Zero-Inflated Models in Stata

Matheus Albergaria and Luiz Paulo Fávero*

matheus.albergaria@usp.br lpfavero@usp.br

***Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo (FEA-USP)**

2016 Brazilian Stata Users Group Meeting

Universidade de São Paulo (USP)

December 2nd, 2016

SECTIONS

Motivation

Our goals today:

- Present a new class of count models (*zero-inflated models*).
- Discuss the *intuition* and *main ideas* related to such models.
- Describe a *step-by-step tutorial* for estimation in Stata.

Motivation

Why should we care?

- Count Models: *increasingly used* in applied research.
- In such models, the dependent variable (Y_i) assumes *non-negative* and *discrete* values ($Y_i = 0, 1, 2, ...$) for a given exposition (e.g., period, area, region, etc.).
- A few examples: patents (Hausman, Hall, and Griliches, 1984), manufacturing (Lambert, 1992), friendships (Marmaros and Sacerdote, 2006), corruption (Fisman and Miguel, 2007), and health (Staub and Winkelman, 2013).

Motivation

Why did we start caring?

- In a recent occasion, we tried to replicate the results of a famous *corruption study* (Fisman and Miguel, 2007).
- We were able to provide a *narrow replication* of the paper's original findings (Albergaria and Fávero, 2017)..
- ..but we could not reject hypotheses favoring the use of *zero-inflated count models* in this setting.

MOTIVATION

Replication of Fisman and Miguel's (2007) Corruption Study

Variable	violations_all	violations_all	violations_all	violations_all	violations_all
staff	0.05*** (0.012)	0.04*** (0.012)	0.05*** (0.011)	0.05*** (0.013)	0.05*** (0.013)
corruption	0.42*** (0.098)	0.52*** (0.178)	0.54*** (0.160)	0.55** (0.215)	0.39 (0.239)
post	−4.53*** (0.165)	−4.53*** (0.165)	−4.35*** (0.163)	−4.55*** (0.165)	−4.54*** (0.166)
lgdppcus		0.07 (0.100)	0.12 (0.114)	47.64* (25.565)	0.05 (0.103)
r_africa			2.73*** (0.466)		
r_middleeast			3.08*** (0.533)		
r_europe			2.05*** (0.498)		
r_southamerica			1.51*** (0.514)		
r_oceana			1.45** (0.677)		
r_asia			1.85*** (0.484)		
lgdppcus2				−10.00* (5.482)	
lgdppcus3				0.91* (0.512)	
lgdppcus4				−0.03* (0.018)	
corruption_post					0.17 (0.205)
Vuong test (uncorrected)	0.008***	0.008***	0.009***	0.010***	0.012***
Vuong test (AIC)	0.021**	0.022**	0.020**	0.026**	0.029**
Vuong test (SIC)	0.082*	0.083*	0.063*	0.096*	0.102
Observations	298	298	298	298	298

Source: Albergaria & Fávero (2017).

Theory

Zero-Inflated Models (ZIM)

- Specific class of Count Models: *Zero-Inflated Models*.
- In these models, the dependent variable is treated as a count variable with an *excess number of zeros*.
- Main Advantage: consider dependent variable with excess zeros as part of the *data generating process* (DGP).

THEORY

ZIM: Basic Intuition

- Zero-inflated models correspond to a combination between a *binary choice model* and a *count model* (Cameron and Trivedi, 2009).
- Such a combination allows for two distinct zero-generating processes: (i) "structural zeros" (*binary distributon*), and (ii) "sampling zeros" (*count distribution*) (Mohri and Roark, 2005).
- One can test the existence of an excessive number of zero counts in the data by Vuong's (1989) test, a likelihood ratio test comparing *standard* and *zero-inflated* count models.

Implementation

Stata Example

- Let's look at a first-order policy issue: the relation between *traffic accidents* and *alcohol prohibition* (Fávero and Belfiore, 2017).
- 2008: Brazilian government instaured a "Dry Law", with *harsher punishment* for drinking drivers.
- We want to estimate the relation between the *number of traffic accidents* (Y) and *population*, given that factors such as *age* and *dry laws* may generate "structural zeros" in this setting.

