https://doi.org/10.31272/jae.i150.1445
https://admics.uomustansiriyah.edu.iq
P-ISSN: 1813–6729 E-ISSN: 2707-1359

JAE
OPEN ACCESS

Comparing the Maximum Likelihood and the Bayesian Method for Estimating the Poisson Regression Model with Application to Lung Cancer Data in Erbil_Iraq

Hardi Z. Abdulrahman
Dept. of Statistics & Informatics, College of Administration and Economics, Salahaddin University, Erbil, Iraq.
Email: hardi.abdulrahman@su.edu.krd, ORCID: https://orcid.org/0009-0002-3750-7561
Kurdistan I. Mawlood
Dept. of Statistics & Informatics, College of Administration and Economics, Salahaddin University, Erbil, Iraq.
Email: kurdistan.mawlood@su.edu.krd, ORCID: https://orcid.org/0000-0002-1612-1996

Article Information

Article History:
Received: 13 / 06 / 2025
Revised: 01 / 11 / 2025
Accepted: 04 / 11 / 2025
Available Online: 01 / 12 / 2025
Pages no: 14 – 27

Keywords:
Poisson Regression, Bayesian Method, Maximum Likelihood, Markov Chain Monte Carlo, Lung Cancer.

Correspondence:
Researcher name:
Kurdistan Ibrahim Mawlood

Email:
kurdistan.mawlood@su.edu.krd

Abstract

The primary objective of this study was to utilise the Bayesian method and maximum likelihood estimation in Poisson regression to model the incidence of lung cancer in Erbil City, Iraq. Poisson regression is commonly used to analyse count data, which makes it suitable for analysing disease rates in medical data. The study compares the performance of the maximum likelihood method with the Bayesian method, which incorporates a prior distribution in the count for parameter estimation. The data set used in this study, which includes cases of lung cancer with potential risk factors, was obtained from Rizgari Hospital in Erbil City.

The Bayesian estimation employs the Markov Chain Monte Carlo technique to generate a posterior distribution. Both processes are evaluated for effectiveness, and goodness of fit is assessed to determine model performance.

The results indicated that both methods effectively identified significant factors of lung cancer and have approximately reached the same factors that have an impact on lung cancer data in Erbil city (according to the MLE method, the most critical factors influencing lung cancer disease are (Grade, Laterality, Surgery, Chemotherapy, Hormone, Isotope, and Family History). For the Bayesian methods, the factors influencing lung cancer in our data set are (Grade, Laterality, Surgery, Chemotherapy, Radio, Hormone, Isotope, and Family History). Bayesian methods give better performance and provide more interpretable quantifications. The results were obtained by using the statistical packages (SPSS v.26, Stata, and R).

1. Introduction

The idea of generalised linear models was formulated by John Nelder and Robert Wedderburn in 1972. Several exponential distributions, such as gamma, Poisson, and binomial, can be used for the response variable, enabling generalised linear models to handle a greater variety of data types. Poisson regression is essential to model count data. It was the first intended to be explicitly used to model counts and remains the basis for the many types of count models available to the analyst [1]. For Poisson regression, the log link is especially appealing since it guarantees that every value predicted for the response variable will be nonnegative. Among those closely related to the Poisson model, there exist models for analysing counted data given as proportions or ratios of counts [2].

Worldwide, Cancer is among the top causes of mortality, which is the uncontrolled growth of cells. Nearly one in five deaths is from cancer, which comes in second. Over the last century, cancer treatment has involved surgery, chemotherapy, and radiation therapy, all of which have proven efficacious. Individually or in combination, these treatment modalities can have a significant impact on tumour growth and may even lead to cures.

The primary goals of this study are to examine the significance of the statistical model known as Poisson Regression in analyzing lung cancer data in Erbil City, identify the diagnostic factors that have the most significant impact on lung cancer patients' survival times, and compare the outcomes of the two methods used to estimate Poisson Regression model—MLE and Bayesian—to identify the model that best fits our data.

2. Literature Reviews

Several studies are presented in this section. Reviewing previous studies makes it easier for the researcher to carry out the study's goals, fills in any gaps left by earlier research, and gives them a broader understanding of the subject.

In 2019, Noaman & Al-Ameer treated the estimation of the shape parameter of the Burr distribution. Estimation methods included the Maximum Likelihood Estimator (MLE), the Percentile Estimator (PER), as well as the Bayesian Estimator employing Jeffreys prior distribution. They used the statistical measure Mean Squared Error (MSE) to compare, considering variation in sample size ($n = 15, 30, 50, 100$) and different sets of initial values for (θ, λ). In their conclusions, they mentioned the Bayesian estimator as the best among all [3].

In 2019, Algama & Abdalteef reviewed and contrasted approaches to variable selection in the Poisson regression model, using both simulations and actual data, which was the goal of their study. To create data in accordance with the Poisson regression model, they employed simulations and the Monte Carlo method [4].

In 2024, Al-Hasani (R.F.M.) examined standard techniques used to obtain estimates of the parameters of the Poisson regression model in the presence of semi-multicollinearity, such as ridge regression and Liu's estimator method. His findings indicated that the Liu estimators' approach outperformed the ridge regression procedure when using (AIC) as a comparison criterion [5].

3. Background Information

The first health issue presented in this section concerns lung cancer; moreover, a description and the basic principles of the Poisson regression model are provided. The section presents two estimation methods (the Bayesian approach and the maximum likelihood estimator) used for estimating the model coefficients. Akaike's Information Criterion and Bayesian Information Criterion were also utilised to choose the most appropriate model between the two estimation methods.

