



## Prediction of Tuberculosis in Iraq Using A ZIPR Model

Afraa A. Hamada<sup>1,\*</sup>

<sup>1</sup>Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, Iraq

Email: [Afraa.Hamada@qu.edu.iq](mailto:Afraa.Hamada@qu.edu.iq)

### Abstract

In this article, the ZeroInflated Poisson Regression model (ZI-PRM) was used to predict the number of tuberculosis patients by estimating the model using the maximum likelihood method and compared with Poisson regression model (PRM). The results showed that the ZIPRM best represented TB data from PRM. The PRM showed that the importance of some variables, although they were not significant as a cause of the TB data. The ZIP model indicates that there will be more TB cases in 2027 than there were in 2023. These findings point to an improvement in the nation's health status.

**Keywords:** ZIP Model; Tuberculosis; Estimation; Maximum Likelihood; Prediction

### 1. Introduction

Research on epidemiology and public health frequently finds that a significant percentage of counting data consists of zeros. For instance, the service use count in a health care utilization research frequently has a high percentage of zeros, which stands for patients who did not use any services throughout the study period. Subjects of interest in the field of substance addiction have varied frequencies of drug or alcohol use, and many patients report not using any drugs or alcohol at all while undergoing treatment. The count scale of this kind of data frequently has a plus zero above the typical count distributions it can support, such Poisson. For instance, if one counts the number of reactions to an illness, one could not count the responses if the person is immune or resistant to the disease. Previous studies have demonstrated that an improbable fit of both zeros and non-zero numbers will arise if the trailing zero is not taken into consideration. Mycobacterium tuberculosis is the bacterium that causes TB, an infectious illness. The brain, kidneys, and spine may also be harmed, even though the lungs are the main organs afflicted. TB is transmitted through the air, with people breathing in bacteria spread when an infected person talks, coughs, or sneezes. The health situation in the world varies from one country to another, and is affected by many factors such as health infrastructure, economic situation, and social and cultural factors. In countries with limited resources, there may be challenges in providing and accessing adequate health care, increasing the prevalence of TB and complicating treatment and prevention processes. Governments, international organizations, civil society, and the commercial sector must work closely together to strengthen infrastructure, raise public awareness, and provide high-quality healthcare in order to combat. In (2015) (Yang S, Berdine G.) Poisson regression was used to evaluate the risk variables connected to the duration of hospital stays for children with asthma [1]. In (2018) (Kamalja & Wagh) used the General Poisson model and the New ZI Regression Mode when the estimators follow the Lindley distribution, and the hypothetical model was applied to real data [2]. In the (2020) (Muche & Mekonnen) aim to identify the best statistical model for estimating and predicting mortality of children under the age of five in Ethiopia, based on data from the Ethiopian Demographic Health Survey conducted in 2016 [3]. Zero-inflated regression models are used in (2021) by L. H. Hashim et al. to analyze rainfall data and differentiate between Zero-Inflated Poisson (ZIP) regression model, which is superior than Poisson regression model [4]. To handle dispersion on poisson regression, Mutia and Darnius (2023) use Zero-Inflated poisson regression testing [4]. This paper aim to estimate the number of TB in iraq by using ZIPRM and compared it with PRM.

**2. Hyper Dispersion [6]**

A situation when the observed variance of the data exceeds what the traditional NB distribution would have anticipated. The Poisson distribution is often used for modeling count data with excess dispersion, or when the variance is greater than the mean. The idea of hyper-dispersion arises when, in some circumstances, the observed data show much more variability than can be explained by a conventional Poisson model. The underlying variability among persons or things is assumed to be sufficiently reflected by the basic NB model. Hyper-dispersion may result if other sources of heterogeneity are not taken into consideration. In the Poisson distribution, the likelihood of finding 0 counts is thought to be independent of the other distribution parameters. Regular patterns or additional zeros in the data that the model cannot explain might cause hyper-dispersion.

