W203, Section 6

Aidan Jackson, Frank Liu, Sam Temlock, Haoyu Zhang

Cross-Team Review for Team 2

1. General Notes.

- Several thoughtful variables are used as the input for the models. However, the number of variables can be reduced to retain higher degree of freedom, and reduce unnecessary collinearity.
- Good structure to the model building process. Clearly identifies the CLM assumptions and validates them using plots for each model. It is clear that there is a story and a natural progression between each model that builds upon each other.
- The rubric requests that all the R code is included within the knitted file, whereas when plotting in the report it is suppressed from showing up.
- Sometimes the language in the report is very professional, whereas other times it may slip to being informal. Generally, it's better to not use personal pronouns such as "I" or "we", which sometimes show up in the writing.
- Occasionally the writing contains subjective claims or ideas that come off as opinions, which should not be included in a professional report. An example of this is when the paper discussed "anecdotal news" regarding COVID among different age groups.
- Generally, the best practice for introducing and talking about figures is to display the figure first and then include the text discussing it immediately afterwards. The Data Exploratory Analysis section could benefit from this by interspersing the text after the relevant figures rather than including all the text at first.
- May want to be more up front that this is a descriptive model rather than an explanatory one. This relates to being clear with potential subjective claims about what the model could be used for.

2. Introduction.

Is the introduction clear? Is the research question specific and well defined? Does the introduction motivate a specific concept to be measured and explain how it will be operationalized. Does it do a good job of preparing the reader to understand the model specifications?

- In general the first few sentences are very well written, and clearly convey the motivation for the general study as well as what variables the report focuses on in particular.

- It could be more explicitly why the goals of the model relate to being descriptive instead of causal
- Good job of relating the motivation and your research question, although could expand further on the operationalization of the variables (such as categorical variables such as age categories)
- It is helpful to hear how the model may be limited, as it puts the work in the right context going forward. Specifically, the example about the relationship between pre-existing conditions and age was included well.
- Again in this section, it doesn't seem correct or worth the space to describe the model as being potentially applicable to other diseases without a more in-depth discussion on what it would take for that to be true.
- Nice description of the research background, it is helpful that it also includes the discussion about limitations on the research caused by the dataset.
- Research question is clear and specific about what is being measured and what is being used as input
    - Current research question expressed focus only on the age group features
    - Should also add the sub questions to describe the other variables that are considered and will be included in the models

3. **The Initial Data Loading and Cleaning.**

Did the team notice any anomalous values? Is there a sufficient justification for any data points that are removed? Did the report note any coding features that affect the meaning of variables (e.g. top-coding or bottom-coding)? Overall, does the report demonstrate a thorough understanding of the data?

- Overall the team selected good data sources that justifies the research purpose.The team did cover anomalous values on certain variable selecting processes.
- Based on the knitted report, there isn't much mention of any data cleaning or anomalous values that were encountered by the team. Would be good to show more work.
- The results from the preliminary data analysis, however, show a thorough investigation of the most important variables such as the outcome variable of covid cases per capita and the age ranges as a fraction of total population. Top vs bottom coding on certain variables is discussed indirectly, such as the limits of the mask mandate policy, but not for general data verification or cleansing.
- As mentioned in the general notes, this section in particular would benefit greatly from being able to see the code that was executed to clean the data. It could also help to more specifically state that the first paragraph will be about how you chose to operationalize some variables to provide more context to the reader, such as those variables which are defined by dates.

- No justification for using mask mandate policies as a policy for “End stay at home/shelter in place”. Can’t see how these are at all related, as they intuitively seem independent. Same with the “stay at home/shelter in place” proxies
- Not enough explanation or justification of the inclusion/exclusion of anomalies
- Need to consider rearranging the discussion order of data explornary according to the models sequence. Most readers would not expect “Closed Business Days” as the very first variable discussed in this section
- The plot of the population density variable needs adjust or data transform for better display purposes. In addition, discussion about the outlier lacks supporting evidence.
- The axis label for the “face mask policy” plots is “0, 10000, 20000, ……”, which can be further clarified
- The variable of “Adults Who Have A Pre−Existing Condition” is a population variable instead of a population rate/density type as others. Need some transform for the model building process.

