Multinomial Regression Tutorial

Barbara Flores

Table of contents

1 Overview **1**
1.1 Generalized Linear Model . . . 1
1.1.1 Link Function . . . 1
1.1.2 GLM Assumptions . . . 1
1.2 Multinomial Regression . . . 1
1.2.1 Examples of research questions . . . 1

2 Multinomial Distribution **2**

3 Model **2**
3.1 General Form . . . 2
3.2 Link function . . . 2
3.3 Model assumptions . . . 3

4 Data Example **3**
4.1 Dataset . . . 3
4.2 Model Fitting . . . 3
4.3 Model Interpretation . . . 4
4.4 Model Assessment . . . 5

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
$n > 0$ number of trials, $k > 0$ number of mutually exclusive events,
$p_1, ...p_k$ 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.
- $\beta_{1j}, \beta_{2j}, ..., \beta_{pj}$ are the coefficients associated with predictor variables $x_{i1}, x_{i2}, ..., x_{ip}$ for category j
- $x_{i1}, x_{i2}, ..., x_{ip}$ 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:

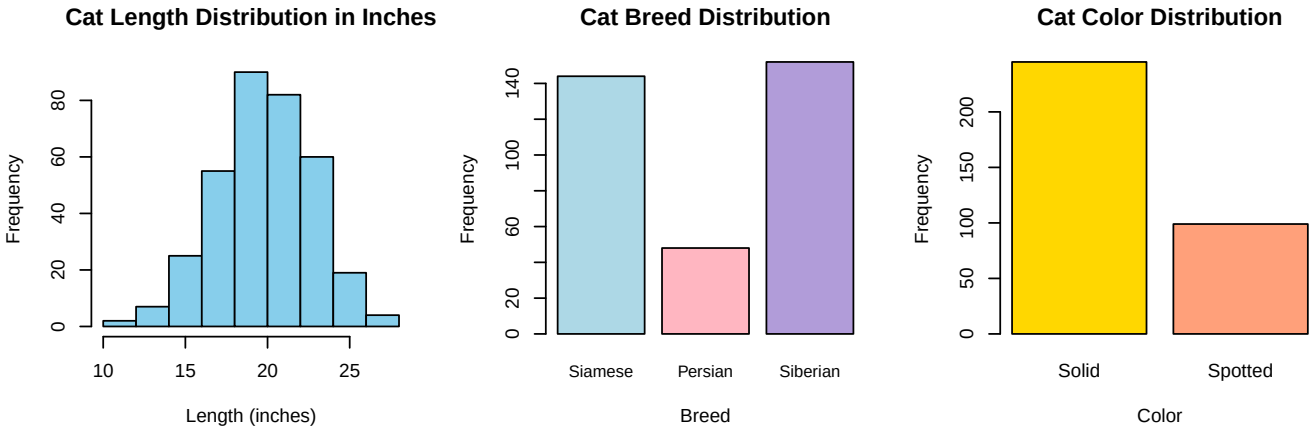
- The dependent variable must be categorical and multinomial.
- Observations must be independent.
- There should be no perfect multicollinearity among independent variables.
- Log-odds probabilities are assumed to be a linear function of independent variables.
- Independence of irrelevant alternatives 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

4.1 Dataset

To better understand the Multinomial Regression Tutorial model, we will review a practical example implemented in R. For this purpose, we will use a simulated database generated using probability distribution functions in R. The dataset is clean, containing no missing data, and comprises three variables: Y, X1, and X2. More details about the data generation are included at the end of this page.

For the purposes of this example, let's consider that this dataset represents various records of cats. Here, Y represents the cat's breed, with 1 = Siamese, 2 = Persian, and 3 = Siberian. Additionally, X1 represents the cat's length in inches, and X2 represents color, where 0 = Solid and 1 = Spotted. The objective of our example will be to define a model such that, based on the cat's length and color, we can predict its breed. The dataset has 344 observations. The variables **breed**, **color**, and **length inches** are distributed as follows:



From the graphs, we can see that there does appear to be a difference in the length among different types of cats and their breeds. For different colors, the distribution of length seems similar for both colors.

4.2 Model Fitting

To implement the Multinomial regression model, we will follow the following steps. First, it is important to have performed data cleaning and transformed the variables into their corresponding types. In our case, we already have a clean database, **cats**, which includes the numeric variables l**ength inches** and the factor variables **breed** and **color**. So, we will proceed with the model fitting.

In the following code, first, we use the library function to load the **nnet** package, which is necessary for fitting multinomial regression models. Then, we rearrange the **breed** variable using **Siamese** as the new reference level with relevel. This step is crucial as it affects how the other categories will be interpreted in the model. By fitting a multinomial model with the predictor variables **color** and **length_inches** using **breed2** as the response variable, we are establishing a baseline for comparison.

Subsequent estimates and comparisons will be made in relation to the reference category, in this case, **Siamese**. This approach aids in interpreting differences between **breed2** categories based on how they differ from 'Siamese'

```
library(nnet)
cats$breed2 <- relevel(cats$breed, ref = "Siamese")
model <- multinom(breed2 ~ color + length_inches, data = cats)
```

```
# weights:  12 (6 variable)
initial  value 377.922627
iter  10 value 319.866408
final  value 319.866256
converged
```

The output indicates successful convergence of the multinomial regression model in 10 iterations, with a notable improvement from the initial value.

