

Metodología de la Encuesta Continua de Hogares

Muestreo II

Licenciatura en Estadística

2023

- ▶ inicio = 1968
- ▶ encuesta multipropósito que releva información (datos) de los hogares y personas
- ▶ indicadores oficiales de:
 - ▶ mercado laboral ¡ML! (e.g. actividad, empleo y desempleo)
 - ▶ ingresos de hogares y personas
 - ▶ pobreza de hogares y personas
 - ▶ condiciones de vida (e.g. educación, salud, etc.)
- ▶ **periodicidad:** ML e ingresos mensuales; resto anual (por el INE)

- ▶ **condición de elegibilidad** = hogares particulares situados en viviendas particulares en todo el país (Inicio =2006)
- ▶ **cross-section** (i.e. muestras mensuales independientes)
- ▶ **diseño muestral**: aleatorio, estratificado, por conglomerados y en dos etapas de selección.

- ▶ **asignación de la muestra:** proporcional por estrato en base a la cantidad de viviendas (provenientes del censo 2011) ajustada por tasa de elegibilidad y respuesta.
- ▶ **tamaño de muestra:**
 - ▶ 3400 hogares mensuales (elegibles respondentes iER_i)
 - ▶ 41000 hogares en el año
- ▶ relevamiento 100% presencial (CAPI)
- ▶ cuestionario único (condiciones de vida + ML + ingresos)

método de estimación (hasta 2019)

- ▶ distintos indicadores/parámetros θ son estimados utilizando **estimadores directos** teniendo en cuenta únicamente a los ER y en base a sistemas de ponderadores $w_i \forall i \in \text{ER}$ que son computados de forma mensual, trimestral y anual.
- ▶ por ejemplo, la estimación del total en la población (U) de una variable cualquiera y ,

$$Y = \sum_{i \in U} y_i$$

es estimado como:

$$\hat{Y} = \sum_{i \in \text{ER}} w_i \times y_i$$

método de estimación (hasta 2019)

- ▶ estimadores **calibrados/regresión** para la producción de las estimaciones de los distintos indicadores θ
- ▶ Vector/set para calibrar/modelar $\mathbf{x}_i = (x_{1i}, \dots, x_{ji})^T$ son variables de pertenencia (indicadoras) a distintos grupos demográficos (e.g. tramos de edad, sexo y departamento)

regression thinking

$$E_m(y_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \mu + \text{edad}_j + \text{sexo}_k + \text{dpto}_l$$

calibration thinking

$$\sum_{i \in ER} w_i \times \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i$$

el ponderador/peso de la persona i queda definido como:

$$w_i = d_i \times \hat{\phi}_i^{-1} \times g_i$$

donde

$$d_i = (\text{Prob}[i \in s])^{-1} = \frac{\text{cantidad viviendas marco en el estrato } h}{\text{tamaño de muestra en el estrato } h} = \frac{N_h}{n_h}$$

$\hat{\phi}_i$ es la propensión de estimada del hogar/persona i de responder.

Se asume MAR y un modelo de igualdad de medias a nivel de estrato.

$g_i = 1 + \left(\sum_{i \in U} \mathbf{x}_i - \sum_{i \in ER} w_i^{\text{nr}} \mathbf{x}_i \right)^T \left(\sum_{i \in s} w_i^{\text{nr}} \times \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{x}_i$ es el ajuste de calibración

la estimación de un total puede verse bajo dos enfoques:

regression thinking

$$\hat{Y}_{RE} = \sum_{i \in U} \hat{y}_i + \sum_{i \in ER} w_i^{nr} \times (y_i - \hat{y}_i)$$

con $\hat{y}_i = \mathbf{x}_i^T \hat{\mathbf{B}}$ en donde los parámetros del modelo son estimado con los datos de la propia muestra y los ponderadores ajustados por no respuesta w_i^{nr} .

