# Predict Global Earth Temperature using Linier Regression

Edwin Swandi Sijabat (23516012)
Program Studi Magister Informatika
Sekolah Teknik Elektro dan Informatika ITB
Jl. Ganesha 10 Bandung 40132, Indonesia
23516012@std.stei.itb.ac.id

*Abstract*— **Global warming is global issue of the world nowadays. Every year global earth temperature tends to increase and data about it recorded continuously. This data can be used to get more valuable information. Predicting is one point that we can do. This paper discuss about how data about global earth temperature can be used to approximate the data that not recorded or how the data can be used to predict about the next global earth temperature. This can be done by implementing linier regression.**

***Keywords—component; global earth temperature, prediction, linier regression.***

## I. Introduction

Global warming is one issue that needs to be considered because this issue can affect world society. This issue can be a very large disaster because it can affect system in earth. Increasing of global earth temperature is the result of the development of industry. The use of machine that has danger waste is growing fast. Pollution levels also increase. This issue affects system in earth.

Every year, NASA analyzes this issue and recodes the information. Observation result shows that temperature in our earth tends increasing. NASA has recorded data from 1880 to 2016. The recorded information shows that 2016 is year when earth surface has highest average temperature. They predict, global earth temperature will be highest in the next year.

To support the prediction, we need to develop a method that can predict global earth temperature using recorded data. We can use a mathematical approach. We can develop prediction method by using linier regression. By using this method, we try to look correlation between year and global earth temperature. We can predict by getting approximate value before doing direct observation. We also can approximate missing value of data recorded.

## II. Methodology

### A. Dataset

One of the most important in this paper is dataset. From dataset, we can get more information. In this case, we use global earth temperature data. Data is taken from NASA website. Data is collected from institute that called NASA's Goddard Institute for Space Studies (GISS). Data contain information about global earth temperature from 2000 to 2016.

TABLE I. Global Earth Temperatur Dataset

| No | Year | Temperature (ºC) |
|---|---|---|
| 1 | 2000 | 0.42 |
| 2 | 2001 | 0.55 |
| 3 | 2002 | 0.63 |
| 4 | 2003 | 0.62 |
| 5 | 2004 | 0.55 |
| 6 | 2005 | 0.69 |
| 7 | 2006 | 0.63 |
| 8 | 2007 | 0.66 |
| 9 | 2008 | 0.54 |
| 10 | 2009 | 0.64 |
| 11 | 2010 | 0.71 |
| 12 | 2011 | 0.6 |
| 13 | 2012 | 0.63 |
| 14 | 2013 | 0.65 |
| 15 | 2014 | 0.74 |
| 16 | 2015 | 0.87 |
| 17 | 2016 | 0.99 |

a. NASA's Goddard Institute for Space Studies (GISS)

### B. Regression

Regression is method to get functional relationship between a variable and the other variable [1]. In this case we want to know global earth temperature in a year. By using regression, we also can match a curve that has low accuracy data. Low accuracy data can be obtained from observation, laboratory experiment, and statistic data.

There are two kinds of variable, response or dependent variable and explanatory or predictor variable. Response variable is variable that contain value we want to know. Predictor variable is variable that will be used to find the response value.

### C. Simple Linier Regression

Simple linier regression is linier regression that has one predictor. This method is suitable for use because dataset is a statistical data that has low accuracy [2]. We only using variable year to get approximate global earth temperature. In linier regression, there is a linier line through a set of data. This linier function is using to predict value of $y$ when $x$ is known.

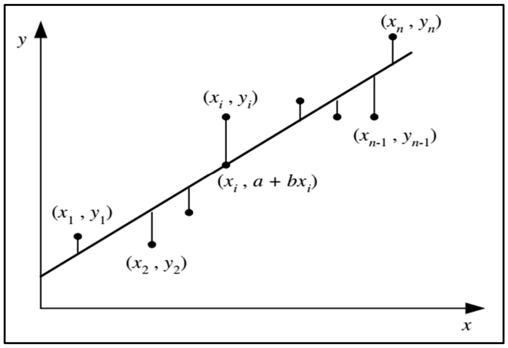

Fig. 1. Linier Regression Curve

In Fig 1, there are several nodes that we have known value of $x$ and $y$ respectively, except $y_i$. We just know value of $x_i$. By using linier regression, we can approximate value of $y_i$ using value is given by linier function.

