Hurdle Model by Site: Knox

2025-11-10

```
#Load in necessary libraries
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.3.3
```

```
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 4.3.3

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 4.3.3
```

```
library(DAAG)
```

```
## Warning: package 'DAAG' was built under R version 4.3.3
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(MuMIn)
```

```
## Warning: package 'MuMIn' was built under R version 4.3.3
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.3
```

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

```
## Warning: package 'carData' was built under R version 4.3.3
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:DAAG':
##
##     vif
```

```r
library(viridis)
```

```
## Warning: package 'viridis' was built under R version 4.3.3
```

```
## Loading required package: viridisLite
```

```r
library(DHARMa)
```

```
## Warning: package 'DHARMa' was built under R version 4.3.3
```

```
## This is DHARMa 0.4.7. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa')
```

```r
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
##
## Attaching package: 'tidyr'
```

```
## The following objects are masked from 'package:Matrix':
##
##     expand, pack, unpack
```

```r
#Import the dataset
data <- read_excel("~/UBCO/Invasion publication/Invasion_publication/Invasion 197198_ID_FINAL.xlsx")

#Remove missing values (SampleID24, 2017)
#Convert NAs to proper data type
data[data == "NA"] <- NA
```

```r
#Remove nas
data <- drop_na(data)

#Set seed
set.seed(123)

#Review data structure
summary(data)
```

```
##       Year          Plot              SampleID         Plant
##  Min.   :2014   Length:285         Min.   : 1.00   Min.   :1.000
##  1st Qu.:2014   Class :character   1st Qu.: 6.00   1st Qu.:1.000
##  Median :2015   Mode  :character   Median :12.00   Median :3.000
##  Mean   :2015                      Mean   :12.38   Mean   :3.232
##  3rd Qu.:2016                      3rd Qu.:18.00   3rd Qu.:4.000
##  Max.   :2017                      Max.   :24.00   Max.   :5.000
##    Distance          Copies_ul
##  Length:285         Length:285
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```r
head(data)
```

```
## # A tibble: 6 x 6
##    Year Plot     SampleID Plant Distance Copies_ul
##   <dbl> <chr>       <dbl> <dbl> <chr>    <chr>
## 1  2014 Aberdeen        1     1 10       2.1
## 2  2014 Aberdeen        1     1 25       0.08
## 3  2014 Aberdeen        1     1 50       0
## 4  2014 Tutt            2     1 10       0.26
## 5  2014 Tutt            2     1 25       5.5
## 6  2014 Tutt            2     1 50       21.8
```

```r
tail(data)
```

```
## # A tibble: 6 x 6
##    Year Plot  SampleID Plant Distance Copies_ul
##   <dbl> <chr>    <dbl> <dbl> <chr>    <chr>
## 1  2017 KM 2        22     5 10       0
## 2  2017 KM 2        22     5 25       0
## 3  2017 KM 2        22     5 50       0
## 4  2017 KM 4        23     5 10       1
## 5  2017 KM 4        23     5 25       0.25
## 6  2017 KM 4        23     5 50       0.23
```

```r
#Ensure copies/uL is numeric
data$Copies_ul<-as.numeric(data$Copies_ul)
```

```r
#Pool replicate samples within each year (x3/SampleID/Year)
data <- data %>%
  group_by(SampleID, Year, Plant, Plot) %>%
  summarize(Copies_ul = sum(Copies_ul, na.rm = TRUE), .groups = "drop")

#Check data properly pooled (285/3 = 95). Should be 95 observations.
summary(data)
```

```
##     SampleID          Year          Plant           Plot
##  Min.   : 1.00   Min.   :2014   Min.   :1.000   Length:95
##  1st Qu.: 6.50   1st Qu.:2014   1st Qu.:2.000   Class :character
##  Median :12.00   Median :2015   Median :3.000   Mode  :character
##  Mean   :12.38   Mean   :2015   Mean   :3.232
##  3rd Qu.:18.00   3rd Qu.:2016   3rd Qu.:4.000
##  Max.   :24.00   Max.   :2017   Max.   :5.000
##    Copies_ul
##  Min.   :  0.000
##  1st Qu.:  0.000
##  Median :  0.280
##  Mean   :  5.641
##  3rd Qu.:  2.740
##  Max.   :142.900
```

```r
#Create clear predictor labels: hostID, inoculation, Site
#Plant 3 and 5 = inoculated (1), plant 1 and 4 = uninoculated (0)
#Plants 1 and 3 = achillea; 4 and 5, = artemesia
data <- data %>%
  mutate(
    hostID = ifelse(Plant > 3, "Artemesia", "Achillea"),
    inoculation = case_when(Plant %in% c(1,4) ~ 0L,
                            Plant %in% c(3,5) ~ 1L,
                            TRUE ~ NA_integer_),
    Site = case_when(Plot %in% c("Aberdeen","Quail","Tutt") ~ "UBCO",
                     Plot %in% c("KM 2","KM 4","KM 5") ~ "Knox",
                     TRUE ~ NA_character_)
  )

#Review
head(data)
```

```
## # A tibble: 6 x 8
##   SampleID  Year Plant Plot     Copies_ul hostID   inoculation Site
##      <dbl> <dbl> <dbl> <chr>        <dbl> <chr>          <int> <chr>
## 1        1  2014     1 Aberdeen      2.18 Achillea           0 UBCO
## 2        1  2015     1 Aberdeen      2.8  Achillea           0 UBCO
## 3        1  2016     1 Aberdeen      1.08 Achillea           0 UBCO
## 4        1  2017     1 Aberdeen      3.49 Achillea           0 UBCO
## 5        2  2014     1 Tutt         27.6  Achillea           0 UBCO
## 6        2  2015     1 Tutt          4.86 Achillea           0 UBCO
```

