Report

Hao He, Haochen Pan, Xu Su, Yeyang Han

2023-04-27

Introduction

Our client is from the Anthropology Department at the Boston University CAS School, and her capstone project focuses on the social behavior and endocrinology of wild orangutans. In her research paper, she examines the impact of multiple factors on the rate of specific behaviors under different social conditions. This report aims to demonstrate the process of performing exploratory data analysis using visualizations to identify patterns in orangutan behavior; to assess the effectiveness of our client's current model, as described in her paper; and to explore alternative methods for potential improvements.

Data Description

Data collection was completed in an Indonesian's rainforest and lasted for a year, from 2014 to 2015. The client with her team observed focal animals and recorded their SDB (Self-Directed Behavior) with records of multiple important variables: **SDB Rate**: self-directed behavior, a rate recorded as #occurrences/10 minutes **Age-Sex of Focal** (4 types of age and sex groups): Adult Female, Adolescent Female, Flanged Male, and Unflanged Male **Social Y or N**: Has the social event occurred **Partner x Name/ Partner x Age-sex**: The ID and an age-sex group of participants in the social event

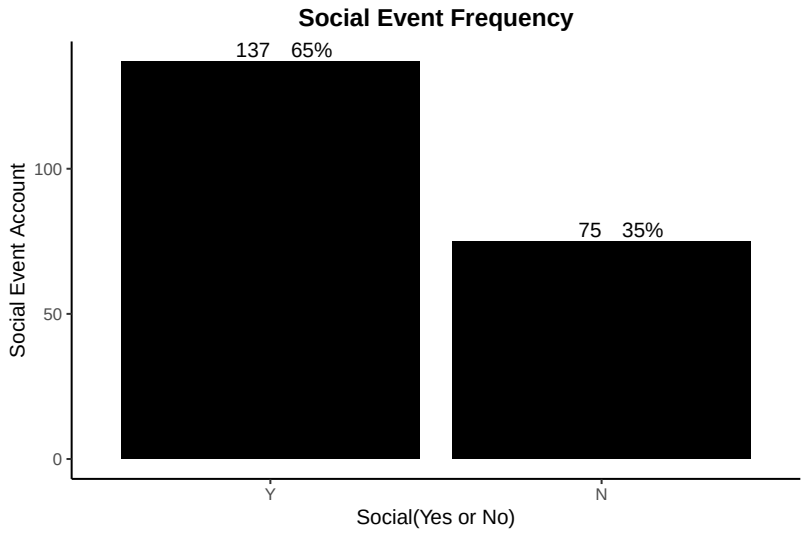
EDA

1. Preprocess data

In the data preparation part, we ensured that both flanged and unflanged male individuals were included as fully mature adults, while also converting the SDB rate to SDB count and storing it in a separate variable. We can further analyze the relationships between variables by utilizing EDA plots first.

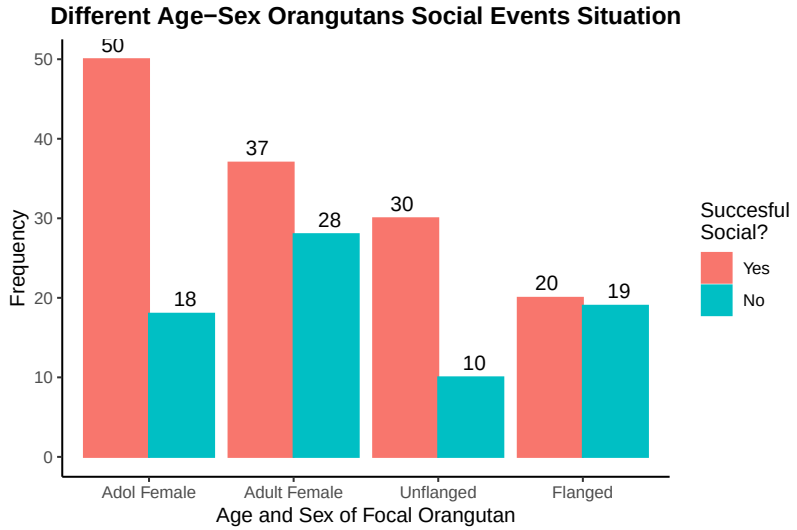
2. Social Event Frequency

The bar plot clearly indicates that the frequency of successful social events is significantly higher at 65%, as compared to the frequency of unsuccessful social events. This suggests that the factors contributing to the success of social events might be different from those leading to an unsuccessful outcome.



