Ant anatomy and behavior research

Qian Xu, Zichun Liu, Zijian Han; TF mentor: Jun Li; Professor Mentor: Masanao Yajima

11/21/2016

1. Background and Data Format:

1.1 Overview

In this project, our client Darcy Gordon from Department of Biology is interested in the relationship between ant brain anatomy and motor information. Our team mainly focused on helping our client analyze the result of her experiment. The experiment focused on 3 types of ants(3 morphological groups) and tested their ability to detect chemical trails with different concentrations.

In the experiment, there are followign factors

- three colonies(PR6,PR7,PR8) of ants and each contains
- three morphological groups:
  - minor N= 17,
  - major N= 18, and
  - supersoldier N= 17 (total across all colonies).
- three chemical trails with different concentrations: 0.0003, 0.001, 0.003

Each individual ant was tested at each of the 3 chemical trails. The whole data set has 3*(17+18+17) = 156 observations.

We have two kinds of outcomes, the first one is binary(detection or not,1 or 0), the second one is the proportion of detected trail crosses (on a scale of 0-1). The proportions are calculated by dividing the total number of turns an ants takes along an artificial trail by the total number of times it encountered the trail when running a behavioral assay. This corresponds to the trail following data sheet our client sent as "proportion detected" which is turns/ (turns + crosses). It gives us information about how well they followed a trail not just if they are able to detect it.

However, the turns and crosses have more information than the proportions. The detection data could be treated that it is from the proportions (0 proportions indicate 0 detection while not 0 proportions indicate 1 detection). After the discussion with our client, we agreed to work on the following two models:

Model 1: Binomial Model with turns and crosses information as responses

$$\begin{aligned} y_i &\sim Bin(n_i, p_i) \\ p_i &= logit^{-1}(\beta_0 + \beta_s x_{is} + \beta_c x_{ic} + \beta_{sc} x_{is} : x_{ic} + \alpha_{j[i]}) \\ \alpha_j &\sim N(0, \sigma_a^2), j = 1, \ldots, J \end{aligned}$$

Where i indexes each trial, and j indexes each ant. In this model we assume turns and crosses are successes and failures of sequence of independent trials, which is not necessarily true. However by accounting for $n_i = trun_i + cross_i$ we are in a way taking into account the indivisual activity level. We assume the turns and crosses are independent.

Model 2: Logistic Regression with binary outcomes as responses

$$
\begin{aligned}
y_i &\sim Bernoulli(p_i) \\
p_i &= logit^{-1}(\beta_0 + \beta_s x_{is} + \beta_c x_{ic} + \beta_{sc} x_{is} : x_{ic} + \alpha_{j[i]}) \\
\alpha_j &\sim N(0, \sigma_a^2), j = 1, \ldots, J
\end{aligned}
$$

Where i indexes each trial, and j indexes each ant. In this model we collapsed the crosses and turns information into binary outcomes, that is, if the number of turns is not zero, then there is a success (outcome 1) for this single trial which includes multiple turns and crosses. We assume these macro trials are independent.

2. Research Questions

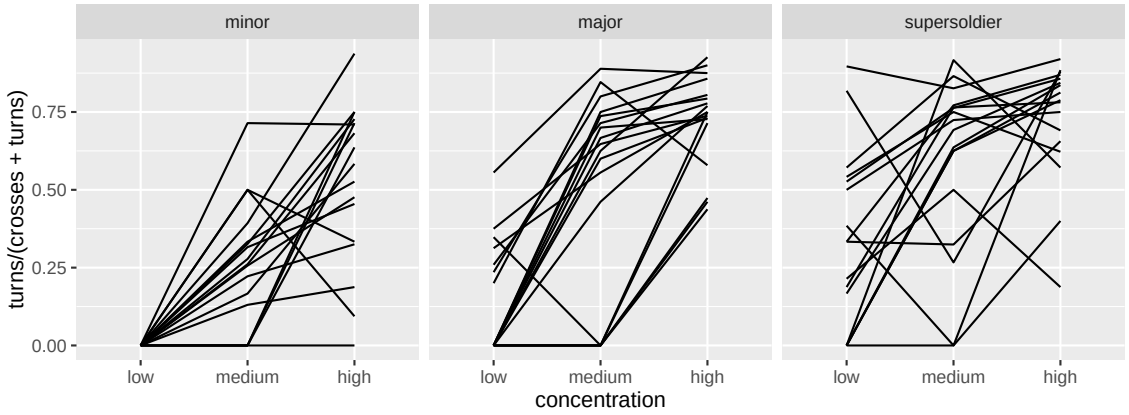
Based on our conversations, client is interested in

