

Testing different imputation methods on PUMS (MCAR) - RandomForest

```
# load dataset: df
load('../..'/Datasets/ordinalPUMS.Rdata')

# take 10,000 samples: df
set.seed(0)
n = 10000
sample <- sample(nrow(df), size = 10000)
df <- df[sample,]

# create MCAR scenario with 45% chance of missing: df_observed
missing_prob = 0.45
df_observed <- df
missing_col = c(1,3,7,9,10,11)
for (col in missing_col) {
  missing_ind <- rbernoulli(n,p = missing_prob)
  df_observed[missing_ind, col] <- NA
}
```

missForest

```
df.imp <- missForest(df_observed, verbose = FALSE)
d1 <- df.imp$xim
df.imp <- missForest(df_observed, verbose = FALSE)
d2 <- df.imp$xim
df.imp <- missForest(df_observed, verbose = FALSE)
d3 <- df.imp$xim
df.imp <- missForest(df_observed, verbose = FALSE)
d4 <- df.imp$xim
df.imp <- missForest(df_observed, verbose = FALSE)
d5 <- df.imp$xim
imputed_sets = rbind(d1, d2, d3, d4, d5)
```

Diagnostics

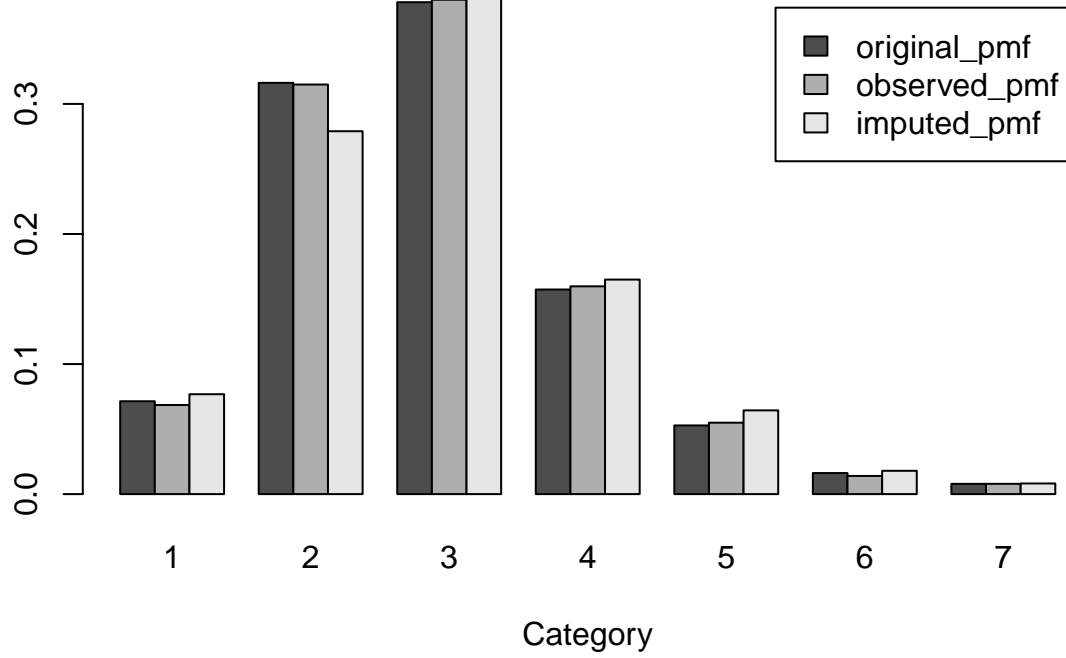
Assess bivariate joint distribution

Assess trivariate joint distribution

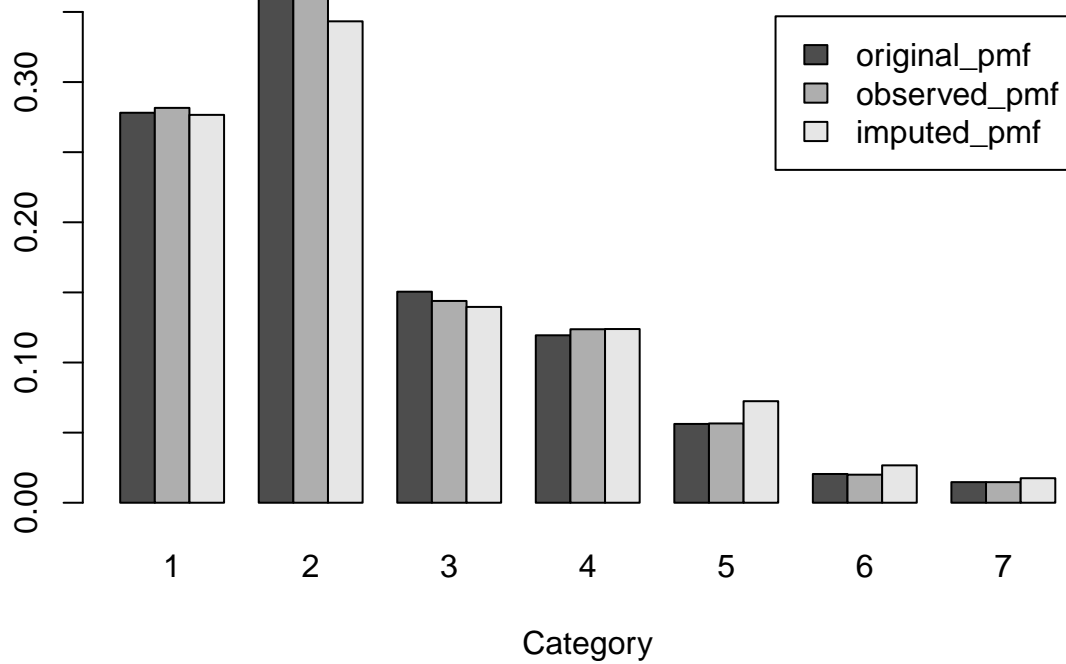
```
## [1] "rmse"
```

```
## [1] 0.2849009
```

