Simple Model for U.S. Income Inequality Data

William H. Press
The University of Texas at Austin

January 2, 2010

1 Introduction

This informal note looks at some aspects of fitting a model to published data on U.S. income inequality. The model can be useful for interpolation; for inferring quantities related to, but not specifically tabulated in, the published data; for displaying and highlighting various trends; and for rough estimations of the direct effects of changes in tax policy. Published data includes that available from IRS [3], and the curated data sets of Piketty and Saez [1, 2].

2 Recent IRS Data Suggest a Functional Form

Figure 1 shows recent (1996–2007) data from the IRS Statistics of Income Division regarding the fraction of Form 1040 returns showing adjusted gross incomes (AGIs) equal or greater than tabulated values. We plot G (AGI) versus P (percentile, accumulating from the high-income end) on a log-log plot so that the high-income tail behavior is shown clearly. The intepretation of the graph is that the individual at percentile point P had an adjusted gross income $G(P)$. In logarithmic coordinates $-\log_{10} P$ counts the number of powers of 10 into the tail of the distribution. A value of 1 indicates at the upper 10% percentile point; 2, upper 1%; 3, upper 0.1%; and so forth.

The "shoulder" in the data at $-\log_{10} P < 1$ (that is, $P > 10\%$) is a consequence of the obvious low-income behavior that $G(P)$ goes to zero as the lowest-income individual is reached at $P = 1$. However, we are here more interested in the high-income side, $P \to 0\%$. There, the straight-line asymptotes for each year shown are striking and indicate a power-law relationship. Specifically, the asymptotic high-income tails are not Gaussian, nor exponential, nor log-normal. They fall off more slowly than any of these.

Also clearly visible in the data, especially the 2007 series, is an inflection point at $\log_{10} P \approx 1.5$ where the power-law steepens to larger incomes for the diminishing tail. A functional form capable of capturing (i) the shoulder, (ii) the exponential tail, with (iii) the possible steepening of the power law is

$$G(P) = (AP^a + BP^b)(1 - P) \tag{1}$$

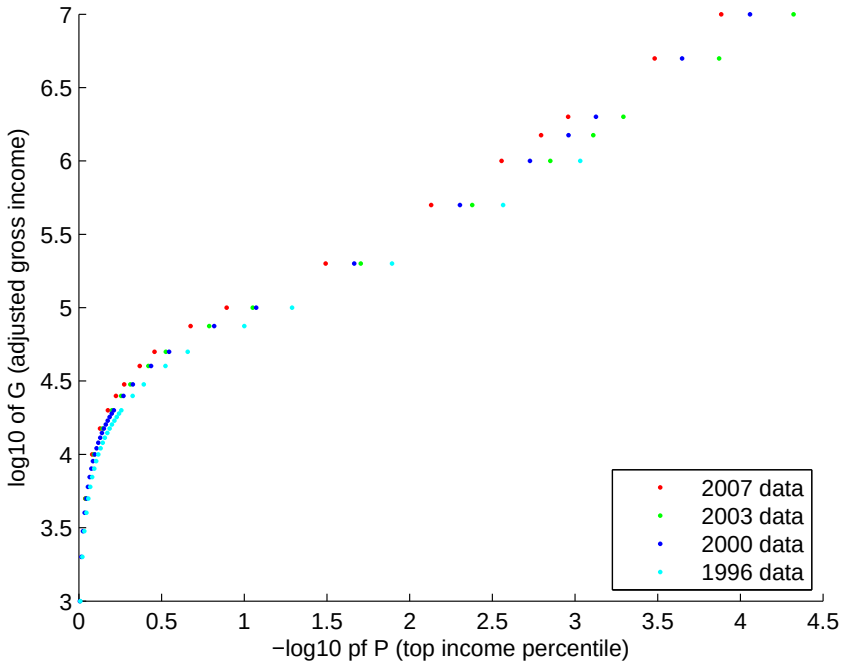

Figure 1: Adjusted Gross Income as a function of (upper) percentile of returns filed. Data from IRS [3]. On the horizontal axis, the value of 1 indicates at the upper 10% percentile point; 2, upper 1%; etc. On the vertical axis, the value 5 indicates an AGI of $100,000; 6, $1,000,000; etc.