calibration thinking

$$\hat{Y} = \sum_{i \in ER} w_i \times y_i$$

donde w_i son los ponderadores calibrados (i.e. cumplen las ecuaciones de calibración)

construcción matriz para la calibración

hogar	persona	sexo	edad	dpto
1	1	H	45	MVD
1	2	M	40	MVD
2	1	H	24	MVD
2	2	H	21	MVD
2	3	M	24	MVD
3	1	H	53	MVD
3	2	M	55	CAN
3	3	M	27	CAN
3	4	M	15	CAN
4	1	H	41	CAN

hogar	persona	X1	X2	X3	X4	X5	X6
1	1	0	1	0	0	1	0
1	2	0	0	0	1	1	0
2	1	1	0	0	0	1	0
2	2	1	0	0	0	1	0
2	3	0	0	1	0	1	0
3	1	0	1	0	0	1	0
3	2	0	0	0	1	0	1
3	3	0	0	0	1	0	1
3	4	0	0	1	0	0	1
4	1	0	1	0	0	0	1

variables indicadoras representando grupos de sexo/edad y de región: X1 hombres (0-24), X2 hombres (25+), X3 mujeres (0-24) X4 mujeres (25+), X5 Mvd, X6 Can

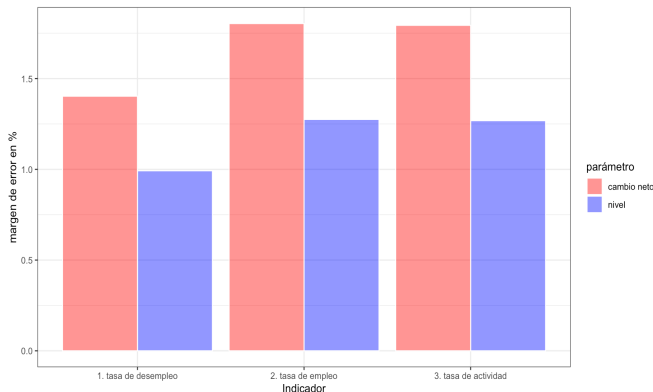
- ▶ variables de calibración \mathbf{x} se encuentran definidas a nivel de persona
- ▶ ECH produce estimaciones a nivel de hogares y personas
- ▶ necesidad de consistencia (sistema de ponderación único **uniweight**)
- ▶ para producir ponderadores w_i iguales para cada individuo del hogar se utiliza el método de integración.
- ▶ las variables X previo a la calibración son sustituidas por el promedio de las mismas a nivel de hogar.

observación: produce consistencia en las estimaciones pero incrementa los errores estándar (SE) producto del aumento en la variabilidad de los ponderadores finales w_i

ECH tiene un tamaño de muestra suficientemente grande para estimar ciertos indicadores (con periodicidad baja ¡e.g. pobreza!) y pequeña para indicadores claves con periodicidad alta (e.g. ML)

- ▶ diseño y tamaño muestral no es suficiente para brindar estimaciones confiables de ML de forma mensual
 - ▶ moe de 1% para las estimaciones mensuales del nivel de las tasas ML
 - ▶ capaz de detectar cambios significativos (al 95%) de un mes a otro si son mayores a 1.3%
- ▶ diseño y tamaño "ideal" para estimar pobreza de forma anual
 - ▶ moe de 0.2 % a nivel total país

moe del nivel y Δ de las tasas mensuales



los moe (al 95%) para las estimaciones de los niveles mensuales y cambios netos, respecto al mes anterior, se encuentran por encima del 1%.

- ▶ encuesta de panel rotativo
- ▶ la muestra de un mes está compuesta por seis paneles (grupos de rotación)
- ▶ relevamiento mixto:
 - ▶ **CAPI**: implantación que releva variables habituales de la ECH por única vez
 - ▶ **CATI**: seguimiento solo ML durante 5 meses +

objetivos

- ▶ mejorar las estimaciones de ML
 - ① reducir los moe para las estimaciones mensuales tanto de nivel como el cambio neto respecto a la ECH 2019
 - ② producir indicadores confiables para una amplia gama de áreas de estimación de forma mensual (e.g. educación, género, departamentos, etc)
- ▶ mantener la calidad de las estimaciones para el resto de los indicadores que brinda la ECH

independientemente del tipo de encuesta, diseño, método de estimación; la mejora en las precisiones depende en mayor medida del tamaño de muestra (i.e. la cantidad de hogares/personas que participan en la estimación)

