# Evaluating the Performance of Zero-Inflated and Hurdle Poisson Models for Modeling Overdispersion in Count Data

**Aswi Aswi[1*], Sri Ayu Astuti[1], and Sudarmin[1]**
[1]Statistics Department, Universitas Negeri Makassar, Indonesia
*Corresponding author: aswi@unm.ac.id

Received: 23 February 2022
Accepted: 30 March 2022
Published: 31 March 2022

**ABSTRACT** – **A Poisson regression model is commonly used to model count data. The Poisson model assumes equidispersion, that is, the mean is equal to the variance. This assumption is often violated. In count data, overdispersion (the variance is larger than the mean) occurs frequently due to excessive zeroes in the response variable. Zero-inflated Poisson (ZIP) and Hurdle models are commonly used to fit data with excessive zeros. Although some studies have compared the ZIP and Hurdle models, the results are inconsistent. This paper aims to evaluate the performance of ZIP and Hurdle Poisson models for overdispersion data through both simulation study and real data. Data were simulated with three different sample sizes, six different means, and three different probabilities of zero with 500 replications. Model goodness-of-fit measures were compared by using Akaike Information Criteria (AIC). Overall, the ZIP model performed relatively the same or better than the Hurdle Poisson model under different scenarios, but both ZIP and Hurdle models are better than the standard Poisson model for overdispersion in count data.**
**Keywords**– **Hurdle Poisson, Overdispersion, Zero-inflated Poisson (ZIP)**

## I. INTRODUCTION

A Poisson regression model is frequently used to model count data. One of the assumptions of the Poisson model (standard Poisson Generalized Linear Model (GLM)) is that the variance is equal to the mean. However, count data frequently deviate from the Poisson distribution due to a larger proportion of zeroes resulting in variance greater than the mean in the observed variable which is called overdispersion [1]. In another word, excessive zeros in the count response may cause overdispersion and result in biased inference [2].

Previous studies have found that if excessive zero is not taken into account, the unreasonable fit will be obtained for both the nonzero and zeros counts [3]. To deal with overdispersion, Hurdle Poisson and Zero Inflated Poisson (ZIP) are generally used [1, 4]. A review, as well as an evaluation of the performance of zero-inflated and Hurdle models for count data, has been conducted [4]. A number of studies have compared the Zero Inflated and Hurdle models. For example, research has concluded that Zero Inflated and Hurdle models are indiscernible in terms of goodness of fit measures namely the Akaike information criterion (AIC) [5, 6]. Research has compared different counts models including ZIP, Hurdle Poisson (HP), zero-inflated negative binomial (ZINB), hurdle negative binomial (HNB), Poisson, and negative binomial (NB) models for analyzing inpatient hospitalization data of patients' psychiatric disorders [7]. They found that the negative binomial model performed much better than the Poisson model. Furthermore, the Hurdle models are better than Zero Inflated Models in terms of AIC, and Vuong's test [8] and hurdle negative binomial model performed the best fit [7]. Another study also compared the ZIP model and Hurdle models by using a simulation study. They conclude that the ZIP model may not fit the data well even when the test indicates a significant witness of zero inflation [9]. On the other hand, research has also been studied regarding Zero-inflated and Hurdle Models of count data with excessive zeros for modeling HIV risk [10]. They concluded that ZIP and zero-inflated negative binomial (ZINB) models performed better than Poisson hurdle (PH) and negative binomial hurdle (NBH) models [10].

A previous study has compared six different count regression models in predicting the number of under-five death in Ethiopia [11]. These six distinct models are Poisson, Negative Binomial (NB), Zero Inflated Poisson (ZIP), Zero-inflated Negative Binomial (ZINB), Poisson-Logit Hurdle, and Negative Binomial-Logit Hurdle (NBLH). They identified that NB and Poisson models are not sufficient for count response with excessive zeros. They also found that NBLH models have a better fit for modeling the count data with overdispersion and excess zero. Furthermore, they compared the Bayesian method and classical method and found that the Bayesian approach in estimating parameters is more robust and precise than the classical (such as Maximum Likelihood Estimation) approach.

As the results of these studies are inconsistent, the relative performance of Zero-inflated Poisson and Hurdle Poisson models for overdispersion data through both simulation study and real data need to be evaluated. The overall aim of this paper is to evaluate which models are likely to be appropriate for overdispersion data under different scenarios with a focus on ZIP and Hurdle Poisson models.

