Fitting multinomial GLM

1. Remove non-informative columns, namely; ShortSourceA, Encounter, Location, Project, Recording, Species.
2. Check and clean the response variable

Class 56 had only one observation, which is too low for multinomial modelling, so it is removed. Dropping it also updates the factor levels accordingly.

1. Further remove classes with less than 8 observations

Some other classes may still be under-represented. This step filters the dataset to include only those categories with at least 8 observations, improving model stability and avoiding glmnet warnings.

1. Fit a LASSO-regularised multinomial model

* model.matrix() creates the design matrix for predictors.
* cv.glmnet() fits a multinomial LASSO model using cross-validation to find the best lambda (regularization strength).
* alpha = 1 specifies LASSO (as opposed to ridge regression)

1. Identify selected variables

Extract all variables with non-zero coefficients in the LASSO model (i.e., those considered important). Removes the intercept term.

1. Further reduce to the 10 most important variables

* Computed the total influence of each variable by summing the absolute values of its coefficients across all classes.
* Selects the top 10 variables with the strongest overall influence.

The 10 selected variables were; DURATION, DCSTDDEV, FREQCOFM, INFLSTDDEVDELTA, INFLMINDELTA, FREQBEGENDRATIO, INFLMEDIANDELTA, INFLMAXDELTA, DCQUARTER3MEAN, DCQUARTER2MEAN

|  |  |
| --- | --- |
| Variable | Meaning |
| DURATION | Whistle duration |
| DCSTDDEV | Standard deviation of dominant frequency contour over time |
| FREQCOFM | Frequency of the centre of mass (mean frequency weighted by amplitude) |
| INFLSTDDEVDELTA | Standard deviation of the frequency delta between inflection points |
| INFLMINDELTA | Minimum frequency change between inflection points |
| FREQBEGENDRATIO | Ratio of frequency at the beginning vs end of the signal |
| INFLMEDIANDELTA | Median of frequency change between inflection points |
| INFLMAXDELTA | Maximum frequency change between inflection points |
| DCQUARTER3MEAN | Mean frequency in the third quarter of the duration (25-75% window) |
| DCQUARTER2MEAN | Mean frequency in the second quarter of the duration (0-25% window) |

1. Fit a mulinomial model using only the top 10 variables

* Subsets the dataset to include only the response and top 10 predictors.
* Fits a clean and interpretable multinomial GLM with just these 10 variables.

**Model Diagnostics**

Accuracy: 0.423

* The model correctly predicted the category for 42.3% of the observations in the dataset.

Misclassification rate: 0.577

* 58% of the predictions were incorrect – the predicted category did not match the actual category.

VIFs

| Variable | VIF | Interpretation |
| --- | --- | --- |
| DCSTDDEV | 1.79 | Low multicollinearity - no issue |
| FREQCOFM | 1.28 | Very low multicollinearity - great |
| FREQBEGENDRATIO | 1.09 | Excellent - almost no multicollinearity |
| DCQUARTER3MEAN | 7.35 | Moderate multicollinearity - worth watching |
| DCQUARTER2MEAN | 6.94 | Moderate multicollinearity |
| INFLSTDDEVDELTA | 33.03 | High multicollinearity |
| INFLMINDELTA | 55.13 | Very high multicollinearity |
| INFLMEDIANDELTA | 34.76 | Very high multicollinearity |
| INFLMAXDELTA | 32.09 | Very high multicollinearity |

4 variables (INFLMINDELTA, INFLMEDIANDELTA, INFLMAXDELTA, INFLSTDDEVDELTA) show extremely high multicollinearity.

They likely capture similar information - e.g., all describe deltas between inflection points.

* Only keep INDMEDIANDELTA

Recheck VIF values. All values are below 10, which is a typical warning threshold – so no serious multicollinearity issue.

The remaining two with VIFs ~7 (DCQUARTER3MEAN, DCQUARTER2MEAN) are measuring frequency in overlapping parts of the call. Remove one of these variables to further reduce multicollinearity.

