

Statistical analysis of the effect of environmental variables on abundance of flounder

Olga Lyashevskaya

2024-05-19

Contents

Data preparation and exploration	2
Distribution of nflounder	3
Correlation analysis	6
Zero-inflation	8
Scale variables	8
Data modelling	9
Negative binomial GLM	9
Negative Binomial GLMM	10
Zero-inflated poisson GLMM	14
Negative binomial GLMM (Hurdle Model).	15
Post-model-fitting procedure.	18

```
# load packages
packages <- c("ggplot2", "MASS", "rmarkdown", "tinytex", "reshape2", "glmmTMB", "DHARMA")
lapply(packages, library, character.only = TRUE)
```

```
## This is DHARMA 0.4.6. For overview type '?DHARMA'. For recent changes, type news(package = 'DHARMA')

## [[1]]
## [1] "ggplot2"      "stats"      "graphics"   "grDevices" "utils"      "datasets"
## [7] "methods"     "base"
##
## [[2]]
## [1] "MASS"        "ggplot2"    "stats"      "graphics"   "grDevices" "utils"
## [7] "datasets"    "methods"    "base"
##
## [[3]]
## [1] "rmarkdown" "MASS"      "ggplot2"    "stats"      "graphics"   "grDevices"
## [7] "utils"      "datasets"  "methods"    "base"
##
## [[4]]
```

```
## [1] "tinytex" "rmarkdown" "MASS" "ggplot2" "stats" "graphics"
## [7] "grDevices" "utils" "datasets" "methods" "base"
##
## [[5]]
## [1] "reshape2" "tinytex" "rmarkdown" "MASS" "ggplot2" "stats"
## [7] "graphics" "grDevices" "utils" "datasets" "methods" "base"
##
## [[6]]
## [1] "glmmTMB" "reshape2" "tinytex" "rmarkdown" "MASS" "ggplot2"
## [7] "stats" "graphics" "grDevices" "utils" "datasets" "methods"
## [13] "base"
##
## [[7]]
## [1] "DHARMa" "glmmTMB" "reshape2" "tinytex" "rmarkdown" "MASS"
## [7] "ggplot2" "stats" "graphics" "grDevices" "utils" "datasets"
## [13] "methods" "base"
```

```
knitr::opts_chunk$set(fig.path = "figure/", dev = "png")
set.seed(123)
```

Data preparation and exploration

```
# load data
df <- read.csv("data.csv")
# describe data
colnames(df)
```

```
## [1] "site" "net" "year" "lat" "long"
## [6] "distshore" "trawl" "area" "chlorophyll" "tempavg"
## [11] "tempstddev" "sal" "bod" "nh3" "po4"
## [16] "depth" "nflounder"
```

```
dim(df)
```

```
## [1] 2763 17
```

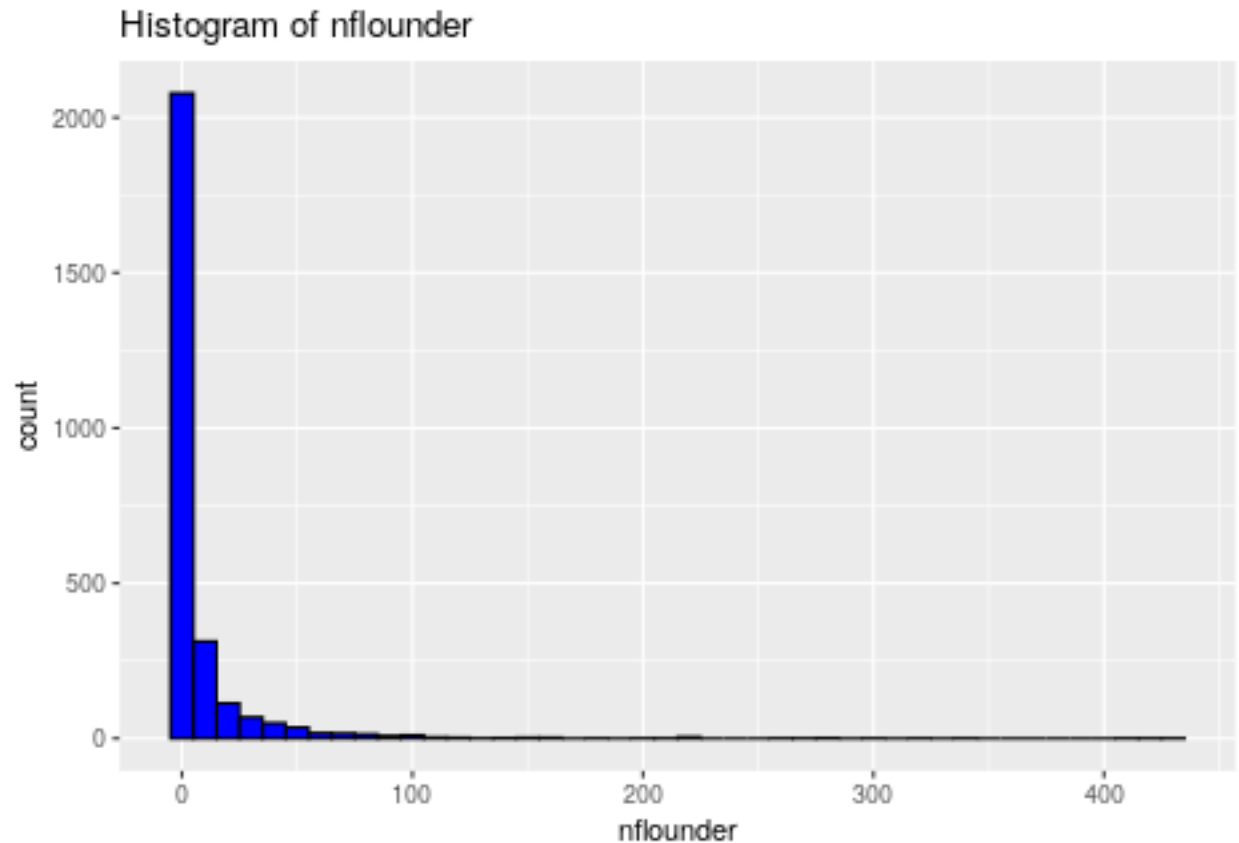
```
df[c("net", "site")]<-lapply(df[c("net", "site")], factor)
summary(df)
```

```
##
##          site          net          year
## Suir Estuary          : 183 BS :1264 Min.   :2001
## Shannon Estuary, Lower : 163 BT : 672 1st Qu.:2008
## Boyne                  : 154 Fyke: 827 Median :2010
## Barrow Suir Nore Estuary : 144          Mean  :2011
## Gweebarra Estuary       : 143          3rd Qu.:2015
## Barrow Nore Suir Estuary, Upper: 106          Max.   :2019
## (Other)                  :1870
##          lat          long          distshore          trawl
## Min.   :51.48 Min.   : -9.966 Min.   : 0.00 Min.   : 0.00
```

```
## 1st Qu.:52.28 1st Qu.: -9.074 1st Qu.: 13.90 1st Qu.: 0.00
## Median :52.66 Median : -8.252 Median : 45.71 Median : 0.00
## Mean :52.98 Mean : -8.025 Mean : 171.84 Mean : 32.82
## 3rd Qu.:53.72 3rd Qu.: -6.956 3rd Qu.: 168.15 3rd Qu.: 0.00
## Max. :55.09 Max. : -6.033 Max. : 3097.40 Max. : 1210.00
##
## area chlorophyll tempavg tempstdev
## Min. : 0.0832 Min. : 1.50 Min. : 7.305 Min. : 0.04534
## 1st Qu.: 3.0464 1st Qu.: 7.40 1st Qu.: 12.773 1st Qu.: 3.21952
## Median : 6.7854 Median : 18.00 Median : 13.558 Median : 3.90394
## Mean : 25.8178 Mean : 37.57 Mean : 13.480 Mean : 3.75874
## 3rd Qu.: 12.2295 3rd Qu.: 50.30 3rd Qu.: 14.455 3rd Qu.: 4.54526
## Max. : 489.4254 Max. : 444.00 Max. : 18.691 Max. : 7.04075
##
## sal bod nh3 po4
## Min. : 4.878 Min. : 0.688 Min. : 0.01500 Min. : 7.909
## 1st Qu.: 7.840 1st Qu.: 1.149 1st Qu.: 0.04100 1st Qu.: 15.595
## Median : 15.609 Median : 1.529 Median : 0.04600 Median : 31.276
## Mean : 15.511 Mean : 1.522 Mean : 0.06381 Mean : 28.421
## 3rd Qu.: 22.959 3rd Qu.: 1.629 3rd Qu.: 0.07000 3rd Qu.: 38.396
## Max. : 33.047 Max. : 3.825 Max. : 0.17300 Max. : 83.600
##
## depth nflounder
## Min. : 0.700 Min. : 0.000
## 1st Qu.: 2.500 1st Qu.: 0.000
## Median : 4.030 Median : 1.000
## Mean : 4.378 Mean : 9.205
## 3rd Qu.: 6.170 3rd Qu.: 5.000
## Max. : 8.400 Max. : 435.000
##
```