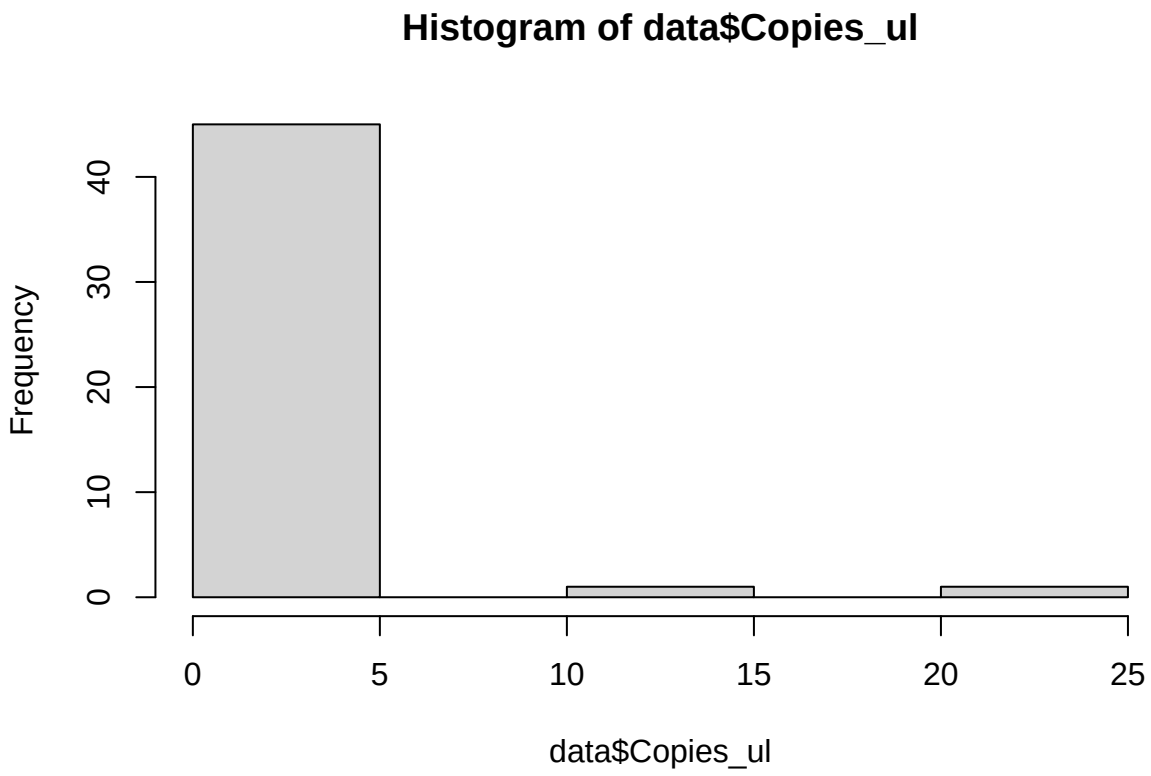
```r
tail(data)
```

```
## # A tibble: 6 x 8
##   SampleID  Year Plant Plot  Copies_ul hostID    inoculation Site
##      <dbl> <dbl> <dbl> <chr>     <dbl> <chr>           <int> <chr>
## 1       23  2015     5 KM 4       0    Artemesia           1 Knox
## 2       23  2016     5 KM 4       0    Artemesia           1 Knox
## 3       23  2017     5 KM 4       1.48 Artemesia           1 Knox
## 4       24  2014     5 KM 5      11.8  Artemesia           1 Knox
## 5       24  2015     5 KM 5       3.5  Artemesia           1 Knox
## 6       24  2016     5 KM 5       0    Artemesia           1 Knox
```

```
#Filter to site (Knox)
SITE <- "Knox"
data <- data %>% filter(Site == SITE)

#Ensure all predictors and response are the correct type. Baseline year for
#intercept estimate
data <- data %>% mutate(
  hostID = factor(hostID),
  Site = factor(Site),
  Plot = factor(Plot),
  inoculation = factor(inoculation),
  Copies_ul = as.numeric(Copies_ul),
  Year = as.integer(Year),
  Year_centered = Year - 2014
)

#Check zero inflation
data <- data %>% mutate(Presence = if_else(Copies_ul > 0, 1, 0))
prop_zero <- mean(data$Presence == 0)
print(prop_zero) #14==38%
```

```
## [1] 0.6170213
```

```
hist(data$Copies_ul)#zero inflation
```

