Notes on imputation procedure for income in expanded child tax credit project

Madison Perry

April 2023

1 Imputation methods:

- Ananat, Glasner, Hamilton, and Parolin (2022)
  - They do cell mean imputation using combinations of # of adults (1-10), number of kids (0-10), and pretax income category (recoded HEFAMINC into 8 groups) as the cells.
  - take the weighted mean of the pre-expansion CTC benefit value for each bin and carry that over to the people in those same bins in the basic monthly CPS.
  - then simulate the additional post-reform benefits that each family is eligible for using the detailed policy rules in the 2021 American research plan
  - subtract tht pre-reform benefit value from the post-reform value to create a net benefit amount for each family.
  - calculate the weighted mean of the net benefit for each of the bins.
  - import that to the monthly CPS panel, matching on # of adults, number of kids, and pretax income category.
- Enriquez, Jones, and Tedeschi (2023)
  - They impute continuous family income from ASEC using weighted random draw (which I think means weighted hotdeck - Hotdecking is a Bayesian bootstrap).

  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.

- Additional considerations
  - Hotdeck imputation is a form of single imputation and is not model-based. Multiple imputation (MI) 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). 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 weighted random draws but it's still model-based.
  - Now, there's two ways I could perform MI with PMM - MI chained and MI monotone. Multiple Imputation by Chained Equations (MICE) is model-based and can use more observables than hotdeck. Frankly I'm not sure how monotone works differently. "Monotone" just refers to the fact that the data's pattern of missingness is monotone, so (?) less iterations are required. MICE needs data to be missing at random and I'm not sure how monotonicity falls into that. Both are 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, which just means it should contain the variables you will use in 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. 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.

Questions:

- Do EJT's stratifiers create cells with sufficiently differentiated average values for the imputed variables?
- Can we evaluate the quality of a given imputation method against others?

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\ln(y_i) - \ln(\hat{y}_i))^2}$$

2 HEFAMINC issue

Table 1: ASEC 2020 mean values by HEFAMINC category

	Total family income	fam_agi	Total family earnings
Family income			
Less than $5,000	33,235	25,727	20,195
$5,000 to $7,499	36,768	29,275	24,867
$7,500 to $9,999	27,321	18,547	14,350
$10,000 to $12,499	34,318	25,585	20,608
$12,500 to $14,999	37,454	26,585	21,068
$15,000 to $19,999	36,856	26,950	21,442
$20,000 to $24,999	40,413	30,653	24,820
$25,000 to $29,999	47,080	37,316	32,115
$30,000 to $34,999	51,083	41,491	35,605
$35,000 to $39,999	57,933	47,633	41,777
$40,000 to $49,999	63,752	53,079	46,700
$50,000 to $59,999	74,873	65,017	57,609
$60,000 to $74,999	88,224	78,068	71,108
$75,000 to $99,999	106,221	96,686	88,424
$100,000 to $149,999	143,269	133,125	122,595
$150,000 and over	230,669	221,125	200,781
Total	107,622	97,792	88,306

It's weird that the self-declared numbers for HEFAMINC are so far off and it's weird that the bottom two categories have means that are higher than the third lowest. I omit the tables for 2021 and 2022 ASEC but they show the same phenomenon. If we tabulate the frequency of people's HEFAMINC responses, they skew high, so there are way fewer people in the lowest HEFAMINC categories. The cell size tables bear this out. If you look at the means by cell within each HEFAMINC category, there are decent differences between cells.

[look at cell distributions document]

I think if we just use HEFAMINC as a way to get differentiation between groups of people and not as a reliable measure of family income, it might be ok but if the variable has no actual relation to real family income, it's not clear what that differentiation would be based on - what's different about the groups that indicate HEFAMINC==1 from those that indicate HEFAMINC==4 if neither category is correct?

I think it's a bad variable with no salient relationship to the values we're trying to impute and I don't want to use it. I will show the distributions generated by the other stratifiers, omitting HEFAMINC in a separate document. I am not sure they are differentiated enough though. I think number of adults (1,2,3,4+) could be a useful stratifier so I'm going to see whether adding that helps.

3 Madison's April 27 status

I computed RMSE and RMSLE for the Enriquez hotdeck imputation. They are really big, but there's always a good chance I've done something wrong.

Table 2: Hotdeck imputation RMSE

	year		
	2020	2021	2022
Total family income	76,760	75,123	84,176
Family AGI	75,535	74,402	84,778
Total family earnings	68,732	66,750	75,167

I also computed RMSE for simple linear predicted values using the stratifiers as regressors (ftotval marital numkid_bin elderly hefaminc). My interpretation of these numbers is that the hotdeck imputation does approximately as bad a job assigning realistic income values as linear prediction.

