# 2025079_dharma_plots

2025-07-09

## Using DHARMa package to evaluate the fit of our negative binomial model

- Model format for all data from all journals **(N = 47353)**
  - MASS::glm.nb(is.referenced.by.count~ da_factor + log(age.in.months) + container.title + container.title*da_factor + log(age.in.months)*da_factor + container.title*log(age.in.months) + log(age.in.months)*da_factor*container.title, data = nsd_yes_metadata, link = log)
  - Data nsd_yes_metadata was filtered to remove all NAs from variables da_factor, age.in.months, and container.title to allow for some of the below visualizations.
  - See below for number of papers from each journal.

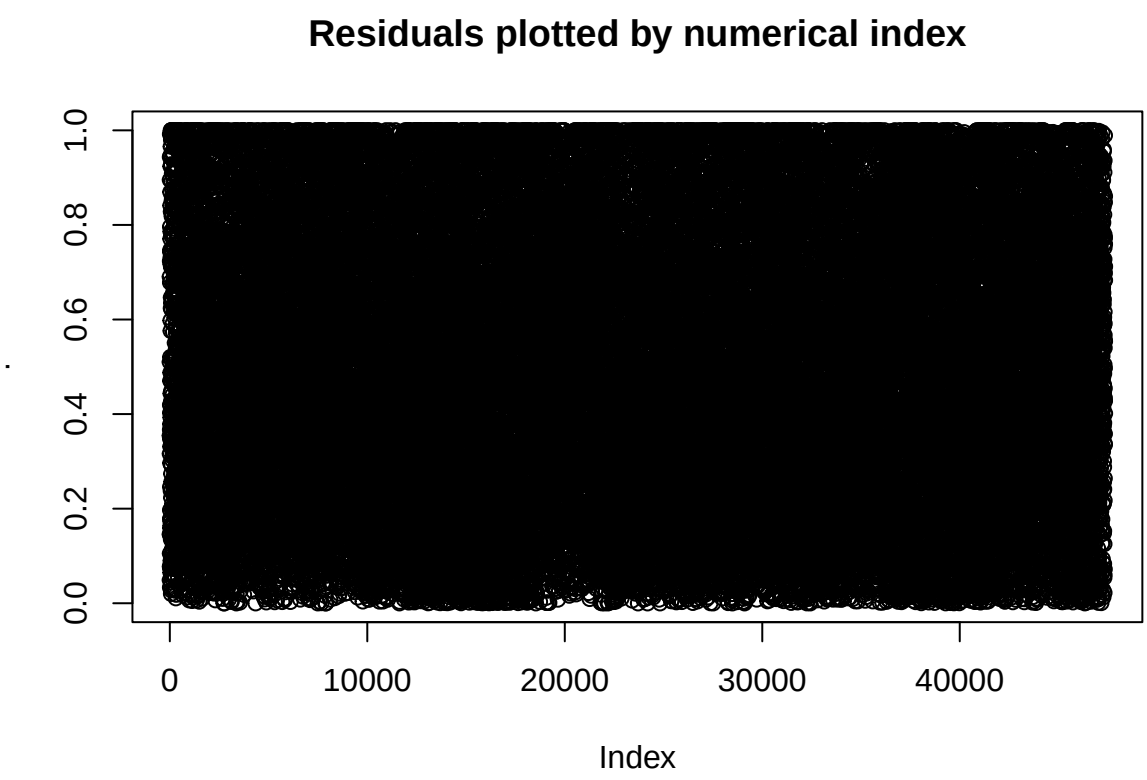
| container.title | n |
|---|---|
| Applied and Environmental Microbiology | 8613 |
| Genome Announcements | 6578 |
| Microbiology Resource Announcements | 5691 |
| Journal of Bacteriology | 4656 |
| Journal of Virology | 4577 |
| Journal of Clinical Microbiology | 4369 |
| Antimicrobial Agents and Chemotherapy | 3223 |
| Microbiology Spectrum | 2900 |
| mBio | 2438 |
| Infection and Immunity | 1854 |
| mSystems | 1406 |
| mSphere | 1041 |
| Journal of Microbiology & Biology Education | 7 |

