The Stata Journal (2019)
19, Number 3, pp. 499–509 DOI: 10.1177/1536867X19874209

Modeling count data with marginalized zero-inflated distributions

Tammy H. Cummings
Institute for Families in Society
University of South Carolina
Columbia, SC
harris68@mailbox.sc.edu

James W. Hardin
Department of Epidemiology and Biostatistics
University of South Carolina
Columbia, SC
jhardin@sc.edu

Abstract. In this article, we present new commands for modeling count data using marginalized zero-inflated distributions. While we mainly focus on presenting new commands for estimating count data, we also present examples that illustrate some of these new commands.

Keywords: st0563, mzip, mzip postestimation, mzigp, mzigp postestimation, mzinb, mzinb postestimation, marginalized, count data, Poisson, generalized Poisson, negative binomial, zero-inflated

1 Introduction

Often, count responses have zero-inflation—a higher prevalence of zeros than is accounted for by the underlying distribution of the regression model to be fit. This discordance can occur for outcome variables in many fields of study, such as medical, public health, and manufacturing. In these cases, estimation based on the distributional assumptions of Poisson, generalized Poisson, and negative binomial models can result in incorrect parameter estimates and biased standard errors. Zero-inflated count data are encountered in the number of defects in manufacturing (Lambert 1992), patient falls in hospitals (Ullah, Finch, and Day 2010), and the number of cubes in the test of tower building for motor development (Cheung 2002), just to name a few. Hardin and Hilbe (2018) describe the two origins of zero outcomes: outcomes for individuals who do not enter into the counting process and outcomes for individuals who enter into the counting process and have a zero outcome. Mullahy (1986) proposed the zero-inflated Poisson (ZIP) model, using a model familiar to researchers (Poisson), to deal with outcomes with an excess of zeros. However, for modeling count data with zero outcomes where overdispersion or underdispersion exists, one should consider other models, such as zero-inflated generalized Poisson (ZIGP) and zero-inflated negative binomial (ZINB) (Famoye and Singh 2006; Greene 1994).

Sometimes analysts want to estimate the marginal mean and be able to interpret estimated coefficients as the population-average parameters. Some authors have proposed different approaches to marginal models, such as Lee et al. (2011), who proposed likelihood-based marginalized models for zero-inflated clustered count data using hurdle models. Kassahun et al. (2014) presented ways to model hierarchical count data that had issues such as overdispersion, correlation, and an excess of zeros by marginalized hurdle and marginalized ZIP (MZIP) normal-gamma models. Others, like Heagerty and Zeger (2000), used a marginalized multilevel model that regressed the marginal mean instead of the conditional mean on the covariates. Long et al. (2014) recently proposed an MZIP regression model that directly models the population mean count, therefore providing the ability to interpret population-wide parameters. Preisser et al. (2016) also proposed a marginalized zero-inflated negative binomial (MZINB) regression model and applied it on dental caries in a school-based fluoride mouth rinse program.

We introduce the new commands `mzip`, for the marginalized zero-inflated Poisson (MZIP) regression model presented in Long et al. (2014), and `mzinb`, for the MZINB regression model presented in Preisser et al. (2016). We also extend that method to include a marginalized zero-inflated generalized Poisson (MZIGP) regression model and its accompanying command.

In this article, we illustrate modeling count data using MZIP, MZIGP, and MZINB regression models. In section 2, we review the three marginalized zero-inflated regression models. In section 3, we present syntax for the new commands. In section 3, we present a synthetic data example and a real world data example. Finally, we summarize in section 5.

