Data Analysis Assignment 03 - Multinomial GLM

Divya Sharma (ds655)

Overview

GLMs (Generalized Linear Models) are a flexible extension of linear regression models. They allow the response variable to have a distribution other than a normal distribution, such as binomial, Poisson, etc.

Like linear regression, GLMs relate the response variable to explanatory variables using a linear function. But they also include a *link function* that transforms the response variable so its distribution matches the assumptions.

The key components of a GLM are:

- The response variable Y and its distribution (e.g. binomial, Poisson)
- The linear predictor - a linear function of the explanatory variables (like in linear regression)
- The link function g() that transforms the expected value of Y to match the linear predictor

The link function connects the linear predictor to the expected response, allowing a non-normal distribution for Y.

For example, a GLM could model student test scores based on hours studied and previous GPA. The response variable test score may follow a Poisson distribution. The linear predictor combines the explanatory variables hours studied and GPA. The log link function transforms the expected test score to match this linear combination.

A multinomial GLM handles a categorical response variable with more than 2 categories. The purpose is to model the probability of different outcomes. For example, a multinomial GLM could predict political party affiliation based on age, income, gender, etc. The response is party, a categorical variable with multiple outcomes. The linear predictor combines the explanatory variables. The logit link connects this to the probability of each party.

Some potential research questions for a GLM model can be:

- How does income level correlate with Republican vs Democrat affiliation?
- Do gender and age interact in predicting party?
- What variables most strongly predict party affiliation?

In summary, a multinomial GLM can model a categorical response to explain the factors influencing different outcomes. The link function handles the non-normal distribution.

Probability Distribution

In a multinomial logistic regression, the outcome variable is assumed to follow a multinomial distribution. This is a generalization of the binomial distribution for categorical variables with more than two categories.

The support of a multinomial distribution is a set of all possible outcomes of the categorical variable. For example, if the outcome variable is "color" with possible values "red", "green", and "blue", the support is {"red", "green", "blue"}.

The parameters of a multinomial distribution are the probabilities associated with each category of the outcome variable. These probabilities must be non-negative and sum up to 1.

For example, if the outcome variable is "color" with possible values "red", "green", and "blue", the parameters could be p_red = 0.3, p_green = 0.5, and p_blue = 0.2. These probabilities represent the likelihood of each color being the outcome.

In the context of a multinomial logistic regression, these probabilities are modeled as functions of the predictor variables. The model estimates a set of coefficients for each category of the outcome variable (except for a reference category), which are used to calculate the log-odds of the probabilities.

Model Overview

The general form of a Generalized Linear Model (GLM) is:

$$g(E(Y)) = \beta 0 + \beta 1 * X1 + \beta 2 * X2$$

where `g()` is the link function, `E(Y)` is the expected value of the response variable `Y`, and `0`, `1`, and `2` are the coefficients of the model.

For a multinomial logistic regression model, the link function is the *logit function*, and the model is:

$$log(P(Y = j)/P(Y = 1)) = \beta 0j + \beta 1j * X1 + \beta 2j * X2$$

The assumptions of the multinomial logistic regression model are:

- The response variable is a categorical variable with unordered categories.
- The observations are independent.
- There is no perfect multicollinearity among the predictor variables.
- The model is correctly specified (i.e., the form of the model is correct, and it includes all relevant predictors).
- The probabilities of the outcomes are a multinomial logit function of the predictor variables.

Data Example

Our simulated Dataset contains 344 observations of 3 variables. The following variables are present -