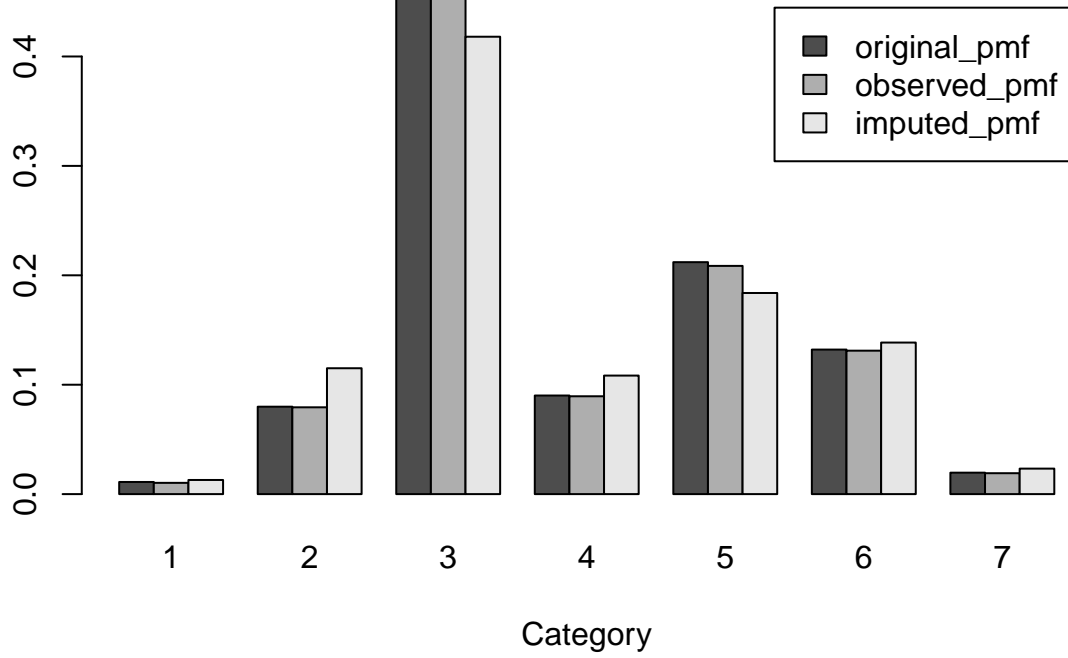
MICE: VEH



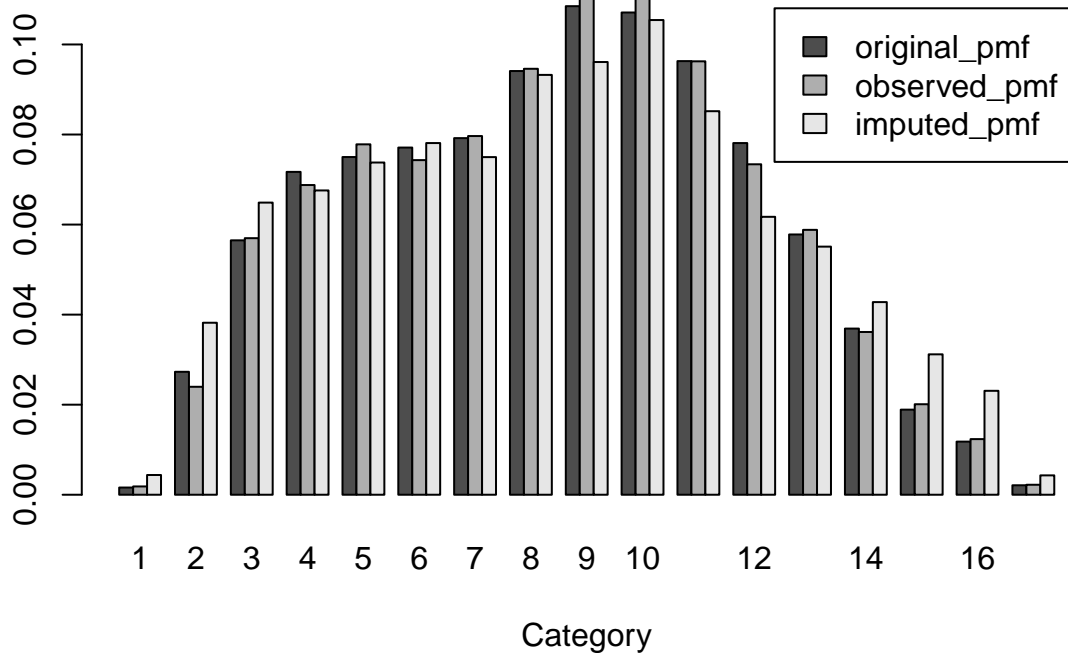
MICE: NP