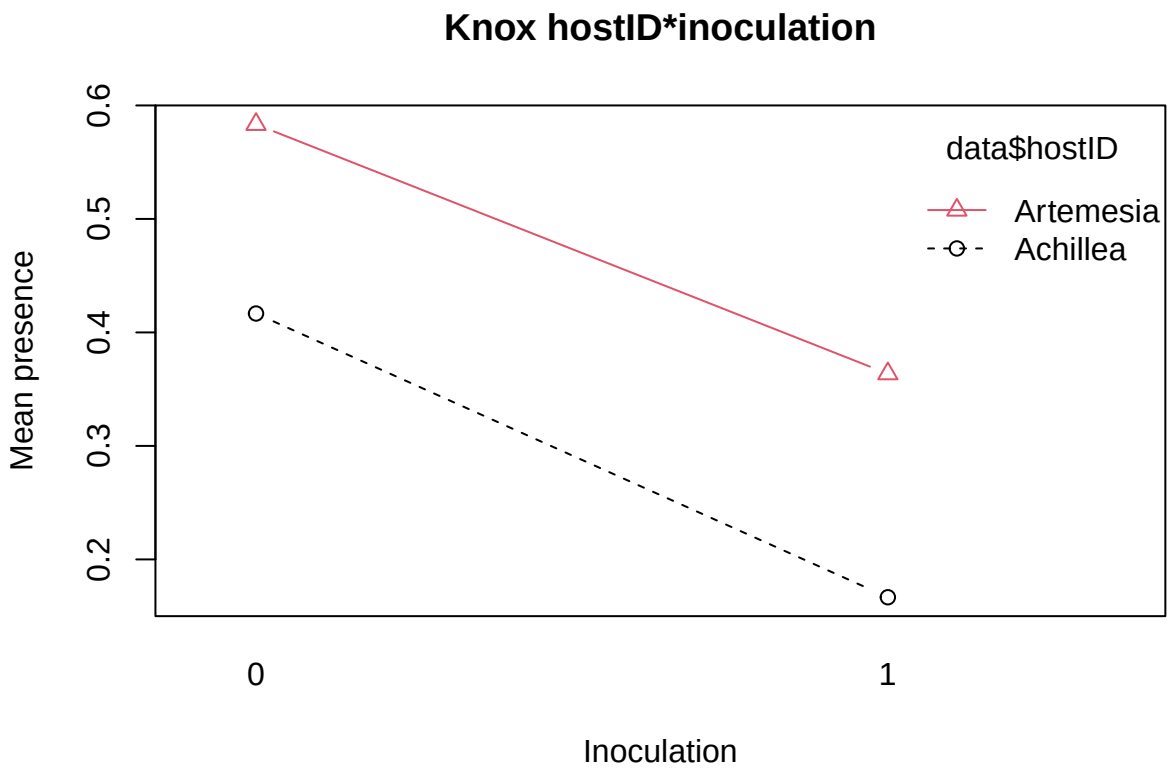



```r
#Zero counts by predictors
zero_df <- data %>% filter(Presence == 0)
View(zero_df)

#Subset non-zero counts, baseline year to 2014
SUBSET <- data %>% filter(Copies_ul > 0) %>% mutate(Year_centered = Year - 2014)

#Save all six one-way plots in one 3x2 layout
png("Appendix_Figure_OneWayPlots.png", width = 2000, height = 2500, res = 200)

layout(matrix(1:6, ncol = 2, byrow = TRUE))
par(mar = c(5, 5, 4, 2))

#(a)
barplot(tapply(data$Presence, data$inoculation, mean, na.rm = TRUE),
        ylim = c(0, 1),
        ylab = "Mean presence",
        main = "Mean Presence by Inoculation (Knox)")
mtext("(a)", side = 3, line = 1, adj = 0, font = 2)

#(b)
barplot(tapply(data$Presence, data$Year, mean, na.rm = TRUE),
        ylim = c(0, 1),
        ylab = "Mean presence",
        main = "Mean Presence by Year (Knox)")
mtext("(b)", side = 3, line = 1, adj = 0, font = 2)
```

