Comparison of some methods for estimating a polynomial regression model using simulation

Mahdi Younes Mohamed, Ahmed Dheyab Ahmed

University of Baghdad, College of Administration and Economics, Department of Statistics

ahmedthieb19@gmail.com

Article Info
Page Number: 693-706
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract

In this research, estimators and parameters of the second-order polynomial regression model were found when the random error distribution was long-tailed symmetric using the Modified Maximum likelihood Method and the Robust M method. These methods proved their efficiency more than the ordinary least squares method through comparison between them using mean square error and simulation for three sample sizes (60, 90, 120).

Keywords: polynomial regression, long-tailed symmetric distribution, Modified Maximum likelihood Method, robust m method, simulation.

Research extracted from master's thesis in statistics tited "Estimation of polynomial regression model when the error is distributed long tailed symmetric"

1 Introduction

Regression [3] is used to explain the relationship between two or more variables. One of these variables is called the dependent variable, and the rest are called explanatory variables. Regression models the data so that we can explain the relationship between the variables. The regression is divided into linear regression and is used when the exponent of the variables is one, i.e., of the first degree, and when there is one explanatory variable in the regression equation called the simple linear regression model, and it is called multiple linear regression if there is more than one explanatory variable. Nonlinear regression is used when the variables have an exponent greater than one, a polynomial, or a logarithmic formula.

2 Polynomial Regression

In polynomial regression [14, 1] the relationship between the explanatory variables Xs and the dependent variable Y is modeled, provided that X is of degree n. Polynomial regression is a type of nonlinear regression and has different fields of application such as economic fields, medical fields and other fields. The formula of the polynomial regression model[7] of degree k is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x^2_{i1} + \ldots + \beta_k x^k_{i1} + e_i \qquad \ldots\ldots\ldots\ldots \qquad (1)$$

It is called a simple regression model if k = 1, and a polynomial regression model of the second degree if k = 2 and of the third degree if k = 3.

The ordinary least squares (OLS) estimations method are not effective in estimating the parameters of the polynomial regression model because of the presence of outliers [5, 6] or extreme values whose source could be an error in reading or recording the data or statistical population; this is when some data values differ significantly from the rest and that the difference is not caused by an error but rather a real situation that exists in reality, so researchers Akkaya & Tiku (2008) re-modeled the multiple linear regression model as follows [13].

$$y_i = \theta_0 + \sum_{j=1}^{q} \theta_j u_{ij} + e_i \qquad 1 \le j \le q \;\; , 1 \le i \le n \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots. \; (2)$$

$$u_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad , \qquad \bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \qquad , \qquad s_j^2 = \sum_{i=1}^{n} \frac{1}{n} (x_{ij} - \bar{x}_j)^2$$

Waebe [17] also concluded in 2008 that the data that represents financial losses and returns are often distributed in a skewed distribution. In 2019, the researcher Kitic [9] explained that many biological applications cannot be analyzed using linear statistics. Puthenpura & Sinha (1968) [15] found that the ordinary least squares method is ineffective in the presence of anomalies, so they suggested using the Modified Maximum Likelihood Method, and AKKaya and Tiku (2018) [1] estimated the parameters of the multiple regression model using the Modified Maximum likelihood Method and the method of least Squares and comparing each of the mean and variance, and it was found that the Modified Maximum Likelihood Method is much better because it was less biased, and in Normolle Danielp (2003) [4] developed the robust M method to be used in nonlinear models, the second order polynomial regression model is in the following formula:

$$y_i = \theta_0 + \sum_{j=1}^{q} \theta_j u_{ij} + \sum_{j=1}^{q} \theta_{jj} u_{ij}^2 + \sum_{j=1}^{q-1} \sum_{k=j+1}^{q} \theta_{jk} u_{ij} u_{ik} + e_i \qquad \ldots\ldots\ldots\ldots\ldots..(3)$$