DOI: 10.12962/j27213862.v5i1.12422

## II. MATERIAL AND METHODS

### A. *Simulation study*

A simulation study was designed to examine the performance of ZIP and Hurdle Poisson models for overdispersion data. Based on the aim of this paper, 500 replicate datasets of count data were generated under 54 scenarios: three different sample sizes ($n$), that are, 30, 300, and 500, six different means ($\mu$), that are, 0.3, 0.6, 1, 6, 12, and 24, and three different probabilities of zero ($\pi$ = 0.3, $\pi$ = 0.5, and $\pi$ = 0.7).

The procedure for generating the synthetic data consists of some steps. First, we considered involving two covariates so that the independent variables $X_1$ and $X_2$ were randomly generated using normal distributions with three different sample sizes, that are, n = 30, n = 300, and n = 500. Second, the dependent variable ($Y$) was generated using Poisson distribution with six different means ($\mu$), that are, 0.3, 0.6, 1, 6, 12, and 24, and three different probabilities of zero ($\pi$ = 0.3, $\pi$ = 0.5, $\pi$ = 0.7). Third, the parameters of standard Poisson Generalized Linear Models (GLMs) under different scenarios were estimated by using Maximum Likelihood Estimation (MLE). Based on the estimated parameters on standard Poisson, the dispersion values were checked by using the R package AER, function *dispersiontest* in R version 3.6.3 [12]. The function *dispersiontest* evaluates the null hypothesis of equidispersion in Poisson distribution versus the alternative hypothesis of overdispersion or underdispersion. Fourth, the parameters of the Zero Inflated Poisson and Hurdle Poisson regression model were estimated by using Maximum Likelihood Estimation. In this step, the AIC values of each model were obtained. Last, the AIC values of standard Poisson, Zero Inflated Poisson, and Hurdle Poisson models were compared to get the best model. R code to generate data is available upon request.

### B. *Poisson Models*

The Poisson regression model is generally used in many fields including public health to model the relationship between independent variables and dependent variables. The dependent variable ($Y$) is usually the number of events where the data is discrete and the number of events is Poisson distributed. One of the characteristics of the Poisson distribution is that the variance equals the mean, that is,

$$V(Y) = E(Y) = \mu.$$

If the variance is larger than the mean, we would have over-dispersion. However, if the variance is less than the mean, we would have under-dispersion, and it is relatively sparse. If the over-dispersion happened and the parameters are estimated by using Poisson regression, then the results will be inefficient [13]. Commonly, excess zeros in the data are one of the causes of over-dispersion. Extra zeros happen when many subjects or participants do not experience any events during the monitoring of the periods.

The Poisson Probability distribution of $Y_i$ is given as follows:

$$P(y_i; \mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}; y_i = 0, 1, 2, \dots$$

where $y_i$ is the number of count data which is a non-negative integer, and $\mu_i$ is the mean of count data.

### C. *Zero Inflated Poisson (ZIP) model*

The zero-inflated model is one of the popular models for modeling count data with extra zeros [14]. The zero-inflated Poisson model [14] is a combination of a regular count Poisson regression model and an element that accommodates the high occurrence of zeros. The zero-inflated model assumes that zero observations have two distinct origins namely structural and sampling where the sampling zeros come from the usual Poisson distribution assuming that zero observations occurred by chance.

The response variables (count data) $Y = (Y_1, Y_2, \dots, Y_n)'$ are independent in the ZIP regression model [14] and

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i \\ \text{Poisson } (\mu_i) & \text{with probability } 1 - p_i \end{cases}$$

If the count data is distributed Poisson, the ZIP model is given as follows [4, 14]:

$$P(Y_i = y_i) = \begin{cases} p_i + (1 - p_i)e^{\mu_i} & \text{if} \quad y_i = 0 \\ (1 - p_i)\dfrac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} & \text{if} \quad y_i = 1, 2, 3, \dots, n \end{cases}$$

where $\mu_i$ is the mean of count data of the standard Poisson distribution, and $p_i$ is the probability of being an excess zero.

### D. Hurdle Poisson (HP) model

A hurdle model [15] is a two-component mixture model consisting of the probability of zero counts and a zero-truncated Poisson distribution (a positive count component). In contrast to zero-inflated which assumed that zero observations have two distinct origins namely structural and sampling, a hurdle model assumes that the whole zero observations only have one origin namely "structural".

