CHL5207 Practicum Weekly Record

Jinyu Luo* 1004935457

2024-06-19

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.

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

2. UHN On-boarding process

   - 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

Next Step

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.
- The project time line before the final presentation.

Meeting Summary

Meeting Summary

- Clarified the reasons of fitting linear mixed model in last year's project.
- GLMM cannot produce the same coefficient as GEE.
- Checked the simulation results which producing reasonable proportion of positive outcomes.
- Proposed trying `trycatch` to allow the simulation run.
- The main topic of the second presentation can be **Endogeneity**.

Next Step

- Run the simulation with "trycatch"

- Add dropouts using `simsurv`
- Check the simulated data, including the mean of estimates, SE, SD, MSE, and Bias.
- Find the 95% Convergence probability.

Date: 2024.03.13

- Increased the total number of simulation to 1100.
- Conducted survival analysis on the original data to extract coefficients.
- Extracted estimations from simulations.

To discuss

In the coding for survival analysis,

1. Do we need to calculate the follow up time from the time that record of visit to the time that the outcome showed positive?
2. Do we need to include the remaining record after the first occurrence of the outcome? In survival data, if death occurs once, then it won't have follow up visit.

Meeting Summary

- Model the informative drop outs which dependent on the outcome and time.
- Calculate the coverage probabilities for each GEE
- The outcome should be an covariate in the survival model

Next Step

- Perform survival analysis and extract coefficients for simulating dropouts
- Find the coverage probabilities for each GEE

Date: 2024.03.20

- Cleaned the data based on the assumptions:
    1. monotone missing pattern - only include visits that had root size measurement and re-set visit numbers.
    2. patients underwent the operation - only include visits that occurred after the date of operation.
- Created an indicator column for drop out.
- Created the follow up time by subtracting the date of operation from the date of visit.
- Used the last visit for parametric survival modeling by assuming exponential time distribution.
- Re-performed the propensity score matching process and GEE with the latest data.
- Incorporate survival model in the simulation to simulate drop outs.

To discuss

During the simulation process for patients' visit, each visit will have a drop out indicator. Which one should we choose?

Next Step

- Make sure to simulate data from true parameter.
- Make sure the simulated parameters have biases close to 0
- mse close to 0.
- Coverage probability close to 95%.
- Mean of standard error estimates should be approx = standard deviation of parameter estimates. ==> empirical SE

Date: 2024.04.03

- Refit the survival model and extracted the coefficient to run the simulation. The time-to-event data is created by selecting the last observation from each participant and labeling them as 1 to represent event of not having 8 visits, 0 otherwise.

- The coverage probability for interaction effect achieved 100%, but the coverage probability of male, visit, and BAV are still 0.

- Large bias for male, BAS, BAV, and visit, but very small bias for the interaction effect.

- The estimation results are similar across different correlation structures.
- Adjusted models had slightly better estimates than Unadjusted models.

- Reviewed paper for discussing simulation and the document of using `simsurv`.

$$h_i(t) = \lambda\gamma t^{\gamma-1}\exp(X_i^T\beta)$$

where
* λ is the scale parameter.
* γ is the shape parameter.
* X_i is covariate which comes from Bernoulli distribution with mean of 0.5.

To discuss

How to construct the time-to-event data for survival analysis using the real data? How to define drop out?

Meeting Summary

- Starts with presentation slides and poster preparation.

Next Step

- Run the simsurv function with only baseline data, set the maxT = 2. Drop patients who die at time 1. People who survived at time 1, fit simsurv function using visit 2.
- Set the maxT = 2, fit the survived patients for the next round.

- Repeat the process
- Increase the number of people for dropping out at each time point.

- Check the proportion of outcomes before drop outs. The proportion of people drop outs who experienced the outcome should be higher than the proportion of people who did not experience the outcome.

- rows are visits, and columns are outcome proportion.

Exam Week

2024.04.17

Current progress

- Created a presentation outline.
- Tried to simulate death separate from the dropout process, but results are not plausible.

