Original Article

Analysis of Infant Congenital Malformation Data using the Bayesian Count Regression

Reza Ali Akbari Khoei[1] (iD) , Anoshirvan Kazemnejad[2,*] (iD) , Farzad Eskandari[3] and Mohammad Heidarzadeh[4]

[1] PhD Candidate of Biostatistics, Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
[2] Professor, Department of Biostatistics, Faculty of MedicalSciences, Tarbiat Modares University, Tehran, Iran
[3] Professor, Department of Statistics, Mathematics and Computer Sciences, Allameh Tabatabai University, Tehran, Iran
[4] Department of Pediatrics, Tabriz University of Medical Sciences, Director of Department of Neonatal Health, Ministry of Health and Medical Education, Tabriz, Iran

***Corresponding author:** Anoshirvan Kazemnejad, Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran. Tel: +982182883875; Email: kazem_an@modares.ac.ir

Abstract

Background: Congenital malformations are one of the most important and common types of anomalies in infants, which are one of the main causes of disability and mortality in children.

Objectives: This study aimed to investigate the risk factors affecting the incidence of congenital malformations, as well as the number of different infant anomalies recorded in neonatal health data in Khoy, Iran, during 2017.

Methods: In this study, all neonates born in the maternity wards of hospitals in Khoy, Iran, during 2017 were evaluated in terms of gender, weight, and parental consanguinity. Hurdle and Zero-inflation approaches were utilized for the double Poisson model. Moreover, the data were collected using some checklists, and the analyses were performed in R-3-6-1 software.

Results: According to the results of the present study, the Hurdle approach was better than Zero-inflation. The birth weight and parental consanguinity affected the incidence of congenital malformations in infants.

Conclusion: Given that a significant proportion of infants are born without any congenital malformations, it is important to use count regression models based on excess zero approaches to assess congenital malformations. It is also necessary to take steps to reduce consanguineous marriages and the number of infants with low-birth-weight to prevent congenital malformations.

Keywords: Congenital malformation, Count regression, Hurdle, Zero-inflation

1. Background

Birth defects are structural or functional anomalies at birth that can lead to physical, mental, and developmental disabilities, as well as other health problems (1). Congenital malformations are one of the most common anomalies in children (2). Any type of anatomical defect at birth that has medical, surgical, or cosmetic consequences is called congenital malformations (3), which are important causes of disability and mortality in infants (4). According to the World Health Organization, three million babies are born with birth defects each year in the world, killing 495,000 babies worldwide (5).

Moreover, 20% of deaths in children under one year of age and 25% of hospitalizations are due to congenital malformation (6). Although various genetic, environmental, and teratogenic factors, such as maternal addiction to alcohol, diabetes, malnutrition, infection, hyperthermia, drug use, and contact with chemicals or radioactive substances are mentioned as factors causing congenital malformations, the cause of the anomaly is unclear in 40-60% of people with congenital malformations (7). In Iran, about 30,000 to 40,000 disabled infants are born each year, and this significant number not only has physical and mental

problems but also creates many psychological and economic problems for their families and the government (8).

The individual's family system is also negatively affected by each person's problems and disabilities. Proper identification of birth defects is the first step to providing useful genetic counseling to a parent couple. Because of the importance of life expectancy in newborns, congenital malformations are the most important issue in health care today (9).

Poisson regression can be used to investigate the factors and risk factors affecting the number of birth defects in newborns. The Poisson regression model is the most common model used in count data analysis. An interesting feature of this distribution is the equality of mean and variance (10). One of the most common problems in count data analysis is the possibility of over-dispersion in this type of data. If the variance is larger than the mean, over-dispersion occurs; accordingly, measures need to be taken to combat over-dispersion. Another issue with count data is the possibility of being extra zeroes (11).