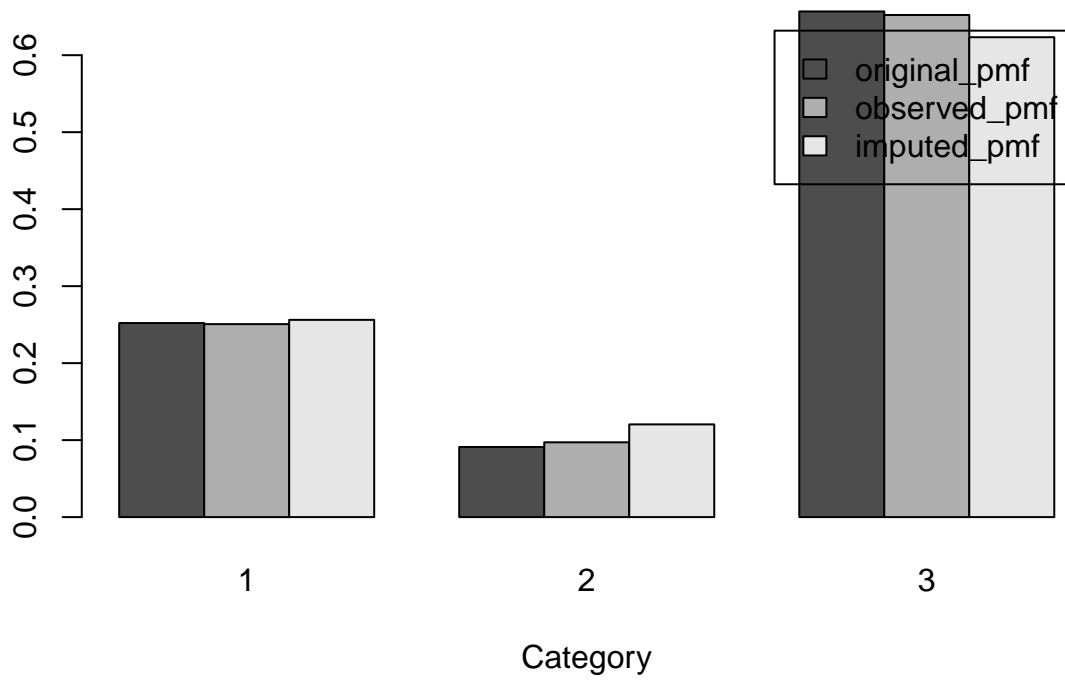
MICE: SCHL



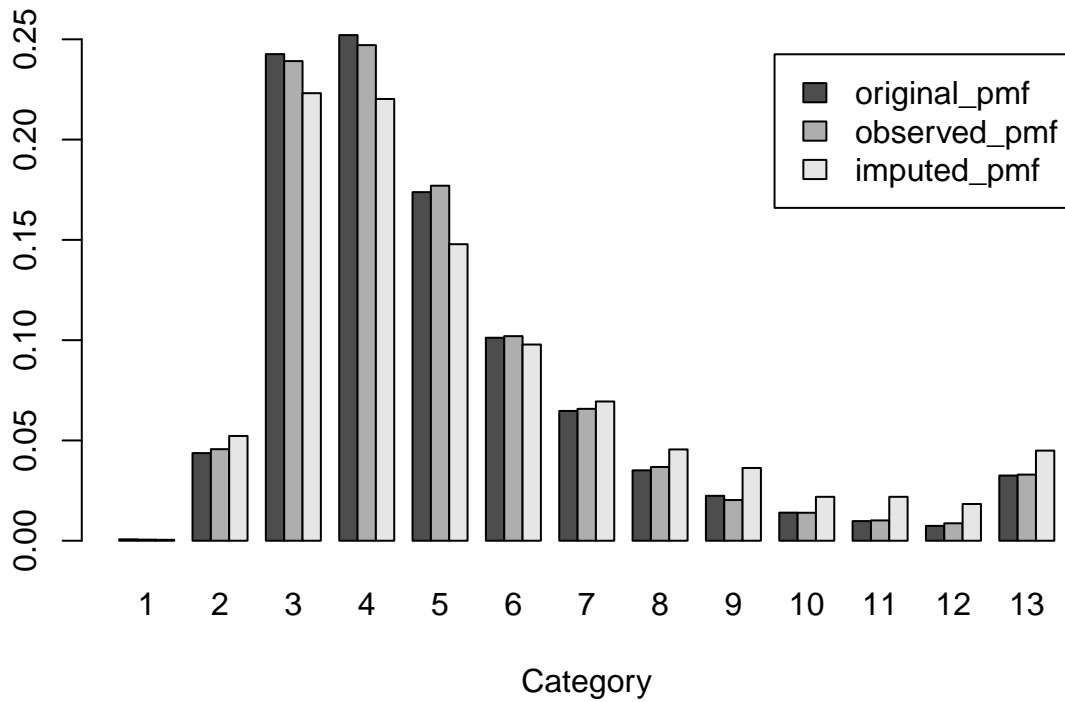
MICE: AGEP



