Available online at www.worldscientificnews.com

World Scientific News

An International Scientific Journal

WSN 140 (2020) 12-25 EISSN 2392-2192

Multiple Linear Regression Using Cholesky Decomposition

Ira Sumiati*, **Fiyan Handoyo** and **Sri Purwani**

Faculty Mathematics and Natural Science, Universitas Padjadjaran,
Jalan Raya Bandung-Sumedang Km. 21 Jatinangor Sumedang 45363, Indonesia

*E-mail address: irasumiati@gmail.com

ABSTRACT

Various real-world problem areas, such as engineering, physics, chemistry, biology, economics, social, and other problems can be modeled with mathematics to be more easily studied and done calculations. One mathematical model that is very well known and is often used to solve various problem areas in the real world is multiple linear regression. One of the stages of working on multiple linear regression models is the preparation of normal equations which is a system of linear equations using the least-squares method. If more independent variables are used, the more linear equations are obtained. So that other mathematical tools that can be used to simplify and help to solve the system of linear equations are matrices. Based on the properties and operations of the matrix, the linear equation system produces a symmetric covariance matrix. If the covariance matrix is also positive definite, then the Cholesky decomposition method can be used to solve the system of linear equations obtained through the least-squares method in multiple linear regression. Based on the background of the problem outlined, such that this paper aims to construct a multiple linear regression model using Cholesky decomposition. Then, the application is used in the numerical simulation and real case.

***Keywords*:** Multiple linear regression, covariance matrix, Cholesky decomposition

1. INTRODUCTION

Mathematical modeling is a field of mathematics that can represent and describe a situation or problem in the real world in the form of mathematical formulas or symbols so that

(Received 08 December 2019; Accepted 24 December 2019; Date of Publication 27 December 2019)

it is easier to learn and do calculations. One mathematical model that is very well-known and is often used to help solve various problem areas, such as engineering, physics, chemistry, biology, economics, social, and other real-world problems, is multiple linear regression. This model is used to measure the effect or linear relationship between two or more independent variables with a dependent variable. Zsuzsanna and Marian [1] used multiple regression to study performance indicators in the ceramics industry, with the dependent variable being the size of earnings, while the independent variable consisted of self-financing capacity, return on equity, level of technical capability, personnel costs per employee, and investment per person employed. The research aims to improve competitiveness, flexibility, adaptability, and the reactivity of companies in the ceramic industry.

Uyanik and & Guler [2] apply multiple linear regression analysis to measure the effect of student learning values (measurement and evaluation, educational psychology, program development, and guidance and counseling techniques) on KPSS exam scores (civil service selection). Desa et al. [3] use multiple regression to determine the effect of personality on work stress. Chen et al. [4] proposed Linear Regression based Projections (LRP) to minimize the ratio between local compactness information and total separation information to find the optimal projection matrix. Nurjannah et al. [5] analyzed multiple linear regression to test the determinants of hypertensive preventive behavior, with the dependent variable being hypertension prevention behavior, while the independent variables consisted of self-efficacy, knowledge, family support, gender, age, and support of health workers.

One of the stages of working on multiple linear regression models is the preparation of normal equations which is a system of linear equations using the least-squares method. If more independent variables are used, the more linear equations are obtained. So that other mathematical tools that can be used to simplify and help solve the system of linear equations are matrices. Based on the nature and operation of the matrix, the linear equation system produces a symmetric covariance matrix. If the covariance matrix is also positive definite, then the Cholesky decomposition method can be used to solve the system of linear equations (normal) obtained through the least-squares method in multiple linear regression. Cholesky Decomposition is a special version of LU decomposition that is designed to handle symmetric matrices more efficiently. Based on the background of the problem outlined, the purpose of this paper is to develop a multiple linear regression model using Cholesky decomposition. Literature completion of multiple linear regression analysis, specifically covariance matrix, using Cholesky decomposition can be seen in [6-12]. Then, the application is used in a case example.

