Journal of Applied Research on Industrial Engineering
www.journal-aprie.com

Poisson Regression Model with the Change Points

Reza Habibi

PAPER INFO

Chronicle:
Received: ****
Revised: ****
Accepted: ***

Keywords:
Keywords: Change Point, Poisson Regression, Mortality Index, Likelihood Ratio, Vital Statistics

ABSTRACT

There are many different fields the change point analysis arises. In those cases, the main problem is locating the unknown change points. The aim of this study is to detect location and time of change point in Poisson regression model. We assume for years before and after the change point k_0, then observation y_t has a Poisson distribution with parameters λ_0, λ_1, respectively. We used several methods for estimation change point in real mortality data by assume Poisson regression model. Using two simulated and real data analysis showed that the change point has been occurred in year 1993 and this confirmed by all methods. Our findings have shown that the change pattern of mortality trend in Iran is related to improvement of health indicators and decreasing mortality rate in Iran.

1. Introduction

The change point problem arises in various practical fields such as Epidemiology, Toxicology, Medical, Economical surveys, Quality control, Statistical process control, Natural events, Demography and Mortality. Demography is a kind of statistical study of populations which is too important for programming in every country. In analytical demography, population analysis is done by applying several mathematical methods to model population distribution and to detect linear and non-linear population changes [1]. Vital statistics such as birth, death, marriage, divorce are important indicators in demography which are affected by different problems duration study years. Increasing or decreasing in the rate of vital statistics cause the change in the pattern of population. Mortality, the death occurs within a population, is affected by population factors such as age, sex, race, occupation and social class. Therefore, changes in population pattern will cause the distribution of mortality in one time period differ from the other time periods.

Let $X_1, \dots, X_{k_0}, X_{k_0+1}, \dots, X_n$ be a sequence of independent observations where $X_1, \dots, X_{k_0}$ come from Poisson distribution according to the formula 1 with rate λ_0 and the remaining observations $X_{k_0+1}, \dots, X_n$ have the Poisson distribution with parameter λ_1, where $\lambda_0 \neq \lambda_1$, that is

$$X_i \sim \begin{cases} poi(\lambda_0) & i = 1,2, \dots, k_0, \\ poi(\lambda_1) & i = k_0 + 1, \dots, n. \end{cases} \quad (1)$$

This problem is referred as the change point detection in statistical literatures. We test the null hypothesis of no change point $H_0: \lambda = \lambda_0$, against $H_1: \lambda = \lambda_1$ which stands a change point has been occurred in unknown point $k_0 = 1, \dots, n-1$. If H_0 is rejected, then we estimate λ_0, λ_1 and k_0.

DOI: 10.22105/jarie.2021.229051.1149

During the last four decades, change point analysis has been received considerable attentions from both theoretical and practical aspects. This problem first proposed in the context of quality control. The most often investigated change point problem is that of the change in the mean of normal variables. Some literatures which deal with change point detection in Poisson distribution random variables are [2]. A best reference in change point detection in parametric families of distributions is [3]. For a comprehensive review in change point analysis see [4] and [5] and references therein.

Testing and estimation the change point in the Poisson distribution has a long history. The maximum likelihood point estimation and driving a Bayesian-based interval estimator for a change point in a Poisson process is studied by [6]. The procedures are evaluated through simulation studies and application to the British coal-mining disaster data [7]. A test for change point hypothesis and estimated position and parameters before and after change point is studied by [8] using simulation and an example on coal-mining disasters. A Bayesian approach to estimation and hypothesis testing for a Poisson process with a change-point in coal-mining disasters data is studied by [8]. We have done the simulations and have seen the good results.

In the current article, change point analysis is done for Iranian mortality rates during 1971 to 2007 under both Poisson and Poisson regression modeling. We first test the existence of change point and then we estimate the location of change point. This article is organized as follows. In section 2, the Bayesian approach is proposed to change point detection in Poisson distribution and then the theoretical results are extended to the Poisson regression models. In section 3, we compare these methods using simulated data and illustrate results of the procedures using mortality data, provided by Iran vital statistics. The conclusions of this study, which are expected to yield new insights regarding potential substantive applications in mortality, are presented in section 4.

2. Change point detection methods

Here, the Bayesian change point detection methods are reviewed. First, they are proposed for Poisson distribution and then they are extended to the Poisson regression model.

