ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2017-57306

Determination of the Affecting Factors of the Number of Babies Born Alive in Multiple Pregnancies with Poisson Models

Çoklu Gebeliklerde Poisson Modelleri ile Canlı Doğan Bebek Sayısını Etkileyen Faktörlerin Belirlenmesi

Gizem ERKAN,[a]
Ozan EVKAYA,[b]
Semra TÜRKAN[c]

[a]The Ministry of Youth and Sports,
[b]Department of Mathematics,
Atılım University
Faculty of Arts & Sciences,
[c]Department of Statistics,
Hacettepe University Faculty of Science,
Ankara

Geliş Tarihi/*Received:* 17.07.2017
Kabul Tarihi/*Accepted:* 16.10.2017

Yazışma Adresi/*Correspondence:*
Ozan EVKAYA
Atılım University
Faculty of Arts & Sciences,
Department of Mathematics, Ankara,
TURKEY/TÜRKİYE
ozanevkaya@gmail.com

This article presented as a 3rd International Researchers, Statisticians and Young Statisticians Congress, 24-26 May 2017, Konya, Turkey.

ABSTRACT Objective: Multiple pregnancies occurred more frequently with being widespread of the assisted reproduction techniques. The recent researches showed that the possibility of multiple pregnancies has been increased by some factors such as twin pregnancy experience in the family, mothers at later ages, social properties and the number of live-born infants. The main aim of this study is to identify the statistically significant factors affecting the multiple pregnancies using count data models. **Material and Methods:** In this study, the number of babies born alive for the pregnant who have multiple pregnancy diagnose are considered in 2015 for a specific location, Ankara province in Turkey. For this purpose, the effects of mother's age, the number of pregnancy, the method of delivery, mother's blood type and previous births of the mother are statistically analyzed using various count data models. **Results:** Quasi Poisson and Conway-Maxwell-Poisson (COM) regression models are used due to under-dispersion problem and these models are compared for the data set. As a result of comparison, the advantage of COM Poisson regression to other count regression models are illustrated with a real data set. According to results of COM Poisson, the number of pregnant and method of delivery (specifically cesarean type) has significant impact on the number of live-born infants at %99 significance level and the blood type of pregnant has significant impact at %95 significance level. However, the cesarean delivery has negative impact on the number of live-born infants. **Conclusion:** The main indicator of existence of under-dispersion is significant so fitting COM-Poisson to the data in this study is meaningful. For the model selection, based on AIC values of Poisson and COM-Poisson models, the latter one is smaller hence fits better. After recovering the problem of under-dispersion for the count data with COM Poisson regression, the number of pregnancy and the method of delivery has been determined as the best predictor. Besides, the blood types were identified as additional explanatory variable, but with lower significance level.

Keywords: Multiple Pregnancy; quasi poisson regression; under-dispersion, COM Poisson regression

ÖZET Amaç: Günümüzde yardımcı üreme tekniklerinin giderek yaygınlaşması ile çoğul gebeliklere daha sık rastlanmaya başlanmıştır. Son yıllarda yapılan araştırmalara göre, ailede daha önceden ikiz gebelik olması, ileri anne yaşı, toplumsal özellikler ve canlı doğan bebek sayısı gibi faktörlerin çoklu gebelik görülme olasılığını artırdığı gözlemlenmiştir. Bu çalışmanın esas amacı çoğul gebelikleri etkileyen istatistiksel olarak anlamlı faktörleri sayım modelleri kullanarak belirlemektir. **Gereç ve Yöntemler:** Bu çalışmada 2015 yılında Ankara ilinde çoklu gebelik tanısı konmuş gebelerin doğum sonuçlarına göre canlı doğan bebek sayısı dikkate alınmıştır. Bu amaçla, annenin yaşı, kaçıncı gebeliği olduğu, doğum yöntemi, kan grubu ve önceki doğum durumu gibi faktörlerin etkileri istatistiksel olarak sayım modelleri kullanılarak analiz edilmiştir. **Bulgular:** Az yayılım problemi nedeniyle Quasi Poisson ve Conway-Maxwell-Poisson (COM) regresyon modelleri kullanılmış ve bu modeller veri seti kullanılarak birbiriyle karşılaştırılmıştır. Bu karşılaştırma sonucunda, COM Poisson regresyon modelinin diğer sayım regresyon modellerine göre üstünlüğü gerçek veri seti kullanılarak gösterilmiştir. COM-Poisson sonuçlarına göre kaçıncı gebeliği olduğu, ve doğum yöntemi (özellikle sezeryan) canlı doğan bebek sayısı üzerinde %99 güven düzeyinde ve hamilenin kan grubu %95 güven düzeyinde etkilidir. Ancak sezeryan ile doğum canlı doğan bebek sayısını negative olatak etkilemektedir. **Sonuç:** COM Poisson yardımıyla az-yayılım probleminin çözülmesiyle birlikte, kaçıncı gebeliği olduğu ve doğum yönteminin canlı doğan bebek sayısının en iyi tahmin edicileri olduğu görülmüştür. Buna ek olarak, kan grubunun istatistiksel olarak daha düşük anlamlılık düzeyinde, bir diğer açıklayıcı değişken olduğu tespit edilmiştir.

