

Modelos para estimar gross-flows en encuestas complejas con ausencia de respuesta

3ESAMP - Jovenes Investigadores

Andrés Gutiérrez
Universidad Santo Tomás

hugogutierrez@usantotomas.edu.co

2011

Considere una población finita como un conjunto de N , donde $N < \infty$, unidades que conforman el universo de estudio. N es comúnmente llamado el tamaño poblacional. Cada elemento perteneciente a la población puede ser identificado por un rótulo. Sea U el conjunto de rótulos, tal que

$$U = \{1, \dots, k, \dots, N\}.$$

Se supone que se accede a la población en dos periodos distintos y se pregunta a cada uno de los miembros acerca de su clasificación en ambos tiempos.

Distribución poblacional de Gross-Flows y no-respuesta

Si se pudiese llegar a toda la población de interés en dos ocasiones sucesivas, se podría establecer la siguiente distribución poblacional, referente a las contingencias poblacionales.

Tiempo $t - 1$	Tiempo t						Complemento fila
	1	2	...	j	...	G	
1	N_{11}	N_{12}	...	N_{1j}	...	N_{1G}	R_1
2	N_{21}	N_{22}	...	N_{2j}	...	N_{2G}	R_2
...
...
i	N_{i1}	N_{i2}	...	N_{ij}	...	N_{iG}	R_i
...
...
G	N_{G1}	N_{G2}	...	N_{Gj}	...	N_{GG}	R_G
Complemento columna	C_1	C_2	...	C_j	...	C_G	M

Cuadro: *Distribución del cambio neto poblacional en dos periodos consecutivos.*

- N_{ij} denota el número total de individuos, en toda la población de interés, que tienen clasificación i en tiempo $t - 1$ y j en el tiempo t .
- Igualmente, R_i denota el número total de individuos, en toda la población de interés, que no respondieron la entrevista en el tiempo t , aunque sí lo hicieron en el tiempo $t - 1$ y tienen clasificación i en tiempo $t - 1$.
- Luego, C_j denota el número total de individuos, en toda la población de interés, que no respondieron la entrevista en el tiempo $t - 1$, aunque sí lo hicieron en el tiempo t y tienen clasificación j en tiempo t .
- Finalmente, M , denota el número total de individuos en toda la población de interés que no respondieron la entrevista en ninguno de los dos tiempos.

- N_{ij} denota el número total de individuos, en toda la población de interés, que tienen clasificación i en tiempo $t - 1$ y j en el tiempo t .
- Igualmente, R_i denota el número total de individuos, en toda la población de interés, que no respondieron la entrevista en el tiempo t , aunque sí lo hicieron en el tiempo $t - 1$ y tienen clasificación i en tiempo $t - 1$.
- Luego, C_j denota el número total de individuos, en toda la población de interés, que no respondieron la entrevista en el tiempo $t - 1$, aunque sí lo hicieron en el tiempo t y tienen clasificación j en tiempo t .
- Finalmente, M , denota el número total de individuos en toda la población de interés que no respondieron la entrevista en ninguno de los dos tiempos.

- N_{ij} denota el número total de individuos, en toda la población de interés, que tienen clasificación i en tiempo $t - 1$ y j en el tiempo t .
- Igualmente, R_i denota el número total de individuos, en toda la población de interés, que no respondieron la entrevista en el tiempo t , aunque sí lo hicieron en el tiempo $t - 1$ y tienen clasificación i en tiempo $t - 1$.
- Luego, C_j denota el número total de individuos, en toda la población de interés, que no respondieron la entrevista en el tiempo $t - 1$, aunque sí lo hicieron en el tiempo t y tienen clasificación j en tiempo t .
- Finalmente, M , denota el número total de individuos en toda la población de interés que no respondieron la entrevista en ninguno de los dos tiempos.