Distribution of nflounder

```
ggplot(df, aes(nflounder)) +
  geom_histogram(binwidth = 10, fill = "blue", color = "black") +
  labs(title = "Histogram of nflounder", x = "nflounder")
```



See how many values fall in each category:

```
# Define the bin width
bin_width <- 10

# Define the breaks for the bins
breaks <- seq(min(df$nflounder), max(df$nflounder), by = bin_width)

# Divide the data into bins
bins <- cut(df$nflounder, breaks = breaks, include.lowest = TRUE, right = FALSE)

# Count the number of values in each bin
bin_counts <- table(bins)

# Print the bin counts
print(bin_counts)
```

```
## bins
##      [0,10)  [10,20)  [20,30)  [30,40)  [40,50)  [50,60)  [60,70)  [70,80)
##      2245      205      94        57        44        20        19        16
##      [80,90) [90,100) [100,110) [110,120) [120,130) [130,140) [140,150) [150,160)
##          10         13          7          5          1          2          2          3
## [160,170) [170,180) [180,190) [190,200) [200,210) [210,220) [220,230) [230,240)
##          2          1          0          0          2          1          4          0
## [240,250) [250,260) [260,270) [270,280) [280,290) [290,300) [300,310) [310,320)
##          0          1          0          1          2          0          1          0
## [320,330) [330,340) [340,350) [350,360) [360,370) [370,380) [380,390) [390,400)
```

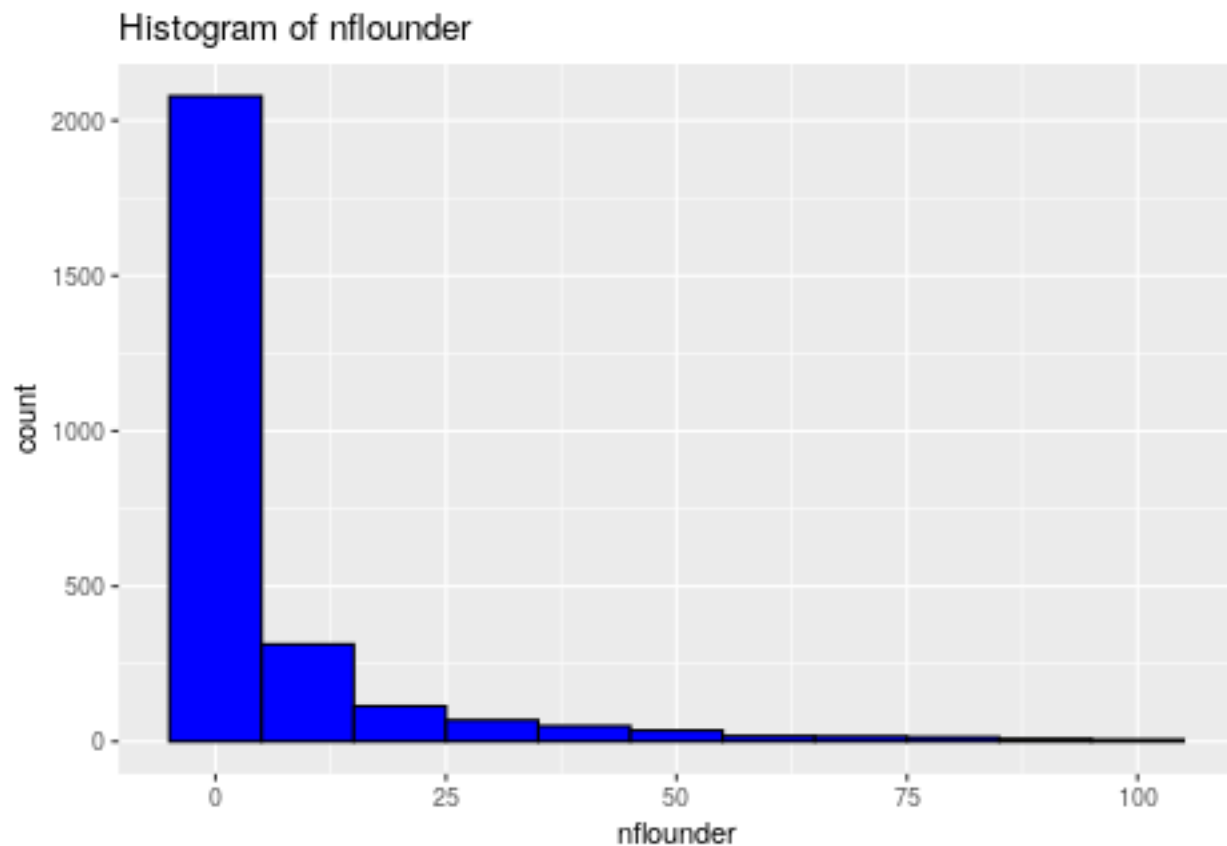
```
##          1          1          0          0          0          0          0          0
## [400,410) [410,420) [420,430]
##          0          2          0
```

