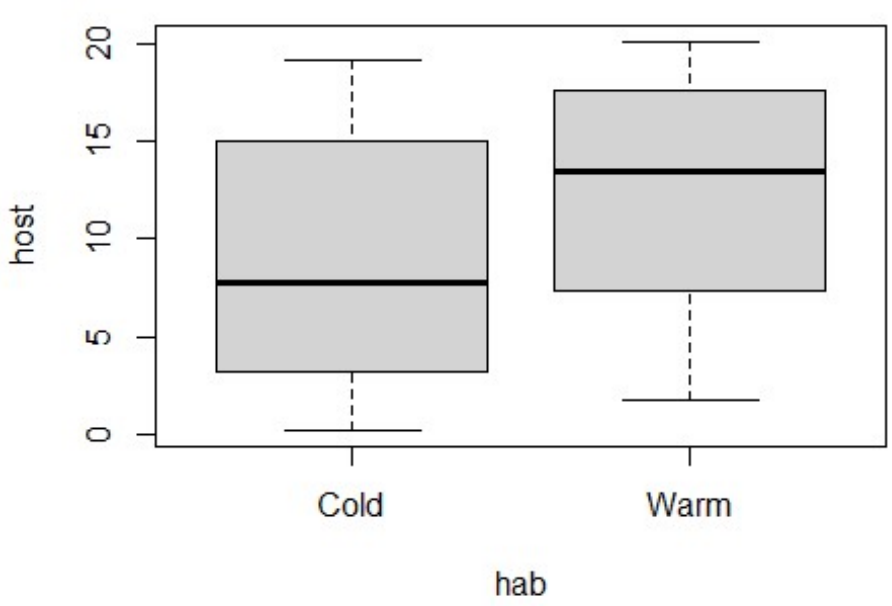
Question 3 - BIOS14 (HT2021)

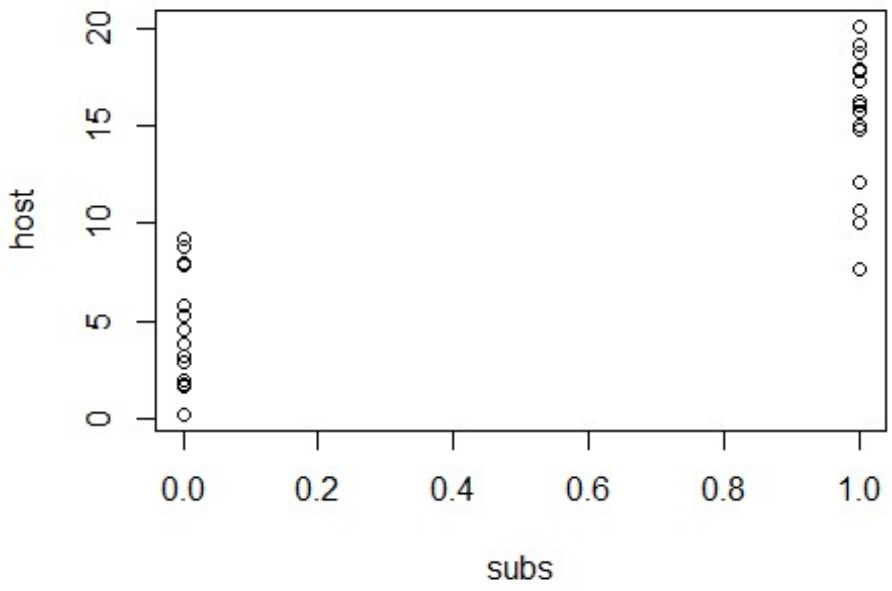
Read and explore data

```
#Question 3

rm(list=ls())

library(car)

## Loading required package: carData

library(DescTools)

## Warning: package 'DescTools' was built under R version 4.1.2

##
## Attaching package: 'DescTools'

## The following object is masked from 'package:car':
##
##     Recode

library(DHARMa)

## Warning: package 'DHARMa' was built under R version 4.1.2

## This is DHARMa 0.4.4. For overview type '?DHARMa'. For recent changes,
type news(package = 'DHARMa')

library(lattice)
library(lmtest)

## Warning: package 'lmtest' was built under R version 4.1.2

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

library(MASS)
library(MuMIn)

## Warning: package 'MuMIn' was built under R version 4.1.2

#read and explore data
moth <- read.csv('moth.csv')
```

```
moth$hab <- factor(moth$hab)

plot(host~hab, data=moth)
```




```
plot(host~subs, data=moth)
```



