Properties of a class of residuals in the zero adjusted inverse Gaussian regression models

Juliana Scudilio Rodrigues
Federal University of São Carlos and
University of São Paulo, Institute of Mathematics and Computer Sciences, São Carlos, SP, Brazil -
juliana-scudilio@uol.com.br

Gustavo Henrique de Araujo Pereira
Federal University of São Carlos, São Carlos, SP, Brazil - gpereira@ufscar.br

Abstract

Residuals play an important role in checking model adequacy and in the identification of outliers and influential observations. In this paper, we studied a class of residuals for the zero adjusted inverse Gaussian regression model. This class of residuals is a function of a residual for the continuous component of the model and the maximum likelihood estimate of the probability of the observation assuming the zero value. Monte Carlo simulation studies are performed to examine the properties of this class of residuals in the zero adjusted inverse Gaussian regression model. Results showed that one residual of this class has some similar properties to the standard normal distribution in the studied model.

Keywords: diagnostic analysis; inflated regression models; quantile residual; ZAIG models.

1. Introduction

Zero inflated regression models are used for fitting response variables that assume the zero value with a probability greater than allowed by a known probability distribution. The response variable of these models may be either discrete or continuous on some interval of the set of positive real numbers and discrete at zero. In the latter case, they are called zero inflated continuous regression models.

Residuals play an important role in checking model adequacy and in the identification of outliers and influential observations. One possible way to perform diagnostic analysis in zero inflated continuous regression models is to carry out residual analysis for the discrete and the continuous component of the model separately. In this case, a residual is used for each component of the model. However, this procedure has a major drawback. Each positive observation has two residuals and it is hard to identify outliers based in a vector of residuals. A possible alternative is to use a single residual to jointly analyze the two components of the model. The randomized quantile residual proposed by Dunn and Smyth (1996) can be used for this purpose. Nevertheless, if the response variable assumes a positive value close to zero, these residual will not assume a high value, in modulus, unless that the probability of a observation assuming the zero value is very low (Pereira et al., 2014). This is not reasonable in many practical situations.

A class of residuals for the zero inflated continuous regression models was proposed by Pereira et al. (2014). This class of residuals is a function of a residual for the continuous component of the model and the maximum likelihood estimate of the probability of the observation assuming the zero value. Residuals of this class do not have the shortcomings of the randomized quantile residual and the authors concluded that one residual of this class has good properties in the zero inflated beta regression model.

The aim of this work is to study the properties of the residuals of the class proposed by Pereira et al. (2014) in the zero adjusted inverse Gaussian regression model. Monte Carlo simulation studies are used for this purpose.

The remainder of the paper is organized as follows. Section 2 defines the zero adjusted inverse Gaussian regression model. The next section presents the residuals considered in this paper. In the following section, Monte Carlo simulation studies are performed to examine the properties of the defined residuals. Concluding remarks are provided in Section 5.

2. Zero inflated regression models

Zero adjusted inverse Gaussian regression models have been used in several different areas like insurance (Resti et al., 2013), public health (Venezuela and Artes, 2014) and medicine (Dias, 2014). The GAMLSS package, which is available for the R statistical software, can be used to fit these models (Stasinopoulos and Rigby, 2007).

The probability density function of the response variable of the zero adjusted inverse Gaussian regression models can be written as follows:

$$f(y;\alpha;\mu;\sigma) = \begin{cases} \alpha, & if \quad y = 0 \\ (1-\alpha)f(y;\mu;\sigma); & if \quad y > 0 \end{cases} \tag{1}$$

where α is the mixture parameter (probability that the variable takes the zero value) and $f(y;\mu;\sigma)$ is the probability density function of the inverse Gaussian distribution defined as

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{(2\pi\sigma^2 y^3)}} \exp\left\{-\frac{1}{2\mu^2\sigma^2 y}(y-\mu)^2\right\} \tag{2}$$

where μ is the mean of the distribution and σ is a dispersion parameter.

