GLM for Binary, Proportions, and Count Data

Søren Lund Pedersen

1 Introduction

In this article, I would like to discuss different types of modeling using Generalized Linear Models (GLMs) for count, binary, and multimodal data. While count data has been discussed in previous articles, here we summarize its key ideas: count data typically involves tallying the number of events occurring in a fixed time interval, such as the number of phone calls made to a call center in an hour. Binary data refers to outcomes that take on one of two possible values (e.g., 0 or 1), and multimodal data involves multiple categories that an outcome can fall into, such as predicting which product a person might buy among several options (A, B, C, or D).

2 Latent Variable Model for Binary Data

In medical studies, researchers often analyze responses to a certain drug, such as determining the lethal dose that kills 50% of a test population. A latent variable model provides a useful framework for binary outcomes, where an unobserved variable Z is linked to the observed binary response Y as follows:

- If $Z > 0$, then $Y = 1$ (response to the treatment).
- If $Z \leq 0$, then $Y = 0$ (no response).

For a binary outcome, we model the probability of "success" (e.g., death) as a function of predictor variables such as dose. Let p be the probability that $Y = 1$. To build a latent variable model for binary data, we typically use a logit or probit link function, which constrains the predicted probability to lie between 0 and 1.

2.1 Logit and Probit Links

Logit Link: The logit link function transforms a probability p to the entire real line:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

The inverse logit function (also known as the logistic function) maps a real number x back to the interval $(0, 1)$:

$$\text{logit}^{-1}(x) = \frac{1}{1+e^{-x}}.$$

Probit Link: The probit link uses the cumulative distribution function (CDF) of the standard normal distribution Φ to transform probabilities:

$$\text{probit}(p) = \Phi^{-1}(p).$$

The inverse of the probit function Φ^{-1} maps a probability to the corresponding quantile of the standard normal distribution. The choice between logit and probit often depends on the desired properties of the link function, such as the steepness of the curve and tail behavior.

3 Odds Ratio

To understand the odds ratio (OR), we first define the odds of an event:

$$\text{Odds} = \frac{p}{1-p}.$$

The odds ratio compares the odds of an event occurring under two different scenarios. For instance, if an unfair coin has a probability $p = 0.6$ of landing heads, the odds of heads are

$$\frac{0.6}{0.4} = 1.5.$$

This means it is 1.5 times more likely to land heads than tails.

The odds ratio provides a measure of how changes in predictors affect the odds of an event:

- $OR = 1$: No change in odds.
- $OR > 1$: Increased odds of the outcome.
- $OR < 1$: Decreased odds of the outcome.

4 Interpretation of Logistic Regression

Logistic regression is a specific type of GLM for modeling binary outcomes using a logit or probit link. Once a logistic regression model is estimated, the coefficients β can be interpreted in terms of odds ratios. For example, suppose β_1 corresponds to a binary predictor indicating home advantage in sports (1 if playing at home, 0 otherwise). If $\beta_1 = 0.69$, the odds ratio comparing home to away games is:

$$\frac{\text{Odds (home)}}{\text{Odds (away)}} = e^{0.69} \approx 2.$$

This means the odds of winning are twice as high when playing at home compared to playing away.

5 Modeling Rates

In many applications, we encounter count data with varying exposure times, areas, or population sizes. Modeling rates allows us to account for this varying exposure. A rate is defined as:

$$r_i = \frac{y_i}{t_i},$$

where

- y_i is the observed count of events.
- t_i is the level of exposure (e.g., time, population).

For Poisson-distributed data, it is assumed that the mean λ_i scales linearly with exposure t_i.

5.1 GLM for Rates

To incorporate rates into a GLM framework for count data, we modify the linear predictor by including an offset for the log of exposure:

$$\log(\lambda_i) = \log(t_i) + \mathbf{X}_i^T \boldsymbol{\beta}.$$

Taking the exponential of both sides, we obtain:

$$\lambda_i = t_i \exp(\mathbf{X}_i^T \boldsymbol{\beta}).$$

This implies that the expected count $\mu_i = \mathbb{E}[Y_i]$ is proportional to the exposure t_i, and dividing by t_i gives us the rate:

$$\log\left(\frac{\mu_i}{t_i}\right) = \mathbf{X}_i^T \boldsymbol{\beta}.$$

Here, $\mathbf{X}_i^T \boldsymbol{\beta}$ models the log-rate, adjusting for varying exposure levels.

6 Overdispersion

A key assumption of the Poisson distribution is that the variance equals the mean:

$$\mathrm{Var}(Y_i) = \mathbb{E}[Y_i].$$

However, in practice, data often exhibit overdispersion (variance greater than the mean) or underdispersion (variance less than the mean). Overdispersion can lead to underestimated standard errors and overly optimistic significance tests.

A common remedy for overdispersion is to use the negative binomial model, which introduces an additional dispersion parameter. The negative binomial allows the variance to exceed the mean, providing a more flexible fit for overdispersed count data.

7 Zero-Inflated Models

The Poisson model assumes that the probability of zero events is dictated solely by the Poisson distribution. However, in many real-world scenarios, there are more zeros than the Poisson model would predict. For example, modeling the number of car crashes at an intersection per hour may result in many hours with zero crashes.

A zero-inflated model addresses this by assuming a two-part process for each observation y_i:

1. A binary process that determines whether the observation is an *excess zero* (always zero) or comes from a standard count distribution.
2. A count process (e.g., Poisson or negative binomial) that models the distribution of non-zero counts.

Formally, the probability mass function of a zero-inflated model can be expressed as:

$$P(Y_i = y) = \begin{cases} \pi_i + (1 - \pi_i) \cdot f_0(0 \mid \mu_i) & \text{if } y = 0, \\ (1 - \pi_i) \cdot f_0(y \mid \mu_i) & \text{if } y > 0, \end{cases}$$

where

- π_i is the probability that the observation belongs to the "always zero" group.
- $f_0(y \mid \mu_i)$ is the probability mass function of the underlying count distribution with mean μ_i.

Zero-inflated models provide a better fit when data have an excess of zero counts.

8 Multinomial Data

For outcomes that can take on more than two categories, we model *multinomial* data. In multinomial settings, each observation falls into one of K categories. A classic example is rolling a die, where the outcome can be one of six faces.

The probability of observing a particular set of counts $(y_1, y_2, \ldots, y_K)$ in n trials with category probabilities $(\pi_1, \pi_2, \ldots, \pi_K)$ is given by the multinomial distribution:

$$P(\mathbf{y} \mid n, \boldsymbol{\pi}) = \frac{n!}{y_1! y_2! \cdots y_K!} \prod_{k=1}^{K} \pi_k^{y_k},$$

subject to $\sum_{k=1}^{K} y_k = n$ and $\sum_{k=1}^{K} \pi_k = 1$.

Multinomial logistic regression extends logistic regression to handle multiple outcome categories, modeling the log-odds of each category relative to a reference category as a linear function of predictors.