2.1. Bayesian approach in Poisson distribution

In spite of likelihood point of view to statistics, the Bayesian approach assumes priors for parameters and uses them to obtain the posterior. We get total information contained in the problem, from the posterior distribution. In a parametric setting, if $\pi(\theta)$ is the probability density of prior of θ and $y = (y_1, \dots, y_n)$ is observed data, then the posterior is given by formula 2 as follows

$$\pi(\theta|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\theta)\pi(\theta)}{\pi(\boldsymbol{y})}, \tag{2}$$

The posterior $\pi(\theta|y)$ is proportional to product of likelihood of $\pi(y|\theta)$ and prior $\pi(\theta)$. The $\pi(y)$ is marginal density of y. We can obtain inferences about the change point and the parameters before and after change point, using a Bayesian approach which specifies prior distributions for the parameters as follows $\pi(\lambda_0) \sim \Gamma(a_0, b_0)$, and

$$\pi(\lambda_1) \sim \Gamma(a_1, b_1),$$
$$\pi(k_0) \sim DU(1,2, \dots, n-1)$$
$$b_0, b_1 \sim IG(0.5,1)$$

where Γ, DU, and IG are gamma, discrete uniform and inverse gamma distributions. Here, for simplicity reasons, it is assumed that $a_0 = a_1 = 0.5$. Therefore, $\theta = (k_0, \lambda_0, \lambda_1, b_0, b_1)$ and the five-dimensional posterior distribution is given by

$$\pi(k_0, \lambda_0, \lambda_1, b_0, b_1|\boldsymbol{y}) \propto \prod_{i=1}^{k_0} {\lambda_0}^{y_i} e^{-\lambda_0} \prod_{i=k_0+1}^{n} {\lambda_1}^{y_i} e^{-\lambda_1} \times$$
$$\times {\lambda_0}^{-0.5} e^{-\lambda_0/b_0} \times {\lambda_1}^{-0.5} e^{-\lambda_1/b_1} \times e^{-\frac{1}{b_0}} {b_0}^{-1.5} \times e^{-\frac{1}{b_1}} {b_1}^{-1.5}$$

which is proportional to

$$\propto \lambda_0^{(\sum_{i=1}^{k_0} Y_i+0.5)-1} e^{-\lambda_0(k_0+\frac{1}{b_0})} \lambda_1^{(\sum_{i=k_0+1}^{n} Y_i+0.5)-1} e^{-\lambda_1(n-k_0+\frac{1}{b_1})} e^{\frac{-\lambda_0}{b_1}} e^{\frac{-\lambda_1}{b_2}} \times$$
$$\times e^{-\frac{1}{b_0}} {b_0}^{-1.5} \times e^{-\frac{1}{b_1}} {b_1}^{-1.5} \quad (3)$$

Markov Chain Monte Carlo (MCMC) [9] is applied to generate samples from posterior distribution similar to EM algorithm in non-Bayesian methods. The Gibbs sampler is one of the most well-known versions of MCMC which simulates samples from the posterior distribution. Sampling from this full conditional posterior distribution ultimately yields draws from the unconditional posterior distribution [9].

For our problem, full conditional distributions are given by formulas 4 as follows

$$f(\lambda_0|k_0,\lambda_1,b_0,b_1,\boldsymbol{y}) = \Gamma\left(\sum_{i=1}^{k_0} Y_i + 0.5, \frac{b_0}{b_0 k_0 + 1}\right),$$
$$f(\lambda_1|k_0,\lambda_0,b_0,b_1,\boldsymbol{y}) = \Gamma\left(\sum_{i=k_0+1}^{n} Y_i + 0.5, \frac{b_1}{b_1(n-k_0)+1}\right),$$
$$f(b_0|k_0,\lambda_0,\lambda_1,b_1,\boldsymbol{y}) = IG\left(0.5, \frac{1}{\lambda_0+1}\right), f(b_1|k_0,\lambda_0,\lambda_1,b_0,\boldsymbol{y}) = IG\left(0.5, \frac{1}{\lambda_1+1}\right),$$
$$f(k_0|\lambda_0,\lambda_1,b_0,b_1,\boldsymbol{y}) \propto \lambda_0^{\sum_{i=1}^{k_0} Y_i} \lambda_1^{\sum_{i=k_0+1}^{n} Y_i} e^{-k_0\lambda_0-(n-k_0)\lambda_1}. \quad (4)$$

