Extrapolation

Overview

OSPC's current micro simulation model is using CPS-matched data file based on 2009 Public Use File (PUF) to generate tax revenue projections on a 10-year budget window. In order to do the calculations in the projection time period (2015-2024), the model needs to extrapolate the variables in PUF to the future years. OSPC follows a two-stage procedure, originally developed by John O'Hare, to extrapolate the PUF variables. The first stage is to 'blow-up' those variables directly using external macro-economic baseline projections, mainly to reflect inflation. Then the second step is to maintain proper distribution of wage and salaries, as well as key macro targets.

The CPS-matched file is a statistically matched file base on 09 PUF. It not only includes filers information from PUF, but also non-filers from Current Population Survey (CPS). The Public Use File of 2009, currently base of the statistical match, is a stratified sample from the IRS full file with all taxpayers' information. In PUF, there are in total 152,526 records. One records was removed before futhur procedures, due to the fact that this record, with Record ID (RECID = 999999) contains aggregated information that would potential distort later steps. All the information in this aggregated record was re-distributed to all other records using a Stage I and Stage II extrapolation routine described below. All the remaining record has a weight variable indicating approximately how many taxpayers this record represents in the full file. In addition to the weight information, each record has 201 variables, roughly 169 of which are used in the OSPC tax calculator. The aggregate values (weighted sum) of each variable are very close to the statistics for all taxpayers provided by IRS.

Stage I

In Stage I, the model blows up PUF variables using per capita adjustment factors. This step will make sure the aggregated values of these variables match certain macroeconomic projections in terms of growth rates. Currently, the external baseline projection used in Stage I is mainly from the CBO economic outlook. In near future, the calculations from OSPC's dynamic model, LOGUS, will replace these external projections.

The per capita adjustment factors applied to each record in PUF are calculated based on both macroeconomic targets growth rates and population growth rates. Both of these rates are simple-compounded rates, derived from multiple data resources, including CBO, IRS and Census Bureau. For example, the projected GDP in year t is projected at X_t, then the OSPC model calculates growth rate r as:

$$r = X_t / X_0 - 1$$

Similarly, population P growth rate p is calculated as:

$$p = P_t / P_0 - 1$$

To make sure the extrapolated variables indeed aggregate at X_{t}, the model calculates the ratio of the two rates as per capital adjustment factor and then uses this value to blow up the target variable for each record in the base year $x_{i,0}$.

Per capita adjustment for each individual record:

$$1 + \varrho = \frac{1 + r}{1 + p}$$

After blow-up, the target variable value for each record and the weight variable would be:

$$x_{i,t} = (1 + \varrho)x_{i,0}$$

$$w_{i,t} = (1 + p)w_{i,0}$$

Here $x_{i,t}$ represents the extrapolated value of targets for one return i at year t, and $w_{i,t}$ represents the stage-I extrapolated value of weight for each return i.

This method of blow-up ensures that the aggregates of future individual variables would be in line with the macro projections based on IRS taxpayer information and CBO baseline. As you can see as follows:

Starting with the weighted sum of a target variable x at year t

$$\begin{aligned}
&\sum_i x_{i,t} w_{i,t} \\
&= \sum_i (1 + \varrho)x_{i,0}(1 + p)w_{i,0} \\
&= \sum \frac{1 + r}{1 + p} x_{i,0}(1 + p)w_{i,0} \\
&= (1 + r)\sum_i x_{i,0} w_{i,0} \\
&= (1 + r)X_0 \\
&= X_t
\end{aligned}$$

Currently in Stage I, OSPC's model has 20 factors, 16 macro factors and 4 population related factors. For macroeconomic economic targets like GDP that do not have corresponding variables in the PUF, they are derived directly from CBO baseline projections. While the other factors, taxable interests for example, are calculated from both CBO projection and IRS statistics. Specifically, the model takes the most recent (currently 2012) statistics of this variable released by IRS as the base of macroeconomic targets projection. Then the model 'ages' those statistics with the yearly growth rates of CBO targets to extrapolate those statistics on the budget window. Some of the values extrapolated here are used as target in Stage II as well. More details will be provided later in this description. With these extrapolated values, the model then using the 2008 data as base, calculates simple growth rates of these extrapolated values for each projection year. Please see appendix A for a complete list of all factors.

As you can see, there are roughly 169 variables that need to be blown up in PUF but only 20 factors and less than 16 per-capita adjustments are available. At the moment for simplicity reasons, many variables are sharing one factor. Most variables are aged at personal income growth rate. Variables that don't follow the trajectory of personal income are applied a factor that fits the best. For example, home mortgage interest deduction is currently using taxable interest income factor for adjustment. The complete table for the correspondence between stage I factors and PUF variables can be found in Appendix B.

