Untangle the Structural and Random Zeros in Statistical Modelings

Tang, W.[a], He, H.[b*], Wang, W. J.[c] and Chen, D. G.[d,e,f]

[a] *Department of Global Biostatistics & Data Science, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA70122, USA*;
[b] *Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA70122, USA*;
[c] *Brightech International, LLC, New Jersey, NJ 08873, USA*;
[d] *School of Social Work, University of North Carolina at Chapel Hill, Chapel Hill, NC27599, USA*;
[e] *Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27517, USA*;
[f] *Department of Statistics, University of Pretoria, South Africa*;

Count responses with structural zeros are common in behavioral and social studies. There are considerable research activities focusing on zero-inflated models such as the zero-inflated Poisson (ZIP) and zero-inflated Negative Binomial (ZINB) models for such zero-inflated count responses. However, when such variables are used as covariates or predictors, the difference between structural and random zeros is often ignored and biased estimates may be resulted.One remedy is to include an indicator for the structural zero in the model as a predictor if observed. However, structural zeros are not observed in most real studies, in which case no statistical method is available to address the biasing issue. This paper is aimed to fill this methodological gap by developing parametric methods to model zero-inflated count outcomes when used as explanatory variables based on the theory of maximum likelihood. The response variable can be any type of data including continuous, binary, count or even zero-inflated count responses. Simulation studies are performed to assess the numerical performance of this new approach when sample size is small and moderate. A real data example is also used to demonstrate the application of this method.

Keywords: generalized linear models; maximum likelihood; structural zeros; zero-inflated Poisson; zero-inflated explanatory variables.

This work was supported by the NIH under grants R33 DA027521, R01GM108337, R01HD075635 and a Faculty Research Support Grant from School of Nursing, University of Rochester.

*Corresponding author. Email: hhe2@tulane.edu

Table S1. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the continuous response when the Main Model is misspecified. The variances are in 10^{-3} for c_x and c_z, 10^{-1} for a_1, and 10^{-2} for the other estimates.

The estimates of the parameters for the Main Model:

	$c_0 = -1$		$c_x = 1$		$c_z = 1$	
N	Est.	CP	Est.	CP	Est.	CP
200	-0.301 (0.68 0.75)	0.0	0.826 (3.00 4.68)	10.5	1.003 (5.91 5.98)	94.8
500	-0.300 (0.27 0.29)	0.0	0.836 (1.10 2.16)	0.4	1.001 (2.37 2.45)	94.8
1000	-0.304 (0.14 0.14)	0.0	0.840 (0.53 1.14)	0.0	0.998 (1.18 1.15)	93.9

The estimates of the parameters for the Auxiliary Model:

	$a_0 = 0.5$		$a_1 = 1$		$b_0 = 0.5$		$b_1 = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	0.469 (7.25 7.33)	94.9	1.081 (1.35 1.62)	93.8	0.488 (1.74 1.79)	95.2	1.013 (1.81 1.87)	94.2
500	0.486 (2.70 2.70)	95.1	1.028 (0.47 0.43)	96.4	0.492 (0.68 0.67)	95.6	1.005 (0.64 0.59)	96.1
1000	0.493 (1.31 1.37)	94.2	1.017 (0.23 0.25)	93.8	0.499 (0.34 0.33)	94.8	1.001 (0.30 0.32)	94.8

Table S2. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the continuous response when the structural zeroes have no effect on the outcome. The variances are in 10^{-3} for c_x and c_z, 10^{-1} for a_1, and 10^{-2} for the others.

The parameter estimates for the Main Model:

	$c_0 = -1$		$c_x = 1$		$c_r = 1$		$c_z = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	-1.001 (3.23 3.41)	95.0	1.001 (4.35 4.38)	95.5	-0.006 (5.17 5.42)	95.2	1.002 (5.03 5.10)	94.4
500	-0.998 (1.20 1.24)	95.0	1.001 (1.50 1.59)	94.4	-0.002 (1.97 2.01)	94.4	1.000 (2.01 2.04)	95.5
1000	-1.003 (0.59 0.58)	95.7	1.001 (0.71 0.74)	94.4	0.001 (0.97 0.96)	95.9	0.998 (1.00 0.99)	94.5

The parameter estimates for the Auxiliary Model:

	$a_0 = 0.5$		$a_1 = 1$		$b_0 = 0.5$		$b_1 = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	0.469 (7.25 7.5)	95.2	1.082 (1.35 1.67)	92.9	0.488 (1.74 1.81)	94.5	1.012 (1.81 1.88)	93.8
500	0.486 (2.7 2.71)	95.2	1.028 (0.47 0.43)	96.5	0.492 (0.68 0.67)	95.9	1.005 (0.64 0.59)	96.1
1000	0.493 (1.31 1.37)	94.2	1.017 (0.23 0.25)	93.9	0.499 (0.34 0.34)	95.1	1.001 (0.3 0.32)	94.3

Table S3. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the Binary response when the Main Model is misspecified. The variances are in 10^{-2} for c_z, a_0, b_0 and b_1, 10^{-1} for the other estimates.