2.2. Extension to the Poisson regression

In this section, we formulate our problem as change point detection in Poisson regression model. Poisson regression is a form of regression analysis used to model counting data. Here, it is assumed that response variable Y (mortality rates) has Poisson distribution and the logarithm of its expected value is modeled by a linear combination of some covariates x. In the simplest case, with a single covariates x, the model is given by

$$\log\big(E(Y|X=x)\big) = a + bx.$$

By obtaining the observations $(\mathrm{Y_i}, \mathrm{x_i}), \mathrm{i} = 1, \ldots, \mathrm{n}$, then the parameters a and b are estimated from maximum likelihood method. Following [10], we let covariate be year itself be in our problem is i.e. $x_i = i, i = 1,...,n$. We test if a change has occurred in parameters a and b (at the same time) and after H_0 is rejected, we estimate the change point k_0.

(a) Likelihood ratio (LR) method. Suppose that the Poisson regression is given by formula 5, i.e.,

$$E(Y|x_i) = \lambda_i, \log(\lambda_i) = a_i + b_i x_i \quad (5)$$

where $x_i = i$, $i = 1,...,n$. The null and alternative hypotheses are

$$\begin{cases} H_0 : (a_i, b_i) = (a,b) \ \ for\ i = 1,...,n \\ H_1 : (a_i, b_i) = (a,b)\ for\ \ i = 1,...,k_0\ and\ (a_i, b_i) = (a^*, b^*)\ for\ i = k_0+1,...,n \end{cases}.$$

Under H_1, the log-likelihood function is given by

$$l(a,b,a^*,b^*,k_0) = C + a\sum_{i=1}^{k_0} y_i + b\sum_{i=1}^{k_0} x_i y_i + a^*\sum_{i=k_0+1}^{n} y_i + b^*\sum_{i=k_0+1}^{n} x_i y_i +$$
$$-\sum_{i=1}^{k_0} e^{a+bx_i} - \sum_{i=k_0+1}^{n} e^{a^*+b^*x_i}$$

in which a, b, a^*, b^* are unknown coefficients and C is known constant. The likelihood function under the null hypothesis is derived from the above formula by letting $\mathrm{a} = \mathrm{a}*$ and $\mathrm{b} = \mathrm{b}*$. Since the formulas of maximum likelihood estimators of the parameters as well as the likelihood ratio have nasty forms they are not proposed and they are derived, numerically, in practice.

(b) The Bayesian method. Notice that in the Poisson regression

$$\lambda_i = \exp(\boldsymbol{\beta}' \boldsymbol{X}_i),$$

where $\boldsymbol{\beta}' = (a, b)$ and $\boldsymbol{X}_i = (1, i)$. The prior for $\boldsymbol{\beta}$ is $N_2(\mu, \sum)$ where μ and $\sum$ are known mean vector and variance matrix of $\boldsymbol{\beta}$, respectively. Here, a, b, a^* and b^* are independent with common distribution $\boldsymbol{N}(0, \sigma^2)$. Denote the density of normal distribution with zero mean and variance σ^2 computed at e, by $n(e, 0, \sigma^2)$. Also, in this article, we let $\sigma^2 = 1$(following [3]). Therefore, the five dimensional posterior distributions are given by formula (6) as follow. The maximum a posteriori (MAP) estimations of parameters are derived by maximizing the posterior with respect the parameters, as follow:

$$\pi(a, b, a^*, b^*, k_0 | \boldsymbol{y}) \propto \exp\big(l(a, b, a^*, b^*, k_0)\big) \times n(a, 0, \sigma^2) \times n(b, 0, \sigma^2) \times n(a^*, 0, \sigma^2) \times n(b^*, 0, \sigma^2), \quad (6)$$

where $l(a, b, a^*, b^*, k_0)$ is given by formula 5.

3. Comparisons

Here, the performances of the likelihood ratio and the Bayesian methods are compared in simulated data and a real data set is analyzed by Bayesian method. To run, calculations and computations, R program and packages are used.