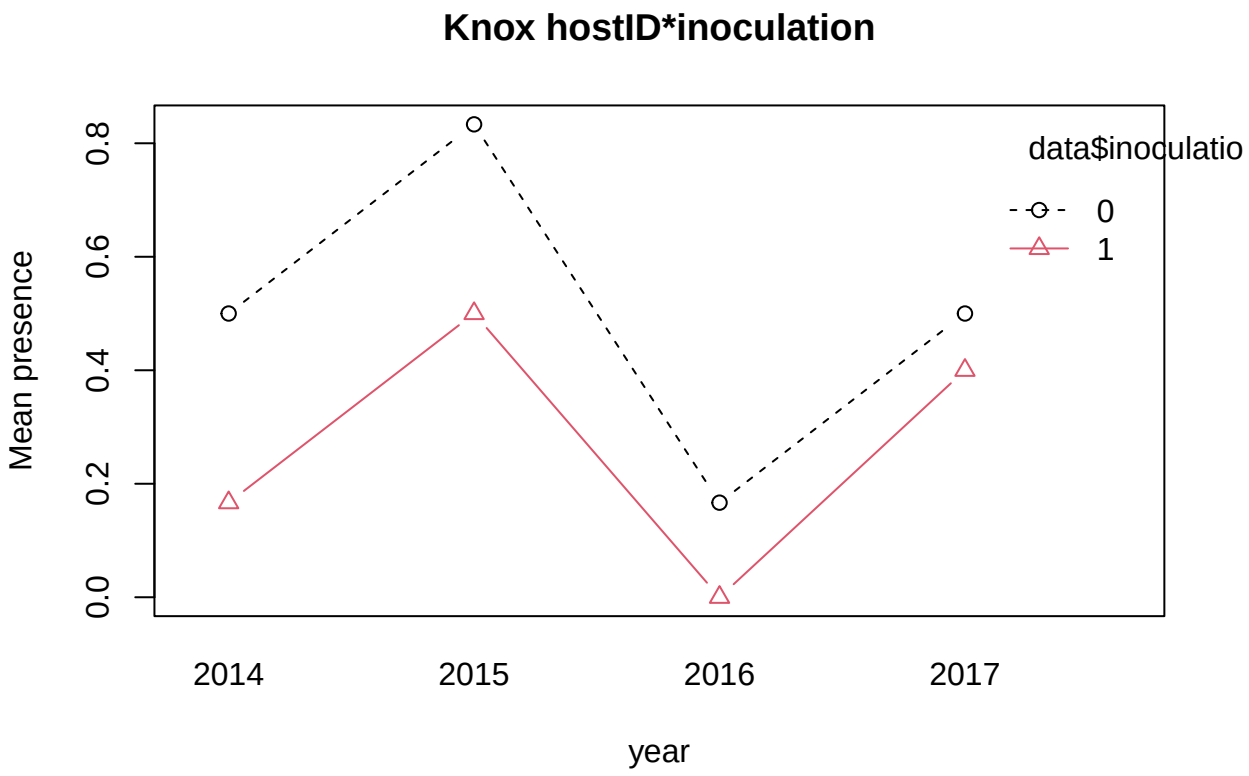
```r
#(c)
barplot(tapply(data$Presence, data$hostID, mean, na.rm = TRUE),
        ylim = c(0, 1),
        ylab = "Mean presence",
        main = "Mean Presence by Host (Knox)")
mtext("(c)", side = 3, line = 1, adj = 0, font = 2)

#(d)
boxplot(Copies_ul ~ inoculation, data = SUBSET,
        xlab = "Inoculation (0/1)",
        ylab = "Copies/uL",
        main = "Abundance by Inoculation (Knox)")
mtext("(d)", side = 3, line = 1, adj = 0, font = 2)

#(e)
boxplot(Copies_ul ~ Year, data = SUBSET,
        xlab = "Year",
        ylab = "Copies/uL",
        main = "Abundance by Year (Knox)")
mtext("(e)", side = 3, line = 1, adj = 0, font = 2)

#(f)
boxplot(Copies_ul ~ hostID, data = SUBSET,
        xlab = "Host",
        ylab = "Copies/uL",
        main = "Abundance by Host (Knox)")
mtext("(f)", side = 3, line = 1, adj = 0, font = 2)

dev.off()
```

```
## pdf
##   2
```

```r
#1 row, 1 column
par(mfrow = c(1,1))

#Check for possible interactions
#inoculation x hostID
interaction.plot(data$inoculation, data$hostID, data$Presence,
                 main= "Knox hostID*inoculation",
                 xlab= "Inoculation",
                 ylab= "Mean presence",
                 legend= TRUE,
                 col= 1:length(unique(data$hostID)),
                 pch= 1:length(unique(data$hostID)),
                 type= "b")
```




```
#Unlikely interaction

#inoculation x Year
interaction.plot(data$Year, data$inoculation, data$Presence,
                 main= "Knox hostID*inoculation",
                 xlab= "year",
                 ylab= "Mean presence",
                 legend= TRUE,
                 col= 1:length(unique(data$inoculation)),
                 pch= 1:length(unique(data$inoculation)),
                 type= "b")
```




```
#Unlikely interaction


#Create binomial (Presence) model
#Note: Plot is often singular or collinear with Site at Knox; omit.
#Start with the simplest model
binom_FIXED_1 <- glm(Presence ~ Year_centered,
                     family = binomial(link = "logit"),
                     data = data)
summary(binom_FIXED_1)
```

```
##
## Call:
## glm(formula = Presence ~ Year_centered, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.3247     0.4929  -0.659     0.51
## Year_centered  -0.1048     0.2717  -0.386     0.70
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 62.557  on 46  degrees of freedom
## Residual deviance: 62.408  on 45  degrees of freedom
## AIC: 66.408
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
binom_FIXED_2 <- glm(Presence ~ Year_centered + inoculation,
                     family = binomial(link = "logit"),
                     data = data)
summary(binom_FIXED_2)
```

```
##
## Call:
## glm(formula = Presence ~ Year_centered + inoculation, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.1888     0.5871   0.322   0.7478
## Year_centered  -0.1258     0.2806  -0.448   0.6538
## inoculation1   -1.0542     0.6287  -1.677   0.0936 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 62.557  on 46  degrees of freedom
## Residual deviance: 59.471  on 44  degrees of freedom
## AIC: 65.471
##
## Number of Fisher Scoring iterations: 4
```

```
binom_FIXED_3 <- glm(Presence ~ Year_centered + inoculation + hostID,
                     family = binomial(link = "logit"),
                     data = data)
summary(binom_FIXED_3) #intercept is the only significant predictor
```

```