Let $y_1, y_2, \ldots, y_n$ be independent random variables, where each y_i, $i = 1, 2, \ldots, n$, is zero adjusted inverse Gaussian distributed with parameters $(\alpha_i, \mu_i, \phi_i)$. The zero adjusted inverse Gaussian regression models are defined by (1) and the following systematic components:

$$\begin{aligned} g_1(\mu_i) &= \eta_{i1} \\ g_2(\sigma_i) &= \eta_{i2} \\ g_3(\alpha_i) &= \eta_{i3} \end{aligned} \tag{3}$$

where $\eta_{i1} = x_{i1}^\top\beta_1$, $\eta_{i2} = x_{i2}^\top\beta_2$, $\eta_{i3} = x_{i3}^\top\beta_3$ are linear predictors; $\beta_1 = (\beta_{11}, \beta_{21}, \ldots, \beta_{p_11})$, $\beta_2 = (\beta_{12}, \beta_{22}, \ldots, \beta_{p_22})$ and $\beta_3 = (\beta_{13}, \beta_{23}, \ldots, \beta_{p_33})$ are vectors of unknown parameters; $x_{i1} = (x_{i11}, x_{i21}, \ldots, x_{ip_11}), x_{i2} = (x_{i11}, x_{i21}, \ldots, x_{ip_11})$ and $x_{i3} = (x_{i13}, x_{i_23}, \ldots, x_{ip_33})$ are known explanatory variables and g_1, g_2 and g_3 are link functions strictly monotonic and twice differentiable.

3. A class of residuals for zero inflated regression models

The quantile residual in a regression model whose response variable is continous and has parameter μ_i and σ_i is defined as:

$$r_{q,i} = \Phi^{-1}\{F(y_i, \hat{\mu_i}, \hat{\sigma_i})\} \tag{4}$$

where $\Phi(\cdot)$ and $F(\cdot)$ are the cumulative distribution function of the standard normal distribution and of the distribution of the response variable, respectively, and $\hat{\mu_i}$ and $\hat{\sigma_i}$ are the maximum likelihood estimators of μ_i and σ_i.

Inspired by the quantile residual, Pereira et al.(2014) proposed a class of residual that do not have the shortcomings of the quantile residual. It is defined as:

$$r_i* = \begin{cases} \Phi^{-1}[(1-\Phi(|r_i|))(1-\hat{\alpha}_i)], & if \quad r_i < 0 \\ \Phi^{-1}[1-(1-\Phi(|r_i|))(1-\hat{\alpha_i})] & if \quad r_i > 0 \end{cases} \tag{5}$$

where r_i is any residual for a regression model with continuous response variable.

Using simple algebraic manipulations, r_i* can be rewritten as:

$$r_i* = \left\{ \begin{array}{cc} \Phi^{-1}[\Phi(r_i)(1-\hat{\alpha}_i)], & if \quad r_i < 0 \\ \Phi^{-1}[\hat{\alpha_i} + \Phi(r_i)(1-\hat{\alpha_i})], & if \quad r_i > 0 \end{array} \right. \tag{6}$$

Note that (6) defines a class of residuals for the positive observations of the zero inflated continuous regression models, because we can use as r_i any residual for a regression model with continuous response.

Pereira et al.(2014) showed that if r_i is exactly standard normal distributed and α_i is known, then r_i* has standard normal distribution. In general, these assumptions are not met, so the authors performed simulation studies for the inflated beta regression models. They concluded that the residual proposed by Pereira et al.(2014) using the quantile residual as r_i presented good properties and is better than the other residuals of this class and the residual proposed by Ospina and Ferrari (2012).

The most used residuals in inverse Gaussian regression and in beta regression are not the same. For this reason, it is important to evaluate this class of residuals in the zero adjusted inverse Gaussian models, using a residual defined for these models as r_i, such as the quantile residual, the standardized Pearson and the standardized deviance.

