11] McKeen, S., Wilczak, J., Grell, G., Djalalova, I., Peckham, S., Hsie, E. Y., ... & McHenry, J. Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. *Journal of Geophysical Research: Atmospheres*. 2005. 110(D21).

[12] Chai, T., Kim, H. C., Lee, P., Tong, D., Pan, L., Tang, Y., ... & Stajner, I. Evaluation of the United States National Air Quality Forecast Capability experimental real-time predictions in 2010 using Air Quality System ozone and NO 2 measurements. *Geoscientific Model Development*. 2013. 6(5): 1831-1850.

[13] Tranmer, M., & Elliot, M. Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR)*. 2008. 5: 30-35.

[14] Ghaedi, M., reza Rahimi, M., Ghaedi, A. M., Tyagi, I., Agarwal, S., & Gupta, V. K. Application of least squares support vector regression and linear multiple regression for modeling removal of methyl orange onto tin oxide nanoparticles loaded on activated carbon and activated carbon prepared from Pistacia atlantica wood. *Journal of Colloid and Interface Science*. 2011. 461: 425-434.