# Binomial Regression: Seed Germination

Alex Logan

2023-08-10

## Data:

The seeds.csv dataset includes the folllowing variables:

```
##   germinated total extract    plant
## 1         10    39    Oa73     bean
## 2         23    62    Oa73     bean
## 3         23    81    Oa73     bean
## 4         26    51    Oa73     bean
## 5         17    39    Oa73     bean
## 6          5     6    Oa73 cucumber
```

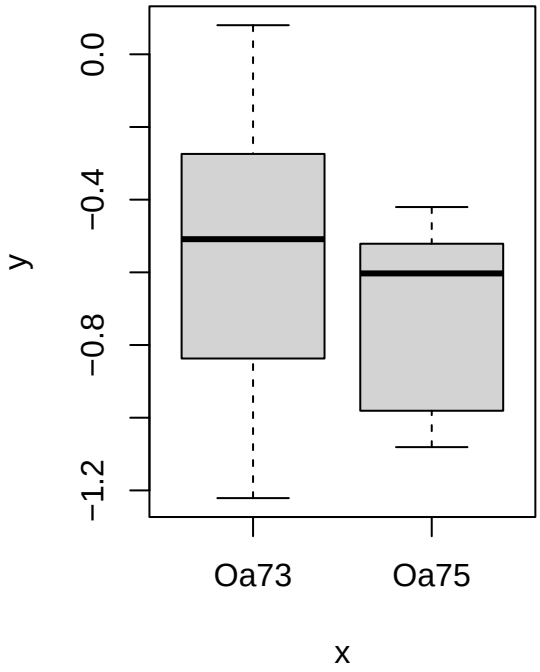
The data arise from a $2^2$ factorial experiment and refer to the number of seeds that successfully germinated on each of 21 plates.
The data give the total number of seeds and the total number of these that germinated successfully, for each plate.

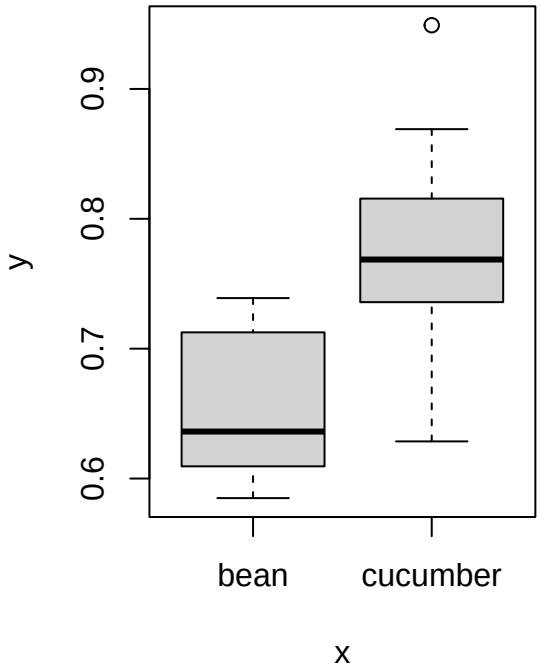
The two factors involved are:
- plant type: (bean,cucumber) and
- root extract: (O.aegyptiaco73, O.aegyptiaco75) for facilitating germination.

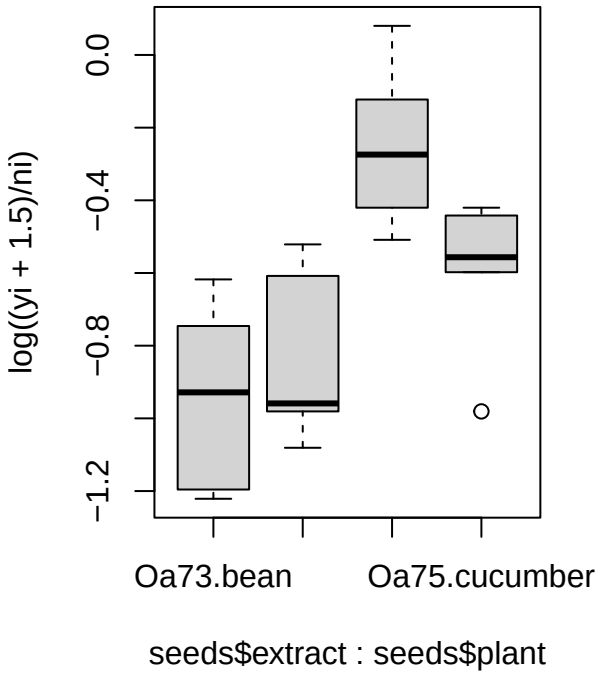
The root extracts are abbreviated as O.a.73 and O.a.75.

Typically, a logistic regression is considered for such data (Binomial), however the probit and complementary log-log (cloglog) link functions may also be considered.

