

Supplementary materials

“Choosing Imputation Models”

Moritz Marbach

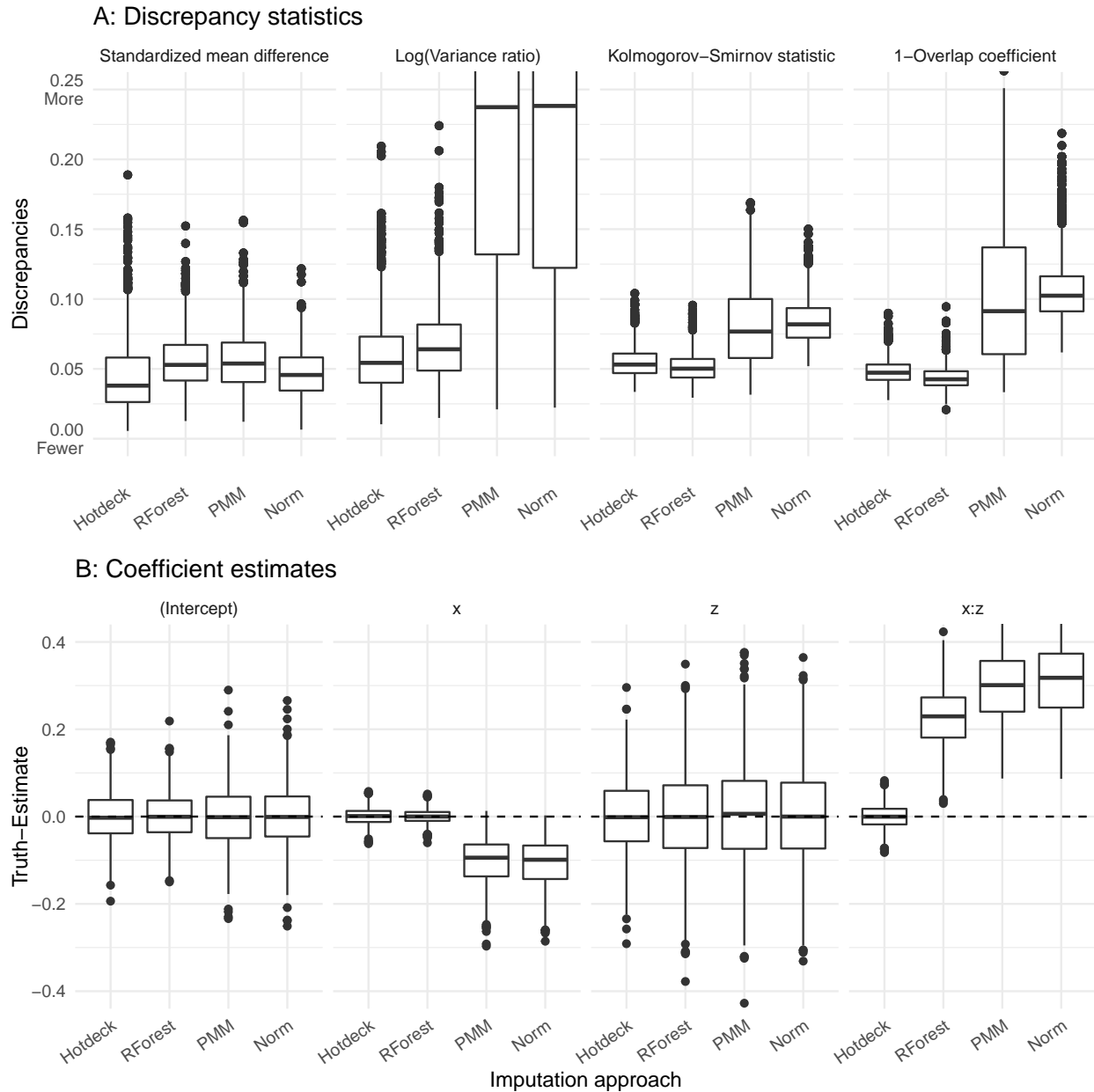
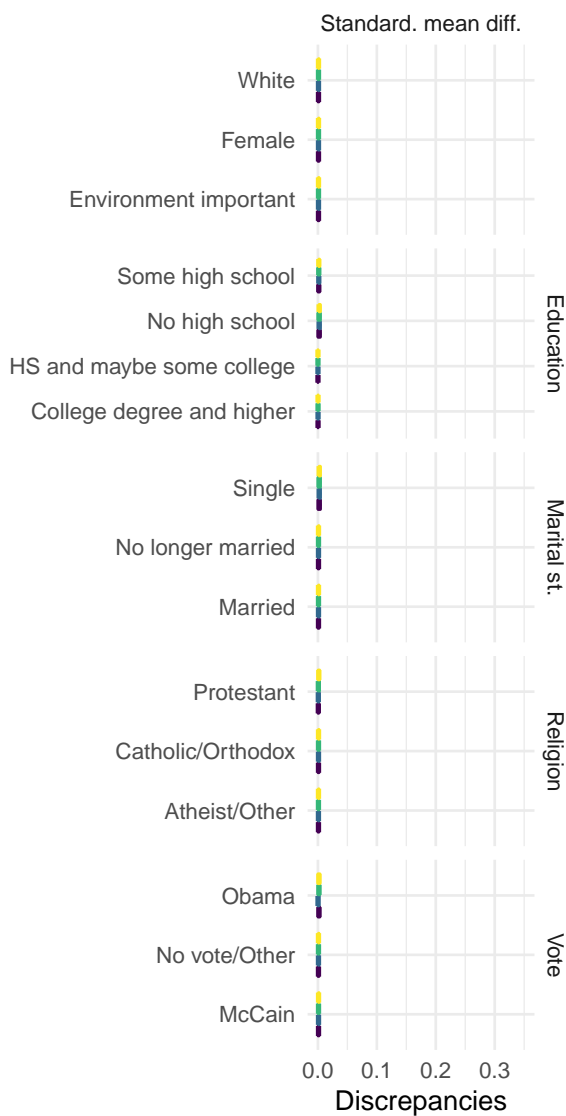


