WORKING PAPER

University of California, Berkeley

SRC

PHONE: (510) 643-6874
FAX: (510) 643-8292
website: http://ucdata.berkeley.edu/rsfcensus

Survey Research Center
2538 Channing Way
Berkeley, CA 94720-5100

Comments welcome

November 2004

Getting the Most Out of the GSS Income Measures

Michael Hout

University of California, Berkeley

GSS Methodological Report #101.

ACKNOWLEDGMENT: I am grateful to Henry Brady and Peter Marsden for comments and to the Russell Sage Foundation for financial support.

GETTING THE MOST OUT OF THE GSS INCOME MEASURES

The GSS monitors social change in the United States by applying the same methods consistently over time. The motto for the project, attributed to Otis Dudley Duncan, is "If you want to measure change, don't change the measure." Attitudes, in particular, are very sensitive to the exact wording of both the question and the answer options (Schuman and Presser 1981); the only way to be sure that observed changes are not an artifacts brought on by changes in questions or answer categories is to use the same questions and answer categories all the time.[1] Consistency works well with most demographic and behavioral variables, too, although respondents may not be as sensitive to the wording of questions about objective conditions.[2]

Income is an exception among the more demographic and behavioral variables because the meaning of its answer categories changes over time. Inflation devalues the answer categories for both GSS income questions – one about family income for all sources for all persons before taxes and the other specifically about the respondent's own earnings from her or his principal occupation last year (also before taxes).[3] This

---

[1] Even with perfectly consistent measurement, artifacts might creep into a study if a significant term were to change its meaning over time.

[2] People report their age more accurately if you ask them their birth date, marital status depends more on whether you offer "living together" as an explicit response option, ancestry probably depends on the countries included in the list of suggestions, race depends on whether the interviewer codes his or her impression or the respondent replies for him or herself (Saperstein 2004).

[3] The respondent's earnings were first asked in the 1974 GSS.

methodological report describes a way to adjust the income data to make year-to-year comparisons meaningful.

Inflation also weakens the higher income categories over time as one year's very high income becomes less distinctive as the value of the dollar erodes. In 1972, the top income bracket the GSS used was $30,000 and over. That bracket was clearly inadequate by 1982 because 16 percent of GSS families made more than $35,000; by 2002 62 percent of GSS families made $30,000 or over. To compensate, the GSS added higher brackets to the income card in 1977, 1982, 1986, 1991, and 1998.

Inflation is not the only complication researchers encounter in using the GSS income data. Researchers want income amounts not categories for most applications. The midpoints of the reporting intervals are appropriate scores for all but the top category. The top category has no upper limit; researchers have to figure out a reasonable value to assign to the open interval at the top of the income distribution.

GSS Methodological Report #64 discussed these issues and described how inflation-adjusted variables (identified as REALINC and REALRINC) were created (Ligon 1989). This methodological report updates #64 and provides instructions for how individual researchers can keep their GSS data up-to-date. I intend it to be a guidebook for users who want a standard approach to make the most out of the GSS income measures. It might also benefit instructors who want a ready reference for students.

My main goal is to describe how to transform the measures of family and personal income into inflation-adjusted variables that can be compared across the years of the GSS. In broad outline, the transformation has two steps: (1) turn categories into dollar amounts and (2) adjust those dollar amounts to remove artificial change attributable to rising wages and prices. To be helpful, I also offer a few other suggestions and present an example.

From Categories to Dollars

The GSS gathered incomes in one set of categories in 1972, another from 1973 to 1976, a third from 1977 to 1980, a fourth from 1982 to 1985, a fifth from 1986 to 1990, a

sixth from 1991 to 1996, and a seventh from 1998 to 2002. The same categories were used to gather information on both family income and personal earnings in each year. Table 1 displays those categories along with their midpoints.

