RES MILITARIS
REVUE EUROPEENNE D ETUDES EUROPEAN JOURNAL OF MILITARY STUDIES
Social Science Journal

Determine and build the best model using time series to predict the number of people with malignant tumors in Missan Governorate

By

Researcher: Lama Tariq Abbas
College of Administration and Economics / Al-Mustansiriya University
Email: Lumaamer67@gmail.com

Teacher Assistant. Azhar Kazem Jabara
College of Administration and Economics / Al-Mustansiriya University
Email: azkdf_2017@uomustansiriyah.edu.iq

Introduction

cancerIt is a disease that occurs due to the presence of abnormal cells that divide without control and are able to spread to the rest of the body's organs100 kinds and not there. One cause of cancer, it may occur due to various factors, and the causes of cancer are still unknown to a large extent, and the issue of cancer in Iraq has had a degree of importance among the medical community..Iraqi since the second half of the last century resulted in the establishment of the Iraqi Cancer Society during The decade of the sixties, and then the establishment of the cancer registry in 1976. After the incidence of the disease began to rise, it became important to prepare a unified study and conduct a scientific and detailed evaluation of the cancer problem in order to develop a comprehensive strategy in terms of planning and implementation to control cancer, especially after the environmental and health problems resulting from it. wars that passed through the country.

Search target

Study chains temporal(Time Series)And that To determine better And more efficient model statistical for a purpose use it to predict in numbers injured tumors malicious in governorate Maysan for the period(2007-2013) -

Time series: Time Series

A time series can be defined as a set of observations of the values of a phenomenon taken at specific times (the intervals between observation and the next may be equal or unequal, and are usually equal). If they are equal, then express them (Ztl, Zt2, ... Zin) at time intervals t, t2, ..., tn thennnRepresents the number of observed valuesstat string (StatisticalSeries) in the form the next

$$t = 012... \pm \pm \pm$$

$$zt=f(t)+aT$$

whereas :
(f(t): represents the regular part expressed by a mathematical function
e:It represents the random part and may be called noise (noise).
Two types of time series can be distinguished:

Stable time series, and unstable time series, as there are two states of stability, which are stability in the average (Stationary in Mean) and stability in variance. (Stationary in

Variance) The stability in the average in the case of the series when it does not show a general trend and can be converted to stable using the differences. As for the stability in the variance, it is the case of the series when it does not show varying fluctuations in the form of the time series, and the variance can be fixed by obtaining the natural logarithm, the square root, or the reciprocals of the series data.

Self-association Autocorrelation

It is an indicator that shows the degree of relationship between the values of the same variable at shift intervals(K)different. Its value ranges between)1 and 1-) E1≤ 1 ≤ PK-Estimates the following formula

$$\hat{p}_k = \sum_{t=1}^{n-k} (z_{t-i})\,(z_{t+k} - z)$$

whereZT Series Views Rate

Z: represents the arithmetic mean and is equal to$\bar{z} = \sum_{i=1}^{n} \frac{z_i}{n}$

Also, the statistical distribution of the autocorrelation coefficients is a normal distribution with zero arithmetic mean and variance(1/N) where N is the sample size:$p_k \sim n\left(o_9 \frac{1}{n}\right)$ The graph of the coefficients of the model(PK) against the displacement intervals (K) where (K=1,2,3,....) is called the autocorrelation function, which is denoted by (ACF)

The autocorrelation function (ACF) as an important means to know the stability of the time series, as it tends either to decline quickly towards zero with increasing shift periods (k) or to break after a number of shift periods (k = q) aJan:

$$p_k = 0 \forall k > q$$

kg 0 = m Since the autocorrelation function of the sample is only estimates of the autocorrelations, its values are likely to be small and not zero, that is:

$$r_k \neq 0 \forall k > q$$

But if the time series is unstable due to an upward or downward trend in the rate, then the function (ACF) for the sample is not interrupted and does not slope slowly towards zero, because the witnesses tend to be on the same direction as the arithmetic mean of the time series for many periods of time, and as a result we get large autocorrelations at long offset periods.

