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, β_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(μ), 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 functions are **Logit, Inverse, Log** etc.

1.3 GLM: Multinomial Regression

Multinomial Regression is used when the response variable is categorical in nature with a multinomial distribution and has more than 2 levels (categories). The link function used in Multinomial Regression is a **Logit** function.

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 outcomes "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 ($\sum_i p_i = 1$)

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 >=2) , 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 π_{i1} is the probability of the reference level being the outcome and π_{ij} is the probability of the selected level being the outcome.

3.2 Link Function

Multinomial regression uses **Logit** as the link function, which is given by the formula:

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

Logit function is used in multinomial regression 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 probability of response variable outcomes.

4. Sample Execution in R

4.1 Dataset

For the sample execution of a Multinomial 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 dataset has been loaded into a dataframe `df` and contains **344** observations of the following 3 variables:

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

The following steps were performed on the dataset before using it for modelling:

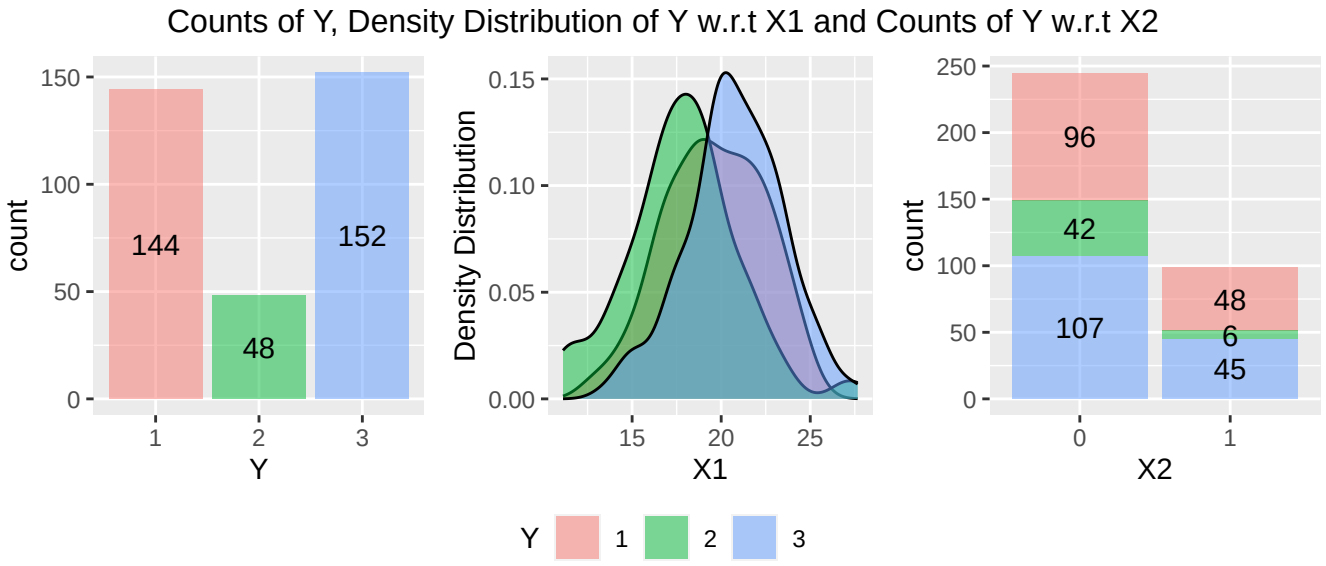
1. original dataset has a serial number column `X` which has been removed as it is not required.
2. The Variables Y and X2 were converted to the correct datatypes (INT -> Factor)

The summary statistics of the data is as follows:

Variable	1	2	3	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	0
Y	144	48	152	NA	NA	NA	NA	NA	NA	NA
X1	NA	NA	NA	11.12058	17.85909	19.9332	19.86254	21.90855	27.66196	NA
X2	99	NA	NA	NA	NA	NA	NA	NA	NA	245

Note: NA is displayed in the summary statistics columns for Y and X2 as they are categorical variables.

The distributions of Y are as follows:



4.2 Model Fitting

We use the `multinom` function from the `nnet` package in R to fit the multinomial model as follows:

```
library(nnet)
df$Y2 <- relevel(df$Y, ref=1)
model <- multinom(Y2 ~X1+X2, data = df)
```

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

Though it was not required, we created a new variable `Y2` in which we specify the category we want to use as reference level. The values which are printed when the model is fit provide information about how the model was optimized (uses maximum likelihood)

4.3 Model Interpretation

We use the `summary` function to view the details of the model fit:

```
summary(model)
```

```
Call:
multinom(formula = Y2 ~ X1 + X2, data = df)

Coefficients:
  (Intercept)         X1       X21
2    3.372821 -0.2237668 -1.292981
3   -2.921269  0.1497374 -0.149719

Std. Errors:
  (Intercept)         X1       X21
2   1.1674792 0.06247086 0.4796478
3   0.8953241 0.04370468 0.2557371

Residual Deviance: 639.7325
AIC: 651.7325
```

For a multinomial logistic regression, the coefficients represent the log-odds of the outcomes relative to a reference category. Sample interpretation of the model summary coefficientsis for Y=2 with comparison for Y=1 as reference category is as follows:

1. raw-coefficients: For each one unit increase in X1, the log odds of Y being in category 2 versus the reference category (category 1) decreases by 0.2237668, holding all other variables constant.

