CHL5207 Practicum Weekly Record

Jinyu Luo* 1004935457

2024-03-06

Meeting 1

Date: 2023.10.31

- Completed GitHub setup and successfully cloned the repository.
- Reviewed the project's history, focusing on achievements and progress made in the previous year.
- Defined and clarified the objectives for this year's study, aligning them with ongoing project goals.

Weekly Progress Summary

1. Literature Review
    a) Hui, S. K., Fan, C.-P. S., Christie, S., Feindel, C. M., David, T. E., & Ouzounian, M. (2018). The aortic root does not dilate over time after replacement of the aortic valve and ascending aorta in patients with bicuspid or tricuspid aortic valves. The Journal of Thoracic and Cardiovascular Surgery, 156(1).

    b) Li, P., Mitani, A., Fan, C.-P. S., & Saha, S. (2023). Modeling longitudinal outcomes in a small matched-pair sample motivated by cardiovascular data: A simulation study. University of Toronto Journal of Public Health, 4(1). https://doi.org/10.33137/utjph.v4i1.41675.

    c) Wan F. (2019). Matched or unmatched analyses with propensity-score-matched data?. Statistics in medicine, 38(2), 289–300. https://doi.org/10.1002/sim.7976.

    d) Papneja, K., Blatman, Z. M., Kawpeng, I. D., Wheatley, J., Oscé, H., Li, B., Lafreniere-Roula, M., Fan, C. P. S., Manlhiot, C., Benson, L. N., & Mertens, L. (2022). Trajectory of Left Ventricular Remodeling in Children With Valvar Aortic Stenosis Following Balloon Aortic Valvuloplasty. Circulation. Cardiovascular imaging, 15(1), e013200. https://doi.org/10.1161/CIRCIMAGING.121.013200.

    e) Wang, M. (2014). Generalized estimating equations in Longitudinal Data Analysis: A review and recent developments. Advances in Statistics, 2014, 1–11. https://doi.org/10.1155/2014/303728.

    f) Austin P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate behavioral research, 46(3), 399–424. https://doi.org/10.1080/00273171.2011.568786.

2. UHN On-boarding process

*University of Toronto, jinyu.luo@mail.utoronto.ca

- Completed the UHN Office of Research Trainee(ORT) e-registration form.

- Addressing the registration issue with Amy through Email.

3. Presentation Slide (1st Draft)
4. Reproduced the simulation process

Questions

1. According to Agresti's textbook in Chapter 12, Generalized Estimating Equations (GEE) are primarily used for categorical outcomes. However, last year's simulation focused on continuous outcomes. Could you explain the rationale behind this approach? Additionally, is it considered appropriate to use GEE for continuous outcomes in this context?
   - It is easier to start with continuous outcome. Although GEE is mainly used for categorical outcome, it is still appropriate to use with continuous outcome.

2. The statement "Quasi-Least Squares (QLS) falls under the framework of Generalized Estimating Equations (GEE)" suggests a close relationship between these two methodologies. Could you clarify what is meant by QLS being under the GEE framework? If QLS is indeed an advancement or a subsequent development of GEE, what is the purpose or benefit of conducting a comparative analysis between them?
   - QLS is an approach based on GEE that estimates the correlation parameters in two stages. It account for the matching correlation which is ignored by GEE. QLS can be applied when GEE estimate is infeasible or when GEE's assumptions do not hold. Therefore, we need to use it to evaluate GEE's performance.

3. Considering the findings from last year's work, are we planning to concentrate primarily on these three specific points for the current year's objectives?
   - The abstract page mentioned that the correlation structure should focus on exchangeable structure.
     - No, the data was generated using exchangeable structure, so it had the least deviation.

   - On page 13, it mentioned that there is need for caution and further investigation into the accuracy of the estimated standard errors to ensure appropriate inference.

   - On page 13, it mentioned that we need to explore the way to account for correlation in the generation process.

Meeting 2

Date: 2023.11.13

- Addressed questions I had last week.

- Clarified the study's objective for this year:
  - Investigate the performance of GEE when the **dropout rate is informative** on the outcome. In other words, the dropout rate depends on outcome and predictors.

- The outcome variable will be binary, a measure of heart function.

- Provided suggestions for presentation slides: there should be 2 slides for introducing the motivation, including treatment and cardiac surgery.

- Assigned new paper about the algorithm of binary outcome generation process: Jiang, W., Song, S., Hou, L., & Zhao, H. (2020). A set of efficient methods to generate high-dimensional binary data with specified correlation structures. The American Statistician, 75(3), 310–322. https://doi.org/10.1080/00031305.2020.1816213.

Weekly Progress Summary

1. Prepared presentation slides.
2. Followed the sample code to produce the binary outcome simulation.
3. On-boarding process:

- UHN network account access information(TID): t127930uhn
- Employee ID: 544196
- UHN email address: Jinyu.Luo@uhn.ca

Meeting 3

Date: 2023.11.22

- Completed UHN E-learning courses and orientation.
- Discussed about presentation contents and modified the slides.