It is the autocorrelation function of the remainderRACF (Residual Autocorrelation Function) is an important way to check the fit of the model by testing the randomness of the residual errors, where they are:

$$p = \begin{cases} 1k = 0 \\ 3k \neq 0 \end{cases}$$

Partial autocorrelation (Partial Autocorrelation (PACF).

It is an indicator that measures the relationship between 2 andk for the same series with the assumption of the stability of the time series values and is defined as the last term of the autoregressive model of degree (AR(P), and the values of the partial autocorrelation coefficient can be found by means of the Autocorrelation Function and according to the formula

$$\mathbf{q^{k+1K+1} = P^{K+1} - \frac{\Sigma_{j=1}^{k} \varphi^{KJ} P^{K+1-J}}{1 - \Sigma_{j=1}^{k} \varphi \hat{} KiP \hat{} i}}$$

The partial autocorrelation function (PACF) in time series analysis and is also used to diagnose the appropriate model from a group of stable random process models, determine its degree, and check its suitability for the sample data by testing the randomness of the residual errors. For a stable time series, the partial autocorrelation function (PACF) tends to decline rapidly towards zero with increasing shift intervals, or to break off after a certain number of shift intervals (k).

Stability of the time seriesStationary Time Series

The time series may be unstable in variance as well as unstable in average (it has no stability in the general trend)Trend), which is one of its elements), which makes it have several circles around which the data fluctuates even when the series is homogeneous. The background (Back Shift Differences Operator) is denoted by ( ) and is∇

$$\nabla z_t = z_t - z_t - 1 = (1 - B)z_t$$

Then the time series becomes stable after taking (d) of the differences i.e.:

$$z_t = \nabla^D z_t d1, \geq$$

As for the instability of the variance, it is treated by taking the natural logarithm of the series data, or by taking the square root of it, or the reciprocal of the data.

Box and Jenkins models (Box & Jenkins (B - J) for time series .autoregressive model (Autoregressive Model (AR

The general form of the degree autoregressive model (p) will take the following form: **(1-2)**

$$Z_t = \Phi_0 + \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \cdots + \Phi_p Z_{t-p} + a_t \qquad (1\text{-}2)$$

$$\Phi_p(B) Z_t = \Phi_0 + a_t \qquad)3(1\text{-}$$

whereas:

zt :Rate the series' views

ϕ_iParameters of the modeli=1,2,3,……

P: the degree of the form

Random errors that are normally distributed with a mean of zero and an equal variance σ_a^2

The autoregressive model can be used to represent stable and unstable time series, and the conditions for $\Phi_p(B) = 0$ achieving the stability of the model must be that the roots of the equation lie outside the limits of the unit circle, that is, $(-1 < \Phi_p < 1)$ that whereasB: Back shift operator, defined as follows:

$$B^k Z_t = Z_{t-k} \quad \forall k = 1, 2, \ldots$$

The autocorrelation function of the regression model (AR(p) is decreasing exponentiallya(exponential darnping) as the displacement periods (k) increase, while the partial autocorrelation function (PACF) discontinues after the period p (Saraf, 1981:17). There are two special cases of the general autoregressive formula (AR(p), namely, the first-order (1)AR and second-order (2)AR autoregressive models, which are commonly used models to represent most time series.

-33). In the case of (1 =p), then equation (2-1) becomes as follows :

$$Z_t = \Phi_0 + \Phi_{1_1} Z_{t-1} + a_t$$

which is a first-order autoregressive model (1)AR. The conditions for achieving stability in the model require that the roots $(\Phi_1(B) = 1 - \Phi_1 B = 0)$ of the equation be:outside the unit circle

$$-1 < \Phi_1 < 1:$$

and that the autocorrelation function (ACF) of the model is as follows:
This equation can be solved and obtained:

