ESTIMACIÓN EN AREAS PEQUEÑAS

Isabel Molina y J. Miguel Marín

Dep. de Estadística, Univ. Carlos III de Madrid

J.N.K. Rao

School of Mathematics and Statistics, Carleton University

INTRODUCCIÓN A LA ESTIMACIÓN EN ÁREAS PEQUEÑAS

ESTIMADORES INDIRECTOS TRADICIONALES

MODELO BÁSICO A NIVEL DE ÁREA

MODELO BÁSICO A NIVEL UNIDAD

MÉTODO EB PARA INDICADORES DE POBREZA

MODELOS PARA DATOS BINARIOS

INFERENCIA BASADA EN EL DISEÑO/MODELO

ELEMENTO	BAJO UN MODELO	BAJO EL DISEÑO
Población	$y \sim P_\theta$	$U = \{1, \ldots, N\}$, $\mathcal{Y} = \{y_1, \ldots, y_N\}$
Muestra	$\mathbf{y} = (y_1, \ldots, y_n)$ y_i i.i.d. as y	$s = (i_1, \ldots, i_n) \in S_\pi$, $\mathbf{y} = (y_{i_1}, \ldots, y_{i_n})$
Distr. Prob.	$P_\theta(\mathbf{y})$	$P_\pi(s)$
Parámetro	θ (e.g., $\theta = E_{P_\theta}(y)$)	$\theta = h(y_1, \ldots, y_N)$
Estimador	$\hat{\theta}(\mathbf{y})$	$\hat{\theta}(s)$

Diseño muestral: (S_π, P_π), $S_\pi \subset \mathcal{P}(U)$ conjunto de muestras, P_π distribución de probabilidad sobre S_π donde $P_\pi(s) > 0$, $\forall s \in S_\pi$, y todas las unidades $j \in U$ están contenidas en alguna muestra $s \in S_\pi$.

INFERENCIA BASADA EN EL DISEÑO

- U población **finita** de tamaño N.
- $y_1, \ldots, y_N$ mediciones para las unidades poblacionales (fijas).
- **Parámetro objetivo:**

$$\delta = h(y_1, \ldots, y_N).$$

- Ejemplo: **media** poblacional

$$\bar{Y} = \frac{1}{N} \sum_{j=1}^{N} y_j.$$

- s **muestra** aleatoria de tamaño n obtenida de una población U bajo un diseño dado.
- $r = U - s$ **no-muestra** (tamaño $N - n$).

ESTIMADOR DE HORVITZ-THOMPSON

- π_j probabilidad de inclusión de la unidad j en la muestra.
- $d_j = 1/\pi_j$ peso muestral de la unidad j.
- **Horvitz-Thompson (HT)** estimador de la media:

$$\hat{\bar{Y}} = \frac{1}{N} \sum_{j \in s} \frac{y_j}{\pi_j} = \frac{1}{N} \sum_{j \in s} d_j y_j.$$

- **Varianza bajo el diseño**:

$$V_\pi(\hat{\bar{Y}}) = \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} (\pi_{j,k} - \pi_j \pi_k) \frac{y_j}{\pi_j} \frac{y_k}{\pi_k},$$

$\pi_{j,k}$ probabilidad de inclusión conjunta de las unidades j y k.

✓ *Hansen, Hurwitz & Madow (1953)*

VARIANZA BAJO EL DISEÑO

- **Estimador insesgado de la varianza bajo el diseño**:

$$\hat{V}_\pi(\hat{\bar{Y}}) = \frac{1}{N^2} \sum_{j \in s} \sum_{k \in s} \frac{\pi_{j,k} - \pi_j \pi_k}{\pi_{j,k}} \frac{y_j}{\pi_j} \frac{y_k}{\pi_k},$$

  ✓ *Särndal, Swensson & Wretman (1992), ecuación (5.8.5)*

- Usando la aproximación $\pi_{j,k} \cong \pi_j \pi_k$, $j \neq k$,

$$\hat{V}_\pi(\hat{\bar{Y}}) \cong \frac{1}{N^2} \sum_{j \in s} \left(\frac{1 - \pi_j}{\pi_j^2} \right) y_j^2 = \sum_{j \in s} d_j (d_j - 1) y_j^2.$$