3.1. Lung Cancer

Lung cancer is the leading cause of cancer-related death worldwide. It is one of the most prevalent and rapidly spreading cancers. Lung cancer causes a wide range of symptoms and indicators depending on where it occurs in the body because it can develop in different parts of the bronchial tree. Lung cancers are histologically classified into two groups: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). These classes differ in how they develop and spread. Among the crucial actions in the diagnosis process is cancer staging, which has several goals, including assisting the physician in suggesting a course of treatment and aiding in the evaluation of treatment results by providing a prognostic indication. There are one to four stages of NSCLC. The less the cancer has spread, the lower the stage. SCLC is divided into two stages: limited (confined to the mediastinum, the hemithorax of origin, or the supraclavicular lymph nodes) and extensive (spread beyond the supraclavicular areas). The stage word refers to the pretreatment, clinical stage.

We have two lungs; each side of your chest has one of them. Your lungs make up most of your respiratory system – the organs and tissues that allow you to breathe. Besides eliminating carbon dioxide and other gases from the body, the lungs "bring oxygen to the body". The process occurs twelve to twenty times per minute.

3.2. Poisson distribution

The Poisson distribution is a discrete probability distribution used in probability theory and statistics that represents the probability of a series of events occurring within a given time interval, assuming these events occur at a known average rate and are independent of the time elapsed since the last event.

According to the Poisson model, the probability of (Y) occurrences in a given interval for a random variable (number of occurrences in a given interval or space) with an average rate of occurrence of λ is as follows:

$$P(Y=y)=\frac{\lambda^{y}e^{-\lambda}}{y!} \qquad Y=0,1,2,\ldots \tag{1}$$

We notice that the Poisson distribution has only one parameter, λ (lambda), which is the mean number of events. The Poisson distribution is equidispersed since the variance equals the mean, which is a feature that distinguishes it from other discrete distributions.

3.3. Poisson Regression Model

The simplest count regression model and the foundation for count data analysis is the Poisson regression. Since counts must be at least zero, coefficients are exponentiated. Poisson regression typically fits count data distributions better than linear regression, which assumes a normal distribution, because count data distributions frequently have a Poisson distribution. Nonetheless, Poisson regression has its own significant limitations. According to this limitation, the variance of the count variable must be approximately equal to its mean. The term "overdispersion" refers to breaking this presumption. The Poisson distribution's PDF is [6]:

$$P(Y_i=y_i)=\frac{e^{-\lambda_i}\ \lambda_i^{y_i}}{y_i!} \qquad Yi = 0, 1, 2, \ldots \tag{2}$$

The Poisson regression model can be written as:

$$E(y_i)=\lambda_i \tag{3}$$

$$g(\lambda_i)=X_i'\beta \tag{4}$$

Where X_i is a vector of explanatory variables and β is a vector of parameters to be estimated.

$$\lambda_i = g^{-1}(X_i'\beta) \tag{5}$$

Substituting the log link function gives:

$$ln\lambda_i = X_i'\beta \tag{6}$$

$$\lambda_i = e^{X_i'\beta} \tag{7}$$

The Poisson family of regression models provides improved and easy-to-implement analyses of count data [7]. A count variable is a variable that has discrete values (0, 1, 2, ...), representing the number of times an event occurs over a predetermined amount of time or space. Since a negative number of times an event occurs is impossible, a count variable can only have positive integer values or zero.

3.4. Poisson Regression Model Assumptions

The Poisson probability distribution function (PDF) is a standard option for count data. The remaining count models are adjustments or variations from the basic Poisson model.[1]

The first step in modelling count data is to look at the data for significant deviations from the assumptions underlying the fundamental Poisson model.

And it is crucial to test each of these assumptions:

1. In a Poisson model, the response variable (Y) must be a count and a positive integer. A different model must be used to model a continuous variable.
2. The response variable must be distributed according to the Poisson distribution.

3. This assumption must be fulfilled for a Poisson distribution, where the response variable's mean and variance are equal or nearly equal (Equidispersion). Such an assumption limits the applicability of the Poisson regression model. When the variance value exceeds the mean value, overdispersion occurs. In cases where the dependent variable's variance exceeds its mean, a phenomenon referred to as overdispersion [8] is observed.
4. There should be no correlation between the observations. Necessary presumption suggests that every observation is unrelated to the others, meaning that none of them may influence the others in any manner.

3.5. Maximum Likelihood Estimation Method (MLE)

The predominant method of finding parameter estimates of a probability distribution is known as Maximum Likelihood Estimation (MLE), which has its foundation concepts established in the 1920s by R.A. Fisher. The invariance property is considered to be an essential property of MLE. A few desirable properties of maximum likelihood estimators are consistency, sufficiency, asymptotic efficiency, and asymptotic normality, among others [9].

In short, maximum likelihood estimation is the process of finding “the probability distribution that would make the observed data most probable” [10].