Lets truncate values above 100 for modelling convenience.

```
original_nrow <- nrow(df)
df <- subset(df, nflounder <= 100)
removed_nrow <- original_nrow - nrow(df)
conditional_var <- var(df$nflounder, na.rm=TRUE)
conditional_mean <- mean(df$nflounder, na.rm=TRUE)
```

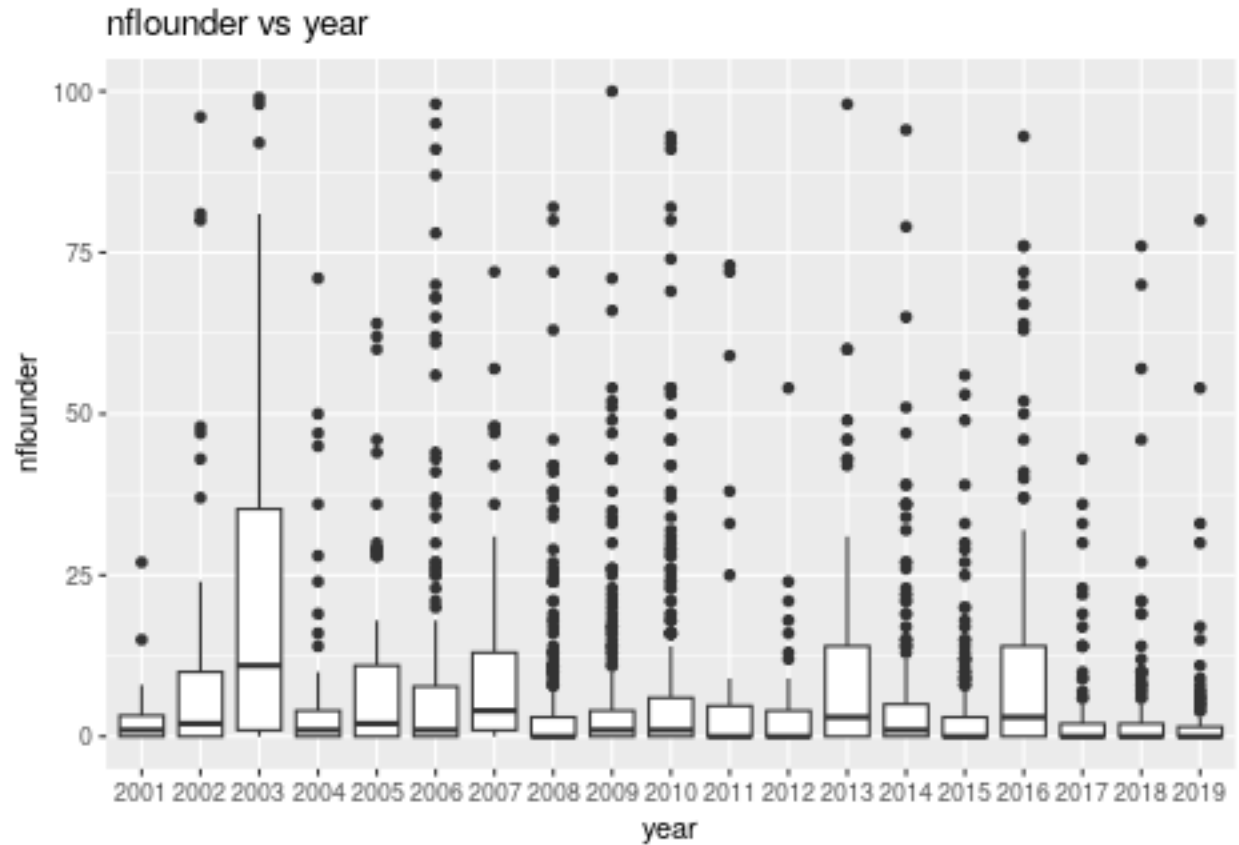
We removed 39 from 2763. Let's visualise distribution of nflounder again.

```
ggplot(df, aes(nflounder)) +
  geom_histogram(binwidth = 10, fill = "blue", color = "black") +
  labs(title = "Histogram of nflounder", x = "nflounder")
```



As we can see data is still highly overdispersed, the conditional variance (208.0536609) exceeds the conditional mean (6.5179883). In situations like this negative binomial is an appropriate distribution to use.

```
ggplot(df, aes(x = factor(year), y = nflounder)) +
  geom_boxplot() +
  # scale_y_log10() +
  labs(x = "year", y = "nflounder", title = "nflounder vs year")
```