- ▶ la ECH pasa a ser una encuesta de panel rotativo con periodicidad mensual, en donde, la muestra de un mes, está compuesta por seis paneles (grupos de rotación).

sistema de rotación

mes de la encuesta (estimación)											
Grupo de rotación (GR)	2021						2022				
	7	8	9	10	11	12	1	2	3	4	5
1	1ra	2do	3ra	4ta	5ta	6ta	abandona la encuesta				
2		1ra	2do	3ra	4ta	5ta	6ta				
3			1ra	2do	3ra	4ta	5ta	6ta			
4				1ra	2do	3ra	4ta	5ta	6ta		
5					1ra	2do	3ra	4ta	5ta	6ta	
6						1ra	2do	3ra	4ta	5ta	6ta
1							1ra	2do	3ra	4ta	5ta
2								1ra	2do	3ra	4ta
3									1ra	2do	3ra
4										1ra	2do
5											1ra
6											

- ▶ cada GR es implantado presencialmente y aplicando un formulario similar a la ECH 2019
- ▶ un GR está conformado por 2000 hogares ER aproximadamente

- ▶ el solapamiento entre las muestras de un mes a otro puede ser explotado también en la etapa de estimación.
- ▶ existen estimadores que utilizan las estimaciones del ML del mes anterior $\hat{\theta}_{t-1}$ para mejorar las precisiones de los estimadores del mes t , ya sea, para el nivel $\hat{\theta}_t$, como para el cambio neto (i.e. $\hat{\Delta} = \hat{\theta}_t - \hat{\theta}_{t-1}$)

dos estrategias distintas:

- ▶ estimadores compuestos K o AK (CPS)
- ▶ estimadores de **calibración/regresión compuestos (CRE)**

Ambos métodos explotan el solapamiento y la información del GR que abandona la encuesta.

- ▶ CRE reduce los moe del nivel y cambio neto mientras siguen cumpliendo la expansión a conteos demográficos
- ▶ Fácil de instrumentar en R utilizando paquetes existentes (e.g. survey). Simplemente se aumentan las variables de calibración.
- ▶ Estimación de la varianza se vuelve compleja y no se encuentra implementada en ningún paquete.

se implementa el estimador de calibración/regresión compuesto (**CRE**)

- ▶ añade variables extras a la ecuación de calibración para reducir el moe de la **estimación del nivel (MR1)**
- ▶ las variables nuevas son construidas con variables indicadoras del mes anterior \mathbf{z}_{t-1} del status de la persona en ML (e.g. ocupado en $t - 1$)

$$\mathbf{z}_i^L = \begin{cases} \mathbf{z}_{i,t-1} & \text{si } i \in \text{ER}_t - \text{ER}_t^b \\ \hat{\mathbf{Z}}/N_{\text{PET}} & \text{si } i \in \text{ER}_t^b \end{cases} \quad (1)$$

- ▶ lo anterior implica que si estimamos el total de la variable \mathbf{z}_i^L coincide con la estimación del mes anterior, i.e,

$$\hat{\mathbf{Z}} = \sum_{i \in \text{ER}} w_i \times \mathbf{z}_i^L.$$

- ▶ añade variables extras a la ecuación de calibración para reducir el moe de la **estimación del cambio (MR2)**

$$\mathbf{z}_i^C = \begin{cases} \mathbf{z}_{i,t-1} + (R^{-1} - 1)(\mathbf{z}_{i,t-1} - \mathbf{z}_{i,t}) & \text{si } i \in \text{ER}_t - \text{ER}_t^b \\ \mathbf{z}_{i,t} & \text{si } i \in \text{ER}_t^b \end{cases} \quad (2)$$

donde $R = 5/6$ es la tasa de solapamiento ponderada.

- ▶ al igual que con el estimador MR1, si estimamos el total de la variable \mathbf{z}_i^C coincide con la estimación del mes anterior

Para las personas incluidas en la **muestra de solapamiento**, las variables z_c , a modo de ejemplo, quedan definidas como:

persona	ocupado en t	ocupado en t-1	Z ocupado C
1	1	1	1
2	1	0	-0.2
3	0	1	1.2
4	0	0	0

importante: se deben generar tantas variables \mathbf{z}^L y \mathbf{z}^C , como parámetros a estimar (e.g. totales de empleo, desempleo y actividad) así como definir las no solo a nivel de todo el país, sino también para distintas áreas de estimación (e.g. sexo, región, etc.)