**Exploratory Data Analysis:**



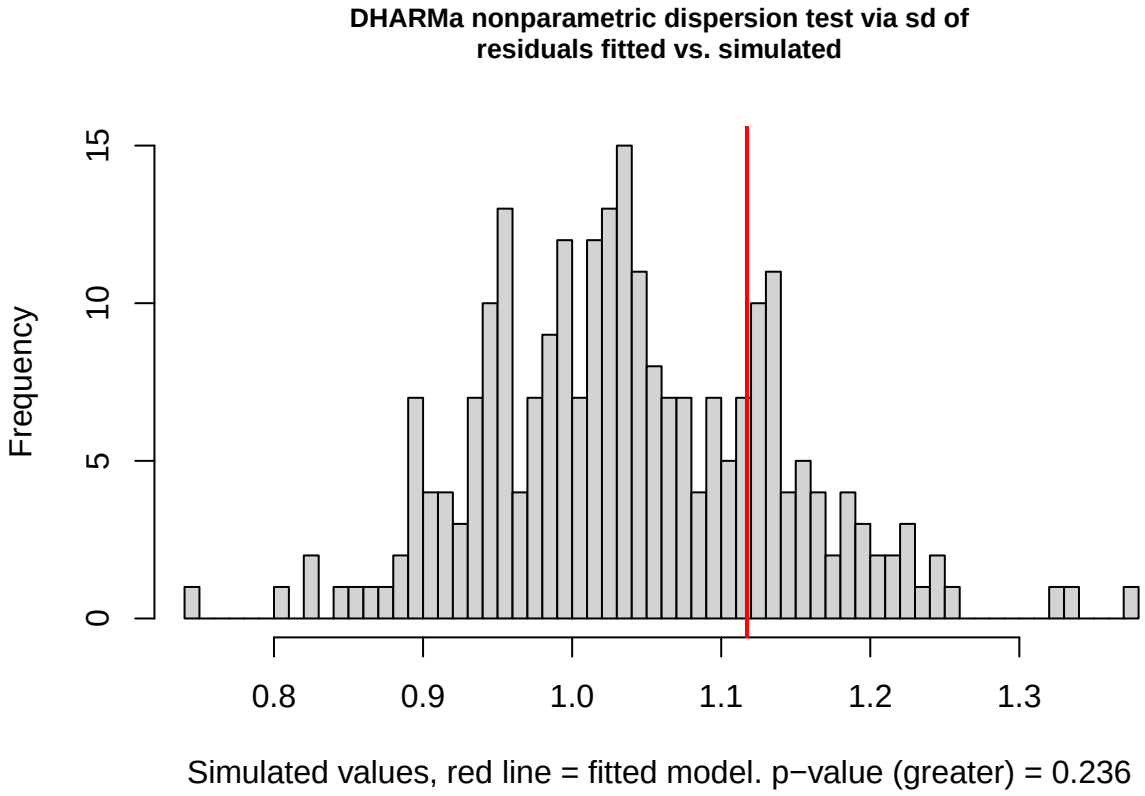
## Logit Link Function Model and Goodnes of Fit:

As per Appendix 1.1, the model without interaction terms shows strong significance in the intercept and plant(cucumber) terms.

As in Appendix 1.2, the Model with interaction term shows strong significance in the intercept and plant(cucumber) terms, also showing a significant interaction effect between plant(cucumber) and extract(Oa75).

Overdispersed model with interaction term shows strong significance in the intercept and plant(cucumber) terms, while no longer showing as significant an interaction effect between plant(cucumber) and extract(Oa75), as seen in Appendix 1.3.

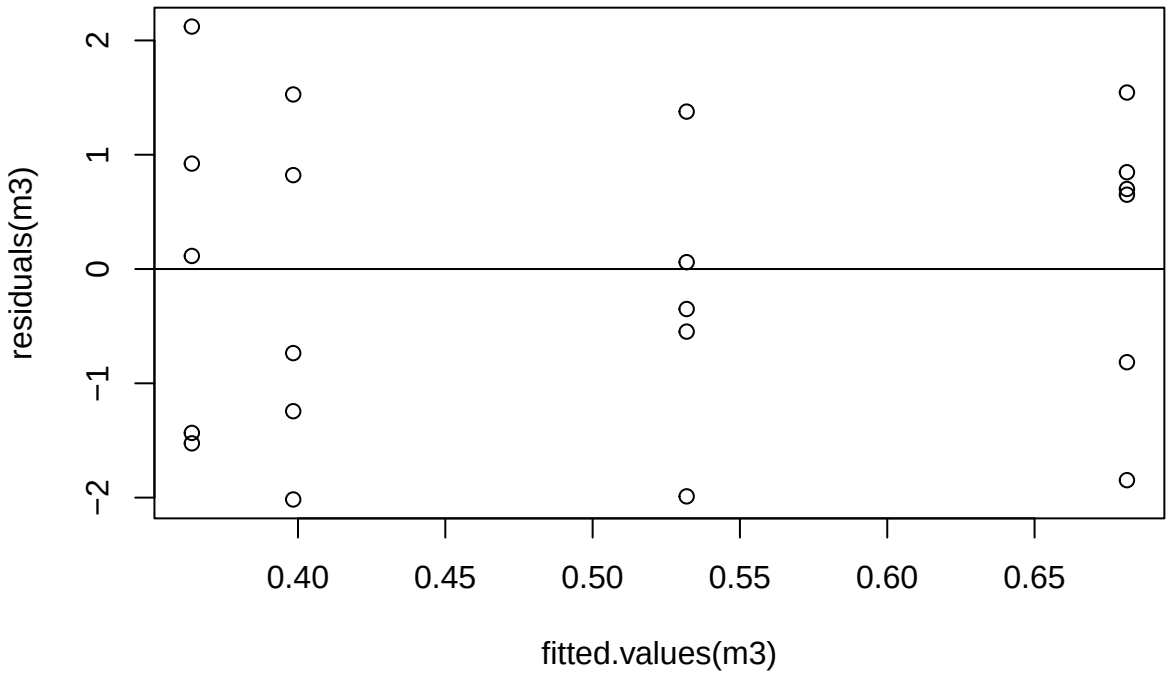
Test for Overdispersion:




```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.0747, p-value = 0.236
## alternative hypothesis: greater
```

