Comparison Between Traditional Time Series Forecasting Models: An applied Study for Primary Schools Students- Iraq/ Karbala Governorate Students as a Sample

Assistant teacher Daham Owaid Matrood
Assistant Teacher of Northern Technical University/Iraq
daham.stat@ntu.edu.iq

Extract: Education is a prerequisite for improving the standard of living, empowering women, protecting children from the harsh effects of child labor and sexual violence, supporting human rights and democracy, protecting the environment. In addition, the education is guiding population growth universal access to primary education for the world's children is one of the Millennium Development Goals (MDG'S) Objectives of "A World Fit for Children" (WFFC). The discrepancy in the quality of the estimated regression models and the inability to use some of them because they do not have the characteristics of good estimators, which leads to a lack of confidence in their predictive or estimation accuracy. The research aims to estimate the general trend regression models using the ordinary least squares method and compare the results of the estimation using the differentiation criteria (AIC, BIC, MSE) and to determine the optimal model as well as to predict the number of primary students in the holy province of Karbala for the time period (2022-2028). The researcher found the best suitable model for prediction, which is the general linear trend model, table 4, which represents the predictive values of primary students.

Keywords: Time Series, Least squares method, Linear trend model, Forecasting Models.

1.Introduction :

The relatively high population growth rate (2.3%) for the year 2020, as well as the obligation of free education all of these combined factors, in addition to other factors, lead to the emergence of a real problem in preparing students for the primary grade, and the obligations associated with this preparation that must be planned and confronted, such as providing adequate preparation of schools, educational staff, free books, and others through predicting the preparation of these students in different ways.

Time series are among the most important modern statistical methods, which make it possible to know the nature of the changes that occur in the values of the phenomenon with time. Thus, determining the causes and results, explain the observed relationships between them, and predict what will happen in the change in the values of the phenomenon in the future in the light of what happened to it in the past. There are a number of methods in time series, including the traditional time series method (linear, quadratic and exponential general trend equation), which will be used to predict the number of primary students in a governorate for the period of time (2004-2022)

2.Research Methodology:

The disparity in the quality of the estimated regression models and the inadequacy of using some of them because they do not have the characteristics of good estimators, which leads to a lack of confidence in their predictive or estimation accuracy, and the failure to use an accurate statistical tool to predict the numbers of students in government institutions according to the researcher's knowledge.

2.1.Objective of Research:

The research aims to:

- Estimation of general trend regression models using the ordinary least squares method.
- The comparison between the estimation results using the differentiation criteria (AIC, BIC, MSE) and determining the optimal model.
- Forecasting the number of primary students in the holy city of Karbala for the period (2023-2028).

2.2.Research Window:

The framework of the research chronologically was the number of primary students for the period (2004-2020), spatially representing the number of primary students in the holy governorate of Karbala.

3.The theoretical side

3.1.Time Series: [9] [8][7]

There are several definitions of time series, including:

- It is a set of observations generated consecutively over time.
- It is a set of related observations recorded in successive periods for a phenomenon.
- Mathematically: We say that the independent time variable is (t_i) and its corresponding values represent the dependent variable (y_i), and that each value in time (y_i) has corresponding values for the dependent variable (y_i).

Then y is a function of time t, that is:

$$Y = F (t) \qquad \ldots(1)$$

3.2.Time series components: [12][6][5][3]

The value of the time series in a specific period determined by the effect of certain changes, which are:

a. Secular Trend:

It is the trend were the time series takes over a long period, and its effect appears after a relatively longer period compared to the effect of the rest of the compounds. The general trend may be in continuous growth (such as population growth). The general trend of the series may be in a contraction (in decrease) as the number of illiterates in a particular society decreases. The general trend of the series, takes the form of increase and decrease or vice versa, such as the price of crude oil at a quarterly rate.