Meeting Summary

- Fixed errors in coding.
- Simulation:
  - At each visit, use the `simsurv` function to compute the probability of death.
  - Simulate dropouts using the full data.
  - Death should be considered as a part of dropouts, so there is no need to simulate death separately.
- Went over the presentation structure:
  - There should be a result section that presents the true parameter coverage probability, bias, and MSE for the main effects and interaction effect.
- Discussed potential questions that might be asked at the Q&A section.
  - Why you are comparing across different correlation structures?
    - GEE is famous for its robustness in modelling the correlation structure. Results are robust even if the correlation structure is mis-specified. However, this relies on large-sample theory, so we want to see how different correlation structure performs on small sample data.
  - Is there a reason to believe that different correlation structures are going to give different results in GEE?
    - When the sample size (the number of patients) is large, there should not be large variability in in estimates and efficiency across different working correlation structures.
- Confirmed meeting schedules during the summer.

Next Step

- Simulate the full data so that everyone has 5 visits with 5 outcomes measured at each visit.
- Finish the presentation slides

2024.04.24

Current progress

- Successfully simulated the full data with dropouts for both adjusted and unadjusted model.
- Coverage probabilities all reached over 90%.
- Extremely large estimates were found in results based on dropout data.
- The estimation results across different working correlation structures are consistent regardless of whether or not having dropouts.

Meeting Summary

- Reviewed the presentation content.
- Found the issue of extremely large estimate: non-convergence in some simulation.

Next Step

- Solve the issue of non-convergence.

2024.05.08

- Addressed the issue of non-convergence in simulation.
- Estimates from the full data had comparable MSE and Bias, both at around 0.05.
- Estimates from data with dropouts showed high mean MSE but low mean Bias.
- Performed analysis on the simulated data.
- Created a poster draft.
- Fitted QLS with real data.

Meeting Summary

- Figured out the math part of QLS correlation estimation.

Continuous outcome: $D_i = X$, which is represented by `Sigma_inv_xi`.

Binary outcome: D_i is the derivative of the link function

$$g^{-1}(\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

Steve's Suggestions on Research Day

- Increase the sample size to compare and see how small does the sandwich estimator can maintain a robust estimation.
- Plot the estimation results to visualize the differences.
- Simulate large sample correct first.
- Document the successfully matched pair and performance for each simulation.
- Calculate the tolerance of bias.

Next Step

Modify Peiyu's function so that the QLS estimation is suitable for binary outcomes.

$$\Pr(\text{BAV} = 1|\text{Age, Sex, BSA}) = \text{logit}^{-1}(\gamma_0 + \gamma_1 \text{Age}_l + \gamma_2 \text{Sex}_l + \gamma_3 \text{BSA}_l)$$

2024.05.15

Progress

- Finished the math work for QLS correlation estimation.
- Modified R code so that the generalized error sum of squares is calculated for binary data.
- Had issue in QLS fit mainly due to the difference in correlation estimations between continuous outcomes and binary outcomes.
- Had difficulty in understanding Peiyu's code.

Meeting Summary

In Peiyu's code, * Q_i is the correlation between pairs.
* R_i is the correlation between time points.
* F_i represents the full correlation matrix for everyone.
* The function `Sigma` subseting for matched pair, so it returns a list of all F_i for each pair.
* t_{i1} is the number of visit of the 1st subject in pair i.
* t_{i2} is the number of visit of the 1st subject in pair i.
* `Qinv` is the inverse of correlation matrix.
* Replace line 48 and 49 by line 375-378 in function `movCWGEEgen`.
* Ignore line 28 to 29.
* Our cluster unit is patient.
* In line 37, `Y_var` is all outcomes of the data.
* Modify `y <- as.matrix(Y_var[mdat$cluster == i])` and `x <- x_dat <- as.matrix(X_mat[data$cluster_id==i,])[ -1]`.
* In function line movCWGEEgen, Dmatrix in line 40 to 42 is the

Next step

- Check CWGEE package developed by Aya.
- Apply function `movCWGEEgen` on the data, but use independent working correlation first to see what's going on.
- Add the sigma into the equation like both are independent. Basically both τ and α equal to 1.

