Brazilian Journal of Probability and Statistics
2012, Vol. 26, No. 3, 306–312
DOI: 10.1214/10-BJPS137

Identifiability of zero-inflated Poisson models

Chin-Shang Li
University of California

Abstract. Zero-inflated Poisson (ZIP) models, which are mixture models, have been popularly used for count data that often contain large numbers of zeros, but their identifiability has not yet been thoroughly explored. In this work, we systematically investigate the identifiability of the ZIP models under a number of different assumptions. More specifically, we show the identifiability of a parametric ZIP model in which the incidence probability $p(x)$ and Poisson mean $\lambda(x)$ are modeled parametrically as $p(x) = \exp(\beta_0 + \beta_1 x)/[1 + \exp(\beta_0 + \beta_1 x)]$ and $\lambda(x) = \exp(\alpha_0 + \alpha_1 x)$ for x being a continuous covariate in a closed interval. A semiparametric ZIP regression model is shown to be identifiable in which (i) $p(x) = \exp(\beta_0 + \beta_1 x)/[1 + \exp(\beta_0 + \beta_1 x)]$ and $\lambda(x) = \exp[s(x)]$, (ii) $p(x) = \exp[r(x)]/\{1 + \exp[r(x)]\}$ and $\lambda(x) = \exp(\alpha_0 + \alpha_1 x)$, or (iii) $p(x) = \exp[r(x)]/\{1 + \exp[r(x)]\}$ and $\lambda(x) = \exp[s(x)]$ for $r(x)$ and $s(x)$ being unspecified smooth functions.

1 Introduction

Numerous disciplines produce count data that contain many zeros; that is, biomedical studies, criminology, environmental economics, political science, and sociology. The zero-inflated Poisson (ZIP) distribution (Singh, 1963, Johnson, Kemp and Kotz, 2005), which is a mixture of a degenerate distribution at zero and a Poisson distribution, has been proposed to deal with the case in which the number of zeros exceeds expected for a regular Poisson distribution. One can view the ZIP model as adding structure to the regular Poisson model. This ZIP model allows separate consideration of those who are not at risk of an event of interest and those who are at risk of the event and may have the event several times during a specific time period (Dietz and Böhning, 1997). This mixture model has become the foundation of much methodological development in zero-inflated count data analysis. Some authors made statistical inferences on the existence of zero inflation in the count data (e.g., El-Shaarawi, 1985, van den Broek, 1995, Ridout, Hinde and Demétrio, 2001, Thas and Rayner, 2005). The seminal work on ZIP regression by Lambert (1992) modeled the parameters of interest simultaneously with linear predictors via appropriate link functions, which are described in case (iv) of the theorem in Section 3. Many authors adopted this basic modeling structure to make a number of important extensions. Hall and Zhang (2004) fit a marginal ZIP regression

Key words and phrases. Count data, semiparametric zero-inflated Poisson (ZIP) regression model.
Received November 2010; accepted December 2010.

model to clustered count data. Hall (2000), Hall and Zhang (2004) and Min and Agresti (2005) used mixed-effects ZIP regression models for repeatedly measured or cluster-correlated data. The ZIP regression models also have been applied to several important clinical studies (e.g., Böhning et al., 1999, Yau and Lee, 2001, Cheung, 2002, Lu, Lin and Shih, 2004). The identifiability of a ZIP model is very important because one can not obtain unique model parameter estimates without it. However, to the best of our knowledge, the identifiability of the ZIP models has not yet been studied in detail. Therefore, motivated by this, in this work we systematically investigate their identifiability.

The remainder of the paper is organized as follows. Section 2 introduces ZIP models. We explore their identifiability in Section 3. Some conclusions are provided in Section 4.

2 ZIP models

Let U be a latent binary variable that indicates an individual's risk state: $U = 0$ if the subject is not at risk of an event of interest (i.e., the subject is in the non-susceptible group); $U = 1$ if the subject is at risk of the event (i.e., the subject is in the susceptible group). Let $p = \Pr(U = 1)$ be an incidence probability. Let Y denote the event count, defined only when $U = 1$, which follows a Poisson distribution whose probability mass function is denoted by $f(y; \lambda|U = 1) = \Pr(Y = y; \lambda|U = 1) = e^{-\lambda}\lambda^y/y!$ for $y = 0, 1, \ldots$ and λ being the Poisson mean. When $U = 0$, we have the function $f(y; \lambda|U = 0) = I_{\{y=0\}}$, a degenerate distribution among individuals, who are not at risk of the event, for $y = 0, 1, \ldots$, where $I_{\{y=0\}}$ is the indicator function, which is 1 if $y = 0$; 0 otherwise. It is noted that $U = 1$ if $Y > 0$, and U is unobserved if $Y = 0$. If we now let $f(y; p, \lambda) = \Pr(Y = y; p, \lambda)$ be the unconditional or marginal distribution of Y, one can write the marginal distribution of Y as follows:

