An Interactive Network security for Evaluating Linear Regression Models in Cancer Mortality Analysis and Self-Correlation of Errors by Using Durbin-Watson Tests in Babylon/ Iraq

Asmaa Shaker Ashoor¹, Ali Abdul Karim Kazem¹, Sharad Gore².
¹University of Babylon/Iraq.
²University of Pune/India.
asmaa_zaid218@yahoo.com
ali.babylon1982@gmail.com
sharaddgore@gmail.com

Abstract. The present study addresses Interactive Network security of the model of simple and frequent linear regression constructing method in describing the phenomenon of deaths of patients with cancer for the time variable in two aspects, one of which provides a theoretical background of the construction of the model of simple linear regression and the adoption of the principle of errors in the description of the standard deviation between the estimated value of the model and its true value, the other one is practical which shows the mechanism of building the model of simple linear regression and multiplexing and analysis of the results using two programs ( SPSS and Minitab ). And the study of the behaviour of errors for each model and drawn in the form of normal probability plot was also used the coefficient of selection $R^2$ and Coefficient Selection Factor(adj) $R^2$ in comparing the models and determining the best model in explaining the growth phenomenon of cancer deaths.

Keywords. Network security, Linear Regression, Self-Correlation, Durbin-Watson Test, SPSS

1. Aim of the Study
The present study aims at building better statistical models in estimating death rates for rates for cancer patients and monitor this phenomenon in terms of the rise and fall of the researcher to contribute to this research study in health empowerment in real future plans and to reduce this phenomenon and working on the policy which works to alleviate this disease and directing the sight of the competent authorities to its seriousness and to reduce the indicators and death rates and improving the level of health services provided to the citizen and enhance the state of health awareness.

2. Introduction
Cancer is one of the most common deadly diseases affecting a human which does not have a cure which works on healing forever. Despite the medical development in the treatment methods and the treatment...
of many diseases that destroy the human mature, but the remaining medicine provides only simple solutions that try to help the patient to live normally during the period of illness through early detection of the disease and chemical treatment of others. Despite all this, the disease remains one of the most dangerous and deadly diseases in humans. Weapons, war remnants, chemicals and pollutants are among the most important methods that have helped spread this disease, especially in last years. As well as the lack of safe measures by the competent authorities to remove these dangers and contaminants have contributed effectively to the growth and spread of this serious disease.

3. Materials and Methods

Linear Regression Model: (Overall trend) [6] [1] represent simple linear regression model throughout the following relationships

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1,2,3, \ldots \ldots n \]

Where \( y_i \) is called the adopted variable (which represents the phenomenon to be interpreted such as the phenomenon of deaths of cancer in Babylon and \( x_i \) with an independent variable which can be the time indicator (days, months, years...), then the model is called the general linear trend model. \( (\beta_0 + \beta_1 x) \) reflects the amount of change in \( y \) value whenever \( x \) value is changed with the amount of one unit.

\( \beta_0 \) and \( \beta_1 \) are the parameters of the model and \( \epsilon_i \) refers to the amount of error in explaining the independent variable \( y_i \).

The above equation represents the true value of \( y_i \) whereas the estimated value which is given with the following relationship:

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \ldots \ldots (2) \]

\( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) represents the estimated values of the parameters of the model and thus art can be obtained each of the following formulas:

\[ \hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \]

\( \bar{x} \) and \( \bar{y} \) are the mean for each \( x \) and \( y \) respectively.

Therefore, the estimated value of the error is as follows:

\[ \hat{\epsilon}_i = y_i - \hat{y}_i \ldots \ldots (3) \]

4. Durbin-Watson Test

The Durbin Watson test is used to determine the existence or the absence of self-association of errors. Errors are generally random and the test formula is as follows:

\[ d = \frac{\sum_{i=1}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \]

If \( d \) is less than 2 so the residues are positively linked, but if \( d \) is greater than 2, the residues are inversely linked.

5. The Total Deviations and the Model Test [5] [2]

The total of deviations for the independent variable \( y_i \) are divided into three parts:

1- Sum of square regression (SSR)

\[ SSR = \sum (\hat{Y}_i - \bar{Y})^2 \quad \ldots \ldots 4 \]

2- Sum of square error (SSE)

\[ SSE = \sum (Y_i - \hat{Y}_i)^2 \ldots \ldots (5) \]

3- Sum of squares total deviations SST
SST = SSR + SSE

In equations 4 and 5, we obtain the value of selection correlation which can be judged on the quality of the model or not which is represented by the following relationship:

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} \quad (6)$$

It is an identified value and it can be within the following domain:

$$0 \leq R^2 \leq 1$$

6. Multiple Regression Model [3] [4]

The multiple linear regression model is another case of describing the relationship between the dependent variable $y_i$ and to $n$ within the views $(x_{1i}, x_{2i}, ..., x_{ki})$ from independent variables that clarify the multiple model in the following relationship:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \epsilon_i \ldots \ldots (7)$$