4. **The Model Building Process.**

Overall, is each step in the model building process supported by EDA? Is the outcome variable (or variables) appropriate? Is there a thorough univariate analysis of the outcome variable? Did the team identify one, or very small number of explanatory variables and perform a thorough univariate analysis of each one? Did the team clearly state why they chose these explanatory variables, does this explanation make sense in terms of their research question? Did the team consider available variable transformations and select them with an eye towards model plausibility and interpretability? Are transformations used to expose linear relationships in scatterplots? Is there enough explanation in the text to understand the meaning of each visualization?

- Overall the model building process is supported by the EDA.The outcome variable is appropriate and can certainly form a worthwhile research topic. Nice job on the evaluation of CLM assumptions included in each model.
- One thing can be improve is that the EDA can also really explain any potential anomalies
- References to figures are off, would be better to form a storyline or talk about figures that are close to the text
- Distributions of each explanatory variable is identified but there is no analysis on operationalization or dealing with non-normal distributions
- There are no arguments regarding transformation of variables which could be important as inputs to the models
- Figures mostly focusing on the histogram plots, would be nice to include scatterplots that show the linear relationship between the control variables and outcome variables.

- There are definitely some explanation in text to understand visualizations, but could expand more in the final report
- Recommend more details and explanations on Figure 4, 8, 9, 12

5. **Regression Models.**

    i. **Model 1 -** Does this model only include key explanatory variables? Do the variables make sense given the measurement goals? Did the team apply reasonable transformations to these variables, to capture the nature of the relationships?

    - Good use of a basic model, only considering key explanatory variables that help describe the research question.
    - Model 1 is analysed with age group variables as input which is splitted as 6 ranges variables. The variable in general makes sense to the measurement goal.
        - However, because the sample size (51 data entries) is considered to be a small data size, introducing more variables will result in losing degree of freedom. As a result, subsampling the age group variables into aggregated or representative variables could yield more trustworthy regression results.
        - Also, maybe reconsider the titles of the model plots
        - No discussion about the coefficient test output or practical significance.
    - No transformation was applied to the age distribution variables since the age group ratio was already of reasonable form for basic model regression.
        - Also, the variables although not in perfect collinearity, do contain a certain degree of correlation. In fact, one age group ratio can be represented by a linear combination of other groups ratios. Even though R didn’t detect the multicollinear here, there may be some near collinear situation here.So at least one age group variable can be dropped for the analysis.

    ii. **Model 2 -** Does this model represent a balanced approach, including variables that advance modeling goals without causing major issues? Does the model succeed in reducing standard errors of the key variables compared to the base model? Does it capture major non-linearities in the joint distribution of the variables?

    - The model is well-balanced, including logical additional predictor variables that do not conflict with the modeling goals nor the variables in the basic models
        - There is a lack of discussion about possible transformation of the newly-introduced variables in model 2, especially for the non-ederly with pre-existing conditions.

    - The standard errors of the most key variables are reduced compared to model 1, according to the regression table in the following section.
    - However, variables included can be seen to be nonlinear from the coefficient test. It may be useful to include some plots of the additional variables to demonstrate their relationship, maybe explore potential transformations.

iii. **Model 3** - Does this model represent a maximalist approach, erring on the side of including most variables? Is it still a reasonable model? Are there any variables that are outcomes, and should therefore still be excluded? Is there too much multicollinearity, to the point that the key causal effects cannot be measured?

- Model 3 includes many extra variables compared to the advanced model. It is hard to judge if it represents a maximalist approach.
    - Regression results here prove the robustness of the advanced model.
- The model is reasonable, with a clear story behind why they included the additional variables about the mandatory mask policy.
    - All new-introduced variables here are reasonable.
    - Considering the inclusion of ethnicity variables.
- Tests per 100k is probably an outcome variable.
    - There may be some collinearity between tests per 100k and cases per 100k (the more people are tested, the more cases can be detected).
    - This variable could potentially be excluded as this could be an outcome variable. At least this discussion needs to be included in this discussion
- There is potential correlation between the age categories and the nonelderly adults with condition, although some data exploratory work could justify this inclusion (also, this model is descriptive, not explanatory)

6. **Plots, Figures, and Table.**

Do the plots, figures and tables that the team has chosen to include successfully move forward the argument that they are making? Do they have a good ratio of “Information to Ink” (Tufte)? Has the team chosen the most effective method (a table or a chart) to display their evidence? Is that table or chart as communicative as it can be? Is every single plot, figure, or table that is included in the report referenced in the main text?

