PRELIMINARY: DO NOT CITE WITHOUT PERMISSION, COMMENTS WELCOME

REVISION I

Using Income Data to Predict Wealth

In memory of Daniel B. Radner

Arthur B. Kennickell
Senior Economist and Project Director
Survey of Consumer Finances
Board of Governors of the Federal Reserve System
Mail Stop 153
Washington, DC 20551
Voice: 202-452-2247
Fax: 202-452-5295
Email: akennickell@frb.gov

January 19, 1999

The author wishes to thank Jenny Wahl for comments, and Amber Lynn Lytle, Kevin Moore and Amy Stubbendick for outstanding research assistance in the work reported here. Barry Johnson has long been an essential player in the work underlying this paper, and he provided essential data used in the analysis. The opinions expressed here are the responsibility of the author alone and do not necessarily reflect those of the Board of Governors of the Federal Reserve System.

Most often, economists are interested in understanding household wealth as a reflection of past saving behavior. As a stock, wealth represents the cumulation of all past saving, transfers, and net shocks to income and consumption. The level of wealth implicitly reflects preferences about risk and intertemporal substitution, expectations about future income and expenses, life expectancy, family structure, institutional factors such as credit availability, and possibly more psychological factors such as cognitive abilities to make choices about the future and desire for autonomy or control. However, inherent in the nature of wealth, there is also a structural relationship between its value and investment returns, though the returns may be difficult to measure, irregularly distributed through time, or even conceptually ambiguous. This second type of functional relationship may be of interest to those who study portfolio allocations and to other who have a particular need to project wealth from a given pattern of income—for example, the Office of Tax Analysis in the Treasury where wealth is projected from income flows reported on tax returns, and at the Federal Reserve where wealth projected from income is a key factor in the sample design for the Survey of Consumer Finances (SCF).

This paper attempts to contribute to the understanding of the relationship between income and wealth using data from the SCF, the Individual Tax File (ITF) at the Statistics of Income Division of the IRS (SOI), and information from *Forbes* magazine about the wealth of the 400 wealthiest people in the U.S. Although the SCF sample over-represents wealthy households, it specifically excludes very prominent individuals, including members of the "*Forbes* 400," whose data might be impossible to protect sufficiently to include in a public dataset. The SCF is stratified by an index defined in terms of income flows which is intended to proxy for households' wealth. If this index is functioning as intended, one would expect that the *Forbes* group would have the very highest values of the index. However, examination of the 1998 SCF sample indicated that a number of people in the *Forbes* were apparently misclassified. A number of factors may explain this error. This paper is driven by, and legally made possible by, a need to understand this problem in order to refine the SCF sample design. Although the investigation is necessarily limited to areas that contribute in technical ways to the survey, it is hoped that the results will shed light on broader issues in the relationship between income and wealth.

The next section provides some background on the data sources used and goes into sufficient detail on the mechanics of the SCF list sample design to provide context for the analysis. The second section provides various summary indications of the relationship between income and wealth. The section also looks at the results of modeling wealth as a function of income and using several data sources. A final section provides a summary and points toward future research.

I. Data

This paper uses data from three sources: the SCF, the ITF, and *Forbes* magazine. The two principal analytical files are one containing linked SCF and ITF data and one containing linked *Forbes* and ITF data. Because the line of investigation must necessarily serve the technical needs of the SCF, the discussion below covers enough detail on the sample design to make apparent the motivation and limitations of the work presented here.

A. Individual Tax File

To create the ITF, every tax year the Statistics of Income Division (SOI) of the IRS selects a sample of all individual tax returns filed during the calendar year which are then specially edited to ensure a high degree of internal consistency.[1] The vast majority of these returns contain income data for the previous year, but the file may contain multiple amended returns for a given taxpayer, and returns for earlier years. The ITF sample is stratified by types of income received and other factors to yield a file that is heavily weighted toward observations with high income and unusual income characteristics. For 1997, the file includes about 126 thousand observations to represent about 121 million returns. The file includes returns filed from taxpayers outside the U.S. including foreign countries, U.S. territories, and APO addresses. The data presented in this paper derive from a subsample of the ITF for 1996 and 1993 including only the most recent return for filers in the U.S. where the age of the filer was at least 18. The file may include multiple returns for a given household—including separate filings for a married couple, filings by unrelated individuals, and filings for children. Generally, when the data are used to make population estimates in this paper, adjustments are made to the weights and data to compensate at least for the increased probability of

[1]See Statistics of Income [1992] and [1996].

selecting married-filing-separately returns and for that fact that the total income of the couple is reported over two returns.

B. Survey of Consumer Finances

The SCF is conducted by the Board of Governors of the Federal Reserve System in cooperation with SOI. Beginning with the 1983 survey, the first of the current series, the SCF has been conducted every three years. Since 1992, data for the surveys have been collected by the National Opinion Research Center at the University of Chicago (NORC). The analysis here uses data from the 1992 and 1995 surveys.

The SCF is widely known as a source for household level data on assets, liabilities, income, pension rights, use of financial services, and other factors related to the financial behavior of households.[2] Many items covered in the survey are narrowly held by relatively wealthy households, but others are broadly held across the whole population. To provide an adequate basis for the analysis of both types of items, the survey employs a dual frame sample including both an area-probability design (see Tourangeau et. al. [1993]), and a special list sample [see Kennickell [1998]) selected from the ITF to oversample wealthy households. To achieve the oversampling, the list sample is stratified by a proxy for wealth, a "wealth index," calculated using income data found on a tax return. Because of the time needed to process tax returns and to edit the ITF, the index must be estimated with income data from the ITF for the previous year.[3]

The wealth index is based on the combination of two separate indices. The first of these, "WINDEX0," derives from the idea of grossing up capital income flows using average rates of return: An ideal version of such an index would be given by $WINDEX0=\Sigma\ (1/r_i)\ Y_i$, where r_i is a rate of return and Y_i is a component of capital income.[4] A particular advantage of this approach is that because the rates of return are explicit, it is straightforward to update the model to compute

---

[2]For an overview of the 1995 SCF, see Kennickell, Starr-McCluer and Sundén [1997]. Other information about the survey, including the data, is available on the Internet at http://www.bog.frb.fed.us/pubs/oss/oss2/scfindex.html.

[3]For example, the 1998 SCF list sample used the 1997 ITF, which contains (almost entirely) income data for 1996.

[4]Greenwood [1983] discusses wealth estimates of this sort.

Figure 1: Definition of WINDEX0, 1998 SCF List Sample

WINDEX0 is defined as the sum of the following variables:

Taxable interest income
- Divided by 0.0750
- Rate on corporate bonds, seasoned issues, all industries
- December 1996 *Federal Reserve Bulletin*, table I.32, line 33

Non-taxable interest income
- Divided by 0.0538
- Rate on Aaa state and local notes and bonds
- December 1996 *Federal Reserve Bulletin*, table I.32, line 30

Dividend income
- Divided by 0.0201
- Dividend-price ratio, common stocks
- December 1996 *Federal Reserve Bulletin*, table I.32, line 39

Absolute value of rents and royalties
- Divided by 0.0692
- Assume follows effective mortgage yield
- December 1996 *Federal Reserve Bulletin*, table I.53, line 7

Absolute value of other types of business, farm, and estate income
- Divided by 0.0487
- Assume average of interest and dividend rates

Sum of absolute values of long term, short term, and other capital gains

Housing equity:
- Median housing value in the 1995 SCF by income groups:

Income ($ thou.)	*Median house value ($ thou.)*
under 60	30
60-120	125
120-250	188
250-1,000	350
1,000-5,000	750
5,000 or more	900

- Multiply by (156.9/152.4) to adjust for inflation (CPI)

Income data are taken from the 1996 ITF

WINDEX0 for the sample in any year. If all capital assets yielded a return that was constant across individuals, then this model would provide an exact measure of wealth. Unfortunately , some assets do not yield regular returns that are easily measurable—for example, principal residences. For assets like IRAs and 401k accounts, what is measured as income may depend on the measurement framework—for example, income from such income would appear on an IRS Form 1040 only when funds are withdrawn from the accounts. Work by Kennickell and McManus also provides evidence that individual returns vary considerably around the average. Moreover, wealthy individuals are often viewed as having a greater than average ability to "time" their receipt of income.

