What is the significance of Negative Binomial Regression, and what is the interpretation of the Stata Annotated Output for this type of regression?

Authored by

stats writer

June 30, 2024

Negative Binomial Regression is a statistical method used to analyze count data, where the outcome variable is a non-negative integer (such as number of events, accidents, or occurrences). It is an extension of the commonly used Poisson Regression, which assumes that the mean and variance of the count data are equal. However, in real-world scenarios, this assumption is often violated, leading to overdispersion in the data. Negative Binomial Regression addresses this issue by allowing for a greater flexibility in the relationship between the mean and variance of the data.

The Stata Annotated Output for Negative Binomial Regression provides valuable information for interpreting the results of this type of regression. It displays the coefficients and their corresponding standard errors, which can be used to determine the significance of each variable in the model. The output also includes the estimated dispersion parameter, which indicates the amount of overdispersion in the data. Moreover, the output provides the predicted counts and confidence intervals for each observation, allowing for a better understanding of the relationship between the predictor variables and the outcome variable. Overall, the Stata Annotated Output for Negative Binomial Regression helps researchers to accurately interpret the results and make informed decisions based on the analysis of count data.

Negative Binomial Regression | Stata 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 the
appropriate model. In other words, we assume that the dependent variable is
over-dispersed and does not have an excessive number of zeros. The first half of
this page interprets the coefficients in terms of negative binomial regression
coefficients, and the second half interprets the coefficients in terms of
incidence rate ratios.

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

generate female = (gender == 1)

nbreg daysabs mathnce langnce female

Fitting Poisson model:

Iteration 0: log likelihood = -1547.9709
Iteration 1: log likelihood = -1547.9709

Fitting constant-only model:

Iteration 0: log likelihood = -897.78991
Iteration 1: log likelihood = -891.24455
Iteration 2: log likelihood = -891.24271
Iteration 3: log likelihood = -891.24271

Fitting full model:

Iteration 0: log likelihood = -881.57337
Iteration 1: log likelihood = -880.87788
Iteration 2: log likelihood = -880.87312
Iteration 3: log likelihood = -880.87312

Negative binomial regression Number of obs = 316
LR chi2(3) = 20.74
Dispersion = mean Prob > chi2 = 0.0001

Log likelihood = -880.87312 Pseudo R2 = 0.0116

--

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

-------------+--

mathnce | -.001601 .00485 -0.33 0.741 -.0111067 .0079048

langnce | -.0143475 .0055815 -2.57 0.010 -.0252871 -.003408

female | .4311844 .1396656 3.09 0.002 .1574448 .704924

_cons | 2.284885 .2098761 10.89 0.000 1.873535 2.696234

-------------+--

/lnalpha | .2533877 .0955362 .0661402 .4406351

-------------+--

alpha | 1.288383 .1230871 1.068377 1.553694

--

Likelihood-ratio test of alpha=0: chibar2(01) = 1334.20 Prob>=chibar2 = 0.000

Iteration Loga

Fitting Poisson model:

Iteration 0: log likelihood = -1547.9709

Iteration 1: log likelihood = -1547.9709

Fitting constant-only model:

Iteration 0: log likelihood = -897.78991
Iteration 1: log likelihood = -891.24455
Iteration 2: log likelihood = -891.24271
Iteration 3: log likelihood = -891.24271

Fitting full model:

Iteration 0: log likelihood = -881.57337
Iteration 1: log likelihood = -880.87788
Iteration 2: log likelihood = -880.87312
Iteration 3: log likelihood = -880.87312

a. Iteration Log - This is the iteration log for the negative binomial model. Note there are three sections; Fitting Poisson model, Fitting constant-only model and Fitting full model. Negative binomial regression is a maximum likelihood procedure and good initial estimates are required for convergence; the first two sections provide good starting values for the negative

binomial model estimated in the third section.

The first section, Fitting Poisson model, fits a Poisson model to the data.
Estimates from the last iteration serve as the starting values for the parameter
estimates in the final section. The second section, Fitting constant-only
model, finds the maximum likelihood estimate for the mean and dispersion
parameter of the response variable. The dispersion parameter is plugged in as
the starting value for the dispersion parameter. Once starting
values are obtained, the negative binomial
model iterates until the algorithm converges. The trace option can be
specified to see how parts from the first two iteration components are used for the final
iteration component.

Model Summary

Negative binomial regression Number of obs = 316d
LR chi2(3) = 20.74e

Dispersion = meanb Prob > chi2 = 0.0001f
Log likelihood = -880.87312c Pseudo R2 = 0.0116g

b. Dispersion - This refers how the over-dispersion is modeled. The default method is mean dispersion.

c. 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.