**3. Poisson Regression model (PRM) [7] [8]**

The model deliberately finds a relationship between a dependent variable (y) that follows the Poisson distribution and several independent variables, and the relationship is as follows:

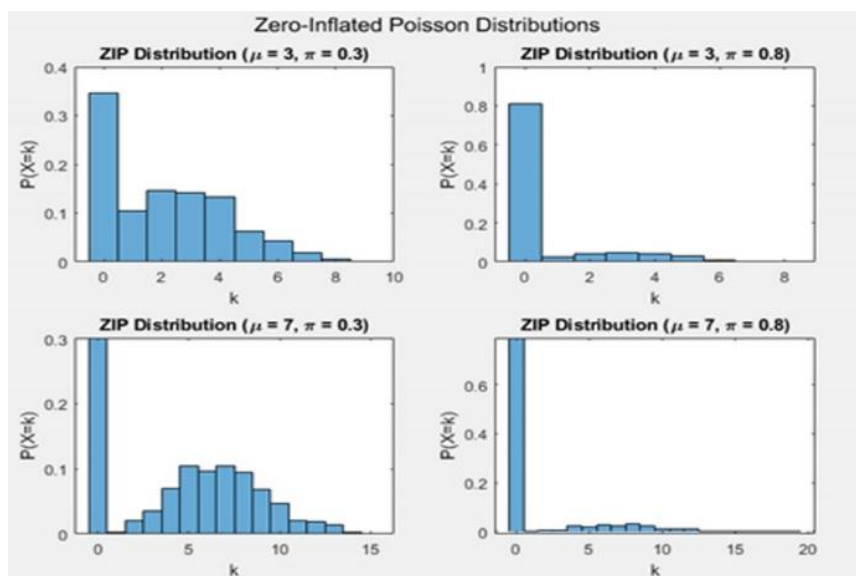
$$p(Y = y_i/M_i) = \frac{\exp(-M_i)M_i^{y_i}}{y_i!} \quad ; y_i = 0, 1, 2, \dots \quad , i = 0,1,2, \dots \tag{1}$$

The Poisson distribution's mean and variance are:

$$E(y_i) = Var(y_i) = M_i \tag{2}$$

**4. Zero Inflated Regression(ZIR) [9] [10]**

When analyzing countable data, like the total number of ER visits a patient receives annually or the total number of fish caught in a single day in a lake, zero models are often used. Count data may include non-negative integer values like 0, 1, 2, ... Some examples of data counting include the number of hospital days a patient has, the number of goals scored in a football game, the number of visits logged in a minute by a Geiger counter, and the number of hypoglycemia episodes a diabetic patient has on a yearly basis. The inflated mass of zero and the countable distribution make up the two mathematical components of the inflated regression models. A Poisson distribution or a Poisson distribution is often used to depict the distribution of numbers for statistical study. It has long been believed that the Poisson regression serves as the foundational counting model for many other enumeration methods. In the Poisson model, the mean, often referred to as the coefficient rate or intensity, is represented by the parameter  $M_i$ , whilst the counting response is represented by the random variable  $y_i$ . In some phenomena, the number of zeros is greater than what would be expected using a Poisson distribution or a Poisson distribution. Data with such an extra number of zero is described as ZI. As an example of graphs of zero inflated Poisson distributions with a mean  $M_i = 3$  or  $7$  and a zero inflation rate  $\pi = 0.3$  or  $0.8$  as shown in Figure (1)



**Figure 1.** ZIPD with a mean of  $M_i = 3$  and  $7$  and a zero inflation rate of  $\pi = 0.3$  and  $0.8$

**5. Zero-Inflated Poisson Model (ZIPM) [11] [12]**

A statistical model known as the inflated zero-probability model is based on inflated zero-likelihood distributions, or distributions that permit frequent observations of zero values. The Diane Lambertis Poisson model with an inflated zero, which relates to the occurrence of a random variable that includes zero inflated data at the adopted unit of time, is one of the famous inflated zero models. The model relies on the number of events within a specific category that will not include inflated zeros for a class of observations that make the ZIPM consists of two processes that produce zeros: the first produces zeros and outputs zero for any matching probability value, the second process is governed by the Poisson distribution, which generates some of the processes that are recorded as an observation equal to zero. The mixture function resulting from the combination of zeros and non-zeros models. In other words, the probability of zeros is P and the probability of non-zeros is 1-P. Thus, the probability mass of the probability of zeros:

