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 scneario with 30% chance of missing: df_observed
missing_prob = 0.3
df_observed <- df
missing_col = colnames(df)[c(1,3,5,7,9,11)]
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$ximp
df.imp <- missForest(df_observed, verbose = FALSE)
d2 <-  df.imp$ximp
df.imp <- missForest(df_observed, verbose = FALSE)
d3 <-  df.imp$ximp
df.imp <- missForest(df_observed, verbose = FALSE)
d4 <-  df.imp$ximp
df.imp <- missForest(df_observed, verbose = FALSE)
d5 <-  df.imp$ximp
imputed_sets = rbind(d1, d2, d3, d4, d5)
```

Diagnostics

Assess bivariate joint distribution

Assess trivariate joint distribution

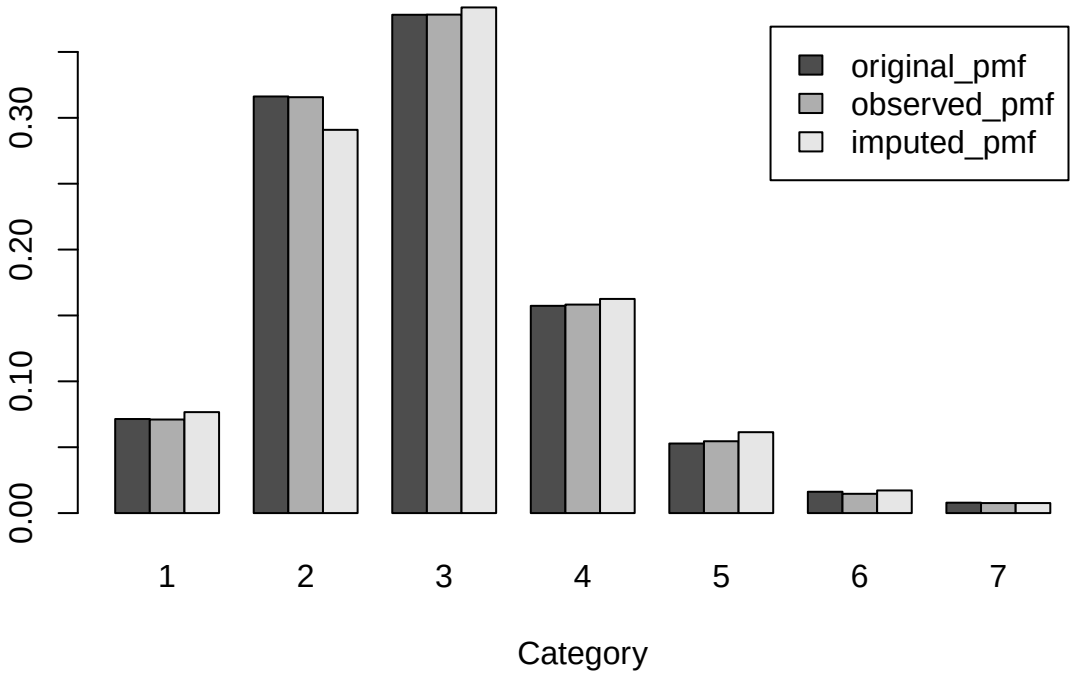
```
## [1] "rmse"
```

```
## [1] 0.2338423
```

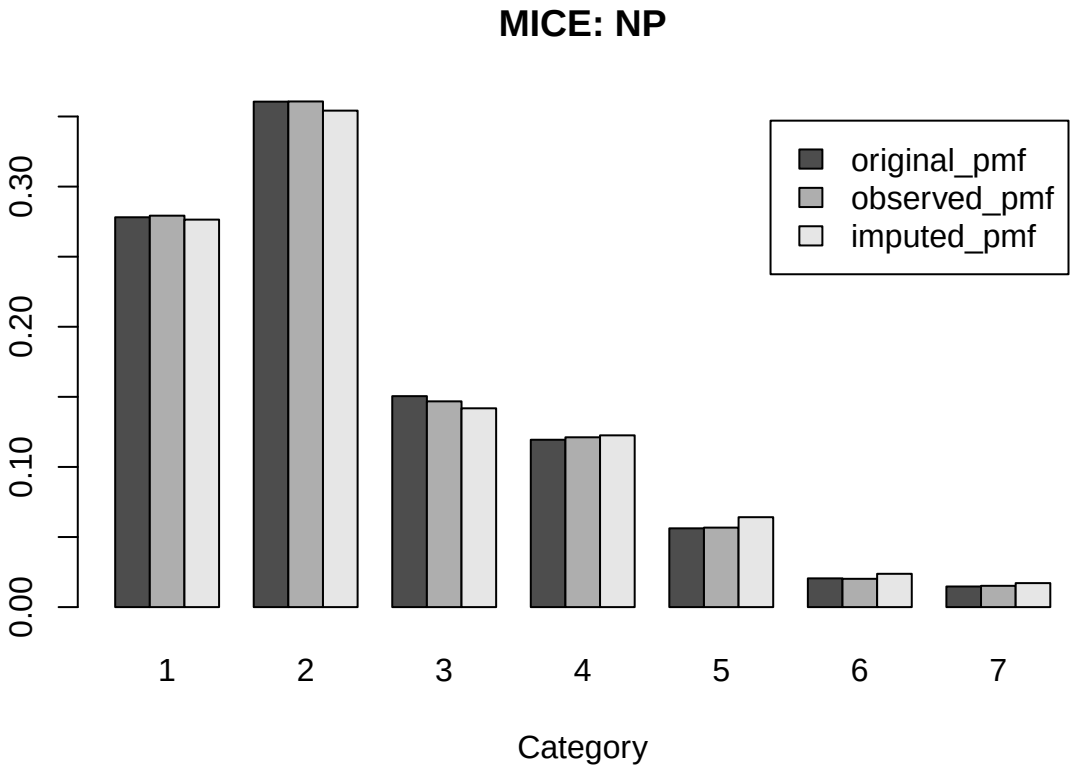
```
# accuracy
acc = sum(numeric_df[missing_matrix] == numeric_impute[missing_matrix])/sum(missing_matrix)
acc
```

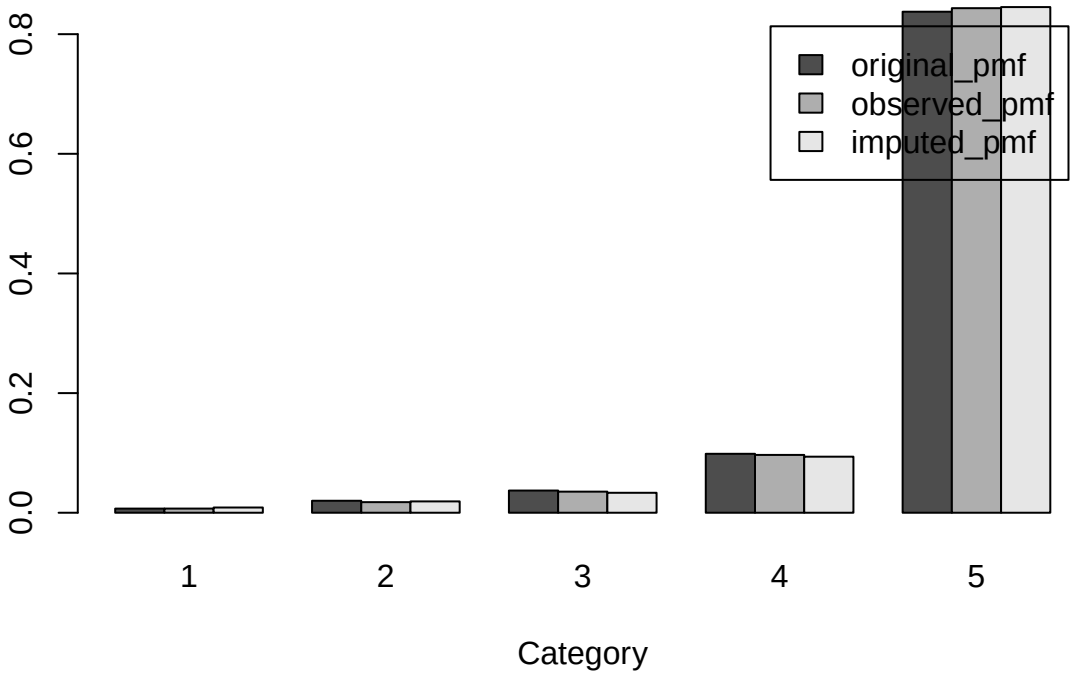
```
