*IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555*
*Vol. 5, No1, February 2015*

# Traffic Accidents Prediction Model

Dr. Sharaf A. Alhomdy
Dept. of Information Technology,
Faculty of Computer and Information Technology, Sana'a University,
Sanaa, Republic of Yemen.

Dr. Abdulwase M. Al-Azzani
Dept. of Computer Science,
Faculty of Computer and Information Technology, Sana'a University,
Sanaa, Republic of Yemen.

Dr Ghaleb H. Al-Gaphari
Dept. of Computer Science,
Faculty of Computer and Information Technology, Sana'a University,
Sanaa, Republic of Yemen.

***Abstract*—Traffic passage accidents problem is one of the greatest challenges in the society, which has acquired significant research interest from academics and research teams in recent years. Therefore, the main objective of this paper is to develop a method to predict the number of traffic passage accidents in the capital city of Yemen and create a certain procedure. It uses a mathematical model that predicts the number of traffic passages in a specific month within a specific period. It is a nonlinear regression model with three variables. This model is implemented programmatically after investigating the model accuracy.**

**The new model is used to predict the number of traffic passage accidents in the capital city of Yemen. Such statistical predictions may help the administration to manage passage accidents problem anywhere, especially in the city crowded places, in a way that may reduce the accidents' rate. The performance and the results of the model were very promising.**

***Keyword: WHO; Least Square Algorithm; Back-Substitution Algorithms; Coefficient; Correlation; Regression.***

## I. INTRODUCTION

One of the problems that affected the society of the capital city in Yemen is the traffic passage accidents. Therefore there is no doubt that the passage accidents problem is extremely costly to the society because it causes death to many peoples. In addition, it affects the population socially and economically in their life style. Therefore, it is necessary to develop methods to reduce the number of passage accidents anywhere, especially in the crowded places of large cities. Sana'a is one of those places. It is very crowded because it is the capital city of Yemen. For vehicle drivers come from all over the country. So, some of them never behave on the bases of the rules of the law related to vehicle driving regulations imposed internationally. For example, some of them exceed the speed limit in the city as a result of chewing Qat, tak drugs, drinking, or using mobile phones while driving. In fact, these reasons increase the number of passage accidents in the capital city. The sooner these reasons are removed, the better the passage system within the capital is likely to be improve and organized.

The term regression simply refers to fitting a curve (which may be a straight line) to the graph of the data. The curve is defined in terms of a functional expression such as [1]:

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \alpha_3 x_i^3 \quad (1)$$

Once the equation (1) of the fitted curve is determined (regression), the degree to which it fits the graph may be calculated (correlation). Then, you can determine whether this is a reasonably good fit or a poor one. If the correlation is good, the equation may be used to predict intermediate values in the data, or it can project the data beyond the largest or the smallest value of the independent variable.

For example, in passage accidents case, we can estimate the number of accidents occur during a specific month in a specific year. This is not one of the values of the independent variable for which we tested the month and the year. But if we have a good mathematical relationship between the number of accidents and the month during which the accidents occur we can simply substitute (month=5, year=2015) then we can predict the number of accidents.

In fact, the problem is how to develop this equation; in this study we developed, tested and programmed such equation. Thus, it is very important to find out a way to decrease the number of passage accidents that may save some people's lives, and improve available resources [2, 3].

The main objective of this study is to:

1. Predict the number of passage accidents in a specific month during a specific period.
2. Show the relationship between times and number of accidents.

The paper has been organized in a flexible manner. Section II focuses on the background that also includes passage accidents problem in a region including Yemen. Section III explains proposal models for prediction algorithm to predict the number of passage accidents. Section IV explains the proposed experiment. Conclusion is in section V. Recommendations and future research assignments will be highlighted in section VI.

## II. BACKGROUND

In this section a survey will be elaborated on previous studies conducted in the matter of passage accidents problem in a region including Yemen and informal explanation of the methods.

### A. *Regression Analysis Parameters*

In general, the regression analysis is a method used for fitting a curve. Such analysis uses many parameters as [1, 5]:

- *Coefficient of Determination*

It is used to determine the degree of estimation goodness using the (2):

$$R2=1-SSE/\ (SSR+SSE) \quad (2)$$

Where SSE is sum of squares errors, and SSR is sum of squares regression [1].

- *Coefficient of Correlation*

The standard method involves calculating what is called the correlation coefficient for the curve fit. The coefficient commonly is denoted by r and given by (3):

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x2-(\sum x)2}\ \sqrt{n\sum y2-(\sum y)2}} \quad (3)$$

The value of r may be either positive or negative, but its absolute value will always be between 0 and 1. If r is zero, there is no correlation between months (x) and number of accidents (y). Hence, y is not dependent on x and x cannot be used as a predictor for y. If r is 1, there is perfect correlation between x and y, also y is completely dependent on x. The closer the absolute value or is to 1, the better the correlation or the better the fit. That is what we are seeking for.

- *Measures of Central Tendency*

There are several measures of central tendency that qualify as averages. These include the mean, median, mode, etc.

### B. *Prior Studies*

In this section a survey will be elaborated on previous studies conducted in the matter of traffic accidents problem in the region including Yemen.