2 Marginalized zero-inflated distributions

2.1 Marginalized ZIP distribution

The widely known ZIP regression model with a count outcome variable, $Y_i\,(i = 1, \ldots, n)$, has the probability p_i that the binary process results in a zero outcome, where $0 \leq p_i < 1$, and the counting process probability of a zero outcome is from the Poisson distribution. Thus, we have a probability mass function (p.m.f.)

$$P(Y_i = y_i) = \begin{cases} p_i + (1 - p_i)\exp(-\mu_i) & y_i = 0 \\ (1 - p_i)\dfrac{\exp(-\mu_i){\mu_i}^{y_i}}{y_i!} & y_i > 0 \end{cases}$$

where $\mu_i = \exp(x_i\beta)$ and $p_i = g^{-1}(z_i\gamma)$ and where $g^{-1}(\cdot)$ is the inverse link function of the linear predictor $z_i\gamma$; our software allows specification of inverse link functions for logit, probit, loglog, and complementary loglog.

For a random sample of observations $y_1, y_2, \ldots, y_n$, the MZIP regression log-likelihood function is given by

$$\mathcal{L} = \sum_{i\in Z} \Big[\ln\left\{p_i + (1 - p_i)\exp(-\mu_i)\right\} \Big] + \sum_{i\notin Z} \Big\{ \ln(1 - p_i) - \mu_i + y_i \ln(\mu_i) - \Gamma(y_i + 1) \Big\}$$

where the mean (μ_i) is rescaled from the ZIP regression model to $\mu_i = \exp\{x_i\beta - \ln(1 - p_i)\}$ and Z is the set of zero outcomes.

2.2 MZIGP distribution

The ZIGP regression model with a count outcome variable, Y_i, where $i = 1, \ldots, n$, has the p.m.f.

$$P(Y_i = y_i) = \begin{cases} p_i + (1 - p_i)\exp(-\mu_i) & y_i = 0 \\ (1 - p_i)\dfrac{\mu_i(\mu_i + \delta y_i)^{y_i - 1}\exp(-\mu_i - \delta y_i)}{y_i!} & y_i > 0 \end{cases}$$

where $\mu_i = \exp(x_i\beta)$, $p_i = g^{-1}(z_i\gamma)$, and δ is the dispersion parameter having $0 \leq \delta < 1$. By applying the same concept from the MZIP regression model in section 2.1 to the ZIGP regression model, we introduce the MZIGP regression model. For a random sample of observations $y_1, y_2, \ldots, y_n$, the MZIGP regression log-likelihood function is

$$\mathcal{L} = \sum_{i \in Z} \Big[\ln\{p_i + (1 - p_i)\exp(-\mu_i)\} \Big] \\ + \sum_{i \notin Z} \Big\{ \ln(1 - p_i) + \ln(\mu_i) + (y_i - 1)\ln(\mu_i + \delta y_i) - \mu_i - \delta y_i - \ln\Gamma(y_i + 1) \Big\}$$

where the mean (μ_i) is rescaled from the ZIGP regression model to $\mu_i = \exp\{x_i\beta - \ln(1 - p_i)\}$, δ is the dispersion parameter having $0 \leq \delta < 1$, and Z is the set of zero outcomes.

2.3 MZINB distribution

The ZINB regression model with a count outcome variable Y_i, where $i = 1, \ldots, n$, has the p.m.f.

$$P(Y_i = y_i) = \begin{cases} p_i + (1 - p_i)\left(\dfrac{1}{1 + \delta\mu_i}\right)^{(1/\delta)} & y_i = 0 \\ (1 - p_i)\dfrac{\Gamma(1/\delta + y_i)}{\Gamma(y_i + 1)\Gamma(1/\delta)}\left(\dfrac{1}{1 + \delta\mu_i}\right)^{(1/\delta)}\left(1 - \dfrac{1}{1 + \delta\mu_i}\right)^{y_i} & y_i > 0 \end{cases}$$

where $\mu_i = \exp(x_i\beta)$, $p_i = g^{-1}(z_i\gamma)$, and δ is the dispersion parameter. Lastly, we apply the same concept from the MZIP regression model in section 2.1 to the ZINB regression model, and we introduce the MZINB regression model. For a random sample of observations $y_1, y_2, \ldots, y_n$, the MZINB regression log-likelihood function is

