What is Poisson Regression and how is it interpreted in SAS?

Authored by

stats writer

June 29, 2024

Poisson Regression is a statistical method used to model count data, which follows a Poisson distribution. It is commonly used in situations where the dependent variable represents the number of events or occurrences within a specific time frame or space.

In SAS, Poisson Regression is interpreted by fitting a model to the data, with the number of events as the dependent variable and one or more independent variables that are thought to affect the event rate. The output of the model includes the estimated coefficients for each independent variable, which can be interpreted as the effect of that variable on the log of the event rate. A significant coefficient indicates that the variable has a significant impact on the occurrence of the event, while a non-significant coefficient suggests that there is no relationship between the variable and the event rate. Additionally, the overall goodness of fit of the model can be evaluated using measures such as deviance or chi-square. Overall, Poisson Regression in SAS provides a useful tool for analyzing and interpreting count data, allowing researchers to identify key factors influencing event rates.

Poisson Regression | SAS Annotated Output

This page shows an example of a 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), and 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
to be 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.

You can download the data set used in this example by clicking
here .

```
data preg;
set "C:temppoisson";
female = (gender = 1);
run;

proc genmod data = preg;
```

model daysabs = mathnce langnce female / link=log dist=Poisson;
run;

The GENMOD Procedure

Model Information

Data Set WORK.PREG
Distribution Poisson
Link Function Log
Dependent Variable DAYSABS number days absent

Number of Observations Read 316
Number of Observations Used 316

Criteria For Assessing Goodness Of Fit

Criterion DF Value Value/DF
Deviance 312 2234.5462 7.1620
Scaled Deviance 312 2234.5462 7.1620
Pearson Chi-Square 312 2774.4139 8.8924
Scaled Pearson X2 312 2774.4139 8.8924
Log Likelihood 1482.2670

Algorithm converged.

Analysis Of Parameter Estimates

Standard Wald 95% Confidence Chi-
Parameter DF Estimate Error Limits Square Pr > ChiSq
Intercept 1 2.2867 0.0700 2.1496 2.4239 1068.59

Model Information

Model Information

Data Seta WORK.PREG
Distributionb Poisson
Link Functionc Log
Dependent Variabled DAYSABS number days absent

Number of Observations Reade 316
Number of Observations Usede 316

a. Data Set - This is the SAS dataset on which the Poisson regression
was performed.

b. Distribution - This is the distribution of the dependent variable.
Poisson regression is a type of
generalized linear model. As such, we need to specify

the distribution of
the dependent variable, dist = Poisson, as
well as the link function, superscript
c.

c. Link Function - This is the link function used for the Poisson
regression. By default, when we specify dist = Poisson, the log link
function is assumed (and does not need to be specified); however, for
pedagogical purposes, we include link = log. When we write our model out, log(μ) = β0 + β1x1 +
... + βpxp, where μ is the count we are
modeling, log( ) defines the link function (i.e., how we transform μ to
write it as a linear combination of the predictor variables).

d. Dependent Variable - This is the dependent variable used in the
Poisson regression.

e. Number of Observations Read and Number of Observations Used

- This is the number of observations read and the number of observation used in the
Poisson
regression. The Number of Observations Used may be less than the
Number of Observations Read if there are missing values for any variables
in the equation. By default, SAS does a listwise deletion of incomplete cases.

Criteria For Assessing Goodness Of Fit

Criteria For Assessing Goodness Of Fit

Criterionf DFg Valueg Value/DFh
Deviance 312 2234.5462 7.1620
Scaled Deviance 312 2234.5462 7.1620
Pearson Chi-Square 312 2774.4139 8.8924
Scaled Pearson X2 312 2774.4139 8.8924
Log Likelihood 1482.2670

Algorithm convergedi.

Prior to discussing the Criterion, DF, Value and Value/DF, we need to discuss the logic of this section.

