Small Area Estimation

David J Dorer

Introduction

Small areas estimation (SAE) is a method for starting with a dataset for a Large Area, for example a Public Use Microdata Area (PUMA) American Community Survey (ACS) PUMS dataset, and creating a "synthetic" dataset for a Small Area, for example a census tract or block group.

One reason for using SAE is because PUMA microdata has more variables with many more categories than are available in ACS tract and block group level Detail "B", Subject "S" and "DP" Data Profile tables.

For example, the ACS Detail table B01001 (AGE BY SEX) has a lowest age category of "Under 5 years." If you would like to have a estimate of the "Pre-K" population (3 and 4 years old), for a census tract you cannot get this information from the B01001 table. However using the Microdata AGEP variable, which is divided into 1 year age categories, you can easily define a "3 to 4 year old" age category for the PUMS data. Then using the Pre-K PUMS variable and applying the Poverty Assessment Toolkit SAE package, you will be able synthesize census tract data with a "Pre-K" variable. You can then tabulate the Pre-K variable in the synthetic tract data with other variables.

Computation Algorithm

The statistical algorithm or method for creating the tract and block group data is called "Iterative Proportional Fitting" or "Raking." The algorithm takes a cross tabulation of PUMA/PUMS categorical variables, the "seed" table, and adjusts the seed table so that its marginals match multiple tract or block group tables. The marginal tables are ACS Detail, Subject and Data Profile tables. The algorithm repeatedly cycles or iterates through the set of marginal or target tables.

An example illustrates the method. Start wit a 5-way cross tabulation "seed" table "Age x Sex x Race x Employed x Tenure (Owner v Renter)" along with and a set of 2 target marginal tables, "Age x Sex" and "Employed" from ACS Detail tables. The seed table is scaled, hence "proportional," so that the resulting updated seed table "Age x Sex" marginal matches the target "Age x Sex" table. Next the the seed table is scaled again so that the updated seed table (second update) "Employed" marginal matches the "Employed" target table. This cycling process is repeated again and again, i.e. "iterated" until the updated seed table marginals are close to the target tables.

The adjusted output seed table is then converted to a "synthetic" dataset with rows, columns and weights. There is one row for each "cell" in the output table and one column for each variable in the output table. The weight for a cell/row is the value of the cell in the adjusted table. The weights do not have to have integer values. This takes some time to get used to this but the value of a survey based estimate is not an nteger even though the data collected on the survey form started out as an integer. You can then make any multiway tabulation using the synthetic data. You can also stack/merge the tract data sets to get tables for other geographies, for example school districts.

Internal and External Validity of Estimates

Any statistical estimates of population sizes or associated rates should be checked for internal and external validity. For example the function `PAT.test.model()` cross checks the small area estimate of a synthetic data cross tabulation at the PUMA geography with the corresponding detail table at the same PUMA geography.

This is an example of an "internal validity" check. An example of an external validity check would be to compare a statistic for a geography that uses data sources other than ACS data. For example a statstic that is published by a State or Federal agency such as a school district. Many schools publish statistics for their school age population. You can compute the same or similar statistic using the PAT package and compare the results.

Some Background and an Example from Epidemiology

Iterative proportional fitting is a generalization of direct/indirect adjustment used in statistics. An early example is the Standardized Mortality Rate, which is the Crude Rate (counts of individual deaths) adjusted for the population age distribution. Standardized death rate calculations started with the first Life Table published in 1693 by Edmond Halley. A life table has the population and number of deaths for each year of age. Using a life table we can compute the age specific death rate for each age by dividing the number of deaths for each age by the population for that age. Age specific death rates depend on the population age distribution.

Suppose we want to see if we are winning the "War on Cancer." We could compare the "cause specific" crude death rate for cancer in 2023 to the same rate for 1980. Is this a fair way to make the comparison? We can do better than that. The age distribution in 1980 was substantially different from the age distribution in 2023. This is largely because of "baby boomers." In 1980 the baby boomers were 43 years younger than in 2023. We know that the cancer death rate increases with age, so we won't know if we are winning the War on Cancer unless we account for the difference in population ages between 2023 and 1980.

We do something simple. We take the 2023 age specific death rates from cancer. This done by starting with the 2023 cause specific (Cancer) life table. The age specific death rate is the population for a single age divided by the number of deaths for that age. Then we multiply the 2023 age specific death rates by the 1980 population for that age group. This is the "expected" number of 1980 deaths assuming the 2023 death rates and corresponds to the 2023 death rate adjusted to the 1980 population age distribution.

We add these adjusted/expected deaths to get a total number of expected deaths for the 2023 rate adjusted to agree with the 1980 population distribution. Next we compare the adjusted 2023 deaths to the 1980 actual number of deaths. In terms of rates, using the total 1980 population we can get the expected number of deaths per 100,000 population and compare that to the actual 1980 death rate per 100,000. We will be winning the War on Cancer if the 2023 rate adjusted to the 1980 population distribution is less than the actual 1980 death rate from cancer.

The ratio of the expected rate based on 2023 rates divided by the actual 1980 rate is called the Standardized Mortality Ratio (SMR). If the SMR < 1.0, then there has been an improvement in cancer deaths between 1980 and 2023. We have adjusted the 2023 life table to the 1980 population distribution. This "scaling" calculation the same as the calculation used with iterative proportional fitting except that IPF is a generalized multidimensional version.

To translate the SMR calculation into the calculations in the PAT package, consider the case of 10 age categories of 10 years each. Construct the 2 x 10 table with rows for each age decade and 2 columns for vital status, "alive" and "dead" based on how many people in each age decade have died. Now adjust this table age marginal to the table of population by age (by decade) of the 1980 population.

Next take the 1-way tabulation of mortality variable in this derived table. This table corresponds to the 2023 age specific life table adjusted to the 1980 population. Dividing the number of deaths in this adjusted 2023 table by the crude death rate for 1980 produces the Standardized Mortality Ratio or SMR. The usual SMR adjusts for both age and sex. For IPF age x sex adjustment corresponds to a calculation that starts with a 3 way Age x Sex x VitalStatus table and then adjust to the 2 way Age by Sex 1980 population marginal.

References

Internal and External Validity https://www.nber.org/system/files/working_papers/w26422/w26422.pdf

Age Adjustment https://www.cdc.gov/nchs/hus/sources-definitions/age-adjustment.htm

3.2 16 Dec 2023 19:54