Portfolio of MSSP Projects

Zichun Liu

March 28, 2017

Contents

1 Overview **2**

2 Measuring the impact of geriatric care among Medicare beneficiaries **3**
2.1 Introduction & Background . . . 3
2.2 Data Format . . . 3
2.3 Modeling & Analysis . . . 3
2.3.1 Methods . . . 3
2.3.2 Results . . . 4
2.4 Conclusions . . . 4
2.5 Discussion . . . 4
2.5.1 Quality of Life . . . 4
2.5.2 Side effect of overuse of geriatric care on cost . . . 4
2.6 Appendix . . . 5

3 Ant anatomy and behavior research **5**
3.1 Introduction & Background . . . 5
3.2 Data Format . . . 5
3.3 Modeling & Analysis . . . 5
3.3.1 Methods . . . 5
3.3.2 Model 1 Results . . . 6
3.3.3 Model 2 Results . . . 6
3.4 Conclusion . . . 6
3.5 Discussion . . . 6
3.5.1 Limitations of our analysis . . . 6
3.5.2 More about random effect: the colony effect . . . 7
3.5.3 Future analysis: fully addressing the numerical issues . . . 7

4 Catalog study on women emancipation **7**
4.1 Introduction & Background . . . 7
4.2 Data Format . . . 7
4.3 Modeling & Analysis . . . 7
4.4 Conclusion . . . 7
4.5 Discussion . . . 7

5 Salmon habitation preference study **7**
5.1 Introduction & Background . . . 7
5.2 Data Format . . . 7
5.3 Modeling & Analysis . . . 7
5.4 Conclusion . . . 7
5.5 Discussion . . . 8

6 Visualization for Boston University Women Soccer Team **8**
6.1 Introduction & Background . . . 8
6.2 Data Format . . . 8
6.3 Visualization . . . 8
6.4 Discussion . . . 8

7 R package: SingleCellTK(ongoing) **8**
7.1 Introduction & Background . . . 8
7.2 Data Format . . . 8
7.3 Modeling & Analysis . . . 8
7.4 Conclusion . . . 8
7.5 Discussion . . . 8

8 R package: Celda(ongoing) **8**
8.1 Introduction & Background . . . 8
8.2 Data Format . . . 9
8.3 Modeling & Analysis . . . 9
8.4 Conclusion . . . 9
8.5 Discussion . . . 9

9 Random Contamination and Select Reponse Styles affecting measures of fit and reliability in factor analysis(ongoing) **9**
9.1 Introduction & Background . . . 9
9.2 Data Format . . . 9
9.3 Modeling & Analysis . . . 9
9.4 Conclusion . . . 9
9.5 Discussion . . . 9

10 Recommender system on Consumer Unit dataset(ongoing) **9**
10.1 Introduction & Background . . . 9
10.2 Data Format . . . 9
10.3 Modeling & Analysis . . . 9
10.4 Conclusion . . . 10
10.5 Discussion . . . 10

11 Acknowledgements **10**

12 Reference **10**

1 Overview

The purpose of this portfolio is to show the work of MSSP students on collaborative consulting projects. Throughout two semesters, students in MSSP formed the consulting groups and solved problems from the clients who were either within BU at different departments or outside of BU. We helped them with their research projects and provided statistical solutions by all kinds of ways. This portfolio consists of separate sections which summarize the background, data, modeling and findings of every project. The links to the codes on GitHub for most of the projects are in the reference section at the end of the portfolio. Due to some restrictions, some data and codes of the projects remain confidential. In this portfolio, I archived eight main projects:

- Ant anatomy and behavior research
- Catalog study on women emancipation
- Salmon habitation preference study
- Visualization for Boston University Women Soccer Team
- 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 Measuring the impact of geriatric care among Medicare beneficiaries

This is a collaborative project conducted by the MSSP consulting team and Trinity Partners, LLC. The outcome was accepted by 22nd ISPOR annual meeting as a poster presentation.