- *Cars Accidents Problem & Solution*

The study started with the statement of the General Manager World Health Organization (WHO)[5, 6]. He said that road safety should not be left to chance based on a proposal submitted by the Sultanate of Oman to the General Assembly of WHO. He revealed that the daily road casualties in the world about 140,000 are infected, of whom more than 3000 people die and about 15000 people suffer permanent disabilities. These infections naturally cause to some families the loss of their breadwinners or buckle under the burdens of carrying the injured maintenance costs. If then these figures are reverse in the year, they would be scary numbers: Road injuries 51,100,000 and Mortality1, 095,000.

The study estimated the death tolls resulting from road collisions in 2000 at about 700,000 deaths and the case in 2002 reported about 1,018,000 deaths. If these rates the same pace vested mortality and disability resulting from road injuries at about 60% by 2020 to occupy the third rank in the list of WHO for having the main reasons of diseases and injuries in the world instead of the ninth position it occupied in 1990.

According to World Health Organization statistics, the low-income countries in Africa and the Mediterranean eastern Mediterranean regions are the highest mortality rates, amounted to 28.3 per 100,000 people in Africa and 26.3 per 100,000 people in the eastern Mediterranean territory, while in high-income countries the ratio is 12, 6 per 100,000 and the lowest recorded ration at the global level in the United Kingdom is 5.4 [6].

Moreover, the researcher estimated the financial cost that is caused by road accidents at the global level at about 520 billion dollars, including 65 billion dollars in low-income countries and middle course, the Arab countries.

Gulf States collectively lose about ten billion dollars, of which Saudi Arabia loses 5,6 billion dollars. It is the Arab Maghreb country. Morocco, that loses annually about $ 33 million. The researcher concluded that most vulnerable road users are males, pedestrians, and motorcyclists.

Then, the researcher mentioned the situation in the Republic of Yemen. He stated that in 1999 road accidents were the top cases staged an emergency at Al-Thawra Modern General Hospital, Sana'a-Yemen. The ratio was 40.89% entries in 2000 and raised to 41.18% in 2001 but dropped to 40.52% in 2002 and raised to 42.75 %, and then in 2003 went down to be increased more in 2004 to 43% of receiving emergency cases. Males also occupy first place of road injuries at 88.1% and children ranked second with 9.3% and 2.6% of women. Males also occupy first place of road injuries at 88.1% and children ranked second with 9.3% and 2.6% of women [5].

The researcher concluded that the positions most vulnerable to injuries are the skull, neck and spinal cord occupies first place at 85.9%, stomach and chest and pelvic 6.1% each. This confirms the high mortality of the 1243 cases of death in 1999 to 1452 cases of death in 2000 and then to 3000 cases of death in 2004, and which confirms this hypothesis that victims died in October 2006 over 256 cases, and the average age of males who passed away is 25 years old. Finally, the researcher mentioned the relationship between traffic accidents and risk factors and the suggested solution [5].

- *Traffic Week & Physical Safety Studies*

Susan's Poll [6] stated that it is not a task of the physical safety, but the moral integrity of the drivers and passengers of public transport. Accidents on the roads alone would require action to stop them. The objective of the Traffic Week is to increase awareness of traffic rules and traffic safety on the road. But also ethics should be considered in terms of public transport. The majority resort to such means, if not daily, may be forced to particular circumstances. Unfortunately, some bus drivers as well as the passengers have lack of feelings, emotions, and morality. Susan's Poll mentioned some people criticisms for inhuman acts emanating from some buses drivers to the elderly, children, disabled and the blind people [5].

## III. Proposed Models

In this section we start analyzing the regression by computing the regression parameters in the simple linear regression based on sufficient data to determine the suitable regression model. But in case there is insufficient data to do so, we start analyzing data to find more than one initial model as shown in Fig. 1. Such models should be tested then the suitable one should be selected. In our experiment, we use determination coefficient R2 between two variables as follows:

h0: R2 =0 that means there is no relationship between traffic accidents and the month.

h1: R2 ≠0 that means there is a relationship between traffic accidents and the month.

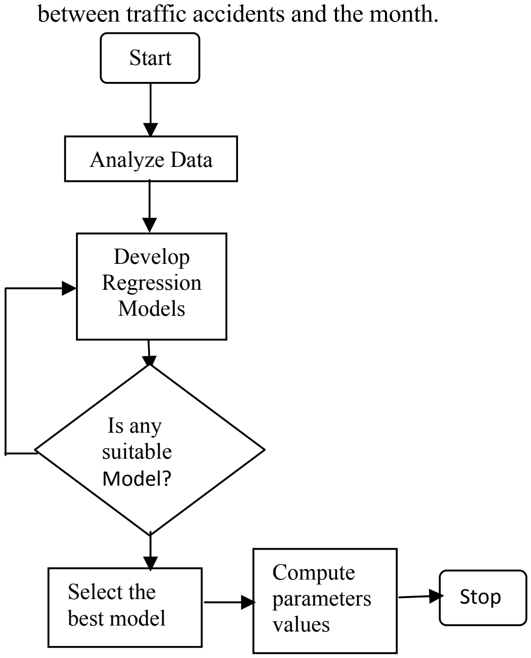


Figure 1: Testing Three Alternatives Models.

