NNPublication Journal of Advance Research in Mathematics And Statistics ISSN: 2208-2409

DOI:https://doi.org/10.53555/4zfphn62 Publication URL: https://nnpub.org/index.php/MS/article/view/1870

ANALYSIS OF THE REGRESSION MODEL FOR ZERO-INFLATION DATA

Sanaa Mohammed Naeem[1*], Shaymaa Qasim Mohsin[2], Zainab Sami Yaseen[3]

[1]Southern technical university/college of health& medical techniques in basrah, Sunamohammed70@gmail.com*
[2]Basra University / College of Administration and Economics, shaymaa.qassim79@gmail.com
[3]Southern technical university / technical institute of Basrah, zainab.s.yassin@stu.edu.iq

Corresponding Author:
Sunamohammed70@gmail.com

Abstract:

The community may contain a large percentage of zero values that cause the community distribution to move away from zero, and this group is referred to as not following a normal distribution, so one of the conditions of the linear regression models is permeated. This type of society can be seen in many applications such as insurance, meteorology, auditing, environment, and manufacturing. The zero-community number is often analyzed via a two-part admixture model: The first part is probabilistic from zero and the second part is regular with a specific probability distribution. Problems of confidence estimation of the zero-classifier population mean under normal models have been present in research. Regression models have also been developed for the zero population groups. However, many of these models are aimed at counting data, although regression models with responses of a continuous type can be seen in application quite often. Moreover, these regression models for homeless populations do not address situations in which the data available for analysis were obtained through complex probability sampling designs.

Different statistical methods and models have been developed for the statistical analysis of such population. Based on the current research, most of the special studies focus on estimating the population mean and developing regression models. This dissertation will also focus on developing regression models.

This dissertation will also focus on developing regression models. Most of the regression models developed for the null population found in research have given more attention to population data in which observations can take only non-negative integer values that arise from counting rather than ordering. They also use maximum possibility methods and pseudo greatest possibility methods to estimate expected responses in Value . Variable / future variables.

research aims:*The main objective of this thesis is to compare estimation methods (both period and point estimation) in generalized linear models associated with complex sampling designs in the zero population. Based on the sample mixture preparation. As well as the application of ZIM regression models on continuous and discrete data.*

Theoretical part:

Regression analysis for the zero-inflated population (ZIP)

Regression analysis is widely used in business, the social sciences, the life sciences and many other disciplines. in these areas, the regression analysis for the non-inflation-free populations was considered as zero. Since the classical hypotheses of linear regression cannot be used in the non-inflation-free populations, Various methods have been proposed to model such communities. There are also large zero-count census data in applications, and various approaches have been developed to generalize regression models to such observations.

The ZIP regression model treats the data as a mixture of zeros and results of a Poisson variable. For the application, a ZIP regression model is used in the data. Variant of this model is:

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

$$P(Y_i = 0) = p_i + (1-p_i)\exp(-\lambda_i),$$

$$P(Y_i = y_i) = (1-p_i)\frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \quad y_i = 1,2,\cdots.$$

$$\log(\lambda) = X\beta$$

$$\log\left(\frac{p}{1-p}\right) = Z\gamma$$

where X and Z are the matrices of covariates. The two sets of covariates may or may not coincide. When they coincide, a simple model can be developed with the consideration that two linear predictors are related in some way. The simplest model, referred to as ZIP(η) regression model, has

$$\log(\lambda) = X\beta \text{ and } \log\left(\frac{p}{1-p}\right) = \eta X\beta$$

where (η) is a scalar parameter, (β) is vector of regression parameters, and X is the vector of covariates.