## [1] 0.4018996
```

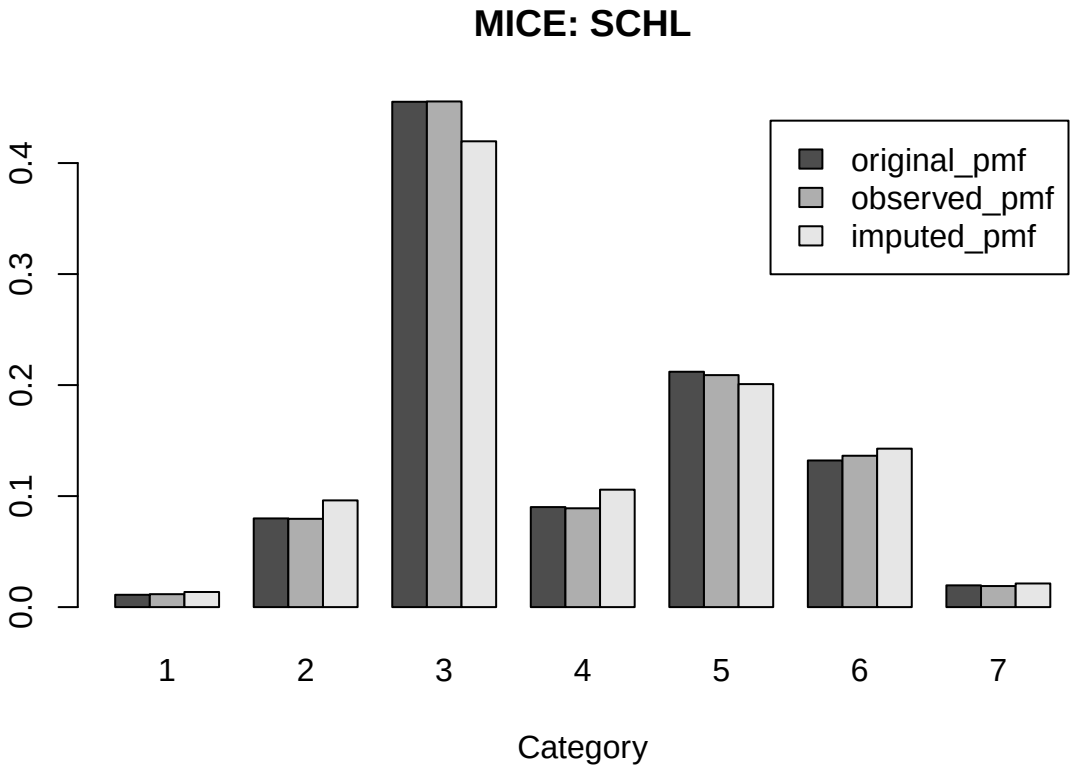
```
# Actual vs Imputed values
col = 1
```

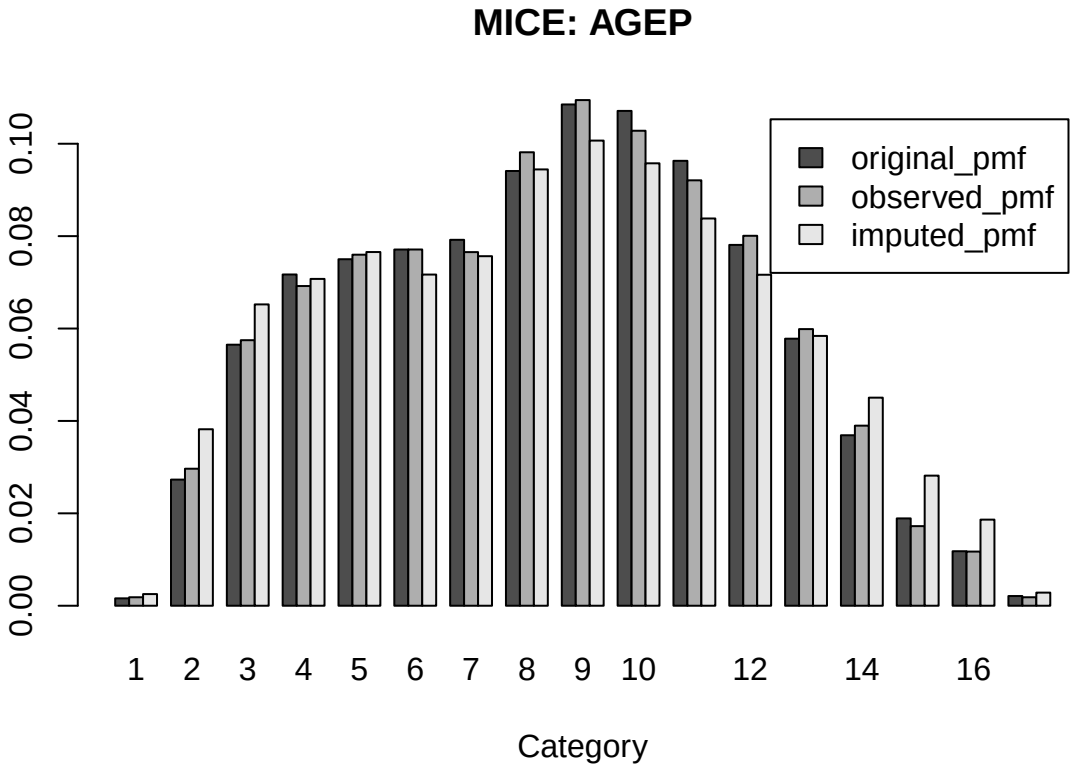


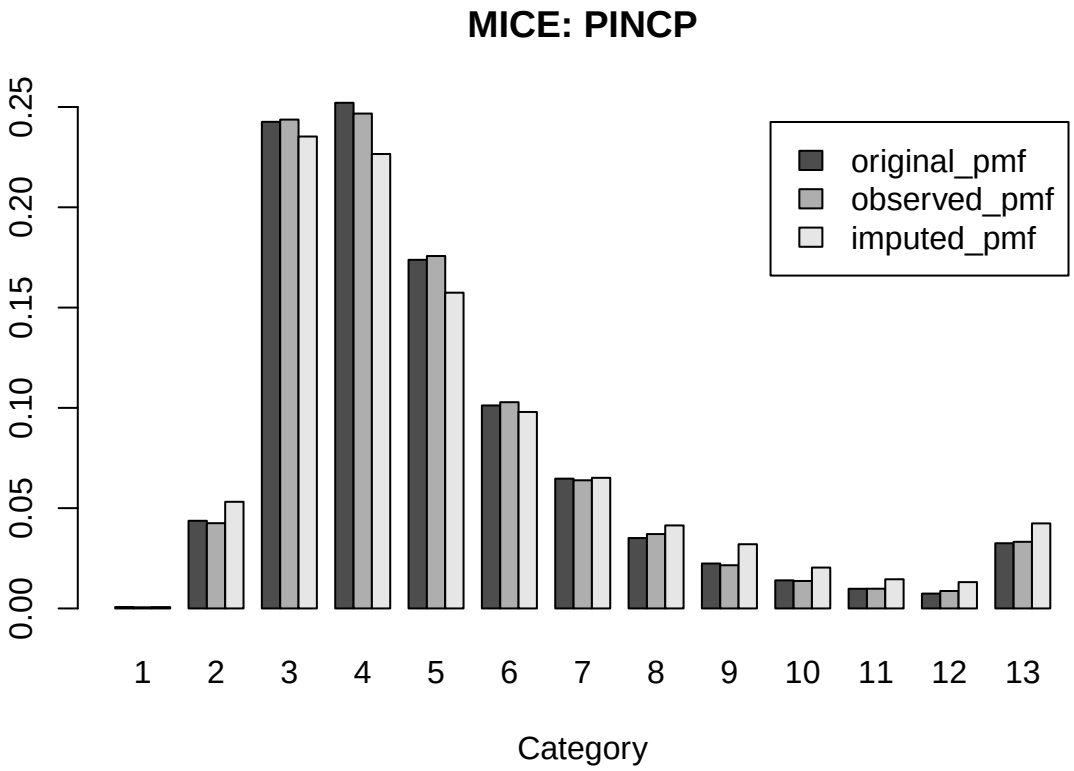

```
missing_indicator = missing_matrix[,col]
plot(as.integer(df[missing_indicator, col]), as.integer(d1[missing_indicator, col]), xlab = 'actual valu