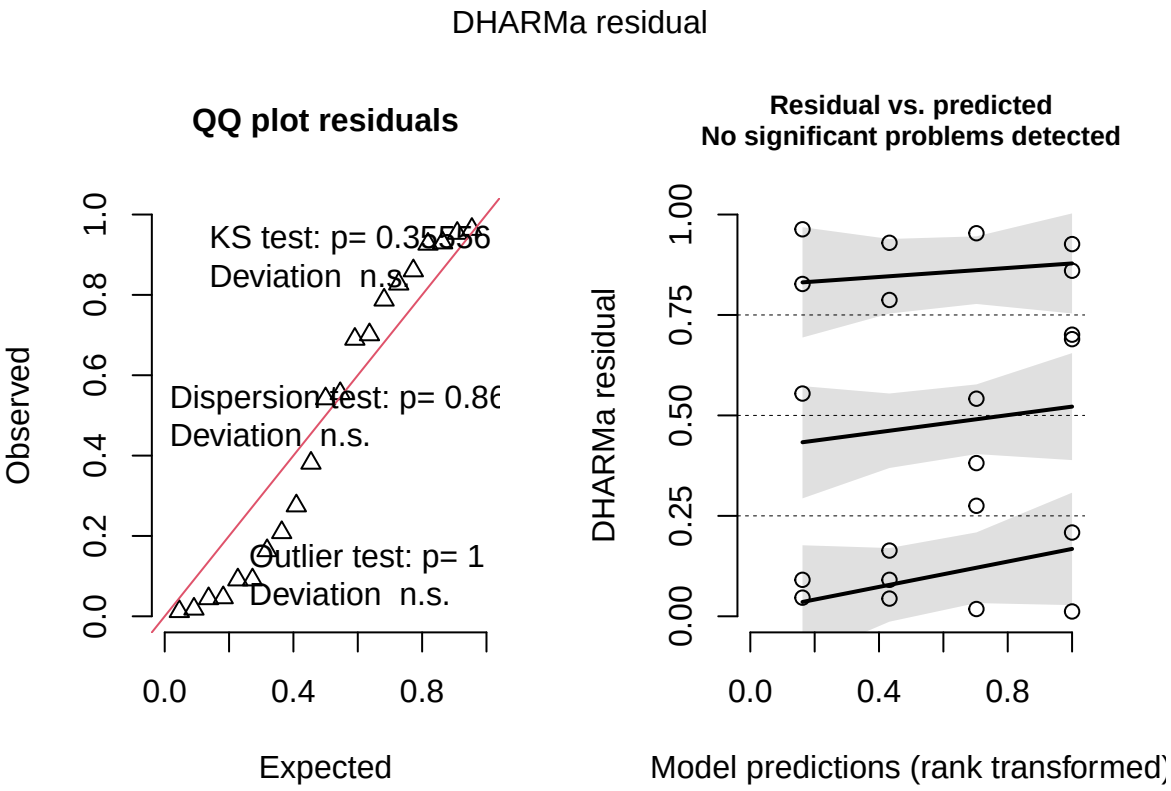
No evidence to suggest the data are significantly overdispersed.

Evaluate Pearson Residuals:




No apparent trend in the Pearson's residuals.

Evaluate Simulated Residuals:




Simulated residuals demonstrate significant departure from the uniform distribution.

## Probit Link Function Model and Goodness of Fit:

The model without interaction term shows strong significance in the intercept and plant(cucumber) terms (Appendix 2.1).

As seen in Appendix 2.2, the model with interaction term shows strong significance in the intercept and plant(cucumber) terms, also showing a significant interaction effect between plant(cucumber) and extract(Oa75).

