Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 26, Issue 4 (2024) ISSN$_{Print}$: 1816-9171

Using nonparametric regression technology to predict the number of people with cancer in Babylon Governorate for the period (2025-2030)

Osama abdulazeez kadhim Al-Quraishi
ousamastat@gmail.com
Ministry of Education, Iraq

sama saadi Ali Alhashimi
samaalhashimi@mtu.edu.iq
Rusafa Management Institute, Baghdad

Received: 10/8/2024
Accepted: 26/9/2024
Available online: 25 /12 /2024

Corresponding Author : Osama abdulazeez kadhim Al-Quraishi

Abstract : Nonparametric regression methods and techniques can be used to evaluate the legitimacy of the assumed parametric model and vice versa, and perhaps to match the regression curve obtained from the techniques. Nonparametric regression methods can be used to assess the validity of a parametric model without assuming a specific functional form for the data. They provide flexibility in modeling relationships and can reveal when parametric assumptions may not hold. The research problem is the increase in the number of people with cancer in Iraq in general and in Babylon Governorate in particular and the accompanying social and economic effects on the individual and the state. The aim of the research is to compare parametric and nonparametric regression models and choose the best one using comparison criteria (Akaike information, Bayesian information, coefficient of determination). The time series data of cancer patients were analyzed using the R-studio program for (50) observations of cancer patients for the period (2019-2023/2), and it was found that the non-parametric regression method (linear spline regression method) is the best, that the number of cancer patients is constantly increasing, and that the expected number until the year (2030) is(45587).

Introduction:

Parametric and nonparametric regression techniques represent two different methods in regression analysis, but this in turn does not mean that using one method precludes the use of the other. Nonparametric regression methods and techniques can be used to evaluate the legitimacy of the assumed parametric model and vice versa, and perhaps to match the regression curve obtained from the techniques. Nonparametric regression suggests a suitable parametric model for use in future studies. The research problem was the human health aspect, which is the incidence of cancerous tumors, as the annual number of infections in Iraq in general is estimated at (2500) infections and (10,000) thousand deaths annually, and these numbers constitute a large economic burden. At the individual and state levels, and a social problem due to the negative repercussions of the disease on the psychological and physical health of the individual, as there is a noticeable increase in the number of cancer patients, especially after 2003, and after the use of various weapons by the coalition forces, and the number of infections in Babylon Governorate in particular in 2023 reached about (27,000) infections. The aim of the research was to predict the number of people with cancer in Babylon Governorate for the period (2025-2030) using one of the parametric or non-parametric regression techniques. And to provide a future database that enables the concerned authorities to develop strategic plans to confront cancer by providing the necessary funds to purchase medicines and medical devices, as well as training and providing medical personnel according to predictive figures.

theoretical side :

1-1: Polynomial Regression:[3][4]

Polynomial regression is a special case of linear regression that is based on the relationship between the independent variable (x) and the dependent variable (y) depending on the degree (nTh) of the polynomial terms. Polynomial regression also represents a nonlinear relationship between the variable (x) and the mean of the variable (y) and is denoted by the symbol E(y/x).

Although polynomial regression represents a non-linear relationship to the data to estimate the problem Linear statistics, meaning that the linear regression function for unknown variables is estimated and calculated From that data, therefore, Polynomial regression is a special case of multiple linear regression, and the independent variables resulting from the polynomial expansion of the "baseline" variables are called higher-order terms.

1-2: Advantages of polynomial regression:[16]

Polynomial models are popular for the following reasons:

1- It has a simple shape

2- It has the characteristics of being understood and known
3- Its shapes are flexible.
4- Closed family. Changes in location and measurement in the raw data lead to the assignment of a multinomial model..

1-3: Polynomial Regression model:[1][3][4]

Polynomial Regression is considered a special case of general linear regression, and it consists of only one independent variable or more than one multiplier independent variable and one dependent variable, and it is expressed by the following relationship:

$y = \theta_0 + \theta_1 x_{i1}^1 + \theta_2 x_{i1}^2 + \cdots + \theta_k x_{i1}^k + ei$ …(1)

The mathematical relationship is as follows:

$y = \theta_0 + \sum_1^n \theta_i \, x_i^n + ei$ …(2)

The polynomial model can be written in the form of matrices as follows:

$\vec{Y} = \mathrm{X}\vec{\theta} + \vec{\varepsilon}$ ... (3)

Whereas:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x & x^2 & \dots & x^m \\ 1 & x^2 & x^3 & \dots & x^{m+1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x^m & x^{m+1} & \dots & x^{m+n} \end{bmatrix} + \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The(ols) estimator is according to the following formula:

$\hat{\theta}_{\mathrm{LS}} = \left(\acute{\mathrm{X}}\mathrm{X}\right)^{-1}\acute{\mathrm{X}}\mathrm{Y}$ … (4)

The vector of parameters of the linear regression model estimated by the least squares (OLS) method is Best Linear Unbiased Estimator.