Part (a)

I have used the **Fisher's exact test** here to compare observed counts with expected counts. Expected counts for one of the groups is less than 5 here, and Fisher's exact test works very well for 2x2 contingency tables with low counts, and for small data sets in general. Hence, I prefer to use this test here.

```
#to create a data frame of the counts of occurrence different species in the
two habitats
cold <- subset(moth, hab=='Cold')
warm <- subset(moth, hab=='Warm')

sp.cold <- rep(0, 2)
for(i in 1:length(cold$subs)){
  if(cold$subs[[i]]==0)
    {sp.cold[1]<-sp.cold[1]+1}
  else
    {sp.cold[2]<-sp.cold[2]+1}
}

sp.warm <- rep(0,2)
for(i in 1:length(warm$subs)){
  if(warm$subs[[i]]==0)
    {sp.warm[1]<-sp.warm[1]+1}
  else
    {sp.warm[2]<-sp.warm[2]+1}
```

```
}

moth.mat <- rbind(sp.cold, sp.warm)
moth.data <- data.frame(moth.mat, row.names=c('Cold', 'Warm'))
colnames(moth.data) <- c('Species 1', 'Species 2')

moth.matrix <- t(moth.mat)
moth.df <- data.frame(moth.matrix, row.names=c('Species 1', 'Species 2'))
colnames(moth.df) <- c('Cold', 'Warm')

#frequency test

#test assumptions

#assumption 1: data in counts - data converted to count form - hence
assumption is true

#assumption 2: categories mutually exclusive: habitats and species are
mutually exclusive categories - assumption true

cat(moth.df>5)

## TRUE TRUE FALSE TRUE

#one of the observed counts <5, fisher's is therefore a better frequency test
here.

#fisher's exact test
fisher.test(moth.data)

##
##  Fisher's Exact Test for Count Data
##
## data:  moth.data
## p-value = 0.06043
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.9422917 44.9348038
## sample estimates:
## odds ratio
##   5.613715

#not significant - can accept null hypothesis
#conclude that occurrence of the two sub-species does not differ between
habitats
```

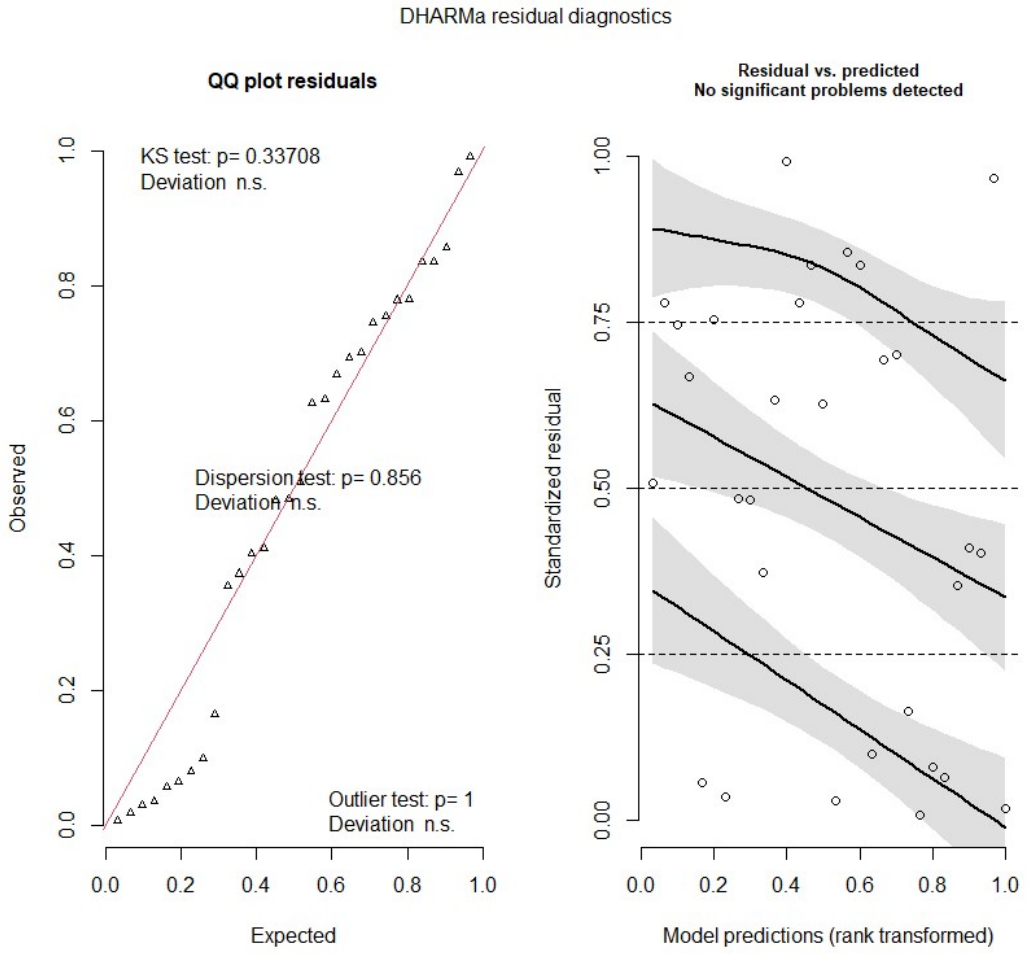
Conclusion: The occurrence of the two sub-species does not differ between habitats

