ENCLOSE: Non-Parametric (Hot-Deck) Imputation

Workflow: Harmonization, Imputation, and Analysis Based on the hotdeck_imputation_updated.Rmd file

Contents

1. Data Preparation & Latent Trait Estimation

Input files:
- `EES10.sav`
- `ZA7649_v2-1-0.sav`

Code: `ENCLOSE_data_harmonization_clean.Rmd`
Output: `ENCLOSE_harmonized_data.RData`

Description:
Harmonizes datasets, aligns variable coding, estimates latent traits.

2. Data Loading & Setup

Input File: * `ENCLOSE_harmonized_data.RData`

Code: `hotdeck_imputation_updated.Rmd` (Sections 1 & 2)

Output: `donor_data` and `recipient_data` dataframes in memory.

Description: Loads the harmonized data file, which contains both the donor (ESS) and recipient (Eurobarometer) data. It splits the combined dataset into two separate dataframes (`donor_data` and `recipient_data`) and defines the common, target, and grouping variables for the matching process.

---

3. Non-Parametric (Hot-Deck) Imputation

Input: `donor_data` and `recipient_data` dataframes

Code: `hotdeck_imputation_updated.Rmd` (Section 3)

Output: `recipient_imputed` dataframe

Description: Performs a non-parametric statistical (hot-deck) imputation using the k-Nearest Neighbor (k-NN) method with k=5. The process is applied on a country-by-country basis. For each record in the recipient dataset, it calculates the Gower distance to find the 5 most similar records in the donor dataset (within the same country) and randomly selects one of them to impute the target variables `ALLOW` and `FEELING`.

---

4. Imputation Quality Analysis

Input: `recipient_imputed` and `donor_data` dataframes

Code: `hotdeck_imputation_updated.Rmd` (Section 4)

Output: * Density plots (original vs. imputed data) * Tables with descriptive statistics and similarity measures (Overlap Index, Hellinger Distance) * Regression coefficient tables and plots * Correlation and convergent validity plots (e.g., scatter plots, QQ-plots)

Description: Evaluates the quality of the imputation through three main analyses: 1. **Marginal Distributions:** Compares the distributions of the imputed variables with the original donor variables. 2. **Relationship Preservation:** Checks if the correlations between variables and the results of regression models are similar between the donor dataset and the recipient dataset with imputed values. 3. **Convergent Validity:** Compares the imputed `FEELING` variable with a similar variable (`FEELING_ZA`) already present in the original recipient dataset.

---

5. Summary Workflow Diagram