Stage II

At the end of Stage I, the OSPC model blows up the base year data through the projection years. Even though all targets set up in Stage I can be hit at this point, some other variables, mainly the ones not included in the targets, might behave oddly if we only target the aggregate values. For example, we targeted total wage in Stage I and this would kick Earned Income Credit (EIC) out of place in future years. But the Stage I simple-compounded rate blow-up process cannot get both total wage and EIC fell in reasonable ranges at the same time, as those two variables impose closely related but different

requirements on wage and population distribution. John found that if wage distribution were maintained same over years, this problem would be fixed automatically.

In Stage II, the OSPC model applies a linear programming (LP) algorithm on the post-Stage-I records to adjust the weights. In this way, all the targeted variables would sum up to the targets, and other non-target variables stay in reasonable ranges. The stage II targets not only include those stage I targets calculated based on both CBO baseline growth rates and IRS statistics, but also certain targets on population and return groups, and wage distribution. For the wage distribution, the model currently assumes the distribution of wage stays the same over the projection time period. This assumption is subject to change as many other resources, including JCT's present model, suggest that the growth rate of higher income class is greater than lower income class. A complete list of Stage II targets is attached as Appendix C.

In this LP model, the object function is based on the absolute value of percentage adjustment on weights. Assume $z_{i,t}$ is the percentage adjustment on weights w for record i between any projection year t and post-stage-I base year weight. One twist on top of the original Stage I blow up here is that we applied different factor to different age groups for the weight variable. From the population growth rates by age in the past few years, the 65+ group grows at a higher rate than the overall population. (citation needed!!!) As PUF doesn't contain any age information, we use social security benefits, represented by variable e02400 in PUF, to approximate the age characteristic of PUF. Specifically, if one filer received social security benefits in 2008, we assume he or she is older than 65 and blow up the weight of this return using the senior population growth rate (APOPSNR). Otherwise, the weight is blown up at the total return growth rate (ARETS). In this way, the weight variable at year t is:

$$w_{i,t} = \begin{cases} w_{i,0}*ARETS & if\ e02400 \le 0 \\ w_{i,0}*APOPSNR & if\ e02400 > 0 \end{cases}$$

Then the percentage adjustment on weights can be expressed as:

$$z_{i,t} = \begin{cases} \dfrac{w_{i,t}*}{w_{i,0}*ARETS} - 1 & if\ e02400 \le 0 \\ \dfrac{w_{i,t}*}{w_{i,0}*APOPSNR} - 1 & if\ e02400 > 0 \end{cases}$$

Here w_{i,t}* is the final optimized weight after stage II, which is unknown at this point.

In this problem, we want to minimize the overall adjustments on all records, in other words, to minimize the sum of the absolute values of all the percentage changes on weights. Thus, this LP problem aims at minimizing the following function:

$$\sum_i |z_{i,t}|$$

To enhance efficiency of the optimization, we decompose the percentage adjustment $z_{i,t}$ into two components: r and s. These two components respectively carry the absolute values of positive and negative elements of original adjustments $z_{i,t}$. Mathematically, they can be expressed as following:

$$r_{i,t} = \begin{cases} z_{i,t}, & z_{i,t} > 0 \\ 0, & else \end{cases}$$

$$s_{i,t} = \begin{cases} z_{i,t}, & z_{i,t} < 0 \\ 0, & else \end{cases}$$

Then it's not hard to see that the original adjustment is the difference of the two components, and its absolute value is the sum of the two components.

$$z_{i,t} = r_{i,t} - s_{i,t}$$

$$|z_{i,t}| = r_{i,t} + s_{i,t}$$

Therefore, the object function turns into:

$$\sum_i (r_{i,t} + s_{i,t})$$

As this LP problem takes post-stage-I data as input and borrows a large number of targets derived in Stage I, the percentage adjustments should be fairly small. In addition, if the adjustment runs too large, there's a big chance that the non-target variables get pulled away from their normal ranges. Thus in the LP model, the tolerance of each $z_{i,t} = r_{i,t} - s_{i,t}$ is also incorporate as a constraint:

$$r_{i,t} + s_{i,t} < \delta$$

For each year, the OSPC model try to find the lowest tolerance that still allows the LP solver generates feasible solutions. This tolerance is different from year to year. Generally speaking, years earlier in the projection period have lower tolerances than years further out in the future. Currently the tolerances for all the years are under 0.45.

In addition to this tolerance constraint, all other constraints of this LP problem can be grouped into three categories:

- Return targets
- Aggregate targets
- Aggregate by Income class targets

The first category of return targets is mainly focused on different filing status of returns, and population of different age groups. Total numbers of returns with different filing status are maintained in the same ratio to total returns through years, and each category grows at the overall return growth rate. Senior filers are extrapolated at the rate for population with age greater than 65, projected by Census. Mathematically, these constraints are

$$\sum_i w_{i,t}(1 + r_{i,t} - s_{i,t}) = W_t$$

where $w_{i,t}$ refers to the twisted post-stage-I weights created for Stage II, and W_t refers to the Stage II return targets.

