Getting Started With Poisson GLM In R

Jiechen Li

Overview

In daily life, we often analyze data that doesn't fit into normal distributions, like the number of questions students asked per day. Generalized Linear Models (GLMs) are perfect for such scenarios. They adapt linear regression to suit various data types, using a distribution from the exponential family, a linear predictor ($\eta = X\beta$), and a link function. Poisson regression, a GLM variant, ideal for count data such as students' daily questions counts. It uses a Poisson distribution and a logarithmic link function to handle non-negative, skewed data, making it a practical tool for everyday data analysis challenges. Example research questions include:

1. Does the total time spent on online courses affect the number of student questions in forums? Are there differences between students in Course A and B?

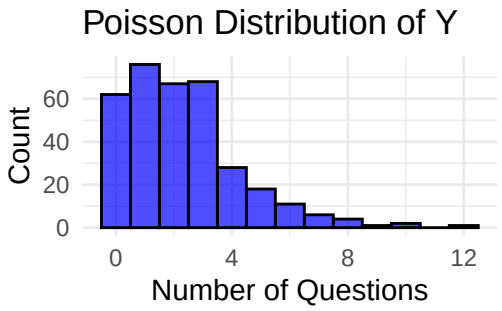
2. Does the frequency of advertising emails impact customer purchases? Is there a variance in purchase behavior based on email frequency?

Probability Distribution

Consider the scenario of analyzing how often students ask questions in an online learning forum. This is perfect for the Poisson distribution, which excels in modeling count data like the number of questions asked. It's restricted to non-negative integers (as question counts can't be negative) and hinges on the parameter λ, the average occurrence rate. A higher λ implies more frequent questioning. The Poisson distribution's Probability Mass Function (PMF) is defined as:

$$Pr[X = k] = \frac{\lambda^x e^{-\lambda}}{k!} \tag{1}$$

In this equation, Y is the random variable representing the number of occurrences of an event. k is the number of times the event is observed to occur (0, 1, 2, ...). λ is the average rate at which events occur per unit time (or space) and is the parameter of the Poisson distribution. e is the base of the natural logarithm (approximately equal to 2.71828).



From the plot, we can tell the characteristics of Y in our dataset — being discrete, non-negative, and right-skewed — strongly suggest a application of the Poisson distribution in our analysis.

The Model

In our exploration, we use the Poisson regression model, perfect for studying occurrences like how often students ask questions in an online forum. This model is neatly described by the equation:

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{2}$$

In this formula, $log(\mu)$ represents the natural logarithm of the mean of the Poisson-distributed response variable Y. The coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ link to our predictor variables $X_1, X_2, \ldots, X_p$.

An essential part of Poisson regression is the log-link function, which ensures our predictions are always positive. This function transforms the sum of our predictors into a prediction for the mean of our response variable.

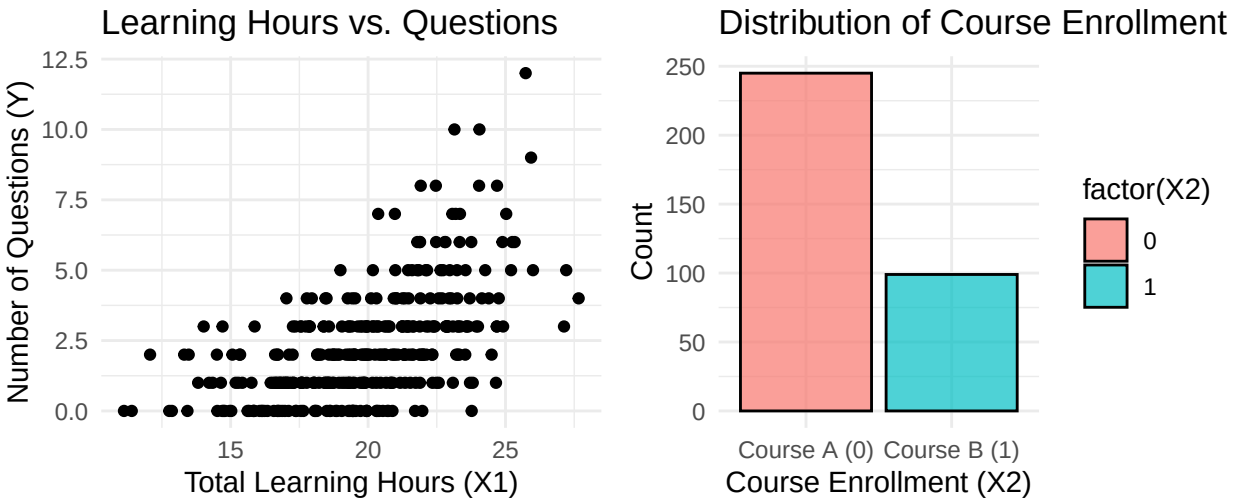
The model rests on a few key assumptions:

- The response variable Y follows a Poisson distribution.
- The observations are independent of each other.
- A linear relationship between the linear predictor and the logarithm of the mean response.
- The variance of the response variable is equal to its mean.