2024.05.29

Progress

Successfully fitted data with QLS estimation.

TO ASK:

Why the dimention of $R_i(\alpha)$ depends on the subject who had more measurement?

Meeting summary

Answer to the question:

For example, let's say the i-th pair include one subject with 4 measurements and another with 2 measurements, then

$$Y_i = \begin{bmatrix} y_{i11} \\ y_{i12} \\ y_{i13} \\ y_{i14} \\ y_{i21} \\ y_{i22} \end{bmatrix}, \; Corr(Y_{ijk}) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \tau & \tau\alpha \\ \alpha & 1 & \alpha & \alpha^2 & \tau\alpha & \tau \\ \alpha^2 & \alpha & 1 & \alpha & \tau\alpha^2 & \tau\alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 & \tau\alpha^3 & \tau\alpha^2 \\ \tau & \tau\alpha & \tau\alpha^2 & \tau\alpha^3 & 1 & \alpha \\ \tau\alpha & \tau & \tau\alpha & \tau\alpha^2 & \alpha & 1 \end{bmatrix}$$

Next Step

- Use Steve's simulation process to fit 2 sets of models:

1. Covariates including age, sex, and BSA.
2. Covariates without age, sex, and BSA.

- Fit Simulated Cohort data using GEE and QLS.
- Compare results.

2024.06.05 Meeting Canceled due to SSC Conference.

Breakdown of Steve's Code

Summary

1. Set up
2. Calculate the mean root size at each visit
3. Dropouts simulation
4. Generate Patient profile dataframe
5. Simulation and model fitting
6. Results Aggregation and Bias Calculation

Issue

- There is no matching process.
- Exposure groups are assigned by proportion instead of logistic regression.

Key Points

- Steve assumed the true correlation structure to be AR1 with $alpha = 1$.
- In fitting with GEE, `wave=factor(time)` is added to the function `geeglm`. By Google, the wave argument is used to specify the within-group time points or repeated measures for each subject. This is particularly useful when dealing with longitudinal data where multiple measurements are taken over time for each subject. Do we need to add this during modeling?