$$\mathcal{L} = \sum_{i \in Z} \ln\left\{ p_i + (1 - p_i)\left(\frac{1}{1 + \delta\mu_i}\right)^{(1/\delta)} \right\} \\ + \sum_{i \notin Z} \Bigg[\ln(1 - p_i) + \ln\Gamma\{(1/\delta) + y_i\} - \ln\Gamma(y_i + 1) - \ln\Gamma\left(\frac{1}{\delta}\right) \\ + (1/\delta)\ln\left(\frac{1}{1 + \delta\mu_i}\right) + y_i\ln\left(1 - \frac{1}{1 + \delta\mu_i}\right) \Bigg]$$

where the mean (μ_i) is rescaled from the ZINB regression model to $\mu_i = \exp\{x_i\beta - \ln(1 - p_i)\}$, δ is the dispersion parameter, and Z is the set of zero outcomes.

3 Syntax

The accompanying software includes the command files and supporting files for prediction and help. In the following syntax diagrams, unspecified options include the usual collection of maximization and display options available for all estimation commands. All marginalized zero-inflated commands include the `ilink(`*linkname*`)` option to specify the link function for the inflation model. Allowable arguments to the `ilink()` option include `logit`, `probit`, `loglog`, or `cloglog`.

Equivalent in syntax to the `zip` command, the basic syntax for specifying an MZIP model for count data is

`mzip` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] `,`
  `inflate(`*varlist*[`,` `offset(`*varname*`)`] | `_cons)` [*options*]

The syntax for specifying an MZIGP distribution for count data is

`mzigp` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] `,`
  `inflate(`*varlist*[`,` `offset(`*varname*`)`] | `_cons)` [*options*]

The syntax for specifying an MZINB distribution for count data is

`mzinb` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] `,`
  `inflate(`*varlist*[`,` `offset(`*varname*`)`] | `_cons)` [*options*]

4 Examples

4.1 Example synthetic marginalized zero-inflated data

Here we illustrate how to generate synthetic marginalized zero-inflated data. We synthesized `trt` from a Bernoulli(0.5) and x_1 from a normal$(0, 1)$. The true parameter values are $\{\gamma_0 = 0.80, \beta_0 = \log(1.75), \gamma_1 = -0.25, \beta_1 = \log(1.25), \gamma_2 = -0.50, \beta_2 = \log(1.45)\}$ (see parameter definitions and references in section 2.1). To highlight the differences between using nonzero-inflated and nonmarginalized zero-inflated models compared with marginalized zero-inflated models, we will fit our data with three separate models—Poisson, ZIP, and MZIP. We will also highlight the use of the average predicted value described in Albert, Wang, and Nelson (2014) to estimate the total effect of the `trt` variable in the ZIP model.

```
. set seed 23982
. set obs 10000
number of observations (_N) was 0, now 10,000
. // Linear predictor for the outcome
. generate trt = rbinomial(1, .5)
. generate x1  = rnormal()
. generate z1  = runiform()
. generate xb  = log(1.75) + log(1.25)*trt + log(1.45)*x1
. // Linear predictor for the zero-inflation
. generate zg = 0.80 - 0.25*trt - 0.50*z1
. // Define the mean of the count distribution and generate y
. generate mu = exp(xb + ln(1+exp(zg)))
. generate y = rpoisson(mu)
. // Mix in zero-outcomes from the inflation to y
. generate p0 = exp(zg)/(1+exp(zg)) // Inflation is in terms of P(Y=0)
. generate u = runiform()
. replace y = 0 if p0 > u
(5,934 real changes made)
```