Test for Overdispersion: Still no evidence to suggest the data are significantly overdispersed (Appendix 2.3).

Evaluate Pearson Residuals: No apparent trend in the Pearson's residuals (Appendix 2.4).

Evaluate Simulated Residuals: Still significant departure from the uniform distribution (Appendix 2.5).

## Complementary Log-Log Link Function Model and Goodness of Fit:

The model without interaction term shows strong significance in the intercept, plant(cucumber) and extract(Oa75) terms (Appendix 3.1).

As seen in Appendix 3.2, the model with interaction term shows strong significance in the intercept and plant(cucumber) terms, also showing a significant interaction effect between plant(cucumber) and extract(Oa75).

Test for Overdispersion: Still no evidence to suggest the data are significantly overdispersed (Appendix 3.3).

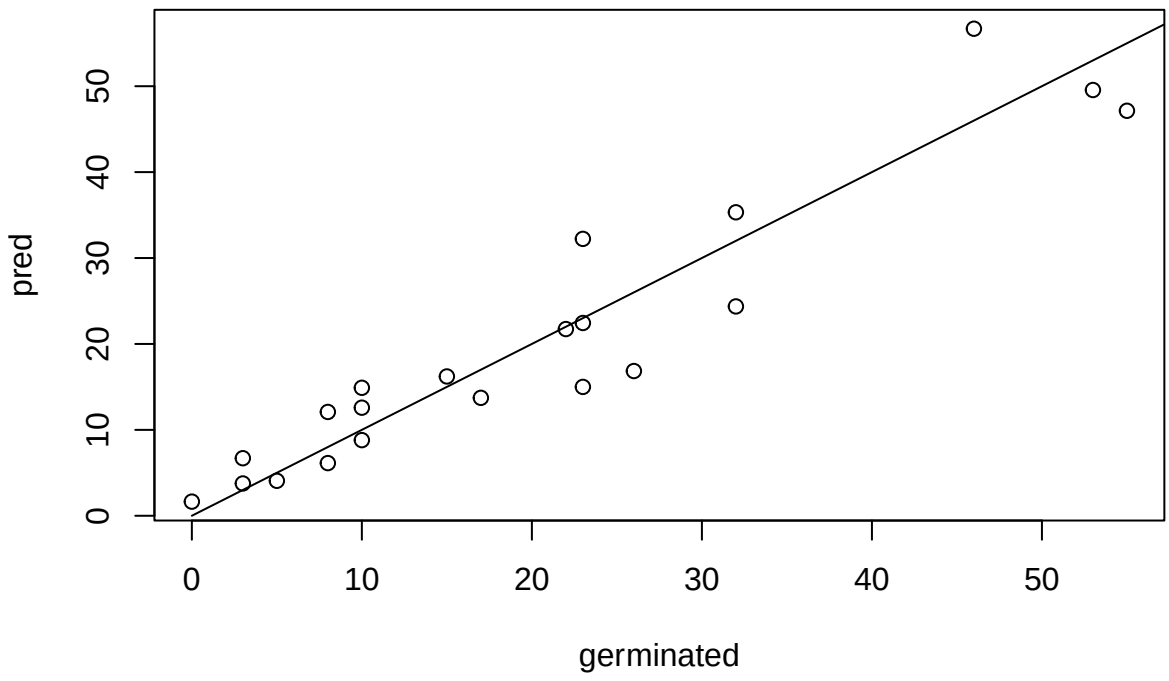
Evaluate Pearson Residuals: No apparent trend in the Pearson's residuals (Appendix 3.4).

Evaluate Simulated Residuals: Still significant departure from Uniform distribution (Appendix 3.5).

None of the alternative link functions appear to model the data better than the logistic link function.

### Predictive Performance

```
## Test model predictive power (Leave-One-Out Cross Validation)
attach(seeds)
pred <- as.numeric()
for(i in 1:length(germinated)){
  temp <- data.frame(germinated=germinated[-i],total=total[-i],extract=extract[-i],plant=plant[-i])
  yntemp <- cbind(temp$germinated,temp$total-temp$germinated)
  fit.cv <- glm(yntemp ~ extract + plant + extract*plant,family="binomial",data=temp)
  predlp <- predict(fit.cv,newdata=data.frame(extract=extract[i],plant=plant[i]))
  pred[i] <- exp(predlp)/(1+exp(predlp))*total[i]
}
plot(germinated,pred)
points(c(0,70),c(0,70),type="l")
```




There appears to be agreement between the predicted and observed data, with some very minor fanning as the response increases. Thus, it can be concluded that the proposed model is a reasonable fit to the data.

In summary, the following model was developed for these data:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$$

where $x_{1i}$ and $x_{2i}$ are indicator variables denoting varieties of extract and plant, respectively.

The model indicates a significant interaction between extract and plant type, with the combination of Oa75 and cucumber leading to, on average, a reduced number of germinated seeds. Specifically, this interaction term reduced the log odds of seed germination by, on average, 0.7781 (SE: 0.3064). Extract does not appear to be significant as a main effect, however, plant is increasing the log odds, on average, by 1.3182 (SE: 0.1775).