**Attention is
placed on Deviance and Scaled Deviance; the argument naturally
extends to Pearson Chi-Square.**

**First, note that the Deviance has an approximate chi-square distribution with
n-p
degrees of freedom, where n is the number of observations and p is the
number of predictor variables (including the intercept), and the expected value of a chi-square random variable is equal
to the degrees of freedom. Then, if our model fits the data well, the ratio of the
Deviance to DF, Value/DF, should be about one. Large ratio
values may indicate model misspecification or an over-dispersed response variable; ratios less than one may also indicate
model misspecification or an under-dispersed response variable. A consequence of
such dispersion issues is that standard errors are incorrectly estimated,**

implying an invalid chi-square test statistic, superscript p. Importantly,
however, assuming our model is correctly specified, the Poisson regression estimates remain unbiased in the presence of over-disperion or under-dispersion.
Two "fixes" are either running the same model as a negative binomial
regression, or correcting the standard errors of the estimates. The standard error
correction corresponds to the approach for the scaled criterion. A naive
explanation when the scale option is specified (scale = dscale), the
Scaled Deviance is forced to equal one. By forcing Value/DF to one (dividing
Value/DF by itself), our model becomes "optimally" dispersed; however,
what actually happens is that the standard errors are adjusted ad hoc. The standard
errors are adjusted by a factor, the square root of Value/DF.

f. Criterion - Below are various measurements used to assess the

model fit.

Deviance - This is the deviance for the model. The deviance is defined as two times the difference of the log likelihood for the maximum achievable model (i.e., each subject's response serves as a unique estimate of the Poisson parameter), and the log likelihood under the fitted model.
The difference in the Deviance and degrees of freedom of two nested models can be used in the likelihood ratio chi-square tests.

Scaled Deviance - This is the scaled deviance.
The scaled deviance is equal to the deviance since we did not specify the scale=dscale option on the model statement.

Pearson Chi-Square - This is the Pearson chi-square statistic. The Pearson chi-square is defined as the squared difference between the observed and predicted values divided by the

variance of the predicted value summed over all observations in the model.

Scaled Pearson X2 - This is the scaled Pearson chi-square statistic.
The scaled Pearson X2 is equal to the Pearson chi-square since we did not specify the scale=pscale option on the model statement.

Log Likelihood - This is the log likelihood of the model. Instead of using the deviance, we can take two times the difference between the log likelihood for nested models to perform a chi-square test.

g. DF and Value - These are the degrees of freedom DF and the respective Value for the Criterion measures. The DF is equal to n-p, where n is the number of observation used and p is the number of parameters estimated.

h. Value/DF - This is the ratio of Value to DF given

in superscript g. Refer to the discussion at the beginning of this section for an interpretation/use of this value.

i. Algorithm Converged - This is a note indicating that the algorithm for parameter estimates has converged, implying that a solution was found.

Analysis Of Parameter Estimates

Analysis Of Parameter Estimates

Standard Wald 95% Confidence Chi-
Parameterj DFk Estimatel Errorm Limitsn Squareo Pr > ChiSqo
Intercept 1 2.2867 0.0700 2.1496 2.4239 1068.59

j. Parameter - Underneath are the predictor variables and the Scale parameter.

k. DF - These are the degrees of freedom DF spent on each of the respective parameter estimates. Note that the DF for the Scale parameter is set to 0. The DF define the distribution used to test

Chi-Square, superscript o.

l. Estimate -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 linear 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.

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 estimates would

be biased, since our model
assumes all subjects had the same observation time. If this was an issue, we would
use the offset option, offset=log_timevar,
where log_timevar corresponds to the logged version of the variable specifying length of time an individual was followed to adjust the Poisson regression estimates.

Intercept - 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.

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.

Scale - This is the Scale value for the
Poisson
model. Since our model was not scaled (Scaled Deviance or Scaled
Pearson X2), the default scale for the Poisson model is set to one. This is
noted by the comment at the bottom of the output: NOTE: The scale parameter was held fixed.

m. Standard Error - These are the standard errors of the individual regression coefficients. They
are used in both the Wald 95% Confidence Limits, superscript n, and the
Chi-Square test
statistic, superscript o.

n. 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 ± (zα/2)*(Standard Error), where zα/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.

o. Chi-Square and Pr > ChiSq - These are the test statistics
and p-values, respectively, testing the null hypothesis that an individual
predictor's regression coefficient is zero, given that the rest of the
predictors are in the model. 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. The probability that a particular Chi-Square test statistic is as extreme as, or more so, than what has been observed under the null hypothesis is defined by Pr>ChiSq.