

# MAR 45% missing - MICE

---

```
# sample MCAR dataset from PUMS
source("../utils/sampleMAR45.R")
n = 10000
missing_col = c(1,3,7,9,10,11)
set.seed(3)

output_list <- sampleMAR45(n)
df <- output_list[['df']]
df_observed <- output_list[['df_observed']]

apply(is.na(df_observed), MARGIN = 2, mean)
```

```
##      VEH      MV      NP      RMSP      ENG      MARHT      SCHL RACNUM      AGEP      WKL      PINCP
## 0.4456 0.0000 0.3998 0.0000 0.0000 0.0000 0.4842 0.0000 0.4670 0.4478 0.4384
```

## MICE

Create 5 imputed dataset

```
library(mice)

##
## Attaching package: 'mice'
## The following objects are masked from 'package:base':
##
##      cbind, rbind

imputed_df <- mice(df_observed,m=5,print=F)
```

```
## Warning: Number of logged events: 50
```

Extract the 5 imputed dataset

```
d1 <- complete(imputed_df, 1)
d2 <- complete(imputed_df, 2)
d3 <- complete(imputed_df, 3)
d4 <- complete(imputed_df, 4)
d5 <- complete(imputed_df, 5)
imputed_sets = rbind(d1, d2, d3, d4, d5)
```

