

What is the annotated output for a Zero-Inflated Poisson Regression in SAS?

Authored by
stats writer

June 30, 2024

RECOMMENDED CITATION

stats writer (2024). *What is the annotated output for a Zero-Inflated Poisson Regression in SAS?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=160482>

The annotated output for a Zero-Inflated Poisson Regression in SAS is a detailed summary of the model's results, including the estimated coefficients, standard errors, p-values, and confidence intervals for each predictor variable. It also includes information on the model's overall fit, such as the log-likelihood statistic and the Akaike Information Criterion (AIC). Additionally, the output provides diagnostic measures, such as the Pearson and deviance goodness-of-fit tests, to assess the model's adequacy. Furthermore, the annotated output may include graphical representations, such as scatterplots and histograms, to visually interpret the relationship between the variables. This comprehensive output is essential for understanding the impact of the independent variables on the dependent variable and for evaluating the overall performance of the Zero-Inflated Poisson Regression model in predicting the outcome variable.

Zero-Inflated Poisson Regression | SAS Annotated Output

NOTE: Zero-inflated Poisson regression using proc countreg or proc genmod is only available in SAS version 9.2 or higher.

This page shows an example of zero-inflated Poisson regression analysis with footnotes explaining the output in Stata. We have data on 250 groups that went to a park for a weekend, fish.sas7bdat. Each group was questioned 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). We explore the relationship of count with with child, camper, and persons.

For a Poisson model, we assume 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 group may have spent the entire weekend fishing, but failed to catch a fish. Another group may have not done any fishing over the weekend and, not surprisingly, caught zero fish. The first group could have caught one or more fish, but did not. The second group was certain to catch zero fish. The second group will be referred

to from this point forward as a "certain zero". Thus, the number of zeroes may be inflated and the number of groups catching zero fish cannot be explained in the same manner as the groups that caught more than zero fish.

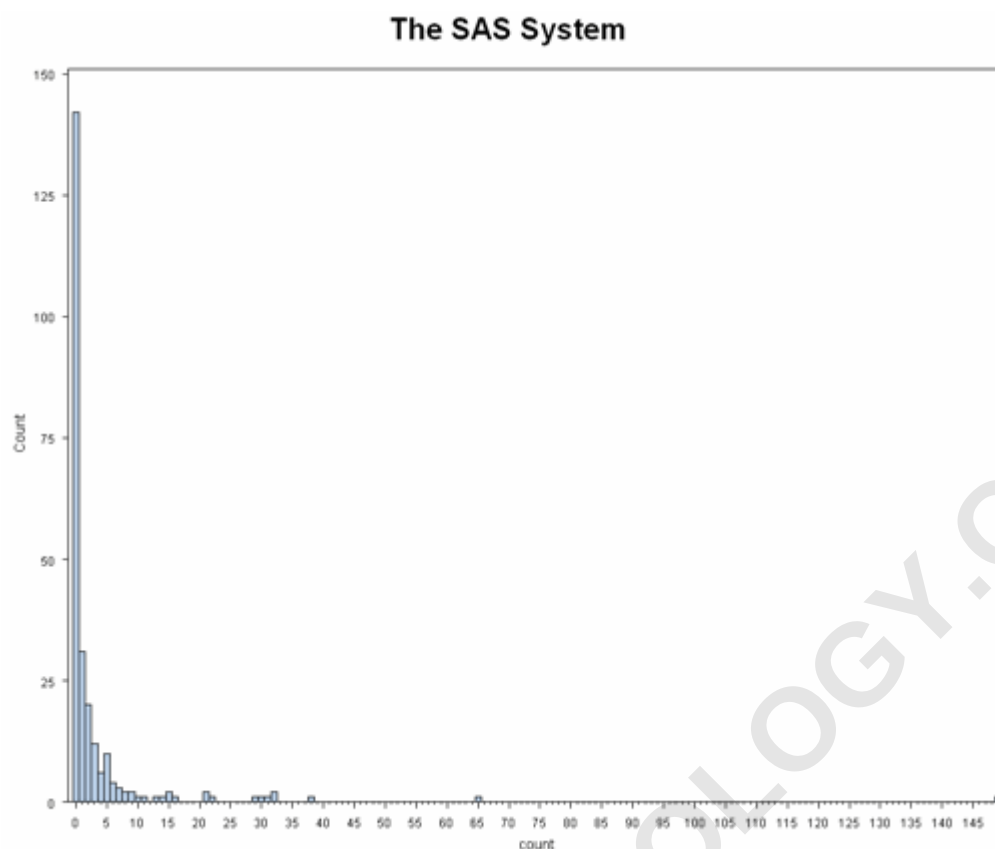
A standard Poisson model would not distinguish between the two processes causing an excessive number of zeroes, but a zero-inflated model allows for and accommodates this complication. When analyzing a dataset with an excessive number of outcome zeros and two possible processes that arrive at a zero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

```
proc means data = fish mean std min max var;
var count child persons;
run;
```

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;
```



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. There are two SAS procedures that can easily run a zero-

inflated Poisson regression:

proc genmod and proc countreg. Both will be shown on this

page, starting with proc genmod.

When running a

zero-inflated Poisson model in proc genmod, you must specify both models: first the

count model in the model line, then the model predicting the certain zeros

in the zeromodel line. In this example, we

are predicting count with child and camper and

predicting the certain zeros with persons.

