International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:6, No:12, 2012

Zero Inflated Strict Arcsine Regression Model

Y. N. Phang, and E. F. Loh

***Abstract*—**Zero inflated strict arcsine model is a newly developed model which is found to be appropriate in modeling overdispersed count data. In this study, we extend zero inflated strict arcsine model to zero inflated strict arcsine regression model by taking into consideration the extra variability caused by extra zeros and covariates in count data. Maximum likelihood estimation method is used in estimating the parameters for this zero inflated strict arcsine regression model.

***Keywords*—**Overdispersed count data**,** maximum likelihood estimation, simulated annealing.

I. Introduction

COUNT data with extra variability are commonly found in various disciplines which include biometrics, economics, agriculture, social sciences, transportation engineering, epidemiology and medicine. The extra zeroes or covariates are usually the contributing factors to the extra variability in count data. Many zero inflated models such as zero inflated Poisson, zero inflated negative binomial, zero inflated generalized Poisson are developed and extensively used in modeling overdispersed count data with excess zeros [1]-[5] . Zero inflated inverse trinomial [6] and zero inflated strict arcsine model [7] are newly developed models which can be used as alternative models in modeling overdispersed count data, but have yet to be researched and used extensively. Models which are developed in modeling overdispersed count data caused by extra zeros and also covariates are zero inflated Poisson regression, zero inflated negative binomial regression and zero inflated generalized Poisson regression [8]-[16]. Strict arcsine (SA) model is designed in modeling data with extra variations, skewed to the left and leptokurtic. Its variance is a cubic variance function of mean. Strict arcsine model was introduced by Mora [17]. Kokonendji [8] compared the strict arcsine distributiuon with Poisson, negative binomial, Poisson inverse Gaussian, and generalized Poisson models by using the moment method to estimate the parameters. Marque and Kokonendji [18] proposed a strict arcsine regression model for regression analysis of count. They applied the model to data concerning cardiovascular mortality among the elderly. Phang and Loh [8] developed zero inflated strict arcsine (ZISA) model and fitted it to both a simulated and a real life data set. The study showed that this developed model can be used as an alternative model in modeling overdispersed count data. In this paper, we developed zero inflated strict arcsine regression model to model overdispersed data where the variations are caused by extra zeros and covariates. We apply the maximum likelihood estimation method through a global optimization routine to estimate the parameters for ZISA regression model.

Section II of the paper discusses the properties of the SA, ZISA, and ZISA regression models, Section III explains fitting the two simulated data sets with the developed zero inflated strict arcsine regression model and in section IV, we apply maximum likelihood estimation in estimating the parameters. The results are discussed in section V. Section VI provides a short conclusion.

Y. N. Phang is with 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 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 was introduced by Letac and Mora [17]. Kokonendji [8] 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:

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

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

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

The recurrence formula of SA is:

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

with

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

The likelihood L is given by:

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

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:

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

where

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

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

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

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

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

B. The Zero Inflated Strict Arcsine Model

Phang and Loh [8] developed the zero inflated strict arcsine model and fitted it to both a simulated and a real life data sets. The study showed that this model can be used in modeling count data with excess zeros.

The probability mass function for ZISA is given by:

$$P_{ZISA}(Y = 0) = \omega + (1-\omega)\exp(-\alpha \arcsin(p))$$

$$P_{ZISA}(Y = y) = (1-\omega)\Pr_{SA}(y), \quad y = 1,2,3,\ldots \quad (9)$$

The likelihood L is given by

$$L_{ZISA} = \prod_{k=0}^{x} \Pr_{ZISA}(k)^{F_k}, \qquad x = 0, 1, 2, ... \quad (10)$$

and the log-likelihood is:

$$\ln L_{ZISA} = \sum_{k=0}^{x} F_k \ln \Pr_{ZISA}(k). \quad (11)$$

C. The Zero Inflated Strict Arcsine Regression Model

The likelihood L and log-likelihood are, respectively, given by;

$$L_{ZISAREG} = \prod_{k=0}^{x} P_{ZISAREG}(k)^{F_k}, \qquad x = 0, 1, 2, \ldots. \quad (12)$$

and

$$\ln L_{ZISAREG} = \sum_{k=0}^{k} F_k \ln \Pr_{ZISAREG}(k) \quad (13)$$

where $\Pr_{ZISAREG}(k)$ is obtainable from equation (9) by replacing α with equation (8) where $\mu = \exp(\beta_0 + \beta_1 x_i)$, i=1, 2… We assume that only log of the mean is depend on the covariates, ω is a constant which does not depend on the covariates.

