

What is the concept and application of Zero-inflated Poisson Regression in Mplus data analysis?

Authored by
stats writer

June 29, 2024

RECOMMENDED CITATION

stats writer (2024). *What is the concept and application of Zero-inflated Poisson Regression in Mplus data analysis?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=158242>

Zero-inflated Poisson Regression is a statistical technique used in data analysis, particularly in Mplus, to address the problem of excess zeros in count data. It combines the Poisson regression model with an additional component to account for the excessive number of zeros in the data. This approach is especially useful when the data has a large number of zeros, which is common in social science research. The concept behind this method is that it assumes two processes contribute to the generation of zeros in the data - one that follows the Poisson distribution and another that produces excessive zeros. By accounting for both processes, this method provides a more accurate estimation of the relationship between the predictors and the outcome variable. This technique is widely used in fields such as health, education, and economics to analyze count data with a large number of zeros.

Zero-inflated Poisson Regression | Mplus Data Analysis Examples

Version info: Code for this page was tested in Mplus version 6.12.

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 and

Logit 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. The associated dataset can be found [here](#).

We have data on 250 groups that went to a park. Each group was questioned before leaving the park about how many fish they caught (count), how many children were in the group (child), how many people were in the group (persons), and whether or not they brought a camper to the park (camper). The outcome

variable of interest will be the number of fish caught. Even though the question about the number of fish caught was asked to everyone, it does not mean that everyone went fishing. What would be the reason for someone to report a zero count? Was it because this person was unlucky and didn't catch any fish, or was it because this person didn't go fishing at all? If a person didn't go fishing, the outcome would be always zero. Otherwise, if a person went to fishing, the count could be zero or non-zero. So we can see that there seemed to be two processes that would generate zero counts: unlucky in fishing or didn't go fishing.

Let's first look at the data. We will start with reading in the data and the descriptive statistics and plots. This helps us understand the data and give us some hint on how we should model the data.

Let's look at the data.

Data:

File is C:fish.dat;

Variable:

Names are

nofish livebait camper persons child xb zg count;

Missing are all (-9999);

Usevariables are

camper persons child count;

Analysis:

type = basic;

Plot:

type = plot1;