$$1 \le j \le q \qquad , \qquad 1 \le i \le n$$

q: Number of the independent variables, the model 3 can be written in matrix equation as follows:

$$Y = U\theta + e \qquad \ldots\ldots\ldots\ldots\ldots\ldots.. (4)$$

Where:

$$\theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_q \\ \theta_{11} \\ \vdots \\ \theta_{qq} \\ \theta_{12} \\ \vdots \\ \theta_{q-1,q} \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}$$

$$U = \begin{pmatrix} 1 & u_{11} & \cdots & u_{1q} & u_{11}^2 & \cdots & u_{1q}^2 & u_{11}u_{12} & \cdots & u_{1q-1}u_{1q} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & u_{n1} & \cdots & u_{nq} & u_{n1}^2 & \cdots & u_{nq}^2 & u_{n1}u_{n2} & \cdots & u_{n,q-1}u_{nq} \end{pmatrix}$$

3 Methods of Estimation

For the purpose of estimating the parameters of the second-order polynomial regression model, three methods will be used, namely, the Ordinary Least Square Method, the modified maximum likelihood method, and the robust M method. The formula of the Ordinary Least Square Method is:

$$\tilde{\theta} = \left(\grave{U}U\right)^{-1}(UY) \qquad \text{......................... (4)}$$

$$\sigma^2 = \frac{Ee^2}{n-c} = \frac{(Y-U\theta)'(Y-U\theta)}{n-c} \quad , \quad c = 1 + 2q + \frac{q(q-1)}{2}$$

3.1 Modified Maximum Likelihood

Assuming that the error follows the long-tailed sympatric distribution, the formula for the derivation of the Modified Maximum likelihood Method is as follows [12, 2]:

$$f(e) = \frac{\Gamma(p)}{\sigma\sqrt{k}\,\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{p-1}{2}\right)}\left\{1 + \frac{e^2}{k\sigma^2}\right\}^{-p} . -\infty < e < +\infty \qquad \text{......................... (5)}$$

$$E(e) = 0 . \ V(e) = \sigma^2 \ , \quad k = 2p - 3 \ , \quad t = \sqrt{\frac{v}{k}\,\frac{e}{\sigma}}$$

σ: (Scale Parameter), p: Shape Parameter, and the Maximum likelihood function is:

$$L = \prod_{i=1}^{n} f(e) \qquad , \qquad z_i = \frac{e_i}{\sigma}$$

$$\ln L = n\ln d - n\ln - p\sum_{i=1}^{n} \ln\left(1 + \frac{{z_i}^2}{k}\right) \qquad \dots\dots\dots\dots\dots\dots\dots \quad (6)$$

$$e_i = y_i - \left(\theta_0 + \sum_{j=1}^{q} \theta_j\, u_{ij} + \sum_{j=1}^{q} \theta_{jj}\, u_{ij}^2 + \sum_{j=1}^{q-1}\sum_{k=j+1}^{q} \theta_{jk}\, u_{ij}\, u_{ik}\right)$$

$$\frac{\partial L}{\partial \theta_0} = \frac{2p}{k\sigma}\sum_{i=1}^{n} \frac{z_i}{1 + \frac{z_i^2}{k}} = 0 \qquad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (7)$$

$$\frac{\partial L}{\partial \theta_j} = \frac{2p}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{q} \frac{u_{ij} z_i}{1 + \frac{z_i^2}{k}} = 0 \qquad \dots\dots\dots\dots\dots (8)$$

$$\frac{\partial L}{\partial \theta_{jj}} = \frac{2p}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{q} \frac{{u_{ij}}^2 z_i}{1 + \frac{z_i^2}{k}} = 0 \qquad \dots\dots\dots\dots\dots\dots (9)$$

$$\frac{\partial L}{\partial \theta_{jk}} = \frac{2p}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{q-1}\sum_{k=j+1}^{q} \frac{u_{ij} u_{ik} z_i}{1 + \frac{z_i^2}{k}} = 0 \qquad \dots\dots\dots\dots\dots (10)$$

$$\frac{\partial L}{\partial \sigma} = \frac{2p}{k\sigma}\sum_{i=1}^{n} \frac{{z_i}^2}{1 + \frac{z_i^2}{k}} = 0 \qquad \dots\dots\dots\dots\dots (11)$$

$$g(z_i) = \frac{z_i}{1+\frac{{z_i}^2}{k}} \qquad \dots\dots\dots\dots\dots\dots\dots\dots\dots (12)$$

