The Development and Use of a Consistent Income Measure for the General Social Survey

Ethan Ligon
NORC
University of Chicago

September, 1989
Revised June, 1994

GSS Methodological
Report No. 64

This research was done for the General Social Survey project directed by James A. Davis and Tom W. Smith. The project is funded by the National Science Foundation Grant SES-8747227.

Annual income is an important demographic variable for a wide variety of researches. Information on household income has been obtained in every General Social Survey (GSS) to date; information on respondent's income has been obtained in every GSS survey since 1974. This means that the cumulative GSS file contains a random national sample of household income data for 24,785 cases (net of item non-response) for 16 years -- a useful set of information.

Unfortunately, changes in the income categories across years and the changes in the nominal income distribution wrought by inflation have caused changes in the income variable over the years. The GSS currently has five variables to express household income (INCOME72, INCOME, INCOME77, INCOME82, and INCOME86) and four to express respondent's income (RINCOME, RINCOM77, RINCOM82, RINCOM86). Only INCOME and RINCOME contain income information for all relevant years, and the income categories used for these two years are quite coarse, and fail to use all the information gathered (in that groups of fine income categories are recoded into a single coarser category) by the GSS. None of the income variables takes into account changes in the value of the dollar due to inflation. Furthermore, all income variables are truncated -- respondents are not allowed to report a negative income, and the top category is open (e.g. greater than or equal to $50,000). This paper will describe the construction and proper use of two new GSS income measures, REALINC and RREALINC, which are subject to none of the aforementioned problems

Description of Real Income Measures

The GSS has income measures defined for two different entities: the respondent and the household. It is important to note the the income concept for these two different sorts of measures differ as well. Measures of household income attempt to measure income from all sources, while measures of respondent's income attempt to measure only the respondent's earnings from a single occupation[1] . The measures described in this paper, REALINC and RREALINC, carry on this tradition. REALINC is a measure of household income in constant (1986) dollars, and RREALINC is a measure of respondent's earnings in constant dollars. Because of the crudity of the underlying data, both income measures are expressed in hundreds of dollars. Expressing these income measures in hundreds of dollars still implies a false precision for higher levels of income -- for many purposes, the user would do well to round to the nearest thousand dollars.

Construction of REALINC[2]

The first issue that one confronts in constructing an income measure which is uniform over time is in trying to correct for changes in the price level. Since we have only categorical data, we could scale both ends of each category (save for the top category, since it has only one end) so as to have all category boundaries in constant dollars. The obvious problem with this approach is that after following this procedure for each of sixteen years, we will have a different set of categories for each year, varying in width and overlapping in all sorts of strange ways. This difficulty can be avoided by assigning a single number to each category, a number which we believe to be somehow representative of the population within that category. Obvious possible choices for such a number are the mean of the category,

[1] Compare the following two questions, for household income (INCOME86) and respondent's earning (RINCOM86) respectively:

INCOME86 In which of these groups idd your total family income, from all sources, fall last year--1988-- before taxes, that is. Just tell me the letter.

[Total income includes interest or dividends, rent, Social Security, other pensions, alimony or child support, unemployment compensation, public ais (welfare), armed forces or veteran's allotment.]

RINCOM86 In which of these groups did your earnings from (OOCUPATION IN Q. XX), for 1988 fall? That is, before taxes or other deductions. Just tell me the letter.

[2] The discussion which follows will refer to REALINC, the income measure for household income: however, construction of REALINC, responent's earnings, is precisely parallel to that REALINC, and applies equally to it.

the median, and the midpoint[3] . We don't have the data to find the median or the midpoint of each category directly (if we did, we wouldn't need these variables anyway), so we simply choose to use the midpoint of each category.