1. Effect of the subcaste and the trail concentration on the success rate of detection for each ants. This can be broken down into
    - Does concentration increase the detection rate.
    - Is there a difference in the detection rate between the subcaste.
    - Is the effect of concentration differeren for different subcaste.
2. Prediction for different subcaste and a trail concentration with the prediction interval.

3. Exploratory Data Analysis

3.1 Model 1: turns and crosses information as responses

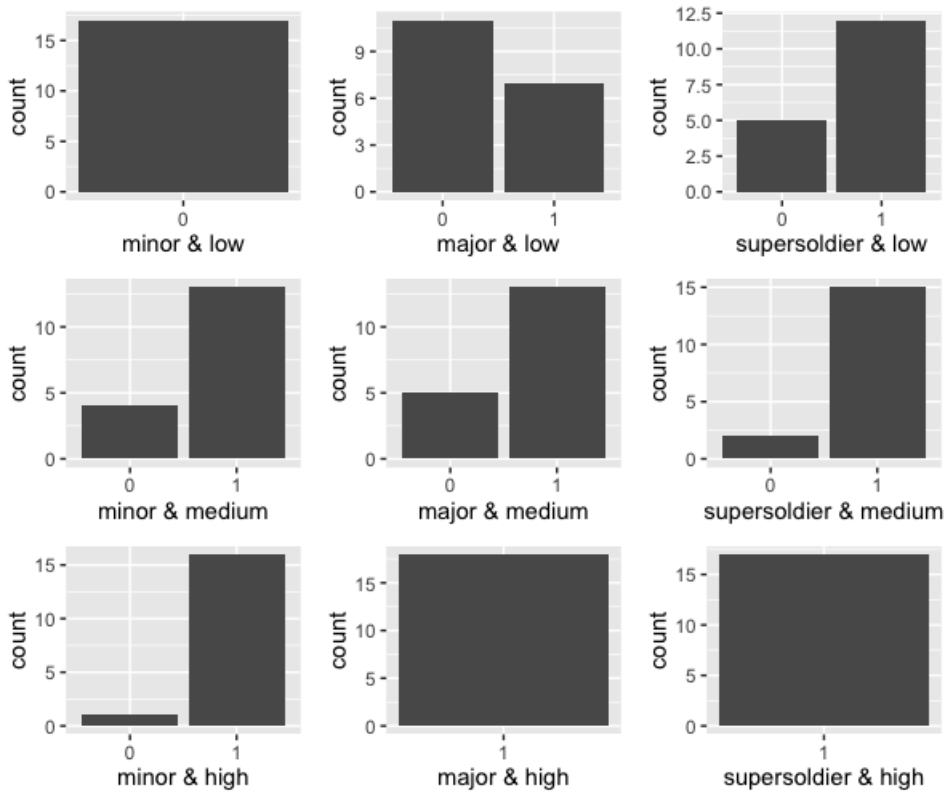
The rate of detection per ant is plotted for the subcaste and the concentration.



The trends are not parallel with each other from low to high concentration when the subcaste is minor, major and supersoldier indicating there is difference in how ants reacts to concentration for different subcaste.

3.2 Model 2: binary outcomes as responses

Count the 0's and 1's for each subcaste level and each concentration level.



Although we lost some information when converting the turns and crosses into the outcomes, we can use the histograms above to get a sense whether there exist interaction terms. In general, the proportion of 1's increases from minor to supersoldier in subcaste and from low to high in concentration. There is no strong evidence showing there exist interaction terms.

However, being more precise, let's focus on the 4 upper-left subplots. There are all 0's in the minor & low class and there are some 1's in the major & low class, but there are more 1's in minor & medium class than that in major & medium class. This phenomenon shows the trail detecting ability is growing faster in minor than that in major subcaste, at least when we increases the concentration from low to medium, but we don't know how strong this effect is. Thus we still want to enroll interaction term when fitting Model 2 and see if the interaction term is significant or not.