Anahtar Kelimeler: Çoklu gebelik; poisson regresyon; quasi poisson regresyon; az yayılım; COM poisson regresyon

Modeling count data could be used in many fields like economics, social sciences and biology etc. Besides, count regression models are useful and arise in many fields when the response variable is a count and the change of count is investigated based on some explanatory variables. Among many different types, classical Poisson regression is the most well-known method. However, it has the limited use in many various disciplines as empirical count data exhibits dispersion problem. Briefly, in many real-world applications, underlying assumption of equidispersion (i.e., an equal mean and variance) has been violated by over or under dispersed data. The excess or less variation results in inaccurate inferences on parameter estimates, standard errors, tests and confidence intervals.

As a result of the violation of equidispersion assumption in classical Poisson model, alternative statistical models are developed for over or under-dispersed data. The Negative Binomial model is one of the popular modeling choices for over-dispersed data.[1] Another alternative might be Quasi-Poisson regression which estimates parameters similar to Poisson regression but with a larger standard error.[2] It is possible to use in both over- and under-dispersion. The third alternative is called Conway-Maxwell-Poisson (COM-Poisson) regression, a useful flexible alternative for capturing both over- and under-dispersion.[3]

There are wide ranges of applications for all Poisson models in the literature. For instance, the predictors of strikes in Turkey between 1964-1998 are determined by using classical Poisson Regression model.[4] As another recent application, the number of divorces in Turkey between 2001- 2009 are modeled with Generalized Poisson, Quasi Poisson and Negative Binomial regressions.[5] Furthermore, identifying black points in traffic routes in Turkey to reduce the number of car accidents via Poisson, Negative Binomial and Empirical Bayesian approaches has been studied.[6]

Nowadays, multiple pregnancies occurred more frequently with being widespread of the assisted reproduction techniques. Besides, some researches showed that the possibility of multiple pregnancies has been increased by some factors such as twin pregnancy experience in the family, mothers at later ages, social properties and the number of live-born infants. As a main motivation of the study, different poisson models for the investigation of the number of live-born infants in multiple pregnancies are investigated with a real life data. In this study, the advantages and limitations of Quasi Poisson and recently Conway-Maxwell-Poisson (COM-Poisson) regression models are also considered for pregnancy data based on under-dispersion problem.

This article is organized as follows; Section 2 describes the count regression models including Poisson, Quasi-Poisson and COM-Poisson regression briefly. The data set on the number of live-born infants with its predictors is described in Section 3, where a sample of under-dispersed data. All mentioned count data models were applied to data set and the results were compared for the best model selection and discussed in section 4. Finally, the main conclusion of the study is summarized with future work in section 5.

MATERIAL AND METHODS

In this section, the count data models including Poisson, Quasi-Poisson and COM-Poisson regression used to analyze the data set on the number of live-born infants with its predictors are briefly given. All background information of used Poisson regression models in this study is summarized with their assumptions.

POISSON REGRESSION MODEL

Poisson model is the special case of Generalized Linear Models (GLM) and widely used count data model evaluated on Poisson distribution having the famous density function as;

$$f(y_i;\vartheta_i)=\frac{e^{-\vartheta_i}\vartheta^{y_i}}{y_i!}$$

where the conditional variance is equal to conditional mean, $\vartheta_i = E(y_i|x_i) = \exp(x_i{}^T\beta)$.

