

What is the process of conducting a Poisson regression analysis and what are the key factors to consider when interpreting the Stata annotated output?

Authored by
stats writer

June 29, 2024

RECOMMENDED CITATION

stats writer (2024). *What is the process of conducting a Poisson regression analysis and what are the key factors to consider when interpreting the Stata annotated output?*.

PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=160321>

Poisson regression analysis is a statistical technique used to analyze count data, where the dependent variable is a count or rate of occurrence. The process involves fitting a Poisson regression model to the data to identify the relationship between the dependent variable and a set of independent variables. This is done using maximum likelihood estimation, which calculates the parameters that best fit the data.

The first step in conducting a Poisson regression analysis is to specify the model, which includes selecting the dependent and independent variables and determining the functional form of the relationship between them. The next step is to check for any violations of assumptions, such as overdispersion, which can affect the accuracy of the results. If overdispersion is present, a negative binomial model may be more appropriate.

Once the model is specified, it can be estimated using statistical software such as Stata. The output of the analysis includes a number of statistics and tables, which can be interpreted to understand the relationship between the variables and the significance of the results. The most important statistics to consider are the coefficient estimates, which indicate the direction and magnitude of the relationship between the variables, and the p-values, which determine the significance of the results.

In addition to the coefficient estimates and p-values, the Stata annotated output also includes information on goodness-of-fit tests, such as the deviance and Pearson chi-squared statistics. These tests assess how well the model fits the data and can help determine if any further adjustments need to be made.

When interpreting the results of a Poisson regression analysis, it is important to carefully consider the assumptions of the model, the significance of the results, and the overall fit of the model to the data. It is also important to keep in mind the limitations of the analysis and the potential for confounding variables. Overall, the process of conducting a Poisson regression analysis and interpreting the Stata annotated output requires careful attention to detail and a thorough understanding of the statistical methods involved.

Poisson Regression | Stata Annotated Output

This page shows an example of Poisson regression analysis with footnotes explaining the output. The data collected were academic information on 316

students. The response variable is days absent during the school year (daysabs), from which we explore its relationship with math standardized tests score (mathnce), language standardized tests score (langnce) and gender (female).

As assumed for a Poisson model our response variable is a count variable, and each subject has the same length of observation time. Had the observation time for subjects varied, the Poisson model would need to be adjusted to account for the varying length of observation time per subject. This point is discussed later in the page. Also, the Poisson model, as compared to other count models (i.e., negative binomial or zero-inflated models), is assumed the appropriate model. In other words, we assume that the dependent variable is not over-dispersed and does not have an excessive number of zeros. For more information on over-dispersion or excessive zeroes please review our Negative Binomial

or Zero-Inflated webpages.

Also, note that each subject in our sample was followed for one school year.

If this was not the case (i.e., some subjects were followed for half a year, some for a year and the rest for two years) and we were to neglect the exposure time, our Poisson regression estimate would be biased, since our model assumes all subjects had the same follow up time. If this was an issue, we would use the exposure option, `exposure(varname)`, where `varname` corresponds to the length of time an individual was followed to adjust the Poisson regression estimates.

The first half of this page interprets the coefficients in terms of Poisson regression coefficients and the second half interprets the coefficients in terms of incidence rate ratios.

We also run the `estat ic` command to calculate the likelihood ratio

chi-square statistic.

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

generate female = (gender == 1)

poisson daysabs mathnce langnce female

Iteration 0: log likelihood = -1547.9709

Iteration 1: log likelihood = -1547.9709

Poisson regression Number of obs = 316

LR chi2(3) = 175.27

Prob > chi2 = 0.0000

Log likelihood = -1547.9709 Pseudo R2 = 0.0536

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

-----+-----
mathnce | -.0035232 .0018213 -1.93 0.053 -.007093
.0000466

langnce | -.0121521 .0018348 -6.62 0.000 -.0157483 -
.0085559

female | .4009209 .0484122 8.28 0.000 .3060348 .495807

_cons | 2.286745 .0699539 32.69 0.000 2.149638

2.423852

estat ic

Model | Obs ll(null) ll(model) df AIC BIC

-----+-----
. | 316 -1635.608 -1547.971 4 3103.942 3118.965

Iteration Log, Model Summary and estat ic

Iteration 0: log likelihood = -1547.9709

Iteration 1: log likelihood = -1547.9709a

Poisson regression Number of obsc = 316

LR chi2(3)b = 175.27

Prob > chi2e = 0.0000

Log likelihood = -1547.9709b Pseudo R2f = 0.0536

estat ic

Model | Obs ll(null)d ll(model)d df AIC BIC

-----+-----

. | 316 -1635.608 -1547.971 4 3103.942 3118.965

a. Iteration Log - This is a listing of the log likelihood at each iteration.