The version of WINDEX0 used in the 1998 sample is given in figure 1. Clearly, this model deviates from a pure rate of return model in several ways. An estimate of the equity in a principal residence computed by income groups in an earlier survey (and updated for inflation) is included, and a measure of the capital gains is included in an attempt to catch assets that might otherwise be missed. The ad hoc use of the absolute value function reflects a perception that many people who accept

Figure 2: Coefficients of WINDEX1, 1998 SCF List Sample

Have taxable interest		Log(long-term losses)	
Log(taxable interest)	*	Have short term losses	
Have nontaxable interest	+	Log(short term losses)	+
Log(nontaxable interest)	*	Have estate income	
Have dividends		Log(estate income)	
Log(dividends)	*	Have pension income	
Have gross Schedule C income		Log(pension income)	
Log(gross Schedule C income)		Have royalties	+
Have partnership/s-corp income		Log(royalties)	*
Log(partnership/s-corp income)	+	Have real estate tax deduction	*
Have Schedule C receipts	+	Log(real estate tax deduction)	*
Log(Schedule C receipts)	+	Have itemized deductions	+
Have negative Schedule C income		Log(itemized deductions)	
Log(negative Schedule C income)		Log(expanded income)	
Have schedule E income		Log(expanded income)**2	*
Log(schedule E income)		Have negative expanded income	+
Have farm income		Log(negative income)	*
Log(abs(farm income))		Filing status head of household	
Have negative farm income		Filing status single	
Log(negative farm income)		Filed from North-central region	
Have gross farm income		Filed from Southern region	
Log(gross farm income)	*	Filed from Western region	+
Have capital gains or losses		Log(age primary filer)	*
Log(abs(gains and losses))		Log(age primary filer)**2	*
Have capital losses		Intercept	*
Log(capital losses)			
Have long-term losses			

Adjusted $R^2 = 0.72$

+ indicates that the estimate is significant at the 5 percent level; * indicates that the estimate is significant at the 1 percent level.
Standard errors used in the significance test are corrected for multiple imputation
All dollar values are taken as absolute values with a floor of one.

losses are more like people who report positive returns than they are like those with little or no returns.

To allow for a more flexible relationship between income and wealth, a second index, "WINDEX1," is computed based on an estimated model first developed by Frankel and Kennickell [1995]. Such a model could, at least implicitly, capture some of the systematic variations in rates of return. Survey wealth measures for list sample respondents were merged with income data from the ITF as described later in the paper under highly controlled conditions designed to ensure that no other use could be made of the data. The end product is a regression of survey wealth on SOI income data. Figure 2 shows the variables used in the calculation of WINDEX1 for the 1998 SCF, where the

model was estimated using 1995 SCF wealth data and 1994 ITF income data.[5] Because the model must be estimated on earlier data and simulated on more current data, there is a risk that rates of return, tax laws affecting the definition of ITF income items, and other institutional factors may have changed in ways that could introduce systematic bias.

To hedge against the possibility of missing important relationships in WINDEX0, and of structural changes that might undermine the validity of WINDEX1, the SCF list sample is stratified by a combination of the two indices,

$$\text{WINDEXM} = \{[\text{WINDEX0}-\text{median}(\text{WINDEX0})]/\text{IQR}(\text{WINDEX0}) + [\text{WINDEX1}-\text{median}(\text{WINDEX1})]/\text{IQR}(\text{WINDEX1})\} / 2,$$

where IQR is the inter-quartile range (75^{th} percentile minus the 25^{th} percentile) of the argument distribution. Strata corresponding to higher values of WINDEXM are oversampled. The sample file is reviewed to exclude members of the "*Forbes* 400." These exclusions are justified by the fact that is it highly unlikely that any such people would agree to be interviewed, and their characteristics that would be collected in the survey are so rare that it would be impossible to disguise their identity to a sufficient degree that their data could be released.

C. Forbes Data

Since 1982, *Forbes* magazine has provided information on the wealth of the 400 wealthiest people in the U.S.[6] Forbes describes their estimates as "highly educated guesses," which are based on a variety of sources. In some cases, individuals provide information to the magazine, and those values are reviewed by their staff for plausibility. In other cases, publicly available information is used to generate an estimate. For this group, businesses and stock holdings represent the great majority of their wealth. Large publicly traded stock holdings are public information, and stock prices are determined in the market. In the case of non-traded business holdings, the compilers base their value estimates on cash flow, earnings, or sales using a variety of techniques appropriate to different types of businesses. Trusts, as they note, are a particular difficulty, and some error is undoubtedly

---

[5]The coefficient values cannot be shown for disclosure reasons. The model was estimated using 1,430 list sample case that had not experienced large changes in structure between 1993 and the time of the interview in 1995. In choosing this model, more complex models with interaction terms, and other features were considered but were rejected by the data.

[6]See Canterbury and Nosari [1985] and the October 13, 1997 issue of *Forbes*.

introduced in making assumptions about the functional ownership of such assets. Their estimates are reviewed by a panel of outside experts in a number of financial and business areas.

The data used in this paper derive mostly from the 1997 listing. To expand the coverage of the top of the distribution and to increase the uncertainty about precisely who is included in the calculations, some cases are taken also from the 1996 listing. In selecting cases for the analysis, several exclusions were applied. Most importantly, the wealth holding must be clearly associated with an individual or a married couple, not with a family. After these and a few other exclusions mostly intended to ensure comparability with the ITF sample, 310 observations remained.

D. Merged Data

Exact matches of information from the SCF and the ITF, and the *Forbes* data and the ITF underlie a key part of the work reported here. Great care is taken with both sets of matched data to ensure both that the data are secure and that the data are used only for the narrow purposes of statistical work geared toward the evaluation and improvement of the SCF sample design. All matched datasets are purged of identifiers, and access is restricted to only this author.

The merged file combining 1994 ITF data and 1995 SCF data contains 1,519 records with information on (largely) 1993 income and 1995 net worth. Households that experienced a change in marital status between 1993 and the time of the survey were excluded from most analyses.

As a rule, this author is forbidden to know the names of respondents selected for the survey. However, because the *Forbes* 400 members are specifically excluded, it was permissible to match *Forbes* wealth data and ITF data along with the computed wealth proxies, for the purpose of improving the SCF sample design. The merged file used here involves the 1997 ITF data (largely on 1996 income) and 1997 Forbes data on wealth.

II. Income and Wealth

Income and wealth are both treated as key indicators of well being, but these variables sometimes give quite different signals. Income (at least as it is usually measured) appears notably less concentrated than wealth: in 1995, the top one percent of the net worth distribution held 35.1 percent of total net worth but the top one percent of the income distribution received only 14.5 percent of total income (Kennickell and Woodburn [1997]). Income is also normally believed to be more

variable than wealth over short periods of time. For example, a fairly common problem in tabulating survey data by income categories is that some people who are quite "rich" in terms of wealth appear in the lowest income group. Most often, this combination of low income and high wealth signals a temporary disturbance of income. However, it may also signal the presence of a person with an unusual ability to manipulate her realized income and who does so in order to minimize income taxes. For some people wealth vary as they use their savings to buffer, while for others it may vary at a lower frequency to meet longer-term contingencies. The relationship between income and wealth is also strongly affected by life cycle effects: Overall, older working people have higher assets levels and income than younger people, but retired people tend to have higher wealth and lower income than younger people. Ultimately, the functional relationship between income and wealth is difficult to estimate: typically, a log-linear regression of wealth on income, age, and many other factors that are typically expected to explain the heterogeneity of wealth holdings will have an R^2 of only about 0.70.

Both income and wealth are highly skewed distributions, but wealth has more mass in the right tail of the distribution than does income. Figure 3 shows a plot of density estimates of income and net worth as measured by the 1995 SCF. The horizontal axis is scaled using a transformation with the convenient property that near zero it is close to linear, and farther away from zero it is close to logarithmic.[7] The figure shows clearly that the distribution of wealth is bimodal, with one mode centered at about zero and one at a higher value. In contrast to wealth, income has a unimodal distribution. There are obvious differences in the scales of the two distributions: median wealth ($56,400) is much higher than median income ($30,800), and wealth has a much larger range of variation (the standard deviation of wealth is over eight times that of income). Both distributions have a long thick right-hand tail.

To highlight the higher-order differences in the two distributions, figure 4 shows a quantile-difference (Q-D) plot, where each distribution has been "standardized" to have a median of zero and standard deviation of one.[8] If the adjusted distributions were identical, the plot would appear as a

---

[7]The transformation is the inverse hyperbolic sine, given by $\log\{\theta y + [\theta^2 y^2 + 1]^{1/2}\}/\theta$, with scale parameter θ of 0.0001. See Burbidge, Magee, and Robb (1988) for a discussion.