Fig .1 Showing the test of three alternative models and selecting the best one. Such models are:

$y_i = \alpha_0 + \alpha_1 x_1$ simple linear model (4)

$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2$ (5)

Polynomial Regression in two variables

$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \alpha_3 x_i^3$ (6)

Polynomial Regression in three variables

### A. Testing of Simple Linear Regression Model

We consider the model as in (4) simple linear model. Then we try to estimate the values of $\alpha_0$ and $\alpha_1$ using least square algorithm.

This algorithm generates unbiased estimation for such parameters as the following steps:

1. Find $\sum x_i$, $\sum x^2$, $\sum y_i$, $\sum x_i y_i$ such that y is a dependable variable that denotes the number of accidents, and x is an independent variable that denotes the month.

2. Solve the system

$$AX=b \quad (7)$$

Such that:

$$A=\begin{pmatrix} m & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \quad x=\begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \quad b=\begin{pmatrix} \sum x_i^0 y_i \\ \sum x_i^2 y_i \end{pmatrix} \quad (8)$$

Where **m** is the number of data items.

3. Use the values presented in a tabular format as shown in Table (1).

Such values are: $\sum x_i$=1176, $\sum y_i$ =1919, $\sum x^2$=38024 and $\sum x_i y_i$ =49310.

4. Substitute for the above variables and solve the system to get $\alpha_0$=33.877 and $\alpha_1$=0.249.

Therefore, the simple linear model is:

$$y=33.877+0.249x_i. \quad (9)$$

To discuss how we determine whether or not we have a good fit with the above equation, we calculate what is called determination coefficient as in (10).

$$R^2=1-SSE/SST \quad (10)$$

$SSE = \sum (y_i - \hat{y}_i)^2$ =1739.471324,

$SSR = \sum (\hat{y}_i - \bar{y}_i)^2$ =571.507843,

$SST=SSR+SSE = \sum (y_i - \bar{y}_i)^2$ =2310.979167

$R^2$=0.247 => r = 0.4973.

Such that y is the dependent variable (number of accidents) and x is the independent variable (month). The parameters estimation model for simple linear regression (SLR) can be shown in Table (2).

Table 2: Parameters Estimation Model for SLR

| Regression Type | Model statistics | | | | Parameters Estimation | | |
|---|---|---|---|---|---|---|---|
| Linear | $R^2$ | F | DF1 | DF2 | Sig. | Constant | B |
| | 0.247 | 15.113 | 1 | 46 | 0.000 | 33.877 | 0.249 |

Since the value of the determination coefficient is not equal to zero, there is a relationship between the number of accidents and the month during which the accidents occur.

Table 1: The Data Collection and Calculated Parameters for Regression.