$$\rho_k = \Phi_1^k \qquad k = 0, 1, 2, \ldots$$

That is the autocorrelation function of a model(1)AR is exponentially declining when Φ_1) is positive and alternately Φ_1 slopes exponentially in the sign when ( ) is negative. The partial autocorrelation function (PACF) of the AR(1) model is:

$$\rho_{11} = \Phi_1$$
$$\rho_{kk} = 0 \qquad k > 1$$

So the partial autocorrelation function (PACF) breaks off after the first offset (1k>). And in the case of (2 = p) in equation (2-1), we obtain a second-order autoregressive model (2)AR, whose formula is as follows:

$$Z_t = \phi_0 + \Phi_{1\ 1} Z_{t-1} + \Phi_{1\ 2} Z_{t-2} + a_t$$

For the model to be (2)AR $(1 - \Phi_1 B - \Phi_2 B^2 = 0)$ is stable it is a must

The roots of the equation lie outside the unit circle, i.e. must be satisfiedthe two parameters (Φ_1, Φ_2) The following conditions:

$$\rho_k = \Phi_1 \rho_k + \Phi_2 \rho_{k-2} \qquad k > 0$$

and in (k=1,2) then: $\rho_2 = \Phi_1 \rho_1 + \Phi_2$

Note that D is the $\rho_1 = \Phi_1 + \Phi_2 \rho_1$ autocorrelation of the modelAR(2) diminishes if

$$\Phi_1^2 + 4\Phi_2 \geq 0$$

أما إذا كانت:

$$\Phi_1^2 + 4\Phi_2 < 0$$

The autocorrelation function (ACF) are wave damping sine waves. The partial autocorrelationsPACFfor model (2)AR can represent:

$$P_{11} = \frac{\Phi 1}{1-\Phi_1}\ ,\ P_{22} = \Phi_2\ ,P_{kk} = 0$$

So the partial autocorrelation function (PACF) of the (2)AR model is interrupted after the second offset, i.e.)(2K>).

Moving media model (Moving Average Model (MA).

The moving media model of degree (f) can be represented using the backscattering operator (B) as follows|

$$Z_t = \Phi_0 + \left(I - \Theta_1 B - \Theta_2 B^2 - \ \dots - \Theta_q B^q\right)a_t$$

And the general form of this form:

$$Z_t = \Phi_0 + a_t - \Theta_1 a_{t-1} - \Theta_2 a_{t-2} - \ \dots - \Theta_q a_{t-q}$$

whereas Θ_i Parameters of the moving media model

$$-1 < \Theta < 1 \text{ أن } , i = 1,2,3, \dots\dots, q$$

q: model score.

And the autocorrelation function of the model (MA) is intermittent or close to zero after displacement (q), while the partial autocorrelation function (PACF) diminishes exponentially.

Mixed model (autoregressive - moving media).

Mixed Autoregressive Moving Average Model (ARMA).)
The model can be written in the general form of degree (p, q) as follows (1976,BJ (

$$Z_t = \Phi_0 + \Phi_1 Z_{t-1} + \cdots + \Phi_p Z_{t-p} + a_t - \Theta_1 a_{t-1} - \ \dots - \Theta_q a_{t-q}$$

By using the backscatter operator (B):

$$\Phi_p(B) Z_t = \Phi_0 + \Theta_q(B) a_t$$

: is a polynomial in (B) for the parameters of the autoregressive model $\Phi_p(B)$

: is a $(\Theta^{d} \cdots\cdots \Theta^{7})\ \Phi_0)$ polynomial in (B) for the parameters of the $:\Theta_q(B)$ moving media model $(\Phi_p(B) = 0)$

In order for stability to be available in this model, the roots of the equation must be present $(\Theta_q(B) = 0)$

It is outside the bounds of the unit circle as well as for the roots of the equation

The integrated mixed model

Autoregressive Integrated Moving Average Models (ARIMA)

