MEMORANDUM

FROM: John O'Hare
TO: Files
DATE: June 15th, 2009
SUBJECT: Extrapolation Methodology

In this memorandum we describe a methodology for extrapolating, or aging, microdata files for several years into the future. While the procedures can be applied to any type of microdata, we pay special attention to the unique challenges faced when dealing with individual income tax data. These data will include, for example, administrative tax return records and samples such as the Statistics of Income (SOI) Public Use File (PUF).

It is useful and occasionally necessary to make a distinction between the methodology we employ to age the data and the development of the macroeconomic forecasts that we would like the data to represent. We treat both here.

After a brief overview, we first describe a methodology that we have used successfully for 20 years in support of microsimulation modeling at the federal and state level. In the second section, we illustrate how some simple macroeconomic forecasts might be developed.

OVERVIEW

Most microdata, whether from administrative sources or surveys, become available with a time lag. This is especially true of individual income tax data where there may be several years between the time the data are collected and eventually released to researchers. In contrast, most public policy models are forward-looking and aim to provide estimates of how a particular tax or spending policy will affect the economy in the near- to mid- term. In tax policy, it is often required to provide estimates of the budgetary effect for 10 years into the future.

Historically, tax policy analysts at the Treasury's Office of Tax Analysis (OTA) and the Congressional Joint Committee on Taxation (JCT) have employed a variety of methods to extrapolate their underlying microsimulation models many years into the future. These methods are generally referred to as "static aging", and the terminology relates to the fact that the procedures impose no intertemporal linkages across records or over time. Put somewhat differently, when static aging is used to extrapolate a microdata file t years into the future, it is not necessary to age the data for the intervening t-1 years.[1]

[1] This is in contrast to dynamic microsimulation models where outcomes for year t will directly influence the outcomes in year t+1 for each micro unit.

Generally speaking, static aging methods are implemented in two distinct steps. In the first step (Stage I), income and deduction items are adjusted to reflect forecasts of inflation from the base year to the simulation year. In the second step (Stage II), numerical optimization procedures are used to re-weight the file with the goal of hitting externally imposed control totals.

In what follows, we assume that a set of forecasts has been obtained and we wish to age our base year data to match these forecasts. The forecasts are generally supplied, or derived from, some external macroeconomic forecast (e.g., economic projections supplied by the Congressional Budget Office (CBO)). In addition, some forecasts of underlying tax variables are supplied by analysts or may be derived from other, special-purpose economic models. Later we outline one way in which targets can be identified and estimated.

EXTRAPOLATING MICRODATA FILES

In most microsimulation models used for policy analysis, the underlying data comes from a survey of relevant economic units or from a sample of administrative records. These samples will be representative of the population of interest (the sample frame) and each record on the file is assigned a sample weight that is derived from the sample design.[2] The sum of the sample weights on the file represents an estimate of the population. Likewise, weighted totals of income items represent an estimate of the population total in the year the sample was chosen.

Microsimulation is a bottom-up approach to policy analysis and aggregate, macroeconomic outcomes are calculated by summing-up the individual outcomes from the constituent micro units that comprise the model. In order for these aggregate outcomes to be closely aligned with an external forecast of the macro economy, we must pay attention to how both the individual variables on the file age adjusted over time and how the sample weights change over the time horizon.

STAGE I

In the first phase of the extrapolation process we adjust the individual data elements on the base year file on a record-by-record basis to correspond with our macroeconomic forecast. Specifically, we multiply each data element on the file by a per capita adjustment factor that ensures that our macro targets will be hit. These adjustment factors (the "Stage I" factors) are calculated as follows. Define the following variables:

X_0: the value of some macroeconomic aggregate in the base year (e.g., wages)
P_0: the population total in the base year (e.g., number of tax returns)
w_{i0}: sample weight of the i^{th} unit in the base year
x_{i0}: observed value of variable x reported by unit i in the base year.

[2] In a stratified sample design, the sample weight is generally calculated as the ratio of the population within a particular strata to the sample size in that strata.

With this notation, we have the following relationships that link the macro and micro variables:[3]

$$\sum w_{i0} = P_0 \qquad \textbf{(1)}$$

$$\sum x_{i0} \bullet w_{i0} = X_0 \qquad \textbf{(2)}$$

Now suppose we have some externally supplied forecast for total population and the aggregate level of our macro variable for some year t:

X_t: the forecasted value of our macroeconomic aggregate in year t
P_t: the forecasted population total in year t.