| year | month | X | Y | XY | X^2 | Y^2 | Yh | (Y-Yh) | (Y-Yh)^2 | (Yh-Yb) | (Yh-Yb)^2 | (Y-Yb) | (Y-Yb)^2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2011 | Jan | 1 | 39 | 39 | 1 | 1521 | 34 | 5 | 25.00 | -5.979 | 35.748 | -0.98 | 0.9584 |
| 2011 | Feb | 2 | 29 | 58 | 4 | 841 | 34 | -5 | 25.00 | -5.979 | 35.748 | -11 | 120.54 |
| 2011 | Mar | 3 | 33 | 99 | 9 | 1089 | 35 | -2 | 4.00 | -4.979 | 24.79 | -6.98 | 48.706 |
| 2011 | Apr | 4 | 28 | 112 | 16 | 784 | 35 | -7 | 49.00 | -4.979 | 24.79 | -12 | 143.5 |
| 2011 | May | 5 | 40 | 200 | 25 | 1600 | 35 | 5 | 25.00 | -4.979 | 24.79 | 0.021 | 0.0004 |
| 2011 | Jun | 6 | 41 | 246 | 36 | 1681 | 35 | 6 | 36.00 | -4.979 | 24.79 | 1.021 | 1.0424 |
| 2011 | Jul | 7 | 43 | 301 | 49 | 1849 | 36 | 7 | 49.00 | -3.979 | 15.832 | 3.021 | 9.1264 |
| 2011 | Aug | 8 | 34 | 272 | 64 | 1156 | 36 | -2 | 4.00 | -3.979 | 15.832 | -5.98 | 35.748 |
| 2011 | Sep | 9 | 37 | 333 | 81 | 1369 | 36 | 1 | 1.00 | -3.979 | 15.832 | -2.98 | 8.8744 |
| 2011 | Oct | 10 | 39 | 390 | 100 | 1521 | 36 | 3 | 9.00 | -3.979 | 15.832 | -0.98 | 0.9584 |
| 2011 | Nov | 11 | 32 | 352 | 121 | 1024 | 37 | -5 | 25.00 | -2.979 | 8.8744 | -7.98 | 63.664 |
| 2011 | Dec | 12 | 33 | 396 | 144 | 1089 | 37 | -4 | 16.00 | -2.979 | 8.8744 | -6.98 | 48.706 |
| 2012 | Jan | 13 | 33 | 429 | 169 | 1089 | 37 | -4 | 16.00 | -2.979 | 8.8744 | -6.98 | 48.706 |
| 2012 | Feb | 14 | 29 | 406 | 196 | 841 | 37 | -8 | 64.00 | -2.979 | 8.8744 | -11 | 120.54 |
| 2012 | Mar | 15 | 35 | 525 | 225 | 1225 | 38 | -3 | 9.00 | -1.979 | 3.9164 | -4.98 | 24.79 |
| 2012 | Apr | 16 | 38 | 608 | 256 | 1444 | 38 | 0 | 0.00 | -1.979 | 3.9164 | -1.98 | 3.9164 |
| 2012 | May | 17 | 34 | 578 | 289 | 1156 | 38 | -4 | 16.00 | -1.979 | 3.9164 | -5.98 | 35.748 |
| 2012 | Jun | 18 | 29 | 522 | 324 | 841 | 38 | -9 | 81.00 | -1.979 | 3.9164 | -11 | 120.54 |
| 2012 | Jul | 19 | 50 | 950 | 361 | 2500 | 39 | 11 | 121.00 | -0.979 | 0.9584 | 10.02 | 100.42 |
| 2012 | Aug | 20 | 40 | 800 | 400 | 1600 | 39 | 1 | 1.00 | -0.979 | 0.9584 | 0.021 | 0.0004 |
| 2012 | Sep | 21 | 34 | 714 | 441 | 1156 | 39 | -5 | 25.00 | -0.979 | 0.9584 | -5.98 | 35.748 |
| 2012 | Oct | 22 | 42 | 924 | 484 | 1764 | 39 | 3 | 9.00 | -0.979 | 0.9584 | 2.021 | 4.0844 |
| 2012 | Nov | 23 | 37 | 851 | 529 | 1369 | 40 | -3 | 9.00 | 0.021 | 0.0004 | -2.98 | 8.8744 |
| 2012 | Dec | 24 | 32 | 768 | 576 | 1024 | 40 | -8 | 64.00 | 0.021 | 0.0004 | -7.98 | 63.664 |
| 2013 | Jan | 25 | 42 | 1050 | 625 | 1764 | 40 | 2 | 4.00 | 0.021 | 0.0004 | 2.021 | 4.0844 |
| 2013 | Feb | 26 | 44 | 1144 | 676 | 1936 | 40 | 4 | 16.00 | 0.021 | 0.0004 | 4.021 | 16.168 |
| 2013 | Mar | 27 | 50 | 1350 | 729 | 2500 | 41 | 9 | 81.00 | 1.021 | 1.0424 | 10.02 | 100.42 |
| 2013 | Apr | 28 | 42 | 1176 | 784 | 1764 | 41 | 1 | 1.00 | 1.021 | 1.0424 | 2.021 | 4.0844 |
| 2013 | May | 29 | 51 | 1479 | 841 | 2601 | 41 | 10 | 100.00 | 1.021 | 1.0424 | 11.02 | 121.46 |
| 2013 | Jun | 30 | 44 | 1320 | 900 | 1936 | 41 | 3 | 9.00 | 1.021 | 1.0424 | 4.021 | 16.168 |
| 2013 | Jul | 31 | 49 | 1519 | 961 | 2401 | 42 | 7 | 49.00 | 2.021 | 4.0844 | 9.021 | 81.378 |
| 2013 | Aug | 32 | 52 | 1664 | 1024 | 2704 | 42 | 10 | 100.00 | 2.021 | 4.0844 | 12.02 | 144.5 |
| 2013 | Sep | 33 | 47 | 1551 | 1089 | 2209 | 42 | 5 | 25.00 | 2.021 | 4.0844 | 7.021 | 49.294 |
| 2013 | Oct | 34 | 34 | 1156 | 1156 | 1156 | 42 | -8 | 64.00 | 2.021 | 4.0844 | -5.98 | 35.748 |
| 2013 | Nov | 35 | 49 | 1715 | 1225 | 2401 | 43 | 6 | 36.00 | 3.021 | 9.1264 | 9.021 | 81.378 |
| 2013 | Dec | 36 | 44 | 1584 | 1296 | 1936 | 43 | 1 | 1.00 | 3.021 | 9.1264 | 4.021 | 16.168 |
| 2014 | Jan | 37 | 35 | 1295 | 1369 | 1225 | 43 | -8 | 64.00 | 3.021 | 9.1264 | -4.98 | 24.79 |
| 2014 | Feb | 38 | 37 | 1406 | 1444 | 1369 | 43 | -6 | 36.00 | 3.021 | 9.1264 | -2.98 | 8.8744 |
| 2014 | Mar | 39 | 44 | 1716 | 1521 | 1936 | 44 | 0 | 0.00 | 4.021 | 16.168 | 4.021 | 16.168 |
| 2014 | Apr | 40 | 44 | 1760 | 1600 | 1936 | 44 | 0 | 0.00 | 4.021 | 16.168 | 4.021 | 16.168 |
| 2014 | May | 41 | 41 | 1681 | 1681 | 1681 | 44 | -3 | 9.00 | 4.021 | 16.168 | 1.021 | 1.0424 |
| 2014 | Jun | 42 | 36 | 1512 | 1764 | 1296 | 44 | -8 | 64.00 | 4.021 | 16.168 | -3.98 | 15.832 |
| 2014 | Jul | 43 | 50 | 2150 | 1849 | 2500 | 45 | 5 | 25.00 | 5.021 | 25.21 | 10.02 | 100.42 |
| 2014 | Aug | 44 | 50 | 2200 | 1936 | 2500 | 45 | 5 | 25.00 | 5.021 | 25.21 | 10.02 | 100.42 |
| 2014 | Sep | 45 | 52 | 2340 | 2025 | 2704 | 45 | 7 | 49.00 | 5.021 | 25.21 | 12.02 | 144.5 |
| 2014 | Oct | 46 | 49 | 2254 | 2116 | 2401 | 45 | 4 | 16.00 | 5.021 | 25.21 | 9.021 | 81.378 |
| 2014 | Nov | 47 | 41 | 1927 | 2209 | 1681 | 46 | -5 | 25.00 | 6.021 | 36.252 | 1.021 | 1.0424 |
| 2014 | Dec | 48 | 31 | 1488 | 2304 | 961 | 46 | -15 | 225.00 | 6.021 | 36.252 | -8.98 | 80.622 |
| total | | 1176 | 1919 | 49310 | 38024 | 79031 | 1920 | -1 | 1707.00 | 1.008 | 584.02 | 0.008 | 2311 |