The MLE in Poisson regression is found by starting with the following form of the likelihood of y given β, assuming that (y_i, ..., y_n) are independent:

The Poisson regression model is defined by:

$$Pr(Y_i = y_i|x_i) = \frac{\exp(-\lambda_i)\ \lambda_i^{y_i}}{y_i!}\ , \qquad Y, i = 0, 1, 2, \ldots \tag{8}$$

Where:

$$\lambda_i = \exp(x_i'\beta) = \exp(\beta_o + \beta_1 x_{1i} + \cdots) \tag{9}$$

The likelihood function for (β) is:

$$L(\underline{\beta}) = \prod_{i=1}^{n}(\frac{\exp(-\lambda_i)\ \lambda_i^{y_i}}{y_i!})$$

$$L(\underline{\beta}) = (\frac{\exp(-\sum_{i=1}^{n}\lambda_i)\ \prod_{i=1}^{n}\lambda_i^{y_i}}{\prod_{i=1}^{n} y_i!}$$

where β is the vector of the parameters of interest and it can be estimated maximizing the log-likelihood function as follows:

$$\ln L\left(\underline{\beta}\right) = -\sum_{i=1}^{n}\lambda_i + y_i\sum_{i=1}^{n}\ln(\lambda_i) - \sum_{i=1}^{n}\ln(y_i!)$$

By depending on Eq. (9) where [$\lambda_i = \exp(x_i'\beta)$]

$$\ln L\left(\underline{\beta}\right) = \sum_{i=1}^{n}[-\exp(x_i'\beta) + (x_i'\beta)y_i - \ln(y_i!)]$$

In addition to being the most popular method for count data, Poisson regression is also being used more and more frequently in estimating multiplicative models of alternative data [11].
It is easy to understand why this estimator is so popular by looking at the score vector and Hessian matrix, which can be expressed as:

$$\frac{\partial lnL}{\partial \underline{\beta}} = s(\beta) = \sum_{i=1}^{n}[y_i - \exp(x_i'\beta)]\, x_i = \frac{\partial l(\beta, y, x)}{\partial \beta} \tag{10}$$

$$H_n(\underline{\beta}, \underline{y}, \underline{x}) = \frac{\partial^2 l(\underline{\beta}, \underline{y}, \underline{x})}{\partial \beta\ \partial \beta'} = -\sum_{i=1}^{n} e^{X_i^T \beta}(X_i^T X_i) \tag{11}$$

3.6. Bayesian Estimation Method

The goal of statistical inference is to obtain estimates of the unknown parameters, λ, conditional on the data, y. The unknown parameters, λ, is primary distinction between Bayesian statistical inference, frequentist statistical inference, and ML estimation. ML estimation assumes that λ is fixed and unknown, while Bayesian estimation assumes that λ is random and has a probability distribution that encapsulates our lack of knowledge regarding the true value λ (i.e., distribution posterior). Formally speaking

$$P(\lambda|y) = \frac{P(\lambda, y)}{P(y)} = \frac{P(y|\lambda)\ P(\lambda)}{P(y)}$$

Where p(y) represents the likelihood of observing y, p(y|λ) represents the likelihood of observing y given unknown parameters. With respect to the observed data y, λ, p(λ|y) represents the posterior distribution of the parameters λ, p(λ, y) represents the joint probability of λ and y, and p(λ) represents the prior distribution of the parameters collectively. Here is the Bayes Theorem equation, this theorem Originally developed by Thomas Bayes, Bayesian methods were further extended and detailed by Pierre Simon Laplace [12].

The prior distributions specified for the model parameters is one characteristic that sets Bayesian estimation apart. Depending on the amount and accuracy of the information we believe we have prior to data collection, priors can either be non-informative or informative. A diffuse prior, another name for a non-informative prior, has a high variance that indicates that the parameter value is highly uncertain. A high prior variance means that the likelihood adds a relatively greater amount of information to the formation of the posterior, and estimation approaches a maximum likelihood estimate [13].

The Bayesian method uses the Markov Chain Monte Carlo (MCMC) sampling method to obtain the parameter estimates and inferences. The Monte Carlo integration using Markov Chain is known as the MCMC. To estimate expectations, the Monte Carlo integration takes samples from the target distribution and then creates sample averages.

Posterior ∝ Likelihood × Prior

We place prior ($\beta_i \sim N(\mu, \sigma^2)$), where $\beta = \left(\beta_o, \beta_1, \dots, \beta_p\right)$ represent the regression coefficients.

$$P(\beta/Y_i) \propto \left(\frac{\exp(-\sum_{i=1}^{n} \lambda_i)\ \prod_{i=1}^{n} \lambda_i^{\,y_i}}{\prod_{i=1}^{n} y_i!}\right) \times \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{j=0}^{p} e^{-\frac{(\beta_j - \mu_j)^2}{2\sigma_j^2}}$$

After omitting constants, we get:

$$P(\beta/Y_i) \propto e^{-\sum_{i=1}^{n} \lambda_i} \quad \prod_{i=1}^{n} \lambda_i y_i \ \times \ \prod_{j=0}^{p} e^{-\frac{(\beta_j - \mu_j)^2}{2\sigma_j^2}}$$

There is no closed-form solution for $P(\beta/Y_i)$, we use Markov Chain Monte Carlo (MCMC) methods to sample from this posterior distribution.

3.7. Deviance Statistics

Deviance statistics is one of the techniques used to measure goodness of fit. Another name for this statistical measure is the "G square statistic."

Deviance is the measure used to quantify how well a generalised linear model fits data. It determines the difference between the data and the values that were fitted. In standard multiple regression, the deviation represents a generalisation of the sum of squares. The following formula gives deviance:

$$D_p = 2\sum_{i=1}^{n} \{y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - \left(y_i - \hat{\lambda}_i\right)\} \tag{12}$$

Where: y_i is the response seen in the ith observation. $\hat{\lambda}_i$ is the fitted value for the ith observation, also known as the predicted mean.

Deviance can be used to compare two models and evaluate how well a GLM fits. Formulating a null model and a complete model is the first step to understanding deviance. The null model consists of a single parameter. $\lambda_i = \lambda$ to denote the predicted value of all outcomes, whereas the saturated (full) model accurately reflects the observed values, $y - \hat{\lambda}$. By comparing it to the full model, the deviance assesses how well a model fits the data. Upon reaching zero, this statistical value indicates that the model fit has increased, or a lower deviance suggests a better fit.