Part (b)

To understand whether the occurrence of either sub-species of moth is influenced by the availability of host plant and habitat, we construct a **generalized linear model for binomial distribution** and observe the variation in occurence with respect to the two aforementioned variables.

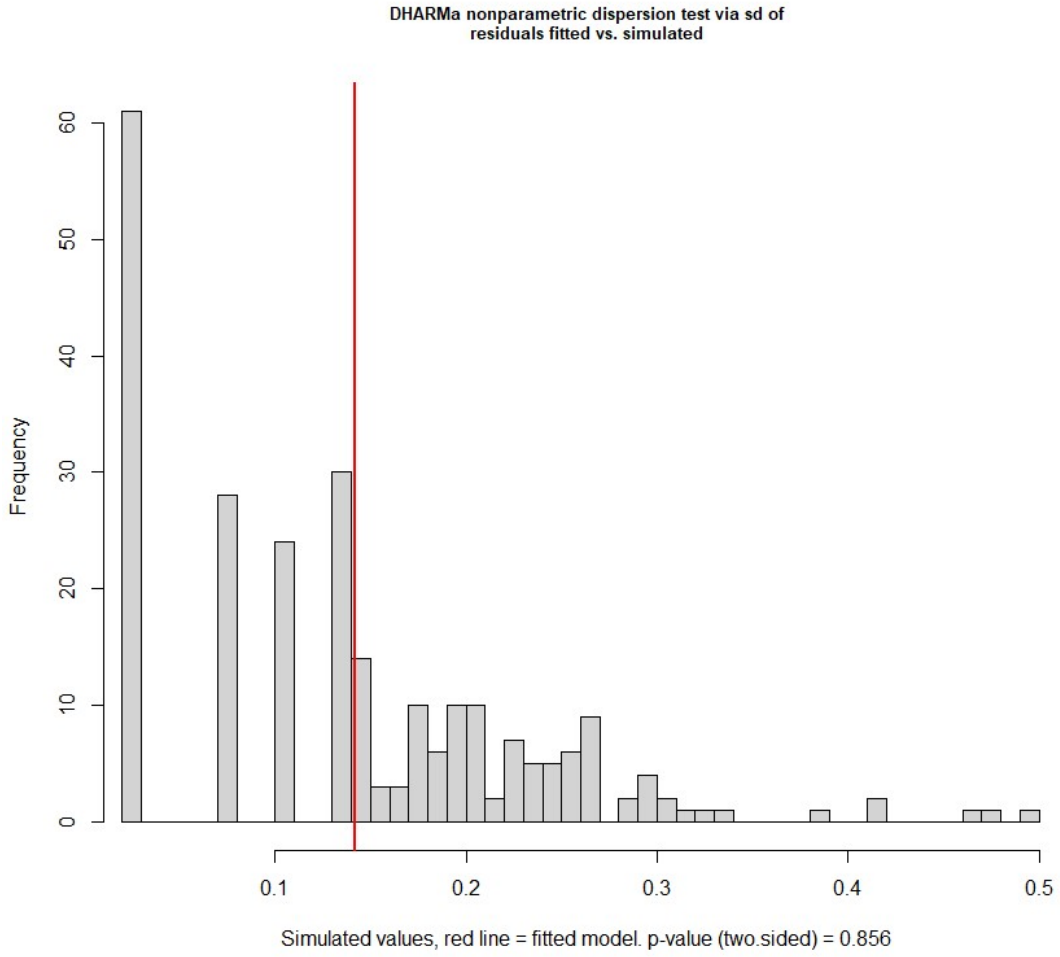
```
#glm with binomial distribution

#setting subspecies as a factor
moth$subs <- factor(moth$subs)

#biologically, habitat might influence the presence of host plants present;
host:hab interaction must therefore not be ignored
model.moth <- glm(subs~host*hab, data=moth, family=binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

#select best model out of the global model
options(na.action='na.fail')#to omit all data missing data
output <- dredge(model.moth)

## Fixed term is "(Intercept)"

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

output

## Global model call: glm(formula = subs ~ host * hab, family = binomial,
data = moth)
## ---
## Model selection table
##        (Int) hab    hst hab:hst df  logLik AICc delta weight
## 3 -1.085e+01     1.1800          2  -4.078 12.6  0.00  0.590
## 4 -9.361e+00   + 0.9446          3  -3.507 13.9  1.34  0.303
## 8 -7.457e+00   + 0.7223       +  4  -3.203 16.0  3.40  0.108
## 2 -6.931e-01   +                 2 -18.205 40.9 28.25  0.000
## 1 -1.216e-16                     1 -20.794 43.7 31.13  0.000
## Models ranked by AICc(x)

#delta AIC values of first two <2, therefore can select either model: picking
model 2 including main effects from host and habitat

sel_mod.moth <- glm(subs~hab+host, data=moth, family=binomial)
#for model without intercept
AIC(sel_mod.moth)

## [1] 13.01355

#check assumptions
```

```
#independent observations
dwtest(sel_mod.moth)

##
##  Durbin-Watson test
##
## data:  sel_mod.moth
## DW = 2.545, p-value = 0.9061
## alternative hypothesis: true autocorrelation is greater than 0

#not significant - assumption true

#glm diagnostics with DHARMa

#simulate scaled residuals