[8]A Q-D plot shows the numerical difference in two distributions at each percentile point of the distributions. It contains the same information as a quantile-quantile (Q-Q) plot rotated by 45 degrees. For this paper, the points of the Q-D plots are computed using the survey weights, and

horizontal line at zero. The actual plot is a roughly linear declining function over most of its range. Until the very top of the two adjusted distributions, the underlying income process tends to become relatively more skewed than net worth. At the very top, the pattern reverses dramatically as the right tail of the adjusted wealth distribution jumps far ahead of that of the income distribution.

Wealth come from cumulated saving from past income, where income is taken to include asset returns (including realized and unrealized capital gains) and transfers. If one could ignore population growth, then these distributions might be taken to represent a steady state of life cycle and other factors relating income and wealth over the whole population. Thus, one would expect to see a relatively fatter right-hand tail for income than wealth in the adjusted distribution since extraordinary income should be the driver of wealth growth. The fact that the relationships differ so strongly at the top of the two distributions could be taken to suggest that the income measurements may be missing very unusual returns, such as very large capital gains that are realized (and, thus, measured) only sporadically, or very large transfers.

To examine these relationships at the very top of the net worth distribution, figure 5 shows density estimates of *Forbes* measures of net worth and ITF measures of total income for the *Forbes* population, where each has been transformed using the same function as in figure 3.[9] Net worth in the full Forbes group varies from about $400 million to about $50 billion, and the group median is about $1 billion. Two facts are particularly salient: wealth for the *Forbes* group is highly skewed, and income is distributed more uniformly. Interestingly, even in this group, a Q-D plot of the standardized distributions (figure 6) shows a very similar pattern to that for the SCF sample.

---

then subjected to some minor smoothing. Note that the transformation of the distributions does not affect the general shape of the associated Q-D plot, only its vertical scale and location.

[9]Because it is a criminal offense to reveal even that a person filed a tax return on the basis of the ITF data, it is necessary to obscure the connection between the samples included in the income and wealth distributions displayed. In figure 5, the income plot includes only the 310 cases described earlier, but the wealth plot includes all Forbes cases except those where the wealth is not clearly attributed to an individual or couple.

Figure 3: Densities of Net Worth and Income, 1995 SCF

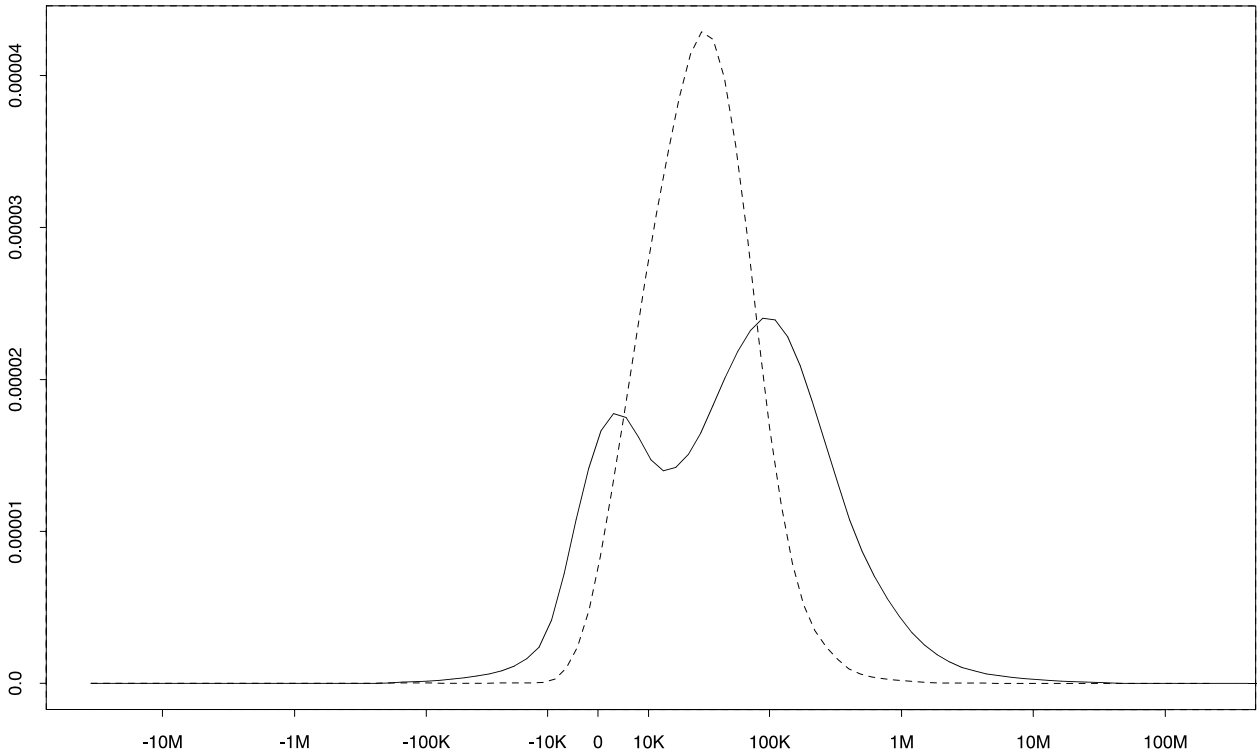


Figure 4: Q-D Plot of Net Worth Minus Income, 1995 SCF

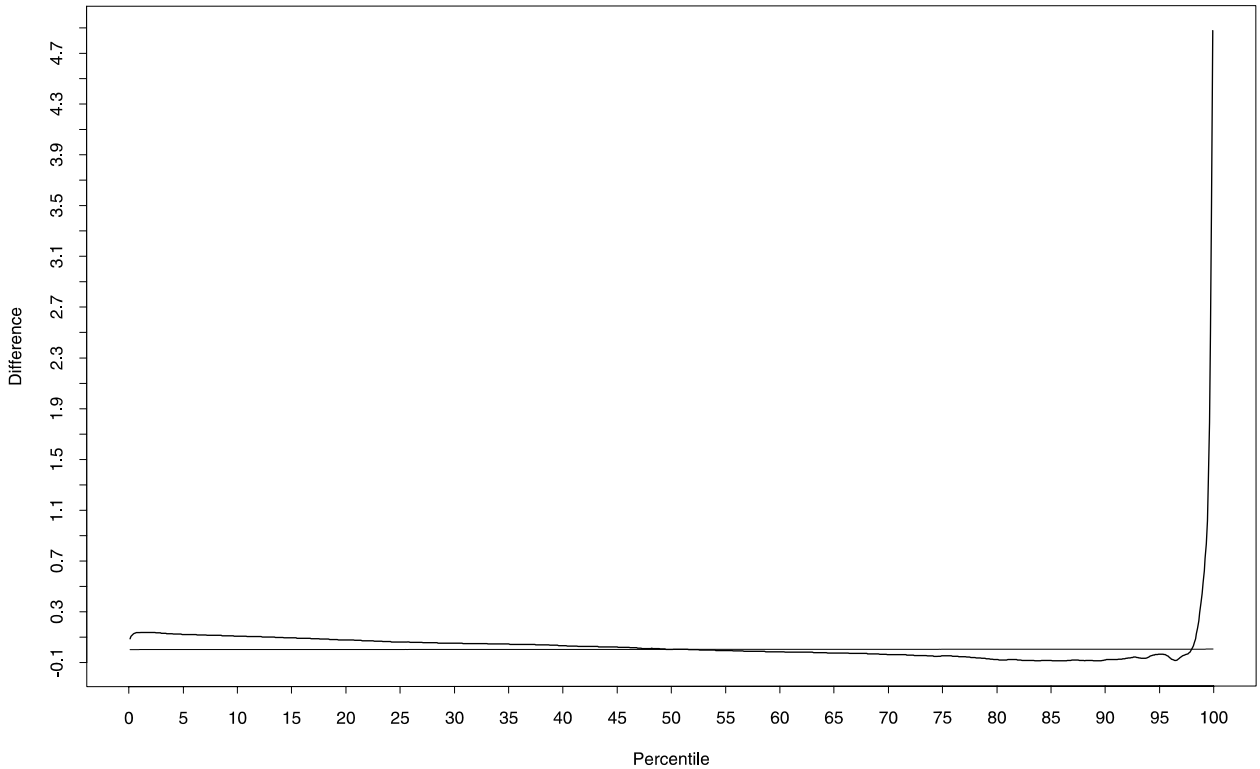


Figure 5: Densities of Net Worth & Income, 1997 *Forbes* Net Worth and 1997 ITF Income