Define the following growth rates from the base year to year t:

$$(1+r) = \frac{X_t}{X_0} \qquad \textbf{(3)}$$

$$(1+p) = \frac{P_t}{P_0} \qquad \textbf{(4)}$$

$$(1+\rho) = \frac{(1+r)}{(1+p)} \qquad \text{(per capita adjustment)} \qquad \textbf{(5)}$$

To perform the Stage I adjustment to the microdata on our base year file we first apply the population growth factor to the base year sample weights on the file, and second, apply the per capita adjustment factor to the reported base year income on each record. That is, we calculate

$$w_{it} = w_{i0} \bullet (1+p) \qquad \textbf{(6)}$$

$$x_{it} = x_{i0} \bullet (1+\rho) \qquad \textbf{(7)}$$

This ensures that our macro targets are hit in year t. That is, for population we calculate:

$$\sum w_{it} = \sum w_{i0} \bullet (1+p)$$

$$= (1+p)\sum w_{i0}$$

[3] In practice, the value of X_0 may represent a different concept that what appears in the microdata. For example, wages on the tax return may not include some types of employee compensation that is measured at the macro level.

$$= \frac{P_t}{P_0} \sum w_{i0}$$

$$= P_t .$$

The first equality is from the definition in **(6)**; the second equality is due to the distributive property; the third equality is from the definition in **(4)**; and the final result is from **(1)**. Similarly, our aggregate macro target is achieved:

$$\sum x_{i,t} \bullet w_{i,t} = \sum x_{i,o}(1+\rho) \bullet w_{i,o}(1+p)$$

$$= (1+\rho)(1+p) \sum x_{i,o} \bullet w_{i,o}$$

$$= (1+r) \sum x_{i,o} \bullet w_{i,o}$$

$$= (1+r) X_0$$

$$= X_t$$

where the first equality defines the Stage I adjustment; the second is from the distributive law; the third is from **(5)**; the fourth is from **(2)**; and the final equality is from **(3)**.

The procedure outlined above is easily extended to accommodate any number of macroeconomic targets coupled with their micro variables and the methodology guarantees that all the targets are reached exactly. We point out that most microsimulation models stop here and perform no other adjustments to the data. However, experience has shown that while this methodology is sufficient to reach any exogenous target, certain endogenous, or jointly determined variables, may need further calibration, especially if the aging is done for more than a few years into the feature. In the next section, we describe a set of additional adjustments that ensure that both the exogenous and endogenous targets are reached.

STAGE II

In the second stage of the extrapolation process, numerical optimization methods are used to adjust the sample weights on the file to ensure all targets are reached. We rely on a linear programming algorithm to solve for a new set of (sample) weights that minimize the absolute value of the percentage change in the weight for each record on the file subject to the constraint that each of the targets are reached.

Mathematically, we let z_i represent the percentage change in the sample weight for record i on the file relative to the base year sample weight:

$$w_{i,t} = w_{i,o}(1 + z_i).$$

Furthermore, to accommodate the absolute value metric, we decompose the z_i into a positive and negative component by defining:

$$r_i = z_i^+ \quad \textbf{(8)}$$

$$s_i = z_i^- \quad \textbf{(9)}$$

$$z_i = r_i - s_i \quad \textbf{(10)}$$

Here, **(8)** and **(9)** define the positive and negative parts, respectively, of the z_i and this leads to the definition of the absolute value:

$$|z_i| = r_i + s_i \quad \textbf{(11)}$$

Our linear programming problem then becomes one of choosing the smallest δ that solves:

$$\min_{\langle r,s \rangle} \sum |z_i| \left(= \sum r_i + s_i \right) \quad \textbf{(12)}$$

subject to:

$$0 \le |z_i| \le \delta \quad \textbf{(13)}$$

and

$$Ax = b \quad \textbf{(14)}$$

where $x \in (r_i, s_i)$ are the variables to be solved for, A is a coefficient matrix that will depend on the specific targets and b is a target vector.

This procedure is flexible enough to handle many types of targets. In practice, we have usually incorporated four types of constraints:

1. Amount Aggregates (AA)
2. Return Aggregates (RA)
3. Amount by Income Class (AB)
4. Returns by Income Class (RB)

An example of an amount aggregate (AA) target might be total capital gains; a return aggregate (RA) might be the total number of taxpayers reporting capital gains[4]; and

[4] Targeting both, of course, would be the same as targeting average capital gains.

targeting both these variables by income class would constitute AB and RB targets, respectively.

Here, in no particular order, are some observations we have made over the years in using this Stage II procedure:

- Solving the linear programming (LP) formulation of this problem means doubling the number of variables that need to be solved for (the r_i's and s_i's). Our view is that this is a small price to pay for the assurance that the optimum is achieved under the LP framework.
- It may take several iterations of the algorithm to achieve a solution that is acceptable by iterating over δ.
- Not all variables in the model will (or should) be targeted. Those that aren't will only receive a Stage I adjustment. In fact, after the LP problem is solved, most returns on the file will receive only a Stage I adjustment. That is, z_i will be zero for most records.
- Experience with this algorithm has show that it is reasonable stable over the forecasting horizon.
- More importantly, those variables that are not targeted in the Stage II extrapolation remain reasonably close to their Stage I values. Put differently, the reweighting does not materially affect the values of these variables.
- In practice, when solving the LP, we usually use the solution for year t as the starting point for obtaining a solution for year t+1. This speeds up the LP algorithm substantially.