We can make a linier equation that gives approximate value for $y$. The true value is contain error, so equation must be written with

$$g(x_i) = f(x) + e_i \quad i = 1,2,3,\dots,n \tag{1}$$

From equation (1), $g(x_i)$ is true value and $e_i$ is error every node. In linier regression, we form linier equation as

$$f(x) = a + bx \tag{2}$$

As explain before, $f(x)$ only gives approximate value. There is deviation value between true value and approximate value. Deviation is calculated by equation

$$r_i = y_i - f(x_i) = y_i - (a + bx) \tag{3}$$

This is equation for square of total deviation.

$$R = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{4}$$

To get a minimum of $R$, we must make differential of $R$.

$$\frac{\partial R}{\partial a} = -2\sum(y_i - a - bx_i) = 0 \tag{5}$$

$$\frac{\partial R}{\partial a} = -2\sum x_i(y_i - a - bx_i) = 0 \tag{6}$$

Each of differentials from equation (5) and equation (6) divided by -2

$$\sum(y_i - a - x_i)^2 = 0 \Leftrightarrow \sum y_i + \sum a + \sum bx_i = 0$$

$$\sum x_i(y_i - a - x_i)^2 = 0 \Leftrightarrow \sum x_i y_i + \sum ax_i + \sum bx_i^2 = 0$$

The equation is simplified again

$$\sum a + \sum bx_i = \sum y_i \tag{7}$$

$$\sum ax_i + \sum bx_i^2 = \sum x_i y_i \tag{8}$$

Or

$$na + b\sum x_i = \sum y_i \tag{9}$$

$$a\sum x_i + b\sum x_i^2 = \sum x_i y_i \tag{10}$$

These equations are linier equation and we can write it in matrix form. We can find value of $a$ and $b$ by solving linier equation system [3].

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \tag{11}$$

Value of $a$ and $b$ also can be computer by equation

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i - (\sum x_i)^2} \tag{12}$$

$$a = \bar{y} - b\bar{x} \tag{13}$$

To know how good our approximate function is, we can measure RMS error (root-mean-square error).

$$E_{RMS} = \left(\frac{1}{n}\sum_{i=1}^{n} |f(x_i) - y_i|^2\right)^2 \tag{14}$$

The smaller the $E_{RMS}$ the better approximate value of the function.

## III. Experiment

In this section we want to implement liner regression to get information of global earth temperature in a year.

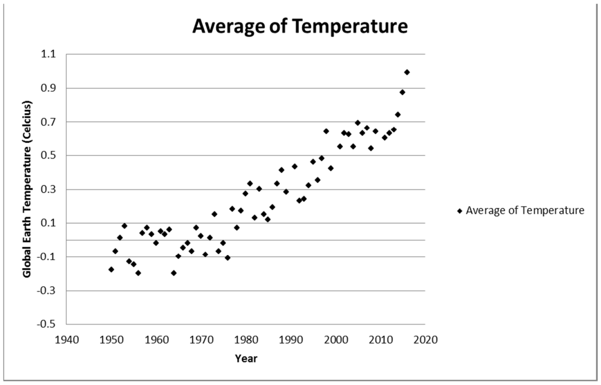

Fig. 2. Global earth temperature data

Distribution data in Cartesian diagram is shown by Fig.1. This diagram shows us that nodes form a line of set of data. This makes us assume that global earth temperature can be predicted by using linier regression.

For starting the experiment, the dataset is divided in to two. 15 rows of them to be data training, and other 2 rows are used as data testing. Data training is used to makes model or function and data test is used to test function.