There is no apparent trend in the Pearson's residuals, however, some lack-of-fit was noted when a form of simulated residuals were considered which suggest that the model may not fit the data appropriately.

To further assess this, the predictive performance of the model was assessed through leave-one-out cross validation, demonstrating that the out-of-sample model predictions were unbiased and therefore suggesting the proposed model is reasonable.

### Appendix

# 1 - Logit Link Function Model

# 1.1 - No Interaction Term Model Summary

```
##
## Call:
## glm(formula = yn ~ plant + extract, family = binomial(link = "logit"),
##     data = seeds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3919  -0.9949  -0.3744   0.9831   2.4766
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.4300     0.1137  -3.781 0.000156 ***
## plantcucumber   1.0647     0.1442   7.383 1.55e-13 ***
## extractOa75    -0.2705     0.1547  -1.748 0.080435 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 39.686  on 18  degrees of freedom
## AIC: 122.28
##
## Number of Fisher Scoring iterations: 4
```

## 1.2 - With Interation Term Model Summary

```
##
## Call:
## glm(formula = yn ~ plant * extract, family = binomial(link = "logit"),
##     data = seeds)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.01617  -1.24398   0.05995   0.84695   2.12123
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.5582     0.1260  -4.429 9.46e-06 ***
## plantcucumber                1.3182     0.1775   7.428 1.10e-13 ***
## extractOa75                  0.1459     0.2232   0.654   0.5132
## plantcucumber:extractOa75   -0.7781     0.3064  -2.539   0.0111 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 33.278  on 17  degrees of freedom
## AIC: 117.87
##
## Number of Fisher Scoring iterations: 4
```

## 1.3 - With Interaction Term Overdispersed Model Summary

```
##
## Call:
## glm(formula = yn ~ plant * extract, family = quasibinomial(link = logit),
##     data = seeds)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.01617  -1.24398   0.05995   0.84695   2.12123
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -0.5582     0.1720  -3.246  0.00475 **
## plantcucumber                1.3182     0.2422   5.444 4.38e-05 ***
## extractOa75                  0.1459     0.3045   0.479  0.63789
## plantcucumber:extractOa75   -0.7781     0.4181  -1.861  0.08014 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.861832)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 33.278  on 17  degrees of freedom
```

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

# 2 - Probit Link Function Model

## 2.1 - No Interaction Term Model Summary

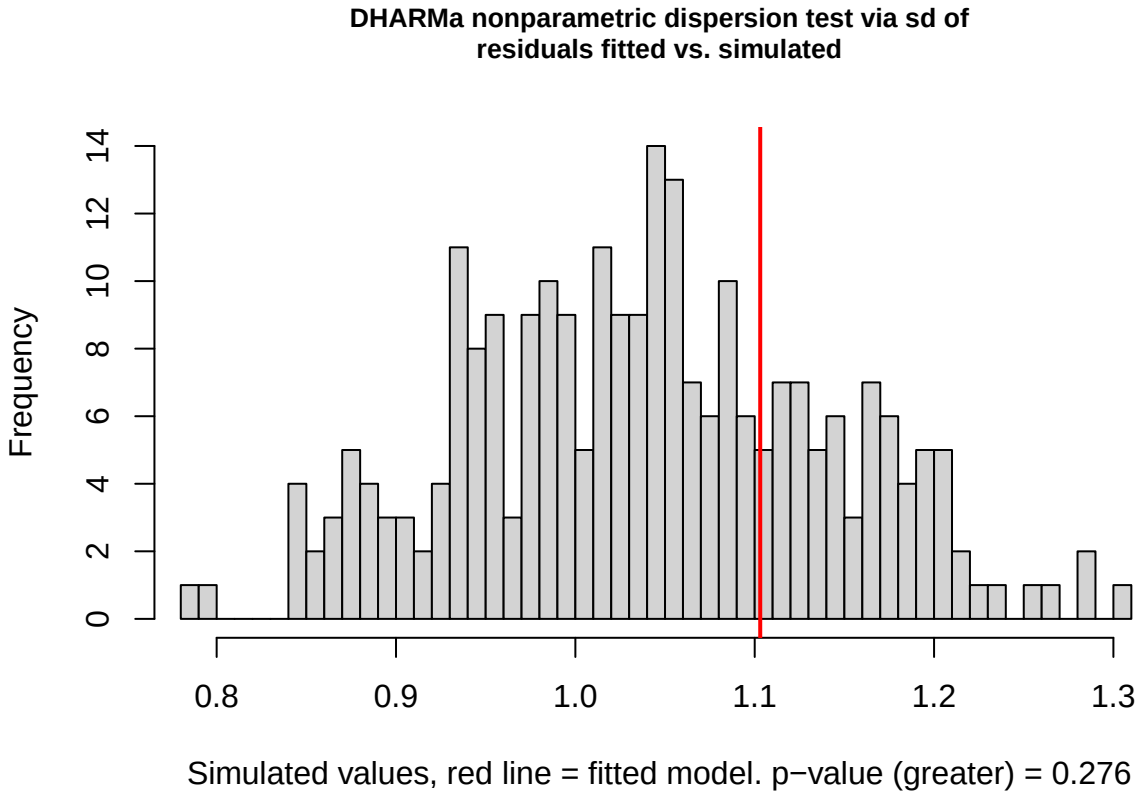
```
##
## Call:
## glm(formula = yn ~ extract + plant, family = binomial(link = "probit"),
##     data = seeds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3963  -1.0018  -0.3783   0.9827   2.4692
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.26797    0.07048  -3.802 0.000144 ***
## extractOa75   -0.16586    0.09541  -1.738 0.082148 .
## plantcucumber  0.66316    0.08894   7.456 8.91e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 39.702  on 18  degrees of freedom
## AIC: 122.3
##
## Number of Fisher Scoring iterations: 4
```

