مجلة كلية التربية الاساسية
كلية التربية الاساسية – الجامعة المستنصرية
Journal of the College of Basic Education Vol.31 (NO. 129) 2025, pp. 80-91

Comparison Between The Maximum Likelihood Method And A Nonparametric Method For Estimating A Poisson Regression Model

dr. [(1)]Azhar Kadhim Jbarah
Statistics Department /Collage of Administration And Economics
Al-mustansiriyah University
azkdfr_2017@uomustansiriyah.edu.iq

[(2)]Hasanain Jalil Neamah Alsaedi
University of Information Technology and Communications
Hasanien.1975@uoitc.edu.iq

[(3)]Taera Najim Abdullah
Statistics Department /Collage of Administration And Economics
Al-mustansiriyah University
tha_alameer@uomustansiriyah.edu.iq

Abstract:

Numerical models are those that deal with specific numbers in health research and studies, such as the number of hospitalized patients, the number of phone calls, the number of cases of a particular disease, and other counts. Among the models that deal with such numbers are the Poisson model, negative binomial regression model, logistic model, and Bernoulli . In this research, the Poisson regression model was used, and the model parameters were estimated using parametric methods (Maximum Likelihood) and non-parametric methods (Spline Regression Method). Due to the complexity of the assumptions in the Poisson regression model, which makes it unsuitable for the least squares method, an alternative method was used to estimate the model for its flexibility and simpler assumptions. A comparison was made between these methods according to specific criteria, relying on real data related to one of the most prevalent diseases today: COVID-19, which affects the human body and causes various inflammations, including respiratory, liver, and kidney inflammation, among others. The data was analyzed using SPSS and R software, leading to the conclusion that the estimation method using the non-parametric approach (cubic spline regression) was better than the parametric method (MLE) based on the results of the MSE

Key words: Poisson regression , Maximum Likelihood Method, Spline regression, Mean square error

Not: The research is not based on a master's thesis or a doctoral thesis.

1- Introduction

Statistical research focuses on many areas, with one of the most important being the health field. This study addresses the infection caused by the COVID-19 virus, a global pandemic that affects the respiratory system, kidneys, and other body organs. It is considered one of the infectious diseases that can be transmitted from an infected person to a healthy one.

This research addresses the use of one of the numerical models, including the Poisson model, which is one of the forms of non-linear regression models. This model assumes that the response variable follows a Poisson distribution with equal mean and variance. There are several methods for estimating the Poisson regression model, including parametric and non-parametric methods, and a comparison between these methods based on specific criteria.

One of the most important previous studies in 2009, researchers Terzi and Cengiz modeled the relationship between the number of recorded respiratory patients and air pollution using a multiple Poisson regression model. They employed the Generalized Additive Linear (GAL) model as a flexible approach to determine the model's form through nonparametric smoothing. In 2010, researchers Feldens and others used Poisson Regression Analysis with Robust Variance to determine the impact of factors associated with painful dental injuries. They found that the Poisson regression model with robust variance was superior to the logistic regression model in estimating the effect of factors influencing dental injuries. In 2013, researcher Hussam Mufid Sabri compared methods for estimating the parameters of the Poisson regression model in the presence of multicollinearity, using the number of congenital heart defects in children as a case study. He also proposed two methods for estimation.

2- Poisson regression model

Poisson regression model It is one of the countable linear logarithmic models. It is also a means in which the dependent variable is modeled when its value is countable. In this model, it is assumed that the response variables follow a specific distribution, which is a Poisson distribution with a parameter α. It is also assumed that the response variable (the dependent variable) The Poisson distribution also has a mean and variance (θ), and the random errors follow the same distribution in the Poisson regression model with its parameter (θ) (6)