Figure A.1: Summary of the 1,000 Monte Carlo simulations. Panel A: Box plots of the four discrepancy statistics comparing the weighted density of observed and imputed values. *Weights are constructed to balance the first and second moments of the covariate densities.* Panel B: Box plots of the bias in the (pooled) coefficient estimates of a linear model $y = x \cdot z + \epsilon$. Each dataset was imputed five times using hot-deck imputation (Hotdeck), random forest imputation (RForest), predictive mean matching imputation (PMM) and normal model imputation (Norm). Some box plots are clipped to increase readability.

A: Binary variables



B: Continuous variables

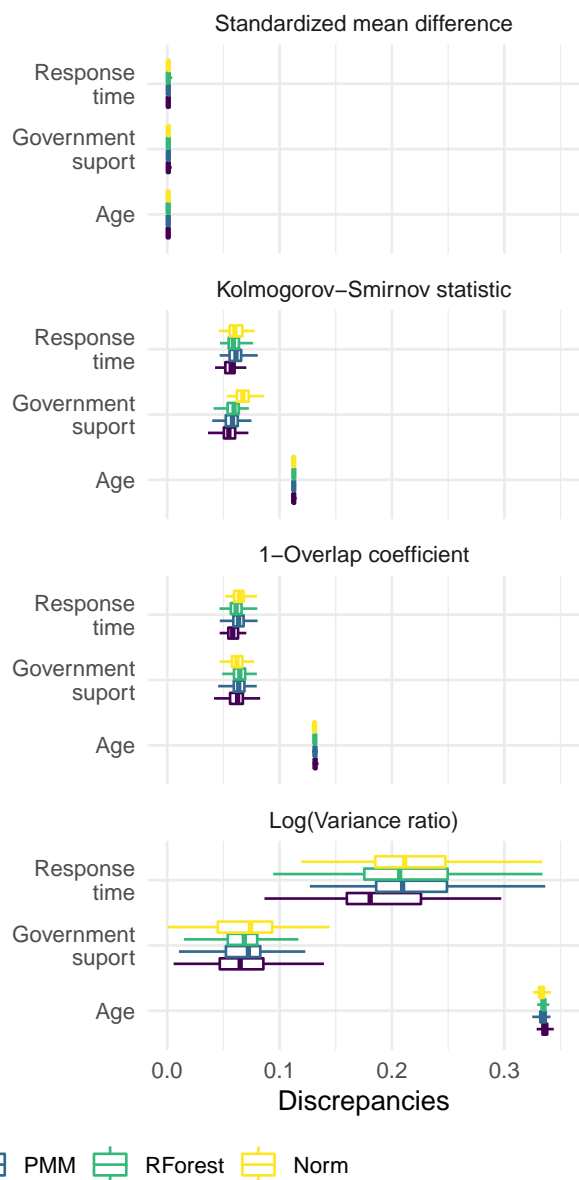
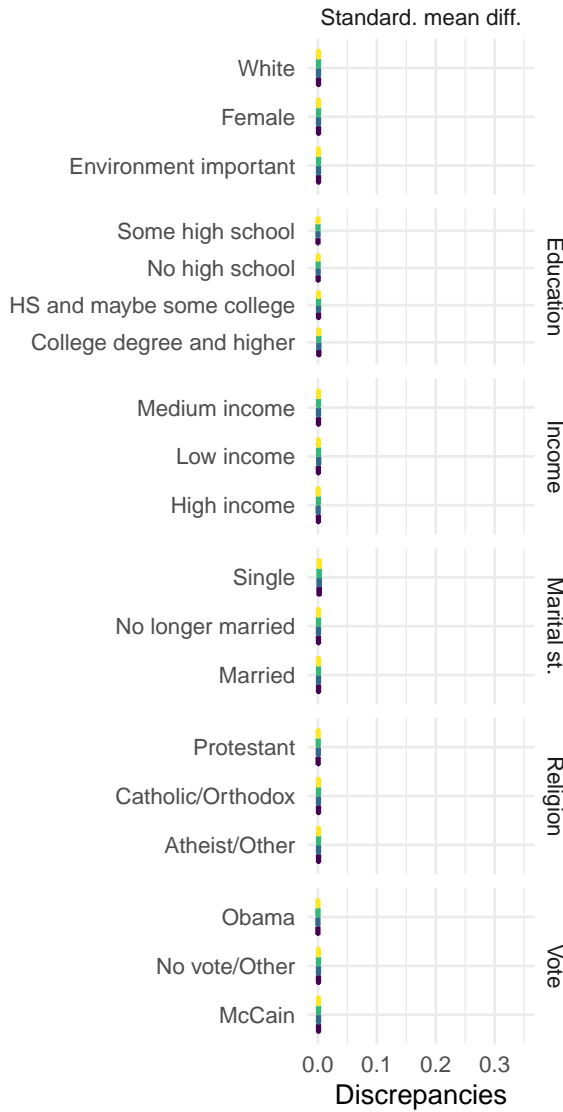
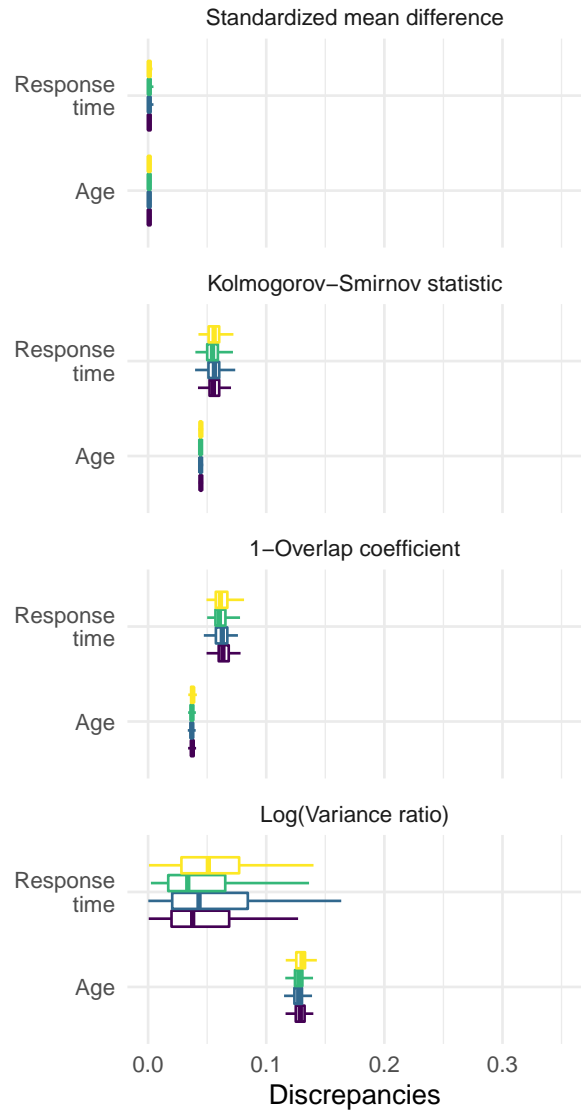


