JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-15, No.-7, July (2020) pp 294-305 ISSN (Print) 0973-8975

ESTIMATING THE AVERAGE RESPONSE FOR THE LINEAR MIXED MODEL USING SOME NON-PARAMETRIC METHODS

Ameena Karem Essa[1], Haifa Taha Abd[2]

[1,2]Department of Statistics , Collage of Management and Economics, University of Mustansiriyah, Bagdad, Iraq

[1]Ameena77kareem@gmail.com, [2]haefaa_adm@uomustansiriyah.edu.iq

Corresponding Author: Ameena Karem Essa

https://doi.org/10.26782/jmcms.2020.07.00024

Abstract

This study aims to test a new treatment that has been developed for type 2 diabetes, by estimating the response of diabetics by experimenting number of mixed linear models, non-parametric, where they were compared by relying on the coefficient of determination and the standard error for the random errors in order to determine the appropriate model and then measure the effectiveness This new treatment is for type 2 diabetes.

Therefore, some non-parametric methods were used in estimating the average response for the mixed linear model. The method of the kernel smoothing function was used by employing the Gaussian and Epanchnikov family functions, as well as some formulas of the Cross Validation method. To estimate Bandwidth as Scott and Silverman. An experiment for a new treatment for type 2 diabetes was chosen as an application of the mixed linear model, by experimenting with this drug on a sample of patients who were divided into three different age groups and performing laboratory tests for a period of three months, and then estimating their response rates to the new drug through four models Different. The results demonstrated that the A mixed non-parametric linear model with (Gaussian) function and the (Scott) package was the best fit model for this study, as it gave the largest determination coefficient and the lowest standard deviation of the error, as well as the new drug, was not effective in regulating blood sugar level for all age groups of patients.

Keywords: Linear mixed model Non-parametric, Kernel Smoothing, Bandwidth.

I. Introduction

Study of the linear regression model is considered to describe the random or fixed effects of explanatory variables on the response variable, as it has been used in several fields, including biological sciences, behavioral, environmental and social sciences, because it provides mathematical formulas that are easy to interpret and predictions, where linear regression models have become a way to predict and know

the changes that occur in The future as scientifically and logical interpretation. As for the mixed linear model, it is a linear regression model that is concerned with studying phenomena that consist of a combination of the effects of the explanatory variables on the response variable at the same time, which are the fixed and random effects. It was introduced by (Rao & Hartly 1967) to represent the phenomena in which this type of a combination appears, and it appears in many fields such as medical, agricultural, economic, and others. As is known, studies in such fields needed frequent observations and long-term observation periods that may extend to months or years at times. It is not hidden to everyone the problem of diabetes and the extent of its impact on human life, as it affects different age groups, so studying this case requires A model in which different measurements of blood sugar levels can be recorded for different age groups and for a relatively short period of time, hence a mixed linear model has been employed to study type 2 diabetes, It must be said that type 2 diabetes is considered the disease of the age until this moment because there is no treatment that will end the disease permanently.

II. Nonparametric Linear Model (NLM)

Simple NLM can be expressed mathematically as follows:

$$Y = m(.)X + \varphi$$

$$E(Y|X)\ m(.)X \tag{1}$$

Whereas: $(Y|X)$: Mean Response

$m(.)$:The non-parametric kernel smoothing function to be estimated according to the non-parametric methods and methods That is to be a derivable function

X: Explanatory variable

III. Linear Mixed Model (LMM)

The mixed linear model with its parameter formula can be expressed mathematically as follows (Gzado 2007):

$$Y = X\beta + Z\theta + \varphi \tag{2}$$

Whereas: Y: Response variable vector.

θ: Fixed parameters vector.

X: Fixed effects matrix.

β: Random parameter vector.

Z: Random effects matrix.

φ: Random error vector.

Then, through equations (1) and (2), the formula for the non-parametric linear mixed model can be expressed as follows:

$$Y = m\{X\beta + Z\theta\} + \varphi$$

$$E\{(Y|X,Z)\} = m\{X\beta\ Z\theta\} \quad (3)$$

The estimation of the parameters of the linear model, whether mixed or not mixed, is done through the use of well-known parametric. methods and is based on many conditions, the most important of which is that the random error follows the normal distribution with an average of zero and a homogeneous variance equal to (σ^2) (Heather, 2008), but in the non-parametric model It does not take into account the nature of the relationship that binds the variables and does not set strict conditions regarding the distribution of the error limit or the distribution of the model variables. It should be noted that there are several types of non-parametric linear models, among which we mention the simple non-parametric model, The aggregate non-parametric model, and The general aggregate non-parametric model. In this paper, the Mixed aggregate linear non-parametric Model was used.