One of the conditions for applying ordinary least squares is that the relationship be linear between the dependent variable and the independent variables. Failure to fulfill this condition leads to biased and ineffective estimates. Therefore, it was necessary to find an alternative method to study the relationship between the variables, as nonparametric methods have been used recently, which have the ability to Compatibility with the data when formulating the model.

1-4: Nonparametric regression:[7]

The non-parametric regression method is one of the many methods that can be used in the modeling or data analysis stage. The regression model can be represented as follows:

$Y_i = m(X_i) + \varepsilon_i$ …(5)

Whereas:

m: *The unknown regression function.*

ε_i *: Random observations errors*

1-5: Basis spline: [6][2][6]

It is a polynomial function of degree p in the variable x, and it is symbolized by the B-Spline, and the first person to deal with it was Nikolai Lobachersk , at the beginning of the nineteenth century, This type of splines is connected to each other by connecting points called knots, and the knots must be an integer and their number is less than the sample size, The concept of splin is a continuous polynomial curve used to approximate the solution of a mathematical problem. This curve depends on the relationship between the Basis function and the control points, splines have been used in many different fields: mathematics, engineering, and computer science. Recently, the use of splines has become common, as they do not require derivative calculations or assumptions when calculating transactions, so they save time and effort when dealing with them, There are more than one type of splines, including linear and quadratic, and the most common splines are cubic splines because of the continuity properties they possess. Cubic splines are written as follows:

$s(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \sum_{i=1}^k (x - ki)^3 + \epsilon$ … (5)

1-6: Estimate spline regression:[10][8][11]

In this part of the research, the introduction to spline regression will be discussed, as the field of the Independent variable represented by the period (a,b) is divided into a group of locations called (nodes) and symbolized by ($T_0, T_1, \dots, T_k, T_{k+1}$) and they are arranged in ascending order, as follows and as follows..

$a = T_0 < T_1 < \cdots < T_k < T_{k+1} = b$ …(6)

The nodes divide the domain of the Independent variable, which is represented by the period (a,b), into (k) sub-periods $(T_r, T_{r+1}), r = 0,1, \dots k$

Whereas: T_r : internal (sub) nodes

In other words, the spline regression represents a cross-sectional polynomial, It is a polynomial of a certain degree within any two adjacent nodes T_r, T_{r+1} for every $r = 0,1, \dots k$
Which have been incorporated into a contract may allow for the existence of cut-off derivatives on those contracts

1- Truncated Power Basis:[12][13]

The chip slope can be constructed using what is called a cutting force basis of degree k with k nodes $T_1, \dots, T_k$
$1, x, .., x^k, (x - T_1)_+^k, \dots \dots, (x - T_k)_+^k$ …(7)
Whereas:
$(x - T_r)_+^k$: It indicates the force k of the positive part of the magnitude ,
That is $(x - T_1)_+^1 = \max\{0, (x - T_1)_+^1\}$.
Noting that the first (k+1) of the basis functions for the basis of the cutting force represented by the formula (7) represents a polynomial of degree that may reach K, while the rest represent all the basis functions of the cutting force of degree K. Therefore, if the degree of the basis of the cutting force (k=0,1,2,3) is called a force basis "fixed, linear, quadratic and cubic" plots respectively
Using the basis of the cutting force represented by the formula(7) The regression of the spline can be explained as follows:
$m(X) = \sum_{s=0}^{k} B_S X^s + \sum_{r=1}^{k} B_{k+r} (X - T_r)_+^k$ …(8)
$whereas$:
β0 ,β1 ,..., βk+K : Related transactions
This can be called spline regression, of degree K with nodes ($T_1, \dots, T_k$) , The spline regression represented by the formula (8) at (k=1,2,3) is called linear, quadratic, and cubic spline regression, respectively.
We can see here that within any subinterval ($T_r, T_r + 1$), we have the following
$m(X) = \sum_{s=0}^{k} B_S X^s + \sum_{r=1}^{k} B_{k+r} (X - T_1)_+^k$ …(9)
This expression, in turn, represents a polynomial of degree k.