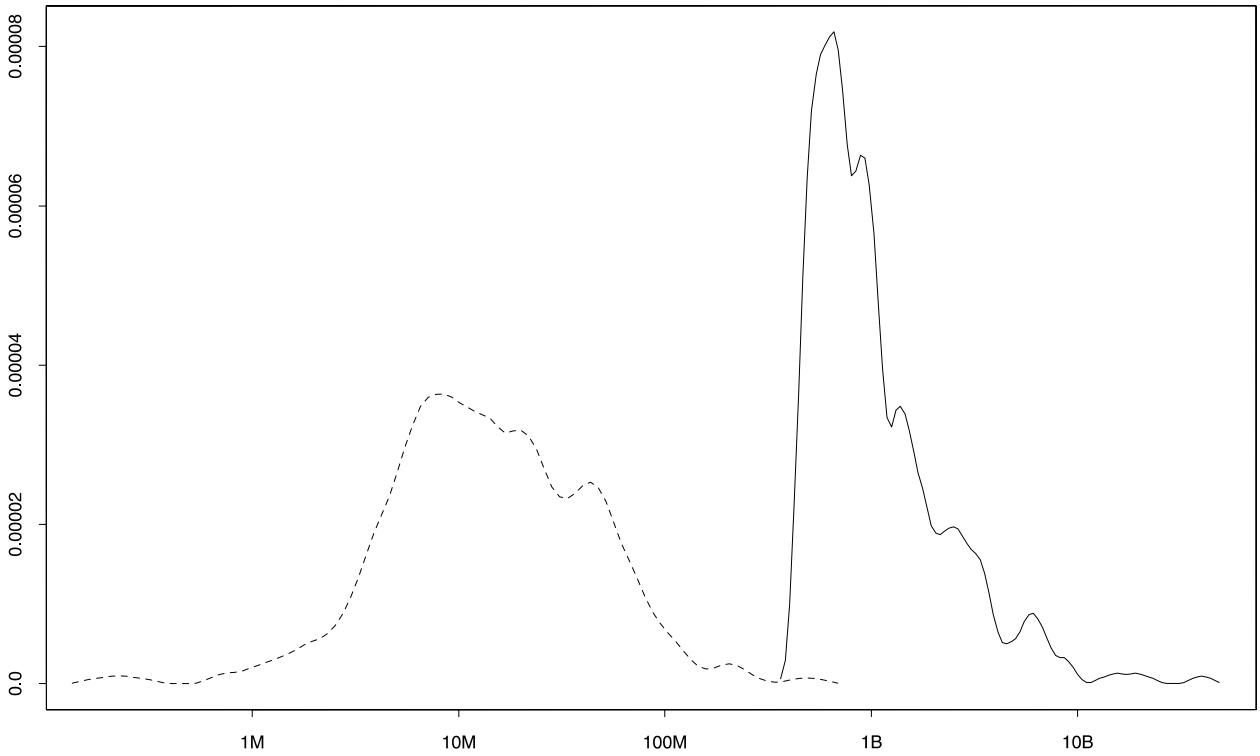
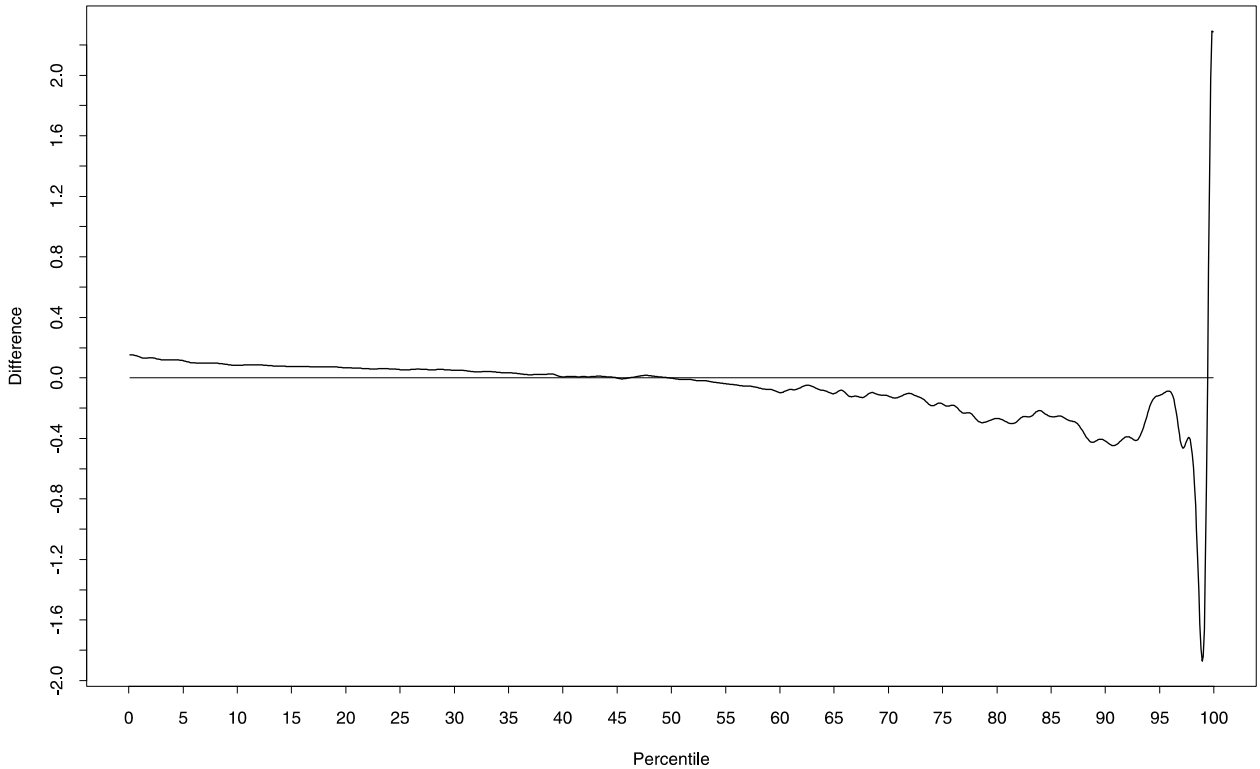


Figure 6: Q-D Plot of New Worth - Income, 1997 *Forbes* Net Worth and 1997 ITF Income



Although comparison of the marginal distributions does give one insight into the underlying structures, this approach tells us nothing about the covariation of income and wealth. Figure 7a shows a summary plot of the distribution of total 1994 household income across wealth classes in the SCF.[10] The horizontal axis is given in unweighted percentiles of the net worth distribution. This choice of axis ensures that an equal number of observations are represented within each interval. The make a connection to wealth levels, the figure also displays the values corresponding to the unweighted decile break points. At each point in the unweighted wealth distribution, the figure gives the weighted 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of the income distribution. To display meaningful variation within the limited space, the vertical axis is scaled in base-10 logarithms. In the log scale, the relative spread of the quantiles of income around the median at each point is approximately symmetrical and constant except up to about the top 20 percent of the wealth groups, suggesting that the conditional distribution of income given wealth may be approximately lognormal. Among the top 20 percent, the pattern is less clear, but it appears that there is an overall increase in skewness in log terms.

As expected, there is substantial movement of the center of the income distribution over wealth groups. Among the bottom 10 percent of the wealth groups—those with negative or minimal net worth—the whole income distribution declines as wealth increases from substantially negative values to the range nearer zero, and income reaches a trough at about the 10^{th} percentile. Households with large negative net worth are not necessarily poor in every sense. Between about the 20^{th} and 80^{th} percentiles of wealth, the scaling of the dollar equivalents on the horizontal scale is approximately logarithmic. Within that interval, the income quantiles rise approximately linearly, suggesting that wealth in that region might reasonably be predicted as a log-linear function of income. At the top of the distribution of wealth, the quantiles of income rises steeply, though in that range of wealth, the horizontal scaling is substantially more compressed in dollar terms. Although one cannot infer

[10]Given the subject of this paper, it might seem more natural to present the distribution of wealth given income. Unfortunately, when extended to analysis of the Forbes population, such an approach would have the consequence of revealing too much information about which individuals are included in the analysis.

directly the distribution of wealth given income from this figure, it is clear that it would be substantially more diffuse than that of income given wealth.

An obvious question is the role of temporary income fluctuations in explaining the variability of income by wealth groups. The SCF contains a question that asks respondents for their "normal" level of income. When this variable replaces actual income, the result (shown in figure 7b) is little different except among the top wealth groups for whom the increased skewness of income for the upper wealth percentiles in figure 7a disappears.

