PRS Analysis for WSL

2025-02-14

# Release Notes

## V0.2

* Transformations of GIS variables (mostly sqrt)
* Machine Learning in mlr3 framework:
  + Testing prediction quality of GIS\_vars → Mediators → PRS\_vars
    - comparing performances + inference
  + Linear models, Random Forests, XGBoost (with parameter tuning)
* Improved code structure by separating data preparation, machine learning, and hypothesis testing
* Hypothethis testing:
  1. Train data X imputation using MissForest (simplified approach as uncertainties not needed for feature selection)
  2. Feature selection based on correlation/VIF analysis
  3. Feature selection implemented due to high VIF
  4. Additionally inspecting all interactions

## V0.1

* Initial release (Technical Setup)
* Data Preparation
  + Conducted sanity checks on data (duplications, type consistency, comparison with df\_varlookup\_for\_lukas.xlsx)
  + Performed type encoding
  + Created 14 dictionaries for translating ordinal categorical variables to numeric (e.g., Very:5, Quite:4, Fair:3, Little:2, Not:1)
* Filtered observations according to criteria 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 using MissForest (each chunk separately) - to be done separately for train/test data
* Initial modeling completed - results to be updated with imputed data and PC1-4

# Data Preparation

# Main Analysis

### Response Variable Selection

Initial approach: - Aggregated MEAN - LA (Fascination) - BA (Being Away) - EC (Extent Coherence) - ES (Compatibility)

**PCA Verification** of this approach. Key findings:

* Data can be well approximated with 3-4 dimensions
* First dimension is close to weighted average of all variables (correlation >0.99)
* EC (Extent Coherence) shows most divergence (see PC2)
* LA (Fascination) and BA (Being Away) show similarity (see PC1-PC3)
* Aggregated PRS variables justified by PCA results (similar rotation values), supporting use of mean

PCA projections will be further investigated as alternative response variables (alongside LA-4).

## Prediction Analysis

Details in [the notebook](notebooks/mlr3.qmd).

Following RESTORE project approach, investigating links between PRS, Mediators, and GIS variables using: - Linear Models - XGBoost (with trees + parameter tuning) - Random Forests

Evaluation: Percentage of explained variance on hold-out data. Missing values imputed with MissForest (no p-values → no assumptions needed).

### Results

* GIS shows limited predictive power for PRS on ES (5% variance explained)
* GIS + Mediators explain 25% of PRS variance
* Mediators alone explain majority of PRS variance
  + GIS primarily helps with ES through tree-based methods
  + Suggests GIS effect is more interaction-based than direct
  + Similar reduction in tree-based methods observed in BA

## Hypothesis Testing

Details in [the notebook](notebooks/hypothethis-tests.qmd).

Process: 1. Train data X imputation using MissForest (simplified approach as uncertainties not needed for feature selection) 2. Feature selection based on correlation/VIF analysis 3. Feature selection implemented due to high VIF

### Results

* Continuous mediator outperforms categorical (scaled to mean 0, sd 1)
* HM\_NOISELVL not significant (pre-p-adjustment)
* Full mice NA-handling likely unnecessary
  + Models use few variables
  + Only LNOISE shows high NA count
  + Information detection still fails
* Significant edges remain in SEM (see all interactions)