Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2016; 2(1): 14-19.

Original Article

The application of Poisson hidden Markov model to forecasting new cases of congenital hypothyroidism in Khuzestan province

Majid Sadeghifar[1*], Maryam Seyed-Tabib[2], Saiedeh Haji-Maghsoudi[2], Kourosh Noemani[3], Fariba Aalipur-Byrgany[3]

[1] Department of Statistics, School of Basic Sciences, Bu-Ali Sina University, Hamadan, Iran
[2] Department of Biostatistics & Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
[3] Department of Non-Communicable Disease, School of Health, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

ARTICLE INFO

Received 10.05.2015
Revised 14.09.2015
Accepted 23.11.2015
Published 15.03.2016

Key words:
Congenital hypothyroidism;
Poisson hidden Markov;
forecast;
Khuzestan

ABSTRACT

Background & Aim: Congenital hypothyroidism (CH) is one of the most common endocrine diseases and is a major cause of preventable mental retardation. Early diagnosis of CH can help prevent future diseases. Although time series techniques are often utilized to forecast future status, they are inadequate to deal with count data with overdispersion. The aim of this study was to apply Poisson hidden Markov model to forecast new monthly cases of CH disease.

Methods & Materials: This study was based on the monthly frequency of new CH cases in Khuzestan province of Iran, from 2008 to 2014. We applied stationary Poisson hidden Markov with different states to determine the number of states for the model. According to the model, with the specified state, new CH cases were forecast for the next 24 months.

Results: The Poisson hidden Markov with two states based on Akaike information criterion was chosen for the data. The results of forecasting showed that the new CH cases for the next 2 years comforted in state two with the frequency of new cases at 6-18. The forecast mode and median for all months were 12 and 13, respectively. Each state is explained by each component of dependent mixture model.

Conclusion: Our estimates indicated that state of frequency of CH case is invariant during the forecast time. Forecast means for the next 2 years were from 13 to 14 new CH cases. Furthermore, forecasting intervals were observed between 7 and 25 new cases. These estimates are valid when the general fertility rate and crude birth rate were been fixed.

Introduction

One of the congenital endocrine disorders is congenital hypothyroidism (CH). CH is the problem of thyroid hormone deficiency. It causes lots of troubles such as mental disease when its treatment is not performed on time. In other wordsCH is a endocrine illness caused by an insufficiency of thyroid hormone in infants (1, 2).

CH in newborn infants has been shown by screening programs. Early diagnosis and treatment are the most important goals of neonatal CH screening programs. Nowadays, in many developed countries, the screening of CH is done by measuring the T4 and thyroid stimulating hormone (TSH) routinely. The diagnosis of CH is determined by testing TSH and free thyroxine (T4) (1, 3).

Important advances and valuable findings have been achieved since the introducing of CH screening program in 1970. American Thyroid

* Corresponding Author: Majid Sadeghifar, Postal Address: Department of Statistics, School of Basic Science, Bu-Ali Sina University, Hamadan, Iran. Email: sadeghifar@basu.ac.ir

Please cite this article in press as: Sadeghifar M, Seyed-Tabib M, Haji-Maghsoudi S, et al. The application of Poisson hidden Markov model to forecasting new cases of congenital hypothyroidism in Khuzestan province. *J Biostat Epidemiol. 2016; 2(1): 14-19*

Association has been the biggest supporting organizer of CH screening from 2005 (4, 5).

The findings of studies conducted around the world show that the incidence of CH in live births varies from 1:3000 to 1:4000. Furthermore, in some cases, the incidence has been reported at 1:2000-1:4000 (1, 6).

Screening for CH was the first implemented by Ordookhani et al. (3) in Tehran in 1987. According to the results of studies in Iran, on average, the prevalence of CH disease in Tehran was 1.914 from 1997 to 2001 and in 2002 and 2009 for Esfahan was 1.370 and 1.748, respectively (6).

In another study, the prevalence of CH in newborns in South Khorasan was reported to be 1 in 549 live births from July 2006 to March 2010 (2). Moreover, the incidence ratio of CH in Zanjan from February 2007 to January 2008 was reported to be 1 in 895 live births (1).