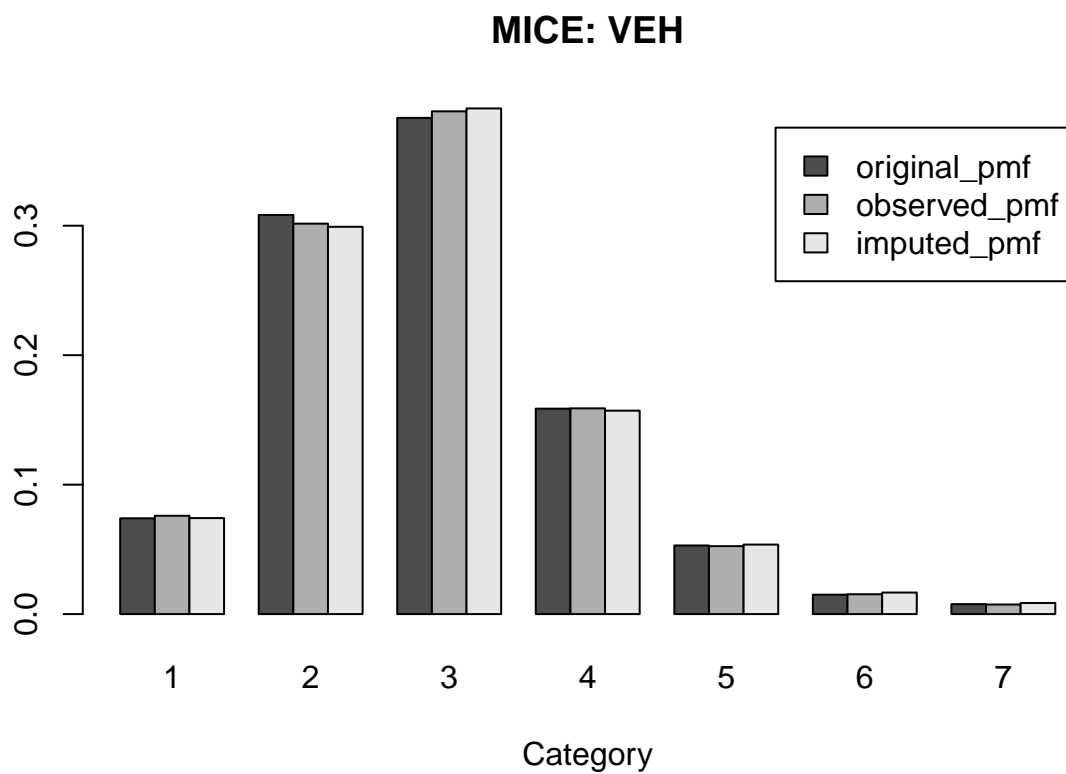
Diagnostics

Assess bivariate joint distribution

Assess trivariate joint distribution

```
# calculate rmse
numeric_df = sapply(df, as.numeric)
normalized_df = t(t(numeric_df-1)/(apply(numeric_df, MARGIN = 2, FUN = max)-1))
numeric_impute = sapply(d1, as.numeric)
normalized_impute = t(t(numeric_impute-1)/(apply(numeric_df, MARGIN = 2, FUN = max)-1))

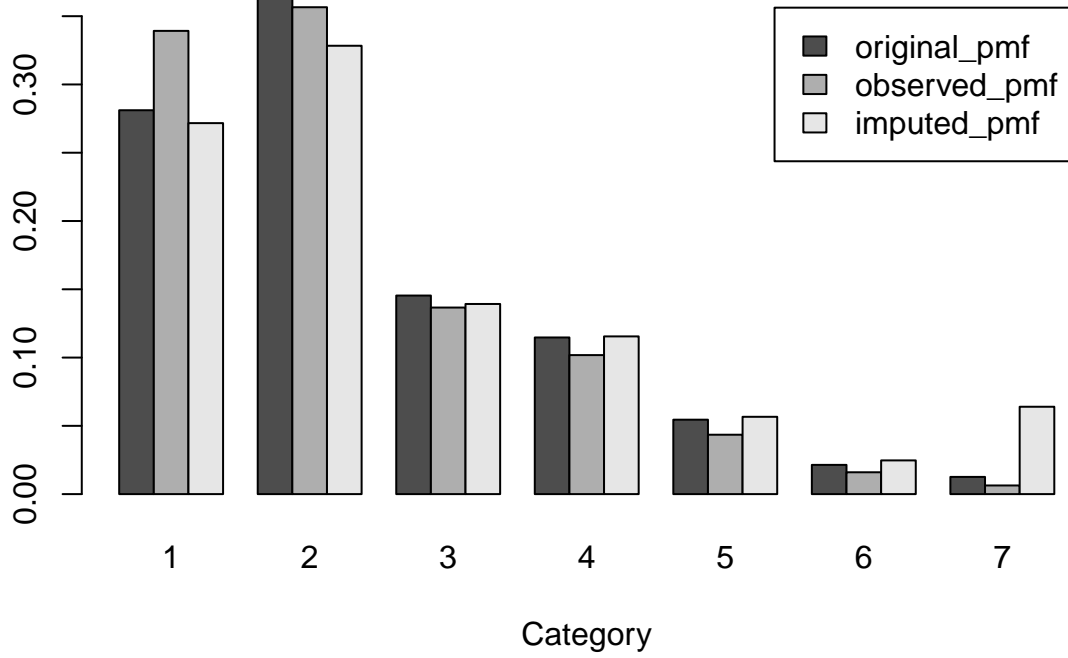
missing_matrix = is.na(df_observed)
```