3.8. Criteria of the Model Selection

To compare and select the best estimation method, the two model selection criteria considered in our study are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The model with the lowest value is chosen as the best model.

3.8.1. Akaike's Information Criterion (AIC) For Selecting the Best Model

To determine which model is best, models are compared using Akaike's information criterion [14]. In comparison to the other models, the selected model has the lowest AIC value. As a result, the AIC is described as:

$$\mathrm{AIC} = -2\mathrm{logL} + 2\mathrm{K} \tag{13}$$

where L: is the log-likelihood and k: is the number of parameters in the model.

3.8.2. The Bayesian Information Criterion

When selecting statistical models, the Bayesian information criterion (BIC) is one of the most widely used and popular tools. Based on the Bayesian probability application's derivation of BIC, the likelihood that BIC will select the actual model increases with dataset size; the model with the smallest BIC is regarded as the best model. Calculating the BIC for models defined by [15]:

$$BIC = -2 * \log L + 2 * \log(N) * k \tag{14}$$

where N is the number of observations and k is the estimated number of parameters.

4. Results and Discussions:

In this section, we review the basics of applying the Poisson regression model to study the key variables affecting the prognosis and survival of lung cancer patients in Erbil. The study employs Bayesian and MLE approaches to identify the factors influencing patient survival through an applied Poisson regression analysis, aiming to determine the most suitable model for analysing lung cancer patient data. Two statistical measures (BIC and AIC) were used to evaluate the best model for this data using three statistical analysis programs: SPSS, Stata R V. 4.3.118, R V. 4.3.1.

4.1. Data Collection

Data were gathered for this study from 454 patients with lung cancer at Rizgary Oncology Centre, affiliated with Rizgary Teaching Hospital in Erbil, Iraq's Kurdistan Region. All lung cancer patients between the ages of 22 and 91 years, of both sexes (176 men and 278 women), had data gathered from January 1, 2019, to August 24, 2020. The survival time, which represents the dependent variable, is measured in months from the first day that the patient is admitted to the hospital to the date of death or the last visit to the hospital, and the data contains (13) variables; their categorisation is shown in Table 1 below.

Table (1) variable categorization

Variable	Name	Categorization	Number	Rate
X_1 = Gender	Gender	Male = (1)	176	0.39
		Female = (2)	278	0.61
X_2 = Grade	grade characterizes the appearance of normal or aberrant cancer cells under a microscope.	Grade one = (1)	31	0.07
		Grade two = (2)	252	0.56
		Grade three = (3)	161	0.35
		Grade four = (4)	10	0.02
X_3 = Laterality	Tumor Laterality	Right = (1)	221	0.49

		Left = (2)	226	0.50
		Bilateral = (3)	6	0.01
		Not applicable = (4)	1	0.00
X_4 = Surgery	Has the patient undergone surgery? Surgical operations to remove cancerous tumors	Yes = (1)	67	0.15
		No = (2)	387	0.85
X_5 = Chemo	Did the patient receive chemotherapy	Yes = (1)	71	0.16
		No = (2)	383	0.84
X_6 = Radio	Did the patient receive radiotherapy?	Yes = (1)	51	0.11
		No = (2)	403	0.89
X_7 = Hormone	Did the patient receive hormone therapy?	Yes = (1)	190	0.42
		No = (2)	264	0.58
X_8 = Isotope	Did the patient receive isotope therapy?	Yes = (1)	444	0.98
		No = (2)	10	0.02
X_9 = Targeted	Did the patient receive targeted therapy?	Yes = (1)	346	0.76
		No = (2)	108	0.24
X_{10} = Family History	Did the patient have a family history of cancer?	Yes = (1)	306	0.67
		No = (2)	148	0.33
X_{11} = Country	The patient's Country	Iraq = (1)	433	0.95
		Arabic =(2)	20	0.04
		Foreign =(3)	1	0.00
X_{12} = Occupation	The patient's Occupation	Jobless =(1)	346	0.76
		Worker =(2)	0	0.00
		Farmer =(3)	0	0.00
		Employee =(4)	70	0.15
		Craftsman =(5)	0	0.00
		Child =(6)	0	0.00
		Retired =(7)	30	0.07
		Another =(8)	8	0.02
X_{13} = Age	The patient's age at diagnosis	(<= 26) =(1)	2	0.004
		(27 - 31) =(2)	17	0.04
		(32 - 36) =(3)	29	0.06
		(37 - 41) =(4)	56	0.12
		(42 -46) =(5)	69	0.15
		(47 - 51) =(6)	72	0.16
		(52 - 56) =(7)	56	0.12
		(57 - 61) =(8)	54	0.12
		(62 - 66) =(9)	37	0.08
		(67 - 71) =(10)	32	0.07
		(72 - 76) =(11)	17	0.04
		(77 - 81) =(12)	6	0.01
		(82 - 86) =(13)	5	0.01
		(87 - 91) =(14)	2	0.004
Y = Time	Period the patient diagnosis (by months)	4 Mo. =(4)	85	0.19
		5 Mo. =(5)	62	0.14
		6 Mo. =(6)	31	0.07
		7 Mo. =(7)	21	0.05
		8 Mo. =(8)	33	0.07
		9 Mo. =(9)	69	0.15
		10 Mo. =(10)	27	0.06
		11 Mo. =(11)	51	0.11
		12 Mo. =(12)	75	0.17

Table 1 shows the 454 patients with lung cancer. Most of them (252) patients had grade II (56%); about the tumor laterality, only six patients (1%) were recorded to have bilateral tumors (or on both sides); most of the patients (85%) had surgery; (383) patients (84%) received chemotherapy; (89%) of the patients received radiotherapy; and (42%) received hormone therapy. In a total of (454) lung cancer cases in our data, (444) patients (98%) received isotope therapy, and (346) cases, or (76%), of them received targeted therapy. In addition, our results indicated that 306 patients (67%) have a family history of cancer.