Table 3: Linear prediction RMSE

	year		
	2020	2021	2022
Total family income	73,396	75,401	77,169
Family AGI	72,574	75,757	79,066
Total family earnings	67,122	67,434	69,791

Additionally, I ran a multiple imputation by chained equations model and have the imputed monthly data from that. I need to run it on the smaller sample I used for computing the fit measures for the hotdeck imputation. MI is a model-based approach so I was able to use more predictor variables and I was able to include spousal employment status as a predictor.

I want to try my idea for hotdeck at the family level because I think it's important to preserve spousal pairings in the data. I started working on this code last week. I think it's important to preserve spousal pairings or include some spouse information in the analysis because people make LFP decisions at the family level. The people who we might think would be most likely to change their LFP status in the short term in response to this money would be women whose spouses work and make more money than them.

Comparing imputation procedures - CTC project

Madison Perry

May 8, 2023

1 General process

I use the 2020, 2021, and 2022 ASECs separately, limiting to only adult observations. I create a clone variable for each of the variables I want to impute. I randomly split the sample in half, using a set seed (so the split is the exact same for all repetitions) and I delete the actual values from the clone variable for one half of the observations. I impute the missing values using one of the procedures below and then I check the "fit" of the imputation method by computing the root mean squared error using the actual and predicted values for the imputed half.

The RMSE is the square root of the average squared difference between the predicted values from a model and the actual values observed in a dataset. Here, the lower the RMSE, the more accurate the imputation model is. All of these RMSEs are pretty large if you consider that they are in terms of dollars - an average deviation of $60,000 is really inaccurate but I'm not sure what else to do to decrease that.

2 Hotdeck

Table 1: Cells: marital numkid_bin elderly hefaminc, weight=marsupwt

	RMSE		
	2020	2021	2022
Total family income	76,760	75,123	84,176
Family AGI	75,535	74,402	84,778
Total family earnings	68,732	66,750	75,167

3 Linear prediction

These RMSE represent the average magnitude of deviation between the redicted values for OOS observations from "regress [var] marital numkid_bin elderly hefaminc, weight=marsupwt":

Table 2: Linear prediction with hotdeck stratifiers as covariates

	RMSE		
	2020	2021	2022
Total family income	73,396	75,401	77,169
Family AGI	72,574	75,757	79,066
Total family earnings	67,122	67,434	69,791

4 Multiple Imputation by Chained Equations

This implementation of the MICE algorithm was slightly different from the normal because the pattern of missingness in the data is monotone, meaning that if a value is missing for one of the imputed variables, the others are also always missing for the same observations. This means that iterations changing the sequence in which variables are imputed is unnecessary. For all three imputed variables, I use the predictive mean matching approach with k-nearest neighbors equal to 3. That specifies the number of closest observations (nearest neighbors) from which to draw imputed values. Note: I can't make use of the weights in this imputation procedure.

Table 3: MICE using i.race i.educ i.sex i.hefaminc c.numadult_bin c.numkid_bin i.marital

	RMSE		
	2020	2021	2022
Total family income	94,206	95,975	98,503
Family AGI	94,051	96,947	102533
Total family earnings	85,153	85,608	87,922

5 Different linear prediction

Table 4: i.race i.educ i.sex i.hefaminc c.numadult_bin c.numkid_bin##i.marital [weight=marsupwt]

	RMSE		
	2020	2021	2022
Total family income	68,419	70,397	71,714
Family AGI	67,880	71,132	74,051
Total family earnings	64,232	64,489	66,605

At this point, I went back and tried to improve the imputation/prediction model more. That work is only shown in do files. The most current model is log-linear and uses RHS: age age^2 hefaminc elderly race##educ_max educ_min veteran HH_type##num_adults num_kid##under6 sex##marital to predict LHS: ln(ftotval), ln(fam_agi), and ln(fearnval).

We raised the issue of whether we should try to impute at the household level and then re-attribute back to the individuals the predicted values for the household of which they are members. It bothers me that we are only able to identify households in the Basic and yet we are imputing a family-level variable. However, the accuracy is even worse when we were trying to impute htotval but that may be bc the predicting is using individual level characteristics.

I want to first act like there's only one family per household and then try imputing family income with a regression at the household level. I want to also try Hotdeck at the family level using stratifiers (elderly, marital, numadults, haskids, HEFAMINC)

Labor force participation by income - Revisited

Madison Perry

May 22, 2023

Problem: I was trying to plot LFPR by income category continuously, using 5 categories I made using imputed values of FTOTVAL. However, the rates for each category were not lining up with the values computed for the same income categories defined using HTOTVAL in the ASEC data.

