taxonomy richness

2025-11-19

Load data

```
laselva <- read_excel("laselva.xlsx", sheet = "alldata")
```

Convert Date to proper Date object

```
laselva$Date <- dmy(paste0("01-", laselva$Date))  # e.g., "Jan-1997" -> "1997-01-01"
```

Extract Year and Month index

```
laselva$Year <- year(laselva$Date)
laselva$Month_idx <- month(laselva$Date)
head(laselva,13)
```

```
## # A tibble: 13 x 13
##    stream Months Date       variable   value ONI_First   SOI  ENSO temp_min
##    <chr>   <dbl> <date>     <chr>      <dbl>     <dbl> <dbl> <dbl>    <dbl>
##  1 carapa      1 1997-01-01 total_rich    31     -0.36   0.8 -0.72     20.2
##  2 carapa      2 1997-02-01 total_rich    20     -0.1    2.9 -0.3      19.6
##  3 carapa      3 1997-03-01 total_rich    18      0.28  -0.7  0.15     21.1
##  4 carapa      4 1997-04-01 total_rich    31      0.75  -1    0.71     20.5
##  5 carapa      5 1997-05-01 total_rich    22      1.22  -2.2  2.34     21.6
##  6 carapa      6 1997-06-01 total_rich    23      1.6   -2.3  2.26     22.7
##  7 carapa      7 1997-07-01 total_rich    27      1.9   -1.2  2.26     22.6
##  8 carapa      8 1997-08-01 total_rich    23      2.14  -2.4  2.2      23.4
##  9 carapa      9 1997-09-01 total_rich    23      2.33  -2.4  2.06     22.8
## 10 carapa     10 1997-10-01 total_rich    22      2.4   -2.4  2.14     22.7
## 11 carapa     11 1997-11-01 total_rich    17      2.39  -2    2.11     22.9
## 12 carapa     12 1997-12-01 total_rich    17      2.24  -1.6  2.3      22.5
## 13 carapa     13 1998-01-01 total_rich    32      1.93  -4.4  2.45     20.8
## # i 4 more variables: rain_51 <dbl>, rain_1.4 <dbl>, Year <dbl>,
## #   Month_idx <dbl>
```

