Version info: Code for this page was tested in SAS 9.3

Zero-inflated Poisson regression is used to model count data that has an excess of zero counts. Further, theory suggests that the excess zeros are generated by a separate process from the count values and that the excess zeros can be modeled independently. Thus, the **zip** model has two parts, a Poisson count model and the logit model for predicting excess zeros. You may want to review these Data Analysis Example pages, Poisson Regression (/stata/dae/poisson-regression/) and Logit Regression (/stata/dae/logistic-regression/).

Please Note: The purpose of this page is to show how to use various data analysis commands. It does not cover all aspects of the research process which researchers are expected to do. In particular, it does not cover data cleaning and verification, verification of assumptions, model diagnostics and potential follow-up analyses.

Examples of zero-inflated Poisson regression

Example 1. School administrators study the attendance behavior of high school juniors over one semester at two schools. Attendance is measured by number of days of absent and is predicted by gender of the student and standardized test scores in math and language arts. Many students have no absences during the semester.

Example 2. The state wildlife biologists want to model how many fish are being caught by fishermen at a state park. Visitors are asked whether or not they have a camper, how many people were in the group, were there children in the group and how many fish were caught. Some visitors do not fish, but there is no data on whether a person fished or not. Some visitors who did fish did not catch any fish so there are excess zeros in the data because of the people that did not fish.

Description of the data

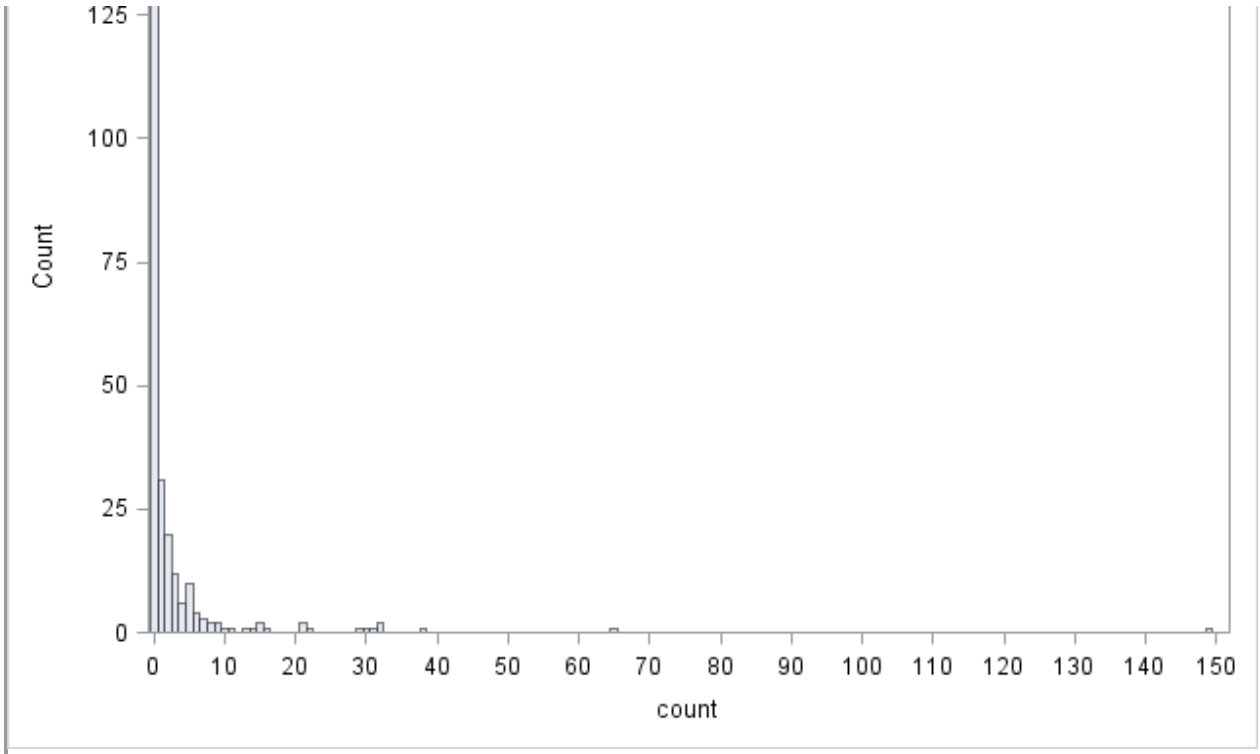
Let's pursue Example 2 from above using the dataset fish (https://stats.idre.ucla.edu/wp-content/uploads/2017/01/fish.sas7bdat).