The *Forbes* data allow us to examine the income-wealth relationship at the very top of the wealth distribution. Using the matched *Forbes* wealth data and ITF income data, figure 8 displays information comparable to that in figure 7. The net worth percentiles on the horizontal axis correspond to the ordering of the 310 observations included. The net worth values corresponding to the percentile labels have been suppressed to blur the ability to identify specific individuals at a given point. It is remarkable how little variation there is in the level of the income quantiles over wealth groups. The median income ranges from about $8 million to about $30 million, and there is similar proportional variation in the other income quantiles. This result stands in contrast to the impression one gets from the SCF data of increasingly rapid increases in income with net worth at the top end of the wealth distribution. Some of the difference may be explained by possible differences in the effective definitions of income in the SCF and the ITF, though in principle there should be little difference since SCF respondents are asked to report the same income items that appear on an IRS Form 1040. However, the result may simply indicate that very wealthy people try quite hard to minimize their income.

Figure 7a: Distribution of Income by Percentiles of Net Worth, SCF

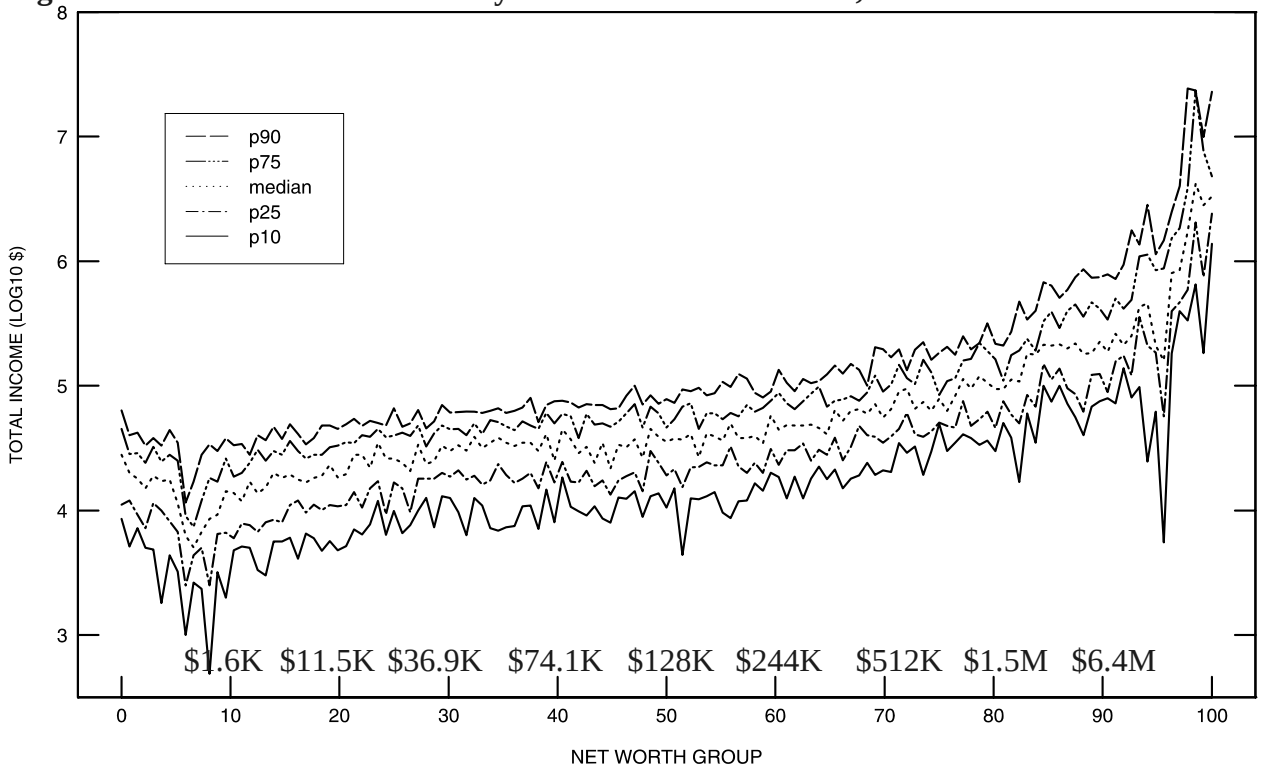


Figure 7b: Distribution of "Normal" Income by Percentiles of Net Worth, 1995 SCF

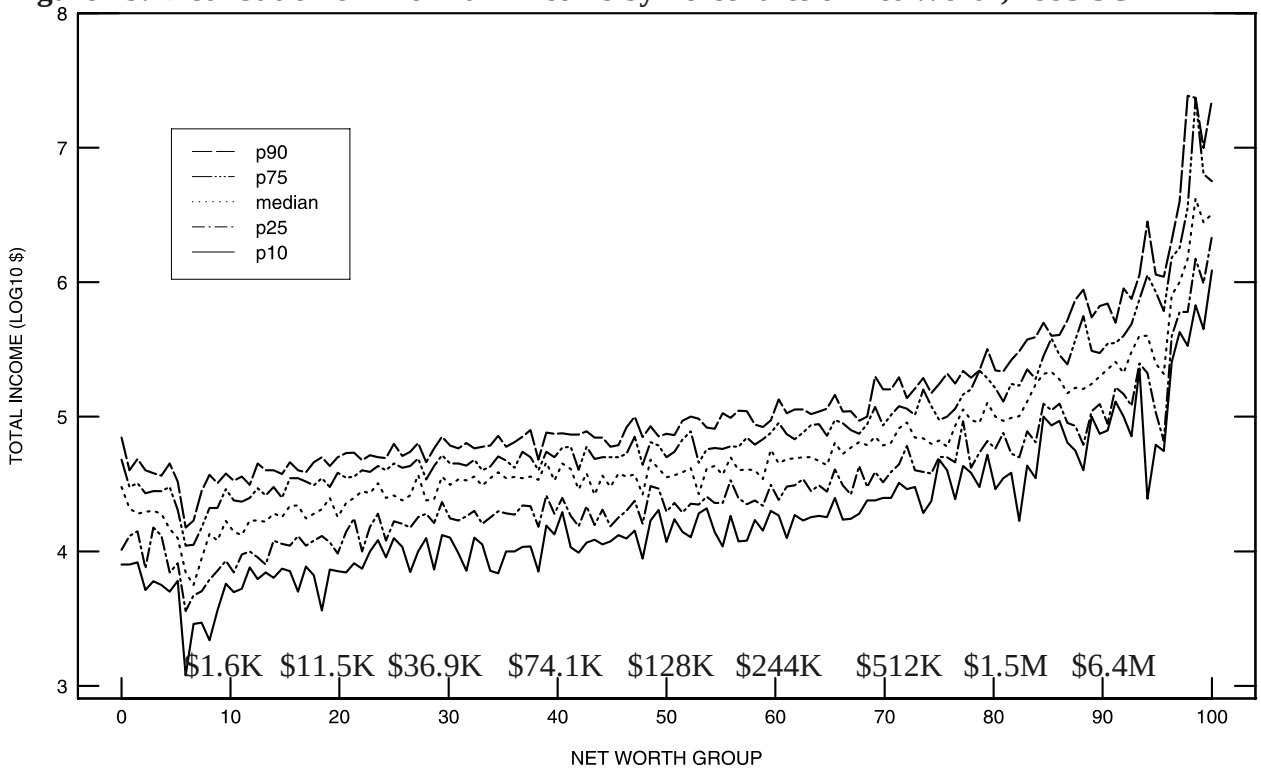
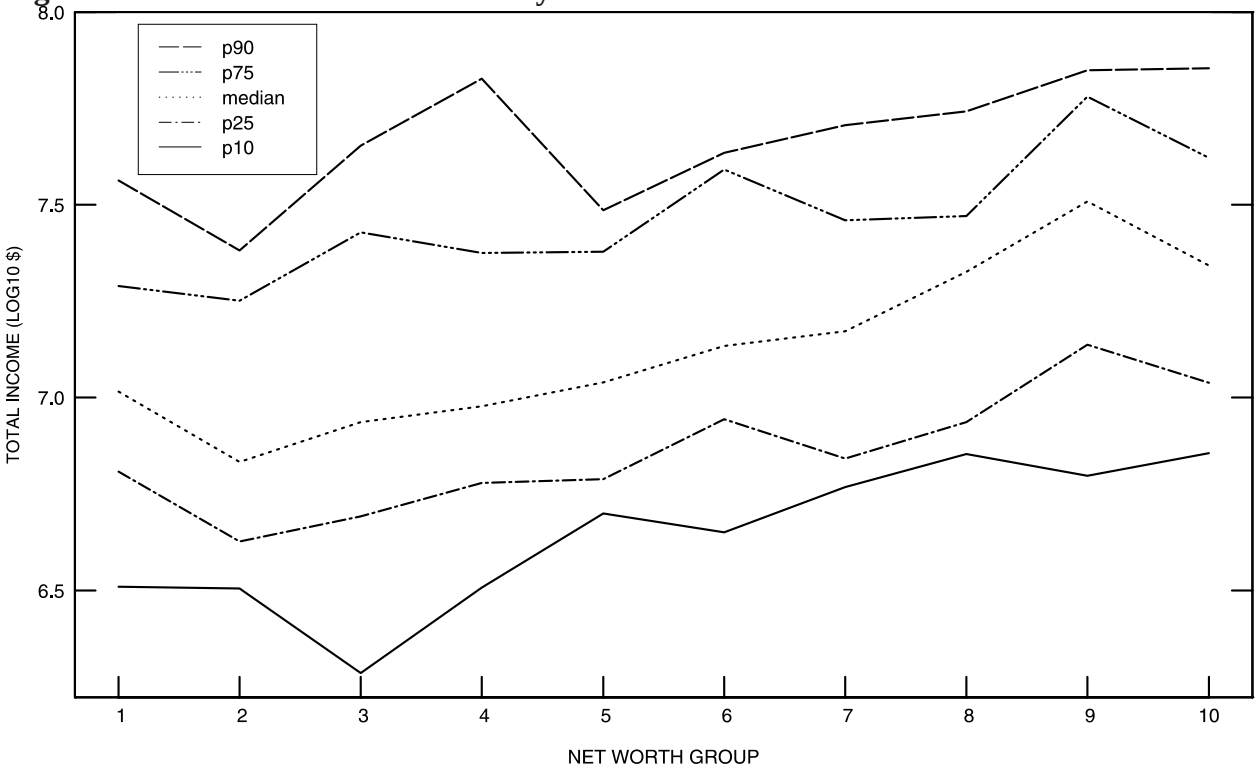