EJEMPLO: INDICADORES DE POBREZA FGT

- E_j medida del poder adquisitivo para individuo j (ej. ingreso neto anual por unidad de consumo).
- z umbral de pobreza: En países de la UE,

$$z = 0.6 \times \text{Mediana}(E_j).$$

- Familia de indicadores de pobreza FGT:

$$F_\alpha = \frac{1}{N} \sum_{j=1}^{N} \left(\frac{z - E_j}{z} \right)^\alpha I(E_j < z), \quad \alpha \geq 0.$$

- $\alpha = 0 \Rightarrow$ Tasa/Incidencia de Pobreza
- $\alpha = 1 \Rightarrow$ Brecha de Pobreza

✓ *Foster, Greer & Thornbecke (1984), Econometrica*

ESTIMADOR DE HORVITZ-THOMPSON

- Indicador de pobreza:

$$F_\alpha = \frac{1}{N}\sum_{j=1}^{N} F_{\alpha j}, \quad F_{\alpha j} = \left(\frac{z - E_j}{z}\right)^\alpha I(E_j < z).$$

- Estimador HT de F_α:

$$\hat{F}_\alpha = \frac{1}{N}\sum_{j\in s} d_j F_{\alpha j}.$$

- Varianza estimada:

$$\hat{V}_\pi(\hat{F}_\alpha) = \frac{1}{N^2}\sum_{j\in s} d_j(d_j - 1)F_{\alpha j}^2.$$

AJUSTES DEL ESTIMADOR DE HT

- g_j factor de ajuste del peso muestral d_j, $j \in s$.
- $w_j = d_j g_j$ peso ajustado, $j \in s$.
- Estimador con pesos ajustados:

$$\hat{\bar{Y}}^A = \frac{1}{N} \sum_{j \in s} w_j y_j.$$

EJEMPLO 1: ESTIMADOR DE RAZÓN

- Estimador HT del tamaño poblacional:

$$\hat{N} = \sum_{j \in s} d_j.$$

- Factor de ajuste constante:

$$g_j = \frac{N}{\hat{N}}, \quad \forall j \in s.$$

- **Estimador de razón**:

$$\hat{\bar{Y}}^R = \frac{\hat{Y}}{\hat{N}}, \quad \hat{Y} = \sum_{j \in s} d_j y_j.$$

EJEMPLO 2: EST. RAZÓN CON VARIABLE AUX.

- $X = \sum_{j=1}^{N} x_i$ total conocido de variable auxiliar x con valores poblacionales:

$$x_1, \ldots, x_N.$$

- Estimador HT de X:

$$\hat{X} = \sum_{j \in s} d_j x_j.$$

- Factor de ajuste:

$$g_j = \frac{X}{\hat{X}}, \quad \forall j \in s.$$

- Estimador de razón con variable auxiliar X:

$$\hat{\bar{Y}}^{RX} = \hat{\bar{Y}} \frac{X}{\hat{X}}.$$

- El estimador de razón anterior se obtiene tomando $x_j = 1$, $\forall j \in U$.

EJEMPLO 3: CALIBRACIÓN

- p variables auxiliares con totales poblacionales conocidos X_k, $k = 1, \ldots, p$.
- **Idea:** Encontrar pesos w_j, $j \in s$, que minimicen la distancia $\mathcal{X}^2$

$$\begin{aligned} \text{mín} \quad & \sum_{j \in s} \frac{\left(w_j - d_j\right)^2}{d_j} \\ s.t. \quad & \sum_{j \in s} w_j x_{jk} = X_k, \ k = 1, \ldots, p. \end{aligned}$$

- **Solución:** $w_j = d_j g_j$, donde $g_j = 1 + \mathbf{x}_j' \hat{\mathbf{T}}^{-1} (\mathbf{X} - \hat{\mathbf{X}})$,

$$\mathbf{x}_j = (x_{j1}, \ldots, x_{jp})', \quad \mathbf{X} = (X_1, \ldots, X_p)', \quad \hat{\mathbf{T}} = \sum_{j \in s} d_j \mathbf{x}_j \mathbf{x}_j'.$$

✓ *Deville & Särndal (1992), JASA*

EJEMPLO 3: CALIBRACIÓN

- Modelo de regresión lineal:

$$y_j = \mathbf{x}_j'\boldsymbol{\beta} + e_j, \quad E(e_j) = 0,\ E(e_j^2) = \sigma_e^2,\ j = 1, \ldots, N.$$

- Estimador de coeficientes de la regresión:

$$\hat{\mathbf{B}} = \hat{\mathbf{T}}^{-1} \sum_{j \in s} d_j \mathbf{x}_j y_j$$

- Estimador de regresión generalizada (GREG):

$$\hat{Y}^A = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})'\hat{\mathbf{B}}.$$

- ¡Coincide con el estimador de calibración!