simply a result of bad luck fishing. We will use the variables **child**, **persons**, and **camper** in our model. Let's look at the data.

```
The MEANS Procedure

Variable            Mean         Std Dev         Minimum         Maximum        Variance
-------------------------------------------------------------------------------------------
count          3.2960000      11.6350281               0     149.0000000     135.3738795
child          0.6840000       0.8503153               0       3.0000000       0.7230361
persons        2.5280000       1.1127303       1.0000000       4.0000000       1.2381687
-------------------------------------------------------------------------------------------
```

```
proc univariate data = fish noprint;
  histogram count / midpoints = 0 to 50 by 1 vscale = count ;
run;
```




```
proc freq data = fish;
  tables camper;
run;
```

```
The FREQ Procedure

                                        Cumulative    Cumulative
```

Analysis methods you might consider

Below is a list of some analysis methods you may have encountered. Some of the methods listed are quite reasonable while others have either fallen out of favor or have limitations.

- Zero-inflated Poisson Regression – The focus of this web page.
- Zero-inflated Negative Binomial Regression – Negative binomial regression does better with over dispersed data, i.e., variance much larger than the mean.
- Ordinary Count Models – Poisson or negative binomial models might be more appropriate if there are no excess zeros.
- OLS Regression – You could try to analyze these data using OLS regression. However, count data are highly non-normal and are not well estimated by OLS regression.

SAS zero-inflated Poisson regression analysis using proc genmod

If you are using SAS version 9.2 or higher, you can run a zero-inflated Poisson model using **proc genmod**.

```
run;
```

The GENMOD Procedure

Model Information

Data Set		WORK.FISH	Written by SAS
Distribution	Zero Inflated Poisson		
Link Function		Log	
Dependent Variable		count	

Number of Observations Read	250
Number of Observations Used	250

Class Level Information

Class	Levels	Values
camper	2	0 1

Criteria For Assessing Goodness Of Fit

Scaled Deviance		2063.2168	
Pearson Chi-Square	245	1543.4597	6.2998
Scaled Pearson X2	245	1543.4597	6.2998
Log Likelihood		774.8999	
Full Log Likelihood		-1031.6084	
AIC (smaller is better)		2073.2168	
AICC (smaller is better)		2073.4627	
BIC (smaller is better)		2090.8241	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	2.4319	0.0413	2.3510	2.5128	3472.23	ChiSq
Intercept	1	1.2974	0.3739	0.5647	2.0302	12.04	0.0005
persons	1	-0.5643	0.1630	-0.8838	-0.2449	11.99	0.0005

The last block of output corresponds to the zero-inflation portion of the model. This is a logistic model predicting the zeroes. The output includes parameter estimates for the inflation model predictors and their standard errors, Wald 95% confidence intervals, Wald Chi-square statistics, and p-values.

