

Testing different imputation methods on PUMS (MAR)

```
# 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 30% chance of missing: df_observed
missing_prob = 0.45
df_observed <- df
missing_col = c(1,3,7,9,10,11)

# Make VEH and WKL MCAR
missing_col_MCAR = c(1,10)
for (col in missing_col_MCAR) {
  missing_ind <- rbernoulli(n,p = missing_prob)
  df_observed[missing_ind, col] <- NA
}

# Make the rest MAR
numeric_df = sapply(df, as.numeric)
normalized_df = t(t(numeric_df-1)/(apply(numeric_df, MARGIN = 2, FUN = max)-1))
missing_col_MAR = c(3,7,9,11)
fully_observed_col = c(2,4,5,6,8)
beta_NP = c(0, 15, -10, 14, 0)
beta0_NP = -1
beta_SCHL = c(2.5, -10, 2.5, 0, 0)
beta0_SCHL = -1
beta_AGE = c(4.5, 6, 0, -10, 0)
beta0_AGE = -1
beta_PINCP = c(-10, 0, 4.5, 4.5, 0)
beta0_PINCP = -1

# missing probability for NP
prob_NP = apply(t(t(normalized_df[, fully_observed_col])*beta_NP), MARGIN = 1, sum)+beta0_NP
prob_NP = exp(prob_NP)/(exp(prob_NP)+1)
indicator = rbernoulli(n, p = prob_NP)
df_observed[indicator, missing_col_MAR[1]] <- NA

# missing probability for SCHL
prob_SCHL = apply(t(t(normalized_df[, fully_observed_col])*beta_SCHL), MARGIN = 1, sum) + beta0_SCHL
prob_SCHL = exp(prob_SCHL)/(exp(prob_SCHL)+1)
indicator = rbernoulli(n, p = prob_SCHL)
df_observed[indicator, missing_col_MAR[2]] <- NA

# missing probability for AGE
prob_AGE = apply(t(t(normalized_df[, fully_observed_col])*beta_AGE), MARGIN = 1, sum) + beta0_AGE
```

```

prob_AGEP = exp(prob_AGEP)/(exp(prob_AGEP)+1)
indicator = rbernoulli(n, p = prob_AGEP)
df_observed[indicator, missing_col_MAR[3]] <- NA

# missing probability for PINCP
prob_PINCP = apply(t(t(normalized_df[, fully_observed_col])*beta_PINCP), MARGIN = 1, sum) + beta0_PINCP
prob_PINCP = exp(prob_PINCP)/(exp(prob_PINCP)+1)
indicator = rbernoulli(n, p = prob_PINCP)
df_observed[indicator, missing_col_MAR[4]] <- NA

# 45.11% missing
apply(is.na(df_observed), MARGIN = 2, mean)

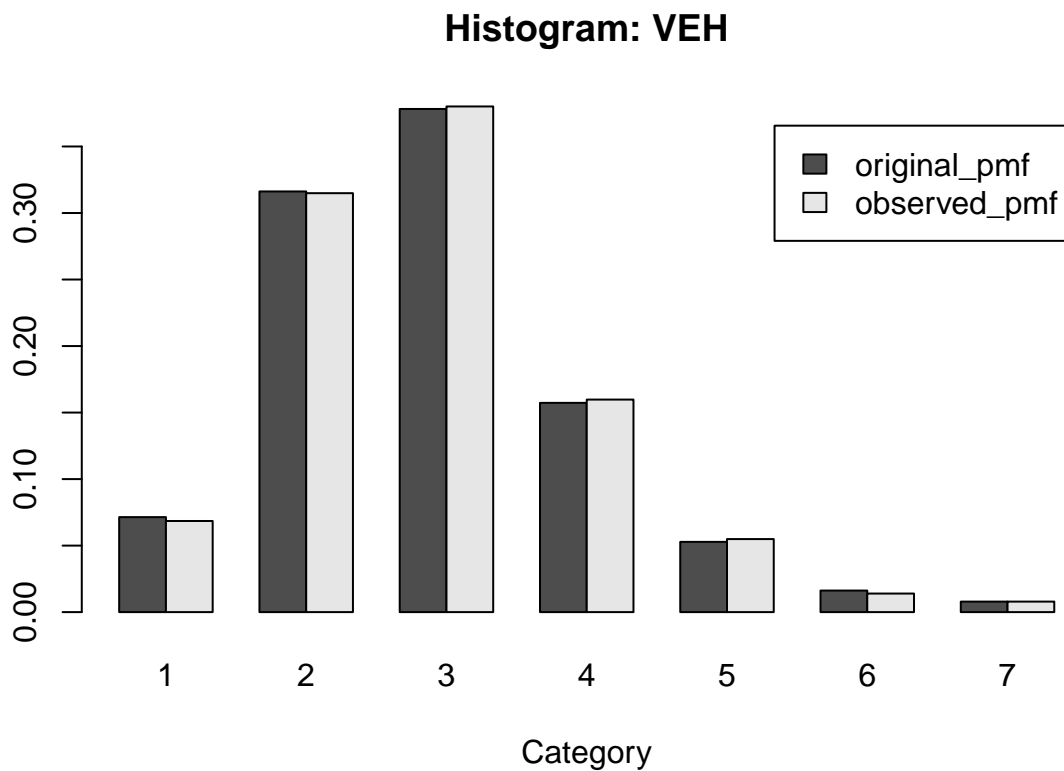
```

```

##      VEH      MV      NP      RMSP      ENG      MARHT      SCHL      RACNUM      AGEP      WKL      PINCP
## 0.4554 0.0000 0.4209 0.0000 0.0000 0.0000 0.4566 0.0000 0.4767 0.4552 0.4417