Figure 8: Distribution of ITF Income by Deciles of *Forbes* Net Worth



Clearly, the relationship between income and wealth is much more complex than can be seen in a simple bivariate distribution. Life cycles effects, precautionary saving and other risk-motivated behavior, permanent income, job loss, inheritances, and many other economic and preference factor are at the heart of the relationship. In some narrow applications, one has a variety of income components along with much more limited data on demographic and other factors, but one needs to make an estimate of the wealth associated with the income—the SCF sample design, the estimation of wealth data for tax simulation models, and the examination of asset and return preferences given income data are a few of the applications of this sort that come to mind. A very useful question for problems of this sort is how well can we do in terms of modeling wealth in terms of detailed income measures.

The WINDEX0, WINDEX1, and WINDEXM models described earlier in the paper are one such approach. These models use ITF income measures to estimate wealth for SCF sampling. At the completion of the 1995 SCF, it was possible to evaluate just how well these models performed in terms of classifying households by their wealth. Figure 9 contains a summary of the performance of these models. For comparison, it also contains the same information for total ITF income. The

graph contains average shifted histogram (ASH) estimates of the densities of each of the variables by net worth groups. Each of the horizontal panels in the figure contains an unweighted decile of the net worth distribution for the 1,519 list sample cases.[11] To remove irrelevant location and scale differences among the indices, the horizontal axis for the distributions is given on a percentile basis. Ideally, one would like to see narrow distributions centered around a diagonal from the lowest decile to the highest.

The figure shows that all of the indices do a fairly good job of distinguishing the very wealthiest groups and the bottom groups, though there is considerable spread even for these groups. In between, there is a positive association between the indices and wealth, but the distributions are fairly dispersed. Some of the dispersion may be accounted for by households that changed composition between their filing of a 1993 tax return and their participation in the 1995 SCF. Temporary fluctuations in income are doubtlessly also important (see Kennickell and McManus [1993]), and the variability of net worth due to imputation (see Kennickell [1991]) is also a contributing factor. Looking at the relative performance of the indices, WINDEX1 is somewhat more peaked on average than WINDEX0, but the differences are not very large. This fact is surprising, given that when the merged data are used to actually estimate the rates of return for WINDEX0 using least squares (or even robust models), the estimates are significantly different from the values assumed in constructing the index.[12] From comparison of the distributions of the indices with those of total income, it is clear that all the models add some refinement over total income alone. The Pearson rank correlations between net worth and the income-based measures are: total ITF income, 0.71; WINDEX0, 0.79; WINDEX1, 0.85; WINDEXM, 0.84.[13].

One might expect there to be a high level of variability in the predictive power of WINDEX1: since for this exercise it is based on coefficients estimated using income data from the 1991 ITF and

---

[11]The unweighted decile points of the net worth distribution in the list sample are about: 10th percentile, $82 thousand; 20th percentile, $280 thousand; 30th percentile, $560 thousand; 40th percentile, $1.0 million; median, $1.7 million; 60th percentile, $2.9 million; 70th percentile, $5.4 million; 80th percentile, $11 million; and 90th percentile, $28 million.

[12]Only the estimated coefficient on taxable interest—14.6, implying a rate of return of about 6.9 percent—is similar to the value used in computing WINDEX0. Some others are negative or too small to be meaningful.

[13]The Spearman correlation using income and wealth data from the SCF is only 0.76.

wealth data from the 1992 SCF, which are applied to 1993 tax data to predict wealth in the 1995 survey. The earlier structure of rates of return is implicitly imbedded in the model coefficients for WINDEX1, and other underlying relationships may have changed in important ways by 1995. However, it is noteworthy that even when the models are reestimated using 1995 SCF wealth data and then resimulated, the results do not change notably.[14]

Using the merged *Forbes*-ITF file, it is possible to evaluate how well the models estimated for the 1998 SCF sample do in terms of classifying the extreme right tail of the wealth distribution. To evaluate the performance of the indices, figures 11a-11d show rank in net worth as a function of the rank of total taxable income (shown for reference), WINDEX0, WINDEX1, and WINDEXM respectively. To protect the privacy of taxpayer information, the values shown have been randomly disturbed; however, this blurring of the data does not make any important differences in the overall interpretation of the results. The solid line in each plot is a loess (local least squares) fit of net worth rank in terms of the income or index rank. Although the points in the graphs are very widely scattered throughout the figure, the loess line suggest there is at least a positive association. The Spearman rank correlations are 0.34 for income, 0.36 for WINDEX0, 0.25 for WINDEX1, and 0.35 for WINDEXM, levels which are substantially below those for the SCF sample.

The deviations of the income and index ranks from the net worth rank may reflect important omitted variables, noise inherent in the use of a single period of income, or structural differences in the relationship of the observed variables for this population. Most likely, the truth is a combination of all three. It is difficult to conceive of a meaningful test for omitted variables in this context, and at this point it is not feasible to look at multiple years of income. Because of *Forbes*-ITF sample is relatively small, it is not possible to check the stability of the coefficients by doing the same sort of detailed modeling that underlies the estimation of WINDEX1. However, it is still possible to decompose classification differences in terms of the dummy variables and age variables included in the model for WINDEX1 shown in figure 2. When the net worth rank minus the corresponding income or index rank is regressed against these dummy variables, very little is significant according

---

[14]See Kennickell [1998] for more details. The Spearman correlation of 1995 wealth and WINDEX1 estimated using the 1995 wealth data is 0.83, which is marginally lower than the comparable figure for the original WINDEX1.

to the standard significance tests. Only in the WINDEX1 rank difference model is anything significant, and there it is only the coefficients on presence of Schedule C income, presence of Schedule F income, and residence in the north-central region of the country. Other tests for structural difference could well yield different results, but the absence of difference at this level makes me skeptical that there are clear systematic differences.

Given the relatively weak performance of the wealth indices in terms of predicting relative wealth levels with the SCF list sample and *Forbes* sample, one would expect a similar problem in discrimination between the two groups—the issue that originally motivated this paper. To examine this proposition, figure 12 provides unweighted ASH plots of the distribution of WINDEX1 in each of the two populations.[15] Despite the selection process that generated the *Forbes* sample used here, the group is still approximately self-weighting. However, because the SCF list sample is a stratified sample with a high rate of oversampling among families with high levels of the index, the relative density of the plot is distorted: there is far too much mass in the right tail relative to what would be found in the full population. Nonetheless, the figure still provides insight into how well the wealth index distinguishes between cases at the *Forbes* 400 level and other less wealthy cases. If the model were performing without error, the two densities would not overlap at all. In fact, about the top quarter of the top of the list sample overlaps with about bottom two-thirds of the *Forbes* group. An important goal for the future is to achieve a greater separation between these two distributions. Work toward this end will focus on the use of multiple years of income data.

