How to Choose Between Negative Binomial and Poisson Regression for Your Count Data

Authored by

stats writer

December 6, 2025

Selecting the appropriate statistical model is perhaps the most critical step in quantitative analysis. For analysts working with discrete numerical outcomes--specifically count data--the choice often narrows down to two established generalized linear models: Poisson regression and Negative Binomial regression. While both are designed to analyze the frequency of events, their underlying assumptions about data variability dictate which model will yield unbiased and efficient estimates.

The primary distinguishing factor lies in how each model handles the relationship between the mean and the variance of the response variable. The choice between them hinges on detecting a phenomenon known as overdispersion. Utilizing the wrong model when overdispersion is present can lead to underestimated standard errors and, consequently, inflated test statistics, resulting in erroneous conclusions about the significance of predictors. Therefore, understanding the theoretical basis and applying robust diagnostic methods are essential steps for any rigorous statistical investigation involving count outcomes.

Understanding Count Data and Modeling Challenges

When the response variable in a statistical analysis represents discrete outcomes, such as the number of occurrences of an event, we deal with count data. Both **Negative binomial regression** and **Poisson regression** are specialized forms of regression models that are ideally suited for this type of data, which consists exclusively of non-negative integers (0, 1, 2, 3, ...).

The application of count data models is widespread across various fields, from epidemiology to marketing analytics. These models allow researchers to quantify the relationship between predictor variables and the expected count of events. However, standard linear regression is inappropriate here because count data typically violates the assumptions of normality and homoscedasticity, necessitating the use of generalized linear models (GLMs) that employ a log link function.

Here are a few common examples of response variables that represent discrete count outcomes encountered in applied statistics:

The number of students who successfully graduate from a specific educational program in a given year.
The frequency of traffic accidents occurring at a particular high-risk intersection over a defined period.
The total number of participants who manage to finish a grueling marathon race.
The volume of customer returns processed in a given month at a major retail store chain.

The Foundations of Poisson Regression

The Poisson regression model is the foundational method for analyzing count data. It is derived

from the Poisson distribution, which is characterized by the assumption of "equidispersion." This assumption dictates that the conditional variance of the response variable must be equal to its conditional mean (i.e., Variance = Mean). This mathematical constraint simplifies model estimation and interpretation, making Poisson regression highly efficient when this assumption holds true.

If the variance is roughly equal to the mean, then a Poisson regression model typically fits a dataset well. This model relies on the idea that events occur independently and at a constant rate within a fixed period. If the observed variability in the data is low, the Poisson regression model provides a highly efficient and accurate fit for the dataset, producing reliable parameter estimates and standard errors.

The primary limitation of Poisson regression becomes apparent when the variance significantly exceeds the mean, a condition known as overdispersion. If we proceed with the Poisson model despite the presence of overdispersion, the standard errors of the estimated coefficients will be biased downwards. This bias leads to overly narrow confidence intervals and inflated test statistics, increasing the risk of Type I errors (falsely rejecting the null hypothesis) and potentially identifying non-significant predictors as statistically significant.

Addressing Overdispersion with Negative Binomial Regression

When the data exhibits significant variability, meaning the variance is substantially greater than the mean, the Negative Binomial regression model offers a robust alternative. This model is essentially an extension of the Poisson framework that explicitly incorporates an additional parameter--often referred to as the dispersion parameter ($alpha$)--to account for the excess variance. This parameter models the unobserved heterogeneity or unmeasured factors that cause the variance to exceed the mean, allowing the model to adjust for overdispersion.

The Negative Binomial distribution can be conceptualized as a Poisson distribution where the mean itself is a random variable, often distributed according to a Gamma distribution. This mixture allows the conditional variance to exceed the conditional mean. By introducing the dispersion parameter, the Negative Binomial regression model provides more conservative and accurate standard errors compared to Poisson regression when overdispersion is present. This ensures that the statistical inferences drawn regarding the predictor variables are reliable, preventing the inflation of significance tests.

If, through diagnostic testing, we determine that the variance is significantly greater than the mean, then a negative binomial regression model is typically able to fit the data better, providing a more reliable representation of the underlying data generating process. Conversely, if the data is close to the equidispersion assumption, the Poisson model is preferred due to its greater parsimony and simpler structure.