4.2. Poisson Regression Analysis

Our study employed Poisson regression as a statistical tool because it is well-suited for count data. In our study, the dependent variable (Y = number of months) is count, and the most widely used regression technique for count data is Poisson regression. To determine how a response variable (event) and several other variables relate to one another, as well as the extent to which these factors impact the lung cancer patient's chances of survival, we constructed a Poisson regression model.

Thirteen explanatory factors were identified in our current study as having an impact on the response variable, which represents patient survival (event): gender, Grade, Laterality, Surgery, Chemotherapy, Radiation, Hormone Therapy, Isotope Therapy, Targeted Therapy, Family History, Country, Occupation, and Age (Binned).

If there is a substantial difference between the count mean and variance (which is equivalent in a Poisson distribution), then we should use another type of regression.

Table (2) descriptive statistics of the dependent Variable:

Variable	Measurement	Count	Min	Max	Mean	Variance
Y = Time	No. of month	454	4	12	7.93	8.60

After verifying the assumptions and checking for approximate equality of mean and variance, we can apply the Poisson regression to our data.

4.2.1. Testing the overdispersion of the data

The Pearson chi-square and deviance values, divided by the number of degrees of freedom, can be used to identify overdispersion. Overdispersion is said to occur if both values are greater than 1. The results of the overdispersion test are shown in Table 3 below.

Table (3) The deviance and Pearson chi-square test

Goodness of Fit			
	Value	df	Value/df
Deviance	442.305	440	1.005
Pearson Chi-Square	432.961	440	0.984

The deviance value and Pearson Chi-square value are (1.005) and (0.984), respectively, according to Table 3, which are approximately 1. This indicates that the Poisson regression model satisfies the equi-dispersion assumption, making it suitable for modelling survival time (in months) following patient diagnosis.

4.2.2. The Omnibus Test of Poisson Regression

To construct the Poisson regression model, the initial step was to ascertain the value of the omnibus test as a simultaneous test. To find out if all predictor variables collectively improve the model compared to the intercept model alone, the omnibus test is a likelihood ratio test. Table 4 below shows the results of the omnibus test.

Table (4) Omnibus Tests of Model Coefficients

Omnibus Test		
Likelihood Ratio Chi-Square	df	Sig.
67.939	13	0.000
Dependent Variable: Time		
Compares the fitted model against the intercept-only model.		

According to Table (4), H_0 is accepted if the model value (sig) > α or if the $G^2 < \chi^2_{table}$. The omnibus test rejects H_0 because the model value (sig) is 0.00 < α (0.05) or the $G^2(67.939) > \chi^2_{tabel}$ (22.362). Thus, it can be said that there is a significant relationship between time and one or more predictor variables.

4.2.3. Estimating the Poisson regression model using the MLE method

The most popular technique for estimating count data models, the maximum likelihood method, was first applied to estimate parameters. To maximise the probability that the specified model produced the observed sample, the parameter should be chosen in accordance with the maximum likelihood principle. Table 5 summarises the Poisson regression model for patients with lung cancer, estimated using the maximum likelihood method at a significance level of $\alpha = 0.05$. The impact of the model's predictors can be better understood by examining the values and signs of the coefficients for the covariates. Covariates with positive coefficients indicate that the dependent variable and the predictors have a positive relationship, and similarly, those with negative coefficients show a negative relationship.

Table (5) Parameter Estimates by the maximum likelihood method

Parameter Estimates								
Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)
			Lower	Upper	Wald Chi-Square	df	Sig.	
(Intercept)	2.752	0.3291	2.107	3.397	69.962	1	0.000	15.679
Gender	0.029	0.0352	-0.040	0.098	0.662	1	0.416	1.029
Grade	-0.056	0.0266	-0.108	-0.004	4.469	1	0.035	0.945
Laterality	-0.076	0.0319	-0.138	-0.013	5.615	1	0.018	0.927
Surgery	-0.098	0.0416	-0.180	-0.016	5.548	1	0.018	0.907
Chemo	-0.085	0.0362	-0.156	-0.014	5.450	1	0.020	0.919
Radio	0.047	0.0358	-0.023	0.117	1.728	1	0.189	1.048
Hormone	-0.078	0.0283	-0.134	-0.023	7.670	1	0.006	0.925
Isotope	0.296	0.0966	0.107	0.486	9.425	1	0.002	1.345
Targeted	-0.025	0.0289	-0.082	0.031	0.777	1	0.378	0.975
Family History	-0.129	0.0371	-0.201	-0.056	12.074	1	0.001	0.879
Country	-0.073	0.0783	-0.227	0.080	0.870	1	0.351	0.930
Occupation	-0.002	0.0092	-0.020	0.016	0.036	1	0.849	0.998
Age (Binned)	0.003	0.0067	-0.010	0.016	0.159	1	0.691	1.003

The Poisson regression model's parameters, estimates, standard errors, Wald Chi-Square, and P-values are displayed in Table 5. To explain the relationship between patient survival time and 13 independent variables, the output displays the findings of the Poisson regression model's fit. The results show that, at the 95.0% confidence level, the P-value is less than 0.05, indicating a statistically significant relationship between the variables. Additionally, because the P-value for the residuals is greater than or equal to 0.05, the model is not essential. The following interpretations can be made using estimates of the parameters.