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

e. LR chi2(3) - This is the test statistic that all regression coefficients in

the model are simultaneous equal to zero. It is calculated as negative two times the difference of the likelihood for the null model and the fitted model. The null model corresponds to the last iteration from Fitting constant-only model. Piecing parts from the iteration log together, the LR chi2(3) value is -2 = 20.74.

f. Prob > chi2 - This is the probability of getting a LR test statistic as extreme as, or more so, than the 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 statistic (20.74) 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, <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, chi2(3).

g. Pseudo R2 - This is McFadden's pseudo R-squared. It is calculated
as 1 - ll(model)/ll(null) = 0.0116. Negative binomial regression does not have an equivalent
to the R-squared
measure found in OLS regression;
however, many people have attempted to create one. Because this statistic does not mean what
R-square means in OLS regression (the proportion of variance for the response variable explained by the predictors), we suggest interpreting this statistic with caution.

Parameter Estimates

--
daysabsf| Coef.g Std. Err.h zi P>|z|i j
-------------+--
mathnce | -.001601 .00485 -0.33 0.741 -.0111067 .0079048
langnce | -.0143475 .0055815 -2.57 0.010 -.0252871 -

.003408

female | .4311844 .1396656 3.09 0.002 .1574448 .704924

_cons | 2.284885 .2098761 10.89 0.000 1.873535 2.696234

----------+----------

/lnalpha | .2533877 .0955362 .0661402 .4406351

----------+----------

alpha | 1.288383 .1230871 1.068377 1.553694

Likelihood-ratio test of alpha=0: chibar2(01) = 1334.20 Prob>=chibar2 = 0.000k

f. daysabs - This is the response variable in the negative binomial regression. Underneath are the predictor variables, the intercept and the dispersion parameter.

g. Coef. - These are the estimated negative binomial regression coefficients for the model. Recall that the dependent variable is a count variable that is either over- or under-dispersed, and the model models the log of the expected count as a function of the predictor

variables. We can interpret
the negative binomial regression coefficient as follows: for a one unit change
in the predictor variable, the log of expected counts of the
response variable changes by the respective regression coefficient, given the other
predictor variables in the model are held constant.

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.**

**_cons - 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. 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.**

**/lnalpha - This is the estimate of the log of the dispersion
parameter, alpha, given on the next line..**

**alpha - This is the estimate of the dispersion
parameter. The dispersion parameter alpha can be obtained by exponentiating /lnalpha.
If the dispersion parameter equals zero, the model reduces
to the simpler poisson model. If the dispersion parameter, alpha, is
significantly greater than zero than the data are over dispersed and are better
estimated using a negative binomial model than a poisson model.**

h. Std. Err. - These are the standard errors for the

regression
coefficients and dispersion parameter for the model. They are used
in both the calculation of the z test
statistic, superscript i, and confidence intervals, superscript
j.

i. 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|.

j. - This is the confidence interval (CI) of an individual negative binomial 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 regression coefficient. It is calculated as Coef. (zα/2)*(Std.Err.), where zα/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 the precision of the point estimate.

k. Likelihood-ratio test of alpha=0 - This is the likelihood-ratio chi-square test that the dispersion parameter alpha is

equal to zero. The test
statistic is negative two times the difference of the log-likelihood from the
poisson model and the negative binomial model, -2 = 1334.1956 with an associated p-value of <0.0001. The large test statistic
would suggest that the response variable is over-dispersed and is not
sufficiently described by the simpler poisson distribution.

Incidence Rate Ratio Interpretation

The following is the interpretation of the negative binomial regression in
terms of incidence rate ratios, which can be obtained by nbreg, irr after
running the negative binomial 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

regression coefficients as a difference between the logs of expected counts to incidence rate ratios. In the discussion above, regression coefficients were interpreted as the difference between the log of expected counts, where formally, this can be written as β= log(μx0+1) - log(μx0), where β is the regression coefficient, μ is the expected count and the subscripts represent where the predictor variable, say x, is evaluated at x0 and x0+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(μx0+1) - log(μx0) = log(μx0+1 / μx0), 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 is technically a rate. Our response variable is the number of days absent over the

school year, which by definition, is a rate. A rate is defined as the number of events per time (or space). Hence, we could also interpret the 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.

Also, 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 regression estimates 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.

nbreg, irr

Negative binomial regression Number of obs = 316
LR chi2(3) = 20.74
Dispersion = mean Prob > chi2 = 0.0001
Log likelihood = -880.87312 Pseudo R2 = 0.0116

---

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

---

mathnce | .9984003 .0048422 -0.33 0.741 .9889547 1.007936
langnce | .9857549 .005502 -2.57 0.010 .9750299 .9965978
female | 1.539079 .2149564 3.09 0.002 1.170516 2.0236933

/lnalpha | .2533877 .0955362 .0661402 .4406351

---

alpha | 1.288383 .1230871 1.068377 1.553694

Likelihood-ratio test of alpha=0: chibar2(01) = 1334.20

Prob>=chibar2 = 0.000

a. IRR - These are the incidence rate ratios for the negative binomial regression model shown earlier.

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
for daysabs would be expected to decrease by a factor of 0.9984,
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 for daysabs would be expected to
decrease by a factor 0.9857, 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 1.539 times greater for
daysabs.