$$\Pr(Y = 0) = P_1 + (1 - P_1)e^{-M_1} \tag{3}$$

The probability function for Non-zeros is:

$$\Pr(Y = y_i) = (1 - P_1) \frac{M_1^{y_i} e^{-M_1}}{y_i!} \tag{4}$$

In other words, the probability function for non-zeros is a general Poisson distribution minus the probability of zeros in the distribution, that is,

$$\Pr(Y = y_i) = \frac{M_1^{y_i} e^{-M_1}}{y_i!} - P_1 \frac{M_1^{y_i} e^{-M_1}}{y_i!} \quad y_i = 1, 2, 3 \dots \tag{5}$$

Where  $(y_i)$  a positive and integer random variable is greater than zero,  $(M_1)$  the expectation parameter of the Poisson process which represents the mean,  $(P_1)$  the inflated zero probability, which is the probability of a process that yields a value of a corresponding variable equal to zero. The expectation of the number of Non-zero events can be expressed as  $M_1$ , and when equation (2) is applied,  $\pi_i$  will be,

$$\begin{aligned} M_1 &= \sum y_i (1 - P_1) \frac{M_1^{y_i} e^{-M_1}}{y_i!} \\ &= (1 - P_1) \sum M_1 \frac{M_1^{y_i-1} e^{-M_1}}{(y_i-1)!} \end{aligned} \tag{6}$$

Hence, the mean for non-zeros is:

$$\mu = (1 - P_1) M_1 \tag{7}$$

To obtain the variance, the following relationship can be applied:

$$E(y_i^2) = E(y_i(y_i - 1)) + E(y_i) \tag{8}$$

$$E(y_i(y_i - 1)) = (1 - P_1) \sum y_i(y_i - 1) M_1^2 \tag{9}$$

$$\begin{aligned} E(y_i^2) &= (1 - P_1) M_1^2 + (1 - P_1) M_1 \\ &= (1 - P_1) M_1 (1 + M_1) \end{aligned} \tag{10}$$

$$\begin{aligned} V(y_i) &= E(y_i^2) - (E(y_i))^2 \\ &= (1 - P_1) M_1^2 + (1 - P_1) M_1 - (1 - P_1)^2 M_1^2 \\ &= (1 - P_1) M_1 (M_1 + 1 - (1 - P_1) M_1) \\ &= (1 - P_1) M_1 (M_1 + 1 - M_1 + P_1 M_1) \\ &= (1 - P_1) M_1 (1 + P_1 M_1) \end{aligned} \tag{11}$$

From equation (11) we see that the variance exceeds the arithmetic mean due to  $(1 + P_1 M_1)$  is greater than (1), so the variance is greater than  $\mu$ . In other words, there is an enlargement of the variance as a result of the inflation of the zero operations of Poisson, it defies a fundamental tenet of the Poisson distribution: that the variance and arithmetic mean are identical.

**6. Parameter Estimations of the (ZIP) parameters [13]**

Suppose that we have observations of size (n) which are  $(Y_1, Y_2, \dots, Y_n)$  independent and identifiably distributed (iid) and have zero inflated Poisson distribution (ZIP) with parameters  $(\pi_i, M_i)$ . Therefore, the Maximum Likelihood approach will be used to estimate the model parameters. The Likelihood function will thus be:

$$L(P_i, M_i/Y = y_i) = \prod_{i=1}^n p(Y = y_i) \tag{12}$$

Assuming that (n) represents the number of observations (variables) that have the value (0) of the  $(Y_i)$  values, and that (n-m) represents the number of observations that do not have zero. And depending on equation (1) and equation (3), the Likelihood function can be obtained as,

$$L(P_i, M_i/Y = y_i) = [(P_i + (1 - P_i)e^{-M_i}]^m \prod_{i=1, y_i \neq 0}^n (1 - P_i)e^{-M_i} \frac{M_i^{y_i}}{y_i!} \tag{13}$$