Subset months (Feb, May, Sep) and remove 2024

```
laselva <- laselva[laselva$Month_idx %in% c(2,5,9) &
                     laselva$Year != 2024, ]
head(laselva, 13)
```

```
## # A tibble: 13 x 13
##    stream Months Date       variable   value ONI_First   SOI  ENSO temp_min
##    <chr>   <dbl> <date>     <chr>      <dbl>     <dbl> <dbl> <dbl>    <dbl>
##  1 carapa      2 1997-02-01 total_rich    20     -0.1    2.9 -0.3      19.6
##  2 carapa      5 1997-05-01 total_rich    22      1.22  -2.2  2.34     21.6
```

```
##  3 carapa      9 1997-09-01 total_rich    23      2.33  -2.4  2.06     22.8
##  4 carapa     14 1998-02-01 total_rich    24      1.44  -3.4  2.27     21.1
##  5 carapa     17 1998-05-01 total_rich    18     -0.13   0.4  0.42     22.3
##  6 carapa     21 1998-09-01 total_rich    25     -1.35   1.7 -1.29     22.5
##  7 carapa     26 1999-02-01 total_rich    21     -1.07   1.6 -1.15     20.3
##  8 carapa     29 1999-05-01 total_rich    23     -1.04   0.4 -1.26     21.7
##  9 carapa     33 1999-09-01 total_rich    41     -1.26  -0.1 -1.33     22.0
## 10 carapa     38 2000-02-01 total_rich    21     -1.07   2.7 -1.37     20.1
## 11 carapa     41 2000-05-01 total_rich    27     -0.64   0.6 -1.18     21.3
## 12 carapa     45 2000-09-01 total_rich    15     -0.63   1.4 -0.52     22.5
## 13 carapa     50 2001-02-01 total_rich    12     -0.44   2.8 -0.78     19.5
## # i 4 more variables: rain_51 <dbl>, rain_1.4 <dbl>, Year <dbl>,
## #   Month_idx <dbl>
```

```
tail(laselva,13)
```

```
## # A tibble: 13 x 13
##    stream  Months Date       variable   value ONI_First   SOI  ENSO temp_min
##    <chr>    <dbl> <date>     <chr>      <dbl>     <dbl> <dbl> <dbl>    <dbl>
##  1 saltito    185 2012-05-01 total_rich     5     -0.01   0.1 -0.28     21.8
##  2 saltito    189 2012-09-01 total_rich    14      0.27   0.4 -0.19     22.3
##  3 saltito    194 2013-02-01 total_rich     5     -0.34  -0.4 -0.12     20.2
##  4 saltito    197 2013-05-01 total_rich     7     -0.41   1.3 -1.14     21.5
##  5 saltito    201 2013-09-01 total_rich    12     -0.18   0.5 -0.13     22.1
##  6 saltito    206 2014-02-01 total_rich     5     -0.27   0.1 -0.05     19.7
##  7 saltito    209 2014-05-01 total_rich     6      0.16   0.9  0        22.2
##  8 saltito    218 2015-02-01 total_rich     3      0.53   0.4  0.15     20.6
##  9 saltito    221 2015-05-01 total_rich     5      1.18  -1.2  1.9      22.7
## 10 saltito    225 2015-09-01 total_rich     3      2.42  -2.7  2.15     22.6
## 11 saltito    230 2016-02-01 total_rich     7      1.58  -3.2  1.32     20.3
## 12 saltito    233 2016-05-01 total_rich     4     -0.07   0.7  0.36     22.7
## 13 saltito    237 2016-09-01 total_rich     6     -0.69   2   -0.54     22.5
## # i 4 more variables: rain_51 <dbl>, rain_1.4 <dbl>, Year <dbl>,
## #   Month_idx <dbl>
```

Make stream a factor

```
laselva$stream <- factor(laselva$stream)
```

Create factor for AR1 (unique time points)

- Date (class Date) no funciona directamente en el AR1 de glmmTMB

```
laselva$time_f <- factor(laselva$Date,
                         levels = sort(unique(laselva$Date)))
head(laselva,13)
```

```
## # A tibble: 13 x 14
##    stream Months Date       variable   value ONI_First   SOI  ENSO temp_min
##    <fct>   <dbl> <date>     <chr>      <dbl>     <dbl> <dbl> <dbl>    <dbl>
##  1 carapa      2 1997-02-01 total_rich    20     -0.1    2.9 -0.3      19.6
##  2 carapa      5 1997-05-01 total_rich    22      1.22  -2.2  2.34     21.6
##  3 carapa      9 1997-09-01 total_rich    23      2.33  -2.4  2.06     22.8
##  4 carapa     14 1998-02-01 total_rich    24      1.44  -3.4  2.27     21.1
##  5 carapa     17 1998-05-01 total_rich    18     -0.13   0.4  0.42     22.3
```