IV. Methods for estimating the non-parametric mixed linear model

Non-parametric estimation methods rely on what is known as non-parametric smoothing, which is the basis for the non-parametric regression, By smoothing the observations of the response variable by specifying weights for those observations where the largest weight of the most impactful observations and the lowest or zero weight of the less-impacted observations (Carroll, Dalaigle and Hall (2007)). The mathematical concept of weights can be clarified as follows:

$$E(Y/X,Z) = \int Y\, f\,(Y,X,Z)dy = \int \left[\frac{f(Y,X,Z)}{\hat{f}(Y)}\right] Y\, dY$$

As the segment $\int[\frac{f(V,Y)}{\hat{f}(Y)}]$ represents the Weights function of the observations of the response variable. There are several techniques and methods for estimating the weight function or the smoothing function m (.).and the most important is the method of kernel functions.

V. Kernel Functions

It is a method by which different curves are reconciled with the response variable curve and find the best smooth curve that matches the response variable curve. There are several methods to reconcile the curves, one of them is the kernel function, which is a weight function used in smooth the response variable data (hardle, 1994) and has the following properties:

Continuity $\quad 0 \leq T(u) \leq \infty.$

Symmetry $\quad T(u) = T(-u).$

Probability density function $\int T(u)du = 1$.

Have known variance $E(u^2) = \int u^2\, T(u)du = \sigma^2{}_t$.

And average = zero $E(u) = \int u\, T(u)du = 0$.

Among the most common kernel functions are the Epanchnikov family function and the Gaussian function, as their formulas differ according to the number of explanatory variables. In the case of the one variable (hardle, 1994), the formula of the Epanchnikov function is:

$$T(u) = \begin{cases} \frac{3}{4}(1-u^2) & |u| < 1 \\ 0 & o.w \end{cases}$$

In the case of two explanatory variables, the formula is:

$$T(u_1,u_2) = \begin{cases} (2\pi)^{-1}(1-{u_1}^2-{u_2}^2) & {u_1}^2+{u_2}^2 < 1 \\ 0 & o.w \end{cases} \quad (4)$$

Two Gaussian function formulas for one explanatory variable and two explanatory variables are, respectively:

$$T(\mathrm{u}) = \begin{cases} (2\pi)^{\frac{-1}{2}} e^{-\frac{u^2}{2}} & 0<u<\infty \\ 0 & o.w \end{cases}$$

$$T(u_1,u_2) = \begin{cases} (2\pi)^{-1} e^{-\frac{{u_1}^2+{u_2}^2}{2}} & 0<u_1,u_2<\infty \\ 0 & o.w \end{cases} \quad (5)$$

When applying the kernel functions in (4) and (5) to the data for the purpose of smooth, the variable (u) is as follows:

$$u_i = \frac{X_i - X_0}{d} \qquad i = 1,2,\dots,n$$

$\mathbf{X_i}$: The measurements represent the explanatory variable.

X_o: Represents observed of the observations of the explanatory variable .

d: The bandwidth (smoothing parameter).

The most important part of the kernel functions is the method of choosing the bandwidth or the smoothing parameter (d).which It precedes the choice any function from the kernel functions and Also shows that (d) represents the free parameter that

has a strong influence on the results of the kernel functions and therefore any estimator from estimators non-parametric regression Which depends on the kernel function as well as the function itself will be very sensitive to changes in the value of (d), which in turn will change shape of the smoothing curve for the response variable (Y). There are several methods for choosing a value (d) Among them:

1-Experimental method.

2- Bootstrap style.

3- Cross Validation method: Of the most accurate and widely used formulas, and depends on excluding one value from the values of observations in each time, and most prominent formulas of this method:

- **Silverman formula**: It is considered one of the famous formulas used to estimate the value of one variable. It is a fixed smoothing parameter and its formula is:

$$\hat{d} = 1.06\ \hat{\sigma}\ n^{-0.2}$$

$$\hat{d} = 1.06\ R\ n^{-0.2}$$

Whereas: R: The difference between the third and first quarters is represented.

