Learning Poisson Regression: A Beginner’s Guide to Analyzing Count Data

Authored by

Mohammed looti

November 9, 2025

Regression is a fundamental statistical method utilized to model the relationship between a response variable and one or more predictor variables. While standard linear regression is suitable for continuous outcomes, many real-world phenomena involve outcomes measured as counts--such as the number of visitors to a website, the frequency of accidents, or the quantity of items purchased. When dealing with such count data, standard linear models often fail to capture the true underlying distribution, necessitating a specialized approach.

Poisson regression is precisely this specialized form of generalized linear model designed specifically for modeling count outcomes. It assumes that the response variable follows a Poisson distribution, which is highly appropriate for non-negative integers representing counts. Understanding when and how to apply this technique is crucial for accurate analysis in fields ranging from public health and economics to sports analytics.

Scenarios for Utilizing Poisson Regression

To illustrate the practical application of this model, consider several distinct situations where the outcome variable is inherently defined by counts. In each case, **Poisson regression** provides the ideal framework for linking potential predictors to the frequency of events.

Example 1: Academic Success. We might use Poisson regression to examine the total number of students who successfully graduate from a specific college program. The predictors could include their cumulative **GPA upon entering the program** (a continuous variable) and their **gender** (a categorical variable). Here, the number of graduates is the count response.

Example 2: Traffic Safety Analysis. Analyzing the number of **traffic accidents** occurring at a specific intersection is another prime use case. Potential categorical predictors might include **weather conditions** (e.g., "sunny," "cloudy," "rainy") and whether a **special city event** is underway ("yes" or "no"). The model helps determine which conditions significantly increase accident frequency.

Example 3: Retail Queue Management. A store manager might examine the **number of people ahead of you in line**. Predictors could involve **time of day** and **day of the week** (which can be treated as continuous or cyclical predictors), and whether a **sale is taking place** (categorical).

Example 4: Athletic Event Completion. Consider a study examining the **number of participants who finish a triathlon**. Key categorical predictor variables might be **weather conditions** and the perceived **difficulty of the course** ("easy," "moderate," "difficult").

Interpreting the Regression Coefficients

The core utility of conducting a Poisson regression is to identify which predictor variables exert a

statistically significant influence on the expected count of the response variable. However, unlike linear regression, the coefficients in a Poisson model are interpreted based on the **log of the expected count**, requiring exponentiation (e raised to the power of the coefficient) to yield meaningful multiplicative changes.

For **continuous predictor variables**, the exponentiated coefficient indicates the percentage change in the expected count associated with a one-unit increase in that predictor. For instance, if the exponentiated coefficient for GPA is 1.125, it means "each additional point increase in GPA is associated with a 12.5% increase in the number of students who graduate."

For **categorical predictor variables**, the interpretation compares the expected count of one category against a reference category (baseline). The exponentiated coefficient represents the percentage change in counts for the specific group compared to the reference group. For example, we might interpret the result as the "number of people who finish a triathlon in sunny weather" being X% higher or lower compared to the "number of people who finish a triathlon in rainy weather."

Assumptions of Poisson Regression

For the results derived from a Poisson regression to be statistically valid and reliable, specific underlying assumptions regarding the data structure and distribution must be satisfied. Violating these assumptions can lead to biased coefficients and incorrect inferences.

The Response Variable Must Be Count Data. This is the most crucial requirement. Unlike traditional linear regression, where the response is continuous, the outcome in Poisson regression must consist of **non-negative integers** (0, 1, 2, 14, etc.). Negative values are not permissible, as they have no meaning in the context of counting events.

Observations Are Independent. Every observation (or data point) in the dataset should be independent of all others. This means that one observation should not be able to provide any information about a different observation. Violation of independence often requires more complex time-series or hierarchical modeling techniques.

The Distribution Follows a Poisson Distribution. The underlying counts are assumed to be generated by a Poisson distribution. Consequently, the observed counts in the data should closely resemble the expected counts predicted by the model. Visual inspection, such as plotting the expected versus observed counts, is a simple method to check this alignment.

The Mean and Variance of the Model Are Equal (Equidispersion). A defining property of the Poisson distribution is that its mean (expected value) is equal to its variance. If this condition is met, the model exhibits equidispersion. However, this assumption is frequently violated in real-

world data, often resulting in **overdispersion** (where the variance exceeds the mean). When overdispersion occurs, alternative models like the Negative Binomial regression may be more appropriate.

Example: Implementing Poisson Regression in R

We will now walk through a practical demonstration of how to execute and interpret a Poisson regression model using the statistical programming language, R.