TABLE II. DATA TRAINING

| No | Year | Temperature (ºC) |
|---|---|---|
| 1 | 2000 | 0.42 |
| 2 | 2001 | 0.55 |
| 3 | 2002 | 0.63 |
| 4 | 2003 | 0.62 |
| 5 | 2004 | 0.55 |
| 6 | 2005 | 0.69 |
| 7 | 2006 | 0.63 |
| 8 | 2007 | 0.66 |
| 9 | 2008 | 0.54 |
| 10 | 2009 | 0.64 |
| 11 | 2010 | 0.71 |
| 12 | 2012 | 0.63 |
| 13 | 2014 | 0.74 |
| 14 | 2015 | 0.87 |
| 15 | 2016 | 0.99 |

TABLE III. DATA TESTING

| No | Year | Temperature (ºC) |
|---|---|---|
| 1 | 2011 | 0.6 |
| 2 | 2013 | 0.65 |

By this experiment we want to make approximate function from data training and try to predict data value of data testing. First, make table that contains $x_i$, $y_i$, $x_i^2$, and $x_iy_i$.

TABLE IV. DATA FOR NORMAL EQUATION

| i | $x_i$ | $y_i$ | $x_i^2$ | $x_iy_i$ |
|---|---|---|---|---|
| 1 | 2000 | 0.42 | 4000000 | 840.00 |
| 2 | 2001 | 0.55 | 4004001 | 1100.55 |
| 3 | 2002 | 0.63 | 4008004 | 1261.26 |
| 4 | 2003 | 0.62 | 4012009 | 1241.86 |
| 5 | 2004 | 0.55 | 4016016 | 1102.20 |
| 6 | 2005 | 0.69 | 4020025 | 1383.45 |
| 7 | 2006 | 0.63 | 4024036 | 1263.78 |
| 8 | 2007 | 0.66 | 4028049 | 1324.62 |
| 9 | 2008 | 0.54 | 4032064 | 1084.32 |
| 10 | 2009 | 0.64 | 4036081 | 1285.76 |
| 11 | 2010 | 0.71 | 4040100 | 1427.10 |
| 12 | 2012 | 0.63 | 4048144 | 1267.56 |
| 13 | 2014 | 0.74 | 4056196 | 1490.36 |
| 14 | 2015 | 0.87 | 4060225 | 1753.05 |
| 15 | 2016 | 0.99 | 4064256 | 1995.84 |
|  | $\sum x_i =$ 30112 | $\sum y_i = 9.87$ | $\sum x_i^2 =$ 60449206 | $\sum x_iy_i =$ 19821.71 |

Second, we make matrix to solve linier equation system. Matrix denoted as matrix A.

$$\begin{bmatrix} 15 & 30112 \\ 30112 & 60449206 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 9.87 \\ 19821.71 \end{bmatrix}$$

By implementing linier equation system, we can get value of $a$ and $b$.

$\mathrm{a} = -42.854;\ b = 0.021675081$

By getting value $a$ and $b$, now we have calculate the function.

$\mathrm{f(x) = a + bx}$

$\mathrm{f(x) = -42.854 + 0.021675081x}$

Now, we try to predict value for data test. We have 2 data test and here the prediction value.

TABLE V. PREDICTION FOR DATA TEST

| i | $x_i$ | $y_i$ | $f(x_i) = a + bx_i$ | error |
|---|---|---|---|---|
| 1 | 2011 | 0.6 | 0.734585 | -0.134585 |
| 2 | 2013 | 0.65 | 0.777935 | -0.127935 |

We also need to know how good the approximate function is. Now, we want to measure the Root Mean Square Error. We compare value of $y_i$ with value of $f(x_i)$. We also measure deviation and square of deviation.

TABLE VI. DATA FOR ROOT MEAN SQUARE ERROR

| i | $f(x_i) = a + bx_i$ | *deviation* | *deviation*$^2$ |
|---|---|---|---|
| 1 | 0.496159 | -0.076159 | 0.005800 |
| 2 | 0.517834 | 0.032166 | 0.001035 |
| 3 | 0.539510 | 0.090490 | 0.008189 |
| 4 | 0.561185 | 0.058815 | 0.003459 |
| 5 | 0.582860 | -0.032860 | 0.001080 |
| 6 | 0.604535 | 0.085465 | 0.007304 |
| 7 | 0.626210 | 0.003790 | 0.000014 |
| 8 | 0.647885 | 0.012115 | 0.000147 |
| 9 | 0.669560 | -0.129560 | 0.016786 |
| 10 | 0.691235 | -0.051235 | 0.002625 |