The above equations contain difficult functions and by applying the modified maximum likelihood steps by placing the equations in the term ‘Variates Order’ by arranging the variables in ascending order and replacing (z_i) in the above equations with $(z_{(i)})$ as follows [(1)]:

$z_{(1)} \leq z_2 \leq z_{(3)} \leq \cdots \leq z_{(n)}$

$$\frac{\partial L}{\partial \theta_0} = \frac{2p}{k\sigma} \sum_{i=1}^{n} g(z_{(i)}) = 0 \qquad \ldots\ldots\ldots\ldots\ldots (13)$$

$$\frac{\partial L}{\partial \theta_j} = \frac{2p}{k\sigma} \sum_{i=1}^{n} \sum_{j=1}^{q} u_{ij} g(z_{(i)}) = 0 \qquad \ldots\ldots\ldots\ldots\ldots (14)$$

$$\frac{\partial L}{\partial \theta_{jj}} = \frac{2p}{k\sigma} \sum_{i=1}^{n} \sum_{j=1}^{q} u_{ij}^{2} g(z_{(i)}) = 0 \qquad \ldots\ldots\ldots\ldots\ldots (15)$$

$$\frac{\partial L}{\partial \theta_{jk}} = -\frac{2p}{k\sigma} \sum_{i=1}^{n} \sum_{j=1}^{q-1} \sum_{k=j+1}^{q} u_{ij} u_{ik} g(z_{(i)}) = 0 \qquad \ldots\ldots\ldots\ldots\ldots (16)$$

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{2p}{k\sigma} \sum_{i=1}^{n} z_i g(z_{(i)}) = 0 \qquad \ldots\ldots\ldots\ldots\ldots (17)$$

Then $g(z_i)$ is substituted by the following linear function[1]:

$$g(z_i) = \alpha_i + \beta_i z_{(i)}$$

To find estimations of α_iUsing the first two terms of the Taylor seriesis to β,

of $g(z_i)$ about $t_{(i)}$

$$g(z_{(i)}) = g(t_{(i)}) + (z_i - t_{(i)}) g'(t_{(i)}) \qquad \ldots\ldots\ldots\ldots\ldots (19)$$

$$\alpha_i = \frac{2\frac{t_{(i)}^{3}}{k}}{\left(1 + \frac{t_{(i)}^{2}}{k}\right)^2} \quad . \quad \beta_i = \frac{1 - \frac{t_{(i)}^{2}}{k}}{\left(1 + \frac{t_{(i)}^{2}}{k}\right)^2}$$

Assuming that q_i represents an estimation of the cumulative function $F(t_i)$ and takes the following formula:

$$q_i = \frac{i}{n+1}$$

Since t_i represents the inverse of the cumulative function, to find its estimation, the following is used:

$$F(t_i) = \frac{\Gamma(p)}{\sigma\sqrt{k}\,\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{p-1}{2}\right)} \int_{-\infty}^{t_{(i)}} \left(1 + \frac{e^2}{k\sigma^2}\right)^{-p} dz \quad \dots\dots (24)$$

$$q_i = \frac{\Gamma(p)}{\sigma\sqrt{k}\,\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{p-1}{2}\right)} \int_{-\infty}^{t_{(i)}} \left(1 + \frac{e^2}{k\sigma^2}\right)^{-p} dz \quad \dots\dots (25)$$