Background Scenario

Imagine a study aimed at predicting the number of scholarship offers received by high school baseball players within a specific geographic county. We hypothesize that the number of offers depends on two predictor variables: the player's **school division** ("A," "B," or "C") and their **college entrance exam score** (measured on a scale of 0 to 100).

The following R code generates a synthetic dataset containing information for 100 baseball players, structured to exhibit characteristics typical of count data.

#make this example reproducible
set.seed(1)

#create dataset
data <- data.frame(offers = c(rep(0, 50), rep(1, 30), rep(2, 10), rep(3, 7), rep(4, 3)),
division = sample(c("A", "B", "C"), 100, replace = TRUE),
exam = c(runif(50, 60, 80), runif(30, 65, 95), runif(20, 75, 95)))

Exploratory Data Analysis (EDA)

Before fitting the Poisson regression model, conducting an exploratory data analysis helps confirm data structure and identify initial trends. We examine the dataset dimensions, view the first few rows, and generate summary statistics using R's built-in functions, along with the **dplyr** library for grouping.

#view dimensions of dataset
dim(data)

100 3

#view first six lines of dataset
head(data)

```
# offers division exam
#1 0 A 73.09448
#2 0 B 67.06395
#3 0 B 65.40520
#4 0 C 79.85368
#5 0 A 72.66987
#6 0 C 64.26416

#view summary of each variable in dataset
summary(data)

# offers division exam
# Min. :0.00 A:27 Min. :60.26
# 1st Qu.:0.00 B:38 1st Qu.:69.86
# Median :0.50 C:35 Median :75.08
# Mean :0.83 Mean :76.43
# 3rd Qu.:1.00 3rd Qu.:82.87
# Max. :4.00 Max. :93.87

#view mean exam score by number of offers
library(dplyr)
data %>%
group_by(offers) %>%
summarise(mean_exam = mean(exam))

# A tibble: 5 x 2
# offers mean_exam
#
#1 0 70.0
#2 1 80.8
#3 2 86.8
#4 3 83.9
#5 4 87.9
```

From this initial inspection of the 100 observations, we can draw several important conclusions about the distribution of our variables:

There are 100 rows and 3 columns in the dataset.

The response variable, **offers**, ranges from a minimum of zero to a maximum of four, with a low overall mean of 0.83.

The division variable is relatively balanced: 27 players from "A," 38 from "B," and 35 from "C."

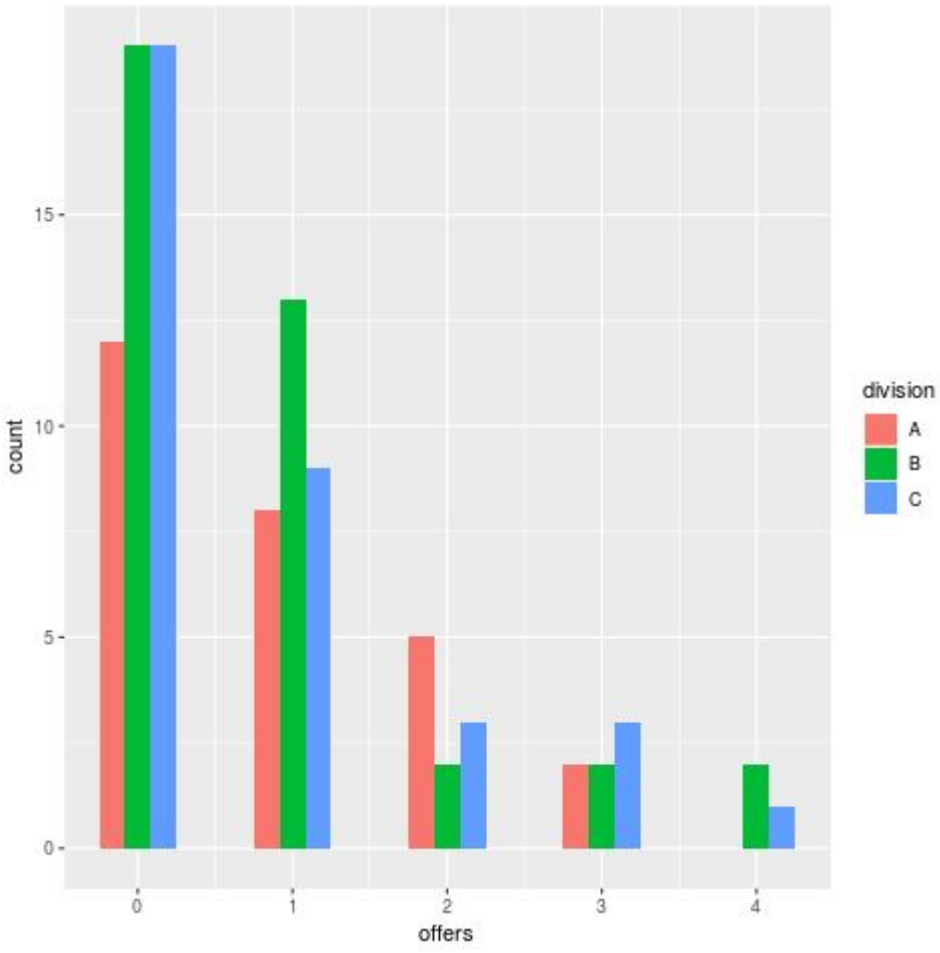
The exam scores range from approximately 60 to 94, centered around a mean of 76.43.