The log-likelihood of Poisson model is given as,

$$lnL(\beta)=\sum_{i=1}^{n} y_i x_i{}^T\beta - \exp(x_i{}^T\beta) - \ln(y_i!)$$

The regression coefficients are estimated by maximum likelihood method, derivation of log-likelihood relative to vector of coefficients β is set equal to zero[7]. Using iterative weighted least square (IWLS) algorithm, the estimated coefficients vector is obtained as,

$$\hat{\beta}=(X^TWX)^{-1}X^TW\tilde{y}$$

where $W=diag(\vartheta_i)$ and $\tilde{y}_i=\log(\vartheta_i)+\frac{y_i-\vartheta_i}{\vartheta_i}$. The main drawback of the Poisson model is the inequality of the mean and variance of the count data, which results in two different dispersion problems, called as over- dispersion and under-dispersion in the literature. In that case, both dispersion problems imply that Poisson Model is insufficient and it is required to improve the model with other alternatives.

QUASI-POISSON REGRESSION MODEL

For relaxing the assumption of equidispersion in Poisson model, Quasi-Poisson model can be considered as an alternative method. In Quasi-Poisson model, the distribution of dependent variable y denoted as $y_i = Poisson(\vartheta_i,\theta)$ with $E(y_i)=\vartheta_i$ and $Var(y_i)=\theta\vartheta_i$.

The main assumption of the Quasi-Poisson is about the variance of the model. To be more precise, $Var(y_i)=\theta\vartheta_i$ is defined as a linear function in terms of ϑ[8]. Using IWLS algorithm as in Poisson regression, the estimated coefficient vector is obtained as,

$$\hat{\beta}=(X^TWX)^{-1}X^TW\tilde{y}$$

where $W=diag(\frac{\vartheta_i}{\theta})$ and $\tilde{y}_i=\log(\vartheta_i)+\frac{y_i-\vartheta_i}{\vartheta_i}$.

CONWAY-MAXWELL-POISSON (COM-POISSON) REGRESSION MODEL

The Conway-Maxwell-Poisson model was introduced in 1962[9]; then only it was evaluated in the context of a GLM[10,11,12]. The corresponding distribution is generalization of the Poisson distribution with two parameters which makes the model flexible enough to describe a wide range of count data.

The COM-Poisson probability distribution function has the following form,

$$f(y,\vartheta,\omega)=\frac{\vartheta^y}{(y!)^\omega Z(\vartheta,\omega)},\ y=0,1,2,\ldots;\ \vartheta>0,\omega>0$$

where $Z(\vartheta,\omega)=\sum_{s=0}^{\infty}{\vartheta^s}/{(s!)^\omega}$ is normalizing constant and ω is dispersion parameter such that $\omega>1$ means under-dispersion and $\omega<1$ means over-dispersion. Taking a GLM approach, Sellers and Shmueli (2010) proposed a COM-Poisson regression model using the link function as[12]

$$logE(y)=X^T\beta=\beta_0+\sum_{j=1}^{p}\beta_j x_j$$

The parameter estimates could be obtained using maximum likelihood method. Maximum likelihood estimates are easily derived for the COM-Poisson distribution due to membership of the exponential family. The log-likelihood function can be written as,

$$\log \mathrm{L}(y_1, \ldots, y_n; \vartheta, \omega) = \log(\vartheta) \sum_{i=1}^{n} y_i - \omega \sum_{i=1}^{n} \log(y_i!) - n log Z(\vartheta, \omega)$$

and the maximum likelihood estimation can thus be achieved by iteratively solving the set of normal equations.[13]

3. NUMERICAL APPLICATION

The main goal is detecting the factors affecting the number of live-born infants using count data models so the number of live-born infants is selected as the response variable. Classical Poisson, Quasi-Poisson and COM-Poisson regression models are investigated to analyze the data and results are compared.

For this study, the multiple pregnancy diagnose data with many explanatory variables are used for the year 2015. Originally, the mentioned data set is retrieved from hospitals of the Ministry of Health in Ankara province, Turkey. The data set with all relevant variables is presented below in Table 1.

