

# MAR 45% missing - Generative Adversarial Imputation Nets (GAIN)

```
# 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      AGE      WKL      PINCP
## 0.4456 0.0000 0.3998 0.0000 0.0000 0.0000 0.4842 0.0000 0.4670 0.4478 0.4384
```

## Generative Adversarial Imputation Nets (GAIN)

reference: <https://arxiv.org/abs/1806.02920>

```
# Load imputed dataset
d1 = read.csv("../GAIN/imputed_dataset/MAR_45percent_1.csv", header = FALSE, sep = ',')
d2 = read.csv("../GAIN/imputed_dataset/MAR_45percent_2.csv", header = FALSE, sep = ',')
d3 = read.csv("../GAIN/imputed_dataset/MAR_45percent_3.csv", header = FALSE, sep = ',')
d4 = read.csv("../GAIN/imputed_dataset/MAR_45percent_4.csv", header = FALSE, sep = ',')
d5 = read.csv("../GAIN/imputed_dataset/MAR_45percent_5.csv", header = FALSE, sep = ',')
colnames(d1) = colnames(df)
colnames(d2) = colnames(df)
colnames(d3) = colnames(df)
colnames(d4) = colnames(df)
colnames(d5) = colnames(df)
imputed_df = rbind(d1, d2, d3, d4, d5)
```

Diagnostics

Assess bivariate joint distribution

Assess bivariate joint distribution

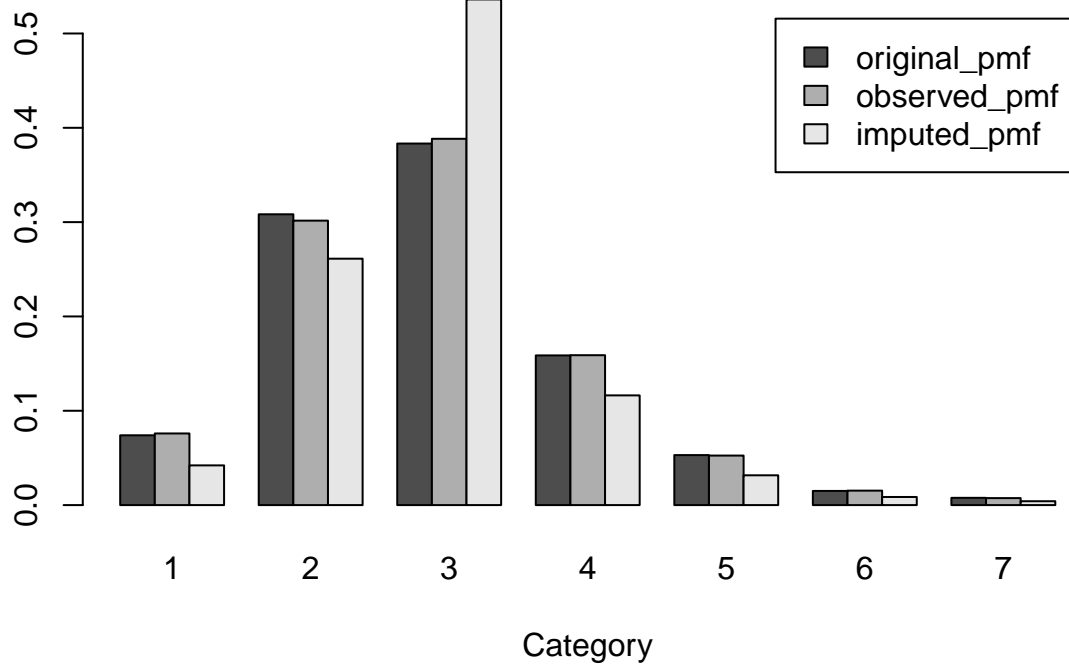
```
# calculate rmse
numeric_df = sapply(df, as.numeric)
normalized_df = t(t(numeric_df-1)/(apply(numeric_df, MARGIN = 2, FUN = max)-1))
numeric_impute = sapply(d1, as.numeric)
normalized_impute = t(t(numeric_impute-1)/(apply(numeric_df, MARGIN = 2, FUN = max)-1))

missing_matrix = is.na(df_observed)

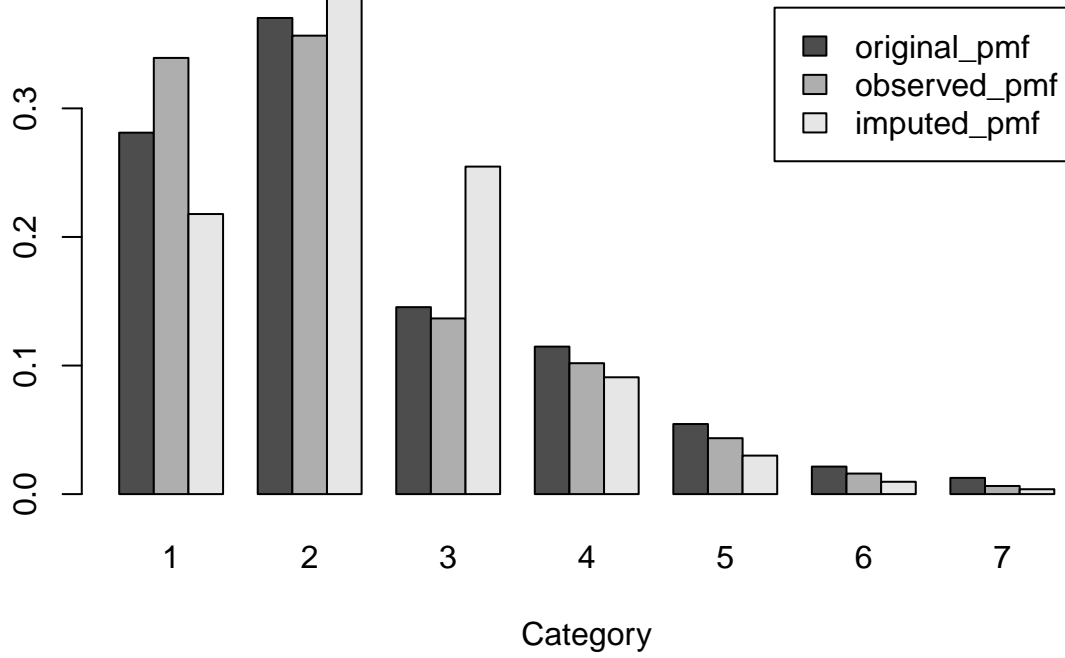
rmse = sqrt(sum((normalized_df[missing_matrix] - normalized_impute[missing_matrix])^2)/sum(missing_matrix))
rmse

## [1] 0.2926325
```