- ▶ **problema:** añadir muchas variables a la ecuación de calibración puede provocar estimaciones inestables producto de la existencia ponderadores w_i extremos
- ▶ **solución (una):** construir variables \mathbf{z} como una combinación lineal ponderada de \mathbf{x}^L y \mathbf{x}^C

$$\mathbf{z}_i = (1 - \alpha)\mathbf{z}_i^L + \alpha\mathbf{z}_i^C$$

donde $\alpha = 2/3$ se elige como un compromiso entre reducción de los moe de los estimadores de nivel y cambio neto.

Los ponderadores w_i^{cre} son obtenidos minimizando la función de distancia lineal entre w^{nr} y los ponderadores w^{cre} , sujetos a la ecuación de calibración

$$\sum_{i \in ER} w_i^{cre} \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_i \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Z} \end{pmatrix}$$

los ponderadores $w_i^{cre} = w_i^{nr} \times g_i$ en donde g_i es el factor de ajuste proveniente de calibración y el mismo es:

$$g_i^{cre} = 1 + (\mathbf{x}_i^T, \mathbf{z}_i^T) \left(\sum_{i \in ER} w_i^{nr} (\mathbf{x}_i^T, \mathbf{z}_i^T)^T (\mathbf{x}_i^T, \mathbf{z}_i^T) \right)^{-1} \left((\mathbf{X}^T, \mathbf{Z}^T)^T - \sum_{i \in ER} w_i^{nr} (\mathbf{x}_i^T, \mathbf{z}_i^T)^T \right)$$

44 grupos de edad/sexo/región (Montevideo e Interior)

- ▶ 14 a 17 años, 18 a 24 años, 25 a 29 años, 30 a 34 años, 35 a 39 años, 40 a 44 años, 45 a 50 años, 50 a 54 años, 55 a 64 años, 65 a 74 años, 75 o + años

Departamento (14 o + años)

Grupo de rotación (14 o + años)

variables auxiliares compuestas

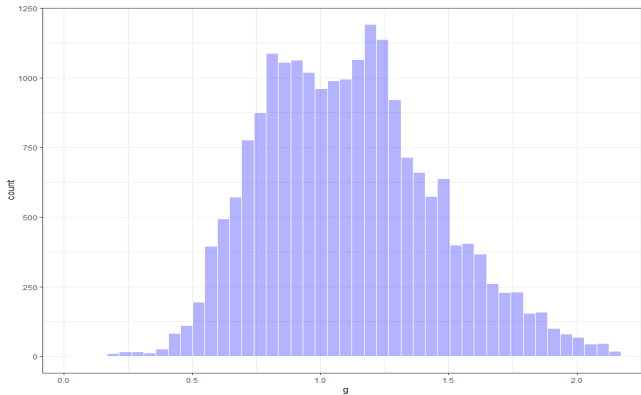
- ▶ Personas ocupadas de 14 o + años a nivel total país
- ▶ Personas desocupadas de 14 o + años a nivel total país
- ▶ Personas inactivas de 14 o + años a nivel total país
- ▶ Hombres ocupados de 14 o + años a nivel total país
- ▶ Hombres desocupados de 14 o + años a nivel total país
- ▶ Hombres inactivos de 14 o + años a nivel total país
- ▶ Mujeres ocupadas de 14 o + años a nivel total país
- ▶ Mujeres desocupadas de 14 o + años a nivel total país
- ▶ Mujeres inactivas de 14 o + años a nivel total país
- ▶ Personas ocupadas de 14 o + años por departamento
- ▶ Personas desocupadas de 14 o + años por departamento
- ▶ Personas inactivas de 14 o + años por departamento