simulationOP <- simulateResiduals(fittedModel=sel_mod.moth, n=250)

plot(simulationOP)
```




```
#QQ plot residuals and KS test - deviation not significant - normality of
residuals true
```

```r
#check for data dispersion
testDispersion(simulationOP)
```




```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.0265, p-value = 0.856
## alternative hypothesis: two.sided
```

```r
#model fitting p-value not significant - good fit

#data slightly under dispersed

#get results
Anova(sel_mod.moth, type=3)
```

```
## Analysis of Deviance Table (Type III tests)
##
```

```
## Response: subs
##      LR Chisq Df Pr(>Chisq)
## hab    1.1426  1     0.2851
## host  29.3970  1  5.897e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#effect of host plant availability is statistically significant

summary(sel_mod.moth)
```

```
##
## Call:
## glm(formula = subs ~ hab + host, family = binomial, data = moth)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.91968  -0.14888  -0.00576   0.03806   2.11141
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.3608     4.5423  -2.061   0.0393 *
## habWarm       2.5870     2.5813   1.002   0.3162
## host          0.9446     0.5187   1.821   0.0686 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.5888  on 29  degrees of freedom
## Residual deviance:  7.0136  on 27  degrees of freedom
## AIC: 13.014
##
## Number of Fisher Scoring iterations: 8
```

I have used type 3 ANOVA for GLM analysis here which controls for the effect of each of the predictors on the response variable when evaluating the effect of a predictor variable.

We can see that effect of host plant availability on moth sub-species occurrence is statistically significant.

From the summary() output, we get that the estimate for host is positive, meaning that occurrence of the particular sub-species (subs=1) increases with increase in density of the particular plant. Further, we can also see that the occurrence of the particular sub-species is slightly lower in cold habitats that in warm habitats. This is however not significant.

