Madison workflow research log - template {ctrl-alt-h}

# Data source description:

I have CPS ASEC for 2005-2022. I have CPS Basic monthly for all months up through March 2023.

For ASEC prior to 2019, the data is only in DAT format from Census so I used the NBER .dct and .do files to read in the data to .dta format. For 2019-2022, data is downloadable as CSV. Appending those CSVs to the previous files carries over the relevant labels. Shigeru wanted me to modify the 2018 NBER dictionary and do files for 2019 onwards so that the whole timespan could be read in from DAT format, which I painstakingly did, but it doesn’t run correctly. The CSV files are fine.

For CPS monthly prior to 2020, I have code from Cole and Tal that reads in the monthly files to. dta from .raw format. That code is located in the “Reasons for Nonparticipation” project folder I inherited.

# Transcribed notes from Madison notebook (3/29):

Can we impute the probability of treatment using observables other than simply whether a tax unit (family) has kids?

Want to identify people who wouldn’t have qualified for CTC under prior rules but would under expansion.

Should we interact postCTC\*kids\*woman?

Can we extend the time period of analysis beyond Ananat etc. to see if there’s changes in LFP, post-cessation of Advance CTC paymnts.

Can we see whether people change labor supply on the intensive margin (# hrs)?

Any way, we need income information from ASEC in the basic monthly because any effect is most likely seen only in the lowest-income families and because income determines the amount of benefit a family is eligible for.

Ananat and Enriquez papers have different imputation methods.

## Imputing income to basic CPS from ASEC (typed 4/24):

### Ananat, Glasner, Hamilton, Parolin (2022)

They use the 2019 CPS ASEC to estimate pre-and post-reform benefit values.

They define bins (10x11x8) defined by the number of adults in the household (1-10), number of children in HH (0-10), and categorical pre-tax income (recode HEFAMINC into 8 categories from 16).

Then they compute the mean pre-reform CTC benefits for each family unit in the ASEC. Then they simulate the additional post-reform benefit values, creating a “net benefit” for each family unit.

Then they calculate the weighted mean of the net benefit value for each of the bins and carry that over to the monthly, matching on the bin criteria. So they don’t impute income directly. They only impute expected net benefit from expansion.

### Enriquez, Jones, and Tedeschi (2023)

They impute continuous family income from ASEC using weighted random draw (which I think means weighted hotdeck). Per email from Enriquez:

The draw between the monthly CPS and the ASEC is based on HEFAMINC (monthly) and FTOTVAL (IPUMS ASEC). We conduct separate draws for each of 16 cells based on marital status (x2), elder status (x2), and number of kids (x4). Han, Meyer, and Sullivan (2022) have an example of this procedure.

For each family income draw, we pull in that same record’s AGI in the ASEC, which is calculated by the Census’ tax model. For non-filers, we assign their family income as their AGI.

(To simulate the CTC values) basic monthly CPS has no filer information, so we use either the single or married-filing-jointy CTC parameters based on marriage status, the random draw of continuous AGI, and number of kids.

They do not adjust the CTC level for incomplete takeup.

They then create a CTC-to-income ratio for each HH. Then, they order HHs in their percentile of this ratio, which is the main regressor.

## Why would I want to do a different imputation method?

The weighted hotdeck is at the individual level, so it doesn’t preserve married pairs or spousal characteristics bc it’s not model-based. Hotdecking limits the number of observables you can use bc you can’t have too many stratifiers or the cell size will be too small. We don’t even know that the particular stratifiers Enriquez et al chose are good stratifiers. Could I use LASSO to figure out which observables are most useful and try those as stratifiers? LASSO can be used for model selection.

## Multiple imputation has advantage over single imputation because it allows for correct variance estimation of the imputed variable.

Within Stata’s MI command, we can do a MI “version” of hotdecking called predictive means matching (PMM). Hotdecking is apparently a Bayesian bootstrap. I think Multiple Imputation with PMM has the advantage of not defining a particular functional form for the model, similar to hotdecking, and it does use draws but it’s still model-based.

Multiple Imputation by Chained Equations (MICE) is model-based and can use more observables. I explore model creation and MICE imputation in “MICE\_enriquez.do” There’s MI chained and MI monotone. Frankly I’m not sure how monotone works. MICE needs data to be missing at random and I’m not sure how monotonicity falls into that.

MICE seems to be a form of conditional mean imputation. It imputes data per variable by specifying an imputation model for each variable.

To create multiple imputations it is recommended to include a large number of predictors in the imputation model, especially variables that will be used in subsequent analyses. Imputation model should be “congenial” to the analysis model.

Both PMM (Predictive Mean Matching) and regression methods can be used for multiple imputation of missing data. [The PMM method ensures that imputed values are plausible and might be more appropriate than the MVN regression method (which assumes a joint multivariate normal distribution) if the normality assumption is violated](https://stats.oarc.ucla.edu/r/faq/how-do-i-perform-multiple-imputation-using-predictive-mean-matching-in-r/). Generally, PMM is preferred to regression if the normality of the underlying model is suspect.

NOTE: Enriquez specifically says they pull all 3 imputed values from the same donor. This is not possible when doing MI PMM. What even are the benefits of keeping the donor the same? The fact that it’s a plausible combo of values? **Try MI Monotone PMM for all 3 imputed values at once and see if that generates plausible combos.**

I am also not sure if my weighted hotdeck pulls from the same donor for each imputed value. Check plausibility of combinations of imputed values.

##### How to test which methods do a better job?

I will remove the actual values for a random subset of the ASEC data and test each method to impute those values. Whichever has the smallest residuals I’ll call the best one.

# 4/25/2023:

Plausibility for the MI monotone PMM was bad, but could partially be bc I didn’t carry the AGI values out to both the head and the spouse, so all spouses had agi==0, which explains why the model fit was garbage for that one. I don’t think the MI PMM method pulls all three values from the same donor. It does each variable separately.

I don’t like that the spousal pairs aren’t preserved through the imputation procedures. I want to impute at the family level.

Collapse to hrhhid hrhhid2 and ph\_seq.