Some time series models may be unstable by themselves, but they become stable after many conversions or differences. Therefore, the model that expresses this process will differ from the original model, as it must include those transformations or differences that were made on the model. These stable models are called integrated mixed models. prepare models (ARIMA) is the most widely used time series model, as it is possible to derive All models, including autoregressive, dynamic or mixed media. These models consist of three parts, the first part of which is an autoregressive model (AR (p) which is usually used in the process of forecasting the time series, while the other part represents the moving media model (MA(q) and the third part (I(d) represents the differences required by the series in order to be stable (Stationary) and therefore it expresses the Auto Regressive Integrated Moving Average models) (Non-seasonal Models according to formula

p,d,g) ARIMA) where
P: is the rank of the autoregressive model (AR(p).
q: is the order of the moving media model (MA(q).
d: is the number of differences that make the series stable. Using the backscatter factor (B) in the following formula

$$\varphi(B)(1-B)^{r}X_t = \varphi_0 + \theta(B)a_t$$

حيث أن:

$$\phi(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$$

$$\theta(B) = (1 - \theta_1 B - \cdots - \theta_q B^q)$$

$$(1-B)^d = \nabla^d$$

$$\nabla^d X_t = Z_t$$: نفرض أن

The general formula for the mixed modelARIMA(P,D,q)

$$Z_t = \phi_0 + \phi_1 Z_{t-1} + \cdots + \phi_p Z_{t-p} + \cdots + dZ_{t-p-d} + a_t - \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}$$

Therefore, models can be consideredARIMA are stable ARMA models with different rank Building the time series model The time series model is built through four stages: diagnosis of the model fit to the data, estimation of the information of the diagnosed model, testing of the fit of the model, future prediction

Model diagnosisidentity

Diagnosing time series models is the most important step in building time series models, and the first stage of the algorithm on which the researchers laid the foundation.Box and Jenkins 1976; The diagnosis phase must precede the data preparation phase. If the data is stable by observing the drawing of the original data and its partial and self-correlations, then the data is prepared for diagnosis. But if the series is not stable in the middle, and the variance, then the instability in the middle is treated by taking the first difference (1 = d), and if it is not stable, we take the second difference (2 = d), and it often settles after the first or second difference. As for the instability in the variance, it is addressed by making the appropriate transformation of the data. After achieving the stability of the time series, the process of defining the model begins, and by that we mean the use of data or any information on how the time series is generated. The goal here is to get an idea of the value of p, d, and Which we need in the general linear model ARIMA whose formula is shown in Equation (7-1) and then obtain preliminary estimates of the model parameters.:

function statement (ACF decreases gradually and exponentially, or the behavior of the diminished sine function, and the statement of the (PACF) function is interrupted after the displacement (P). The appropriate model for the data is (AR(P).

function statement ((ACF) is cut off after the shift (q) and the statement of the (PACF) function gradually decreases exponentially or the behavior of the diminishing sine function, then the appropriate model for the data is (MA (q).

The statement of the function (ACF) and (PACF) is gradually decreasing in an exponential way or the behavior of the declining sine function, the model and the fit for the data is (q,ARMA (p)

AppreciationEstimation

The process of estimating the model is the second stage of studying and analyzing the time series, and it comes after the process of diagnosing the appropriate model for the time series, and in order for the model to achieve the main goal of its construction, which is prediction, we must ensure the quality of its estimate and its suitability for the time series, and there are several methods for estimating the parameters of the model most notable)

1. Ordinary least squares method (.Method of Ordinary Least Square (0.LSE).

This method is based on the principle of reducing the sum of the squares of the estimation error, and bringing it to its minimum.

2. The method of greatest possibilityMaximum Likelihood Method

The parameter matrix of the model to be estimated is chosen according to the principle of maximizing the possible function.