TABLE 1: The variables used in the study.

Explanatory Variables	Factor Levels
Age	19-53
Blood type	$Rh^{0,A,B,AB}$
The number of pregnancy	1-4 *
The method of delivery Previous births summary	1: Normal 2: Cesarean 3: Vacuum
Previous birth summary	1: Birth 2: Abortion

*For the total number of pregnancy data, the number of observations has been reduced from 1223 to 1170, since having 5 or more pregnancy diagnose has been assumed as extreme values in the study.

The descriptive statistics are tabulated for only numerical variables among all interested ones in Table 2 to understand the main properties of them at the first step. Besides, blood type, the method of delivery and previous birth summary are categorical variables used in this study, having 7, 3 and 2 levels, respectively as they defined in Table 1. To sum up, the most observed method of delivery is cesarean type (1023 observations) among the others for multiple pregnancy diagnoses. Moreover, the dominating blood types are 0+ and A+, multiple pregnancy data includes 828 abortions for previous birth summary of diagnoses.

TABLE 2: The descriptive statistics for numerical variables.

	Variables	
Statistics	***Age***	***The number of Pregnancy***
Minimum	19	1
Maximum	53	4
Median	32	2
Mean	32.3179	1.9128
Variance	30.2307	0.9060
Standard Deviation	5.4982	0.9518

As a pre-process of data set, the number of observations has been reduced from 1223 to 1170 by considering the number of pregnancy is 5 or more as extreme values in the study. All calculations in this study were derived using RStudio-1.0.143 with packages AER and COMPoissonReg.[14,15]

The bar chart of the number of live-born infants in the data set is employed for the first inspection. As Figure 1 suggests most of the multiple pregnancies resulted in having twins.

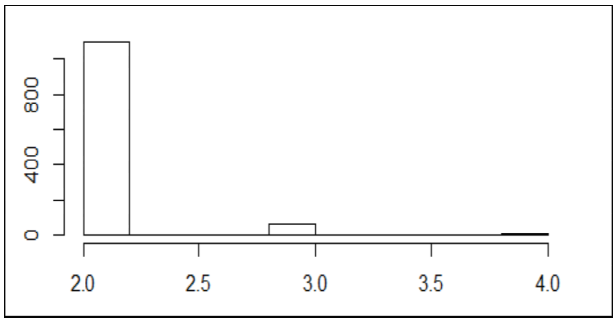

FIGURE 1. Distribution of the number of live-born infants in the multiple pregnancies.

RESULTS AND DISCUSSION

All above mentioned count regression models are considered to detect the significant predictor or predictors of the dependent variable and their effects. The best model is chosen based on both Akaike Information Criteria (AIC) and then p values and all results of three different count regression models are presented in Table 3.

The reason of implementation of the Quasi- and COM-Poisson models directly depends on the dispersion test applied for classical Poisson model. In this respect, it is assessed the hypothesis that (equidispersion) assumption holds against the alternative one where the variance is of the form,

$Var(y) = \mu + \alpha * tr(\mu)$

the coefficient α is estimated by an auxiliary OLS regression and tr is transformation function and common specifications of that are $tr(\mu) = \mu^2$ or $tr(\mu) = \mu$. Based on dispersion test mentioned above, $\alpha = -0.9715843 < 0$ was observed which represents the under-dispersion problem. Besides, coefficient estimates and their p-values do not support any explanatory variable for being a suitable predictor for the dependent variable because of insufficient Poisson model result in Table 3.

When both Quasi- and COM-Poisson regressions were proceeded, Table 3 exhibits that among all explanatory variables, number of pregnant (NumPreg) and method of delivery (specifically cesarean type) has significant impact on the number of live-born infants at %99 significance level (since their p-values are less than 0.001). Specifically, cesarean delivery has negative impact on the number of live-born infants, ie. whenever it increases 1 unit, which reduces the response variable with –1.9924 unit in COM-Poisson model. On the other hand, NumPreg has positive impact on the number of live-born infants in the multiple pregnancies as expected. To illustrate, one can say that 1 unit increase on the number of pregnancy has resulted in 0.68 unit increase on having probability of multiple pregnancy diagnose.