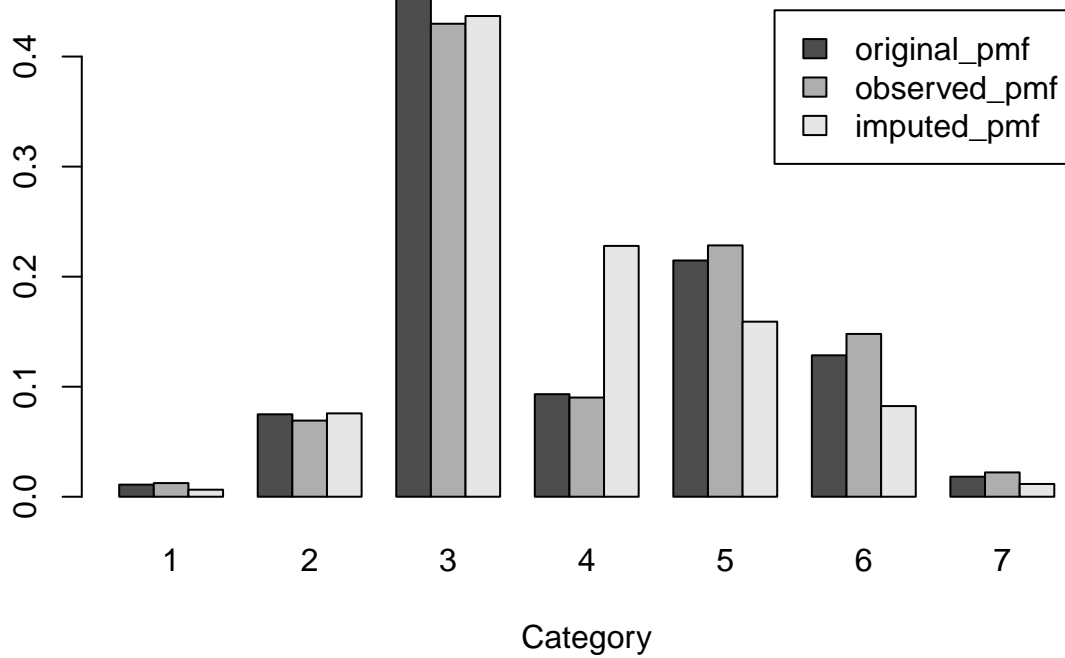
## MICE: VEH



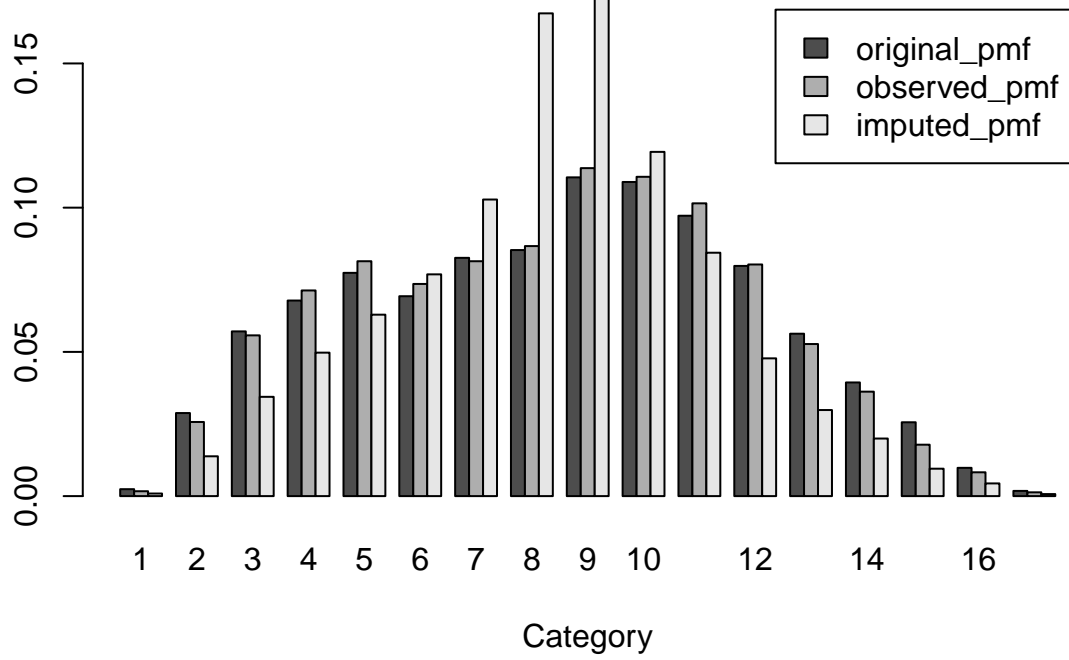
### MICE: NP