Huang and Li [13] presented a new formulation of the Cholesky decomposition for the power spectral density (PSD) or evolutionary power spectral density (EPSD) matrix, then the application of the proposed scheme is used for Gaussian stochastic simulations. Wang and Ma [14] discussed the Cholesky decomposition of the Hermitian positive definite quaternion matrix. For the first time, the structure-preserving Gauss transformation is defined, and then a novel structure-preserving algorithm, which is applied to its real representation matrix. He and Xu [15] investigated the problem of estimating Cholesky decomposition in a normal independent conditional model with missing data. Explicit expressions for maximum likelihood estimators and unbiased estimators are derived. Madar [16] presented two novel and explicit parametrizations of the Cholesky factor from a nonsingular correlation matrix. One used the semi-partial correlation coefficient, and the second used the difference between successive ratios of two determinants. Feng et al. [17] proposed a modified Cholesky decomposition to model the structure of covariance in multivariate longitudinal data analysis. This decomposition entry has a simple structure and can be interpreted as a general moving average coefficient matrix and an innovation covariance matrix. Lee at al. [18] proposed a class of flexible, nonstationary, heteroscedastic models that exploits the structure allowed by combining the AR and MA modeling of the covariance matrix that we denote as ARMACD (autoregressive moving average using Cholesky decomposition), then applied it to the study of lung cancer to illustrate the power of the proposed method. Nino-Ruiz et al. [19] proposed the Kalman posterior filter ensemble (EnKF) based on the modified Cholesky decomposition. The main idea behind the approach is to estimate the analytical distribution moments based on the model realization ensemble. Okaze and Mochida [20] proposed a new method for producing turbulent fluctuations in wind velocity and scalars, such as temperature and contaminant concentrations, based on Cholesky decomposition of the time-averaged turbulent flux tensors of the momentum and the scalar for inflow boundary condition of large-eddy simulation (LES).

Kokkinos and Margaritis [21] identified the combination of matrix decomposition and cross-validation versions, then analyzed it theoretically and experimentally to find which is the fastest using Singular Value Decomposition (SVD), Eigen Value Decomposition (EVD), Cholesky Decomposition, and QR Decomposition, which produces reusable matrices (orthogonal, Eigen, singular, and upper triangle). Helmich-Paris et al. [22] introduced the Cholesky-decomposed density (CDD) matrix relativistic second-order Murber-Plesset energy disruption theory (MP2). Work equations are formulated in the usual MP2 intermediate form when using resolution-of-the-identity approximation (RI) for two-electron integrals. Naf et al. [23] proposed a complex dependency introduced by combining latent variables through the lower triangular matrix so that each component is the sum of a generalized independent hyperbolic (GHyp) random variables. This is done through the Cholesky decomposition of the dispersion matrix, which depends on latent random vectors. Nino-Ruiz et al. [24] discussed the efficient parallel implementation of the Kalman ensemble filter based on modified Cholesky decomposition. The proposed implementation started with decomposing the domain into sub-domains. In each sub-domain, a thin estimate of the inverse background error covariance matrix is calculated through modified Cholesky decomposition; estimates are calculated simultaneously on separate processors. Furthermore, the systematic writing in this paper includes: Section 2 discusses the basic theory of matrix, Section 3 describes the research methods used, Section 4 presents the results and discussion of applying Cholesky decomposition to multiple linear regression, and Section 5 presents the conclusion.

2. MATRIX

A matrix is an arrangement of numbers or symbols which are located in rows and columns so that they form a square shape. Matrices are generally denoted by capital letters and the elements are located in square brackets:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

If A is any matrix $m \times n$, then the transpose of A is denoted by A^T and is defined by the matrix $n \times m$ obtained by exchanging rows and columns from A, so the first column of A^T is the first row of A, the second column of A^T is the second row of A, and so on, as follows:

$$A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nm} \end{bmatrix}.$$

Suppose matrix A with size $n \times n$ is said to be symmetric if $A^T = A$. For example a_{ij} is the ij element of matrix A, then for the symmetric matrix $a_{ij} = a_{ji}$ applies to every i and j.

The matrix A of size $n \times n$ is said to be positive definite if for any vector $x \neq 0$, quadratic is $x^T Ax > 0$. Whereas it is said to be semi-definite if $x^T Ax \geq 0$.

3. METHODS

Broadly speaking, in this study, the method used is multiple linear regression and Cholesky decomposition. The literature on the analysis of multiple linear regression and Cholesky decomposition can be seen in Rawlings et al. (1998), Sarstedt and Mooi (2014), Darlington and Hayes (2017), Thomas (2017), Kumari and Yadav (2018), Schmidt and Finan (2018), Aster (2019).

