# Poisson Regression

Mandy Abernathy & Henri Medeiros Dos Reis

MSSC 6250

May 9, 2023

# Introduction

Poisson Regression is a popular statistical technique that is widely used in various fields such as healthcare, finance, and social sciences, where the data often exhibit count-based or discrete characteristics. This paper will provide an overview of Poisson Regression, including its history, intuition, the pros and cons of this method, the actual model, simulation, and discussion.

Poisson Regression is a type of generalized linear model (GLM) that is used to model the relationship between a dependent variable that follows a Poisson distribution in one or more independent variables. It was introduced by French mathematician Siméon Denis Poisson in the early 19th century and has since become a fundamental tool in statistical modeling. Poisson Regression is an extension of ordinary linear regression, which assumes that the dependent variable follows a normal distribution. However, Poisson Regression is specifically designed for count-based data, where the response variable represents the number of occurrences of an event within a fixed time or space interval.

The intuition behind Poisson Regression is to model the expected value of a count-based response variable as a function of one or more predictor variables. The key idea is that the expected value of the response variable is equal to the variance, which is a defining property of the Poisson distribution. Poisson Regression models the relationship between the mean of the Poisson-distributed response variable and the predictor variables using a link function, which allows for nonlinear relationships between the variables. Commonly used link functions for Poisson Regression include the log link, identity link, and square root link.

Like any statistical method, Poisson Regression has its advantages and limitations. Some of the benefits of Poisson Regression include:

- Suitable for count-based data: Poisson Regression is specifically designed for modeling count-based data, making it ideal for situations where the response variable represents

discrete events, like the number of hospital visits, the number of accidents, or the number of customer purchases.

- Handles over-dispersion: Poisson Regression can handle over-dispersion, which occurs when the variance of the response variable is greater than the mean. This makes it a versatile method for modeling data with high levels of variance.

- Interpretable results: Poisson Regression provides interpretable results in the form of rate ratios, which allow for an easier interpretation of the effect of predictor variables on the response variable.

On the other hand, Poisson Regression also has some limitations, including:

- Assumes independence: Poisson Regression assumes that the events are independent of each other, which may not be valid in some cases where events are correlated and can lead to higher bias.

- Limited to count-based data: Poisson Regression is limited to modeling count-based data and is not suitable for continuous or categorical data.

- Sensitivity to outliers: Poisson Regression can be highly sensitive to outliers, which may affect the model's performance and accuracy.

Poisson Regression is a powerful statistical method for modeling count-based data, with a long history and wide applications. It provides an intuitive approach to modeling the relationship between a count-based response variable and predictor variables, and offers several advantages such as suitability for count-based data, ability to handle over-dispersion, and interpretable results. However, it also has some limitations, including assumptions of independence, limited applicability to count-based data, and sensitivity to outliers. Understanding the intuition, ideas, and pros and cons of Poisson Regression can help researchers effectively utilize this method in their statistical machine learning studies. In the subsequent

sections of this paper, we will go further into the mathematical formulation, estimation techniques, simulation, and applications of Poisson.

# Model

When analyzing count-based data, a Poisson Regression model is a natural choice because Poisson random variables take on nonnegative integer values. We assume that the response variable is a count described by a Poisson distribution and the observations are independent. A random variable $Y$ is Poisson with mean $\lambda > 0$ if

$$\Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k = 0, 1, 2, \ldots.$$

The expected value and variance of a Poisson random variable are both equal to $\lambda$. This is a very useful property since we can construct models where both the mean and variance change.

In Poisson regression, the log link function can be applied to model the mean as a function of the $p$ predictors:

$$\lambda(X_1, \ldots, X_p) = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}.$$

Therefore, with predictors $\mathbf{X} = (X_1, \ldots, X_p)$ the Poisson regression model for observation $i$ is given by

$$\Pr(Y_i = y_i | \mathbf{X}_i) = \frac{e^{-\lambda(\mathbf{X}_i)}\lambda(\mathbf{X}_i)^{y_i}}{y_i!}, \ \lambda(\mathbf{X}_i) = e^{\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}}.$$

The coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are found using maximum liklihood estimation with the following likelihood:

$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!}.$$

There are no closed forms for the maximum likelihood estimates, so numerical methods like gradient descent must be used to determine the coefficients. With the fitted coefficients, we can use the link function to predict the expected value of the response variable.