The results of Iran's studies show the prevalence of CH screening in this country is higher than other countries that be founded in other studies (1).

Increasing use is being made of time series designs in biomedical data, thanks to the availability of series of administrative or medical data collected routinely including mortality or morbidity counts, environmental measures, changes in socio-economic or demographic indices (7).

When there is a sequence of unbounded count data, applying usual time series models such as autoregressive integrated moving average which are based on continuous outcomes are not suitable for unbounded counts. Because of the serial dependency between the data, some models that can consider this dependency and have the capability of forecasting are needed (8).

One use of Poisson hidden Markov model (PHMM) is the modeling of dynamics of counts data with unobservable underlying processes. Count data for time series have some characteristics such as serial dependence and overdispersion. There are some approaches for modeling these data (9).

Markov models can be drawn on to tackle uncertainty in the absence of historical data. This is in the light of the fact that the probability of observing future state depends only on the probability of observed condition states at the present (10).

HMMs are considered to be powerful tools for this purpose (11). In these models, the distribution of the response based on the current state of the chain can be obtained (12).

Recently, HMMs are being used in many areas of study such as recognition, bioinformatics, finance, DNA decoding, and economics. These models are attractive because of their simplicity.

When there is a sequence of data, Markov chains are stochastic transitions between states and the state at the new step is only dependent on the previous state (13). HMMs are used as the flexible models for both univariate and multivariate time series (14).

Poisson model is a standard way to deal with count data because variables with this distribution are unbounded. Besides, the particular property of this distribution is that the mean and variance are equal. But often this property does not hold. In practice, in this situation, usually, the variance is greater than the mean, which is referred to as overdispersion. To handle this problem, one suggestion is to use mixture models. Mixture models accommodate unobserved heterogeneity and have a finite number of components. In this situation, there is a mixture of finite distributions with a distinct distribution for the variable that has been observed (14). One method for overcoming with the overdispersion due to heterogeneity is using mixture model while the dependency between observations has been considered. One simple way is relaxing to serial dependency, and using markov property (each observation is dependent on only one previews observation). Poisson–hidden markov model is obtained by allowing this assumption (14).

Although a hidden Markov chain has been generalized and extended to different areas, it has not been common to apply this model to medical studies.

In this study, we aim to use Poisson hidden Markov to forecast count data. Due to the importance of knowing the future status of the incidence of diseases in medical and health areas, this study makes use of Poisson hidden

Markov as the forecast method as the nature of the data in these areas usually is count, and the data collected for this purpose have dependency.

Methods

We used the dataset containing the frequency of new cases of CH monthly from 2008 to 2014 in Khozestan province of Iran. These data belong to the infants registered in Shargh laboratory whose tests of TSH were positive (venous sampling).

HMM can be considered a kind of mixture model. The Markov model is comprised two parts. One part is unobserved parameter of process (finite state Markov model C_t) and another part is observed sequence of random variables depending on the first part (X_t). X_t does not depend on previous states of C_t. X_t depends only on the current state. We define as follows:

$$P(C_t|C^{(t-1)}) = P(C_t|C^{t-1}) \qquad t = 2,3,\ldots$$

Where:

$C^{(t-1)} = (C_1, C_2, \ldots, C_{t-1})$

$P(X_t|X^{(t-1)}, C^{(t)}) = P(X_t|C_t) \qquad t \in N$

$C^{(t)} = (C_1, C_2, \ldots, C_t)$

$X^{(t-1)} = (X_1, X_2, \ldots, X_{t-1})$

$P_i(x) = P(X_t = x| C_t = i)$

P_i indicates probability mass function of X_t at time t lie in i^{th} state.

For Poisson observation model, we have as follows:

$$P_i(x) = p(X_t = x|C_t = i) = \frac{e^{\lambda_i}\lambda_i^x}{x!}$$

The likelihood function for this model is shown by:

$$L_T = P(X^{(T)} = x^{(T)}) = \delta P(x_1)\, \Gamma\, P(x_2)\, \Gamma\, P(x_T)1'$$

In this function, initial distribution is δ and P(x) is a diagonal matrix. The elements of P(x) are $P_i(x_1)$s. And Γ is transition probability matrix. Parameters are obtained by maximum likelihood estimation.