- ▶ para la creación de los vectores \mathbf{z}_i^L y \mathbf{z}_i^C se imputan los datos faltantes del grupo de rotación de nacimiento.
- ▶ pueden existir datos faltantes de las variables indicadoras de ML en $t - 1$ para los nuevos miembros del hogar o integrantes que comienzan a formar parte de la PET.
- ▶ para los nuevos miembros se imputan los datos de $t - 1$ utilizando donantes aleatorios.
- ▶ para los integrantes que comienzan a participar en la PET se les imputa cero.

tratamiento de los ponderadores finales menores a uno

- ▶ la función de distancia lineal para la calibración puede permitir excepcionalmente que existan algunos ponderadores menores a uno e incluso negativos
- ▶ desde el punto de vista teórico, la existencia de ponderadores negativos no repercute en el insesgamiento del estimador
- ▶ por convención e incluso por "interpretación", los ponderadores de las unidades (hogares y personas) deben ser mayores o iguales a uno.
- ▶ una vez realizada la calibración compuesta a todas las unidades que tengan ponderadores menores a uno, se les asigna un ponderador final igual a 1.

factores de ajuste g



- ▶ se simularon 5000 réplicas bajo un SIR
- ▶ para cada réplica se calcularon las estimaciones de nivel $\hat{\theta}_t$ y cambio neto $\hat{\Delta}_t = \hat{\theta}_t - \hat{\theta}_{t-1}$ de ML
- ▶ se computaron las eficiencias relativas, la cuales, se definen como el cociente entre varianzas

cambio en el método de estimación en ML

- ▶ RCE vs RE nivel ocupados **90%**
- ▶ RCE vs RE nivel desocupados **91%**
- ▶ RCE vs RE cambio neto ocupados **54%**
- ▶ RCE vs RE cambio neto desocupado **85%**

se utilizan técnicas de remuestreo o réplicas (Bootstrap)

- ▶ se parte de los ER del mes y se generan 1000 sistemas de ponderadores bootstrap

$$w_{kjh}^b = \frac{n_h}{n_{h-1}} m_{hj}^b \times w_{kjh}^{nr}$$

- ▶ para cada uno de los sistemas de ponderadores bootstrap se realiza la calibración compuesta. Actualmente los conteos o benchmarks $\hat{\mathbf{Z}}$ se consideran fijos.
- ▶ peor escenario = subestima la varianza

Bootstrap coordinado

- 1 cuando las UPM son las mismas de un mes a otro, se les asigna los factores de multiplicidad del mes anterior.
- 2 cuando solo algunas UPM se mantiene en los dos meses ($5/6$ aprox) se transfieren los factores de multiplicidad de las UPM en común. EL resto ($1/6$) pertenecientes al GR de nacimiento, se les asigna un factor de multiplicidad de forma aleatoria correspondiente a las UPM que abandonan la encuesta.

- 1 a los ponderadores bootstrap se les aplica el factor de ajuste por no respuesta (único).
- 2 utilizando los ponderadores finales bootstrap del mes anterior se computan los totales $\hat{\mathbf{Z}}^{(b)}$ para cada una de las réplicas. Esto se hace para agregarle el componente estocástico a los conteos compuestos.
- 3 los ponderadores bootstrap generados en el paso 1 son calibrados a los conteos demográficos y a los conteos de las variables auxiliares compuestas para la réplica bootstrap correspondiente.

- ▶ la matriz de pesos Bootstrap permite obtener estimaciones de la varianza para los estimadores cross-section (nivel); y pueden ser utilizados con cualquier software que estime varianza para diseños muestrales complejos.
- ▶ las estimaciones de las varianzas para los cambios netos son más complejas y no pueden ser llevadas a cabo con softwares tradicionales.
- ▶ se indica a los usuarios utilizar los pesos Bootstrap para la estimación de niveles y aplicar formulas que tengan en cuenta la correlación entre muestras y el porcentaje de solapamiento. Por ejemplo, la estimación del error estándar (SE) del cambio neto $\hat{\Delta} = \hat{\theta}_t - \hat{\theta}_{t-1}$ es:

$$\widehat{SE}(\hat{\Delta}) = \sqrt{(1 - R\rho)} \sqrt{\widehat{SE}^2(\hat{\theta}_t) + \widehat{SE}^2(\hat{\theta}_{t-1})}$$