We tested the model coefficients using the Wald chi-square test. According to the results in Table 5, we found that the estimated coefficients of Grade, Laterality, Surgery, Chemo, Hormone, Isotope, Family History, and the intercepts' p-values are below 0.05; therefore, they are significant. Most of the independent variables contribute negatively, but the radiotherapy, isotope, and age (binned) variables contribute positively.

The coefficients in Poisson regression represent changes in the log of the expected count, or, in our research, time (period of the patient's diagnosis with lung cancer).

Using parameter estimates, we can interpret the following:

- Our data shows that the coefficients of the following variables in the Poisson regression model (Grade, Laterality, Surgery, Chemotherapy, Hormone, Isotope, and Family History) are significant and impact the length of time that patients with lung cancer survive.
- Based on Poisson regression's estimated coefficients, age (binned), radiation, and isotope are thought to have an impact that lowers the likelihood that the patient will survive.
- For (Gender, Radio, Targeted, Country, Occupation, and Age (Binned)), the corresponding P-values are greater than 0.05 and are (0.416, 0.189, 0.378, 0.351, 0.849, and 0.691), respectively. Therefore, for those variables whose Poisson regression coefficients are not significant.
- Intercept: The Poisson regression estimate is (2.752) units, which is the response variable's log of the expected count when every other variable in the model is assessed at zero.

The Poisson regression model by maximum likelihood estimation method with all factors as follows:

$$\ln(y) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

$\ln(y)$ = (2.752 + 0. 029 Gender – 0. 056 Grade – 0. 076 Laterality – 0. 098 Surgery – 0. 085 Chemotherapy + 0. 047 Radiotherapy – 0. 078 Hormone + 0. 296 Isotope – 0. 025 Targeted – 0. 129 Family History – 0. 073 Country –0.002 Occupation + 0.003 Age-binned)

Additionally, with just the important variables, we can create the Poisson regression model:

$\ln(y)$ = (2.752 – 0. 056 Grade – 0. 076 Laterality – 0. 098 Surgery – 0. 085 Chemotherapy – 0. 078 Hormone + 0. 296 Isotope – 0. 129 Family History)

4.2.4. Estimating the Poisson regression model using the Bayesian method

The second method to estimate parameters in our research is the Bayesian method. Since the mid-2000s, Bayesian estimation has gained popularity. The great majority of statisticians followed what is known as the frequentist interpretation of statistics, though there were surely Bayesians before this. According to Bayesian theory, each parameter in a model has its own distribution and is a random variable in and of itself. The model's predictor coefficients can differ entirely from one another and are all random variables.

One of the key characteristics of Bayesian statistics is the possibility that data information may influence parameter values. This is referred to as prior information, and its mathematical representation is an initial distribution. Every iteration in the overall estimation process updates the posterior distribution for each predictor based on the corresponding prior distribution. To obtain an updated posterior distribution, the likelihood is multiplied by the prior using Markov Chain Monte Carlo techniques for posterior sampling, and an N(0,10) prior is placed on the coefficients.

Table (6) Parameter Estimates by the Bayesian method

Time	Mean	Std. dev.	MCSE	Median	Equal-tailed 95% cred. Interval	
Constant	2.663	0.086	0.022	2.664	2.482	2.820
Gender	0.031	0.032	0.002	0.031	-0.032	0.094
Grade	-0.063	0.025	0.004	-0.063	-0.115	-0.012
Laterality	-0.083	0.029	0.005	-0.087	-0.131	-0.020
Surgery	-0.087	0.032	0.006	-0.088	-0.150	-0.027
Chemo	-0.098	0.029	0.005	-0.100	-0.154	-0.041
Radio	0.059	0.030	0.006	0.058	0.003	0.119
Hormone	-0.082	0.027	0.003	-0.081	-0.135	-0.030
Isotope	0.286	0.044	0.012	0.292	0.188	0.360
Targeted	-0.024	0.029	0.004	-0.025	-0.083	0.031
Family History	-0.134	0.036	0.005	-0.135	-0.203	-0.058
Country	-0.094	0.072	0.017	-0.094	-0.235	0.054
Occupation	-0.002	0.009	0.001	-0.002	-0.020	0.017
Age Binned	0.002	0.006	0.001	0.002	-0.010	0.014

In the above table (6), each row represents a regression coefficient for one of our predictors. Table (6)'s findings lead us to the conclusion that the estimated coefficients for Laterality, Surgery, Chemotherapy, Hormones, Isotopes, Family History, Country, and Constant are statistically significant, as the value 0 does not fall within their credible intervals. Again, in this model too, most independent variables have an adverse effect, while Radiotherapy, Isotopes, and Age Binned have a positive impact. We can interpret the following using parameter estimates:

- Our findings indicate that the coefficients of the following variables (Grade, Laterality, Surgery, Chemo, Radio, Hormone, Isotope, Family History, and Constant), have an impact on time and are significant in the Poisson regression model.
- Radio, Age (binned), and Isotope are thought to have an impact that decreases the patient's time risk based on estimated coefficients of Poisson regression.

- For (Gender, Targeted, Country, Occupation, and Age (Binned)), the corresponding 95% credible intervals include (0) and are, respectively, {(–0.032, 0.094), (–0.083, 0.031), (–0.235, 0.054), (–0.020, 0.017), and (–0.010, 0.0140)}. For those variables, we must accept the null hypothesis because their Poisson regression coefficients are not significant.
- Intercept: It is the logarithm of the response variable's expected count when every other variable in the model is assessed at zero.

