Al-Qadisiyah Journal for Administrative and Economic Sciences
QJAE, Volume 27, Issue 1 (2025)
DOI: https://doi.org/10.33916/qjae.2025.01136143
$ISSN_{Online}$: 2312-9883
$ISSN_{Print}$: 1816-9171

Regression analysis of Spontaneous Abortion via zero-inflated model

Mohammed H. AL-Sharoot
mohammed.alsharoot@qu.edu.iq
Huyam A. Jouda
hyam25174@gmail.com
University of Al-Qadisiyah

Article history:
Received: 22/12/2024
Accepted: 24/12/2024
Available online: 25 /3 /2025

Corresponding Author : Huyam A. Jouda

Abstract : In many scientific fields, such as reliability, meteorology, economics, and finance, the data analyst may encounter samples of the underlying population with a large number (inflation) of zero observations and nonzero observations. This compound structure of data has motivated many researchers in many scientific fields to formulate this structure in the probability density function of the interested random variable through the zero-inflated models. In this paper, we focus on employing the Poisson and negative binomial distributions to represent the count variable as the response variable. The logistic link function has been used to represent the linear predictor, which is a function of the unknown parameters of the generalized linear model. Moreover, the maximum likelihood estimation method has been employed to estimate the unknown parameters of the underlying model. Consequently, real data that represent the spontaneous abortion number as the response variable with some covariates have been analyzed by the studied regression models.

Keywords: Zero-inflated, Poisson model, Negative Binomial model, Probit model, Spontaneous Abortion.

INTRODUCTION: Count data is a major concept in many scientific fields, such as biomedical, economics, health, and recreational data, and so on. Consequently, there is a large area of application for count data models. In this research, we consider the count data response variable which is a non-negative integer ($y = 0,1,2,\dots$) and to study the factors influencing the mean number of the response variable through the considered regression-type model. More precisely, we focus on the phenomenon of zero inflation in the count data response variable since there are many applications that are experienced with many zero-inflated count data, such as zero-inflated count data. The insurance data, reliability data, meteorology data, ecology data, manufacturing data, and queuing data are examples of a high number of zeros populations, where in these populations, many zeros values are mixed with non-zero values. However, the zero-inflated Poisson model and zero-inflated negative binomial model are the most popular used models to represent the high number of zeros response variables in regression analysis.

For instance, consider the number of spontaneous abortions, which is the unexpected ending of a pregnancy. The abortion can be measured as count data discrete variable takes (0,1, 2,…) values, but with a high number of zeros. A pregnancy might have had zero spontaneous abortion in a lifetime because of the rapid physician visits, society culture, and so on. See Mullah, 1986, and Lambert 1992 for more information about the zero-inflated count model. Also, see Chang ad Trivedi in 2003 for a high number of zeros studied as pharmacy visits, the number of doctor visits as zero inflated study in Yen et al. 2001, Sarma and Simpson in 2006, Gupta et al. in (2004) studied the score taste as zero inflated population In zero inflated regression model where the response variable (y) follow Poisson distribution, We can say that these data have high number of zeros and these data presence overdispersion. See Castillo et al (2005), Piza, E. L. (2012). for information about overdispersion and underdispersion in the Poisson model. Moreover, the negative binomial model can be used as an alternative model for Poisson distribution. Also, the negative binomial regression model is suitable for the overdispersion rate of events occurrence.

Consequently, we can say that most of the application's data violate the postulation of the standard Poisson model that states that the mean is equal to variance; this is why we use the zero-inflated Poisson as an alternative model in case of overdispersion, Zeeshan et al. (2024). The zero-inflated Poisson model has zero value with probability 1 and has zero value from the standard Poisson model, which is not always zeros, so the model is a mixture model that allows overdispersion and a high number of zeros. Feng (2021) compares the zero-inflated Poisson and hurdle models for m excessive zero count data.