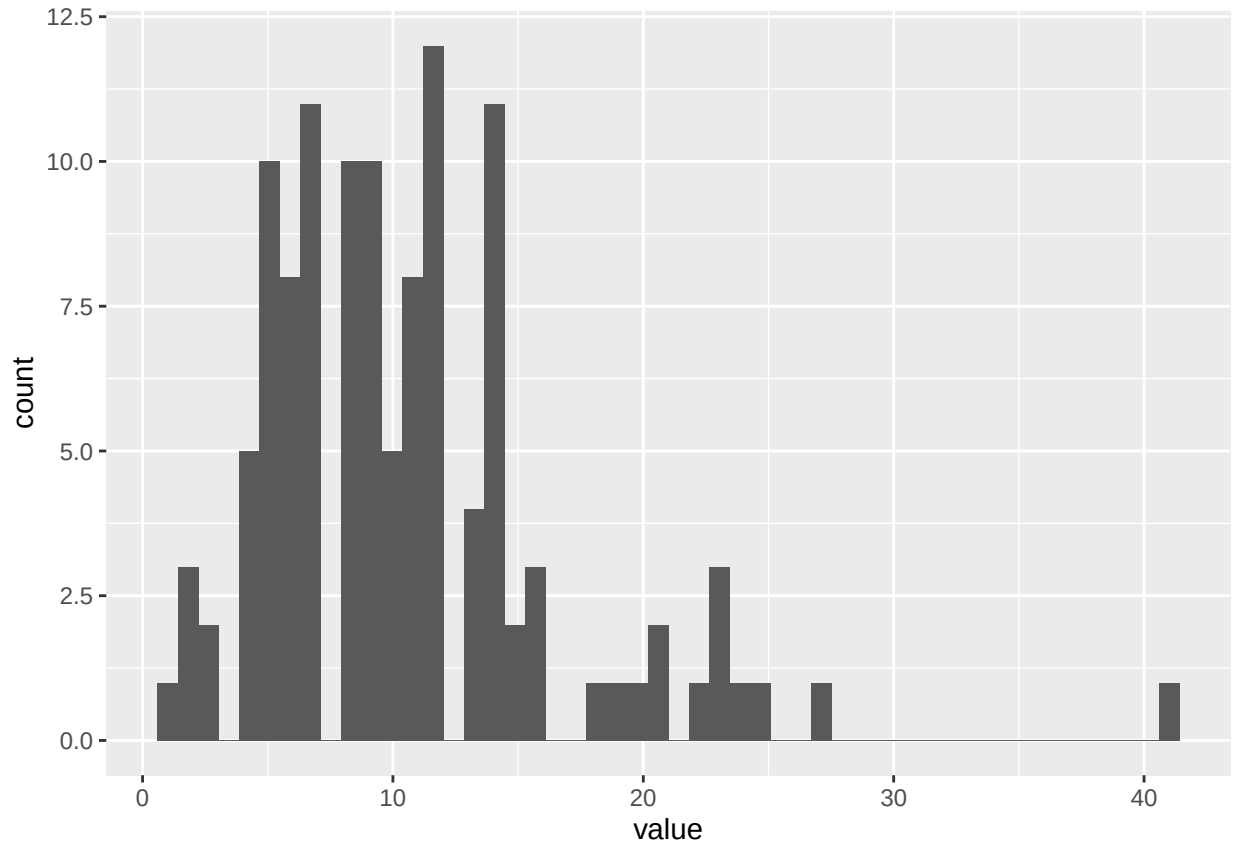
```
##  6 carapa     21 1998-09-01 total_rich    25     -1.35   1.7 -1.29     22.5
##  7 carapa     26 1999-02-01 total_rich    21     -1.07   1.6 -1.15     20.3
##  8 carapa     29 1999-05-01 total_rich    23     -1.04   0.4 -1.26     21.7
##  9 carapa     33 1999-09-01 total_rich    41     -1.26  -0.1 -1.33     22.0
## 10 carapa     38 2000-02-01 total_rich    21     -1.07   2.7 -1.37     20.1
## 11 carapa     41 2000-05-01 total_rich    27     -0.64   0.6 -1.18     21.3
## 12 carapa     45 2000-09-01 total_rich    15     -0.63   1.4 -0.52     22.5
## 13 carapa     50 2001-02-01 total_rich    12     -0.44   2.8 -0.78     19.5
## # i 5 more variables: rain_51 <dbl>, rain_1.4 <dbl>, Year <dbl>,
## #   Month_idx <dbl>, time_f <fct>
```

