Multinomial Regression Tutorial

Barbara Flores

Table of contents

1 Overview

1.1 Generalized Linear Model

Generalized Linear Models (GLMs) are a class of statistical models in which the relationship between a dependent variable and one or more independent variables follows a specific probability distribution, such as the Normal, Binomial, Poisson, Gamma, and Multinomial distributions. The general formula of a Generalized Linear Model (GLM) is expressed as follows:

$$g(\mu) = X\beta$$

where

- $g(\mu)$: The link function connecting the mean μ of the response variable distribution with the linear combination of predictor variables $X\beta$
- X: The design matrix containing the values of the predictor variables.
- β: The vector of coefficients associated with each predictor variable.

The main purpose of GLMs is to provide a statistical framework for modeling the relationship between a dependent variable and one or more independent variables, allowing flexibility in the distribution of the dependent variable. The choice of probability distribution and the link function allows the adaptation of the model to different types of data and relationships between variables.

1.1.1 Link Function

The purpose of a link function in GLMs is to connect the distribution of the outcome variable to the linear predictors, defining the relationship between the linear predictor and the expected value (mean) of the response variable. It indicates how the expected value of the response variable relates to the linear combination of explanatory variables. In logistic regression, for example, the link function is the logit function, which transforms the probability of the outcome variable into a log-odds format. This allows us to model a binary response variable using a linear combination of predictors. Within a GLM, the link function transforms the linear combination of predictor variables into a format that is appropriate for modeling the mean of the response variable. This conversion is crucial, ensuring that the forecasted values conform to the particular range determined by the characteristics of the response distribution.

1.1.2 GLM Assumptions

GLMs consider the following assumptions:

- The data are independently distributed.
- The dependent variable Y follows a specific distribution.
- There is a linear relationship between the transformed expected response, in terms of the link function, and the explanatory variables.

In particular, in this tutorial, we will address a type of GLM, Multinomial Regression, which is detailed in the following sections

1.2 Multinomial Regression

Multinomial regression is a form of GLM employed when the outcome variable encompasses multiple categories. In this regression, the multinomial distribution is utilized to model the outcome variable.

1.2.1 Examples of research questions

Some examples of research questions that could be answered with Multinomial regression are:

- **Inference problem:** Within the context of a university student having the choice to enroll in a major among the academic departments of 'Social Sciences and Humanities,' 'Health Sciences,' 'Natural Sciences and Mathematics,' and 'Arts,' what sociodemographic variables significantly influence the selection of academic departments?

- **Prediction problem:** Based on a new customer's viewing history on a movie streaming service, which includes variables such as the number and type of movies watched, total duration in minutes, etc., what option is the customer most likely to choose at the conclusion of the 30-day trial period: cancel their subscription, subscribe to the monthly plan, or subscribe to the annual plan?

2 Multinomial Distribution

The multinomial distribution models the probability of observing a specific vector of event counts across k different categories, after conducting n independent trials. The Probability Mass Function (PMF) describes the joint probability of observing x_1 events in category 1, x_2 events in category 2, and so forth. The PMF is expressed by the following formula:

Probability Mass Function

$$Pr(X_1 = x_1, ..., X_k = x_k) = \frac{n!}{x_1!...x_k!} p_1^{x_1} ... p_k^{x_k}$$

Now, let's delve into the support of the multinomial distribution, which outlines the valid range of counts for each category, providing insights into the possible outcomes of the experiment:

Support

$$x_i \, \epsilon \{0, .., n\}, i \, \epsilon \{1, .., k\}, with \sum_i x_i = n$$

Each $x_i$represents the count of occurrences for event i, and it can range from 0 to the total number of trials n. The constraint $\sum_i x_i = n$ ensures that the total count of occurrences across all categories equals the total number of trials, emphasizing the nature of the distribution.

Finally, the parameters of this probability distribution model, adhere to the following constraints:

Parameters

$$n > 0 \text{ number of trials, } k > 0 \text{ number of mutually exclusive events,}$$
$$p_1, ...p_k \text{ event probabilities } (\sum_i p_i = 1)$$

3 Model

3.1 General Form

The general equation for Multinomial Regression can be expressed as follows:

$$log(\frac{\pi_{ij}}{\pi_{i1}}) = \beta_{0j} + \beta_{1j} x_{i1} + ... + \beta_{pj} x_{ip}$$

- π_{ij} represents the probability of an observation i belonging to category j of the response variable.
- π_{i1} is the probability of the reference category for observation i.

- β_{0j} is the intercept for category j.
- $\beta_{1j}, \beta_{2j}, \ldots, \beta_{pj}$ are the coefficients associated with predictor variables $x_{i1}, x_{i2}, \ldots, x_{ip}$ for category j
- $x_{i1}, x_{i2}, \ldots, x_{ip}$ represent the predictor variables for observation i.

This equation models the log-odds ratio of the probability of belonging to category j compared to the reference category, providing a flexible framework for analyzing outcomes with more than two categories.

3.2 Link function

In this general equation, the link function is represented by a logit function: $log(\frac{\pi_{ij}}{\pi_{i1}})$. In multinomial regression, the most commonly used link function is the logit function, which is defined as the natural logarithm of the ratio of the probability of belonging to a specific category to the probability of belonging to the reference category.
The choice of the logit function as the link function in multinomial regression is grounded in its ability to **map probabilities onto a range covering all real numbers.** By transforming probabilities into log-odds, the logit function establishes a linear relationship between predictors and the log-odds of each category relative to the reference category. This not only simplifies interpretation but also ensures that model estimates are situated on a suitable scale for regression analysis. Furthermore, the logit function provides a level of **interpretability** that aligns with changes in predictor variables. Its **symmetry in odds** ratios across categories promotes consistency in comparing the effects of predictors on different outcome categories. **Widely adopted and standardized,** the logit function enjoys common usage in statistical literature and software packages, enhancing its practicality and facilitating cross-study comparisons. Additionally, the logit function demonstrates **numerical stability**, especially when handling extreme probabilities, thereby contributing to the reliability and robustness of the estimation process.

3.3 Model assumptions

The assumptions of multinomial logistic regression include:

- The dependent variable must be categorical and multinomial.
- Observations must be independent.
- There should be no perfect multicollinearity among independent variables.
- Log-odds probabilities are assumed to be a linear function of independent variables.
- Independence of irrelevant alternatives is assumed, meaning that the probabilities of choosing one category over another are not affected by the inclusion or exclusion of alternative categories.

4 Data Example