Crucially, the grouped summary reveals an initial positive trend: players who received more scholarship offers generally achieved higher exam scores (e.g., mean exam score for 0 offers was 70.0 vs. 87.9 for 4 offers).

Visualizing the count data through a grouped histogram confirms the distribution pattern characteristic of a Poisson distribution: a significant portion of observations fall at or near zero.

#load *ggplot2* package
library(ggplot2)

#create histogram
ggplot(data, aes(offers, fill = division)) +
geom_histogram(binwidth=.5, position="dodge")



As expected, the majority of players received either zero or one offer. This pattern, where a decent chunk of response values are zero, is typical for datasets appropriate for Poisson regression modeling.

Fitting and Interpreting the Poisson Regression Model

Next, we fit the model using R's **glm()** function (Generalized Linear Model), explicitly setting the distribution family to **family = "poisson"** to specify the correct link function and error structure. The model predicts *offers* based on *division* and *exam score*.

```
#fit the model
model <- glm(offers ~ division + exam, family = "poisson", data = data)

#view model output
summary(model)

#Call:
#glm(formula = offers ~ division + exam, family = "poisson", data = data)
#
#Deviance Residuals:
# Min 1Q Median 3Q Max
#-1.2562 -0.8467 -0.5657 0.3846 2.5033
#
#Coefficients:
# Estimate Std. Error z value Pr(>|z|)
#(Intercept) -7.90602 1.13597 -6.960 3.41e-12 ***
#divisionB 0.17566 0.27257 0.644 0.519
#divisionC -0.05251 0.27819 -0.189 0.850
#exam 0.09548 0.01322 7.221 5.15e-13 ***
#---
#Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
#(Dispersion parameter for poisson family taken to be 1)
#
# Null deviance: 138.069 on 99 degrees of freedom
#Residual deviance: 79.247 on 96 degrees of freedom
#AIC: 204.12
#
#Number of Fisher Scoring iterations: 5
```

The output provides comprehensive details, including the coefficient estimates, standard errors, z-scores, and p-values for each predictor. We can interpret these results as follows:

The **exam score** variable is highly statistically significant ($p < 0.0001$). Its coefficient is **0.09548**. Since this is the log-count increase, we exponentiate it: $e^{0.09548} \approx$ **1.10**. This indicates a 10% multiplicative increase in the expected number of scholarship offers for every one-point rise in the entrance exam score.

The coefficient for **divisionB** is **0.17566**, meaning the expected log count of offers for a player in Division B is 0.17566 higher than for a player in the reference group, Division A. Exponentiating yields $e^{0.1756} \approx$ **1.19**, suggesting players in Division B receive 19% more offers than those in Division A. However, the p-value ($p = 0.519$) shows this difference is **not statistically significant**.

The coefficient for **divisionC** is **-0.05251**, meaning the expected log count for number of offers for a player in division C is lower than for a player in division A. Exponentiating: $e^{-0.05251} \approx$ **0.94**. This suggests players in Division C receive 6% fewer offers than those in Division A. This difference is also **not statistically significant** ($p = 0.850$).

Finally, we evaluate the model fit using the deviance statistics. We focus on the residual deviance, which is **79.247** on **96** degrees of freedom. Using these values, we can perform a Chi-Square goodness-of-fit test to see if the model adequately explains the observed data variation.

```
pchisq(79.24679, 96, lower.tail = FALSE)

# 0.8922676
```

The resulting p-value for this test is **0.89**, which is substantially greater than the conventional significance level of 0.05. Therefore, we conclude that the data fits the specified Poisson regression model reasonably well, suggesting no immediate need to switch to an alternative model due to poor fit.