- N_{ij} denota el número total de individuos, en toda la población de interés, que tienen clasificación i en tiempo $t - 1$ y j en el tiempo t .
- Igualmente, R_i denota el número total de individuos, en toda la población de interés, que no respondieron la entrevista en el tiempo t , aunque sí lo hicieron en el tiempo $t - 1$ y tienen clasificación i en tiempo $t - 1$.
- Luego, C_j denota el número total de individuos, en toda la población de interés, que no respondieron la entrevista en el tiempo $t - 1$, aunque sí lo hicieron en el tiempo t y tienen clasificación j en tiempo t .
- Finalmente, M , denota el número total de individuos en toda la población de interés que no respondieron la entrevista en ninguno de los dos tiempos.

El tamaño total de la población de interés, debe satisfacer la siguiente expresión:

$$N = \sum_i \sum_j N_{ij} + \sum_j C_j + \sum_i R_i + M$$

Mediante la definición de las siguientes características de interés, es posible definir los parámetros de interés.

$$y_{1ik} = \begin{cases} 1, & \text{si el individuo } k\text{-ésimo responde en } t - 1 \text{ y tiene clasificación } i; \\ 0, & \text{en otro caso.} \end{cases} \quad (1)$$

$$y_{2jk} = \begin{cases} 1, & \text{si el individuo } k\text{-ésimo responde en } t \text{ y tiene clasificación } j; \\ 0, & \text{en otro caso.} \end{cases} \quad (2)$$

Por lo anterior, el producto de las anteriores cantidades, definido como $y_{1ik}y_{2jk}$, provee una nueva característica de interés que toma el valor uno, si el individuo contestó en ambos periodos y está clasificado en la celda ij y cero para cualquier otro caso. De esta forma, se tiene que

$$N_{ij} = \sum_{k \in U} y_{1ik}y_{2jk} \quad (3)$$

Además se definen las siguientes características dicotómicas

$$z_{1k} = \begin{cases} 1, & \text{si el individuo } k\text{-ésimo responde en } t - 1; \\ 0, & \text{en otro caso.} \end{cases} \quad (4)$$

$$z_{2k} = \begin{cases} 1, & \text{si el individuo } k\text{-ésimo responde en } t; \\ 0, & \text{en otro caso.} \end{cases} \quad (5)$$

Por lo tanto, se tiene que

$$N_{ij} = \sum_{k \in U} y_{1ik} y_{2jk} \quad (6)$$

$$R_i = \sum_{k \in U} y_{1ik} (1 - z_{2k}) \quad (7)$$

$$C_j = \sum_{k \in U} y_{2jk} (1 - z_{1k}) \quad (8)$$

$$M = \sum_{k \in U} (1 - z_{1k})(1 - z_{2k}) \quad (9)$$

Estimadores de los parámetros

Siendo w_k el peso de muestreo del k -ésimo individuo inducido por una estrategia de muestreo, entonces las siguientes expresiones representan estimadores de los parámetros de interés

$$\hat{N}_{ij} = \sum_{k \in S} w_k y_{1ik} y_{2jk} \quad (10)$$

$$\hat{R}_i = \sum_{k \in S} w_k y_{1ik} (1 - z_{2k}) \quad (11)$$

$$\hat{C}_j = \sum_{k \in S} w_k y_{2jk} (1 - z_{1k}) \quad (12)$$

$$\hat{M} = \sum_{k \in S} w_k (1 - z_{1k})(1 - z_{2k}) \quad (13)$$

para N_{ij} , R_i , C_j y M , respectivamente.

Modelo para la ausencia de respuesta: primera etapa

Acogemos el modelo de Stasny (1987), quien supone que las observaciones son el resultado de un proceso de dos etapas, donde en la primera etapa (no observable), los individuos son ubicados dentro de las celdas de una matriz $G \times G$ de acuerdo con las probabilidades de una cadena de Markov.

Modelo para la ausencia de respuesta: primera etapa

Sea η_i la probabilidad inicial de que un individuo esté en el estado i en el tiempo $t - 1$ ($\sum_i \eta_i = 1$), y p_{ij} la probabilidad de transición desde el estado i al estado j ($\sum_j p_{ij} = 1$).