$$\begin{aligned} f(y; p, \lambda) &= \Pr(U = 0) f(y; \lambda|U = 0) + \Pr(U = 1) f(y; \lambda|U = 1) \\ &= (1 - p) I_{\{y=0\}} + p \frac{e^{-\lambda}\lambda^y}{y!}, \qquad y = 0, 1, \ldots. \end{aligned} \tag{2.1}$$

The marginal distribution $f(y; p, \lambda)$ is a mixture distribution, which has mixing proportions $1 - p$ and p with component distributions $f(y; \lambda|U = 0)$ and $f(y; \lambda|U = 1)$. It can be seen from (2.1) that the ZIP distribution is reduced to a regular Poisson distribution as the mixing proportion $p = 1$, which means all subjects at risk of the event of interest. Both p and λ can be dependent on the same or different covariates. We consider the case in which they depend on the same covariates. The mixing proportion p is dependent on covariates x, denoted by $p(x)$ and the degenerate function $f(y; \lambda|U = 0)$ is denoted by $f(y; \lambda(x), x|U = 0) = I_{\{y=0\}}$. The Poisson mean λ depends on x, denoted by $\lambda(x)$

and the probability mass function $f(y;\lambda|U=1)$ is denoted by $f(y;\lambda(x),x|U=1)=e^{-\lambda(x)}[\lambda(x)]^y/y!$. We assume that x is one-dimensional and not a constant within the sample. Hence, we consider the ZIP model as follows:

$$\begin{aligned} &f(y;p(x),\lambda(x),x)\\ &\quad=[1-p(x)]f\big(y;\lambda(x),x|U=0\big)\\ &\qquad+p(x)f\big(y;\lambda(x),x|U=1\big)\\ &\quad=[1-p(x)]I_{\{y=0\}}+p(x)\frac{e^{-\lambda(x)}[\lambda(x)]^y}{y!},\qquad y=0,1,\ldots. \end{aligned} \tag{2.2}$$

Let $\mathcal{X}$ be the design space that is assumed to be the closed interval $[a_0,a_1]$ for convenience. Assume $\mathcal{F}=\{f(y;\lambda(x),x|U=1):\lambda(x)>0$ for $x\in\mathcal{X}$ and $y=0,1,\ldots\}$ is the class of conditional event count distributions, that is, Poisson distributions, for individuals, who are at risk of the event. Denote the space of incidence probability functions by $\mathcal{P}=\{p(x):0<p(x)\le 1$ for $x\in\mathcal{X}\}$. Let

$$\begin{aligned} \mathcal{H}=\big\{&f(y;p(x),\lambda(x),x):f(y,p(x),\lambda(x),x)=[1-p(x)]I_{\{y=0\}}+p(x)\\ &f\big(y;\lambda(x),x|U=1\big),x\in\mathcal{X},y=0,1,\ldots,p(x)\in\mathcal{P},\\ &f\big(y;\lambda(x),x|U=1\big)\in\mathcal{F}\big\} \end{aligned}$$

denote the class of ZIP models.

Following Redner and Walker (1984), we define the identifiability of the ZIP model in (2.2) as follows.

Definition. The class of ZIP models $\mathcal{H}$ is identifiable if for any two members of $\mathcal{H}$ given by $f(y;p(x),\lambda(x),x)=[1-p(x)]I_{\{y=0\}}+p(x)f(y;\lambda(x),x|U=1)$ and $f(y;p^*(x),\lambda^*(x),x)=[1-p^*(x)]I_{\{y=0\}}+p^*(x)f(y;\lambda^*(x),x|U=1)$, then $f(y;p(x),\lambda(x),x)=f(y;p^*(x),\lambda^*(x),x)$ if and only if $p(x)=p^*(x)$ and $\lambda(x)=\lambda^*(x)$ for $x\in\mathcal{X}$ and $y=0,1,\ldots$.

3 Identifiability

We now investigate the identifiability of the ZIP model in (2.2) under a number of different assumptions. We consider the following cases. $p(x)$ is constant, or a parametric form or unspecified smooth function is given for the covariate effect on $p(x)$ via a logit link. $\lambda(x)$ is a constant parameter, or a parametric form or unspecified smooth function is given for the covariate effect on $\lambda(x)$ through a log link. The results are stated in the following theorem.