We can justify the choice of the midpoint as the measure of central tendency within each category by noting that the midpoint of each category seems to not be very different from the eman of the category. In order to test this claim, we calculated means and midpoints of income categories in the 1980 March Current Population Survey (see Table I). This table shows that the only large difference between mean and midpoint occur in the most extrememem categories. The persistent bias which apparently places the midpoint above the mean may be the result of rounding on the part of the respondent--note that if the respondent reound income to the nearest thousand dollars, this will bias the mean downward within a category for most categories.

It is easy to use the midpoint of a category as the measure of central tendency within that category if the category is closed. However, income data collected by the GSS has a top category which is open (e.g. greater than or equal to $50,000). Clearly, no midpoint is even defined for this category. To cope with this difficulty, we need to find some other measure of central tendency for the top catefory. Three possibilities present themselves. We could simply use the lower boundry of the category, or we could fit some distribution to the data, and impute the meanf rom this. Fortunatly, we are not without guidance' standard demographic practice[4] dictates that we fit a Pareto curve to the upper end of our distibution, and find the mean[5] of this interval.

---

[3] A less obvious choice we could make would be to fit some sort of curve to the income distribution. WE have used this approach in the estimation of the mean of the top (open) category (*infra*), but have rejected this approach for other categories because of its greater complexity, because the midpoint of a category is probably close to what we'd get with the more complex method (particularly for fine categories), and because it is far from clear what sort of curve ought to be fit to the distribution for any part of the distribution save the high end, where the Pareto distribution is a good fit.

[4] See Shryock & Siegel 1980, Miller 1966

[5] Parker and Fenwick (1983) present evidence that the median of the open-ended category is a more appropriate chioce. However, since we must impute incomes in the top category, computation of both the mean and median must be done indirectly. Derivation of a formula for the mean is straight-forward (*infra*), but derivation of a median appears not to be. The formula used by Parker & Fenwick is:

Median=10^(0.301/v)Xi

where Xi is the lower boundry of the open-ended category and v is calculated below. Just how the number 0.301 was arrived at is unclear; presumably it was estimated empirically. West (1985A) disputes Parker and Fenwick's claim that the median is a superior measure of central tendency, particularly for time series data.

Table I: Comparison of Means and Midpoints

Category	Midpoint	Mean
%Difference		
< $999	NA	-$361.59
NA		
1000-2999	$2000	2255.28
12.7%		
3000-3999	3500	3499.07
0.0%		
4000-4999	4500	4447.81
5.2%		
5000-5999	5500	5455.81
4.4%		
6000-6999	6500	6447.94
5.2%		
7000-7999	7500	7457.86
4.2%		
8000-9999	9000	8936.33
6.4%		
10000-124999	11250	11188.26
2.5%		
12500-14999	13750	13701.46
1.9%		
15000-17499	16250	16167.12
3.3%		
17500-19999	18750	18687.84
2.5%		
20000-22499	21250	21134.15
4.6%		
22500-24999	23750	23686.04
2.6%		
25000-34999	30000	29207.30
7.9%		
35000-49999	42500	40878.04
10.8%		
> 49999	NA	61504.88
NA		

The Pareto distribution was derived by Vilfredo Pareto as a good empirical approximation of the upper end of the income distribution. It is simple in form: Y = AxV, where Y is the number of people with an income greater than or equal to x, and A and v are parameters to be estimated. Given this, it is simple to derive a formula for the mean.

Let X be the lower limit of the open-ended category. Then Y is equal to the number of people with an income which places them in this top category. Since the Pareto curve only describes the upper end of the income distribution, we set some lower limit, say q. The total number of people we're interested in is then just the area under the distribution from q to infinity, that is:

$$Y = \int_q y\,dx$$

where x is the income level, and y = f(x), the number of people with an income of x. Y is then the cumulative number of people with an income greater than or equal to q.

Similarly,

$$\int_q xy\,dx$$

is the sum of the incomes of all those with an income greater than q; hence

$$\bar{x} = \frac{\int_q xy\,dx}{Y} = \frac{\int_q xy\,dx}{\int_q y\,dx}$$

is the mean income of those with an income greater than or equal to q.