2.1 Introduction & Background

In 2015, there are almost 50 million people over 65 and 70 percent of them have two or more chronic conditions, which contributes to non-stopping, continuing costs of medicine and other medical services. In the previous studies geriatricians, doctors who meet the unique healthcare need of older adults, have been shown to have an influence on elderly patients, especially those with severe chronic diseases.

To form a better view of the impact of geriatrician on the aging population, we modeled the Medicare claim data between January 2010 and January 2015 to infer the effect of geriatric care on **longevity, cost, and the use of medical resources**. This paper measures the impact of geriatric care on patient outcomes

2.2 Data Format

The study cohorts were created from claims spanning 2010-2014, representing a random 5% sample of Medicare FFS beneficiaries that were aged 65+ in 2010. Patients with at least one geriatric care claim between (but not before) 2011-2014 were propensity score matched to those without any geriatric care claims from 2010-2014.

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.

2.3 Modeling & Analysis

2.3.1 Methods

To draw casual relationship from the Medicare data, a matched cohort analysis, **propensity score matching**, was performed, we used comparing patient outcomes among those with and without geriatrician care. **Survival rates, healthcare resource utilization (HCRU) and costs** were compared between matched cohorts.

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.

Survival Analysis were performed to investigate the impact of geriatric care on longevity. The Kaplan-Meier estimate is a famous nonparametric maximum likelihood estimate of the survival function. The estimate is a step-wise empirical function which can be written as:

$$\hat{s}(t) = 1, if \quad t < t_i$$
$$\hat{s}(t) = \prod_{t_i \leq t} [1 - \frac{d_i}{Y_i}], if \quad t \geq t_i$$

Where we have order the event times as a sequence of ascending, e.g. $0 < t_1 < t_2 < ... < t_D$. di is how many of individuals with an observed event at time t_i and Y_i is the number of individuals who are at risk at time t_i.

And to make an accurate and solid comparison between the case and control groups in terms of the survival analysis, we constructed a Weibull regression model. The survival function for the Weibull distribution is:

$$S_T(t) = e^{-\lambda t^P}$$

Where λ is reparameterized in terms of predictor variables and regression parameters, and p is the shape parameter which is held fixed for a parametric model.

Cost analysis used T-test after propensity score matching to compare the cost between case and control groups.

Resource Utilization analysis used Linear Regression Models after propensity score matching to compare the medical resources usage. The response variables are: Log-transformed claim numbers and log-transformed length of stay in hospital; The independent variable is whether there is geriatric treatment.

2.3.2 Results

118,382 patients receiving geriatric care were identified, along with a random sample of 535,526 patients not receiving geriatric care. Samples of the cohorts were taken and propensity score matching was performed to create 10,000 matched pairs. Kaplan-Meier plots comparing the two matched cohorts indicated that, overall, the geriatric care group had shorter survival times than their counterparts. Among sicker patients, however, the analysis suggested that survival times were relatively longer in the geriatric care group. In terms of HCRU, patients receiving geriatric care were, on average, likely to have more inpatient, outpatient and physician office claims (62%, 18% and 49% more, respectively). Mean monthly costs were comparable in the two matched cohorts (p=0.53).

2.4 Conclusions

Geriatric care has the potential to improve the longevity of those afflicted by multiple chronic conditions. Geriatricians focus on coordination of care increases the amount of healthcare resource utilization but costs are on par with patients not receiving geriatric care, possibly the result of support that is more efficient and preventative in nature. While only a small proportion of elderly patients receive geriatric care, this study suggests that more patients and the healthcare system at large could benefit from promoting the involvement of and increasing access to geriatricians.

2.5 Discussion

2.5.1 Quality of Life

Theres another aspect of life thats not covered in our data quality. How much pain or distressing physical symptoms do they experience on a daily basis? Are they depressed, anxious or do they sleep well at night? Most importantly, their functional status, are they highly dependent or able to manage ordinary, mundane daily activities? Its also of great interest and importance to investigate whether geriatricians can make a difference for elderly in this aspect, for life has much more dimensions than just time and money. Therefore, for future analysis, another angle is to assess the quality of life of those elderly patients treated by geriatricians.