Part (c)

Illustration for part (a)

```
#plot count data as bar graph to visualize difference in counts
barplot(as.matrix(moth.data),beside=TRUE,col=c('lightseagreen',
'lightcoral'), main='Distribution of different species of moth\n over
different habitats\n', ylab='Counts')

legend('top', fill =c('lightseagreen', 'lightcoral'), legend=c('Cold
habitat', 'Warm habitat'))
```

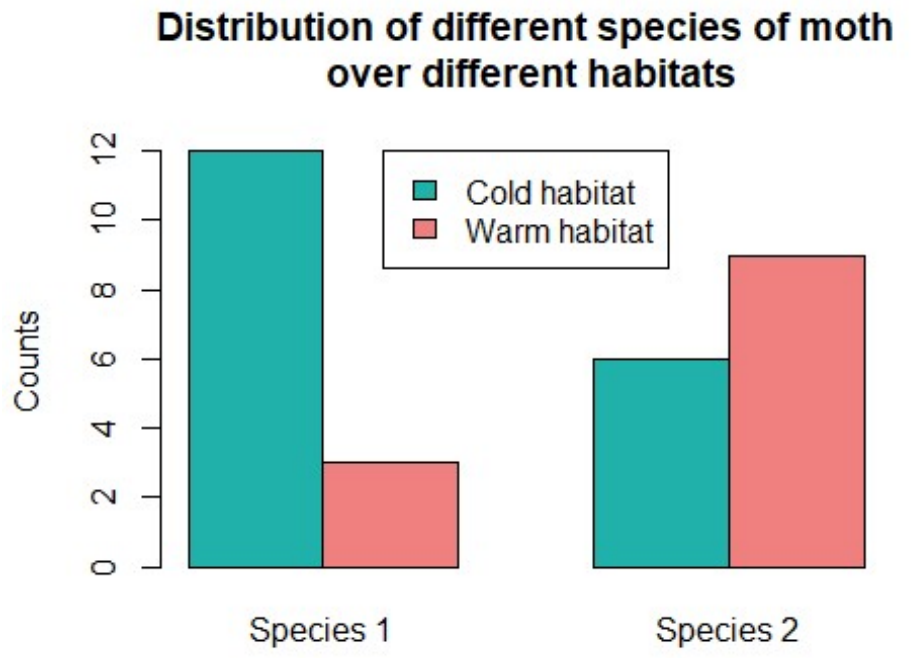
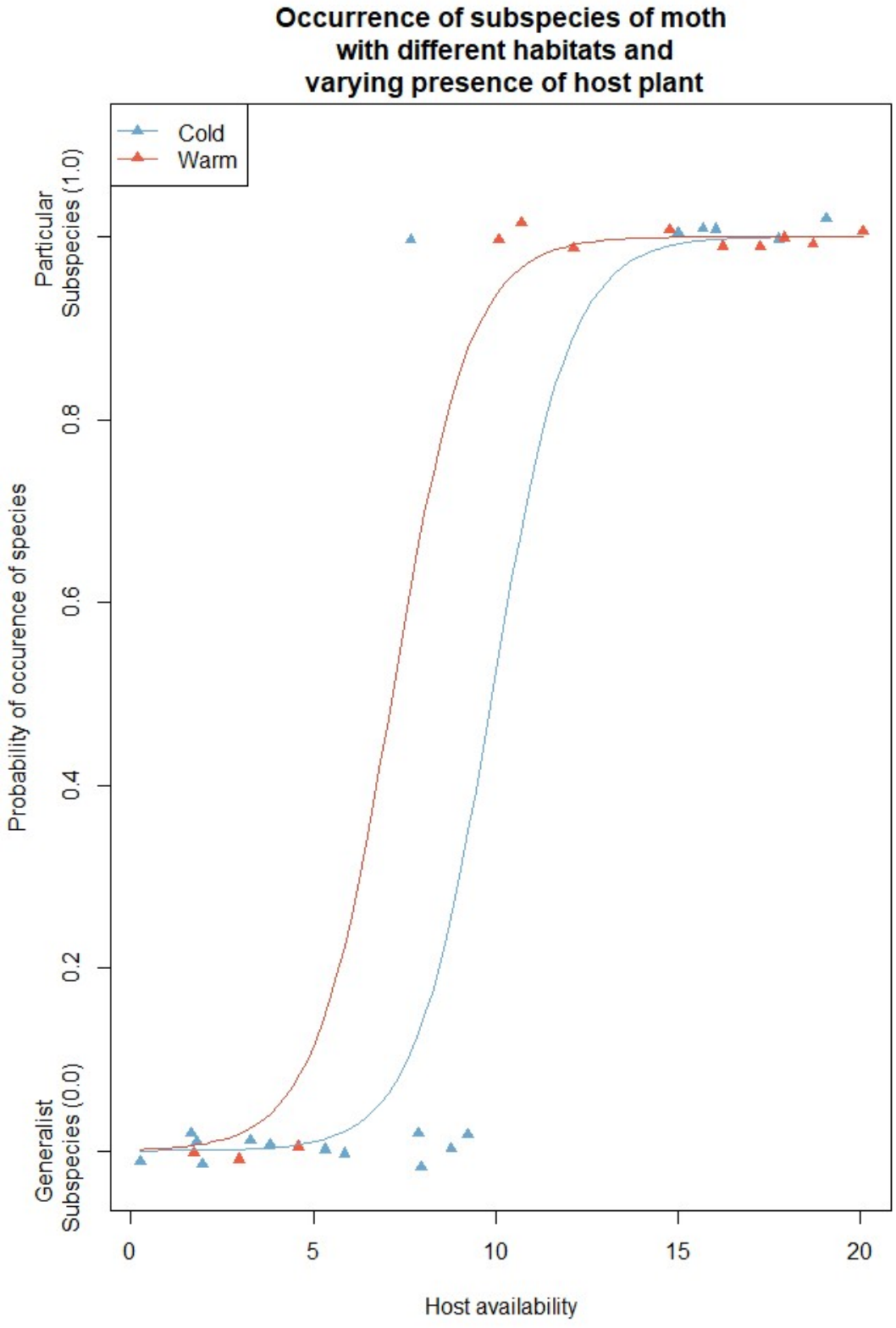