In this case, there are more zeroes in the data, compared to those in the ordinary Poisson distribution. Under these circumstances, the use of Poisson distribution will lead to erroneous inferences

and misleading results. In order to deal with the extra zeros, Hurdle and Zero-inflated approaches have been introduced by Mullahy [12] and Lambert [13], respectively. To deal with over-dispersion, there are some models, such as negative-binomial [14], Conway-Maxwell Poisson [15], Hermit [16], and generalized Poisson [17]. Zero-inflated negative-binomial regression, Hurdle negative-binomial regression [18], Zero-inflated Conway-Maxwell Poisson regression [19], Zero-inflated hermit [20], and Zero-inflated generalized Poisson regression [21] were used in order to simultaneously investigate over-dispersion and extra zeroes. Another distribution used for count data is the double Poisson distribution. This distribution is a special case of a double exponential family first introduced by Efron [22]. Hurdle and Zero-inflated approaches for this distribution are introduced by Gurmu [23] and Karen [24], respectively.

2. Objectives

This study aimed to investigate the causes of congenital malformations among newborns in Khoy, Iran, during 2017. To achieve this goal, several count regression models were utilized, such as Zero-inflated double Poisson and Hurdle double Poisson using a Bayesian Approach. Moreover, Deviance Information Criterion (DIC) index was used to compare them with the Zero-inflated negative binomial and Hurdle negative binomial. Finally, the best model was selected, and the risk factors affecting the number of congenital malformations of newborns were determined based on this model.

3. Methods

3.1. Data

The neonatal data set in Khoy, Iran, were used in this study, which was gathered under the supervision of the neonatal health department of the Ministry of Health and Medical Education. This national data set contains the demographic characteristics of the mother and the infants, types of neonatal anomalies, and involved risk factors. The data set recorded in 2017 has been studied in the present study. In order to evaluate and fit the regression models, the number of anomalies for each infant was considered the response variable. The gender, birth weight, and parental consanguinity were considered predictive variables and entered into regression models.

3.2. Double Poisson Distribution

Double Poisson distribution introduced by Efron as follows [22]:

$$f_{\theta,\mu}(y) = c(\theta,\mu)\sqrt{\theta}\left(\frac{e^{-\mu}\mu^y}{y!}\right)^{\theta}\left(\frac{e^{-y}y^y}{y!}\right)^{1-\theta} \qquad 1$$

Where μ and θ are location and dispersion

parameters, respectively, and $c(\theta,\mu)$ is constant normalizer. Moreover, E(Y)=μ and Var(Y)=$\frac{\mu}{\theta}$ signify the mean and variance of this distribution, respectively. Due to this variance, if $\theta = 1$, double Poisson converts to ordinary Poisson distribution.

3.3. Zero-inflated Double Poisson Distribution

Zero-inflated double Poisson distribution is obtained as follows [25]:

$$P(Y = y)$$
$$= \begin{cases} p + (1-p)c(\theta,\mu)\sqrt{\theta}e^{-\mu\theta} & y = 0 \\ (1-p)c(\theta,\mu)\sqrt{\theta}\left(\frac{e^{-\mu}\mu^y}{y!}\right)^{\theta}\left(\frac{e^{-y}y^y}{y!}\right)^{1-\theta} & y = 0,1,2,\dots \end{cases} \qquad 2$$

Regarding the parameters in Equation (2), the influence of covariates can be obtained as below:

$$\mu_i = \exp(\beta_0 + \beta_1 X_i + \cdots + \beta_p X_i) \qquad 3$$

$$logit(p_i) = \frac{\exp(\gamma_0 + \gamma_1 W_i + \cdots + \gamma_p W_i)}{1 + \exp(\gamma_0 + \gamma_1 W_i + \cdots + \gamma_p W_i)} \qquad 4$$

3.4. Hurdle Double Poisson Distribution

Hurdle double Poisson distribution is obtained as follows:

$$P(Y = y) = \begin{cases} p & y = 0 \\ (1-p)\frac{c(\theta,\mu)\sqrt{\theta}\left(\frac{e^{-\mu}\mu^y}{y!}\right)^{\theta}\left(\frac{e^{-y}y^y}{y!}\right)^{1-\theta}}{(1 - c(\theta,\mu)\sqrt{\theta}e^{-\mu\theta})} & y = 1,2,\dots \end{cases} \qquad 5$$

Where positive values have zero-truncated double Poisson distribution [12].