Let the response variables of the $i^{th}$ count data $Y_i$, $i = 1, 2, 3, \ldots, n$, where $n$ indicates the total number of count data. If the count data is distributed Poisson, the probability distribution for the Hurdle Poisson is given as follows:

$$P(Y_i = y_i) = \begin{cases} p_i & \text{if} \quad y_i = 0 \\ (1 - p_{i_i}) \dfrac{\mu_i^{y_i} e^{-\mu_i} / y_i!}{1 - e^{-\mu_i}} & \text{if} \quad y_i = 1, 2, 3, \ldots, n \end{cases}$$

### E. Comparing Models

The goodness-of-fit of these two models was compared using the Akaike Information Criterion (AIC) [16]. A smaller value of AIC indicates a better model fit. AIC is based on the Maximum Likelihood Estimator (MLE) method. The formula of AIC is given as follows:

$$\text{AIC} = -2L + 2q$$

where L represents the log-likelihood, and q is the number of model parameters that is the number of variables and the intercept in the model.

## III. RESULTS AND DISCUSSION

### A. Simulated Data

The dispersion values of the simulation study under 54 different scenarios namely three different sample sizes; 30, 300, and 500, six different means ($\mu$): 0.3, 0.6, 1, 6, 12, and 24, and three different probabilities of zero: $\pi = 0.3$, $\pi = 0.5$, and $\pi = 0.7$ can be obtained based on the estimated parameters on standard Poisson Generalized Linear Models (GLM) and it can be seen on Table 1.

**Table 1**. The dispersion values of the generated data under different scenarios

| | $\mu$ | Dispersion values | | |
|---|---|---|---|---|
| | | $\pi = 0.3$ | $\pi = 0.5$ | $\pi = 0.7$ |
| n = 30 | 24 | 8.518190 | 14.66514 | 18.40991 |
| | 12 | 4.984096 | 7.539053 | 9.92305 |
| | 6 | 2.654237 | 3.976187 | 5.872537 |
| | 1 | 1.018792 | 1.278355 | 1.824422 |
| | 0.6 | 0.887774 | 0.925816 | 1.240138 |
| | 0.3 | 0.943989 | 0.996254 | 1.268558 |
| n = 300 | 24 | 8.322195 | 13.16923 | 17.92736 |
| | 12 | 4.775567 | 7.199179 | 9.656975 |
| | 6 | 2.975343 | 4.166912 | 5.245435 |
| | 1 | 1.469863 | 1.613538 | 1.723377 |
| | 0.6 | 1.269158 | 1.302904 | 1.272920 |
| | 0.3 | 1.123450 | 1.158888 | 1.135914 |
| n = 500 | 24 | 8.264605 | 13.18284 | 18.17418 |
| | 12 | 4.711403 | 7.225716 | 9.657460 |
| | 6 | 2.907330 | 4.105305 | 5.347267 |
| | 1 | 1.352680 | 1.602214 | 1.815279 |
| | 0.6 | 1.175130 | 1.341328 | 1.428738 |
| | 0.3 | 1.059547 | 1.153011 | 1.231163 |

Based on Table 1, it can be seen that the dispersion values do not depend on the number of samples, but depend on $\mu$ and $\pi$ values. The higher the $\mu$ and $\pi$ values, the higher the dispersion values. The AIC values for three models namely standard Poisson, ZIP, and Hurdle Poisson under different scenarios are given in Table 2.