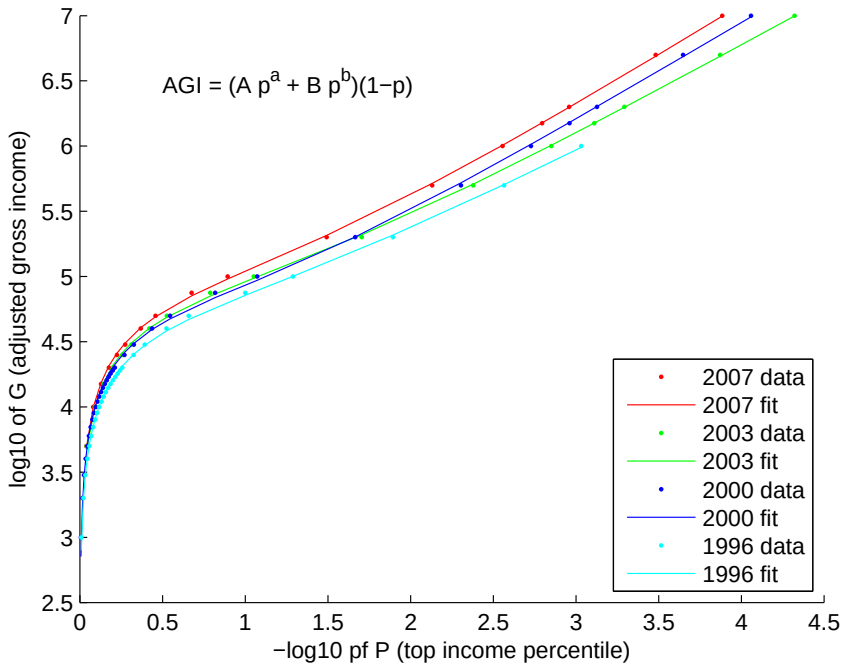

Figure 2: Same data as Figure 1, now showing fits to the functional form of equation (1).

This form has four free parameters A, B, a, b. Since (A, a) and (B, b) enter symmetrically, we can choose to make (A, a) the asymptotic power law and (B, b) the intermediate-range correction.

It is straightforward to fit each year's data for the four parameters in the model. Results are shown in Figure 2. Although the fit does not capture the inflection region perfectly, it is remarkably good overall. In particular, it does a good job of capturing the relatively small differences among the distributions for the four years shown. The fitted parameters for the years shown are given in Table 1. One sees in the fitted parameters year-to-year variability in the asymptotic power-law exponent a and in the relative significance of the steepening. For example, comparing 1996 to 2007, we see an increasing ratio B/A (more steepening) as well as an increasing (in magnitude) exponent a (larger asymptotic incomes as $P \to 0$). However, as the Table indicates, these trends are not monotonic in the years shown.

We note that equation (1) is optimized for study of the distribution of income on the high-income side. The model could be extended for better accuracy at low incomes by replacing the simple linear factor $(1-P)$ by a more complicated function of $(1-P)$, for example, $(1-P)^\alpha$, with an additional free parameter.