```

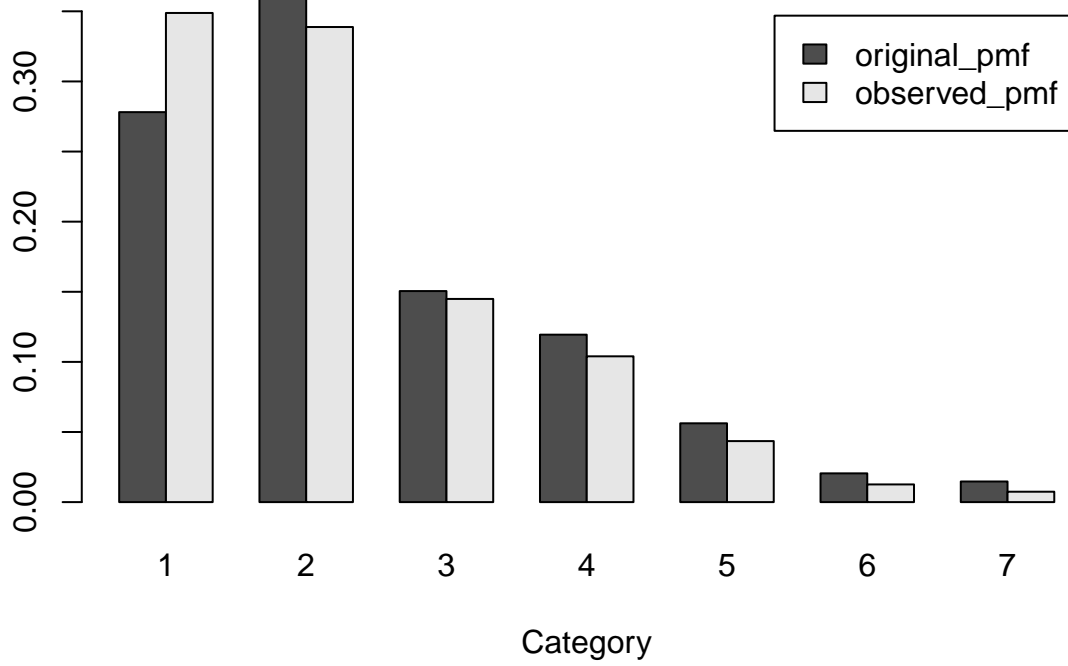
Histogram for univariate distribution



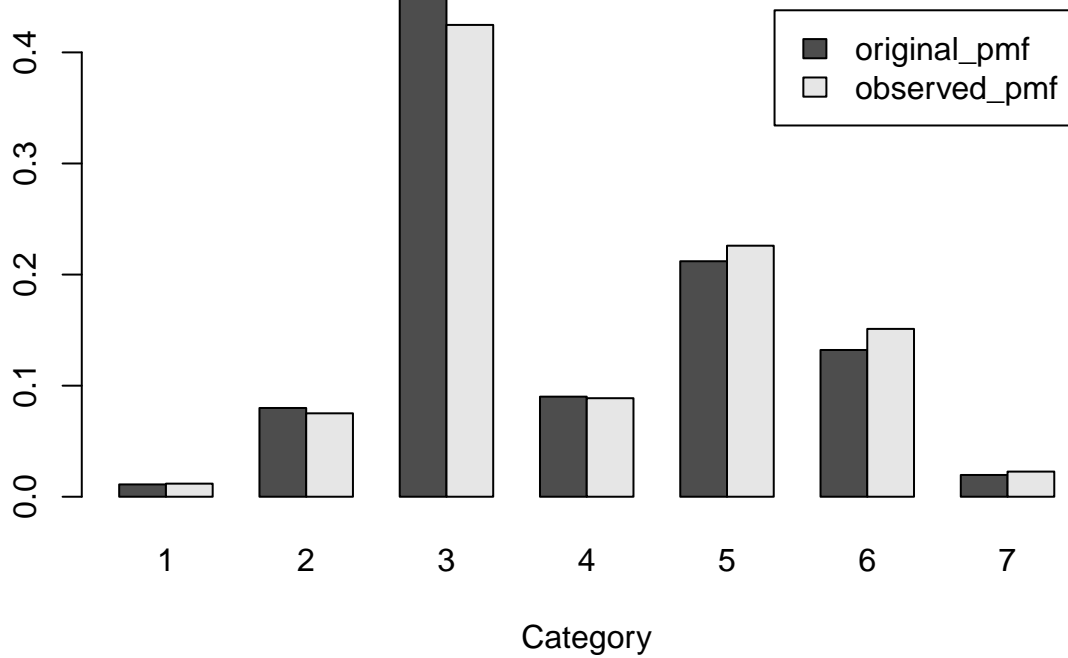