By making some integrations, we get an estimation of the value of t, and by substituting $\alpha_i + \beta_i z_{(n)}$ into equations 13, 14, 15, 16, 17

$$\frac{\partial L}{\partial \theta_0} = \frac{2p}{k\sigma} \sum_{i=1}^{n} \left(\alpha_i + \beta_i z_{((i))}\right) = 0 \quad \dots\dots (26)$$

$$\frac{\partial L}{\partial \theta_j} = \frac{2p}{k\sigma} \sum_{i=1}^{n} \sum_{j=1}^{q} u_{ij} \left(\alpha_i + \beta_i z_{(i)}\right) = 0 \quad \dots\dots (27)$$

$$\frac{\partial L}{\partial \theta_{jj}} = \frac{2p}{k\sigma} \sum_{i=1}^{n} u_{ij}{}^2 \left(\alpha_i + \beta_i z_{(i)}\right) = 0 \quad \dots\dots (28)$$

$$\frac{\partial L}{\partial \theta_{jk}} = \frac{2p}{k\sigma} \sum_{i=1}^{n} \sum_{j=1}^{q-1} \sum_{k=j+1}^{q} u_{ij} u_{ik} \left(\alpha_i + \beta_i z_{(i)}\right) \quad \dots\dots (29)$$

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{2p}{k\sigma} \sum_{i=1}^{n} z_i \left(\alpha_i + \beta_i z_{(i)}\right) \quad \dots\dots (30)$$

By solving the above equations, we get the Modified Maximum likelihood estimations as follows[1, 11]:

$$\theta = K + D\tilde{\sigma} \quad \dots\dots (31)$$

$$K = (W'\mathfrak{B}W)^{-1}(W'\mathfrak{B}Y) = (K_\ell) \quad \dots\dots (32)$$

$$D = (W'\mathfrak{B}W)^{-1}(W'\alpha I) \quad \dots\dots (33)$$

$$\tilde{\sigma} = \frac{B + \sqrt{B^2 + 4nc}}{2\sqrt{n(n-c)}}$$

$$\alpha = \mathrm{diag}(\alpha_i) \quad , \qquad I' = [1.1\dots\dots 1] \quad , \qquad \mathfrak{B} = \mathrm{diag}(\beta_i$$

3.2 Robust M method

It is one of the methods that depends on minimizing or reducing the residual function: (20, 18)

$$\hat{\theta}_M = \min_{\beta} \rho(y_i - \sum_{i=1}^{n} \rho(\theta_j u_{ij}) \qquad \dots\dots (34)$$

$$\hat{\theta}_M = \min_{\beta} \sum_{i=1}^{n} \psi\left(\frac{(y_i - \left(\theta_0 + \sum_{j=1}^{q} \theta_j u_{ij} + \sum_{j=1}^{q} \theta_{jj} u_{ij}^2 + \sum_{j=1}^{q-1} \sum_{k=j+1}^{q} \theta_{jk} u_{ij} u_{ik}\right)}{\sigma}\right)$$

To find σ, following formula is used:

$$\sigma = 1 \cdot 483[\mathrm{medain}|e_i - \mathrm{medain}(e_i)|] \quad \dots\dots (35)$$

Using the Huber function:

$$\rho(e) = \begin{cases} \frac{e^2}{6} \\ \frac{c\,|e| - c^2}{2} \end{cases} \qquad \dots\dots (36)$$

$$\psi(e) = \begin{cases} e & \text{if } |e_i| \le c \\ \mathrm{Csign}\ (e) & \text{if } |e_i| > c \end{cases} \qquad \dots\dots (37)$$

c=1.345, and the estimator $\hat{\beta}_M$ is found from the following formula: (13, 19,):

$$\hat{\theta}_M = (U'\, W\, U)^{-1}\ U'WY \qquad \dots\dots (39)$$

w_i stands for the weight function and is calculated by the following formula:

$$w_i = \frac{\psi \frac{(y_i - \sum_{i=1}^{n}(\theta_j u_{ij})}{\sigma}}{\frac{(y_i - \sum_{i=1}^{n}(\theta_j u_{ij})}{\sigma}}$$