Goal was to reconcile the multiple differences in the calculation of LFPR by income between the March memo and my recent computations.

The only reason I imputed FTOTVAL (total family income) instead of HTOTVAL (total household income) is because it's what Enriquez et al do and one of the major variables they leverage for imputation is HEFAMINC (defined as total family income, binned). Perhaps the error is lower for the imputation method when they are using family-level variables. I will try using both their Hotdeck method and our loglinear method to impute HTOTVAL and compute the RMSE for each method using the ASEC test set.

Table 1: Madison's table: RMSEs for loglinear imputation model, by variable

	Year		
Variable	2019	2020	2021, 2022
FTOTVAL	68,730	70,739	72,366
FAM_AGI	68,754	73,353	76,735
FEARNVAL	62,786	62,998	65,105
HTOTVAL	70,271	72,559	73,886

This table shows that the log-linear model does slightly worse at imputing accurate values of HTOTVAL than it does with FTOTVAL and others. Still need to run the check on the hotdeck method's ability to impute HTOTVAL.

I also can redo Shigeru's memo computations using income categories based on FTOTVAL and see if those LFPR values match up with what I computed using the monthly data and imputed FTOTVAL.

Here we see that the thresholds are slightly lower for each quintile of total family income. I think this is just because households can contain multiple families, so total household income will generally be greater than or equal to total family income, depending on whether a household is single-family or

Table 2: Shigeru's table: Minimum total household income for each quintile ($)

	Year			
Quintile	2018	2019	2020	2021
1	0	0	0	0
2	26,019	28,791	30,790	31,401
3	50,001	53,801	56,359	58,089
4	79,716	86,555	89,204	93,012
5	130,001	142,501	145,710	152,500

Table 3: Madison's table: minimum total FAMILY income for each quintile ($)

	Year			
Quintile	2018	2019	2020	2021
1	0	0	0	0
2	21,321	23,508	25,106	25,912
3	41,001	45,006	48,810	48,910
4	68,403	74,801	76,801	79,650
5	116,800	128,002	130,219	137,965

not.

Since the thresholds of the quintiles are pulled down a little when we use family income, we may want to consider re-defining the fixed categories to closer reflect these values. As we see in the following few tables, setting the bottom category at 0-$25,000 results in the LFPR for that category being significantly higher and not really representing the LFPR of the poorest quintile of people.

Table 4: Shigeru's table: LFPRs by household income (% as of March)

	Year			
Income group ($)	2019	2020	2021	2022
– 24,999	29.3	27.8	24.4	24.1
25,000 – 49,999	50.7	48.6	45.1	45.7
50,000 – 99,999	66.2	63.8	61.9	62.4
100,000 – 149,999	74.3	72.5	71.2	72.1
150,000 –	76.7	75.5	75.9	76.2
Overall (ASEC, 15+)	62.1	61.4	60.4	61.4
Overall (Official NSA)	63.0	62.6	61.5	62.4

Notes: This table shows labor force participation rates for people 15 and up.

Table 5: LFPRs by actual family income, ASEC (% as of March)

	Year			
Income group ($)	2019	2020	2021	2022
– 24,999	35.8	33.6	31.0	29.5
25,000 – 49,999	58.7	56.3	54.2	53.8
50,000 – 99,999	70.2	67.2	65.8	65.7
100,000 – 149,999	77.6	75.5	74.4	74.3
150,000 –	80.3	78.7	79.3	79.6
Overall (ASEC, 18+)	65.3	64.1	63.4	63.8
Overall (Official NSA)	63.0	62.6	61.5	62.4

Notes: Sample is all participants 18+ from the March Supplements

Table 6: LFPRs by IMPUTED family income, ASEC (% as of March)

	Year		
Income group ($)	2020	2021	2022
– 24,999	40.9	40.9	39.9
25,000 – 49,999	56.3	55.7	56.1
50,000 – 99,999	65.2	64.8	64.5
100,000 – 149,999	74.9	74.8	74.7
150,000 –	78.9	78.5	79.1
Overall (ASEC, 18+)	64.3	63.5	63.9
Overall (Official NSA)	62.6	61.5	62.4

Notes: Sample is all participants 18+ from the March Supplements. Income groups are defined using the predicted values of ftotval for the same observations that appear in the ASEC.