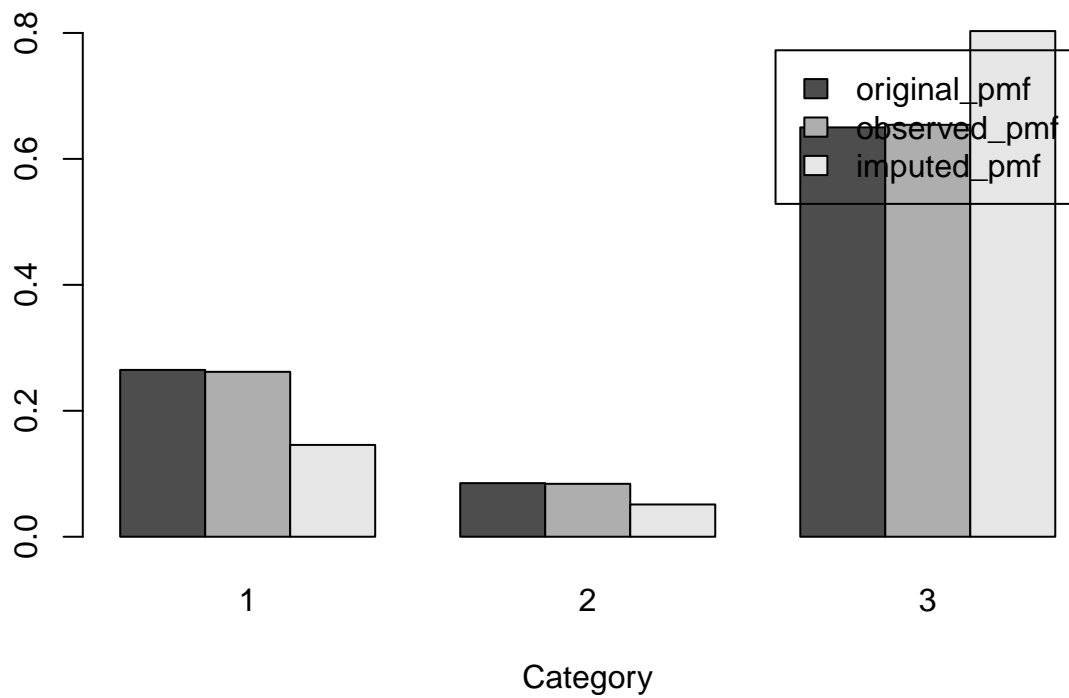
### MICE: SCHL



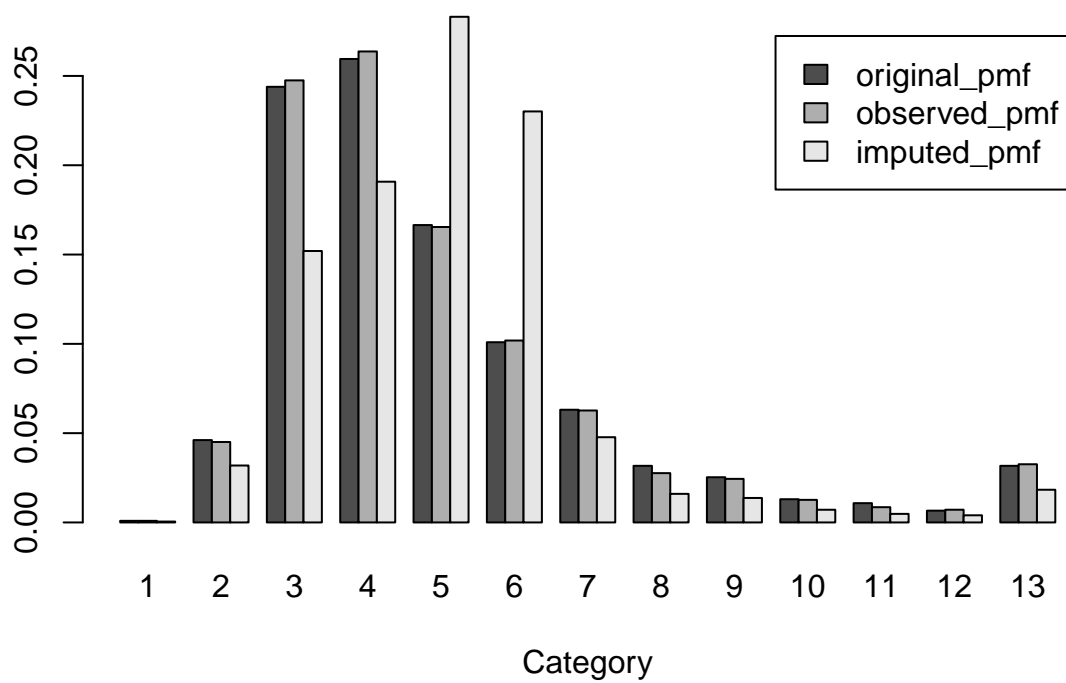
### MICE: AGEP