```
proc genmod data = fish;
```

```
model count = child camper /dist=zip;
```

```
zeromodel persons /link = logit ;
```

```
run;
```

The GENMOD Procedure

Model Information

Data Set WORK.FISH

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 2063.2168

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

Standard Wald 95% Confidence Wald

Parameter DF Estimate Error Limits Chi-Square Pr > ChiSq

Intercept 1 1.5979 0.0855 1.4302 1.7655 348.96 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

Model Information and Goodness of Fit-Proc Genmod

Model Informationa

Data Set WORK.FISH

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

Devianceb 2063.2168

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 Likelihoodc -1031.6084

AIC (smaller is better)d 2073.2168

AICC (smaller is better) 2073.4627

BIC (smaller is better)e 2090.8241

Algorithm converged.f

a. Model Information -

This block of output includes the distribution assumed for the overall model, the function linking the model predictions to the outcome, the dependent variable, and the number of observation read and used (or not dropped due to missingness in the outcome or predictors).

b. Deviance - The deviance of the model is equal to $-2 \times (\text{full log likelihood})$. Here, we can see that $-2 \times -1031.6084 = 2063.2168$.

c. Full 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 model are simultaneously zero.

d. AIC - This is the Akaike Information Criterion. It is calculated as

AIC = -2 Log Likelihood + 2(s), where s is the total number of predictors in the model. Here, s = 5 so AIC = -2*-1031.6084 + 5*2 = 2073.2168. AIC is used for the comparison of models from different samples or non-nested models. It penalizes for the number of predictors in the model. Ultimately, the model with the smallest AIC is considered the best.

e. BIC - This is the Bayesian information criterion. Like the AIC, it is based on the log likelihood and penalizes for the number of predictors in the model. The smallest BIC is most desirable.

f. Convergence - The fitting of this model is an iterative procedure.

With each iteration of the model, the parameters are updated and the log-likelihood is calculated. This process continues until the log-likelihood

is no longer improved and has been maximized. If this point is reached,
 then the model is said to have converged. If this point is not reached,
 then SAS will note the failure to converge.

Parameter Estimates-Proc Genmod

Analysis Of Maximum Likelihood Parameter Estimates

Standard Wald 95% Confidence Wald

Parameter *g* DF Estimate *h* Error *i* Limits *j* Chi-Square *k* Pr
 > ChiSq *l*

Intercept	1	1.5979	0.0855	1.4302	1.7655	348.96	<.0001
child	1	-1.0428	0.1000	-1.2388	-0.8469	108.78	<.0001
camper	1	0.8340	0.0936	0.6505	1.0175	79.35	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates

Standard Wald 95% Confidence Wald

Parameter *term* DF Estimate *n* Error *i* Limits *j* Chi-Square *k* Pr

> ChiSql