Quick summary

```
summary(laselva$value)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     7.0     9.0    10.6    13.0    41.0
```

```
ggplot(laselva, aes(value)) +
  geom_histogram(bins = 50)
```

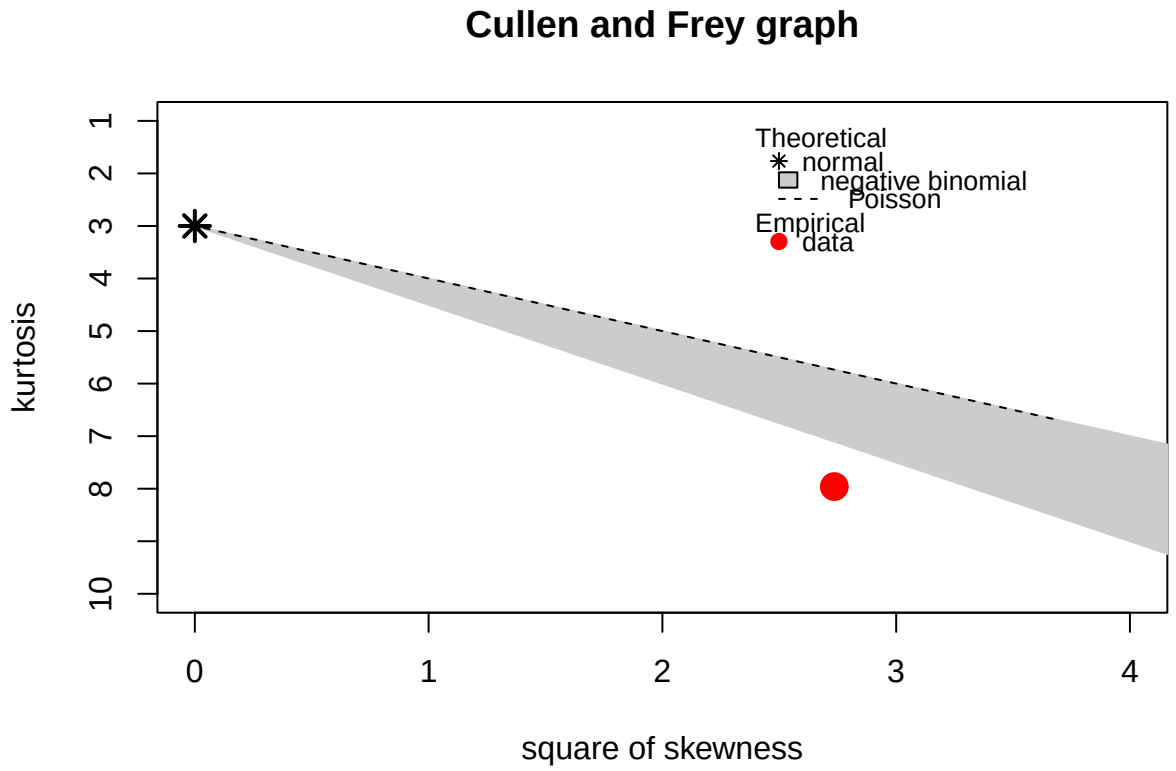


Distribution check

- We checked the distribution of the count variable (value) using fitdistrplus::descdist().
- The summary shows:
  - Mean = 10.6
  - Variance (SD^2) $6.04^2 = 36.48$

  - So variance » mean
- Since the variance was much larger than the mean (variance » mean), **the data violated the Poisson assumption of mean variance**. This level of overdispersion is typical of ecological count data, and **therefore a negative binomial** model provides a more appropriate and flexible error structure for the analysis.
- https://cran.r-project.org/web/packages/GlmSimulatoR/vignettes/count_data_and_overdispersion.html
- https://pacificpapermill.com/slideshows/PPM___Poisson_Regression.pdf

```
fitdistrplus::descdist(laselva$value, discrete = TRUE)
```




```
## summary statistics
## ------
## min:  1   max:  41
## median:  9
## mean:  10.60169
## estimated sd:  6.041367
## estimated skewness:  1.653916
## estimated kurtosis:  7.959617
```