Theorem 3.1. *Let x be a continuous covariate in the design space $\mathcal{X}=[a_0,a_1]$, where $-\infty<a_0<a_1<\infty$.*

(i) *The ZIP model in* (2.2) *is identifiable if* $p(x)$ *is specified as* $\text{logit}[p(x)] = \log\{p(x)/[1-p(x)]\} = \beta_0 + \beta_1 x$ *and* $\lambda(x)$ *is specified as* $\log[\lambda(x)] = \alpha_0 + \alpha_1 x$.

(ii) *The ZIP model in* (2.2) *is identifiable if* $p(x)$ *is specified as* $\text{logit}[p(x)] = \log\{p(x)/[1-p(x)]\} = \beta_0 + \beta_1 x$ *and* $\log[\lambda(x)] = s(x)$, *where* $s(x)$ *is an unspecified smooth function instead of the linear form* $\alpha_0 + \alpha_1 x$.

(iii) *The ZIP model in* (2.2) *is identifiable if* $\text{logit}[p(x)] = r(x)$, *where* $r(x)$ *is an unspecified smooth function instead of the linear form* $\beta_0 + \beta_1 x$, *and* $\lambda(x)$ *is specified as* $\log[\lambda(x)] = \alpha_0 + \alpha_1 x$.

(iv) *The ZIP model in* (2.2) *is identifiable if* $\text{logit}[p(x)] = r(x)$, *where* $r(x)$ *is an unspecified smooth function instead of the linear form* $\beta_0 + \beta_1 x$, *and* $\log[\lambda(x)] = s(x)$, *where* $s(x)$ *is an unspecified smooth function instead of the linear form* $\alpha_0 + \alpha_1 x$.

Proof. To explore the identifiability of the ZIP model under each case, we need to show that $f(y; p(x), \lambda(x), x) = f(y; p^*(x), \lambda^*(x), x)$ if and only if $p(x) = p^*(x)$ and $\lambda(x) = \lambda^*(x)$ for $x \in \mathcal{X}$ and $y = 0, 1, \ldots$. Because the "if" part is clearly true in all cases, we focus on "only if." To show the "only if" part, suppose that $f(y; p(x), \lambda(x), x) = f(y; p^*(x), \lambda^*(x), x)$. With some algebra, one then can have the following ratio

$$\frac{p(x)}{p^*(x)} = \frac{I_{\{y=0\}} - e^{-\lambda^*(x)}[\lambda^*(x)]^y/y!}{I_{\{y=0\}} - e^{-\lambda(x)}[\lambda(x)]^y/y!}. \tag{3.1}$$

(i) If $p(x)$ is specified as $p(x) = \exp(\beta_0 + \beta_1 x)/[1 + \exp(\beta_0 + \beta_1 x)]$ and $\lambda(x)$ is specified as $\lambda(x) = \exp(\alpha_0 + \alpha_1 x)$, the left-hand side (LHS) of the ratio in (3.1) is dependent on x and the right-hand side (RHS) of the ratio in (3.1) depends on x and y. Thus, the ratio is a positive function of x, denoted by $c(x)$, for $x \in \mathcal{X}$. The quantities $p^*(x) = \exp(\beta_0^* + \beta_1^* x)/[1 + \exp(\beta_0^* + \beta_1^* x)]$ and $\lambda^*(x) = \exp(\alpha_0^* + \alpha_1^* x)$ must satisfy the following equations

$$e^{-\lambda^*(x)}[\lambda^*(x)]^y = [1 - c(x)]y! I_{\{y=0\}} + c(x)e^{-\lambda(x)}[\lambda(x)]^y, \qquad y = 0, 1, \ldots,$$
$$p^*(x) = p(x)/c(x).$$

If we can show $c(x) = 1$ for $x \in \mathcal{X}$, we then show the identifiability of the parametric ZIP regression model. Because $e^{-\lambda^*(x)}\lambda^*(x) = c(x)e^{-\lambda(x)}\lambda(x)$ as $y = 1$ and $e^{-\lambda^*(x)}[\lambda^*(x)]^2 = c(x)e^{-\lambda(x)}[\lambda(x)]^2$ as $y = 2$, it follows that $\lambda^*(x) = \lambda(x)$ for $x \in \mathcal{X}$ and, hence, $c(x) = 1$. The proof is complete.