We know, from the definition of the Pareto distribution, that

$$Y = Ax\text{^}(-v)$$

Hence, by the fundamental theorem of calculus,

$$y = \frac{dY}{dx} = \frac{-Av}{x\text{^}(v+1)}$$

Substituting this expression for y into the expression given above for x yields (all integrals are henceforth assumed to be evaluated from q to unless explicitly denoted otherwise):

$$\bar{x} = \frac{\int x(-Avx\text{^}(-v-1))\,dx}{\int(-Avx\text{^}(-v-1))\,dx} = x\left(\frac{v}{v-1}\right)$$

Using this expression to find the mean income of the top (open) category simply involves setting x to the lower bound of the open category, X.

Note that A cancels out of this expression, so that the only

parameter that we need to estimate is v. A number of different strategies have devised to perform this estimation. Quandt (1966) has examined a number of these methods[6] ; Likes (1969) derived the minimum variance unbiased estimator (MVUE). However, the methods used by Quandt and Likes do not lend themselves to situations where income data are collapsed, as they are in the GSS. The method most commonly used in such situations is the quartile method, or some variant of it. Koutrouvelis (1981) shows that the quartile estimator of v is consistent, and finds using Monte Carlo techniques that the quartile method yields estimates nearly as good as Like's MVUE when quartiles are optimally spaced.

Though the existing categorization of income data in the GSS is undoubtedly non-optimal, we have persisted in use of the quartile method because of its low computational cost and the pervasiveness of its use in the demographic and economic literature.

Estimation of v is straight-forward. Let

a = Log_{10} of the lower bound of the category preceding the top category.
b = Log_{10} of the lower bound of the top category.
c = Log_{10} of the sum of the frequencies in the top two categories.
d = Log_{10} of the frequency in the top category.

Then

$$v = \frac{c - d}{b - a}$$

is the quartile estimator of v.

So, for example, if we wanted to calculate the mean income of the top category of GSS respondents in 1989, we would calculate the following:

X_T = Number in top category = 30

X_{T-1} = Number in next category = 24

a = $\text{Log}_{10}(50{,}000) = 4.6991$

b = $\text{Log}_{10}(60{,}000) = 4.7782$

c = $\text{Log}_{10}(X_T + X_{T-1}) = 1.7324$

d = $\text{Log}_{10}(X_T) = 1.4771$

[6] n particular, Quandt compares a maximum likelihood method with a method of moments, and the quartile method. The last, of course, does lend itself to use with censored data.

$$v = \frac{c - d}{b - a} = 3.2275$$

$$\bar{x} = 60{,}000 \frac{v}{v - 1} = 86936.03$$

Mean of Income in the Bottom Category

Estimation of the mean of the bottom category presents some of the same difficulties that estimation of the mean of the top does. The Current Population Survey allows respondents to report net negative income, so that the bottom category is, for the census, unbounded just as the top category is. The GSS makes no provision for the reporting of negative income, and therefore the bottom category is *de facto* bounded. We therefore follow the same practice with the bottom category as with all other closed categories, and use the midpoint of the category as the measure of central tendency within the category. In fact, this is probably not a very accurate measure; both the mean and the median can generally be expected to lie above the midpoint of the lowest category[7] . For the sake of simplicity and to avoid *ad hoc* data manipulation, we have avoided any manipulation of the bottom category. Those interested in studying low-income families may wish to perform their own adjustments to the bottom category, or may find that the poverty measures POVLINE and INCDEF[8] are more appropriate for such a study.