Modelo para la ausencia de respuesta: segunda etapa

En esta etapa cada individuo en la celda ij de la matriz puede:

- ser no respondiente en el tiempo $t - 1$ y perder la clasificación por fila,
- o ser no respondiente en el tiempo t y perder la clasificación columna,
- o bien, ser no respondiente en ambos tiempos y perder ambas clasificaciones.

Modelo para la ausencia de respuesta: segunda etapa

En esta etapa cada individuo en la celda ij de la matriz puede:

- ser no respondiente en el tiempo $t - 1$ y perder la clasificación por fila,
- o ser no respondiente en el tiempo t y perder la clasificación columna,
- o bien, ser no respondiente en ambos tiempos y perder ambas clasificaciones.

Modelo para la ausencia de respuesta: segunda etapa

En esta etapa cada individuo en la celda ij de la matriz puede:

- ser no respondiente en el tiempo $t - 1$ y perder la clasificación por fila,
- o ser no respondiente en el tiempo t y perder la clasificación columna,
- o bien, ser no respondiente en ambos tiempos y perder ambas clasificaciones.

Modelo para la ausencia de respuesta: segunda etapa

Sea $\psi(i, j)$ la probabilidad inicial de que un individuo en la celda ij responda en el tiempo $t-1$, $\rho_{RR}(i, j)$ la probabilidad de transición de ser respondiente en el tiempo $t-1$ a ser respondiente en el tiempo t , y $\rho_{MM}(ij)$ la probabilidad de transición de ser no respondiente en el tiempo $t-1$ a ser no respondiente en el tiempo t .

En el contexto de este modelo de dos etapas, las probabilidades para la matriz de datos son como se muestra en la tabla

Tiempo $t - 1$	Tiempo t						
	1	2	...	j	...	G	Comp. fila
1							
2							
⋮							
i				$\{\eta_i p_{ij} \psi(i, j) \rho_{RR}(i, j)\}$		$\{\sum_j \eta_i p_{ij} \psi(i, j) (1 - \rho_{RR}(i, j))\}$	
⋮							
G							
Comp. col.	$\{\sum_i \eta_i p_{ij} (1 - \psi(i, j)) (1 - \rho_{MM}(i, j))\}$					$\sum_i \eta_i p_{ij} \sum_j (1 - \psi(i, j)) \rho_{MM}(i, j)$	

Cuadro: probabilidades para la distribución del cambio neto en dos periodos consecutivos.

Dado que los datos son observados sólo después de la segunda etapa, se quiere realizar inferencias para las probabilidades tanto de la cadena de Markov que genera los datos como la cadena que genera la ausencia de respuesta.

Función de verosimilitud

La función de verosimilitud para los datos observados bajo este modelo de dos etapas es proporcional a

$$L = \prod_i \prod_j [\psi(i, j) \rho_{RR}(i, j) \eta_i p_{ij}]^{N_{ij}} \times \prod_i \left[\sum_j \psi(i, j) (1 - \rho_{RR}(i, j)) \eta_i p_{ij} \right]^{R_i} \\ \times \prod_j \left[\sum_i (1 - \psi(i, j)) (1 - \rho_{MM}(i, j)) \eta_i p_{ij} \right]^{C_j} \times \left[\sum_i \sum_j (1 - \psi(ij)) \rho_{MM}(i, j) \eta_i p_{ij} \right] \\ (14)$$

Con $i, j = 1, \dots, G$.

La log-verosimilitud para los datos observados en toda la población de interés se puede reescribir como

$$l = \sum_{k \in U} f_k(\psi, \rho_{RR}, \rho_{MM}, \eta, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2) \quad (15)$$

$$\begin{aligned} &= \sum_{k \in U} \left[\sum_i \sum_j y_{1ik} y_{2jk} \ln(\psi(i, j) \rho_{RR}(i, j) \eta_i p_{ij}) \right. \\ &+ \sum_i y_{1ik} (1 - z_{2k}) \ln \left(\sum_j \psi(i, j) (1 - \rho_{RR}(i, j)) \eta_i p_{ij} \right) \\ &+ \sum_j y_{2jk} (1 - z_{1k}) \ln \left(\sum_i (1 - \psi(i, j)) (1 - \rho_{MM}(i, j)) \eta_i p_{ij} \right) \\ &\left. + (1 - z_{1k})(1 - z_{2k}) \ln \left(\sum_i \sum_j (1 - \psi(i, j)) \rho_{MM}(i, j) \eta_i p_{ij} \right) \right] \quad (16) \end{aligned}$$