Implementation

Data Description (file "acidentes.dta")

```
. desc

----------------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
----------------------------------------------------------------------------------------------
accidents       byte    %8.0g                 Number of traffic accidents over last week
population      float   %9.5f                 Urban population (in millions)
age             float   %9.2f                 Average age for drivers with valid drivers licenses
drylaw          float   %9.0g      leiseca    Whether the city adopts 'Dry Law' standards after 10:00 p.m.
----------------------------------------------------------------------------------------------
```

IMPLEMENTATION

Data Tabulation

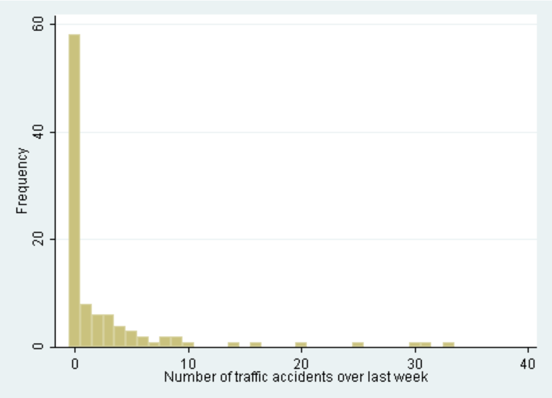
```
. tab accidents
```

Number of traffic accidents over last week	Freq.	Percent	Cum.
0	58	58.00	58.00
1	8	8.00	66.00
2	6	6.00	72.00
3	6	6.00	78.00
4	4	4.00	82.00
5	3	3.00	85.00
6	2	2.00	87.00
7	1	1.00	88.00
8	2	2.00	90.00
9	2	2.00	92.00
10	1	1.00	93.00
14	1	1.00	94.00
16	1	1.00	95.00
20	1	1.00	96.00
25	1	1.00	97.00
30	1	1.00	98.00
31	1	1.00	99.00
33	1	1.00	100.00
Total	100	100.00	

Implementation

Histogram

```
. hist accidents, discrete freq
(start=0, width=1)
```



IMPLEMENTATION

Zero-Inflated Poisson Model Estimates

```
. zipcv accidents population, inf(age drylaw) vuong nolog

Zero-inflated Poisson regression                Number of obs     =        100
                                                Nonzero obs       =         42
                                                Zero obs          =         58

Inflation model = logit                         LR chi2(1)        =      37.72
Log likelihood  = -256.0484                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
   accidents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
accidents    |
  population |   .5039652   .0863993     5.83   0.000     .3346256    .6733047
       _cons |   .9329778   .1987482     4.69   0.000     .5434386    1.322517
-------------+----------------------------------------------------------------
inflate      |
         age |   .2252293   .0584096     3.86   0.000     .1107485    .3397101
      drylaw |   1.725743   .5531873     3.12   0.002     .6415157     2.80997
       _cons |  -11.72936   3.030402    -3.87   0.000    -17.66884   -5.789881
------------------------------------------------------------------------------
        Vuong test of zip vs. standard Poisson:  z =     4.19  Pr>z = 0.0000
                                                               Pr<z = 1.0000
                   with AIC (Akaike) correction:  z =     4.13  Pr>z = 0.0000
                                                               Pr<z = 1.0000
                  with BIC (Schwarz) correction:  z =     4.04  Pr>z = 0.0000
                                                               Pr<z = 1.0000
```