After having derived measures of central tendency for each category in each year of the GSS, the next step is to scale all these measures to correct for inflation. We have used the Consumer Price Index (CPI) for this scaling with 1986 as the base year, not necessarily because it is the best index for this purpose, but becuase of its widespread use, and because it doesn't differ from much from other indices which might have been more suitable. The user should be aware, incomparing real income across years, that the CPI is commenly believed by economists to overstate increase in the cost of living by as much as two to three percent per year[9]; hence, as a measure of general economic wellbeing over time, REALINC is biased downward. See Table II for the precise figures used to weight the REALINC and RREALINC variables (note that since the GSS asks for income retrospectively, all reported incomes are for the year previous to the year of the interview; e.g. the 1972 GSS asked about 1971 income).

[7] <Cite imputed CPS data here>

[8] See Ligon (1988) for an explanation and discussion of these variables; see Davis and Smith (1989) for codebook information on these variables.

[9] <Find Cite here>

Table II:CPI Weights

Year	Weight
1971	2.707
1972	2.621
1973	2.467
1974	2.223
1975	2.037
1976	1.926
1977	1.809
1978	1.681
1979	1.511
1980	1.331
1981	1.206
1982	1.136
1983	1.101
1984	1.056
1985	1.019
1986	1.000
1987	0.965
1988	0.968

Comparison of CPS and GSS Income Measures

One can get a sense of how reliable the GSS income measures are by comparing some summarystatistics against similarstatistics derived from CPS data. Figure 1 compares mean household income calculated from GSS data to comparable figures derived from CPS data. The GSS time series is much less smooth, presumably due to its much smaller sample size, but generally appears to follow the CPS series quite well.

It is more difficult to compare GSS data on respondent's earnings

to CPS data, because there is no comparable measure collected by the census bureau. The principle problem is that the census bureau collects earning data for the "householder,"[10] while the GSS collects data for the respondent, and only collects such data if the respondent is employued. For this reason, figure 2 uses data collected by the Bureay of Labor Statistices, using data collected from employers[11].

REALINC and RREALINC essentially pretend to make a continuous variable out of categorical data. The time series presented above indicate that using the income yields figures for the GSS that compare well to CPS figures; however, we can illustate some of the pitfalls of making this pretense by comparing not the mean income figures, but rather median income figures. Time series for GSS and CPS median household income are presented in Figure 3. Note that they don't compare nearly as well as the mean figures given in Figure 1[12] . Changes in the nominal income distribution due to inflation are not controlled for by changes in the categories for each year; "spikes" in the distribution due to measurement error cause the median to be a very poor measure of central tendency. This same measurement error makes the calculation of the standard error of REALINC and RREALINC problematic, and this should be taken into account by the user when performing any tests of significance involving these variables.

Conclusion

Hitherto, it has been difficult to incorporate income data into research using the GSS across time because of inconsistencies in categorization and changes introduced by increases in the price level. The construction of the income and earnings measures REALINC and RREALINC should make such researches much more straightforward.

Three steps are involved in the construction of these measures: first, use of category midpoints as a measure of central tendency within each income category; second, calculation of the mean income in the top category through use of the Pareto distribution; and third, scaling income and earning data across years into constant (1986) dollars. Aggregate statistics generated by this procedure agree well with other data sources. However, the

[10] The person for whom earning data is collected by the census has changed somehwat in recent years. Prior to 19??, the census bureau used "head of household;" it now uses "householder." The former of these was determined in large part by the gender of the household members; the latter is less so.

[11] <Thing about data not including armed forces, self-employed, etc..

[12] West (1985B) has a good discussion on the pitfalls of using medians as a measure of central tendency for categorized data across time.

user should be aware of the issues involved in estimation of the standard errors of these variables; conventional estimators of the standard error (i.e. statistics which assume no measurement error in REALINC and RREALINC) will be biased and inconsistent.