## 2.2 - With Interaction Term Overdispersed Model Summary

```
##
## Call:
## glm(formula = yn ~ extract * plant, family = binomial(link = "probit"),
##     data = seeds)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.01617  -1.24398   0.05995   0.84695   2.12123
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -0.34787    0.07769  -4.478 7.54e-06 ***
## extractOa75                 0.09031    0.13827   0.653   0.5137
## plantcucumber               0.81936    0.10868   7.539 4.72e-14 ***
## extractOa75:plantcucumber -0.48172    0.18989  -2.537   0.0112 *
```

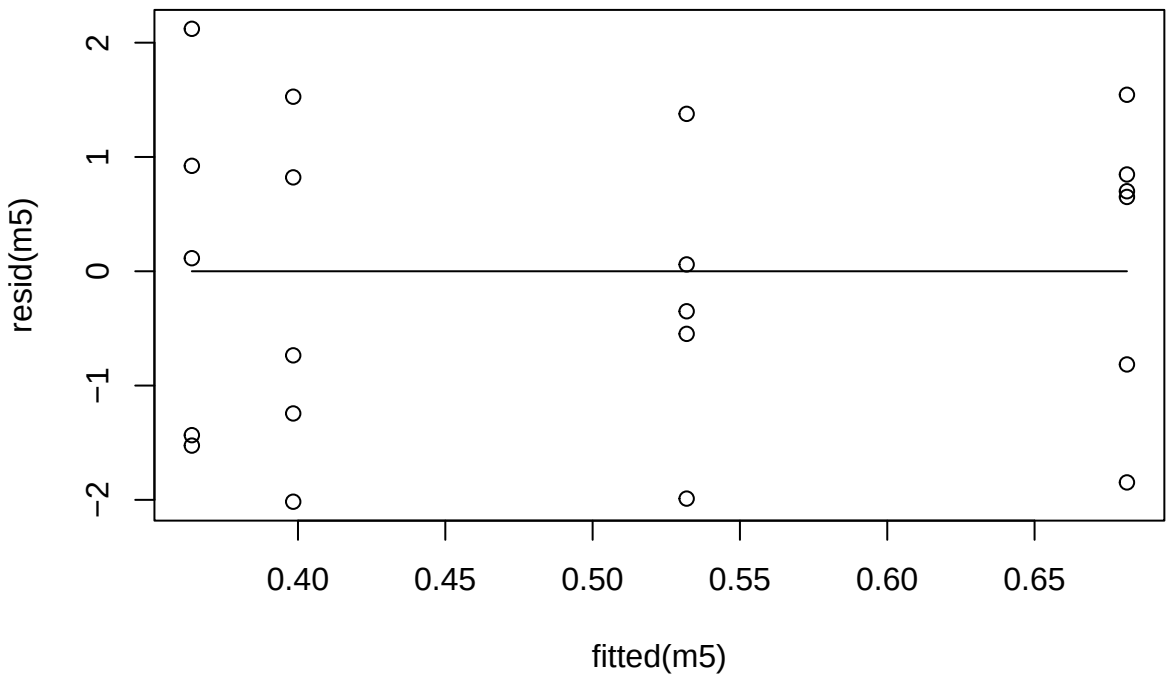
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 33.278  on 17  degrees of freedom
## AIC: 117.87
##
## Number of Fisher Scoring iterations: 3
```

## 2.3 - Overdispersion Test Results

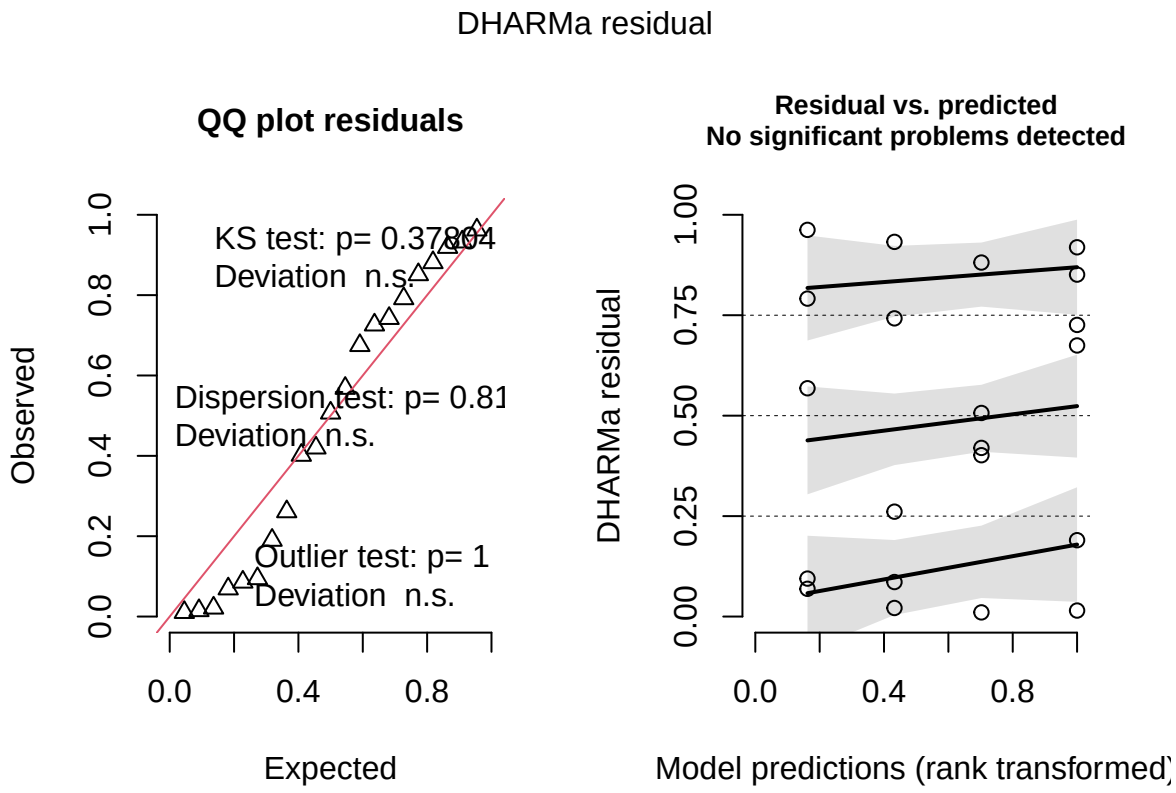