Therefore, the logarithm of the Likelihood function in equation (13) is as follows:

$$\text{Ln}L = m\text{Ln}(P_i + (1 - P_i)e^{-M_i}) + (n - m)\text{Ln}(1 - P_i) - (n - m)M_i + n\bar{y}\text{Ln}(M_i) - \text{Ln}(\prod_{i=1}^n y_i!) \tag{14}$$

By deriving equation (14) for the parameters  $(\pi_i, M_i)$  and equating the derivative to zero, we get the following equations:

$$\frac{m(1-\pi_i)e^{-M_i}}{\pi_i+(1-\pi_i)e^{-M_i}} + n - m - \frac{n\bar{y}}{M_i} = 0 \tag{15}$$

$$\frac{m(1-e^{-M_i})m(1-\pi_i)}{\pi_i+(1-\pi_i)e^{-M_i}} - n + m = 0 \tag{16}$$

**7. Applied Side**

One of the first illnesses that humans have ever encountered is TB, an infectious disease brought on by the bacteria Mycobacterium tuberculosis. Although its primary effect is on the lungs, tuberculosis may also damage the brain, kidneys, and spine [14] [15]. Data representing the frequency and percentage of people with tuberculosis were obtained from the Ministry of Health - Directorate of Planning and resources Development - department of Health and vital Statistics for two years (2022 and 2023), and the explanatory variables which are (Sex, Age, place of life (urban - rural), blood sugar ratio, alcohol consumption (yes - no), number of working hours, chronic asthma, smoking (yes - no), occupation (civilian - military - factories), exposure to chemicals (yes - no)). In order to ensure that this data is subject to the Poisson distribution, it was tested using (Easy Fit) program to test are the data have Poisson distribution or not by using Kolmogorov Smirnov and Anderson darling tests, results as table (1)

**Table 1:** real data statistic

| Test               | $\alpha$ | Sig.   |
|--------------------|----------|--------|
| Kolmogorov Smirnov | 0.01     | 0.7891 |
| Anderson darling   |          | 3.7848 |

The null hypothesis, which states that the sample's statistical distribution follows the Poisson distribution, is not rejected because the Sig. value of (0. 7891, 3. 7848) is greater than the significance level of 0.01 for both the test values (Kolmogorov Smirnov = 0. 7891) and the test value (Anderson darling = 3. 7848). This is demonstrated in Table 2.

Table (2) shows the descriptive statistics for the real data as follows:

**Table 2:** Descriptive Statistics for real data

| Statistic | Mean   | Variance | Person Chi-Square |
|-----------|--------|----------|-------------------|
| Value     | 3.9899 | 23.5568  | 0.8976            |

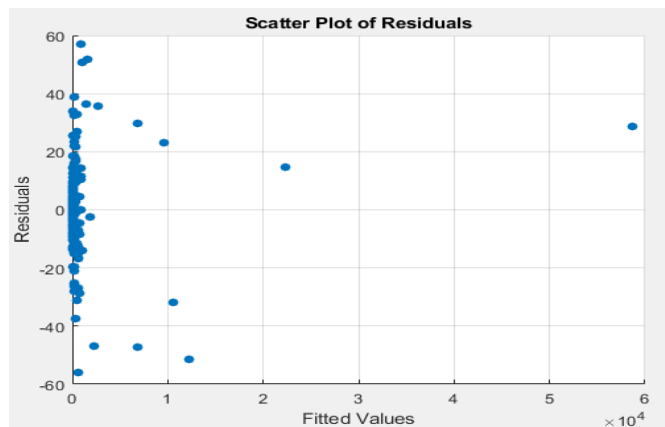
Table (2) makes it evident that the variance exceeds the real data's arithmetic mean and that the Person Chi-Square statistic value of (0.8976) is greater than (1). These findings point to the presence of an excessive dispersion issue in the real data and the ZIPD of the data.