Table 1 about here

The midpoints of the closed intervals are appropriate scores for those categories. The midpoint of the open-ended top category is undefined because the top category has no upper limit. Researchers solve the "top code" problem in a variety of ways. Common practices include adding either a constant amount, such as $10,000, or a constant percentage, say 30 percent, to the lower limit of the top category. A more rigorous approach involves extrapolating from the next-to-last category's midpoint using the frequencies of both the next-to-last and last (open-ended) categories a formula based on the Pareto curve. The formula is:

$$M_{top} = L_{top}\frac{V}{V-1} \qquad [1]$$

where $V = \dfrac{\ln(f_{top-1} + f_{top}) - \ln(f_{top})}{\ln(L_{top}) - \ln(L_{top-1})}$,

L_{top} is the lower limit of the top category, L_{top-1} is the lower limit of the category before the top one, f_{top} is the frequency in the top category, and f_{top-1} is the frequency in the category before the top one. The top codes shown in Table 1 were calculated using formula [1]. In most years, formula [1] implies a top code that is two or more times bigger than the midpoint of the last closed interval. Exploratory analyses with six dependent variables indicate that the Pareto curve results in top codes that consistently overpredict positive outcomes whether I use the dollar amount or the natural logarithm of the dollar amount as the predictor. This suggests that formula [1] results in dollar amounts that are too high. To compensate, I reduced the gap between the result for formula [1] and the lower limit of the top category (L_{top}) by half. The new formula is:

$$M^{*}_{top} = L_{top} + \tfrac{1}{2}\left(M_{top} - L_{top}\right) = \tfrac{1}{2}L_{top}\left(1 + \frac{V}{V-1}\right) \qquad [2]$$

where V is defined as before. Figures 1 and 2 show the patterns graphically.

Figures 1 and 2 about here

A researcher can never be certain about the top-code attributions. Experimenting with several approaches may be appropriate. If the income measure is an independent variable in a multivariate analysis, a dummy variable for top-code cases can give additional guidance as to which adjustment is most appropriate; you are looking for the adjustment that returns a statistically insignificant coefficient for the top-coded dummy variable. As top-coding affects more cases in the year before a new set of categories is introduced than in the year a set of categories debuts, it might be appropriate in many instances to include an interaction between the top-code dummy variable and time measured either as years since a revision or years until the next revision. I include an example below.

Removing the Influence of Inflation

The value of the dollar changes over time. Each year since 1972 a typical bundle of goods and services cost more than it did the year before. The Bureau of Labor Statistics tracks these changes and publishes the Consumer Price Index for Urban consumers (CPI-U) every month. The big factor in rising prices is inflation; the true value of the dollar is falling. But some price increases should be chalked up to improvements in the goods and services themselves. For example, cars are safer, more fuel-efficient and cleaner than they were in 1972. So comparing the price of a new car in 1972 with the price of a new car, comparably equipped, in 2002, combines the falling value of the dollar and the rising quality of the car. Economists and others have debated how best to separate falling dollar values from rising quality. The consensus index is the CPI-U research series – CPI-U-RS (http://www. bls.gov/cpi/cpirsdc.htm). The research series still shows the value of the dollar to be less in 2002 than it was in 1972, but by a smaller factor than the total price change recorded in the CPI-U. The BLS website norms the CPI-U-RS to equal 100 in December 1977. For this report, I changed the base year to 2000 on the assumption that more recent prices are more meaningful as a basis of

comparison than a reference to prices that prevailed 27 years ago. Table 2 shows both time series.[4] I extended the CPI-U-RS back in time using an experimental series described on the BLS website.

Table 2 about here

The GSS income questions refer to total family income and respondent's earnings in the calendar year before the interview, for example, the 2002 GSS inquired about income in 2001. Thus the appropriate CPI-U-RS is the one from the year before the survey.

To remove the impact of inflation from GSS income midpoints, divide each midpoint by the CPI-U(RS) for the year in question. For example the CPI-U-RS for 1990 is 0.783. For individuals and families interviewed in the 1991 GSS whose 1990 income fell in the $25,000-$29,999 category, we find out the equivalent purchasing power of that income in 2000 dollars by dividing $27,500 by .783 – $35,135. Similarly, because the CPI-U-RS for 1979 is much smaller – 0.455 – the equivalent purchasing power of an income in the $25,000-$29,999 category in the 1980 GSS is $27,500/.455 = $60,447. Figure 2 shows the inflation adjusted categories.