Correlation analysis

```
df_numeric <- df[sapply(df, is.numeric)]
cor_matrix <- cor(df_numeric, use = "complete.obs")
print(cor_matrix)
```

```
##          year          lat          long  distshore      trawl
## year      1.000000000  0.0049817860 -0.06511398  0.088010356 -0.119769828
## lat       0.004981786  1.0000000000  0.04506691 -0.005444478  0.093568489
## long     -0.065113977  0.0450669064  1.000000000 -0.224120769  0.043158598
## distshore 0.088010356 -0.0054444775 -0.22412077  1.000000000  0.144581861
## trawl     -0.119769828  0.0935684891  0.04315860  0.144581861  1.000000000
## area      0.067678064 -0.0007925651 -0.10255599  0.164351090 -0.028732391
## chlorophyll 0.070847279 -0.0322708516  0.14170391  0.020114161  0.003544562
## tempavg    0.093554648 -0.2230929909 -0.04791776  0.084518093 -0.022062488
## tempstdev -0.156715923 -0.0234881935 -0.05046801  0.096050784  0.025851643
## sal       -0.204827187  0.3989884868 -0.12004630 -0.050253948  0.029244586
## bod       -0.125186675 -0.3154381712  0.25662766 -0.113689564  0.051555411
## nh3       -0.269058242 -0.1454581609  0.24194535 -0.078002487  0.013807491
## po4       -0.159703341 -0.1719509276  0.50093553  0.038484470 -0.036196786
## depth     0.118780600 -0.3806972918  0.20077301  0.045172606 -0.005677570
## nflounder -0.096151776 -0.1291115877  0.18078237 -0.132022226 -0.002051093
##          area  chlorophyll  tempavg  tempstdev      sal
```

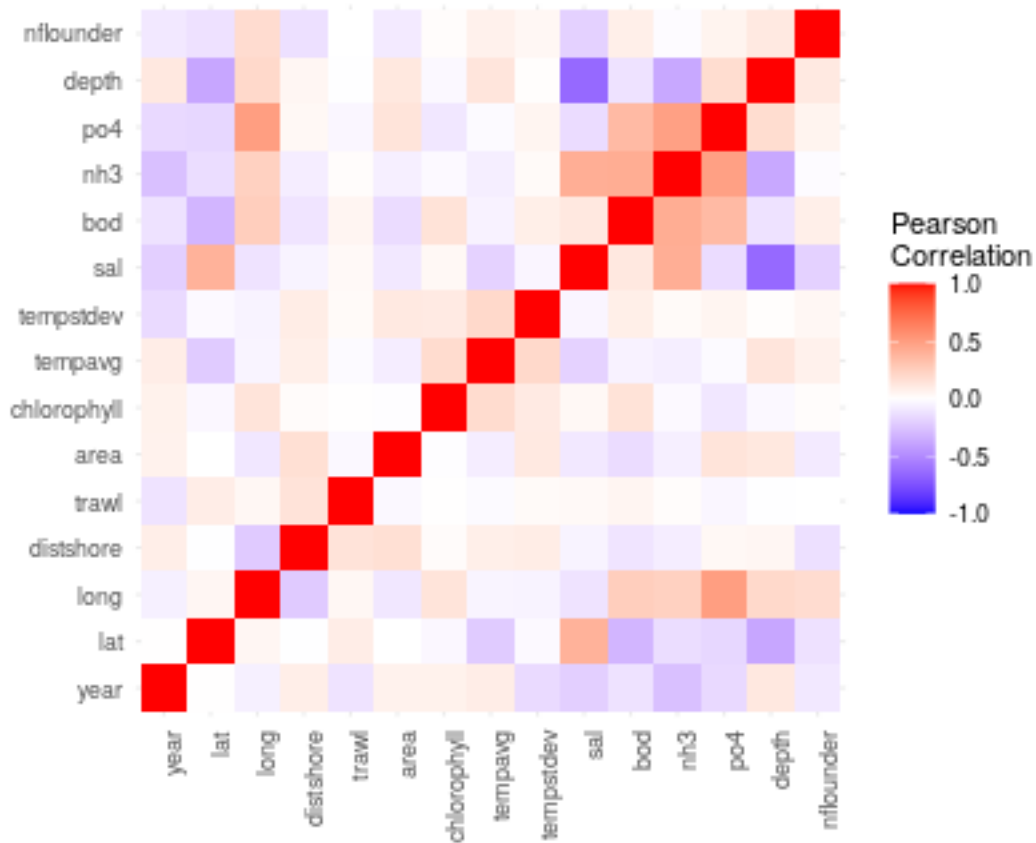
```
## year      0.0676780642  0.070847279  0.09355465 -0.15671592 -0.20482719
## lat       -0.0007925651 -0.032270852 -0.22309299 -0.02348819  0.39898849
## long      -0.1025559916  0.141703905 -0.04791776 -0.05046801 -0.12004630
## distshore  0.1643510905  0.020114161  0.08451809  0.09605078 -0.05025395
## trawl     -0.0287323909  0.003544562 -0.02206249  0.02585164  0.02924459
## area      1.0000000000 -0.007407639 -0.07788099  0.11644657 -0.09950945
## chlorophyll -0.0074076388  1.000000000  0.18577262  0.10878705  0.03726607
## tempavg   -0.0778809900  0.185772616  1.00000000  0.20038897 -0.19076514
## tempstdev  0.1164465703  0.108787049  0.20038897  1.00000000 -0.04111293
## sal       -0.0995094464  0.037266065 -0.19076514 -0.04111293  1.00000000
## bod       -0.1491648863  0.148704747 -0.05496218  0.08648001  0.12076372
## nh3       -0.0667732092 -0.026391337 -0.07254021  0.02594419  0.41822587
## po4       0.1422456757 -0.101320955 -0.02263963  0.05345896 -0.14777738
## depth     0.1227755778 -0.028296778  0.13414873  0.01176289 -0.64814422
## nflounder -0.0897024377  0.015470871  0.07377867  0.04200609 -0.19505829
##           bod      nh3      po4      depth      nflounder
## year      -0.12518667 -0.26905824 -0.15970334  0.11878060 -0.096151776
## lat       -0.31543817 -0.14545816 -0.17195093 -0.38069729 -0.129111588
## long      0.25662766  0.24194535  0.50093553  0.20077301  0.180782366
## distshore -0.11368956 -0.07800249  0.03848447  0.04517261 -0.132022226
## trawl     0.05155541  0.01380749 -0.03619679 -0.00567757 -0.002051093
## area     -0.14916489 -0.06677321  0.14224568  0.12277558 -0.089702438
## chlorophyll 0.14870475 -0.02639134 -0.10132096 -0.02829678  0.015470871
## tempavg   -0.05496218 -0.07254021 -0.02263963  0.13414873  0.073778667
## tempstdev  0.08648001  0.02594419  0.05345896  0.01176289  0.042006088
## sal       0.12076372  0.41822587 -0.14777738 -0.64814422 -0.195058290
## bod       1.00000000  0.42517649  0.35803600 -0.12377171  0.085604235
## nh3       0.42517649  1.00000000  0.48816749 -0.37503797 -0.015704750
## po4       0.35803600  0.48816749  1.00000000  0.18052167  0.059434892
## depth    -0.12377171 -0.37503797  0.18052167  1.00000000  0.113621304
## nflounder 0.08560423 -0.01570475  0.05943489  0.11362130  1.000000000
```

```
# Melt the correlation matrix into a long format
```

```
cor_matrix_melted <- melt(cor_matrix)
```

```
# Plot the heatmap using ggplot2
```

```
ggplot(data = cor_matrix_melted, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
    midpoint = 0, limit = c(-1,1), space = "Lab",
    name="Pearson\nCorrelation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 1,
    size = 9, hjust = 1),
    axis.title.x = element_blank(), # Remove x-axis title
    axis.title.y = element_blank(),) +
  coord_fixed()
```



Zero-inflation

```
# Calculate the proportion of values that are 0
zero_proportion <- mean(df$nflounder == 0)
```

The proportion of zeros in nflounder is 0.4533774.