IMPLEMENTATION

Overdispersion Test

```
. tabstat accidents, stats(mean var)
```

variable	mean	variance
accidents	3.01	42.9999

Implementation

Zero-Inflated Negative Binomial Model Estimates

```
. zinbcv accidents population, inf(age drylaw) vuong nolog zip

Zero-inflated negative binomial regression        Number of obs   =        100
                                                  Nonzero obs     =         42
                                                  Zero obs        =         58

Inflation model = logit                           LR chi2(1)      =      10.87
Log likelihood  = -164.4035                       Prob > chi2     =     0.0010
```

accidents	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
accidents						
population	.8661751	.2621428	3.30	0.001	.3523847	1.379966
_cons	.0253062	.5403137	0.05	0.963	-1.033689	1.084301
inflate						
age	.2882047	.0998951	2.89	0.004	.0924139	.4839954
drylaw	2.85907	1.076625	2.66	0.008	.7489239	4.969217
_cons	-16.23734	5.726858	-2.84	0.005	-27.46178	-5.012905
/lnalpha	.2399887	.3137446	0.76	0.444	-.3749393	.8549167
alpha	1.271235	.398843			.687331	2.351179

```
Likelihood-ratio test of alpha=0: chibar2(01) =   183.29 Pr>=chibar2 =  0.0000
Vuong test of zinb vs. standard negative binomial:  z =     3.88  Pr>z = 0.0001
                                                                  Pr<z = 0.9999
                       with AIC (Akaike) correction:  z =     3.31  Pr>z = 0.0005
                                                                  Pr<z = 0.9995
                      with BIC (Schwarz) correction:  z =     2.57  Pr>z = 0.0051
                                                                  Pr<z = 0.9949
```

Implementation

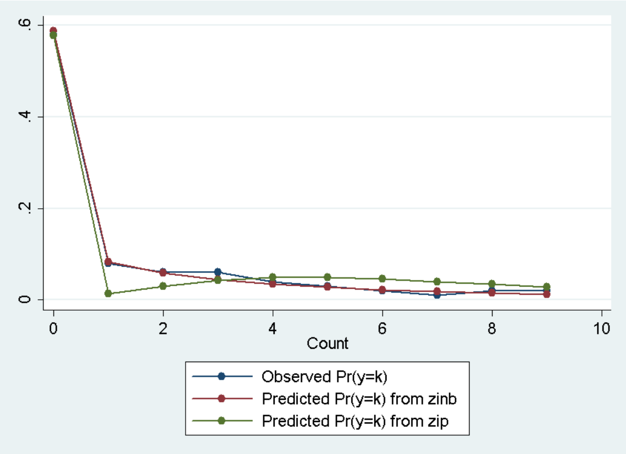
Model Comparison

	(1) accidents	(2) accidents
accidents		
population	0.504***	0.866***
	(0.0864)	(0.262)
_cons	0.933***	0.0253
	(0.199)	(0.540)
inflate		
age	0.225***	0.288**
	(0.0584)	(0.0999)
drylaw	1.726**	2.859**
	(0.553)	(1.077)
_cons	-11.73***	-16.24**
	(3.030)	(5.727)
lnalpha		
_cons		0.240
		(0.314)
N	100	100
ll	-256.0	-164.4