Using the Inflation-Adjusted Midpoints

For the most part, the numbers that result from coding the income intervals to midpoints and then adjusting the midpoints for inflation – "real-dollar incomes" for short – can be used just like any other continuous variable. The real-dollar incomes themselves might be used directly, or a transformation of them might be better. For example, researchers often transform income data by taking the natural logarithm of the real-dollar amount. This practice is based on the supposition that people respond to proportional changes in income, not absolute changes. For example, in a logistic regression analysis of the

[4]The REALINC and REALRINC variables in some GSS cumulative data files follow these procedures and apply the CPI-U normed to 1986.

relationship between social class identification and income, the log-odds of being middle or upper class rather than working or lower class might increase linearly or proportionally as income increases. If the log-odds of a identifying as middle class or higher increase by pretty much the same amount – call it β – as income rises from $8,000 to $16,000 to $24,000 to $32,000 ,etc., then no transformation is called for. But if, instead, the log-odds of identifying increase by some other constant – call it γ – as income doubles from $8,000 to $16,000 to $32,000 to $64,000, then log-income is more appropriate than absolute income for multivariate analyses.

Very low incomes pose a different kind of challenge. As Figure 3A shows, a logistic regression model that treats class identification in the 2002 GSS as a simple linear function of personal earnings the year before predicts too much middle- or upper-class identification at low earnings and too little middle- or upper-class identification at high earnings. The data are noisy at the low end because few people report $1-$9,999 earnings. The dots show the observed percentage middle or upper class for each income category; the line shows the expected percentages under the model.[5] To improve the model's fit, I simply treated everyone with less than $10,000 annual income as if their income was $10,000, i.e., for the purposes of the analysis I recoded incomes less than $10,000 to $10,000. I then added two dummy variables to the model – one for having no earnings in 2001 and one for having between $1 and $9,999. This three-parameter treatment of the income-class relationship improved the fit at the high end as well as in the low range directly affected by the transformation.[6]

Figure 3 about here

[5]Note that the y-axis is scaled to the logit scale; that is, each percentage on the y-axis is 1.2 points away from the adjacent ones on the logit scale.

[6] Even better fit could be obtained by adding other income – the difference between total family income and personal earnings – to the model.

Missing data

Some people refuse to answer questions about income. In the GSS 5 percent of respondents refuse to state their family income and 4 percent refuse to state their personal earnings. Researchers have to choose between leaving those cases out of their analysis and guessing the income of the missing cases. The Census Bureau employs sophisticated "hot-deck" methods to attribute incomes to people and families that lack valid data (see http://www.census.gov/ cps for details). I doubt that the hot-deck attributions are appropriate for the smaller samples gathered by the GSS. Even if they are appropriate in principle, they cannot be used because the public release of the GSS does not have enough geographic detail to use the hot-deck methods.

Alternatives to hot-deck methods involve plugging in averages or predicted values based on regression equations. SPSS programs of the early 1970s offered the option of substituting the mean for missing data. That proved to be a bad idea because cases with missing data were not typical, so their average was poorly approximated by the average among valid cases. Mean substitution also deflated the variance of the independent variable. At that time, standardized coefficients were more important than they are today, and deflating the variance of an independent variable deflates its standardized coefficient relative to the unstandardized coefficient.

Researchers who have a good idea of what determines earnings (or any other independent variable of interest) can use that information to predict values for the cases with missing data as long as at least one of the missing-data predictors is not a factor in explaining the dependent variable of interest. When using attributed data – whether it is a hot-deck attribution, an average of some kind, or a regression-based attribution – it is a good idea to also include a dummy variable equal to one for cases with attributed data and zero otherwise in the regression equations.

Conclusion

Income data in the GSS require slightly more preparation than most other objective variables. This report describes techniques for removing distortions due to inflation,

assigning top codes for open intervals, and hints about how to handle missing data. The example of how personal income correlates with subjective social class identification illustrates the main techniques.

References

Ligon, Ethan, "The Development and Use of a Consistent Income Measure for the General Social Survey," GSS Methodological Report No. 64. Chicago: NORC.

Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys.* New York: Academic Press.

Table 1
GSS Income Categories and Their Midpoints, Including Codes for the Open-Ended Top Category

1972			1973-1976		
Lower limit	Upper limit	Midpoint	Lower limit	Upper limit	Midpoint
$0	$1,999	$1,000	$0	$999	$500
$2,000	$3,999	$3,000	$1,000	$2,999	$2,000
$4,000	$5,999	$5,000	$3,000	$3,999	$3,500
$6,000	$7,999	$7,000	$4,000	$4,999	$4,500
$8,000	$9,999	$9,000	$5,000	$5,999	$5,500
$10,000	$12,499	$11,250	$6,000	$6,999	$6,500
$12,500	$14,999	$13,750	$7,000	$7,999	$7,500
$15,000	$17,499	$16,250	$8,000	$9,999	$9,000
$17,500	$19,999	$18,750	$10,000	$14,999	$12,500
$20,000	$24,999	$22,500	$15,000	$19,999	$17,500
$25,000	$29,999	$27,500	$20,000	$24,999	$22,500
$30,000		$37,500	$25,000		$31,250
1977-1980			1982-1985		
Lower limit	Upper limit	Midpoint	Lower limit	Upper limit	Midpoint
$0	$999	$500	$0	$999	$500
$1,000	$2,999	$2,000	$1,000	$2,999	$2,000
$3,000	$3,999	$3,500	$3,000	$3,999	$3,500
$4,000	$4,999	$4,500	$4,000	$4,999	$4,500
$5,000	$5,999	$5,500	$5,000	$5,999	$5,500
$6,000	$6,999	$6,500	$6,000	$6,999	$6,500
$7,000	$7,999	$7,500	$7,000	$7,999	$7,500
$8,000	$9,999	$9,000	$8,000	$9,999	$9,000
$10,000	$12,499	$11,250	$10,000	$12,499	$11,250
$12,500	$14,999	$13,750	$12,500	$14,999	$13,750
$15,000	$17,499	$16,250	$15,000	$17,499	$16,250
$17,500	$19,999	$18,750	$17,500	$19,999	$18,750
$20,000	$22,499	$21,250	$20,000	$22,499	$21,250
$22,500	$24,999	$23,750	$22,500	$24,999	$23,750
$25,000	$49,999	$37,500	$25,000	$34,999	$30,000
$50,000		$62,500	$35,000	$49,999	$42,500
			$50,000		$62,500
1986-1990			1991-1996		
Lower limit	Upper limit	Midpoint	Lower limit	Upper limit	Midpoint
$0	$999	$500	$0	$999	$500
$1,000	$2,999	$2,000	$1,000	$2,999	$2,000
$3,000	$3,999	$3,500	$3,000	$3,999	$3,500
$4,000	$4,999	$4,500	$4,000	$4,999	$4,500
$5,000	$5,999	$5,500	$5,000	$5,999	$5,500

$6,000	$6,999	$6,500	$6,000	$6,999	$6,500
$7,000	$7,999	$7,500	$7,000	$7,999	$7,500
$8,000	$9,999	$9,000	$8,000	$9,999	$9,000
$10,000	$12,499	$11,250	$10,000	$12,499	$11,250
$12,500	$14,999	$13,750	$12,500	$14,999	$13,750
$15,000	$17,499	$16,250	$15,000	$17,499	$16,250
$17,500	$19,999	$18,750	$17,500	$19,999	$18,750
$20,000	$22,499	$21,250	$20,000	$22,499	$21,250
$22,500	$24,999	$23,750	$22,500	$24,999	$23,750
$25,000	$29,999	$27,500	$25,000	$29,999	$27,500
$30,000	$34,999	$32,500	$30,000	$34,999	$32,500
$35,000	$39,999	$37,500	$35,000	$39,999	$37,500
$40,000	$49,999	$45,000	$40,000	$49,999	$45,000
$50,000	$59,999	$55,000	$50,000	$59,999	$55,000
$60,000		$75,000	$60,000	$74,999	$67,500
			$75,000		$93,750