##
## Call:
## glm(formula = Presence ~ Year_centered + inoculation + hostID,
##     family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.2311     0.6785  -0.341   0.7334
## Year_centered    -0.1212     0.2865  -0.423   0.6722
## inoculation1     -1.0777     0.6424  -1.678   0.0934 .
## hostIDArtemesia   0.8259     0.6357   1.299   0.1939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 62.557  on 46  degrees of freedom
## Residual deviance: 57.741  on 43  degrees of freedom
## AIC: 65.741
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
#80/20 CV accuracy
cv.binary(binom_FIXED_1)
```

```
##
## Fold:  3 10 2 6 5 4 9 8 7 1
## Internal estimate of accuracy = 0.617
## Cross-validation estimate of accuracy = 0.617
```

```
cv.binary(binom_FIXED_2)
```

```
##
## Fold:  10 4 6 8 7 1 2 5 3 9
## Internal estimate of accuracy = 0.702
## Cross-validation estimate of accuracy = 0.489
```

```
cv.binary(binom_FIXED_3)
```

```
##
## Fold:  8 1 7 10 6 5 4 3 9 2
## Internal estimate of accuracy = 0.66
## Cross-validation estimate of accuracy = 0.574
```

```
#Try an interaction
binom_FIXED_4 <- glm(Presence ~ Year_centered * inoculation,
                     family = binomial(link = "logit"),
                     data = data)
summary(binom_FIXED_4) #An interaction does not lead to a better model fit
```

```
##
## Call:
## glm(formula = Presence ~ Year_centered * inoculation, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   0.4050     0.6941   0.583    0.560
## Year_centered                -0.2700     0.3720  -0.726    0.468
## inoculation1                 -1.5532     1.0568  -1.470    0.142
## Year_centered:inoculation1    0.3433     0.5710   0.601    0.548
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 62.557  on 46  degrees of freedom
## Residual deviance: 59.108  on 43  degrees of freedom
## AIC: 67.108
##
## Number of Fisher Scoring iterations: 4
```

```r
#Check sample ID as a random intercept
#Random intercept per plant (SampleID)
binom_RI <- lme4::glmer(
  Presence ~ Year_centered + inoculation + hostID + (1 | SampleID),
  family = binomial(link = "logit"),
  data = data)
```

```
## boundary (singular) fit: see help('isSingular')
```

```r
summary(binom_RI) #intercept is the only significant predictor
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Presence ~ Year_centered + inoculation + hostID + (1 | SampleID)
##    Data: data
##
##      AIC      BIC   logLik deviance df.resid
##     67.7     77.0    -28.9     57.7       42
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.3464 -0.7855 -0.4892  0.8647  2.3078
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  SampleID (Intercept) 0        0
## Number of obs: 47, groups:  SampleID, 12
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.2311     0.6785  -0.341   0.7334
## Year_centered    -0.1212     0.2865  -0.423   0.6722
## inoculation1     -1.0777     0.6424  -1.678   0.0934 .
## hostIDArtemesia   0.8259     0.6357   1.299   0.1939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Yr_cnt inclt1
## Year_centrd -0.633
## inoculatin1 -0.405  0.064
## hostIDArtms -0.469  0.000 -0.078
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```r
#Singular fit

#Create model without hostID
#Random intercept per plant (SampleID)
binom_RI_2 <- lme4::glmer(
  Presence ~ Year_centered + inoculation + (1 | SampleID),
```

```
  family = binomial(link = "logit"),
  data = data)
```

```
## boundary (singular) fit: see help('isSingular')
```

