

Zero-Inflated Poisson Regression Modeling of Plant Protein Consumption

I. J. David¹, P. O. Ikwuoché¹, R. L. Kolo²

¹Department of Mathematics and Statistics, Federal University Wukari, Nigeria,
e-mail: davidij@fuwukari.edu.ng

²Department of Biological Science, Federal University, Gashua, Nigeria

SUMMARY

This research fitted a discrete distribution for modeling count data. Specifically, Zero-Inflated Poisson (ZIP) regression was used to model plant protein consumption by 400 randomly sampled individuals in Wukari. The data was collected by questionnaire. The ZIP regression model was used based on its ability to model data with excess zeros present in the collected data. Variables considered and used for the analysis are Age, Body Mass Index, Blood Pressure, Occupation, Gender, Weight, Height, Body Reaction, and Consumption Class. The parameters of the ZIP model were estimated using the maximum likelihood estimation technique. The model was tested for Goodness of Fit (GoF) using deviance, scaled deviance, Pearson- χ^2 and scaled Pearson- χ^2 statistics. The results obtained showed that Age, Gender, and Reaction were significant at 5%, and the GoF tests revealed that the Zero-Inflated Poisson regression produces a good fit and is a good model for overcoming the overdispersion effect.

Key words: Zero-inflated; Poisson regression model; Protein; Maximum likelihood estimation; Goodness of fit

1. Introduction

In many studies count data may possess an excess amount of zeros. Data with a large proportion of zero counts, also called zero-inflated (ZI) data, appear quite frequently in various fields of research including health research, agricultural research, ecology, manufacturing, econometrics and others (Ridout et al., 1998). In health research, zero-inflated data are often found when examining counts of plant protein consumed by individuals, where some individuals consume little or no plant protein. To handle the count data, Lambert (1992) discussed this issue

and proposed a zero-inflated Poisson (ZIP) regression model, with an application to defects in manufacturing. In this context, the distribution of zero and positive counts in the data set which is being investigated is very important. Overdispersion (greater variability) implies that there is more variability around the model's fitted values than consistency (Zafakali and Ahmad, 2013). The Poisson distribution is the basis model that is commonly applied to study count data, but the equivalence of mean and variance assumption for the Poisson process is invalid for many processes, such as in microbiological data, because in most cases there are no occurrences of micro-organisms (total coliform, *E. coli* and HPC at 37°) in drinking water, which causes the microbiological data set generally to have more zeros than expected (Zafakali and Ahmad, 2013; Sharp et al., 2006). For this reason, overdispersion, which is a situation where the variance is greater than the mean, seems to be the main feature for microbiological datasets. A number of mixed and compound Poisson distributions have been developed and compared for handling overdispersion and excess zero counts (Li, 2015; Akinpelu et al., 2016; Cao et al., 2017; Altun, 2018; Sohn and Li, 2018; Xu et al., 2021; and Workie and Azene, 2021).

2. Methods

In this section the method of estimating the sample size and the Zero-Inflated Poisson Regression Model (ZIPRM) are presented. The data analysis was performed using SAS 9.3, and all conclusions were based on a 5% significance level. The fitted model's goodness of fit is evaluated using measures such as the deviance, scaled deviance, Pearson Chi-square, scaled Pearson Chi-square, log-likelihood (LL), and full log-likelihood (FLL).

2.1. Sample size estimation

Based on an estimated projected population size of 362,862, we estimated our sample size for the project from the formula of David et al. (2020). The formula is as follows:

$$SS = \frac{N}{1 + N(e)^2}. \quad (1)$$

where SS is the sample size, N is the population size, and e is the tolerable error in investigating the population. Therefore,

$$SS = \frac{362862}{1 + 362862(0.05)^2} = 399.998 \approx 400$$

Hence, 400 questionnaires were administered to individuals in three different locations in Wukari, Taraba state, which were purposely selected due to their population densities before using a simple random sampling technique to administer the data instrument. Variables used for the analysis are subjects' plant protein consumption per week, Gender (Male and Female), Occupation (Artisans [AT], Business [BS], Civil servants [CS], Farmers [FA], and Students [STD]), Age, Height, Weight, Body Mass Index [BMI], and Body Reaction to Plant Protein Consumption (Heartburn [HB], Indigestion [IG], Other reaction [OR], and Running stomach [RS]).