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

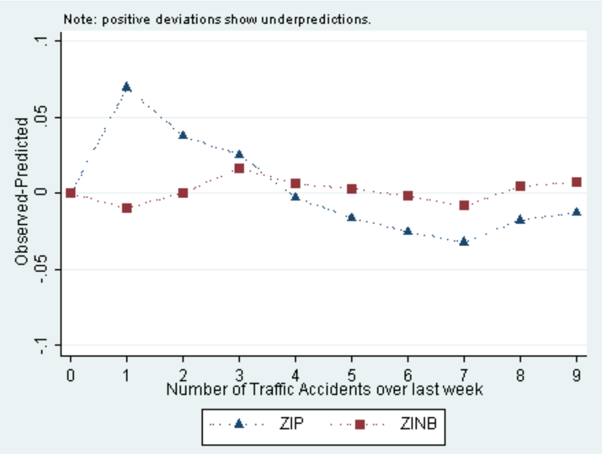
Implementation

Observed and Predicted Probabilities



Implementation

Error Terms' Deviations



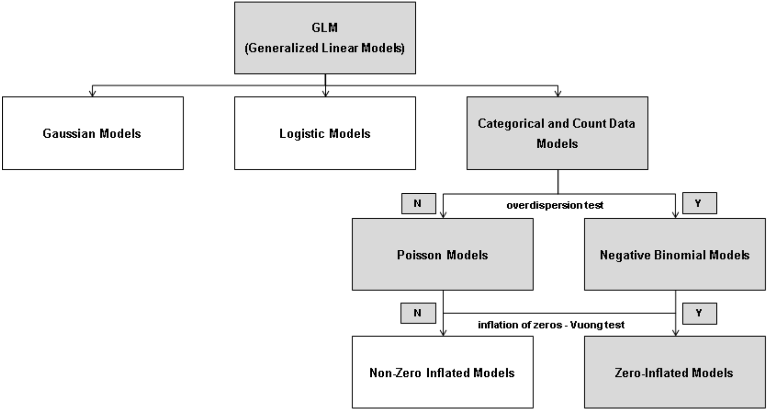
CONCLUSIONS

Count Data Models: Decision Table

Verification	Count Data Regression Model			
	Poisson	Negative Binomial	Zero-Inflated Poisson (ZIP)	Zero-Inflated Negative Binomial (ZINB)
Overdispersion in Outcome Variable	NO	YES	NO	YES
Inflation of Zeros in Outcome Variable	NO	NO	YES	YES

CONCLUSIONS

Zero-Inflated Models as a Special Class of Generalized Linear Models (GLM)



Conclusions

- Zero-Inflated Models: still employed *with parsimony* by Stata users today.
- Stata 14 has a *full command suite* for the estimation of zero-inflated models.
- Several research opportunities in the near future, both in *theoretical* and *applied* terms (e.g., initial public offerings, product innovations, etc.) (Blevins et al., 2015).

References

Albergaria, M., Fávero, L. P. (2017). Narrow replication of Fisman and Miguel's (2007a) 'Corruption, norms, and legal enforcement: evidence from diplomatic parking tickets'. *Journal of Applied Econometrics*, forthcoming.

Blevins, D. P., Tsang, E. W., Spain S. M. (2015). Count-based research in management: suggestions for improvement. *Organizational Research Methods*, 18(1), 47–69.

Cameron, A. C., Trivedi, P. K. (2009). *Microeconometrics using Stata*. Stata Press Books.

References

Desmarais, B., Harden, J. J. (2013). Testing for zero inflation in count models: bias correction for the Vuong test. *Stata Journal*, 13(4), 810–835.

Fávero, L. P., Belfiore, P. (2017). *Data science for business and decision making*. Boston: Elsevier, forthcoming.

Fisman, R., Miguel, E. (2007). Corruption, norms, and legal enforcement: evidence from diplomatic parking tickets. *Journal of Political Economy*, 115(4), 1020–1048.

References

Hausman, J. A., Hall, B. H., Griliches, Z. (1984). Econometric models for count data with an application to the patents-R&D relationship. *Econometrica*, 52(4), 909–938.

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

Marmaros, D., Sacerdote, B. (2006). How do friendships form? *Quarterly Journal of Economics*, 121(1), 79-119.

References

Mohri, M., Roark, B. (2005). *Structural zeros versus sampling zeros*. Technical Report CSEE-05-003, OGI School of Science Engineering, Oregon Health Science University.

Staub, K. E., Winkelmann, R. (2013). Consistent estimation of zero-inflated count models. *Health Economics*, 22(6), 673–686.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2), 307–333.

Thank You

Matheus Albergaria and Luiz Paulo Fávero

matheus.albergaria@usp.br lpfavero@usp.br

Appendix

Appendix A: Technical Details

Probability Function for the Zero-Inflated Poisson Model

$$\begin{cases} p(Y_i = 0) = p_{logit_i} + (1 - p_{logit_i})e^{-\lambda_i} \\ \mathrm{p}(\mathrm{Y}_i = m) = (1 - p_{logit_i})\dfrac{e^{-\lambda_i}\lambda_i^m}{m!}, \text{for } m = 1, 2, .. \end{cases} \Gamma$$