**Table 2**. The AIC values for three Poisson, ZIP, and Hurdle Poisson models under different scenarios

| $N$ | $\mu$ | $\pi$ | AIC | | |
|---|---|---|---|---|---|
| | | | Poisson | ZIP | Hurdle |
| 30 | 24 | 0.3 | 492.2 | 170.1703 | 170.1703 |
| | | 0.5 | 672.6 | 141.7778 | 141.7778 |
| | | 0.7 | 610.4 | 96.42247 | 96.42247 |
| | 12 | 0.3 | 321.5 | 159.7532 | 159.7532 |
| | | 0.5 | 375.6 | 125.7752 | 125.7752 |
| | | 0.7 | 348.7 | 90.5955 | 90.59551 |
| | 6 | 0.3 | 196.4 | 139.5688 | 139.5597 |
| | | 0.5 | 194.2 | 111.9903 | 111.9866 |
| | | 0.7 | 208.1 | 87.13797 | 87.13771 |
| | 1 | 0.3 | 77.88 | 78.32632 | 80.1327 |
| | | 0.5 | 70.51 | 73.81056 | 73.05206 |
| | | 0.7 | 67.44 | 55.55161 | 61.12456 |
| | 0.6 | 0.3 | 60.61 | 59.57188 | 66.21250 |
| | | 0.5 | 52.93 | 54.71073 | 56.71159 |
| | | 0.7 | 45.91 | 49.93129 | 48.91696 |
| | 0.3 | 0.3 | 43.43 | 47.94528 | 49.37687 |
| | | 0.5 | 41.47 | 44.90892 | 41.84016 |
| | | 0.7 | 38.32 | 39.43287 | 39.31652 |
| 300 | 24 | 0.3 | 4903 | 1651.239 | 1651.239 |
| | | 0.5 | 5885 | 1340.221 | 1340.221 |
| | | 0.7 | 5742 | 912.1925 | 912.1925 |
| | 12 | 0.3 | 2954 | 1522.599 | 1522.599 |
| | | 0.5 | 3358 | 1247.346 | 1247.346 |
| | | 0.7 | 3119 | 855.6506 | 855.6506 |
| | 6 | 0.3 | 1895 | 1370.011 | 1369.988 |
| | | 0.5 | 1961 | 1139.049 | 1139.061 |
| | | 0.7 | 1691 | 797.2207 | 797.2382 |
| | 1 | 0.3 | 727.4 | 697.3378 | 698.8206 |
| | | 0.5 | 621.1 | 573.9492 | 573.8655 |
| | | 0.7 | 451.1 | 396.5298 | 397.1351 |
| | 0.6 | 0.3 | 545 | 539.1004 | 541.6867 |
| | | 0.5 | 449.2 | 441.9961 | 441.1014 |
| | | 0.7 | 303.3 | 296.3232 | 297.6241 |
| | 0.3 | 0.3 | 371.4 | 369.272 | 374.1139 |
| | | 0.5 | 308.1 | 306.6627 | 308.1935 |
| | | 0.7 | 208.3 | 209.8180 | 209.2676 |
| 500 | 24 | 0.3 | 8125 | 2728.905 | 2728.905 |
| | | 0.5 | 9931 | 2219.141 | 2219.141 |
| | | 0.7 | 9717 | 1544.938 | 1544.938 |
| | 12 | 0.3 | 4914 | 2514.001 | 2514.001 |
| | | 0.5 | 5608 | 2064.325 | 2064.325 |
| | | 0.7 | 5217 | 1442.085 | 1442.085 |
| | 6 | 0.3 | 3148 | 2270.878 | 2270.878 |
| | | 0.5 | 3246 | 1871.285 | 1871.277 |
| | | 0.7 | 2867 | 1332.327 | 1332.328 |
| | 1 | 0.3 | 1197 | 1163.241 | 1163.259 |
| | | 0.5 | 1027 | 960.8013 | 960.85 |
| | | 0.7 | 774.3 | 676.3094 | 676.2504 |
| | 0.6 | 0.3 | 894.4 | 892.4016 | 892.9387 |
| | | 0.5 | 746.3 | 731.4642 | 730.7281 |
| | | 0.7 | 544.3 | 521.8026 | 521.801 |
| | 0.3 | 0.3 | 583.9 | 576.8232 | 588.4093 |
| | | 0.5 | 483.8 | 481.9361 | 483.0538 |
| | | 0.7 | 361.4 | 356.0369 | 356.0082 |

Based on Table 1, when $\mu$ = 6, 12, and 24 whatever the values of $\pi$, the ZIP and Hurdle models have the same performance in terms of AIC. Furthermore, ZIP and Hurdle models also perform the same when zero probability = 0.7 and $\mu$ = 0.3. Also, when zero probability = 0.5, ZIP tends to be the same as Hurdle models. However, when n = 30 and $\mu$ = 0.6, the dispersion value is close to 1, and in this case, Poisson is better than the ZIP and Hurdle Poisson models. It can be also concluded that ZIP performed relatively better than Hurdle Poisson when zero probability ($\pi$) = 0.3 or if zero

probability $\pi = 0.7$ and $\mu = 1$. Overall, when the dispersion values were more than 1 (overdispersion), the ZIP model performed relatively the same or better than the Hurdle Poisson model under different scenarios, based on AIC.

### *B. Case study*

The two models namely ZIP and Hurdle Poisson were applied to a case study of positive Covid-19 in Makassar city for every district on February 14, 2021. This time was selected as it consists of many zeros in the count data. The count data ($Y$) of positive Covid-19 were gathered from the Makassar city health office. We also considered including two covariates to mimic the simulation study. Data on the height of the area in each district ($X_1$) and data on the percentage of population density in each district ($X_2$) were also used. The highest number of confirmed covid cases in Makassar city is Panakukkang district (4 cases). There are 10 districts with zero positive Covid-19 cases, namely Biringkanaya, Bontoala, Manggala, Mariso, Rappocini, Sangkarang, Tallo, Tamalanrea, Ujung Tanah, and Wajo districts. The dispersion values of the real data on Covid-19 and two covariates used in this case study, the dispersion values, and AIC values were given in Table 3.