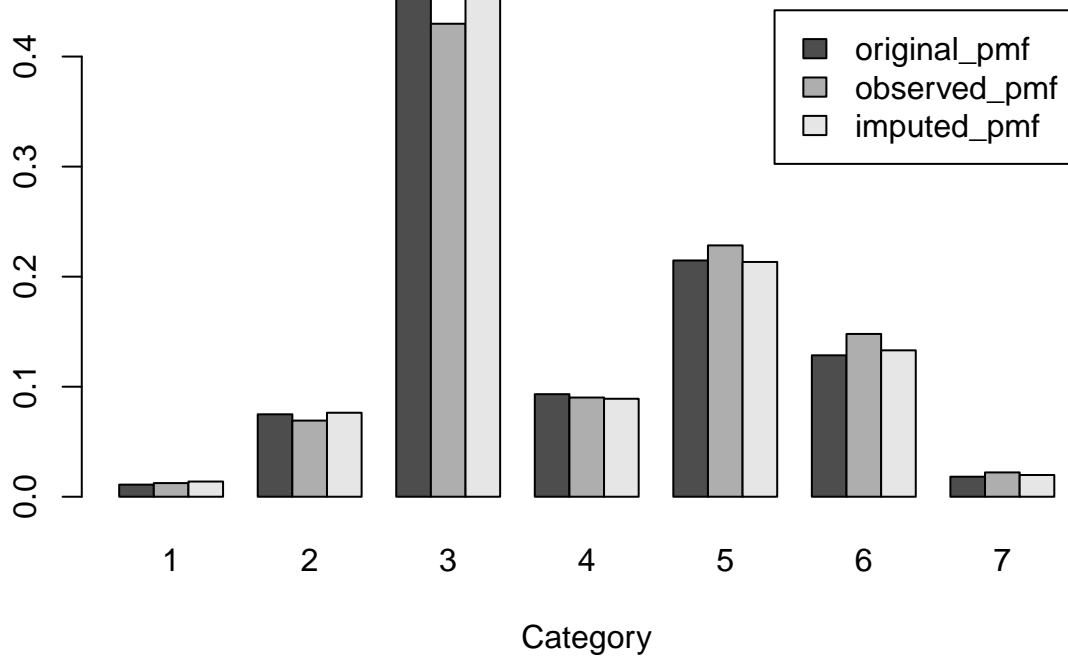
```
rmse = sqrt(sum((normalized_df[missing_matrix] - normalized_impute[missing_matrix])^2)/sum(missing_matrix == 1))
rmse
```

```
## [1] 0.3439604
```