3.3 Simulation

The default values that were used in the simulation for the parameters of shape and scale are (P=3,5,7) and $(\sigma^2 = 1)$. As for the default values for the parameters of the regression model, they are as follows:

cof	θ_0	θ_1	θ_2	θ_3	θ_{11}	θ_{22}	θ_{33}	θ_{12}	θ_{13}	θ_{23}
value	77.21	-8.79	-7.43	-0.05	-3.06	**-3.52**	**-1.73**	**-4.68**	**-2.08**	**-1.17**

The formula for generating a random variable is:

$$e = \frac{t\sigma}{\sqrt{\frac{v}{k}}}$$

t is the random data that is generated according to the MATLA program $t = \text{trad}(2p - 1)$ and three sample sizes were chosen (60,90,120) and the experiment was repeated 1000. To compare between the estimation methods, the mean square error (MSE) was used, as follows:

1. Parameters of the model by the formula:

$$MSR(\hat{\theta}) = \frac{\sum_{i=1}^{r}(E(\hat{\theta}) - \theta)^2}{R}$$

2. For the model by the formula:

$$MSR(\hat{\theta}) = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n - c}$$

R: The number of times the experiment was repeated. The program was written in MATLA language. The simulation results are as follows:

Table (1) MSE of model parameters for all estimation methods when P=3

n	**60**			**90**				**120**	
parameters	**ols**	**MMLE**	**M-EST**	**ols**	**MMLE**	**M-EST**	**ols**	**MMLE**	**M-EST**
θ_0	0.085391	322.5215	0.051626	0.057637	295.2409	0.035067	0.040806	284.3584	0.024551
θ_1	0.018077	149.1451	0.011372	0.012634	128.7969	0.007377	0.009336	123.0894	0.005422
θ_2	0.019532	59.17818	0.011887	0.011726	49.9603	0.007178	0.009996	47.09173	0.006127
θ_3	0.019172	31.17377	0.011932	0.012874	27.35147	0.007545	0.009287	26.12682	0.005553
θ_{11}	0.025042	8.438129	0.015185	0.015988	5.337014	0.009738	0.011281	4.849357	0.00668
θ_{22}	0.026141	9.723491	0.016087	0.015613	7.85317	0.009554	0.011065	7.3537	0.006409
θ_{33}	0.025014	9.277775	0.015472	0.015902	6.999886	0.009823	0.011597	6.511795	0.00698

θ_{12}	0.021078	17.96225	0.01328	0.012657	13.48186	0.00746	0.008233	12.29989	0.005029
θ_{13}	0.022571	32.6825	0.013568	0.013362	27.42217	0.007887	0.008764	26.11237	0.005318
θ_{23}	0.020516	12.88293	0.012343	0.013378	10.11803	0.0082	0.009412	9.588027	0.005855

Table (2) MSE of model parameters of all estimation methods when P=5

n	60			90				120	
parameters	ols	MMLE	M-EST	ols	MMLE	M-EST	ols	MMLE	M-EST
θ_0	0.086411	0.0000043	0.061121	0.057641	0.0000042	0.039307	0.041381	0.000004	0.027741
θ_1	0.020628	0.0000059	0.014594	0.011939	0.0000056	0.008486	0.009276	0.0000055	0.006327
θ_2	0.020502	0.0000034	0.014246	0.011668	0.0000032	0.007982	0.008546	0.0000031	0.006166
θ_3	0.018804	0.00000023	0.013377	0.01308	0.00000019	0.009334	0.008984	0.00000018	0.00632
θ_{11}	0.023547	0.00000025	0.016251	0.015179	0.00000021	0.011023	0.011124	0.00000019	0.007638
θ_{22}	0.026713	0.0000011	0.018741	0.015229	0.0000011	0.010891	0.011657	0.000001	0.008042
θ_{33}	0.026161	0.00000039	0.01863	0.015638	0.00000034	0.010951	0.011665	0.00000031	0.008122
θ_{12}	0.020664	0.0000006	0.015085	0.013248	0.00000057	0.009508	0.008805	0.00000054	0.006123
θ_{13}	0.023107	0.00000082	0.016449	0.011968	0.00000072	0.008335	0.009459	0.0000007	0.006643
θ_{23}	0.020105	0.00000027	0.01424	0.01385	0.00000023	0.010174	0.009347	0.00000022	0.006599