Having created an outcome with our specified associations, we can fit some models (below) to see how closely the sample data match the specifications. The first model using our marginalized zero-inflated synthesized data with a Poisson distribution shows that using the robust variance estimator does a good job adjusting for the overdispersion due to the excess zeros (compared with the marginalized ZIP results at the end of this section).

```
. poisson y trt x1, nolog irr robust
Poisson regression                              Number of obs     =     10,000
                                                Wald chi2(2)      =     598.07
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -27929.481               Pseudo R2         =     0.0488
```

y	IRR	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
trt	1.27774	.0387465	8.08	0.000	1.204011	1.355984
x1	1.423367	.0220454	22.79	0.000	1.380808	1.467238
_cons	1.718827	.0388726	23.95	0.000	1.644302	1.796729

```
Note: _cons estimates baseline incidence rate.
```

However, when we fit our ZIP model to our sample data, we see a worse match to our synthetic-data specifications. The estimated coefficients for both of the nonzero-inflated components are not close to the values from our synthesized data. However, we can use a program to calculate the difference and ratio versions of the average predicted value.

```
. capture program drop GetAPV
. program define GetAPV
  1.     syntax varlist(min=1 max=1)
  2.     quietly {    // There is no error checking in this program
  3.         local trt `varlist´
  4.         tempvar bu mu0 mu1
  5.         generate `bu´ = `trt´
  6.         replace `trt´=0
  7.         predict double `mu0´
  8.         replace `trt´=1
  9.         predict double `mu1´
 10.         replace `trt´ = `bu´
 11.         tempname bb
 12.         bootstrap, reps(200) : mean `mu1´ `mu0´
 13.         mat `bb´ = r(table)
 14.         noisily display as txt _n "APV for `trt´"
 15.         noisily display as txt    "APV(difference) = "
>             as result %9.0g (`bb´[1,1] - `bb´[1,2])
 16.         noisily display as txt    "APV(ratio)      = "
>             as result %9.0g (`bb´[1,1] / `bb´[1,2])
 17.     }
 18. end
```

The ratio version of the average predicted value depicted above illustrates the total estimated effect of the `trt` variable. This same effect is what is estimated by the Poisson and MZIP models. That is, when the value of `trt` is changed, it affects the rate and probability of zero-outcomes.

```
. zip y trt x1, inflate(trt z1) irr nolog
Zero-inflated Poisson regression                Number of obs     =     10,000
                                                Nonzero obs       =      3,895
                                                Zero obs          =      6,105
Inflation model = logit                         LR chi2(2)        =    2548.01
Log likelihood  = -15115.84                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           y |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y            |
         trt |    1.08098   .0154567     5.45   0.000     1.051106    1.111703
          x1 |   1.444892   .0106511    49.93   0.000     1.424167    1.465919
       _cons |   4.727234   .0528784   138.87   0.000     4.624722    4.832018
-------------+----------------------------------------------------------------
inflate      |
         trt |  -.2906165    .041755    -6.96   0.000    -.3724548   -.2087781
          z1 |  -.5053849   .0725662    -6.96   0.000    -.6476122   -.3631577
       _cons |   .8245697   .0476765    17.30   0.000     .7311255    .9180138
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
Note: _cons estimates baseline incidence rate.

. GetAPV trt
APV for trt
APV(difference) =  .5228705
APV(ratio)      =  1.287507
```

Finally, we fit the data with the MZIP regression model with requested exponentiated coefficients. As expected, because the data are generated according to this model, they are well estimated.

```
. mzip y trt x1, inflate(trt x1) eform nolog
Marginalized Zero-inflated Poisson regression       Number of obs   =      10000
                                                    Nonzero obs     =       3895
Inflation link : logit                              Zero obs        =       6105
                                                    LR chi2(2)      =     681.62
Log likelihood = -15140.09                          Prob > chi2     =     0.0000

------------------------------------------------------------------------------
           y |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y            |
         trt |   1.290013   .0369601     8.89   0.000     1.219569    1.364526
          x1 |    1.43612   .0203707    25.52   0.000     1.396744    1.476606
       _cons |   1.706826   .0373099    24.46   0.000     1.635244    1.781541
-------------+----------------------------------------------------------------
inflate      |
         trt |  -.2920746   .0416405    -7.01   0.000    -.3736884   -.2104607
          x1 |   .0104196   .0210467     0.50   0.621    -.0308311    .0516703
       _cons |   .5702831   .0301419    18.92   0.000      .511206    .6293601
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
```