Welsh et al. discussed different regression models for zero-inflated count data with the application in the abundance of rare species. Considering the number of trees with hollows as a covariate, different regression models were used to estimate the mean abundance of possums. Their comments on di_erent models that they considered are summarized below. In zero-inflated Poisson distribution, non-zero counts are assumed to follow a zero truncated Poisson distribution. In practice, count data are often over dispersed. It has been established that if non-zero observations are over dispersed and simultaneously correlated due to the sampling designs or the data collection procedures, the parameter estimates using ZIP regression may be seriously biased. In this context, Ridout et al. proposed a score test for testing a ZIP regression model against zero-inflated negative binomial alternatives. Yau et al. developed zero-inflated negative binomial mixed regression model for over dispersed count data with extra zeros. Cui and Yang developed zero-inflated generalized Poisson (ZIGP) regression mixture model as another alternative to ZIP regression. ZIGP regression mixture model handles the zero inflated and Poisson dispersion in the same distribution.

Fletcher et al. described an approach that combines the ordinary and the logistic regression models to the skewed data with many zeros. The objective of their study was to present a special case of conditional model developed by Welsh et al. with the assumption that the positive abundance has log-normally distributed error term. This study suggests to use parametric bootstrapping to construct the confidence intervals. According to this approach, the expected value of the response is given by

$$\begin{aligned} E(Y) &= P(Z=1)E(Y \mid Z=1) + P(Z=0)E(Y \mid Z=0) \\ &= P(Z=1)E(Y \mid Z=1) + 0 \\ &= \pi\mu \end{aligned}$$

Where $\pi = P(Z=1)$ and $\mu = E(Y \mid Z=1)$ The point estimate of expected response is given by

$$\hat{E}(y) = \hat{\pi}\hat{\mu}$$

$$where\ \hat{\pi} = \frac{\exp(x'\hat{\beta})}{1+\exp(x'\hat{\beta})}$$

$$\hat{\mu} = \exp\left(w'\hat{\theta} + \hat{\sigma}^2\right).$$

Zero-inflated mixed regression

The linear regression model is widely used to study the linear relationship between the response variable and the explanatory variables under the assumption that the error terms are normally distributed with zero mean and constant variance. For the Y component vector, a linear regression model with k fixed vectors and unknown parameters:

$$Y = \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

$$\varepsilon \sim N(0, \sigma^2 I_n).$$

.

The generalized linear models of amplified zero-sums can be obtained using a specific mixed model as follows:
$g(y_i; \alpha, x_i', \beta, \sigma) = \alpha f(y_i; \alpha, x_i', \beta, \sigma) I(y_i \neq 0) + (1 - \alpha) I(y_i = 0)$
$\tau = E(Y|X = x_0) = \alpha\mu(x_0'\beta) = \alpha\psi^{-1}(x_0'\beta)$

Three types of distributions were used for the dependent variable (models), which are:

- **Gamma model**
- **log-normal model**
- **Poisson model**

Estimation methods:
To estimate model parameters and distribution (point estimation and interval estimation):

1) Maximum pseudo-likelihood function:

For convenience, let $\lambda = (\tau ; \beta ; \sigma)$. Since reparameterization does not change the likelihood function, pseudo-likelihood function $\hat{\imath}\lambda = \hat{\imath}(\tau ; \beta ; \sigma)$.. Let

$$\Omega_n(\lambda) = -\frac{\partial^2 \hat{\ell}}{\partial \lambda^2}$$

$\Delta_n(\lambda) = \text{Var}\,(\partial\hat{\ell}/\,\partial\lambda)$

where $\Delta_n(\lambda)$ explains the variations due to the probability sampling design and the model g(.) at true value λ_0 of λ. Rewrite $\Delta_n(\lambda)$ and $\Omega_n(\lambda)$ in the partition matrix form as:

$$\Omega_n(\lambda) = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}_{(k+2)\times(k+2)},$$

$$\Delta_n(\lambda) = \begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix}_{(k+2)\times(k+2)},$$

where δ_{11} and w_{11} are 1x1 and δ_{22} and w22 are (k + 1) x (k + 1) submatrices.