2. A one unit increase in X1 is associated with a ~20% decrease of the odds of the outcome being category 2 as compared to category 1. The value is derived using the following calculation:

$$e^{-0.2237668} = 0.8 \; |1 - 0.8 = 0.2 \; |20\%$$

3. The log odds of being in the category 2 vs Category 1 will decrease by 1.29 if moving from 0 to 1 for X2
4. using similar calculation as in point 2, The odds of being in category 2 vs Category 1 are ~73% lower if X2 is 1 as compared to 0.
5. The intercept is 3.37, so the baseline log odds of Y=2 vs Y=1 is 3.37.

Other information in Summary:

1. Std. Errors: They measure the variability in the estimate for the coefficient. Smaller standard errors mean the estimate is more precise.
2. Residual Deviance: This is a measure of how well the model fits the data. Lower values indicate a better fit
3. AIC: This is a measure of the relative quality of statistical models. Lower values indicate a better model. .

4.4 Model Assessment

We assess the model in the following categories:

4.4.1 Assumptions

1. Linearity: can be verified by fitting separate Logistic Regression Models for each of the combinations
2. Independence: is verified by study design and domain knowledge

4.4.2 Model Fit

To assess the model fit we use metrics based on the likelihood like **Deviance** and **AIC** which can be viewed and interpreted in the model summary as explained in section 4.3

4.4.3 Predictions

To asses the predictions we use the `confusionMatrix` function from `caret` package to study different parameters like sensitivity, Kappa score, F1 score etc.

To generate these values, we make the model predict the outcome variable based on the initial dataset and compare it against the actual values of the response variable.

```
confusionMatrix(predict(model), df$Y2, mode = "everything")
```

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3
         1  74  30  46
         2   4   7   5
         3  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: 1 Class: 2 Class: 3
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
```

Interpretation of selected statistics from the above results:

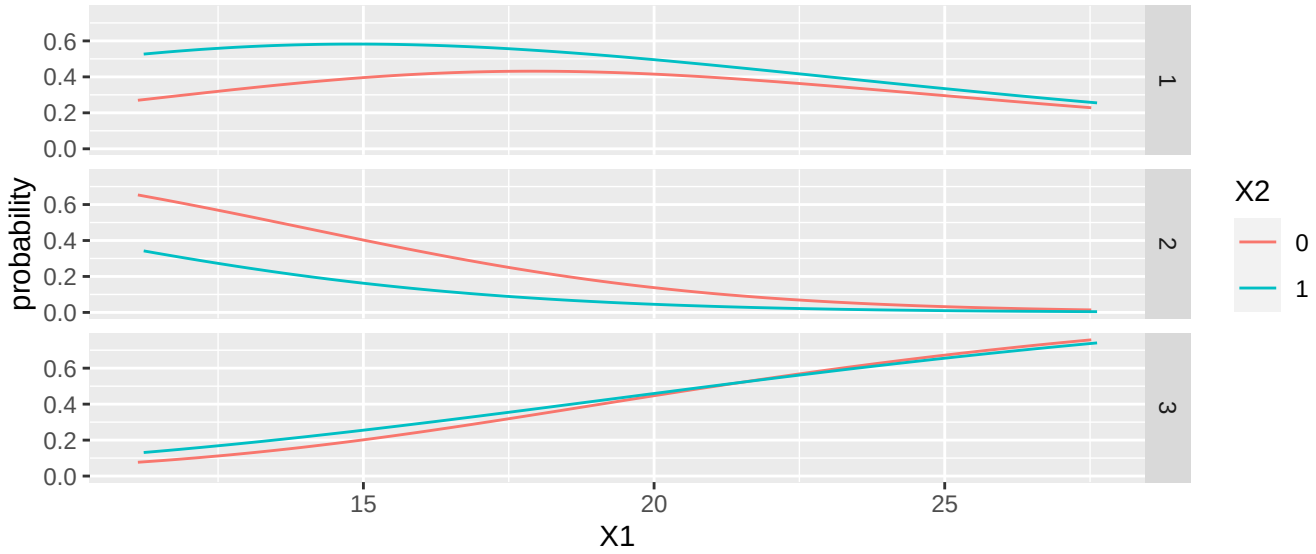
1. Confusion Matrix : Shows the comparison of the actual values vs the values predicted by the model. ideally the left diagonal values should be maximum and other elements should be minimum.
2. Accuracy: Our model has an accuracy of approximately 53% which is comparatively low
3. Kappa Score : It calculates the degree of agreement between the model's predictions and the actual class labels while accounting for the possibility of chance agreement. a score of ~0.19 indicates that the model is better than random chance prediction, however the usual desirable range is atleast 0.4+
4. Response Variable category-wise statistics
    1. Sensitivity: measure of how many true cases are identified as true, while the scores are low for category 1 and 3, it is very low for category 2
    2. Specificity: measure of how many negative cases are identified as negative, here an inverse pattern to sensitivity is seen where the score is good for category 2 but lower for category 1 and 3
    3. F1 Score: is a measure of overall accuracy which takes into account the precision and recall, while it is in acceptable range for category 1 and category 3 , it is very low for category 2

Overall the model isn't very strong and the possible reasons maybe the data imbalance and low sample size.

4.5 Model Results Visualization

To visualize the results of our model, we predict the values of our response variable(Y) on generated data for the independent variables (X1, X2) such that we cover their range of possible values.

We make the `predict` function generate the output as a probability so that it is easier for us to interpret.



The graph above shows the probabilities of Y being in the possible categories given the values of the independent variables. The probabilities (colored lines) will always add up to 1 for given values of the independent variables.

As implied by the sample interpretation of the coefficients in section 4.3, the probability of Y being category 2 (Plot 2) as compared to category 1 decreases with an increase in X1 and if X2 is 1 as compared to 0.