The estimates of the parameters for the Main Model:

	$c_0 = -1$		$c_x = 1$		$c_z = 1$	
N	Est.	CP	Est.	CP	Est.	CP
200	-0.242 (0.31 0.30)	1.4	0.676 (0.43 0.37)	57.5	0.998 (3.82 3.78)	94.7
500	-0.236 (0.12 0.12)	0.0	0.656 (0.16 0.13)	23.2	0.984 (1.47 1.51)	94.6
1000	-0.240 (0.06 0.06)	0.0	0.654 (0.08 0.06)	2.4	0.975 (0.72 0.74)	93.0

The estimates of the parameters for the Auxiliary Model:

	$a_0 = 0.5$		$a_1 = 1$		$b_0 = 0.5$		$b_1 = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	0.469 (7.25 7.33)	94.9	1.081 (1.35 1.62)	93.8	0.488 (1.74 1.79)	95.2	1.013 (1.81 1.87)	94.2
500	0.486 (2.70 2.70)	95.1	1.028 (0.47 0.43)	96.4	0.492 (0.68 0.67)	95.6	1.005 (0.64 0.59)	96.1
1000	0.493 (1.31 1.37)	94.2	1.017 (0.23 0.25)	93.8	0.499 (0.34 0.33)	94.8	1.001 (0.30 0.32)	94.8

Table S4. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the binary response when the structural zeroes have no effect on the outcome. The variances are in 10^{-2} for c_z, a_0, b_0 and b_1, 10^{-1} for the others.

The parameter estimates for the Main Model:

	$c_0 = -1$		$c_x = 1$		$c_r = 1$		$c_z = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	-1.087 (3.48 3.63)	96.3	1.085 (1.42 1.50)	97.2	0.052 (4.90 5.21)	95.7	1.039 (4.52 4.86)	94.3
500	-1.039 (1.14 1.24)	95.2	1.036 (0.46 0.51)	94.7	0.026 (1.69 1.75)	95.8	1.020 (1.70 1.82)	94.0
1000	-1.026 (0.55 0.57)	94.6	1.021 (0.22 0.22)	95.7	0.024 (0.82 0.87)	94.8	1.008 (0.83 0.84)	95.3

The parameter estimates for the Auxiliary Model:

	$a_0 = 0.5$		$a_1 = 1$		$b_0 = 0.5$		$b_1 = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	0.468 (7.26 7.43)	95.6	1.081 (1.34 1.67)	93.2	0.489 (1.73 1.82)	94.8	1.012 (1.80 1.87)	94.2
500	0.486 (2.70 2.71)	95.1	1.027 (0.47 0.44)	96.2	0.493 (0.68 0.68)	95.4	1.005 (0.64 0.59)	95.7
1000	0.493 (1.31 1.37)	94.0	1.017 (0.23 0.26)	93.7	0.499 (0.34 0.34)	94.8	1.001 (0.30 0.32)	94.5

Table S5. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the Poisson response. The variances are in 10^{-3} for c_z and c_z, 10^{-1} for a_1, and 10^{-2} for the other estimates.

The estimates of the parameters for the Main Model:

	$c_0 = -1$		$c_x = 1$		$c_r = 1$		$c_z = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	-0.989 (0.86 0.83)	94.7	0.999 (0.23 0.23)	95.2	0.980 (1.34 2.46)	93.1	0.997 (1.08 1.14)	93.3
500	-0.994 (0.17 0.18)	92.8	0.999 (0.03 0.02)	92.3	0.989 (0.45 0.72)	92.9	1.000 (0.18 0.21)	93.5
1000	-0.997 (0.05 0.06)	93.8	1.000 (0.00 0.01)	93.8	0.995 (0.16 0.15)	96.3	1.000 (0.04 0.05)	93.6

The estimates of the parameters for the Auxiliary Model:

	$a_0 = 0.5$		$a_1 = 1$		$b_0 = 0.5$		$b_1 = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	0.490 (5.21 5.15)	95.7	1.081 (0.89 0.97)	95.4	0.488 (1.46 1.49)	94.9	1.004 (1.70 1.65)	96.0
500	0.494 (2.07 1.94)	95.1	1.039 (0.33 0.30)	96.8	0.493 (0.58 0.58)	95.2	0.994 (0.61 0.55)	95.5
1000	0.500 (0.97 1.10)	93.3	1.038 (0.16 0.17)	94.1	0.497 (0.28 0.28)	94.1	0.990 (0.29 0.29)	94.1

Table S6. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the Poisson response when the Main Model is misspecified. The variances are in 10^{-3} for c_x and c_z, 10^{-1} for a_1, and 10^{-2} for the other estimates.