2) Pseudo-Likelihood Ratio Statistics

To obtain the confidence intervals for the parameter of interest τ, pseudo-likelihood ratio statistic is defined and its limiting distribution is derived under the hypothesis H0 : $\tau = \tau_0$.
Let λ_0 be the true value of λ, $\hat{\lambda}_0$ 0 be the maximum pseudo-likelihood estimate of λ under the null model and $\hat{\lambda}$ be the maximum pseudo-likelihood estimate of λ under the full model.
Define:

$$D_0 = -2\left(\hat{\ell}(\lambda_0) - \hat{\ell}(\hat{\lambda}_0)\right)$$

$$D_1 = -2\left(\hat{\ell}(\lambda_0) - \hat{\ell}(\hat{\lambda})\right)$$

$$Q_n(\lambda) = \Delta_n^{1/2}\left\{\Omega_n^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & w_{22}^{-1} \end{pmatrix}\right\}\Delta_n^{1/2}$$

$$a_n^2(\lambda) = \{\text{tr}(Q_n(\lambda))\}^{-1}$$

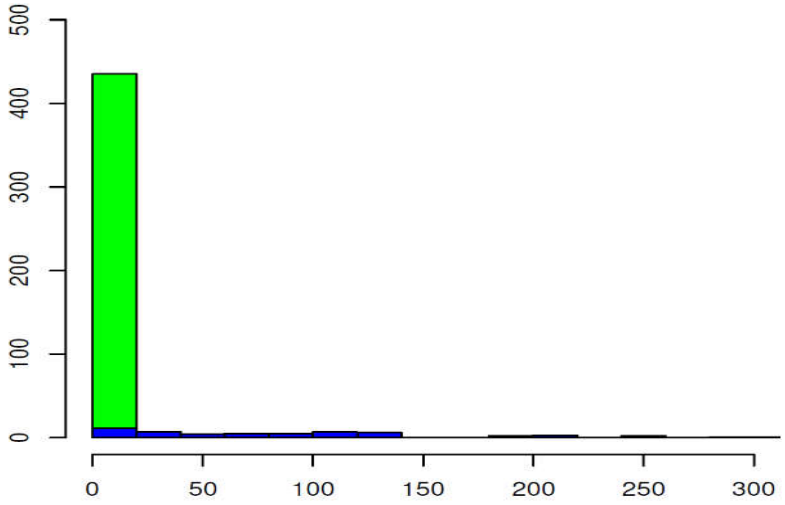
where tr(.) is the trace operator. Pseudo-likelihood ratio statistic for τ at $\tau = \tau_0$ is defined by:

$$D(\tau_0) = a_n^2(D_1 - D_0)$$

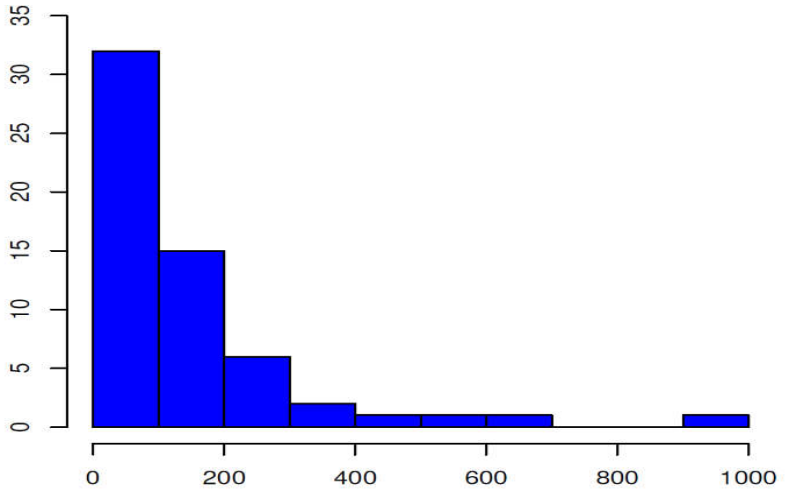
$$D(\tau_0) = 2a_n^2\{\hat{\ell}(\hat{\lambda}) - \hat{\ell}(\hat{\lambda}_0)\}$$

The applied part:

