Notes on the PSRC Data Weighting Process

Mark Bradley

July 10, 2014

Step 1: Do the first expansion based only on sampling probabilities for the different sampling groups. (These are groups that had equal sampling probabilities for all households within each group.)   We have 13 groups, with the following numbers of households in each group, in total and by county.:

Segment	Block group type	Hholds	Percent	King	Kitsap	Pierce	Snohomish
REG_LI	Regular Low Income BG	141	2.3	57	10	38	36
REG_MI	Regular Medium Income BG	836	13.7	316	82	222	216
REG_HI	Regular High Income BG	1789	29.4	906	104	393	386
OS_LI	Oversample Low Income BGI	746	12.2	308	64	286	88
OS_MI	Oversample Medium Income BG	373	6.1	207	52	72	42
OS_HI	Oversample High Income BG	413	6.8	316	57	10	30
UVOS_LI	Urban Village Low Income BG	573	9.4	573	0	0	0
UVOS_MI	Urban Village Medium Income BG	621	10.2	621	0	0	0
UVOS_HI	Urban Village High Income BG	268	4.4	268	0	0	0
BREG_MI	Bellevue Regular Low+Med Inc. BG	59	1.0	59	0	0	0
BREG_HI	Bellevue Regular High Income BG	134	2.2	134	0	0	0
BOS_MI	Bellevue Oversample Low+Med Inc	78	1.3	78	0	0	0
BOS_HI	Bellevue Oversample High Inc BG	63	1.0	63	0	0	0
Total	**Total**	**6094**	**100.0**	**3906**	**369**	**1021**	**798**

The data source for this step will be the most recent estimate of the number of households at the block group level- from the 5-year 2008-2012 ACS.

Step 2: Impute incomes for the households with missing income. This will be based on a model estimated on the 90% of households that are not missing income. The main variables used in the model will likely be:

- The income distribution in the residence block group, from the most recent 5 year ACS
- The number of working adults in the household with college post graduate education
- The number of working adults in the household with college undergraduate education
- The number of other full time workers in the household
- The number of other part time workers in the household
- The number of children in the household in different age groups
- Age group effects for adults in the household.
- Home ownership dummy
- Housing type dummy for detached single family

The most likely form of the model is ordered logit across income categories. A linear regression model is another option, although this would require using the midpoints of the reported income categories as the dependent variable.

The resulting model will be applied to the households with missing income to assign a categorical value (using a random Monte Carlo method in the case of the ordered logit model).

Step 3: Use iterative proportional fitting to adjust the expansion weights to fit the data to observed distributions along a number of dimensions. The targets for fitting will likely be:

- Planning district geography
- Household size (1, 2, 3, 4+)
- Number of workers (0, 1, 2+)
- Income group (categories to be determined)
- Number of vehicles (0, 1, 2+)
- Age of head of householder (Under 35, 35-64, 65 or older) by family/non-family HH
- Presence of children under 18 (yes, no)

An option is to do the IPF separately within each county, using the most recent year of ACS data (2012?), which is available at the county level. The exception is the planning district geography, which requires using tract-level data from the 3-year ACS (2010-2012).

Another option is to do the IPF separately within each planning district. This provides some more geographic accuracy in the weighting, but would require using the 3-year ACS for setting all the targets, and will likely produce more cases of small cell sizes and large expansion weights. (We could combine target categories in such cases where necessary.)

Even if we use the second option and the 3-year ACS, the expansion factors can be adjusted uniformly within each county to match the latest estimates of the number of households at the county level.

In regard to the income bands used for the expansion, we suggest using the 5 "broad" categories below that were used for the follow-up income question posed to people who refused to answer the more detailed income question. Using these categories, we only need to use imputed incomes for the 471 households who didn't answer either income question.

Household income 2013: Broad categories, all respondents (derived)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Under $25,000	769	12.6	12.6	12.6
	$25,000-$49,999	1146	18.8	18.8	31.4
	$50,000-$74,999	1046	17.2	17.2	48.6
	$75,000-$99,999	867	14.2	14.2	62.8
	$100,000 or more	1795	29.5	29.5	92.3
	Prefer not to answer	471	7.7	7.7	100.0
	Total	6094	100.0	100.0	