III. Model Fitting

A. Simulation

The newly developed ZISA regression model is fitted to two simulated data sets where data set 1, sample size, n=2000, p=0.7, β_0=1.0, β_1=0.5 and ω=0.10 and data set 2, sample size, n=5000, p=0.85, β_0=-2.5, β_1=-025 and ω=0.15. The results are shown in Table I and Table II.

TABLE I
SIMULATION OF ZISA REGRESSION: p=0.70, ω=0.10, β_0=1.0, β_1=0.5

	Observed frequency			Expected frequency		
	X_A	X_B	TOTAI	X_A	X_B	TOTAL
0	205	126	331	198.94	131.62	330.57
1	203	83	286	193.36	88.61	281.98
2	198	133	331	195.19	137.99	333.18
3	144	149	293	147.17	150.50	297.67
4	94	135	229	98.23	134.11	232.34
5	59	109	168	62.50	106.23	168.73
6	37	81	118	39.06	78.47	117.53
7	23	58	81	24.33	55.57	79.90
8	14	41	55	15.17	38.35	53.52
9	9	28	37	9.50	26.04	35.55
10	5	19	24	5.98	17.52	23.50
11	3	13	16	3.79	11.71	15.50
12	2	8	10	2.41	7.80	10.21
13	1	6	7	1.54	17.52	6.73
14	1	4	5	0.99	3.44	4.44
15	2	7	9	1.82	6.83	8.65

-loglikelihood =4425.91

$\chi^2 = 0.44$

$\hat{p}_{ZISAREG} = 0.7006, \quad \hat{\varpi} = 0.1032, \quad \hat{\beta}_{0ZISAREG} = 1.0400,$

$\hat{\beta}_{1\,ZISAREG} = 0.4334$

Mean = 3.2315
Variance =7.9659

TABLE II
SIMULATION OF ZISA REGRESSION: p=0.70, ω=0.15, β_0=-2.5, β_1=-0.25

	Observed frequency			Expected frequency		
	X_A	X_B	TOTAL	X_A	X_B	TOTAL
0	4786	4832	9618	4784.82	4834.91	9619.73
1	175	138	313	176.44	136.21	312.65
2	4	2	6	3.82	2.25	6.07
3	21	17	38	20.88	16.10	36.98
4	1	1	2	0.90	0.53	1.43
5	7	5	12	6.66	5.13	11.79
6	0	0	0	0.34	0.20	0.54
7	3	2	5	2.810	2.16	4.97
8	0	0	0	0.160	0.09	0.25
9	1	1	2	1.350	1.04	2.39
10	2	2	4	1.820	1.38	3.20

-loglikelihood =1897.50

$\chi^2 = 1.31$

$\hat{p}_{ZISAREG} = 0.8456, \quad \hat{\varpi} = 0.1410, \quad \hat{\beta}_{0ZISAREG} = -0.5249$

$\hat{\beta}_{1\,ZISAREG} = -0.2710$

Mean = 0.06
Variance =0.1782

B. Parameter Estimation

In this study, maximum likelihood estimation method is used in estimating the parameters for the newly developed model. This method is used because it has desirable mathematical and optimality properties such as , it becomes minimum variance unbiased estimators as the sample size increases and the asymtotically normal characteristic under certain regularity condition can be used to generate confidence interval and hypothesis tests for the parameters. Simulated annealing[19], which is a global optimization routine is applied in attaining the parameters. The advantage of this approach is that derivatives of the likelihood function are not needed. To check that a global optimum is achieved, various seeds from the random generator RANMAR and temperature reduction factor are used. Convergence is evaluated at each step by the difference between two function values lower than 10^{-6} (which is the convergence criteria). The maximum likelihood estimates are validated by substituting these estimates into the likelihood score equations.