Fit negative binomial GLMM

- We use a negative binomial distribution because 'value' is **overdispersed count data**.
  - (variance » mean, assessed with fitdistrplus::descdist()).
- glmmTMB is ideal because it accommodates overdispersion, random effects, and correlation structures.

```r
laselva <- laselva %>%
  group_by(stream) %>%
  arrange(Months, .by_group = TRUE) %>%
  mutate(time_factor = factor(row_number())) %>%
  ungroup()

# Ajustar el modelo AR1 usando time_factor
mod1 <- glmmTMB(
  value ~ ONI_First + SOI + ENSO + temp_min + rain_51 + stream +
    ar1(time_factor + 0 | stream),  # AR1 por stream
  family = nbinom2,
  data = laselva
)

summary(mod1)
```

```
##  Family: nbinom2  ( log )
## Formula:
## value ~ ONI_First + SOI + ENSO + temp_min + rain_51 + stream +
##     ar1(time_factor + 0 | stream)
## Data: laselva
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##     655.0     682.7    -317.5     635.0       108
##
## Random effects:
##
## Conditional model:
##  Groups Name         Variance Std.Dev. Corr
##  stream time_factor1 0.2374   0.4872   0.96 (ar1)
## Number of obs: 118, groups:  stream, 2
##
## Dispersion parameter for nbinom2 family (): 5.7e+08
##
## Conditional model:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.63522    0.76634   3.439 0.000584 ***
## ONI_First      0.04275    0.10472   0.408 0.683066
## SOI           -0.03227    0.02875  -1.122 0.261717
## ENSO          -0.09092    0.10314  -0.881 0.378072
## temp_min      -0.01040    0.03223  -0.323 0.747013
## rain_51       -0.02336    0.02126  -1.099 0.271764
## streamsaltito -0.10427    0.47642  -0.219 0.826752
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results and limitations of the AR1 model

- The AR1 model fits but shows an unrealistically large dispersion parameter (5.7×10), and **AR1 assumptions are questionable because sampling was quarterly and not equally spaced across the year**.

Why AR1 is not ideal here?

- Temporal spacing is irregular:
  - Feb → May → Sep → (next Feb)
  - **AR1 assumes constant t (equally spaced time steps). When spacing is uneven, AR1 overestimates correlation and produces unstable estimates (REFERENCE).**
- Only three sampling points per year.
  - AR1 needs longer, regularly sampled time series to be statistically meaningful.
- Each year begins again at "time = 1" This breaks the AR1 chain, making the correlation structure artificial. AR1 assumes uniform time steps and a continuous series, which is not the case here.

More appropriate alternative: Random intercept for Year

```
mod2 <- glmmTMB(
  value ~ ONI_First + SOI + ENSO + temp_min + rain_51 + stream +
    (1 | Year),  # random intercept for Year
  family = nbinom2,
  data = laselva
)

summary(mod2)
```

```
##  Family: nbinom2  ( log )
## Formula:
## value ~ ONI_First + SOI + ENSO + temp_min + rain_51 + stream +      (1 | Year)
## Data: laselva
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##     697.1     722.1    -339.6     679.1       109
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  Year   (Intercept) 0.1448   0.3806
## Number of obs: 118, groups:  Year, 20
##
## Dispersion parameter for nbinom2 family (): 21.8
##
## Conditional model:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.75566    0.81828   3.368 0.000758 ***
## ONI_First     -0.05673    0.13726  -0.413 0.679381
## SOI           -0.04780    0.03880  -1.232 0.217946
## ENSO          -0.01627    0.12659  -0.129 0.897717
## temp_min      -0.01874    0.03813  -0.492 0.623032
## rain_51       -0.01892    0.02759  -0.686 0.492801
## streamsaltito -0.08103    0.07087  -1.143 0.252870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Why model2 is more appropriate

1) Accounts for repeated measures / grouped structure

- Multiple observations occur within each year, and those observations are not independent.