3. 1. Multiple Linear Regression

Multiple linear regression analysis is an analysis that measures the effect/relationship linearly between two or more independent variables (X_1, X_2, ..., X_p) with the dependent variable (Y). Multiple linear regression models can be presented in the form of general equations as follows:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_p X_{pj} + \varepsilon_j \qquad (1)$$

with $j = 1, 2, \ldots, n$. Equation (1) can be written in matrix notation as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{13} & X_{2n} & \cdots & X_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad (2)$$

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times(p+1)} \boldsymbol{\beta}_{(p+1)\times 1} + \boldsymbol{\varepsilon}_{n\times 1},$$

where **Y** is the independent variable column vector, **X** is the independent variable matrix, **β** is the regression coefficient estimator column vector, and **ε** is the residual/error column vector.

Using the least-squares method, regression coefficients are obtained by minimizing the residual squares, so that the normal equation is obtained as follows:

$$\begin{aligned} n\beta_0 + \beta_1 \sum X_{1j} + \beta_2 \sum X_{2j} + \cdots + \beta_p \sum X_{pj} &= \sum Y_j, \\ \beta_0 \sum X_{1j} + \beta_1 \sum X_{1j}^2 + \beta_2 \sum X_{1j} X_{2j} + \cdots + \beta_p \sum X_{1j} X_{pj} &= \sum X_{1j} Y_j, \\ \beta_0 \sum X_{2j} + \beta_1 \sum X_{1j} X_{2j} + \beta_2 \sum X_{2j}^2 + \cdots + \beta_p \sum X_{2j} X_{pj} &= \sum X_{2j} Y_j, \\ &\vdots \\ \beta_0 \sum X_{pj} + \beta_1 \sum X_{1j} X_{pj} + \beta_2 \sum X_{2j} X_{pj} + \cdots + \beta_p \sum X_{pj}^2 &= \sum X_{pj} Y_j. \end{aligned} \tag{3}$$

Equation (3) can be written in matrix notation as follows:

$$\begin{bmatrix} n & \sum X_{1j} & \sum X_{2j} & \cdots & \sum X_{pj} \\ \sum X_{1j} & \sum X_{1j}^2 & \sum X_{1j} X_{2j} & \cdots & \sum X_{1j} X_{pj} \\ \sum X_{2j} & \sum X_{1j} X_{2j} & \sum X_{2j}^2 & \cdots & \sum X_{2j} X_{pj} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum X_{pj} & \sum X_{1j} X_{pj} & \sum X_{2j} X_{pj} & \cdots & \sum X_{pj}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \sum Y_j \\ \sum X_{1j} Y_j \\ \sum X_{2j} Y_j \\ \vdots \\ \sum X_{pj} Y_j \end{bmatrix} \tag{4}$$

$$\mathbf{\Sigma X} \cdot \boldsymbol{\beta} = \mathbf{\Sigma Y}.$$

The **ΣX** matrix is called the covariance matrix.

3. 2. Cholesky Decomposition

Cholesky Decomposition is a special version of LU decomposition that is designed to handle symmetric matrices more efficiently. For example, *A* is a definite symmetric and positive matrix, $a_{ij} = a_{ji}$, then *A* can be written

$$A = LL^T \tag{5}$$

where *L* is the bottom triangle matrix which is defined as follows:

$$L = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix}. \tag{6}$$

Using the Cholesky decomposition, elements of *L* are valued as follows:

$$l_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2} \text{ dan } l_{ki} = \frac{1}{l_{ii}} \sqrt{a_{ki} - \sum_{j=1}^{i-1} l_{ij} l_{kj}} \tag{7}$$

where the first subscript is the row index and the second is the column index, with $k = 1, 2, \ldots, n$ and $i = 1, 2, ..., k - 1$.

Steps to solve $Ab = c$, where A is symmetric and positive definite, using Cholesky decomposition is given as follows: decomposition A becomes $A = LL^T$, then solution b is obtained by: (a) forward substitution: solution d uses $Ld = c$, then (b) back substitution: solution b uses $L^T b = d$.

In this study, Cholesky decomposition was used to solve $\mathbf{\Sigma X} \cdot \boldsymbol{\beta} = \mathbf{\Sigma Y}$.

4. RESULTS AND DISCUSSION

Case examples through numerical simulations used in this study are looking for the influence of five independent variables (X_1, X_2, X_3, X_4, X_5) on the dependent variable (Y) with data of 30 samples, as shown in the Table 1.

Table 1. Simulation Data.

