International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:7, 2013

Zero Truncated Strict Arcsine Model

Y. N. Phang and E. F. Loh

***Abstract*—**The zero truncated model is usually used in modeling count data without zero. It is the opposite of zero inflated model. Zero truncated Poisson and zero truncated negative binomial models are discussed and used by some researchers in analyzing the abundance of rare species and hospital stay. Zero truncated models are used as the base in developing hurdle models. In this study, we developed a new model, the zero truncated strict arcsine model, which can be used as an alternative model in modeling count data without zero and with extra variation. Two simulated and one real life data sets are used and fitted into this developed model. The results show that the model provides a good fit to the data. Maximum likelihood estimation method is used in estimating the parameters.

***Keywords*—**Hurdle models, maximum likelihood estimation method, positive count data.

I. Introduction

ZERO truncated model is commonly used in modeling count data without zero. Examples of these data are number of hospital stay; number of times a voter has voted among the people who has voted during the general election. The most popular ones are zero truncated Poisson (ZIP) [1]-[3] and zero truncated negative binomial models (ZTNB). ZTNB is recommended for count data without zero but with extra variations. It has been used to analyze the abundance of rare species [4], [5]. Reference [6] analyzed the overdispersed positive count data of ischemic stroke hospitalizations using truncated negative binomial mixed regression model. Reference [7] applied the truncated model to a number of recreational fishing trips taken from a sample of Alaskan fishermen. Reference [8] applied it to the [9] data set on contract strikes. Reference [10] discussed zero-truncated Poisson-Lidley distribution and its applications. Reference [11] gave modeling Sage data with a truncated gamma-Poisson model. Another application of zero truncated models is it can be used as a building block for hurdle models. In this study, we developed a new model named zero truncated strict arcsine model which can serve as an alternative model in modeling data without zero and with extra variations. We fit the developed model into one real life and two simulated data sets. The chi-square value shows that it provides a good fit to the three data sets. Maximum likelihood estimation method is used in estimating the parameters in this model. The properties and characteristics of strict arcsine and zero truncated strict arcsine models are provided in Section II. Section III studies the parameter estimation method used. Section IV discusses the application and the results of this developed model to three data sets. A short concluding remark is given in Section V.

Y. N. Phang is with the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Kampus Bandaraya Melaka,Malaysia (e-mail: phang@melaka.uitm.edu.my).

E. F. Loh is with the Academy of Language Studies, Universiti Teknologi MARA, Kampus Bandaraya Melaka, Malaysia (e-mail: david_loh@ melaka.uitm.edu.my).

II. Properties of the Distributions

A. The Strict Arcsine Distribution

The SA distribution is introduced by [12]. Reference [13] studied the properties of the strict arcsine distribution and found that the SA distribution is overdispersed, skewed to the right and leptokurtic.

The probability mass function of SA is given by

$$\Pr_{SA}(x) = \frac{A(x;\alpha)}{x!}\, p^x \exp\{-\alpha \arcsin(p)\}, \quad x = 0, 1, 2, \ldots\ldots \quad (1)$$

where $0< \alpha$, $0 < p < 1$, and $A(x; \alpha)$ is defined as

$$A(x;\alpha) = \begin{cases} \prod_{k=0}^{z-1}(\alpha^2 + 4k^2) & \text{if } x = 2z; \text{ and } A(0, \alpha) = 1 \\ \alpha \prod_{k=0}^{z-1}(\alpha^2 + (2k+1)^2) & \text{if } x = 2z+1; \text{ and } A(1, \alpha) = \alpha \end{cases} \quad (2)$$

The recurrence formula of SA is

$$\Pr(x+1) = \frac{A(x+1;\alpha)}{A(x;\alpha)} \cdot \frac{p}{x+1} \Pr(x), \quad x = 0, 1, 2, \ldots\ldots \quad (3)$$

with

$$\Pr(0) = \exp(-\alpha \arcsin(p)) \text{ and } \Pr(1) = \alpha \exp(\text{-arcsin(p)}).$$