4. Model 1 Fitting: turns and crosses information as responses

4.1 linear separation problem

Let's enroll the interaction term in the logistic mixed effect model. We will find something strange happens: the standard error of `subcasteminor:concentrationlow` is too large. It is called the linear separation problem. This phenomenon happens due to the sudden change in response when we change the value of discrete predictors. Use our case as an example, we could get the reference level is subcaste major and concentration high. In this situation, there should be a lot of data points where turns are not equal to 0. We also know that when in the subcaste minor and low concentration situation, no ant has any turns, so the standard error of `subcasteminor:concentrationlow` explodes. We can also interpret this phenomenon in this way: some linear combination of the predictors could nearly separate the response classes. It is where the term "linear separation" comes from.

```
fit1 <- glmer(cbind(turns,crosses) ~ subcaste * concentration + (1|ID),
            data= trail.following, family = "binomial")
round(summary(fit1, correlation=FALSE)$coefficient,2)
```

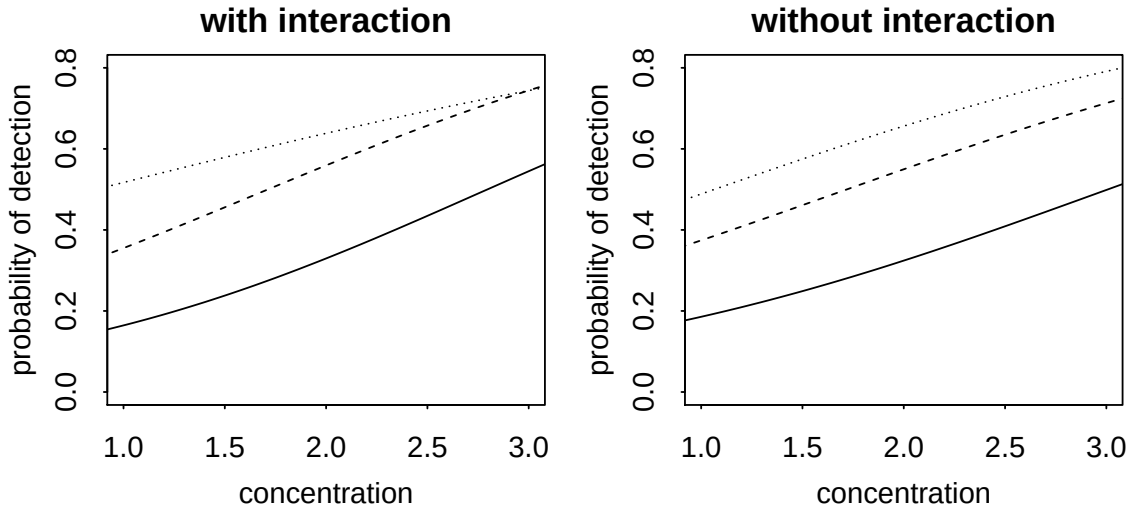
```
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                 1.06       0.22    4.73     0.00
## subcasteminor                              -1.01       0.32   -3.16     0.00
## subcastesupersoldier                        0.05       0.31    0.17     0.87
## concentrationlow                           -3.16       0.22  -14.48     0.00
## concentrationmedium                        -1.01       0.17   -5.77     0.00
## subcasteminor:concentrationlow            -16.83    1167.59   -0.01     0.99
## subcastesupersoldier:concentrationlow       1.30       0.27    4.77     0.00
## subcasteminor:concentrationmedium          -0.24       0.24   -0.97     0.33
## subcastesupersoldier:concentrationmedium    0.35       0.23    1.54     0.12
```

4.2 Remedy: convert the format of concentration to numeric

One necessary condition of linear separation is that all the predictors should be discrete, so one simple remedy method is to use continuous concentrations: 0.0003, 0.001, 0.003. Because we are considering interaction term, the scales of predictors should be almost the same. I multiplied each concentration by 1000.

```
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                               -2.77       0.26  -10.63     0.00
## subcastemajor                              1.16       0.35    3.31     0.00
## subcastesupersoldier                       2.26       0.34    6.58     0.00
## concentration_num                          0.98       0.08   13.00     0.00
## subcastemajor:concentration_num           -0.05       0.10   -0.45     0.65
## subcastesupersoldier:concentration_num    -0.41       0.09   -4.45     0.00
```

We will find that now there is no linear separation. The following plot shows us that there is the supersoldier ant has a less sleep increase of detection probability when concentration goes up, which indicates we should have the interaction term. The negative value of `subcastesupersoldier:concentration_n` coefficient agrees this.