Notice how the accuracy is worst for the bottom income category. This led us to examine the distribution of the predicted values and see that our model under-allocates people to the lowest income category (so it falsely attributes too-high incomes to some people). We played with estimating the model over different subsets of the data to see if it would make more accurate predictions, but no dice. There just isn't a covariate that we have that successfiully captures the variation in income in the lowest fixed category. Also, it doesn't make sense to arbitrarily limit the sample to 18+, so in the next iteration, we will use 16+, in keeping with BLS.

Table 7: LFPRs by family income imputed Basic (% as of March)

	Year			
Income group ($)	2020	2021	2022	2023
− 24,999	40.4	39.6	38.9	37.4
25,000 − 49,999	54.4	53.3	54.0	52.4
50,000 − 99,999	66.4	64.7	64.1	63.5
100,000 − 149,999	75.7	74.0	74.3	73.8
150,000 −	79.1	79.4	78.4	79.7
Overall (March Basic)	62.4	61.2	61.4	61.6
Overall (Official NSA)	63.0	62.6	61.5	62.4

Notes: This table uses the monthly basic data in March with family income categories made using imputed FTOTVAL

Table 8: LFPRs by family income (% annual average)

	Year			
Income group ($)	2019	2020	2021	2022
− 24,999	42.2	39.7	40.5	38.3
25,000 − 49,999	56.4	53.7	53.7	53.4
50,000 − 99,999	67.1	65.6	64.5	64.0
100,000 − 149,999	76.1	74.9	73.8	73.7
150,000 −	80.0	78.9	79.4	79.3
Overall (Year avg.)	63.4	61.6	61.4	61.3
Overall (Official NSA)	63.0	62.6	61.5	62.4

Notes: This table uses the CPS monthly basic data with family income categories made using imputed FTOTVAL. For 2019, I only used months August-December, simply because I chose that cutoff when I was imputing so that's what I had values ready for.

Labor force participation by imputed family income category

Madison informal notes

June 1, 2023

1 Verifying model w/ ASEC

1.1 No exclusion

Table 1: LFPRs by actual family income category, ASEC (% as of March)

	Year		
Income group ($)	2020	2021	2022
– 24,999	35.9	34.1	33.7
25,000 – 49,999	53.6	53.2	52.7
50,000 – 99,999	65.4	65.5	65.6
100,000 – 149,999	73.3	72.8	72.8
150,000 –	76.2	76.9	77.2
Overall (ASEC, 16+)	62.7	62.0	62.5
Overall (Official NSA)	62.6	61.5	62.5

Notes: Sample is all participants 16+ from the March Supplements

Table 2: LFPRs by predicted family income, ASEC (% as of March)

	Year		
Income group ($)	2020	2021	2022
– 24,999	38.0	38.0	37.3
25,000 – 49,999	56.0	55.2	55.6
50,000 – 99,999	63.1	63.0	63.1
100,000 – 149,999	72.8	73.0	72.7
150,000 –	75.4	75.3	75.7
Overall (ASEC, 16+)	62.7	62.0	62.5
Overall (Official NSA)	62.6	61.5	62.4

Notes: Sample is all participants 16+ from the March Supplements. Income groups are defined using the predicted values of ftotval for the same observations that appear in the ASEC.

I used different characterisitcs to identify the groups with the least attachment to the labor force so I can leave them out bc they are unlikely to be affected either way by the policy.

Note that these numbers in Table 2 are slightly different from ones I showed yesterday because I had to remove a variable from the imputation model that was not present in the Basic monthly dataset. Since the predictive model is different, some people's income values were differently imputed, resulting in some ending up in different predicted income categories. Hence, different LFPRs for the predicted categories.

1.2 Exclude 2-person retiree households

Table 3: LFPRs by actual family income category

	Year		
Income group ($)	2020	2021	2022
− 24,999	39.0	37.1	36.7
25,000 − 49,999	60.8	60.9	60.0
50,000 − 99,999	71.0	71.2	71.5
100,000 − 149,999	77.3	76.6	77.0
150,000 −	78.9	79.2	79.7
Overall	67.6	66.9	67.4

Notes: Sample is all participants 16+ from the March Supplements, excluding individuals from 2-person households wherin both people are retired.

Table 4: LFPRs by predicted family income category

	Year		
Income group ($)	2020	2021	2022
− 24,999	40.3	40.2	39.3
25,000 − 49,999	62.4	61.9	62.6
50,000 − 99,999	70.2	69.9	70.1
100,000 − 149,999	76.1	76.5	76.8
150,000 −	76.6	76.2	76.8
Overall	67.6	66.9	67.4

Notes: Sample is all participants 16+ from the March Supplements, excluding individuals from 2-person households wherin both people are retired. Income groups are defined using the predicted values of ftotval for the same observations that appear in the ASEC.