Forecasting distribution can be obtained from this equation:

$$P(X_{t+h} = x|X^{(T)} = x^{(T)}) = \phi_T\Gamma^h P(x)1'$$

Where:

$$\phi_T = \alpha_T/\alpha_T 1'$$

And:

$$\alpha_T = \alpha_{T-1}\, \Gamma P(x_t)$$

In the first step, we fitted several stationary PHMMs to determine the number of the state (14).

Akaike information criterion (AIC) and Bayesian information criterion (BIC) are used to compare the models with different state. In this case, the number of the states was considered as transition matrix states, and the proportion of those cases transferred from one state to other states and those states that do not change is computed as elements of this matrix.

Forecasting mean, mode and probability of staying on the new state were calculated for the next 2 years (monthly).

All analyses were done using R3.2.3 software (R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/). We used available codes provided by Zucchini and MacDonald.

Results

Data contained the frequency of new CH cases for 72 months. Minimum and maximum numbers were 6 and 31, respectively. Median new cases were 13.

The mean and variance of new CH cases were 14.47 and 31.15 which indicates strong overdispersion relative to the Poisson distribution and the inappropriateness of that distribution as a model. Figure 1 shows trend plot of this data.

After fitting PHMMs with the different state (from one to three states), the comparison of these models showed the model with two states has less AIC and BIC among others. Improvement in AIC was not seen in three states model so that we chose the model with two states (Table 1).

Minimum and maximum of new CH cases in cluster one and two were 19, 31 and 6, 18, respectively, so these two clusters are state space for Markov chain.

We used this matrix and other parameters from table 1, as initial values for building PHMM.

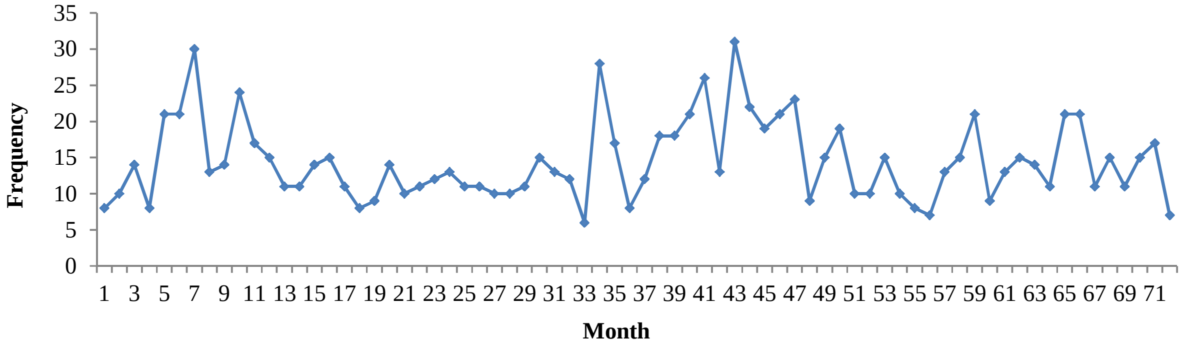

Figure 1. Frequency of new cases with congenital hypothyroidism in Khuzestan province of Iran from 2008 to 2014

Table 1. Comparison of stationary PHMMs with different states

Number of states	AIC	BIC
1	466.17	468.40
2	433.00	442.11
3	443.00	463.00

AIC: Akaike information criterion, BIC: Bayesian information criterion

Table 2. Parameter estimation by maximum likelihood method for two state Poisson hidden Markov

Parameter	State 1	State 2
Lambda*	21.46	11.95
Transition matrix	0.66	0.34
	0.12	0.88
P	0.26	0.74

*Lambda: Mean of distribution in each state

Estimation parameters of Poisson hidden Markov for transition matrix, probability, and lambda (the parameter for Poisson distribution) were obtained by maximum likelihood method.

The probability of jumping from state one to two in one step was 0.34 and the reverse was 0.12. The probability of remaining in state one and two were 0.66 and 0.88, respectively. The mean of distribution in two states were 21.46 and 11.95 (Table 2).