✓ *Deville & Särndal (1992), JASA*

ESTIMACIÓN EN DOMINIO/ÁREA

- U particionada en D dominios $U_1, \ldots, U_D$ de tamaños $N_1, \ldots, N_D$.
- s_d muestra de tamaño n_d obtenida de U_d.
- Tamaño muestral total $n = \sum_{d=1}^{D} n_d$.
- $r_d = U_d - s_d$ complemento de la muestra, de tamaño $N_d - n_d$.

Ejemplo: Encuesta de condiciones de vida en 2006

Tamaño muestral total: $n = 34,389$ personas.
Resumen de tamaños muestrales por provincia×género:

(Barcelona,M)	(Córdoba,M)	(Tarragona,V)	(Soria,M)
1483	230	129	17

ESTIMADORES TRADICIONALES DIRECTOS

- Parámetro objetivo: $\delta_d = h_d(\{y_j; j \in U_d\})$.
- Ejemplo: media del dominio d-ésimo

$$\bar{Y}_d = \frac{1}{N_d} \sum_{j \in U_d} y_j.$$

- **Estimador directo:** Usa solo los datos del área específica.
- Ejemplo: Estimador HT de $\bar{Y}_d$,

$$\hat{\bar{Y}}_d^{DIR} = \frac{1}{N_d} \sum_{j \in s_d} d_j y_j.$$

- Estimador de la varianza: Usando $\pi_{j,k} \cong \pi_j \pi_k$, $j \neq k$,

$$\hat{V}_\pi(\hat{\bar{Y}}_d^{DIR}) = \frac{1}{N_d} \sum_{j \in s_d} d_j (d_j - 1) y_j^2.$$

ESTIMADORES DIRECTOS

INDICADORES OBJETIVO:

- Aditivos en las observaciones individuales.

REQUERIMIENTOS de DATOS:

- Pesos muestrales d_j, $j \in s_d$ para las unidades muestreadas en el área.
- Para el estimador de HT de la media y para el estimador de Hájek del total, el tamaño poblacional del dominio N_d.

ESTIMADORES DIRECTOS

VENTAJAS:

- **Sin supuestos de modelo** (no paramétrico).
- Se pueden usar pesos muestrales $\Rightarrow$ Aproximadamente **insesgados y consistentes** bajo el diseño cuando $n_d \uparrow$.
- Aditividad (propiedad **"benchmarking"**):

$$\sum_{d=1}^{D} \hat{Y}_d^{DIR} = \hat{Y}^{DIR}.$$

DESVENTAJAS:

- $V_\pi(\hat{\bar{Y}}_d^{DIR}) \uparrow$ cuando $n_d \downarrow$. Muy **ineficiente** para dominios pequeños.
- No se pueden calcular para áreas no muestreadas ($n_d = 0$).

LÍMITES DE DESAGREGACIÓN

RECOMENDACIONES:

(i) Usar estimadores directos a nivel nacional y para desagregaciones con CV estimado por debajo de un límite especificado para todas las áreas.

(ii) Para mayores desagregaciones, usar estimadores indirectos en las áreas con sesgo absoluto relativo por debajo de un límite dado.

(iii) Para áreas donde los estimadores indirectos exceden el límite de sesgo, no obtener estimaciones. Siempre es posible modificar el reparto del tamaño muestral total entre las áreas para tener un número mínimo de observaciones en cada área.

ESTIMADORES INDIRECTOS

- **Estimador indirecto:** Estimador que comparte información con otras áreas (**"borrows strength"**) estableciendo relaciones de homogeneidad entre ellas (modelo con parámetros **comunes**).

