Identify the distribution for each model

Morgan Gray

Contents

1 Introduction

The objective of this analysis was to determine the most suitable probability distribution for each of six generalized linear mixed models (GLMMs) fit to plant observation data. Two response variable types were considered: species richness (count data) and percent cover abundance (continuous data). For each response variable, models were fit for three plant groups: native species, native forb species (a subset of native species), and non-native species.

> **Note**
>
> The subsequent model fitting, selection, and performance assessments are not described here.

2 Methods

I conducted a systematic analysis combining visual inspection, statistical tests, and model comparisons to identify the most suitable probability distributions for count and continuous data. Modular functions were developed to handle each assessment component, including visualization generation and statistical testing. This approach provided multiple lines of evidence for selecting the most appropriate distribution for subsequent analyses.

The evaluation framework included:

- Visual inspection via histograms and quantile-quantile (QQ) plots
- Distribution fitting tests specific to data type:
  - Count data: goodness-of-fit tests and pairwise Vuong tests

– Continuous data: Shapiro-Wilks test

The influence of data transformations (standardized, log, and square root) was also explored. Log transformation was applied to positively skewed data or data with non-negative values, while square root transformation was used to reduce skewness in count-like or non-negative data. A constant of 1 was added to all values prior to log transformation for distributions requiring strictly positive values.

2.1 Histograms

The analysis for each data type began with visual inspection of histograms generated from the raw data. These visualizations revealed key distributional features, including central tendency, spread, symmetry, and potential outliers.

For count data, histogram shapes indicated potential alignment with common distributions like Poisson (right-skewed with single peak) or negative binomial (right-skewed with longer tail). For continuous data, histograms were generated for raw, standardized, log-transformed, and square-root transformed values.

2.2 Goodness-of-fit

For richness (count) data, discrete goodness-of-fit tests were applied to evaluate Poisson and negative binomial distributions. P-values from these tests were interpreted as follows:

- $p > 0.05$: The data were consistent with the distribution (failure to reject the null hypothesis).
- $p < 0.05$: The data significantly deviated from the distribution (rejection of the null hypothesis).

For abundance (continuous) data, the Shapiro-Wilk test was used to evaluate normality of both raw and transformed data, using the same significance thresholds.

2.2.1 Vuong tests

For count data, Vuong tests were used to directly compare the fit of competing models (i.e., Poisson vs. negative binomial). The test produces a z-statistic and p-value, where:

- Positive z-statistics ($p < 0.05$): First model provides better fit
- Negative z-statistics ($p < 0.05$): Second model provides better fit
- $p > 0.05$: Models are statistically indistinguishable

2.2.2 Quantile-quantile plots

QQ plots compared observed data against theoretical distributions. While not providing discrete test statistics, these visualizations supported statistical test results and offered insights when data failed to fit common distributions. I evaluated plot patterns focusing on:

- Overall adherence to the diagonal reference line
- Nature of deviations (random vs. systematic)
- Patterns at distribution extremes

Points falling along the diagonal reference line indicated good agreement between observed and theoretical distributions. Deviations from the line, particularly systematic patterns, suggested departures from the theoretical distribution. In ecological count data, deviations at the extremes, often due to rare species (excess zeros) or highly abundant counts, were examined for their implications for model selection (e.g., zero-inflated models or negative binomial models).

For count data, QQ plots were reviewed for Poisson and negative binomial distributions. When initial tests were inconclusive (e.g., for non-native counts), QQ plots for alternative distributions (normal, log-normal, gamma) were examined based on histogram shapes.

In ecological count data, deviations are often observed at the extremes, particularly due to rare species (excess zeros) or highly abundant counts. These patterns can aid in model selection. For example, consistent deviations at low values might suggest the need for a zero-inflated model, while heavy tails might favor a negative binomial over a Poisson distribution.

2.2.3 Additional plots for continuous data

For abundance (continuous) data, QQ plots were generated for normal distributions fit to raw, standardized, square-root transformed, and log-transformed values. In addition to QQ plots, density plots (overlaid on histograms), cumulative distribution function (CDF) plots, and probability-probability (PP) plots were used to visualize the comparison between empirical and theoretical distributions.

3 Results

3.1 Richness

Preview the first 10 rows of the data table for richness (rich) to see the column names and formats.