Forecasting result for next 24 months (2015 and 2016) showed that the frequency of new cases of this disease lies in state 2 (count of 6 to 18) with a probability > 0.7, for all months.

Monthly forecast mean varies from 13.07 to 14.38. Forecast mode and median for this period were 12 and 13, respectively. Forecast interval of new cases for the 1st month was interval of 7, 23, 2nd month 7, 24 and 7, 25 for other months (Table 3).

Table 3. Result of forecasting state, mean, median, mode, and interval for two states Poisson hidden Markov

Month		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
State 1	Probability	0.12	0.18	0.22	0.23	0.24	0.25	0.25	0.25	0.26	0.25	0.25	0.25
State 2		0.88	0.82	0.78	0.77	0.76	0.75	0.75	0.75	0.74	0.75	0.75	0.75
Forecast mean		13.07	13.67	13.99	14.17	14.27	14.32	14.35	14.36	14.37	14.38	14.38	14.38
Forecast median		12	13	13	13	13	13	13	13	13	13	13	13
Forecast mode		12	12	12	12	12	12	12	12	12	12	12	12
Forecast interval		(7, 23)	(7, 24)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)
Month		**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
State 1	Probability	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
State 2		0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
Forecast mean		14.38	14.38	14.38	14.38	14.38	14.38	14.38	14.38	14.38	14.38	14.38	14.38
Forecast median		13	13	13	13	13	13	13	13	13	13	13	13
Forecast mode		12	12	12	12	12	12	12	12	12	12	12	12
Forecast interval		(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)	(7, 25)

Discussion

The results of this study showed that the new CH cases in the forecast period are approximately invariant. Lying in state two of these chains with a count of 6-18 showed that new CH cases were decreasing over time. Interval forecasts showed that the frequency of the new CH cases can vary from 7 to 25. These estimates are reliable if other effective factors such as fertility rate and crude birth rate are been constant during the two next years.

In our literature review, we found a limited number of studies conducted on the application of HMMs for forecasting next states.

In some epidemiologic surveillance data, determining epidemic and non-epidemic status is important for health policy makers and the information available to them to proceed. In this situation, forecasting next status is possible by HMMs that offer a set of advantages such as flexibility and ability to handle missing data.

Monitoring epidemiologic surveillance data using HMM is an important area in the field. The results showed a mixture of two dynamics with a low and high level. Low level refers to the non-epidemic dynamic (the incidence rate of change based on seasonal pattern) and high level refers to the epidemic dynamic (the sharp increase of the incidence rate in irregular intervals). They argued that this method could present a clear distinction between epidemic and non-epidemic rates. It is for the purposes of an illustration that this model can be used since it meets a standard epidemiologic objective, that is, the identification of the timing of epidemic periods (8).

Junko Murakami used Bayesian approach to estimate parameter of PHMMs instead of expectation maximization and Markov chain Monte Carlo approaches. The results of their study showed that the Bayesian approach was superior to maximum likelihood estimation for the data with small size and small observation space (15).

Using HMM is applicable in many studies. Previous studies applied HMMs for other purposes than forecasting. In a study conducted by Roberta Paroli et al., PHMMs were used in non-life insurance. Results suggested that PHMMs area more general approach compared to Poisson distribution and Poisson process to model claim number in non-life insurances (16).

In another study conducted by Olteanu and Ridgway (17), the suggestion was made to apply HMMs to time series of counts with excess zeros. The real-life data example showed that the ZIP-HMM performs better than the HMM when there is strong overdispersion in zero.

Also, Green and Richardson (18) presented a new methodology to extend the HMM to the spatial domain where it is used to analyze spatial heterogeneity of count data on rare outcomes using rare phenomena. They also propose hierarchical Poisson model, a new model within the HMM random field of the framework.

One limitation of this study was that we focused only on the frequency of new cases per month without considering other effective factors. On the other hand, we had data for 6 years. This sample size was not large enough for forecasting 24 months. Our suggestion for future studies is to apply the PHMMs in the presence of covariates because several factors play important roles in the advent of diseases.