The Poisson regression model can be written as follows

$$z = e^{x\underline{\theta}+e} \qquad (1)$$

Since

z: represents the vector of the dependent variable, whose rank is (m*1).
x : represents the matrix of explanatory variables whose rank is (m*(p+1))
$\underline{\theta}$: represents a vector of model parameters whose order is (m*1)
e: represents the vector of random errors, which has the order (m*1)
It can be written in array form as follows

$$Z = \exp(x\underline{\theta} + e) \qquad (2)$$

2- The most important assumptions of the Poisson regression model
Poisson regression assumes several assumptions, including [6]
First assumption:
The dependent variable (response variable (Z)) follows a Poisson distribution with parameter θ when θ >0
As in the following figure
1- The dependent variable (response variable (Z)) follows a Poisson distribution with parameter θ when θ >0
2- Also, one of its assumptions is independence, meaning that all of its observations (z_j, x_j) are independent from each other, meaning that they have a different distribution.
3- The mean and variance of the distribution parameter The dependent variable is the same as the mean and variance of the values of the response variable according to the following formula.

$$E(Z_i) = e^{x^T\underline{\theta}} \qquad (3)$$

By integrating the first assumption with the second, we obtain the conditional probability function as follows

$$f(z_i \backslash x_i) = \frac{e^{e^{x_i^T\underline{\theta}} * e^{zx_i^T\underline{\theta}}}}{z_i!} \qquad (4)$$

After verifying the three assumptions mentioned above regarding the Poisson regression model, this model achieves the exponential (linear-logarithmic) conditional mean and variance functions, as shown in the following formula:

$$E(z \backslash x) = \alpha = \exp(x^T\underline{\theta}) \qquad (5)$$

$$var(z \backslash x) = \alpha = \exp(x^T\underline{\theta}) \qquad (6)$$

3- Estimating the parameters of the Poisson regression model
The parameters of the Poisson regression model are estimated according to parametric methods, the most commonly used is (the maximum likelihood method) and the most common nonparametric methods are (the kernel method and the partial spline method).

4-Maximum Likelihood method [3],[7]

The parameters of the Poisson regression model are estimated by maximizing the observations of the distribution of the dependent variable (Z), as the maximum possibility function represents the product of the multiplication of the conditional probability mass functions for each of the observations of the dependent variable (Z), which follows the Poisson distribution, as shown in the first hypothesis and as follows:

$$\log L(\underline{\theta},\underline{Z},\underline{X}) = \log \prod_{i=1}^{n} f(z_i/x_i,\underline{\theta}) = \sum_{i=1}^{n} \log f(z_i/x_i,\underline{\theta}) \quad (7)$$

$$= \sum_{i=1}^{n} \frac{exp(-\sum_{i=1}^{n} e^{(x_i^T\theta)}_{(e^{\sum_{i=1}^{n} z_i x_i^T\theta})}}{\sum_{i=1}^{n} z_i!} \quad (8)$$

$$\ln l(\theta) = -\sum_{i=1}^{n} e^{(x_i^T\theta)+\sum_{i=1}^{n} z_i x_i^T\theta-\sum_{i=1}^{n}\ln(z_i!)} \quad (9)$$

By equating equation (9) with the zero value, which is one of the conditions for the process of maximizing the function

$$\frac{\partial lnl(\theta)}{\partial\theta} = -\sum_{i=1}^{n} x_i^T e^{(x_i^T\theta)} + \sum_{i=1}^{n} z_i x_i \quad (10)$$

Thus, we note that formula (10) is non-linear and can be solved using iterative numerical methods (Stephen Raphson's method). Iterative creativity is easier to find the appropriate solution. If this method is decided to give an initial estimated value for the parameters of the regression model $\hat{\theta}_0$and thus we obtain a second-order approximation. The logarithm of the possibility function around the common values $\hat{\theta}_0$ has the following formulas:

$$l^*(\theta_0) = l(\hat{\theta}_0) + S_k(\hat{\theta}_0)^T(\theta+\hat{\theta}_0) + 1/2(\theta+\hat{\theta}_0)^T H_k(\hat{\theta}_0)(\theta+\hat{\theta}_0) \quad (11)$$
$$\approx l(\hat{\theta}_0)$$

The Newton-Raphson iterative formula, considering $\hat{\theta}_0$ as an initial estimated value, has the following formula:

$$\hat{\theta}^{t+1} = \hat{\theta}^t - [H_n(\hat{\theta}_t)]^{-1} s_n(\hat{\theta}_t) \quad (12)$$

The estimation process is stopped when the difference value (θ^(t+1)-(θ_t)ˆ) is very close to zero.