The general trend measures the average change for each period, and it may be a straight line or non-linear, such as the exponential curve (an irregular or unstable measure). The trend depends on long changes in the time series, which reflect the amount of growth and development related to positive or negative contraction for a long period. It represents the general trend line for the phenomenon and for the entire period.

b. Seasonal variations:

They are the changes that occur regularly every year, this is due to the natural conditions throughout the year, such as holidays or the beginning of the school year, and their duration is often less than a year, and they occur because of changes in the climate, social customs, or religious and national events.

c. Cyclical Variation:

What is meant by periodicity is the changes that occur in the values of the time series, which differ from seasonal changes in that they occur in a period of time longer than a year, in addition to that, as usual, they do not occur in a regular period, for example, the recession cycle or the global economic inflation cycle.

d. Irregular Variations (Random):

What is meant by randomness is the changes that occur in the values of the time series, which are the result of either by chance, abnormal conditions and in this case they cannot be predicted or determined, or are the result of certain sudden events such as (wars, earthquakes, ... etc.), and in this The condition is unpredictable but identifiable.

3.3.Time Series Models: [12][10][13][3]

There are many models used in forecasting, and the research dealt with the following models, some models of the general linear trend and its agencies.

1. **Linear Trend:**

It means that the observations of the studied phenomenon increase or decrease by a fixed amount during a specific period. In light of this, the general trend equation takes the form of a straight line according to the following formula:

$Y = B_0 + B_1 * t_i + u_i$ ... (2)

2. **Trend Quadratic:**

That is, the studied phenomenon has a general, non-linear trend as in economic phenomena, the series of observations takes the form of a curve of the second degree, and the formula of the model is as follows:

$Y_t = B_0 + B_1 * t + B_2 * {t_i}^2 + u_i$ ... (3)

3. **Exponential Trend:**

It is the model, which the values of the time series observations take an exponential form, i.e. non-linear, and its mathematical formula is as follows:

Yt = B0 $*{B_1}^{ti}$ *ui …(4)

By taking the natural logarithm of both sides of the above equation, we get:

Ln(Yt)= Ln(B0) + ti Ln(B1) +Ln(ui) …(5)

Thus, the exponential equation becomes a linear equation of the first degree. Where:

Yt: the value of the phenomenon studied.
ti: : time.
(B2, B1, B0): Parameters of the model.
Ui: random error.

4.General trend measurement: [13][2][1]

There are several methods for measuring the general trend, such as the moving averages method, smoothing by hand method, least squares method and other methods.

- **Least squares method:**

It stands for OLS, an acronym for Ordinary Least Square It is the most widely used method for estimating parameters. This method is base on the principle of (minimizing the sum of the squares of errors); as it seeks to find the parameters that make the sum of the squares of error as few as possible.

$Y = B_0 + B_1 * t_i + u_i$

Since:
Y_t: The vector values of the time series t.
b_0: The point of intersection of the trend line with the y-axis.
b_1: The slope of the general trend line.
e_i: Random error.

$\hat{u}_i = Y_i - \hat{Y}_i$ …(6)

$$\hat{u}_i = \sum_{t=1}^{n} \left(Y_i - \hat{B}_0 - \hat{B}_1 t \ \right) 2 \qquad \dots (\ 7\ )$$

By partially differentiating the above equation with respect to $\hat{B}_1, \hat{B}_0$respectively, and setting them to zero, we get an unbiased estimate of the parameters $\hat{B}_1, \hat{B}_0$, that is:

$$\frac{\partial ei}{\partial \hat{B}_0} = -2 \sum_{i=1}^{n} \left(Yi - \hat{B}_0 - \hat{B}_1 t_i\right) = 0 \qquad \dots (8\ )$$

By dividing both sides of the equation by the sample size n, we get an estimate of ($\hat{B}_0$) as in the following formula:

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{t} \qquad \dots . (\ 8\ )$$

$$\frac{\partial \hat{u}_i}{\partial \hat{B}_1} = -2 \sum_{i=1}^{n} \left(Yi - \hat{B}_0 - \hat{B}_1 t_i\right) \times t_i \ = 0 \qquad \dots (9)$$