Figure A.2: Density of the discrepancy statistics comparing observations with imputed and observed values of the variable **income** across the 100 imputed datasets. Outliers are removed and some upper whiskers are clipped to increase readability.

A: Binary variables



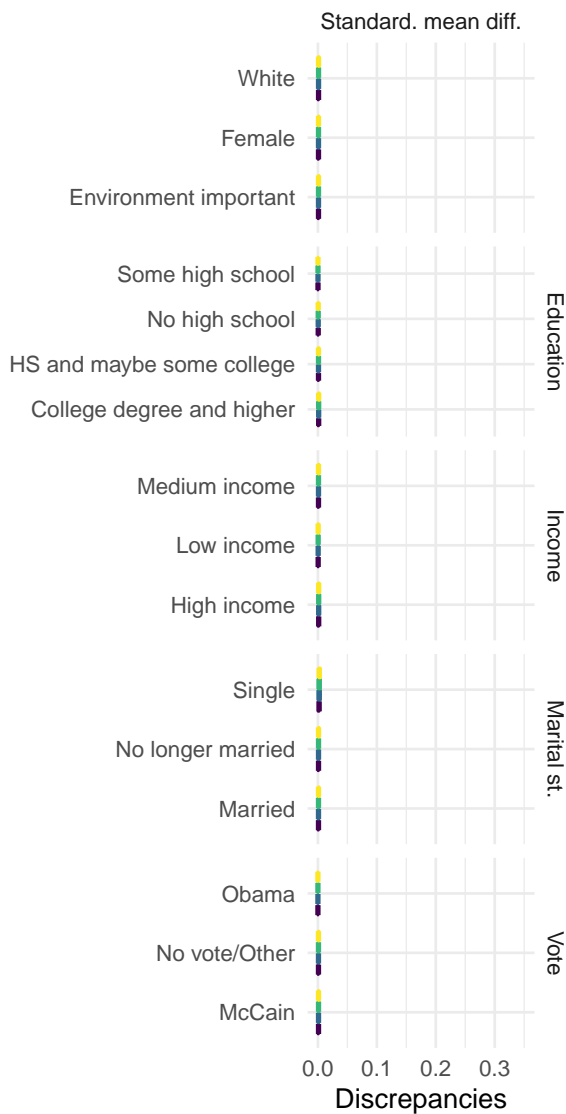
B: Continuous variables



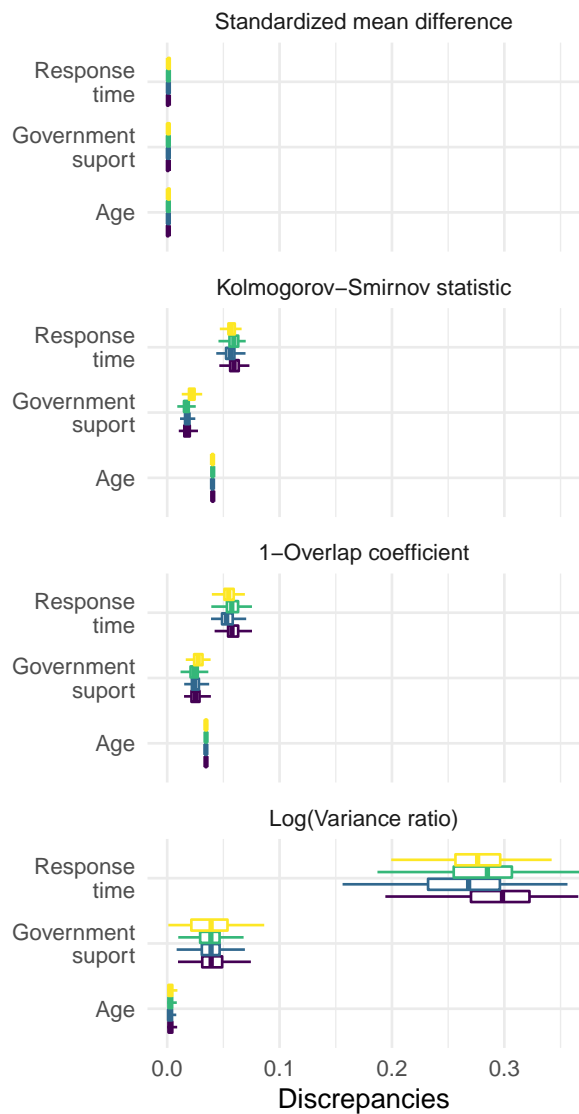
Imputation model: Hotdeck PMM RForest Norm

Figure A.3: Density of the discrepancy statistics comparing observations with imputed and observed values of