All of the predictors in both the count and inflation portions of the model are statistically significant. This model fits the data significantly better than the null model, i.e., the intercept-only model. To show that this is the case, we can run the null model (a model without any predictors) and compare the null model with the current model using chi-squared test on the difference of log likelihoods.

```
run;
```

The GENMOD Procedure

Model Information

Data Set	WORK.FISH Written by SAS
Distribution	Zero Inflated Poisson
Link Function	Log
Dependent Variable	count

Number of Observations Read	250
Number of Observations Used	250

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance		2254.0459	
Scaled Deviance		2254.0459	
Pearson Chi-Square	248	1918.7890	7.7371
Scaled Pearson X2	248	1918.7890	7.7371

```
                BIC (smaller is better)                2265.0888


         Algorithm converged.


                    Analysis Of Maximum Likelihood Parameter Estimates

                                    Standard     Wald 95% Confidence          Wald
Parameter     DF    Estimate          Error           Limits            Chi-Square    Pr > ChiSq

Intercept      1      2.0316         0.0349        1.9631       2.1000       3388.16          ChiSq

Intercept      1      0.2728         0.1277        0.0225       0.5232          4.56        0.0327
```

The log likelihoods for the full model and null mode are -1031.6084 and -1127.0229, respectively. The chi-squared value is 2*(-1031.6084 – -1127.0229) = 190.829. Since we have three predictor variables in the full model, the degrees of freedom for the chi-squared test is 3. This yields a p-value <.0001. Thus, our overall model is statistically significant.

We may want to compare the current zero-inflated Poisson model with the plain poisson model, which can be done with the Vuong test. Currently, the Vuong test is not a standard part of **proc genmod**, but a macro program that performs the Vuong test is available from SAS here (http://support.sas.com/kb/42/514.html). Usage of the macro program requires the **%include** statement, in which we list the location of the macro. This macro program takes quite a few arguments, as shown below. We rerun the models to get produce these required input

```
  class camper;
  model count = child camper /dist=zip;
  zeromodel persons;
  output out=outzip pred=predzip pzero=p0;
  store m1;
run;
proc genmod data = outzip order=data;
  class camper;
  model count = child camper /dist=poi;
  output out=out pred=predpoi;
run;
%vuong(data=out, response=count,
       model1=zip, p1=predzip, dist1=zip, scale1=1.00, pzero1=p0,
       model2=poi, p2=predpoi, dist2=poi, scale2=1.00,
       nparm1=3,   nparm2=2)
```

The Vuong Macro
Model Information

Data Set	out
Response	count
Number of Observations Used	250

Scale	1.00
Zero-inflation Probability	p0
Log Likelihood	-1031.6084
Model 2	poi
Distribution	POI
Predicted Variable	predpoi
Number of Parameters	2
Scale	1.00
Log Likelihood	-1358.5929

Vuong Test

H0: models are equally close to the true model

Ha: one of the models is closer to the true model

Vuong Statistic	Z	Pr>|Z|	Preferred Model
Unadjusted	3.5814	0.0003	zip
Akaike Adjusted	3.5705	0.0004	zip
Schwarz Adjusted	3.5512	0.0004	zip

Clarke Sign Test

H0: models are equally close to the true model

Ha: one of the models is closer to the true model

```
Unadjusted           13.0000    0.1137        zip
Akaike Adjusted       2.0000    0.8496        zip
Schwarz Adjusted      2.0000    0.8496        zip
```

For the Vuong test, a significant z indicates that the zero-inflated model is better. Here we see that the preferred model is a zero-inflated Poisson model over a regular Poisson model. The positive values of the z statistics for Vuong test indicate that it is the first model, the zero-inflated poisson model, which is closer to the true model.

We can use the **estimate** statement to help understand our model. We will compute the expected counts for the categorical variable **camper** while holding the continuous variable **child** at its mean value using the **atmeans** option, as well as calculate the predicted probability that an observation came from the zero-generating process. In the estimate statement, we provide values at which to evaluate each coefficient for both the Poisson model and the zero-inflation model. The sets of coefficients of the two models are separated by the @ZERO keyword.

```
  zeromodel persons /link = logit ;
  estimate "camper = 0" intercept 1 child .684 camper 1 0 @ZERO intercept 1 persons 2.528;
  estimate "camper = 1" intercept 1 child .684 camper 0 1 @ZERO intercept 1 persons 2.528;
run;
```