Data Example

Dataset Introduction

Our simulated dataset contains 344 entries and 4 columns, which represented by variable Y, with predictors X_1 and X_2. Now, let's take Research Question 1 to consider a plot to understand why Poisson regression is a suitable choice in such case. In this research scenario, we are interested in understanding how the total number of hours spent on online learning courses by students (X_1) impacts the number of questions asked by students in an online forum (Y). We are also examining whether there are differences in question behavior between students who are enrolled in Course A ($X_2 = 0$) and Course B ($X_2 = 1$)



1. **Scatter Plot**: The plot helps us visually assess whether there's any apparent pattern or trend between these two variables. For example, we can see if an increase in learning hours is associated with an increase in the number of questions.
2. **Bar Plot**: This plot helps us understand the balance or distribution of students between the two courses.

Fit the Poisson GLM

Having grasped the fundamentals and the relevance of Poisson regression for our analysis, let's now dive into modeling it to unveil the insights and results our data holds.

```
library(MASS)
poisson_model <- glm(Number_of_Questions ~ Total_Learning_Hours + Course_Enrollment,
                     family = poisson(link = "log"), data = df)
summary(poisson_model)
```

```
Call:
glm(formula = Number_of_Questions ~ Total_Learning_Hours + Course_Enrollment,
    family = poisson(link = "log"), data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5871  -0.9245  -0.0760   0.5069   3.1776

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.98497    0.27586 -10.821  < 2e-16 ***
Total_Learning_Hours  0.17639    0.01278  13.807  < 2e-16 ***
Course_Enrollment     0.54754    0.07192   7.613 2.67e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 619.27  on 343  degrees of freedom
Residual deviance: 362.70  on 341  degrees of freedom
AIC: 1148.9

Number of Fisher Scoring iterations: 5
```

Interpretation of Coefficient Estimates

Next, we will explain the coefficient estimates, revealing how each variable impacts student questions in our Poisson regression model.

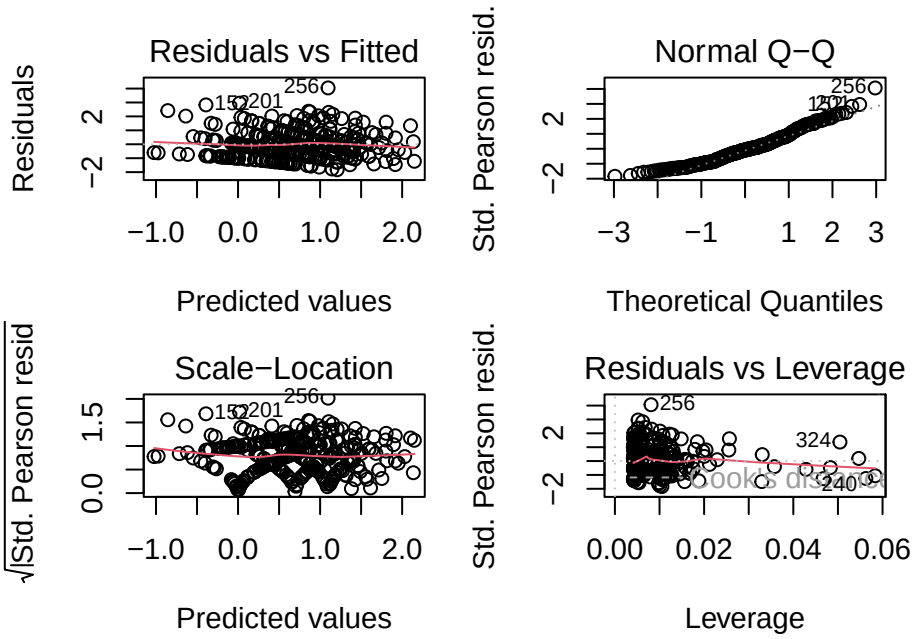
1. **Intercept** (β_0): Reflects the expected question count when no hours are spent on learning and no course enrollment, serving as a baseline for comparison.
2. **Total Learning Hours** (β_1): Each additional hour spent on online learning increases the expected log count of questions by 0.17639. This translates to about a 1.192-fold increase in the expected count of questions for each extra hour spent learning.
3. **Course Enrollment** (β_2): Enrollment in Course B (versus Course A) increases the expected log count of questions by 0.54754, suggesting that students in Course B ask about 1.731 times more questions than those in Course A, holding other variables constant.

