Planned Missing-Data Designs and Statistical Matching: A Smart Response to Minimising Total Survey Error?

Femke De Keulenaer and Robert Manchin

Gallup Europe

GALLUP

Gallup World Poll
(since 2006, 140+ countries, annual data collection)

GALLUP® World

Search Gallup.com | Search

HOME | POLITICS | ECONOMY | WELLBEING | WORLD | ELECTION 2012 | SOCIAL & ECONOMIC ANALYSIS | MANAGEMENT CONSULTING

May 25, 2012

Economic Crisis Has Lasting Effect on Wellbeing Worldwide

Wellbeing in negative territory in Middle East and North Africa

by Gale Muller and Julie Ray

March 15, 2011

Worldwide, Good Jobs Linked to Higher Wellbeing

Those who are self-employed tend to have the lowest wellbeing

by Jon Clifton and Jenny Marlar

June 22, 2011

Low-Income Britons Struggle With Their Wellbeing

Physical health, healthy behaviors, access to basics all worse for low-income Britons

by Anna Manchin

"In an ideal world,..." ⟷ TSE trade-offs

Well-being = multidimensional concept, with many predictors

- financial security
- health
- social contacts, etc.

↓

LONG QUESTIONNAIRE

↓

- Higher costs, smaller samples
- Lower response rates
- More item non-response, more measurement error

Two data sources: A and B *(same target population)*

Y_1	$\dots$	Y_Q	X_1	$\dots$	X_P	Z_1	$\dots$	Z_R
y^A_{11}	$\dots$	y^A_{1Q}	x^A_{11}	$\dots$	x^A_{1P}			
	$\dots$			$\dots$			**A**	
y^A_{a1}	$\dots$	y^A_{aQ}	x^A_{a1}	$\dots$	x^A_{aP}			
	$\dots$			$\dots$				
$y^A_{n_A 1}$	$\dots$	$y^A_{n_A Q}$	$x^A_{n_A 1}$	$\dots$	$x^A_{n_A P}$			
			x^B_{11}	$\dots$	x^B_{1P}	z^B_{11}	$\dots$	z^B_{1R}
				$\dots$			$\dots$	
B			x^B_{b1}	$\dots$	x^B_{bP}	z^B_{b1}	$\dots$	z^B_{bR}
				$\dots$			$\dots$	
			$x^B_{n_B 1}$	$\dots$	$x^B_{n_B P}$	$z^B_{n_B 1}$	$\dots$	$z^B_{n_B R}$

A and B share a set of *variables X*

	Y_1	$\ldots$	Y_Q	X_1	$\ldots$	X_P	Z_1	$\ldots$	Z_R
	y^A_{11}	$\ldots$	y^A_{1Q}	x^A_{11}	$\ldots$	x^A_{1P}			
		$\ldots$			$\ldots$				
A	y^A_{a1}	$\ldots$	y^A_{aQ}	x^A_{a1}	$\ldots$	x^A_{aP}			
		$\ldots$			$\ldots$				
	$y^A_{n_A1}$	$\ldots$	$y^A_{n_AQ}$	$x^A_{n_A1}$	$\ldots$	$x^A_{n_AP}$			
				x^B_{11}	$\ldots$	x^B_{1P}	z^B_{11}	$\ldots$	z^B_{1R}
					$\ldots$			$\ldots$	
B				x^B_{b1}	$\ldots$	x^B_{bP}	z^B_{b1}	$\ldots$	z^B_{bR}
					$\ldots$			$\ldots$	
				$x^B_{n_B1}$	$\ldots$	$x^B_{n_BP}$	$z^B_{n_B1}$	$\ldots$	$z^B_{n_BR}$

But the *variables Y and Z* are not jointly observed

	Y_1	$\ldots$	Y_Q	X_1	$\ldots$	X_P	Z_1	$\ldots$	Z_R
	y^A_{11}	$\ldots$	y^A_{1Q}	x^A_{11}	$\ldots$	x^A_{1P}			
		$\ldots$			$\ldots$				
A	y^A_{a1}	$\ldots$	y^A_{aQ}	x^A_{a1}	$\ldots$	x^A_{aP}			
		$\ldots$			$\ldots$				
	$y^A_{n_A1}$	$\ldots$	$y^A_{n_AQ}$	$x^A_{n_A1}$	$\ldots$	$x^A_{n_AP}$			
				x^B_{11}	$\ldots$	x^B_{1P}	z^B_{11}	$\ldots$	z^B_{1R}
					$\ldots$			$\ldots$	
B				x^B_{b1}	$\ldots$	x^B_{bP}	z^B_{b1}	$\ldots$	z^B_{bR}
					$\ldots$			$\ldots$	
				$x^B_{n_B1}$	$\ldots$	$x^B_{n_BP}$	$z^B_{n_B1}$	$\ldots$	$z^B_{n_BR}$

Objective = investigate relationship between Y and Z

“Micro” vs. “Macro” approach

Pre-processing steps in applying statistical matching

- Choice of **target variables** Y and Z (*not jointly observed*)

  Example: well-being and net household income

- Identification of all **common variables** X (*with the same marginal/joint distribution*)

  + harmonisation step, if necessary

  Example: gender, age, marital status, level of education, employment status, health problems etc.