From this, we get an estimate of the marginal slope parameter of the general trend line:

$$\hat{B}_1 = \frac{\sum_{t=1}^{n} t y_i - n\bar{t}\bar{y}}{\sum_{t=1}^{n} t^2 - n\bar{t}^2} \qquad \dots . (\ 10)$$

5.The criterion of comparison between models: [11][4]

The general trend models compared using several proxy criteria:

5.1 Mean Square Error:

It is the average of the squares of deviations from the real values, and whenever the quantity of the mean squares of error is close to zero, this indicates that the estimated values of the series are close to the real observations of the time series, and its mathematical formula is:

$$MSE = \frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 \qquad \dots (\ 11\ )$$

5.2 Mean Absolute Deviation:

It is the second criterion for measuring prediction efficiency, and defined as the average of the absolute valuesof random errors, and its mathematical formula is:

$$MEA = \frac{1}{n} \sum_{i=1}^{n} |y_t - \hat{y}_t|^2 \qquad \dots (12)$$

5.3 Mean Absolute percentage Error:

It is the third criterion for measuring the efficiency of forecasting, and it defined as summing the product of dividing the absolute value of random errors by the values of the time series, multiplying the result by 100, and dividing the result by the number of series observations. The mathematical formula is:

$$MAPE = \frac{\left[\frac{\sum_{i=1}^{n} |y_t - \hat{y}_t|}{y_t}\right] * 100}{n} \qquad \dots (13)$$

Since

Y_t: The vector values of the time series t.

$\hat{y}_t$: The estimated values of the time series t.

n: is the sample size.

6.Practical side:

After the process of obtaining data for the time period (2004 - 2020) represented by the numbers of primary students, the data was analyzed using the statistical program (Minitab-19) and my agencies:

6.1.Choosing the best general trend model:

It is possible to know the best model by performing the following statistical analysis:

a.Linear trend model:

The results of the statistical analysis of the number of primary students showed the following:

- • The estimated model (significant) found to be suitable for prediction, as the calculated F-test value appeared equal to (1236.82) with significance (0.000), which is less than the significant level (0.05).
- The value of the parameter fixed boundary appeared equal to (b_0 = 116801), while the value of the parameter marginal slope appeared equal to (b_1 = 9474), and with significance (0.000), which is less than the level of significance (0.05), and accordingly, the estimated general linear trend model can be written as follows:
  y_t= 116801+9474 t_i
- The value of the coefficient of determination (R^2) appeared equal to (0.99). This is an indication of the quality of the estimated regression equation on its explanatory ability for the relationship between the two variables of time and the numbers of students (t_i,yi), and it was found that the time variable exerts its influence on the variable number of students by ((99%).
- The value of the mean square error (MSE=26127867) appeared.
- The value of the absolute mean of deviations appeared (MEA = 4012).
- Mean Absolute Relative Errors (MAPE=2).

Table.1. It shows the results of the statistical analysis, not the general linear trend model, not the student counter

F	F	$MAPE$	MEA	MSE	R^2	b_1	b_0
0.000	*1236.82*	2	4012	26127867	*0.*99	9474	116801

b.The general quadratic trend model:

The results of the statistical analysis of the number of primary students showed the following:

- The value of the parameter fixed term appeared equal to (b_0=112758) with significance (0.000), which is less than the level of significance (0.05). While the value of the marginal slope parameter appeared equal to (b_1= 10751,) with the significance of the probability value (00.0), which is less than the level of significance (0.05). Thus, the value of the marginal slope parameter appeared equal to (b_2= 70.9) with significance (0.265), which is greater than the level of significance (0.05), which leads to the absence of any significant effect of the time variables $(t_i{}^2)$on the variable number of students (y_i). Therefore, the model is not suitable for prediction, so it cannot be relied upon despite the appearance of the calculated F-test value equal to (633.41) with a significance (0.000), which is less than a significant level (0.05) and by deleting the variable $(t_i{}^2)$ we get a simple linear regression model.