## B. *Testing of Two Variables Regression Model*

A linear regression may be inaccurate, so we consider nonlinear regression that may exhibit what is called a parabolic shape. A parabolic shape has the general equation as in (5).

Now we should compute the values of three parameters $\alpha_0$, $\alpha_1$, and $\alpha_2$ using least square algorithm as follows:

1. Compute $\sum x_i$, $\sum x_i^2$, $\sum x_i^3$, $\sum x_i^4$, $\sum y_i$ , $\sum x_i y_i$.

2. Solve the system

$$AX=b \qquad (11)$$

$$A=\begin{pmatrix} m & \sum x & \sum x^2 \\ \sum x & \sum x^2 & \sum x^3 \\ \sum x^2 & \sum x^3 & \sum x^4 \end{pmatrix} \quad x=\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} \quad b=\begin{pmatrix} \sum x \\ \sum x y_i \\ \sum x^2 y_i \end{pmatrix} \qquad (12)$$

3. Extend Table 1 to accommodate $x_i^3$, $x_i^4$ and $x_i^2 y_i$.

4. Use the values: m=48, $\sum x_i$=1176, $\sum x_i^2$=38024, $\sum x_i^3$ =1382976, $\sum x_i^4$=53651864, $\sum y_i$ =1919, $\sum x_i y_i$ =49310 and $\sum x_i^2 y_i$ = 1625698 to get the regression parameters values.

5. Substitute for the above variables and solve the system to end up with $\alpha_0$=31.882, $\alpha_1$=0.488, and $\alpha_2$=0.005. Therefore, the non- linear model is:

$$Y = 31.882 + 0.488x_i - 0.005x_i^2 . \quad (13)$$

To discuss how we determine whether or not we have a good fit with the above equation, we calculate what is called determination coefficient

$$R^2=1-SSE/SST \quad (14)$$

$SSE = \sum (y_i - \hat{y}_i)^2$ =170736538,

$SSR = \sum (\hat{y}_i - \bar{y}_i)^2$ =570.1260862,

$SST=SSR+SSE=\sum(y_i - \bar{y}_i)^2$ =2310.979167

$R^2$=0.262 => r= ± 0.511859.

Such that y is the dependent variable (number of accidents) and x is the independent variable (month). The model parameters estimation can be shown in Table (3).

Table 3: Parameters Estimation two Variables

| Regression Type | Model statistics | | | | Parameters Estimation | | | |
|---|---|---|---|---|---|---|---|---|
| Quadratic | R2 | F | DF1 | DF2 | Sig. | Constant | b1 | b2 |
| | 0.262 | 7.983 | 2 | 45 | 0.001 | 31.882 | 0.488 | -0.005 |

Since the value of the determination coefficient is not equal to zero, there is a relationship between the number of accidents and the month during which the accidents occur.

*C. Testing of Three Variables Regression Model*

A non-linear regression may have the general equation as in (7). Now we should compute the values of four parameters $\alpha_0,\ \alpha_1,\ \alpha_2$ and $\alpha_3$ using least square algorithm as follows:

1. Compute $\sum x_i$, $\sum x_i^2$, $\sum x_i^3$, $\sum x_i^4$, $\sum x_i^5$, $\sum x_i^6$, $\sum y_i$, $\sum x_i y_i$, $\sum x_i^2 y_i$, $\sum x_i^3 y_i$.

2. Solve the system

$$AX=b \quad (15)$$

Such that:

$$A=\begin{pmatrix} m & \sum x & \sum x^2 & \sum x^3 \\ \sum x & \sum x^2 & \sum x^3 & \sum x^4 \\ \sum x^2 & \sum x^3 & \sum x^4 & \sum x^5 \\ \sum x^3 & \sum x^4 & \sum x^5 & \sum x^6 \end{pmatrix} \quad x=\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} \quad b=\begin{pmatrix} \sum y_i \\ \sum x y_i \\ \sum x^2 y_i \\ \sum x^3 y_i \end{pmatrix} \quad (16)$$

3. Expand Table (1) to include $x_i^5$, $x_i^6$ and $x_i^3\, y_i$

4. Use the values

m=48, $\sum x_i$=1176, $\sum x_i^2$=38024, $\sum x_i^3$ =1382976, $\sum x_i^4$=53651864, $\sum x_i^5$=2168045376, $\sum x_i^6$=90109584824, $\sum x_i y_i$ =49310,

$\sum x_i^2 y_i$=1625698, $\sum x_i^3 y_i$=59554232 and $\sum y_i$ =1919.