(ii) When $p(x)$ is specified as $p(x) = \exp(\beta_0 + \beta_1 x)/[1 + \exp(\beta_0 + \beta_1 x)]$ and $\lambda(x) = \exp[s(x)]$, the LHS of the ratio in (3.1) depends on x and the RHS of the ratio in (3.1) is dependent on x and y. So, the ratio is a positive function of x, denoted by $c(x)$, for $x \in \mathcal{X}$. The quantities $p^*(x) = \exp(\beta_0^* + \beta_1^* x)/[1 + \exp(\beta_0^* +$

$\beta_1^* x)]$ and $\lambda^*(x) = \exp[s^*(x)]$ must satisfy the following equations

$$e^{-\lambda^*(x)}[\lambda^*(x)]^y = [1 - c(x)]y! I_{\{y=0\}} + c(x)e^{-\lambda(x)}[\lambda(x)]^y, \qquad y = 0, 1, \ldots,$$
$$p^*(x) = p(x)/c(x).$$

By using the arguments of proving case (i), we can show $c(x) = 1$ for $x \in \mathcal{X}$ and, hence, the identifiability of the semiparametric ZIP regression model. In addition, $p(x)$ is a constant when $\beta_1 = 0$, so, one can see that the ratio in (3.1) must be a positive constant c, which is not dependent on x or y because the LHS of the ratio is independent of x and the RHS of the ratio depends on x and y. One can show with the above arguments that $c = 1$. It then turns out that the semiparametric ZIP regression model is identifiable.

(iii) When $p(x) = \exp[r(x)]/\{1 + \exp[r(x)]\}$ and $\lambda(x)$ is specified as $\lambda(x) = \exp(\alpha_0 + \alpha_1 x)$, the LHS of the ratio in (3.1) depends on x and the RHS of the ratio in (3.1) is dependent on x and y. Thus, the ratio is a positive function of x, denoted by $c(x)$, for $x \in \mathcal{X}$. The quantities $p^*(x) = \exp[r^*(x)]/\{1 + \exp[r^*(x)]\}$ and $\lambda^*(x) = \exp(\alpha_0^* + \alpha_1^* x)$ must satisfy the following equations

$$e^{-\lambda^*(x)}[\lambda^*(x)]^y = [1 - c(x)]y! I_{\{y=0\}} + c(x)e^{-\lambda(x)}[\lambda(x)]^y, \qquad y = 0, 1, \ldots,$$
$$p^*(x) = p(x)/c(x).$$

By using the arguments of proving case (i), one can show that $c(x) = 1$ for $x \in \mathcal{X}$ and, hence, the semiparametric ZIP regression model is identifiable. Additionally, because $\lambda(x)$ is a constant when $\alpha_1 = 0$, the LHS of the ratio in (3.1) is dependent on x and the RHS of the ratio in (3.1) depends on y. Then, the ratio must be a positive constant c. By using the above arguments, we can show $c = 1$ and, hence, the identifiability of the semiparametric ZIP regression model.

(iv) One can show with the arguments of proving the previous cases that when $p(x) = \exp[r(x)]/\{1 + \exp[r(x)]\}$ and $\lambda(x) = \exp[s(x)]$ the semiparametric ZIP regression model is identifiable. □

4 Conclusions

We have demonstrated the identifiability of the ZIP models under some assumptions. Of note, for the parametric ZIP regression model, $p(x) = \exp(\beta_0 + \beta_1 x)/[1 + \exp(\beta_0 + \beta_1 x)]$ and $\lambda(x) = \exp(\alpha_0 + \alpha_1 x)$, given in case (i) of Theorem 3.1, it has three special cases: (1) $\alpha_1 = 0$ and $\beta_1 = 0$ (i.e., $p(x)$ and $\lambda(x)$ are constant); (2) $\beta_1 = 0$ (i.e., $p(x)$ is a constant); (3) $\alpha_1 = 0$ (i.e., $\lambda(x)$ is a constant). The class of mixtures of Poisson regression models with constant mixing probabilities is subsumed by the parametric ZIP models given in the first two special cases; its identifiability was discussed by, for example, Wang et al. (1996).

The semiparametric ZIP regression models, given in cases (ii)–(iv) of the Theorem 3.1, have been shown to be identifiable. The model in case (ii) of Theorem 3.1

with $p(x)$ being a constant is a semiparametric ZIP regression model, which is a special case of the semiparametric ZIP regression model by Li (2011). The semiparametric ZIP regression model can be utilized to test the lack of fit of a postulated parametric form for the covariate effect on the Poisson mean, for example, testing the linear effect of the covariate on the Poisson mean of the ZIP regression model in which $p(x)$ is a constant and $\lambda(x)$ is specified as $\log[\lambda(x)] = \alpha_0 + \alpha_1 x$.