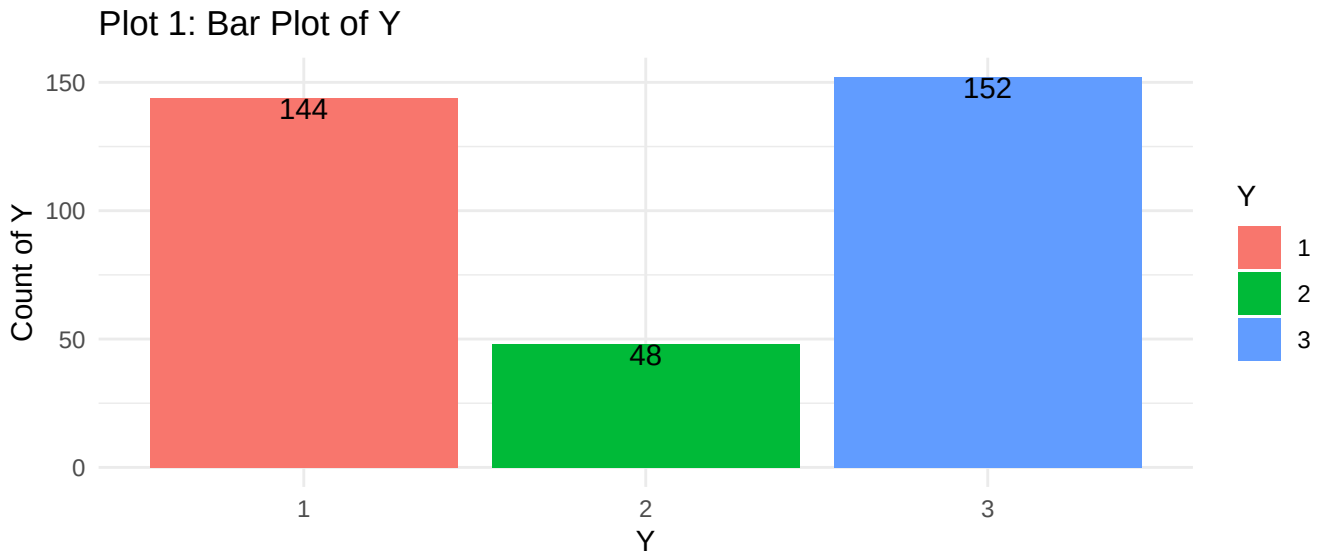
- `Y`, the multinomial dependent variable with three values - 1, 2, and 3
- `X1`, a continuous predictor with values ranging from 11.12 to 27.66
- `X2`, a categorical predictor with values 0 and 1

Sample Data

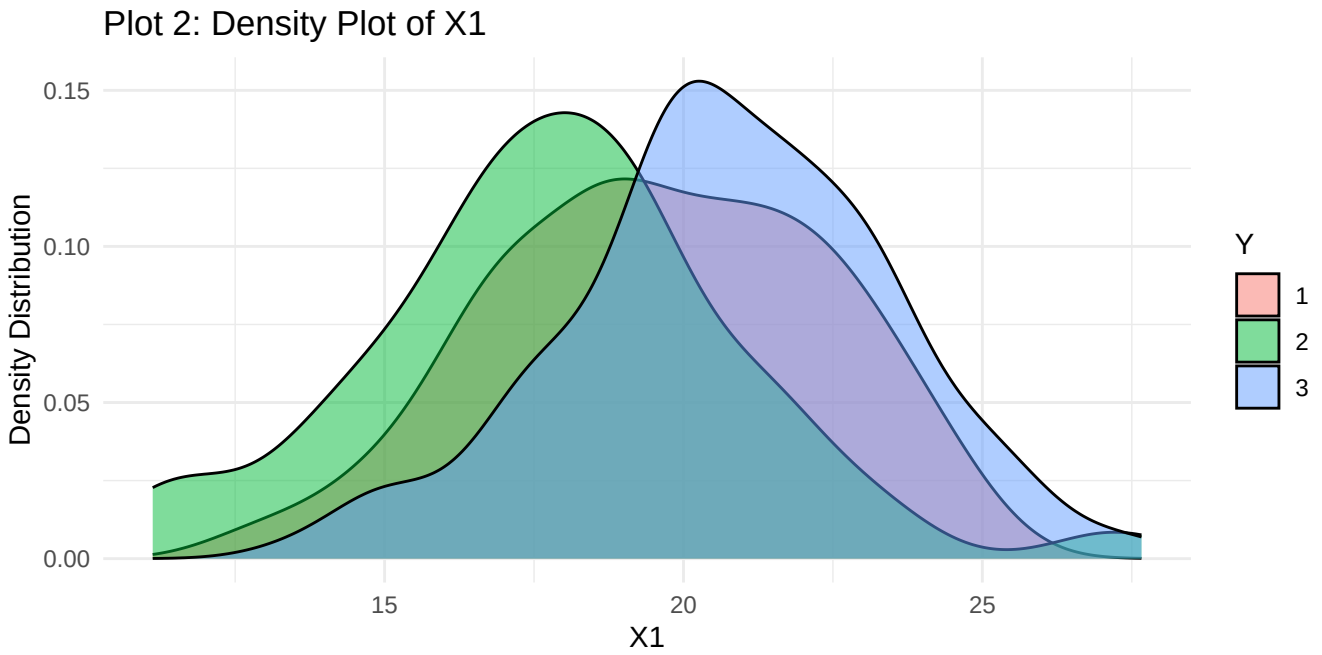
Y	X1	X2
2	21.34653	0
3	24.88894	1
1	24.65451	0
2	16.71951	0
1	17.55165	0
2	18.90001	1