Assess bivariate joint distribution

Assess trivariate joint distribution

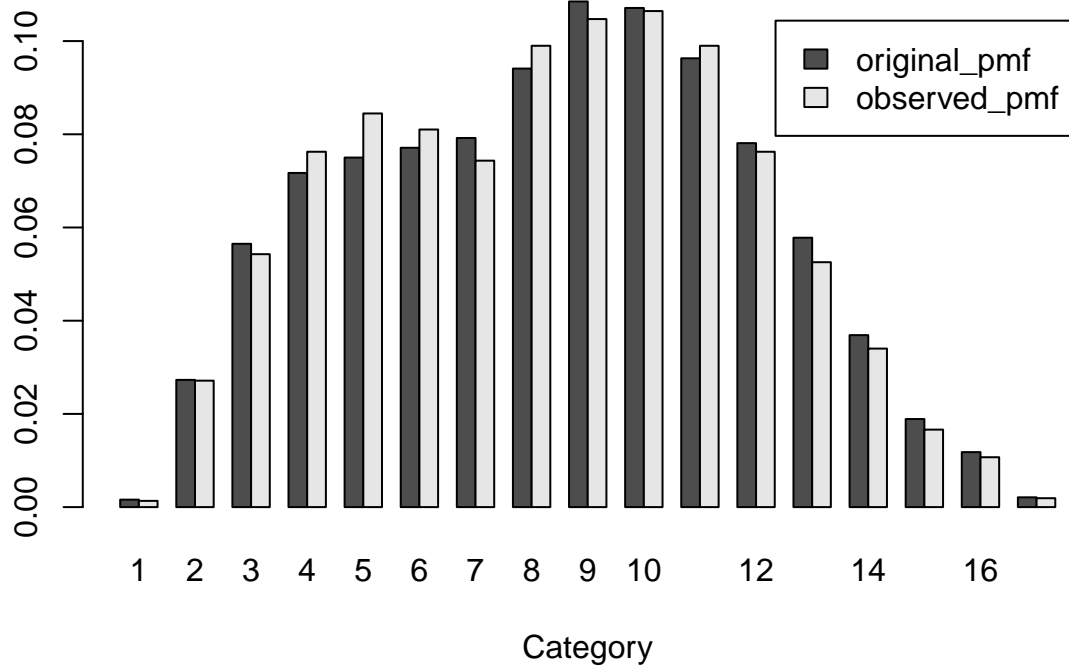