Es necesario maximizar la anterior función, se tiene que la ecuaciones de verosimilitud están dadas por

$$\sum_{k \in U} \mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{0} \quad (17)$$

donde los vectores \mathbf{u}_k , a los que se les conoce como *scores*, se definen como

$$\mathbf{u}_k(\boldsymbol{\theta}) = \frac{\partial f_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (18)$$

De esta forma, considerando a w_k como la ponderación inducida por la estrategia de muestreo, entonces un estimador insesgado para esta expresión se define como $\sum_{k \in S} w_k \mathbf{u}_k(\boldsymbol{\theta})$. Las siguientes expresiones se conocen con el nombre de ecuaciones de pseudo-verosimilitud y constituyen un método efectivo para encontrar estimadores de los parámetros del modelo que dependan de los pesos de la estrategia de muestreo.

$$\sum_{k \in S} w_k \mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{0} \quad (19)$$

Luego, es posible obtener un estimador $\hat{\boldsymbol{\theta}}$, al resolver el sistema de ecuaciones resultante.

ATENCIÓN: IMPOSIBLE

Sin embargo, en el modelo hay $4G^2 + G$ parámetros con $G + 1$ restricciones para los parámetros en la anterior función de verosimilitud y dado que la función de verosimilitud está supeditada a la tabla de contingencia, pues el modelo en dos etapas es inducido por los datos poblacionales observados, entonces sólo se cuenta con $(G + 1)^2$ celdas de conteos observados o estimados.

Luego, se debe reducir el número de parámetros del modelo. De otra forma no existiría solución a las ecuaciones de verosimilitud.

Para este modelo reducido, se considera que la probabilidad inicial de que un individuo sea respondiente en el tiempo $t - 1$ es la misma para todas las clasificaciones contempladas en la encuesta, es decir $\psi(i, j) = \psi$.

También, las probabilidades de transición entre respondientes y entre no respondientes no dependen de la clasificación del individuo en la encuesta, es decir $\rho_{MM}(i, j) = \rho_{MM}$ y $\rho_{RR}(i, j) = \rho_{RR}$.

Bajo los supuestos del modelo A, las estimaciones resultantes del método de máxima pseudo-verosimilitud para ψ , ρ_{MM} y ρ_{RR} están dadas por

$$\hat{\psi}_{mpv} = \frac{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j + \hat{M}} \quad (20)$$

$$\hat{\rho}_{RR,mpv} = \frac{\sum_i \sum_j \hat{N}_{ij}}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i} \quad (21)$$

$$\hat{\rho}_{MM,mpv} = \frac{\hat{M}}{\sum_j \hat{C}_j + \hat{M}} \quad (22)$$

respectivamente.

Bajo los supuestos del modelo A, las estimaciones resultantes del método de máxima pseudo-verosimilitud para η_i y p_{ij} se obtienen mediante la iteración hasta convergencia de las siguientes expresiones

$$\hat{\eta}_{i,mpv}^{(v+1)} = \frac{\sum_j \hat{N}_{ij} + \hat{R}_i + \sum_j \left(\frac{\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)}}{\sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)}} \right)}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j} \quad (23)$$

$$\hat{p}_{ij,mpv}^{(v+1)} = \frac{\hat{N}_{ij} + \frac{\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)}}{\sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)}}}{\sum_j \hat{N}_{ij} + \sum_j \left(\frac{\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)}}{\sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)}} \right)} \quad (24)$$

respectivamente. En donde el superíndice (v) determina el valor de la estimación de los parámetros de interés en la iteración v -ésima.