1998-2002

Lower limit	Upper limit	Midpoint
$0	$999	$500
$1,000	$2,999	$2,000
$3,000	$3,999	$3,500
$4,000	$4,999	$4,500
$5,000	$5,999	$5,500
$6,000	$6,999	$6,500
$7,000	$7,999	$7,500
$8,000	$9,999	$9,000
$10,000	$12,499	$11,250
$12,500	$14,999	$13,750
$15,000	$17,499	$16,250
$17,500	$19,999	$18,750
$20,000	$22,499	$21,250
$22,500	$24,999	$23,750
$25,000	$29,999	$27,500
$30,000	$34,999	$32,500
$35,000	$39,999	$37,500
$40,000	$49,999	$45,000
$50,000	$59,999	$55,000
$60,000	$74,999	$67,500
$75,000	$89,999	$82,500
$90,000	$109,999	$100,000
$110,000		$137,500

Table 2
Comsumer Price index Research Series (CPI-U-RS), 1971-2003

	CPI-U-RS		*2000 as base year*	
Year	Year-end	Average	Year-end	Average
1971		68.2	--	27.2
1972		70.3	--	28.0
1973		74.7	--	29.8
1974		82.1	--	32.7
1975		88.9	--	35.5
1976		94.0	--	37.5
1977	100.0	100.0	39.5	39.9
1978	107.7	104.3	42.5	41.6
1979	119.2	114.1	47.0	45.5
1980	131.9	126.7	52.1	50.5
1981	142.8	138.6	56.4	55.3
1982	149.8	146.8	59.1	58.5
1983	155.3	152.9	61.3	61.0
1984	161.1	159.0	63.6	63.4
1985	166.9	164.3	65.9	65.5
1986	168.5	167.3	66.5	66.7
1987	175.3	173.0	69.2	69.0
1988	182.2	179.3	71.9	71.5
1989	189.9	187.0	74.9	74.6
1990	200.7	196.3	79.2	78.3
1991	205.4	203.4	81.1	81.1
1992	210.5	208.5	83.1	83.1
1993	215.3	213.7	85.0	85.2
1994	219.9	218.2	86.8	87.0
1995	224.9	223.5	88.8	89.1
1996	231.8	229.5	91.5	91.5
1997	235.4	234.4	92.9	93.5
1998	238.7	237.7	94.2	94.8
1999	245.2	242.7	96.8	96.8
2000	253.4	250.8	100.0	100.0
2001	257.3	257.8	101.5	102.8
2002	263.4	261.9	103.9	104.4

Source: http://www. bls.gov/cpi/cpirsdc.htm
Note: The GSS asks about income for the year before the survey, so the appropriate CPI-U-RS value is the one for the year before the survey.