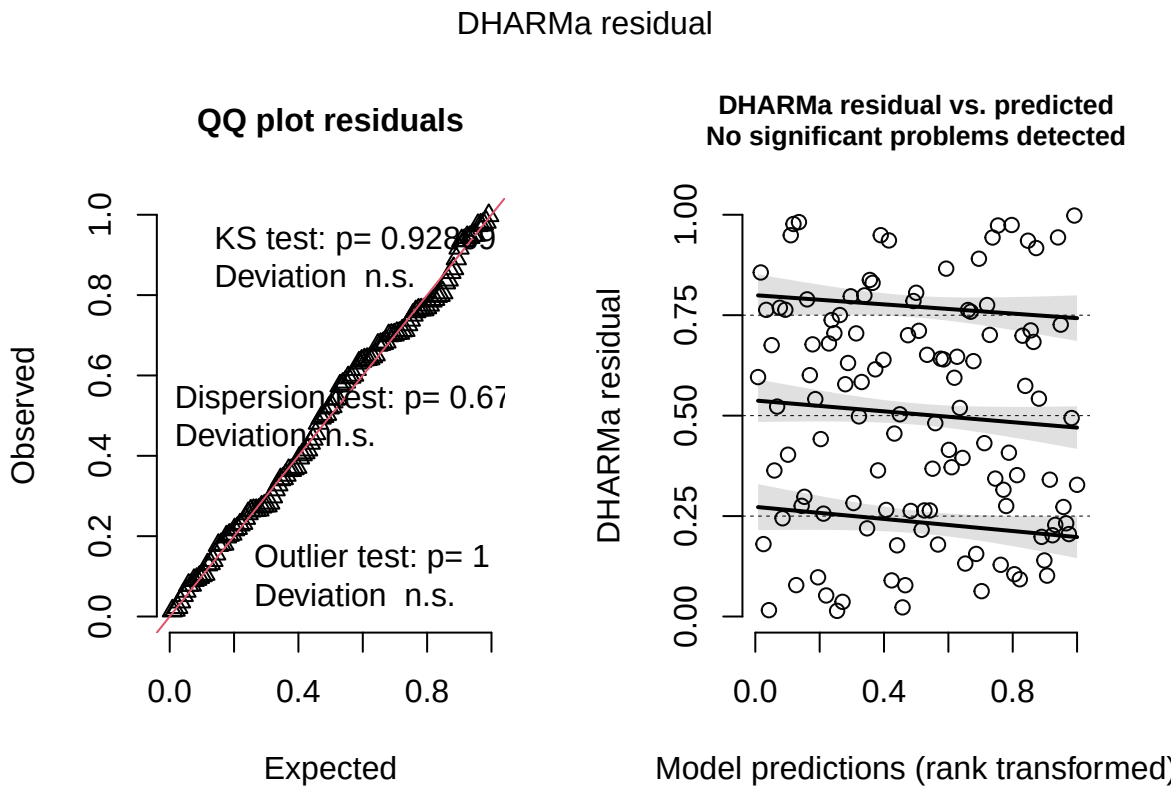
2) Captures unmeasured annual variation
   - Each year may differ due to unrecorded events (storms, droughts).
   - The random intercept lets each year shift the baseline taxon_rich/abundance/…
3) Controls temporal dependence indirectly
   - **Quarterly sampling is not evenly spaced, so classical AR1 is inappropriate.**
   - A random intercept groups observations within each year so that measurements taken in the same year are more similar to each other than to those from other years, partially accounting for temporal autocorrelation without misusing an AR1 structure.