- Choice of **matching variables** (*linked to matching framework: e.g. parametric/non-parametric/mixed*)

Example data from Gallup World Poll (Bulgaria)

- Examples of statistical matching in R environment (*StatMatch*)
- Artificial data set derived from Gallup World Poll (±2000 respondents in Bulgaria)
- Data set split randomly in two (equal) parts:
  - rec.A and don.B share the variables X.vars
  - the respondents' WB score (y.var) is available in rec.A
  - net household income (z.var) is available in don.B
- Selecting "best" matching variables:
  - relationship between Y and X explored in rec.A
  - relationship between Z and X explored in don.B
  - RESULT: "best" predictors are gender, age, level of education, employment status and health problems

Non-parametric micro approaches

	Recipient				Donor			
ID	*Gender*	*Age category*	*Education (in years)*	**Wellbeing**	*Gender*	*Age category*	*Education (in years)*	**Income**
1	Man	15-29	11	**5**	Man	15-29	12	**9,130.4**
2	Man	30-49	16	**9**	Man	30-49	16	**16,956.5**
3	Woman	30-49	11	**7**	Woman	30-49	11	**5,151.3**
4	Woman	50-64	12	**7**	Woman	50-64	11	**6,521.7**
5	Woman	64+	14	**8**	Woman	64+	14	**12,234.5**

Random hot deck (donation classes)

Random selection of each donor from a “suitable” subset/donation class

Non-parametric micro approaches

	Recipient				Donor			
ID	*Gender*	*Age category*	*Education (in years)*	**Wellbeing**	*Gender*	*Age category*	*Education (in years)*	**Income**
1	Man	15-29	11	**5**	Man	15-29	12	**9,130.4**
2	Man	30-49	16	**9**	Man	30-49	16	**16,956.5**
3	Woman	30-49	11	**7**	Woman	30-49	11	**5,151.3**
4	Woman	50-64	12	**7**	Woman	50-64	11	**6,521.7**
5	Woman	64+	14	**8**	Woman	64+	14	**12,234.5**

Nearest neighbour distance hot deck

Function that searches, for each case in the recipient file, the nearest neighbour in the donor file

(to reduce computation effort, combined with donation classes)

- Distance functions: Manhattan, Gower's dissimilarity etc.
- Constrained and unconstrained matching (size of rec.A and don.B)

Conditional Independence (CI) assumption

- Traditional SM methods, that use the set of common variables X to match A and B, implicitly assume the **conditional independence of Y and Z given X**:

  $f(x,y,z) = f(y \mid x) \times f(z \mid x) \times f(x)$

- ***Very strong assumption that usually does not hold in practice***

- Solution: incorporate **auxiliary information** about the relationship between Y and Z
  1. a 3rd file where (x,Y,Z) or (Y,Z) are jointly observed
  2. plausible value for inestimable parameter of (Y, Z|X) or (Y,Z)
- Alternative: **assessing "uncertainty"** (interval of plausible values)

Mixed methods

Three steps:

1) Estimation of parameters of two regression models (regression step)

rec.A: $Y=\alpha+\beta X$

don.B: $Z=\delta+\gamma X$

(= more parsimonious)

2) Data sets filled with “intermediate values”

$z_a = z_a + e_a$ $(a=1,\ldots,n_A)$

$y_b = y_b + e_b$ $(a=1,\ldots,n_B)$

adding a random residual to predicted values

3) Each record in A is filled with a donor from B according to constrained distance hot deck (matching step)

Mahalanobis distance, considering both “intermediate” and “live” values

(= “protection” against model misspecification)

Results of matching – some examples

	Descriptives of income variable				Corr. WB & income
	Mean	SD	Skewness	Kurtosis	
"complete" file	9582.1	(7675.0)	*3.23*	*19.94*	**.348**
"matched" files					
(non-par) **distance hot deck**	9478.5	(7437.8)	*3.07*	*18.96*	**.172**
(mixed) **ML** estimates, rho.yz=0 **(= CI)**	9151.0	(6171.6)	*2.41*	*2.41*	**.200**
(mixed) **ML** estimates, rho.yz=.22 **(auxiliary info)**	9177.2	(6219.6)	*2.41*	*2.39*	**.318**
(mixed) MS estimates, rho.yz=.35	9151.4	(6171.1)	*1.44*	*2.41*	**.344**

Weak correlation (incorrect CI assumption)

More "normal" distribution due to parametric approach

GALLUP

Donation classes/matching variables:
gender, age, education, employment and health problems

Discussion

- Opportunities
  - Reducing cost and response burden
  - Potential for using data with less measurement error etc.
- Challenges
  - Selection of best matching approach: (non-)parametric, distance function, etc.
  - “Advanced” topics: complex survey designs, etc.
  - Selection of matching variables (conditional independence)
  - *How to deal with uncertainty in matching results?*
  - *Need for “good” auxiliary information*

Thank you!

Contact:

Femke De Keulenaer

Gallup Europe, Brussels, Belgium

femke_de_keulenaer@gallup-europe.be

GALLUP