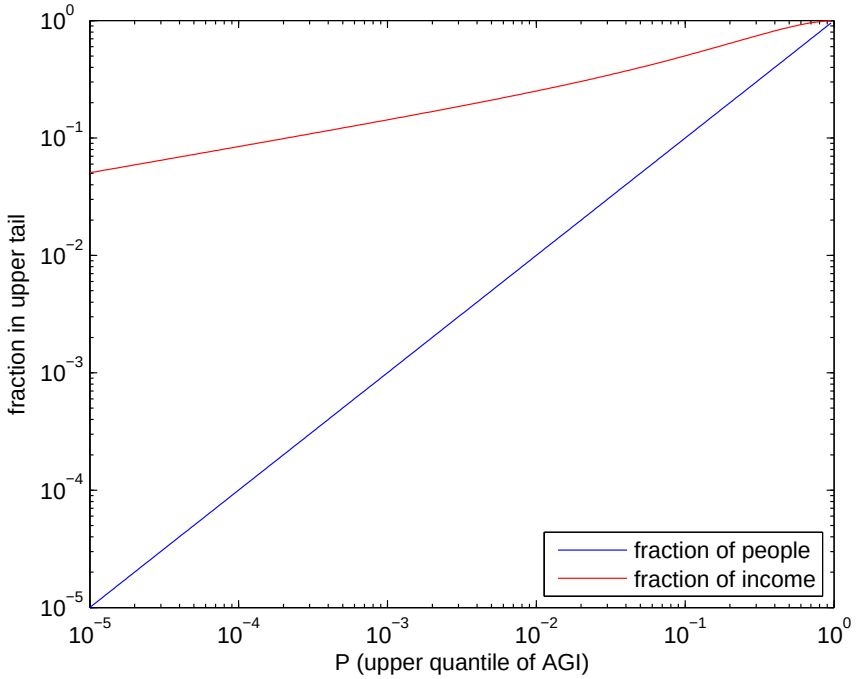

Figure 3: Fractions of people (tax returns) and total income (AGI) as a function of upper quantile P. IRS 2007 data. Higher income is to the left.

However, as Figure 2 shows, the simple linear factor, with no free low-income parameters, fits remarkably well.

year	A	B	a	b
1996	10187	29956	-0.647	-0.063
2000	8854	41219	-0.749	-0.044
2003	9465	40951	-0.698	-0.123
2007	9192	46130	-0.778	-0.161

Table 1: Values of fitted parameters for the fits shown in Figure 2.

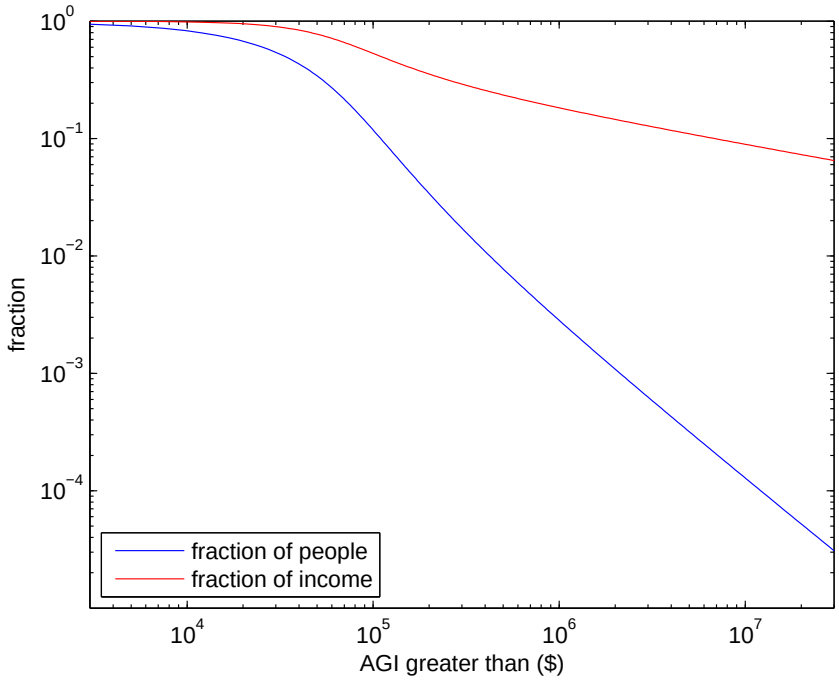

Figure 4: Fractions of people (tax returns) and total income (AGI) as a function of income. IRS 2007 data. Higher income is to the right.

