PRS Analysis for WSL

Lukas Graz

Release Notes

V0.1

- Initial release (Technical Setup)
- Data Preparation
  - Conducted sanity checks on data (duplications, type consistency, and comparison with `df_varlookup_for_lukas.xlsx`)
  - Performed type encoding
  - Created 14 dictionaries for translating ordinal categorical variables to numeric (e.g., mapping: Very:5, Quite:4, Fair:3, Little:2, Not:1)
- Filtered observations according to criteria described in `_INFO_for_Lukas.docx`
- Missing Values
  - Checked patterns of missingness for PRS-Variables, Mediators, and GIS-Variables
  - For PRS-Variables: Compared imputation methods: MissForest, Column-wise Mean, Observation-wise Mean
  - Imputed missing values for all relevant variables using MissForest (each chunk separately) - later to be done separately for Train/Test data
- Initial Modeling Completed - results to be updated with imputed data and PC1-4

Main Analysis

Which Response to Use?

Initial Idea was to use: - Aggregated MEAN - PRS1 (Fascination) - PRS2 (Being Away) - PRS3 (Extent Coherence) - PRS4 (Compatibility)

Verify if this is a good approach with **PCA**. Findings so far:

- PCA suggests that the data can be well approximated with 3-4 dimensions.
- Unsurprisingly, the first dimension is close to a weighted average of all variables. Projecting on PC1 yields a correlation >0.99.
- PRS3 (Extent Coherence) differs the most from the others (see PC2)
- PRS1 (Fascination) and PRS2 (Being Away) are rather similar (see PC1-PC3)
- The aggregated PRS variables are also justified given the PCA results (similar values in rotation). Therefore, it is justified to use the mean.

We will continue to investigate the PCA projections as alternative response variables (along with PRS1-4).