7.1. PRM

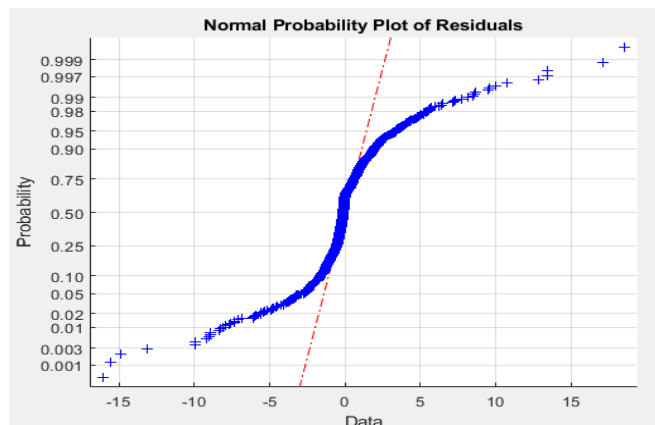
Real data were analyzed using the PR in the NCSS program, and the results were as follows:

**Table 3:** Poisson regression results

| Variable                | Coefficient  | Value    | Sig.     |
|-------------------------|--------------|----------|----------|
|                         |              | Constant | 4.5652   |
| Sex                     | $\beta_1$    | 0.0034   | <0.111   |
| Age                     | $\beta_2$    | 0.1133   | <0.001** |
| Place of life           | $\beta_3$    | 0.8787   | <0.231   |
| blood sugar ratio       | $\beta_4$    | 0.1495   | <0.211   |
| alcohol consumption     | $\beta_5$    | -0.0014  | <0.071   |
| number of working hours | $\beta_6$    | 2.9459   | <0.001** |
| chronic asthma          | $\beta_7$    | 0.0011   | <0.512*  |
| Smoking                 | $\beta_8$    | -0.0751  | <0.001** |
| Occupation              | $\beta_9$    | -0.1297  | <0.667   |
| exposure to chemicals   | $\beta_{10}$ | 0.6778   | <0.781   |



**Figure 2.** Residuals for the PRM for standardized data



**Figure 3.** Normality Test for the Residuals of the PRM for standardized data

Figures (3) and (4) explained that the residuals of the PRM while using standardized data do not adhere to the normal distribution.

The following is the Poisson regression equation:

$$\text{Log}(y) = 4.5652 + 0.0034X_1 + 0.1133X_2 + 0.8787X_3 + 0.1495X_4 - 0.0014X_5 + 2.9459X_6 + 0.0011X_7 - 0.0751X_8 - 0.1297X_9 + 0.6778X_{10} \tag{17}$$

Equation (17) indicates that the variables (age - blood sugar ratio - number of working hours - asthma - occupation - exposure to chemical substances) were not significant, while in medical reality these variables are important factors that greatly influence the incidence of pulmonary tuberculosis, and the coefficients for these factors also appeared. Variables with a sign contrary to reality, for example, the smoking variable showed significance, while the sign of the coefficient for this variable was negative, which is contrary to reality, meaning that the more smoking, the lower the probability of contracting pulmonary tuberculosis. The same situation for drinking alcohol. This indicates the inaccuracy of the PRM to estimate the studied phenomena which contain zero inflated.

