Portfolio of MSSP Projects

Zichun Liu

March 28, 2017

1 Overview

As an important graduation requirement of MSSP, the portfolio is made up of project analysis reports that have been done within the one and a half year of the master program. In this portfolio, I archived eight main projects:

- Ant anatomy and behavior research
- Catalog study on women emancipation
- Salmon habitation preference study
- Measuring the impact of geriatric care among Medicare beneficiaries
- R package: Single cell toolkit
- Random Contamination and Select Response Styles affecting measures of fit and reliability in factor analysis
- R package: celda - Cellular Latent Dirichlet Allocation(ongoing)
- Recommender system on Consumer Unit dataset(ongoing)

2 Ant anatomy and behavior research

2.1 Introduction & Background

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.

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.

2.2 Data Format

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).

2.3 Modeling & Analysis

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

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

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

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

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.

2.3.1 Exploratory Data Analysis

Model 1: turns and crosses information as responses

The rate of detection per ant is plotted for the subcaste and the concentration. .
.
.
.
.
Pic 1.
.
.
.
.
.

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.

Model 2: binary outcomes as responses

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

.
.
.
.
Pic 2.
.
.
.
.
.

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 1s 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, lets focus on the 4 upper-left subplots. There are all 0s in the minor & low class and there are some 1s in the major & low class, but there are more 1s 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 dont 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.

2.4 Model 1 Fitting

2.4.1 linear separation problem

Lets 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. .
.
.
.
.
R output.
.
.
.
.
.

2.4.2 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. .
.
.
R output.
.
.
.

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. .
.
.
R plot.
.
.
.

2.4.3 Prediction

If we dont 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.
.
.
.
R output.
.
.
.

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.

2.5 Model 2 Fitting

2.5.1 Linear separation problem and remedy

Using predictor concentration with 3 levels low, medium, high, we came across the linear separation again. And as what we have done in the last section, use numeric concentration and again scale them (times 1000) and refit the model:
.
.
.
R output.
.
.
.

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.

2.5.2 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. .
.
.
R output.
.
.
.

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.

2.6 Discussion

2.6.1 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 dont include colony effect in our analysis. See Section 2.6.2.

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, wed like to include ant ID as random effect only as what we did in Section 2.1-2.5.

2.6.3 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 wont provide p-values which is required in many fields. We will be happy to make suggestions if our client's field dont have a strict rule about providing p-values.

3 Measuring the impact of geriatric care among Medicare beneficiaries

3.1 Introduction & Background

This is a collaboration project with the Trinity Partners.

Rising medical cost due to aging population is one of the top concerns in the US. There are currently almost 50 million people over 65 and 70 percent of them have two or more chronic conditions, which contribute to non-stopping, continuing costs of medicine and other medical services. This population cost Medicare nearly 650 billion dollars just looking at year 2015.

As the result, doctors who focus on elderly patients, geriatricians are gaining interest, both for the system and general public, to see whether geriatric care can provide better outcomes, either clinical or economical, for treating the elderly population.

Thus, our research focus on three main questions:

- Would geriatric care increase the longevity of the elderly population?
- Would geriatric care reduce the costs for the Medicare system?
- Would geriatric care reduce the medical resources utilization?

3.2 Data Format

The data we have was originated from Medicare claim data from Jan 1st 2010 to Dec 31st 2014, in which each observation is a claim report in the system.

For the purpose of studying the impact of geriatric care on patients, we converted the claim-based data to patient-based data. The number of claims in 2010 and CCI are calculated from the information between 2010-01-01 and 2010-12-31. Other variables are calculated or extracted from information between 2011-01-01 and 2014-12-31.

2 pair of sample datasets have been used in the analysis, being one for survival analysis and the other for cost and resource analysis.

3.3 Methods

3.3.1 Charlson Comorbidity Index

The Charlson Comorbidity Index (CCI) is the predicted one-year mortality for a patient with one or more chronic conditions (Charlson, M,1987). Given the fact that there are nearly 70% of Medicare beneficiaries in the U.S. have multiple chronic diseases, the CCI is applied and calculated for each patient in order to control for patients health condition among case and control groups.

The CCI takes into account 17 chronic conditions and assigns each of them with a weight of 1, 2, 3, or 6, based on the associated risk of dying. Then the weights are summed for each patient in order to attain a total score. The range of a patients CCI is 0-30.

3.3.2 Propensity Score Matching

Propensity score is the probability of being assigned treatment conditional on observed covariates, which we denote as $P(T = 1|X)$. Here X indicates a matrix of all pre-treatment covariates, T indicates binary treatment corresponding to geriatric care. T is 1 when there is geriatric care and T is 0 when there is not.

Propensity score matching entails forming matched sets of subjects in case and control groups, who share a similar value of the propensity score. Subjects in case and control groups that are paired are close to each other, which is done by using nearest neighbor matching method.

3.3.3 Causal Inference

Causal inference for individual i is the comparison of the outcome $Y_i(1)$ if he/she receive the treatment and the outcome $Y_i(0)$ if he/she does not receive the treatment. In general, there are two measurement of causal inference, Average Treatment Effect(ATE) and Average Treatment Effect only focus on Treated group(ATT).

ATE is measured by $E(Y_i(1) - Y_i(0)|x, i \in All.groups)$, where x is control variables. ATT is measured by $E(Y_i(1) - Y_i(0)|x, i \in Treated.group)$

Propensity score matching is a popular method to addressing this problem. After using propensity score matching, we can find individuals in control group who have similar situation with individuals in case group. Therefore, we have the data of $Y_i(0)(i \in Treated.group)$. Thus, propensity score can be very useful for us to measure average treatment effect by only focusing on treated group(ATT).

3.3.4 Exploratory Data Analysis

3.4 Findings

3.5 Discussion

3.6 Appendix

4 Ant anatomy and behavior research