(a) Simulated data: LR method. Here, we assume that $a = 2, b = 4$ for $i = 1, \dots, k_0$ and $a^* = 7$, $b^* = 9$ for $i = k_0 + 1, \dots, n = 100$, for the Poisson regression model. The performances of the procedures are compared by MSE of the change point estimators. The results are presented in Table 1. Comparing estimates of $(\hat{a}, \hat{b})$ and $(\hat{a}^*, \hat{b}^*)$ with actual values of (a, b) and (a^*, b^*), it is seen that the MSE of $(\hat{a}, \hat{b})$ and $(\hat{a}^*, \hat{b}^*)$ are too small in each cells of Table 1. Another likelihood-based approach is the Bayesian information criterion (BIC). Notice that the results of BIC and LR methods are the same using simulated observations, since the BIC is a monotone function of LR results; see [3].

Table 1. *LR results: simulation Poisson regression model*

k_0	$\hat{k}$	$\hat{a}$	$\hat{b}$	$MSE(\frac{\hat{k}}{n}) \times 10^5$	$\hat{a}^*$	$\hat{b}^*$
10	*10*	*2*	*4*	*3.6*	*8*	*9*
15	*16*	*2*	*4*	*2.5*	*7*	*9*
20	*21*	*1.98*	*4.2*	*40*	*6.15*	*9.15*
25	*24*	*2*	*4*	*8*	*6*	*9*
30	*30*	*2*	*4*	*0.047*	*6*	*9*
35	*35*	*2*	*4*	*0.013*	*6*	*9*
40	*42*	*2*	*4*	*0.018*	*6.25*	*9*
45	*44*	*2.1*	*3.98*	*18*	*7.08*	*9.1*
50	*51*	*2*	*4*	*0.01*	*7.45*	*9*
55	*54*	*1.9*	*4.08*	*3.4*	*6.98*	*9.03*
60	*61*	*2.1*	*3.98*	*4*	*7.1*	*9.1*
65	*64*	*2*	*4*	*0.051*	*7*	*9*
70	*69*	*2.1*	*4.08*	*2.5*	*7.02*	*9.12*
75	*75*	*2*	*4*	*0.092*	*7*	*9*
80	*81*	*2*	*4*	*0.032*	*7*	*9*
85	*85*	*2*	*4*	*0.034*	*7*	*9*
90	*90*	*2*	*4*	*0*	*7*	*9*
95	*95*	*2*	*4*	*0.14*	*7*	*9*

(b) Simulated data: Bayesian approach. Here, the Bayesian result of running the Bayesian change point detection method in the Poisson regression.

Table.2 *Bayesian results: simulation Poisson regression model*

k_0	$E(\hat{k})$	$\hat{a}$	$\hat{b}$	$MSE(\frac{\hat{k}}{n}) \times 10^5$	$\hat{a}^*$	$\hat{b}^*$
10	*10*	*2*	*4*	*3.2*	*8*	*9*
15	*15*	*2*	*4*	*1.5*	*7*	*9*
20	*21*	*1.98*	*4.2*	*39.38*	*6.15*	*9.15*
25	*25*	*2*	*4*	*7.1*	*6*	*9*
30	*30*	*2*	*4*	*0.084*	*6*	*9*
35	*35*	*2*	*4*	*0.075*	*6*	*9*
40	*40*	*2*	*4*	*0.18*	*6.25*	*9*
45	*44*	*2.1*	*3.98*	*18.73*	*7.08*	*9.1*
50	*50*	*2*	*4*	*0.199*	*7.45*	*9*
55	*54*	*1.9*	*4.08*	*3.19*	*6.98*	*9.03*
60	*61*	*2.1*	*3.98*	*3.52*	*7.1*	*9.1*
65	*65*	*2*	*4*	*1.01*	*7*	*9*
70	*69*	*2.1*	*4.08*	*1.96*	*7.02*	*9.12*
75	*75*	*2*	*4*	*0.88*	*7*	*9*
80	*80*	*2*	*4*	*0.094*	*7*	*9*
85	*85*	*2*	*4*	*0.45*	*7*	*9*
90	*90*	*2*	*4*	*0.88*	*7*	*9*
95	*95*	*2*	*4*	*0.82*	*7*	*9*

The following figure (Fig.1) compares the MSE of LR and Bayesian methods. It is seen that, both of methods works similar although, the Bayesian method is larger than the LR methods, when the change point is close to the starting points.