ESTIMATED SAMPLE STATISTICS

Means

CAMPER PERSONS CHILD COUNT

1 0.588 2.528 0.684 3.296

Covariances

CAMPER PERSONS CHILD COUNT

CAMPER 0.242

PERSONS -0.026 1.233

CHILD -0.014 0.515 0.720

COUNT 0.730 2.856 -1.670 134.832

Correlations

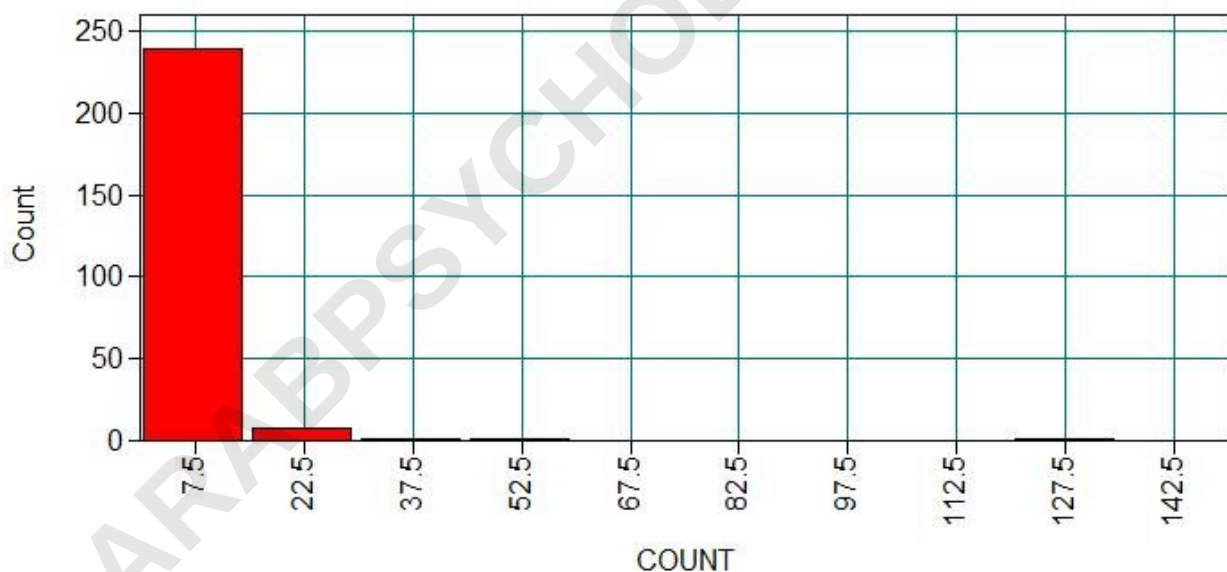
CAMPER PERSONS CHILD COUNT

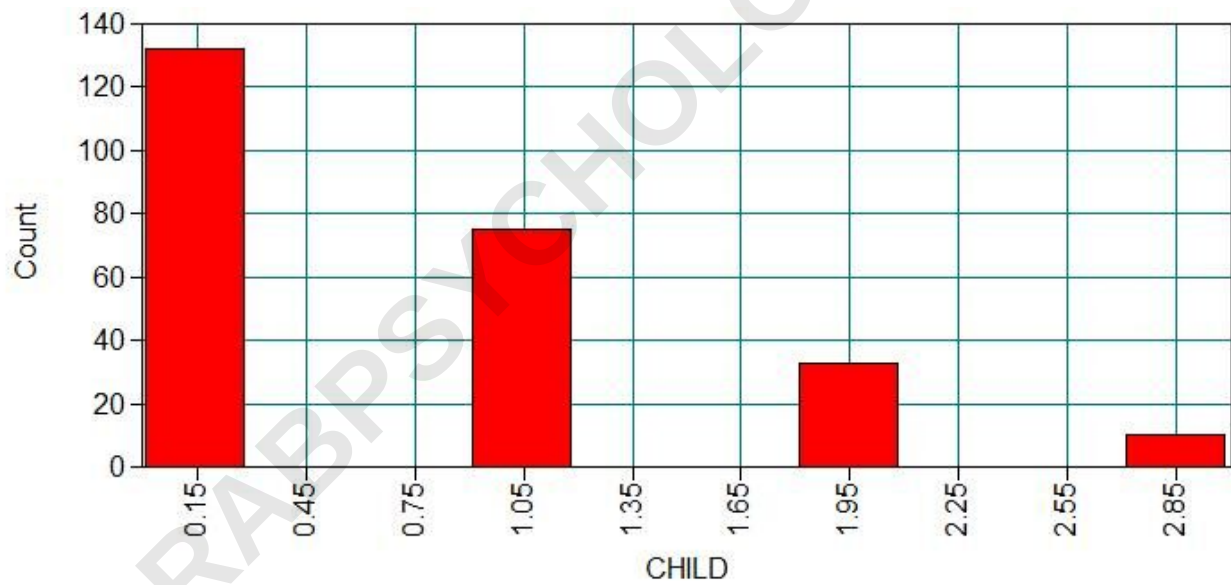
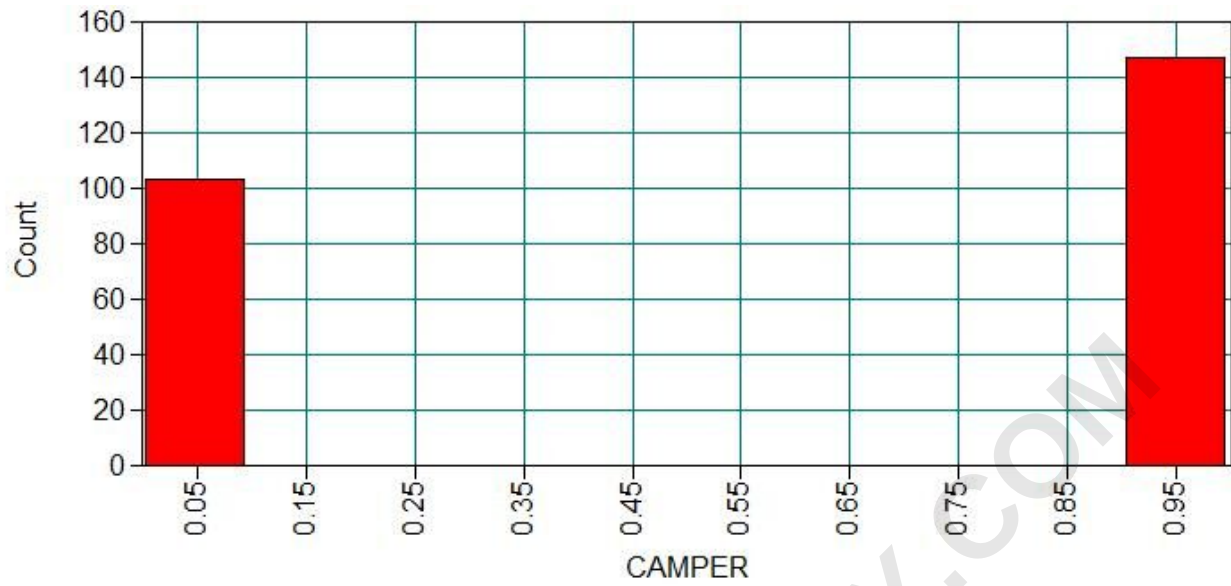
CAMPER 1.000