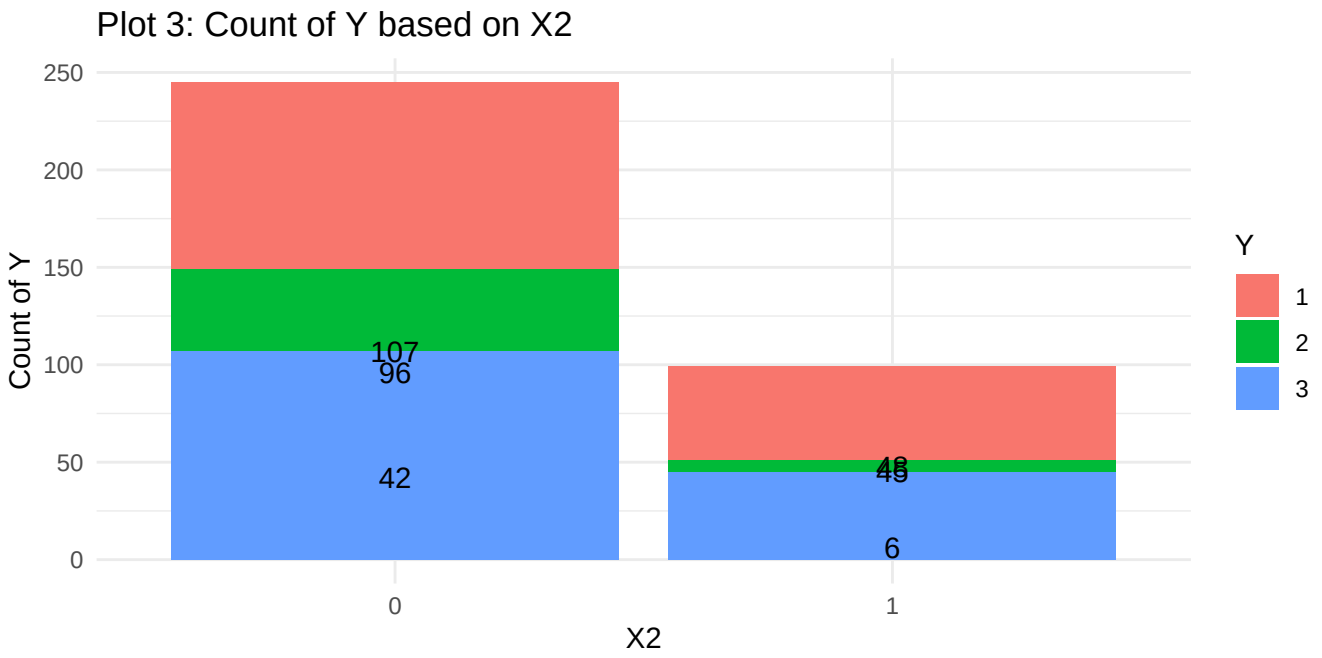
Here are some summary statistics for the dataset:

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



- Out of 344 observations, 144 are Y category 1, 48 are Y category 2, and 152 are Y category 3, as can be observed in Plot 1.
- The distribution of Y = 3 peaks at X1 > 20, the distribution of Y = 1 peaks at X1 = 20, and the distribution of Y = 2 peaks at X1 < 20, as can be observed in Plot 2.
- X2 has 99 1 values and 245 0 values. The lowest distribution within these is of Y = 2, as expected by the overall distribution of the Y variable (Plot 3)





Model

Fitting a multinomial Generalized Linear Model (GLM) involves estimating the relationship between a categorical dependent variable (Y in this case) and one or more predictor variables (X1 and X2 in this case). In a multinomial GLM, the dependent variable Y is not binary, but rather it can take on multiple categories. The model estimates the log odds of being in each category of Y compared to a reference category, as a linear function of the predictor variables.

Model Fitting

Code:

```
library(nnet, quietly = T)
model <- multinom(Y ~ X1 + X2, data = base_data)
```

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

Assessing the model involves checking the *goodness of fit* and validating the assumptions of the model. For a multinomial logistic regression model, there are no assumptions about the distribution of predictors (like normality or linearity), but it assumes that the outcomes are correctly specified and that the observations are independent. The goodness of fit can be observed in the model summary, as shown below :

```
summary(model)
```

```
Call:
multinom(formula = Y ~ X1 + X2, data = base_data)

Coefficients:
  (Intercept)         X1        X2
2    3.372821 -0.2237668 -1.292981
3   -2.921269  0.1497374 -0.149719

Std. Errors:
  (Intercept)         X1        X2
2   1.1674792 0.06247086 0.4796478
3   0.8953241 0.04370468 0.2557371

Residual Deviance: 639.7325
AIC: 651.7325
```

- Coefficients: These are the estimated parameters for the model. For example, the coefficient for X1 in category 2 is -0.2237668. This means for each one unit increase in X1, the log odds of Y being in category 2 versus the reference category (category 1) decrease by 0.2237668, holding all other variables constant.
- Std. Errors: These are the standard errors of the coefficients. They measure the variability in the estimate for the coefficient. Smaller standard errors mean the estimate is more precise.
- Residual Deviance: This is a measure of how well the model fits the data. Lower values indicate a better fit. In this case, the residual deviance is 639.7325.
- AIC: This is the Akaike Information Criterion, a measure of the relative quality of statistical models. Lower values indicate a better model. In this case, the AIC is 651.7325.