Histogram: NP



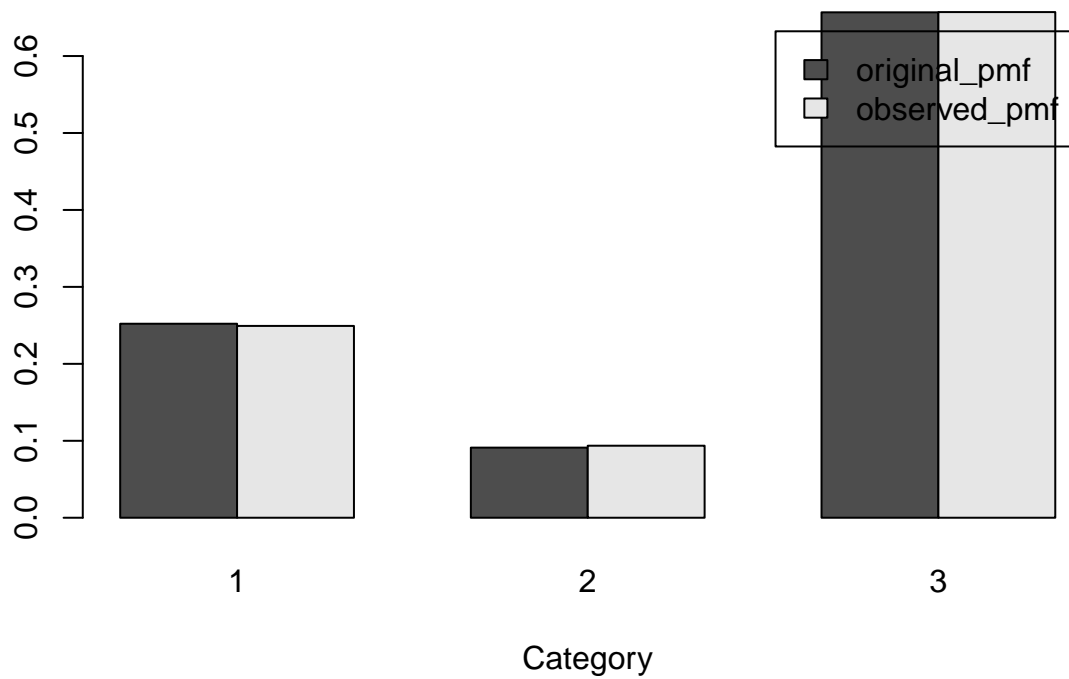
Histogram: SCHL



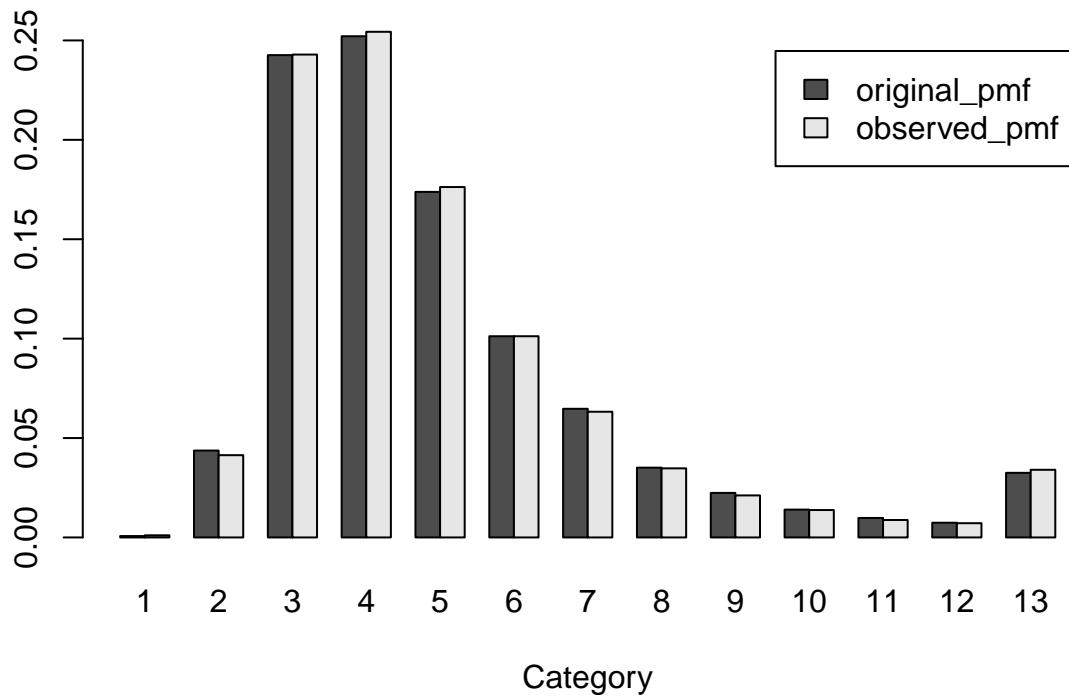