No.	X_1	X_2	X_3	X_4	X_5	Y
1.	301	36	1043	26	12	20
2.	303	75	1052	31	27	16
3.	338	68	1031	28	25	19
4.	442	25	1043	19	35	16
5.	340	34	1177	16	4	21
6.	391	5	1079	18	36	22
7.	334	6	1145	17	0	22
8.	415	7	1183	15	10	26
9.	428	25	1026	25	10	21
10.	302	35	1091	26	35	29
11.	304	55	1076	21	42	29
12.	398	54	1048	14	26	24
13.	326	59	1010	39	37	24

14.	323	42	1050	29	14	23
15.	421	1	1008	18	34	20
16.	443	97	1060	20	15	24
17.	403	2	1077	36	43	21
18.	308	13	1115	21	14	27
19.	444	95	1003	21	5	15
20.	440	38	1136	30	47	28
21.	337	54	1137	39	38	21
22.	443	33	1137	14	50	18
23.	427	28	1067	33	11	26
24.	355	43	1019	20	14	26
25.	378	23	1004	13	16	28
26.	406	71	1020	27	2	19
27.	445	16	1000	18	25	15
28.	430	44	1030	31	34	29
29.	321	3	1067	35	44	23
30.	350	17	1174	20	30	17

Using the least squares method in the Table 1, the covariance matrix equation is obtained $\mathbf{\Sigma X} \cdot \boldsymbol{\beta} = \mathbf{\Sigma Y}$ as follows:

$$\begin{bmatrix} 30 & 11296 & 1104 & 32108 & 720 & 735 \\ 11296 & 4335950 & 415914 & 12074579 & 267976 & 277201 \\ 1104 & 415914 & 61222 & 1168025 & 27549 & 25056 \\ 32108 & 12074579 & 1168025 & 34454876 & 768876 & 788388 \\ 720 & 267976 & 27549 & 768876 & 18992 & 18631 \\ 735 & 277201 & 25056 & 788388 & 18631 & 24263 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 669 \\ 250579 \\ 24234 \\ 716918 \\ 16133 \\ 16549 \end{bmatrix}.$$

Using the Cholesky decomposition method on the covariance matrix $\mathbf{\Sigma X}$ the following triangular matrix L is obtained:

$$\begin{bmatrix} 5.4772 & 0 & 0 & 0 & 0 & 0 \\ 2062.3580 & 287.4534 & 0 & 0 & 0 & 0 \\ 201.5619 & 0.7695 & 143.5068 & 0 & 0 & 0 \\ 5862.0919 & -52.7155 & -94.1337 & 281.2717 & 0 & 0 \\ 131.4534 & -10.8817 & 7.3959 & -5.6650 & 38.8174 & 0 \\ 134.1920 & 1.5619 & -13.8892 & 1.8377 & 28.8821 & 72.2674 \end{bmatrix}.$$

Applying steps to solve $\mathbf{\Sigma X} \cdot \boldsymbol{\beta} = \mathbf{\Sigma Y}$ with the Cholesky decomposition method the $\boldsymbol{\beta}$ solution is obtained as follows:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 21.1016 \\ -0.0140 \\ -0.0145 \\ 0.0056 \\ 0.0243 \\ 0.0162 \end{bmatrix}.$$

So that the regression model for the simulation data is obtained as follows:

$$\hat{Y} = 21.1016 - 0.0140X_1 - 0.0145X_2 + 0.0056X_3 + 0.0243X_4 + 0.0162X_5.$$

The regression model above represents that the independent variables X_3, X_4 and X_5 give a positive influence on the dependent variable Y, while the independent variables X_1 and X_2 give a negative influence. That is, if the independent variables X_3, X_4 and X_5 increase, the dependent variable Y increases, conversely if the independent variables X_1 and X_2 increase, the dependent variable Y decreases.

After applying multiple linear regression using Cholesky decomposition in simulation data, the method proposed is applied to the real case, namely Denver neighborhoods, where X_1 is the percentage of population change over the last few years, X_2 is the percentage of children (under 18 years) in the population, X_3 is the percentage of free school lunch participation, X_4 is the percentage change in household income over the past few years, X_5 is the crime rate (per 1000 population), and Y is the total population (in thousands).

The data of the independent and dependent variables of Denver neighborhoods are presented in Table 2.

Table 2. Denver Neighborhoods Data.