Scale variables

We shall scale some of the variables to avoid numerical overflow.

```
var_to_scale <- c("distshore", "area", "chlorophyll")
df[, var_to_scale] <- scale(df[, var_to_scale])
```

Consider discarding some of the variables, e.g. trawl, in which the proportion of 0's (should be NA's?) 0.8509545.

```
df <- subset(df, select = -trawl)
```


Data modelling

Negative binomial GLM

We fit a negative binomial generalized linear model with log link to the full dataset, excluding site, trawl and net for the moment. This model is useful for count data that exhibit overdispersion (the var exceeds the mean). Let Y_i denote the count response variable for the i -th observation. The negative binomial distribution for Y_i is parameterised by the mean λ_i and the dispersion parameter θ :

$$Y_i \sim \text{NB}(\lambda_i, \theta)$$

where the probability mass function is given by:

$$P(Y_i = k) = \binom{k + \theta - 1}{k} \left(\frac{\theta}{\theta + \lambda_i} \right)^\theta \left(\frac{\lambda_i}{\theta + \lambda_i} \right)^k, \quad k = 0, 1, 2, \dots$$

Log Link Function

The relationship between the mean λ_i and the explanatory variables \mathbf{X}_i is modeled using a log link function:

$$\log(\lambda_i) = \mathbf{X}_i \boldsymbol{\beta}$$

where:

- λ_i is the expected count for the i -th observation.
- \mathbf{X}_i is a vector of explanatory variables for the i -th observation.
- $\boldsymbol{\beta}$ is a vector of coefficients to be estimated.

Linear Predictor

The linear predictor is given:

$$\eta_i = \mathbf{X}_i \boldsymbol{\beta}$$

where $\eta_i = \log(\lambda_i)$. Therefore, the model can be rewritten as:

$$\lambda_i = \exp(\mathbf{X}_i \boldsymbol{\beta})$$

```
m.glmm.fixed <- glmmTMB(
  nflounder ~ year + lat + long + distshore + area + chlorophyll + tempavg + tempstdev + sal + bod + nh3 + po4 + depth,
  data = df,
  family = nbinom2()
)
summary(m.glmm.fixed)

## Family: nbinom2 ( log )
## Formula:
## nflounder ~ year + lat + long + distshore + area + chlorophyll +
##      tempavg + tempstdev + sal + bod + nh3 + po4 + depth
## Data: df
```

```
##
##      AIC      BIC   logLik deviance df.resid
## 12986.0 13074.7 -6478.0 12956.0      2709
##
##
## Dispersion parameter for nbinom2 family (): 0.275
##
## Conditional model:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) 142.085983  18.135502   7.835 4.70e-15 ***
## year        -0.067561   0.008911  -7.582 3.42e-14 ***
## lat         -0.031889   0.069301  -0.460  0.64541
## long         0.264338   0.045712   5.783 7.35e-09 ***
## distshore   -0.337031   0.039239  -8.589 < 2e-16 ***
## area        -0.086590   0.034720  -2.494  0.01263 *
## chlorophyll -0.049502   0.043955  -1.126  0.26008
## tempavg      0.041370   0.024032   1.721  0.08516 .
## tempstdev    0.071309   0.031546   2.260  0.02379 *
## sal         -0.095546   0.010504  -9.096 < 2e-16 ***
## bod          0.075791   0.079467   0.954  0.34021
## nh3          4.690307   2.031163   2.309  0.02093 *
## po4         -0.009337   0.004179  -2.234  0.02546 *
## depth       -0.078342   0.028648  -2.735  0.00624 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Negative Binomial GLMM

Lets add random effect of site and net to the best model. This is an additional term that accounts for unobserved heterogeneity in the data. Random effects are used when data are collected in clusters or groups, and there may be variability between these groups that is not captured by the observed explanatory variables.

Random Effects

The random effects u_j are assumed to follow a normal distribution with mean zero and variance σ_u^2 :

$$u_j \sim \mathcal{N}(0, \sigma_u^2)$$

Linear Predictor

The linear predictor is given by:

$$\eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + u_j$$

where $\eta_{ij} = \log(\lambda_{ij})$. Therefore, the model can be rewritten as:

$$\lambda_{ij} = \exp(\mathbf{X}_{ij}\boldsymbol{\beta} + u_j)$$

```
# a model of crossed random effects (site and net)
m.glmm.random1 <- glmmTMB(
  nflounder ~ year + lat + long + distshore + area + chlorophyll + tempavg + tempstdev + sal + bod + nh3 + po4 + depth + (1 | site) + (1 | net),
  data = df,
  family = nbinom2()
)
summary(m.glmm.random1)
```