Diagnostic Method 1: Analyzing Residual Plots

One intuitive technique used to assess model fit involves creating a residual plot. This plot graphs the standardized residuals against the predicted values from the fitted regression model. Residual analysis is crucial for evaluating how well the model predicts the observed outcomes and whether systematic errors remain.

We can create a residual plot of the standardized residuals vs. predicted values from a regression model. If the majority of the standardized residuals fall within the standardized range of -2 and +2, it suggests that the model is adequately capturing the variance structure, and a Poisson regression model is likely appropriate.

However, if many residuals fall outside of this range, particularly beyond $pm 3$, it is a strong visual sign of overdispersion. In such cases, a Negative Binomial regression model will likely provide a better fit by successfully modeling the excess variability and producing smaller, more tightly clustered standardized residuals.

Diagnostic Method 2: The Likelihood Ratio Test

The most formal and statistically rigorous technique for comparing these two models is the Likelihood Ratio Test (LRT). This test compares the goodness-of-fit of the simpler, restricted model (Poisson) against the more complex, unrestricted model (Negative Binomial). The Poisson model is considered "nested" within the Negative Binomial model, where the dispersion parameter is constrained to zero under the null hypothesis.

We can fit a Poisson regression model and a negative binomial regression model to the same dataset and then perform a Likelihood Ratio Test. The test evaluates whether the log-likelihood function of the Negative Binomial model is significantly greater than that of the Poisson model.

If the p-value of the test is less than some standard significance level (e.g. 0.05), we reject the null hypothesis of equidispersion and conclude that the negative binomial regression model offers a significantly better fit, justifying the use of the more complex model.

Case Study Example: Negative Binomial vs. Poisson Regression

The following example demonstrates how to use both of these diagnostic techniques in R to determine whether a Poisson regression or negative binomial regression model is better to use for a given dataset.

Suppose we want to model how many scholarship offers a high school baseball player in a given county receives based on their school division ("A", "B", or C") and their college entrance exam

score (measured from 0 to 100).

Use the following steps to determine if a negative binomial regression model or Poisson regression model offers a better fit to the data.

Step 1: Preparing the Dataset

Before fitting any models, we must first generate and inspect the simulated dataset. This dataset includes observations for 1,000 unique high school baseball players, capturing their exam scores, divisional ranking, and the count of scholarship offers received.

The following code creates the dataset we will work with, which includes data on 1,000 baseball players:

```
#make this example reproducible
set.seed(1)

#create dataset
data <- data.frame(offers = c(rep(0, 700), rep(1, 100), rep(2, 100),
rep(3, 70), rep(4, 30)),
division = sample(c('A', 'B', 'C'), 100, replace = TRUE),
exam = c(runif(700, 60, 90), runif(100, 65, 95),
runif(200, 75, 95)))

#view first six rows of dataset
head(data)

offers division exam
1 0 A 66.22635
2 0 C 66.85974
3 0 A 77.87136
4 0 B 77.24617
5 0 A 62.31193
6 0 C 61.06622
```

Step 2: Fitting Both Regression Models

The next step involves fitting both the Poisson regression model and the Negative Binomial regression model using the same set of predictor variables. The following code shows how to fit both models to the data:

```
#fit Poisson regression model
p_model <- glm(offers ~ division + exam, family = 'poisson', data = data)

#fit negative binomial regression model
library(MASS)

nb_model <- glm.nb(offers ~ division + exam, data = data)
```