Remove DCQUARTER2MEAN

* Now all variables in the model have very small VIF values

So now our model is

category ~ DURATION + DCSTDDEV + FREQCOFM + FREQBEGENDRATIO + INFLMEDIANDELTA + DCQUARTER3MEAN

This formula tells us that the response variable category is predicted using six acoustic features related to duration, frequency variation, and pitch contour structure.

**Inference approach to evaluating model performance**

1. Coefficient Significance

For class 2:  
Statistically significant predictors for Class 2:

* DURATION (negative effect)
* FREQCOFM (negative effect)
* FREQBEGENDRATIO (strong negative effect)
* INFLMEDIANDELTA (positive effect)

Not significant:

* DCSTDDEV
* DCQUARTER3MEAN

For class 7:  
Strongest predictors for Class 7:

* DURATION: strong negative effect
* FREQBEGENDRATIO: strong negative effect

Not statistically significant:

* DCSTDDEV
* FREQCOFM
* INFLMEDIANDELTA
* DCQUARTER3MEAN

*Significance Table*

A black numbers and letters on a white background

AI-generated content may be incorrect.

Interpretation

* For classes 2, 3, 7, 8, and 13, the intercepts are highly significant, suggesting the baseline log-odds of being in these classes (relative to the reference category) are meaningfully different from zero.
* Class 16's intercept is not significant, meaning its baseline log-odds are not statistically distinguishable from the reference category (possibly due to low class size or overlap in features).

1. Likelihood Ratio Test

Fitting a full model and a reduced model (model without one of the predictors) and comparing if the removed variable improves model fit.

DURATION – highly important

DCSTDDEV – very important (less significant than other variables)

FREQCOFM – highly important

FREQBEGENDRATIO – highly important

INFLMEDIANDELTA – highly important

DCQUARTER3MEAN – highly important

1. Confidence intervals for coefficients

DURATION

* Significant negative effect in many classes: 2, 3, 7, 8, 13, 24, 29, 31, 38, 40
* Often excluded 0 → a consistent predictor of lower odds in certain classes
* In some classes (e.g., 16, 19, 26, 27), it includes 0 → not significant

DCSTDDEV

* Intervals frequently include 0
* Suggests it is not a reliable predictor - effect is highly uncertain across most classes
* Occasionally has wide intervals, suggesting instability

FREQCOFM (Frequency Center of Mass)

* Strong negative effect in classes 2, 13, 33, 42, 51, 60 — intervals entirely < 0
* Often includes 0 → less consistent but has strong effects for some classes

FREQBEGENDRATIO

* Very consistent: Most classes have strong negative coefficients
* CI excludes 0 in nearly every class - e.g., 2, 3, 7, 8, 13, 20, 21, 24, 29, 31, 33, 34, 38, 40, 41, 42, 43, 51, 52, 60
* One of the most reliable and important predictors

INFLMEDIANDELTA

* Significant positive effect in several classes: 2, 8, 13, 24, 29, 33, 34, 38, 41, 51, 60
* In some others, includes 0 → moderately reliable

DCQUARTER3MEAN

* CIs often include 0, indicating weak or uncertain effect in most classes
* Occasionally just barely significant (e.g., class 3, 29)

1. Residuals

A graph of a number of bars

AI-generated content may be incorrect.

Interpretation

Classes 16, 21 and 60 are almost entirely misclassified.

* Potentially due to low sample size, overlapping features with other classes or lack of unique signal in selected variables

Classes 23, 53 and 57 stand out as having strong predictive performance

* This model finds their acoustic profiles more distinctive or consistent

Overall Summary:  
This inference-driven evaluation highlights that:

* FREQBEGENDRATIO, DURATION, and INFLMEDIANDELTA are consistently useful predictors.
* Some variables (e.g., DCSTDDEV, DCQUARTER3MEAN) are less reliable and may be candidates for removal.
* Residual analysis revealed specific classes where the model struggles or excels.
* These insights guide both interpretation and future model refinement (e.g., feature engineering or class balancing).