```
## Family: nbinom2 ( log )
## Formula:
## nflounder ~ year + lat + long + distshore + area + chlorophyll +
##      tempavg + tempstdev + sal + bod + nh3 + po4 + depth + (1 |
##      site) + (1 | net)
## Data: df
##
##      AIC      BIC   logLik deviance df.resid
## 12633.4 12733.9 -6299.7 12599.4      2707
##
## Random effects:
##
## Conditional model:
## Groups Name      Variance Std.Dev.
## site  (Intercept) 1.086e+00 1.0421893
## net   (Intercept) 1.813e-08 0.0001346
## Number of obs: 2724, groups: site, 74; net, 3
##
## Dispersion parameter for nbinom2 family (): 0.356
##
## Conditional model:
##      Estimate Std. Error z value Pr(>|z|)
## (Intercept) 146.856572  22.283365   6.590 4.39e-11 ***
## year        -0.079193   0.009943  -7.965 1.66e-15 ***
## lat          0.297084   0.177035   1.678  0.0933 .
## long         0.268307   0.133346   2.012  0.0442 *
## distshore    -0.340952   0.052580  -6.484 8.91e-11 ***
## area         0.010225   0.045513   0.225  0.8222
## chlorophyll  -0.117903   0.045901  -2.569  0.0102 *
## tempavg      0.054199   0.025686   2.110  0.0349 *
## tempstdev    0.043631   0.036127   1.208  0.2272
## sal          -0.092916   0.021349  -4.352 1.35e-05 ***
## bod          0.462236   0.242099   1.909  0.0562 .
## nh3          -4.194333   4.324354  -0.970  0.3321
## po4          0.014944   0.011991   1.246  0.2127
## depth        -0.083901   0.084408  -0.994  0.3202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# alternatively we consider a model of nested random effects (net within a site since nets were reused)
m.glmm.random2 <- glmmTMB(
  nflounder ~ year + lat + long + distshore + area + chlorophyll + tempavg + tempstdev + sal + bod + nh3 + po4 + depth + (1 | site) + (1 | net),
  data = df,
  family = nbinom2()
)
```

```

)
summary(m.glmm.random2)

## Family: nbinom2 ( log )
## Formula:
## nflounder ~ year + lat + long + distshore + area + chlorophyll +
##      tempavg + tempstdev + sal + bod + nh3 + po4 + depth + (1 |      site/net)
## Data: df
##
##      AIC      BIC    logLik deviance df.resid
## 12620.4 12720.9 -6293.2 12586.4      2707
##
## Random effects:
##
## Conditional model:
## Groups      Name      Variance Std.Dev.
## net:site (Intercept) 0.1387  0.3724
## site      (Intercept) 0.9891  0.9945
## Number of obs: 2724, groups: net:site, 191; site, 74
##
## Dispersion parameter for nbinom2 family (): 0.367
##
## Conditional model:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 143.642489  22.690961   6.330 2.45e-10 ***
## year        -0.077686   0.010172  -7.637 2.22e-14 ***
## lat          0.302926   0.174863   1.732  0.0832 .
## long         0.278606   0.131623   2.117  0.0343 *
## distshore    -0.319119   0.053400  -5.976 2.29e-09 ***
## area         -0.007699   0.046434  -0.166  0.8683
## chlorophyll  -0.098694   0.046527  -2.121  0.0339 *
## tempavg       0.049323   0.025896   1.905  0.0568 .
## tempstdev     0.051795   0.036576   1.416  0.1567
## sal          -0.096830   0.021197  -4.568 4.92e-06 ***
## bod           0.433694   0.238992   1.815  0.0696 .
## nh3          -3.737041   4.264675  -0.876  0.3809
## po4           0.015004   0.011836   1.268  0.2049
## depth        -0.075507   0.083328  -0.906  0.3649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

# year/site for a model with random variation in slopes through years across sites

```

```

m.glmm.random3 <- glmmTMB(
  nflounder ~ lat + long + distshore + area + chlorophyll + tempavg + tempstdev + sal + bod + nh3 + po4 + depth + (1 | site)
  data = df,
  family = nbinom2()
)
summary(m.glmm.random3)

```

```

## Family: nbinom2 ( log )
## Formula:
## nflounder ~ lat + long + distshore + area + chlorophyll + tempavg +

```

```
##      tempstdev + sal + bod + nh3 + po4 + depth + (year | site) +      (1 | net)
## Data: df
##
##      AIC      BIC    logLik deviance df.resid
## 12635.0 12741.3 -6299.5 12599.0      2706
##
## Random effects:
##
## Conditional model:
## Groups Name      Variance Std.Dev.  Corr
## site  (Intercept) 8.497e+04 2.915e+02
## year      2.105e-02 1.451e-01 -1.00
## net  (Intercept) 4.263e-06 2.065e-03
## Number of obs: 2724, groups: site, 74; net, 3
##
## Dispersion parameter for nbinom2 family (): 0.368
##
## Conditional model:
##      Estimate Std. Error z value Pr(>|z|)
## (Intercept) -17.23186    10.24417  -1.682   0.0925 .
## lat          0.38670     0.18178   2.127   0.0334 *
## long         0.20767     0.13643   1.522   0.1279
## distshore    -0.34925     0.05252  -6.650 2.94e-11 ***
## area         0.02209     0.05015   0.440   0.6597
## chlorophyll  -0.10359     0.04803  -2.157   0.0310 *
## tempavg      0.02584     0.02631   0.982   0.3260
## tempstdev    0.04094     0.03928   1.042   0.2973
## sal          -0.09750     0.02211  -4.409 1.04e-05 ***
## bod          0.48614     0.25131   1.934   0.0531 .
## nh3          -4.83232     4.37853  -1.104   0.2697
## po4          0.02072     0.01246   1.663   0.0963 .
## depth       -0.08201     0.08681  -0.945   0.3448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lets compare fixed and random effect models.

```
anova(m.glmm.fixed, m.glmm.random1, m.glmm.random2, m.glmm.random3)
```

```
## Data: df
## Models:
## m.glmm.fixed: nflounder ~ year + lat + long + distshore + area + chlorophyll + , zi=~0, disp=~1
## m.glmm.fixed:      tempavg + tempstdev + sal + bod + nh3 + po4 + depth, zi=~0, disp=~1
## m.glmm.random1: nflounder ~ year + lat + long + distshore + area + chlorophyll + , zi=~0, disp=~1
## m.glmm.random1:      tempavg + tempstdev + sal + bod + nh3 + po4 + depth + (1 | , zi=~0, disp=~1
## m.glmm.random1:      site) + (1 | net), zi=~0, disp=~1
## m.glmm.random2: nflounder ~ year + lat + long + distshore + area + chlorophyll + , zi=~0, disp=~1
## m.glmm.random2:      tempavg + tempstdev + sal + bod + nh3 + po4 + depth + (1 | , zi=~0, disp=~1
## m.glmm.random2:      site/net), zi=~0, disp=~1
## m.glmm.random3: nflounder ~ lat + long + distshore + area + chlorophyll + tempavg + , zi=~0, disp=~1
## m.glmm.random3:      tempstdev + sal + bod + nh3 + po4 + depth + (year | site) + , zi=~0, disp=~1
## m.glmm.random3:      (1 | net), zi=~0, disp=~1
##      Df      AIC      BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
```

```
## m.glmm.fixed    15 12986 13075 -6478.0    12956
## m.glmm.random1 17 12633 12734 -6299.7    12599 356.660    2    <2e-16 ***
## m.glmm.random2 17 12620 12721 -6293.2    12586 12.963    0    <2e-16 ***
## m.glmm.random3 18 12635 12741 -6299.5    12599 0.000    1    1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We choose the model with lowest AIC which is `m.glmm.random2`, a model of nested random effects (net within a site).

Zero-inflated poisson GLMM

Now lets try a zero-inflated Poisson model with a single zero inflation parameter applying to all observations using `ziformula~1`.