# Simulation

Now, we will apply Poisson regression to a simulated data set and compared it's performance to linear regression and multinomial logistic regression. First, we generate one data set of `n = 50` observations from the following Poisson model,

$$\Pr(Y = y|x) = \frac{e^{-\lambda(x)}\lambda(x)^y}{y!}$$

where $\lambda(x) = \exp\{0.15 + 0.3x\}$.

Figure 1: Expected Values, 95% CI

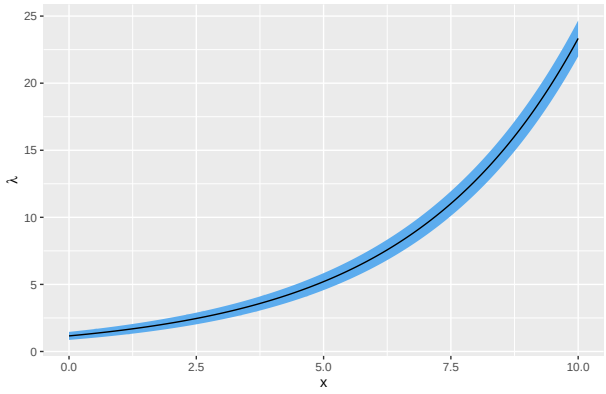
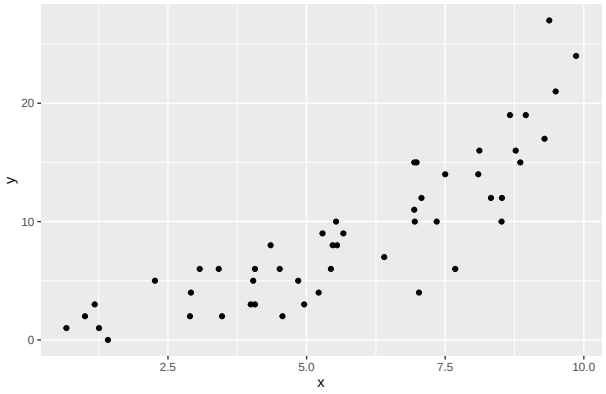



Figure 2: Simulated Poisson Data




Now, we will explore the model fits for Poisson Regression, linear regression, and multinomial logistic regression. Poisson Regression results the model

$$\lambda(x) = 0.41 + 0.27x.$$

Both coefficients are statistically significant with $p = 0.016$ for $\hat{\beta}_0$ and $p < 2 \times 10^{-16}$ for $\hat{\beta}_1$. This model suggests that the expected response when $x = 0$ is $\lambda = 0.41$ and that when $x$ increases by 1, the response will increase by a factor of $e^{0.27} = 1.32$ on average. The model has an AIC of 237.0 and a residual deviance of 45.6.

Linear regression results in the model

$$\hat{y} = -3.17 + 2.13x.$$

Both coefficients are statistically significant with $p = 0.01$ for $\hat{\beta}_0$ and $p = 2.34 \times 10^{-15}$ for $\hat{\beta}_1$, and the model has an $R^2$ value of 0.73. This model suggests that when $x$ increases by 1,

the response will increase by 2.13 on average and when $x = 0$ the expected response is -3.17. While the coefficients are significant and seem to provide a good fit, this raises concerns since the linear model will lead to negative response values for small values of $x$.

Lastly, multinomial logistic regression results in a model with the coefficients listed in Table 1.

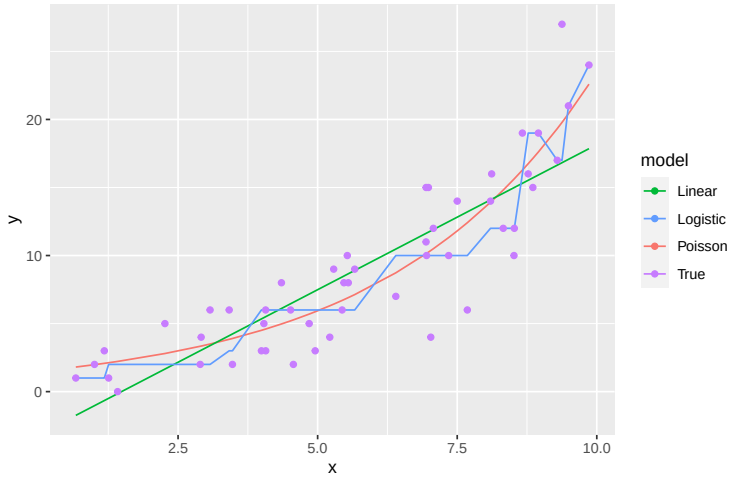
Table 1: Logisitic Model Coefficients

| class | 1 | 2 | 3 | 4 | 5 | 6 | ... | 24 | 27 |
|---|---|---|---|---|---|---|---|---|---|
| intercept | 4.3 | -3.2 | -4.4 | -9.4 | -5.2 | -7.2 | ... | -189.7 | -131.7 |
| x | -3.4 | 2.6 | 3.0 | 4.1 | 3.2 | 3.8 | ... | 24.6 | 18.5 |