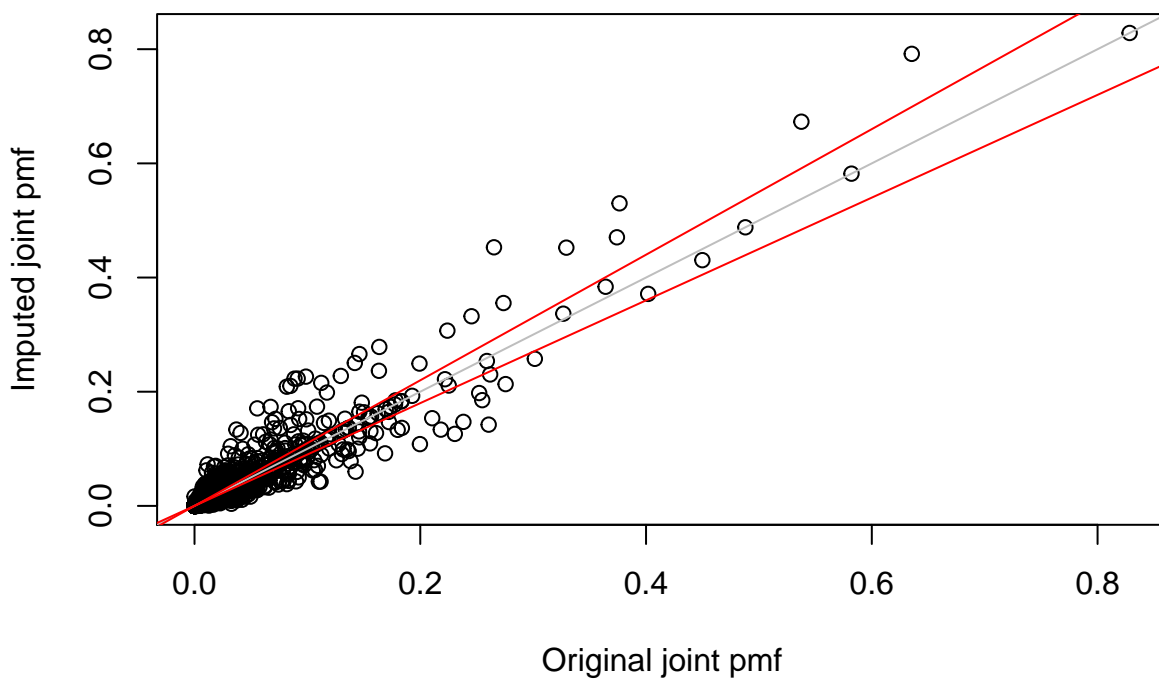
### MICE: WKL



### MICE: PINCP



### Bivariate pmf



### Trivariate pmf