Table (3) MSE of parameters model of all methods of estimation when P = 7

	60			90					
ameters	ols	MMLE	M-EST	ols	MMLE	M-EST	ols	MMLE	M-E
	0.093566	0.000017	0.067695	0.057695	0.000016	0.040593	0.043648	0.000015	0.125
	0.019611	0.00000024	0.014469	0.011803	0.00000022	0.008648	0.00929	0.00000021	0.023
	0.019475	0.00000014	0.014445	0.01223	0.00000013	0.008657	0.009222	0.00000012	0.024
	0.020782	0.000000005	0.014946	0.012491	0.00000000043	0.009111	0.00855	0.000000004	0.025
	0.025861	0.00000001	0.019291	0.016656	0.00000000095	0.012238	0.011016	0.0000000085	0.029
	0.026257	0.000000024	0.019154	0.015856	0.0000000023	0.011338	0.011888	0.000000022	0.032

	0.026164	0.000000015	0.01895	0.014241	0.0000000014	0.010442	0.011402	0.000000013	0.03
	0.021792	0.000000013	0.016064	0.012719	0.0000000011	0.009415	0.009505	0.00000001	0.02
	0.020102	0.000000016	0.015166	0.012458	0.0000000014	0.009123	0.009525	0.000000014	0.02
	0.022106	0.0000000051	0.016621	0.012354	0.00000000041	0.009239	0.009671	0.0000000037	0.02

Table (4) MSE of model and all methods of estimation

	n	**ols**	**MMLE**	**M**	**The best**
P=3	**60**	1.202293	609.4715	1.127583	**M**
	90	1.130553	510.5617	1.079906	**M**
	120	1.084538	476.5779	1.048152	**M**
P=5	**60**	1.195609	0.994181	1.137852	**MMLE**
	90	1.128175	1.003263	1.091816	**MMLE**
	120	1.092455	0.996361	1.07367	**MMLE**
P=7	**60**	1.20235	0.999345	1.149155	**MMLE**
	90	1.123508	1.002293	1.090656	**MMLE**
	120	1.083188	0.992137	1.058376	**MMLE**

4 Analysis of the results

- **The mean square error (MSE) of the model parameters**

At P = 3 for all sample sizes, the best method is to estimate the mean squared error of θ_3 is the N method and for the rest of the parameters the M method is the minimum MSE

At P=5 and for all sample sizes, the MMLE method gives minimum MSE

At P=7 and for all sample sizes, the MMLE method gives minimum MSE

- **The mean square error (MSE) of the model**

At P=3 for all sample sizes, M method gives the minimum mean square error MSE

At P=7,5 for all sample sizes, the MMLE method gives the minimum mean MSE

5 Practical framework

Data on the level of sugar in the body were collected through one of the laboratories licensed by the Iraqi Ministry of Health (Gilgamesh Laboratory) for the year 2022 and for a sample size of 90 individuals, the following variables were calculated:

c-peptide: This test measures the level of c-peptide in the blood, which is a substance made in the pancreas along with insulin that works to control the level of glucose (blood sugar). Glucose is the main source of energy in the body. Insulin and c-peptide are produced by the pancreas at the same time and in approximately equal amounts. Therefore, this test can be a good way to measure insulin because it tends to stay in the body for a longer period and if the body does not produce enough insulin, this will be a sign of diabetes.