```

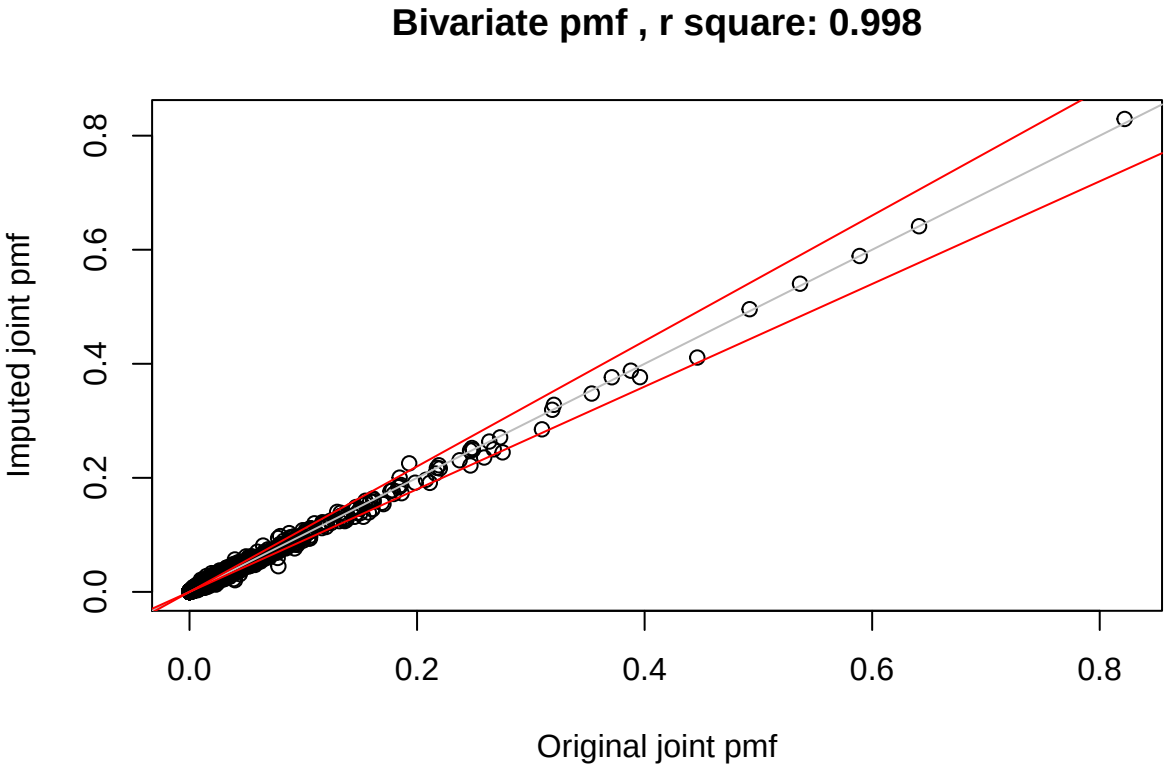

MICE: NP
original_pmf
observed_pmf
imputed_pmf
0.00
0.10
0.20
0.30
1
2
3
4
5
6
7
Category

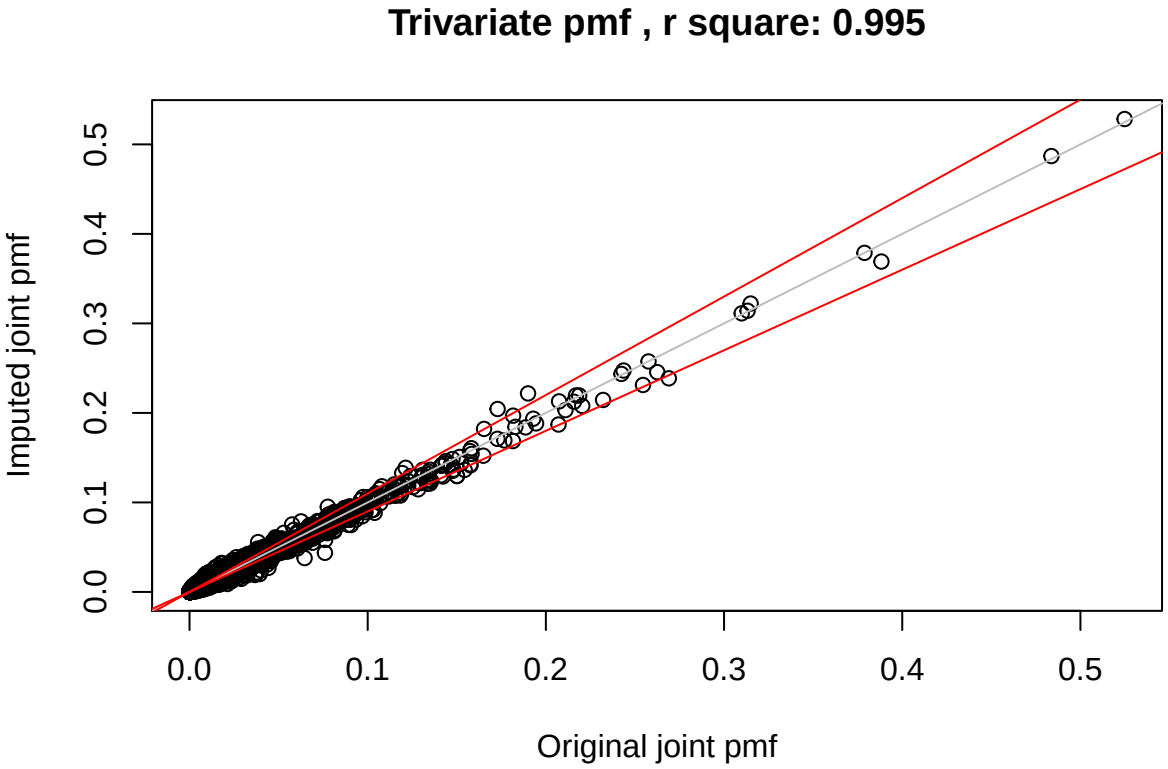

MICE: ENG
original_pmf
observed_pmf
imputed_pmf
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
Category

