What is zero-inflated negative binomial regression and how is it used in Mplus data analysis?

Authored by

stats writer

June 29, 2024

Zero-inflated negative binomial regression is a statistical method used in Mplus data analysis to model count data with excessive zeros. It combines two models - a binomial model for predicting the excessive zeros and a negative binomial model for predicting the non-zero counts. This approach is useful when the data has a large number of zero values, which cannot be adequately captured by a traditional negative binomial regression model. It allows for the identification of factors that influence both the probability of a zero count and the count values, providing a more accurate and comprehensive analysis of the data. This technique is commonly used in social science research to analyze data on discrete outcomes, such as the frequency of a certain behavior or event.

Zero-inflated Negative Binomial Regression | Mplus Data Analysis Examples

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

Zero-inflated negative binomial regression is for modeling count variables with excessive zeros and it is usually for overdispersed count outcome variables. Furthermore, 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.

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 checking, verification of assumptions, model diagnostics or
potential follow-up analyses.

Examples of zero-inflated negative binomial regression

Example 1.

School administrators study the attendance behavior of high school juniors at two schools.
Predictors of the number of days of absence include gender of the student and standardized
test scores in math and language arts.

Example 2.

The state wildlife biologists want to model how many fish are being caught by fishermen
at a state park. Visitors are asked how long they stayed, 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 is 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

