Family Income vs. Household Income

Madison CPS notes

July 10, 2023

1 Difference between variables

Total household income (htotval) is defined as the sum of the following items: farm income, self-employment income, wages and salaries, annuities, child support, disability income, dividend income, retirement distributions, education income, financial assistance income, interest income, other income, public assistance income, pension income, rental income, social security, supplemental security income, survivor income, unemployment compensation, veteran payments, and worker's compensation. Total family income (ftotval) is the family-level sum of the exact same items. Energy assistance, food stamps, COVID economic impact payments, and Child Tax Credit payments are not included in htotval, but these household-level variables do exist. There are no family-level variables for energy assistance and food stamps.

Families are different from households. Households may be single-family or multi-family. As an example of how the two relate, consider the universe of individuals with family income less than $25,000. Everyone with a household income under $25k also has a family income under $25k because families are defined as being a level below household. Household income is equivalent to the sum of the family incomes for each family in the household[1]. This allows, for example, the existence of cases where two families each have family incomes below 25k which sum to a household income greater than 25k.

Consider four cases:

1. In single-family household, family income $<$ \$25,000, household income $<$ \$25,000
2. In single-family household, family income $<$ \$25,000, household income $\geq$ \$25,000
3. In multi-family household, family income $<$ \$25,000, household income $<$ \$25,000
4. In multi-family household, family income $<$ \$25,000, household income $\geq$ \$25,000

Household income being the sum of all family incomes for each family in a household means that there are no instances of the second type[2]. The difference between a household and a family is only relevant in the case of multifamily households.

[1] In years 2020, 2021, and 2022, there are 11, 13, and 17 observations, respectively, where this is not true. This makes for a total of 41 out of 367,781 observations.

[2] The singular individual in 2020 in a single family household with family income less than household income puzzles me. Household income is $239,860 and family income is $9,600, yet there is only one individual in the household. She is a widow in her 70s, listed as the householder of her house so I don't think it's a group-home situation. She is listed as being NILF - retired but has a listed wage/salary income of $220,000. My inclination is to ignore this one observation.

Table 1: Frequencies - Family income<$25,000 Universe

Persons	Year		
	2020	2021	2022
1. Single-family, hh<25k	12,883	14,559	13,224
2. Single-family, hh≥25k	1	0	0
3. Multi-family, hh<25k	1,216	1,484	1,266
4. Multi-family, hh≥25k	3,738	4,166	3,628
Total, hh<25k	14,099	16,043	14,491
Overall total	17,838	20,209	18,118

If you wanted to make categories based on family income only, cases 1, 3, and 4 (Overall total) would all be in a bin for income below 25k. If you based your categories on household income, the below-25k group would only hold cases 1 and 3 (Total, hh<25k).

2 Picking the level of analysis

If you're trying to pick a variable to use as a measure of income to which a person has access,[3] irrespective of their individual labor force participation status, family income may be more accurate than household income because of multi-family households. In those cases, household income may be more than twice what a particular family actually makes. We want to impute income information to the monthly surveys using the March supplement. We can switch the level of analysis after imputation to some extent - if we impute family income, we could decide to aggregate up to the household level.

3 Picking the level for imputation

'Level' refers to the type of variables we use in the imputation model - individual-specific information, family-specific information, etc. Individuals in the same family unit will have all of the same family-level characteristics so if we run an imputation model at the family level, all members of the same family will have the same predicted family income value. If we use a mix of family and individual level variables, members of the same family would have slightly different predicted values. It makes intuitive sense to use variables in the imputation model that are the same level as the variable you are trying to impute. We are choosing to impute "FTOTVAL" because we want to make use of the family-level variable "HEFAMINC"[4] in the imputation model. Note that there are no Census-provided markers for sub-household family units in the CPS monthly data - this requires me to generate a variable indicating to which family unit a person belongs within a household using variable "perrp". There is no need to collapse the data to the level at which you're imputing. Every family member will simply have the same family level information, so each of the individual observations will be treated the same.

[3] Imagine a non-working spouse - they nominally have an income of zero but that's not really how we should think of them.

[4] defined "family income - combined income of all family memvers during the last 12 months. Includes money from jobs, net income from business, farm or rent, pensions, dividends, interest, social security payments, and any other money income received by family members who are 15 years or older".