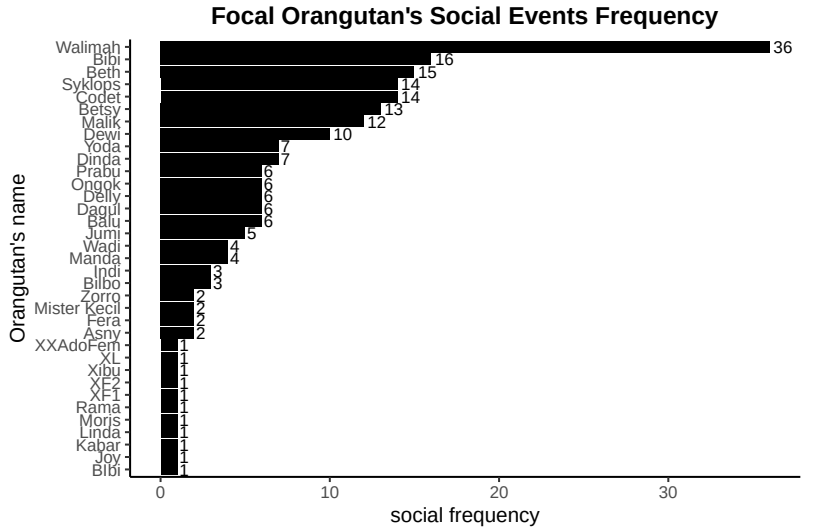
3. Different Age-Sex Orangutans Social Events Situation

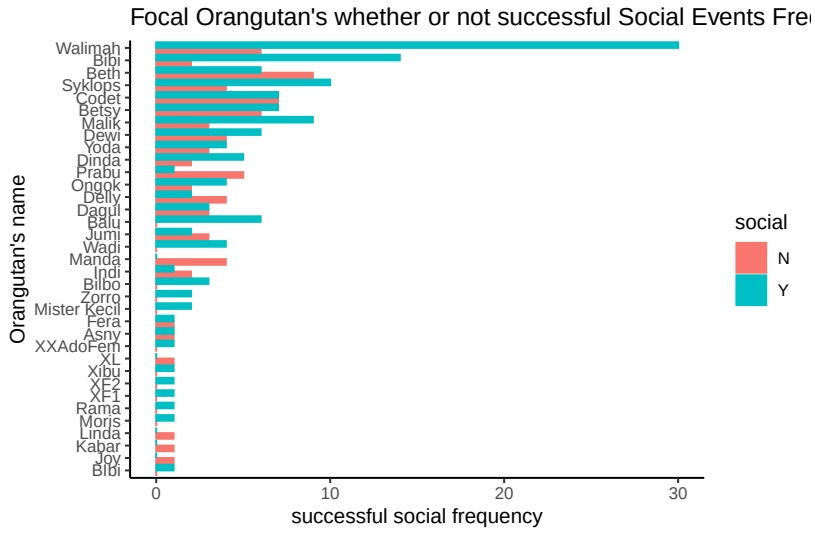
To gain more insights into the frequency of successful social events, we created a new bar plot that separates the data by age and sex groups. Interestingly, the plot reveals that adolescent females have the highest frequency of successful social events.



4.Focal Orangutan's Social Events Frequency

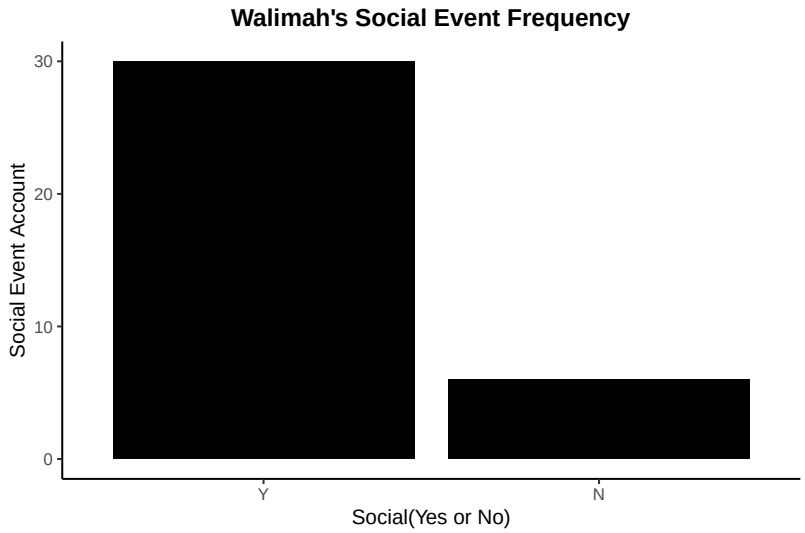
The bar plot represents the distribution of social frequency for each focal orangutan, with a notable finding that Walimah, the adolescent female orangutan, has the highest frequency and success rate compared to the other focal orangutans. This finding suggests that Walimah might play a more significant role in the social dynamics of the group, potentially influencing social behaviors and interactions among other orangutans.