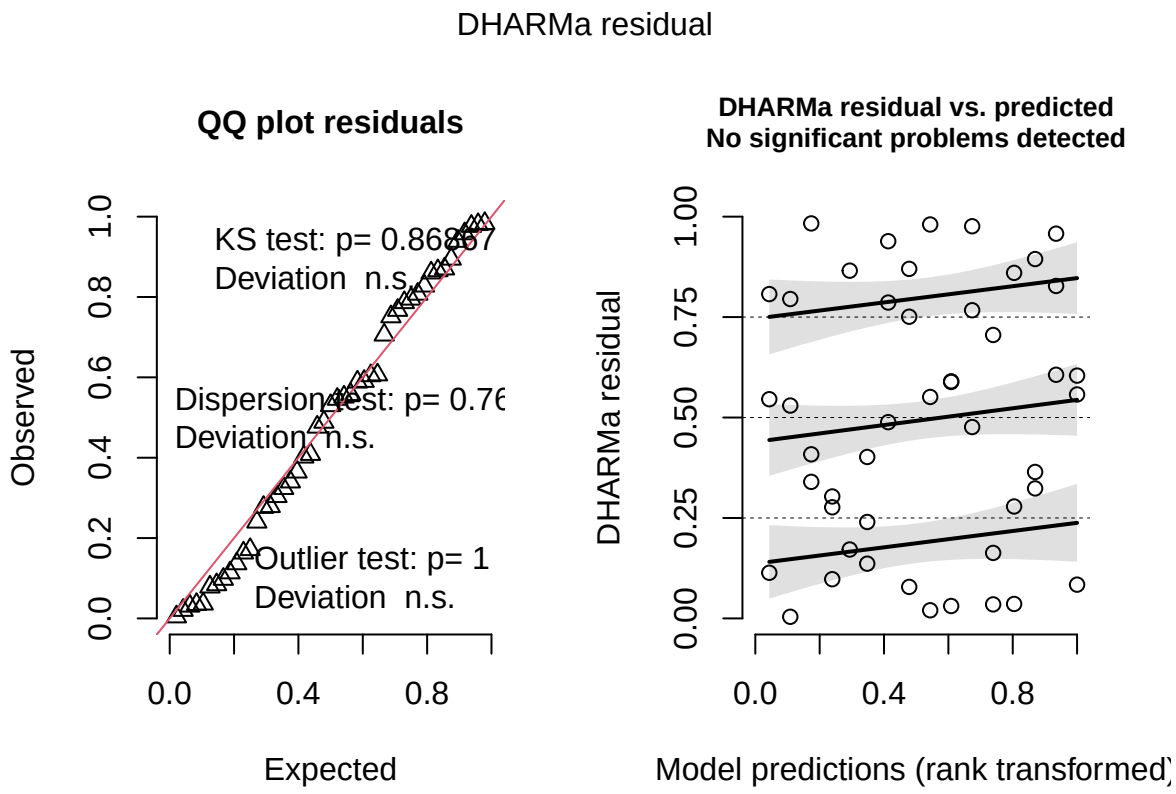
```
summary(binom_RI_2) #intercept is the only significant predictor
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Presence ~ Year_centered + inoculation + (1 | SampleID)
##    Data: data
##
##      AIC      BIC   logLik deviance df.resid
##     67.5     74.9    -29.7     59.5       43
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.0990 -0.7793 -0.5720  0.9690  1.8617
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  SampleID (Intercept) 0        0
## Number of obs: 47, groups:  SampleID, 12
##
## Fixed effects:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.1888     0.5871   0.322   0.7478
## Year_centered  -0.1258     0.2806  -0.448   0.6538
## inoculation1   -1.0542     0.6287  -1.677   0.0936 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Yr_cnt
## Year_centrd -0.717
## inoculatin1 -0.498  0.061
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

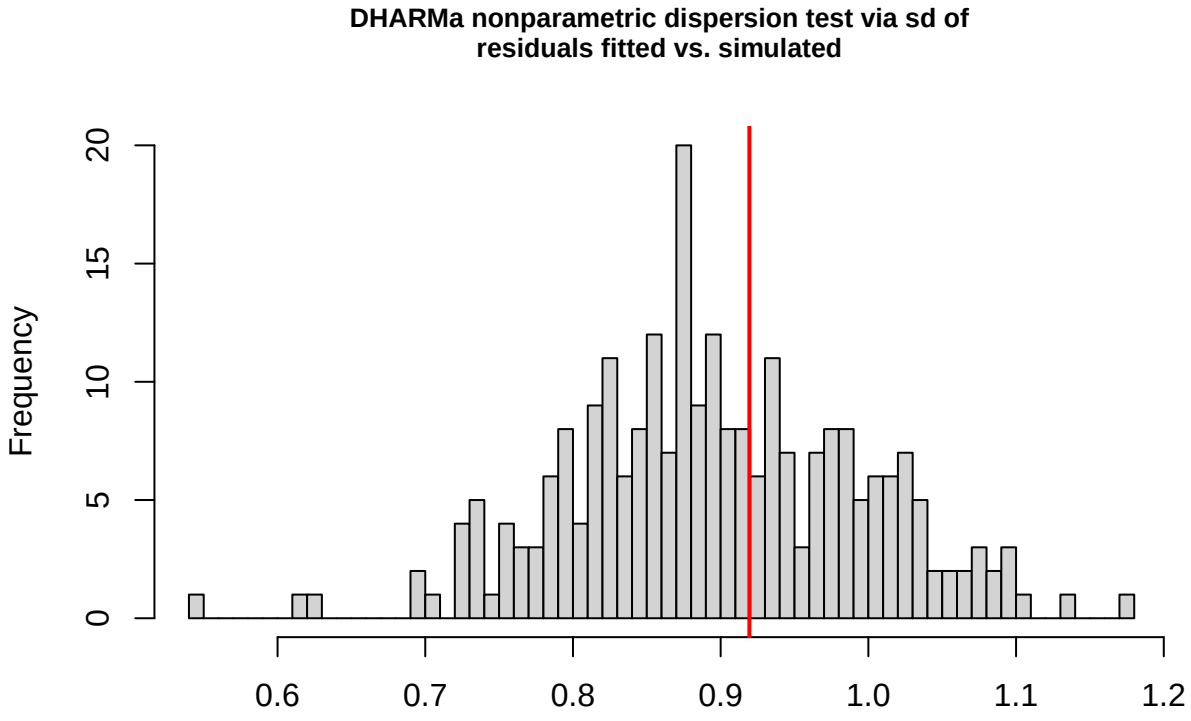
```
#Sample ID explains zero variance
#model with and without hostID are near identical, exclude host ID