the variable **government support** across the 100 imputed datasets. Outliers are removed and some upper whiskers are clipped to increase readability.

A: Binary variables



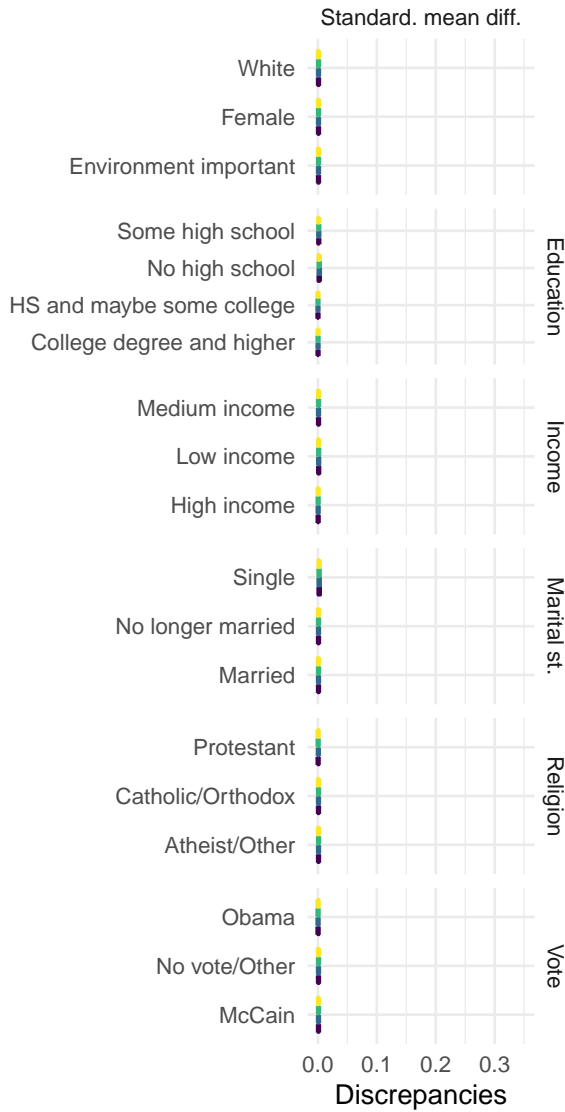
B: Continuous variables



Imputation model: Hotdeck PMM RForest Norm

Figure A.4: Density of the discrepancy statistics comparing observations with imputed and observed values of the variable **religion** across the 100 imputed datasets. Outliers are removed and some upper whiskers are clipped to increase readability.