| i | $f(x_i) = a + bx_i$ | *deviation* | $deviation^2$ |
|---|---|---|---|
| 11 | 0.712910 | -0.002910 | 0.000008 |
| 12 | 0.756260 | -0.126260 | 0.015942 |
| 13 | 0.799611 | -0.059611 | 0.003553 |
| 14 | 0.821286 | 0.048714 | 0.002373 |
| 15 | 0.842961 | 0.147039 | 0.021621 |
| | | | $\sum = 0.089936$ |

By using data from table, we can measure the RMS error for knowing how good the function is. The function is good if it have small value.

$$E_{RMS} = \left(\frac{1}{n}\sum_{i=1}^{n}|f(x_i) - y_i|^2\right)^{1/2}$$

$$E_{RMS} = \left(\frac{1}{15}\sum_{i=1}^{15}|f(x_i) - y_i|^2\right)^{1/2}$$

$$E_{RMS} = \left(\frac{1}{15}0.089936\right)^{1/2}$$

$$E_{RMS} = 0.005996$$

From experiment above, we get $E_{RMS}$ value and the value is small. The smaller $E_{RMS}$, the better result will be obtained. There are several non linier functions that we can use for comparison.

*1) Simple Power Equation*

Equation for this method is

$$y = Cx^b \quad (15)$$

We can change that non liner equation to a linier equation. By change the non linier equation, we can use it in linier regression. Now, we try to change the form.

$$y = Cx^b \leftrightarrow \ln(y) = \ln(C) + b\ln(x) \quad (16)$$

We make new variables and form new equation

$$Y = \ln(y)\,; a = \ln(C)\,; X = \ln(x)\,; C = e^a \quad (17)$$

So we can form it to linier regression form.

$$Y = a + bX \quad (18)$$

TABLE VII. DATA FOR $y = Cx^b$ EQUATION

| i | $X_i = \ln(x_i)$ | $Y_i = \ln(y_i)$ | $X_i^2$ | $X_iY_i$ |
|---|---|---|---|---|
| 1 | 7.600902 | -0.867501 | 57.773718 | -6.593787 |
| 2 | 7.601402 | -0.597837 | 57.781317 | -4.544400 |
| 3 | 7.601902 | -0.462035 | 57.788913 | -3.512348 |
| 4 | 7.602401 | -0.478036 | 57.796506 | -3.634220 |
| 5 | 7.602900 | -0.597837 | 57.804095 | -4.545295 |
| 6 | 7.603399 | -0.371064 | 57.811682 | -2.821345 |
| 7 | 7.603898 | -0.462035 | 57.819264 | -3.513270 |
| 8 | 7.604396 | -0.415515 | 57.826844 | -3.159744 |
| 9 | 7.604894 | -0.616186 | 57.834420 | -4.686031 |
| 10 | 7.605392 | -0.446287 | 57.841993 | -3.394189 |
| 11 | 7.605890 | -0.342490 | 57.849563 | -2.604944 |
| 12 | 7.606885 | -0.462035 | 57.864692 | -3.514650 |
| 13 | 7.607878 | -0.301105 | 57.879809 | -2.290771 |
| 14 | 7.608374 | -0.139262 | 57.887362 | -1.059558 |
| 15 | 7.608871 | -0.010050 | 57.894912 | -0.076472 |
| | $\sum \ln(x_i) = 114.069387$ | $\sum \ln(y_i) = -6.569277$ | $\sum X_i^2 = 867.455091$ | $\sum X_iY_i = -49.951024$ |

Now, we must find value of $a$ and $b$. We use linier equation system.

$$\begin{bmatrix} 15 & 114.069387 \\ 114.069387 & 867.455091 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -6.569277 \\ -49.951024 \end{bmatrix}$$

By implementing linier equation system, we can get value of $a$ and $b$.