[15]The behavior of WINDEX0 and WINDEXM is similar. The WINDEX1 plot is chosen to highlight the most flexible part of wealth modeling exercise.

Figure 10: Densities of WINDEX0, WINDEX1, WINDEXM and Income, by Deciles of Net Worth

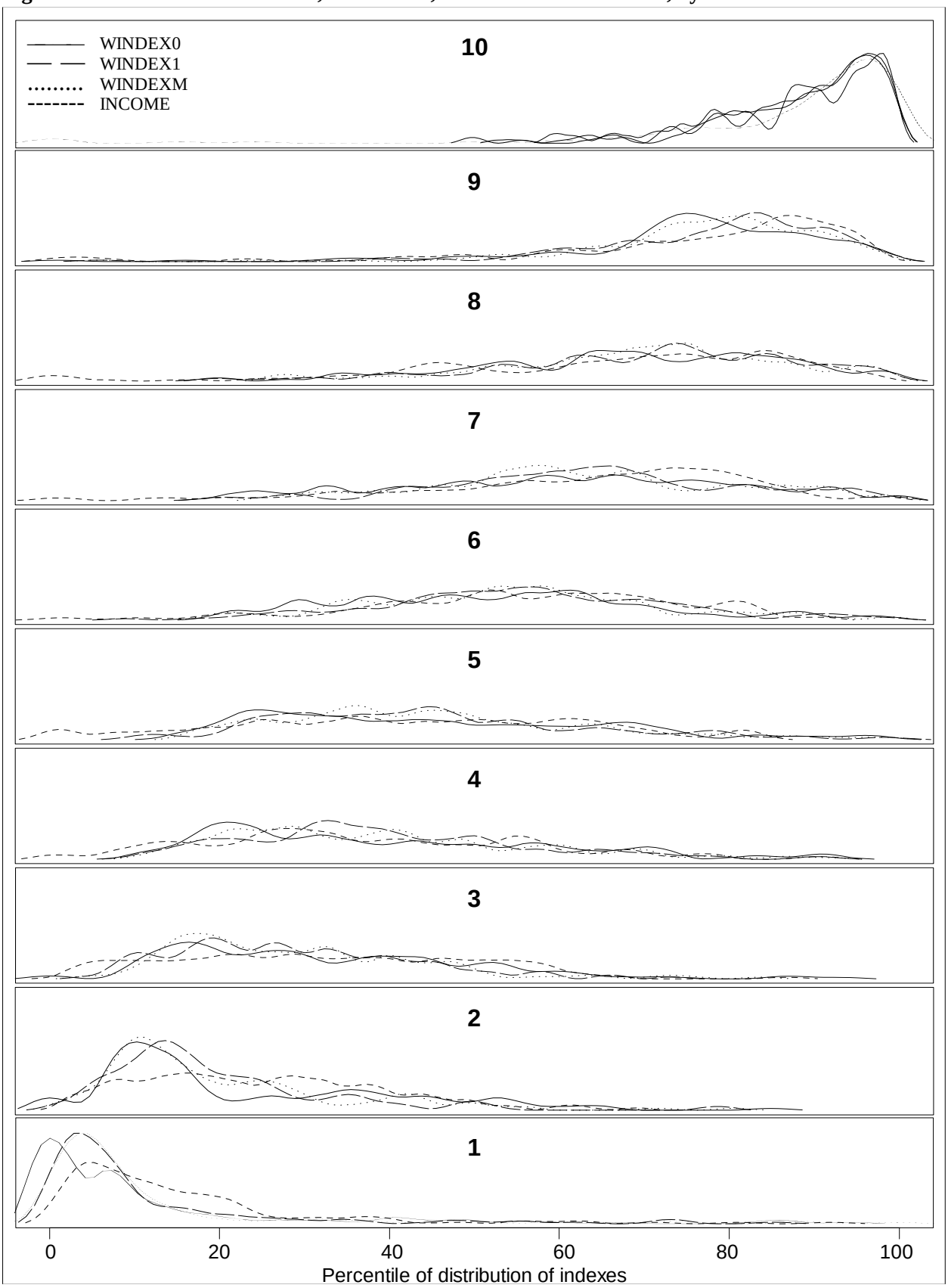


Figure 11a: Rank of Net Worth by Rank of Total Income, *Forbes* Sample

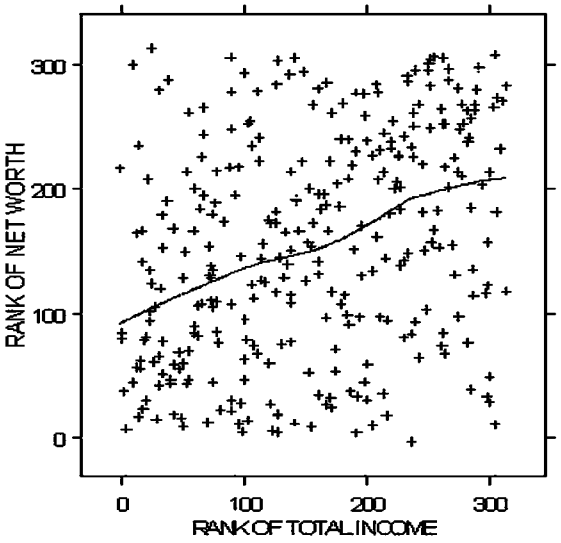


Figure 11b: Rank of Net Worth by Rank of WINDEX0, *Forbes* Sample

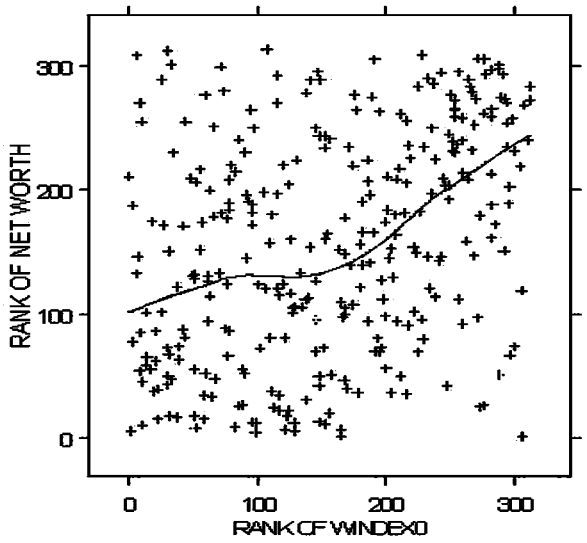


Figure 11c: Rank of Net Worth by Rank of WINDEX1, *Forbes* Sample

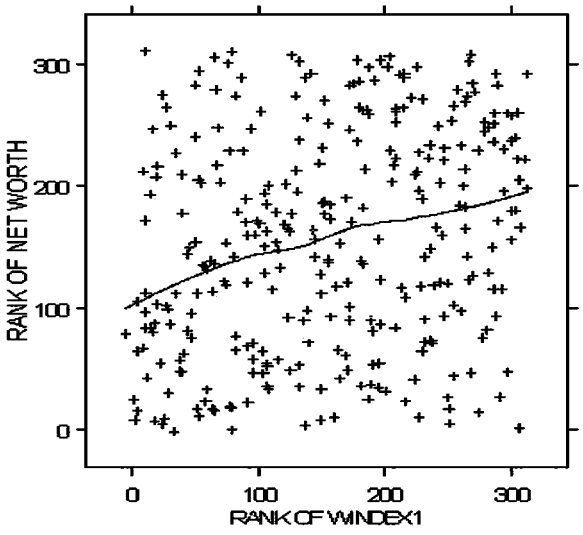


Figure 11d: Rank of Net Worth by Rank of WINDEXM, *Forbes* Sample

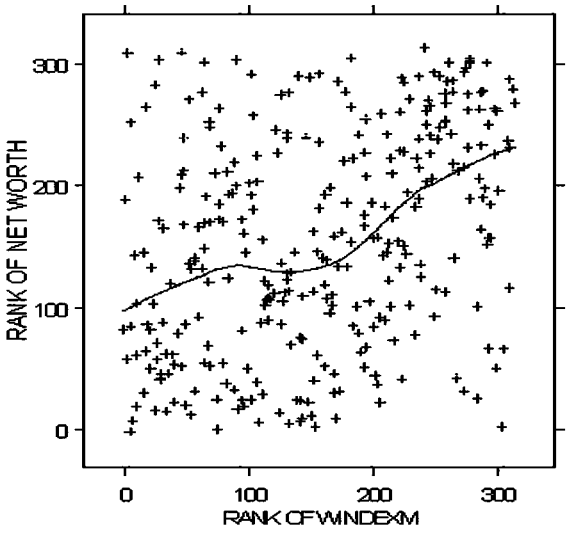
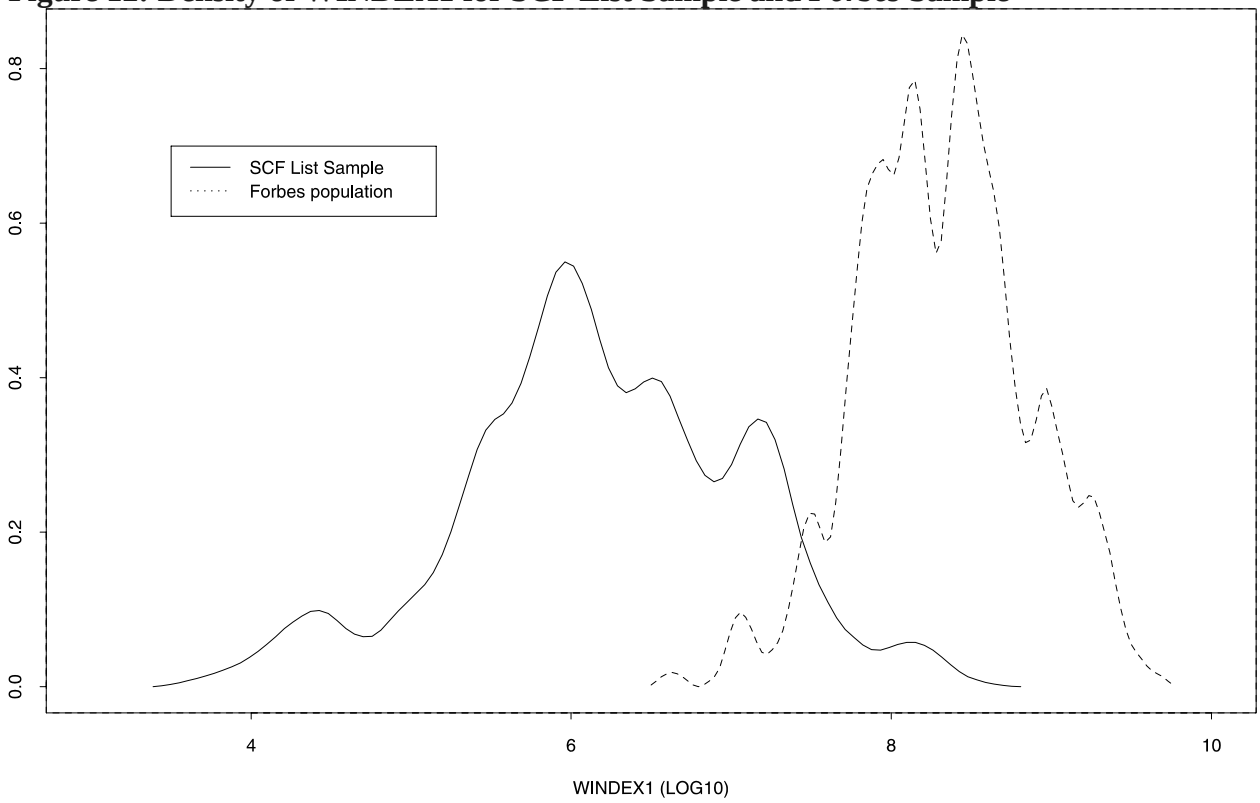


Figure 12: Density of WINDEX1 for SCF List Sample and *Forbes* Sample



III. Summary and Future Research

This paper has presented a range of descriptive results on the relationship between income and wealth using data from three sources: the SCF, the ITF, and *Forbes* Magazine. Because of the legal and ethical constraints on the use of some of these data, the focus of the paper is limited to a few issues that are relevant to improving the design of the SCF and understanding better the quality of the information collected.

The distribution of wealth is clearly bimodal, and it has a much longer right tail than that of income in both the SCF and *Forbes* data. However, when income and wealth are standardized to have the same median and standard deviation, income has a fatter right tail than wealth. The distribution of income conditional on wealth appears to have about the same log variance across most of the wealth distribution. The important goal of the paper is estimation of functional relationships between income and wealth. Detailed modeling using SCF and ITF data yields a model that is at least more effective than using income alone as a proxy for wealth. However, it is clear that the relationship is not strong, and more work is needed.

One serious limitation in the work reported here is the use of only one year of income data in each of the matches. It is generally recognized that over time, income often deviates from a longer run trend. In the case of wealthy people who have greater flexibility in the timing of their income, this issue may be particularly important. In general, it is likely that pooling multiple years of income could add substantially to our ability to predict wealth in terms of income flows. Practical considerations make it unlikely that we will be able to get reliable information on multiple years of income directly from SCF respondents. Although it is possible, in principle, to link up taxable income for respondents who file tax returns, there are some serious obstacles to the creation of a public or private research file. Legal and ethical considerations limit our ability to match survey and tax data only to the list sample, and even then, there are strong restrictions on the information that can be matched and how it can be used. Moreover, there would be very strong resistance to anything that would undermine even the impression that the survey undermined the privacy of individuals.