1- **Regression Spline Smoother** :[14][15]

The basis of the cutting force shown in formula (7) is indicated as follows:
Φ (x)= $[1, x, .., x^k, (x - T_1)_+^k, \dots \dots, (x - T_k)_+^k]^T$ …(10)
Whereas:
p= Number of participating basis functions =K+k+ 1
β= $Associated\ transactions = [\beta_0, \beta_1, \dots, \beta_K, \beta_{K+1,\dots,} \beta_{K+K}]^T$
Therefore, the slope of the spline shown by the formula (7) can be reformulated as in the following figure:
$m(X) = \Phi(x)^T \beta$ …(11)
The model shown in formula (5) can be reformulated as follows:
$Y = XB + \varepsilon_i$ …(12)
Whereas:
$y = (y_1, \dots, y_n)^T$
X= $[\Phi_p(X_1), \dots, \Phi_P(X_n)]^T$ ….(13)
ε= $(\varepsilon_1, \dots, \varepsilon_n)^T$
as $\Phi_p(X)$ Basis & X She is of full rank , So $(X^T X)$ is inverse when (n≥ p) , It is possible to clarify the normal estimator for parameter B , Which solves the approximation linear model shown in equation (13) , This is done using the least squares method (ols) as follows :
$\hat{\beta} = (X^T X)^{-1} X^T Y$ …(14)
This leads to the spline regression fit to the function m(X) shown in Equation (5)as follows
$\hat{m}_p(X) = \Phi_p X^T (X^T X)^{-1} X^T Y$ …(15)
This is often called the spline regression smoother for the function m .
In particular, the m values calculated at design time points k for each (i=1,2,…,n) are related as shown in the following response vector :
$\hat{Y}_p = X^T (X^T X)^{-1} X^T Y = A_p Y$ …(16)
Whereas:
$\hat{Y}_p = (\hat{Y}_1, \dots, \hat{Y}_n)^T$
$\hat{Y}_i = \hat{m}_p(X_i)$ i=1,2,…,n
$A_p = the\ spline\ regression\ smoothing\ matrix = X (X^T X)^{-1} X^T$

2 - The applied aspect:

2-1: Description of study data:

The data on the monthly number of cancer patients, which represents the dependent variable (G), in Babylon Governorate were analyzed using the R-studio program for (50) observations taken from the Babylon Health Department - Planning Department - Health and Life Statistics for the period (2019 - 2023/2), which represents the independent variable (x).

2-2: polynomial regression model.

Table(1)

Comparison standards between polynomial regression

	Quadratic	**Cubic**	Fourth	Fifth	Six
AIC(Akaike Information)	871.53	871.12	866.15	860.22	861.11
BIC(Bayesian Information)	881.30	878.79	879.66	871.84	878.35
R^2(coefficient of determination)	0.62	0.64	0.65	0.69	0.67

Looking at the table (1), we see that the Fifth model has the smallest comparison criteria, the largest coefficient of determination; therefore it is the best model for prediction:

2-4: Comparison between linear regression and cubic regression models:

Table(2)

Comparison standards between linear regression and Quartet regression models

	Fifth	**liner**
AIC	860.22	864.37
BIC	871.84	877.14
R^2	0.69	0.66

Looking at the table (2), we see that the Fifth model has the smallest comparison criteria, the largest coefficient of determination; therefore it is the best model for prediction:

$y = 6248 - 321.5X + 24.31X^2 - 0.546\,X^3 - 0.0007X^4 - 0.00004X^5$ …(17)

The calculated F (125.014) appeared with a probability value (1.38e-006) smaller than the significance level (.05) and therefore the model is suitable for prediction.

The following table shows the significance of the estimated parameters:

Table (3)

Significance of estimated parameters

	Parameter estimation	t-test	p-value
θ_0	6248	3.409	3.39e-06 ***
θ_1	−321.5	1.474	0.4714
θ_2	24.31	−0.9629	0.0425 *
θ_3	-0.546	1.209	0.0234 *
θ_4	- 0.0007	−1.437	0.0144 *
θ_5	-0.00004	-1.265	0.0125 *

Looking at Table (3), we see the fixed limit (θ_0) has a significant effect because it has a probability value (sig= 3.39e-06) that is smaller than the significance level (.05) . As for the first marginal slope ,(θ_1) does not have a significant effect because it has a probability value (sig=0.4714) greater than the significance level (.05). As for the second, third, fourth, and fifth marginal slopes ($\theta_2, \theta_3, \theta_4, \theta_5$) have a significant effect because they have probability values (0.0425, 0.0234, 0.0144, 0.0125) smaller than a significance level of (0.05).