Implementation

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]  0.5  0.0  0.0  0.0  0.0  0.0
## [2,]  0.0  0.5  0.0  0.0  0.0  0.0
## [3,]  0.0  0.0  0.5  0.0  0.0  0.0
## [4,]  0.0  0.0  0.0  0.5  0.0  0.0
## [5,]  0.0  0.0  0.0  0.0  0.5  0.0
## [6,]  0.0  0.0  0.0  0.0  0.0  0.5
```

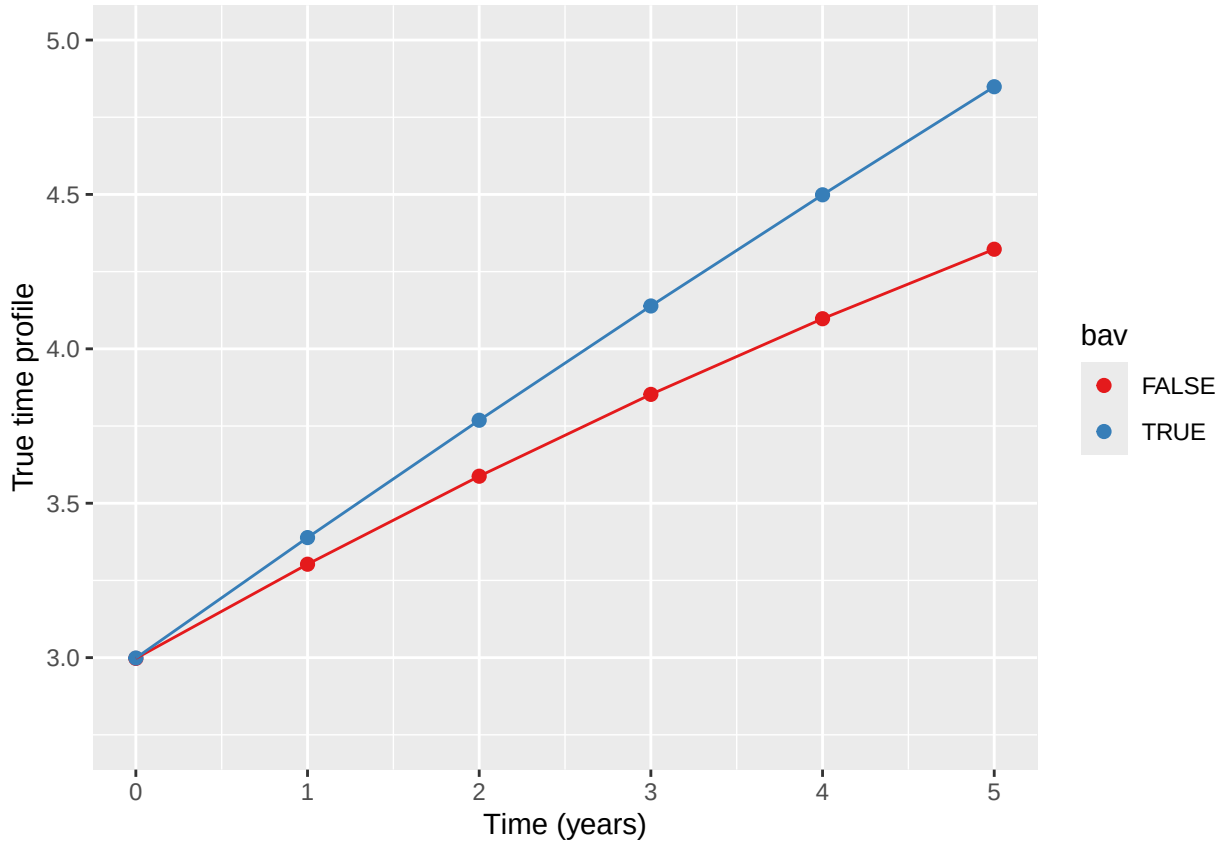
```
##         [,1]   [,2]  [,3]  [,4]   [,5]    [,6]
## [1,] 1.00000 0.6000 0.360 0.216 0.1296 0.07776
## [2,] 0.60000 1.0000 0.600 0.360 0.2160 0.12960
## [3,] 0.36000 0.6000 1.000 0.600 0.3600 0.21600
## [4,] 0.21600 0.3600 0.600 1.000 0.6000 0.36000
## [5,] 0.12960 0.2160 0.360 0.600 1.0000 0.60000
## [6,] 0.07776 0.1296 0.216 0.360 0.6000 1.00000
```

`corr_mat` corresponds to $R_i(\alpha)$.

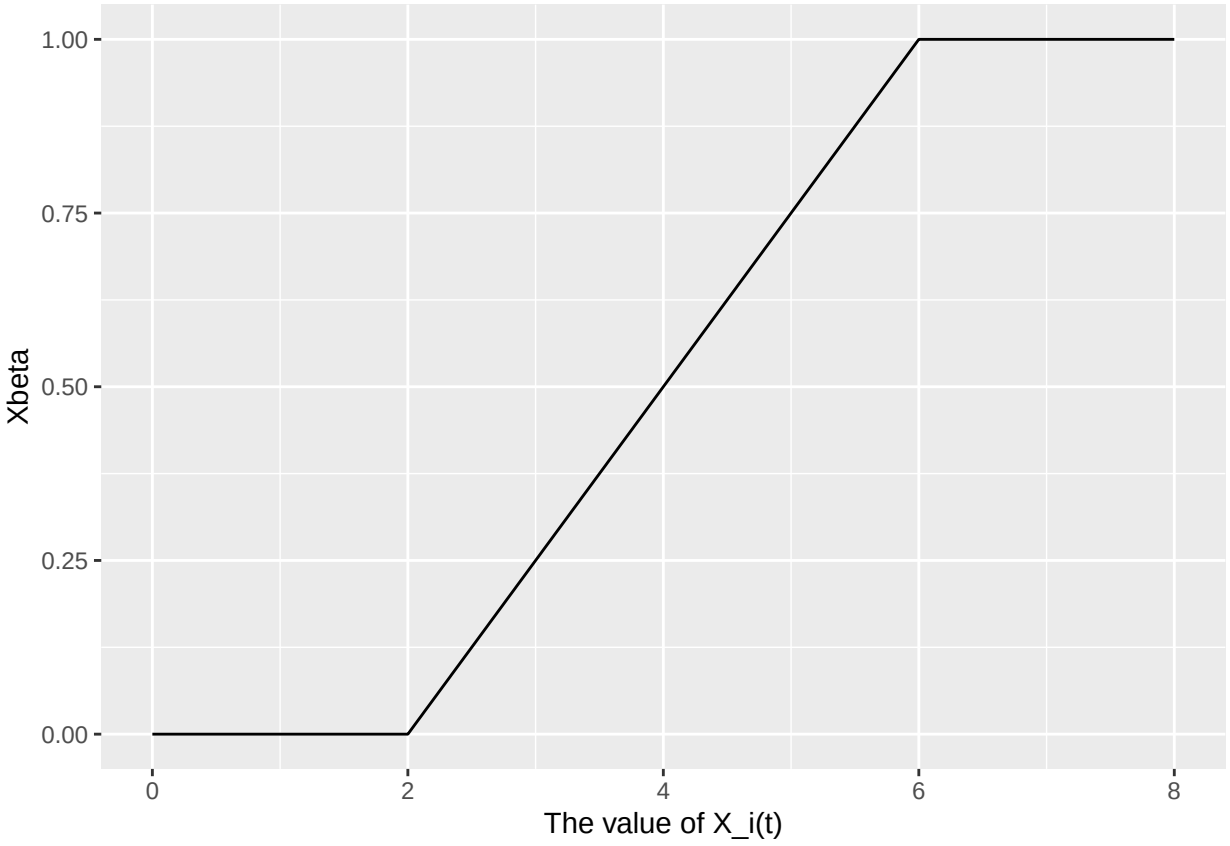
$$R_i(\alpha) = \begin{bmatrix} 1 & 0.6 & 0.36 & 0.216 & 0.1296 & 0.07776 \\ 0.6 & 1 & 0.6 & 0.36 & 0.216 & 0.1296 \\ 0.36 & 0.6 & 1 & 0.6 & 0.36 & 0.216 \\ 0.216 & 0.36 & 0.6 & 1 & 0.6 & 0.36 \\ 0.1296 & 0.216 & 0.36 & 0.6 & 1 & 0.6 \\ 0.07776 & 0.1296 & 0.216 & 0.36 & 0.6 & 1 \end{bmatrix}$$