$e^{2.663} \approx 14.339$

The Poisson regression model by Bayesian estimation method with all factors as follows:

$$\ln(y) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

$\ln(y)$ = (2.663 + 0.031 Gender − 0.063 Grade − 0.083 Laterality − 0.087 Surgery − 0.098 Chemo + 0.059 Radio − 0.082 Hormone + 0.286 Isotope − 0.024 Targeted − 0.134 Family History − 0.094 Country − 0.002 Occupation + 0.002 Age (Binned))

Furthermore, we can use just the significant variables to write the Poisson regression equation:

log(E[Time]) = (2.663 − 0.063 Grade − 0.083 Laterality − 0.087 Surgery − 0.098 Chemo + 0.059 Radio − 0.082 Hormone + 0.286 Isotope − 0.134 Family History)

Where, log(E[Time]): The natural logarithm of the expected value (or mean) of the response variable (Time).

4.3. Comparing models

After modelling by both methods, we now compare them to identify which of them is the best. For this purpose, we used the Akaike information criterion and the Bayesian information criterion. The model with the lowest values for both requirements will be the best among the approaches. The study's findings suggest that the Bayesian method is superior to maximum likelihood estimation in terms of efficiency. the results are shown in the table (7):

Table (7) Comparing Models with AIC and BIC

Method	Maximum likelihood	Bayesian
No. of parameters	14	14
- Log Likelihood	-1133.82	**-1096.69**
AIC	2295.64	**2221.38**
BIC	2304.83	**2230.57**
MSE	8.769	**7.315**

Depending on the AIC and BIC values, which are displayed in Table 7 and used to compare two models, we can know which of the two models best fits our data (Maximum likelihood or Bayesian)? According to the findings, the Bayesian model is the most appropriate model for our lung cancer study data because its (-log likelihood = 1096.69), which is smaller than the (-log likelihood) of the maximum likelihood method, which equals (1133.82), and therefore the AIC value of the Bayesian method equals (2221.38). The BIC equals (2230.57), whose values are the lowest when compared to the Maximum likelihood models AIC of (2295.64) and BIC of (2304). Furthermore, the MSE value of the Bayesian method is smaller than the MSE value of the maximum likelihood approach, as shown in Table 7's final row.

Now we can say the best model for our data is the Bayesian model, and according to the Bayesian method, the most significant factors in our study are (Laterality, Surgery, Chemo, Hormone, Isotope, Family History, Country).

In addition, the best model for our data is

$$\ln(\text{Time}) = (2.663 - 0.063\,\text{Grade} - 0.083\,\text{Laterality} - 0.087\,\text{Surgery} - 0.098\,\text{Chemo} + 0.059\,\text{Radio} - 0.082\,\text{Hormone} + 0.286\,\text{Isotope} - 0.134\,\text{Family History})$$

5. Conclusion

After examining the data on lung cancer in Erbil City and based on the findings from the practical part of the study, the following conclusions have been reached:

1. According to the Poisson model's omnibus test of model effects, the model fits the variables that were chosen; however, the statistical model is statistically significant when the p-values are less than 0.05, meaning that the variables in the model are essential and have an effect.
2. The Poisson model's over-dispersion is unable to produce sufficient results. The equidispersion assumption is satisfied by the Poisson regression model in our data, making it appropriate for estimating the survival duration of patients with lung cancer
3. Considering the two methods for our data set, the Poisson model found the same statistically significant prognostic factors that affected lung cancer, except radiotherapy, which is considerable and had an impact in the Bayesian method but was not substantial in the maximum likelihood method.
4. The MLE methods findings indicate that the most critical factors influencing lung cancer disease are (Grade, Laterality, Surgery, Chemotherapy, Hormone, Isotope, and Family History).
5. According to the Bayesian methods findings, the factors influencing lung cancer in our data set are (Grade, Laterality, Surgery, Chemotherapy, Radio, Hormone, Isotope, and Family History).
6. The Poisson regression model estimation using the Bayesian method appears to be the most suitable model after the models' performance in analysing the lung cancer data in Erbil City was evaluated after fitting the model to the data using MLE and Bayesian methods.

6. Supplementary material

(None)

7. Author's Contributions

Hardi Zrar Abdulrahman, who designed the research and conducted the analyses. And Kurdistan Ibrahim Mawlood, who writes, edits, and interprets the results.

8. Funding

(None)

9. Data availability statement

Recorded dataset from the Rizgary Oncology Centre, affiliated with Rizgary Teaching Hospital in Erbil, Iraq's Kurdistan Region.

10. Acknowledgements

We would like to thank the Rizgary Oncology Centre for providing the data.

11. Conflict of interest

We declare that there is no conflict of interest.

References

[1] Hilbe, J. M., 2014. Modeling Count Data. Arizona: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139236065
[2] McCullagh, P. a. J. A. N., 1989. Generalized linear models. second ed. New York: Chapman and Hall. DOI: http://doi.org/10.1201/9780203753736
[3] Noaman, D. I. A. & Al-Ameer, A. H. A. A., 2019. Comparison of Classical and Bayesian methods to Estimate the shape parameter and Reliability function in Burr type X or two parameter of exponential Rayleigh distribution under different Loss function. Journal of Administration and Economics, Issue 119, pp. 42-58. DOI: http://doi.org/10.31272/JAE.42.2019.119.3
[4] Algama, D. Z. Y. & Abdalteef, A. M., 2019. Variable selection in Poisson regression model using penalized likelihood methods. Journal of Administration and Economics, Issue 118, pp. 285-294. DOI: http://doi.org/10.31272/JAE.42.2019.118.1
[5] Al-Hasani, R. F. M., 2024. Comparison Between Estimators Leu Regression Method and Ridge Regression Method of the Poisson Regression Model in The Presence of Multicollinearity Problem. Journal of Administration and Economics, 49(144), pp. 36-46. DOI: https://doi.org/10.31272/jae.i144.1236