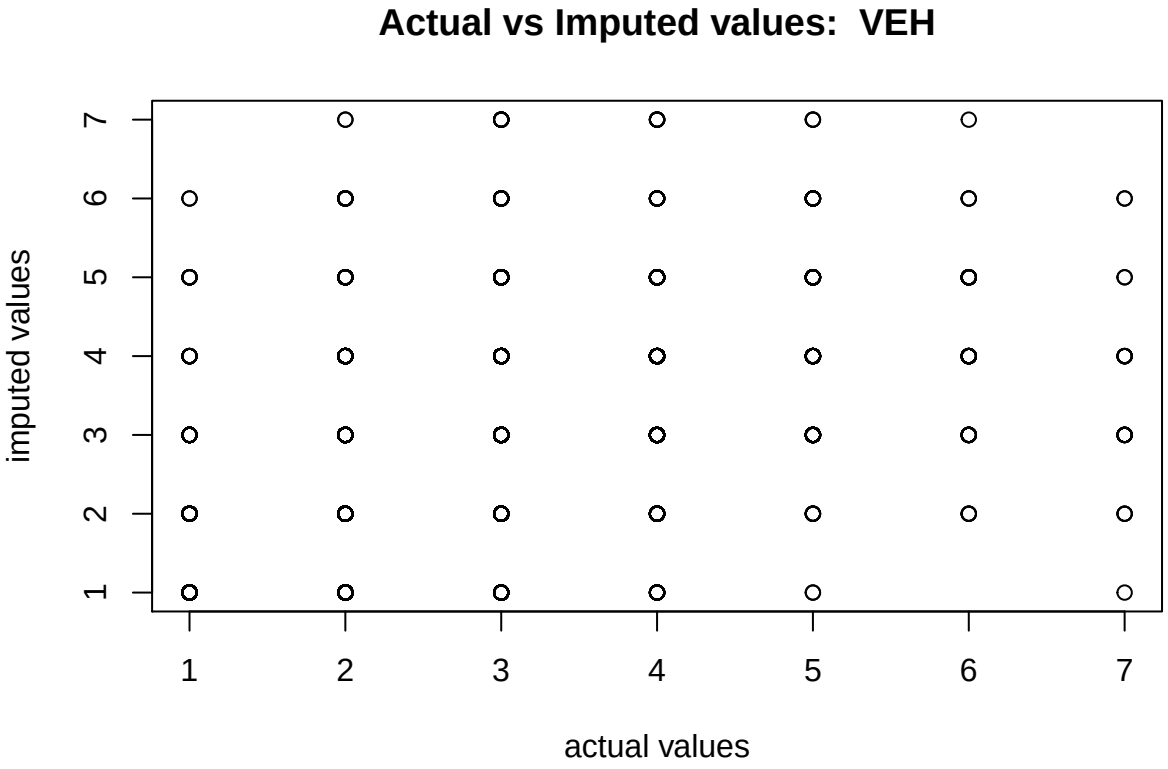

MICE: SCHL
original_pmf
observed_pmf
imputed_pmf
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7
Category


MICE: AGEP
original_pmf
observed_pmf
imputed_pmf
0.00
0.02
0.04
0.06
0.08
0.10
1
2
3
4
5
6
7
8
9
10
12
14
16
Category


MICE: PINCP
original_pmf
observed_pmf
imputed_pmf
0.00
0.05
0.10
0.15
0.20
0.25
1
2
3
4
5
6
7
8
9
10
11
12
13
Category


Bivariate pmf , r square: 0.998
Imputed joint pmf
0.0
0.2
0.4
0.6
0.8
Original joint pmf


Trivariate pmf , r square: 0.995
Imputed joint pmf
0.0
0.1
0.2
0.3
0.4
0.5
Original joint pmf


Actual vs Imputed values: VEH
imputed values
1
2
3
4
5
6
7
actual values