2.2. Zero-Inflated Poisson (ZIP) regression

In ZIP regression, the responses $Y = (y_1, y_2, \dots, y_n)$ are independent, and excess zero counts are assumed to occur with probability p_i and follow a Poisson distribution with mean λ_i , with probability $1 - p_i$ where, $i = 0, 1, 2, \dots, n$. The ZIP model can thus be seen as a mixture of two component distributions, a zero part and non-zero components, given by:

$$p(Y = y_i / \lambda, p_i) = \begin{cases} p_i + (1 - p_i)^{-\lambda_i} & , y_i = 0 \\ (1 - p_i) \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} & , y_i = 1, 2, \dots \end{cases} \quad (2)$$

The first part of the equation above is the zero part of the model and the second part is the non-zero counts part of the model. The two components together constitute the zero-inflated model. The mean and variance of the zero-inflated Poisson model are:

$$E(y_i) = \lambda_i(1 - p_i) \quad (3)$$

$$\text{var}(y_i) = \lambda_i(1 - p_i)(1 + \lambda_i p_i) \quad (4)$$

Note that this distribution approaches the Poisson distribution as $p_i \rightarrow 0$. The Poisson mean vector $\lambda = (\lambda_1, \dots, \lambda_n)$ has the canonical link $\log(\lambda)$ for a Poisson regression model, which satisfies.

$$\text{Log}(\lambda) = B\beta \quad (5)$$

$$\lambda = e^{B\beta} \quad (6)$$

The canonical logit link is considered for the parameter vector $p = (p_1, \dots, p_n)$ in regression, that is,

$$\text{Logit}(p) = \log\left(\frac{p}{1-p}\right) = G\gamma \quad (7)$$

$$p = \frac{e^{G\gamma}}{1+e^{G\gamma}} \quad (8)$$

$$(1-p) = \frac{1}{1+e^{G\gamma}} \quad (9)$$

Here B and G are covariate matrices. The covariate matrix B is responsible for the Poisson outcome in Y and the covariate matrix G is responsible for excess zeros in Y . Also, B and G can be identical, or may have some common covariates depending on the study. Therefore,

$$F(y_i) = \begin{cases} p_i + (1-p_i)e^{-\lambda_i} = \frac{e^{G_i\gamma}}{1+e^{G_i\gamma}} + \frac{e^{-e^{B_i\beta}}}{1+e^{G_i\gamma}} & ; \text{ when } y_i = 0 \\ (1-p_i) \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = \frac{e^{-e^{B_i\beta}} (e^{B_i\beta})^{y_i}}{(1+e^{G_i\gamma})^{y_i!}} & ; \text{ when } y_i = 1, 2, \dots \end{cases} \quad (10)$$

The likelihood function is given by

$$l(y_i; \gamma, \beta, \lambda, B, p) = \left(\left(\frac{e^{G_i\gamma}}{1+e^{G_i\gamma}} + \frac{e^{-e^{B_i\beta}}}{1+e^{G_i\gamma}} \right) \left(\frac{e^{-e^{B_i\beta}} (e^{B_i\beta})^{y_i}}{(1+e^{G_i\gamma})^{y_i!}} \right) \right) \quad (11)$$

Hence, the log-likelihood function for estimating the ZIP regression parameters is

$$ll = \sum_{y_i=0} \log(e^{G_i\gamma} + e^{-e^{B_i\beta}}) - \sum_{i=1}^n \log(1+e^{G_i\gamma}) + \sum_{y_i>0} (y_i B_i \beta - e^{B_i\beta}) - \sum_{y_i>0} \log(y_i!) \quad (12)$$

The parameter estimation can be carried out by employing the EM algorithm and/or the Newton–Raphson algorithm to obtain the maximum likelihood estimates as used for this study.

3. Results

This section presents results for the ZIP regression model (ZIPRM) fitting. The goodness of fit test (GoFT) results are first presented, where measures like the deviance, scaled deviance, Pearson Chi-square, scaled Pearson Chi-square, log-likelihood, and full log-likelihood were applied. Also, the frequency distribution of subjects' plant protein consumption, the full ZIPRM and the reduced ZIP regression model are all presented.

Table 1. ZIP Goodness of Fit Test Result

Criterion	DF	Value	Value/DF	P-value
Deviance	374	1098.3195	2.9367	
Scaled Deviance	374	354.2966	0.94732	
Pearson Chi-Square, χ^2	374	332.4202	0.8888	
Scaled Pearson χ^2	374	332.4202	0.8888	0.9401
Log Likelihood		830.4187		
Full Log Likelihood		-549.1598		

Table 1 presents the GoFT for ZIPRM for plant protein consumption count data. It can be observed that the deviance value of 1098.3195 is in excess of the degree of freedom (DF), but after the overdispersion effect has been removed the scaled deviance value is close to the DF, and when divided by the DF the value obtained is close to one. Likewise, the Pearson- χ^2 value of 332.4202 is equal to the scaled Pearson- χ^2 value of 332.4202 after overdispersion has been removed. The values obtained are close to N (400) and the obtained value divided by the degree of freedom produced a value close to one (0.8888). This implies that the Zero-Inflated Poisson model is a good fit, and based on the p-value of 0.9401, which is close to one, it is concluded that there is no overdispersion in the data. Also, the LL is positive and the FLL value of -549.1598 is less than one, which supports the conclusion drawn from the scaled deviance and scaled Pearson- χ^2 .