**Table 3**. The dispersion values of the real data on Covid-19 and two covariates, the dispersion values, and AIC values

| $n$ | $\mu$ | $\pi$ | Dispersion | AIC values | | |
|---|---|---|---|---|---|---|
| | | | | Poisson | ZIP | Hurdle |
| 15 | 0.73 | 0.7 | 1.52 | 43.22 | 42.41 | 42.60 |

Based on Table 3, the results show that the AIC values for ZIP and Hurdle models are relatively similar namely 42.41 and 42.60 respectively, but less than the AIC value of the standard Poisson distribution. This is in line with the results of our simulation study stating that when the overdispersion occur, the ZIP model performed relatively the same or better than the Hurdle Poisson model under different scenarios, but ZIP and Hurdle models are better than the standard Poisson model for all given $n$ based on AIC. These results are in agreement with some other studies stating that Zero Inflated and Hurdle models are similar in terms of AIC [5, 6]. Another study also concluded a result somewhat similar to that found in our study that the ZIP model performed better than the Poisson hurdle (PH) model [10].

This study used the classical approach in estimating the parameters of Poisson, the Zero Inflated Poisson, and Hurdle Poisson regression. It is acknowledged that using other approaches such as the Bayesian approach may have different results.

## IV. CONCLUSION

Overall, the performance of Zero Inflated Poisson is relatively the same or better than the Hurdle Poisson model under different scenarios, but ZIP and Hurdle models are better than the standard Poisson for count data with overdispersion. The real data used in this study have the same conclusion as the simulation study. Considering other models to fit data with excessive zeros such as Zero Inflated Negative Binomial and Hurdle Negative Binomial as well Bayesian approach for estimating the parameters of the model could be possible future work.

## REFERENCES

[1] C. E. Rose, S. W. Martin, K. A. Wannemuehler, and B. D. Plikaytis, "On the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data," *Journal of biopharmaceutical statistics,* vol. 16, no. 4, pp. 463-481, 2006.

[2] Z. Yang, J. W. Hardin, and C. L. Addy, "Testing overdispersion in the zero-inflated Poisson model," *Journal of statistical planning and inference,* vol. 139, no. 9, pp. 3340-3353, 2009.

[3] S. E. Perumean-Chaney, C. Morgan, D. McDowall, and I. Aban, "Zero-inflated and overdispersed: what's one to do?," *Journal of statistical computation and simulation,* vol. 83, no. 9, pp. 1671-1683, 2013.

[4] C. X. Feng, "A comparison of zero-inflated and hurdle models for modeling zero-inflated count data," *Journal of statistical distributions and applications,* vol. 8, no. 1, pp. 1-19, 2021.

[5] L. Xu, A. D. Paterson, W. Turpin, and W. Xu, "Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data," *PloS one,* vol. 10, no. 7, pp. e0129606-e0129606, 2015.

[6] F. Tüzen, S. Erbaş, and H. Olmuş, "A simulation study for count data models under varying degrees of outliers and zeros," *Communications in statistics. Simulation and computation,* vol. 49, no. 4, pp. 1078-1088, 2020.

[7] S. Sharker, L. Balbuena, G. Marcoux, and C. X. Feng, "Modeling socio-demographic and clinical factors influencing psychiatric inpatient service use: a comparison of models for zero-Inflated and overdispersed count data," *BMC medical research methodology,* vol. 20, no. 1, pp. 232-232, 2020.

[8] Q. H. Vuong, "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica,* vol. 57, no. 2, pp. 307-333, 1989.

[9] Y. Min and A. Agresti, "Random effect models for repeated measures of zero-inflated count data," *Statistical modelling,* vol. 5, no. 1, pp. 1-19, 2005.

[10] M.-C. Hu, M. Pavlicova, and E. V. Nunes, "Zero-Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial," *The American journal of drug and alcohol abuse,* vol. 37, no. 5, pp. 367-375, 2011.

[11] M. S. Workie and A. G. Azene, "Bayesian zero-inflated regression model with application to under-five child mortality," *Journal of big data,* vol. 8, no. 1, pp. 1-23, 2021.

[12] R Core Team, "R: A language and environment for statistical computing," ed. Vienna, Austria: R Foundation for Statistical Computing, 2019.

[13] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data* (Econometric Society monographs ; 30). Cambridge: Cambridge University Press, 1998.

[14] D. Lambert, "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing," *Technometrics,* vol. 34, no. 1, pp. 1-14, 1992.

[15] J. Mullahy, "Specification and testing of some modified count data models," *Journal of econometrics,* vol. 33, no. 3, pp. 341-365, 1986.

[16] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," ed. New York, NY: Springer New York, 1998, pp. 610-624.

 