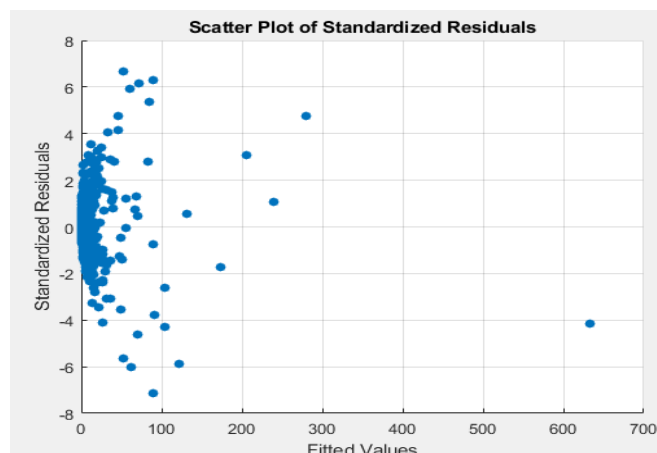
**7.2. ZIPRM**

The real data were analyzed using the ZIPRM by the MatLab program, and the results were as follows:

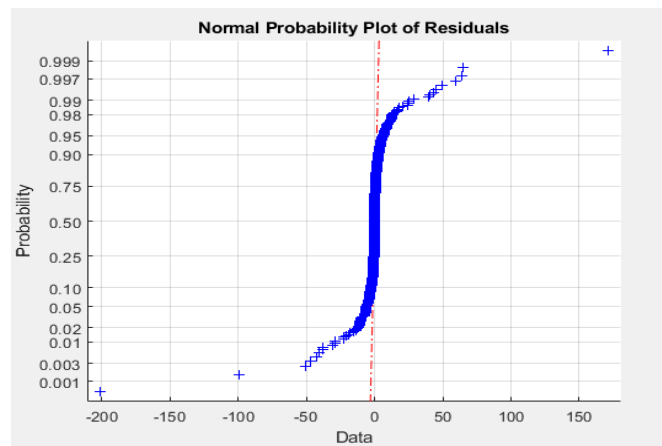
**Table 4:** Results of the ZIPRM

| Variable                       | Coefficient  | Value  | Sig.     |
|--------------------------------|--------------|--------|----------|
|                                | Constant     | 5.4428 | <0.001** |
| <b>Sex</b>                     | $\beta_1$    | 0.5674 | <0.111   |
| <b>Age</b>                     | $\beta_2$    | 0.5579 | <0.001** |
| <b>Place of life</b>           | $\beta_3$    | 1.4794 | <0.001** |
| <b>blood sugar ratio</b>       | $\beta_4$    | 1.6689 | <0.001** |
| <b>alcohol consumption</b>     | $\beta_5$    | 2.1554 | <0.001** |
| <b>number of working hours</b> | $\beta_6$    | 3.3356 | <0.001** |
| <b>chronic asthma</b>          | $\beta_7$    | 2.3355 | <0.001** |
| <b>Smoking</b>                 | $\beta_8$    | 4.8997 | <0.001** |
| <b>Occupation</b>              | $\beta_9$    | 1.5579 | <0.001** |
| <b>exposure to chemicals</b>   | $\beta_{10}$ | 1.7899 | <0.001** |

\*\*Refer to high significance



**Figure 4:** Residuals for the ZIPRM



**Figure 5:** Normality test for the Residuals of the ZIPRM for standardized data

Figures (4) and (5) showed that the residuals of the ZIPRM follow the Normal distribution for standardized data

The equation for a ZIPRM is as follows:

$$\text{Log}(y) = 5.4428 + 0.5674X_1 + 0.5579X_2 + 1.4794X_3 + 1.6689X_4 + 2.1554X_5 + 3.3356X_6 + 2.3355X_7 + 4.8997X_8 + 1.5579X_9 + 1.7899X_{10} \tag{18}$$

Equation (18) indicates the significance of the variables indicates that the variables (age - blood sugar ratio - number of working hours - asthma - occupation - exposure to chemical substances) with sign compatible with medical reality, this indicates the inaccuracy of the ZIPRM to estimate the studied phenomenon which contain zero inflated.

### 7.3. The Comparison between the PRM and the ZIPRM

The model was compared using Akaiki's Criteria Information and Bayes Akaiki's Criteria Information, and MSE as shown in table (5) as follows:

**Table 5:** Comparison between Models

| Model | AIC      | BIC      | MSE    |
|-------|----------|----------|--------|
| PM    | 2899.53  | 2845.243 | 2.3477 |
| ZIPM  | 1123.985 | 1162.547 | 0.0092 |

Table (5) showed that the superiority of the ZIPRM over the PRM, which indicates that the ZIPRM addressed the problem of excessive dispersion of zeros in the data.