4.2 Example real-world study

We use the popular German health reform data for the year 1984 as example data. The goal of our example is to understand the number of visits made to a physician during 1984. Our predictor of interest is whether the patient is highly educated based on achieving a graduate degree (`edlevel4`), for example, MA/MS, MBA, PhD, or a professional degree. Confounding predictors are age (`age`) ranging from 25–64 and income in German marks (`hhninc`) divided by 10. Almost half the time (42%), the patients did not visit the doctor (excess zero counts). Therefore, a zero-inflated model would be appropriate to model this data. We model the data using our MZIGP and MZINB regression models, which we explained earlier.

```
. use rwm1984, clear
(German health data for 1984; Hardin & Hilbe, GLM and Extensions, 4th ed)
. generate hh = hhninc/10
. mzigp docvis edlevel4 age hh, inflate(edlevel4 age hh) nolog
Marginalized Zero-inflated Gen Poisson regression Number of obs   =       3874
                                                  Nonzero obs     =       2263
Inflation link : logit                            Zero obs        =       1611
                                                  LR chi2(3)      =     155.79
Log likelihood = -8295.035                        Prob > chi2     =     0.0000

------------------------------------------------------------------------------
      docvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
docvis       |
    edlevel4 |  -.2510622   .1146271    -2.19   0.029    -.4757272   -.0263972
         age |   .0219337   .0019299    11.37   0.000     .0181512    .0257162
          hh |  -.5784546   .1655214    -3.49   0.000    -.9028707   -.2540385
       _cons |   .3343386   .1081728     3.09   0.002     .1223237    .5463534
-------------+----------------------------------------------------------------
inflate      |
    edlevel4 |   .7483066   .3594542     2.08   0.037     .0437892    1.452824
         age |  -.0237062    .007483    -3.17   0.002    -.0383727   -.0090397
          hh |  -.6331557   .7837937    -0.81   0.419    -2.169363    .9030518
       _cons |  -.1265331   .4166084    -0.30   0.761    -.9430705    .6900044
-------------+----------------------------------------------------------------
 /atanhdelta |   .7732168   .0170575    45.33   0.000     .7397848    .8066488
-------------+----------------------------------------------------------------
       delta |   .6487961   .0098774                      .6290151    .6677374
------------------------------------------------------------------------------

. estat ic
Akaike´s information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
           . |      3,874  -8372.932  -8295.035       9    16608.07   16664.43
-----------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] BIC note.
```

From the output, variables `edlevel4` and `age` appear to affect zero counts, with younger graduate patients less likely to see a physician at all during the year. Patients not at the graduate level made about 22% [exp(−0.251)] fewer visits than graduate school patients. All three variables (`edlevel4`, `age`, `hh`) affect the nonzero counts significantly at $\alpha = 0.05$. Also note that the dispersion parameter $\delta = 0.6488$ is statistically significant, showing the overdispersion in the data.

```
. mzinb docvis edlevel4 age hh, inflate(edlevel4 age hh) nolog
Marginalized Zero-inflated neg bin regression     Number of obs   =      3874
                                                  Nonzero obs     =      2263
Inflation link : logit                            Zero obs        =      1611
                                                  LR chi2(3)      =    161.29
Log likelihood = -8330.529                        Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      docvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
docvis       |
    edlevel4 |  -.2981929   .1278499    -2.33   0.020    -.5487741   -.0476117
         age |   .0258583   .0022949    11.27   0.000     .0213604    .0303562
          hh |  -.7939298   .1620665    -4.90   0.000    -1.111574   -.4762853
       _cons |   .2101778   .1185294     1.77   0.076    -.0221355    .4424911
-------------+----------------------------------------------------------------
inflate      |
    edlevel4 |   1.136279   .3597673     3.16   0.002      .431148     1.84141
         age |  -.0554078   .0128414    -4.31   0.000    -.0805764   -.0302392
          hh |   .1139156   .9168216     0.12   0.901    -1.683022    1.910853
       _cons |   .1137756   .4906687     0.23   0.817    -.8479174    1.075469
-------------+----------------------------------------------------------------
    /lnalpha |   .6231516   .0651368     9.57   0.000     .4954857    .7508174
-------------+----------------------------------------------------------------
       alpha |   1.864796   .1214669                      1.641295    2.118731
------------------------------------------------------------------------------

. estat ic
Akaike´s information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
           . |      3,874  -8411.173  -8330.529       9    16679.06   16735.42
-----------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] BIC note.
```