A: Binary variables



B: Continuous variables

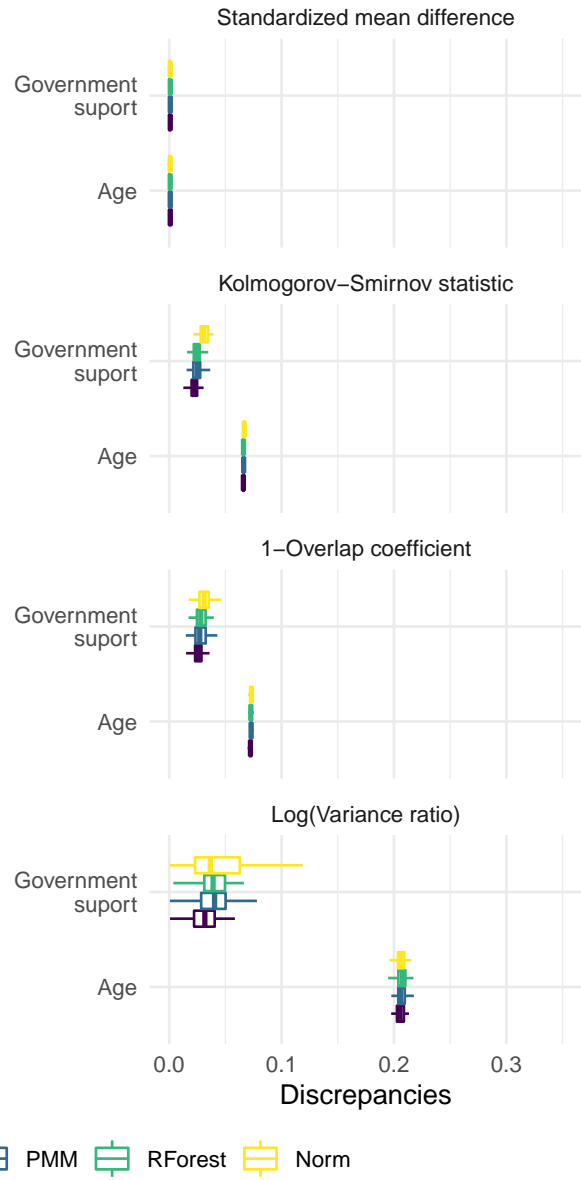