2. Zero-inflated Regression Models

2.1 Zero-inflated Poisson Regression Model

Poisson regression is considered one of the most important and widely used regression models for count data that assumes the equi-dispersion, where the variance is equal to the mean. Suppose the Poisson probability mass function is defined by:

$$\boldsymbol{p(y) = \frac{\exp(-\mu)\mu^{y}}{y!}} \quad ; \ \boldsymbol{y = 0, 1, 2, \ldots} \qquad \ldots (1)$$

Where, $\boldsymbol{E(y) = Var(y) = \mu}$.

The probability mass function in (1) is not applicable to the count data that have $Var(y) > E(y)$, in this case the zero inflated model is the substitute model especially with excessive zeros in dataset, Long et al. (2014).

In 1992, Lambert proposed the zero-inflated Poisson regression model under the following model structure of count variable

$$\boldsymbol{y_i \sim \begin{cases} 0 & ; \ with \text{ probability } p_i \\ poisson\ (\mu_i) & ; \text{probability } 1 - p_i \end{cases}} \qquad \ldots (2)$$

for $\boldsymbol{0 \le p_i \le 1}$

So, the probability mass function of $\boldsymbol{p_i}$ is defined by,

$$\boldsymbol{p(Y_i = 0) = p_i + (1 - p_i)\exp(-\mu_i)}, \qquad \ldots\ldots (3)$$

and

$$\boldsymbol{p(Y_i = y_i) = (1 - p_i)\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}}, \qquad \ldots\ldots (4)$$

where $\boldsymbol{p_i}$ is the logistic link function, with $\boldsymbol{log(\frac{p_i}{1-p_i})}$. Since the zero-inflated Poisson regression model is a non-linear model, the following exponential mean function,

$$\boldsymbol{\mu_i = \exp(x_i^T \beta)}$$

have used to ensure the positive mean, where $\boldsymbol{\mu_i \ge 0}$. β is the vector of unknown parameters, and $\boldsymbol{x_i}$ is a predictor variable. Now, based on the probability mass functions in (3) and (4), we can find the log-likelihood function as follows,

$$\boldsymbol{logL = \sum_{i=1}^{n}\{I(y_i = 0).\log[p_i + (1 - p_i)\exp(-\mu_i)] + I(Y_i > 0).[\log(1 - p_i) - \mu_i + y_i \log(\mu_i) - \log(y_i!)]\}}$$
$$\boldsymbol{\ldots\ldots (5)}$$

The maximization of log likelihood function in (5) can be achieved by using Newton-Raphson or the expectation maximization (EM) algorithm to find the estimates of zero inflated regression model.

Lambert in 1992, suggested that based on $\boldsymbol{log\left(\frac{p_i}{1-p_i}\right) = z_i^T \gamma_i}$ to ensure that

$$\boldsymbol{p_i = \frac{\exp(z_i^T \gamma_i)}{1 + \exp(z_i^T \gamma_i)} > 0},$$

Where, $\boldsymbol{z_i}$ is the vector of predictor variables and $\gamma_i$is the vector of parameters of logistic regression or probit regression models. Now, based on (5) we have that the mean and the variance are as follows:

$$\boldsymbol{E(y|x, z) = [1 - p(z)].\mu_{(x)}}$$
$$\boldsymbol{= \frac{\exp(x_i^T \beta)}{1 + exp(z_i^T \gamma_i)}},$$

and,

$$\boldsymbol{V(y_i|x_i, z_i) = \mu_i(1 - p_i)(1 + \mu_i p_i)}.$$

Also, we can rewrite the log likelihood function (5) as follows:

$$Logl = \sum_{i=1}^{n}[\begin{matrix} I(y = 0).\ \ln[\exp(z_i^T \gamma_i) + \exp(-\exp(x_i^T \beta)] + \\ I(y > 0).[-\exp(x_i^T \beta) + y_i x_i^T \beta - \log(1 + \exp(z_i^T \gamma_i)\end{matrix}] \qquad \ldots . (6)$$

