The development of a data linkage strategy for Statistics Flanders

Argho Aranya Sangma

Types of population data collection

- Population register: Continuously updated administrative database for individuals with demographic data.
- Population census: Periodically updated data with detailed information conducted by Government Authority.
- Survey: A detailed and extensive data collection from a representative sample of population measuring unobserved information such attitude, happiness and other life-satisfaction measures.
- Ideal Data collection: Collection of all the above-mentioned data on an individual basis. But unfortunately, there is no such data collection method.

Why Data Linkage

- Idea: Joint analysis of variables from two data-sets not jointly observed
- Record Linkage: Exact matching of individual level data from two data sources (requires personal identification number)
- Statistical matching: Similar records based on common variables are considered statistical matches.

Sample A (recipient)	Sample B (Donor)	Matched data-set
X, Y		
	X, Z	X, Y, Z

Table: Statistical matching situation.

The data sets can be overlapping of non-overlapping.

Data-sets

- Recipient: Survey data from 2016
- Donor: Population register and individual education data from 2016

Common variables (X)	Survey variables (Y)	Variables for imputation (Z)
Demographic, Age, gender, origin, region, Parent's origin, education	Various survey questionnaire measuring happiness index	No. of children, Civil status (married/non-married)

Key Question

What can we learn from variables not jointly observed?

Objectives

1. Check for marginal and joint distributions of variables in donor, recipient and matched data so that survey efficiency and sample selection can be improved. (Cross-sectional)
2. After matching with population register, look for same individuals in a recent population register for relative changes, and compare to the survey questionnaire. (Longitudinal-study)

Research questions:

1. Do people with higher self-declared happiness index tend to have more children in the near future?
2. Do unmarried individuals with higher self-declared happiness index has more probability to marry in the near future?

Step-wise approach

Selection of proper matching technique is only one of the steps, and not often very essential.

- Consistency of common variables: The common variables should have same definition in both data-sets
- Similarity: The common variables should show similar marginal distribution
- Categorical Frequency: Some population may be over-represented (e.g., Belgian, European) than African, Asians, Americans. For low frequency, similar groups can be merged into another category.
- Quality of common variables: Using variables with a lot of missing values or imputed values to be avoided.
- Power of explanatory variables: We may consider to perform a linear regression with no. of children as dependent variables and common variables as independent variables from population register.

Step-wise approach

- Target population: For fertility behavior we may stick to individuals between 20 to 60 years old for males and 20 to 50 years old for females, which will also give better age distribution.
- Distance metric selection: Common variables has continuous (Age) and categorical variables. Gower's distance (D'Orazio, 2021) is a good measurement of distance. 0 implies lowest distance and 1 implies highest distance. Ideally, we want a distance of 0 for a statistical match, but no threshold value is defined yet.

$$d_{G,ij} = 1 - s_{G,ij} = \frac{\sum_{t=1}^{p} \delta_{ijt}\, d_{ijt}}{\sum_{t=1}^{p} \delta_{ijt}}$$

Statistical matching methods

- **Random hot-deck**: In case of multiple matching pairs, a random record is taken for imputation.
  - Advantage: The value exists for real.
  - Disadvantage: The value may not be close to the actual value.
- **Average hot-deck**: From multiple records with lowest distance, the average value is imputed.
  - Advantage: The value is close to the real values, only if all the records with smallest distance have close values.
  - Disadvantage: The value is not real and will not preserve the variance and co-variance structure. Only to be used if all other records with same distance have close values.