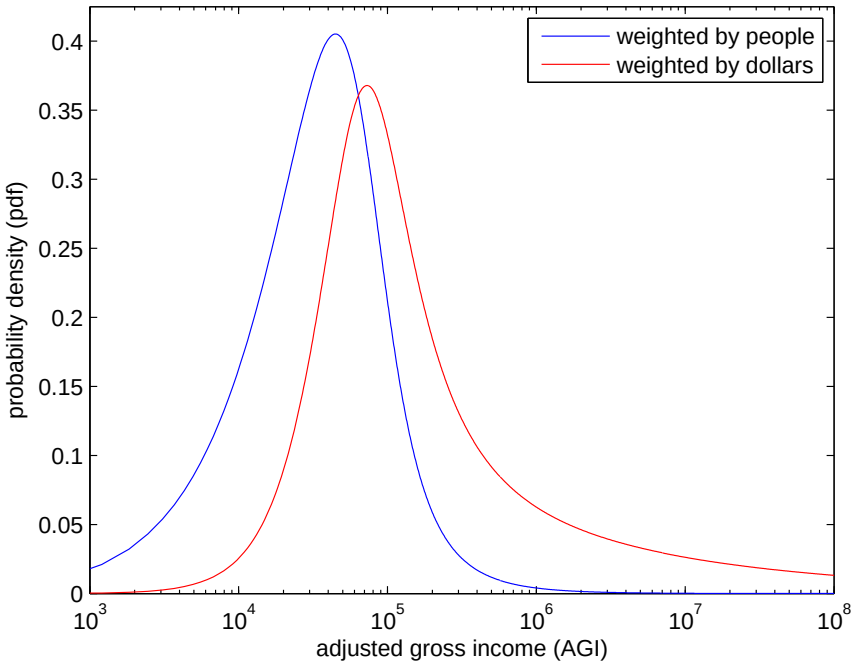

Figure 5: Probability density functions for people (number of returns) and total adjusted gross income (AGI) as a function of log AGI. IRS 2007 data. Higher income is to the right.

3 Other Quantities Are Easily Calculated

The model's simple functional form makes it easy to calculate other quantities of interest. For example, if we define the integral

$$C(P) \equiv \int_0^P G(P')dP' = A\left(\frac{1}{a+1} - \frac{P}{a+2}\right)P^{a+1} + B\left(\frac{1}{b+1} - \frac{P}{b+2}\right)P^{b+1}, \tag{2}$$

then the fraction of aggregate income due to percentile P or less (that is, higher income) is $C(P)/C(1)$. We can plot this as a function either of P or of $G(P)$, as for the 2007 data in Figures 3 and 4. From the figures, we see, for example that in 2007, about 2 people in 10,000 were responsible for about 10% of aggregate income, earning just under \$10,000,000 each.

If we take the derivatives of the curves in Figure 4 with respect to the abscissa $d\log G$, we get the probability density functions for people and aggregate income, as a function of AGI, that are shown in Figure 5. As in Figure 4, we see in Figure 5 that individuals with incomes greater than several million dollars, though few in number, generate a significant fraction of aggregate income.