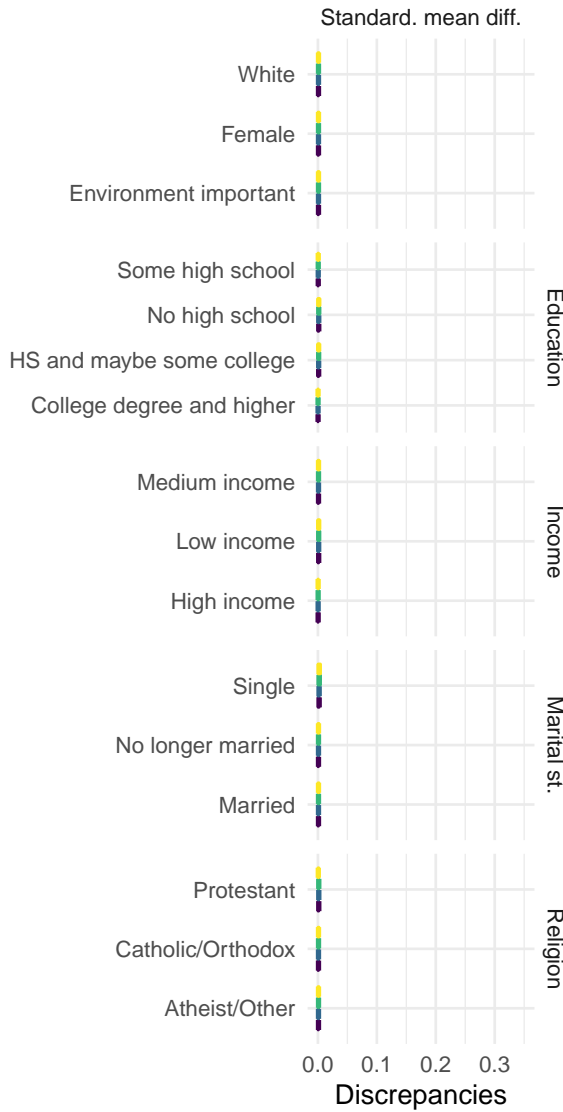
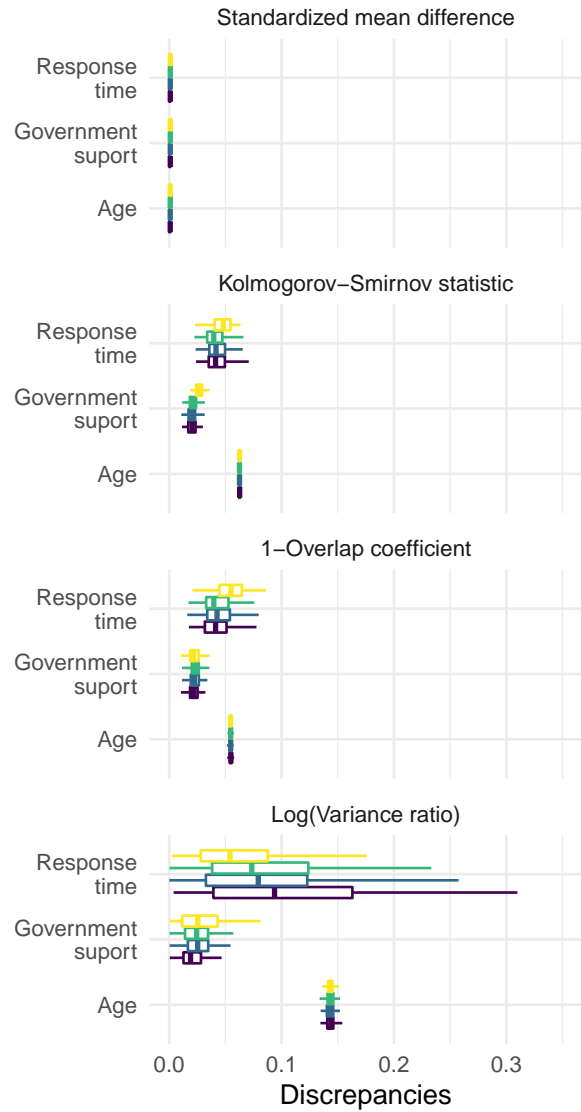