Besides supporting the result of Quasi-Poisson model, COM-Poisson shows that possible effects of blood type of pregnant in Table 3 at %95 significance level. Moreover, ν value, the main indicator of existence of under-

dispersion, is significant so fitting COM-Poisson to count data in this case is meaningful. After analyzing the results of modeling, it is considered goodness-of-fit evaluation and comparison between models. For the model selection, AIC values are evaluated. It is a very simple way to compare models with different numbers of parameters. When comparing models, the smaller the AIC, the better the fit.[16] Based on AIC values of Poisson and COM-Poisson models, the latter one fits better since,

$AIC_{COM-Poisson} < AIC_{Poisson}$

Therefore it is concluded that COM-Poisson regression model is better than classical Poisson and Quasi-Poisson models for multiple pregnancies data. In other words, COM-Poisson model is the best model fit for determination of the affecting factors on the number of babies born alive of multiple pregnancies data.

TABLE 3: Parameter estimates, standard error and AIC value of each model.

Regressors	Poisson		Quasi-Poisson		COM-Poisson	
	Estimate	StdError	Estimate	StdError	Estimate	StdError
	(Pr(> |z|))		(Pr(>|z|))		(Pr(>|z|))	
(Intercept)	0.7738	0.1766	0.7738	0.0299	19.3289	1.5162
	(1.18e-05***)		(2e-16***)		(3.201e-37***)	
Age	-0.0002	0.0038	-0.0002	0.0006	0.0216	0.0234
	(0.9680)		(0.813)		(0.3567)	
NumPreg	0.0207	0.0245	0.0207	0.0042	0.6799	0.1351
	(0.3986)		(7.42e-07***)		(4.852e-07***)	
cesarean	-0.1020	0.0616	-0.1020	0.0105	-1.9924	0.2624
	(0.0981.)		(2e-16***)		(3.093e-14***)	
Vacuum /Forceps	-0.1028	0.7109	-0.1028	0.1205	-0.1452	3.1968
	(0.8850)		(0.3940)		(0.9638)	
0 RH +	0.0088	0.1169	0.0088	0.0198	2.9820	0.9345
	(0.9400)		(0.6570)		(0.001418*)	
A RH -	0.0034	0.1433	0.0034	0.0243	2.3609	1.0910
	(0.9813)		(0.8900)		(0.03047*)	
A RH +	0.0079	0.1146	0.0079	0.0194	2.8824	0.9236
	(0.9452)		(0.6850)		(0.001804*)	
AB RH -	-0.0004	0.3095	-0.0004	0.0525	0.2474	2.0155
	(0.9989)		(0.9940)		(0.9023)	
AB RH +	0.0027	0.1366	0.0027	0.0232	2.9720	1.0197
	(0.9844)		(0.9080)		(0.003563*)	
B RH -	0.0072	0.1868	0.0072	0.0317	2.0368	1.3368
	(0.9692)		(0.8200)		(0.1276)	
B RH +	-0.0009	0.1233	-0.0009	0.0209	2.7101	0.9586
	(0.9944)		(0.9670)		(0.004699**)	
abortion	-0.0008	0.0505	-0.0008	0.0086	-0.2449	0.2631
	(0.9870)		(0.9230)		(0.3519)	
AIC value	3142.3		-		568.1178	

Statistical Significance Levels: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1.

CONCLUSION

The objective of this study is to provide tenable guidance on how to find an appropriate model for multiple pregnancies data. Due to under-dispersion, COM-Poisson model is also considered to analyze the data. The possible predictors of the number of live-born infants were investigated by using

different Poisson Models based on the real life data to compare with COM-Poisson model. For this reason, Classical Poisson, Quasi-Poisson and COM-Poisson regression models were compared in terms of their coefficient estimations and significance level with AIC values. Since Classical Poisson regression exhibited the under-dispersion so that Quasi-Poisson and COM-Poisson models were considered alternatively. Both Quasi- and COM-Poisson model fit better to the count data and identify the significant predictor for the response variable. Accordingly, the number of live-born infants in the multiple pregnancies is affected by the number of pregnancy and the cesarean delivery. In addition to that, blood types might be important at some lower significance level under the COM-Poisson model, selected based on AIC value.