4.3 Model Interpretation

After fitting the model, we print a summary to examine the statistics and parameters of the model using the following line:

```
summary(model)
```

```
Call:
multinom(formula = breed2 ~ color + length_inches, data = cats)

Coefficients:
         (Intercept) colorSpotted length_inches
Persian     3.372821    -1.292981    -0.2237668
Siberian   -2.921269    -0.149719     0.1497374

Std. Errors:
         (Intercept) colorSpotted length_inches
Persian    1.1674792    0.4796478    0.06247086
Siberian   0.8953241    0.2557371    0.04370468

Residual Deviance: 639.7325
AIC: 651.7325
```

From the obtained results, we can observe the following:
The **coefficients** indicate the direction and magnitude of the impact of each predictor on the odds of being in different **breed** categories. For instance, compared to **Siamese** (the reference category), **Persian** shows an odds increase of 3.37 in the intercept, a decrease of 0.22 in **length inches**, and a decrease of 1.29 if the color is **Spotted**.
For every one-unit increase in **length inches** the odds of the outcome being **Persian** (as opposed to 'Siamese') decrease by approximately 20%. Given:

$$e^{-0.2237668} = 0.8$$

$$1 - 0.8 = 0.2$$

The **Standard Errors** provide a measure of the precision of the estimated coefficients. The smaller the standard error, the more accurate the estimated coefficient.
The **Residual Deviance** is a measure of how much better the model fits compared to a null model. Lower values indicate a better fit.
AIC (Akaike Information Criterion) is a criterion assessing the model's quality, taking into account complexity. A lower AIC is preferred, suggesting a better balance between model fit and complexity. In general, this term is used to compare different models.

4.4 Model Assessment

The next step in our tutorial is to assess the model's performance. We utilize the confusionMatrix function, which compares the model's predictions obtained with **`head(predict(model))`** against the actual breed categories in the dataset (**`cats$breed2`**). This step provides a detailed evaluation, helping us understand the model's accuracy and its capability to correctly classify observations into different breed categories. For this task, we'll need to load the **`caret`** package beforehand. Finally, the necessary codes for this assessment can be found below

```
library(caret)

confusionMatrix(predict(model), cats$breed2, mode = "everything")
```

```
Confusion Matrix and Statistics

          Reference
Prediction Siamese Persian Siberian
  Siamese       74      30       46
  Persian        4       7        5
  Siberian      66      11      101

Overall Statistics

               Accuracy : 0.5291
                 95% CI : (0.4748, 0.5828)
    No Information Rate : 0.4419
    P-Value [Acc > NIR] : 0.0007099

                  Kappa : 0.1913

 Mcnemar's Test P-Value : 1.1e-05

Statistics by Class:

                     Class: Siamese Class: Persian Class: Siberian
Sensitivity                  0.5139        0.14583          0.6645
Specificity                  0.6200        0.96959          0.5990
Pos Pred Value               0.4933        0.43750          0.5674
Neg Pred Value               0.6392        0.87500          0.6928
Precision                    0.4933        0.43750          0.5674
Recall                       0.5139        0.14583          0.6645
F1                           0.5034        0.21875          0.6121
Prevalence                   0.4186        0.13953          0.4419
Detection Rate               0.2151        0.02035          0.2936
Detection Prevalence         0.4360        0.04651          0.5174
Balanced Accuracy            0.5669        0.55771          0.6317
```

The confusion matrix and accompanying statistics reveal insights into the multinomial logistic regression model's performance. With an overall accuracy of approximately 52.91%, the model demonstrates a moderate ability to correctly predict cat breeds based on length and color. However, the relatively low precision and recall values for individual breed categories, particularly for Persian cats, suggest challenges in accurately distinguishing among the breeds. Further refinement of the model or exploration of additional features may enhance its predictive capabilities. The Kappa statistic of 0.1913 indicates only fair agreement beyond what would be expected by chance. These findings highlight the importance of considering both overall accuracy and individual class performance when evaluating the model's effectiveness. To visualize the outcomes of our model, we project the values of our response variable (breed) onto synthesized data for the independent variables (length_inches, color), spanning their potential values. The predict function is utilized to yield the output as a probability, facilitating interpretation.The graph below illustrates the probabilities of the breed falling into the feasible categories based on the values of the independent variables. The probabilities (represented by colored lines) will always sum to 1 for a given value of the independent variables. The likelihood of the breed being Persian, in comparison to Siamese, diminishes with an increase in length inches and if the color is Spotted rather than Solid.

```r
plot_df <- data.frame(
  length_inches = rep(seq(min(cats$length_inches), max(cats$length_inches), 0.1), 2),
  color = rep(c("Solid", "Spotted"), each = 166)
)

preds <- cbind(plot_df, predict(model, newdata = plot_df, type = "probs"))

preds_long <- gather(preds, "level", "probability", 3:5)
preds_long$color <- as.factor(preds_long$color)

ggplot(preds_long, aes(x = length_inches, y = probability, col = color)) +
  geom_line() +
  facet_grid(level ~ .)
```