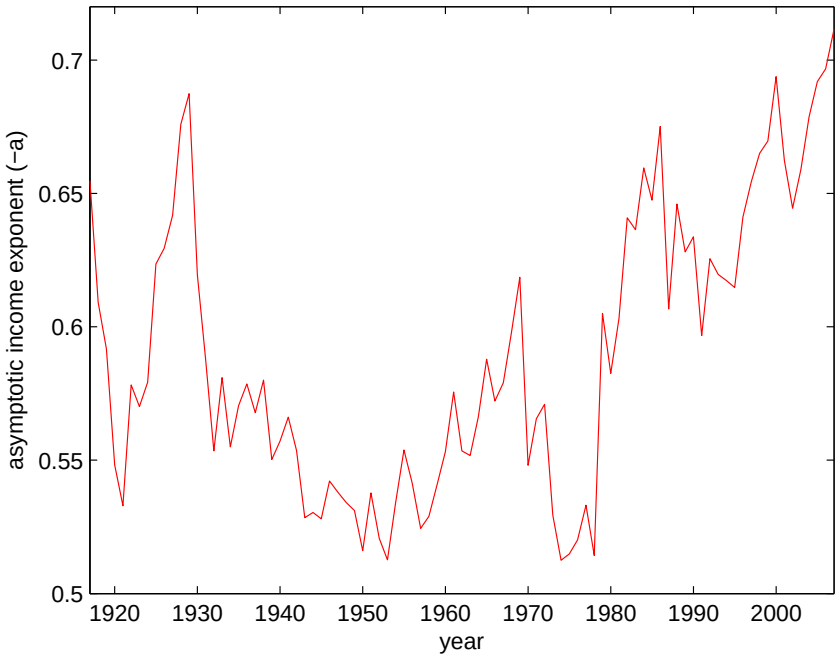

Figure 6: Asymptotic exponent for high incomes, fitted from Saez data [2]. The U-shaped curve, echoing Saez' presentation of the data in other formats, shows substantially greater income equality in the period 1945–1975 than in 1980–present.

4 Modeling Saez's Historical Data

Piketty and Saez [1, 2] have maintained a carefully curated collection of income time series (1917–2007) that include upper quantile income fraction data for the upper 10%, 5%, 1%, and 0.01%. In our notation, these correspond to data values for $C(P)/C(1)$ for four values of P for each year Y. Because $C(P)/C(1)$ is a ratio, these data do not determine A and B separately, but only their ratio. For each year, the model thus has 3 free parameters to be fitted from 4 data points. As before, the fitting is computationally straightforward.

Figure 6 plots the key exponent $-a$, indicating the asymptotic power law at the high-income end, as a function of year. As a colleague has remarked, "Wow, something did happen around 1980!" To see the effect across the whole range of incomes, we can plot the full income distributions $G(P)$ (equation (1)) for multiple years. Figure 7 shows this for the years 1945–1975 (blue) and separately for 1980–2006 (red). The year 2007, which is the most extreme, is shown in green.

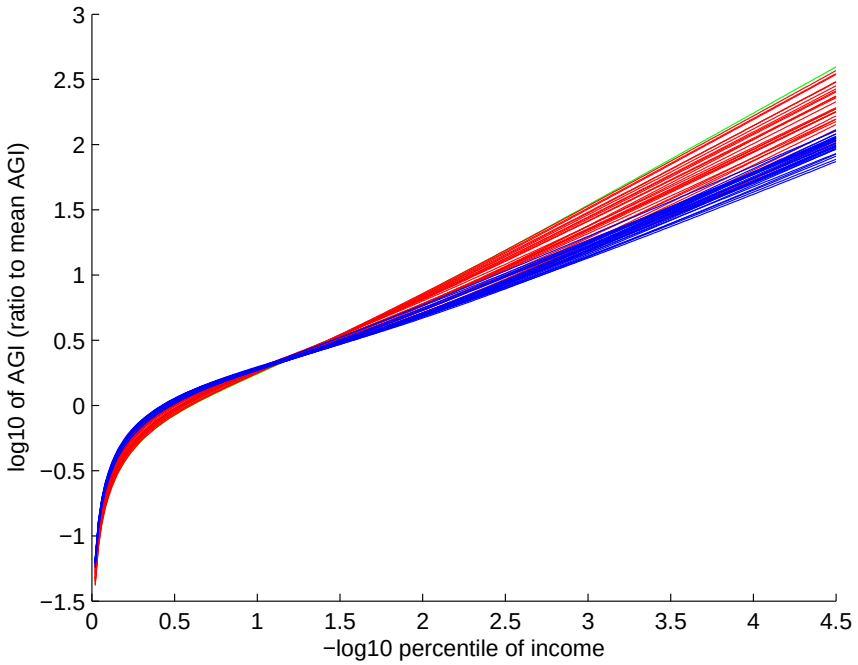

Figure 7: Income relative to mean income as a function of upper quantile P for years 1945–1975 (blue), 1980–2006 (red), and 2007 (green, highest curve). The format of this figure is the same as for Figures 1 and 2, except that the ordinate is relative to the mean income of that year.