Test Statistics: Paired t-test

Pair	Before	After	Difference
1	2	2	0
2	2	5	3
3	1	3	2
4	0	0	0
5	2	2	0
6	3	5	2
7	1	1	0
8	3	3	0
9	2	3	1
10	1	2	1
		Mean	0.9

Mean = 0.9

Standard deviation = 1.10

Standard error = $\frac{Standard\ deviation}{\sqrt{n}} = \frac{1.10}{\sqrt{10}} = 0.348$

t = $\frac{mean}{standard\ error} = \frac{0.9}{0.348} = 2.58$

- t value to be compared for significance
- To be calculated for three groups, lower, middle and higher happiness index
- 3 models, low, medium and high happiness index

Assumptions:

Difference is normally distributed

Test Statistics: Mcnemar test

Pair	Before	After
1	0	0
2	0	1
3	1	1
4	0	0
5	0	1
6	1	1
7	0	0
8	0	1
9	1	1
10	0	0

Before	After		
	0	1	Sum
1	0 (A)	3 (B)	3
0	4 (C)	3 (D)	7
Sum	4	6	10

A: Married becomes non-married (0)
B: Married remains married (3)
C: Non-married remains non-married
D: Non-married gets married (3)

Follows chi-sq statistics

$$\chi^2 = \sum_{i=1}^{k} \frac{(fo-fh)^2}{fh}$$

fo: observed f in category i
fh: expected f in category i
Consider changes in each cell
A: 1 to 0, D: 0 to 1
Chi-sq simplified to $\frac{(A-D)^2}{A+D}$.

Correction: $\chi^2 = \frac{(|A-D|-1)^2}{A+D}$.
With df = 1

We can consider 3 models,
Low, medium and high happiness index

Alternatives

1. Construct a liner model with no. of children as dependent and happiness index as independent variable on the matched data-set.
   - H0: There is no change in fertility with happiness.
   - H1: Higher happiness index is associated with more fertility

2. Construct a Logistic regression model with marriage factor as dependent and happiness index as independent variable.
   - H0: The probability of marriage is not affected by happiness index.
   - H1: The probability of marriage is positively affected by happiness index.

Thank You

Gower's Distance (D'Orazio, 2021)

$$d_{G,ij} = 1 - s_{G,ij} = \frac{\sum_{t=1}^{p} \delta_{ijt}\, d_{ijt}}{\sum_{t=1}^{p} \delta_{ijt}}$$

$1 - s_{ijt}$ is distance on *t*th variable

- s_{ijt} is similarity between i (from survey) and j (from census) for *t*th variable
- δ_{ijt} stands for missingness for the point
- Distance in continuous variables normalized between 0 to 1

Table 1 – Calculation of the Gower's similarity by type of variable

Type of variable	s_{ijt}	δ_{ijt}	**Note**
Binary symmetric	1 if $x_{it} = x_{jt}$ 0 if $x_{it} \neq x_{jt}$ 0 if x_{it} or x_{jt} or both are missing	1 if both the variables are nonmissing 0 if x_{it} or x_{jt} or both are missing	s_{ijt} corresponds to the *simple matching coefficient*
Binary asymmetric	1 if $x_{it} = x_{jt} = 1$ 0 otherwise 0 if x_{it} or x_{jt} or both are missing	1 if both the variables are nonmissing 0 if $x_{it} = x_{jt} = 0$ 0 if x_{it} or x_{jt} or both are missing	s_{ijt} corresponds to the *Jaccard index*
Categorical nominal (more than two categories)	1 if $x_{it} = x_{jt}$ 0 if $x_{it} \neq x_{jt}$ 0 if x_{it} or x_{jt} or both are missing	1 if both the variables are nonmissing 0 if x_{it} or x_{jt} or both are missing	s_{ijt} is the simple matching on the untransformed variable or to the *Dice* (*Czekanovsky-Sorerensen*) *measure* applied to the dummies obtained by transforming the original variable
Measured on an interval or ratio scale	$1 - \lvert x_{it} - x_{jt} \rvert / R_t$ 0 if x_{it} or x_{jt} or both are missing	1 if both the variables are nonmissing 0 if x_{it} or x_{jt} or both are missing	$R_t = \max(x_t) - \min(x_t)$ is the range of the kth variable $1 - s_{ijt}$ is *the Manhattan* or *city-block distance* scaled by the range