No.	X_1	X_2	X_3	X_4	X_5	Y
1.	1.8	30.2	58.3	27.3	84.9	6.9
2.	28.5	38.8	87.5	39.8	172.6	8.4
3.	7.8	31.7	83.5	26	154.2	5.7
4.	2.3	24.2	14.2	29.4	35.2	7.4
5.	-0.7	28.1	46.7	26.6	69.2	8.5
6.	7.2	10.4	57.9	26.2	111	13.8
7.	32.2	7.5	73.8	50.5	704.1	1.7
8.	7.4	30	61.3	26.4	69.9	3.6
9.	10.2	12.1	41	11.7	65.4	8.2
10.	10.5	13.6	17.4	14.7	132.1	5
11.	0.3	18.3	34.4	24.2	179.9	2.1
12.	8.1	21.3	64.9	21.7	139.9	4.2
13.	2	33.1	82	26.3	108.7	3.9
14.	10.8	38.3	83.3	32.6	123.2	4.1
15.	1.9	36.9	61.8	21.6	104.7	4.2
16.	-1.5	22.4	22.2	33.5	61.5	9.4
17.	-0.3	19.6	8.6	27	68.2	3.6
18.	5.5	29.1	62.8	32.2	96.9	7.6
19.	4.8	32.8	86.2	16	258	8.5
20.	2.3	26.5	18.7	23.7	32	7.5
21.	17.3	41.5	78.6	23.5	127	4.1
22.	68.6	39	14.6	38.2	27.1	4.6

23.	3	20.2	41.4	27.6	70.7	7.2
24.	7.1	20.4	13.9	22.5	38.3	13.4
25.	1.4	29.8	43.7	29.4	54	10.3
26.	4.6	36	78.2	29.9	101.5	9.4
27.	-3.3	37.6	88.5	27.5	185.9	2.5
28.	-0.5	31.8	57.2	27.2	61.2	10.3
29.	22.3	28.6	5.7	31.3	38.6	7.5
30.	6.2	39.7	55.8	28.7	52.6	18.7
31.	-2	23.8	29	29.3	62.6	5.1
32.	19.6	12.3	77.3	32	207.7	3.7
33.	3	31.1	51.7	26.2	42.4	10.3
34.	19.2	32.9	68.1	25.2	105.2	7.3
35.	7	22.1	41.2	21.4	68.6	4.2
36.	5.4	27.1	60	23.5	157.3	2.1
37.	2.8	20.3	29.8	24.1	58.5	2.5
38.	8.5	30	66.4	26	63.1	8.1
39.	-1.9	15.9	39.9	38.5	86.4	10.3
40.	2.8	36.4	72.3	26	77.5	10.5
41.	2	24.2	19.5	28.3	63.5	5.8
42.	2.9	20.7	6.6	25.8	68.9	6.9
43.	4.9	34.9	82.4	18.4	102.8	9.3
44.	2.6	38.7	78.2	18.4	86.6	11.4

Source: The Piton Foundation, Denver, Colorado. (https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html, accessed on December 21, 2019 at 7.31)

Using the least squares method in the Table 2, the covariance matrix equation is obtained $\mathbf{\Sigma X} \cdot \boldsymbol{\beta} = \mathbf{\Sigma Y}$ as follows:

$$\begin{bmatrix} 44 & 344.6 & 1199.9 & 2266.5 & 1186.3 & 4779.6 \\ 344.6 & 9104.04 & 9848.1 & 17710.42 & 10783.03 & 51872.68 \\ 1199.9 & 9848.1 & 36097.31 & 66192.05 & 32240.44 & 119070.74 \\ 2266.5 & 17710.42 & 66192.05 & 145561.27 & 61246.76 & 294756.18 \\ 1186.3 & 10783.03 & 32240.44 & 61246.76 & 33957.43 & 141332.5 \\ 4779.6 & 51872.68 & 119070.74 & 294756.18 & 141332.5 & 993000.66 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 309.8 \\ 2100.91 \\ 8671.13 \\ 15856.25 \\ 8266.76 \\ 27872.73 \end{bmatrix}.$$

Using the Cholesky decomposition method on the covariance matrix $\mathbf{\Sigma X}$ the following triangular matrix L is obtained:

$$\begin{bmatrix} 6.6332 & 0 & 0 & 0 & 0 & 0 \\ 51.9504 & 80.0324 & 0 & 0 & 0 & 0 \\ 180.8917 & 5.6314 & 57.8254 & 0 & 0 & 0 \\ 341.6877 & -0.5047 & 75.8560 & 151.8432 & 0 & 0 \\ 178.8414 & 18.6442 & -3.7266 & 2.8379 & 40.0451 & 0 \\ 720.5518 & 180.4233 & -212.4874 & 426.5056 & 177.3455 & 427.4854 \end{bmatrix}.$$

