

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

Olga Lyashevskaya

2024-03-24

Contents

Data preparation and exploration	2
Distribution of nflounder	3
Correlation analysis	6
Zero-inflation	7
Scale variables	7
Data modelling	8
Negative binomial GLM	8
Negative Binomial GLMM	10
Zero-inflated negative binomial GLM	10
Zero-inflated negative binomial GLMM	10
Negative binomial GAM	10
Negative binomial GAMM	10

```
# load packages
packages <- c("ggplot2", "MASS", "mgcv", "rmarkdown", "tinytex", "lme4")
lapply(packages, library, character.only = TRUE)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```
## Loading required package: Matrix
```

```
##
```

```
## Attaching package: 'lme4'
```

```
## The following object is masked from 'package:nlme':
```

```
##
```

```
##      lmList
```

```
## [[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] "mgcv" "nlme" "MASS" "ggplot2" "stats" "graphics"
## [7] "grDevices" "utils" "datasets" "methods" "base"
##
## [[4]]
## [1] "rmarkdown" "mgcv" "nlme" "MASS" "ggplot2" "stats"
## [7] "graphics" "grDevices" "utils" "datasets" "methods" "base"
##
## [[5]]
## [1] "tinytex" "rmarkdown" "mgcv" "nlme" "MASS" "ggplot2"
## [7] "stats" "graphics" "grDevices" "utils" "datasets" "methods"
## [13] "base"
##
## [[6]]
## [1] "lme4" "Matrix" "tinytex" "rmarkdown" "mgcv" "nlme"
## [7] "MASS" "ggplot2" "stats" "graphics" "grDevices" "utils"
## [13] "datasets" "methods" "base"
```

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

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] "tempstdev" "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
##

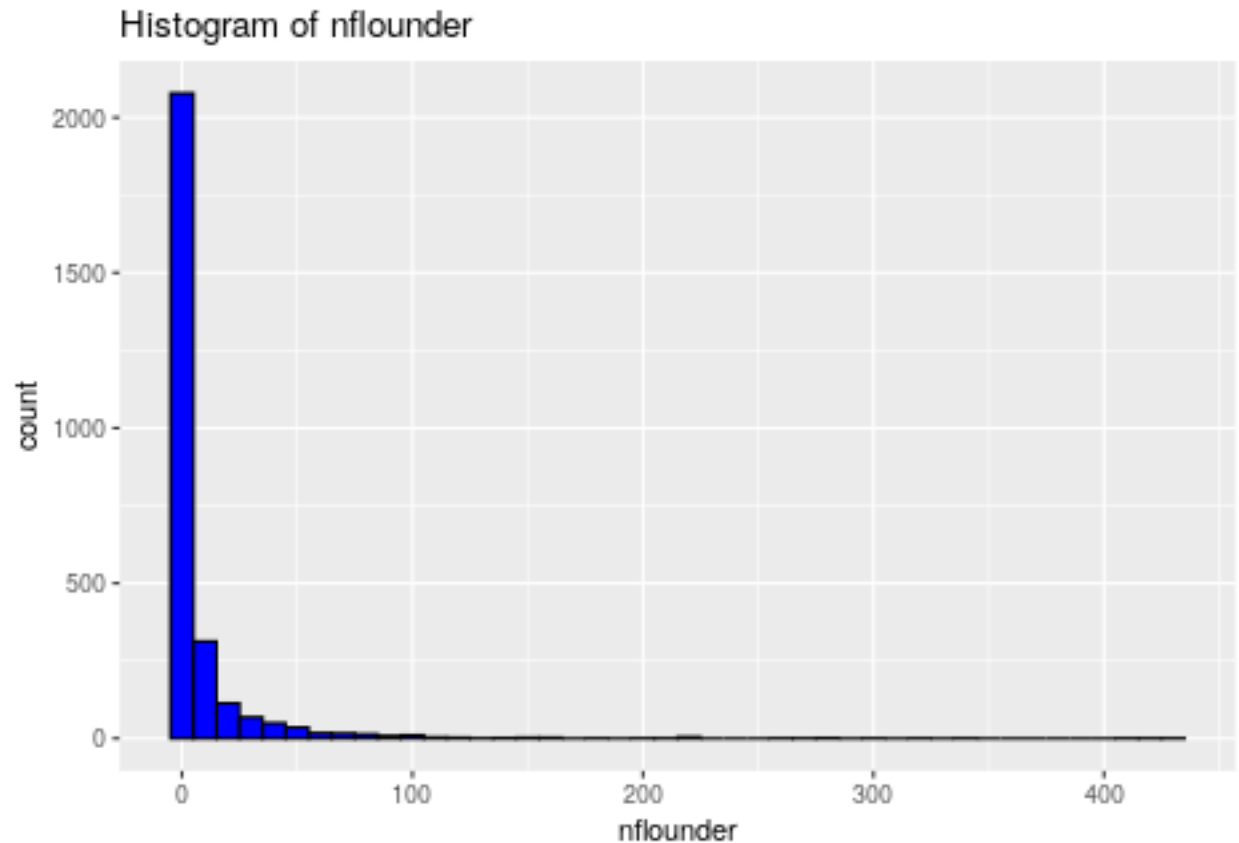
```

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)
```