2.5.2 Side effect of overuse of geriatric care on cost

For all case sample we have had resampled from the raw data, cost first decrease to the bottom until the number of months having geriatric care is around 25, and then cost increase after that. This also indicates that geriatric care wont increase cost for patients or even decrease cost at first, but overuse of geriatric care can have side effects on cost.

2.6 Appendix

3 Ant anatomy and behavior research

3.1 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) from 3 colonies and tested their ability to detect chemical trails with different concentrations.

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.

3.2 Data Format

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

There are 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). 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).

3.3 Modeling & Analysis

3.3.1 Methods

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

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

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.

3.3.2 Model 1 Results

linear separation problem At first we enroll the interaction term in the logistic mixed effect model. We found something strange happens: the standard error of $subcasteminor : concentrationlow$ is extremely 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.

Remedy 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 plot.
.
.
.

Prediction If we dont specify the ant ID in prediction, we need to add variance from random effect and get larger prediction intervals. If we use ant ID in prediction, we will get small prediction intervals. The corresponding results are generated through 'predictInterval' function in R.

3.3.3 Model 2 Results

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.

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. .

.
.
R plot.
.
.
.

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 generated through 'predictInterval' function in R.

3.4 Conclusion

From both models, we can see obviously that the increase in concentration will increase the detection rate. And different ants have a distinction in detecting chemical trails: Supersoldiers have the highest detection rate and Minors have the lowest detection rate. What's more, we may tell that the effect of concentration is different for different subcaste from the interaction term in our model.

3.5 Discussion

3.5.1 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 2.5.2.

3.5.2 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 2.1-2.4.

3.5.3 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.

4 Catalog study on women emancipation

asdf

4.1 Introduction & Background

asdf

4.2 Data Format

adf

4.3 Modeling & Analysis

asdf

4.4 Conclusion

adf

4.5 Discussion

asdf

5 Salmon habitation preference study

asdf

5.1 Introduction & Background

asdf

5.2 Data Format

asdf

5.3 Modeling & Analysis

asdf

5.4 Conclusion

asdf

5.5 Discussion

asdf

6 Visualization for Boston University Women Soccer Team

asdf

6.1 Introduction & Background

fa
asdf

6.2 Data Format

asdf

6.3 Visualization

asdf

6.4 Discussion

asdf

7 R package: SingleCellTK(ongoing)

asdf

7.1 Introduction & Background

sadf

7.2 Data Format

asdf

7.3 Modeling & Analysis

adsf

7.4 Conclusion

asdf

7.5 Discussion

asdf

8 R package: Celda(ongoing)

asdf

8.1 Introduction & Background

asdf

8.2 Data Format

asdf

8.3 Modeling & Analysis

asdf

8.4 Conclusion

asdf

8.5 Discussion

adsf

9 Random Contamination and Select Reponse Styles affecting measures of fit and reliability in factor analysis(ongoing)

asdf

9.1 Introduction & Background

asdf

9.2 Data Format

asdf

9.3 Modeling & Analysis

asdf

9.4 Conclusion

adsf

9.5 Discussion

adsf

10 Recommender system on Consumer Unit dataset(ongoing)

adsf

10.1 Introduction & Background

asdf

10.2 Data Format

asdf

10.3 Modeling & Analysis

asdf

10.4 Conclusion

asdf

10.5 Discussion

asdf

11 Acknowledgements

Projects included in the portfolio received support and guidance from:

- Professor Eric D. Kolaczyk in the Mathematics and Statistics at Boston University.
- Professor Masanao Yajima in the Mathematics and Statistics at Boston University.
- Professor Gregg A. Harbaugh in the School of Education at Boston University
- Teaching fellow Jun Li in the Mathematics and Statistics at Boston University.

12 Reference

Projects available at https://github.com/lloydliu717/Portfolio