To sum up, the intent of this study is to encourage new ideas about how to approach the modeling of multiple pregnancies data. In multiple pregnancies, number of pregnancy and cesarean delivery are important factors which affect the live-born infants positively and negatively, respectively based on COM-Poisson model. For the improvement of the study, different techniques might be alternative tools for under-dispersion count data like Extended Poisson Process Models (EPPM). Moreover, other models can be considered based on the observed important predictors for the number of live-born infants.

Acknowledgements

The authors would like to thank two anonymous reviwers for their valuable comments and suggestions to improve and clarify this manuscript.

Conflict of Interest

Authors declared no conflict of interest or financial support.

Authorship Contributions

Idea/Concept: *Gizem Erkan, Ozan Evkaya, Semra Türkan;* ***Design:*** *Gizem Erkan, Ozan Evkaya, Semra Türkan;* ***Control/Supervision:*** *Gizem Erkan, Ozan Evkaya, Semra Türkan;* ***Data Collection and/or Processing:*** *Gizem Erkan, Ozan Evkaya, Semra Türkan;* ***Analysis and/or Interpretation:*** *Gizem Erkan, Ozan Evkaya, Semra Türkan;* ***Literature Review:*** *Gizem Erkan, Ozan Evkaya, Semra Türkan;* ***Writing the Article:*** *Gizem Erkan, Ozan Evkaya, Semra Türkan;* ***Critical Review:*** *Gizem Erkan, Ozan Evkaya, Semra Türkan.*

REFERENCES

1. Hilbe JM. Negative Binomial Regression. 2nd ed. London: Cambridge University Press; 2011. p.541.
2. Ismail N, Jemain AA. Handling overdispersion with negative binomial and generalized poisson regression models. Casualty Actuarial Society Forum Winter 2007;103-58.
3. Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P. A useful distribution for fitting discrete data: revival of the Conway-maxwell-poisson distribution. Journal of the Royal Statistical Society Series C (Applied Statistics) 2005;54(1):127-42.
4. Sahin H. [Poisson regression application: the determinants of strikes in Turkey 1964- 1998]. Dogus Universitesi Dergisi 2002;(5):173-80.
5. Pamukcu E, Colak C, Halisdemir N. [Modeling of the number of divorce in Turkey using the generalized poisson, quasi-poisson and negative binomial regression]. Turkish Journal of Science and Technology 2014;9(1):89-96.
6. Dereli MA, Erdoğan S, Soysal OM, Çabuk A, Uysal M, Tiryakioğlu İ, et al. [Determination of traffic accident black spot based on geographical information system: the empirical bayes application]. Electronic Journal of Map Technologies 2015;7(2):36-42.
7. Altun E, Turkan S. [Zero-inflated upper truncated poisson regression model: an application to miscarriage data]. International Journal of Mathematics & Computation 2017;28(1):17-24.

8. Ver Hoef JM, Boveng PL. Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data?]. Ecology 2007;88(11):2766-72.
9. Conway RW, Maxwell WLA. Queuing model with state dependent service rates. Journal of Industrial Engineering 1962;12:132-6.
10. Zou Y, Geedipally SR, Lord D. Evaluating the double poisson generalized linear model. Accid Anal Prev 2013;59:497-505.
11. Guikema SD, Coffelt JP. A flexible count data regression model for risk analysis. Risk Anal 2008;28(1):213-23.
12. Sellers KF, Shmueli G. A flexible regression model for count data. Ann Appl Stat 2010;4(2):943-61.
13. Sellers KF, Borle S, Shmueli G. The COM-poisson model for count data: a survey of methods and applications. Appl Stoch Models Bus Ind 2012;28(2):104-16.
14. Christian K, Achim Z. Applied Econometrics with R. 1st ed. New York: Springer-Verlag; 2008. p.222.
15. Kimberly S, Thomas L, Andrew R. COMPoissonReg: Conway- Maxwell Poisson (COM-Poisson) Regression. R package version 0.4.1. 2017;15.
16. Zaihra T. Modeling the number of research papers produced by graduate students using zero-inflated models. CS-BIGS 2012;5(1):44-50.