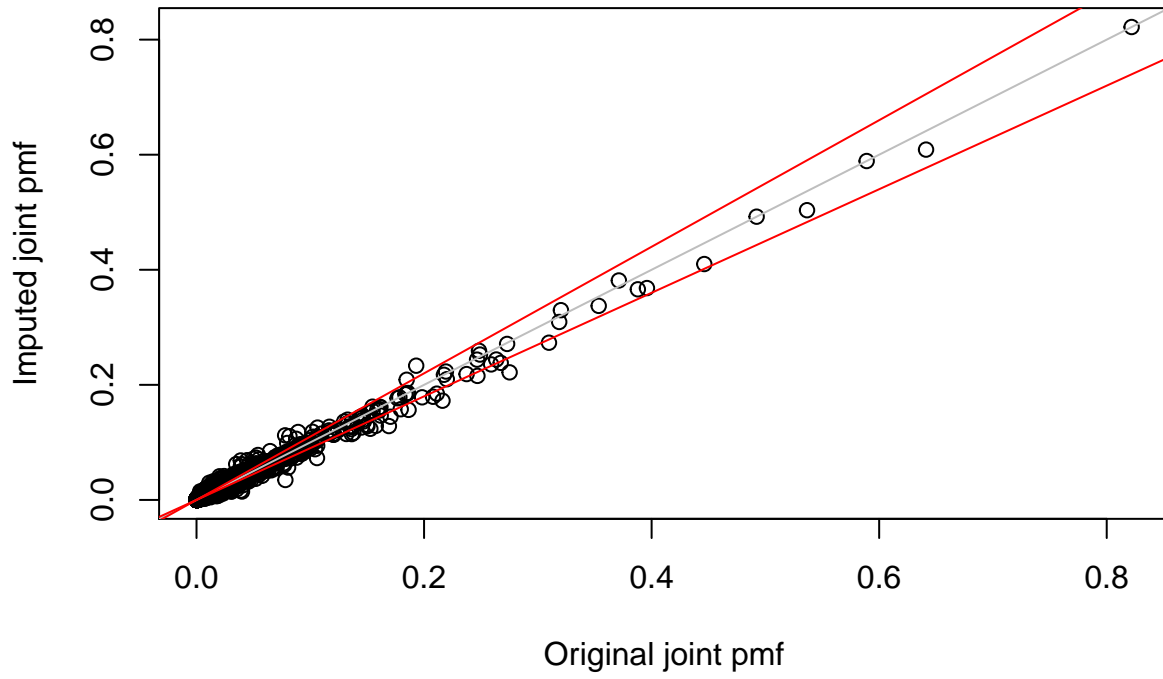
MICE: WKL



MICE: PINCP



Bivariate pmf



Trivariate pmf