```
m.glmm.random.zero <- glmmTMB(
  nflounder ~ year + lat + long + distshore + area + chlorophyll + tempavg + tempstdev + sal + bod + r
  data = df,
  ziformula=~1,
  family = poisson()
)
```

```
## Warning in (function (start, objective, gradient = NULL, hessian = NULL, :
## NA/NaN function evaluation
## Warning in (function (start, objective, gradient = NULL, hessian = NULL, :
## NA/NaN function evaluation
```

```
summary(m.glmm.random.zero)
```

```
## Family: poisson ( log )
## Formula:
## nflounder ~ year + lat + long + distshore + area + chlorophyll +
## tempavg + tempstdev + sal + bod + nh3 + po4 + depth + (1 | site/net)
## Zero inflation: ~1
## Data: df
##
##      AIC      BIC  logLik deviance df.resid
## 26443.9 26544.3 -13204.9 26409.9    2707
##
## Random effects:
##
## Conditional model:
## Groups   Name      Variance Std.Dev.
## net:site (Intercept) 0.5892  0.7676
## site      (Intercept) 28.6579  5.3533
## Number of obs: 2724, groups: net:site, 191; site, 74
##
## Conditional model:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.548e+02 1.552e+01 -16.415 < 2e-16 ***
## year        -4.676e-02 2.152e-03 -21.729 < 2e-16 ***
## lat          6.265e+00 2.798e-01 22.390 < 2e-16 ***
```

```
## long      -1.380e+00  1.405e-01  -9.820  < 2e-16 ***
## distshore -2.792e-01  2.195e-02 -12.718  < 2e-16 ***
## area      9.999e-03  1.305e-02   0.766  0.44351
## chlorophyll -2.942e-02  9.872e-03  -2.980  0.00288 **
## tempavg    4.913e-02  6.204e-03   7.918  2.40e-15 ***
## tempstdev -1.869e-02  7.958e-03  -2.349  0.01882 *
## sal       -1.532e-01  8.080e-02  -1.896  0.05792 .
## bod        4.488e+00  1.013e+00   4.430  9.44e-06 ***
## nh3       -2.895e+01  1.053e+01  -2.749  0.00598 **
## po4        6.538e-02  5.017e-02   1.303  0.19253
## depth     5.195e-01  3.564e-01   1.458  0.14497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.57605    0.04689  -12.28  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model has higher AIC, hence we prefer binomial models.

Negative binomial GLMM (Hurdle Model).

In contrast to zero-inflated models, hurdle models treat zero-count and nonzero outcomes as two completely separate categories, rather than treating the zero-count outcomes as a mixture of structural and sampling zeros.

```
m.glmm.random2.hurdle <- update (m.glmm.random2,
                                data=df,
                                ziformula=~.,
                                family=truncated_nbinom2())
summary(m.glmm.random2.hurdle)
```

```
## Family: truncated_nbinom2 ( log )
## Formula:
## nflounder ~ year + lat + long + distshore + area + chlorophyll +
##           tempavg + tempstdev + sal + bod + nh3 + po4 + depth + (1 | site/net)
## Zero inflation: ~.
## Data: df
##
##           AIC          BIC    logLik deviance df.resid
## 12602.9 12797.9 -6268.4 12536.9      2691
##
## Random effects:
##
## Conditional model:
## Groups Name Variance Std.Dev.
## net:site (Intercept) 0.08004 0.2829
## site (Intercept) 0.70240 0.8381
## Number of obs: 2724, groups: net:site, 191; site, 74
##
```

```
## Zero-inflation model:
## Groups Name Variance Std.Dev.
## net:site (Intercept) 0.1106 0.3325
## site (Intercept) 0.4699 0.6855
## Number of obs: 2724, groups: net:site, 191; site, 74
##
## Dispersion parameter for truncated_nbinom2 family (): 0.28
##
## Conditional model:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.123e+02 2.642e+01 4.252 2.12e-05 ***
## year -6.124e-02 1.239e-02 -4.945 7.63e-07 ***
## lat 2.450e-01 1.710e-01 1.433 0.151945
## long 8.322e-02 1.251e-01 0.665 0.505784
## distshore -3.746e-01 7.180e-02 -5.218 1.81e-07 ***
## area -1.690e-02 5.471e-02 -0.309 0.757369
## chlorophyll -6.165e-02 6.006e-02 -1.026 0.304671
## tempavg 6.411e-02 3.245e-02 1.976 0.048191 *
## tempstdev 4.108e-04 4.514e-02 0.009 0.992738
## sal -8.562e-02 2.212e-02 -3.871 0.000108 ***
## bod 2.784e-01 2.219e-01 1.255 0.209532
## nh3 -4.264e+00 4.294e+00 -0.993 0.320644
## po4 1.256e-02 1.109e-02 1.132 0.257515
## depth -8.218e-02 8.350e-02 -0.984 0.325046
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.625e+02 2.596e+01 -6.260 3.86e-10 ***
## year 7.956e-02 1.234e-02 6.446 1.15e-10 ***
## lat -3.740e-02 1.371e-01 -0.273 0.785048
## long -4.484e-01 1.057e-01 -4.243 2.21e-05 ***
## distshore 1.929e-01 5.682e-02 3.394 0.000689 ***
## area -3.049e-02 5.444e-02 -0.560 0.575440
## chlorophyll 9.591e-02 5.440e-02 1.763 0.077881 .
## tempavg -2.900e-02 2.948e-02 -0.984 0.325230
## tempstdev -9.100e-02 4.212e-02 -2.161 0.030731 *
## sal 7.956e-02 1.699e-02 4.683 2.83e-06 ***
## bod -2.746e-01 1.924e-01 -1.428 0.153400
## nh3 1.106e+00 3.358e+00 0.329 0.741781
## po4 -4.584e-05 9.307e-03 -0.005 0.996070
## depth 1.147e-01 6.511e-02 1.762 0.078119 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lets compare the models.