$$a = -0.556764564;\ b = 0.01562375$$

So, we can get value of $C$.

$$C = e^a = 0.57306$$

TABLE VIII. PREDICTION FOR DATA TEST USING $y = Cx^b$ EQUATION

| i | $x_i$ | $y_i$ | $f(x_i) = Cx^b$ | error |
|---|---|---|---|---|
| 1 | 2011 | 0.6 | 0.734585 | -0.134585 |
| 2 | 2013 | 0.65 | 0.777935 | -0.127935 |

*2) Exponential Model*

Equation for this method is

$$y = Ce^{bx} \quad (19)$$

We can change that non liner equation to a linier equation. By change the non linier equation, we can use it in linier regression. Now, we try to change the form.

$$y = Ce^{bx} \leftrightarrow \ln(y) = \ln(C) + bx\ln(e) \quad \ln(e) = 1 \quad (20)$$

We make new variables and form new equation

$$Y = \ln(y)\,; a = \ln(C)\,; X = x; C = e^a \quad (21)$$

So we can form it to linier regression form.

$$Y = a + bX \quad (22)$$

TABLE IX. DATA FOR $y = Ce^{bx}$ EQUATION

| i | $\mathbf{X_i = x_i}$ | $\mathbf{Y_i = ln(y_i)}$ | $\mathbf{X_i^2}$ | $\mathbf{X_iY_i}$ |
|---|---|---|---|---|
| 1 | 2000 | -0.867501 | 4000000 | -1735.001135 |
| 2 | 2001 | -0.597837 | 4004001 | -1196.271839 |
| 3 | 2002 | -0.462035 | 4008004 | -924.994990 |
| 4 | 2003 | -0.478036 | 4012009 | -957.505709 |
| 5 | 2004 | -0.597837 | 4016016 | -1198.065350 |
| 6 | 2005 | -0.371064 | 4020025 | -743.982681 |
| 7 | 2006 | -0.462035 | 4024036 | -926.843132 |
| 8 | 2007 | -0.415515 | 4028049 | -833.939496 |
| 9 | 2008 | -0.616186 | 4032064 | -1237.301768 |
| 10 | 2009 | -0.446287 | 4036081 | -896.590789 |
| 11 | 2010 | -0.342490 | 4040100 | -688.405521 |
| 12 | 2012 | -0.462035 | 4048144 | -929.615345 |
| 13 | 2014 | -0.301105 | 4056196 | -606.425657 |
| 14 | 2015 | -0.139262 | 4060225 | -280.613066 |
| 15 | 2016 | -0.010050 | 4064256 | -20.261477 |
|  | $\sum X_i = 30112$ | $\sum Y_i = -6.569277$ | $\sum X_i^2 = 60449206$ | $\sum X_iY_i = -13175.817954$ |

Now, we must find value of $a$ and $b$. We use Gauss elimination.

$$\begin{bmatrix} 15 & 30112 \\ 30112 & 60449206 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -6.569277 \\ -13175.817954 \end{bmatrix}$$

By process of Gauss elimination we found value of $a$ and value of $b$.

$$a = -64.43354;\ b = 0.031879$$

So, we can get value of $C$.

$$C = e^{a} = 0.978643$$

TABLE X. PREDICTION FOR DATA TEST USING $y = Ce^{bx}$ EQUATION

| i | $\mathbf{x_i}$ | $\mathbf{y_i}$ | $\boldsymbol{f(x_i) = Ce^{bx}}$ | error |
|---|---|---|---|---|
| 1 | 2011 | 0.6 | 0.722621 | -0.122621 |
| 2 | 2013 | 0.65 | 0.770194 | -0.120194 |

### 3) *Saturated growth rate model*

Equation for this method is

$$y = \frac{Cx}{d+x} \tag{23}$$

$$\frac{1}{y} = \frac{d}{C}\frac{1}{x} + \frac{1}{C} \tag{24}$$

Define new variables

$$Y = \frac{1}{x}; a = \frac{1}{C}; b = \frac{d}{C}; X = \frac{1}{x} \tag{25}$$

So we can form it to linier regression form.