5. Substitute for the above variables and solve the system to get:

$\alpha_0$=36.463, $\alpha_1$=-0.579, $\alpha_2$=0.049, and $\alpha_3$=0.001

Therefore, the nonlinear model is:

$$y = 36.463 - 0.579x_i + 0.049x_i^2 + 0.001x_i^3 \quad (17)$$

To discuss how we determine whether or not we have a good fit with the above equation, we calculate what is called determination coefficient

$$R^2 = 1-SSE/SST \quad (18)$$

$SSE = \sum (y_i - \hat{y}_i)^2$ =1617.7493,

$SSR = \sum (\hat{y}_i - \bar{y}_i)^2$ =627.8376,

$SST=SSR+SSE = \sum (y_i - \bar{y}_i)^2$ = 2310.9792

$R^2$=0.310 => r = ± 0.557 such that y is the dependent variable (number of accidents) and x is the independent variable (month). The model parameters estimation for three variables Regression Model can be shown in Table (4).

Table 4: Parameters Estimation Three Variables

| Regression Type | Model statistics | | | | Parameters Estimation | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Cubic | R2 | F | DF1 | DF2 | Sig. | Constant | b1 | b2 | b3 |
| | 0.310 | 6.601 | 3 | 44 | 0.001 | 36.463 | -0.579 | 0.049 | 0.001 |

Since the value of the determination coefficient is not equal to zero, there is a relationship between the number of accidents and the month during which the accidents occur.

## IV. PROPOSED EXPERIMENT

This section proposes a method to detect and classify traffic accidents. On the basis of the above investigation, it is reasonable that the Three Variables Regression Model is a suitable model in our experiment. The primary goal of implementing regression model is to predict the number of traffic accidents. This goal is achieved in steps which provide the core capability of generalizing large numbers of specific facts into new knowledge. Implementation process includes Least Square Algorithm which arranges data in matrices: **A** and **b**. Both matrices were used by Gauss Elimination and Back-Substitution Algorithms to generate the regression model. The later algorithm is used in the prediction algorithm. In fact, Least Square Algorithm in turn includes the following detailed steps:

1. Reading the Training Data Set which is collected from different police stations in Sana'a the capital of Yemen as shown in function Fig. 2:

```
Functions f = new Functions(l,max);
  for (int j = 0; j < max; j++)
    f[j] = j + 1;
      public double this[int i]
    {
      get{ return x[i];}
      set{ x[i] = value;}
    }
```

Figure 2: Reading Data Set Algorithm

2. Generating the matrix as shown in Fig. 3.

$$\left[ S= m, \sum x_i, \sum x_i^2, \sum x_i^3, \sum x_i^n \right] \quad (19)$$

♦Computing the $(x[i])^n$.

♦Computing summation of $x_i$

♦Saving the result in the variable sum.

3. Generating the matrix

$$\left[ b=\sum y_i, \sum x_i y_i, \sum x_i^2 y_i, \sum x_i^n y_i, \sum x_i^n \right] \quad (20)$$

♦ Computing (x[i],y[i]) sum.

♦ Setting **sumy** to zero.

♦Filling the array **x[i]** using the function **p.**

♦Computing **sumy**=**x[i]*y[i]** for all i=0,1,2,. , m

♦Filling the matrix **b[i]** using the function **sumy(i)** for every i=0,1,2,…,n+1.

4. Generating the Extended Matrix as indicated in Fig. 4.

```
public void FillS(int n)
 {
   for(int i=0;i<2*n+1;i++)
         s[i]=sumx(i);
 }
public double sumx(int n)
 {
   double sum=0;
   for(int k=0;k<y.Length;k++)
    {
       sumxi[k] = x[k];
       for(int i=0;i<y.Length;i++)
        {
          sumxi[i]=p(n,i);
          sum +=sumxi[i];
        }
       return(sum);
    }
 }
public double p(int n,int i)
{
  double help=1;
  for(int k=1;k<=n;k++)
   help *=x[i];
  return(help);
}
```

Figure 3: Generating the matrix S Algorithm

```
public void FillA(int n)
 {
   for(int i=0;i<=n+1;i++)
    for (int j = 0; j <= n+1 ; j++)
     {
      if (j < n + 1)
        a[i, j] = s[i + j];
      else
        a[i, j ] = b[i];
     }
 }
```

Figure 4: Extended Matrix Algorithm

$$A/b = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{in} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ & & & & \\ a_{n1} & a_{n2} & \dots & a_{nm} & b_n \end{pmatrix}$$

Figure 5: Extended Matrix

Such that the last column represents matrix **b** items while other columns represent matrix S items.