Residuals diagnostics

```
res <- simulateResiduals(mod2, n = 1000)
```

plots Residuals

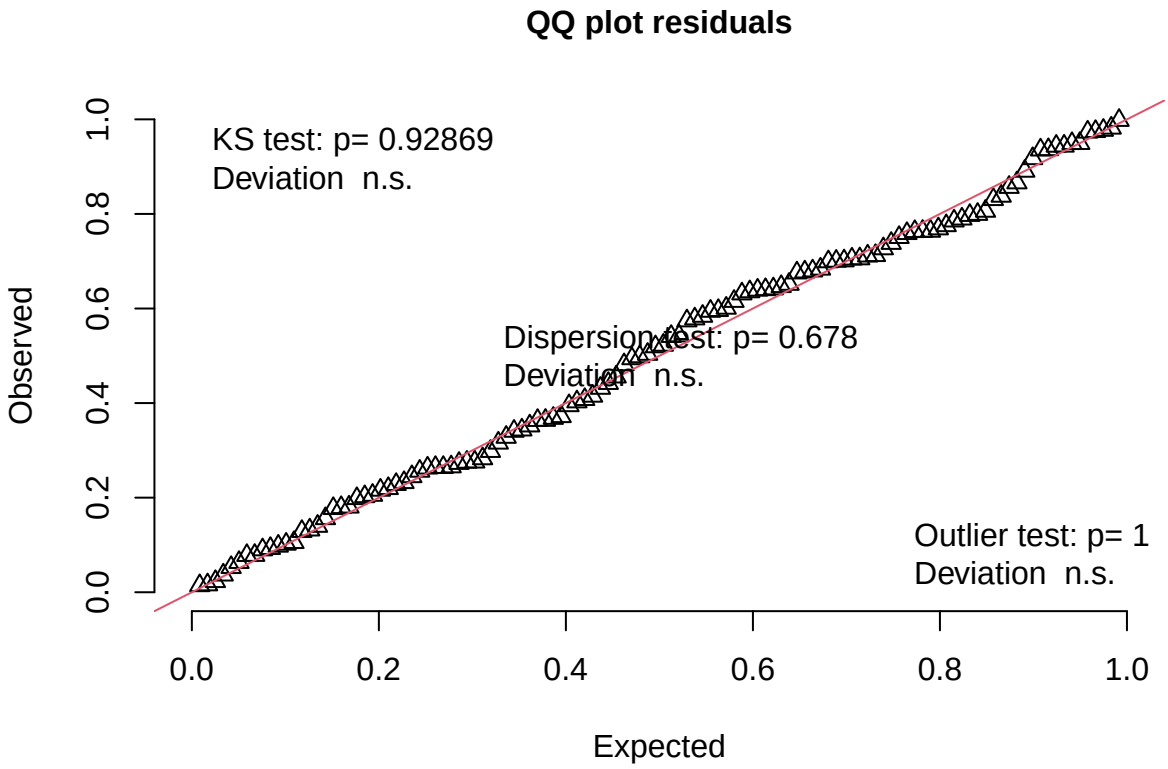
```
plot(res)
```



Uniformity test

- The Kolmogorov–Smirnov test on the DHARMa scaled residuals shows no deviation from the expected uniform distribution (D = 0.050, p = 0.929). **This indicates that the model's residuals are well behaved and that there is no evidence of major distributional misspecification**.

```
testUniformity(res)
```