Histogram: AGEP



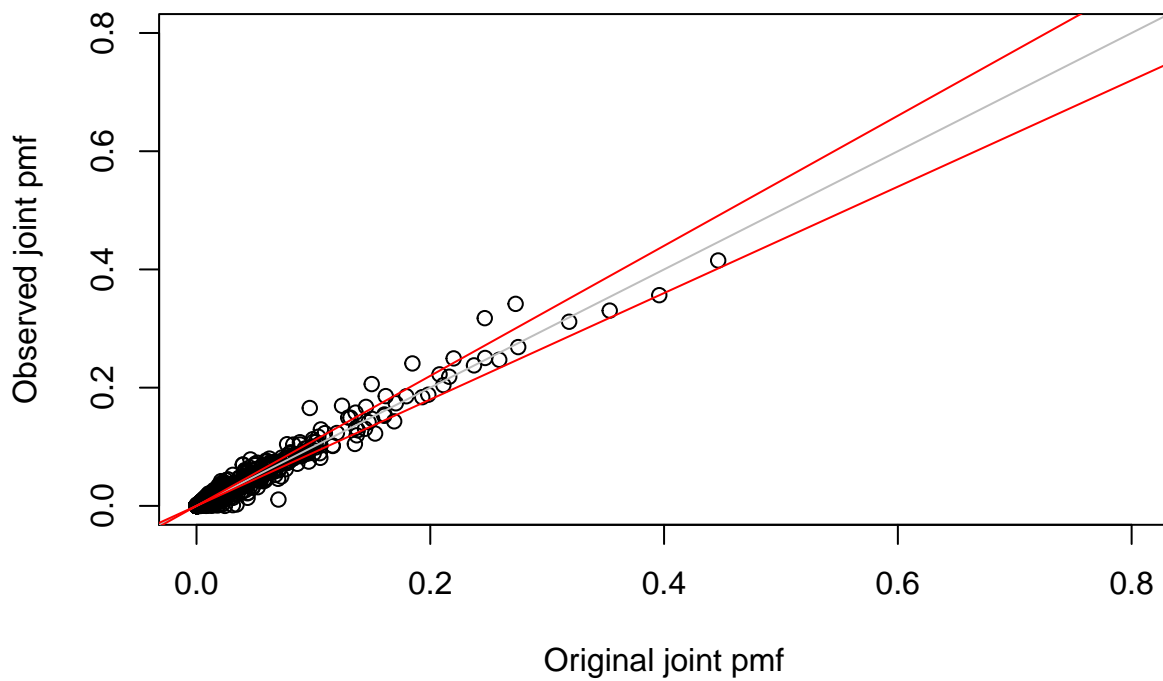
Histogram: WKL



Histogram: PINCP



Bivariate pmf



Trivariate pmf