The estimates of the parameters for the Main Model:

	$c_0 = -1$		$c_x = 1$		$c_z = 1$	
N	Est.	CP	Est.	CP	Est.	CP
200	-0.574 (0.39 3.46)	2.9	0.939 (0.10 1.81)	2.9	0.980 (0.97 4.30)	52.3
500	-0.713 (0.10 2.12)	1.2	0.968 (0.01 0.58)	1.2	0.986 (0.17 1.43)	34.8
1000	-0.786 (0.04 0.97)	0.2	0.980 (0.00 0.16)	0.2	0.987 (0.04 0.70)	27.0

The estimates of the parameters for the Auxiliary Model:

	$a_0 = 0.5$		$a_1 = 1$		$b_0 = 0.5$		$b_1 = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	0.469 (7.46 7.60)	94.8	1.117 (1.42 1.68)	94.4	0.484 (1.76 1.84)	95.0	1.008 (1.97 2.05)	93.8
500	0.484 (2.77 2.69)	95.7	1.054 (0.49 0.46)	96.8	0.490 (0.68 0.70)	94.3	0.996 (0.71 0.65)	96.1
1000	0.496 (1.35 1.48)	94.3	1.047 (0.24 0.29)	92.6	0.497 (0.34 0.33)	94.3	0.991 (0.34 0.36)	93.8

Table S7. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the Poisson response when the structural zeroes have no effect on the outcome. The variances are in 10^{-3} for c_z and c_z, 10^{-1} for a_1, and 10^{-2} for the others.

The parameter estimates for the Main Model:

	$c_0 = -1$		$c_x = 1$		$c_r = 1$		$c_z = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	-0.999 (0.89 0.81)	96.5	1.000 (0.24 0.23)	96.6	-0.006 (2.59 2.05)	97.2	0.999 (1.31 1.33)	94.0
500	-0.999 (0.17 0.16)	95.1	1.000 (0.03 0.02)	95.1	0.000 (0.83 0.62)	95.7	1.000 (0.20 0.19)	95.2
1000	-1.001 (0.05 0.06)	94.0	1.000 (0.01 0.01)	95.3	0.000 (0.37 0.19)	97.8	1.000 (0.04 0.05)	94.3

The parameter estimates for the Auxiliary Model:

	$a_0 = 0.5$		$a_1 = 1$		$b_0 = 0.5$		$b_1 = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	0.470 (7.45 7.65)	94.8	1.116 (1.41 1.7)	93.8	0.485 (1.75 1.83)	94.5	1.008 (1.97 2.06)	93.9
500	0.483 (2.77 2.69)	95.5	1.054 (0.49 0.46)	96.8	0.49 (0.68 0.70)	94.3	0.997 (0.71 0.65)	96.0
1000	0.496 (1.35 1.5)	94.3	1.047 (0.24 0.29)	92.8	0.497 (0.34 0.33)	94.5	0.991 (0.34 0.37)	94.0

Table S8. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the ZIP response when the Main Model is misspecified. The variances are in 10^{-2} for c_{zp}, a_0, b_0 and b_1, 10^{-1} for the other estimates except for c_{0b}.

The estimates of the parameters for the poisson component of the Main Model:

	$c_{0p} = -1$		$c_{xp} = 1$		$c_{zp} = 1$	
N	Est.	CP	Est.	CP	Est.	CP
200	-0.214 (0.38 0.47)	4.3	0.609 (0.38 0.70)	29.5	0.969 (2.46 3.35)	90.6
500	-0.211 (0.14 0.16)	0.0	0.676 (0.08 0.17)	2.9	0.947 (0.82 1.15)	85.0
1000	-0.210 (0.07 0.09)	0.0	0.699 (0.03 0.08)	0.1	0.946 (0.38 0.58)	78.1

The estimates of the parameters for the zero-inflated component of the Main Model:

	$c_{0b} = -1$		$c_{xb} = 1$		$c_{zb} = 1$	
N	Est.	CP	Est.	CP	Est.	CP
200	-0.230 (0.24 0.26)	52.0	0.730 (1.67 1.36)	79.4	1.038 (2.07 2.47)	92.5
500	-0.162 (0.08 0.07)	16.9	0.704 (0.33 0.24)	56.7	0.945 (0.63 0.67)	92.7
1000	-0.149 (0.04 0.04)	1.9	0.701 (0.16 0.11)	33.2	0.916 (0.30 0.30)	90.1