#Compare AIC values
AIC(binom_FIXED_1, binom_FIXED_2, binom_FIXED_3, binom_RI, binom_RI_2)
```

```
##               df      AIC
## binom_FIXED_1  2 66.40811
## binom_FIXED_2  3 65.47134
## binom_FIXED_3  4 65.74069
## binom_RI       5 67.74069
## binom_RI_2     4 67.47134
```

```
#No statistical differences in AIC based model fit (delta AIC<2)
#Year and inoculation selected for the final models as these are the two
#predictors of interest.

#Estimated coefficients are identical with and without the random
#effects structure (RE explain zero variance). The random effects
#structure was retained to ensure continuity with the effects structure
#of the UBCO analysis.


#Residual checks (DHARMa diagnostics)
binom_RI_res_2 <- simulateResiduals(binom_RI)
plot(binom_RI_res_2) #good
```

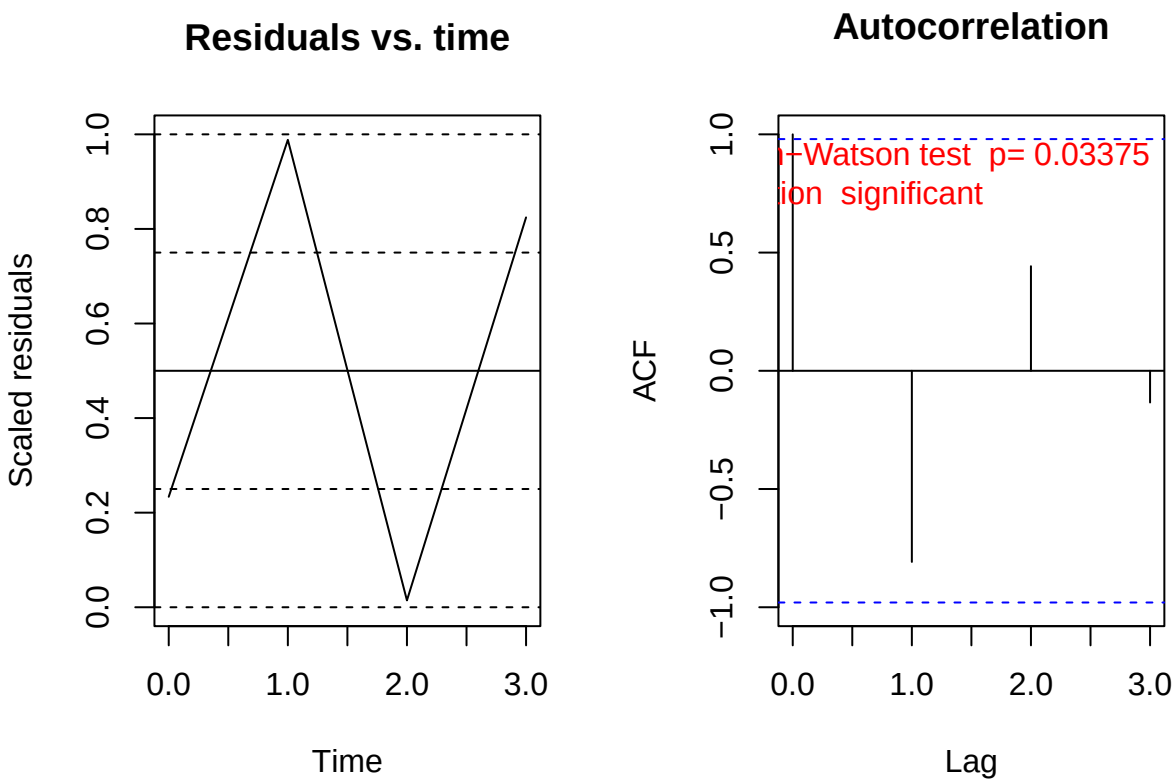



```
print(testDispersion(binom_RI_res_2)) #good
```



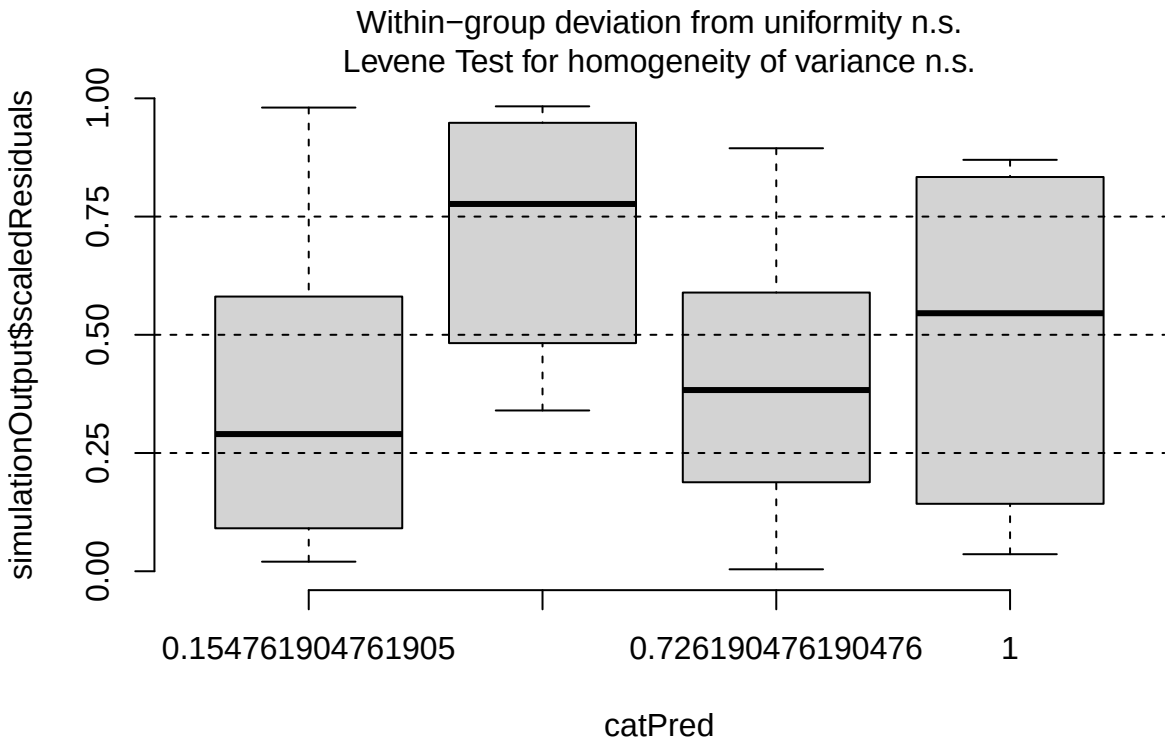

```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.026, p-value = 0.768
## alternative hypothesis: two.sided
```