4. Simulation studies

Monte Carlo simulation studies were performed using a zero adjusted inverse Gaussian model, in which,

$$\begin{array}{rcl} \log\{\mu_i\} & = & \beta_{11} + \beta_{21}x_{i21} + \beta_{31}x_{i31} \\ \log\{\sigma_i\} & = & \beta_{12} \\ \log\left\{\frac{\alpha_i}{1-\alpha_i}\right\} & = & \beta_{13} + \beta_{23}x_{i23} + \beta_{33}x_{i33} \end{array} \tag{7}$$

The covariates values were generated as independent draws from the standard uniform distribution and remained constant throughout the simulations. Four scenarios were considered. For all scenarios considered, $\beta_{11} = 0.5 = \beta_{21}$, $\beta_{31} = 1$, which resulted in $\mu \,\epsilon\, (1.65; 7.39)$. In the first scenario (Scenario I) $\beta_{13} = -0.9$, $\beta_{23} = -0.5$, $\beta_{33} = 1$, which resulted in $\alpha \,\epsilon\, (0.20; 0.50)$. In the second scenario (Scenario II) $\beta_{13} = -1.4$, $\beta_{23} = -0.6$, $\beta_{33} = 1$, which resulted in lower values of α (0.12 to 0.35). In the following scenario (Scenario III) $\beta_{13} = -0.5$, $\beta_{23} = 0.3$, $\beta_{33} = 1$, which resulted in higher values of α (0.38 to 0.69). For these scenarios (I, II and III) we considered $\sigma = 0.1$. In the last scenario, we considered the values of β_{13}, β_{23} and β_{33}, used in Scenario I, but $\sigma = 0.05$. All results are based on 5000 Monte Carlo replications and n equals to 50, 100 and 200. Simulations were performed using the R statistical software. For all scenarios the residual r_i* was calculated using the quantile residual as r_i (equation 7).

We calculated the percentage of residuals lower than -3, -2 and -1 and higher than 1,2 and 3 for each of the n observations. If r_i* is a good residual, it is expected that this percentage is close to the theoretical value of the standard normal distribution for all n observation. For example, if r_i* is a good residual, it is expected that the percentage of residuals higher than 2, is close to $1 - \Phi(2) = 2.28\%$. If $y_i = 0$, r_i* is not defined. In this case, we assume that r_i* is not lower than $-k$ and not higher than k, for any $k \,\epsilon R^+$. Tables 1 to 4 present the descriptive statistics for the observed percentage of residuals in each interval.

In the first scenario, for $n = 50$, the observed percentage in each interval is not far from the theoretical value of the standard normal distribution for all observations. The percentage of residuals lower than -2, for example, varies between 1.68 and 2.94, close to the theoretical value of 2.28. When sample size increases from n = 50 to n = 100, the values for all statistics become closer to the theoretical value of the standard normal distribution. However, when sample size increases to $n = 200$, an improvement is observed only for the mean and median. For Scenarios II and IV, results are similar to Scenario I. These findings suggest that a reduction in α and σ does not affect considerably the properties of the residual in study.

In Scenario III, the percentage of residuals lower than -3 and -2, and higher than 2 and 3 is also close to the theoretical values of the standard normal distribution. However, for the intervals starting at -1 and 1, the minimum and the first quartile values are far from the theoretical values than in other scenarios. This is not

Table 1: Descriptive statistics for the observed percentage of residuals in each interval - Scenario I

