Analysing Pollination Syndromes Among Pea Species in Western Australia

Statistical Methods

There are many pea species endemic to Australia, with the Southwest of Western Australia a hotspot for pea diversity. Of interest is the adaptations of the structure of some pea flowers which appear adapted to bird pollination while other close relatives appear adapted to insect pollination, predominantly bees. Scientists have been intrigued by pinpointing distinct pollination "syndromes" (insect vs bird vs mammal) associated with insects, birds, or mammals. This fascination stems from the belief that a comprehensive grasp of these syndromes serves as proof of adaptive evolution. The dataset we are using was created by researchers conducted a survey on various species of peas in Western Australia, examining those pollinated by birds and bees. They recorded the number of seeds (peas) in a selection of pods (technically termed "funicles"), collecting multiple samples from numerous plants across several species spanning various mountains.

The dataset contains 7 variables including number of funicles in the pod ("Funicle Number", numeric/integer), number of seeds in the pod ("Seed", numeric/integer), the species of pea ("Species", categorical), the pollination syndrome of the plant ("Syndrome", categorical), the name of the mountain that the plant was located ("Mountain", categorical), the ID of the individual plant ("Individual", character), and the method ("Method", categorical). On inspection of the dataset there were various issues relating to the cleanliness of the data which had to be addressed. Firstly observations with NAs were removed and observations with "unknown" syndrome were also removed as they wouldn't contribute anything to the analysis. There were also errors in funicle number measurements for numerous observations with a decimal number recorded. These observations were also removed from the dataset leaving a sample size of 1635. Using this data we attempt to explore the relationship between seed production and the different types of pollination syndromes while accounting for random effects.

Consideration was given to models using funicle number (potential for seeds in a pod) and seed as the response variable. Models using funicle number showed a large amount of overdispersion, raising questions over model fit while models using seed (number of seeds) as the response variable proved to be a better fit. There was still evidence of overdispersion within the models which I tried to account for by using different distributions including poisson models and negative binomial models. In the final model I tried to capture 2 different random effects, variation between the mountains with species nested within mountains (not all species grow on all mountains) and the variation in individuals (multiple measures for each individual). Adding species as an explanatory variable alongside syndrome were trialed but removing species provided better model fit probably due to the high correlation between the 2 variables. The final model was represented as

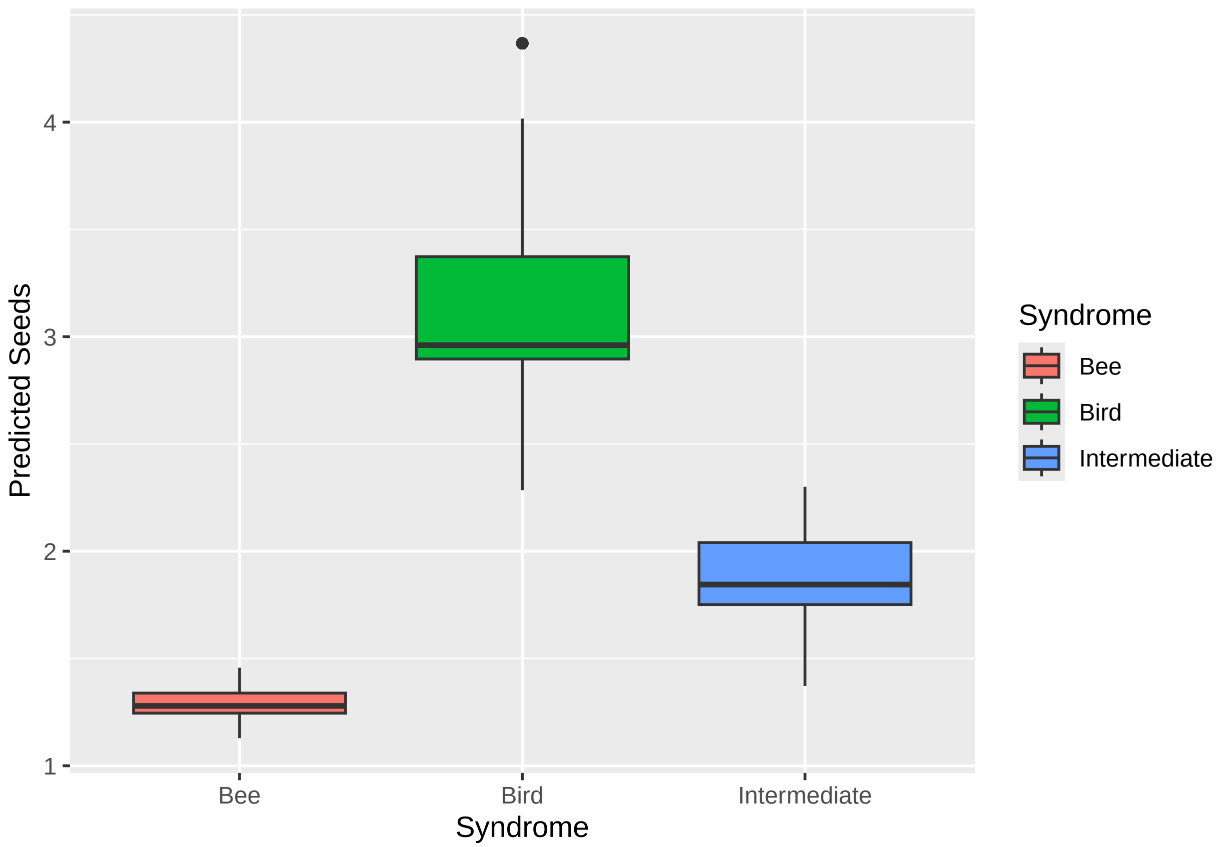
$$\text{Seeds} \sim \text{Syndrome } + (1|\text{Mountain,Species}) + (1|\text{Individual})$$