4 Building the model

When imputing it may be best to use a log-linear model. The left-hand side variable would be ln(ftotval) and predicted values generated by the model are ln(predicted ftotval) so I must exponentiate those values again to get predicted family income in dollars again. Since the LHS is a log, without intervention, only those with ftotval>0 can be taken in and used to predict and all predicted values will also be >0. This means our model will systematically over-predict family income because it is incapable of predicting a zero for family income but clearly we do have those cases in the ASEC data. Inability to incorporate those cases likely compromises the accuracy of the imputation. My solution is to replace ln(ftotval) with zero if ftotval≤0. This is in-keeping with Shigeru's practice of replacing all negative income values with zero in constructing h_tot_income_calc in the March memo. Doing this allows those observations to be part of the information set that the regression uses to predict income values, which also rectifies some of the issues of over-prediction at the lowest end of the income distribution.

Family-level imputation regression:

forvalues x = 2019/2022 {
regress lnftotval c.mean_age##i.max_educ c.mean_age2##i.min_educ c.share_elderly c.numadult##i.hefaminc i.hrhtype##c.share_married c.share_black##i.gestfips c.share_hisp c.share_immg c.kid_ratio c.share_under6 if is_asec==1 & year==‘x’
}

My methods of evaluating the quality of predictions are root mean squared errors and comparing the labor force participation rates by income group and by income quintile, using actual and predicted income. The second one should show us if the right people are being predicted into the right income ballparks - otherwise, the LFPRs will be distorted from the actual.

Table 2: RMSEs from family-level model

	Year		
	2020	2021	2022
ftotval	70,758	66,231	72,854
fam_agi	71,178	67,900	93,665
fearnval	72,497	69,457	75,642

These values are very high. Essentially, my predicted values are off from the actual by $70,000, on average.

Table 3: LFPR by actual family income

	Year		
% as of March	2020	2021	2022
<25k	33.9	31.7	31.8
25-50k	54.5	53.3	54.2
50-100k	64.5	64.6	65.0
100-150k	71.5	70.7	71.5
150k+	74.7	75.3	75.4
Overall	61.4	60.5	61.4

Notes: Sample is all participants 15+ from the March Supplements, dropping PEMLR==0. Weights used are marsupwt.

Table 4: LFPR by predicted family income

	Year		
% as of March	2020	2021	2022
<25k	48.2	48.0	48.2
25-50k	58.0	56.3	58.8
50-100k	61.5	60.8	61.4
100-150k	68.7	68.7	68.2
150k+	73.7	73.5	73.7
Overall	61.4	60.5	61.4

Notes: Sample is all participants 15+ from the March Supplements, dropping PEMLR==0. Weights used are marsupwt.

Using the fixed category definitions isn't sufficient here because we can't rule out the possibility that these differences are merely a result of predicted values falling just above or below a threshold, making the groups different in size and skewing the average. Here I show LFPRs by actual and predicted income quintile, so we know that the groups must be the same size:

Table 5: LFPR by actual family income quintile

	Year		
% as of March	2020	2021	2022
1	39.2	36.3	36.8
2	56.8	57.0	58.2
3	65.8	65.8	65.6
4	71.6	70.8	71.4
5	74.9	75.6	75.7
Overall	61.4	60.5	61.4

Notes: Sample is all participants 15+ from the March Supplements, dropping PEMLR==0. Weights used are marsupwt.

Table 6: LFPR by predicted family income quintile

	Year		
% as of March	2020	2021	2022
1	49.4	48.8	49.4
2	58.5	56.7	58.9
3	62.6	62.8	62.5
4	71.1	70.4	70.5
5	73.6	74.7	74.9
Overall	61.4	60.5	61.4

Notes: Sample is all participants 15+ from the March Supplements, dropping PEMLR==0. Weights used are marsupwt.

The differences in LFPRs for the predicted and actual income groups are telling me that people from the top and middle quintile are given predicted values that place them in the lowest and second quintiles and vice versa.

5 Contribution of each variable to the model's fit

Using only the 2022 ASEC, I show the current imputation model's fit and how omission of each term contributes to the adjusted R^2 statistic.
Full model: Adj $R^2 = 0.3041$
Omitting the variable or term listed below gives the corresponding R^2

- mean age x max educ : 0.3029
- (mean age)^2 x min educ : 0.3038
- # adults x hefaminc : 0.3029
- household type x share married : 0.3023
- share Black x state : 0.3036
- mean age : 0.3002
- max educ : 0.2989
- (mean age)^2 : 0.3019
- min educ : 0.3031
- share elderly : 0.3041
- number of adults : 0.2891
- hefaminc : 0.2401
- household type : 0.3023

- share married : 0.2995
- share Black : 0.3035
- state : 0.3020
- share hispanic : 0.3041
- share immigrant : 0.3025
- ratio of kids to adults : 0.3041
- share of kids under 6 : 0.3041