where $Y \sim ZIP(\lambda, p_{logit_i})$, with

$$p_{logit_i} = \frac{1}{1 + e^{-(\gamma + \delta_1 W_{1i} + \delta_2 W_{2i} + \ldots + \delta_q W_{qi})}}$$

and $\lambda_i = e^{(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki})}$

APPENDIX

Appendix A: Technical Details

Log-Likelihood Function for the Zero-Inflated Poisson Model

$$LL = \sum_{Y_i=0} ln[p_{logit_i} + (1 - p_{logit_i})e^{-\lambda}]$$

$$+\sum_{Y_i>0} ln[(1 - p_{logit_i}) - \lambda_i + (Y_i)ln(\lambda_i) - ln(Y_i!)] = max$$

Appendix

Appendix A: Technical Details

Probability Function for the Zero-Inflated Negative Binomial Model

$$\begin{cases} p(Y_i = 0) = p_{logit_i} + (1 - p_{logit_i})(\frac{1}{1+\phi u_i})^{\frac{1}{\phi}} \\ \mathrm{p}(Y_i = m) = (1 - p_{logit_i})[\binom{m+\phi^{-1}-1}{\phi^{-1}-1}(\frac{1}{1+\phi u_i})^{\frac{1}{\phi}}(\frac{\phi u_i}{\phi u_i+1})^{m}], \text{for } m = 1, 2, .. \end{cases}$$

Γ

where $Y \sim ZINB(\phi, u, p_{logit_i})$, ϕ denotes the inverse of the shape parameter of a Gamma distribution, with p_{logit_i} defined as before and

$$u_i = e^{(\alpha+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_k X_{ki})}$$

Appendix

Appendix A: Technical Details

Log-Likelihood Function for the Zero-Inflated Negative Binomial Model

$$LL = \sum_{Y_i=0} ln\left[p_{logit_i} + (1 - p_{logit_i})\left(\frac{1}{1+\phi u_i}\right)^{\frac{1}{\phi}}\right]$$

$$+\sum_{Y_i>0} ln\left[(1 - p_{logit_i}) + Y_i ln\left(\frac{\phi u_i}{1+\phi u_i}\right) - \frac{ln(1+\phi u_i)}{\phi}\right.$$

$$\left. + ln\Gamma(Y_i + \phi^{-1}) - ln\Gamma(Y_i + 1) - ln\Gamma(\phi^{-1})\right] = max$$

Appendix

Appendix B: Stata do-file

```
*****************************************************************************
*Stata do-file for the Presentation "Zero-Inflated Models in Stata"
*by Matheus Albergaria and Luiz Paulo Fávero
*2016 Brazilian Stata Users Group Meeting
*Universidade de São Paulo (USP), São Paulo, Brazil
*December 2nd, 2016

*This do-file was written by Luiz Paulo Fávero and Matheus Albergaria
*The data file is "Accidents.dta".
*For more details, see Fávero, L.P.; Belfiore, P. (2017).
*"Data science for business and decision making". Boston: Elsevier, forthcoming.
*****************************************************************************

*Open Dataset
use C:\Acidentes.dta

*Data Description
desc

*Descriptive Statistics
tab accidents
hist accidents, discrete freq
```

Appendix

Appendix B: Stata do-file

```
*Count Data Models' Estimation
*Zero-Inflated Poisson (ZIP)
zipcv accidents population, inf(age drylaw) vuong nolog
*Overdispersion Test
tabstat accidents, stats(mean var)
*Zero-Inflated Negative Binominal (ZINB)
zinbcv accidents population, inf(age drylaw) vuong nolog zip

*Model Comparison
eststo: quietly zip accidents population, inf(age drylaw) vuong
prcounts lambda_inflate, plot
eststo: quietly zinb accidents population, inf(age drylaw) vuong
prcounts u_inflate, plot
esttab, scalars(ll) se

*Observed and Predicted Probabilities (Graph)
graph twoway (scatter u_inflateobeq u_inflatepreq
lambda_inflatepreq u_inflateval, connect (l l l))

*Comparison of Mean Observed and Predicted Count (Table + Graph)
countfit accidents population, zip zinb noestimates
```