- The plots/figures/tables included all help to move forward the argument and explain different observations highlighted.
- There is a good rationale for including all the plots that have been plotted.
- Although all plots are referenced in the main text, many of the figures lack a detailed explanation and analysis regarding the relationships displayed within them. There

could be some expansion on the explanation of the figures in the text and inclusion of more discussion on the insight gleaned from them.

- There could be a better flow between the text and the figures. Sometimes too many figures are introduced at the same time in one paragraph of text, making it difficult to follow.

7. **Assessment of the CLM.**

Has the team assessed each of the CLM assumptions (including random sampling)? Did they use visual tools or statistical tests, as appropriate? Did they respond appropriately to any violations?

- Most of the CLM assumptions are discussed, although the random sampling assumption as part of the data being IID could be more directly stated. It didn't appear that IID is discussed explicitly.
- There is no mention of collinearity between the chosen variables in the model.
- If a plot is used to make a judgement on an assumption, it would be helpful to include that particular plot in this section rather than in the model building section proceeding it.
- To make the argument about homoscedasticity stronger, a Breusch-Pagan test could be used in addition to visually assessing through a plot.
- Violations were sometimes reported, such as with the population mean of zero of the error term, but discussion of how this would impact the usefulness of the model was not included.
- Using lm() for the regression model doesn't guarantee the linearity of the actual model relative to the data.
- "As we expected, the more variables we add, the model became more accurate, because it is closer to the real problem." This discussion about the model limits is vague, consider specific analysis according to your research question. A discussion about degrees of freedom between number of variables and number of datapoints would also help this.
- Good structure to list potential CLM assumptions violations case by case, although it would be better to name the assumptions specifically for the reader.

8. **A Regression Table.**

Are the model specifications properly chosen to outline the boundary of reasonable choices? Is it easy to find key coefficients in the regression table? Does the text include a discussion of practical significance for key effects?

- Model specifications are properly chosen as commented on in the sections above.
- The regression table is clear and significance within the coefficients is highlighted.

- It may be useful to provide more explanation/comments on the analysis of the significant variables (such as factors that contribute).
- There is no explanation of why the adjusted R2 decreases from model 1 and model 2 (a key discrepancy).
- More elaboration on practical significance of key effects could be provided.
- Would be helpful to add some explanation on the degree of freedom associated with residual standard error, and how a reduction of degree of freedom may affect the result.

9. An Omitted Variables Discussion.

Did the report miss any important sources of omitted variable bias? Are the estimated directions of bias correct? Was their explanation clear? Is the discussion connected to whether the key effects are real or whether they may be solely an artifact of omitted variable bias?

- No comment on any omitted variables discussion given that this model is a descriptive one rather than an explanatory one. However, as commented on by the instructor during our last live session, reports should be updated to include discussion of omitted variables even within the descriptive models, so this can be included in the final report. The requirements for this section with regards to a descriptive model may have to be followed up on with the instructor.

10. Conclusion.

Does the conclusion address the research question? Does it raise interesting points beyond numerical estimates? Does it place relevant context around the results? Are there any other errors, faulty logic, unclear or unpersuasive writing, or other elements that leave you less convinced by the conclusions?

- The model and conclusion clearly refers back to and addresses the result of the analysis with regards to the research question.
- The conclusion raises interesting suggestions beyond the numerical estimates such as the impact of certain variables towards the prevention of the spread of COVID-19 cases (i.e., practical significance).
- The conclusion does not reiterate/highlight any potential limiting factors, errors, or discrepancies that exist within the model that could serve as caveats to the general conclusion regarding the effectiveness of the model.
- The conclusion seems to “exaggerate” the function of the descriptive model by stating this model is able to predict the expectation of change of certain policies. It is important to be deliberate in terms of the application of a descriptive model.