

What is the annotated output for a zero-inflated Poisson regression in Stata?

Authored by
stats writer

June 30, 2024

RECOMMENDED CITATION

stats writer (2024). *What is the annotated output for a zero-inflated Poisson regression in Stata?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=160456>

A zero-inflated Poisson regression in Stata is a statistical model used to analyze count data with an excessive number of zeros. The annotated output for this type of regression includes two parts: the zero-inflated model component and the Poisson model component. The zero-inflated model component provides information on the excess zeros, such as the probability of a zero count and the zero-inflation parameter. The Poisson model component includes the coefficients for the predictors and the estimated count mean. The annotated output also includes diagnostic tests and graphs to assess the fitting of the model. Overall, the annotated output for a zero-inflated Poisson regression in Stata provides a comprehensive summary of the model's performance and helps interpret the results accurately.

Zero-Inflated Poisson Regression | Stata Annotated Output

This page shows an example of zero-inflated Poisson regression analysis with footnotes explaining the output in Stata. The data collected were academic information on 316 students at two different schools. The response variable is days absent during the school year (daysabs). We explore its relationship with math standardized test scores (mathnce), language standardized test scores (langnce), and gender (female).

As assumed for a Poisson model, our response variable is a count variable, each subject has the same length of

observation time, and the variance of the response variable is relatively close to the mean of the response variable. In a dataset in which the response variable is a count, the number of zeroes may seem excessive. With the example dataset in mind, consider the processes that could lead to a response variable value of zero. A student might be absent zero days during the school year if he never gets sick and never skips school. Another student might be absent zero days during the school year because her parents insist she go to school every day, regardless of illness or desire to skip school. These two students will look identical in the response variable, but they have arrived at the same outcome through two different processes. The first student could have been absent during the school year (had he become ill or opted to skip school), but was not. The second student was certain to be absent zero days. The second student will

be referred to from this point forward as a "certain zero". Thus, the number of zeroes may be inflated and the number of students absent for zero days cannot be explained in the same manner as the number of students that were absent for more than zero days. Some students were absent zero days for the same reasons other students were absent one, two, or three days (health and truancy) and while some students were absent zero days for a different set of reasons.

A standard Poisson model would not distinguish between the two processes causing an excessive number of zeroes, but a zero-inflated model allows for and accommodates this complication. When analyzing a dataset with an excessive number of outcome zeros and two possible processes that arrive at a zero outcome, a zero-inflated model should be considered.