[6] Cameron, A. C. & Trivedi, P. K., 2012. Regression analysis of count data. In: Event history analysis with R. s.l.:Cambridge university press.Brostrom. DOI: https://doi.org/10.1017/CBO9780511814365
[7] Coxe, S., West, S. G. & Aiken, L. S., 2013. The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. Personality Assessment, pp. 121-136. DOI: https://doi.org/10.1080/00223890802634175
[8] Guo, J. Q. & Li, T., 2002. Poisson regression models with errors-in-variables: implication and treatment. Statistical Planning and Inference, 104(2), pp. 391-401. DOI: https://doi.org/10.1016/S0378-3758(01)00250-6
[9] Dobson, A. J. & Barnett, A., 2018. An introduction to generalized linear models. Boca Raton, Florida, USA: Chapman and Hall/CRC. DOI: http://dx.doi.org/10.1007/978-1-4899-7252-1
[10] Myung, I. J., 2003. Tutorial on maximum likelihood estimation. Mathematical Psychology, 1(47), pp. 90-100. DOI: https://doi.org/10.1016/S0022-2496(02)00028-7
[11] Silva, J. M. C. S. & Tenreyro, S., 2006. THE LOG OF GRAVITY. The Review of Economics and Statistics, 88(4), p. 641–658. DOI: http://doi.org/10.1162/rest.88.4.641
[12] Hogg, R. V., McKean, J. W. & Craig, A. T., 2013. Introduction To Mathematical Statistics. Seventh ed. Boston: Pearson Education India. DOI: https://doi.org/10.1080/10543406.2013.756334
[13] Kim, S.-Y.et al., 2013. Single and Multiple Ability Estimation in the SEM Framework: A Non-Informative Bayesian Estimation Approach. Multivariate behavioral research, 4(48), p. 563–591. DOI: https://doi.org/10.1080/00273171.2013.802647
[14] Akaike, H., 1998. Information Theory and an Extension of the Maximum Likelihood Principle. New York, Springer New York, pp. 199-213. DOI: http://dx.doi.org/10.1007/978-1-4612-1694-0_15
[15] LEE, E. T. & JOHN, W. W., 2003. Statistical methods for survival data analysis. Third ed. New Jersey: John Wiley & Sons. DOI: https://doi.org/10.1002/0471458546

https://doi.org/10.31272/jae.i150.1445
https://admics.uomustansiriyah.edu.iq
P-ISSN: 1813–6729 E-ISSN: 2707-1359

JAE
OPEN ACCESS

مقارنة بين طريقتي الاحتمال الأعضم و بيز لتقدير نموذج الانحدار بواسون مع التطبيق على بيانات سرطان الرئة في أربيل_العراق

هردى زرار عبدالرحمن
قسم الإحصاء والمعلوماتية، كلية الإدارة والاقتصاد، جامعة صلاح الدين، اربيل،العراق.
Email: hardi.abdulrahman@su.edu.krd, ORCID: https://orcid.org/0009-0002-3750-7561

كوردستان ابراهيم مولود
قسم الإحصاء والمعلوماتية ، كلية الإدارة والاقتصاد، جامعة صلاح الدين، اربيل،العراق.
Email: kurdistan.mawlood@su.edu.krd, ORCID: https://orcid.org/0000-0002-1612-1996

المستخلص

ركزت الفكرة الأساسية لهذه الدراسة على استخدام الطريقة البايزية وتقدير الأحتمال الأعضم في انحدار بواسون لنمذجة معدل الإصابة بسرطان الرئة في مدينة أربيل، العراق. يُستخدم انحدار بواسون بشكل شائع لتحليل بيانات الإحصاء، مما يجعله مناسبًا لتطبيقه في تحليل معدلات الإصابة في البيانات الطبية. تقارن الدراسة تقدير الأحتمال الأعضم مع الطريقة البايزية، التي تتضمن التوزيع المسبق في الإحصاء لتقدير المعاملات. تم الحصول على مجموعة بيانات هذه الدراسة، التي تشمل حالات سرطان الرئة مع عوامل الخطر المحتملة، من مستشفى رزكاري في مدينة أربيل.

تستخدم طريقة التقدير البايزي نهج سلسلة ماركوف مونت كارلو لإنشاء توزيع خلفي، ويتم تشخيص فعالية كلتا الطريقتين واختبار جودة الملاءمة. أشارت النتائج إلى أن كلتا الطريقتين نجحتا في تحديد العوامل المهمة لسرطان الرئة، و أنهما تقريباً تصلا إلى نفس العوامل المؤثرة على بيانات سرطان الرئة في مدينة أربيل. وتُعطي الطرق البايزية أداءً أفضل وتوفر تقديرات كمية أكثر قابلية للتفسير. تم الحصول على النتائج باستخدام الحزم الإحصائية (SPSS الإصدار 26، وStata الإصدار 18، وR الإصدار 4.3.1).

معلومات البحث

تواريخ البحث:
التقديم: 13 / 06 / 2025
المراجعة: 01 / 11 / 2025
قبول النشر: 04 / 11 / 2025
نشر الكتروني: 01 / 12 / 2025
تسلسل الصفحات: 14 – 27

الكلمات المفتاحية:
انحدار بواسون، الطريقة البيزية، الأحتمال الأعضم، سلسلة ماركوف مونت كارلو، سرطان الرئة.

المراسلة:
أسم الباحث: كوردستان ابراهيم مولود

Email:
kurdistan.mawlood@su.edu.krd