Forecastingforecasting

Forecasting is the last step in the study and analysis of time series models, and it is the main objective of the study. After determining the appropriate model for the data, it is used to know the values of the future phenomenon and for periods (L) and the prediction of the number of steps (L) can be calculated according to the formula:

$$\hat{Z}_{t+L} = E[Z_{t+L} | Z_t, Z_{t-1}, Z_{t-2}, \dots\dots] \quad \text{for } L \geq 1$$

So if the form (1)AR, the best predictor of the number of steps (L) is:

$$\hat{Z}_{t+L} = \Phi^L Z_{t-1+L} \qquad L \geq 1$$

But if the form (2)AR, the best predictor of the number of steps (L) is:

$$\hat{Z}_{t+L} = \Phi_1^L Z_{t-1+L} + \Phi_2^L Z_{t-2+L} \qquad L \geq 1$$

In the case of moving media (MA(q), the best predictor of the number of steps (L) is:

$$\hat{Z}_{t+L} = a_{t+L} - \theta_1^L a_{t-1+L} - \theta_2^L a_{t-2+L} - \cdots \theta_q^L a_{t-q+L}$$

In the case of the mixed model (ARMA(p , q), the best predictor of the number of steps (L) is:

$$\hat{Z}_{t+L} = \Phi_1^L Z_{t-1+L} + \Phi_2^L Z_{t-2+L} \qquad L \geq 1$$

The applied side

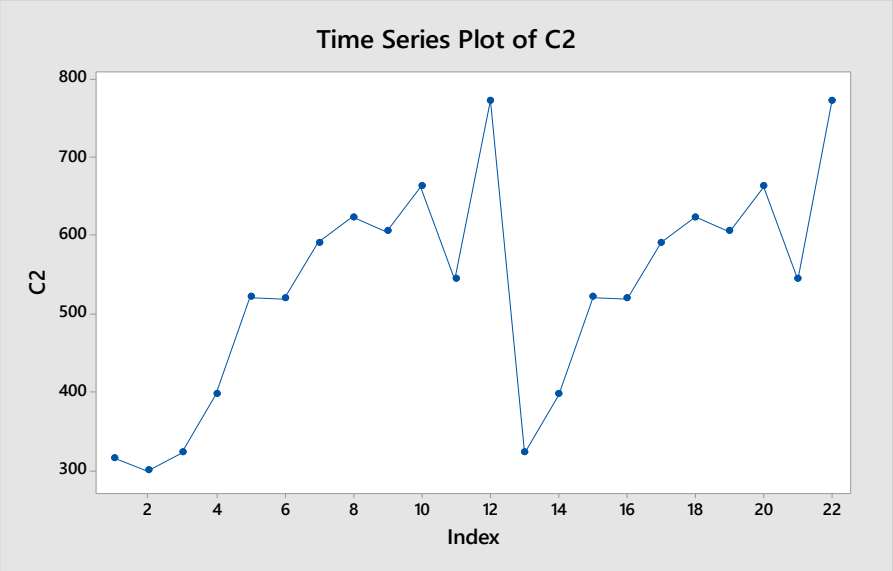
1 .Data collection

Data were collected that consisted of a time series consisting of (22) watching. The period dates from January 2007 until December 2013. This data counts the number of patients with malignant tumors in the city of Amarah taken from the Maysan Health Department, as shown in the following table:

Stability of the time series

The drawing of the time series is the first step in the analysis process, as it is possible through the drawing to initially identify some of the characteristics of the series, and the time series of the data has been drawn As shown in Figure (1).

Shape No. (1) *Time series drawing*



And to get acquainted with the initial model for describing the series under study, by extracting the autocorrelation function and partial autocorrelation as shown in Figure No. (2).