4.3 Prediction

If we don't specify the ant ID in prediction, we need to add variance from random effect and get larger prediction intervals. The following prediction intervals are computed from simulation, providing empirical 2.5%, 50%(median), 97.5% quantiles as well as the empirical means of the success probabilities, which are listed as the last number in each category.

```
## INDICES: minor:low
##  2.5%   50% 97.5%
## 0.006 0.079 0.481 0.077
## ------------------------------------------------------
## INDICES: minor:medium
##  2.5%   50% 97.5%
## 0.013 0.149 0.657 0.145
## ------------------------------------------------------
## INDICES: minor:high
##  2.5%   50% 97.5%
## 0.076 0.550 0.933 0.542
## ------------------------------------------------------
## INDICES: major:low
##  2.5%   50% 97.5%
## 0.019 0.213 0.763 0.209
## ------------------------------------------------------
## INDICES: major:medium
##  2.5%   50% 97.5%
## 0.037 0.344 0.863 0.338
## ------------------------------------------------------
## INDICES: major:high
##  2.5%   50% 97.5%
## 0.196 0.775 0.976 0.768
## ------------------------------------------------------
## INDICES: supersoldier:low
##  2.5%   50% 97.5%
## 0.045 0.425 0.895 0.416
## ------------------------------------------------------
## INDICES: supersoldier:medium
##  2.5%   50% 97.5%
## 0.070 0.525 0.927 0.514
## ------------------------------------------------------
## INDICES: supersoldier:high
##  2.5%   50% 97.5%
## 0.179 0.771 0.975 0.765
```

If we use ant ID in prediction, we will get small prediction intervals. The results are in the "trail_following_pred_intervals.csv", headed with "crossturn".

5. Model 2 Fitting: binary outcomes as responses

5.1 Linear separation problem and remedy

Using predictor "concentration" with 3 levels low, medium, high, we came across the linear separation again:

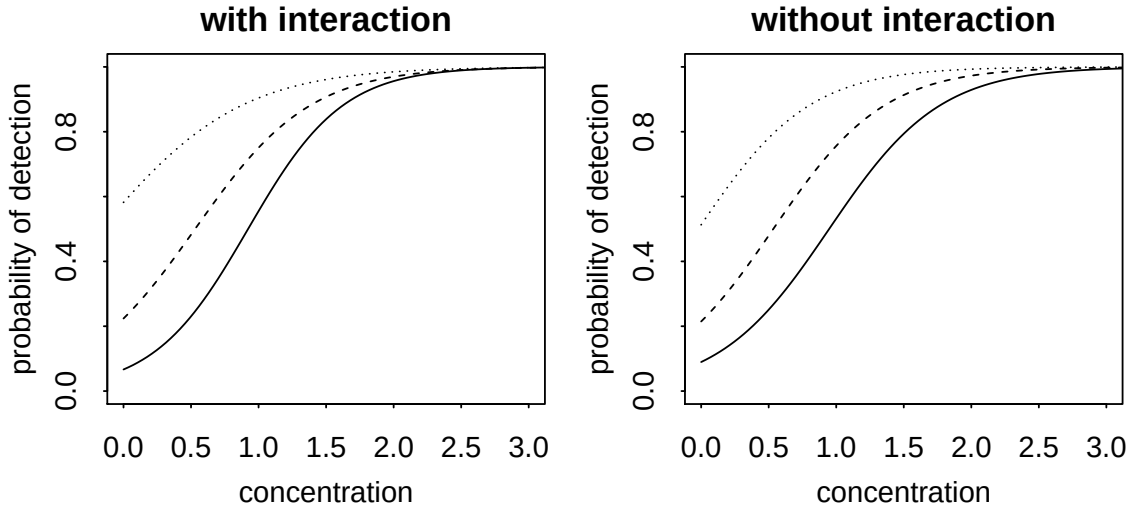
```
fit4 <- glmer(outcome ~ subcaste * concentration + (1|ID),
              data= trail.following, family = "binomial")
round(summary(fit4, correlation=FALSE)$coefficient,2)
```

```
##                                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                   17.96    1512.13    0.01     0.99
## subcasteminor                                -14.23    1512.13   -0.01     0.99
## subcastesupersoldier                           5.79   26557.33    0.00     1.00
## concentrationlow                             -18.64    1512.13   -0.01     0.99
## concentrationmedium                          -16.57    1512.13   -0.01     0.99
## subcasteminor:concentrationlow               -19.12 5699102.97    0.00     1.00
## subcastesupersoldier:concentrationlow         -3.81   26557.33    0.00     1.00
## subcasteminor:concentrationmedium             14.59    1512.13    0.01     0.99
## subcastesupersoldier:concentrationmedium      -4.41   26557.33    0.00     1.00
```