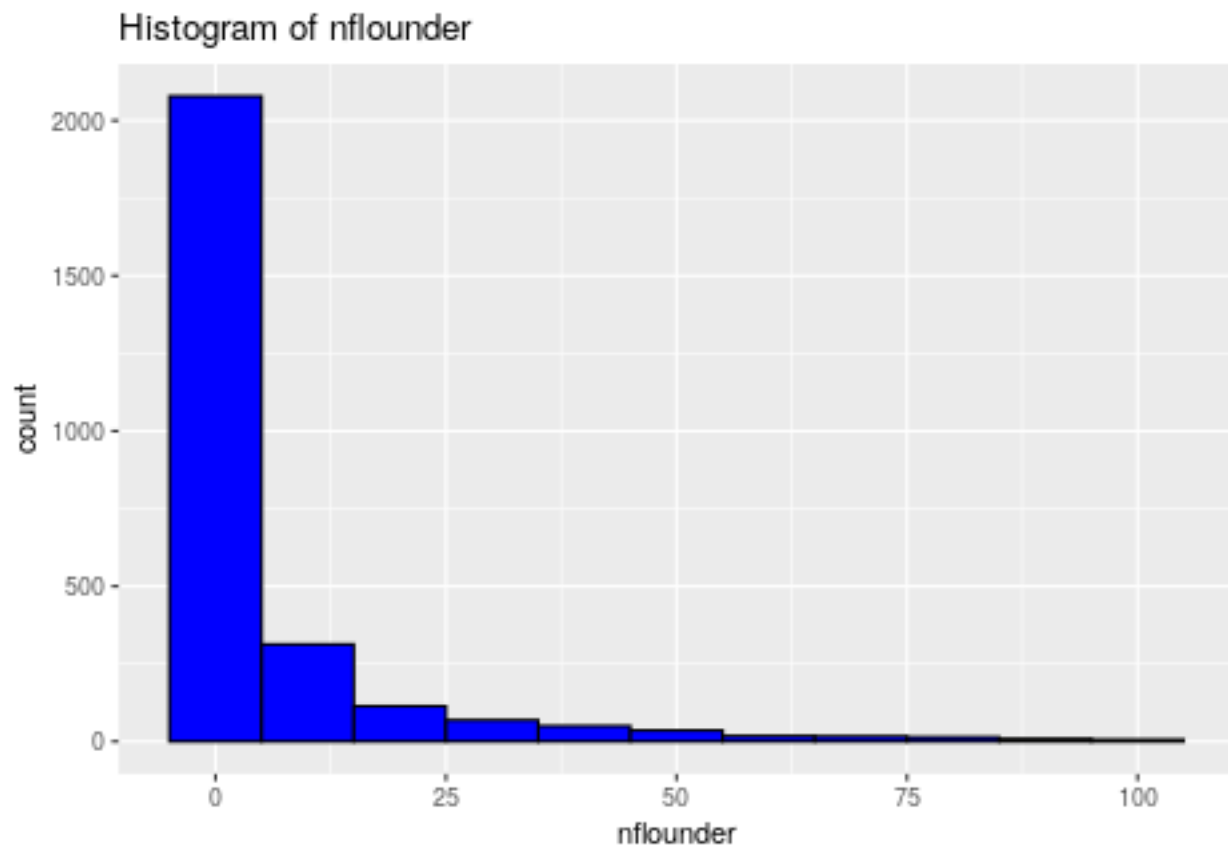
```
##          1          1          0          0          0          0          0          0
## [400,410) [410,420) [420,430]
##          0          2          0
```