### Plot Residuals

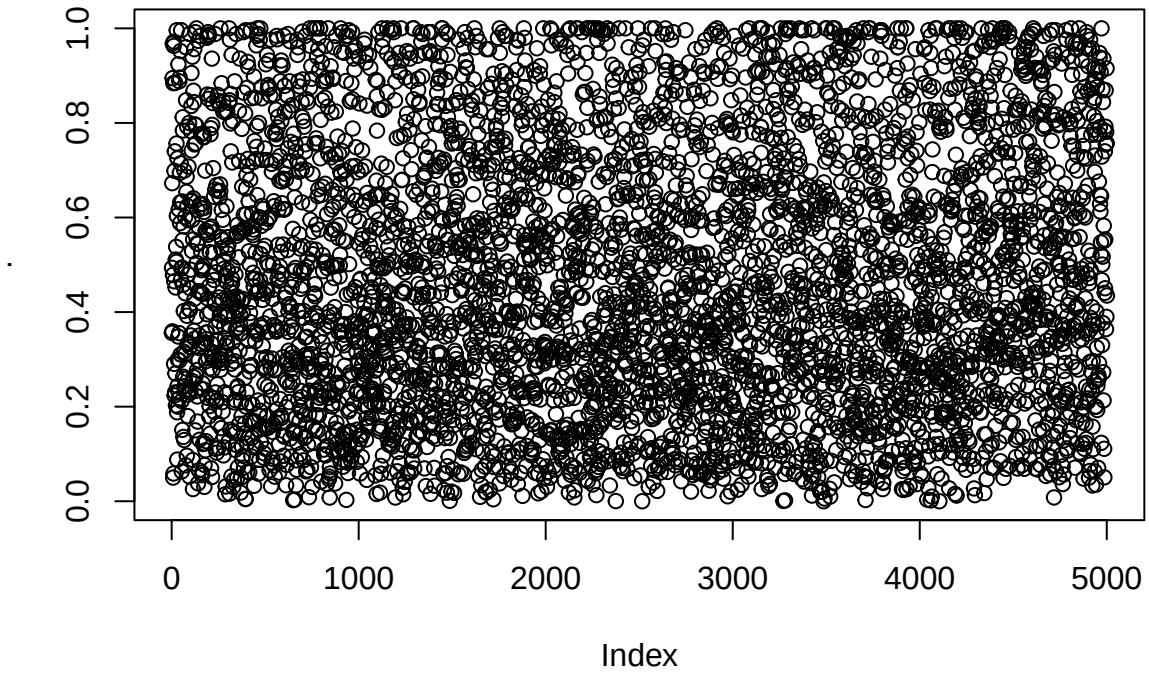
- Graph is very busy with all 47K observations
- see 2nd plot for 5K observations for a smaller group to examine (index 25000-30000)
- Residuals look like an amorphous blob as suggested by Abner@CSCAR

```
simulationOutput <- simulateResiduals(fittedModel = total_model, plot = F)

residuals(simulationOutput)  %>%
  plot(main = "Residuals plotted by numerical index")
```




```
residuals(simulationOutput)[25000:30000] %>%
  plot(main = "Residuals for observation indices 25000 - 3000")
```