Figure A.5: Density of the discrepancy statistics comparing observations with imputed and observed values of the variable **response time** across the 100 imputed datasets. Outliers are removed and some upper whiskers are clipped to increase readability.

A: Binary variables



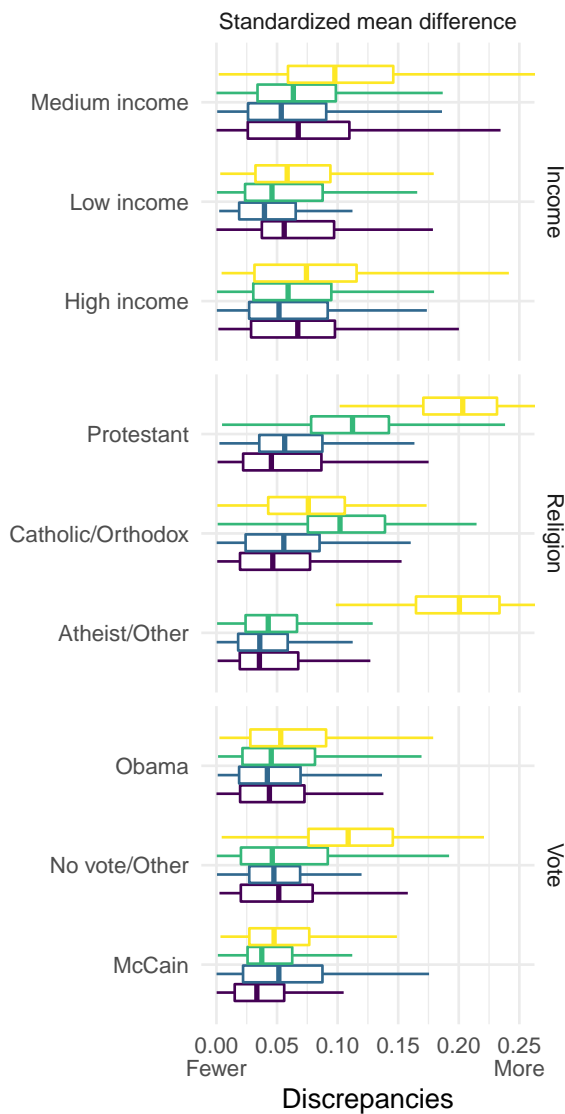
B: Continuous variables



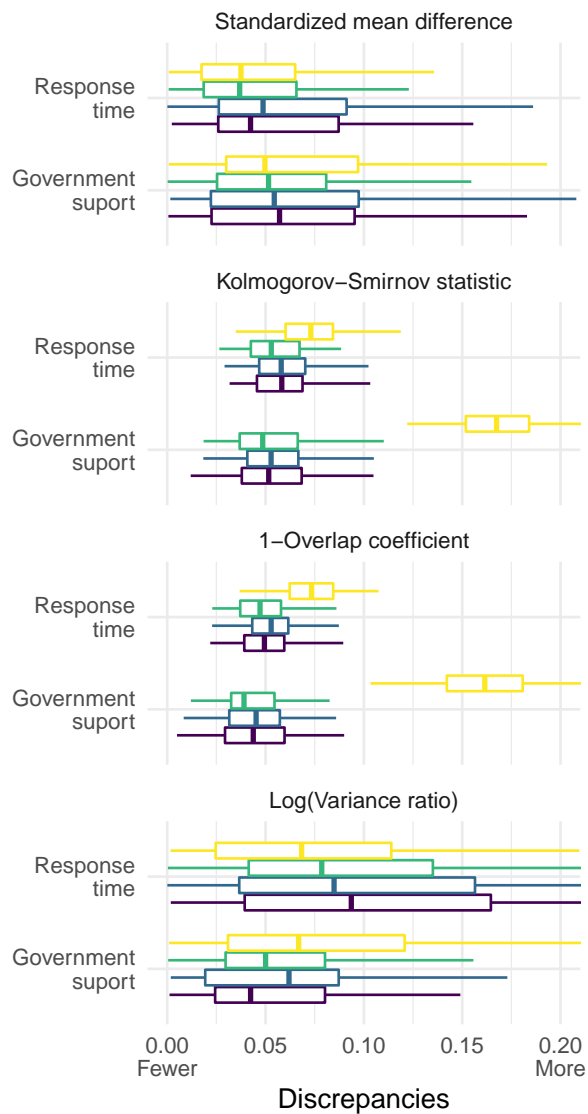
Imputation model: Hotdeck PMM RForest Norm

Figure A.6: Density of the discrepancy statistics comparing observations with imputed and observed values of the variable **vote** across the 100 imputed datasets. Outliers are removed and some upper whiskers are clipped to increase readability.

A: Binary variables



B: Continuous variables



Imputation model: Hotdeck PMM RForest Norm

Figure A.7: Additional balancing of interactions. Box plots of the discrepancy statistics comparing the weighted density of observed and imputed values across the 100 imputed datasets. Outliers are removed and some upper whiskers are clipped to increase readability.