Conclusion

Modelling overdispersion in the count data in addition to variability is possible using poisson hidden markov models where the poisson parameter switches to an unobserved Markov chain. As the application of PHMMs to forecast, the future was not undertaken in other previous studies of forecasting health and treatments parameters, this study merely put forward an alternative method to common forecast methods, in the light of their shortcomings.

Acknowledgments

We are deeply grateful to Professor Walter Zucchini for his invaluable help. We would like to thank the Deputy of Health of Ahvaz University of Medical Sciences for providing the data.

References

1. Valizadeh M, Mazloomzadeh S, Niksirat A,

Shajari Z. High incidence and recall rate of congenital hypothyroidism in Zanjan province, a health problem or a study challenge? Int J Endocrinol Metab 2011; 9(4): 338-42.

2. Namakin K, Sedighi E, Sharifzadeh G, Zardast. Prevalence of congenital hypothyroidism in south Khorasan province (2006-2010). J Birjand Univ Med Sci 2012; 19(2): 191-9. [In Persian].
3. Ordookhani A, Mirmiran P, Najafi R, Hedayati M, Azizi F. Congenital hypothyroidism in Iran. Indian J Pediatr 2003; 70(8): 625-8.
4. Rastogi MV, LaFranchi SH. Congenital hypothyroidism. Orphanet J Rare Dis 2010; 5: 17.
5. Yordam N, Calikoglu AS, Hatun S, Kandemir N, Oguz H, Tezic T, et al. Screening for congenital hypothyroidism in Turkey. Eur J Pediatr 1995; 154(8): 614-6.
6. Feizi A, Hashemipour M, Hovsepian S, Amirkhani Z, Klishadi R, Rafee Al-Hosseini M, et al. Study of the efficacy of therapeutic interventions in growth normalization of children with congenital hypothyroidism detected by neonatal screening. Iran J Endocrinol Metab 2011; 13(6): 681-9. [In Persian].
7. Fazekas M. Application time series models on medical research [Online]. [cited 2014]; Available from: URL: http://m.ludita.uni-nke.hu/repozitorium/handle/11410/109?show=full
8. Cooper B, Lipsitch M. The analysis of hospital infection data using hidden Markov models. Biostatistics 2004; 5(2): 223-37.
9. Lu Y, Zeng L. A nonhomogeneous Poisson hidden Markov model for claim counts. ASTIN Bulletin 2012; 42(1): 181-202.
10. Nam L, Kaito K, Kobayashi K. A Poisson hidden Markov model for deterioration prediction of road asset system [Online]. [cited 2010]; Available from: URL: http://library.jsce.or.jp/jsce/open/00039/201011_no42/pdf/161.pdf
11. Antonucci A, de Rosa R. Time series classification by imprecise hidden Markov models [Online]. [cited 2011]; Available from: URL: http://people.idsia.ch/~alessandro/papers/antonucci2011e.pdf
12. Viviano LCM. Discrete or continuous-time hidden Markov models for count time series [Online]. [cited 2006]; Available from: URL: http://old.sis-statistica.org/files/pdf/atti/rs08_spontanee_10_4.pdf
13. Inge A. Hidden Markov models. Theory and simulation [Thesis]. Stockholm, Sweden: Stockholm University; 2013.
14. Zucchini W, MacDonald IL. Hidden Markov models for time series: an introduction using R. Boca Raton, FL: CRC Press; 2009.
15. Murakami J. Bayesian posterior mean estimates for Poisson hidden Markov models. Computational Statistics & Data Analysis 2009; 53(4): 941-55.
16. Paroli R, Redaelli G, Valizadeh M. Poisson hidden markov models for time series of overdispersed insurance counts [Online]. [cited 2000]; Available from: URL: http://www.actuaries.org/astin/colloquia/porto_cervo/paroli_redaelli_spezia.pdf
17. Olteanu M, Ridgway J. Hidden Markov models for time series of counts with excess zeros [Online]. [cited 2012]; Available from: URL: https://hal.inria.fr/file/index/docid/655588/filename/esannV2.pdf
18. Green PJ, Richardson S. Hidden Markov models and disease mapping. Journal of the American Statistical Association 2002; 97(460): 1055-70.