Regarding the parameters in Equation (5), the influence of covariates is based on (3) and (4).

3.5. Parameter Estimation Method and Comparative Criterion for Bayesian Models

Various prior distributions can be considered for regression coefficients. Madigan et al. used Laplace and Normal priors for logistic regression coefficients [26]. Eskandari et al. used the Bayesian logistic regression model to emphasize Laplace prior via the Laplace-Metropolis algorithm [27]. In this study, the posterior distribution is obtained, and statistical inference is based on the posterior distribution using the non-informative prior distributions for the model parameters as well as the likelihood function for each model. The Markov Chain Monte Carlo method is used to calculate the posterior distribution. The following methods are used to compare the models used.

3.5.1. Deviance Information Criteria

One of the common indices for making comparisons among Bayesian models is the DIC, which

has been introduced by Spiegelhalter et al. ([28]).

$$DIC = \overline{D}(\theta) + p_D \qquad 16$$

Where $\overline{D}(\theta) = E(D(\theta)|y))$ is the posterior mean of deviance and p_D denotes the difference between the mean of posterior deviance and that of the posterior mean for parameters.

The model with the lowest DIC value is selected as the better model.

3.5.2. Geweke
Geweke statistic has a standard normal distribution and its non-significant mean convergence of Markov chain ([29]).

3.5.3. Raftery- Lewis
If the value of this index is less than 5, the Markov chain will be convergent ([30]).

4. Results

This section utilized the data of congenital malformations of newborns during 2017 in Khoy, Iran. This data set was collected under the supervision of the Department of Neonatal Health of the Ministry of Health and Medical Education. Table 1 tabulates the frequency of the number of malformations for each infant.

As can be seen, a significant proportion of infants are born without any congenital malformations.

Table 1. Frequency for response variable

Number of malformation	N (%)
0	6360(99.6)
1	1(0.0002)
2	20(0.003)
3	2(0.0003)
4	3(0.0005)

The mean age of the mothers was obtained at 27.8 years, and the majority (bout 52%) of the neonates were male. Moreover, 83% of the parents had no consanguineous marriages. The frequency of malformation was within the range between 0 and 4; moreover, the mean and variance values were approximately determined at 0.01 and 0.02 respectively, indicating the inequality of mean and variance leading to the over-dispersion in data.

Data analysis was performed using OpenBUGS software, R2OpenBUGS, and coda statistical packages in the R programming language. The statistical value of the Geweke was not significant for any of the parameters of the models used, indicating the convergence of the chains. All values for the Raftery-Lewis test were between 1 and 5, indicating that the sample size was sufficient for the convergence of the chains. Therefore, Markov chains for all models have reached their stationary and convergence. The DIC value for the models is shown in Table 2. Hurdle double Poisson model with a value of 12,401.48 indicated that this model performed slightly better. Therefore, the output for this model is shown in Table 3. In Bayesian models, the analysis is based on

Table 2. Deviance information criteria

HUDP*	ZIDP	HUNB	ZINB
-12401.48	-12401.09	-12401.25	-12400.81

*HUDP: Hurdle Double Poisson, ZIDP: Zero Inflated Double Poisson, HUNB: Hurdle Negative Binomial, ZINB: Zero Inflated Negative Binomial