Ahora, una pregunta de interés estaría determinada por la escogencia de los valores iniciales $\{\hat{\eta}_i^{(0)}\}$ y $\{\hat{\rho}_{ij}^{(0)}\}$. En general, cualquier conjunto de valores es válido, siempre y cuando respeten las restricciones iniciales. Es decir,

$$\sum_i \hat{\eta}_i^{(0)} = 1$$

$$\sum_j \hat{\rho}_{ij}^{(0)} = 1$$

Sin embargo, notando que bajo el hipotético caso de que todos los individuos fuesen respondientes en ambos periodos, entonces se tendría que $M = 0$, $R_i = 0$ (para todo $i = 1, \dots, G$) y $C_j = 0$ (para todo $j = 1, \dots, G$). Por lo tanto, sus estimaciones por muestreo serían también nulas. Luego, apelando a las expresiones de los estimadores resultantes, un escogencia prudente está dada por

$$\hat{\eta}_i^{(0)} = \frac{\sum_j \hat{N}_{ij}}{\sum_i \sum_j \hat{N}_{ij}}$$

$$\hat{p}_{ij}^{(0)} = \frac{\hat{N}_{ij}}{\sum_j \hat{N}_{ij}}$$

- 1 Utilizando la aproximación de Taylor de primer orden, es muy sencillo probar que los estimadores del modelo de Markov son aproximadamente insesgados (*design unbiased*) para su contraparte poblacional.
- 2 La varianza se puede hallar siguiendo procedimientos de MPV o enfoques computacionales como Jackknife o Bootstrap.

- 1 Utilizando la aproximación de Taylor de primer orden, es muy sencillo probar que los estimadores del modelo de Markov son aproximadamente insesgados (*design unbiased*) para su contraparte poblacional.
- 2 La varianza se puede hallar siguiendo procedimientos de MPV o enfoques computacionales como Jackknife o Bootstrap.

Utilizando el método de la linealización de Taylor, la varianza asintótica, basada en el diseño de muestreo, de un estimador de máxima pseudo-verosimilitud está dado por

$$V_p(\hat{\theta}_s) \approx [J(\theta_U)]^{-1} V_p \left[\sum_{k \in s} d_k u_k(\theta_U) \right] [J(\theta_U)]$$

donde $V_p[\sum_{k \in s} d_k u_k(\theta_U)]$ es la matriz de covarianzas, basada en el diseño de muestreo, del estimador del total poblacional de *scores* y

$$J(\theta_U) = \left. \frac{\partial \sum_{k \in U} u_k(\theta)}{\partial \theta} \right|_{\theta = \theta_N}$$

Ahora, un estimador para $V_p(\hat{\theta}_s)$ está dado por

$$\hat{V}_p(\hat{\theta}_s) = [\hat{J}(\hat{\theta}_s)]^{-1} \hat{V}_p \left[\sum_{k \in S} d_k u_k(\hat{\theta}_s) \right] [\hat{J}(\hat{\theta}_s)]$$

donde $\hat{V}_p[\sum_{k \in S} d_k u_k(\hat{\theta}_s)]$ es un estimador consistente para la matriz de covarianzas del estimador del total poblacional de *scores* y

$$\hat{J}(\hat{\theta}_s) = \left. \frac{\partial \sum_{k \in S} d_k u_k(\theta)}{\partial \theta} \right|_{\theta = \hat{\theta}_s}$$

Por lo tanto, la distribución asintótica de $\hat{\theta}_s$ es normal multivariada puesto que

$$\hat{V}_p \left[\sum_{k \in s} d_k u_k(\hat{\theta}_s) \right]^{-1/2} (\hat{\theta}_s - \theta_N) \sim NM(\mathbf{0}, \mathbf{I})$$

Lo anterior ofrece una base para la correcta inferencia cuando se utilizan muestras grandes, justo como en el caso de encuestas de tipo laboral.