However, it is likely that narrower progress could be made by expanding the content of the ITF file used for the SCF sample design to include multiple years of income data. Earlier, Kennickell and McManus [1993] attempted to match multiple years of ITF data to examine the variability of the wealth index that could be attributable to income variability. It turned out that a critical problem is that many cases in a given year are not present in adjoining years. The main ITF is not a panel, but the Kefitz-like design of that sample effectively makes it more likely that cases that have either stable income or a spike in income in the following year are likely to be retained. Other observations with large declines in income are substantially less likely to be retained.[16] Thus, to address the question of the effects of income variability, it will be necessary to supplement the ITF file with data from the IRS Master File of individual returns. This step appears to be practical, but it raises many procedural and technical issues that place it beyond the scope of this paper. I very much hope to continue working toward this goal.

---

[16]This fact may have important implications for tax modeling. I would encourage the Office of Tax Analysis and SOI to consider retaining a higher proportion of cases that experience large declines in income.

Bibliography

Burbidge, John B., Lonnie Magee, and A. Leslie Robb [1988] "Alternative Transformations to Handle Extreme Values of the Dependent Variable," *Journal of the American Statistical Association*, v. 83, no. 401, pp.123-127.

Canterbury, E. Ray and E. Joe Nosari [1985] "The Forbes Four Hundred: The Determinants of Super-Wealth," *Southern Economic Journal*, v. 51 (April), pp. 1073-1982.

Frankel, Martin and Arthur B. Kennickell [1995] "Toward the Development of an Optimal Stratification Paradigm for the Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, 1995 Annual Meetings of the American Statistical Association, Orlando, FL.

Greenwood, Daphne [1983] "An Estimation of U.S. Family Wealth and its Distribution from Micro-Data, 1973," *Review of Income and Wealth*, March 1983, pp. 23-44.

Kennickell, Arthur B. [1998] "List Sample Design for the 1998 SCF," working paper, Board of Governors of the Federal Reserve System.

_____ [1991] "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," *Proceedings of the Section on Survey Research Methods*, 1991 Annual Meetings of the American Statistical Association, Atlanta, GA.

_____, Martha Starr-McCluer, and Annika Sundén [1997] "Family Finances in the U.S.: Recent Evidence from the Survey of Consumer Finances," *Federal Reserve Bulletin*, vol. 83 (January), pp. 1-24.

_____, _____, and R. Louise Woodburn [1996] "Weighting Design for the 1992 Survey of Consumer Finances," working paper, Board of Governors of the Federal Reserve System.

_____ and Douglas A. McManus [1993] "Sampling for Household Financial Characteristics Using Frame Information on Past Income," *Proceedings of the Section on Survey Research Methods*, 1993 Annual Meetings of the American Statistical Association, San Francisco, CA.

Tourangeau, Roger, Robert A. Johnson, Jiahe Qian, Hee-Choon Shin, and Martin R. Frankel [1993] "Selection of NORC's 1990 National Sample," working paper, National Opinion Research Center at the University of Chicago, Chicago, IL.