Table 5: LFPRs by predicted family income category, Basic March

	Year		
Income group ($)	2020	2021	2022
− 24,999	47.8	45.4	44.9
25,000 − 49,999	61.6	61.0	61.8
50,000 − 99,999	67.6	66.4	66.5
100,000 − 149,999	76.2	75.0	74.9
150,000 −	60.5	61.2	61.3
Overall	64.1	63.0	63.3

Notes: Sample is all participants 16+ from the March Basic, excluding individuals from 2-person households wherin both people are retired. Income groups are defined using the predicted values of ftotval for the Basic monthly observations.

Having applied the imputation model to the Basic monthly data, Table 5 shows the same computations as above, only on the people who appear in the March Basic data.

My only explanation for the differences would be that the people in the Basic are a subset of the people in

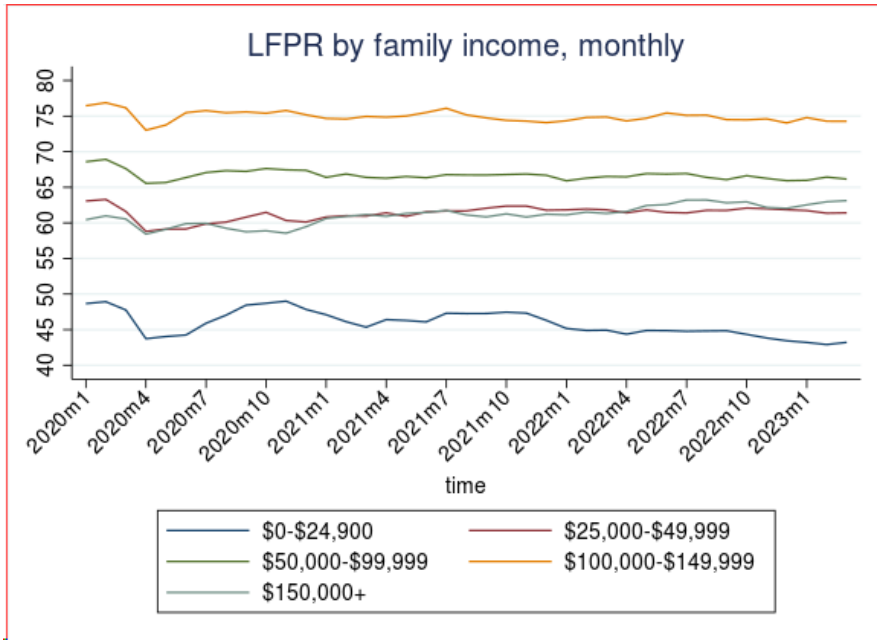

Figure 1:

the ASEC. ASEC has additional groups added in order to improve national representation and, for some reason, the subset of those individuals who appear in the March Basic survey have slightly different patterns in labor force participation by income category.

2 LFPR by income, monthly frequency

Labor force participation by imputed family income quintile

Madison informal notes

June 7, 2023

Why did we switch from using fixed thresholds to define income categories to now using quintiles? The following tables show the percent of each year's sample that falls into each fixed income category:

Actual	year		
Family Income	2020	2021	2022
− 24,999	14	15	14
25,000 − 49,999	18	18	17
50,000 − 99,999	29	29	28
100,000 − 149,999	17	17	17
150,000 −	22	22	24

Predicted	year		
Family Income	2020	2021	2022
− 24,999	11	12	12
25,000 − 49,999	21	22	21
50,000 − 99,999	30	29	29
100,000 − 149,999	18	17	17
150,000 −	19	19	21

Notice that the fixed thresholds don't cut the sample into equal proportions and that the proportional split changes slightly from year to year. The imputation model also allocates fewer people to the lowest income category than the actual ASEC data shows.

The two tables below show the lower threshold values for each quintile.

	Actual		
	year		
	2020	2021	2022
Quintile			
1	1	1	1
2	32,760	32,764	32,760
3	54,871	54,878	54,869
4	84,786	84,780	84,795
5	130,499	130,488	130,500

	Predicted		
	year		
	2020	2021	2022
Quintile			
1	6,036	6,848	4,889
2	32,835	32,835	32,835
3	54,635	54,636	54,640
4	84,306	84,306	84,306
5	129,237	129,230	129,232

1 Verifying model w/ ASEC

Now, I re-run the computations of labor force participation by income group, using quintiles instead of the fixed categories.

