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

February 24, 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)
+ }
```

2 Running Models on Multiple Hot Decking Result

2.1 Using Zelig

In version ≥ 5.0 of `Zelig`, the output from `hot.deck` will have to be converted into a format that looks like Amelia's. You can do this as follows:

```
> out <- hd2amelia(out)
```

Then, with the output in the appropriate format, we can use `Zelig` to do the modeling.

```
> library(Zelig)
> z <- zelig(AI ~ lagAI + pctchgPCGNP + PCGNP + pctchgLPOP + LPOP + MIL2 + LEFT +
+     BRIT + POLRT + CWARCOW + IWARCOW2, data=out, model="normal", cite=FALSE)
> summary(z)

Model: Combined Imputations

             Estimate Std.Error z value Pr(>|z|)
(Intercept)  5.41e-01  1.29e-01    4.19  2.8e-05
lagAI        4.51e-01  2.96e-02   15.24  < 2e-16
pctchgPCGNP  8.01e-03  6.32e-03    1.27   0.2046
PCGNP       -2.22e-05  3.45e-06   -6.44  1.2e-10
pctchgLPOP  -6.95e-01  8.80e-01   -0.79   0.4294
LPOP         7.62e-02  9.50e-03    8.01  1.1e-15
MIL2         1.08e-01  4.38e-02    2.48   0.0133
LEFT        -1.69e-01  5.73e-02   -2.95   0.0032
BRIT        -1.27e-01  3.12e-02   -4.08  4.4e-05
POLRT       -7.22e-02  1.10e-02   -6.55  5.8e-11
CWARCOW      6.56e-01  5.20e-02   12.61  < 2e-16
IWARCOW2     1.95e-01  5.94e-02    3.28   0.0010

For results from individual imputed datasets, use summary(x, subset = i:j)
Next step: Use 'setx' method
```

Note that the summary indicates that the results have been combined across 5 multiply imputed datasets.

2.2 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
> 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)  5.409996e-01 1.290335e-01  2.879718e-01  7.940274e-01      4 %
lagAI        4.508905e-01 2.959084e-02  3.824401e-01  5.193408e-01     77 %
pctchgPCGNP  8.012387e-03 6.315472e-03 -6.877721e-03  2.290250e-02     80 %
PCGNP       -2.221411e-05 3.447866e-06 -2.925084e-05 -1.517739e-05     40 %
pctchgLPOP  -6.951217e-01 8.796181e-01 -2.488792e+00  1.098549e+00     40 %
LPOP         7.616794e-02 9.503724e-03  5.707302e-02  9.526286e-02     31 %
MIL2         1.084923e-01 4.380054e-02  1.818793e-02  1.987968e-01     45 %
LEFT        -1.691424e-01 5.734644e-02 -2.903893e-01 -4.789549e-02     54 %
BRIT        -1.273566e-01 3.118678e-02 -1.885943e-01 -6.611887e-02      8 %
POLRT       -7.216079e-02 1.101875e-02 -9.529812e-02 -4.902347e-02     52 %
CWARCOW      6.564780e-01 5.204289e-02  5.542129e-01  7.587430e-01     10 %
IWARCOW2     1.952307e-01 5.944839e-02  7.595791e-02  3.145036e-01     30 %
```

2.3 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)  5.742e-01 1.359e-01    4.2253 166.800 3.918e-05
2        lagAI  4.642e-01 2.652e-02   17.5027   9.449 1.614e-08
3  pctchgPCGNP  3.075e-03 2.763e-03    1.1130  41.265 2.721e-01
4        PCGNP -2.076e-05 3.375e-06   -6.1519  30.333 8.717e-07
5   pctchgLPOP -5.469e-01 8.769e-01   -0.6237  12.257 5.443e-01
6         LPOP  7.282e-02 1.095e-02    6.6484  17.945 3.119e-06
7         MIL2  9.239e-02 4.552e-02    2.0296  19.209 5.648e-02
8         LEFT -1.652e-01 5.297e-02   -3.1192  23.319 4.772e-03
9         BRIT -1.261e-01 3.162e-02   -3.9873 275.638 8.567e-05
10       POLRT -7.495e-02 1.002e-02   -7.4776  28.581 3.353e-08
11     CWARCOW  6.381e-01 5.880e-02   10.8516  45.156 3.597e-14
12    IWARCOW2  1.814e-01 5.474e-02    3.3137 154.540 1.147e-03
```

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.