For case (iii) of theorem, although the $r(x)$ can be estimated with any smoothing technique (e.g., splines, kernels, and local polynomial kernel regression), as mentioned in Li (2011), the $r(x)$ can be approximated by regression splines, for example, cubic splines with either the truncated power ("plus-function") basis or the B-spline basis (Schoenberg, 1946; Curry and Schoenberg, 1966; de Boor, 2001) that has better numerical properties than the truncated power basis. The expectation-maximization (EM) algorithm can be applied to easily estimate the model parameters. Moreover, the semiparametric ZIP regression model with $\lambda(x)$ being a constant, which is a special case of the semiparametric ZIP regression model in case (iii), can be applied to test the adequacy of a parametric form for the effect of the covariate on $p(x)$; for example, one can test the linear effect of the covariate on $p(x)$ of the ZIP regression model in which $p(x)$ is specified as $\text{logit}[p(x)] = \log\{p(x)/[1 - p(x)]\} = \beta_0 + \beta_1 x$ and $\lambda(x)$ is a constant.

For case (iv) of theorem, both $r(x)$ and $s(x)$ can be estimated by regression splines (e.g., cubic splines) or any other smoothing techniques. Furthermore, as a practical use, this semiparametric ZIP regression model can be used to test the lack of fit of any of the ZIP models given in the previous cases. In related work, Houseman, Coull and Shine (2004) discussed the identifiability of semiparametric negative binomial models for a time series of pathogen counts.

Acknowledgments

The author expresses his thanks to the editor and a referee whose helpful comments improved the presentation. This publication was made possible by Grant number UL1 RR024146 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research.

References

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L. and Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society, Ser. A* **162**, 195–209.

Cheung, Y. B. (2002). Zero-inflated models for regression analysis of count data: A study of growth and development. *Statistics in Medicine* **21**, 1461–1469.

Curry, H. B. and Schoenberg, I. J. (1966). On Pólya frequency functions IV: The fundamental splines and their limits. *Journal d'Analyse Mathematique* **17**, 71–107. MR0218800

de Boor, C. (2001). *A Practical Guide to Splines*, revised ed. *Applied Mathematical Sciences* **27**. New York: Spring-Verlag. MR1900298

Dietz, K. and Böhning, D. (1997). The use of two-component mixture models with one completely or partly known component. *Computational Statistics* **12**, 219–234.

El-Shaarawi, A. H. (1985). Some goodness-of-fit methods for the Poisson plus added zeros distribution. *Applied and Environmental Microbiology* **49**, 1304–1306.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics* **56**, 1030–1039. MR1815581

Hall, D. B. and Zhang, Z. (2004). Marginal models for zero inflated cluster data. *Statistical Modelling* **4**, 161–180. MR2062098

Houseman, E. A., Coull, B. A. and Shine, J. P. (2006). A nonstationary negative binomial time series with time-dependent covariates: Enterococcus counts in Boston harbor. *Journal of the American Statistical Association* **101**, 1365–1376. MR2307571

Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). *Univariate Discrete Distributions*, 3rd ed. Hoboken: Wiley. MR2163227

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.

Li, C. S. (2011). A lack-of-fit test for parametric zero-inflated Poisson models. *Journal of Statistical Computation and Simulation* **81**, 1081–1098. MR2783890

Lu, S. E., Lin, Y. and Shih, W. C. J. (2004). Analyzing excessive no changes in clinical trials with clustered data. *Biometrics* **60**, 257–267. MR2044122

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling* **5**, 1–19. MR2133525

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26**, 195–239. MR0738930

Ridout, M., Hinde, J. and Demétrio, G. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* **57**, 219–223. MR1833310

Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. *Quarterly of Applied Mathematics* **4**, 45–99, 112–141.

Singh, S. (1963). A note on inflated Poisson distribution. *Journal of the Indian Statistical Association* **1**, 140–144. MR0162309

Thas, O. and Rayner, J. C. W. (2005). Smooth tests for the zero-inflated Poisson distribution. *Biometrics* **61**, 808–815. MR2196169

van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics* **51**, 738–743. MR1349912

Wang, P., Puterman, M. L., Cockburn, L. and Le, N. (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics* **52**, 381–400.

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.

Department of Public Health Sciences
Division of Biostatistics
University of California, Davis
California 95616
USA
E-mail: cssli@ucdavis.edu