Table 3. Bayesian estimates of parameters

Logistic part Levels	Mean	S.D	Lower CI*	Upper CI
Intercept	2.41	0.71	1.20	3.89
Gender				
Male	Reference	-----------	------------	------------
Female	0.29	0.40	-0.47	1.08
Parental Relativity				
No	Reference	-----------	-----------	-----------
Yes	0.59	0.21	0.12	1.01
Birth Weight (kg)				
<1	Reference	--------------	---------------	------------
1-1.5	1.08	1.35	-1.22	4.24
1.5- 2.5	0.52	0.69	-0.91	1.76
>2.5	-2.12	0.71	-0.98	-3.37
Count model part				
Intercept	0.72	0.30	0.13	1.27
Gender				
Male	Reference	-----------	------------	-------------
Female	0.04	0.35	-0.65	0.70
Parental Relativity	Reference	--------------	--------------	-------------
Yes	-0.21	0.36	-0.85	0.51
Theta	0.88	0.09	0.66	1

CI: Credible Interval

Credible Intervals. These intervals play a role similar to confidence intervals and p-value in classic statistics. If this interval contains zero, it indicates the non-significance of the covariate on the response.

5. Discussion

This study was conducted on the data of neonatal malformations in the Iranian Maternal and Newborn Network. Since a significant proportion of infants are born without any congenital malformations, the response variable in this study obtained many zeroes. Therefore, Hurdle and Zero-inflated approaches were used to examine the effect of risk factors on the occurrence and number of anomalies. Therefore, in this study, due to the effects of extra zeros and over-dispersion on data, Hurdle and Zero-inflated double Poisson regression models were used, and the results were compared with those obtained from the Hurdle and Zero-inflated negative binomial regression models.

According to the DIC index values, the Hurdle double Poisson model performed better than other models. Therefore, the output is examined and interpreted based on this model. If the value of all model covariates is zero, the intercept will be equal to the mean of the response. In the model used, the covariates will be zero when they are all in the reference category. In the final model, both intercepts represent the mean response for male infants under one kg without parental consanguinity.

In the logistic regression, OR=exp (b). In this study, the values of the odds ratio were obtained using the mean column in Table 3 (OR=exp [Mean]). The odds ratio for the risk factor of the parental consanguinity was exp (0.59)=1.80. This means that the infants of consanguineous parents had an 80% higher chance of malformation, which was similar to the results of a study conducted by Rittler in which the infant's parental consanguinity was one of the causes of malformation (31). In addition, the obtained results were consistent with the findings of a study performed by Mosayebi (32). The odds ratio for infants above 2.5 kg was obtained at exp (-2.12). This means that infants under one kg have 1/exp (-2.12)=8.33 times more chance to get congenital malformations. These results were in line with the findings of a study carried out by Tulandi (33). They were also consistent with the findings of a study performed by Taksande in which maternal age and low birth weight were the risk factors for congenital malformations (34).

6. Conclusion

According to the obtained results, parental consanguinity and low birth weight of the infant are considered important factors in the occurrence of congenital malformations in infants. To reduce the chance of getting congenital malformations in infants, culturalization is necessary in this regard to reduce consanguineous marriages; moreover, the necessary measures should be taken to increase weight at birth. These results were obtained using the Hurdle double Poisson regression model, which had a better performance among the models used. Therefore, the use of this model is recommended to analyze the data of neonatal anomalies. The results of this study, taking into account the comprehensive data, can be used to implement health and preventive measures to reduce the number of neonatal birth defects.

Footnotes

Authors' Contribution: Reza Ali Akbari Khoei wrote the manuscript and performed the statistical analysis and study planning. Anoshirvan Kazemnejad edited the manuscript and performed the statistical analysis and study planning. Farzad Eskandari performed the statistical interpretation of the results. Mohammad Heidarzadeh conducted the clinical counseling for the results of the study.

Conflict of Interests: The authors declare no conflict of interest.

Ethical Approval: The study protocol was approved by the Research Ethics Committee of Tarbiat Modares University, Tehran, Iran (IR.MODARES.REC.1397.222).

Funding/Support: There was no funding.