PRIMERA APLICACIÓN DE REGRESIÓN SINTÉTICA

Encuesta de Radio 1945:

- Objetivo: estimar la mediana del número de emisoras de radio que son escuchadas durante el día en 500 condados de EE.UU.
- Encuesta por correo: En cada uno de los 500 condados, se muestrearon 1000 familias y se les envió un cuestionario por correo. Tasa de respuesta solo 20 % y cobertura incompleta.
- x_d num. mediano de emisoras escuchadas durante el día (encuesta por correo) en el condado d-ésimo, para $d = 1, \ldots, 500$. Sesgado debido a la falta de respuesta y cobertura incompleta.
- Encuesta con entrevistas personales en 85 condados: muestra probabilística de 85 condados, éstos son submuestreados de forma intensiva, realizando entrevistas personales.

✓ *Hansen, Hurwitz & Madow, 1953, p. 483;* ✓ *Rao & Molina, 2015*

PRIMERA APLICACIÓN DE REGRESIÓN SINTÉTICA

Encuesta de Radio 1945:

- y_d num. mediano de emisoras que escuchadas durante el día (entrevista personal) en el condado d, $d = 1, \ldots, 85$. Se consideran como las medianas verdaderas.
- $\text{corr}(y, x) = 0.70$
- Regresión Lineal:

$$y_d = \beta_0 + \beta_1 x_d + e_d, \quad d = 1, \ldots, 85.$$

- Estimadores indirectos para los 500-85 condados restantes:

$$\hat{y}_d^{SYN} = 0.52 + 0.74 x_d \quad \text{(Estimador sintético de regresión)}$$

- No tiene en cuenta la posible heterogeneidad entre condados.

ESTIMADORES SINTÉTICOS

Definición:

A partir de una encuesta, se obtiene un estimador insesgado para un área grande; cuando esta estimación se utiliza para calcular estimaciones para subáreas bajo el supuesto de que las áreas pequeñas tienen las mismas características que el área grande, identificamos estas estimaciones como estimaciones sintéticas.

✓ *González (1973)*

Ejemplo SIMPLE:

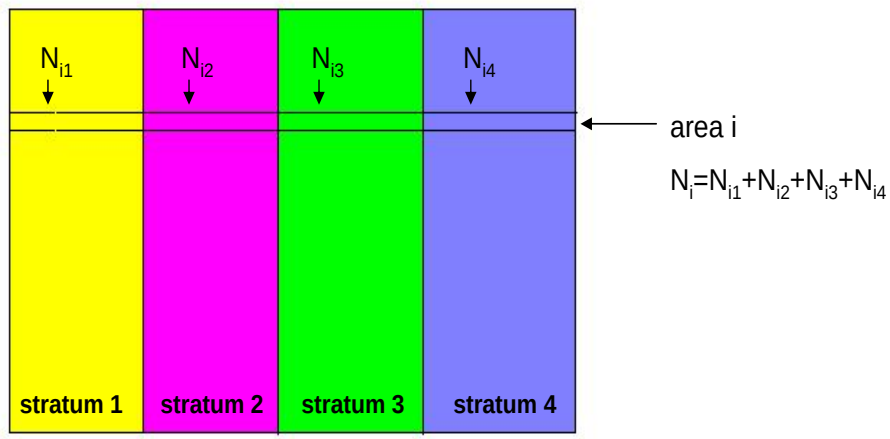
- Objetivo: $\bar{Y}_d$ media del dominio d.
- Se asume: $\bar{Y}_d = \bar{Y}$.
- Estimador sintético de $\bar{Y}_d$:

$$\hat{\bar{Y}}_d^{SYNT} = \hat{\bar{Y}}.$$

ESTIM. POST-ESTRATIFICADO SINTÉTICO

- J post-estratos ($j = 1, \ldots, J$) que se cruzan con los dominios.
- N_{dj} tamaño poblacional del cruce entre el dominio d y el post-estrato j.
- Total del dominio d:

$$Y_d = \sum_{j=1}^{J} N_{dj} \bar{Y}_{dj}$$



- Suposición (modelo implícito):

$$\bar{Y}_{dj} = \bar{Y}_{+j} = Y_{+j}/N_{+j}, \ \forall d, j$$

ESTIMADOR POST-ESTRATIFICADO SINTÉTICO

- Estimador post-estratificado sintético:

$$\hat{Y}_d^{SYN} = \sum_{j=1}^{J} N_{dj} \hat{\bar{Y}}_{+j}^R, \quad \hat{\bar{Y}}_{+j}^R = \hat{Y}_{+j} / \hat{N}_{+j}.$$