HBa1c: Hemoglobin A1c Test, or cumulative glucose test, which measures the amount of sugar in the blood bound to hemoglobin. The importance of this examination comes from the fact that it shows the average amount of glucose in the blood that is related to hemoglobin during the past three months, due to the fact that the life of red blood cells in the bloodstream is three months.

HBa1c hemoglobin test results	**Interpretation of results**
Less than 42 mmol/mol (5.6%)	non diabetic
Between 42 and 47 mmol/mol (% 5.7 - 6.4%)	Prediabetes
48 mmol/mol 6.5% or more	patients with type 2 diabetes

RBS random blood sugar: a random blood sugar test at any time of the day. This analysis measures the level of glucose in the blood, regardless of when the food was last eaten. More than one measurement can also be taken throughout the day. For normal people, random blood sugar levels do not change over time. Throughout the day, having levels that vary greatly throughout the day means there is a problem.

UREA: Urea is a natural waste product produced by the human body after eating. The liver breaks down the protein in the food, thus producing urea. The percentage of urea varies from one individual to another depending on age, gender and other factors. The normal percentage of urea in the blood is as follows: Men (8-24 milligrams/dL), Women (6-21) milligrams/dL), Children up to 17 years old (7-20 milligrams/dL).

Kolmogorov-smirnov tast was used and it was found that the data distribution is not long-tailed symmetric, so the data was processed by taking the standard degree of the dependent variable Y. Data was tested again and it was found that the distribution of error terms is long-tailed symmetric. The calculated value of the test D=0.1411 is less than the tabular values, significant level $K_{0.05} = 0.1434$and at $K_{0.01} = 0.1718$

It is proved that the tabular value is greater than the calculated one, which means accepting the null hypothesis H_0, meaning that the distribution of data is long-tailed symmetric.

6 Data Analysis Methods

The modified maximum likelihood method (MMLE) was used as it gives minimum MSE for parameters and the model than other methods and for all sample sizes, The mean square error was found by simulation, and this method needs initial values so these values were found using the minimum squares method.

Table (5) Estimated value of parameters of ordinary least squares method when p=7

Cof.	θ_0	θ_1	θ_2	θ_3	θ_{11}
Value	**-0.02303**	**0.003143**	**-0.24897**	**0.277306**	**0.013646**
Cof.	θ_{22}	θ_{33}	θ_{12}	θ_{13}	θ_{23}
Value	**0.053395**	**0.070051**	**-0.03018**	**-0.45665**	**0.3352**

Table (6) the estimated value of the parameters of the modified maximum likelihood method when p = 7

Cof.	θ_0	θ_1	θ_2	θ_3	θ_{11}
Value	**-0.02223**	**0.003171**	**-0.24895**	**0.277254**	**0.013605**
Cof.	θ_{22}	θ_{33}	θ_{12}	θ_{13}	θ_{23}
Value	**0.05333**	**0.07001**	**-0.03021**	**-0.45654**	**0.335203**

$$\hat{y} = -0.02223 + 0.003171u_{i1} - 0.24895u_{i2} + 0.277254u_{i3} + 0.013605u_{i1}^2 + 0.05333u_{i2}^2 + 0.07001u_{i3}^2 - 0.03021u_{i1}u_{i2} - 0.45654u_{i1}u_{i3} + 0.335203u_{i2}$$

7 Conclusions

1- The relationship between the independent variable RBC and the dependent variable C-PeP is positive. That is, an increase in RBC leads to an increase in C-PeP
2- The relationship between the independent variable HBAIC and the dependent variable C-PeP is negative, meaning that any increase in HBAIC leads to a decrease in C-PeP
3- The relationship between the independent variable UREA and the dependent variable C-PeP is positive. That is, an increase in UREA leads to an increase in C-PeP

8 Recommendations

- Using the modified maximum likelihood method to estimate the parameters of the polynomial regression model when the distribution of error terms is long-tailed symmetric
- I recommend relying on c-peptide to find out if a person has diabetes or not.
- Using some algorithms such as genetic algorithm and Iterative Reweighting Algorithm to estimate the regression model when the distribution of error terms is long-tailed symmetric.
- Using other variables different from the one that was used.