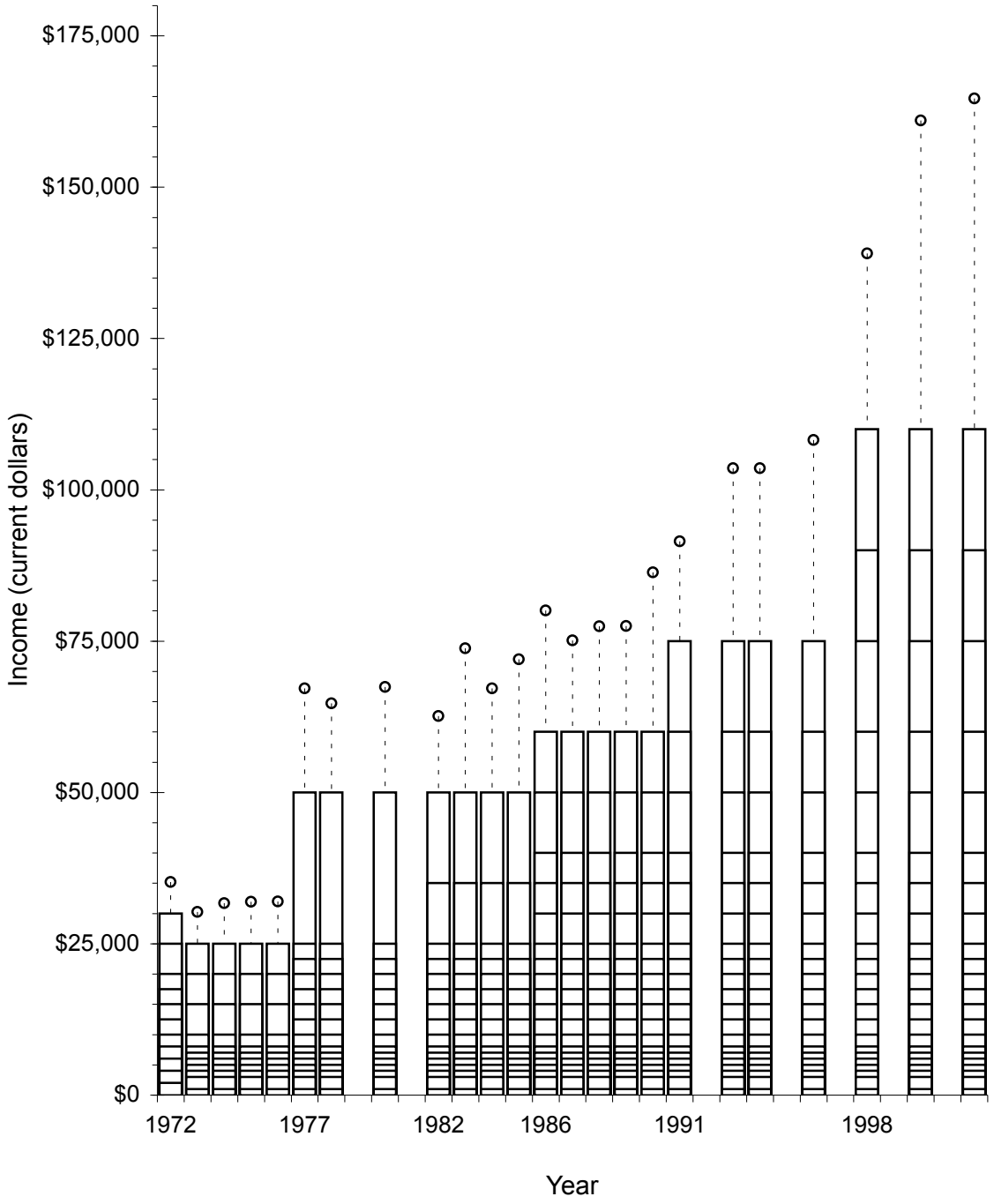


Figure 1. GSS Family Income Data by Year

Note: Top codes assigned by modified Pareto-distribution method (eq. 2); see text for details.