Poisson regression uses maximum likelihood estimation, which is an iterative procedure to obtain parameter estimates. If you are familiar with other regression models that use maximum likelihood (e.g., logistic regression), you may notice this iteration log behaves differently. Specifically, the log likelihood at iteration 0 does not correspond to the likelihood for the empty (or null) model. This is evident when we look under ll(null) from the estat ic command, which provides the log likelihood for the empty model. The log likelihood for the fitted model is given in the last iteration of the iteration log and under ll(model) from estat ic; note that both values are equal (unlike ll(null) and the log likelihood from

iteration 0). The log likelihood for the fitted model is then used with $ll(null)$ to calculate the Likelihood ratio chi-square test statistic.

b. Log Likelihood - This is the log likelihood of the fitted model. It is used in the calculation of the Likelihood Ratio (LR) chi-square test of whether all predictor variables' regression coefficients are simultaneously zero and in tests of nested models.

c. Number of obs - This is the number of observations used in the Poisson regression. It may be less than the number of cases in the dataset if there are missing values for some variables in the model. By default, Stata does a listwise deletion of incomplete cases.

d. LR chi2(3), $ll(null)$ and $ll(model)$ from estat ic - This is the LR test statistic for the omnibus test that at least one predictor variable regression coefficient is not equal to zero in the

model. The degrees of freedom (the number in parenthesis) of the LR test statistic is defined by the number of predictor variables (3). LR $\chi^2(3)$ is calculated as $-2 \times \text{log likelihood} = -2 \times -175.274 = 175.274$.

e. Prob > χ^2 - This is the probability of getting a LR test statistic as extreme as, or more so, than the one observed under the null hypothesis; the null hypothesis is that all of the regression coefficients are simultaneously equal to zero. In other words, this is the probability of obtaining this chi-square test statistic (175.274) if there is in fact no effect of the predictor variables. 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, $p < 0.00001$, would lead us to conclude that at least one of the regression coefficients in the 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, $\chi^2(3)$.

f. Pseudo R2 - This is McFadden's pseudo R-squared. It is calculated

as $1 - \text{ll}(\text{model})/\text{ll}(\text{null}) = 0.0536$. Poisson regression does not have an equivalent to the R-squared found in OLS regression;

however, many have tried to derive an equivalent measure. There are a variety of pseudo-R-square statistics. Because this statistic does not mean what R-square means in OLS regression (the proportion of variance of the response variable explained by the predictors), we suggest interpreting this statistic with caution.

Parameter Estimates

```
-----
daysabs | Coef. Std. Err. z P>|z|
-----+-----
mathnce | -.0035232 .0018213 -1.93 0.053 -.007093
        | .0000466
langnce | -.0121521 .0018348 -6.62 0.000 -.0157483 -
        | .0085559
female | .4009209 .0484122 8.28 0.000 .3060348 .495807
_cons | 2.286745 .0699539 32.69 0.000 2.149638
        | 2.423852
```

g. daysabs - This is the response variable in the Poisson regression. Underneath daysabs are the predictor variables and the intercept (**_cons**).

h. Coef. - These are the estimated Poisson regression coefficients for the model. Recall that the dependent variable is a count variable, and Poisson regression models the log of the expected count as a function of the predictor variables. We can interpret the Poisson regression coefficient as follows: for a one unit change in the predictor variable, the difference in the logs of expected counts is expected to change by the respective regression coefficient, given the other predictor variables in the model are held constant.

mathnce - This is the Poisson regression estimate for a one unit increase in

math standardized test score, given the other variables are held constant in the model. If a student were to increase her mathnce test score by one point, the difference in the logs of expected counts would be expected to decrease by 0.0035 unit, while holding the other variables in the model constant.

langnce - This is the Poisson regression estimate for a one unit increase in language standardized test score, given the other variables are held constant in the model. If a student were to increase her langnce test score by one point, the difference in the logs of expected counts would be expected to decrease by 0.0122 unit while holding the other variables in the model constant.

female - This is the estimated Poisson regression coefficient comparing females to males, given the other variables are held constant in the model. The difference in the logs of expected counts is expected to be 0.4010 unit higher for females compared

to males, while holding the other variables constant in the model.

_cons - This is the Poisson regression estimate when all variables in the model are evaluated at zero. For males (the variable female evaluated at zero) with zero mathnce and langnce test scores, the log of the expected count for daysabs is 2.2867 units. Note that evaluating mathnce and langnce at zero is out of the range of plausible test scores. If the test scores were mean-centered, the intercept would have a natural interpretation: the log of the expected count for males with average mathnce and langnce test scores.

i. **Std. Err.** - These are the standard errors of the individual regression coefficients. They are used both in the calculation of the z test statistic, superscript j, and the confidence interval of the regression coefficient, superscript

k.

j. z and $P > |z|$ - These are the test statistic and p-value, respectively,

that the

null hypothesis that an individual predictor's regression coefficient is zero given that the rest of the predictors are in the model. 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. The probability that a particular z test statistic is as extreme as, or more

so, than what has been observed under the null hypothesis is defined by $P > |z|$.