$\hat{\sigma}$: The variance estimator represents the observations of the $X_i$variable

In the case of the two variables, we will have a square matrix:

$$D = \begin{bmatrix} d_1 & d_{12} \\ d_{21} & d_2 \end{bmatrix} \qquad (6)$$

Whereas: d_1: The smoothing parameter for the explanatory variable X.

d_2: The smoothing parameter for the explanatory variable Z.

d_{12} , d_{12}: The common smoothing parameter for the X and Z variants.

- **Scott formula**: Scott developed the following formula for estimating (d) in the case of two variables:

$$\hat{d}_j = \hat{\sigma}_j\ n^{\frac{-1}{6}}$$

$$\hat{d}_j = \hat{\sigma}_j\ n^{\frac{-1}{6}}$$

$$\hat{d}_{ij} = \hat{\sigma}_{ij}\ n^{\frac{-1}{6}}$$

$$\hat{D} = n^{\frac{-1}{6}}\ \hat{\sum}^{0.5}$$

It can be seen that in the case of $\hat{\sigma}_{ij} = 0$, the preamble matrix will be:

$$D = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}$$

(Flashier) has indicated that two facts must be considered in testing the smoothing parameter: The first fact is that when choosing a large value for the smoothing parameter, this will lead to large the data smoothing with a large bias and little variance, and the second is that when choosing a smoothing parameter with a small value will lead to smoothing The data greatly, little bias, great variance.

Of the foregoing, the estimated value of the smoothing function using the kernel functions will be as follows:

$$E\{Y|X,Z\} = m\{X\beta + Z\theta\} = \frac{1}{n}\sum_{i=1}^{n} K\left[\left(\frac{X_i - X}{h_1}\right) * \left(\frac{Z_i - Z}{h_2}\right)\right] \quad (7)$$

X_i: The measurements represent the illustrative variable with a fixed effect

X: Represents one observation from explanatory variable observations.

Z_i: The measurements represent the illustrative variable with a random effect.

Z_i: Represents one observation from explanatory variable observations.

Through the algorithm (Back fitting), this model can be estimated as follows:

1. Determination of Initial Values:

$$\overline{Y} \cdot m^{(0)} \equiv 0$$

2. Repeat the process:

$$r_i = Y - \hat{C} - \sum_{k=1}^{i-1} m^{k+1} - \sum_{k=i-1}^{k} {m_k}^i$$

3. When achieving values convergence, the final estimate is reached:

$$m^{i+1}(.) = D(r_i)$$

VI. The Applied Frame

It is known that glucose or blood sugar is a type of simple or single diabetes that a person gets from food and uses them to produce the energy needed for the body. Glucose carries the chemical formula (C6H12O6) i.e. six atoms of carbon and oxygen and twelve atoms of hydrogen, and it is often absorbed easily to enter the bloodstream and travel through it to reach the cells.

The rate of sugar varies from one person to another according to age, depending on a number of factors, including the time of the examination or eating

soon before the examination, or taking some medications, or exercising, or any other effort before the examination, or infection with other diseases. In addition to a number of medications that control normal blood sugar levels. It is worth noting that the blood sugar level of the second type in the adult human body, according to the World Health Organization, from 72 to 126 mg / dL before eating [world health organization].

So the mixed linear non-parametric model was applied to data related to blood sugar measurements after receiving a new treatment chosen by one of the doctors in his clinic in Baghdad, through tests for the percentage of urea in the blood and recorded for a sample size (24) patients distributed in three age groups (groups) (30-20), (50-30), (50+) for a period of three months, and the three groups were represented by numbers (3,2,1), respectively, as shown in Table (1):

Table (1): Diabetics data by age groups

Patient	Stage	Urea ratio	The percentage	Patient	Stage	Urea ratio	The percentage
1	1	120	25	13	2	118	410
2	1	180	40	14	2	75	360
3	1	165	55	15	2	20	98
4	1	168	57	16	2	22	111
5	1	130	28	17	3	27	125
6	1	110	21	18	3	42	185
7	1	90	15	19	3	60	215
8	1	82	13	20	3	28	128
9	2	384	88	21	3	16	87
10	2	250	68	22	3	18	96
11	2	343	70	23	3	63	238
12	2	460	115	24	3	61	220