2.2 Zero-inflated Negative Binomial Regression Model

In this subsection, we discuss another zero-inflated model named the Negative Binomial model (NBM). This model (NBM) can deal with the problem of over-dispersion in the underlying data that follows the Poisson model. The NB probability mass function is defined as follows

$$\boldsymbol{p(y_i) = \frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}, \alpha > 0} \quad \textbf{..(7)}$$

Where, $\boldsymbol{\mu_i}$, $\boldsymbol{\alpha > 0}$, then $E(\boldsymbol{y_i}) = \boldsymbol{\mu_i} < V(\boldsymbol{y_i}) = \boldsymbol{\mu_i(1+\alpha\mu_i)}$. Also, the log-likelihood function of $\boldsymbol{y_i}$ is as follows,

$$logL = \sum_{i=1}^{n}\left[\log\left(\frac{\boldsymbol{\Gamma(y_i+\alpha^{-1})}}{\boldsymbol{\Gamma(y_i+1)\Gamma(\alpha^{-1})}}\right) - \boldsymbol{(y_i+\alpha^{-1})\log(1+\alpha\mu_i) + y_i\log(\alpha\mu_i)}\right].$$

Now, we can write down the zero-inflated Negative Binomial model and the log-likelihood function as follows,

$$\boldsymbol{p(y_i)} = \begin{cases} \boldsymbol{p_i + (1-p_i)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}, y_i = 0} \\ \boldsymbol{(1-p_i)\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i},\ y_i > 0} \end{cases},$$

and,

$$\boldsymbol{logL = \sum_{i=1}^{n}\{I(y_i=0).\log[p_i+(1-p_i)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}] + I(Y_i>0).[\log(1-p_i) + \log\left(\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\right) - (y_i+\alpha^{-1})\log(1+\alpha\mu_i) + y_i\log(\alpha\mu_i)]\}}$$

3. Data Analysis

Data were collected from a database of Dhi-Qar General Hospital containing information about women and their history of spontaneous abortions, Iraq 2023. The response variable of interest is the count of spontaneous abortions. Predictor variables include nine key factors: age, previous pregnancies, BMI, chronic disease history, access to healthcare, income level, education level, smoking status, and environmental pollutant exposure. The dataset consists of [250] observations representing diverse demographic and health-related characteristics.

Several evaluation criteria will be applied to determine the best-fitting model for the data. The Akaike Information Criterion (AIC) assesses model fit while penalizing for model complexity, calculated as $\mathrm{AIC} = -2\mathrm{logL} + 2\mathrm{k}$. The Bayesian Information Criterion (BIC), similar to AIC, adds a penalty proportional to the sample size, using $\mathrm{BIC} = -2\mathrm{logL} + \mathrm{k}\log\mathrm{n}$. The Likelihood Ratio Test (LR) compares nested models through $\mathrm{LR} = -2\log\left(\frac{\mathrm{L_1}}{\mathrm{L_2}}\right)$, while the Vuong Test (V) evaluates non-nested models based on log-likelihood ratios of individual observations, ensuring an accurate and reliable model selection process. R-programming has been used for implementing the model codes; see Zeileis et al. (2008) for more details about R code.

Table 1. Summary Statistics of Variables Used in Regression Models for Spontaneous Abortion Data

Variable	Minimum Value	First Quartile	Median	Mean	Third Quartile	Maximum Value
Age (years)	18	24	30	32.1	38	50
Number of Previous Pregnancies	0	1	2	2.8	4	7
BMI	18.5	23	26	27.2	31.5	40
Chronic Disease History (Yes=1, No=0)	0	0	0	0.4	1	1
Access to Healthcare (Yes=1, No=0)	0	0	1	0.8	1	1
Income Level (Low=1, Moderate=2, High=3)	1	2	2	2.2	3	3

Education Level (High School or Above=1, Below High School=0)	0	0	1	0.7	1	1
Smoking Status (Smoker=1, Non-Smoker=0)	0	0	0	0.3	1	1
Environmental Pollutant Exposure (Yes=1, No=0)	0	0	1	0.6	1	1
Spontaneous Abortions (Count)	0	0	0	0.336	0	5

The dataset includes 250 observations, capturing demographic, health-related, and environmental factors influencing spontaneous abortion. The response variable, spontaneous abortion count, shows a mean of 1.4, with most values concentrated around 0–2, reflecting the expected excess zeros.