Constant dollars

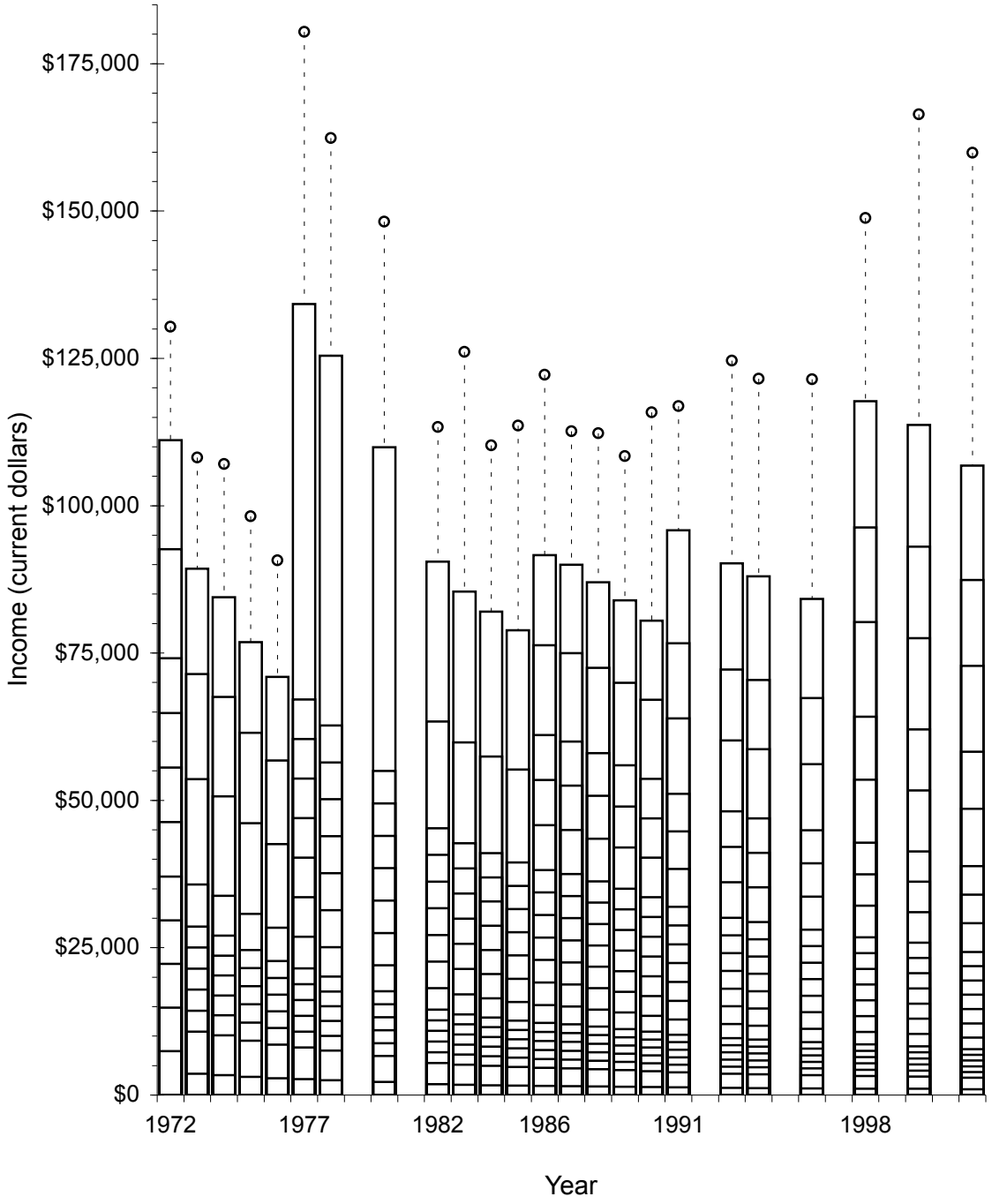
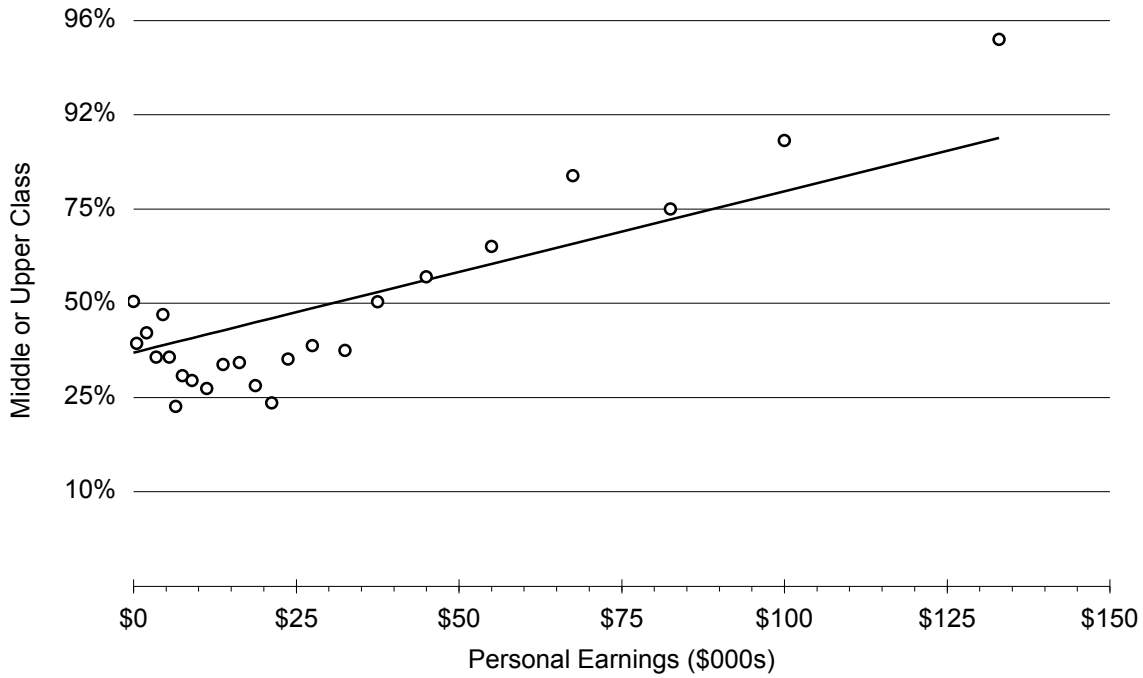