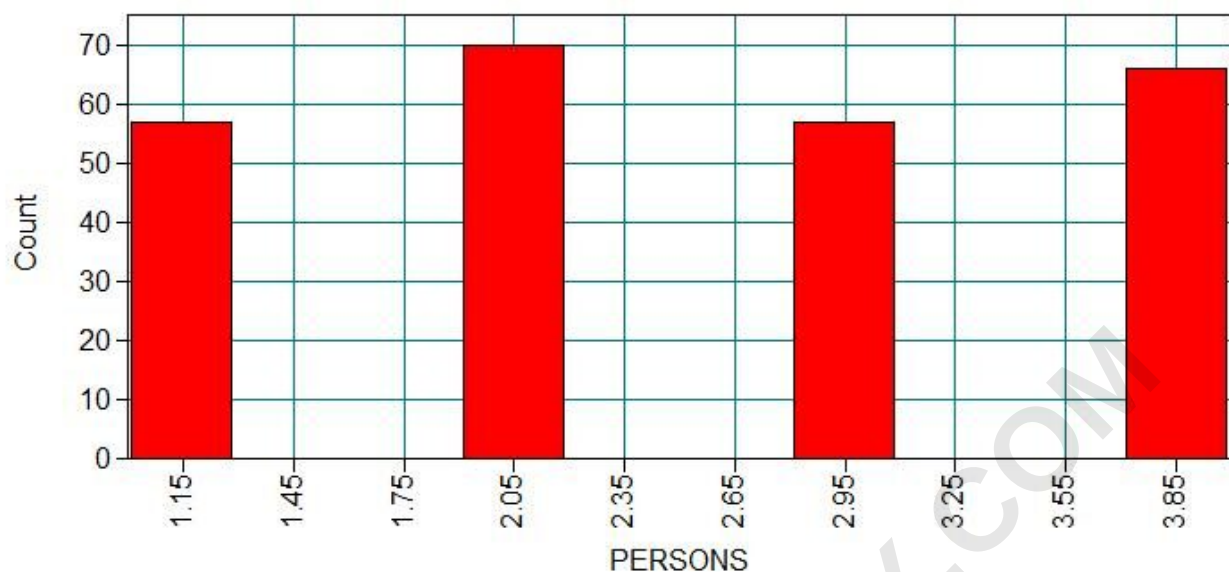
PERSONS -0.048 1.000

CHILD -0.034 0.546 1.000

COUNT 0.128 0.221 -0.170 1.000







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

In the syntax below, we have indicated that count is a count

variable by using the count statement. The (i) option is used to indicate that we are specifying a zero-inflated poisson model.