6.2.Logarithmic general trend model (exponential function):

The results of the statistical analysis of the number of primary students showed the following:

- It turns out that the estimated model (significant) is suitable for prediction, as the calculated F-test value appeared equal to (582.46) with a significant probability value (0.000), which is less than the level of significance (0.05).
- The value of the parameter fixed limit appeared equal to (b_0= 11.7496), while the value of the parameter marginal slope appeared equal to (b_1=0.04872), and with significance (0.000) for both parameters, which is less than the level of significance (0.05). Therefore, the general trend model of the estimated exponential function can be written as follows:
  $\ln y_t$= 11.7496+0.04872 t_i
- The value of the coefficient of determination (R^2) appeared equal to (0.98), and this is an indication of the quality of the estimated regression equation on its explanatory ability for the relationship between the two variables of time and the numbers of students (t_i , yi). In addition, it found that the time variable exerts its influence on the variable number of students by (98%).
- The mean value of the error squares appeared equal to (MSE=7046852).
- The value of the absolute mean of deviations appeared (MEA = 6724).
- Mean Absolute Relative Errors (MAPE=3)

Table.2. It shows the results of the statistical analysis, not the general linear trend model, not the student counter

F	F	$MAPE$	MEA	MSE	R^2	b_1	b_0
0.000	*582.46*	3	6724	70468531	*0.98*	0.04872	11.7496

6.3. Choosing the best-estimated trend model:

By comparing the estimated models, it was found that the general linear trend model is the best for having the smallest comparison criteria (MAPE, MAPE, MSE).

6.4. Prediction:

After the model showed the general linear trend as the best, it was used in the following predictions:

- **Predicting the number of future students:**

Looking at Table (3), we note the predictive values for the period (-2022-2028) are annual upward values, and it is expected that the number of students predicted will reach (353662) students until 2022.

Table.3. Predictive Values for Primary Pupil Count for the period (2022-2028)

Year	2028	2027	2026	2025	2024	2023	2022
Predicted number	353662	344188	334713	325239	*315765*	306290	*296816*

