Pre-Analysis Plan

By Anran Zhao, Dylan Myaing, Michelle Cheng, Helena Moore, Matt Lau, Paula Delgado

1. What is an observation in your study?

An observation in this study could be a county in the United States. Each observation will represent a county's socioeconomic, health, and disaster-related data. The datasets we collected could also provide further information by county on educational attainment, poverty rates, health outcomes, and natural disaster impacts.

2. Are you doing supervised or unsupervised learning? Classification or regression?

This analysis will use supervised learning since we have labeled data (e.g., health outcomes) and are trying to predict specific outcomes based on other variables like poverty rates and educational attainment. The task is primarily a regression problem because we are predicting continuous outcomes such as long-term health impacts rather than classifying into discrete categories. Thus, part of our analysis will definitely include trying to find the best fit line to predict the output.

3. What models or algorithms do you plan to use in your analysis? How?

We plan to use a combination of regression models and techniques to assess the relationship between socioeconomic factors and long-term health impacts in disaster prone areas.

- Linear Regression: This will establish a baseline model for the relationship between socioeconomic factors (e.g., poverty, education) and health outcomes.
- Random Forest Regression: This helps capture more complex, non-linear relationships among variables, which could be missed by linear models.
- LASSO Regression: This will manage potential multicollinearity and perform feature selection in the case there are various correlated socioeconomic variables.
- Principal Component Analysis (PCA): This will reduce dimensionality if there are many correlated numeric variables (e.g., GDP, unemployment), simplifying the feature set and improving model interpretability.

4. How will you know if your approach "works"? What does success mean?

Success will be measured by the model's ability to predict long-term health outcomes based on socioeconomic and disaster-related factors. We will evaluate model performance using:

- R^2: To assess how much variance in health outcomes is explained by the model.
- Root Mean Square Error (RMSE): To evaluate prediction accuracy.
- Accuracy, Sensitivity/Specificity, F1 Score: We can use these metrics in case of classification tasks, such as if we shift to predicting binary outcomes ("high vs low vulnerability").

If we see strong performance on these metrics across both training and test datasets, without signs of overfitting, we can consider the approach successful.

5. What are weaknesses that you anticipate being an issue? How will you deal with them if they come up?

Anticipated weaknesses include:

- Data Sparsity or Missing Data: Some counties may lack data for certain variables. We can address this by imputing missing values or excluding counties with significant missing data.
- Geographical Aggregation Issues: Some data may only be available at the state level rather than county level, leading to mismatched granularity. We might need to aggregate county-level data up to the state level or find additional data sources at the county level.
- Multicollinearity: Socioeconomic variables (e.g., GDP, unemployment, poverty rates) may be highly correlated, which we can handle using LASSO regression or PCA to reduce dimensionality.

If our approach underperforms (e.g., low R^2 or high RMSE), we will re-evaluate feature selection, consider non-linear models such as Random Forests, or explore interactions between additional variables.

6. Feature Engineering: How will you prepare the data specifically for your analysis?

We plan to perform the following feature engineering steps to refine the data for more effective analysis and modeling:

- One-hot Encoding: For categorical variables like disaster types or regions to include them as model-friendly features.

- Normalization/Standardization: For continuous variables such as GDP and unemployment rates to ensure comparability across features.
- Dimensionality Reduction: Use of PCA if many socioeconomic variables are highly correlated.
- Interaction Terms: Creation of interaction terms between key predictors (e.g., poverty rate * educational attainment) to capture potential synergies.

7. Results: How will you communicate or present your results?

We will present results using:

- Regression Coefficients Table: For linear models, showing how each predictor influences long-term health impacts.
- Feature Importance Plots: For Random Forests, to highlight variables contributing most to predictions.
- Performance Metrics: Including R^2, RMSE, accuracy, sensitivity/specificity, depending on the model used.
- Visualizations: Scatter plots or heatmaps to illustrate relationships between key predictors (e.g., poverty rate) and health outcomes across counties.