Figures

Figure 1

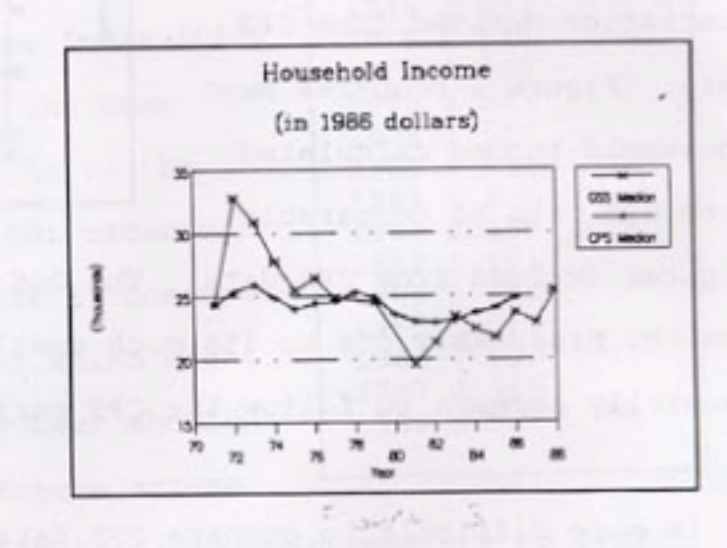

Figure 3

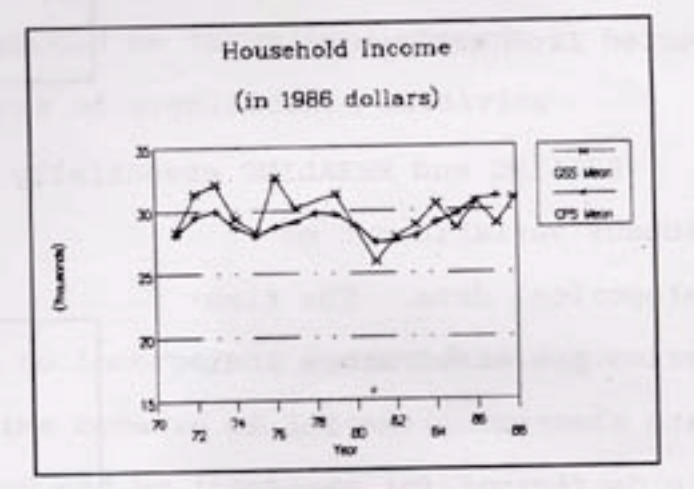

Figure 2 is unavailable

References

Bowman, M.J., "A Graphical Analysis of Personal Income Distribution in the U.S.," American Economics Review 35 607-628, 1945.

Koutrouvelis, I.A., "Large Sample Quantile Estimation in Pareto I Laws," Communications in Statistics: Theory and Methods A10 189-201, 1981.

Davis, James Allen and Smith, Tom W., General Social Surveys, 1972-1989: Cumulative Codebook, Chicago: National Opinion Research Center, 1989.

Ligon, Ethan, "Rationale and Construction of Poverty Measures in the General Social Survey," GSS Methodological Report No. 57, 1988.

Likes, J., "Minimum Variance Unbiased Estimates of the Parameters of Power-function and Pareto Distribution," Statistische Hefte 10 104-110, 1969.

Miller, Herman P., Income Distribution in the United States, Washington: U.S. Bureau of the Census, 1966.

Parker, R. N. & Fenwick, R., "The Pareto Curve and its Utility for Open-Ended Income Distributions in Survey Research," Social Forces 61 872-885, 1983.

Quandt, R.E., "Old and New Methods of Estimation of the Pareto Distribution," Metrika 10 55-82, 1966.

Shryock, H. & Siegel, Jacob, The Methods and Materials of Demography, Washington: U.S. Bureau of the Census, 1980.

U.S. Bureau of the Census, Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Accuracy of Data on Population Characteristics as Measured by Census-CPS match, Series ER 60, No. 5, Washington, U.S. Bureau of the Census, 1964.

West, Sandra A., "Estimation of the Mean from Censored Income Data," Bureau of Labor Statistics Report, 1985A.

West, Sandra A., "Standard Measures of Central Tendency for Censored Earnings Data from the Current Population Survey," Bureau

of Labor Statistics Report, 1985B.