## Plots for Residuals - QQ and DHARMa

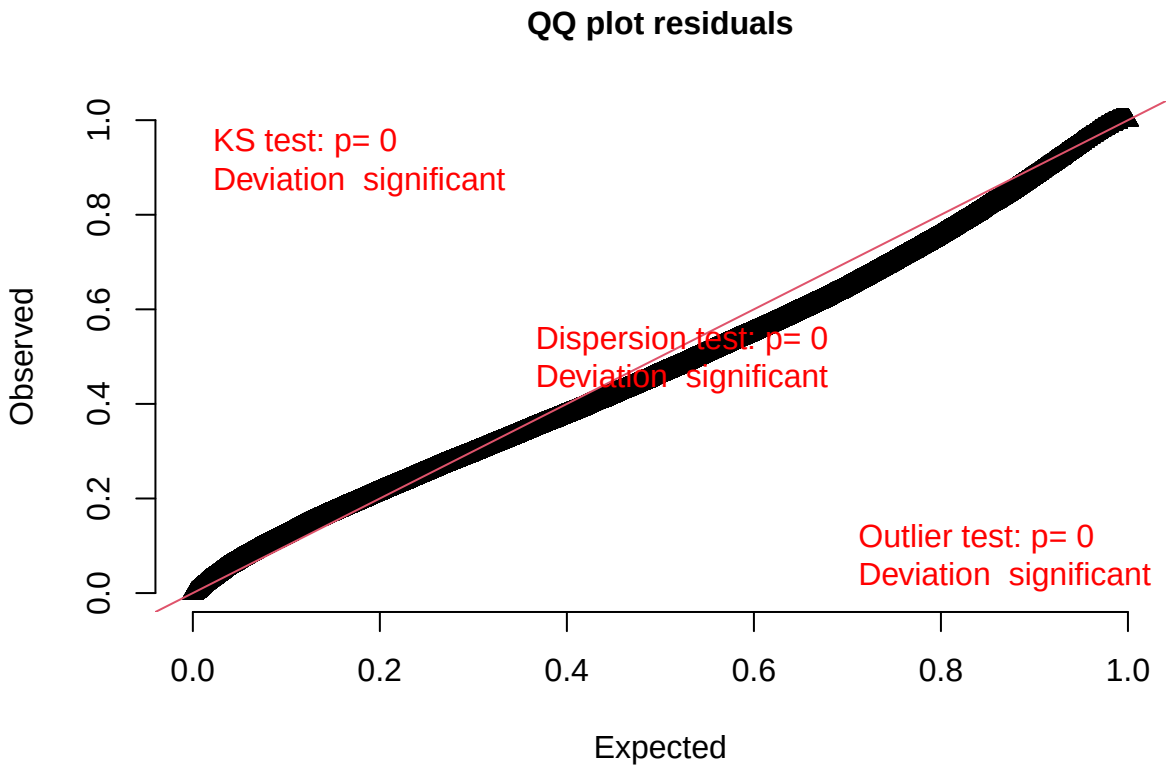
### QQ Plot

- KS Test = Two sample Kolmogorov-Smirnov (KS) Test
  - This function tests the overall uniformity of the simulated residuals in a DHARMa object - Deviation is significant between the expected residuals and the actual observed residuals.
  - "If the P value is small, conclude that the two groups were sampled from populations with different distributions." -Prism help page
- Dispersion Test
  - This function performs simulation-based tests for over/underdispersion
  - Over / underdispersion means that the observed data is more / less dispersed than expected under the fitted model.
  - Deviation is significant between the observed data and fitted model.
- Outlier Test
  - This function tests if the number of observations outside the simulation envelope are larger or smaller than expected
  - Methods generate a null expectation, and then test for an excess or lack of outliers. Per default, testOutliers() looks for both, so if you get a significant p-value, you have to check if you have to many or too few outliers.

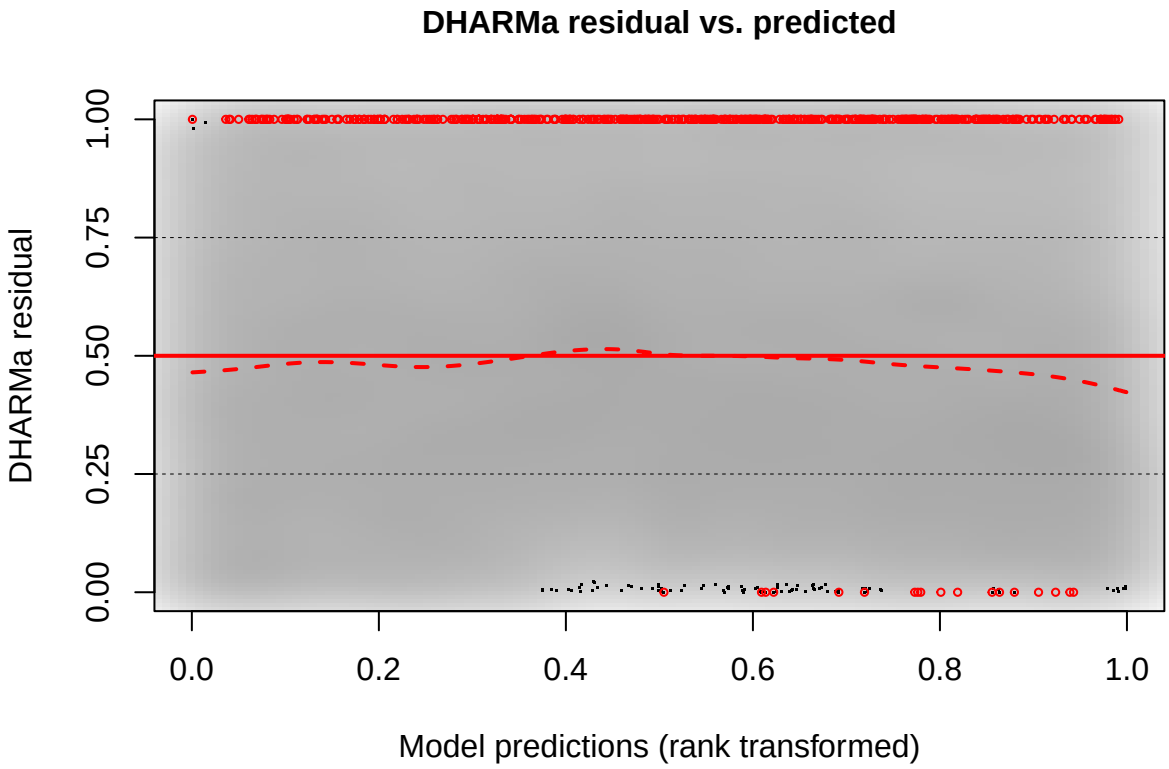
– See Outlier test for distribution of outliers. - Many at 1.0 residual.

```
plotQQunif(simulationOutput)
```