In this section, data on the cost of hospitalization (in dollars) for hospitalized patients are discussed. Details about the data set are described, the data set includes 483 individuals with 5 different variants. These included hospitalization cost, age, sex (1 = male and 0 = female), race (1 = Caucasian and 0 = African American), and general health status as measured by the SF-36. SF-36 is a short form of health survey for patient validation. It measures eight different health-related criteria. These include bodily functions, social function, physical role, emotional role, mental health, energy, pain and general health perceptions. Each parameter is scored from 0 to 100. The response variable (cost of recovery) includes 59 observations with non-zero positive values and 424 observations with zero values. Therefore, approximately 12% of the observations are non-zero and approximately 88% of the observations are zero for the response variable. Figure 4-1 shows the hospitalization cost graph for 483 patients. It is clear that the graph is strongly skewed to the right due to the large proportion of zero values and the additional skew in the positive acceptance cost.

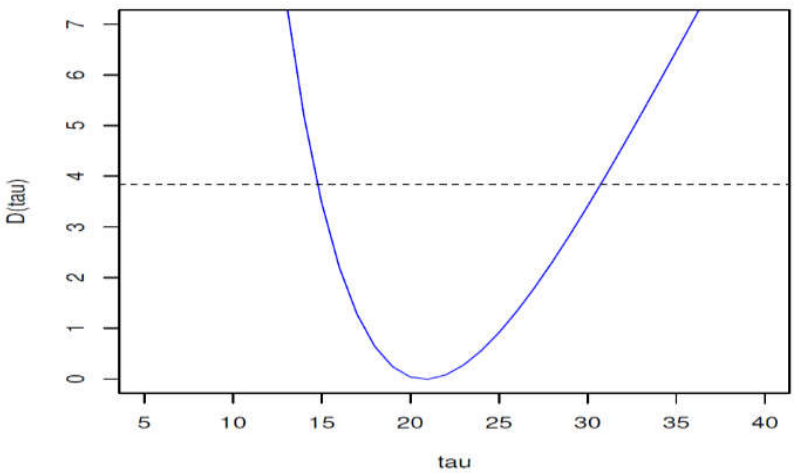


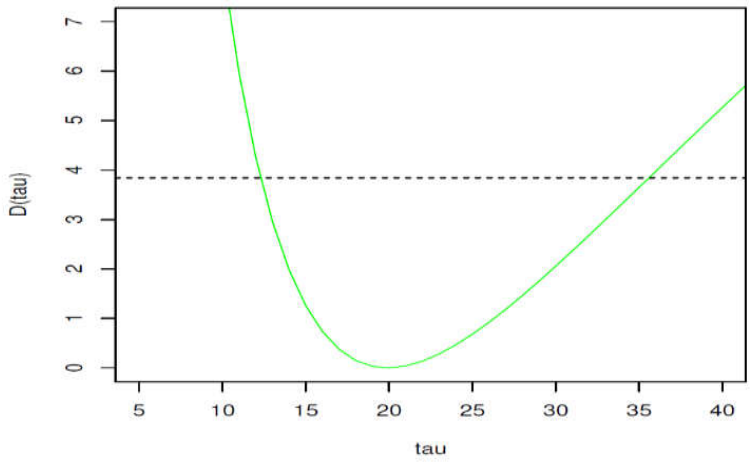
Since the distribution of non-zero observations is not clear in the above graph, a separate histogram for non-zero observations is shown in the figure below. It can be seen that the histogram is strongly skewed to the right for non-zero observations. From the graph, we can see that the non-zero responses have a log-normal distribution.



The table below shows that the pseudo-maximum likelihood method gives a shorter confidence interval than the two other popular methods. The plots below show the relative maximum potential and estimated likelihood potential, respectively.

$\hat{t}$	upper bound	lower limit	Method
547 .17	.04321	05114 .	Maximum pseudo-likelihood ratio
092 .15	365 .18	819 .11	Maximum likelihood ratio
492 .16	476 .19	508 .13	Maximum likelihood estimated MLE





The table below represents the criterion values of the absolute mean relative error for the three estimation methods:

Method	Point estimate	confidence interval
Maximum pseudo-likelihood ratio	0.0107	0.0450
Maximum likelihood ratio	0.042	0.0496
Maximum likelihood estimated MLE	0.074	0.088

Through the results obtained and shown in the tables above, it was shown that the Maximum pseudo-likelihood ratio method is the best because it has the least absolute difference for points and periods.