The likelihood L is given by

$$L_{SA} = \prod_{k=0}^{x} \Pr_{SA}(k)^{F_k}, \quad x = 0, 1, 2, \ldots\ldots \quad (4)$$

and the log-likelihood is

$$\ln L_{SA} = \sum_{k=0}^{x} F_k \ln \Pr_{SA}(k) \quad (5)$$

The likelihood score functions are given below

$$\frac{\partial \ell_{SA}}{\partial \alpha} = \sum_{k=0}^{x} F_k \frac{\partial \log A(k,\alpha)}{\partial \alpha} - \arcsin(p) \quad , \; x = 0, 1, 2, \ldots \qquad (6)$$

where

$$\frac{\partial \log A(x,\alpha)}{\partial \alpha} = \begin{cases} \sum_{k=0}^{z-1} \frac{2\alpha}{(\alpha^2 + 4k^2)} \; , \text{ if } x = 2z; \text{ and } \frac{\partial \log A(0,\alpha)}{\partial \alpha} = 0 \\ -\frac{1}{\alpha^2} \sum_{k=0}^{z-1} \frac{2(2k+1)^2 - 2\alpha^2}{\left[\alpha^2 + (2k+1)^2\right]^2} \; , \text{ if } x = 2z+1; \text{ and } \frac{\partial \log A(1,\alpha)}{\partial \alpha} = 1 \end{cases}$$

$$\frac{\partial \ell_{SA}}{\partial p} = \sum_{k=0}^{x} F_k \left[\frac{k}{p} - \frac{\alpha}{\sqrt{1-p^2}} \right] , \qquad x = 0, 1, 2, \ldots \qquad (7)$$

$$\mu = \alpha p (1-p^2)^{-1/2} = E[X]$$

$$\alpha = \frac{\mu (1-p^2)^{-1/2}}{p} \qquad (8)$$

$$\sigma^2 = \alpha p (1-p^2)^{-3/2} = VAR[X]$$

B. The ZeroTruncated Strict Arcsine Model

The zero truncated strict arcsine model (ZTSA) is developed by dividing the probability mass function of strict arcsine by 1-P(0).

The PMF for zero truncated strict arcsine model is given by

$$P_{ZTSA}(Y = y) = \Pr_{SA}(y) / (1 - \Pr_{SA}(0)) , \quad y = 1, 2, 3, \ldots \qquad (9)$$

The likelihood L is given by

$$L_{ZTSA} = \prod_{k=1}^{x} \Pr_{ZTSA}(k)^{F_k} , x = 1, 2, \ldots \qquad (10)$$

and the log-likelihood is

$$\ln L_{ZTSA} = \sum_{k=1}^{x} F_k \ln \Pr_{ZTSA}(k) \qquad (11)$$

III. Parameter Estimation

The maximum likelihood method together with simulated annealing [14], a global optimization routine, is used to obtain the unknown parameters for this model. Simulated annealing has been used in various combinatorial optimization problems. It is particularly good in solving problems involved circuit design [15]. The applied concept is based on the manner in which liquid freeze or metals recrystalize in the process of annealing. At high temperature, molecules are free to move, but the mobility of the molecules drops as the temperature decreases and the molecules tend to line themselves up in a rigid structure which in fact is a stage of minimum energy. The advantage of this approach is that derivatives of the likelihood function are not needed.

We find the parameter estimates for a real life and two simulated data sets using the above-mentioned method. Table I shows the real life data set from [6]. Table II shows simulated data set with $p = 0.40$, $\alpha = 8.00$, and sample size=500. Table III shows the data set simulated with $p = 0.70$, $\alpha = 5.00$ and sample size=1000.

TABLE I
Empirical Distribution of Ischaemic Stroke Hospitalizations

	Observed frequency	Expected frequency
1	554	553.88
2	99	96.30
3	18	21.59
4	4	4.60
5	2	1.22
6	1	0.41

-loglikelihood =407.29

$\chi^2 = 2.10$

$\hat{p}_{ZTSA} = 0.3361$, $\hat{\alpha}_{ZTSA} = 1.0347$

TABLE II
Simulation of Truncated Strict Arcsine with $p = 0.40$, $\alpha = 8.0$ and Sample Size=500.

	Simulated frequency	Expected frequency
1	62	62.02
2	102	99.86
3	107	108.53
4	89	90.62
5	62	62.62
6	38	37.60
7	21	20.31
8	11	10.11
9	5	4.72
10	2	2.10
11	1	1.51
12	0	0
	-loglikelihood	1497.57
	χ^2	0.40

$\hat{p}_{ZTSA} = 0.3605$ $\hat{\alpha}_{ZTSA} = 8.9318$

IV. Results

The newly developed zero truncated strict arcsine model provides very good fit for the three data sets. The chi-values χ^2

= 2.10 for the first real life data set (Table I), $\chi^2 = 0.40$ for the second simulated data set(Table II) and χ^2 =2.47 for the third simulated data set (Table III) indicate that zero truncated strict arcsine model can be used as an alternative model in modeling overdispersed positive count data. The results also show that the estimation method used provides very good estimates which are very close to the set parameters.