Figure 2. GSS Family Income Data by Year After Adjustment for Inflation

Note: Top codes assigned by modified Pareto-distribution method (eq. 2); see text for details. Incomes adjusted for inflation using CPI-U-RS.

A. Linear logit model



B. Logit model with modification for low incomes

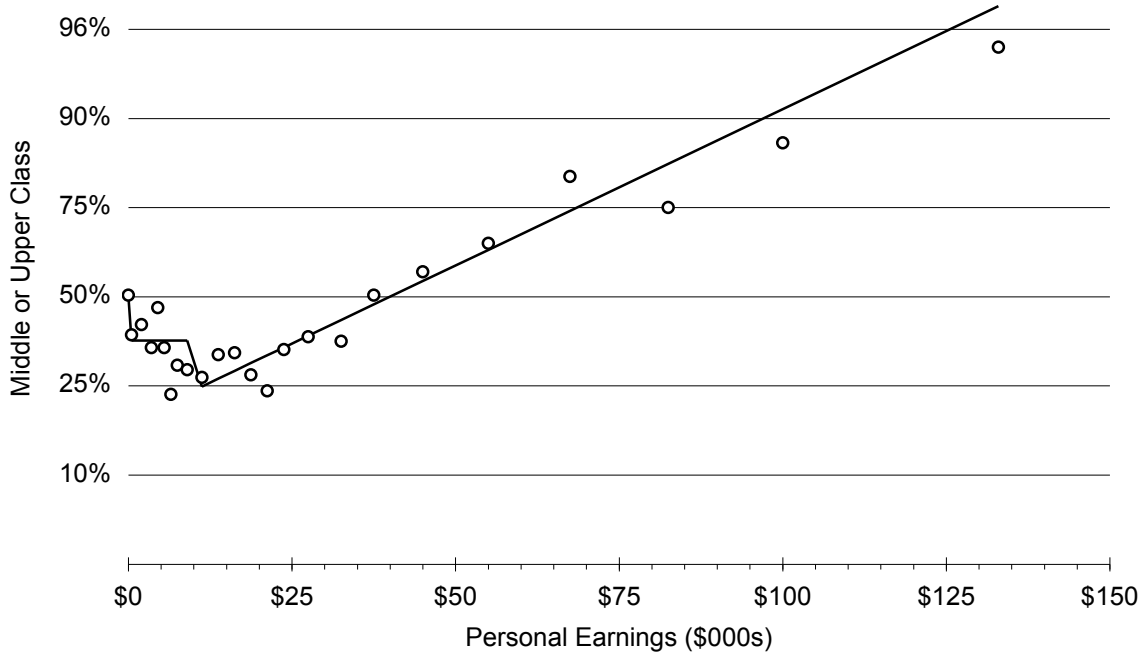


Figure 3. Percentage Identifying as Middle or Upper Class (vs. Working or Lower Class) by Personal Earnings: Observed Percentages and Percentages Expected from (A) Linear and (B) Modified Logit Models

Note: The modifications for low earnings are: (1) all earnings below $10,000 recoded to $10,000, (2) dummy variable for zero earnings, (3) dummy variable for $1-$9,999 in earnings.