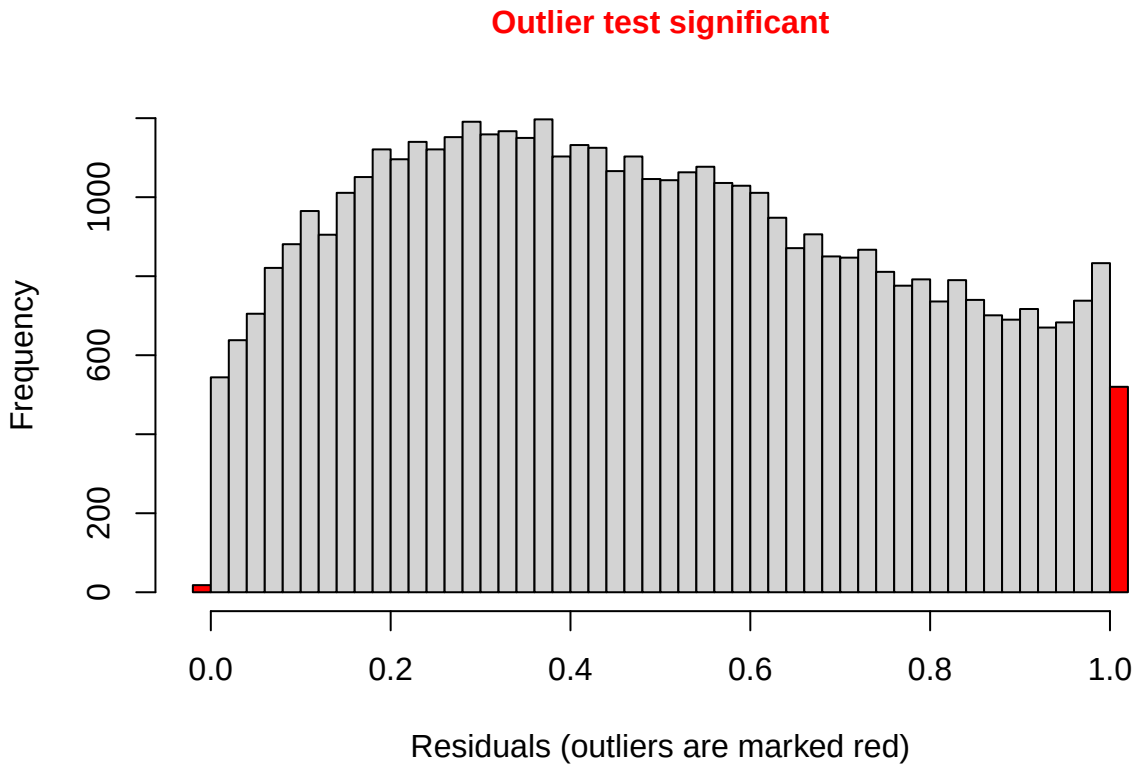
```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```




```
plotResiduals(simulationOutput)
```




```
testOutliers(simulationOutput, type = "bootstrap")
```




```
##
##  DHARMa bootstrapped outlier test
##
## data:  simulationOutput
## outliers at both margin(s) = 538, observations = 47353, p-value <
## 2.2e-16
## alternative hypothesis: two.sided
##  percent confidence interval:
##  0.006957848 0.008756045
## sample estimates:
## outlier frequency (expected: 0.00784047473232953 )
##                                         0.01136148
```