Conclusions and Recommendations:

- The zero mixed regression model (ZIM) was developed as generalized linear models in complex sampling designs via a two-component mixture model where the probability distribution of the non-zero component is assumed to be parametric.
- ZIM regression model is applicable in both continuous type and discrete type data.
- An issue of complex probability sample designs not previously addressed in ZIM regression is addressed, and is applicable to samples who have a large proportion of zeros.
- Maximum pseudo-likelihood method is the best method for estimating the parameters of the studied model.
- ZIM regression is applied to the commonly used models: normal, lognormal, and gamma, which means that the non-zero component follows these probability distributions. The proposed pseudo-maximum possibility method is derived for all three models (normal, lognormal, and gamma) and detailed calculations of the practical application of each model are provided in this thesis.
- In the future, the application of ZIM regression could be extended to counting data. For example, the response variable follows a Poisson or negative binomial distribution, ZIM regression extends to logistic regression with binary responses.

REFERENCES

[1]. Abadir, K. M. and Magnus, J. R. (2005). Matrix algebra, Cambridge University Press.
[2]. Chai, H. S. and Bailey, K. R. (2008). Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero. Statistics in Medicine, 27, 3643-3655.
[3]. Chen, H., Chen, J., and Chen, S. (2010). Con_dence intervals for the mean of a population containing many zero values under unequal-probability sampling. Canadian Journal of Statistics, 38, 582-597.
[4]. Cui, Y. and Yang, W. (2009). Zero-inated generalized poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros. Journal of Theoretical Biology, 256, 276-285.
[5]. Dobbie, M. J. and Welsh, A. H. (2001). Modelling correlated zero-inflated count data. Australian and New Zealand Journal of Statistics, 43, 431-444.
[6]. Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random e_ects:A case study. Biometrics, 56, 1030-1039.
[7]. Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: A case study. Biometrics, 56, 1030-1039.
[8]. Lee, A. H., Wang, K., and Yau, K. K. W. (2001). Analysis of zero-inflated poisson data incorporating extent of exposure. Biometrical Journal, 43, 963-975.
[9]. Murray, M. D., Harris, L. E., Overhage, J. M., Zhou, X., Eckert, G. J., Smith, F. E.,Buchanan,N. N., Wolinsky, F. D., McDonald, C. J., and Tierney, W. M. (2004). Failure of computerized treatment suggestions to improve health outcomes of outpatients with uncomplicated hypertension: Results of a randomized controlled trial. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 24, 324-337.

[10]. Ridout, M., Hinde, J., and Dem_etrio, C. G. B. (2001). A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. Biometrics, 57, 219-223.
[11]. Rizzo, M. L. (2007). Statistical Computing with R, Chapman & Hall/CRC.
[12]. Welsh, A. H. and Zhou, X. H. (2006). Estimating the retransformed mean in a heteroscedastic two-part model. Journal of Statistical Planning and Inference, 136, 860-881.
[13]. Welsh, A. H., Cunningham, R. B., Donnelly, C. F., and Lindenmayer, D. B. (1996). Modelling the abundance of rare species: Statistical models for counts with extra zeros. Ecological Modelling, 88, 297-308.
[14]. Yau, K. K. W. and Lee, A. H. (2001). Zero-inflated poisson regression with random effects to evaluate an occupational injury prevention programme. Statistics in Medicine, 20, 2907-2920.
[15]. Yau, K. K. W., Wang, K., and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biometrical Journal, 45, 437-452.
[16]. Zhou, X. and Cheng, H. (2008). A computer program for estimating the retransformed mean in heteroscedastic two-part models. Computer Methods and Programs in Biomedicine, 90, 210-216.