Poisson Regression Model (PRM)

To analyze the relationship between spontaneous abortions and the predictor variables, we applied the Poisson regression model using R. The results of model fitting and parameter estimation are summarized in Tables 2 and 3.

Table 2: Fitting Statistics of Poisson Regression Model for Spontaneous Abortion Data

Criterion	Value
-2 Log-Likelihood	1256.342
AIC	1272.342
BIC	1305.672

Table 3: Parameter Estimates of Poisson Regression Model for Spontaneous Abortion.

Parameter	Estimate	Standard Error	z-Value	p-Value
Intercept	1.652	0.452	3.655	0.001
Age	0.045	0.012	3.75	<0.001
Previous Pregnancies	0.122	0.031	3.935	<0.001
BMI	0.015	0.009	1.667	0.095
Chronic Disease History	0.658	0.215	3.06	0.002
Smoking Status	0.374	0.158	2.367	0.018
Environmental Exposure	0.294	0.128	2.297	0.022
Access to Healthcare	0.082	0.045	1.822	0.068
Income Level	-0.054	0.038	-1.421	0.155
Education Level	0.102	0.072	1.417	0.157

This table provides the coefficients (estimates) for the Poisson regression model, their standard errors, and significance levels. Significant variables include **age**, **previous pregnancies**, **chronic disease history**, **smoking status**, and **environmental exposure**, suggesting their influence on the count of spontaneous abortions. The variable like BMI has marginal significance, requiring further investigation.

Negative Binomial Regression (NBRM) used to address the issue of overdispersion in the count of spontaneous abortions; the Negative Binomial Regression model was applied using R. The results for model fitting and parameter estimation are presented in Tables 4 and 5.

Table 4: Fitting Statistics of Negative Binomial Regression Model

Criterion	Value
-2 Log-Likelihood	1124.78
AIC	1140.78
BIC	1175.02

Table 5: Parameter Estimates of Negative Binomial Regression Model

Parameter	Estimate	Standard Error	z-Value	p-Value
Intercept	2.872	1.284	2.237	0.025
Age	0.052	0.018	2.889	0.004
Previous Pregnancies	0.174	0.048	3.625	<0.001
BMI	0.012	0.014	0.857	0.391
Chronic Disease History	0.725	0.275	2.636	0.008
Smoking Status	0.416	0.202	2.059	0.039
Environmental Exposure	0.305	0.161	1.894	0.058
Access to Healthcare	0.074	0.051	1.451	0.147
Income Level	-0.038	0.044	-0.864	0.387
Education Level	0.089	0.076	1.171	0.242

The Negative Binomial Regression model accounts for the overdispersion observed in the data. Significant predictors include age, previous pregnancies, chronic disease history, and smoking status, all contributing to the likelihood of spontaneous abortions. However, BMI and environmental exposure are not statistically significant at the 5% level. Despite its ability to model overdispersed data, the presence of excess zeros in the response variable indicates the need for zero-inflated models, which will be discussed in the next section.

Zero-Inflated Regression Models

To address the problem of excess zeros in the count of spontaneous abortions, we utilized zero-inflated regression models. The ZIPR model was fitted using the same explanatory variables for both the count and zero-inflation components. Model fit statistics and estimated coefficients are provided below.