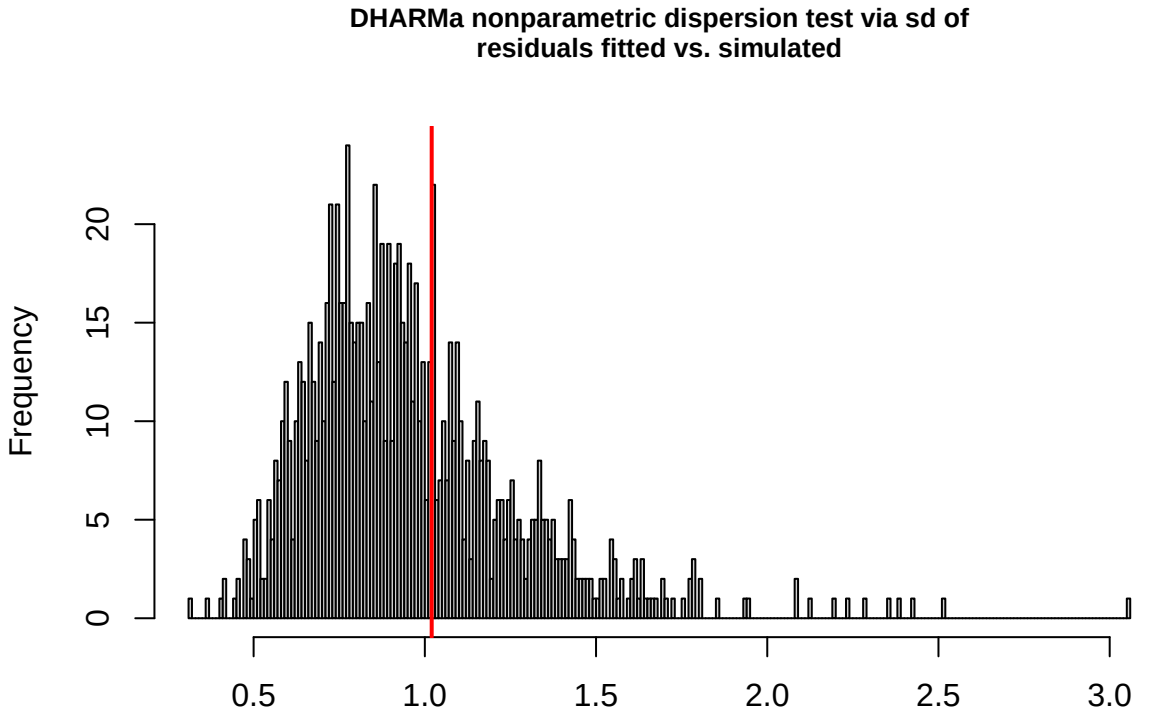



```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.050081, p-value = 0.9287
## alternative hypothesis: two-sided
```

Dispersion test

- The DHARMa dispersion test indicates that the model does not show evidence of overdispersion. The dispersion value is close to 1 (1.0662) and the high p-value (0.678) means we cannot reject the null hypothesis of correct dispersion. This suggests that the negative binomial GLMM adequately captures the variance structure of the count data.

```
testDispersion(res)
```




```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.0662, p-value = 0.678
## alternative hypothesis: two.sided
```

Model with <

```
mod3 <- glmmTMB(
  value ~ ONI_First + SOI + ENSO + temp_min + rain_1.4 + stream +
    (1 | Year),  # random intercept for Year
  family = nbinom2,
  data = laselva
)

summary(mod3)
```

```
##  Family: nbinom2  ( log )
## Formula:
## value ~ ONI_First + SOI + ENSO + temp_min + rain_1.4 + stream +      (1 | Year)
## Data: laselva
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##     696.8     721.7    -339.4     678.8       109
```

```
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  Year   (Intercept) 0.145    0.3808
## Number of obs: 118, groups:  Year, 20
##
## Dispersion parameter for nbinom2 family ():   22
##
## Conditional model:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.559952   0.869982   2.942  0.00326 **
## ONI_First     -0.047061   0.138218  -0.340  0.73349
## SOI           -0.051418   0.039067  -1.316  0.18812
## ENSO          -0.032662   0.129020  -0.253  0.80015
## temp_min      -0.016909   0.038172  -0.443  0.65778
## rain_1.4       0.009279   0.010513   0.883  0.37747
## streamsaltito -0.081169   0.070755  -1.147  0.25131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```