Multinomial logistic regression addresses some of the problems with linear regression because it provides non-negative, integer-valued responses. However, because the response values are counts rather than classes, we see that there are also problems with a multinomial logistic regression model. First, the interpretability of the coefficients does not translate well to counts since ours baseline class would simply be a response value of 0 rather than a more concrete class or category. The magnitudes of the coefficients and their standard errors also become increasingly large for larger response values. This model has an AIC of 265.4 and a residual deviance of 185.4, so it does not fit the data as well as the Poisson Regression model.

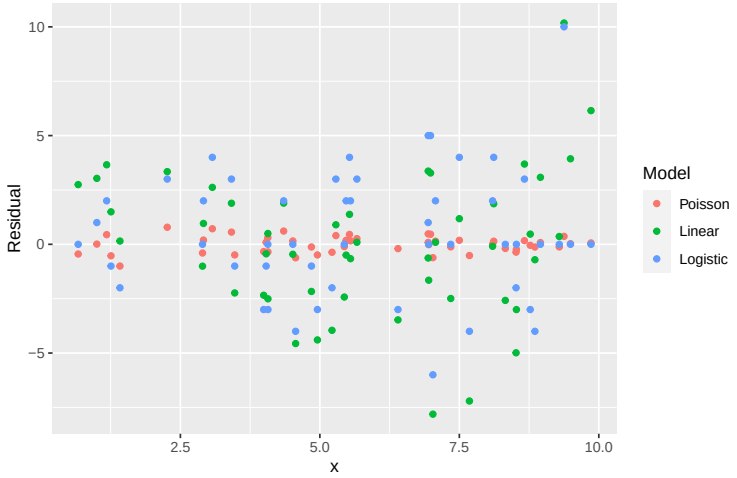
The simulated data set along with the three model fits are shown in the Figure 3.

Figure 3: Simulated Data Set and Fitted Models



We can see that all three regression models capture the general behavior of the data, however, the Poisson regression model provides the best visual fit. The linear model leads to some negative predictions and does not account for the increase in variance. The multinomial logistic regression model is not smooth and likely overfit to the training data, so it would not perform well on unseen data. If the logistic or linear models were used to do prediction for $x$ greater than 10, they would most likely perform very poorly. Also, the residuals of the Poisson model do not increase, while the residuals of the other models increase significantly as $x$ increases (Figure 4).

Figure 4: Model Residuals




Finally, we see how these models perform on a larger data set with several predictors by fitting them to a training set and applying them to unseen data. We generate observations from a Poisson model with mean given by

$$\lambda(X_1, X_2, X_3, X_4) = \exp\left\{\log(2) + 0.5X_1 + 0.2X_2 + X_3 - 0.3X_4 + \epsilon\right\}.$$

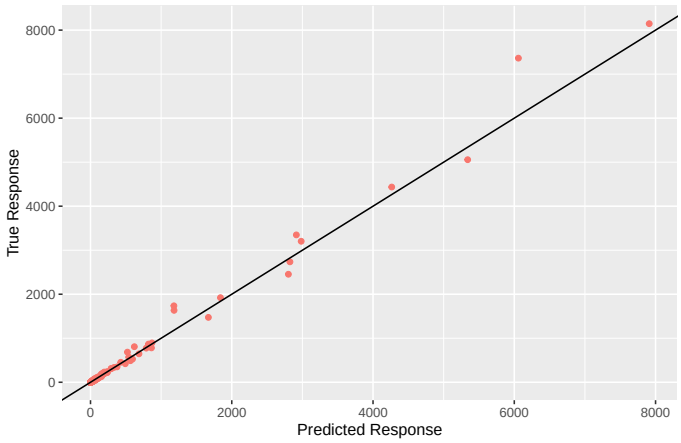
The coefficients of the fitted Poisson model are very close to the true values. Taking a closer look to the Poisson model, we have the following fitted mean:

$$\lambda(X_1, X_2, X_3, X_4) = \exp\left\{0.7337 + 0.5962X_1 + 0.1682X_2 + 0.9549X_3 - 0.3072X_4\right\}.$$

All coefficients are significant, and this model captures the original equation almost perfectly. Now, we test the model and compare the results with linear regression and multinomial logistic regression.