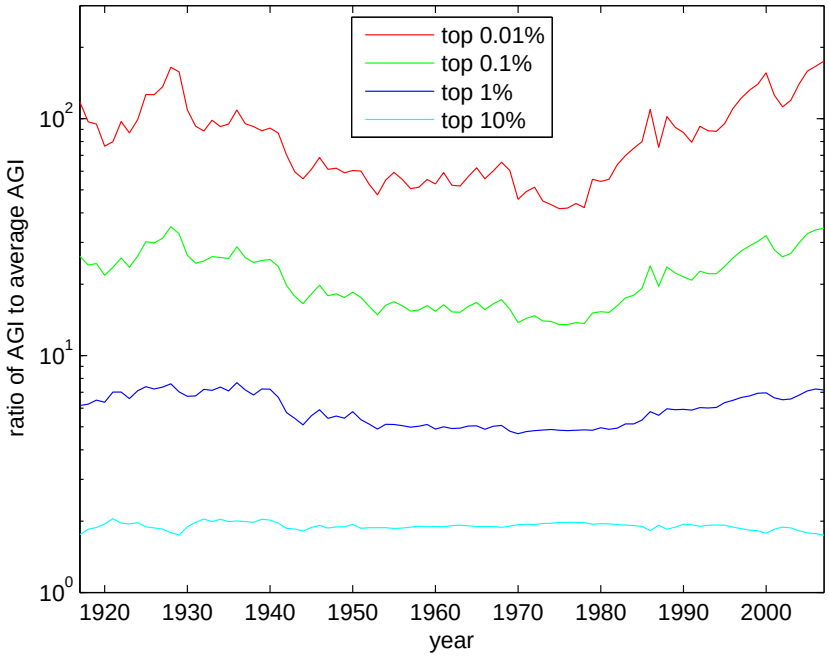

Figure 8: Income as a multiple of mean income for individuals at selected top quantile points, by year, from fits to Saez data [2].

It is interesting to look at some vertical slices through Figure 7 at selected quantiles, displayed by year. This is shown in Figure 8. One sees that individuals at the top 10% quantile point earn about 2 times the mean fairly consistently across the whole time period shown. As we move to progressively more rarified high-income returns, we see progressively greater differences between the periods 1945–1975 and 1980–present. This echos the main findings of Piketty and Saez [1].

Table 2 gives the average income multiples separately for the periods 1945–1975, 1980–2006, and 2007. In the first period, an individual at the top 0.01% quantile point earned 55.2 times the mean. In 2007, the multiple for this quantile was 174.7.

quantile point	multiple of mean income		
	1945–1975	1980–2006	2007
top 0.01%	55.2	103.4	174.7
top 0.1%	16.1	23.9	34.4
top 1%	5.1	6.1	7.2
top 10%	1.9	1.9	1.7

Table 2: Multiples of mean income for individuals at selected quantile points for the periods 1945–1975, 1980–2006, and 2007.

5 A Surtax to Change the Shape of the Distribution?

Suppose that we have decided, somehow, that the period 1945–1975 was a Golden Age of equitable income distribution, and that the period 1980–2007 is an aberration to be corrected by tax policy. This is entirely simplistic, of course, but it leads to an interesting calculation: We can ask what surtax rate, as a function of adjusted gross income, is necessary to reshape the 1980–2007 distribution to the same shape as that of 1945–1975. The surtax rate can be positive or negative, so as to make the net result income neutral. Note that our modest proposal does not seek to return income *levels* to their 1945–1975 values, but only to reshape the distribution curve on current aggregate income. (That ought to be radical enough!)