$X_{1i}, X_{2i}, ..., X_{ki}$. They are called the explanatory variables (independent) for the dependent variable $y_i$ and $\beta_0, \beta_1, ..., \beta_k$ are sample parameters, this system can be written in the following form:

"$Y = X \beta + \epsilon \ldots \ldots (8)"$

Matrix form

$$Y = (Y_1, Y_2, ..., Y_n), \quad X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

From the matrix written above, we can obtain the quadratic model from the second mark in the following form:

"$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \epsilon_i"$

And the cubic model from the third mark in the following:

"$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3 + \epsilon_i"$

7. Applied side

Data were collected by researchers for deaths due to cancer for different ages of the Statistics Division in the department of health of Babylon and for a sample of the period 2010-2018 and the number of deaths due to cancer out of the total number of deaths and we build simple and multiple regression models cancer deaths without going into classification of numbers for deaths by cancer type, but the trend to analyze statistical numbers for cancer in general with statistical analysis of the results and study the behavior of errors for each model. Thus determining the best result in the interpretation of indicators of death phenomenon of the disease.

| Year | Total deaths number | Deaths due to cancer |
|------|---------------------|---------------------|
| 2010 | 6892                | 681                 |
| 2011 | 6804                | 704                 |
| 2012 | 7107                | 733                 |
| 2013 | 7261                | 743                 |
| 2014 | 7898                | 711                 |
8. Simple Linear Regression Model

The simple linear regression model (general trend) is constructed between the deaths number of cancer in Babylon province as a dependent variable $dci$ and time $yi$ as an independent explanatory variable that affects the number of deaths by hypothesis and according to statistical theory, we can clarify the following:

- The errors of the model are independent and do not follow any statistical model that links errors to each other.
- The distribution of errors follows natural distribution in order to be the basis of the process of estimation and construction of models scientifically basis, where the model was estimated through the method of the squares with minimum extraction of statistical tests of the parameters and model with the efficiency measures of the model and the results are as follows:

9. The regression equation is

$$dc = 660.3 + 20.17y$$

S =36.8883   R-sq =71.92%   R-sq(adj)= 67.91%

Analysis of Variance

| Source       | DF | SS        | MS         | F        | P       |
|--------------|----|-----------|------------|----------|---------|
| Regression   | 1  | 24401.7   | 24401.7    | 17.93    | 0.004   |
| Error        | 7  | 9525.2    | 1360.7     |          |         |
| Total        | 8  | 33926.9   |            |          |         |

Accuracy Measures

| MAPE         | 3.20 |
| MAD          | 25.15 |
| MSD          | 1058.36 |

Durbin – Watson statistics = 2.937

10. Statistical Analysis of the Model

The regression equation and the variance analysis table indicated the following data:

1- There is a strong correlation between the variables on the number of deaths for cancer ($dc_i$) and the independent variable ($yi$) with the positive relationship indicating that the increase of time in the unit leads to an increase in the number of deaths from this disease.

2- According to the value of the determination coefficient which amounts to 71.92%, we notice that the explanatory force for the model is high and this denotes that the time explains these changes concerning the number of death cases which amounts to 71.92%.

3- The sample is approved from the statistics point of view according to Test F which amounts to 17.93 and the P-value (0.004) whose value is less than (0.05). This indicates the efficiency of the sample.

4- The value of Durbin–Watson Test (which amounts to 2.937 and which is bigger than 2) denotes that there is an inverse connection.

5- Time (as a dependent variant) is positive and this indicates that the increase in time one unit leads to increase in the number of death cases concerning cancer disease as the results of the estimated model are showed in Figure (1) below.

However, we will adopt the multiple regression sample (Quadratic and Cubic as we will see later) to get a better and more valuable model in accordance with the statistical theory and in order to reach a larger possibility than the normal model in explaining the phenomenon of death concerning the cancer disease.
11. Errors behavior model
In order to ensure the efficiency/no efficiency of the simple linear trend model, the errors behavior sample has been employed and drawn through (Normal probability plot). We find that the model errors are very close to the line of trend which signifies the acceptability of the model and the nonexistence of the extreme value whether in its beginning or its end. The acceptability of the model is relatively accepted and the distribution of errors through (Histogram) is not distant from the bell natural distribution figure. However, the errors model is more distant from zero somehow and Figure (2) explains the errors behavior for the Linear estimated model.

12. The Quadratic model
Quadratic regression model has been established for the variant of death cases concerning cancer disease as a dependent variant \((dc)\) according to the time variant as an independent variant (explanatory) \((y)\) as below:

The regression equation is
\[
dc = 677.1 + 11.0 \, y + 0.920 \, y^2
\]

\[
S = 39.2950 \quad \text{R-sq} = 72.69\% \quad \text{R-sq(adj)} = 63.59\%
\]

| Analysis of Variance |
|----------------------|
| Source          | DF | SS     | MS    | F     | P     |
| Regression      | 2  | 24662.3| 12331.2| 7.99  | 0.020 |
| Error           | 6  | 9264.6 | 1544.1|       |       |
| Total           | 8  | 33926.9|       |       |       |

![Figure 1](image1.png)

Figure 1. The Linear Trend Model estimated for the Death Variant due to Cancer Disease.