n	Residuals	Theoretical	Simulation					
	Interval	Value	Min	1st Qu	Median	Mean	3rd Qu	Max
50	<-3	0.13	0.02	0.07	0.08	0.10	0.12	0.20
	<-2	2.28	1.68	2.11	2.32	2.31	2.53	2.94
	<-1	15.87	14.54	15.78	16.12	16.04	16.44	17.38
	>1	15.87	14.88	15.79	16.17	16.19	16.60	17.72
	>2	2.28	1.40	2.03	2.19	2.18	2.34	2.90
	>3	0.13	0.00	0.03	0.06	0.06	0.10	0.20
100	<-3	0.13	0.02	0.08	0.12	0.12	0.16	0.26
	<-2	2.28	1.74	2.12	2.26	2.29	2.46	2.86
	<-1	15.87	14.62	15.52	15.98	15.97	16.40	17.20
	>1	15.87	14.62	15.68	16.10	16.06	16.38	17.44
	>2	2.28	1.64	2.06	2.22	2.20	2.38	2.70
	>3	0.13	0.02	0.06	0.10	0.10	0.12	0.22
200	<-3	0.13	0.02	0.10	0.12	0.13	0.16	0.26
	<-2	2.28	1.68	2.12	2.24	2.26	2.38	2.88
	<-1	15.87	14.16	15.62	15.94	15.96	16.32	17.16
	>1	15.87	14.56	15.58	15.94	15.95	16.33	17.36
	>2	2.28	1.60	2.10	2.26	2.26	2.40	2.84
	>3	0.13	0.02	0.08	0.12	0.12	0.16	0.28

a major problem in practice since the limits for outliers identification are usually 2 or 3.

5. Concluding remarks

In this work, we studied the properties of a class of residuals in the zero adjusted inverse Gaussian regression models using Monte Carlo simulation techniques. Four scenarios were simulated for different sample sizes. In all scenarios, one residual of this class has some similar properties to the standard normal distribution in the zero adjusted inverse Gaussian regression models. In fact, the probability of this residual assuming values greater than 2 or 3 is close to the theoretical values of the standard normal distribution. These findings suggest that this residual is good for outliers detection in the zero adjusted inverse Gaussian regression models.

Acknowledgments

The authors gratefully acknowledge partial financial support from Capes and Fapesp.

References

Dias, M. F. (2014). "Asymetric inflated zero models" (Master dissertation, Universidade de São Paulo).

Dunn, P. K. and Smyth, G. K. (1996), "Randomized quantile residuals", Journal of Computational and Graphical Statistics 5(3), 236-244.

Ospina, R. and Ferrari, S. L. P. (2012), "A general class of zero-or-one inflated beta regression models", Computational Statistics and Data Analysis 56(6),1609-1623.

Pereira, G. H. A; Botter, D. A. ; Sandoval, M. C. (2014). "A new residual in zero inflated regression models." Proceedings of the 21th Simpósio Nacional de Probabilidade e Estatística.

Table 2: Descriptive statistics for the observed percentage of residuals in each interval - Scenario II

n	Residuals	Theoretical	Simulation					
	Interval	Value	Min	1st Qu	Median	Mean	3rd Qu	Max
50	<-3	0.13	0.00	0.06	0.08	0.08	0.12	0.20
	<-2	2.28	1.64	2.07	2.24	2.25	2.48	2.78
	<-1	15.87	14.60	15.85	16.26	16.20	16.59	17.24
	>1	15.87	14.98	15.78	16.22	16.23	16.44	18.16
	>2	2.28	1.48	2.04	2.22	2.20	2.38	3.00
	>3	0.13	0.00	0.06	0.08	0.07	0.10	0.14
100	<-3	0.13	0.02	0.08	0.10	0.11	0.14	0.28
	<-2	2.28	1.80	2.14	2.28	2.28	2.44	2.80
	<-1	15.87	14.88	15.68	16.01	16.02	16.38	17.38
	>1	15.87	14.78	15.70	16.09	16.07	16.46	17.22
	>2	2.28	1.72	2.04	2.23	2.20	2.37	2.76
	>3	0.13	0.00	0.08	0.10	0.11	0.14	0.26
200	<-3	0.13	0.04	0.10	0.12	0.12	0.16	0.32
	<-2	2.28	1.72	2.12	2.26	2.28	2.41	3.00
	<-1	15.87	14.46	15.53	15.93	15.93	16.30	17.00
	>1	15.87	14.54	15.58	15.94	15.95	16.30	17.32
	>2	2.28	1.52	2.12	2.24	2.25	2.40	2.80
	>3	0.13	0.02	0.08	0.12	0.12	0.14	0.32

