Homework 5

Denis Ostroushko

Question 1

1 - A

Residents within the same neighborhood share the same environment, infrastructure, facilities, and access to food and resources, which can affect their health and health outcomes in the same way. Moreover, depending on what is in the neighborhood, a certain type of people want to more into, or continue to live in the neighborhood. For example, if there are a lot of restaurants, bars, and social activities, younger people may want to live in such place, which would make health outcomes correlated. Moreover, in the U.S., people within the same zip code tend to be of very similar socio-economic status, which would also inherently make the outcomes correlated.

1 - B

Depends of what we mean by generalizability in this context. To me, this question was not clearly enough stated by the instructor.

Qualifying criteria for the study was quite rigid. I interpret that this question has two possible ways to answer it. First, if we want to know "Do these results generalize to the people who lived in the community prior to gentrification, and continue to live in the same community though the years?", then the answer is yes, we can perhaps apply these results to a very specific sub-group of people in the overall population?

If we want to know "Do these results generalize to the overall population? I.e. to a randomly picked community?" then the answer is no. The requirements for the study inclusion is quite restrictive. In the real world, most people are push out of the communities due to gentrification, and the profile of people changes quite a lot. The effect of this is driven by various factors, such as how easy is it to relocate, or what other areas are nearby? So, these results probably do not generalize to the community that is undergoing the process of gentrification.

1 - C

Reason one: gentrification status is time invariant. Interpreting effect of gentrification is interpreting a marginal effect, which is applicable to a GEE model.

Reason two: the sample size is large and the number of clusters is big, which aids sandwich variance estimator.

1 - D

1 - D - (i)

Logit works with probabilities so I would assume Binomial (or Bernoulli when n = 1) family.

1 - D - (ii)

Two clusters are independent, i.e. a person from neighborhood i are not correlated with outcome of a person from neighborhood j. Within clusters, correlation is constant pair-wise between all subjects within a given cluster.

1 - D - (iii)

The only other correlation structure that I would consider appropriate for this case is unstructured correlation structure.

1 - E

1 - E - (i)

When we consider baseline, model equation reduces to $logit(.) = \beta_{k0} + \beta_{k3} * C_i$, which means that there are no differences between gentrified and non-gentrified neighborhoods at baseline. This may lead to bias because neighborhoods are picked to be gentrified for some reasons; the effect of this reasons can be captured by including gentrification model effect at baseline.

If there is already some effect of gentrification at baseline, the overall effect of gentrification between two follow up periods can be overstated.

1 - E - (ii)

Expression of odds ratio in terms of model coefficients is $exp(\beta_{k2} * I(T_i = 1))$

1 - F

Honestly, I dot see a straightforward way to improve the paper based on what we have learned. They have a nice table one, they transparently state the model that they use (outside of what we already addressed in previous parts of this question). Interpretations are clear enough and concise. I would appreciate a more detailed summary of confounder adjustment for various health outcomes, but not in the main text of the paper.

1 - G

1 - G - (i)

Including more people within the same block/neighborhood can lead to more data, which usually leads to more precision in the estimate. Including more neighborhoods can introduce more levels of baseline and demographic data, but can also introduce more variability and patterns in the data.

Thus, holding the number of clusters constant, and increasing sample size within each cluster will lead to more precision.

Holding sample size constant but increasing the number of clusters will not necessarily increase precision in the estimates

1 - G - (ii)

1. Number of people who qualify for criteria for inclusion, i.e. EHR records within the same address. This is a strict requirements, and we have no control over this factor
2. Availability of the data in EHR. If a person did not go to a health facility, and was not recorded, they would be unobservable to us, which makes effective sample size smaller.
3. Size of the city and the number of distinct blocks that we can include in the study. Here we have some control, we can pick a city where the number of unique communities is large and diverse enough.

Question 2

2 - A

2 - A - (i)

Missing data was handled through full information maximum likelihood, which is different from the complete case analysis. This method accounts for missingness while making use of all available data.

Additionally, similar idea to a complete case analysis was applied such that a child with less than 75% complete data was discarded. So, this may lead to some bias in the estimates, such as children who maybe were more depressed were less likely to complete enough data to be included in the study.

2 - A - (ii)

Yes, but not very transparently. 95.6% of children had at least 75% complete data, but we do not know exactly how much data there could have been, and how many observations and in which period had missing data.

2 - A - (iii)

Could have used data imputation such as MICE or nearest neighbor imputation.

2 - B

2 - B - (i)

It sounds like the levels are school and individual children. There were 31 schools and 3659 children in the study.

2 - B - (ii)

It is unclear. However, I presume it is a crossed design because there is a high chance that some children moved schools in the total observational time period.

2 - B - (iii)

If the design is indeed crossed and children were switching schools, then inclusion of multiple levels can help explain additional variation in the data that is due to an extra level of random effects.

If the design is nested, and by some random chance all students remained in one school only over the observational period, then the inclusion of school level does not help explain additional variation, because we already accounted for the lower level (child level).

2 - B - (iv)

The model presented in Table 1:

$$\begin{aligned} E[Depression_{ij}] = \beta_0 + \beta_1 * (Socioeconomic\ status_{ij}) + \beta_2 * I(Sex = M) + \beta_3 * (Social\ Media\ Use_{ij}) + \\ \beta_4 * (Video\ Gaming_{ij}) + \beta_5 * (Television_{ij}) + \beta_6 * (Computer\ Use_{ij}) + \\ a_i + b_i * (Social\ Media\ Use_{ij}) + c_i * (Video\ Gaming_{ij}) + \\ d_i * (Television_{ij}) + e_i * (Computer\ Use_{ij}) \end{aligned}$$

where a_i, b_i, c_i, d_i, e_i are random effects

2 - C

1. In the results section, state the mode type, assumed family for outcome, and a link function for GLMM
2. Table 1 with summary statistics for the collected sample
3. Visualize trajectories of depression development scores for various levels of interacting variables. It would be helpful in visualizing spiraling trajectories that authors are talking about.

2 - D

Table 1 coefficient for television use between person effect has to be bounded by -0.40 lower bound, not 0.40