5- Nonparametric methods for estimating the Poisson regression model[3]

There are many nonparametric methods that can be used to estimate the Poisson regression model, and we will suffice with one method

Spline regression method

6- Spline regression method [2],[1]

The Spline regression method is a nonparametric method. The researcher Osullivan was the first to use spline bootstrap models for the Poisson model using the penalized potential function. The logarithm of the expected value of

the response variable for the Poisson regression model is represented by the following formula:

$$Log(\mu(x)) = s(x) \quad (14)$$

As s(x) is a function that is modeled using cubic splines according to the formula

$$s(x) = \sum_{i=1}^{K} \theta_i B_i(x) \quad (15)$$

whereas

B_i: Basis functions for cubic splines of the function

θ_i Slice parameters

Equation (15) can be written in another form

$$s(x) = B\theta \quad (16)$$

B: represents the matrix of basis functions for cubic splines (n*k)

We conclude that the logarithm of the maximum potential function is written according to the following formula

$$l(\theta) = \sum_{i=1}^{n} [\, z_i B_{xi}\, \theta - \exp(B_{xi}\, \theta) - \log(z_i!) \quad (17)$$

To estimate the vector of parameters θ, we use the penalty potential function method minus the penalty term $(\int s(x))^2 dx$, which is used to measure the smoothing amount,

Since

s(x))^2 :represents the second derivative of the function S.

The logarithm of the partial potential function is written in the following form

$$l_{pen}(\theta) = l(\theta) - 1/2 \int s(x))^2 dx \quad (18)$$

To summarize the logarithm of the penal potential function, it can be approximated to be according to the formula

$$l_{pen}(\theta) = l(\theta) - \frac{\lambda}{2}\, \theta^T P \theta \quad (19)$$

Since

λ: represents the bootstrap parameter

To estimate the vector of parameters θ, we derive the logarithm of the penal potential function shown in equation (19) and set the derivative equal to zero. The solution to the system of equations is found based on the multi-stage returned weights least squares (IRWLS) method.

$$(B^T W B + \lambda P) = B^T W Z^* \quad (20)$$

Since

W: A diagonal matrix with dimension (n*n) and written according to the following formula

$$W = diag[\exp(B_0)] \quad (21)$$

Z^*: represents a vector of values of the response variable with dimension (n*1).

$$Z^* = W^{-1}(Y - \mu) + B\theta \quad (22)$$

The vector of coefficients of the partial slices ($\hat{\theta}_{Open}$) is estimated in the stage (t+1) according to the following formula:

$$\hat{\theta}_{t+1} = (B^T \hat{\theta}_t B + \lambda P)^{-1} B^T \widehat{W}_t \hat{Z}_t^* \quad (23)$$

Whereas, $\widehat{W}_t$ ، $\hat{Z}_t^*$are calculated based on estimating the vector of coefficients of the penalty segments at stage t.

4- Selection of the smoothing parameter λ(1)

To choose the optimal value for the smoothing parameter λ, based on the Akaikes information criterion (AIC), according to the following formula:

$$AIC = Dev(Z; \theta, \lambda) + 2tr(S_\lambda) \quad (24)$$

Since

Dev(Z;θ,λ): The deviation statistic represents one of the measures of goodness of fit and is calculated according to the following formula:

$$Dev = \sum_{i=1}^{n} [Z_i \log(\frac{Z_i}{\hat{\mu}})] \quad (25)$$

S_λ: represents the smoothing matrix and is calculated according to the following formula

$$S_\lambda = B(B^T W B + \lambda P)^{-1} B^T W \quad (27)$$

In addition to the Bayesian Information Criterion (BIC), which is better than the Akiyake criterion, especially in large samples, it is calculated as in the following formula:

$$BIC(\lambda) = Dev(Z; \theta, \lambda) + LN(N) + 2tr(S_\lambda) \quad (28)$$

7- **Comparison criteria**

One of the most important comparison criteria for the Poisson regression model and methods for estimating its parameters.