```
##          [,1]   [,2]  [,3]  [,4]   [,5]    [,6]
## [1,] 0.25000 0.1500 0.090 0.054 0.0324 0.01944
## [2,] 0.15000 0.2500 0.150 0.090 0.0540 0.03240
## [3,] 0.09000 0.1500 0.250 0.150 0.0900 0.05400
## [4,] 0.05400 0.0900 0.150 0.250 0.1500 0.09000
## [5,] 0.03240 0.0540 0.090 0.150 0.2500 0.15000
## [6,] 0.01944 0.0324 0.054 0.090 0.1500 0.25000
```

$$\Sigma = \begin{bmatrix} 0.25000 & 0.1500 & 0.090 & 0.054 & 0.0324 & 0.01944 \\ 0.15000 & 0.2500 & 0.150 & 0.090 & 0.0540 & 0.03240 \\ 0.09000 & 0.1500 & 0.250 & 0.150 & 0.0900 & 0.05400 \\ 0.05400 & 0.0900 & 0.150 & 0.250 & 0.1500 & 0.09000 \\ 0.03240 & 0.0540 & 0.090 & 0.150 & 0.2500 & 0.15000 \\ 0.01944 & 0.0324 & 0.054 & 0.090 & 0.1500 & 0.25000 \end{bmatrix}$$