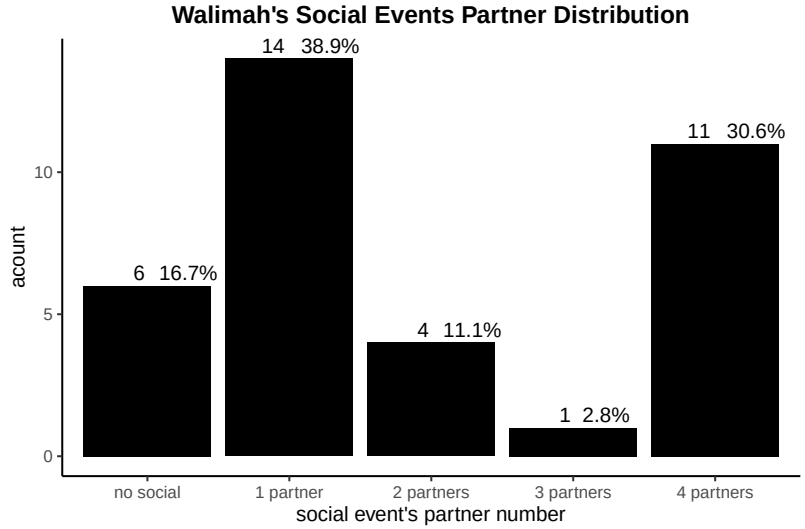


5. Walimah

Our analysis was further focused on Walimah, and we created a new bar plot to explore the frequency of successful social events for her. The plot clearly indicates that Walimah has a significantly higher rate of successful social events than other orangutans.



We also created a new bar plot to explore the partners and frequency of orangutans to Walimah. We noticed that even the 1 partner shows the most frequency, there are always situations that there are about 4 partners involved in the social event, which may be another reason that Walimah is a special case to see.



Modeling

Model replication

To model the count data of SDBs per 10-minute interval, Generalized Linear Mixed Model (GLMM) with a Poisson distribution were initially fitted, taking into account the discrete nature of the data. However, convergence issues were observed with the `glmer()` function from the `lme4` package. To overcome this, we recommended using Stan for Bayesian modeling, which provided a flexible and computationally efficient framework for fitting complex hierarchical models with count data, and allowed for more accurate estimation of model parameters and uncertainty quantification.

To address the client's concern regarding the influence of various factors, we first fitted GLMMs using the two-way interaction between focal age-sex class and social state (social or solitary) with a Poisson distribution.

However, overdispersion in the data prompted us to use the negative binomial GLMM, which provided a better fit compared to the Poisson GLMM by better capturing the variability in the count data. A subsequent sanity check was performed on the negative binomial GLMM, including a three-way interaction between focal age-sex class, social state (social or solitary), and count of social partners for each age-sex class.

Poisson

```
## stan_glmer
##  family:       poisson [log]
##  formula:      SDB ~ AgeSexFocal * SocialYN + (1 | FocalID) + (1 | EventID)
##  observations: 1533
## ------
##                                   Median MAD_SD
## (Intercept)                       -0.072  0.290
## AgeSexFocalAdult Female           -1.458  0.436
## AgeSexFocalFlanged                -1.211  0.465
## AgeSexFocalUnflanged              -1.137  0.452
## SocialYNN                         -0.750  0.339
## AgeSexFocalAdult Female:SocialYNN  0.892  0.488
## AgeSexFocalFlanged:SocialYNN       0.975  0.556
## AgeSexFocalUnflanged:SocialYNN     0.248  0.658
##
## Error terms:
##  Groups  Name        Std.Dev.
##  EventID (Intercept) 0.9922
##  FocalID (Intercept) 0.5538
## Num. levels: EventID 213, FocalID 36
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

Negative Binomial