The second category contains macroeconomic targets other than total wages. These constraints are set up as following:

$$\sum_i x_{i,t} w_{i,t}(1 + r_{i,t} - s_{i,t}) = X_t$$

Both target variable for record i, $x_{i,t}$, and weight variable $w_{i,t}$ are the post-stage-I results.

The third category of aggregate by income class includes 12 targets. These targets are the weighted sum of wages and salaries ranked and grouped by adjusted gross income (AGI) classes. Currently, the distribution of wages and salaries is maintained the same as the base year through the entire projection period.

After setting up the object function and all the constraint, the OSPC model runs the CLP solver to find

solutions for this LP problem. To finish the Stage II adjustment, the model applies the solution from CLP solver, $r_{i,t} - s_{i,t}$, to the post-stage-I weights and gets the final adjusted weights.

To sum up, the OSPC tax calculator applies the Stage I factors to blow up most of the PUF variables and uses the adjusted weights generated in Stage II for all projection years. This description outlines the current extrapolation routine, which is subject to change in near future for various reasons. Many factors would change the routine significantly, which includes a new version of PUF released by IRS. This description will be updated accordingly.

Appendix A Stage I factors

Var_name	Long name
AGDPN	GDP
AWAGE	Wage and Salaries
AINTS	Interest
ADIVS	Dividends
ATXPY	Personal Income
ASCHCI	Business Income
ASCHCL	Business Loss
ACGNS	Capital Gain
ASCHF	Schedule F Income
ASCHEI	Schedule E Income
ASCHEL	Schedule E Loss
AUNCOMP	Unemployed Compensation
ASOCSEC	Social Security
ACPIM	Medical CPI
ABOOK	Book Income
AIPD	Interest Paid

Appendix B PUF variables and blow-up factors

For each e-variable below, please refer to the spreadsheet for complete definition.

Factors	PUF variables
AGDPN	e03240
AWAGE	e00200
AINTS	e00300
	e00400
ADIVS	e00600
	e00650
ATXPY	e00700
	e00800
	e01400
	e01500
	e01700
	e03150
	e03210
	e03220

e03230
e03300
e03400
e03500
e07230
e07240
e07260
e07300
p08000
e09700
e09800
e09900
e10700
e10900
e59560
e59680
e59700
e59720
e11550
e11070
e11100
e11200
e11300
e11400
e11570
e11580
e11581
e11582
e11583
e10605
e18400
e18500
e19550
e19800
e20100
e19700
e20550
e20600
e20400
e20800
e20500
e21040
e32800
e33000
e53240
e53280

	e53410
	e53300
	e53317
	e53458
	e58950
	e58990
	p60100
	p61850
	e60000
	e62100
	e62900
	e62720
	e62730
	e62740
	p65300
	p65400
	e68000
	e82200
	t27800
	s27860
	p27800
	t27860
	s27860
	e87530
	e87550
ASCHCI	e00900
	e03260
	e30400
	e30500
ASCHCL	e00900
ACGNS	e01000
	e01100
	e01200
	p22250
	e22320
	e22370
	p23250
	e24515
	e24516
	e24518
	e24535
	e24560
	e24598
	e24615
	e24570
ASCHEI	e02000

	p25350
	p25470
	p25700
	e25820
	e25850
	e25860
	e25940
	e25980
	e25920
	e25960
	e26110
	e26170
	e26190
	e26160
	e26180
	e26270
	e26100
	e26390
	e26400
	e27200
ASCHEL	e02000
ASCHF	e02100
AUCOMP	e02300
ASOCSEC	e02400
	e02500
ACPIM	e03270
	e03290
	e17500
ABOOK	e07300
	e07400
AIPD	e19200

Appendix C Stage II targets

- US Population
- Total Returns
- Single Returns
- Joint Returns
- Head of Household Returns
- Number of Returns w/ Gross Security Income
- Number of Dependent Exemptions
- Taxable Interest Income
- Ordinary Dividends
- Business Income (Schedule C)
- Business Loss (Schedule C)
- Net Capital Gains in AGI
- Taxable Pensions and Annuities
- Supplemental Income (Schedule E)
- Supplemental Loss (Schedule E)

- Gross Social Security Income
- Unemployment Compensation
- Wages and Salaries: Zero or Less
- Wages and Salaries: $1 Less Than $10,000
- Wages and Salaries: $10,000 Less Than $20,000
- Wages and Salaries: $20,000 Less Than $30,000
- Wages and Salaries: $30,000 Less Than $40,000
- Wages and Salaries: $40,000 Less Than $50,000
- Wages and Salaries: $50,000 Less Than $75,000
- Wages and Salaries: $75,000 Less Than $100,000
- Wages and Salaries: $100,000 Less Than $200,000
- Wages and Salaries: $200,000 Less Than $500,000
- Wages and Salaries: $500,000 Less Than $1 Million
- Wages and Salaries: $1 Million and Over