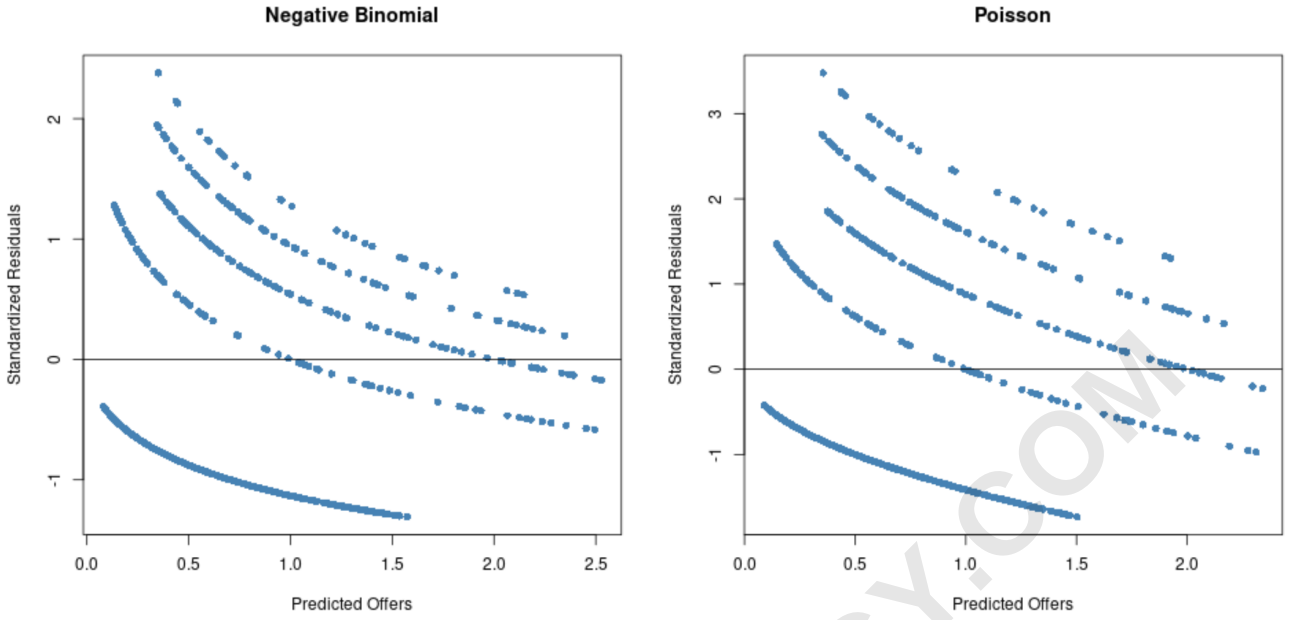
Step 3: Interpreting Residual Plots

We now generate the residual plots for both fitted models to visually assess the spread of the standardized residuals and check for signs of overdispersion.

The following code shows how to produce residual plots for both models.

```
#Residual plot for Poisson regression
p_res <- resid(p_model)
plot(fitted(p_model), p_res, col='steelblue', pch=16,
xlab='Predicted Offers', ylab='Standardized Residuals', main='Poisson')
abline(0,0)

#Residual plot for negative binomial regression
nb_res <- resid(nb_model)
plot(fitted(nb_model), nb_res, col='steelblue', pch=16,
xlab='Predicted Offers', ylab='Standardized Residuals', main='Negative Binomial')
abline(0,0)
```



From the plots we can clearly see that the residuals are much more spread out for the Poisson regression model (notice that some residuals extend beyond 3) compared to the negative binomial regression model.

This excessive spread in the Poisson plot is a strong sign of overdispersion. Since the residuals of the Negative Binomial regression model are smaller and more tightly clustered around zero, this is a visual confirmation that it is the more appropriate model.

Step 4: Executing the Likelihood Ratio Test

Lastly, we perform a formal Likelihood Ratio Test to determine if the improvement in fit provided by the Negative Binomial model is statistically significant:

pchisq(2 * (logLik(nb_model) - logLik(p_model)), df = 1, lower.tail = FALSE)

'log Lik.' 3.508072e-29 (df=5)

The p-value of the test turns out to be **3.508072e-29**. Since this value is infinitesimally small and significantly less than the standard significance level of 0.05, we decisively reject the hypothesis of equidispersion.

Thus, we would conclude that the negative binomial regression model offers a significantly better fit to the data compared to the Poisson regression model, and should be used for inference.

Conclusion: Selecting the Optimal Model

The choice between Poisson and Negative Binomial regression is a critical statistical decision based on the fundamental characteristics of the data's variance structure. While Poisson regression is the simplest and most efficient model under the restrictive assumption of equidispersion, its vulnerability to excess variability necessitates careful diagnostic checking.

For count data exhibiting significant heterogeneity, the Negative Binomial model is indispensable. By formally accounting for the excess variance through the dispersion parameter, it ensures that the standard errors are correctly estimated, thereby safeguarding the integrity of the statistical inferences. Always rely on formal diagnostic tools, such as residual plots and the Likelihood Ratio Test, to make an informed decision and guarantee the validity of your count data analysis.