```
## stan_glmer.nb
##  family:       neg_binomial_2 [log]
##  formula:      SDB ~ AgeSexFocal * SocialYN + (1 | FocalID) + (1 | EventID)
##  observations: 1533
## ------
##                                   Median MAD_SD
## (Intercept)                        0.098  0.287
## AgeSexFocalAdult Female           -1.460  0.432
## AgeSexFocalFlanged                -1.275  0.457
## AgeSexFocalUnflanged              -1.231  0.414
## SocialYNN                         -0.760  0.326
## AgeSexFocalAdult Female:SocialYNN  0.834  0.498
## AgeSexFocalFlanged:SocialYNN       0.910  0.544
## AgeSexFocalUnflanged:SocialYNN     0.266  0.644
##
## Auxiliary parameter(s):
##                       Median MAD_SD
## reciprocal_dispersion 0.659  0.082
##
```

```
## Error terms:
##  Groups  Name        Std.Dev.
##  EventID (Intercept) 0.8170
##  FocalID (Intercept) 0.5466
## Num. levels: EventID 213, FocalID 36
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

Model diagnosis

Dispersion

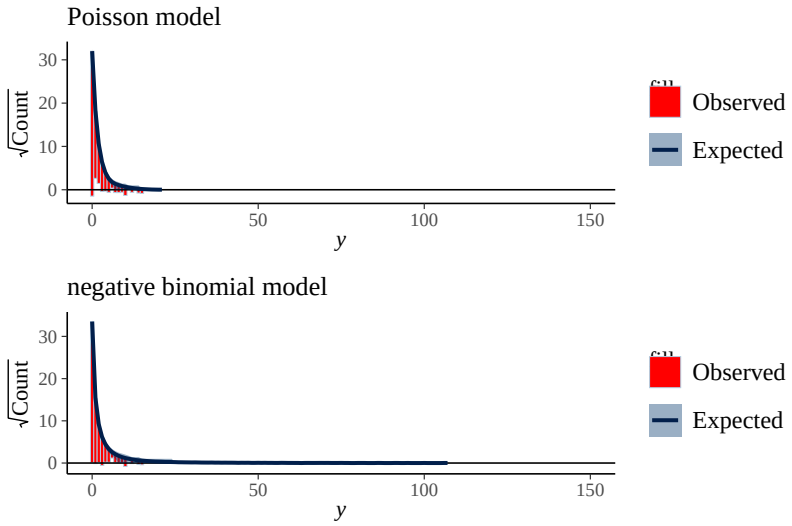
```
## # Overdispersion test
##
##        dispersion ratio =    1.379
##   Pearson's Chi-Squared = 2103.454
##                 p-value =  < 0.001

## Overdispersion detected.

## # Overdispersion test
##
##        dispersion ratio =    1.516
##   Pearson's Chi-Squared = 2310.307
##                 p-value =  < 0.001

## Overdispersion detected.
```

Based on the summary of both models, the Poisson model was found to have overdispersion as the dispersion ratio was approximately 1.36, indicating that the variance of the data is greater than the mean. To address this issue, we fitted a GLMM using a negative binomial distribution instead. The negative binomial model has a dispersion parameter of approximately 0.77, which is expected to be lower than that of the Poisson model.