5. Gauss Elimination Algorithm Concept [1]:

Gauss Elimination Algorithm converts the matrix **A** (equation system) into an upper triangle matrix as follows:

♦Put the coefficients of both sides into one Matrix

$$A/b = X \Rightarrow \begin{pmatrix} a_{11} & a_{12} & \dots & a_{in} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ & & & & \\ a_{n1} & a_{n2} & \dots & a_{nm} & b_n \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \\ x_n \end{pmatrix}$$

Figure 6: Coefficients Matrix

♦Eliminate $\mathbf{x_1}$ from within all equations except the first one by putting their coefficients=0, using matrix Factorial $M_{ik}=\mathbf{a_{ik}/a_{kk}}$, multiplying $\mathbf{M_{ik}}$ row 1 and subtracting it from row **i** till we get:

$$\left(A/b\right) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{in} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ & & & & \\ a_{n1} & a_{n2} & \dots & a_{nm} & b_n \end{pmatrix}$$

Figure 7: Eliminated Matrix

Such algorithm:

♦Accepts the equation degree **n** and the coefficients $(A\backslash b)_{i,j}$ as inputs.

♦Computes $M_{i,j}=(A\backslash b)_{i,j}/(A\backslash b)_{i,j}$ for i=0,1,2,…

n+1 and for j=i+1 ,i+2,…,i+1.

♦Computes $=(A\backslash b)_{i,k} = (A\backslash b)_{i,k} - M_{i,j} * (A\backslash b)_{i,k}$

♦Saves the result {Ai,j,bi,n} in the matrix a[i,k] to be used by Back Substitution Algorithm.

6. Back Substitution Algorithm

This algorithm computes the coefficients of the regression model based on the following steps:

♦Computes $\mathbf{X_n = b_n/A_{n.n}}$

♦Computes $\mathbf{X_{n-1} = (b_{n-1} - A_{n-1.n} * X_n)/A_{n-1,n-1}}$ using $\mathbf{X_n}$ obtained so far.

♦Compute $\mathbf{X_i} = (\mathbf{b} - \sum A_{i,j}x_j)/A_{i,j}$ for i=n-1, n-2,..,1 ,0 in a bottom up process.

```
public void  BackSub(int n)
 {
  for (int i = 0; i < b.Length; i++)
     b[i] = a[i, n+1];
  if(a[n,n] !=0)
    sol[n]=b[n]/a[n,n];
  for(int i=n-1;i>=0;i--)
   {
    double sum = 0;
            for(int j=i+1;j<=n;j++)
     {
              sum +=a[i,j]*sol[j];
     sol[i]=(b[i]-sum)/a[i,i];
     }
  `
```

Figure 8: Back Substitutes Algorithm

In fact, the above algorithm performs the following:

♦Computing $X_n=b_n/A_n$ based on the coefficients computed by Gauss Elimination Algorithm.

♦Setting $X_i=\mathbf{b_i}$ for every i=n-1,n-2,…,1,0.

♦Computing $X_i=(X_i-A_{ij}*X_j)/A_{,j})$, for every i=n-1,n-2,…,1,0.

♦Saving the computed result $\mathbf{X=(x_0, x_1,\dots,x_n)^t}$ in the array **sol[i]**. Fig. 8. shows the algorithm details.

7. Prediction Algorithm

After obtaining the regression model as in 17:

$$y = 36463 - 0.579x_i + 0.049x_i^2 + 0.001x_i^3 \quad (21)$$

By estimating the dependent variable $\mathbf{y_i}$ through:

♦Computing **variable x** month between the base year and **prediction month** based on the formula as shown in 22:

**x=12*(target year-base year)+prediction month** (22)

♦Substituting values in the regression model as in 21 to obtain estimated $y_i$ as the number of traffic accidents for each month.

♦Computing moment M using the formula:

$$M=(\square / y_i)*100 \quad (23)$$

Such that □ is the mean of the actual accident value & $y_i$ is the estimated value.

♦Tabulating the result for 4 years, as appeared in Table (5).

Table 5: Predicted Moments.

| | 2015 | 2016 | 2017 | 2018 | Total | Average |
|---|---|---|---|---|---|---|
| Jan | 108.60 | 92.51 | 101.19 | 75.82 | 378.12 | 94.53 |
| Feb | 81.73 | 80.50 | 104.66 | 79.96 | 346.86 | 86.71 |
| Mar | 93.91 | 96.11 | 117.48 | 95.00 | 402.49 | 100.62 |
| Abr | 80.27 | 103.14 | 97.52 | 95.05 | 375.98 | 93.99 |
| May | 115.25 | 91.15 | 117.11 | 88.75 | 412.26 | 103.06 |
| Jun | 118.47 | 76.76 | 99.98 | 78.21 | 373.42 | 93.36 |
| Jul | 124.36 | 130.61 | 110.27 | 109.22 | 474.45 | 118.61 |
| Aug | 98.22 | 103.09 | 115.99 | 110.02 | 427.31 | 106.83 |
| Seb | 106.58 | 86.44 | 104.00 | 115.48 | 412.50 | 103.13 |
| Oct | 111.82 | 105.33 | 74.71 | 110.08 | 401.49 | 100.48 |
| Nov | 91.19 | 91.54 | 107.04 | 93.38 | 383.14 | 95.79 |
| Dec | 93.33 | 78.11 | 95.66 | 71.77 | 338.86 | 84.72 |
| | | | | | | 1181.83 |

♦Computing the moment component for a specific period for every month from the table (5) using the following formula:

$$Mc = \frac{Mx}{Average} * 1200 \quad (24)$$

♦Finally the algorithm returns the result as shown in Table (6).