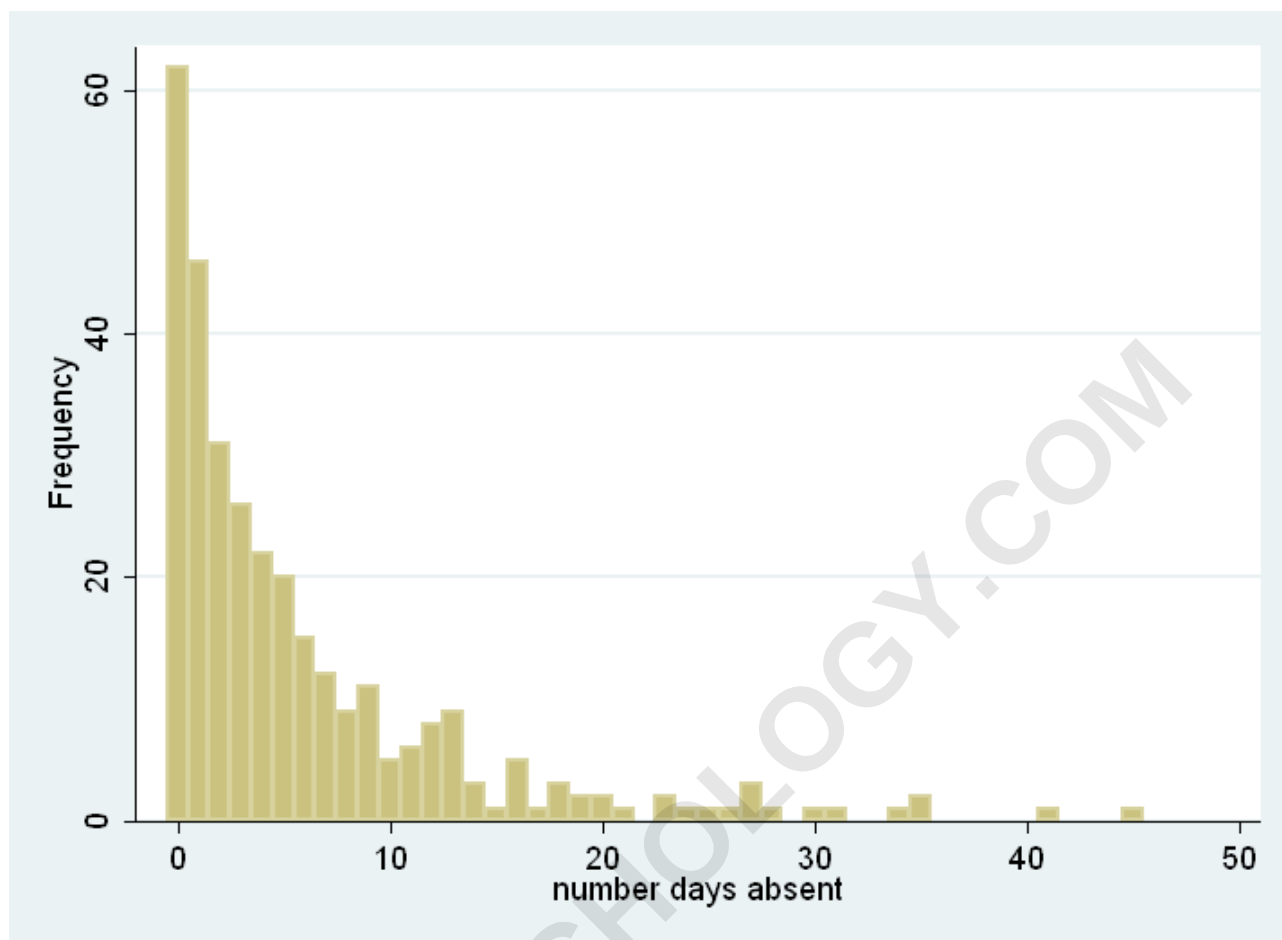
We can look at a histogram of the response variable to try to gauge if the number of zeros is excessive.

(If two processes generated the zeroes in the response variable but there is not an excessive number of zeroes, a zero-inflated model may or may not be used.)

**use <https://stats.idre.ucla.edu/stat/stata/notes/lahigh>,
clear**

generate female = (gender == 1)

histogram daysabs, discrete freq



While zero is the most common number of days absent, it is difficult to see from this histogram if the number of zeroes is excessive. Thus, we can run a zero-inflated Poisson model and test whether it better predicts our response variable than a standard Poisson model.

The zero-inflated Poisson regression generates two separate models and then combines them. First, a logit model is generated for the

"certain zero" cases

described above, predicting whether or not a student would be in this group.

Then, a Poisson model is generated to predict the counts for those students who

are not certain zeros. Finally, the two models are combined. When

running a zero-inflated Poisson model in Stata, you must specify both models: first the

count model, then the model predicting the certain zeros. In this example,

we are predicting count with mathnce, langnce and female, and

predicting the certain zeros with mathnce and langnce.

zip daysabs mathnce langnce female, inflate(mathnce langnce)

Fitting constant-only model:

Iteration 0: log likelihood = -1494.2292

Iteration 1: log likelihood = -1388.076

Iteration 2: log likelihood = -1385.31

Iteration 3: log likelihood = -1385.2992

Iteration 4: log likelihood = -1385.2992

Fitting full model:

Iteration 0: log likelihood = -1385.2992

Iteration 1: log likelihood = -1351.1453

Iteration 2: log likelihood = -1350.7936

Iteration 3: log likelihood = -1350.7935

Zero-inflated Poisson regression Number of obs = 316

Nonzero obs = 254

Zero obs = 62

Inflation model = logit LR chi2(3) = 69.01

Log likelihood = -1350.794 Prob > chi2 = 0.0000

| Coef. Std. Err. z P>|z|

-----+-----
daysabs |
mathnce | -.0002822 .0018619 -0.15 0.880 -.0039314
.0033671
langnce | -.0095004 .0019094 -4.98 0.000 -.0132428 -
.005758
female | .2473714 .0488047 5.07 0.000 .1517161 .3430268

```
_cons | 2.297473 .0693793 33.11 0.000 2.161492  
2.433453
```

```
-----+-----
```

```
inflate |
```

```
mathnce | .0188534 .011087 1.70 0.089 -.0028766  
.0405834
```

```
langnce | .0110383 .0111014 0.99 0.320 -.0107201  
.0327966
```

```
_cons | -2.963912 .5315156 -5.58 0.000 -4.005664  
-1.922161
```

```
-----
```

Iteration History

Fitting constant-only model:a

Iteration 0: log likelihood = -1494.2292

Iteration 1: log likelihood = -1388.076

Iteration 2: log likelihood = -1385.31

Iteration 3: log likelihood = -1385.2992

Iteration 4: log likelihood = -1385.2992

Fitting full model:b

Iteration 0: log likelihood = -1385.2992

Iteration 1: log likelihood = -1351.1453

Iteration 2: log likelihood = -1350.7936

Iteration 3: log likelihood = -1350.7935

a. Fitting constant-only model - This is a listing of the log likelihoods at each iteration

for the logistic model predicting whether

or not a student is a certain zero. Remember that

logistic regression

uses maximum likelihood estimation, which is an iterative procedure. The first

iteration (called Iteration 0) is the log likelihood of the "null" or "empty"

model; that is, a model with intercept only model for the count model and intercept set to zero for inflated model or logistic model. At the next iteration (called

Iteration 1), the variables specified for predicting certain zeroes are included in the model. In this example, the predictors

for the constant-only model are mathnce and langnce.

At each

iteration, the log likelihood increases because the goal is to maximize the log

likelihood. When the

difference between successive iterations is very small, the model is said to have "converged" and the iterating stops. For more information on this process for binary outcomes, see

Regression Models for Categorical and Limited Dependent Variables by J. Scott Long (page 52-61).

b. Fitting full model - This is a listing of the log likelihoods at each iteration for the full model, combining the constant-only model with the count model.

Again, the fitting of this model is an iterative procedure. Note that the log likelihood of Iteration 0 for the full model is equal to the log likelihood at which the constant-only model had converged. This illustrates that the full model begins with the fitted constant-only model stopped and improves on it with the count model.

Model Summary

Zero-inflated Poisson regression Number of obse = 316

Nonzero obsf = 254

Zero obsg = 62

Inflation modelc = logit LR chi2(3)h = 69.01

Log likelihoodd = -1350.794 Prob > chi2i = 0.0000

c. Inflation model - This indicates that the inflated model is a logit

model, predicting a latent binary outcome: whether or not a student is a certain zero. This also informs the interpretation of the parameter estimates.

d. Log Likelihood - This is the log likelihood of the fitted full model. It

is used in the Likelihood Ratio Chi-Square test of whether all predictors'

regression coefficients in the count model are simultaneously zero.

e. Number of obs - This is the number of observations in the dataset

for which all of the response and predictor variables are non-missing.

f. Nonzero obs - This is the number of observations in the dataset for which the response variable is not equal to zero.

g. Zero obs - This is the number of observations in the dataset for which the response variable is equal to zero.

h. LR chi2(3) - This is the Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression

coefficient in the count model is not equal to zero. The number in the parentheses indicates the degrees of freedom of the Chi-Square distribution used to test the LR Chi-Square

statistic and is defined by the number of predictors in the model (3). The LR Chi-Square statistic can be calculated by

$$-2 * (L(\text{null model of full model}) - L(\text{fitted model of full model})) = -2 * ((-1385.2992) - (-1350.7935)) = 69.01.$$

i. Prob > chi2 - This is the probability of getting a LR test

statistic as extreme as, or more so, than the observed

statistic under the null hypothesis; the null hypothesis is that all of the regression coefficients for count model are simultaneously equal to zero. In other words, this is the probability of obtaining this chi-square statistic (69.01) or one more extreme if there is in fact no effect of the predictor variables in the count model. This p-value is compared to a specified alpha level, our willingness to accept a type I error, which is typically set at 0.05 or 0.01. The small p-value from the LR test, <0.00001 , would lead us to conclude that at least one of the regression coefficients in the count model is not equal to zero. The parameter of the chi-square distribution used to test the null hypothesis is defined by the degrees of freedom in the prior line, $\chi^2(3)$.