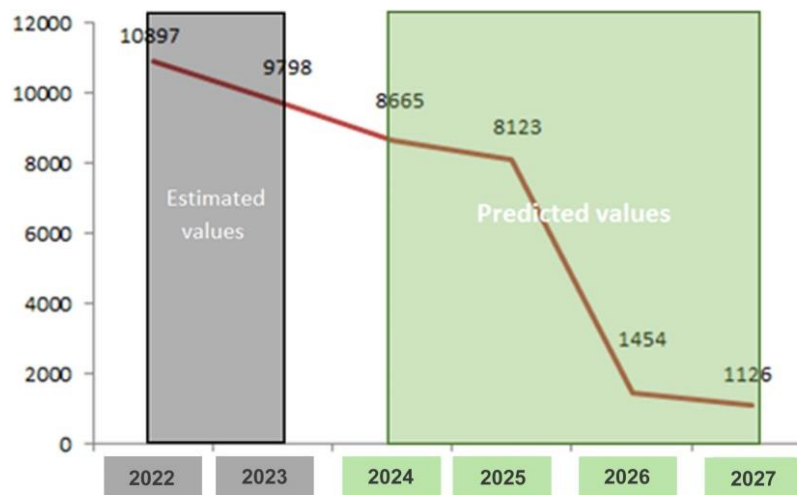
### 8. Predicting Tuberculosis Mortality using the ZIPRM

Table (6) shows the predictive values of Tuberculosis for the years 2021-2024 according to the ZIPRM.

Table 6: Estimated and Predictive Values of Tuberculosis according to the ZIPRM

| 2022  | 2023 | 2024 | 2025 | 2026 | 2027 |
|-------|------|------|------|------|------|
| 10897 | 9798 | 8665 | 8123 | 1454 | 1126 |

Table (6) showed that we are predictions there is a decrease in the number of Tuberculosis in 2022 compared to 2023, as well as a decrease in deaths of Tuberculosis in 2023, to reach 1126 Tuberculosis deaths in 2024.



**Figure 6:** Estimated and Predictive Values of Tuberculosis according to the ZIPRM

## 9. Results Discussion

The ZIPRM is better in representing frequency data for Tuberculosis. The factors that have a significant on Tuberculosis (age - blood sugar ratio - number of working hours - asthma - occupation - exposure to chemical substances) with sign compatible with medical reality, this indicates the inaccuracy of the ZIPRM to estimate the studied phenomena which contain zero inflated. According to the Zero Inflated Poisson Regression. The PRM showed the significance of some variables, despite their insignificance as a reason for the Tuberculosis. There is increase in the number of Tuberculosis in 2027 compared to 2023. These results indicate the development of the health status of the country.