```
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
```

g. Parameter (count model) -

These are the independent variables, as well as the intercept, predicting the count. SAS presents a Scale parameter as well, but it is set to 1

and not estimated because we are assuming our Poisson distribution has a scale parameter of 1.

h. Estimate (count model) -

These are the regression coefficients related to the count model. The coefficients for Intercept, child, and camper are interpreted as you would interpret coefficients from a standard Poisson model: the expected number of fish caught changes by a factor of $\exp(\text{Estimate})$ for each unit increase in the corresponding predictor.

Predicting Number of Fish Caught for the Non-"Certain

Zero" Groups

child - If a group were to increase its child count by one, the expected number of fish caught would decrease by a factor of $\exp(-1.0428) = 0.3524664$ while holding all other variables in the model constant. Thus, the more children in a group, the fewer caught fish are predicted.

camper - The expected number of fish caught in a weekend for a group with a camper is $\exp(0.8340) = 2.302510$ times the expected number of fish caught in a weekend for a group without a camper while holding all other variables in the model constant. Thus, if a group with a camper and a group without a camper are not certain zeros and have identical numbers of children, the expected number fish caught for the group with a camper is 2.302510 times the expected number of fish caught by the group without a camper.

Intercept - If all of the predictor variables in the model are evaluated at zero, the predicted number of fish caught would be calculated as $\exp(\text{Intercept}) = \exp(1.5979) = 4.942642$. For groups without a camper or children (the variables camper and child evaluated at zero), the predicted number of fish caught would be 4.942642.

i.

Standard Error -

These are the standard errors of the individual regression coefficients. They are used in both the calculation of the Wald Chi-Square test statistic, superscript k.

j. **Wald 95% Confidence Limits** -

This is the Wald Confidence Interval (CI) of an individual poisson regression coefficient, given the other predictors are in the model. For a given predictor variable with a level of 95% confidence, we'd say that we are 95% confident that

upon repeated trials, 95% of the CI's would include the "true" population poisson regression coefficient. It is calculated as Estimate $(z_{\alpha/2}) \times (\text{Standard Error})$, where $z_{\alpha/2}$ is a critical value on the standard normal distribution. The CI is equivalent to the Chi-Square 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 information on where the "true" parameter may lie and the precision of the point estimate.

k. Wald Chi-Square -

The Chi-Square test statistic is the squared ratio of the Estimate to the Standard Error of the respective predictor. The Chi-Square value follows a standard chi-square distribution with degrees of freedom given

by DF, which is used to test against the alternative hypothesis that the Estimate is not equal to zero.

I. $Pr > ChiSq$ -

This is the probability the Wald Chi-Square test statistic (or a more extreme test statistic) would be observed under the null hypothesis that its associated coefficient is zero, given that the rest of the predictors are in the model. For a given alpha level, $Pr > ChiSq$ determines whether or not the null hypothesis can be rejected. If $Pr > ChiSq$ is less than alpha, then the null hypothesis can be rejected and the parameter estimate is considered significant at that alpha level.

Predicting Number of Fish Caught for the Non-"Certain Zero" Groups

child - The Wald Chi-Square test statistic for the predictor child is $(-1.0428/0.1000)^2 = 108.78$ with an associated p-value of $<.0001$. If we set our alpha level to

0.05, we would reject the null hypothesis and conclude that the regression coefficient for child has been found to be statistically different from zero given the other variables are in the model.

camper -The Wald Chi-Square test statistic for the predictor camper is $(0.8340/0.0936)^2 = 79.35$ with an associated p-value of $<.0001$. If we set our alpha level to 0.05, we would reject the null hypothesis and conclude that the regression coefficient for camper has been found to be statistically different from zero given the other variables are in the model.

Intercept - The Wald Chi-Square test statistic for the intercept, Intercept, is $(1.5979/0.0855)^2 = 348.96$ 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 Intercept has been found to be statistically different from zero given the other

variables are in the model
and evaluated at zero.

Predicting Membership in the "Certain Zero" Group

persons- The Wald Chi-Square test statistic for the predictor persons is $(-0.5643/0.1630)^2 = 11.99$ with an associated p-value of 0.0005. If we again set our alpha level to 0.05, we would reject the null hypothesis and conclude that the regression coefficient for persons has been found to be statistically different from zero given the other variables are in the model.

Intercept -The Wald Chi-Square test statistic for the intercept, Intercept, is $(1.2974/0.3739)^2 = 12.04$ with an associated p-value of 0.0005. With an alpha level of 0.05, we would reject the null hypothesis and conclude that Intercept has been found to be statistically different from zero given the other variables are in the model.

m. Parameter (inflation model) -

These are the independent variables, as well as the intercept, predicting the certain zeroes.

n. Estimate (inflation model) -

These are the regression coefficients related to the inflation model. The coefficients for Intercept and persons are interpreted as you would interpret coefficients from a standard logistic model: the odds change by a factor of $\exp(\text{Estimate})$ for each unit increase in the corresponding predictor.

Predicting Membership in the "Certain Zero" Group

persons- If a group were to increase its persons value by one, the odds that it would be in the "Certain Zero" group would decrease by a factor of $\exp(-0.5643) = 0.5687581$. In other words, the more people in a group, the less likely the group is a certain zero.

Intercept - If all of the predictor variables in the model are evaluated at zero, the odds of being a "Certain Zero" is $\exp(1.2974) = 3.659769$. This means that the predicted odds of a group with zero persons is 3.659769 (though remember that evaluating persons at zero is out of the range of plausible values-every group must have at least one person).

Code and output from Proc Countreg

Proc countreg can also be used to run a zero-inflated Poisson regression in SAS. The code and output can be found below. The superscripts in the output below corresponds to the equivalent portion of the proc genmod output. Proc countreg presents t values rather than Wald Chi-square test statistics. The t values seen below can be squared to get the Wald Chi-squares seen in the proc genmod output. The p-values are equivalent.

```
proc countreg data = fish method = qn;  
model count = child camper / dist= zip;
```

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

The COUNTREG Procedure

Model Fit Summary

Dependent Variable count

Number of Observations 250

Data Set WORK.FISH

Model ZIP

ZI Link Function Logistic

Log Likelihood -1032

Maximum Absolute Gradient 4.61991E-7

Number of Iterations 15

Optimization Method Dual Quasi-Newton

AICd 2073

SBCe 2091

Algorithm converged.f

Parameter Estimates

Standard Approx

Parameter, m DF Estimate, n Error t Value Pr > |t|

Intercept 1 1.597889 0.085538 18.68