![Figure 2](image2.png)

Figure 2. The Errors behavior for the Linear Simple Estimated model for the Variant of the Number of Death Cases Concerning Cancer Disease.
Accuracy Measures
MAPE 3.26
MAD 25.50
MSD 1029.40
Durbin – Watson statistics = 2.973

13. The Statistical Analysis of the Quadratic Model
The results of the quadratic model for the regression equation refer to the following findings:

1- There is a related connection between the variant of the number of death cases of cancer disease (as dependent) and the time variant (as independent) and the relation between them is direct i.e. the increase of one unit leads to the increase in the number of death cases with the amount of 11.0

2- The determination coefficient value has been increased slightly with reference to the simple model where its value amounts to 72.69% and this means that time explains somehow the changes in the number of death cases for the cancer disease with the percentage of 72.69%

3- The model is accepted from the statistical point of view and according to test F which is (7.99) and the P-value 0.020 which is less than 0.50 which indicates the acceptability of the estimated model.

4- The influence of the other variant (the quadratic of time) was direct as well due to its positive signal which leads to the increase of death cases continually with the amount of 0.920

5- The value of Durbin–Watson Test (2.973) is surely larger than 2 which indicates that the connection in inverse for the model.

Figure 3. The General Quadratic Trend Estimated model for the variant of Death Cases for Cancer Disease in Babylon Governorate.

The following procedures concerning the sequential analysis of the cubic regression are explained below:

Sequential Analysis of Variance

| Source  | DF | SS  | F     | P   |
|---------|----|-----|-------|-----|
| Linear  | 1  | 24401.7 | 17.93 | 0.004 |
| Quadratic | 1  | 260.6  | 0.17  | 0.695 |
| Cubic   | 1  | 568.2  | 0.33  | 0.592 |

It is noted that the results of the model have been improved. However, the cubic model does not explain most of the results and as such the model is non-accepted statistically and accordingly the quadratic model is better in explaining the phenomenon of death cases of cancer.

14. Errors Behaviour model
To ensure the efficiency/non-efficiency of the cubic model, errors model and their drawn through (normal probability plot) have been employed to find that the errors model is very close to the line of
the trend which indicates the high efficiency and the distribution of errors through (Histogram) is relatively distant from the bell natural distribution figure. However, the errors model in accordance with the estimated value was distant from the zero somehow in (versus fits) and Figure (6) explains the errors cubic estimated model.

![Figure 6. the Errors Behaviour Cubic estimated model for the Variant of Death Cases of Cancer Disease in Babylon Governorate.](image)

15. Results
After the procedures of the practical analysis, the following results are arrived at:

1- The Linear model is accepted statistically according to the value of F Test whose P-value is 0.004 and determination coefficient is 71.92% which denotes the efficient acceptability of the model.

2- The quadratic model is better among the three models according to F Test, P-value 0.020, and determination coefficient 72.69 which are the highest among its counterparts in the linear model and closer to 1.

3- The cubic model has failed to explain the phenomenon of death cases because its P-value which reaches 0.061 is larger than 0.05 and means the non-acceptance of the model.

Acknowledgments
The researcher would like to express his thanks and appreciation to Babylon Health Department/training and Human Development Centre for easing the researcher's tasks and providing him with formal data about death number of cancer disease. Thanks and appreciation are due to the conference organizing members who allowed this study the chance to be presented where it is going to be scientifically assessed.

References
[1] Al-Timimi, Raad Fadhil, Regression and Sequential time, Applied advanced Statistical Methods Using Minitab System, Al-Jazira Office, Baghdad, 2013.
[2] Al-Rawi, khashia Mahmoud, Introduction to Regression Analysis, Directorate of Dar Al-Kutub for printing and publishing, 2009.
[3] Al-Shimari, Nathir Abass and Al-Azawi, Dijlah Ibrahim, Standard Economy and its Appications, Baghdad: Al-Jazira for printing and publishing, 2007.
[4] Al-Aagili, Hanaa Muhsin, Satistical Analysis in Education and Psychology, Al-Mustansiriya University, Dar Al-Hikma for printing and publishing, 1991.
[5] Morrison,D.F.,"Multivariate Statistical Method", McGraw Hill, New York, 1976.
[6] Norusis, M., User Gide Spss/pc,"Chicago(manual), 1986.
Spss Manual 7.5,(1997) "Advanced Statistical", NC.
[7] Asmaa Shaker Ashoor , Ali Abdul Karim Kazem Naji, Evaluation of the Birth Rate of Children in Babil Governorate for the Period 2010-2017 by using Network Security through Linear Regression Models/Iraq, Journal of Engineering and Applied Sciences, vol.13, Issue.13SI, 2018