Simulación: probabilidades de la primera etapa

#Prob de clasificiación en $t-1$
 $\eta = c(0.9, 0.05, 0.05)$

#Prob de clasificiación en t para C1
 $p_1 = c(0.8, 0.15, 0.05)$

#Prob de clasificiación en t para C2
 $p_2 = c(0.3, 0.6, 0.1)$

#Prob de clasificiación en t para C3
 $p_3 = c(0.1, 0.1, 0.8)$

Simulación: probabilidades de la segunda etapa

#Creación de la respuesta en $t-1$ para todos
 $\psi=0.8$

#Creación de la respuesta en t para
los que respondieron en $t-1$
 $\rho_{RR}=0.9$

#Creación de la Norespuesta en t para
los que no respondieron en $t-1$
 $\rho_{MM}=0.7$

Simulación: distribución de los GF y respuestas

> Distribución de los Gross-Flows

		U2		
U1		E1	E2	E3
E1	7230	1341	471	
E2	122	297	45	
E3	35	50	409	

> Distribución de la No respuesta

		Resp2	
Resp1		NR	R
NR	1323	624	
R	823	7230	

Simulación: distribución en ambos periodos

	E1	E2	E3	Ri
E1	5210	976	340	745
E2	82	220	33	43
E3	26	35	308	35
Cj	426	94	53	M=1374

SESGADO: diseños de muestreo con probabilidades desiguales

```
> bias.eta
      E1      E2      E3
[1,] -30.33216 578.9645 -32.98562

> bias.pij
      E1      E2      E3
E1  0.2394436 -2.449707  3.518022
E2 -7.9304497  4.419315 -2.724544
E3 11.7506739 -2.613555 -1.142140

> bias.psi
[1,] 6.5485
> bias.rhoRR
[1] -0.783732
> bias.rhoMM
[1,] 0.5862668

> bias.GF
      E1      E2      E3
E1  0.3509447 -1.941254 -2.545795
E2 -1.3540533 -2.106892  5.734192
E3 26.9894021 -6.359187  2.709465
```

INSESGADO: diseños de muestreo con probabilidades iguales

```
> bias.eta
      E1      E2      E3
[1,] -0.0009578087  0.009218608  0.008021948

> bias.pij
      E1      E2      E3
E1 0.0003030149 -0.002318661  0.0021077431
E2 0.0019337093 -0.006874866  0.0354480706
E3 0.0380055022 -0.041378244  0.0004215928

> bias.psi
[1,] 0.000832875
> bias.rhoRR
[1] 0.00002102504
> bias.rhoMM
[1,] 0.00681159

> bias.GF
      E1      E2      E3
E1 0.0008590477 -0.001801076  0.004115975
E2 0.0134191260 -0.010503022 -0.015733773
E3 0.0216589589 -0.029735065 -0.008010320
```

INSESGADO: diseños de muestreo con probabilidades desiguales

```
> bias.eta
      E1      E2      E3
[1,] -0.0005621629 0.02546946 -0.01535053

> bias.pij
      E1      E2      E3
E1 0.0008028766 0.0003441968 -0.01387862
E2 -0.0524155375 0.0225986342 0.02165481
E3 -0.0153727443 -0.0658441165 0.01015211

> bias.psi
[1,] -0.0007014436
> bias.rhoRR
[1] -0.001197169
> bias.rhoMM
[1,] 0.004372776

> bias.GF
      E1      E2      E3
E1 9.447637e-05 0.0001960196 0.001703437
E2 -2.712068e-02 -0.0201865530 -0.004235081
E3 5.646701e-02 0.0175990017 0.013445806
```

- Establecer más modelos reducidos.
- Estudios empíricos sobre los resultados encontrados.
- Aplicación práctica en una encuesta del Brasil.

- Establecer más modelos reducidos.
- Estudios empíricos sobre los resultados encontrados.
- Aplicación práctica en una encuesta del Brasil.

- Establecer más modelos reducidos.
- Estudios empíricos sobre los resultados encontrados.
- Aplicación práctica en una encuesta del Brasil.

OBRIGADO