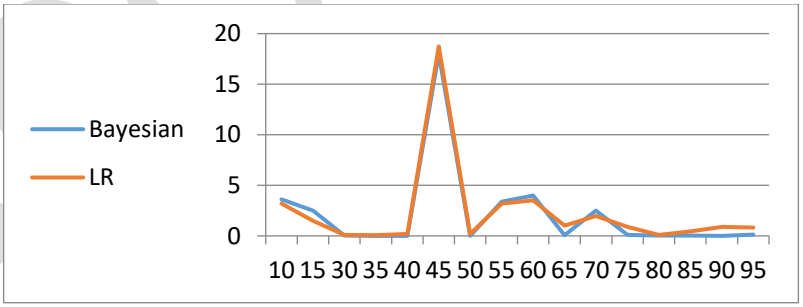

Fig. 1. *Comparison of Bayesian and LR results*

Hereafter, the sensitivity of results to the priors is studied (see, Fig.2). To this end, notice that an small change in the parameter σ^2. It is seen that results are very sensitive to the value of σ^2.

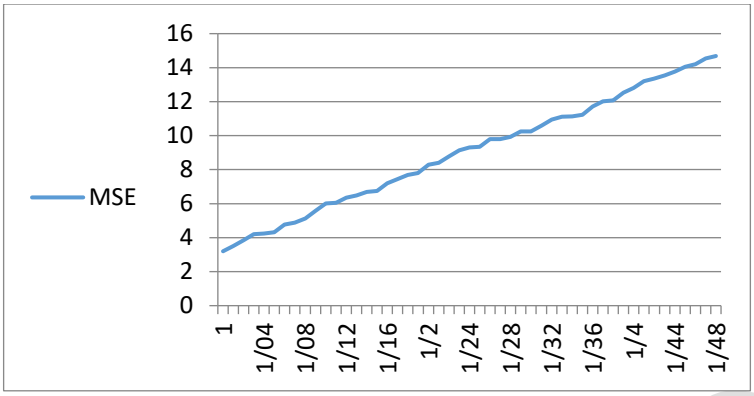

Fig. 2. *MSE of Bayesian-estimate with respect to σ*

(c) Real data set: The mortality data. Vital statistics such as birth, death, marriage, divorce and migration are the best direct indexes to estimate population changes in developed countries. These data have been registered by National Organization for Civil Registration (NOCR) in all countries. In Iran, these data are collected in Statistical Yearly Book published by Statistical Center of Iran (SCI), a national government organization [11]. Another direct procedure for recording information about the members of a given population is census. The census data is commonly used for research, business marketing, planning, sampling surveys and other studies. Census is conducted every ten years by SCI in Iran.

From SCI reports, it is seen that a total of 9.7 million deaths registered in during 1971 to 2007 in Iran. About 5.9 million (61.1 %) were in male and 3.8 million (38.9%) in female with sex ratio 1.6 that from them 5.86 million (60.4%) were in urban and 3.84 million (% 39.6) in rural. Crude death rate in all studied population according to census statistics during 1971 to 2010 was 5.1 per 1000 people. The results showed that the CDR rate in Iran was 13 in years 1970-75 and dropped to 5 in 2005-2010 (see Fig.3). Life expectancy at birth increased during the third (57.7) and fourth (59.6) censuses, but during recent ten years, it was increased more rapidly and reached to 71 in years 2005-2010 (see Fig. 4). This descriptive result suggests the existence of change point among the data. To check these possibilities, we first test the existence of change point and then we estimate it.

