# Using Multiple Hot Deck Data Sets for Inference

Skyler Cranmer
Ohio State University

Jeff Gill
Washington University St. Louis

Natalie Jackson
The Huffington Post

Andreas Murr
University of Oxford

David A. Armstrong II
University of Wisconsin-Milwaukee

March 27, 2020

This document will walk you through some of the methods you could use to generate pooled model results that account for both sampling variability and across imputation variability. The package `hot.deck` does not come with a set of functions to do inference, so we will show you how you could use the data generated by `hot.deck` in combination with `glm.mids` (and similarly `lm.mids`) from the `mice` package, `zelig` from the `Zelig` package and by using `MIcombine` from the `mitools` package on a list of model objects.

## 1 Generating Imputations

The data we will use come from Poe, Tate and Keith (1999) dealing with democracy and state repression. First we need to call the `hot.deck` routine on the dataset.

```
> library(hot.deck)
> data(isq99)
> out <- hot.deck(isq99, sdCutoff=3, IDvars = c("IDORIGIN", "YEAR"))
```

This shows us that there are still 47 observations with fewer than 5 donors. Using a different method or further widening the `sdCutoff` parameter may alleviate the problem. If you want to see the frequency distribution of the number of donors, you could look at:

```
> numdonors <- sapply(out$donors, length)
> numdonors <- sapply(out$donors, length)
> numdonors <- ifelse(numdonors > 5, 6, numdonors)
> numdonors <- factor(numdonors, levels=1:6, labels=c(1:5, ">5"))
> table(numdonors)
```

```
numdonors
   1    2    3    4    5   >5
  18   10   11    6   20 4596
```

Before running a model, three variables have to be created from those existing. Generally, if variables are deterministic functions of other variables (e.g., transformations, lags, etc...) it is advisable to impute the constituent variables of the calculations and then do the calculations after the fact. Here, we need to lag the `AI` variable and create percentage change variables for both population and per-capita GNP. First, to create the lag of `AI`, `PCGNP` and `LPOP`. To do this, we will make a little function.

```
> tscslag <- function(dat, x, id, time){
+         obs <- apply(dat[, c(id, time)], 1, paste, collapse=".")
+         tm1 <- dat[[time]] - 1
+         lagobs <- apply(cbind(dat[[id]], tm1), 1, paste, collapse=".")
+         lagx <- dat[match(lagobs, obs), x]
+ }
> for(i in 1:length(out$data)){
+     out$data[[i]]$lagAI <- tscslag(out$data[[i]], "AI", "IDORIGIN", "YEAR")
+     out$data[[i]]$lagPCGNP <- tscslag(out$data[[i]], "PCGNP", "IDORIGIN", "YEAR")
+     out$data[[i]]$lagLPOP <- tscslag(out$data[[i]], "LPOP", "IDORIGIN", "YEAR")
+ }
```

Now, we can use the lagged values of `PCGNP` and `LPOP`, to create percentage change variables:

```
> for(i in 1:length(out$data)){
+     out$data[[i]]$pctchgPCGNP <- with(out$data[[i]], c(PCGNP-lagPCGNP)/lagPCGNP)
+     out$data[[i]]$pctchgLPOP <- with(out$data[[i]], c(LPOP-lagLPOP)/lagLPOP)
+ }
```

## 1.1 Using MIcombine

You can use the `MIcombine` command from the `mitools` package to generate inferences, too. Here, you have to produce a list of model estimates and the function will combine across the different results.

```
> # initialize list
> out <- hd2amelia(out)
> results <- list()
> # loop over imputed datasets
> for(i in 1:length(out$imputations)){
+     results[[i]] <- lm(AI ~ lagAI + pctchgPCGNP + PCGNP + pctchgLPOP + LPOP + MIL2 + LEFT +
+     BRIT + POLRT + CWARCOW + IWARCOW2, data=out$imputations[[i]])
+ }
> summary(mitools::MIcombine(results))

Multiple imputation results:
      MIcombine.default(results)
                  results           se        (lower        upper) missInfo
(Intercept)  5.414609e-01 1.426877e-01  2.575031e-01  8.254187e-01     24 %
lagAI        4.645526e-01 3.265179e-02  3.868938e-01  5.422114e-01     81 %
pctchgPCGNP  8.638575e-03 6.374589e-03 -6.692200e-03  2.396935e-02     83 %
PCGNP       -2.109636e-05 4.282693e-06 -3.046329e-05 -1.172943e-05     64 %
pctchgLPOP  -5.322637e-02 1.553785e+00 -3.752742e+00  3.646289e+00     82 %
LPOP         7.343262e-02 9.143452e-03  5.520244e-02  9.166281e-02     26 %
MIL2         1.173896e-01 3.967121e-02  3.772848e-02  1.970508e-01     31 %
LEFT        -1.565735e-01 5.995653e-02 -2.855496e-01 -2.759742e-02     60 %
BRIT        -1.310100e-01 3.276232e-02 -1.957653e-01 -6.625472e-02     18 %
POLRT       -7.242826e-02 1.121835e-02 -9.618840e-02 -4.866813e-02     55 %
CWARCOW      6.110300e-01 6.852079e-02  4.669423e-01  7.551177e-01     52 %
IWARCOW2     1.974241e-01 5.633958e-02  8.567129e-02  3.091769e-01     21 %
```

## 1.2 Using mids

The final method for combining results is to convert the data object returned by the `hot.deck` function to an object of class `mids`. This can be done with the `datalist2mids` function from the `miceadds` package.

```
> out.mids <- miceadds::datalist2mids(out$imputations)
> s <- summary(mice::pool(mice::lm.mids(AI ~ lagAI + pctchgPCGNP + PCGNP + pctchgLPOP + LPOP + MIL2 + LEFT +
+ BRIT + POLRT + CWARCOW + IWARCOW2, data=out.mids)))
> print(s, digits=4)
```

```
          term   estimate std.error statistic     df   p.value
1  (Intercept)  5.519e-01 1.396e-01    3.9518 133.45 1.251e-04
2        lagAI  4.447e-01 1.739e-02   25.5774 123.48 0.000e+00
3  pctchgPCGNP  4.608e-03 4.052e-03    1.1371  12.14 2.774e-01
4        PCGNP -2.041e-05 4.390e-06   -4.6490  10.74 7.531e-04
5   pctchgLPOP -2.228e-01 9.770e-01   -0.2281  61.34 8.203e-01
6         LPOP  7.637e-02 8.378e-03    9.1162 576.43 0.000e+00
7         MIL2  1.332e-01 3.561e-02    3.7395 447.38 2.084e-04
8         LEFT -1.560e-01 5.152e-02   -3.0275  29.45 5.087e-03
9         BRIT -1.352e-01 3.545e-02   -3.8125  53.31 3.587e-04
10       POLRT -7.534e-02 9.583e-03   -7.8626  44.89 5.561e-10
11     CWARCOW  6.283e-01 5.487e-02   11.4495 144.68 0.000e+00
12    IWARCOW2  2.056e-01 5.428e-02    3.7879 265.44 1.880e-04
```

# References

Poe, Steven, C. Neal Tate and Linda Camp Keith. 1999. "Repression of the Human Right to Personal Integrity Revisited: A Global, Cross-National Study Covering the Years 1976-1993." *International Studies Quarterly* 43:291–313.