$$Y = a + bX \tag{26}$$

TABLE XI. DATA FOR $y = \frac{Cx}{D+X}$ EQUATION

| i | $\mathbf{X_i = \frac{1}{x_1}}$ | $\mathbf{Y_i = \frac{1}{y_1}}$ | $\mathbf{X_i^2}$ | $\mathbf{X_iY_i}$ |
|---|---|---|---|---|
| 1 | 0.000500000 | 2.380952381 | 0.000000250 | 0.001190476 |
| 2 | 0.000499750 | 1.818181818 | 0.000000250 | 0.000908637 |
| 3 | 0.000499500 | 1.587301587 | 0.000000250 | 0.000792858 |
| 4 | 0.000499251 | 1.612903226 | 0.000000249 | 0.000805244 |
| 5 | 0.000499002 | 1.818181818 | 0.000000249 | 0.000907276 |
| 6 | 0.000498753 | 1.449275362 | 0.000000249 | 0.000722831 |
| 7 | 0.000498504 | 1.587301587 | 0.000000249 | 0.000791277 |
| 8 | 0.000498256 | 1.515151515 | 0.000000248 | 0.000754933 |
| 9 | 0.000498008 | 1.851851852 | 0.000000248 | 0.000922237 |
| 10 | 0.000497760 | 1.562500000 | 0.000000248 | 0.000777750 |
| 11 | 0.000497512 | 1.408450704 | 0.000000248 | 0.000700722 |
| 12 | 0.000497018 | 1.587301587 | 0.000000247 | 0.000788917 |
| 13 | 0.000496524 | 1.351351351 | 0.000000247 | 0.000670979 |
| 14 | 0.000496278 | 1.149425287 | 0.000000246 | 0.000570434 |
| 15 | 0.000496032 | 1.010101010 | 0.000000246 | 0.000501042 |
|  | $\sum X_i = 0.007472150$ | $\sum Y_i = 23.690231087$ | $\sum X_i^2 = 0.000003722$ | $\sum X_iY_i = 0.011805613$ |

Now, we must find value of $a$ and $b$. We use Gauss elimination.

$$\begin{bmatrix} 15 & 0.007472150 \\ 0.007472150 & 0.000003722 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 23.690231087 \\ 0.011805613 \end{bmatrix}$$

By process of Gauss elimination we found value of $a$ and value of $b$.

$$a = 1.579348737;\ b = 0.000005073$$

So, we can get value of $C$.

$$C = 1/1.579348737 = 0.633172381$$

$$d = 0.000005073 * 0.633172381 = 0.000003212$$

TABLE XII. PREDICTION FOR DATA TEST USING $y = \frac{Cx}{d+x}$ EQUATION

| i | $\mathbf{x_i}$ | $\mathbf{y_i}$ | $\boldsymbol{f(x_i) = \frac{Cx}{d+x}}$ | error |
|---|---|---|---|---|
| 1 | 2011 | 0.6 | 0.734585 | -0.134585 |
| 2 | 2013 | 0.65 | 0.777935 | -0.127935 |

## IV. Conclution

In this paper, we want to show what correlation between global earth temperature and year is. Global earth temperature tends increasing every year. This data forms a set of nodes that tends to be linier. By this situation, we can make prediction using linier regression.

Experiment is done with four approaches. First, we use linier regression. Second, we use simple power equation. Third, we use exponential model. Fourth, we use saturated growth rate model. Approach that generates smallest error is exponential model.

## References

[1] S. Chatterjee and A. S. Hadi, Regession Analysis By Example, 5th ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2012.

[2] R. Munir, Metode Numerik. Penerbit Informatika, 2013.

[3] S. C. Chapra, Applied Numerical Methods, 3rd ed. 1221 Avenue of the Americas, New York: McGraw-Hill, 2012.

[4] https://climate.nasa.gov/vital-signs/global-temperature, accessed on May, 10th 2017

## Statement

I hereby declare that this paper is written by me in my own words, not an adaptation or translation of another paper, and certainly not a result of plagiarism.

Bandung, 29 April 2012

EdwinSwandi Sijabat
23516012