```r
#Test for temporal autocorrelation
testTemporalAutocorrelation(
  DHARMa::recalculateResiduals(binom_RI_res_2, group = data$Year_centered),
  time = sort(unique(data$Year_centered))
)
```

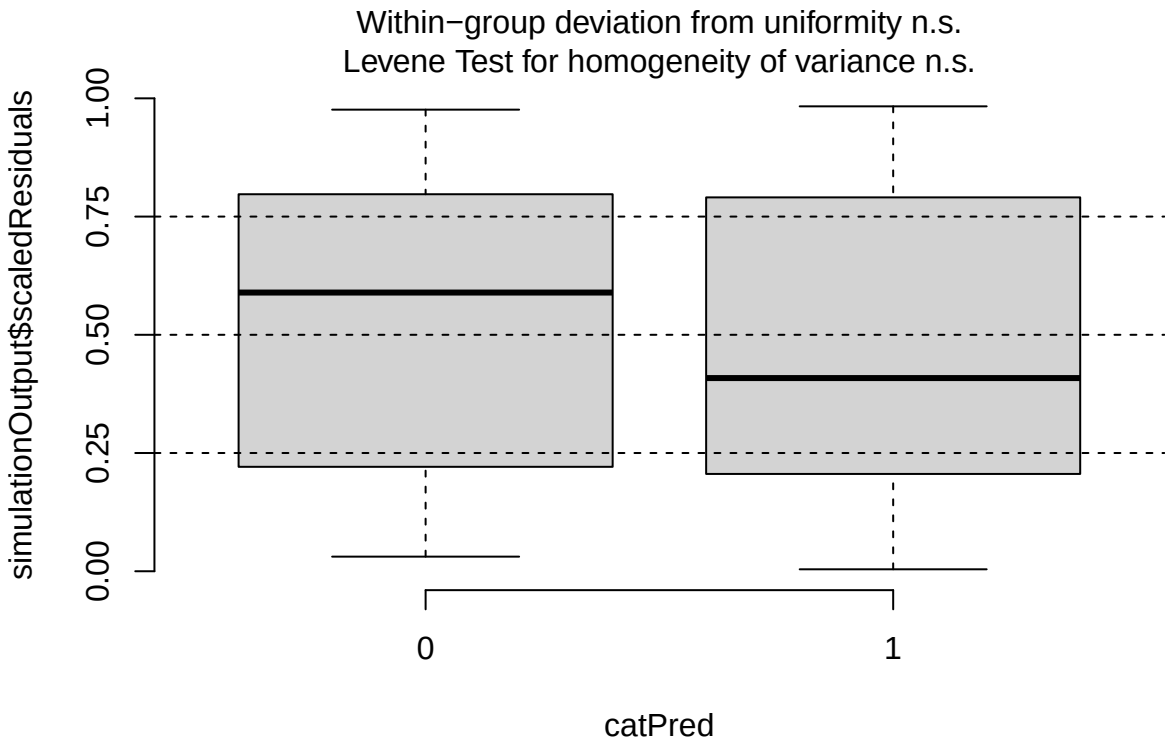



```
##
##  Durbin-Watson test
##
## data:  simulationOutput$scaledResiduals ~ 1
## DW = 3.3473, p-value = 0.03375
## alternative hypothesis: true autocorrelation is not 0
```

```
#Nonlinear time pattern that is not captured by the model

#Check residuals for each predictor
plotResiduals(binom_RI_res_2, data$Year_centered)
```




```
plotResiduals(binom_RI_res_2, data$inoculation)
```




```
#Check for colinearity among predictors
performance::check_collinearity(binom_RI_2) #no evidence of colinearity
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##           Term  VIF  VIF 95% CI adj. VIF Tolerance Tolerance 95% CI
##  Year_centered 1.00 [1.00, Inf]     1.00      1.00     [0.00, 1.00]
##    inoculation 1.00 [1.00, Inf]     1.00      1.00     [0.00, 1.00]
```

```
#Move forward with the mixed effects model (inoculation + year + (1|SampleID))

#Ensure there are adequate observations per predictor for each model
data %>%
  count(Year, inoculation)
```

```
## # A tibble: 8 x 3
##    Year inoculation     n
##   <int> <fct>       <int>
## 1  2014 0               6
## 2  2014 1               6
## 3  2015 0               6
## 4  2015 1               6
## 5  2016 0               6
```

```
## 6  2016 1               6
## 7  2017 0               6
## 8  2017 1               5
```

```
SUBSET %>%
  count(Year, inoculation)
```

```
## # A tibble: 7 x 3
##    Year inoculation     n
##   <int> <fct>       <int>
## 1  2014 0               3
## 2  2014 1               1
## 3  2015 0               5
## 4  2015 1               3
## 5  2016 0               1
## 6  2017 0               3
## 7  2017 1               2
```

```
#The abundance aspect of the Knox site was excluded due to insufficient
#non-zero observations over time

#Summarize final selected model
summary(binom_RI_2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Presence ~ Year_centered + inoculation + (1 | SampleID)
##    Data: data
##
##      AIC      BIC   logLik deviance df.resid
##     67.5     74.9    -29.7     59.5       43
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.0990 -0.7793 -0.5720  0.9690  1.8617
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  SampleID (Intercept) 0        0
## Number of obs: 47, groups:  SampleID, 12
##
## Fixed effects:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.1888     0.5871   0.322   0.7478
## Year_centered  -0.1258     0.2806  -0.448   0.6538
## inoculation1   -1.0542     0.6287  -1.677   0.0936 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Yr_cnt
## Year_centrd -0.717
```

```
## inoculatin1 -0.498  0.061
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```