8- Mean Squared Error (MSE)(8)

This criterion is calculated using the following formula

$$Mse = \frac{\sum_{i=1}^{k} (\hat{C}_i - C)}{k}$$

Where

$\hat{C}$:represents the estimated value using the estimation methods.

C: represents the true value.

K: represents the number of repetitions in the estimation process.

9-The application side

A set of real data was analyzed to build a Poisson regression model and estimate the model parameters according to parametric and nonparametric

methods. The data was obtained from Kirkuk General Hospital with a sample size of 110 patients, A patient represented by the number of times they have contracted COVID-19, which serves as the response variable (Z), and four independent variables that represent: x1:sex,age, and X2: The patient's age in years represents,X3: Blood sugar level represents,X4: Heart rate represents.

10-Test data

Through the Easyfit program, the fit of the Poisson distribution for the dependent variable has been verified. It has been found that the dependent variable Z follows a Poisson distribution with parameter $\alpha = 1.527$.

Table (1) and the Kolmogorov-Smirnov test results indicate a good fit for the number of COVID-19 cases, showing that the Poisson distribution is the most prominent distribution that the dependent variable Z can follow.

Table (1)Goodness-of-fit tests

[#3] Poisson				
Kolmogorov-Smirnov				
Sample Size				110
Statistic				0.188
p-Value				0.64
Rank				1
□	0.2	0.1	0.05	0.01
Critical Value	0.565	0.642	0.708	0.828
Reject?	No	No	No	No

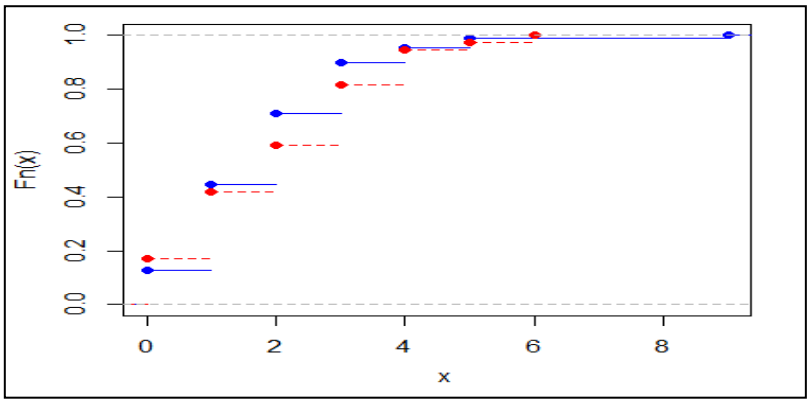

Figure (1):Goodness of fit for the cumulative distribution function of the Poisson distribution and disease incidence count data

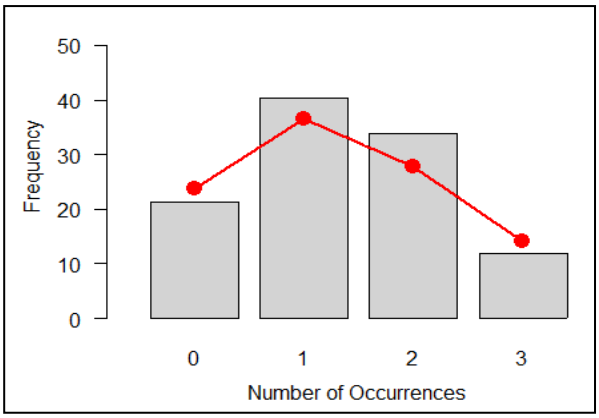

Figure (2):Goodness of fit for the disease incidence count data

11- Detecting the presence of multicollinearity.

To detect the presence of multicollinearity among the explanatory variables, we will find the Variance Inflation Factor (VIF) for the explanatory variables, as shown in Table 2.