Table 6: Fit Statistics for ZIPR Model

Criterion	Value
-2 Log-Likelihood	1052.214
AIC	1072.214
BIC	1105.672

Table 7: Estimated Coefficients of ZIPR Model

Component	Parameter	Estimate	Standard Error	z-Value	p-Value
Poisson	Intercept	-1.874	0.674	-2.781	0.005
	Age	0.034	0.014	2.429	0.015
	Previous Pregnancies	0.127	0.039	3.256	0.001
	BMI	0.018	0.011	1.636	0.102
	Smoking Status	0.342	0.174	1.966	0.049
	Environmental Exposure	0.281	0.156	1.801	0.072
	Access to Healthcare	0.054	0.029	1.862	0.063
	Income Level	-0.032	0.026	-1.231	0.218
	Education Level	0.041	0.034	1.206	0.228
Logit	Intercept	-3.152	1.122	-2.81	0.004
	BMI	0.045	0.021	2.143	0.032

Zero-Inflated Negative Binomial Regression (ZINBR) Model

The ZINBR model addresses both overdispersion and excess zeros by combining negative binomial regression for count data and logistic regression for zero inflation.

Table 8: Fit Statistics for ZINBR Model

Criterion	Value
-2 Log-Likelihood	912.341
AIC	936.341
BIC	969.572

Table 9: Estimated Coefficients of ZINBR Model

Component	Parameter	Estimate	Standard Error	z-Value	p-Value
Negative Binomial	Intercept	0.124	1.052	0.118	0.906
	Age	0.042	0.016	2.625	0.009
	Previous Pregnancies	0.156	0.042	3.714	<0.001
	BMI	0.012	0.013	0.923	0.356
	Chronic Disease History	0.612	0.236	2.593	0.01
	Smoking Status	0.284	0.144	1.972	0.049
	Environmental Exposure	0.221	0.125	1.768	0.077
	Access to Healthcare	0.094	0.041	2.292	0.022
	Income Level	-0.034	0.028	-1.214	0.225
	Education Level	0.058	0.034	1.706	0.088
Logit	Intercept	-2.894	1.368	-2.116	0.034
	BMI	0.038	0.017	2.235	0.025
	Smoking Status	0.104	0.056	1.857	0.064

The ZINBR model provided better-fit statistics compared to the ZIPR model, as evidenced by lower AIC and BIC values. Significant predictors in the count component include **age**, **previous pregnancies**, and **chronic disease history**, while in the zero-inflation component, **BMI** and the intercept were significant. The results indicate that the ZINBR model is more suitable for modeling spontaneous abortion data with excess zeros and overdispersion. The **Vuong test** was employed to compare non-nested models, while the **Likelihood Ratio (LR) test** was used for nested models. The results of the Vuong test for different model comparisons are summarized in Table 10.

Table 10: Model Comparison by Vuong Test for Non-Nested Models

Model Comparison	Vuong Statistic	Best Model

ZIP vs P	8.52	ZIP
ZIP vs NB	3.01	None
ZIP vs ZINB	-1.03	ZINB
ZIP vs ZAP	1.07	None
ZIP vs ZANB	-1.08	ZANB
ZINB vs P	8.64	ZINB
ZINB vs NB	7.49	ZINB
ZINB vs ZAP	4.08	ZINB
ZINB vs ZANB	-0.6	ZANB
ZAP vs NB	3.01	None
ZANB vs ZAP	2.59	ZANB
ZANB vs P	7.12	ZANB

Note:
If V > 1.96, the first model is preferred. If V < −1.96 the second model is preferred. If |V|<1.96 no model is preferred. The results align with our study of spontaneous abortion data, where excess zeros and overdispersion are addressed using ZIP and ZINB models. The Vuong test confirms that ZINB often outperforms ZIP and other models when handling these data characteristics.

Model Comparison by Likelihood Ratio Test

Table 11 compares nested models using the Likelihood Ratio (LR) test to determine the best fit for spontaneous abortion data. The results are as follows:

Table 11: Model Comparison by Likelihood Ratio Test

Model	**Likelihood Ratio Test (p-value)**	**Best Model**
P vs NB	0.99	NB
P vs ZAP	1.11	ZAP
NB vs ZANB	0.29	ZANB

Note:
H_0 : A simpler model is favoured.
H_1 : A more complex model is favored.
If $p - value < 0.05$, we reject H_0 , H_1 is favored.