Contrast Estimate Results

Label	Mean Estimate	Mean Confidence Limits		L'Beta Estimate	Standard Error	Alpha
camper = 0	2.4220	1.9724	2.9741	0.8846	0.1048	0.05
camper = 0 (Zero Inflation)	0.4677	0.3838	0.5536	-0.1292	0.1756	0.05
camper = 1	5.5768	4.8823	6.3701	1.7186	0.0679	0.05
camper = 1 (Zero Inflation)	0.4677	0.3838	0.5536	-0.1292	0.1756	0.05

Contrast Estimate Results

Label	L'Beta Confidence Limits		Chi-Square	Pr > ChiSq
camper = 0	0.6792	1.0899	71.28	<.0001
camper = 0 (Zero Inflation)	-0.4735	0.2150	0.54	0.4619
camper = 1	1.5856	1.8516	641.42	<.0001
camper = 1 (Zero Inflation)	-0.4735	0.2150	0.54	0.4619

In the Mean Estimate column, we find predicted counts of fish from the Poisson model, ignoring the zero-inflation model, for both **camper** = 0 and **camper** = 1, as well as the predicted probability of belonging to the zero-generating process from the zero-inflation model. The zero-inflation model does not include **camper** as a predictor, so the probability of zero for both zero-inflation models is the same. To get the expected counts of **fish** from the mixture of the two models, simply multiply the expected counts from the Poisson model by the probability of getting a non-zero from the zero-inflation model (1 – p(zero)). Thus, the expected counts of fish for **camper** = 0 and **camper** = 1 including zero-inflation are 2.422*(1-0.4677) = 1.289 and 5.5768*(1-0.4677) = 2.968, respectively.

SAS zero-inflated Poisson analysis using proc countreg

Proc countreg is another option for running a zero-inflated Poisson regression in SAS (again, version 9.2 or higher). This procedure allows for a few more options specific to count outcomes than **proc genmod**. The **proc countreg** code for the original model run on this page appears below. We indicate **method** = **qn** to specify the quasi-Newton optimization process that matches the **proc genmod** results.

```
  zeromodel count ~ persons;
run;
```

The COUNTREG Procedure

Class Level Information

Class	Levels	Values
camper	2	0 1

Model Fit Summary

Dependent Variable	count
Number of Observations	250
Data Set	MYLIB.FISH
Model	ZIP
ZI Link Function	Logistic
Log Likelihood	-1032
Maximum Absolute Gradient	3.69075E-7
Number of Iterations	13
Optimization Method	Quasi-Newton
AIC	2075

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	2.431911	0.041271	58.93	<.0001
child	1	-1.042838	0.099988	-10.43	<.0001
camper 0	1	-0.834022	0.093627	-8.91	<.0001
camper 1	0	0	.	.	.
Inf_Intercept	1	1.297439	0.373850	3.47	0.0005
Inf_persons	1	-0.564347	0.162962	-3.46	0.0005

SAS Zero-inflated Poisson analysis using proc nlmixed

For those using a version of SAS prior to 9.2, a zero-inflated negative binomial model is doable, though significantly more difficult. Please see this code fragment: Zero-inflated Poisson and Negative Binomial Using Proc Nlmixed (/sas/code/zero-inflated-poisson-and-negative-binomial-using-proc-nlmixed/).

Things to consider

- Since **zip** has both a count model and a logit model, each of the two models should have good predictors. The two models do

exposure into your model by using the **exposure()** option.

- It is not recommended that zero-inflated poisson models be applied to small samples. What constitutes a small sample does not seem to be clearly defined in the literature.
- Pseudo-R-squared values differ from OLS R-squareds, please see FAQ: What are pseudo R-squareds? (https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/) for a discussion on this issue.

See also

- Annotated output for the poisson command (/sas/output/poisson-regression/)
- SAS Online Manual
  - proc genmod (https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#genmod_toc.htm)

References

- Cameron, A. Colin and Trivedi, P.K. (2009) Microeconometrics using stata. College Station, TX: Stata Press.
- Long, J. Scott (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