As what we have done in the last section, use numeric concentration and again scale them (times 1000) and refit the model:

```
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                -2.66       0.97   -2.75     0.01
## subcastemajor                               1.40       1.16    1.21     0.23
## subcastesupersoldier                        3.01       1.29    2.33     0.02
## concentration_num                           2.88       1.10    2.62     0.01
## subcastemajor:concentration_num            -0.50       1.38   -0.36     0.72
## subcastesupersoldier:concentration_num     -0.95       1.54   -0.62     0.54
```

The interaction terms are not significant, but the signs before the coefficients of the interactions are the same as those in Section 4.2. Keeping the interaction term will not hurt the predictions. The plots below also showed that there is some difference between the left and right plot, although not significant.



5.2 Prediction

We can get similar prediction interval for each combination of subcaste and concentration level without ant ID, via the same simulation method. The prediction intervals are also in the "trail_following_pred_intervals.csv", headed with "binary".

```
## INDICES: minor:low
##  2.5%   50% 97.5%
## 0.012 0.145 0.710 0.145
## ------------------------------------------------------
## INDICES: minor:medium
##  2.5%   50% 97.5%
## 0.096 0.557 0.939 0.560
## ------------------------------------------------------
## INDICES: minor:high
##  2.5%   50% 97.5%
## 0.702 0.997 1.000 0.997
## ------------------------------------------------------
## INDICES: major:low
##  2.5%   50% 97.5%
## 0.041 0.363 0.887 0.363
## ------------------------------------------------------
## INDICES: major:medium
##  2.5%   50% 97.5%
## 0.191 0.754 0.975 0.754
## ------------------------------------------------------
## INDICES: major:high
##  2.5%   50% 97.5%
## 0.714 0.997 1.000 0.997
## ------------------------------------------------------
## INDICES: supersoldier:low
##  2.5%   50% 97.5%
## 0.160 0.711 0.970 0.712
## ------------------------------------------------------
## INDICES: supersoldier:medium
##  2.5%   50% 97.5%
## 0.378 0.903 0.993 0.903
## ------------------------------------------------------
## INDICES: supersoldier:high
##  2.5%   50% 97.5%
## 0.503 0.998 1.000 0.998
```

Note that the prediction and confidence interval is different from that in Section 4.3 because the definition of "success" is different: in Section 4, turns means successes while in Section 5, when outcome is 1, there is a success.

6. Remarks

6.1 Limitations of our analysis

There are some limitations of our analysis, because of these two assumptions we made: 1. We are assuming the turns are independent from one another, which we have known they are not true. 2. We are assuming homogeneity across the colonies, because we don't include colony effect in our analysis. See Section 6.2.

6.2 More about random effect: the colony effect

We tried putting the colony effect in the model but we encountered numerical issues again and had some problems in convergence. Thus we were unable to include the colony effect by now. If you want to add colony effect urgently, we are happy to work on applying the regularization methods mentioned below to address the numerical issues and include colony effect into analysis. In other case, we'd like to include ant ID as random effect only as what we did in Section 1-5.

6.3 Future analysis: fully addressing the numerical issues

To fully address the numerical issues including linear separation, we would probably need to use some regularization methods in Bayesian framework. The Bayesian regularization methods won't provide p-values which is required in many fields. We will be happy to make suggestions if your field don't have a strict rule about providing p-values.