The estimates of the parameters for the Auxiliary Model:

	$a_0 = 0.5$		$a_1 = 1$		$b_0 = 0.5$		$b_1 = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	0.468 (7.33 7.47)	94.9	1.095 (13.70 1.60)	94.3	0.486 (1.75 1.82)	95.1	1.013 (1.86 1.92)	93.9
500	0.488 (2.72 2.74)	95.3	1.036 (4.80 0.40)	96.6	0.493 (0.68 0.68)	95.5	1.003 (0.66 0.61)	96.6
1000	0.493 (1.33 1.42)	93.7	1.025 (2.30 0.30)	93.5	0.498 (0.34 0.33)	94.8	0.998 (0.32 0.35)	94.1

Table S9. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the ZIP response when the structural zeroes have no effect on the outcome. The variances are in 10^{-2} for c_{zp}, a_0, b_0 and b_1, 10^{-1} for the others except c_{rb} and c_{0b}.

The parameter estimates for the Main Model: the Poisson component

	$c_{0p} = -1$		$c_{xp} = 1$		$c_{rp} = 1$		$c_{zp} = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	-1.081 (3.08 3.16)	94.7	1.025 (0.93 0.98)	95.4	0.019 (4.02 4.31)	93.3	1.008 (3.55 4.31)	93.6
500	-1.053 (0.91 0.92)	95.2	1.023 (0.21 0.21)	94.5	0.044 (1.26 1.29)	93.9	1.010 (1.11 1.20)	94.6
1000	-1.016 (0.38 0.37)	94.0	1.005 (0.07 0.07)	94.8	0.005 (0.54 0.53)	94.5	1.000 (0.50 0.49)	95.5

The parameter estimates for the Main Model: the zero-inflated component

	$c_{0b} = -1$		$c_{xb} = 1$		$c_{rb} = 1$		$c_{zb} = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	-1.531 (2.53 2.34)	97.6	1.290 (6.16 5.85)	96.9	-0.034 (2.39 2.4)	97.5	1.337 (6.23 7.37)	95.4
500	-1.266 (0.61 0.59)	96.2	1.146 (1.43 1.39)	97.3	0.083 (0.67 0.66)	95.4	1.139 (1.38 1.50)	95.9
1000	-1.098 (0.25 0.23)	96.7	1.062 (0.57 0.54)	96.8	0.030 (0.29 0.28)	96.2	1.043 (0.57 0.55)	96.3

The parameter estimates for the Auxiliary Model:

	$a_0 = 0.5$		$a_1 = 1$		$b_0 = 0.5$		$b_1 = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	0.456 (7.31 7.68)	94.7	1.107 (1.37 1.68)	92.9	0.486 (1.72 1.84)	94.7	1.014 (1.83 1.92)	94.2
500	0.488 (2.72 2.75)	95.2	1.036 (0.48 0.45)	96.3	0.493 (0.68 0.68)	95.8	1.003 (0.66 0.62)	96.4
1000	0.493 (1.33 1.43)	93.9	1.025 (0.23 0.27)	93.5	0.498 (0.34 0.34)	94.6	0.998 (0.32 0.35)	94.4

Table S10. The estimates (Estimate), standard errors (Std Err) and p-values (P-value) from the Auxiliary Model of the DAD for the 2009-2010 NHANES study

The estimates of the parameters for the Auxiliary Model of the DAD:

Parameter		the poisson compoint			the zero-inflated component		
		Estimate	Std Err	P-value	Estimate	Std Err	P-value
Intercept		-0.4750	0.1076	¡.0001	-0.3484	0.2304	0.1304
Gender	Male vs. Female	0.4729	0.0310	<.0001	-0.6696	0.0718	<.0001
AGE		0.0037	0.0009	<.0001	0.0278	0.0022	<.0001
Race/Ethnicity	Mexican American	0.0763	0.0937	0.4152	-0.5728	0.1892	0.0025
	Other Hispanic	0.1144	0.0987	0.2466	-0.2905	0.1955	0.1373
	Non-Hispanic White	0.3514	0.0855	<.0001	-0.9031	0.1733	<.0001
	Non-Hispanic Black	0.2792	0.0907	0.0021	-0.4708	0.1837	0.0104
	NOther Race	0.0000	0.0000	.	0.0000	0.0000	.
Education		0.0641	0.0129	<.0001	-0.2150	0.0297	<.0001