References

1. Akkaya, A.D., and Tiku, M.L. (2008) " Robust estimation in multiple linear regression model with non-Gaussian noise". Automatica 44, 407-417
2. Akkaya, A.D., and Tiku, M.L. (2010)." Estimation in multifactor polynomial regression under non- normality" . Pak. J. Statist. Vol. 26(1), 49-68
3. Al-Mashhadani, M. H., Hormuz, A. H. (1989). Statistics. College of Administration and Economics, University of Baghdad.
4. Daniel P. Normolle .(2003)."An Algorithm for Robust nonlinear Analysis of Radioimmunoassays and Othre Bioassays "Ann Arbor,MI 48109-2029 , 313-936-1013
5. David M. Rocke and David L.Woodruff, (1998) ," Some Statistical Tools for data Mining Applications", University of California. ,Davis.
6. Filzmoser, P.(2004) , " Amultivariate Outlier detection Method", Department of statistics and Probability Theory. Vienna, Austria, Volume 1,PP. 18-22
7. Islam,T., Shaibur,M.R. and Hossain,S.S.(2009)." Effectivity of Modified Maximum Likelihood Estimators Using Selected Ranked Set Sampling Data". AUSTRIAN JOURNAL OF Statistics Volume 38, Number 2, 109–120
8. Jajo, N. K. (1989). Robust estimator of the linear regression model. Master's thesis in Statistics, Second College of Education, Ibn Al-Haytham, University of Baghdad.
9. Kiliç , Muhammet (2020) , ―Using Genetic Algorithms For Parameter Estimation Of A Two-
10. M.L. Tiku, M.Q. Islam, and A.S. YILDIRIM,(2001), NONNORMAL REGRESSION. I. SKEW DISTRIBUTION, Commun. Stat. Theory Methods 30, no. 6 , pp. 1021–1045
11. M.L. Tiku, R.P. Suresh, (1992) , A new method of estimation for location and scale parameters, J. Stat. Plann. Inference 30 (2) 281–29
12. M.L. Tiku, S. Kumra, (1985) , Expected values and variances and covariances of order statistics for a family of symmetric distributions (Student's t), selected tables in mathematical statistics 141–270
13. N. R. Draper and H. Smith, (1998), Applied Regression Analysis, Third Edition, Wiley Interscience Publication, United States,
14. Ostertagova, E.(2012)"Modelling using polynomial regression" Procedia Engineering 48 500 – 506

15. Qumsiyeh, S.B. (2007)."Non-normal bivariate distribution estimation and hypothesis testing" . A thesis submitted of the graduate school of natural and applied sciences of middle east technical 1-4
16. Ungwon Yu, Soyoung Yang, Jinhong Kim, Youngjae Lee, Kil-Taek Lim, Seiki Kim, Sung-Soo Ryu, Hyeondeok Jeong.(2001)" A Confidence Interval-Based Process Optimization Method Using Second-Order Polynomial Regression Analys". Processes
17. Waeber , Rolf , Embrechts , Paul , Roy , Parthanil And Lysenko , Natalia (2008) , ―Multivariate Skew-Normal Distributions And Their Extremal Properties ― , Master Thesis Swiss Federal Institute Of Technology, Eth Zurich
18. Y. Susanti and H. Pratiwi, (2011), "Robust Regression Model for Predicting the Soybean Production in Indonesia", Canadian Journal on Scientific and Industrial Research, 2, No..318-328
19. Y. Susanti, H. Pratiwi, and T. Liana, (2009), Application of M- estimation to Predict Paddy Production in Indonesia, presented at IndoMS International Conference on Mathematics and Its Applications (IICMA), Yogyakarta,
20. Yuliana and Y. Susanti, (2008), Estimasi M dan sifat-sifatnya pada Regresi Linear Robust, Jurnal Math-Info, 1, No. 8-16