**Agenda**:

* Discuss thesis motivation (identify gap in literature)
* Discuss my many questions about my method

**Motivation for thesis**: I am not sure if this is a strong enough motivation, or too much like ‘copying’ what has already been done

* Changed data pipeline to prevent the need to impute subgroup MMR with no ground truth
* There exists a modelled MMR estimate from the WHO, with some countries missing
  + My thesis at the moment uses many more socioeconomic variables to estimate the ratio than used by the WHO (allowing a more in-depth predictive power assessment to be carried out and be used to inform policy)
  + Many methods (from my literature search) exclude variables with >90% missing data
    - I do not (at the moment). Should I? This would reduce the number of features from 730 to 540.
    - Could potentially allow MMR to be predicted from any of a subset of variables?
* If these are not strong enough motivations, should I solely focus on splitting the data by year, and using my model for prediction?
* Check types of models
  + More specific than solely regression
* Need to point out difference
* Can do other variables
* Argue for whether machine learning is necessary
* Have a model that works at country-level
* Train model for low income countries if I see differences between the levels
  + Could be better to do model for each level
  + Just middle income

**Overall Flowchart of Experimental Data-Processing Process:**

Do I add in ‘removing variables with >90% missing’?

Optional bootstrapping

Merged dataset

Optional normalisation

For each of the folds (to prevent data leakage).

Standardise the national and modelled estimates, then average the standardised estimates. Rescale to a meaningful unit using the modelled estimate’s original mean and standard deviation. (see Equations 1a and 1b).

The train dataset is split randomly 80:20 into a train and validation set (with a specific area only belonging to either the train or validation set). This is performed for 5 80:20 permutations of the data.

For each of the datasets

Not thresholded and uncorrelated

If applied, strong Pearson correlation coefficients between feature columns were used to signal that one column could be used to impute the values of the other column. A first order regression model was fit to the column with more data and used to impute the values of the column with less data. See Section 4.23 for exact thresholds used to qualify columns for this type of imputation.

Used either no thresholding or 95%

Iteratively remove rows and columns until all rows and columns have less than the threshold proportion of missing data.

Not thresholded and correlated

Thresholded and uncorrelated

Tresholded and correlated

This process should produce 4 datasets

The dataset was split into four mini-datasets, each containing data belonging to all areas within one of the World Bank’s defined income levels, Each mini-dataset was split into train, test sets in a 90:10 proportion. All data for a specific area was either placed in the train or test set to make the sets independent. Merging the mini-datasets produced one complete train and test set.

Imputation

Iterative imputation (Bayesian ridge regression or support vector machine)

Regression model (orders 1, 2, 3)

Miss Forest

k-Nearest Neighbours

No imputation

Merging national and modelled MMR estimates

Cross-Validation and encoding

Split each of the datasets into 90:10 train/test split.

Optional correlation-based imputation

Iterative thresholding of entire merged dataset

Removing samples that have an NAN value for the modelled and national MMR estimate.

**Overall Flowchart of the Model Training Process:**

**ha**

PyTorch’s Multi-Layer Perceptron

Scikit Learn’s Support Vector Machine

Scikit Learn’s k-Nearest Neighbours Regressor

Scikit Learn’s AdaBoost

XGBoost

LightGBM

Scikit Learn’s Random Forest

Iterative imputation (Bayesian, SVM)

No imputation

Regression: orders 1, 2, 3

Miss Forest

k-Nearest Neighbours

For the non-imputed dataset (4base\*1imputation\*5folds), apply the ML models

For each imputed dataset (4base\*10imputation\*5folds), apply the ML model

Data Pre-processing and imputation

Symmetrical MAPE

Apply Optuna to each model for 300 trials to test hyperparameter combinations, minimising mean squared error

Compare combinations using the Friedman test

Symmetrical MAPE

Root Mean Squared Error

R-Squared Error

For each dataset, test each model on all the folds on the test set. Take mean and standard deviation to judge if the folds’ results are significantly different

Root Mean Squared Error

R-Squared Error

Somehow combine the folds’ models?

Compare all combined models on a test set?

Compare combinations using the Friedman test

Take the best performing imputation method and use multiple imputation to get error range

Take the best performing models and test ensemble based

Sensitivity analysis?

Use SHAP to find predictive power of variables

**Questions about experimental process**:

* Should I average the national and modelled MMR estimates, or should I just drop the national estimate from my data?
  + If I drop all the missing national estimate data, there are no further missing modelled estimates (so keeping national data in would not prevent rows from being dropped).
* Should I remove variables with more than 90% missing data?
  + Sources say that with this amount of missing data, bias is unavoidable
  + *Should I do this through the iterative thresholding approach?*
* At the moment, I apply the correlation-based imputation before the train/test split. However, would this cause data leakage, as data that would go into the train set could affect the test set? Should I include this method after the train/test split? Should I include it at all, given that a similar method is applied using scikit learn’s iterative imputer?
  + Similarly, should encoding be done on the train set before or after splitting into cross-validation sets?
* Should I include the bootstrapping and/or normalisation before or after imputation?
  + Does this go before or after the train/test split?
* Do I need to be keeping all these datasets simultaneously, or can I introduce tests throughout and only carry forward the best performing data/model combinations
  + As this is a lot of datasets to keep track of, especially considering the folds
* How to compare performance of models and imputation methods given there are so many combinations and variables? Statistically significant difference?
  + Friedman test?
  + Linear mixed effects model?

**Questions about thesis structure:**

* I will be including information about how different patterns in missing data are categorised to justify my choice of imputation method
  + Will this go in the background information or as part of the methods?
* I will be speaking about how I explored the pattern of missing data within my own data set
  + Does the description of the way I’ve explored the data go in the methods, but the results from the exploration go in the results section?
    - What if I use these results to help justify my choice of imputation method?
  + Similar question for how I should discuss my merging of the modelled and national estimates. I justify my methodology using exploration of the data. Which part should go in methods versus results?
* In terms of data sources, can I solely reference the page from which I downloaded the data, or do I need to cite the page’s original source of data?
* Should I put my exploration and manipulation of the target variables in the methods or results, and within the
* I include lists in the method to show the steps more clearly?

**Questions about previous student’s thesis:**

* They used the Kolmogorov-Smirnov test to represent empirical distributions of branch lengths and evolutionary model parameters using the best approximate parametric distribution for the purpose of simulating data
  + Is this method relevant for my thesis?

**General Questions:**

* Can I have the download links for the datasets you gave me at the beginning of the year?