For a multinomial logistic regression, the coefficients represent the log-odds of the outcomes relative to a reference category. For example, if X1 had a coefficient of 0.5 for category 2, it means that a one unit increase in X1 leads to an increase in the log-odds of category 2 versus the reference category by 0.5, holding all other predictors constant.

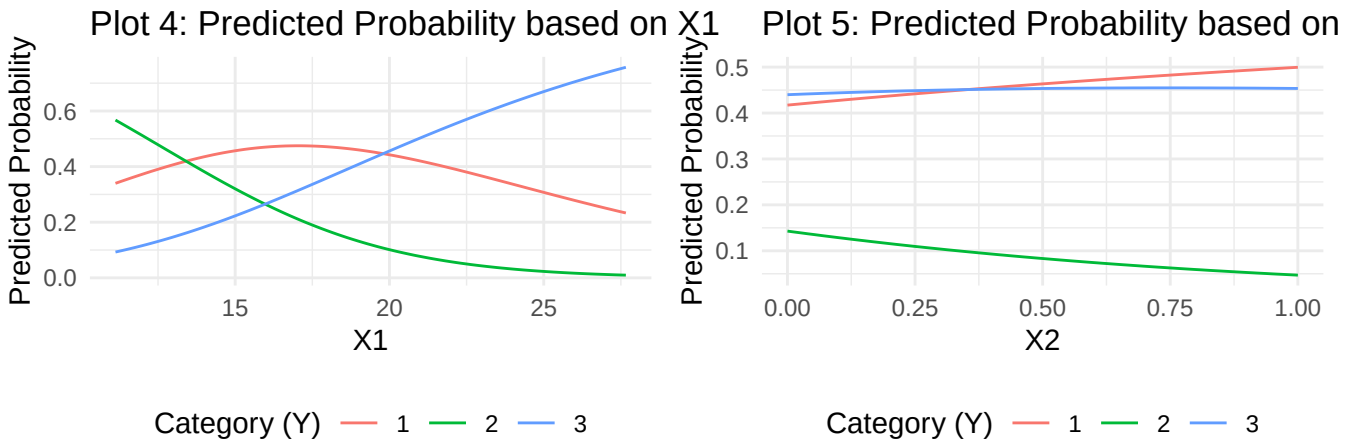
For category Y=2 compared to the reference Y=1:

- The intercept is 3.37, so the baseline log odds of Y=2 vs Y=1 is 3.37.
- The coefficient for X1 is -0.22, meaning a 1 unit increase in X1 decreases the log odds of Y=2 vs Y=1 by 0.22.
- The odds ratio is exp(-0.22) = 0.80. So a 1 unit increase in X1 decreases the relative odds of Y=2 vs Y=1 by 0.80 times or 20%.

For Y=3 compared to Y=1:

- The intercept is -2.92, so the baseline log odds of Y=3 vs Y=1 is -2.92.
- The coefficient for X1 is 0.15, meaning a 1 unit increase in X1 increases the log odds of Y=3 vs Y=1 by 0.15.
- The odds ratio is exp(0.15) = 1.16. So a 1 unit increase in X1 increases the relative odds of Y=3 vs Y=1 by 1.16 times or 16%.
- The X2 coefficient is significant and negative for both Y=2 and Y=3, indicating X2=1 reduces the odds of these categories compared to Y=1.

In summary, X1 has opposite effects on the relative odds of Y=2 and Y=3, while X2 significantly reduces the odds of both compared to the reference Y=1. The intercepts give the baseline log odds.



These two graphs illustrate the predicted probabilities of each category of the response variable (y=1, y=2, and y=3) as functions of the predictors X1 and X2, respectively.

In the left graph (Plot 4), the x-axis represents the X1 variable and the y-axis represents the predicted probabilities. Each line represents a different category of the response variable (y=1, y=2, or y=3). The X2 variable is held constant at its mean value. This graph shows how the predicted probabilities change with varying values of X1, while keeping X2 constant.

Similarly, the right graph (Plot 5) shows how the predicted probabilities change with varying values of X2, while keeping X1 constant.

By examining the slopes and positions of the lines, we can understand the relationships between the predictors and the response variable categories.