Table (2)Results of Detecting Multicollinearity Issues

Collinearity Statistics	
Tolerance	VIF
.928	1.077
.930	1.075
.880	1.137
.987	1.013

Indicates table (2) that the Variance Inflation Factor (VIF) values are less than ten, which suggests that there is no impact from multicollinearity among the explanatory variables.

12-Estimating A Poisson regression model

Depending On the estimation methods, the Poisson regression model will be estimated using the parametric methods in question

- Estimation according to the maximum likelihood method

The statistical program SPSS was used to estimate the parameters of the Poisson regression model according to the Maximum likelihood method, and the estimation results are shown in Table (3).

Table (3)Results of estimates of the maximum likelihood method for the Poisson regression model

Parameter	**Parameter estimate**	S.E	Wald chi-squar	sig	Likelihood Ratio Chi-squar	Sig.
0θ	0.198	0.422	0.219	0.639		
1θ						0.757
$_2\theta$	-0.026	0.069	0.143	0.705	1.883	
$_3\theta$	0.000	0.0023	0.013	0.908		
$_4\theta$	-0.169	0.1550	1.196	0.274		
	-0.028	0.856	0.111	0.739		

In the table below, the parameters of the Poisson regression model were estimated, as the results show that the model parameters are not significant at the level of 0.05 according to the Wald test, meaning that there is no effect for the five variables on the number of times infected with Corona disease, and this is also confirmed by the Likelihood Ratio Chi test square was found to have a non-significant value at the same level.

Table(4)Goodness of fit test

Deviance	value/df	person	value/df	Scal person	value/df	Scale parameters
19.221	105	18.906	105	105	105	0.180

The table below indicates the goodness-of-fit measures (Deviance, Pearson) Showing that the data exhibit under-dispersion, as illustrated by the results in the table below. Since the values of the Pearson statistic and the Deviance statistic were lower than the degrees of freedom, the analysis was conducted based on the Pearson statistic, and the scale parameters were estimated by dividing the Pearson statistic by the degrees of freedom.

Table (5)Results of estimates of the cubic spline regression method for the Poisson regression model

Parameter	**Parameter estimate**	S.E	Wald stat	sig	Smoothing parameter
B_0	1.278	0.109	11.674	0.0006	0.544
B_1	0.104	0.088	0.096	0.756	
B_2	0.102	0.025	4.080	0.043	
B_3	0.345	0.037	9.247	0.002	
B_4	0.190	0.030	6.206	0.012	

Table (5) displays the results of the estimated values according to the spline regression method for the Poisson model, which indicates the significance of the estimated parameters below the 0.05 significance level based on the wald-test value.

Table (6)Results of quality and goodness of fit measures

Method	M.S.E	R^2
MLK	0.93	**0.55**
SPLIN	**0.416**	**0.65**

The above table indicates that the results of the Pseudo R-Square regression coefficient according to the cubic spline regression method equal 0.65, which shows the quality of the estimated Poisson regression equation. The table also indicates that the Mean Square Error value according to the cubic spline regression method was lower than the maximum likelihood method, reaching 0.416

13-Conclusions

1. It was found that the Poisson regression model estimated using the maximum likelihood method was generally non-significant.
2. Based on the results of the Mean Square Error criterion in Table (6), we conclude that the non-parametric method for estimating the Poisson regression model is efficient and serves as a good alternative to the maximum likelihood method.
3. In light of the results from the comparison of Poisson regression model estimations and the cubic spline regression method, we conclude that the spline regression method has all its parameters significant, in contrast to the maximum likelihood method.

4. The results showed that the values of the deviance statistic and Pearson statistic for the Poisson regression models estimated using the maximum likelihood method indicated that the response variables suffer from under-dispersion, as their values were lower than the degrees of freedom, as shown in Table (3).