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)
```

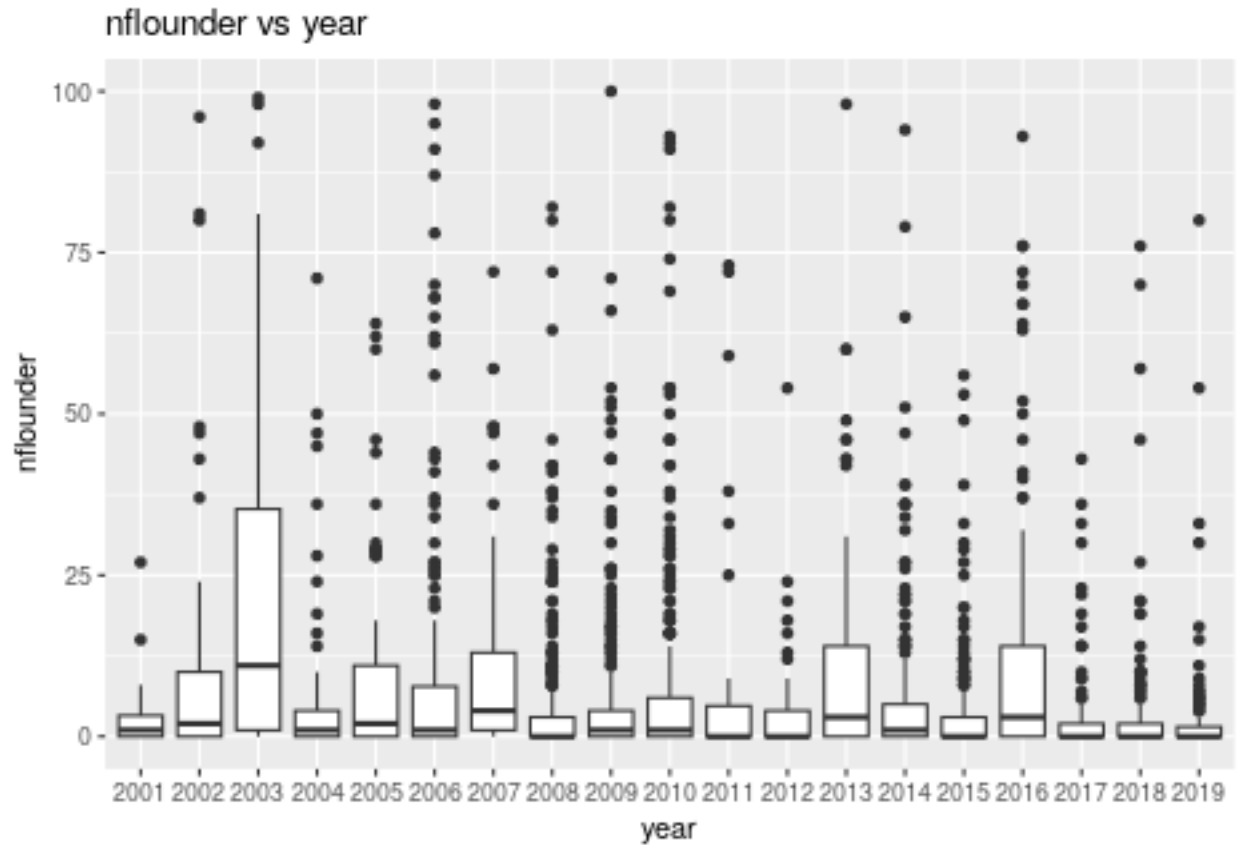
We removed 39 from 2763. Let's visualise distribution of nflounder again.

```
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
```

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

Data modelling

Negative binomial GLM

Lets fit a negative binomial generalized linear model for full dataset, excluding site and net for the moment.

