# Analytics Assignment 3: GLM - Multinomial Regression

Revanth Chowdary Ganga (rg361)

## 1. Introduction

### 1.1 Generalized Linear Models

Generalized Linear Models (GLMs) are a class of statistical models that extend linear regression to handle a broader range of response variable distributions such as binomial, Poisson, and gamma, making them suitable for diverse types of data. Unlike traditional linear regression, GLMs are not constrained by the assumption of normality.

The Primary components of a GLM are:

1. **Response Variable**: with a distribution such as Binomial, Poisson etc.
2. **Linear Predictor**: A Linear combination of the predictor variables (similar to Linear Regression)
3. **Link Function**: to connect the linear predictor to the expected value of the response variable.

The General form of a GLM is:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Where $g(\mu)$ is the link function, $\beta_i$ are the coefficients associated with the corresponding predictor variables $X_i$

### 1.2 Link Function

A link function is used to connect the linear predictor to the expected value of the response variable i.e. it describes how the mean of the response variable is related to a linear combination of the predictor variables. The link function, denoted as g($\mu$), transforms the linear predictor into a scale appropriate for the response variable. The choice of the link function depends on the nature of the response variable and the distribution assumed for it some examples of link funtions are **Logit, Inverse, Log** etc.

### 1.3 GLM: Multinomial Regression

Multinomial Regression is used when the response variable is categorical in nature and has more than 2 levels (categories). The link function used in Multinominal Regression is a **Logit** function which is defined as follows:

$$logit(p) = log(\frac{p}{1-p})$$

some sample research questions which can be answered by multinomial regression include:

1. What Occupation are people most likely to chose based on their parents occupation and their own education
2. What food preferences will an animal have based on its size and habitat

# 2. Probability Distribution

## 2.1 Assumed Probability Distribution

Multinomial distribution assumes that the response variable has a multinomial distribution, which is a generalization of the binomial distribution for categorical variables with more than 2 categories.

The Probability Mass Function of a multinomial distribution is given by the equation:

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

where n is the total number of observations or trails, $x_i$ is the counts for each category and $p_i$ are the probabilities for each category.

## 2.2 Support

The **Support** for multinomial regression is given by:

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

This implies that the number of times each outcome $x_i$ can occur is in the range 0 and the total number of observations n, and the sum of the count of all outcomes should add up to the total number of observations. k is the total number of possible categories of the outcome variable.

## 2.3 Parameters

The parameters for multinomial regression are that the number of trials "n" should be greater than 0 and the number of mutually exclusive events "k" should be greater than 0 with the probabilities of these events occurring $p_i$ taking values between 0 and 1 and sum of these probabilities should add up to 1 ($\sum_i p_i = 1$)

## 2.4 Example

Example if we try to predict the chances of picking a shape out of the circle, square and triangle and if we have 10 trials (n>0),

the **support** would be that each of these shapes can be picked anywhere between 0 and 10 times but the sum of the number of times the shapes are picked should add upto 10 (e.g. Circle-3, Square-5, Triangle-2)

(assuming the example given above is the input data) the probabilities $p_{circle} = 0.3$, $p_{square} = 0.5$, $p_{triangle} = 0.2$ would add up to 1.

# 3. Model

## 3.1 General Form

For computing the multinomial regression, one of the outcomes is set as the "reference" or "baseline" level and a logistic regression is performed between all the other levels with respect to the baseline (where j >1) , the general equation of each of these is of the form:

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

where $\pi_{i1}$ is the probability of the reference level being the outcome.

## 3.2 Link Function

As mentioned in Section-1, multinomial regression uses **Logit** as the link function. Logit function is used for the following reasons:

1. Range of Probabilities: Logit ensures that the predicted probabilities are between 0 and 1 (Parameter requirements of selected Regression)
2. Symmetry: The logit function is symmetric around 0.5, making it well-suited for modeling the odds of an event occurring
3. Interpretability: The function provides coefficients that represent the log-odds. This makes the interpretation of coefficients more intuitive in terms of how the odds of being in a particular category change with changes in the predictor variables.

## 3.3 Assumptions

Multinomial Regression works on the following assumptions:

1. Response Variable: The response variable is a categorical variable with multinomial distribution.
2. Independence: The observations are independent of each other.
3. Linearity: There is a linearly mapable (via link function) relationship between the predictor variables and the response variable

# 4. Sample Execution in R

## 4.1 Dataset

For a sample execution of a Multinominal GLM in R we will be using a **simulated** Dataset. Since this is a simulated Dataset, the data is clean and has no missing values, so we will not be performing any cleaning of data.

The initial data has been read and stored in a variable called `df`. The dataset contains 2 predictor variable and 1 response variable,

```
  X Y       X1 X2
1 1 2 21.34653  0
2 2 3 24.88894  1
3 3 1 24.65451  0
4 4 2 16.71951  0
5 5 1 17.55165  0
6 6 2 18.90001  1
```

| Variable | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| Y | 1.00000 | 1.00000 | 2.0000 | 2.0232558 | 3.00000 | 3.00000 |
| X1 | 11.12058 | 17.85909 | 19.9332 | 19.8625400 | 21.90855 | 27.66196 |
| X2 | 0.00000 | 0.00000 | 0.0000 | 0.2877907 | 1.00000 | 1.00000 |