As expected, a Poisson model performs very well on the testing data with almost all predictions matching the true response (Figure 5).

Figure 5: Poisson Predictions




The linear model does not predict very well on the test data, especially for small and large response values (Figure 6). A significant number of the predicted values are negative, and the model does not capture the increasing variance of the data. The multinomial logisitic regression model performs fairly well for smaller response values, but has significant prediction errors for some data points (Figure 7). Since there are fewer data points for larger response values, the logistic model is likely overfit to the training data for these observations.

Figure 6: Linear Model Predictions

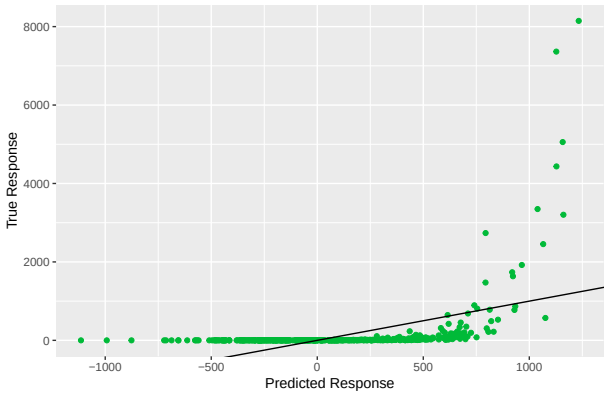
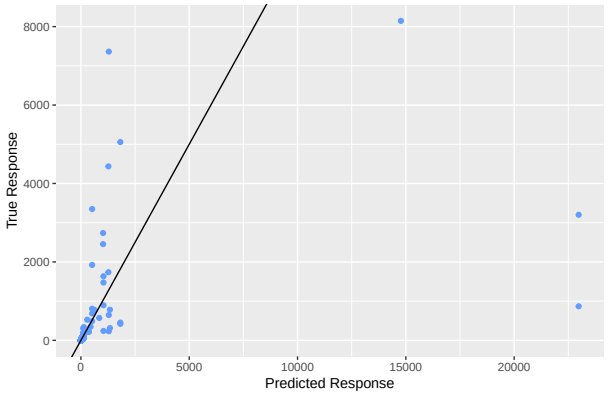



Figure 7: Logistic Model Predictions



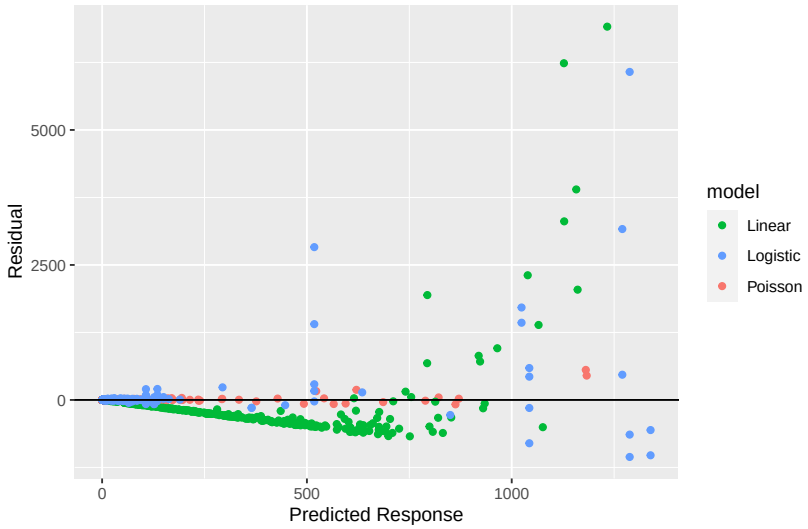

The testing mean squared errors (MSE) for these three models are given in Table 2. As expected, the Poisson model performs well, while the test MSE for the linear model and logistic models are very large because of the data points with significant prediction errors.