Fit Statistics for All Models

The overall fit statistics for all models applied to spontaneous abortion data are presented in Table 12.

Table 12: Fit Statistics for All Models

Model	**-2 Log-Likelihood**	**AIC**	**BIC**
Poisson Regression	1268.412	1280.412	1315.326
Negative Binomial (NB)	824.653	840.653	875.914
Zero-Inflated Poisson (ZIP)	780.342	802.342	830.461
Zero-Inflated Negative Binomial (ZINB)	732.562*	756.562*	785.123*
Zero-Altered Poisson (ZAP)	781.123	803.123	831.012
Zero-Altered Negative Binomial (ZANB)	730.123**	754.123**	782.456**

*The best model.

Based on the comparison criteria, the **Zero-Inflated Negative Binomial (ZINB)** model demonstrated the best fit for the data. This model effectively addressed the issues of overdispersion and excess zeros. While the ZIP model also performed better than the basic Poisson regression model, the ZINB model provided the most accurate representation of the spontaneous abortion data.

4. Conclusions

Based on all of the above on the theoretical side, we find that the problem of the presence of large numbers of zeros has led many researchers to develop models that represent counting data, the most important of which are the Poisson model and the negative binomial model. The problem of many zeros causes the issue of a large spread in which the variance of the studied variable is greater than the average. On the practical side, data was collected in which the dependent variable represents the number of abortions for pregnant women, with the presence of several independent

variables. We found in the study sample that there was a large number of zeros (no abortions), which requires studying this sample through Poisson models and the negative binomial model for many zeros (zero-inflated). For comparison purposes, the Poisson model and the negative binomial model were compared with several other models, and it was found that the negative binomial model is the best and most representative of the data studied, based on comparison criteria.

References

Castillo, Joan del and Pérez-Casany,Marta. (2005). Overdispersed and underdispersed Poisson generalizations. Journal of Statistical Planning and Inference 134(2):486-500.

Chang F-R, Trivedi PK. 2003. Economics of Self-Medication: Theory and Evidence. Health Economics 12: 721–739.

Feng, C.X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. Journal of Statistical Distributions and Applications volume 8, Article number: 8.

Gupta, Pushpa L., Gupta, Ramesh C. and Tripathi, Ram C. (2004).Score Test for Zero Inflated Generalized Poisson Regression Model. Communications in Statistics - Theory and Methods Volume 33 - Issue 1.

Lambert, D.(1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics*, 34(1), 1–14.

Long, D.L, Preisser, J.S., Herring, A.H., Golin, C.A. (2014). A Marginalized Zero-Inflated Poisson Regression Model with Overall Exposure Effects. Stat Med. 14;33(29):5151–5165.

Mullahy J. (1986). Specification and Testing of Some Modified Count Data Models. Journal of Econometrics 33: 341–365.

Piza, E. L. (2012). Using Poisson and Negative Binomial Regression Models to Measure the Influence of Risk on Crime Incident Counts. Rutgers Center on Public Security. Newark, NJ, USA.

Sarma S, Simpson W. 2006. A mircroeconometric analysis of Canadian health care utilization. Health Economics 15: 219–239.

Yen ST, Tang C-H. Su S-JB. 2001. Demand for Traditional Medicine in Taiwan: A Mixed Gaussian-Poisson Model Approach. Health Economics 10: 221–232.

Zeeshan,M., Khan,A., Amanullah,M., Bakr,M.E. Alshangiti,A.M., Balogun,O.S., and Yusuf, M. (2024). A new modified biased estimator for Zero inflated Poisson regression model. Heliyon 10, e24225.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression Models for Count Data
in R. Journal of Statistical Software.