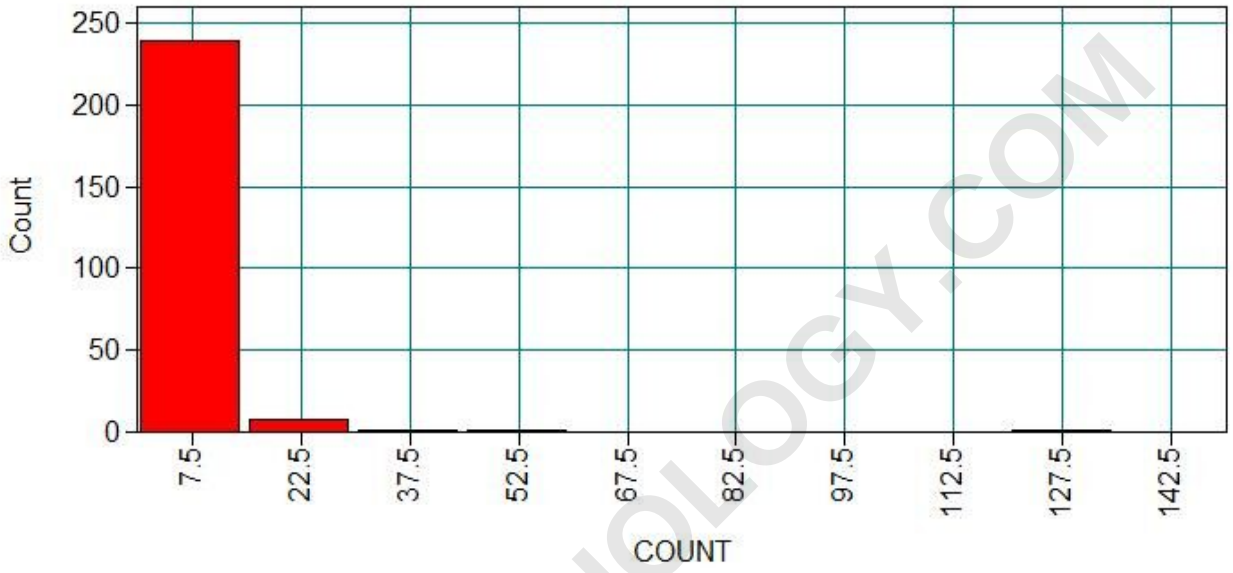
_____ _____ _____ _____

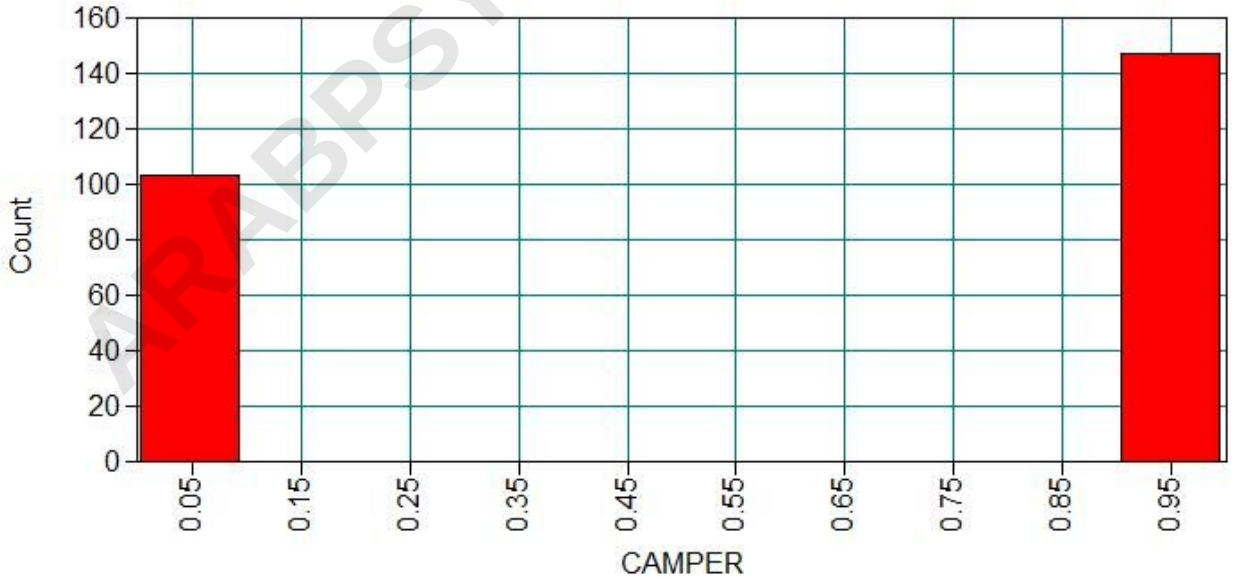
CAMPER 1.000

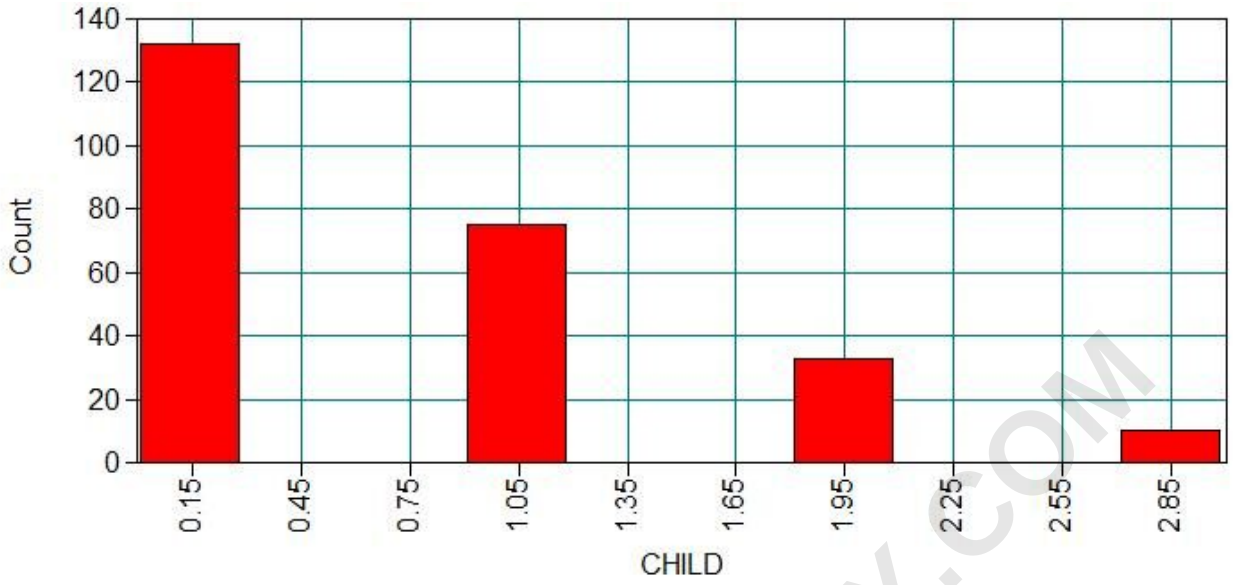
PERSONS -0.048 1.000

CHILD -0.034 0.546 1.000

COUNT 0.128 0.221 -0.170 1.000







Analysis methods you might consider

Before we show how you can analyze this with a zero-inflated negative binomial analysis, let's

consider some other methods that you might use.

Zero-inflated negative binomial regression

In the syntax below, we have indicated that count is a count
variable by using the count statement. The (nbi) option is
used to indicate 2 things: that we are modeling our count variable with a
negative binomial distribution, and that we are specifying a zero-inflated model.
Without the (nb) option we would be specifying a (zero-inflated) poisson
model, and without the (i) option, we would be estimating a negative
binomial 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.
Two regression equations are specified in the model statement: the first
equation is the negative binomial 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(nbi);

Usevariables are camper persons child count;

Analysis:

estimator = ml;

Model:

count on child camper;

count#1 on persons;

MODEL RESULTS

Two-Tailed

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

COUNT ON

CHILD -1.515 0.196 -7.747 0.000

CAMPER 0.879 0.269 3.265 0.001

COUNT#1 ON

PERSONS -1.666 0.679 -2.454 0.014

Intercepts

COUNT#1 1.603 0.836 1.916 0.055

COUNT 1.371 0.256 5.353 0.000

Dispersion

COUNT 2.679 0.471 5.683 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 negative binomial 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 both child and camper are significant predictor 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 1.515. For each additional person, the log odds of membership to the excess zero-

generating process
decreases by1.666.

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(nbi);
Usevariables are camper persons child count;
Model:
count on child camper;
count#1 on persons;

MODEL FIT INFORMATION

Number of Free Parameters 6

Loglikelihood

H0 Value -432.891
H0 Scaling Correction Factor 1.762

for MLR

Information Criteria

Akaike (AIC) 877.782
Bayesian (BIC) 898.911
Sample-Size Adjusted BIC 879.890
(n* = (n + 2) / 24)

MODEL RESULTS

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

COUNT ON
CHILD -1.515 0.241 -6.280 0.000
CAMPER 0.879 0.470 1.869 0.062

COUNT#1 ON
PERSONS -1.666 0.431 -3.871 0.000

Intercepts
COUNT#1 1.603 0.665 2.410 0.016
COUNT 1.371 0.389 3.520 0.000

Dispersion

COUNT 2.679 0.577 4.645 0.000

The robust standard errors attempt to adjust for heterogeneity in the model. Robust standard errors tend to be larger than "regular" standard errors for parameters in the negative binomial part of the model and smaller for parameters in the logit part of the model.

We see that now camper is not statistically significant.

Things to consider

Here are some issues that you may want to consider in the course of your research analysis.

References