Table 2: Test MSE

| Poisson | Linear | Logistic |
|---|---|---|
| 5788 | 360229 | 2009551 |

By analyzing the residuals, we can better understand the conditions where these model fail to perform well and the conditions where they may perform better. Figure 8 shows how the residuals of each model display different patterns for non-negative fitted values that are less than 1500 (some outliers are not plotted to reduce the scale of the plot).

Figure 8: Model Residuals




The Poisson model accurately captures the significant increase in variance, so the Poisson residuals are all relatively small. If there is prior knowledge that a data set or its variance is related to a Poisson process or something similar, a Poisson Regression model is the clear choice.

The residuals of the linear model show that there is significant nonlinearity and the assumption of constant variance is violated. The linear model errors are not constant and increasing, and switch from entirely negative to positive. In some cases, a log transformation may address the increase in variance, but the increase in variance of this sample is too drastic and a transformed linear model also does not perform well. If there was a smaller number of predictors, smaller range of response values, or smaller change in variance, a linear model

could have comparable performance to a Poisson model. However, Poisson Regression is an especially useful alternative when the variance of a count-based response variable is far from constant.

The residuals of the multinomial logistic model are well-behaved for smaller predicted values, but drastically increase for larger predicted values. The logistic model could lead to more accurate predictions if the range in response values was smaller. Multinomial logistic regression performs better than the linear model since the residuals have constant variance for response values less than 500, and the predicted values are non-negative integers. If choosing between these two models, logistic regression may be the appropriate choice if it is important to avoid negative or continuous-valued predictions. However, unlike Poisson Regression and linear regression, it is not straightforward to interpret a multinomial logistic model for count-based data. If interpretation is necessary, another model would be preferred.

# Discussion

In this paper, we provided an overview of Poisson Regression, a statistical method widely used to model count-based data. We discussed the history and intuition behind Poisson Regression, its pros and cons, and the model formulation. We also presented simulation results and discussed some of the applications of Poisson Regression.

One of the key advantages of Poisson Regression is its suitability for count-based data. It is specifically designed to model the relationship between the count-based response variable and one or more predictor variables, which makes it an ideal method for many real-world applications such as healthcare, finance, and social sciences. Additionally, Poisson Regression can handle over-dispersion, making it a versatile method for modeling data with high levels of variance.

One of the limitations of Poisson Regression is its assumption of independence among events. This may not be valid in some cases where events are correlated, which may lead to higher bias. Another limitation is its sensitivity to outliers, which may affect the model's

performance and accuracy. In addition, Poisson Regression is limited to modeling count-based data and is not suitable for continuous or categorical data.

Despite its limitations, Poisson Regression has a wide range of applications. For example, it can be used to model the number of hospital visits by patients with a certain disease, the number of accidents on a highway segment, or the number of customer purchases at a retail store. Poisson Regression has also been used in social sciences to model the number of crime incidents in a certain area or the number of votes received by a political candidate.

In terms of estimation techniques, Poisson Regression relies on maximum likelihood estimation to estimate the model parameters. The maximum likelihood estimation method provides a way to estimate the model parameters that maximize the likelihood of the observed data.

An extension of Poisson Regression is Regularized Poisson Regression. It reduces overfitting of the model and seeks to minimize both the prediction error and the coefficients in the mean function $\lambda(x)$, similar to ridge regression. As with other models, cross-validation can then be used to determine the best value for the shrinkage parameter in the regularization term. Negative binomial regression is another method that can be used to model discrete events or counts. It is a useful alternative to Poisson Regression in cases where the variance is not equal to the mean or there are many responses equal to zero. It is

In conclusion, Poisson Regression is a powerful statistical method for modeling count-based data. It offers several advantages such as suitability for count-based data, ability to handle over-dispersion, and interpretable results. However, it also has limitations such as assumptions of independence, limited applicability to count-based data, and sensitivity to outliers. This paper and simulation highlight the utility of Poisson Regression and relevance to many areas of research.

# Bibliography

1. Pardoe, Iain. (Accessed 2023, May 4). "12.3 - Poisson Regression." 12.3 - Poisson Regression — STAT 462, https://online.stat.psu.edu/stat462/node/209/.

2. Roback, Paul, and Julie Legler. (Accessed 2023, May 4). "Beyond Multiple Linear Regression." Chapter 4 Poisson Regression, 26 Jan. 2021, https://bookdown.org/roback/bookdown-BeyondMLR/ch-poissonreg.html.

3. OARC Stats (Accessed 2023, May 4)., https://stats.oarc.ucla.edu/r/dae/poisson-regression/.