Applying steps to solve $\mathbf{\Sigma X} \cdot \boldsymbol{\beta} = \mathbf{\Sigma Y}$ with the Cholesky decomposition method the $\boldsymbol{\beta}$ solution is obtained as follows:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 5.9685 \\ -0.0313 \\ -0.0161 \\ 0.0262 \\ 0.0814 \\ -0.0164 \end{bmatrix}.$$

So that the regression model for Denver neighborhoods is obtained as follows

$$\hat{Y} = 5.9685 - 0.0313X_1 - 0.0161X_2 + 0.0262X_3 + 0.0814X_4 - 0.0164X_5 .$$

The regression model above represents that the percentage of free school lunch participation and the percentage change in household income over the past few years give a positive influence on the total population in Denver, while the percentage of population change over the past few years, the percentage of children (under 18 years) in population, and crime rates give a negative influence on the total population in Denver.

5. CONCLUSIONS

This paper aims to develop a multiple linear regression model using Cholesky decomposition. One of the stages of working on multiple linear regression models is the preparation of normal equations which is a system of linear equations using the least-squares method. If more independent variables are used, the more linear equations are obtained. So that other mathematical tools that can be used to simplify and help to solve the system of linear equations are matrices. Based on the properties and operations of the matrix, the linear equation system produces a symmetric covariance matrix. If the covariance matrix is also positive definite, then the Cholesky decomposition method can be used to solve the system of linear equations obtained through the least-squares method in multiple linear regression. The application to numerical simulation and real case are discussed in this study support the statement that the Cholesky decomposition is a special version of LU decomposition that is designed to handle symmetric matrices more efficiently. This study only looks for multiple linear regression models using Cholesky decomposition. Regarding the classical assumption test, significance test, and evaluation on the regression model are not discussed in this study.

References

[1] Zsuzsanna, T. & Marian, L. 2012. Multiple regression analysis of performance indicators in the ceramic industry. *Procedia Economics and Finance*, vol. 3, pp. 509-514.

[2] Uyanik, G. K. & Guler, N. 2013. A study on multiple linear regression analysis. *Procedia Social and Behavioral Sciences*, vol. 106, pp. 234-240.

[3] Desa, A., Yusoof, F., Ibrahim, N., Kadir, N. B. A. & Rahman, R. M. A. 2014. A study of the Relationship and Influence of Personality on Job Stress among Academic Administrators at a University. *Procedia-Social and Behavioral Sciences*, vol. 114, pp. 355-359.

[4] Chen, S., Ding, C. H. Q. & Luo, B. 2018. Linear regression based projections for dimensionality reduction. *Information Sciences*, vol. 467, pp. 74-86.

[5] Nurjannah, Rahardjo, S. S. & Sanusi, R. 2019. Linear Regression Analysis on the Determinants of Hypertension Prevention Behavior. *Journal of Health Promotion and Behavior*, vol. 4, no. 1, pp. 22-31.

[6] Sun, X. & Sun, D. 2005. Estimation of the Cholesky decomposition of the covariance matrix for a conditional independent normal model. *Statistics & Probability Letters*, vol. 73, pp. 1-12.

[7] Wang, J. & Liu, C. 2006. "Generating Multivariate Mixture of Normal Distributions Using a Modified Cholesky Decomposition. *Proceedings of the Winter Simulation Conference*, pp. 342-347.

[8] Pourahmadi, M., Daniels, M. J. & Park, T. 2007. Simultaneous modelling of the Cholesky decomposition of several covariance matrices. *Journal of Multivariate Analysis*, vol. 98, pp. 568-587.

[9] Alaba. O. O., Olubuseyo, O. E. & Oyebisi, O. 2013. Cholesky Decomposition of Variance-Covariance Matrix Effect on the Estimators of Seemingly Unrelated Regression Model. *Journal of Science Research*, vol. 12, pp. 371-380.

[10] Chen, Z. & Leng, C. 2015. Local Linear Estimation of Covariance Matrices Via Cholesky Decomposition. *Statistica Sinica*, vol. 25, pp. 1249-1263.