Similarly, from the output, variables `edlevel4` and `age` appear to affect zero counts, with younger graduate patients less likely to see a physician at all during the year. Patients not at the graduate level made about 26% [exp(−0.298)] fewer visits than graduate school patients. All three variables (`edlevel4`, `age`, `hh`) affect the nonzero counts significantly at $\alpha = 0.05$. Also note that the dispersion parameter $\delta = 1.865$ is statistically significant, showing the overdispersion in the data. The Akaike information criterion and Bayesian information criterion statistics are slightly lower in the MZIGP regression model, indicating a much better fit than the MZINB regression model.

5 Summary

In this article, we introduced supporting programs for modeling count data using marginalized zero-inflated distributions. We illustrated the use of the new command `mzip` using synthesized data, and we illustrated the new commands `mzigp` and `mzinb` using real-world German health data from 1984.

6 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 19-3
. net install st0563     (to install program files, if available)
. net get st0563         (to install ancillary files, if available)
```

7 References

Albert, J. M., W. Wang, and S. Nelson. 2014. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Statistical Methods in Medical Research* 23: 257–278.

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.

Famoye, F., and K. Singh. 2006. Zero-inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science* 4: 117–130.

Greene, W. H. 1994. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper Series EC-94-10, Department of Economics, Stern School of Business, New York University.

Hardin, J. W., and J. M. Hilbe. 2018. *Generalized Linear Models and Extensions*. 4th ed. College Station, TX: Stata Press.

Heagerty, P. J., and S. L. Zeger. 2000. Marginalized multilevel models and likelihood inference. *Statistical Science* 15: 1–19.

Kassahun, W., T. Neyens, G. Molenberghs, C. Faes, and G. Verbeke. 2014. Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeros. *Statistics in Medicine* 33: 4402–4419.

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

Lee, K., Y. Joo, J. J. Song, and D. W. Harper. 2011. Analysis of zero-inflated clustered count data: A marginalized model approach. *Computational Statistics & Data Analysis* 55: 824–837.

Long, D. L., J. S. Preisser, A. H. Herring, and C. E. Golin. 2014. A marginalized zero-inflated Poisson regression model with overall exposure effects. *Statistics in Medicine* 33: 5151–5165.

Mullahy, J. 1986. Specification and testing of some modified count data models. *Journal of Econometrics* 33: 341–365.

Preisser, J. S., K. Das, D. L. Long, and K. Divaris. 2016. Marginalized zero-inflated negative binomial regression with application to dental caries. *Statistics in Medicine* 35: 1722–1735.

Ullah, S., C. F. Finch, and L. Day. 2010. Statistical modelling for falls count data. *Accident Analysis and Prevention* 42: 384–392.

About the authors

Tammy H. Cummings is a senior research associate in the Institute for Families in Society at the University of South Carolina in Columbia, SC.

James W. Hardin is a professor in the Department of Epidemiology and Biostatistics and the associate dean of Faculty Affairs and Curriculum for the Arnold School of Public Health at the University of South Carolina in Columbia, SC.