```
m.glm <- glm.nb(nflounder ~. - site - net , data = df)
summary(m.glm)

##
## Call:
## glm.nb(formula = nflounder ~ . - site - net, data = df, init.theta = 0.2756588988,
##       link = log)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 144.949318   18.089464    8.013 1.12e-15 ***
## year        -0.069201    0.008955   -7.728 1.09e-14 ***
## lat         -0.022559    0.060047   -0.376  0.70715
## long         0.268999    0.041393    6.499 8.10e-11 ***
## distshore   -0.327702    0.045287   -7.236 4.62e-13 ***
## trawl        -0.049986    0.040031   -1.249  0.21178
## area        -0.087073    0.042107   -2.068  0.03865 *
## chlorophyll -0.051274    0.041786   -1.227  0.21981
## tempavg      0.039821    0.024510    1.625  0.10423
## tempstdev    0.069279    0.032497    2.132  0.03302 *
## sal         -0.096513    0.007967  -12.114 < 2e-16 ***
## bod          0.097436    0.079297    1.229  0.21916
## nh3          4.812033    1.581512    3.043  0.00234 **
## po4         -0.010353    0.004002   -2.587  0.00968 **
## depth       -0.076482    0.025549   -2.993  0.00276 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.2757) family taken to be 1)
##
##      Null deviance: 3043.2  on 2723  degrees of freedom
## Residual deviance: 2597.3  on 2709  degrees of freedom
## AIC: 12987
##
## Number of Fisher Scoring iterations: 1
##
##              Theta:  0.27566
##            Std. Err.:  0.00931
##
## 2 x log-likelihood: -12954.50200
```

Choose a model by AIC in a Stepwise Algorithm

```
m.glm.stp<-stepAIC(m.glm, direction="both")
```



```
## Warning in glm.nb(formula = nflounder ~ year + lat + long + trawl + area + :
## alternation limit reached

## Warning in glm.nb(formula = nflounder ~ year + long + trawl + area +
## chlorophyll + : alternation limit reached

## Warning in glm.nb(formula = nflounder ~ year + long + trawl + area + tempavg +
## : alternation limit reached

## Warning in glm.nb(formula = nflounder ~ year + long + area + tempavg +
## tempstdev + : alternation limit reached

## Warning in glm.nb(formula = nflounder ~ year + long + area + tempavg +
## tempstdev + : alternation limit reached
```

```
summary(m.glm.stp)
```

```
##
## Call:
## glm.nb(formula = nflounder ~ year + long + distshore + area +
##      tempavg + tempstdev + sal + nh3 + po4 + depth, data = df,
##      init.theta = 0.2749853449, link = log)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 140.807458  17.848231   7.889 3.04e-15 ***
## year        -0.067722   0.008849  -7.653 1.96e-14 ***
## long         0.248007   0.037670   6.584 4.59e-11 ***
## distshore   -0.345917   0.044278  -7.812 5.61e-15 ***
## area        -0.094406   0.041363  -2.282 0.022465 *
## tempavg      0.034858   0.023539   1.481 0.138633
## tempstdev    0.069446   0.031955   2.173 0.029761 *
## sal         -0.101265   0.007299 -13.874 < 2e-16 ***
## nh3          5.636543   1.405786   4.010 6.08e-05 ***
## po4         -0.008177   0.003752  -2.180 0.029294 *
## depth       -0.081203   0.023551  -3.448 0.000565 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.275) family taken to be 1)
##
##      Null deviance: 3037.5  on 2723  degrees of freedom
## Residual deviance: 2596.8  on 2713  degrees of freedom
## AIC: 12983
##
## Number of Fisher Scoring iterations: 1
##
##              Theta:  0.27499
##            Std. Err.:  0.00928
##
## 2 x log-likelihood: -12958.79300
```

Likelihood ratio test, p-value tells if the likelihoods can be considered different. Non-significant p-value, use model with smaller df.

```
anova(m.glm, m.glm.stp)
```

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: nflounder
##
## 1
## 2 (site + net + year + lat + long + distshore + trawl + area + chlorophyll + tempavg + tempstdev + s
##      theta Resid. df    2 x log-lik.   Test    df LR stat.   Pr(Chi)
## 1 0.2749853    2713    -12958.79
## 2 0.2756589    2709    -12954.50 1 vs 2      4 4.291258 0.3680211
```

Negative Binomial GLMM

Add random effect of site and net to the best model.

Zero-inflated negative binomial GLM

Zero-inflated negative binomial GLMM

Negative binomial GAM

Negative binomial GAMM

Consider adding random effect of site and net

linearity assumptions -> homogeneity variance -> influential points ->