**Sampling of tax return data**

**Introduction:** in order to collect a representative sample for tax returns for microsimulation the following methodology may be used. See Appendix for detailed explanation of the data distribution issues.

**Note:** Tax form refers to the type of form, say Form 1A, 1B, 2, etc. Tax return refers to the tax return submitted by an individual (or withheld tax) which could be of any form type.

**Methodology:** For each tax form do the following:-

1. Where there are multiple tax return forms or data sources for each taxpayer
   1. **(OPTION 1)** **– When it is possible to merge all tax return data**
      1. Link all the tax forms by taxpayer ID so that there is for each taxpayer, one row of all data from multiple tax forms;
      2. In order to do this, include all Tax IDs in one table. Merge other tax forms into this table;
      3. In MYSQL this would require a FULL OUTER JOIN Command;
      4. In STATA this would require a merge command and **all records are to be included** even if \_merge is 1, 2 or 3;
      5. Alternatively, take all tax returns from the tax form that has the most taxpayers. Merge into this all the other tax forms data using the FULL OUTER JOIN or merge command in STATA including all rows.
   2. **(OPTION 2) – When it is not possible to merge all tax return data**
      1. For each of the tax forms, include only two fields, the **Tax ID** and the **aggregate/total income**. Merge all these tax records as above using the FULL OUTER JOIN to make sure that no records are dropped during the merge.
2. For each category of tax forms store the total tax collection from all the tax returns. This information is used to check if the sample can replicate the total collection from all the tax returns.
3. For each individual, add up the aggregate incomes from each of the tax forms. Call this the **gross total income**;
4. Sort the tax forms on the gross total income;
5. Find the **lowest** and **highest gross total income**. Store these values;
6. Find the 10th to 100th percentiles of the gross total income–inc\_0, inc\_10, inc\_20, ..inc\_100 where inc\_0 is the lowest aggregate income; Note that inc\_0 should be a positive number i.e. inc\_0 = max(0, inc\_0);
7. Store the number of tax returns in each 10-percentile block;
8. For the last block (inc\_90<= aggregate income < inc\_100), **select all the tax returns** and **store the number of records in the block**;
9. For returns in the block (0<= aggregate income < inc\_10):-
10. Note the number of tax returns in this bracket. Call this num\_tax\_returns\_10 and **store this information** (they will be different for each block);
11. Randomly select **10,000** tax returns. If the number of tax returns in this group is less than 10,000 select all returns in that group.
12. Repeat for (inc\_10<= aggregate income < inc\_20)…until (inc\_80<= aggregate income < inc\_90);
13. Find the minimum of the negative incomes. **Store** this number;
14. Randomly select 10,000 tax returns. If the number of tax returns in this group is less than 10,000 select all returns in that group. **Store the number** of tax returns with negative incomes;
15. Note that, along with the selected tax returns data it is necessary to provide the total number of tax returns in each bracket from which the sample was taken (including those negative incomes) so that the correct weights may be calculated;
16. The above methodology should give about 100,000 tax records.
17. **(OPTION 1)** **– When it is possible to merge all tax return data,** the above method should give us the entire information from all tax forms.
18. **(OPTION 2) – When it is not possible to merge all tax return data**, the above method would give us the Tax IDs and gross total income. Use the selected Tax IDs to capture data from all the tax returns **using the INNER JOIN merge command in MySQL. In STATA this would be records with \_merge=3.**

**Appendix: Distribution of Incomes**

Incomes at the upper end tends to follow a pareto distribution which has a probability distribution functions

where is the pareto parameter. Diamond and Saez (2011) estimate this parameter to be 1.5 for tax returns in the United States.

**Figure – 1: Distribution of Incomes - an illustration**

A close up of a map

Description automatically generated

In Figure-1, a typical income distribution is illustrated where the incomes are normally distributed below a certain level of income and above that level is pareto distributed with the parameter of 1.5. In such a distribution the top 10% of individuals earn 28% of all the incomes of the population. The skewness in the distribution is also reflected in the fact that the median income is less than the mean income.

**Table 1: Income Distribution - illustration**

|  |  |  |
| --- | --- | --- |
| Percentile | Percentile position in population | Percentile income |
| 10% | 949 | 220,000 |
| 20% | 1,898 | 400,000 |
| 30% | 2,847 | 550,000 |
| 40% | 3,796 | 690,000 |
| 50% | 4,745 | 840,000 |
| 60% | 5,694 | 990,000 |
| 70% | 6,643 | 1,050,000 |
| 80% | 7,592 | 1,400,000 |
| 90% | 8,542 | 1,860,000 |
| 100% | 9,491 | 5,000,000 |

The distribution of tax contribution by incomes would be further skewed due to a progressive income tax. Figure 2 below shows that for a modestly progressive tax the contribution of tax could be highly skewed.

A screenshot of a cell phone

Description automatically generated

**Table 2: Distribution of Tax Revenue – an illustration**

|  |  |  |  |
| --- | --- | --- | --- |
| **Percentile** | **Revenue Percentile** | **Income level** | **Income Percentile** |
| 10% | 97,850,665 | 580,000 | 32% |
| 20% | 195,701,329 | 790,000 | 47% |
| 30% | 293,551,994 | 960,000 | 58% |
| 40% | 391,402,658 | 1,130,000 | 68% |
| 50% | 489,253,323 | 1,310,000 | 77% |
| 60% | 587,103,988 | 1,520,000 | 84% |
| 70% | 684,954,652 | 1,860,000 | 90% |
| 80% | 782,805,317 | 2,540,000 | 95% |
| 90% | 880,655,981 | 3,580,000 | 98% |
| 100% | 978,506,646 | 5,000,000 | 100% |

**Conclusion:** The discussion above indicates the importance of appropriately sampling the distribution of incomes in order tocapture the most information from the tax return for analysis. Ideally the goal is to obtain as much of a sample from the tax return population to equate the tax revenue from the samples across each income percentile group.