Table 2 shows the frequency distribution of a subject's bean consumption. It can be seen that the subjects that do not consume beans at all make up 53.25% of the group, while 1.7% consume beans 6 times per month. Also, subjects with the highest bean consumption make up 2% of the 400 subjects.

Table 2. Frequency Distribution Table

NBIntake	Frequency	Percent	CF	CP
0	213	53.25	213	53.25
1	46	11.50	259	64.75
2	16	4.00	275	68.75
3	23	5.75	298	74.50
4	8	2.00	306	76.50
5	23	5.75	329	82.25
6	7	1.75	336	84.00
7	8	2.00	344	86.00
8	8	2.00	352	88.00
10	16	4.00	368	92.00
11	8	2.00	376	94.00
12	16	4.00	392	98.00
18	8	2.00	400	100.00

CF is cumulative frequency. CP is cumulative percent

From the results in Table 3 it can be observed that Female (F) gender of consumers of plant protein is significant at 5% with a p-value of < 0.0001 , and this implies that F consumers of plant protein differ significantly from Male (M) consumers. Also, Age is found to be a significant factor in plant protein consumption at 5%, with a p-value of 0.0138. As regards consumers' body reactions to plant protein consumption, Heartburn (HB) and Other Reaction (OR) are not significant, with p-values of 0.2199 and 0.5917 respectively at $\alpha = 5\%$. This implies that HB and OR do not differ significantly from Running Stomach (RS). However, Indigestion (IG) is significant with a p-value of 0.0249 at 5%, and this implies that IG differs significantly from RS. The fitted ZIPRM for the plant protein consumption data is given in equation (13). It can be observed that F, Age, and IG all have a negative contribution to the weekly consumption of plant protein by the subjects. To confirm that the outcome is valid, all factors found not to be significant (Occupation, Height, Weight, and BMI) were removed from the model, and a reduced ZIPRM was fitted.

Table 3. ZIP Parameter Estimate for Plant Protein Consumption

Maximum Likelihood Zero-Inflation Parameter Estimates								
Parameter		DF	Estimate	Std. Error	Wald 95% CI		Wald Chi-Square	P-value
Intercept		1	-20.3084	32.0294	-83.0848	42.4679	0.40	0.5260
Gender	F	1	-1.4744	0.3164	-2.0945	-0.8543	21.72	<.0001
Gender	M	0	0.0000	0.0000	0.0000	0.0000	*	*
Occupation	ART	1	0.2870	0.6800	-1.0458	1.6199	0.18	0.6730
Occupation	BUS	1	0.1628	0.7332	-1.2743	1.5999	0.05	0.8243
Occupation	CS	1	0.3397	0.6021	-0.8404	1.5198	0.32	0.5726
Occupation	FA	1	0.4978	0.7695	-1.0103	2.0059	0.42	0.5177
Occupation	STD	0	0.0000	0.0000	0.0000	0.0000	*	*
Age		1	-0.0532	0.0216	-0.0956	-0.0109	6.06	0.0138
Height		1	13.7048	19.6255	-24.7606	52.1701	0.49	0.4850
Weight		1	-0.1644	0.2459	-0.6464	0.3175	0.45	0.5037
BMI		1	0.4520	0.6610	-0.8435	1.7475	0.47	0.4941
Reaction	HB	1	-0.4697	0.3828	-1.2200	0.2807	1.51	0.2199
Reaction	IG	1	-0.9924	0.4424	-1.8595	-0.1254	5.03	0.0249
Reaction	OR	1	-0.2902	0.5411	-1.3507	0.7702	0.29	0.5917
Reaction	RS	0	0.0000	0.0000	0.0000	0.0000	*	*

* denotes redundant, and *p*-values in bold are significant.

$$\lambda_i = \exp(-20.31 - 1.4744F + 0.29ART + 0.163BUS + 0.034CS + 0.498FA - 0.05Age + 13.7Height - 0.16Weight + 0.45BMI - 0.47HB - 0.99IG - 0.29OR). \quad (13)$$

Table 4 presents the reduced model of the ZIP regression analysis, that is, removing the variables that were not significant in the full model represented by equation (13).