TABLE III
SIMULATION OF TRUNCATED STRICT ARCSINE WITH $p = 0.70$, $\alpha = 5.00$ AND SAMPLE SIZE=1000

	Simulated frequency	Expected frequency
1	74	72.63
2	132	129.31
3	157	159.04
4	154	156.43
5	131	133.71
6	103	104.44
7	77	76.92
8	55	54.49
9	39	37.59
10	27	25.48
11	18	17.05
12	12	11.31
13	8	7.46
14	6	4.90
15	4	3.20
16	3	6.03
	-loglikelihood	2355.75
	χ^2	2.47

$$\hat{p}_{ZTSA} = 0.6778, \quad \hat{\alpha}_{ZTSA} = 5.2538$$

V. CONCLUDING REMARKS

Zero truncated models are commonly used in modeling positive count data. It is used as a base in developing hurdle model. Zero truncated strict arcsine model is found to be suitable in modeling positive count data with extra variations. The results indicate that this model fits well into two simulated and one real life data sets. This study also shows that maximum likelihood estimation method together with simulated annealing is a good choice in estimating the parameters in this model.

ACKNOWLEDGMENT

This research is funded by the Fundamental Research Grant Scheme(FRGS), Ministry of Higher Education Malaysia, that is managed by the Research Management Institute, Universiti Teknologi MARA (600-RMI/ST/FRGS 5/3/Fst (217/2010).

REFERENCES

[1] G. J. McLachlan, "On the EM algorithm for overdispersed count data," Statistical Methods in Medical Research 6, 1997,76-98
[2] J. N. S. Mathews and D. R. Appleton, "An application of the truncated Poisson distribution to immunogold assay," Biometrics, 49, 1993, 617-621.
[3] M. D. Creel andj. B. Loomis, " Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in California," American Journal of Agricultural Economics, 72(2), 1990, 434-441.
[4] A. H. Welsh, R. B. Cunningham, C. G. Donelty, D. B. Lindenmayer, "Modeling the abundance of rare species: statistical models for counts with extra zeros". Ecological Modeling 88, 1996, 297-308.
[5] A. H. Welsh, R. B. Cunningham, R. L. Chambers, "Methodology for estimating the abundance of rare animals: seabird nesting on north eash Herald Cay". Biometrics, 56, 2000, 22-30.
[6] A. H. Lee, k. Wang, K. K. W. Yau and P. J. Somerford, " Truncated negative binomial mixed regression modelling of ischaemic stroke hospitalizations," Statistics in Medicine, 22, 2003, 1129-1139.
[7] J. T. Grogger and R. T. Carson, ' Models for truncated counts." Journal of Applied Econometrics, 1991, 6, 225-238.
[8] S. Gurmu, " Tests for detecting overdispersion in the positive Poisson regression model," Journal of the Business and Economic Statistics; 1991,9, 215-222.
[9] J. Kennan, "The duration of contract strikes in U. S. manufacturing," Journal of Econometric, 1985, 28, 5-28.
[10] M. E. Ghitany, D. K. Al-mutairi and S. Nadarajah, " Zero-truncated Poisson-Lindley distribution an its application," Mathematics and Computers in Simulation, 2008, 79, 279-287.
[11] H. H. Thygesen and A. Zwinderman, "Modeling Sage data with a truncated gamma-Poisson model," BMC Bioinformatics 2006, 7, 157.
[12] G. Letac and M. Mora, "Natural real exponential families with cubic variance functions," The Annals of Statistics, 18, 1990, 1-37.
[13] C. C. Kokonendji and M. Khoudar, "On Strict Arcsine Distribution" Communications in Statistics. Theory Methods,33(5), 2004, pg993-1006
[14] W. L. Goffe., G. Ferrier and, J. John Rogers, "Global optimization of statistical functions with simulated annealing. Journal of Econometric, 60 (1/2), 1994, 65-100.
[15] S. Kirkpatrick, C. D. Gelatt Jr., M. P. Vecchi, "Optimization by Simulated Annealing",*Science*, 220, 4598, 671-680, 1983.