The data were analyzed using the (R3.5.1) program to estimate the level of response of patients in the three groups after receiving a new treatment, and through the non-parametric assessment of the mixed model using the kernel functions (Epanchnikov) defined in equation (4), and employing equations (7) and (8)), Using the packages $(d_1 = 2.5, d_2 = 0.5)$, and the results were as shown in Table (2):

Table (2) :Estimated data for diabetics according to the age groups of the first model, with a bandwidth of 2.5, 0.5

Patient	Stage	The percentage of sugar	Patient	Stage	The percentage of sugar
1	1	119.5	13	2	410.1
2	1	175.5	14	2	350.7
3	1	178.7	15	2	103
4	1	185.6	16	2	109.6
5	1	130.2	17	3	126.5
6	1	106.3	18	3	179
7	1	87.2	19	3	126.5
8	1	81.2	20	3	179.1
9	2	384.1	21	3	130.2
10	2	288.3	22	3	96.5
11	2	309.2	23	3	237.1
12	2	459.8	24	3	217.9

From the results of Table (2) and after taking the new treatment by the patients, and using the (Epanchnikov) function firmly (d_1 = 2.5, d_2 = 0.5), we find that the estimated values of the average response for diabetics in the first group (5) from (8) patients They respond to the new treatment and in varying proportions, while the second and third groups responded to four patients, and this indicates the effect of the age group on the results.

Through the non-parametric estimation of the mixed model using the kernel functions (Gaussian) and knowledge in equations (5), and by employing equations (7) and (8), the results were as shown in Table(3):

Table (3): Estimated data for diabetics according to the age groups of the second model, with a bandwidth of 2.5, 0.5 .

Patient	Stage	The percentage of sugar	Patient	Stage	The percentage of sugar
1	1	110.4	13	2	296.3
2	1	100.8	14	2	251.8
3	1	120	15	2	113.6
4	1	132.6	16	2	100.8
5	1	116.8	17	3	116.8
6	1	104	18	3	132.7
7	1	84.9	19	3	193.5
8	1	84.9	20	3	120
9	2	193.5	21	3	88

10	2	196.4	22	3	94.4
11	2	231.9	23	3	193.5
12	2	294.7	24	3	113.6

From the results of Table (3) concerning estimates of the patient's response averages using the (Gaussian) function with firmness $(d_1 = 2.5, d_2 = 0.5)$,we notice a clear response to the new treatment, in the first group 6 of 8 patients responded while the response of the patients of the second and third groups was satisfactory, The response reached 7 of 8 patients.

The non-parametric estimation of the mixed model using the cores functions (Epanchnikov) with packages $(d_1 = 3.5, d_2 = 0.6)$The results of the estimated values of diabetic response averages for the three groups were shown in Table (4).

Table (4): Estimated data for diabetics according to the age groups of the third form, with a bandwidth of 3.5, 0.6

Patient	Stage	The percentage of sugar	Patient	Stage	The percentage of sugar
1	1	132.3	13	2	355.1
2	1	173.8	14	2	129.8
3	1	194.5	15	2	459.5
4	1	195.4	16	2	470.6
5	1	130.6	17	3	132.2
6	1	111.6	18	3	149.6
7	1	115.1	19	3	184.4
8	1	101.2	20	3	236.6
9	2	105.4	21	3	166
10	2	371.7	22	3	97.4
11	2	415.5	23	3	104.8
12	2	237.7	24	3	254

The results of Table (4) show that estimating the average response of patients according to this model led to the improvement of only one patient from the first group, four patients from the second group, and 3 patients from the third group.

As for the non-parametric estimate of the mixed model using the cores functions (Gaussian) with packages $(d_1 = 3.5, d_2 = 0.6)$,the results were shown in Table (5).

Table (5): Estimated data for diabetics according to the age groups of the fourth form with the bandwidth 3.5, 0.6

Patient	Stage	The percentage of sugar	Patient	Stage	The percentage of sugar
1	1	122	13	2	308.7
2	1	163.4	14	2	85.8
3	1	184	15	2	437.6
4	1	184	16	2	415.7
5	1	122	17	3	114.8
6	1	101.8	18	3	132.2
7	1	101.5	19	3	167
8	1	91	20	3	219
9	2	90.8	21	3	149.6
10	2	349.5	22	3	80
11	2	393.7	23	3	87
12	2	308.3	24	3	236.5

The results of Table (5) show that estimating the average response of patients according to this model was an improvement of 3 patients from the first group, four patients within the second group, 6 patients within the third group.