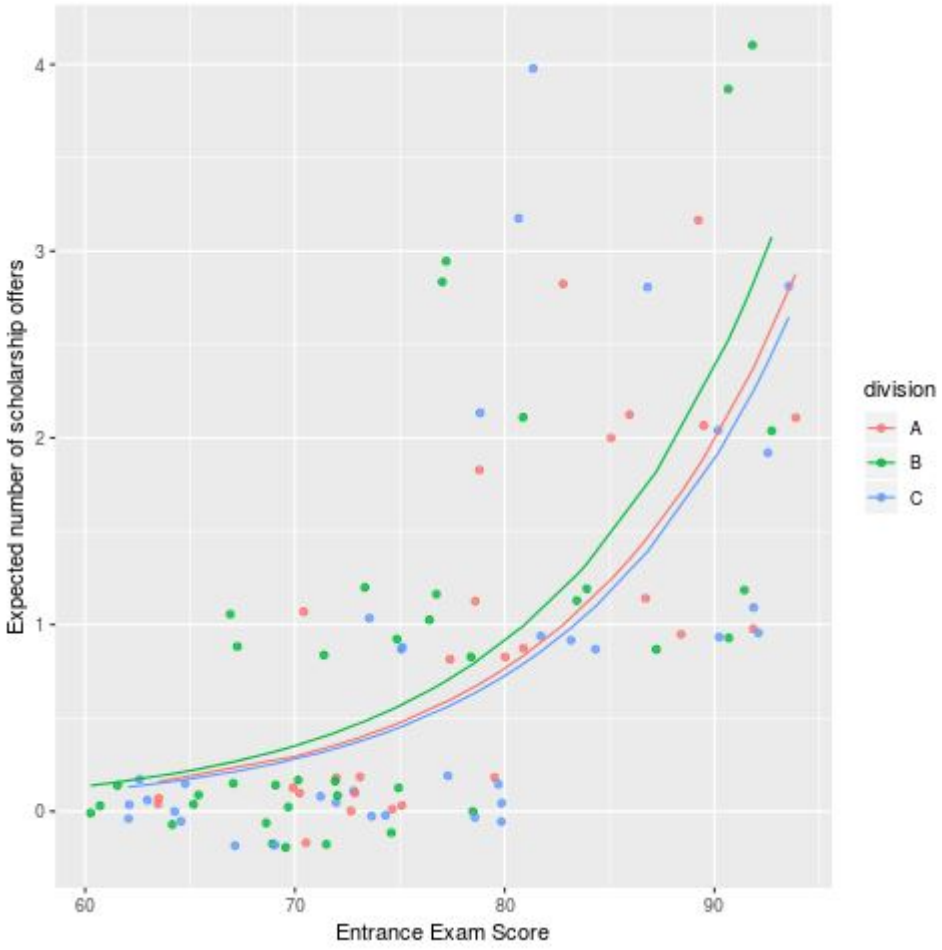
Visualizing the Predicted Outcomes

To enhance the understanding of the model's predictions, we can visualize the relationship between exam scores, divisions, and the expected number of scholarship offers. We use the **predict()** function in R to calculate the fitted values for each observation.

```
#find predicted number of offers using the fitted Poisson regression model
data$phat <- predict(model, type="response")

#create plot that shows number of offers based on division and exam score
ggplot(data, aes(x = exam, y = phat, color = division)) +
```

```
geom_point(aes(y = offers), alpha = .7, position = position_jitter(h = .2)) +
geom_line() +
labs(x = "Entrance Exam Score", y = "Expected number of scholarship offers")
```



The plot clearly illustrates the model's predictions: players achieving higher entrance exam scores are associated with a greater expected number of scholarship offers. Furthermore, the visualization confirms the non-significant trend observed in the coefficients--players from division B (the green line) are expected to receive slightly more offers across all exam scores than players from either division A or division C.

Reporting the Final Results

The final step in any statistical analysis is clearly and concisely summarizing the findings for stakeholders. When reporting the results of a Poisson regression, it is standard practice to present the exponentiated coefficients (Incidence Rate Ratios) alongside their significance levels.

A Poisson regression model was employed to predict the number of scholarship offers received by baseball players based on their school division and college entrance exam scores. The analysis demonstrated that the entrance exam score was a highly significant predictor. For each additional point scored on the entrance exam, there is a 10% increase in the expected number of offers received ($p < 0.0001$). Conversely, the differences in offers attributable to school division were

found to not be statistically significant when controlling for exam score.

Additional Resources for Further Study

For those interested in delving deeper into count data modeling, further research into Generalized Linear Models (GLMs), overdispersion correction methods (like Negative Binomial regression), and zero-inflated models is recommended.