1.1 No exclusions

Table 1: LFPRs by actual family income quintile

	Year		
Income group ($)	2020	2021	2022
Lowest	40.4	38.4	38.0
Second	55.9	56.4	55.8
Middle	63.5	64.0	64.4
Fourth	71.4	70.5	70.5
Highest	75.9	76.4	76.3
Overall (ASEC, 16+)	62.7	62.0	62.5
Overall (Official NSA)	62.6	61.5	62.5

Notes: Sample is all participants 16+ from the March Supplements.

Table 2: LFPRs by predicted family income quintile

	Year		
Income group ($)	2020	2021	2022
Lowest	44.4	43.5	43.1
Second	57.0	56.8	57.3
Middle	61.6	61.3	61.5
Fourth	69.3	69.3	69.2
Highest	76.8	77.0	77.0
Overall (ASEC, 16+)	62.8	62.1	62.5
Overall (Official NSA)	62.6	61.5	62.5

Notes: Sample is all participants 16+ from the March Supplements. Income groups are defined using the predicted values of ftotval for the same observations that appear in the ASEC.

1.2 Exclude 2-person retiree households

Table 3: LFPRs by actual family income quintile, ASEC (% as of March)

	Year		
Income quintile ($)	2020	2021	2022
Lowest	44.4	42.4	41.9
Second	63.6	64.1	63.7
Middle	69.9	70.6	71.1
Fourth	76.0	75.3	75.6
Highest	78.9	79.0	79.4
Overall	67.6	66.9	67.4

Notes: Sample is all participants 16+ from the March Supplements, excluding individuals from 2-person households wherein both people are retired.

Table 4: LFPRs by predicted family income quintile, ASEC (% as of March)

	Year		
Income group ($)	2020	2021	2022
Lowest	47.5	46.9	46.3
Second	64.7	64.1	65.1
Middle	69.3	69.2	69.3
Fourth	75.0	75.2	75.2
Highest	78.4	78.3	78.6
Overall	67.8	67.1	67.6

Notes: Sample is all participants 16+ from the March Supplements. Income groups are defined using the predicted values of ftotval for the same observations that appear in the ASEC.

Why is the gap between the LFPRs for the bottom actual and predicted rates bigger when we use quintiles than when we use fixed income thresholds?

Having applied the imputation model to the Basic monthly data, Table 5 shows the same computations as above, only on the people who appear in the March Basic data.

Table 5: LFPRs by predicted family income quintile, Basic March

	Year		
Income group ($)	2020	2021	2022
Lowest	52.5	50.8	50.9
Second	63.1	61.4	62.3
Middle	66.5	65.6	66.0
Fourth	73.5	73.2	73.0
Highest	78.5	77.6	77.5
Overall	66.8	65.6	66.1

Notes: Sample is all participants 16+ from the March Basic, excluding individuals from 2-person households wherin both people are retired. Income groups are defined using the predicted values of ftotval for the Basic monthly observations.

My only explanation for the gap between the ASEC estimates and Basic estimates would be that the people in the Basic are a subset of the people in the ASEC. ASEC has additional groups added in order to improve

national representation and, for some reason, the subset of those individuals who appear in the March Basic survey have slightly different patterns in labor force participation by income category.