Rootogram

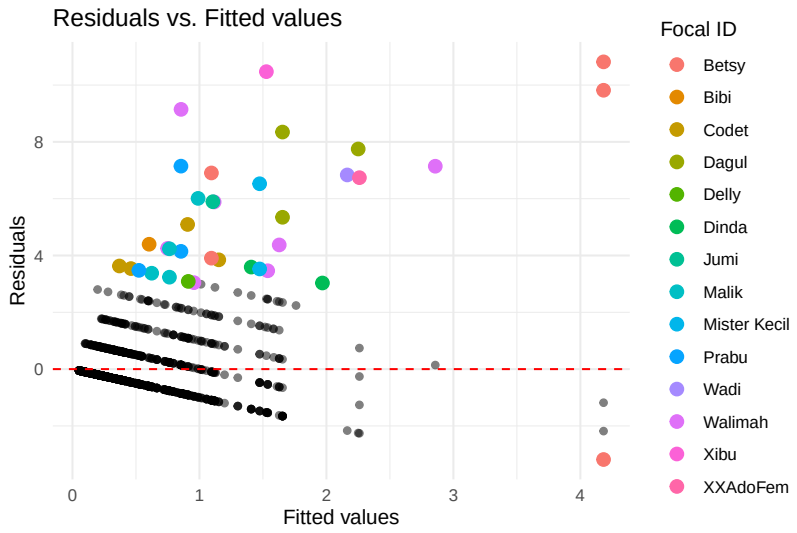
However, the use of residuals plots alone is not a good indicator of model fit for GLMMs with count data. As an improved approach to assess the fit of a count regression model, Rootgrams are proposed. Below is a side-by-side Rootogram comparison of the Poisson and negative binomial models.

```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```



Because this is a hanging rootogram, the rootogram can be thought of as relating to the fitted counts — if a bar doesn't reach the zero line then the model over predicts a particular count bin, and if the bar exceeds the zero line it under predicts. Based on the analysis, it is evident that the Poisson model is underfitting some of the data points, resulting in negative residuals in the rootogram. On the other hand, the Negative Binomial model shows fewer data points deviating from the expected frequency, indicating a better fit with less underfitted data points. Therefore, the Negative Binomial model is recommended as a better fit based on the dispersion and rootogram assessment tools. Based on the results of the negative binomial model, it is evident that, holding the social state of the focal orangutan constant, adult orangutans tend to have a higher expected SDB compared to the expected SDB of adolescent females.

However, it should be noted that both models suggest potential outliers or influential observations in the data, as indicated by the residual plots and QQ plot. To investigate this further and understand their impact on the model, the original data points corresponding to the residuals with large deviations from the expected frequencies were examined using the Negative Binomial model. A threshold of 3 was chosen manually for large deviations in residuals. Among the 36 outliers identified, the focal Wahlimah was found to have the highest counts (7) and may be a dominant factor that impacts the model (See Table 1 in Appendix). This finding is consistent with the previous exploratory data analysis process.




```
## fixed-effect model matrix is rank deficient so dropping 4 columns / coefficients
```

```
## stan_glmer.nb
##  family:       neg_binomial_2 [log]
##  formula:      SDB ~ AgeSexFocal * SocialYN * `#AdoFemPresent` + (1 | FocalID) +
##     (1 | EventID)
##  observations: 1533
## ------
##                                          Median MAD_SD
## (Intercept)                               0.374  0.331
## AgeSexFocalAdult Female                  -1.878  0.515
## AgeSexFocalFlanged                       -2.282  0.619
## AgeSexFocalUnflanged                     -2.258  0.593
## SocialYNN                                -0.972  0.354
## `#AdoFemPresent`                         -0.551  0.284
## AgeSexFocalAdult Female:SocialYNN         1.209  0.530
## AgeSexFocalFlanged:SocialYNN              1.918  0.658
## AgeSexFocalUnflanged:SocialYNN            1.235  0.781
## AgeSexFocalAdult Female:`#AdoFemPresent`  0.745  0.400
## AgeSexFocalFlanged:`#AdoFemPresent`       1.529  0.538
## AgeSexFocalUnflanged:`#AdoFemPresent`     1.673  0.656
##
## Auxiliary parameter(s):
##                       Median MAD_SD
## reciprocal_dispersion 0.663  0.083
##
## Error terms:
##  Groups  Name        Std.Dev.
##  EventID (Intercept) 0.7967
##  FocalID (Intercept) 0.5131
## Num. levels: EventID 213, FocalID 36
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg

## fixed-effect model matrix is rank deficient so dropping 4 columns / coefficients

## stan_glmer.nb
##  family:       neg_binomial_2 [log]
##  formula:      SDB ~ AgeSexFocal * SocialYN * `#UnflangedPresent` + (1 | FocalID) +
##     (1 | EventID)
##  observations: 1533
## ------
##                                             Median MAD_SD
## (Intercept)                                  0.115  0.348
## AgeSexFocalAdult Female                     -1.451  0.482
## AgeSexFocalFlanged                          -1.400  0.545
## AgeSexFocalUnflanged                        -1.316  0.513
## SocialYNN                                   -0.792  0.365
## `#UnflangedPresent`                         -0.107  0.405
## AgeSexFocalAdult Female:SocialYNN            0.834  0.523
## AgeSexFocalFlanged:SocialYNN                 1.063  0.569
## AgeSexFocalUnflanged:SocialYNN               0.363  0.716
## AgeSexFocalAdult Female:`#UnflangedPresent`  0.008  0.658
## AgeSexFocalFlanged:`#UnflangedPresent`       0.703  0.938
```

```
## AgeSexFocalUnflanged:`#UnflangedPresent`     0.554  0.808
##
## Auxiliary parameter(s):
##                       Median MAD_SD
## reciprocal_dispersion 0.657  0.079
##
## Error terms:
##  Groups  Name        Std.Dev.
##  EventID (Intercept) 0.8256
##  FocalID (Intercept) 0.5974
## Num. levels: EventID 213, FocalID 36
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg

## fixed-effect model matrix is rank deficient so dropping 4 columns / coefficients

## stan_glmer.nb
##  family:       neg_binomial_2 [log]
##  formula:      SDB ~ AgeSexFocal * SocialYN * `#AdultFemPreent` + (1 | FocalID) +
##     (1 | EventID)
##  observations: 1533
## ------
##                                          Median MAD_SD
## (Intercept)                               0.104  0.357
## AgeSexFocalAdult Female                  -1.539  0.513
## AgeSexFocalFlanged                       -0.991  0.579
## AgeSexFocalUnflanged                     -1.120  0.542
## SocialYNN                                -0.769  0.372
## `#AdultFemPreent`                        -0.008  0.281
## AgeSexFocalAdult Female:SocialYNN         0.925  0.556
## AgeSexFocalFlanged:SocialYNN              0.649  0.642
## AgeSexFocalUnflanged:SocialYNN            0.170  0.700
## AgeSexFocalAdult Female:`#AdultFemPreent`  0.152  0.527
## AgeSexFocalFlanged:`#AdultFemPreent`      -0.558  0.654
## AgeSexFocalUnflanged:`#AdultFemPreent`    -0.146  0.485
##
## Auxiliary parameter(s):
##                       Median MAD_SD
## reciprocal_dispersion 0.656  0.081
##
## Error terms:
##  Groups  Name        Std.Dev.
##  EventID (Intercept) 0.8331
##  FocalID (Intercept) 0.5604
## Num. levels: EventID 213, FocalID 36
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg

## fixed-effect model matrix is rank deficient so dropping 6 columns / coefficients
```

```
## stan_glmer.nb
##  family:       neg_binomial_2 [log]
##  formula:      SDB ~ AgeSexFocal * SocialYN * `#FlangedPresent` + (1 | FocalID) +
##     (1 | EventID)
##  observations: 1533
## ------
##                                            Median MAD_SD
## (Intercept)                                 0.002  0.296
## AgeSexFocalAdult Female                    -1.665  0.462
## AgeSexFocalFlanged                         -1.175  0.457
## AgeSexFocalUnflanged                       -1.119  0.439
## SocialYNN                                  -0.657  0.352
## `#FlangedPresent`                           0.406  0.418
## AgeSexFocalAdult Female:SocialYNN           1.033  0.530
## AgeSexFocalFlanged:SocialYNN                0.814  0.539
## AgeSexFocalUnflanged:SocialYNN              0.130  0.656
## AgeSexFocalAdult Female:`#FlangedPresent`   0.599  0.719
##
## Auxiliary parameter(s):
##                        Median MAD_SD
## reciprocal_dispersion 0.660  0.085
##
## Error terms:
##  Groups  Name        Std.Dev.
##  EventID (Intercept) 0.8245
##  FocalID (Intercept) 0.5061
## Num. levels: EventID 213, FocalID 36
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

We fitted a series of more sophisticated negative binomial models to examine the effect of social partners on the SDB use the three-way interactions between the focal orangutan's age-sex class, social state, and the count of social partners of each age-sex class.

In all four models, the main effect of SocialYN consistently shows a negative relationship between the presence of social partners (Y) and the rate of SDB. This suggests that, on average, focal orangutans exhibit lower rates of SDB when social partners are present, compared to when they are absent. This general observation is not consistent with client's predictions.