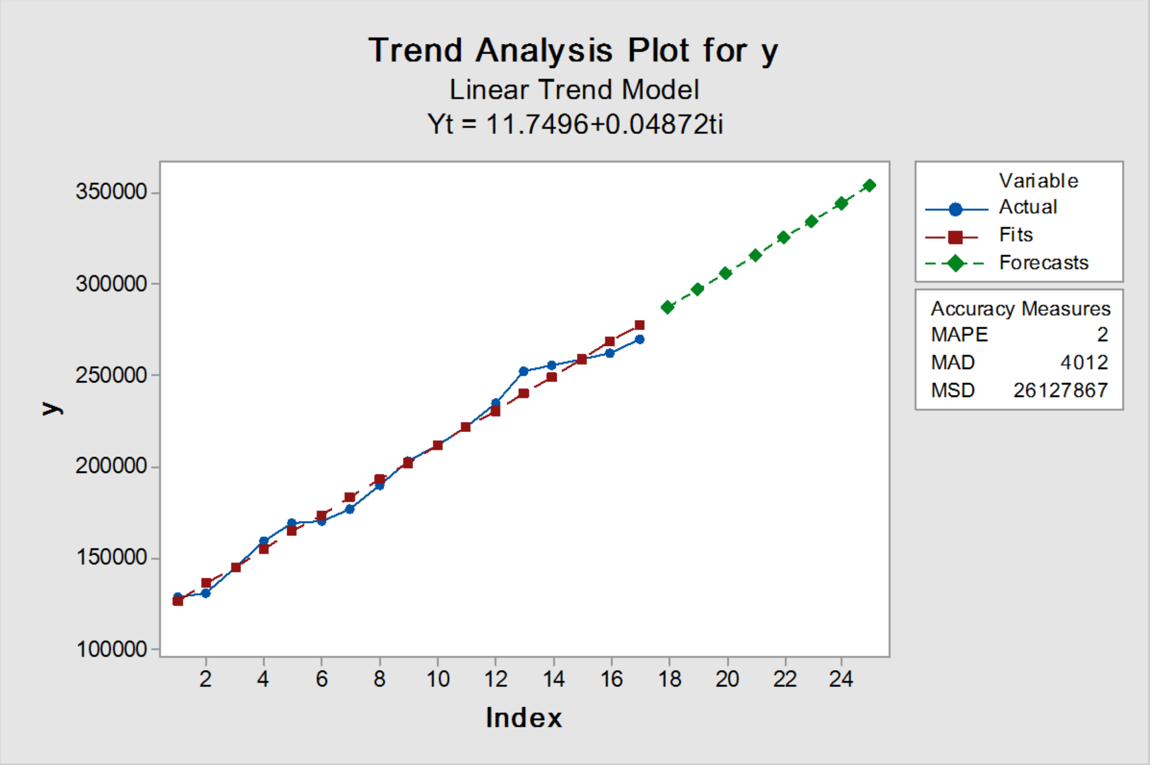

Figure.1. Plotting the predictive values of the series of primary students' numbers

7. Conclusions:

After conducting a statistical analysis using the Mintab-17 program for a series of primary students in the Holy Karbala governorate, the researcher reached several conclusions, the most important of which are:

- Only general trend models (linear + exponential) appeared suitable for forecasting.
- The quadratic model appeared inappropriate for prediction.
- It turns out that the general linear trend model is the best suitable model for forecasting.
- Predicting the number of primary students for the period (2028-2022).

8. Recommendations:

- Benefit from the research results from the relevant authorities.
- Predicting the numbers of primary students in the Holy Karbala governorate according to the sectors, as one district includes a group of partitions.
- Using other time series methods in the field of prediction and comparison with the methods used in the research.
- Conducting similar studies for all governorates of Iraq.

References

Al-Tamimi. Zahra & Hassan Abbas. eta, 2014, “The Introduction to Regression Analysis”, Dar Al-Kutub for Printing and Publishing, Mosul.

Suleiman, Osama Rabie Suleiman.2007, ‘Researchers Guide to Statistical Analysis of Data Using the (Minitab) Program’, Menoufia University ,Egypt.

Al-Senussi, Abu Al-Qasim& Al-Mousawi, Kamal Gallab,2005 “Time-Series Analysis of Medicinal Drug Disbursement Data”. a research published in Al-Satell magazine,Libya

Al-Shamrani, Muhammad Musa, 2013 “A comparison between some traditional statistical methods and Jenkins Box models in analyzing time series data”, Umm Al-Qura Journal for Educational and Psychological Sciences, Volume Five, First Issue, KSA.

Al-Sarraf, Nizar Mustafa & Shoman, Abdul Latif Hassan, 2013, “Time Series and Indices”, Dr.’s House for Administrative and Economic Sciences, first edition, Baghdad, Iraq.

Al-Ani, Ahmed, Hussein Battal.2017, "The use of ARIMA models in economic forecasting", research published by the College of Administration and Economics, University of Anbar.

Ghoneim, Othman Muhammad,2011, Planning Standards (their philosophy, types, preparation methodology and applications in the field of urban planning), Dar Al-Safa Publishing and Distribution, Amman, Jordan.

Mazen, Abdel Rahman Al-Heity,2013, Geography of Services, Foundations and Concepts, Arab Society Library for Publishing and Distribution, Amman, Jordan.

Al-Haymes, Fatima Faisal,1982 ‘Spectral Analysis of Time Series with Application in Geology’, College of Administration and Economics, University of Baghdad,.

AL-Nasser ,A. H & JUma,A. A (2013) "Interoduction To Applied Time Series Analysis ", AL-jazeera Bureau for printing and publishing.

John .O. Rawlings & others, (1998),"Applied regression analysis Aresearch Tool” North Carolina State University , USA , Second Edition.

Robert.S.Pindyck & Daniel. I. Rubinfeld, (2000), "Econometric models and economic forecasts), New York , McGraw-Hill Book company . Second Edition.

William R.Bell and Steven C. Hillmer,(1984) (Seasonal Adjustment of Economic Time Series) Statistical Research Series Census/SRD/RR- March 30,.