Table 4 presents the parameter estimates for the reduced model. It can be observed that all the parameters that are significant in the complete model are also significant in the reduced model, except for levels of body reaction to plant protein consumption, for which none was significant at 5%. This implies that the most significant factors with an effect on plant protein consumption are subjects' gender (Females) and Age, which are highly significant at 5%. Their effects are positive, which contradicts the outcome of the full model. The reduced fitted ZIPRM for the plant protein consumption data is given in equation (14).

$$\lambda_i = \exp(-3.12 + 0.5994F + 0.0995Age - 0.0698HB + 0.0349IG + 0.0048OR). \quad (14)$$

Table 4. ZIP Parameter Estimate for the Reduced Model

Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Std. Error	Wald 95% CI		Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.1245	0.2688	-3.6513	-2.5976	135.10	<.0001
Gender	F	1	0.5994	0.0786	0.4455	0.7534	58.23	<.0001
Gender	M	0	0.0000	0.0000	0.0000	0.0000	.	.
Age		1	0.0995	0.0053	0.0891	0.1098	356.57	<.0001
Reaction	HB	1	-0.0698	0.0869	-0.2401	0.1005	0.65	0.4218
Reaction	IG	1	0.0349	0.0898	-0.1411	0.2108	0.15	0.6975
Reaction	OR	1	0.0048	0.1347	-0.2592	0.2688	0.00	0.9716
Reaction	RS	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

4. Conclusion

In this research, ZIPRM was fitted to counts with inflated zeros from a study on plant protein consumption on 400 subjects who were selected randomly. The results obtained revealed that Gender (F) and Age are factors with a high influence on plant protein consumption at a 1% level of significance. Other factors like Occupation, Height, Weight, BMI, and Body Reaction to plant protein consumption were found not to influence the subjects' plant protein consumption. The ZIPRM was found to be a good fit for the data based on the GoFT results presented in Table 1. Hence, the results obtained are valid, since there was no overdispersion effect and the inflated zeros were well modeled by the ZIPRM. The result is in agreement with results obtained by Li (2015), Akinpelu et al. (2016), Cao et al. (2017), Altun (2018), Sohn and Li (2018), Xu et al. (2021), and Workie and Azene (2021) on the application of zero-inflated regression models in handling excess zeros in any count data set.

REFERENCES

- Akinpelu K.P., Yusuf O.B., Akpa O.M., Gbolahan A.O. (2016): Zero Inflated Regression Models with Application to Malaria Surveillance Data. *International Journal of Statistics and Application* 6(4): 223-234.

- Cao Y., Zhang A., Li H. (2017): Microbial composition estimation from sparse count data. ArXiv. Available at:arXiv:1706.02380.
- David I., Adubisi O., Farouk B., Adehi M. (2020): Assessing MSMEs Growth Through ROSCA Involvement Using Paired t-Test and One Sample Proportion Test. *Journal of Social and Economic Statistics*, 9(2): 30-42, <https://doi.org/10.2478/jses-2020-0011>
- Li H. (2015): Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application* 2: 73–94.
- Sohn M.B., Li H. (2018): A GLM-based latent variable ordination method for microbiome samples. *Biometrics*, 74: 448–457. [PubMed: 28991375]
- Xu T., Demmer R.T., Li G. (2021): Zero-inflated Poisson factor model with application to microbiome read counts. *Biometrics*, 77(1): 91-101, <https://doi.org/10.1111/biom.13272>
- Altun E. (2018): A new zero-inflated regression model with application. *Journal of Statisticians: Statistics and Actuarial Sciences*, 11(2): 73-80.
- Workie M.S., Azene A.G. (2021): Bayesian zero-inflated regression model with application to under-five child mortality. *Journal of Big Data* 8(4), <https://doi.org/10.1186/s40537-020-00389-4>
- Lambert D. (1992): Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1): 1–14.
- Ridout M., Demétrio C.G., Hinde J. (1998): Models for count data with many zeros. *Proceedings of the XIXth International Biometric Conference* 19(1): 179–192.
- Sharp E., Parson S., Jefferson B. (2006): Coagulation of NOM: linking character to treatment. *Water Science and Technology*, 53(7): 67-76.
- Zafakali N.S., Wan M.A., Ahmad W. (2013): Modelling and Handling Overdispersion Health Science data with Zero-inflated Poisson Model. *Journal of Modern Applied Statistical Methods*, 12(1): 255-260.