In complex model 1, 2, and 3, adult females and flanged males show a positive interaction with SocialYNN, indicating that the reduction in SDB when social partners are present is less pronounced for these age-sex classes. In model 4, the interaction is positive for adult females but weaker for flanged males.

In summary, the presence of social partners generally appears to be associated with lower rates of self-directed behavior across different age-sex classes of focal orangutans. However, the influence of specific social partners varies depending on the age-sex class of the focal orangutan and the presence or absence of other social partners.

Model Extensions

In addition to replicating the client's model, we also fitted new models to investigate the influence of the buffer effect and the impact of a specific focal orangutan, Wahlimah, on the rate of self-directed behavior.

Buffer Effect in Adolescent Female Orangutans In this case, the buffer effect refers to the expectation that adolescent female orangutans will have lower SDB rates when socializing with both other adolescent and adult social partners, while being associated with elevated SDB rates when socializing with only adult social partners. Two models were fitted to investigate this effect: the first model included adolescent social partners, while the second model included both adolescent and adult social partners.

```
## stan_glmer.nb
##  family:       neg_binomial_2 [log]
##  formula:      SDB ~ at_least_one_adol + (1 | FocalID) + (1 | EventID)
##  observations: 353
## ------
##                       Median MAD_SD
## (Intercept)           -0.074  0.191
## at_least_one_adolTRUE -0.210  0.388
##
## Auxiliary parameter(s):
##                       Median MAD_SD
## reciprocal_dispersion 0.565  0.109
##
## Error terms:
##  Groups  Name        Std.Dev.
##  EventID (Intercept) 0.7093
##  FocalID (Intercept) 0.2797
## Num. levels: EventID 69, FocalID 8
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
## Warning: There were 2 divergent transitions after warmup. See
## https://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
## to find out why this is a problem and how to eliminate them.
```

```
## Warning: Examine the pairs() plot to diagnose sampling problems
```

```
## stan_glmer.nb
##  family:       neg_binomial_2 [log]
##  formula:      SDB ~ at_least_one_adol + at_least_one_adult + (1 | FocalID) +
##     (1 | EventID)
##  observations: 353
## ------
##                        Median MAD_SD
## (Intercept)            -0.445  0.259
## at_least_one_adolTRUE  -0.381  0.394
## at_least_one_adultTRUE  0.653  0.288
##
## Auxiliary parameter(s):
##                       Median MAD_SD
## reciprocal_dispersion 0.563  0.100
##
## Error terms:
##  Groups  Name        Std.Dev.
##  EventID (Intercept) 0.6637
```

```
##  FocalID (Intercept) 0.2410
## Num. levels: EventID 69, FocalID 8
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

In the first model (buffer.adol.glmm), the coefficient for the predictor at_least_one_adol is negative (-0.192). This suggests that when adolescent female orangutans socialize with at least one adolescent social partner, their SDB rate is lower than when they do not socialize with any adolescent social partners. Although the standard deviation for this coefficient is relatively large, there is still some support for the notion that the coefficient is non-zero, and its interpretation aligns with our expectations based on previous research.

In the second model (buffer.both.glmm), the coefficient for at_least_one_adol is also negative (-0.389), and the coefficient for at_least_one_adult is positive (0.658). This suggests that socializing with at least one adult social partner is associated with an elevated SDB rate in adolescent female orangutans, while socializing with at least one adolescent social partner still appears to have a buffering effect, reducing the rate of SDB. Again, the standard deviation for these coefficients is relatively large, and there is not strong evidence that they are non-zero. However, their interpretation makes sense in the context of the buffer effect hypothesis.

In short, these results provide some support for the buffer effect, but it is important to note that the evidence for non-zero coefficients is not particularly strong. Nonetheless, the coefficients' interpretations align with our expectations, and further research may be needed to confirm their significance.

Walimah-specific Model To explore if the buffer effect exists in Walimah, a specific model was built based on the buffer effect model using a subset of Walimah's data. The random effect of focal ID was removed in this model to focus solely on the potential buffer effect in Walimah.