Without the (i) option, we would be estimating a poisson model without zero-inflation. Also, we use the usevariables statement to indicate that we are not using all of the variables in the data set in the current model.

We have omitted the missing statement because we have no missing data in this data set. The default estimation method is MLR - maximum likelihood parameter estimates with standard errors and a chi-square test statistic that are robust to non-normality and non-independence of observations when used with type = complex. The MLR standard errors are computed using a sandwich estimator. This is what we generally call robust standard errors. To get the "regular" standard errors, we use the estimator = ml on the analysis statement. (In the next example, we will omit the analysis statement and obtain the robust standard errors.)

Two regression equations are specified in the model

statement: the first equation is the poisson model, predicting the count of fish using child and camper. The second equation is the logit model, indicated by count#1, predicting membership to the zero generating process using persons.

Data:

File is C:fish.dat;

Variable:

Names are

nofish livebait camper persons child xb zg count;

Count is count(i);

Usevariables are camper persons child count;

Analysis:

estimator = ml;

Model:

count on child camper;

count#1 on persons;

MODEL FIT INFORMATION

Number of Free Parameters 5

Loglikelihood

H0 Value -1031.608

Information Criteria

Akaike (AIC) 2073.217

Bayesian (BIC) 2090.824

Sample-Size Adjusted BIC 2074.974

($n^* = (n + 2) / 24$)

MODEL RESULTS

Two-Tailed

Estimate S.E. Est./S.E. P-Value

COUNT ON

CHILD -1.043 0.100 -10.430 0.000

CAMPER 0.834 0.094 8.908 0.000

COUNT#1 ON

PERSONS -0.564 0.163 -3.463 0.001

Intercepts

COUNT#1 1.297 0.374 3.470 0.001

COUNT 1.598 0.086 18.680 0.000

In the **MODEL FIT INFORMATION** portion of the output, you will find the log likelihood for the final model as well as a number of fit statistics. In the **MODEL RESULTS** section of the output you will find the poisson regression coefficients (estimates) for each of the variables, standard errors and the ratio of the estimate to its standard error. This can be used as a Z test, where values greater than 2 are considered to be statistically significant. Following these are logit coefficients for predicting excess zeros. In the above output, we see that child and camper are both significant predictors of count, and persons is a significant predictor in the logit model. Thus for each additional child, the log count of number of fish count decreases by 1.043. For each additional person, the log odds of membership to the excess zero-generating process decreases by 0.564.

Now let's rerun the model without the analysis statement in order to obtain robust standard errors.

Data:

File is C:fish.dat;

Variable:

Names are

nofish livebait camper persons child xb zg count;

Count is count(i);

Missing are all (-9999);

Usevariables are camper persons child count;

Model:

count on child camper;

count#1 on persons;

MODEL RESULTS

Two-Tailed

Estimate S.E. Est./S.E. P-Value

COUNT ON

CHILD -1.043 0.389 -2.684 0.007

CAMPER 0.834 0.407 2.050 0.040

COUNT#1 ON

PERSONS -0.564 0.288 -1.957 0.050

Intercepts

COUNT#1 1.297 0.493 2.632 0.008

COUNT 1.598 0.293 5.456 0.000

Robust standard errors tend to be larger than "regular" standard errors. Still we see that child, camper, and persons are still significant predictors within their respective models.

Things to consider

See Also

References