Results

The model was fit using the glmer() function from the lme4 package with the family argument set poisson. The model produces the following summary information:

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: Seed ~ Syndrome + (1 | Mountain/Species) + (1 | Individual)
##    Data: Peas_clean
##
##      AIC      BIC   logLik deviance df.resid
##   4811.2   4843.6  -2399.6   4799.2     1612
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7797 -0.5387 -0.1724  0.5505  2.8080
##
## Random effects:
##  Groups           Name        Variance  Std.Dev.
##  Individual       (Intercept) 0.0273156 0.16527
##  Species:Mountain (Intercept) 0.0000000 0.00000
##  Mountain         (Intercept) 0.0006884 0.02624
## Number of obs: 1618, groups:
## Individual, 179; Species:Mountain, 10; Mountain, 5
##
## Fixed effects:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.24981    0.04966   5.031 4.88e-07 ***
## SyndromeBird          0.84743    0.06708  12.632  < 2e-16 ***
## SyndromeIntermediate  0.37709    0.05712   6.602 4.05e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation of Fixed Effects:
##             (Intr) SyndrB
## SyndromeBrd -0.691
## SyndrmIntrm -0.789  0.565
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help(’isSingular’)
```

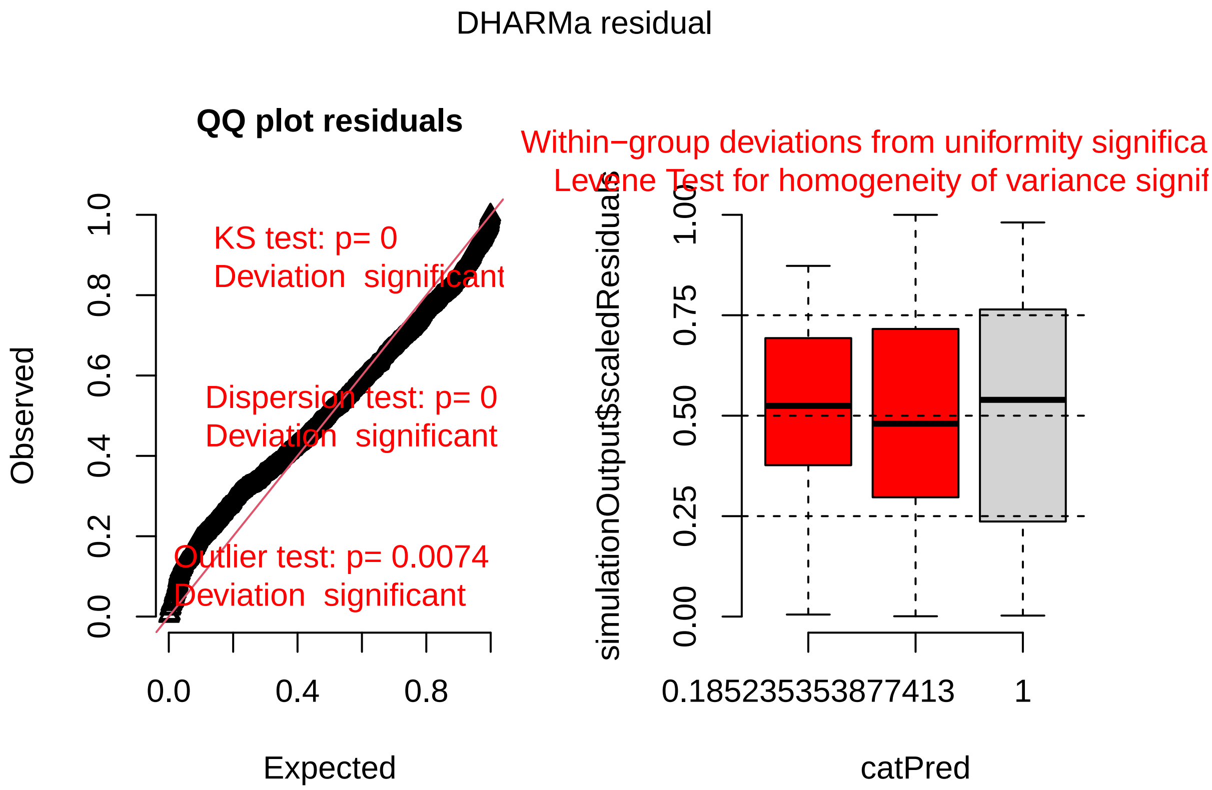
From the above summary you can see the intercept, which in this case relates to the bee pollination syndrome, has a coefficient of 0.24979. If the plant instead is classified with intermediate pollination syndrome the estimate for seed production log count rises by 0.37712, and if the plant is classified with bird pollination syndrome the estimate rises by 0.84745 from the reference or intercept group. Exponentiating these coefficients we get estimates for seed production as 1.28 for the intercept, an increase of 1.46 for intermediate pollination syndrome and an increase of 2.33 for the bird pollination syndrome. These results suggest that pea species that use bird pollination syndrome produce the most amount of peas per pod while those that use bee pollination produce the least. The fixed effects can be visualised using the following plot:



Looking at the coefficients of the random effects for the model you can see that the random effects didn't contribute much to the model with a total of roughly 0.028 variance explained by the random effects. The random effects contained within different pods of the same individual (0.02732) was by far the most out of the random effects.

When assessing model assumptions the DHARMa pakage was used. The following plot shows the results from this.

```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```



```
## Object of Class DHARMa with simulated residuals based on 250 simulations with refit = FALSE . See ?DH
##
## Scaled residual values: 0.225297 0.6836163 0.5820531 0.3785728 0.2104553 0.1266121 0.6377677 0.17632
```

As you can see above, the QQ residual plot shows some deviation from the line, especially along the tails, providing evidence of overdispersion. The other plot provided also shows some issues arising from within group deviations from uniformity suggesting issues with homogeneity of variance. Although these issues have been deemed significant, I'm happy enough with the model however this should be taken into account when drawing conclusions from the model.

Problems encountered

One of the problems encountered during this project was inaccurate and incomplete data. As alluded to in the statistical methods section, from the original dataset of 2013 observations 399 observations were removed due to NAs, unknown syndrome types, and inaccurate measurements. There was evidence of overdispertion in the data even when trying to account for it using different distributions for the response variable and trying different link functions. There was also evidence of issues with homogeneity of variance which should be taken into account when inferring from the model. The final model didn't show much in the way of random effects, especially through the species of the pea and mountain the plant was located, and so maybe trying a model that only includes individuals as a random effect would have been sufficient.