```
## stan_glmer.nb
##  family:       neg_binomial_2 [log]
##  formula:      SDB ~ at_least_one_adol + at_least_one_adult + (1 | EventID)
##  observations: 171
## ------
##                          Median MAD_SD
## (Intercept)               0.124  0.245
## at_least_one_adolTRUE    -1.539  0.750
## at_least_one_adultTRUE -0.054  0.306
##
## Auxiliary parameter(s):
##                         Median MAD_SD
## reciprocal_dispersion 0.761  0.190
##
## Error terms:
##  Groups  Name        Std.Dev.
##  EventID (Intercept) 0.3618
## Num. levels: EventID 36
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

The model results can be interpreted as follows: 1. The coefficient for `at_least_one_adol` is negative (-1.503), suggesting that when Walimah socializes with at least one adolescent social partner, her SDB rate is lower than when she does not socialize with any adolescent social partners. This aligns with the previously

observed buffer effect. 2. The coefficient for `at_least_one_adult` is close to zero (-0.050) and has a relatively small MAD_SD (0.305), indicating that the presence of adult social partners does not have a strong impact on Walimah's SDB rate. This is different from the general pattern observed in adolescent female orangutans, where socializing with adult social partners was associated with elevated SDB rates.

These results suggest that the buffer effect is also present in focal Walimah, as socializing with adolescent social partners is associated with a reduction in SDB rate. However, the impact of adult social partners on Walimah's SDB rate appears to be less pronounced compared to the general pattern observed among adolescent female orangutans.

Conclusion

In conclusion, we suggest utilizing GLMMs with a negative binomial distribution for modeling the impact of variables on the SDB rate instead of Poisson models or linear mixed models. Our recommendation is based on evidence of overdispersion, and the better observed fit for the negative binomial model using the rootogram method, which indicated fewer negative residuals. In addition, we found that there is no evidence for a buffer effect for the model which focuses only on Walimah. Results for the Walimah specific model are not representative of all orangutangs as there may be variability between orangutangs.

Appendix

Data Preprocessing

Both flanged and unflanged males are sexually mature adults, but they exhibit different physical and behavioral characteristics. It is important to note that unflanged males are not adolescents; they are fully mature adults. Besides, we convert the SDB rate to the count of the SDB and save it in a variable `SDB`.

```
# variable selection
sdb %>% dplyr::select(AveSDBEventID,
         `Age-Sex of Focal`,
         `Social Y or N`,
         `Focal Orangutan ID`,
         `Event ID`,
         `SDB Rate`,
         `#AdoFemPresent`,
         `#UnflangedPresent`,
         `#AdultFemPreent`,
         `#FlangedPresent`) -> mix_sdb

# convert social state, focal id and event id to factor
mix_sdb<- mix_sdb %>%
  mutate( `Age-Sex of Focal` = factor(mix_sdb$`Age-Sex of Focal`),
          `Social Y or N` = factor(mix_sdb$`Social Y or N`, levels = c("Y", "N")),
         `Focal Orangutan ID` = factor(mix_sdb$`Focal Orangutan ID`),
         `Event ID` = factor(mix_sdb$`Event ID`))


# add a new column that converts the sdb rate to the count of sdb.
mix_sdb %>%
  mutate(SDB = `SDB Rate` *10) %>% relocate(SDB) -> mix_sdb
```

Rootogram Code

```
# Create rootogram plots
poi.root <- ppc_rootogram(mix_sdb$SDB, posterior_predict(poi.fit), style = "hanging") +
  ggtitle("Poisson model") + scale_fill_manual(values = "red") +
  coord_cartesian(xlim = c(0, 150))

nb.root <- ppc_rootogram(mix_sdb$SDB, posterior_predict(nb.fit), style = "hanging") +
  ggtitle("negative binomial model") + scale_fill_manual(values = "red") +
  coord_cartesian(xlim = c(0, 150))

# Put the plots side by side
grid.arrange(poi.root, nb.root, nrow = 2)
```

Outliers from Negative Binomial Model

```
table1<- outliers_data %>% count(FocalID) %>% arrange(desc(n)) %>%
  kable(caption = "Potential outliers ranked by count") %>%
```

Table 1: Potential outliers ranked by count

FocalID	n
Walimah	7
Betsy	5
Codet	4
Malik	4
Dagul	3
Prabu	3
Dinda	2
Mister Kecil	2
Bibi	1
Delly	1
Jumi	1
Wadi	1
Xibu	1
XXAdoFem	1

```
  kable_styling(latex_options = "basic")

cat(table1)
```