Table 3: Descriptive statistics for the observed percentage of residuals in each interval - Scenario III

n	Residuals	Theoretical	Simulation					
	Interval	Value	Min	1st Qu	Median	Mean	3rd Qu	Max
50	<-3	0.13	0.02	0.06	0.10	0.09	0.12	0.22
	<-2	2.28	1.72	2.09	2.33	2.32	2.50	3.20
	<-1	15.87	12.06	14.77	15.40	15.41	16.20	17.74
	>1	15.87	13.00	15.02	15.80	15.54	16.15	17.18
	>2	2.28	1.60	2.00	2.28	2.21	2.42	2.84
	>3	0.13	0.00	0.02	0.04	0.05	0.10	0.18
100	<-3	0.13	0.04	0.08	0.12	0.12	0.15	0.26
	<-2	2.28	1.78	2.16	2.30	2.33	2.49	2.98
	<-1	15.87	13.44	15.29	15.73	15.68	16.16	17.26
	>1	15.87	13.32	15.32	15.89	15.76	16.25	17.24
	>2	2.28	1.48	2.12	2.26	2.25	2.39	2.82
	>3	0.13	0.00	0.06	0.10	0.09	0.12	0.22
200	<-3	0.13	0.00	0.10	0.12	0.13	0.16	0.28
	<-2	2.28	1.84	2.16	2.30	2.30	2.44	2.98
	<-1	15.87	13.74	15.38	15.80	15.80	16.28	17.22
	>1	15.87	13.36	15.48	15.90	15.84	16.22	17.42
	>2	2.28	1.62	2.14	2.26	2.26	2.39	2.92
	>3	0.13	0.02	0.08	0.12	0.12	0.14	0.32

Table 4: Descriptive statistics for the observed percentage of residuals in each interval - Scenario IV

n	Residuals	Theoretical	Simulation					
	Interval	Value	Min	1st Qu	Median	Mean	3rd Qu	Max
50	<-3	0.13	0.00	0.06	0.08	0.09	0.12	0.20
	<-2	2.28	1.64	2.09	2.30	2.28	2.48	2.86
	<-1	15.87	14.60	15.75	16.08	16.05	16.43	17.40
	>1	15.87	14.86	15.70	16.10	16.13	16.55	17.78
	>2	2.28	1.50	2.08	2.22	2.22	2.40	2.88
	>3	0.13	0.00	0.04	0.08	0.07	0.10	0.20
100	<-3	0.13	0.02	0.08	0.12	0.12	0.16	0.26
	<-2	2.28	1.74	2.10	2.24	2.28	2.44	2.86
	<-1	15.87	14.66	15.57	15.98	15.98	16.40	17.18
	>1	15.87	14.68	15.70	16.06	16.04	16.38	17.40
	>2	2.28	1.66	2.08	2.25	2.22	2.39	2.76
	>3	0.13	0.02	0.06	0.10	0.10	0.14	0.22
200	<-3	0.13	0.02	0.10	0.12	0.12	0.16	0.28
	<-2	2.28	1.68	2.10	2.24	2.25	2.39	2.90
	<-1	15.87	14.10	15.64	15.96	15.96	16.32	17.26
	>1	15.87	14.50	15.58	15.93	15.95	16.33	17.36
	>2	2.28	1.64	2.12	2.27	2.26	2.40	2.84
	>3	0.13	0.02	0.10	0.12	0.12	0.16	0.26

Resti, Y., Ismail, N., & Jamaan, S. H. (2013). ”Estimation of claim cost data using zero adjusted gamma and inverse Gaussian regression models”. Journal of Mathematics and Statistics, 9(3), 186-192.

Stasinopoulos, D. M., & Rigby, R. A. (2007). ”Generalized additive models for location scale and shape (GAMLSS) in R”. Journal of Statistical Software, 23(7), 1-46.

Venezuela, M. K., & Artes, R. (2014). ”Estimating equations and diagnostic techniques applied to zero-inflated models for panel data”. Electronic Journal of Statistics, 8(1), 1641-1660.