mathnce - The z test statistic testing

the slope for mathnce on daysabs is zero,

given the other variables are in the model, is $(-0.0035/0.0018) -1.93$, with an

associated p-value of 0.053. If we set our alpha level at 0.05, we would

fail to reject the null hypothesis and conclude the Poisson regression

coefficient for mathnce is not statistically different from zero given langnce and female are in the model.

langnce - The z test statistic testing the slope for langnce on daysabs is zero, given the other variables are in the model, is $(-0.0122/0.0018)$ -6.62, with an

associated p-value of <0.0001 . If we set our alpha level at 0.05, we would

reject the null hypothesis and conclude the Poisson regression

coefficient for langnce is statistically different from zero given mathnce and female are in the model.

female - The z test statistic testing the difference between the log of expected counts between males and females on daysabs

is zero, given the other variables are in the model, is $(0.4009/0.04841)$ -8.28,

with an associated p-value of <0.0001 . If we set our alpha level at 0.05,

we would reject the null hypothesis and conclude that

the coefficient for female is statistically different from zero given mathnce and langnce are in the model.

_cons - The z test statistic testing _cons is zero, given the other variables are in the model and evaluated at zero, is (2.2867/0.0670) -32.69, with an associated p-value of <0.0001. If we set our alpha level at 0.05, we would reject the null hypothesis and conclude that _cons on daysabs has been found to be statistically different from zero given mathnce, langnce and female are in the model and evaluated at zero.

k. - This is the 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 $\text{Coef.} \pm (z_{\alpha/2}) * (\text{Std.Err.})$, 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 information on where the "true" parameter may lie and the precision of the point estimate.

Incidence Rate Ratio Interpretation

The following is the interpretation of the Poisson regression in terms of incidence rate ratios, which can be obtained by poisson, irr after running the Poisson model or by specifying the irr option when the full model is specified. This part of the interpretation applies to the output below.

Before we interpret the coefficients in terms of incidence rate ratios, we

must address how we can go from interpreting the Poisson regression coefficients as a difference between the logs of expected counts to incidence rate ratios. In the discussion above, Poisson regression coefficients were interpreted as the difference between the log of expected counts, where formally, this can be written as $\beta = \log(\mu_{x+1}) - \log(\mu_x)$, where β is the regression coefficient, μ is the expected count and the subscripts represent where the predictor variable, say x , is evaluated at x and $x+1$ (implying a one unit change in the predictor variable x). Recall that the difference of two logs is equal to the log of their quotient, $\log(\mu_{x+1}) - \log(\mu_x) = \log(\mu_{x+1} / \mu_x)$, and therefore, we could have also interpreted the parameter estimate as the log of the ratio of expected counts: This explains the "ratio" in incidence rate ratios. In addition, what we referred to as a count can also be called a rate. By

definition a rate is the number of events per time (or space), which our response variable qualifies as.

Hence, we could

also interpret the Poisson regression coefficients as the log of the rate ratio:

This explains the "rate" in incidence rate ratio. Finally, the rate at which

events occur is called the incidence rate; thus we arrive at being able to

interpret the coefficients in terms of incidence rate ratios from our interpretation above.

poisson daysabs mathnce langnce female, irr

Iteration 0: log likelihood = -1547.9709

Iteration 1: log likelihood = -1547.9709

Poisson regression Number of obs = 316

LR chi2(3) = 175.27

Prob > chi2 = 0.0000

Log likelihood = -1547.9709 Pseudo R2 = 0.0536

daysabs | IRRa Std. Err. z P>|z| b

```
-----+-----  
mathnce | .996483 .0018149 -1.93 0.053 .9929321  
1.000047  
langnce | .9879214 .0018127 -6.62 0.000 .984375  
.9914806  
female | 1.493199 .072289 8.28 0.000 1.35803 1.641823  
-----
```

a. IRR - These are the incidence rate ratios for the Poisson model shown earlier. We obtain at the incidence rate ratio by exponentiating the Poisson regression coefficient

mathnce - This is the estimated rate ratio for a one unit increase in math standardized test score, given the other variables are held constant in the model. If a student were to increase his mathnce test score by one point, his rate ratio for daysabs would be expected to decrease by a factor of 0.9965, while holding all other variables in the model constant.

langnce - This is the estimated rate ratio for a one unit increase in language standardized test score, given the other variables are held constant in the model. If a student were to increase his langnce test score by one point, his rate ratio for daysabs would be expected to decrease by a factor 0.9880, while holding all other variables in the model constant.

female - This is the estimated rate ratio comparing females to males, given the other variables are held constant in the model. Females compared to males, while holding the other variable constant in the model, are expected to have a rate 1.493 times greater for daysabs.

b. - This is the CI for the rate ratio given 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 upon repeated trials,

95% of the CI's would include the "true" population incidence rate ratio, given the other variables are in the model.

References:

- 1. Regression Analysis of Count Data, Second Edition by A. Colin Cameron and Pravin K. Trivedi**
- 2. Modeling Count Data by Joseph M. Hilbe**

ARABPSYCHOLOGY.COM