References

1. Durban M.,Currie,I.and Eilers,P.,2001,'Using P-Splines to Smooth Two-Dimensional Poission Data'Proceedings of the 17th International Workshop on Statistical Modeling.Chania,Greece,Eds,M.Stasinopoulos and G.Touloumi,207-214.
2. Li,C.S.and,Tu,W.,2007,"ASplin-Based Lack-Of-Fit, Test for independent variable Effect in poisson Regression Notional Institutes of health,vol,6,issue 1,pp.(239-247) .
3. Mansson, K & Shukur , G (2011) A Poisson Ridge Regression Estimator " , Economic Modeling, Vol. 28, Issue 4 ,pp. 1475-1491
4. M. V. Myers, Generalized Linear Model with Applications in Engineering and Sciences,2th Edition (New Jersey: John Wiley & Sons, 2010) .
5. Prahutama, Alan. "Modelling infant mortality rate in Central Java, Indonesia use generalized poisson regression method." Journal of Physics: Conference Series. Vol. 1025. No. 1. IOP Publishing, 2018.
6. Rodriguez,G.,2007,"generalized linear models",Princeton university,Revised September 2007.
7. Sultan, Maha Hassan, 2017, "The Maximum Likelihood Method and Some Nonparametric Methods in Estimating the Poisson Regression Model," Master's Thesis, College of Administration and Economics, Al-Mustansiriya University.
8. Wagh, Yogita S., and Kirtee K. Kamalja. "Comparison of methods of estimation for parameters of generalized Poisson distribution through simulation study." Communications in Statistics-Simulation and Computation 46.5 (2017): 4098-4112.

مجلة كلية التربية الاساسية
كلية التربية الاساسية – الجامعة المستنصرية
Journal of the College of Basic Education Vol.31 (NO. 129) 2025, pp. 80-91

المقارنة بين طريقتي الامكان الاعظم واحدى الطرائق اللامعلمية في تقدير انموذج انحدار بواسون

م.د.ازهار كاظم جبارة (1)
كلية الادارة والاقتصاد/الجامعة المستنصرية/قسم الاحصاء
azkdfr_2017@uomustansiriyah.edu.iq
م.م.حسنين جلال نعمه (2)
جامعه تكنلوجيا المعلومات والاتصالات
Hasanien.1975@uoitc.edu.iq
م. م. ثائرة نجم عبدالله (3)
كلية الادارة والاقتصاد/الجامعة المستنصرية /قسم الاحصاء
tha_alameer@uomustansiriyah.edu.iq

مستخلص البحث:
تعتبر النماذج العددية من النماذج التي تتعامل مع اعداد خاصة في الابحاث والدراسات الصحية مثل اعداد المرضى الراقدين او عدد المكالمات الهاتفية او عدد الاصابات بمرض معين وغيرها من الاعداد ومن النماذج التي تتعامل مع الاعداد انموذج بواسون وانموذج انحدار ذي الحدين السالب والانموذج اللوجستي وانموذج برنولي ،تم في هذا البحث استخدام انموذج انحدار بواسون وتقدير معالم الانموذج بالطرق المعلمية (Maxliklihood) والطرق اللامعلمية (Spline Regression Method)، ونتيجة لصعوبة وتعقيد الفرضيات الخاصة بطريقة المربعات الصغرى في تقدير انموذج انحدار بواسون ، تم استعمال بديل اخر لتقدير الانموذج وذلك لمرونته وبساطة فرضياته، و تمت المقارنة بين تلك الطرق وفق المعايير الخاصة ،اذ تم الاعتماد على بيانات حقيقية تخص اهم الامراض المنتشرة في وقتنا الحالي وهو مرض كورونا الذي يصيب جسم الانسان ويسبب التهابات عدة منها التهاب الجهاز التنفسي والتهاب الكبد والكلى والتهابات اخرى وتم تحليل تلك البيانات وفق برنامج Spss وبرنامج R والتوصل وحسب النتائج الى ان طريقة التقدير وفق طريقة اللامعلمية (انحدار الشرائح التكعيبية) كانت افضل من الطريقة المعلمية (MLE) من خلال ما فرزته نتائج معيار MSE .

الكلمات المفتاحية: انحدار بواسون ،طريقة الامكان الاعظم،طريقة الشرائح، متوسط مربعات الخطأ.