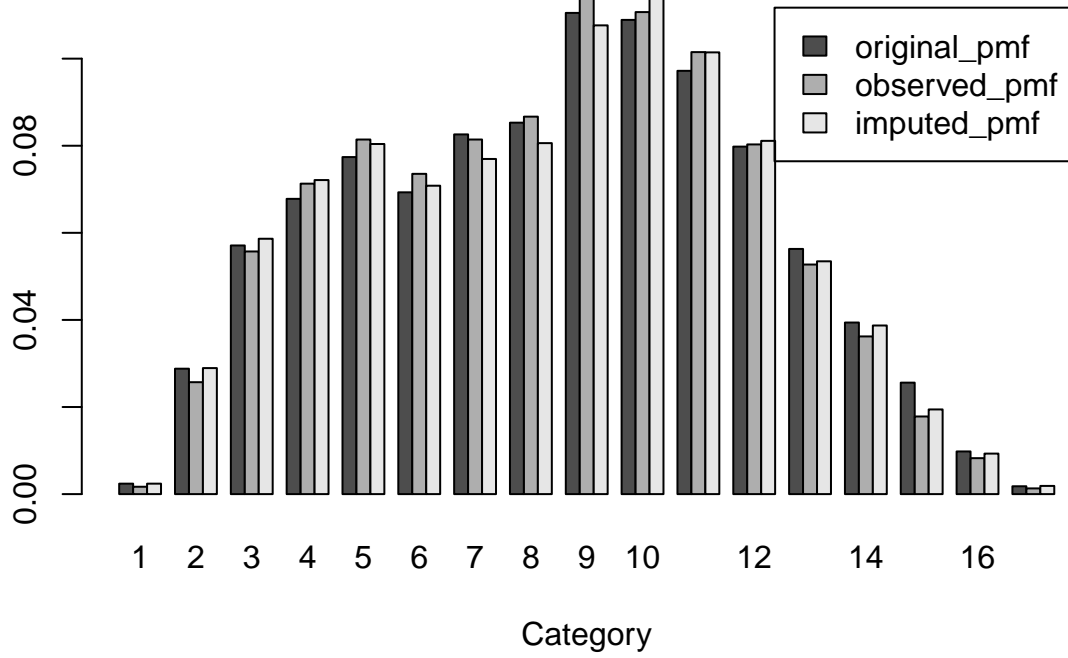
### MICE: NP



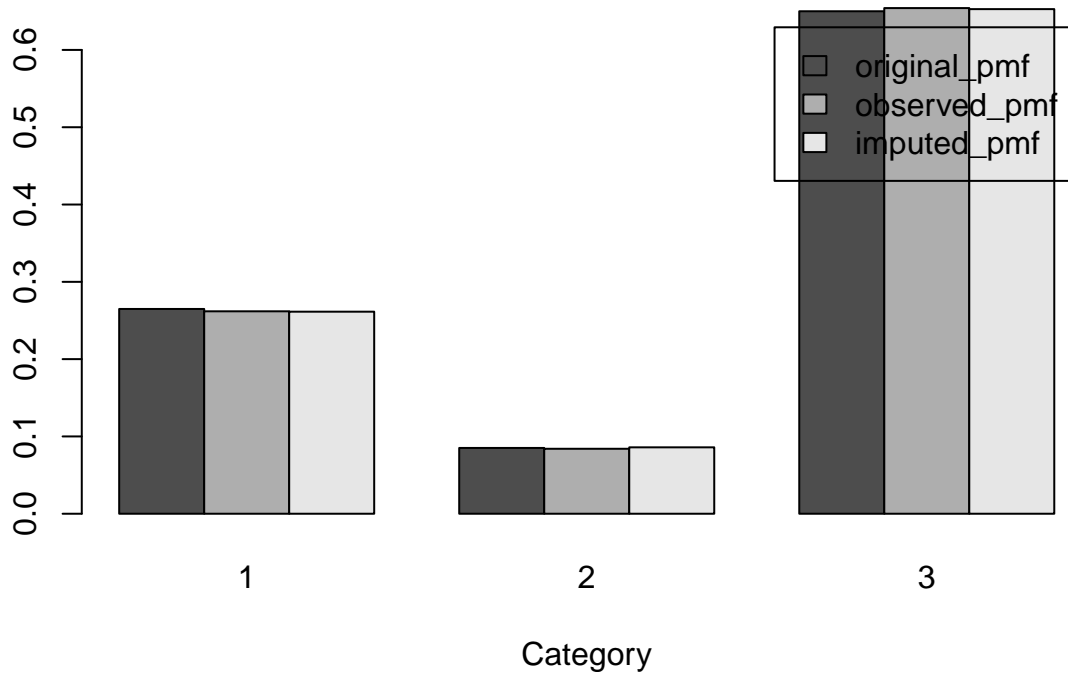
### MICE: SCHL



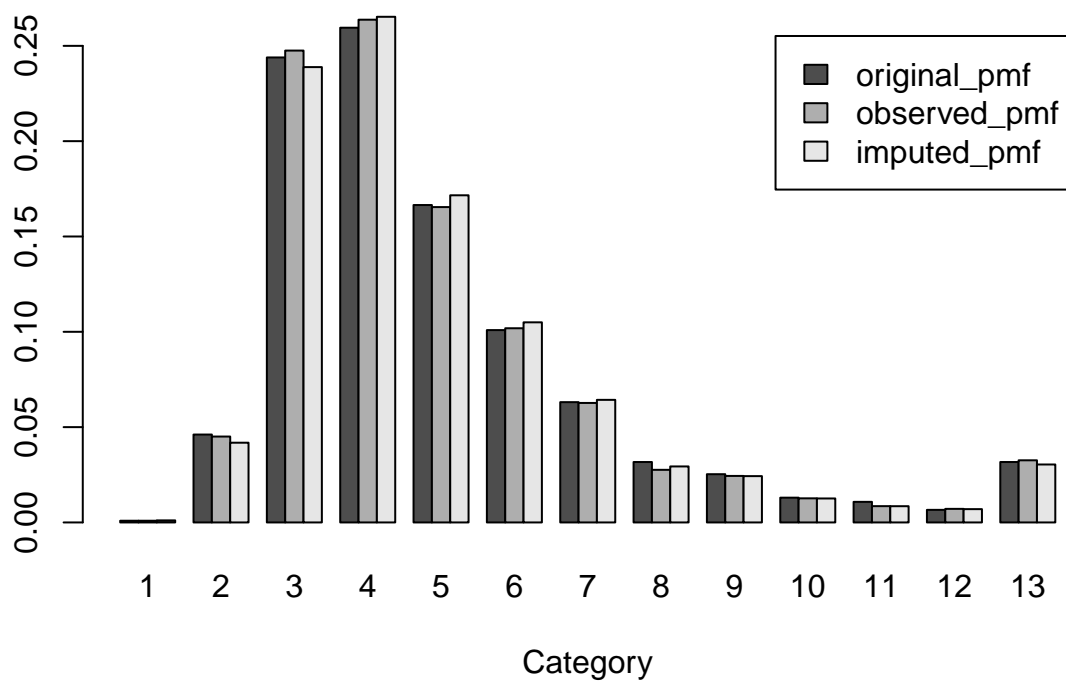
