What is the SAS annotated output for Negative Binomial Regression?

Authored by

stats writer

June 30, 2024

The SAS annotated output for Negative Binomial Regression is a comprehensive report generated by the SAS software that provides a detailed summary of the results obtained from the analysis. It includes information on the model fit, parameter estimates, hypothesis testing, and goodness-of-fit measures. The output also includes annotated notes, which explain the interpretation and significance of each statistic and result, making it easier for the user to understand and interpret the findings accurately. This annotated output is a valuable resource for researchers and analysts to identify the key factors influencing the response variable and make informed decisions based on the regression analysis.

Negative Binomial Regression | SAS Annotated Output

**This page shows an example of negative binomial 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 negative binomial 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 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 negative binomial model, as compared to
other count models (i.e., Poisson or zero-inflated models), is assumed to be the
appropriate model. In other words, we assume that the dependent variable is
ill-dispersed (either under- or over- dispersed) and does not have an excessive number of zeros.

The dataset can be downloaded here.

```
options nofmterr;
data lahigh;
set "C:templahigh";
female = (gender = 1);
run;

proc genmod data = lahigh;
model daysabs = mathnce langnce female / link=log dist=negbin;
run;
```

The GENMOD Procedure

Model Information

Data Set WORK.LAHIGH
Distribution Negative Binomial
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 312 356.9348 1.1440
Scaled Deviance 312 356.9348 1.1440
Pearson Chi-Square 312 337.0888 1.0804
Scaled Pearson X2 312 337.0888 1.0804
Log Likelihood 2149.3649

Algorithm converged.

Analysis Of Parameter Estimates
Standard Wald 95% Confidence Chi-
Parameter DF Estimate Error Limits Square Pr > ChiSq
Intercept 1 2.2849 0.2099 1.8735 2.6962 118.52 <.0001
mathnce 1 -0.0016 0.0048 -0.0111 0.0079 0.11 0.7413

langnce 1 -0.0143 0.0056 -0.0253 -0.0034 6.61 0.0102
female 1 0.4312 0.1397 0.1574 0.7049 9.53 0.0020
Dispersion 1 1.2884 0.1231 1.0684 1.5537

Model Information

Model Information

Data Seta WORK.LAHIGH
Distributionb Negative Binomial
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 negative binomial
regression was performed.

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

dependent variable, dist =
negbin, as well as the link function, superscript c.

c. Link Function - This is the link function used for the negative
binomial regression. By default, when we specify dist = negbin, 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
negative binomial 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 312 356.9348 1.1440
Scaled Deviance 312 356.9348 1.1440
Pearson Chi-Square 312 337.0888 1.0804
Scaled Pearson X2 312 337.0888 1.0804
Log Likelihood 2149.3649

Algorithm converged.i

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. Also, this section is more applicable to Poisson
regression where issues of dispersion are relevant. Negative binomial
regression handles dispersion issues by modeling the dispersion parameter of the
response variable.

First, note the Deviance has an approximate chi-square distribution with n-p degrees of freedom, where n is the number of
observations, 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 regression estimates remain unbiased in the presence of
ill-dispersion. A "fix" is to adjust the standard error of the estimates. The standard error correction corresponds to the approach for the
scaled criterion. A naive explanation is that 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
specific factor, namely 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 negative binomial 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 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
equals 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 estimation has converged, implying that a solution has been found.

Analysis of Parameter Estimates

Analysis Of Parameter Estimates

Standard Wald 95% Confidence Chi-
Parameterj DFk Estimatel Errorm Limitsn Squareo Pr > ChiSqo
Intercept 1 2.2849 0.2099 1.8735 2.6962 118.52 <.0001
mathnce 1 -0.0016 0.0048 -0.0111 0.0079 0.11 0.7413
langnce 1 -0.0143 0.0056 -0.0253 -0.0034 6.61 0.0102
female 1 0.4312 0.1397 0.1574 0.7049 9.53 0.0020
Dispersion 1 1.2884 0.1231 1.0684 1.5537

j. Parameter - Underneath are the intercept, the predictor variables and the dispersion parameter.

k. DF - These are the degrees of freedom DF spent on each of the respective parameter estimates. The DF define the distribution used to test Chi-Square, superscript o.

l. Estimate -These are the estimated negative binomial regression coefficients for the model. Recall that the dependent variable is a count variable, and the regression models the log of the expected count as a linear function of the predictor variables. We can interpret each regression coefficient as follows: for a one unit change in the predictor variable, the difference in the logs of expected counts of the response variable is expected to change by the respective regression coefficient, given the other predictor variables in the

model are held constant.

Also, each subject in our sample was followed for one academic 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 neglect exposure time, the regression estimates would be biased since our model assumes
all subjects had the same observation time. If this is an issue, we could use
the offset option in the model statement, offset=logvarname, where
logvarname
corresponds to the log of a variable specifying length of time an individual was followed
to adjust the regression estimates. The log of the time-followed variable must
be calculated in an earlier data step.

Intercept - This is the negative binomial 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 2.2849 units. Note that
evaluating mathnce and langnce at zero is out of the range of
plausible test scores. Had the test scores been 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 negative binomial 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 0.0016 unit, while holding the other variables in the
model constant.

**langnce - This is the negative binomial 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 0.0143 unit while holding the other variables in the
model constant.**

**female - This is the estimated negative
binomial 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 0.4312
unit higher for females compared to males, while holding the other variables
constant in the model.**

**Dispersion - This is the estimate of the
dispersion parameter. If the dispersion parameter**

equals zero, the
model reduces to the simpler Poisson model; if dispersion is greater than zero the
response variable is over-dispersed; and if dispersion is less than zero
the response variable is under-dispersed.

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 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 CIs
would include the "true" population negative binomial 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 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 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 central 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.