2 LFPR by income, monthly frequency

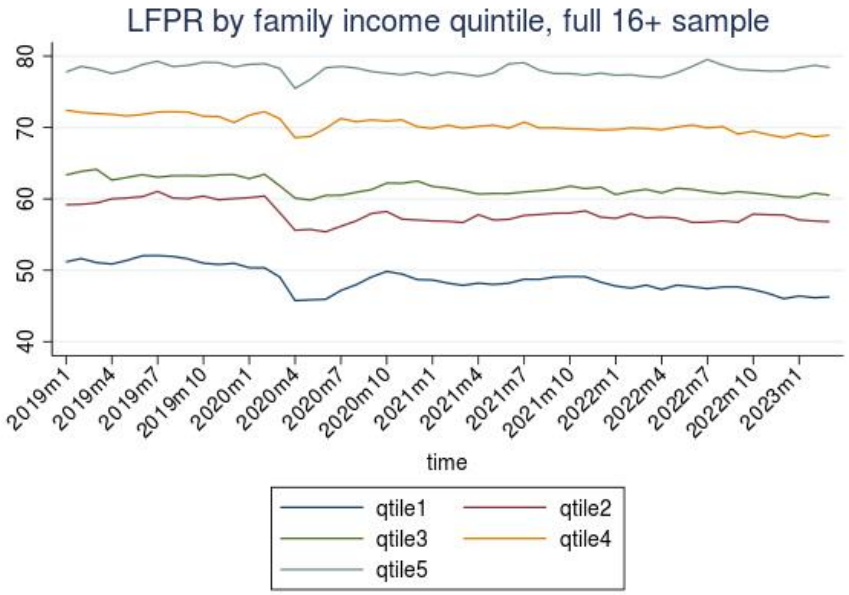

Figure 1: Full 16+ CPS Basic sample

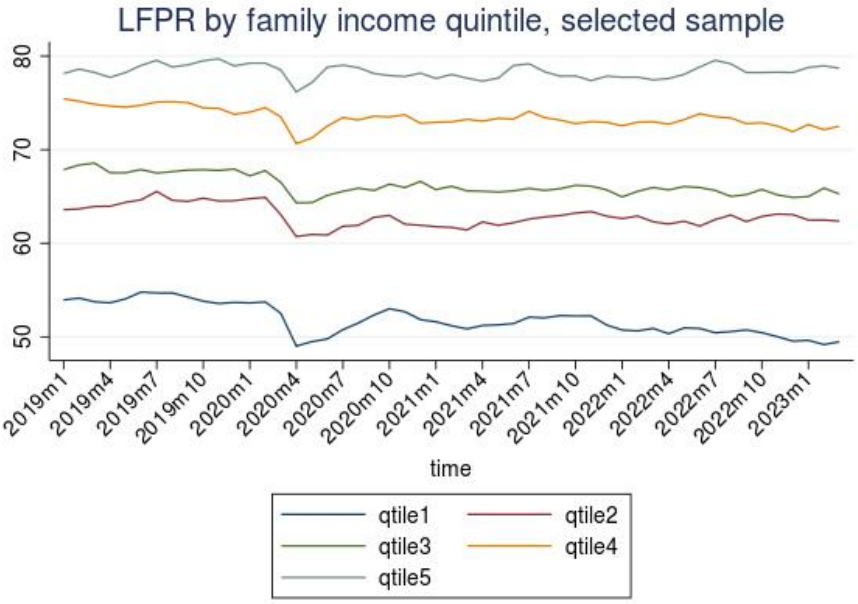

Figure 2: 16+ CPS Basic sample, excluding 2-person retiree households

Labor force participation by imputed family income quintile

Madison informal notes

June 12, 2023

The following tables show the percent of each year's sample that falls into each fixed income category:

Actual	year		
Family Income	2020	2021	2022
− 24,999	14.0	15.1	14.6
25,000 − 49,999	18.6	18.8	17.8
50,000 − 99,999	29.1	29.0	28.1
100,000 − 149,999	16.9	16.5	16.8
150,000 −	21.3	20.7	22.7

Predicted	year		
Family Income	2020	2021	2022
− 24,999	10.6	11.7	11.6
25,000 − 49,999	22.8	24.2	22.6
50,000 − 99,999	31.2	30.2	29.8
100,000 − 149,999	17.8	16.5	16.5
150,000 −	17.6	17.5	19.5

Notice that the fixed thresholds don't cut the sample into equal proportions and that the proportional split changes slightly from year to year. The imputation model also allocates fewer people to the lowest income category than the actual ASEC data shows.

The two tables below show the lower threshold values for each quintile.

	Actual		
	Year		
	2020	2021	2022
Quintile			
1	1	1	1
2	31,979	32,304	32,698
3	52,286	52,608	53,428
4	79,765	80,515	82,758
5	121,403	125,150	130,180

	Predicted		
	Year		
	2020	2021	2022
Quintile			
1	6,028	6,650	4,532
2	32,164	32,497	32,755
3	52,130	52,309	53,267
4	79,184	79,774	82,286
5	119,882	123,119	129,165

1 Verifying model w/ ASEC

Now, I re-run the computations of labor force participation by income group, using quintiles instead of the fixed categories.

1.1 No exclusions

Table 1: LFPRs by actual family income quintile

	Year		
Income quintile	2020	2021	2022
Lowest	39.9	38.4	38.1
Second	56.2	56.6	56.2
Middle	63.0	64.0	64.7
Fourth	71.2	70.6	71.0
Highest	76.0	76.7	77.0
Overall (ASEC, 16+)	62.8	62.1	62.5
Overall (Official NSA)	62.6	61.5	62.5

Notes: Sample is all participants 16+ from the March Supplements.