```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.0611, p-value = 0.276
## alternative hypothesis: greater
```

## 2.4 - Pearson Residuals



## 2.5 - Simulated Residuals




# 3 - Complementary log-log Link Function Model

## 3.1 - No Interaction Term Model Summary

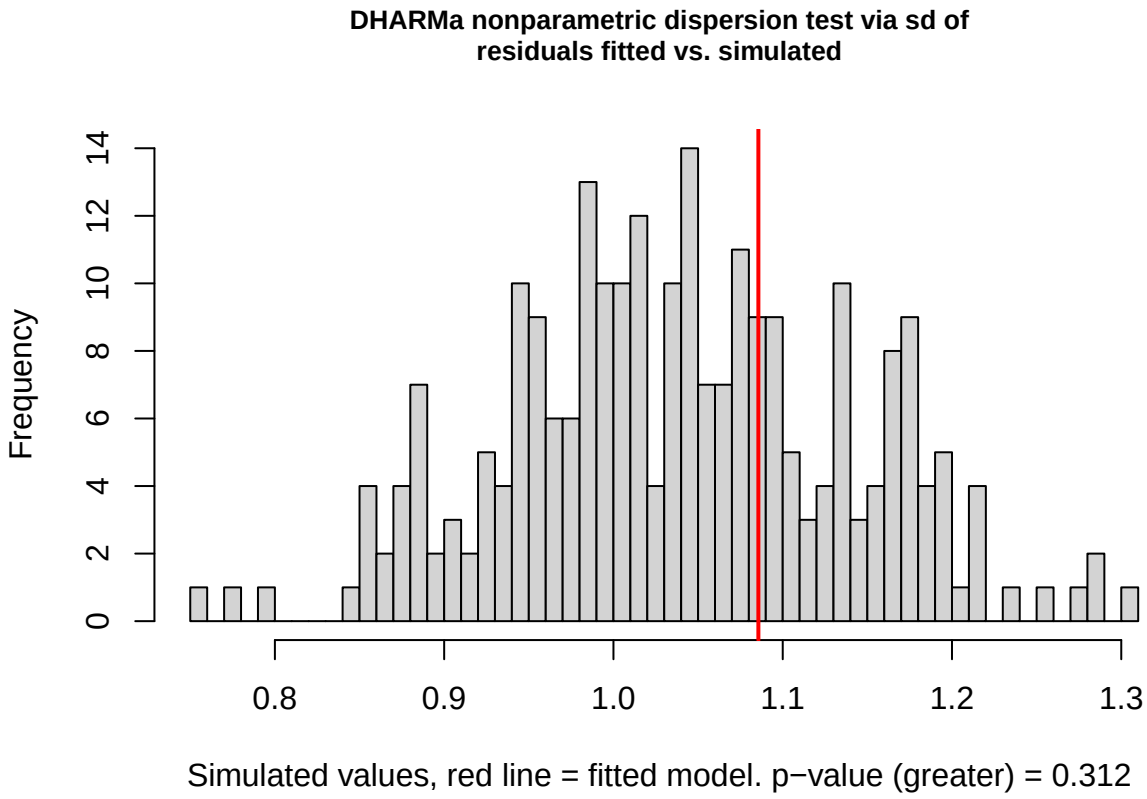
```
##
## Call:
## glm(formula = yn ~ plant + extract, family = binomial(link = "cloglog"),
##     data = seeds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3206  -0.8813  -0.3608   0.9501   2.4933
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.69380    0.08849  -7.841 4.48e-15 ***
## plantcucumber  0.77066    0.10417   7.398 1.38e-13 ***
## extractOa75   -0.21918    0.11004  -1.992   0.0464 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 38.653  on 18  degrees of freedom
## AIC: 121.25
##
## Number of Fisher Scoring iterations: 4
```

## 3.2 - With Interaction Term Model Summary

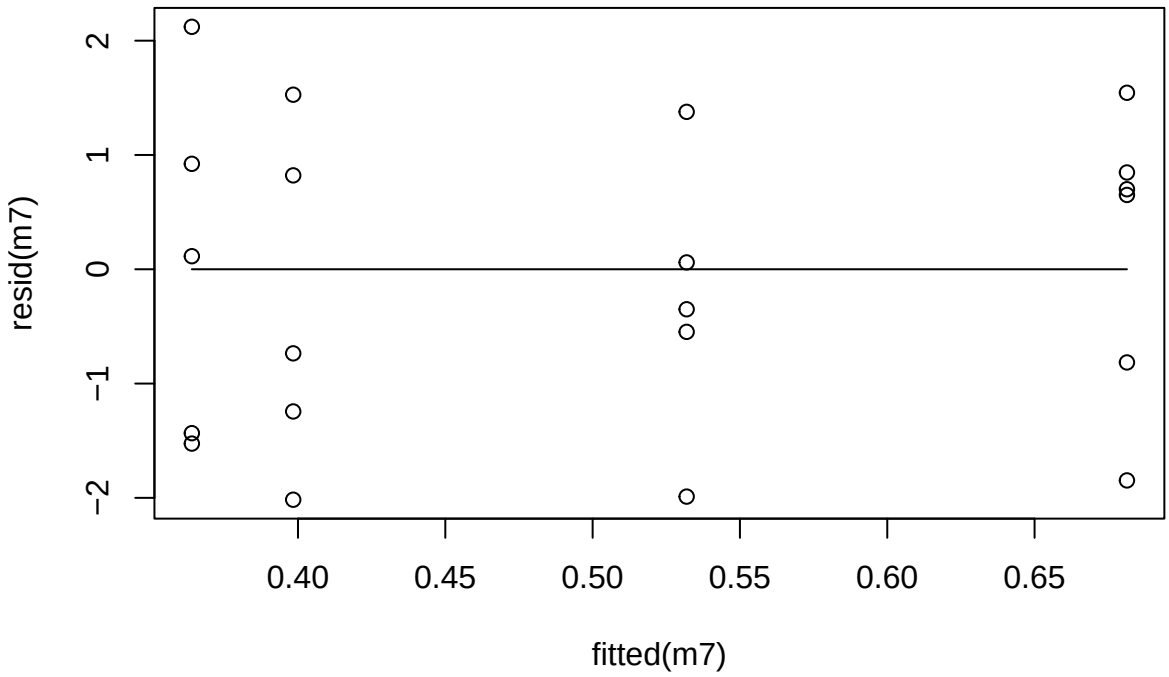
```
##
## Call:
## glm(formula = yn ~ plant * extract, family = binomial(link = "cloglog"),
##     data = seeds)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.01617  -1.24398   0.05995   0.84695   2.12123
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.7929     0.1014  -7.823 5.17e-15 ***
## plantcucumber                0.9272     0.1258   7.373 1.67e-13 ***
## extractOa75                  0.1159     0.1764   0.657   0.5112
## plantcucumber:extractOa75   -0.5258     0.2251  -2.336   0.0195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 33.278  on 17  degrees of freedom
## AIC: 117.87
##
## Number of Fisher Scoring iterations: 4
```

## 3.3 - Overdispersion Test Results




```
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 1.0446, p-value = 0.312
## alternative hypothesis: greater
```

### 3.4 - Pearson Residuals



## 3.5 - Simulated Residuals