[11] Younis, G. 2015. "Practical Method to Solve Large Least Squares Problems Using Cholesky Decomposition." *Geodesy and Cartography*, vol. 41, no. 3, pp. 113-118.

[12] Babaie-Kafaki, S. & Roozbeh, M. 2017. A revised Cholesky decomposition to combat multicollinearity in multiple regression models. *Journal of Statistical Computation and Simulation*. Volume 87, Issue 12, Pages 2298-2308. doi:10.1080/00949655.2017.1328599

[13] Huang, G., Liao, H. & Li, M. 2013. New formulation of Cholesky decomposition and applications in stochastic simulation. *Probabilistic Engineering Mechanics*, vol. 34, pp. 40-47.

[14] Wang, M. & Ma, W. 2013. A structure-preserving algorithm for the quaternion Cholesky decomposition. *Applied Mathematics and Computation*, vol. 223, pp. 354-361.

[15] He, D. & Xu, K. 2014. Estimation of the Cholesky decomposition in a conditional independent normal model with missing data. *Statistics and Probability Letters*, vol. 88, pp. 27–39.

[16] Madar, V. 2015. Direct formulation to Cholesky decomposition of a general nonsingular correlation matrix. *Statistics and Probability Letters*, vol. 103, pp. 142-147.

[17] Feng, S., Lian, H. & Xue, L. 2016. A new nested Cholesky decomposition and estimation for the covariance matrix of bivariate longitudinal data." *Computational Statistics and Data Analysis*, vol. 102, pp. 98-109.

[18] Lee, L., Baek, C. & Daniels, M. J. 2017. ARMA Cholesky factor models for the covariance matrix of linear models. *Computational Statistics and Data Analysis*, vol. 115, pp. 267-280.

[19] Nino-Ruiz, E. D., Mancilla, A. & Calabria, J. C. 2017. A Posterior Ensemble Kalman Filter Based On A Modified Cholesky Decomposition. *Procedia Computer Science*, vol. 108C, pp. 2049-2058.

[20] Okaze, T. & Mochida, A. 2017. Cholesky decomposition–based generation of artificial inflow turbulence including scalar fluctuation. *Computers and Fluids*, vol. 159, pp. 23-32.

[21] Kokkinos, Y. & Margaritis, K. G. 2018. Managing the computational cost of model selection and cross-validation in extreme learning machines via Cholesky, SVD, QR and eigen decompositions. *Neurocomputing*, vol. 295, pp. 29-45.

[22] Helmich-Paris, B., Repisky, M. & Visscher, L. 2019. Relativistic Cholesky-decomposed density matrix MP2. *Chemical Physics*, vol. 518, pp. 38-46.

[23] Naf, J., Paolella, M. S. & Polak, P. 2019. Heterogeneous tail generalized COMFORT modeling via Cholesky decomposition. *Journal of Multivariate Analysis*, vol. 172, pp. 84-106.

[24] Nino-Ruiz, E. D., Sandu, A. & Deng, X. 2019. A parallel implementation of the ensemble Kalman filter based on modified Cholesky decomposition. *Journal of Computational Science*, vol. 36, pp. 100654 (1-8).

[25] Rawlings, J. O., Pantula, S. G. & Dickey, D. A. 1998. Applied Regression Analysis: A Research Tool, Second Edition. New York: Springer-Verlag.

[26] Sarstedt, M. & Mooi, E. 2014. A Concise Guide to Market Research. Heidelberg: Springer-Verlag. doi: 10.1007/978-3-642-53965-7_7

[27] Darlington, R. B. & Hayes, A. F. 2017. Regression Analysis and Linear Models: Concepts, Applications, and Implementation. New York: Guilford Press.

[28] Thomas McSweeney, 2017. Modified Cholesky Decomposition and Applications. A dissertation submitted to the University of Manchester for the degree of Master of Science in the Faculty of Science and Engineering. Manchester Institute for Mathematical Sciences.

[29] Kumari, K. & Yadav, S. 2018. Linear regression analysis study. *J Pract Cardiovasc Sci* vol. 4, pp. 33-36.

[30] Schmidt, A. F. & Finan, C. 2018. Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, vol. 98, pp. 146-151.

[31] Richard C. Aster, 2019. Parameter Estimation and Inverse Problems (Chapter Two: Linear Regression). Elsevier. doi:10.1016/B978-0-12-804651-7.00007-9