Since $C(1)$ (equation (2)) is the mean income, $G(P)/C(1)$ is the income distribution normalized to the mean, so that

$$\int_0^1 \frac{G(P)}{C(1)} dP = 1 \tag{3}$$

for any model parameters. Let $G_p(P)$ be the "present" model with parameters A, B, a, b that are the means of the period 1980-2007; while $G_t(P)$ is the "target" model whose parameters average the period 1945–1975. Figure 9 plots

$G_t(P)/C_t(1)$ and $G_p(P)/C_p(1)$ in the by-now-familiar format. Although the logarithmic axes obscure this point, equation (3) says that the area between the curves on the right (red above blue) is the same as the area on the left (blue above red) when plotted in (P, G) coordinates.

Thus, if we apply a positive or negative surtax rate at each value of P that brings the red curve down or up to the blue curve target, the overall result will be aggregate income neutral. The implied rate is

$$r = 1 - \frac{G_t(P)/C_t(1)}{G_p(P)/C_p(1)}, \tag{4}$$

to be applied to AGI before any other taxation. In the first instance this surtax rate is a function of the percentile P; but we can use the 2007 model $G(P)$ to convert the independent variable to (2007) AGI. The result is shown in Figure 10.

The surtax rate is seen to vary from about 40% on AGIs in the \$10,000,000 range to -40% (negative tax) for AGIs less than \$10,000. The cross-over between positive and negative tax is at about \$180,000 ($P \approx 5\%$). For AGIs of \$350,000 ($P \approx 1.5\%$), the surtax is about 10%, while for AGIs of \$1,000,000 ($P \approx 0.25\%$), the surtax is on the order of 20%. It is interesting that so massive a reshaping of income distribution can in principle be achieved by levying a positive surtax on only 5% of returns, and a surtax $> 20\%$ on only 0.25% of returns. This would of course be little consolation for those few taxpayers with eight-digit incomes, who might well view a 40% additional tax as confiscatory.

There are many reasons that Figure 10 is not the basis for a practical proposal, even aside from the question of whether its premise is desirable. One main reason is that although it is income neutral, it is not tax revenue neutral. It would be interesting to see the calculation equivalent to Figure 10, but tax-revenue neutral, something beyond the simple models given here. It seems likely that the surtax in Figure 10 actually generates increased tax revenues, because the incomes reduced by positive surtax will generally not change tax bracket, while the funds generated will be applied by negative surtax rate to individuals with incomes in the range of dense tax-bracket changes. A tax-revenue neutral scheme with a somewhat smaller incidence on the super-rich might therefore be possible, if this is desirable.

Acknowledgement

This exploration was motivated by a talk by former NERA President Richard Rapp to *Salon XI* in Santa Fe, NM, December 29, 2009.