To demonstrate the best model among the four models, based on an of the coefficient of determination (R^2), and the Residual standard errors, as shown in Table (6):

Table (6): coefficient of determination and residual standard errors for non-parametric models

d_1, d_2	Model	R^2	Residual SD
2.5,0.5	1	0.6713421	11.8337
	2	0.6887554	11.26714
3.5,0.6	3	0.4210699	31.02931
	4	0.5748934	17.89282

Through table No. (6), it was found that the highest determination coefficient was for the second model, the determining coefficient was 68.87%, followed by the first model 67.13%, the fourth model 57.49%, and finally the third model 42.11%. As for the lowest residual standard errors, the second model was 11.26714, followed by the first model and then The fourth model and the third model. This indicates that the

second model has the advantage of having the highest determination coefficient and the lowest residual standard errors.

VII. Conclusions

1- Through Table (2), which refers to the estimated values of the response average of diabetics for the three groups, using the (Epanchnikov) function with packages $(d_1 = 2.5, d_2 = 0.5)$, it was found that (13) patients responded to the new treatment, as the measurement of sugar was lower Than in the old treatment.

2- Through Table (3) for Estimates of patient response averages Using Gaussian function With packages $(d_1 = 2.5, d_2 = 0.5)$, It turns out that (20) patients improved as a result of receiving the new treatment

3- Through Table (4), which refers to the estimated values of response average for diabetics for the three groups, using the (Epanchnikov) function with packages $(d_1 = 3.5, d_2 = 0.6)$, it was found that (8) patients improved as a result of receiving the new treatment.

4- Through Table (5)of Estimates of patient response averages using Gaussian function with package $(d_1 = 3.5, d_2 = 0.6)$, it was found that (13) patients improved when receiving the new treatment.

5- Table (6) shows that the second model is the best. The value of the determination coefficient has reached (0.6887554), which is the largest among the four models. The value of the standard deviations for lower errors was (11.26714), then the first model solved secondly, and the third model came last, This means that the non-parametric mixed linear model using the Gaussian function is best for estimating the mean response of diabetics to the new treatment. Likewise, the first package was better than the second package because of the results of the coefficient of determination and standard deviations of errors for the two models best-arranged It was in favor of this package.

6- Through the better model, it is clear that the new treatment for diabetes did not help in reducing the complications and risks of high blood sugar for patients under treatment and for different age groups.

VIII. Acknowledgement

The authors would like to thank the Doctor for kindly help with provided us with diabetes data after receiving the new and the reviewers for the thoughtful comments and suggestions.

Reference

I. Carroll, R.J., Delaigle, A. and Hall, P. (2007). "Nonparametric Regression Estimation from Data Contaminated by a Mixture of Berkson and Classical Errors". Journal of the Royal Statistical Society, 69, 859-878.

II. Czado, C, (2007). "Linear Models with Random Effects". Lecture notes, web paper.

III Hardle, W., (1994), "Applied Nonparametric Regression". Humboldt - University, Berlin, Germany.

IV Heather, T., (2008). "Introduction to Generalized Linear Models". ESRC National Centre for Research Methods, University of Warwick, UK**.**

V Jiang, J. (2007)." Linear and Generalized Linear Mixed Models and Their Applications". Springer, New York.

VI McCullagh, P., and Nelder, J. A‖ Generalized Linear Models‖, (1989). 2nd ed. London: Chapman& Hall.

VII Nelder, J., Wedderburn.W., (1972)." Generalized Linear Models‖, Journal of the Royal Statistical Society. Series A (General), 135(3), 370-384.

VIII Racine J, Li, Q., (2004). "Nonparametric Estimation of Regression Functions with both Categorical and Continuous Data." Journal of Econometrics, 119(1), 99–130.

IX Racine JS, Li Q, Zhu X (2004). "Kernel Estimation of Multivariate Conditional Distributions.". Annals of Economics and Finance, 5(2), 211–235.

X - Wand M, Ripley, B., (2008)." Kern Smooth Functions for Kernel Smoothing R package", version 2.22-22, URL http://CRAN.R-project.org/package=KernSmooth.

XI - Watson, G., (1964). "Smooth Regression Analysis." Sankhya, 26(15), 359–372.

XII- Yin, Z., Liu, F.& Xie, Y., (2016). "Nonparametric Regression Estimation with Mixed Measurement Errors ". Applied Mathematics, (7), 2269-2284.