We obtained the parameter estimates for the two simulated data sets using the above-mentioned method. Table I and II show the fitting of the two simulated data sets and the estimated parameters..

IV. RESULTS

The results show that this newly developed zero inflated strict arcsine regression model provide good fit to the two simulated data set with small chi-square values which are 0.44 for the first data set and 1.31 for the second data set. The estimated parameters obtained using maximum likelihood estimation method is close to the values set in simulating the data.

V. CONCLUDING REMARKS

The model is developed by considering one explainable variable. Only log of the mean is assumed to depend on the covariate. Future research may take into consideration more variables. The small chi-square values for the two simulated data sets indicate that this model can be used as an alternative model in modeling overdispersed count data where the extra variability are caused by extra zeros and explainable variables. The estimated parameters also show that the estimating method used in estimating the parameters for ZISA regression model is valid.

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] M. Ridout, C. G. B. Demetrio, and J. Hinde, "Models for count with many zeros", in: Invited Paper Presented at the 19th International Biometric Conference, CapeTown, South Africa, 1998, 178.

[2] S. Gurmu and P. K. Trivedi, "Excess zeros in count models for recreational trips", Journal of Business and Economic Statistics, 14, 1996, 469-477.

[3] K. K. W. Yau and K. C. H. Yip, "On modeling claim frequency data in general insurance with extra zeros". Insurance: Mathematics and Economics Vol. 36, Issue 2, 2005, 153-163.

[4] M. L. Dalrymple, I. L. Hudson, and R. P. K. Ford, "Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS", Computational Statistics & Data Analysis, 41, 2003, 491-504

[5] R. Winkelmann, Econometric Analysis of Count Data. Springe Verlag, Berlin, Heidelberg, 2008.

[6] R. Winkelmann, "Health care reform and the number of doctor visits – An econometric analysis," Journal of Applied Econometrics 19, 2004, 455-472

[7] Y. N. Phang, "Statistical inference for a family of discrete distribution with cubic variance functions", Unpublished PhD thesis, University Malaya, Malaysia, 2007

[8] Y. N. Phang, and E. R. Loh. Proceedings: IASC 2008: Joint Meeting of 4th World Conference of the IASC and 6th Conference of the IASC and 6th conference of the Asian Regional Section of the IASC on Computational Statistic and Data Analysis, Yokohama, Japan, 2008

[9] D.Lambert, "Zero-inflated Poisson regression, with an application to random defects in manufacturing". Technometrics, 34, 1992, 1-14

[10] A. C. Cameron and P. K. Trivedi, "Regression analysis of count data". Cambridge University Press. 1998

[11] D. B. Hall, "Zero inflated Poisson and binomial with random effects: a case study," Biometrics, 56, 2000, 1030-1039

[12] D. Bohning, E. Dietz, P. Schlattman, L. Mendonca and U. Kirchner, "The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology". Journal of the Royal Statistical Society, Series A, 1999,162-209

[13] K. K. W. Yau, K. Wang, and A. H.and Lee, "Zero-inflated negative binomial mixed regression Modeling of overdispersed count data with extra zeros". Biometrical Journal 45, 4, 2003, 437-452.

[14] F. Famoye and P. S. Karan, " Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data," J of Data Science 4, 2006, 117-130.

[15] A. C. Mehmet, "Zero-inflated regression models for modeling the effect of air pollutants on hospital admissions", Polish Journal of Environment Studies, Vol. 21, No. 3, 2012, 565-568.

[16] B. M. Golam Kibria, " Applicaations of some discrete regression models for count data", Pakistan Journal of Statistics and Operation research, Vol11 No. 1, 2006, 1-16.

[17] G. Letac and M. Mora, "Natural real exponential families with cubic variance functions," The Annals of Statistics, 18, 1990, 1-37.

[18] C. C. Kokonendji and M. Khoudar, "On Strict Arcsine Distribution" Communications in Statistics. Theory Methods,33(5), 2004, pg993-1006

[19] 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