Figure (2) *Autocorrelation*

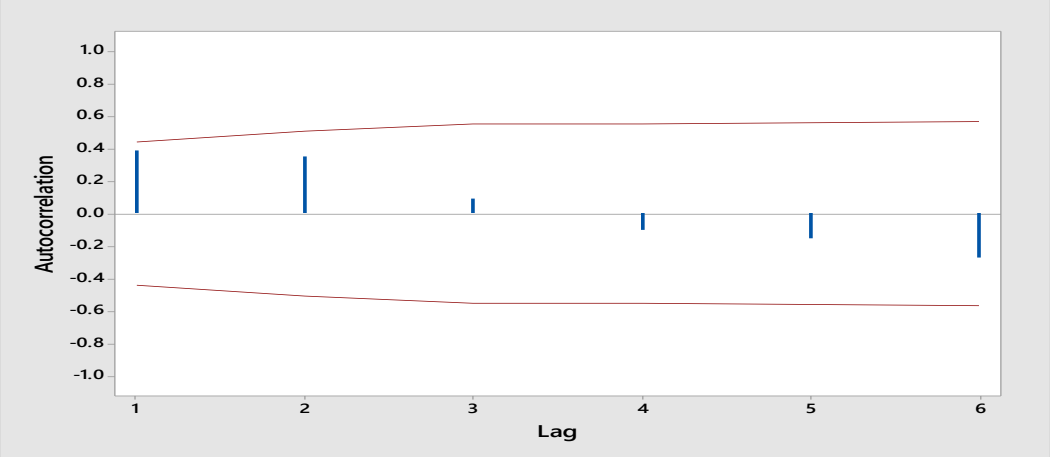


Figure (3) *partial correlation function*

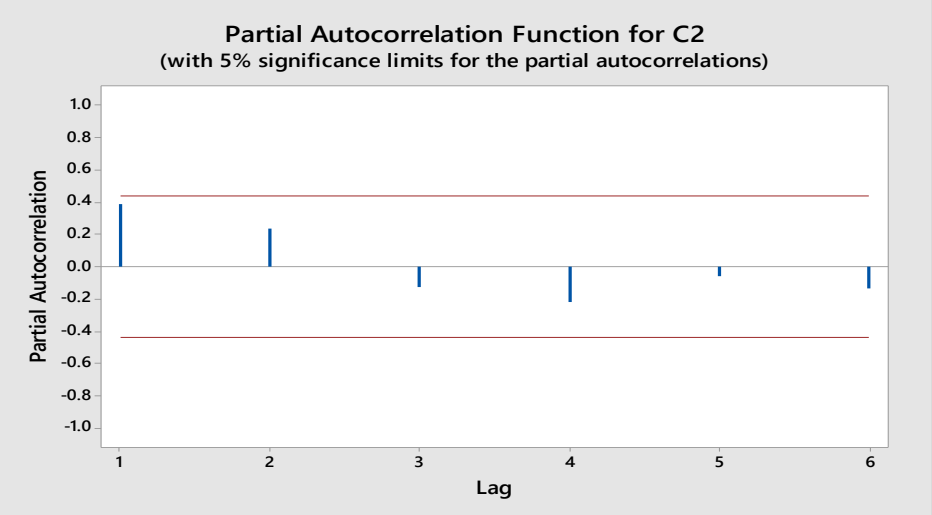


Figure (2 and 3) shows that there is a general increasing trend for this series, and this indicates that the time series is stable

2- Choose the appropriate model

After obtaining the stable time series, the model was determinedARIMA(1,2,1)It is the appropriate model based on the correlation and partial autocorrelation functions, and for the purpose of determining the appropriate model and making sure of it, we applied several models(ARIMA)Depending on several measures, including the mean squares of the residuals (rmsf) and the average ratios of the absolute values of the residuals(MAE)As shown in Table No. (1).

Models
A) ARMA(1,0,0)
(B) ARMA(0,0,1)
(C) ARMA(1,0,1)

Table (1) *shows the approved comparison criteria for the series*

Model	RMSE	MAE
(A)	92.6231	62.1872
(B)	123,545	88,906
(C)	99,201	63,698

It was shown by the results that the modelARIMA(1,0,0)takes lessRMSEAnd lessMAETherefore, we will use it in estimation and prediction