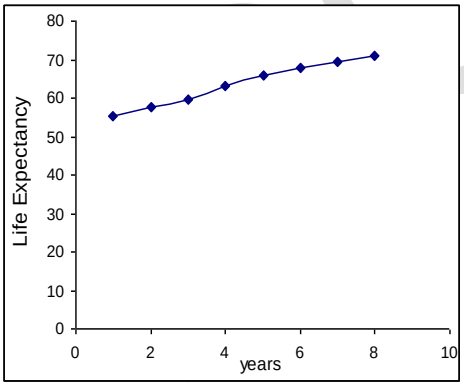

Fig.3. *Life expectancy in Iran 1970-2010*

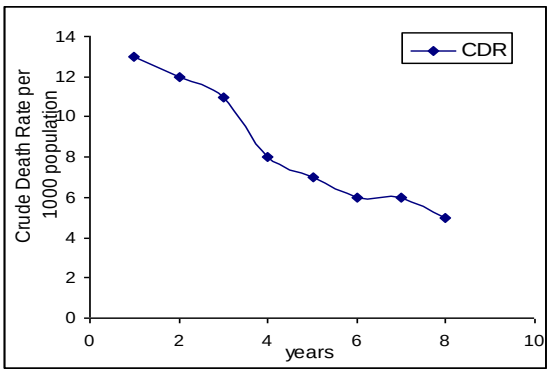

Fig.4. *Crude death rate in Iran1970-2010*

Here, we assume a Poisson regression model for Iranian mortality rates and we let the single covariate be year. We use LR and BIC to detect change point and that is successfully detected in $k_0 = 22$. The null hypothesis is rejected because

$$BIC_{H_0} = 34160 > \min_{1 \le k_0 \le n-1} BIC_{H_1}(k_0) = 2195.5 .$$

The maximum value of LR is 31968.7 with change point at $k_0 = 22$. Results showed that each of the BIC and LR is better for detecting the change point in Poisson regression with similar results. Using the Bayesian setting, the MAP estimator of k_0 is derived. It is seen again the point 22 is best selection

for change point. The variance of this estimator using a bootstrap technique is derived 0.1 which shows this estimator works well.

4. Concluding remarks

In this article, we used likelihood-based and Bayesian methods to detect the change point in Iranian mortality rates. First, we supposed the rates are independent Poisson distributed random variables and then we got the Poisson regression model. It is shown that a change has been occurred on mortality rates at $k_0 = 22$ which stands for year 1993. This result corresponds to result of SCI, although their results are descriptive. All methods work well under Poisson regression model. For future, we are adding some new covariates to our model rather than only years as well as we are training other priors.

5. Acknowledgments

The author is grateful to the referee for several suggestions for improving the article.

References

[1] Assareh, H., Noorossana, R. & Mengersen, K. L. (2013). Bayesian change point estimation in Poisson-based control charts. *Journal of Industrial Engineering International*, *9*(2), 11-23.

[2] Benson, A., & Friel, N. (2018). Adaptive MCMC for multiple change point analysis with applications to large datasets. *Electronic Journal of Statistics*, *12*(1), 3365-3396.

[3] Chen, J. & Gupta, A. K. (2000). *Parametric statistical change point analysis*. Boston: Birkhauser. USA.

[4] Chernoyarov, O. V., Kutoyants, Y. A. & Top, A. (2018). On multiple change-point estimation for Poisson process. *Communications in Statistics - Theory and Methods*, *47*(1), 35-48.

[5] Jarrett, R. G. (2000). A note on the intervals between coal-mining disasters. *Biometrika*, *66*(3), 191-197.

[6] Khosravi, A., Taylor, R., Nagavi, M. & Lopez, A. D. (2007). Mortality in the Islamic Republic of Iran During Years 1964-2004. *Bulletin of World Health Organization*, *85*(2), 607-614.

[7] Mohammad, I. (2006). Trends and patterns of mortality in china, Japan and India: a comparative analysis. *The Social Sciences*, *1*(3), 149-153.

[8] Ng, H. K. & Midi, H. & Ng, K. H. (2017). Change point detection of robust individuals control chart. *International Journal of Industrial Engineering: Theory, Applications and Practice*, *24*(2), 73-85.

[9] Nyambura, S., Mundia, S., Waititu, A. (2016). Estimation of change point in Poisson random variables using the maximum likelihood method. *American Journal of Theoretical and Applied Statistics*, *5*(3), 219-224.

[10] Pina-Monarrez, M. R. (2018). Generalization of the Hotelling's T^2 decomposition method to the r-chart. *International Journal of Industrial Engineering: Theory, Applications and Practice*, *25*(2), 80-95.

[11] Shaochuan, L. (2019). A Bayesian multiple change point model for marked Poisson processes with applications to deep earthquakes. *Stochastic Environmental Research and Risk Assessment*, *33*(1), 59–72.