PUBH 7430 – Assignment 5

You should submit your assignment on Canvas as a PDF file, titled "PubH7430_LastNameFirstName_A5.pdf" with your own name replacing LastNameFirstName. Keep in mind that while working together on homework assignments is permitted, each student is expected to independently write up homework assignments, including any code, in their own words.

Questions

1. **Application of GEE models [24 pts].** The article "Association between Gentrification and Health and Healthcare Utilization" which appeared in the *Journal of Urban Health* in 2022 summarizes the results of a longitudinal study evaluating the association between gentrification and health and health care utilization. Many of the analyses were carried out using techniques we have discussed in this class. Download the article and the supplementary materials at `https://doi.org/10.1007/s11524-022-00692-w`, and answer the following questions:

   (a) [2 pts] What aspect of the study design requires that correlation be accounted for in the analysis?

   (b) [2 pts] The analysis was restricted to individuals with EHR evidence of having lived at the same address in the same block group during both the baseline and the follow-up study periods. Do you think this restriction affected the conclusions and/or generalizability of the study?

   (c) [2 pts] In the statistical analysis section the authors say "GEE was used due to the longitudinal nature of the study where each individual had one observation and another observation at follow-up." Based on what we have learned in class, give two reasons why the authors might have chosen GEE models instead of mixed effects models.

   (d) Consider the two GEE models described in the statistical analysis section.

      i. [2 pts] For the models estimating the difference in the change of health indicators and inpatient and emergency department encounters only the link function (logit link) is specified. What would be an appropriate distribution from the exponential family to assume in this case (e.g. what "family" would you select in fitting this model)?

      ii. [2 pts] The authors accounted for the "within-subject correlation" "by assuming a compound symmetry covariance structure." What does this imply about how "clusters" were defined in this study. What assumption is being made about the correlation between clusters?

      iii. [2 pts] Besides the compound symmetry (exchangeable) structure. What other kind/s of working correlation structure might have been appropriate?

   (e) Consider the parameterization of the models given in supplement 2

      i. [2 pts] Notice that these models do not include a main effect for gentrification (i.e. a $\beta * (G_i = 1)$ term. What does this assume about the outcomes at baseline for individuals in the two types of

neighborhoods (neighborhoods that did not gentrify vs. neighborhoods that did gentrify)? Could this assumption bias the results of the study?.

ii. [2 pts] Now consider the following sentence presented in the "Health Outcomes" part of the "Results" section, "Individuals living in neighborhoods that gentrified as compared to individuals living in neighborhoods that did not gentrify had a lower odds of obesity (OR=0.89; 95% CI: 0.81-0.99). If this estimate represents the odds ratio at follow-up, what expression in terms of the beta terms from the model in supplement 2 are being presented?

(f) [2 pts] Given what you've learned in the class comment on changes you might make to improve the paper.

(g) In the results section of the paper the authors discuss how changing the definition of gentrification changed the width of the confidence interval potentially due to a greater number of block groups eligible (larger sample size). Suppose a new study is proposed in a different state where individuals within block groups are surveyed at a single time point about their health status. Information on the gentrification status (gentrified or not-gentrified) of the block group is also obtained. The primary study question is whether individuals in gentrified neighborhoods have worse or better outcomes than individuals in non-gentrified neighborhoods. Suppose you have to make a decision about which sampling scheme to prioritize with both sampling schemes recruiting approximately the same number of people:
Option 1: Inclusion of more block groups
Option 2: Inclusion of more people within the same block groups

i. [4 pts] Which scenario would we expect to lead to the most precise inference about the association between gentrification and health outcomes.

ii. [2 pts] In thinking through your response to the question in part i you may have thought through how these options would influence the effective sample size of the study. What are three factors in the settings of this study that would influence the effective sample size. (In your response make your answer specific to the study such as talking about block groups and people instead of "clusters" and "observations").

2. **Application of mixed effects models [16pts]** Revisit the "Association of Screen Time and Depression in Adolescence" article we talked about during the first day of class, now having completed a full course on correlated data. This time pay attention to the methods and results section.

(a) Missing data

i. [1pts] What approach do the authors use to account for missing data?

ii. [2pts] Do the authors provide any information on how much of their data was missing?

iii. [2pts] What are some alternatives the authors could have used to account for missing data?

(b) Model description

i. [2pts] The authors say they fit a multilevel model to evaluate the association of screen time. What are the levels in the model?

ii. [2pts] Is this study an example of a crossed cluster design, nested cluster design, or is it unclear from the data description? Explain.

iii. [2pts] Given what we learned about in the class about including random effects in a nested or crossed design, why might the authors have included random effects for multiple levels in this study.

iv. [2pts] Given the description in the paper, attempt to write out the multilevel model the authors used to assess the association of the four types of screen time with depression. (You do not need to include your work here unless you wish to do so). Is the model clearly described. How would you improve this description?

(c) [3pts] Given what you've learned in the class comment on other ways you would improve the paper.

(d) [1pt extra credit] Identify a typo in the paper. There's at least one.