Each coefficient's significance is affirmed by very small p-values, underscoring the robustness of these findings.

Plot Results & Assess Models

Let's now examine the plots and evaluate our model to ensure its accuracy and to identify any potential areas for improvement.

```r
par(mfrow=c(2,2),
    cex=0.9,
    mar=c(4, 4, 2, 1) + 0.1)
plot(poisson_model)
```



1. **Residuals vs Fitted Plot**: Checks for patterns suggesting non-linearity or heteroscedasticity in the model.
2. **Q-Q Plot of Standardized Deviance Residuals**: Assesses if the residuals align with a normal distribution. It is an important consideration for model fit.
3. **Scale-Location Plot**: Suggests possible heteroscedasticity, as the plot reveals a trend instead of a flat, horizontal line, indicating unequal variance across predictor ranges.
4. **Residuals vs Leverage Plot**: Identifies outliers or influential data points that might skew the model's results.

Through Cook's distance, I pinpointed three potential outliers "256", "324" and "240". After excluding them in "poisson_model2", p-values revealed no notable change over "poisson_model". Therefore, we will continue to use the "poisson_model" for our analysis.

Now, we will compute and display the exponentiated coefficient estimates and their confidence intervals from our Poisson model. This step helps in understanding the magnitude and significance of the predictors' effects.

```
cbind(exp(coef(poisson_model)), exp(confint(poisson_model)),
      summary(poisson_model)$coefficient[, 4])
```

```
                                2.5 %     97.5 %
(Intercept)           0.05054097 0.02931505 0.08644191 2.745181e-27
Total_Learning_Hours 1.19290683 1.16349686 1.22325083 2.320517e-43
Course_Enrollment     1.72899709 1.50073326 1.98970956 2.671218e-14
```

1. **Intercept (Baseline):** The baseline (when Total Learning Hours and Course Enrollment are zero) shows an expected count of questions to be about 0.05054. We are 95% confident that the true expected count is between 0.02932 and 0.08644. This is statistically significant, with a p-value much less than 0.001.

2. **Total Learning Hours:** For each additional hour spent on learning, the expected count of questions increases by a factor of approximately 1.193. In practical terms, this means that with each additional learning hour, the number of questions asked increases by about 19.3%. The 95% confidence interval for this factor ranges from 1.163 to 1.223, and this effect is statistically significant (p<.001).
3. **Course Enrollment:** Comparing students enrolled in Course B to those in Course A, the expected count of questions for Course B students is about 1.729 times higher. This suggests a significant difference in question-asking behavior between the two courses. The 95% confidence interval for this multiplier is from 1.501 to 1.990, indicating a high level of confidence in this finding, and the result is statistically significant (p<.001).

Finally, we will conduct a dispersion test using the AER package to check for overdispersion in our Poisson model. This step is crucial for validating the model's assumptions and ensuring the reliability of our findings.

```
library(AER)
test <- dispersiontest(poisson_model)

test_statistic <- test$statistic
test_p_value <- test$p.value
dispersion_estimate <- test$estimate
```

```
  Test_Statistic   P_Value Dispersion_Estimate
z     -0.8936425 0.8142434            0.936963
```

From the results of dispersion test, we can interpret as follows:

1. **Test Statistic (z-value):** The z-value of -0.8936425, being below 1, suggests mild underdispersion in our model. However, the p-value is more critical for determining statistical significance.
2. **P-value:** The high p-value of 0.8142434 indicates insufficient evidence to reject the null hypothesis, suggesting no significant overdispersion in the model.
3. **Dispersion Estimate:** At 0.936963, the dispersion estimate is close to 1, pointing to slight underdispersion but not statistically significant.
4. **Research Question Implications:** The lack of significant overdispersion or underdispersion in the Poisson model means it's suitably fitting for analyzing the impact of online learning hours and course enrollment on student questions in forums, allowing reliable interpretation of results.

Conclusion

In this tutorial, we have journeyed through the practical application and interpretation of Poisson regression for count data. We have delved into its essential aspects, like the distribution it is based on, the role of predictor variables, and the use of link functions. Most importantly, we have highlighted how crucial it is to evaluate the model's fit and check for overdispersion, ensuring our results are both accurate and trustworthy.