Table 6: Moment Components

| Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 95.98 | 88.04 | 102.1 | 95.43 | 104.7 | 94.79 | 120.4 | 108.4 | 104.71 | 102.0 | 97.2 | 86.0 | 1200 |

♦ Estimating the number of accidents that occur during the month in any specific year using the formula:

$$\mathbf{Y=y_i*M_c/100} \quad (25)$$

Such that $y_i$ is the estimated value according to 21. & $M_c$ as computed in table (6).

Example to predict the number of accidents (Y) for May 2015, then we compute $y_i$ as in 21 and also pick the value of $M_c$ from table (6) as:

Y=34.918*104.7/100 = 36.559 accidents.

## V. CONCLUSION

This paper develops & builds a nonlinear regression (with three variables) mathematical model to predict the number of traffic passage accidents in the capital city of Yemen and to create a certain procedure. This model is implemented programmatically after investigating the model accuracy. The result of the experiment has shown good performance in terms of estimating the number of traffic accidents occur a during a specific month.

Therefore, these statistical measures give a good indication for the government to improve the traffic system utilities, such improvement could reduce the number of traffic accidents. The traffic accidents reduction in turn reduce number of deaths, injures and different types of loses in general.

## VI. RECOMMENDATIONS AND FUTURE RESEARCH ASSIGNMENTS

The government should raise the awareness of traffic rules of law and traffic safety through:

- Teaching vehicle drivers traffic rules and traffic safety before issuing them driving licenses.
- Specifying some media programs to show traffic accidents happened in the country.
- Including some topics related to traffic safety in school curriculum contents.
- Planning and maintaining different roads.
- Restricting drivers to implement the international standard of vehicle speed, especially within crowded cities.
- Setting sings along the roads to reflect the nature of the road.
- Imposing penalties on those who violate the traffic rules of law.
- Suspending issuing driving-licenses to those who are under age.
- Suspending old vehicles from operating within cities.

By improving data collection methodology and data classification, in future we can use such data as a training data set to compare some algorithms performance such as Bayesian classifier, decision tree, and artificial neural networks.

## REFERENCES

[1] John O. Rawlings, Sastry G. Pantula and David A. Dickey, Applied Regression Analysis: A Research Tool, 2nd Ed. Springer-Verlag New York Berlin Heidelberg, USA, 1998.

[2] David Sims, CDS Urban Development Expert,"Sana'a Priority Capital Investment Plan", CDS Urban Development Expert, 2010–2017.

[3] Doris Schopper Jean-Dominique Lormand and Rick Waxweiler,"Developing Policies to Prevent Injuries and Violence", World Health Organization 2006.

[4] https://openknowledge.worldbank.org/bitstream/handle/10986/2981/

[5] Aows N. Altef and Mojtaba Zourbakhsh, "An Overview of Urban Transport in Sana'a (Yemen)", Research Journal of Applied Sciences, Engineering and Technology 6(15): 2773- 2776, 2013.

[6] World Bank Document,"Republic of Yemen Urban Transport in Sana'a Strategy Note", Sep. 2010, Report No. 49176-YE.

[7] http://yemenpost.net/Detail123456789.aspx?ID=3&SubID=1483

[8] http://www.albaldnews.com/news3963.html

[9] http://www.qary.net/news/99/274818/

[10] http://anawen.net/show112419.html.

[11] http://www.albawaba.com/business/unemployment-yemen-476720

[12] http://documents.worldbank.org/curated/en/2010/09/14106644/

[13] Lary Christensen and Charles M. Stoup, Introduction to Statistics for the Social and Behavioral Science, Brooks/Cole Publisher Company, 1986.

[14] Michael McMillan, Data Structures and Algorithms, Cambridge University Press, 2007.

[15] Deitel and Deitel, C# 2010 For Programmers Fourth Edition, Pearson Education, Inc, 2010.

## AUTHORS PROFILE

**Dr. Sharaf Abdulhak Alhomdy**, born in 20/01/1971, Alsena, Taiz, Republic of Yemen. Hold a Ph.D. in Computer Scince, Pune University, India, 2009. Assistant Prof. & Vice-Dean for students affairs, Faculty of Computer and Information Technology, Sana'a University, Republic of Yemen.
E-mail: sharafalhomdy@gmail.com

**Dr. Abdulwase M. Al-Azzani**, born in 12/08/1971, Saber,Taiz, Republic of Yemen. Hold a Ph.D. in Computer Scince, Technology University, Iraq, 2004. Assistant Prof. at the Dep. of Computer Science, Faculty of Computer and Information Technology, Sana'a University, Republic of Yemen.
E-mail: E-mail: amalezzani71@gmail.com.

**Dr Ghaleb H. Al-Gaphari**, born in 20/01/1965, Taiz, Republic of Yemen. Professor & Vice-Dean for Academic Affairs, Faculty of Computer and Information Technology, Sana'a University, Yemen.
E-mail: drghalebh@gmail.com