Financial Disclosure: There was no funding.

Informed consent: Regarding the research design, there was no need to obtain informed consent.

References

1. Centers for Disease Control and Prevention (CDC). Update on overall prevalence of major birth defects - Atlanta, Georgia, 1978-2005. *Morb Mortal Wkly Rep*. 2008;**57**(1):1-5. doi: 10.1001/jama.299.7.756. [PubMed: 18185492].
2. Kehr PH, Karger C. John Herring (ed): Tachdjian's pediatric orthopaedics, 4th edn. *Eur J Orthop Surg Traumatol*. 2009;**19**(8):605. doi: 10.1007/s00590-009-0508-9.
3. Fama EF, French KR. Dissecting anomalies with a five-factor model. *Rev Finan Stud*. 2016;**29**(1):69-103. Doi: 10.1093/rfs/hhv043.
4. Garry VF, Harkins ME, Erickson LL, Long-Simpson LK, Holland SE, Burroughs BL. Birth defects, season of conception, and sex of children born to pesticide applicators living in the Red River Valley of Minnesota, USA. *Environ Health Perspect*. 2002;**110**(Suppl 3):441-9. doi: 10.1289/ehp.02110s3441. [PubMed: 12060842].
5. Shawky RM, Sadik DI. Congenital malformations prevalent among

Egyptian children and associated risk factors. *Egypt J Med Hum Genet*. 2011;**12**(1):69-78. doi: 10.1016/j.ejmhg.2011.02.016.

6. Baxter R, Hastings N, Law A, Glass EJ. A rapid and robust sequence-based genotyping method for BoLA-DRB3 alleles in large numbers of heterozygous cattle. *Anim Genet*. 2008;**39**(5):561-3. doi: 10.1111/j.1365-2052.2008.01757.x. [PubMed: 18637877].

7. Karbasi SA, Golestan M, Fallah R, Mirnaseri F, Barkhordari K, Bafghee MS. Prevalence of congenital malformations. *Acta Med Iran*. 2009;**47**(2):149-53.

8. Dastgiri S, Stone DH, Le-Ha C, Gilmour WH. Prevalence and secular trend of congenital anomalies in Glasgow, UK. *Arch Dis Child*. 2002;**86**:257-63. doi: 10.1136/adc.86.4.257. [PubMed: 11919098].

9. Corsello G, Giuffrè M. Congenital malformations. *J Matern Neonatal Med*. 2012;**25**(Suppl 1):25-9. doi: 10.3109/14767058.2012.664943. [PubMed: 22356564].

10. Cameron A, Trivedi P. Microeconometrics: methods and applications. Cambridge: Cambridge University Press; 2005.

11. Dalrymple ML, Hudson IL, Ford RP. Finite mixture, zero-inflated poisson and hurdle models with application to SIDS. *Comput Stat Data Anal*. 2003;**41**(3-4):491-504. doi: 10.1016/S0167-9473(02)00187-1.

12. Mullahy J. Specification and testing of some modified count data models. *J Econ*. 1986;**33**(3):341-65. doi: 10.1016/0304-4076(86)90002-3.

13. Lambert D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*. 1992;**34**(1):1-14. doi: 10.1080/00401706.1992.10485228.

14. Hilbe JM. Negative binomial regression. 2nd ed. Cambridge: Cambridge University Press; 2011.

15. Guikema SD, Goffelt JP. A flexible count data regression model for risk analysis. *Risk Anal*. 2008;**28**(1):213-23. doi: 10.1111/j.1539-6924.2008.01014.x. [PubMed: 18304118].

16. Giles DE. Hermite regression analysis of multi-modal count data. *Econ Bull*. 2010;**30**(4):2936-45.

17. Consul PC, Famoye F. Generalized poisson regression model. *Commun Stat Theory Methods*. 1992;**21**(1):89-109. doi: 10.1080/03610929208830766.