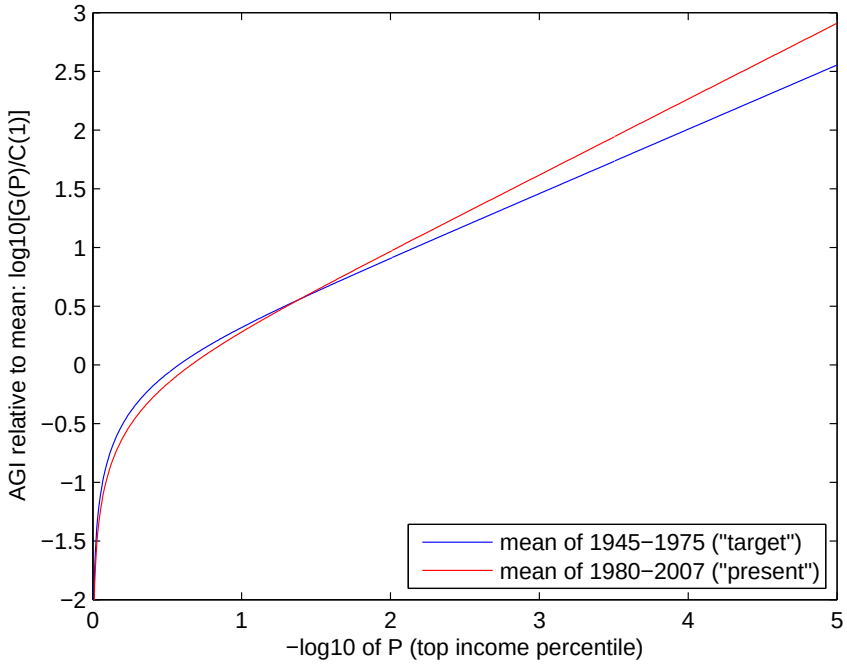

Figure 9: Distributions of AGI relative to mean AGI, averaged for the periods 1945–1975 and 1980–2007.

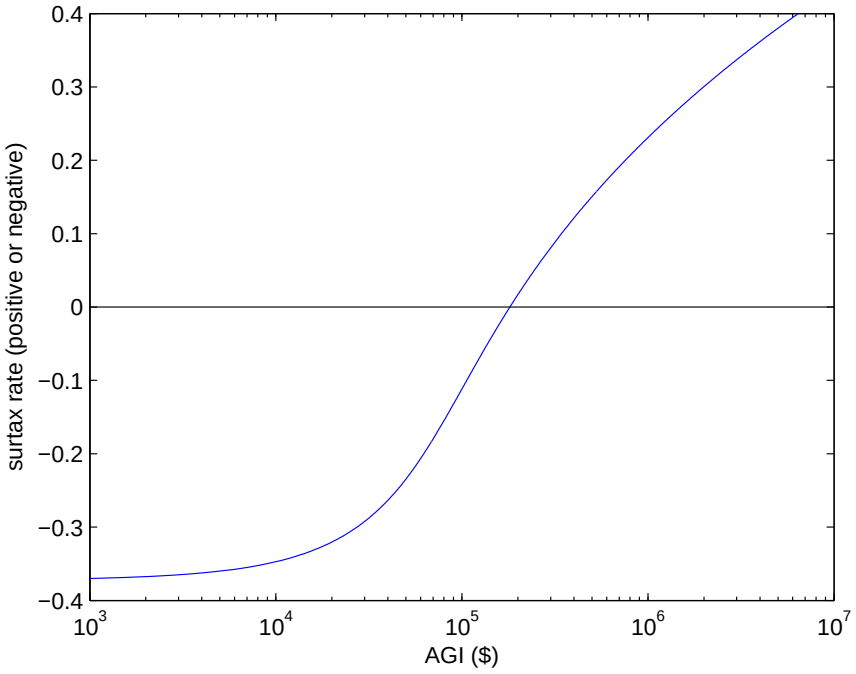

Figure 10: Surtax rates that would map the 1980–2007 income distribution curve into the shape of the 1945–1975 curve. The abscissa is 2007 AGI. The surtax rate varies between about 40% on AGIs in the \$10,000,000 range to −40% (negative tax) for AGIs less than \$10,000.

References

[1] Piketty, Thomas and Saez, Emmanuel (2003) "Income Inequality in the United States, 1913–1998", *Quarterly Journal of Economics*, 118, 1–39; updated version at http://elsa.berkeley.edu/∼saez/piketty-saezOUP04US.pdf (Oxford University Press, 2007).

[2] Saez, Emmanuel (2009) Updated data through 2007 available at http://elsa.berkeley.edu/∼saez/ .

[3] Statistics of Income Division, Internal Revenue Service, (1996–2007) "Individual Income Tax Returns Publication 1304 (Complete Report)" available at http://www.irs.gov/taxstats/indtaxstats/ .