2-5 : **Estimate spline regression:**

1- **Determine the number of nodes and their locations** :

The node was determined through the drawing below and choosing more than one value. The point that constitutes an inversion in the series is considered a node, and the most prominent of these points are (5,25 ,30 , 46).

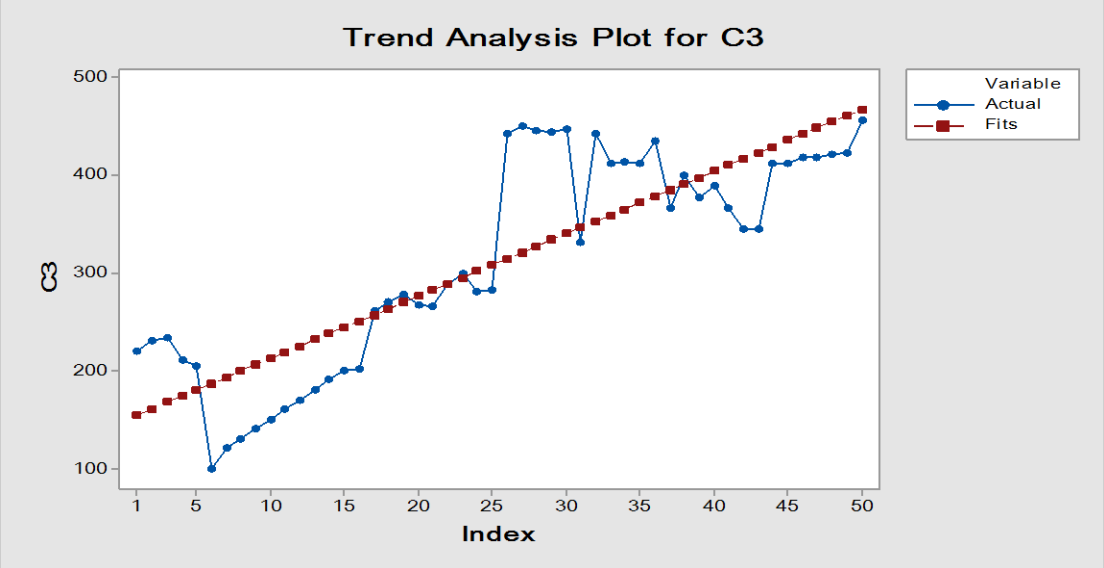

Figure (1) shows the series Coups(notes) points

2- **Determine the best model**

Table(4)

Comparison standards between Spline regression

	Liner	Quadratic	**Cubic**
AIC(Akaike Information)	853.233	857.456	854.654
BIC(Bayesian Information)	866.651	873.566	868.604
R^2(coefficient of determination)	0.792	0.742	0.731

Looking at the table (4), we see that the linear spline model has the smallest comparison criteria, and largest coefficient of determination, therefore it is the best model for prediction.

$y = 6524.7 + 3351.4X + 5731.7\,S13487.3\,S2 + 2464.2\,S3 + 378.3\,S4$ …(18)

Whereas:

k_i : notes

Si=(X-Ki) ,

S1=xi-5 ,

S2=Xi-25

S3=Xi-30 ,

S4=Xi-46

The calculated F (116.02) appeared with a probability value (2.5e-007) smaller than the significance level (.05) and therefore the model is suitable for prediction,

The following table shows the significance of the estimated parameters:

Table (6)

Significance of the estimated parameters

parameter	**Estimates**	t-test	P-value
θ_0	6524.7	7.133	2.57e-10 ***
θ_1	3351.4	4.780	2.02e-05 ***
θ_2	5731.7	1.539	0.13452
θ_3	-3487.3	9.247	9.26e-12 ***
θ_4	2464.2	2.289	0.01134 *

θ_5	378.3	3.795	0.00751 **

Looking at Table (6), we see the fixed limit (θ_0) has a significant effect because it has a probability value (sig= 2.57e-10) that is smaller than the significance level (.05) . As for the second marginal slope ,(θ_2) does not have a significant effect because it has a probability value (sig= 0.13452) greater than the significance level (.05). As for the first, third, fourth, and fifth marginal slopes ($\theta_1, \theta_3, \theta_4, \theta_5$) have a significant effect because they have probability values (2.02e-05, 9.26e-12, 0.01134, 0.00751) smaller than a significance level of (0.05).