18. Gurmu S, Trivedi PK. Excess zeros in count models for recreational trips. *J Bus Econ Stat*. 1996;**14**(4):469-77. doi: 10.1080/07350015.1996.10524676.

19. Barriga GD, Louzada F. The zero-inflated Conway–Maxwell–Poisson distribution: Bayesian inference, regression modeling and influence diagnostic. *Stat Methodol*. 2014;**21**:23-34. doi: 10.1016/j.stamet.2013.11.003.

20. Satheesh Kumar C, Ramachandran R. On some aspects of a zero-inflated overdispersed model and its applications. *J Appl Stat*. 2020;**47**(3):506-23. doi: 10.1080/02664763.2019.1645098.

21. Gupta PL, Gupta RC, Tripathi RC. Score test for zero inflated generalized poisson regression model. *Commun Stat Theory Methods*. 2005;**33**(1):47-64. doi: 10.1081/STA-120026576.

22. Efron B. Double exponential families and their use in generalized linear regression. *J Am Stat Assoc*. 1986;**81**(395):709-21. doi: 10.1080/01621459.1986.10478327.

23. Gurmu S. Generalized hurdle count data regression models. *Econ Lett*. 1998;**58**(3):263-8. doi: 10.1016/S0165-1765(97)00295-4.

24. Yip KC, Yau KK. On modeling claim frequency data in general insurance with extra zeros. *Insur Math Econ*. 2005;**36**(2):153-63. doi: 10.1016/j.insmatheco.2004.11.002.

25. Aragon DC, Achcar JA, Martinez EZ. Maximum likelihood and Bayesian estimators for the double Poisson distribution. *J Stat Theory Pract*. 2018;**12**(4):886-911. doi: 10.1080/15598608.2018.1489919.

26. Madigan D, Genkin A, Lewis DD, Fradkin D. Bayesian multinomial logistic regression for author identification. AIP Conference Proceedings. *Ame Instit Physics*. 2005;**803**(1):509-16. doi: 10.1063/1.2149832.

27. Eskandari F, Meshkani MR. Bayesian logistic regression model choice via laplace-metropolis algorithm. *J Iran Stat Soc*. 2006;**5**(1):9-24.

28. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B Stat Methodol*. 2002;**64**(4):583-616. doi: 10.1111/1467-9868.00353.

29. Geweke J. Statistics, probability and chaos: comment: inference and prediction in the presence of uncertainty and determinism. *Stat Sci*. 1992;**7**:94-101. doi: 10.2307/2245993.

30. Carlo M. Comment: one long run with diagnostics: implementation strategies for markov chain. *Stat Sci*. 1992; **7**(4):493-7.

31. Rittler M, Liascovich R, López-Camelo J, Castilla EE. Parental consanguinity in specific types of congenital anomalies. *Am J Med Genet*. 2001;**102**(1):36-43. doi: 10.1002/1096-8628(20010722)102:1<36::AID-AJMG1394>3.0.CO;2-M. [PubMed: 11471170].

32. Mosayebi Z, Movahedian AH. Pattern of congenital malformations in consanguineous versus nonconsanguineous marriages in Kashan, Islamic Republic of Iran. *EMHJ East Mediterr Health J*. 2007;**13**(4):868-75.

33. Tulandi T, Martin J, Al-Fadhli R, Kabli N, Forman R, Hitkari J, et al. Congenital malformations among 911 newborns conceived after infertility treatment with letrozole or clomiphene citrate. *Fertil Steril*. 2006;**85**(6):1761-5. doi: 10.1016/j.fertnstert.2006.03.014. [PubMed: 16650422].

34. Taksande A, Vilhekar K, Chaturvedi P, Jain M. Congenital malformations at birth in central India: a rural medical college hospital based data. *Indian J Hum Genet*. 2010;**16**(3):159-63. doi: 10.4103/0971-6866.73412. [PubMed: 21206705].