Parameter Estimates

| Coef. | Std. Err. | m | zn | P > | z | o | p

```

-----+-----
daysabsj |
mathnce | -.0002822 .0018619 -0.15 0.880 -.0039314
.0033671
langnce | -.0095004 .0019094 -4.98 0.000 -.0132428 -
.005758
female | .2473714 .0488047 5.07 0.000 .1517161 .3430268
_cons | 2.297473 .0693793 33.11 0.000 2.161492
2.433453
-----+-----
inflatek |
mathnce | .0188534 .011087 1.70 0.089 -.0028766
.0405834
langnce | .0110383 .0111014 0.99 0.320 -.0107201
.0327966
_cons | -2.963912 .5315156 -5.58 0.000 -4.005664
-1.922161
-----

```

j. daysabs - This is the response variable predicted by the full model.

k. inflate - This portion of the output refers to the

logistic model predicting whether or not a student is a certain zero.

I. Coef. - These are the regression coefficients.

The coefficients in the daysabs section of the output are interpreted as

you would interpret coefficients from a standard Poisson model: the expected number of days absent changes by $\exp(\text{Coef.})$ for each unit increase in the corresponding predictor.

Predicting Days Absent for For Students Not in the "Certain Zero" Group

mathnce - If a subject

were to increase his mathnce score by one point, the expected number of

days absent in a year would decrease by a factor of $\exp(-.0002822) = .99983821$ while holding all other variables in the model constant.

Thus, the higher a student's mathnce score, the fewer predicted days absent.

langnce - If a subject were to increase his langnce score by one point, the expected number of days absent in a year would decrease by a factor of $\exp(-.0095004) = .99046$ while holding all other variables in the model constant. Thus, the higher a student's langnce score, the fewer predicted days absent.

female - The expected number of days absent in a year for a female student is $\exp(.2473714) = 1.2764187$ times the expected number of days in a year for a male student while holding all other variables in the model constant. If female student and male student are not certain zeros and have identical mathnce and langnce scores, the expected number of days absent for the female student would be 1.2764187 times the expected number of days absent for the male student.

_cons - If all of the predictor variables in the model are evaluated at zero, the predicted number of days absent

would be calculated as $\exp(_cons) = \exp(2.297473)$. For males (the variable `female` evaluated at zero) with zero `mathnce` and `langnce` scores, the predicted number of days absent would be 9.9677412. This may seem very high, considering the mean number of days absent is less than 6, but note that evaluating `mathnce` and `langnce` at zero is out of the range of plausible scores.

Predicting Membership in the "Certain Zero" Group

mathnce - If a subject were to increase her `mathnce` score by one point, the odds that she would be in the "Certain Zero" group would increase by a factor of $\exp(0.0188534) = 1.0190322$. In other words, the higher a student's `mathnce` score, the more likely the student is a certain zero.

langnce - If a subject were to increase her `langnce` score by one point, the odds that she would be in the "Certain Zero"

group would increase by a factor of $\exp(0.0110383) = 1.0110994$. In other words, the higher a student's langnce score, the more likely the student is a certain zero.

_cons - If all of the predictor variables in the model are evaluated at zero, the logit for being in the "Certain Zero" group is $\exp(-2.963912) = 0.0516166$. This means that the predicted odds of a student with mathnce and langnce scores of zero being a certain zero are 0.0516166 (though remember that evaluating mathnce and langnce at zero is out of the range of plausible scores). Odds of 0.05 are very close to 0, meaning that it is very unlikely students (if there were any) with mathnce and langnce scores of 0 would be in "certain zero" group.

m. Std. Err. - These are the standard errors of the individual

regression coefficients for the two models. They are used in both the calculation of the z test statistic, superscript n, and the confidence interval of the regression coefficient, superscript p.

n. z - The test statistic z is the ratio of the Coef. to the Std. Err. of the respective predictor. The z value follows a standard normal distribution which is used to test against a two-sided alternative hypothesis that the Coef. is not equal to zero.

o. $P > |z|$ - This is the probability the z test statistic (or a more extreme test statistic) would be observed under the null hypothesis that a particular predictor's regression coefficient is zero, given that the rest of the predictors are in the model. For a given alpha level, $P > |z|$ determines whether or not the null hypothesis can be rejected. If $P > |z|$ is less than alpha, then the null hypothesis can be rejected and the parameter

estimate is considered statistically significant at that alpha level.