```
##  (Intercept)         time      bavTRUE time:bavTRUE
##  3.039083333  0.265000000 -0.019541667  0.105000000
```



Ri	Term	Intercept	Time	BAV	TimeBAV
AR1	Mean	3.004	0.292	-0.004	0.097
AR1	Standard Error	0.067	0.063	0.103	0.095
AR1	Bias	-0.035	0.027	0.016	-0.008
AR1	Relative Bias	-0.012	0.102	0.811	-0.081
AR1	Relative Bias	0.006	0.005	0.011	0.009
Exchangeable	Mean	3.004	0.291	-0.004	0.097
Exchangeable	Standard Error	0.068	0.065	0.104	0.100
Exchangeable	Bias	-0.035	0.026	0.016	-0.008
Exchangeable	Relative Bias	-0.012	0.100	0.805	-0.079
Exchangeable	Relative Bias	0.006	0.005	0.011	0.010
Independence	Mean	3.004	0.262	-0.003	0.094
Independence	Standard Error	0.069	0.069	0.106	0.108
Independence	Bias	-0.035	-0.003	0.016	-0.011
Independence	Relative Bias	-0.012	-0.010	0.840	-0.104
Independence	Relative Bias	0.006	0.005	0.011	0.012

2024.06.12

Progress

- Went through each steps in Steve's simulation process and summarized his steps in previous page.
- Re-modeled the dropout process using the real data by setting the time to event to be the total number of visit and the event to be whether or not the total number of visit is greater than 6.

$$\text{Total Number of Visit} = \begin{cases} >= 6, & \text{the patient stayed in the study} \\ \text{otherwise}, & \text{dropped out} \end{cases}$$

The best set of coefficients for simulation comes from the model with bav, age, and sex. *Best* is defined to be the distribution of dropouts at each visit that is the closest to the distribution to the real data.

```
# Call:
#   phreg(formula = Surv(total_visit, event) ~ bav + age + sex, data = surv_data,
#         dist = "weibull", shape = 0)
#
# Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
# bav                 0.781    -0.084     0.920     0.322     0.795
# age                64.696    -0.021     0.979     0.010     0.032
# sex                 0.701    -0.081     0.922     0.278     0.771
#
# log(scale)                    0.611              0.568     0.282
# log(shape)                    0.260              0.092     0.005
#
# Events                    74
# Total time at risk           421
# Max. log. likelihood      -197.49
# LR test statistic         4.47
# Degrees of freedom        3
# Overall p-value           0.214905
```

- Implemented Steve's code and re-documented his simulation results.
- Created a document of simulation process using Steve's method for binary outcome.

TO DISCUSS

1. Steve assumed to be AR1 with correlation coefficient $\alpha = 0.6$. I think this is more reasonable than assume the true correlation structure to be exchangeable because the outcome is identified to be positive either the root size is greater than 45mm or the growth is greater than 5mm, meaning that the next measurement is more associated with the last measurement than the baseline measurement. Therefore, a decreasing correlation trend is expected instead of constant correlation.
2. **How to understand the matrix `sigma_mat`?** Based on the code, its calculation is just the normal matrix multiplication but not the Kronecker product between matrices.
3. The true set of GEE coefficients are determined by applying regular regression on longitudinal outcomes generated by the equations defined by Steve. **Why?**

- Based on the code, Steve first calculated the mean root size at each visit and then simulated longitudinal outcomes using multivariate normal distribution with the mean determined here.

4. Do we need to add `wave=factor(time)` during the GEE fitting process?

$$\text{Y (Root Size)} = \begin{cases} \text{BAV}, & 3.1 + 0.35 \times visit - 0.005 \times (t - 4.5)^2 \\ \text{TAV}, & 3.2 + 0.225 \times visit - 0.01 \times (t - 4.5)^2 \end{cases}$$

5. Need to check the use of cBern.

```
prob_y = c(0.08, 0.2, 0.1, 0.8, 0.88, 0.01)
y = as.vector(cBern(n=1, p=prob_y, rho = 0.6, type = "DCP"))
```

Does these simulated y correlates to each other?

6. The goal of this study for the final report.

Meeting Summary

Meeting Canceled due to illness

2024.06.19

Addressed issue with Git connection and updated the most recent simulation profile and meeting summary.