Weekly Progress Summary

1. Picked up laptop on November 24th.
2. Tried to setup Ubuntu and Rdocker.
3. Scheduled a meeting with Sudip and Steve to fix mounting issue.

Next Step

1. Produce Table 1 by BAV and TAV groups.
2. Make Rscript of describing weekly jobs.
3. Read Sudip's paper.

Meeting 4

Date: 2023.12.15

- Fixed the issue of ubuntu connection and drive mounting.
- Created R script for recording weekly meeting.
- Read papers in the shared file.

Issues need to be addressed later

1. Identify confounding variables.
   2.Reduce the number of potential predictors because the target number is too low.
2. Understand which variable is changing over time.

What are the situations of applying correlation analysis?
- Changing in root over time (measured outcome multiple times).
- Repeated measure of outcome –> so they are correlated.

Distinguish between: 1. correlation between outcomes.
2. correlations between features.

Next Step

1. Produce Table 1
2. Perform EDA
3. Read papers in L drive folder
4. Summarize Matched or unmatched PPS analysis:

- what are the similarities and differences from our situation?

5. Use the data:

a) Match patients with baseline features using logistic regression.
   - ensuring equal number of patients in each group.
   - Each patient should have repeated measurements.

b) Dichotomize the outcome using cutoff point, then fit GEE.

Meeting 5

Date: 2024.01.10

- Completed EDA and Table 1.
- Finished Paper reading.
- Finished the comparison between paper DOI: 10.1002/sim.7976 and last year final paper.
- Tested the mean difference before and after matching.
- Reviewed the concept of statistical power, correlation between outcomes, and correlations between features

Questions to ask

- What should we do if the mean difference before matching is NOT statistically significant while the difference after matching is statistically significant after matching?
- How to deal with missing values?

Meeting Summary

- Use standardized difference to compare before and after matching.
- Plot the distribution of propensity score (density) after matching when we have the probability of being BAV vs. TAV. The overlapped region are matched pairs.

Next Step

1. Recreate TableOne for matched-pair.
2. Confirm with Sudip about how to dichotomize root size to binary and then Plot the proportion.
3. Descriptive analysis on longitudinal measurements (refer to correlated data).
4. Check whether missing in root, missing by records or missing by patient.
    - Check how many patient don't have baseline root and how many patient have at least one root missing.

Fit GEE with the binary data during the meeting.

Date: 2024.01.17

- Re-performed data cleaning mainly for filtering out records without root size information.
- Re-performed EDA using methods learned from lecture 1 of correlated data.
- Confirmed with Sudip about the cutoff point for dichotomizing root to binary.
  - The absolute value of the aortic root size $>$ 4.5cm or growth $>$ 5mm over time.
  - Sudip is on vacation, so I haven't ask about relevant literature for supporting this cutoff criteria.

TODO:
* Fix issues within the matching process.
* Plot the distribution of propensity score (density) after matching.
* Recreate TableOne for matched pairs.
* Confirm with Sudip whether or not setting the absolute value of growth $>$ 5mm to be 1 as well.
* Confirm with Sudip that whether it is correct to take the cutoff point using the root size measured at the last visit.

Questions to ask

1. When the number of visits are not consistent across group and individuals, how to construct covariance matrix?
2. How to calculate growth? Which measurement result should we use?

Meeting Summary

- Went through the entire data clening process, mainly in clarifying the process of removing records with NA at the column of root.
- Corrected the calculation for root size.
  - Each record from the same patient should have different growth and so different outcome variable.
  - The outcome variable should be corrected accordingly.

Next Step

1. Recalculate the growth for each record usinng y_{ij} - y_{i j-1}.
2. Correct the outcome column using the updated growth.
3. Create a TableOne for age, sex, bsa by exposure groups before and after using info from the first visit.
4. Read paper in the shared folder.
5. Find all coefficients.

Date: 2024.01.24

- Recalculated growth for each record.
- Corrected the outcom column.
- Created two tableone for the initial visit and the last visit.
- Performed pair-matching using both matchIt and optmatch.
- Extracted coefficients from the logistic regression model.
- Separated Coding files for data cleaning, EDA, Propensity Score Matching

Meeting Summary

- Went through the data cleaning process
- Solved the issues in the matching process
- Pointed out the need to create pair ID for the data generation process

Next Step

- Extract matched pairs and create pair IDs for them.
- Create a long table for matched pairs.
- Create Table One for matched pairs.
- Generate binary outcome using visit, age, baseline BSA, exposure group and their interaction effect with visit.
- Fit the data with GEE and QLS.
- Read GEE Wang 2014.