Predicting Days Absent for For Students Not in the "Certain Zero" Group

mathnce - The z test

statistic for the predictor **mathnce** is $(-0.0002822/0.0018619) = -0.15$ with an associated p-value of 0.880. If we set our alpha level to 0.05, we would fail to reject the null hypothesis and conclude that the regression coefficient for **mathnce** has not been found to be statistically different from zero given **langnce** and **female** are in the model.

langnce -The z test

statistic for the predictor **langnce** is $(-0.0095004/0.0019094) = -4.98$ with an associated p-value of < 0.001 . If we set our alpha level to 0.05, we would reject the null hypothesis and conclude that the regression coefficient for

langnce has been found to be statistically different from zero given mathnce and female are in the model.

female -

The z test

statistic for the predictor female is $(0.2473714/0.0488047) = 5.07$ with an associated p-value of < 0.001 . If we again set our alpha level to 0.05, we would reject the null hypothesis and conclude that the difference between males and females has been found to be statistically different given that mathnce and langnce are in the model.

_cons - The z test

statistic for the intercept, _cons, is $(2.297473/0.0693793) = 33.11$ with an associated p-value of < 0.001 . If we set our alpha level at 0.05, we would reject the null hypothesis and conclude that _cons has

been found to be statistically different from zero given mathnce, langnce and female are in the model and evaluated at zero.

Predicting Membership in the "Certain Zero" Group

mathnce -

The z test

statistic for the predictor mathnce is $(0.0188534/0.011087) = 1.70$ with an associated p-value of 0.089. If we set our alpha level to 0.05, we would fail to reject the null hypothesis and conclude that the regression coefficient for mathnce has not been found to be statistically different from zero given langnce is in the model.

langnce - The z test

statistic for the predictor langnce is $(0.0110383/0.0111014) = 0.99$ with an associated p-value of 0.320. If we set our alpha level to 0.05, we would fail to reject

the null hypothesis and conclude that the regression coefficient for langnce has not been found to be statistically different from zero given mathnce is in the model.

_cons -The z test

statistic for the intercept, **_cons**, is $(-2.963912/0.5315156) = -5.58$ with an associated p-value of < 0.001 . With an alpha level of 0.05, we would reject the null hypothesis and conclude that **_cons** has been found to be statistically different from zero given mathnce and langnce are in the model and evaluated at zero.

p. - This is the Confidence Interval (CI) for an individual coefficient given that the other predictors are in the model. For a given predictor with a level of 95% confidence, we'd say that we are 95% confident that the "true" coefficient lies between the lower and upper limit of

the interval. It is calculated as the Coef. $(z\alpha/2) \cdot (\text{Std.Err.})$, where $z\alpha/2$ is a critical value on the standard normal distribution.

The CI is equivalent to the z test statistic: if the CI includes zero,

we'd fail to reject the null hypothesis that a particular regression coefficient

is zero given the other predictors are in the model. An advantage of a CI is

that it is illustrative; it provides a range where the "true" parameter may lie.

For more information

In cases where there is a question as to which count model to use, the countfit command is helpful for comparing the range of count models. You can download countfit from within Stata by typing search countfit

(see

How can I used the search command to search for programs and get additional help? for more information about using search).

In times past, the Vuong test had been used to test whether a zero-inflated Poisson model or a Poisson model (without the zero-inflation) was a better fit for the data. However, this test is no longer considered valid. Please see [The Misuse of The Vuong Test For Non-Nested Models to Test for Zero-Inflation by Paul Wilson](#) for further information.

ARABPSYCHOLOGY.COM