2-6: **Comparison between the linear spline regression model and Polynomial**

Table(7)

Comparison standards between Spline regression and Polynomial

Model	AIC	BIC	R^2
l **fifth-order**	860.22	871.84	0.69
Linear Spline	854.21	865.63	0.79

Looking at Table (7), we conclude that the linear spline regression model is the best.Therefore, forecasts are made for the period (2025-2030) by taking the arithmetic mean of the monthly forecasts using equation (18).

Table. (8)

Predictive values for people with malignant tumors for the period (2025-2030)

years	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**
Predictive values	30687	34067	36447	39827	42207	45587

Conclusions:

1- After testing the parametric regression models, it was found that they were all significant and adequate for prediction

2- It turns out that the best parametric regression model is a fifth-order polynomial regression model

3- It turns out that the best nonparametric regression model is the linear spline regression model

4- The advantage of the nonparametric regression model compared to the parametric regression model

5- The number of predictive cancer cases in Babylon Governorate is constantly increasing, and this suggests a health risk

Recommendations:

1- We recommend conducting similar research to build statistical models based on different models or new methods, for example (Generalized Additive Models (GAM).

2- We recommend that the concerned authorities take appropriate measures to confront this dangerous disease and adopt strategic planning based on forecasting to confront and control the disease.

Sources

1- Abdul-Karim Iddrisu , Emmanuel A. Amikiya,(A predictive model for daily cumulative COVID-19 cases in Ghana version 1; peer review: awaiting peer review),2021

2- Abdulwasaa MA, Abdo MS, Shah K, et al. : Fractal-fractional mathematical modeling and forecasting of new cases and deaths of covid-19 epidemic outbreaks in india. Results Phys. 2021; 20: p. 103702.

3- Al-Gazzar, Farouk Fathi Al-Sayed, et al., (2021) "The multinomial regression model as a treatment for standard problems, an applied study on the relationship between the economic growth rate and the inflation rate in the Egyptian economy," Journal of Financial and Commercial Research, Volume 22, Issue 4

4- Al-Sarraf, Nizar Mustafa and Shoman, Abdul Latif Hassan, (2013) "Time Series and Index Numbers," Doctor's House for Administrative and Economic Sciences, first edition, Baghdad, Iraq.

5- Delaigle A. & Meister A. (2007), "Nonparametric Regression Estimation in the Heteroscedastic Errors in Variables Problem", Journal of the American Statistical Association, vol. 102 , No. 480 .

6- Gálvez, A., & Iglesias, A. (2013). Firefly algorithm for explicit B-spline curve fitting to data points. Mathematical Problems in Engineering, 2013

7- Härdle, Wolfgang, (1994), "*Applied Nonparametric Regression*". Cambridge: Cambridge University Press.

8- Ibrahim.N.A and Suliadi.(2009)"Nonparametric Regression for Correlated Data".WSEAS TRANSACTION ON MATHEMATICS ,ISSN : 1109-2769,Issue7,Volume8

9- kadhim Al-Quraishi, O. A. (2021, May). Choosing the best eestimated regression equation for data subject to geometric distribution (Student data as a case study). In *Journal of Physics: Conference Series* (Vol. 1879, No. 3, p. 032045). IOP Publishing

10-Lee, Jong S., & Cox, Dennis, D., (2010), "*Robust Smoothing: Smoothing Parameter Selection and Application to Fluorescence Spectroscopy*". Computational Statistics & Data Analysis, 54

11-Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). Semiparametric Regression.CambridgeUniversity Press, New York.

12-Schimek, Michael G., (2000), "*Smoothing and Regression: Approaches, Computation, and Application*". Wiley Series in Probability and Statistics, John Wiley & Sons, Inc.
13-Wand,M. P. (2000), "A Comparison of Regression Spline Smoothing Procedures," computational statistical 15,443-462
14-Wu, Hulin, & Zhang, Jin-Ting, (2006), "*Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling*
15-Yao, F., & Lee, Thomas C.M., (2008), "*On Knot Placement for Penalized Spline Regression*". Journal of the Korean Statistical Society, 37.
16-https://en.wikipedia.org/wiki/Polynomial_and_rational_function_modeling