```
anova(m.glmm.random2, m.glmm.random2.hurdle)
```

```
## Data: df
## Models:
## m.glmm.random2: nflounder ~ year + lat + long + distshore + area + chlorophyll + , zi=~0, disp=~1
```



```
## m.glmm.random2:      tempavg + tempstdev + sal + bod + nh3 + po4 + depth + (1 | , zi=~., disp=~1
## m.glmm.random2:      site/net), zi=~0, disp=~1
## m.glmm.random2.hurdle: nflounder ~ year + lat + long + distshore + area + chlorophyll + , zi=~., disp=
## m.glmm.random2.hurdle:      tempavg + tempstdev + sal + bod + nh3 + po4 + depth + (1 | , zi=~0, disp=
## m.glmm.random2.hurdle:      site/net), zi=~., disp=~1
##
##          Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m.glmm.random2      17 12620 12721 -6293.2    12586
## m.glmm.random2.hurdle 33 12603 12798 -6268.4    12537 49.528      16 2.725e-05
##
## m.glmm.random2
## m.glmm.random2.hurdle ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model `m.glmm.random2.hurdle` is significantly better. From now on we continue with the negative binomial GLMM Hurdle model. This model assumes the following:

- Count data: The response variable Y represents counts of events or occurrences.
- Conditional distribution: The counts are assumed to follow a Negative Binomial distribution.
- Random effects: The model incorporates random effects to account for unobserved heterogeneity among groups or individuals.
- Excess zeros: The excess zeros in the data are modeled separately from the counts using a hurdle model approach.

Parameterization

Let Y_{ij} denote the count response variable for the i -th observation in the j -th group or individual. The model is parameterised as follows:

1. Count Model:

$$Y_{ij} \sim \text{NegBin}(\lambda_{ij}, \theta)$$

where λ_{ij} is the mean of the Negative Binomial distribution and θ is the dispersion parameter.

2. Zero-Inflation Model:

$$Z_{ij} \sim \text{Bernoulli}(\pi_{ij})$$

where Z_{ij} represents a binary indicator for excess zeros and π_{ij} is the probability of observing excess zeros.

Link Function

The relationship between the mean count λ_{ij} and the predictors X_{ij} is modeled using a log link function:

$$\log(\lambda_{ij}) = X_{ij}\beta + u_j$$

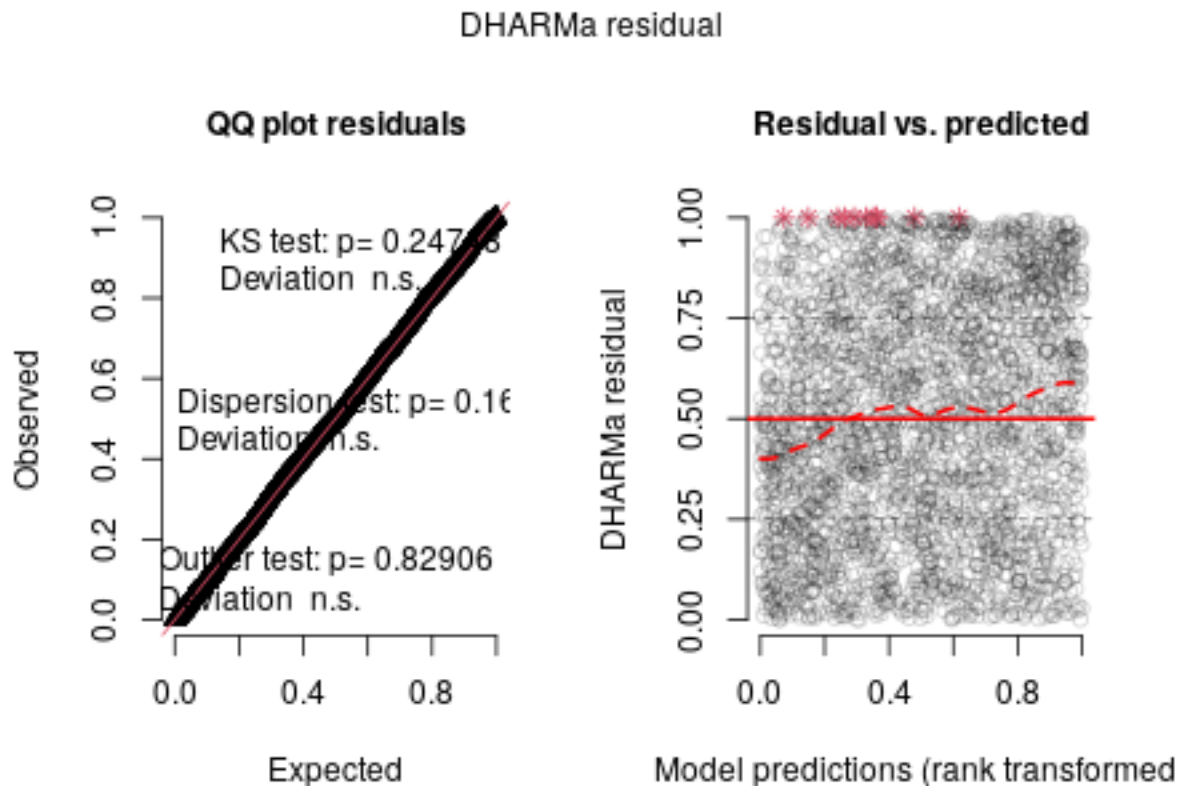
where X_{ij} is a vector of fixed effect predictors, β is a vector of fixed effect coefficients, and u_j represents the random effect for the j -th group or individual.

Interpretation

The Negative Binomial GLMM estimates the effects of the predictors on both the count of events and the presence of excess zeros. The fixed effect coefficients (β) quantify the impact of the predictors on the mean count (λ_{ij}), while the random effects (u_j) capture the variability among groups or individuals.

Post-model-fitting procedure.

```
plot(simulateResiduals(m.glmm.random2.hurdle))
```



Plot effects. This is not working for hurdle.

```
# move to the top  
library(effects)
```

```
## Loading required package: carData
```

```
## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.
```

```
effects_ok <- (requireNamespace("effects") && getRversion() >= "3.6.0")  
if (effects_ok) {  
  all_effects <- allEffects(m.glmm.random2)}
```

```
plot(all_effects)
}
```

```
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
## Warning in Effect.glmmTMB(predictors, mod, vcov. = vcov., ...): overriding
## variance function for effects/dev.resids: computed variances may be incorrect
```