### MICE: AGEP



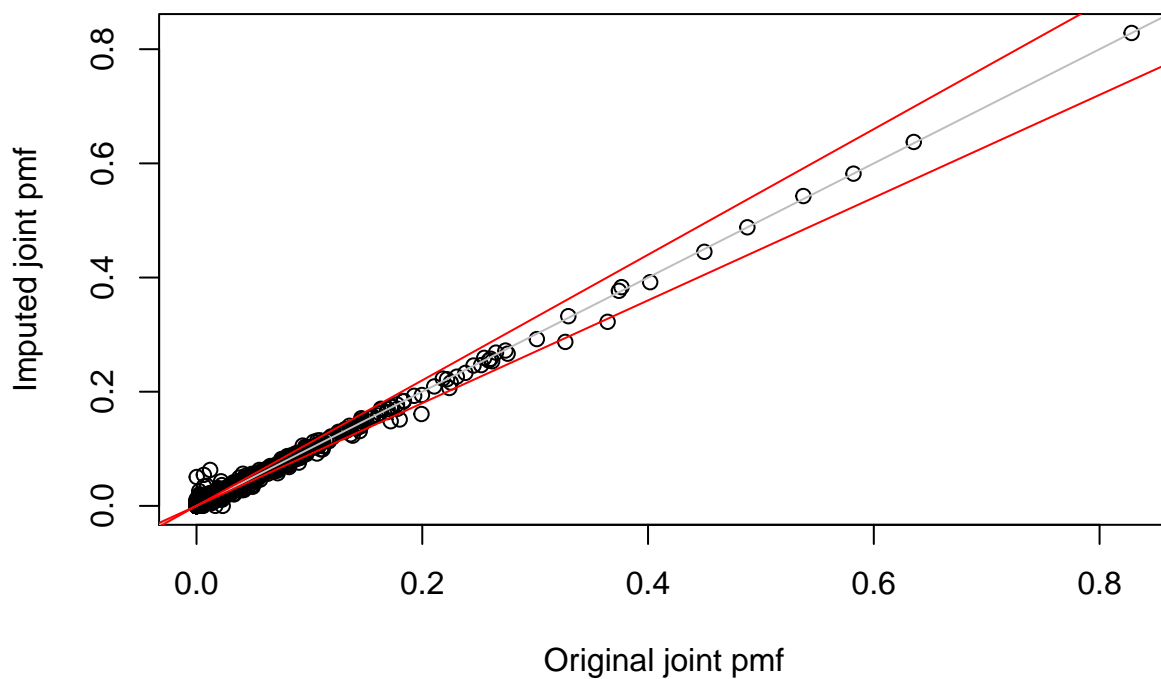
### MICE: WKL



### MICE: PINCP



### Bivariate pmf



Trivariate pmf