## References

- [1] Yang S, Berdine G. Poisson regression." *The Southwest Respiratory and Critical Care Chronicles*". 2015;3(9):61-4.
- [2] Kamalja KK; Wagh YS, " Zero-inflated models and estimation in zero-inflated Poisson distribution". *Communications in Statistics-Simulation and Computation*. 2018;47(8):2248-65.
- [3] Muche, Fenta , Setegn; Fenta, Mekonnen ,Haile (2020), "Risk Factors of child mortality in Ethiopia: Application of multilevel two-part model", *PLOS ONE*, e0237640. <https://doi.org/10.1371/journal.pone.0237640>
- [4] L. H. Hashim, K. H. Hashim and Mushtak A. K. Shiker , (2021), " An Application Comparison of Two Poisson Models on ZeroCount Data ", *Iraqi Academics Syndicate International Conference for Pure and Applied Sciences, Journal of Physics: Conference Series 1818 (2021) 012165*, IOP Publishing, doi:10.1088/1742-6596/1818/1/012165.
- [5] Mutia Sari, Open Darnius, (2023), " Zero-Inflated Poisson Regression Testing In Handling Overdispersion On Poisson Regression ", *Journal of Mathematics Education and Application (JMEA)* Vol. 2, No 2, Juni 2023, pp. 96-107 DOI: <https://doi.org/10.30596/jmea.v2i2.13591>.
- [6] Cindy Cahyaning Astuti , Agnes Ona Bliti Puka , Akbar Wiguna , (2023), "ZERO-INFLATED NEGATIVE BINOMIAL MODELING IN INFANT DEATH CASE DUE TO PNEUMONIA IN EAST JAVA PROVINCE ", *Journal of Mathematics and Its Applications*, Volume 17 Issue 4 Page 1835–1844
- [7] Kim-Hung Pho & Buu-Chau Truong,(2024), " The Zero-Inflated Poisson - Probit regression model: a new model for count data",*Communications in Statistics - Simulation and Computation* <https://doi.org/10.1080/03610918.2024.2311797>.

- [8] Muhammad Zeeshan a, Aamna Khan a, Muhammad Amanullah a, M.E. Bakr b, Arwa M. Alshangiti b, Oluwafemi Samson Balogun c, M. Yusuf, (2024), " A new modified biased estimator for Zero inflated Poisson regression model ", HELEION, Volume 10, Issue 3.
- [9] Sadie Beckett, Joshua Jee, Thapelo Ncube, Sophia Pompilus, Quintel Washington, Anshuman Singh, Nabendu Pal, (2014), " Zero-inflated Poisson (ZIP) distribution: parameter estimation and applications to model data from natural calamities ", *Involve* 7(6): 751-767 (2014). DOI: 10.2140/involve.2014.7.751
- [10] 12Sangung Park and Sunghae Jun,(2023), " Zero-Inflated Patent Data Analysis Using Compound Poisson Models ", *Appl. Sci.* 2023, 13, 4505. <https://doi.org/10.3390/app13074505>.
- [11] Chi Zhang, G. Tian, Tao Li, (2024), "Multivariate zero-and-one inflated Poisson model with applications", *Journal of Computational and... Mathematics*. DOI:10.1016/J.CAM.2019.112356 Corpus ID: 199679828
- [12] David Giles, (2010), "Notes on the Zero-Inflated Poisson Regression Model ", Department of Economics, University of Victoria.
- [13] Wan Muhamad Amir W. Ahmad , Nursyabiha Zafakali , Nor Azlida Aleng , Mohd Shafiq Ibrahim , Ruhaya Hasan5 and Kasypi Mokhtar, (2019), "MODIFIED ZERO INFLATED POISSON REGRESSION ANALYSIS AND ITS APPLICATION TO PUBLIC HEALTH DATA", *Advances and Applications in Statistics*, Volume 55, Number 1, Pages 1-18.
- [14] Zahra Madadi , Farhad Pishgar 1, Erfan Ghasemi , Alireza Khajavi , Sahar Moghaddam, Farshad Farzadfar , (2021), " Human resources for health density and its associations with child and maternal mortality in the Islamic Republic of Iran ", *Eastern Mediterranean Health Journal*, 27(1), 16-22.
- [15] José Nildo de Barros Silva Júnior<sup>1</sup> , Rodrigo de Macedo Couto<sup>1</sup> , Layana Costa Alves<sup>1</sup> , Daiane Alves da Silva<sup>1</sup> , Isabela de Lucena Heráclio<sup>1</sup> , Daniele Maria Pelissari<sup>1</sup> , Kleydson Bonfim Andrade<sup>2</sup> , and Patrícia Bartholomay Oliveira<sup>1</sup> , (2021)," Trends in tuberculosis incidence and mortality coefficients in Brazil, 2011–2019: analysis by inflection points ", *Rev Panam Salud Publica* 47, 2023 | [www.paho.org/journal](http://www.paho.org/journal) | <https://doi.org/10.26633/RPSP.2023.152>.