Date: 2024.02.07

- Extracted matched pairs to form the long table.
- Created pair IDs and subject IDs.
- Created Table One for the long table of matched pairs and compared to the Table One of the original dataset.
- Fitted the matched pair data with GEE, but failed to fit with QLS.
- Re-read Peiyu's final paper and the code file for GEE-QLS analyses.
- Tutorials and Papers:
  - Xie, Jichun and Shults, Justine, ""Implementation of quasi-least squares With the R package qlspack"" (June 2009). UPenn Biostatistics Working Papers. Working Paper 32. https://biostats.bepress.com/upennbiostat/art32
  - Schwartz, S. (2022, December 1). Quantitative Methods in R - Vol.5 Multilevel. 16 GEE, Binary Outcome: Respiratory Illness. https://cehs-research.github.io/eBook_multilevel/gee-binary-outcome-respiratory-illness.html.
  - Balise, R. (2023, April 4). Intro to Categorical Data . 8 Models for Matched Pairs. https://raymondbalise.github.io/Agresti_IntroToCategorical/Matched.html#x8.1

Questions / Problems need to be solved

- How does the correlation estimation process changed from continuous outcome to binary outcome?
- Xie's paper argued that the estimates for independence structure are identical for QLS and GEE, do we need to still consider this as a comparison?
- AR1 structure is appropriate for studies in which the measurements are equally spaced in time, but measurements in our dataset are not equally spaced in time. Is it meaningful to implement QLS estimation with AR1 structure?

Meeting Summary

- Answered the above questions.
- Found paper and code for simulation.

Next Step

1. Generate all covariates.

- `male` $\sim$ Binomial (size = 1, prob = 0.65, n = 250) using #rbinom.
- `age` $\sim$ rnorm(n=250, mean = 60, sd = 10)
- `BSA` $\sim$ rnorm(n=250, mean = 2, sd =0.2)

$$\text{logit Pr}(BAV = 1) = -0.4 - 0.1 \times \text{Age} + 1.2 \times \text{male} + 3 \times \text{BSA} = m$$

where $\Pr(BAV = 1) = \frac{\exp(m)}{1+\exp(m)}$ represents the propensity score.

2. Generate BAV using rnorm(size = 1, prob = pscore, n = 250)
3. Set visit to be 5.
4. Generate the outcome variable:

$$\text{logit Pr}(Y = 1) = -1 - 0.05 \times \text{visit} - 2 \times \text{BAV} - 0.05\text{Age} + 1.5 \times \text{Male} + 0.5 \times \text{BSA} + 0.5 \times \text{Visit} \times \text{BAV}$$

5. Fit GEE with exchangeable structure to the full data using n = 250
6. Use Propensity score matching to get matched data.
7. Fit Gee with exchangeable structure to the matched data.

Date: 2024.02.13

Completed Items

- Generated all covariates for 250 patients, with each has 10 observations.
- Generated the outcome variable for each visit per patient.
- Fitted GEE with exchangeable structure to the full data using n = 250.
- Performed propensity score matching.
- Fitted GEE with the matched data.

Questions

1. How to select value for the common correlation ρ?
2. Is it correct to assume age and BSA to be constant over time?
3. How to control the interval between each visit?

Meeting Summary

Two Challenges need to be solved:

1. GEE relies on large sample size. We assume the correct correlation structure for small sample data and aim to compare different correlation structures using GEE.
2. For each correlation structures, under GEE, fit two GEE models:
    - Independent
    - AR1 (The true structure)
    - Exchangeable

Need to fit two models: 1. Adjusted Model which is fitted with all covariates
+ age, sex, BSA are confounders which affect both outcome and exposure + The matching process removes the association between confounder and exposure

2. Unadjusted model which is fitted only with BAV and the interaction effect between BAV and visit
    - This is typically used in matched-pair study
    - We want to assess whether the matching process is necessary for this data.
    - We are mainly interested in the interaction effect.

Next Step

1. Write the simulation and model fitting process to R script.
2. Repeat the simulation 1000 time.
3. Save estimates to a matrix.
    - est <- gee$coef[2]
    - se <- gee
4. Fit with different correlation structures.

Reading Week: No meeting.

Date: 2024.02.28

- Created R script for simulation process
- Failed to simulate samples

Meeting Summary

Fixed problems occurred in the simulation process.

Next Step

1. Fit 6 models with different correlation structures
2. Extract coefficients

Date: 2024.03.06

- Fitted 6 models
- Added a GLMM at the end of simulation.
- Encountered the issue of rank deficient in the process of simulation for adjusted GEEs
- Tried to solve the problem:
  1. replaced `BAV:visit` with `BAV+visit+BAV*visit`. The problem retains.
  2. created a column of interaction using `interaction()` function and use this column as the main effect. The problem retains.
- Possible reasons and solutions for rank deficiency:
  - **Collinearity among predictors**: If two or more predictors in your model are highly correlated, it can lead to rank deficiency in the model matrix. So, is it appropriate to remove main effects of visit and BAV?
  - **Sparse data or categories with very few observations**: This might be a potential reason because our data is too small.

To discuss

- How to solve the issue of rank deficiency?
- Need to figure out the mathematical theory of the simulation process and the survival analysis.

Meeting Summary