Table 2: LFPRs by predicted family income quintile

	Year		
Income quintile	2020	2021	2022
Lowest	44.1	43.5	43.3
Second	56.9	56.8	57.3
Middle	61.2	61.1	61.7
Fourth	68.8	69.1	69.8
Highest	76.5	76.9	77.0
Overall (ASEC, 16+)	62.8	62.1	62.5
Overall (Official NSA)	62.6	61.5	62.5

Notes: Sample is all participants 16+ from the March Supplements. Income groups are defined using the predicted values of ftotval for the same observations that appear in the ASEC.

1.2 Exclude 2-person retiree households & people with missing predicted values

Table 3: LFPRs by actual family income quintile, ASEC (% as of March)

	Year		
Income quintile	2020	2021	2022
Lowest	44.4	42.4	41.9
Second	63.5	64.4	64.3
Middle	69.9	70.9	71.6
Fourth	76.0	75.6	76.0
Highest	79.2	79.5	80.1
Overall	67.8	67.1	67.6

Notes: Sample is all participants 16+ from the March Supplements, excluding individuals from 2-person households wherein both people are retired.

Table 4: LFPRs by predicted family income quintile, ASEC (% as of March)

	Year		
Income quintile	2020	2021	2022
Lowest	47.0	46.9	46.5
Second	64.5	64.1	65.4
Middle	69.0	69.1	69.5
Fourth	74.5	75.0	75.5
Highest	78.3	78.2	78.7
Overall	67.8	67.1	67.6

Notes: Sample is all participants 16+ from the March Supplements. Income groups are defined using the predicted values of ftotval for the same observations that appear in the ASEC.

Having applied the imputation model to the Basic monthly data, Table 5 shows the same computations as above, only on the people who appear in the March Basic data.

Table 5: LFPRs by predicted family income quintile, Basic March

	Year		
Income quintile	2020	2021	2022
Lowest	52.5	50.9	51.1
Second	63.1	61.7	62.5
Middle	66.3	66.0	66.3
Fourth	73.3	73.3	73.1
Highest	78.5	77.9	77.5
Overall	66.8	65.6	66.1

Notes: Sample is all participants 16+ from the March Basic, excluding individuals from 2-person households wherin both people are retired. Income groups are defined using the predicted values of ftotval for the Basic monthly observations.

My only explanation for the gap between the ASEC LFPRs and Basic LFPRs would be that the people in the Basic are a subset of the people in the ASEC. ASEC has additional groups added in order to improve national representation and, for some reason, the subset of those individuals who appear in the March Basic survey have slightly different patterns in labor force participation by income category.

2 LFPR by income, monthly frequency

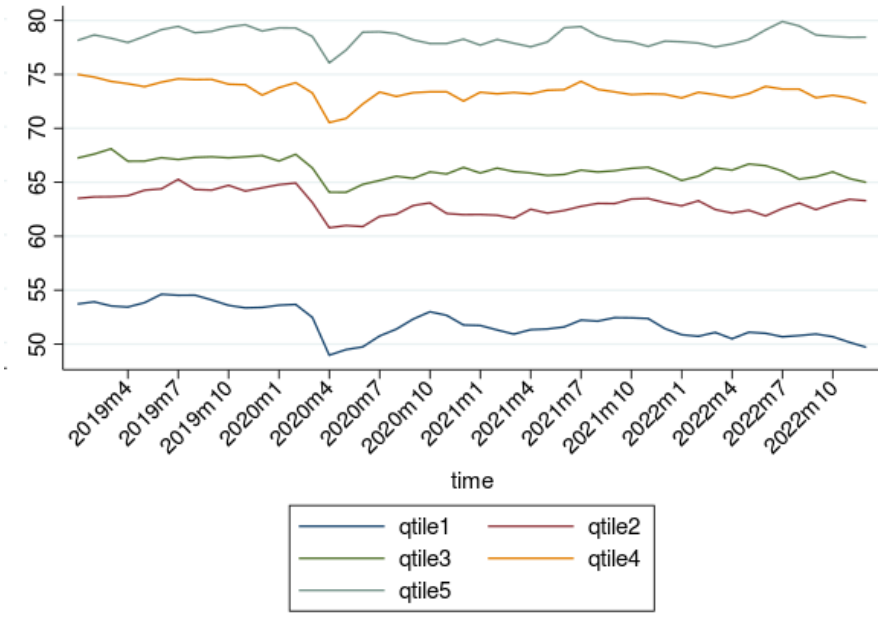

Figure 1: 16+ CPS Basic sample, excluding 2-person retiree households

Madison's questions:

- Why did we switch from using fixed thresholds to define income categories to now using quintiles?
- Why is the gap between the LFPRs for the bottom actual and predicted rates bigger when we use quintiles than when we use fixed income thresholds?
- What should I do next?