Illustration for part (b)

```
#illustrate model results

pred.moth <- predict(sel_mod.moth, moth, type='response')

moth$subs <- as.numeric(moth$subs)-1

moth$subs.jit <- jitter(moth$subs, amount=0.02)

plot(moth$subs.jit~moth$host, type='n', main='Occurrence of subspecies of
moth\n with different habitats and\n varying presence of host plant',
xlab='Host availability', ylab='Probability of occurence of species',
ylim=c(-0.02,1.1), yaxt='n')

axis(2, at=c(0.0, 0.2, 0.4, 0.6, 0.8, 1.0), labels=c('Generalist\n Subspecies
(0.0)', 0.2, 0.4, 0.6, 0.8, 'Particular \n Subspecies (1.0)'))
```

```
points(moth$subs.jit[moth$hab=='Cold']~moth$host[moth$hab=='Cold'], pch=17,
col='skyblue3')

points(moth$subs.jit[moth$hab=='Warm']~moth$host[moth$hab=='Warm'], pch=17,
col='tomato2')

coeff.moth <- coefficients(sel_mod.moth)
coeff.moth

## (Intercept)     habWarm        host
##  -9.3608103   2.5870354   0.9445565

x <- seq(min(moth$host), max(moth$host), len=100)

y.cold <- coeff.moth[1]+coeff.moth[3]*x
predy.cold <- exp(y.cold)/(1+exp(y.cold))

y.warm <- coeff.moth[1]+coeff.moth[2]+coeff.moth[3]*x
predy.warm <- exp(y.warm)/(1+exp(y.warm))

lines(x, predy.cold, col='skyblue3')
lines(x, predy.warm, col='tomato3')

legend('topleft', lty=1, col=c('skyblue3', 'tomato2'), legend=c('Cold',
'Warm'), pch=c(17,17))
```


Occurrence of subspecies of moth
with different habitats and
varying presence of host plant
Cold
Warm
Particular
Subspecies (1.0)
Generalist
Subspecies (0.0)
Probability of occurence of species
0.2
0.4
0.6
0.8
0
5
10
15
20
Host availability