3- Estimating the model

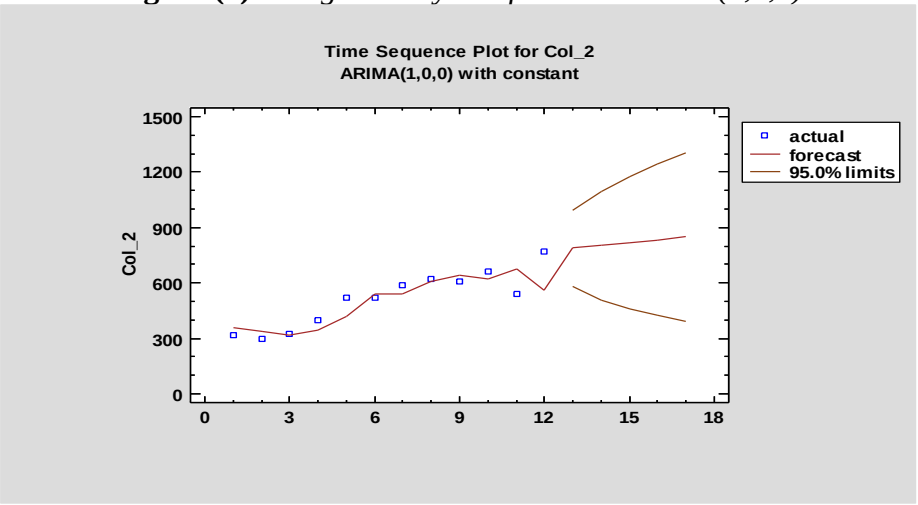
One of the stages of building a time series model is estimating the model by applying the usual least squares method (Ordinary least Squares) on the time series data, where the following data were obtained through Table (2).

Schedule (2)

Parameter	Estimate	Stnd. Error	T	P-value
AR(1)	0.342669	0.13118	2.6122	0.011527

Below is a time series plot of real and predictive data:

Figure (4) *String stability test for modelARMA(1,0,0)*



4-Forecasting

Forecasting is the final stage of the time series analysis using mixed models, and it is the ultimate goal of time series analysis. Table (3) gives future predictions for data

Table (3) *predictive values*

Residual	forecast	Data	Period
-41.4836	356,484	315.0	1
-36.8584	335,858	299.0	2
1.96258	320,037	322.0	3
53.2199	342.78	396.0	4
104,048	415,952	520.0	5
-19.5652	538,565	519.0	6
52.4236	537,576	590.0	7
15.2178	607,782	623.0	8
-35,413	640,413	605.0	9
39.3856	622,614	662.0	10
-135,977	678,977	543.0	11
210,692	561,308	772.0	12

Upper 95.0% Limit	Lower 95.0% Limit	forecast	Period
996,143	579.35	787,746	13
1096.39	510,243	803,316	14
1175.66	461,766	818,712	15
1243.82	424,052	833,936	16
1304.73	393,252	848,989	17

Conclusions and recommendations

Conclusions

1- The time series showed that it is stable in mean and variance
2- The appropriate model is the mixed modelARIMA(1,0,0) where it has the least RMSE, MAE

Recommendations

1- We recommend the use of (Box Jenkins models) because it is considered the best in prediction, especially in health data.
2- Considering the results of this research, which shows an increase in the number of people with malignant tumors over time, which requires taking the necessary measures by the competent authorities to limit this phenomenon, especially since most of the governorate's hospitals lack early detection devices for this disease and the treatment requirements for it.
3- Paying attention to increasing health units, visiting hospitals, providing means of treatment, and combating malignant cancerous diseases

Reference

1. Al-Jubouri, Walid Dahan Salibi, (2010) Predicting the level of inflation in monthly consumer prices in Iraq

Using Bivariate Time Series", Master's Thesis in Statistics, College of Administration and Economics, Al-Mustansiriya University
2. Al-Khudairi, Muhammad Qaduri Abd, (1996) a comparative study of estimation and prediction methods for some seasonal Boxes and Jenkins models, master's thesis in statistics, University of Baghdad, College of Administration and Economics..
3. Al-Sarraf, Nizar Mustafa, (1981), Time Series Analysis Using Statistical Refinement of Economic Forecasts in Iraq, Master's Thesis in Statistics, University of Baghdad, College of Administration and Economics.