- $\hat{Y}_{+j}, \hat{N}_{+j}$ estimadores directos fiables de Y_{+j}, N_{+j}.
- Se necesita homogeneidad dentro de cada post-estrato.
- Caso especial: Cuando $y \in \{0, 1\}$, la proporción del dominio P_d es Y_d/N_d, donde $N_d = \sum_{j=1}^{J} N_{dj}$.
- Estimador post-estratificado sintético de P_d:

$$\hat{P}_d^{SYN} = \frac{1}{N_d} \sum_{j=1}^{J} N_{dj} \hat{P}_{+j}^R$$

ECM DEL ESTIMADOR SINTÉTICO

- El estimador post-estratificado sintético $\hat{Y}_d^{SYN}$ depende de los estimadores directos $\hat{Y}_{+j}/\hat{N}_{+j}$ para el post-estrato j. Por tanto, la varianza bajo el diseño de los estimadores sintéticos es pequeña en relación a la de los estimadores directos para un dominio pequeño.

- Pero los estimadores sintéticos dependen en gran medida de las hipótesis de homogenidad y pueden tener un sesgo grande cuando no sean ciertas.

- Por tanto, como medida de error, conviene dar el error cuadrático medio (ECM), que incluye sesgo y varianza.

ESTIMADOR del ECM

- ECM aproximado:

$$\text{ECM}_d(\hat{Y}_d^{SYN}) \approx E_d(\hat{Y}_d^{SYN} - \hat{Y}_d^{DIR})^2 - \hat{V}_d(\hat{Y}_d^{DIR})$$

- ECM estimado:

$$\hat{\text{ECM}}_d(\hat{Y}_d^{SYN}) = (\hat{Y}_d^{SYN} - \hat{Y}_d^{DIR})^2 - \hat{V}_d(\hat{Y}_d^{DIR}).$$

- $\hat{\text{ECM}}_d(\hat{Y}_d^{SYN})$ es aproximadamente insesgado pero inestable.
- Promedio sobre dominios: *(✓ González & Wakesberg, 1973)*

$$\hat{\text{ECM}}_a(\hat{\bar{Y}}_d^{SYN}) = \frac{1}{D}\sum_{\ell=1}^{D}\frac{1}{N_\ell^2}(\hat{Y}_\ell^{SYN} - \hat{Y}_\ell^{DIR})^2 - \frac{1}{D}\sum_{\ell=1}^{D}\frac{1}{N_\ell^2}\hat{V}_d(\hat{Y}_\ell^{DIR})$$

- Limitación: $\hat{\text{ECM}}_a(\hat{\bar{Y}}_d^{SYN})$ es estable pero es igual para todas las áreas.

ESTIMADOR SINTÉTICO

INDICADORES OBJETIVO:

- Para estimador sintético de regresión, indicadores generales. Para post-estratificados sintéticos, parámetros aditivos.

REQUERIMIENTOS DE DATOS:

- Para el estimador sintético de regresión, valores agregados de p variables auxiliares a nivel de dominio.
- Para estimadores sintéticos post-estratificados, indicador de post-estrato en la encuesta y tamaños poblacionales de cruces entre post-estratos y dominios.

ESTIMADOR SINTÉTICO

VENTAJAS:

- Pueden tener una varianza muy pequeña.
- Permiten estimar en áreas no muestreadas.

ESTIMADORES SINTÉTICOS

DESVENTAJAS:

- No tienen en cuenta la posible heterogeneidad entre áreas; por tanto, pueden estar seriamente sesgados bajo el diseño.
- El modelo debe verificarse cuidadosamente (por ejemplo, mediante gráficos de residuos y contrastes de significatividad de la varianza de los efectos aleatorios).
- Si se conoce el modelo, ¡no se usan los datos de la variable de interés obtenidos de la encuesta!
- No tienden al estimador directo al aumentar el tamaño muestral del dominio.
- No existen estimadores del ECM estables y distintos para cada área.
- Es necesario realizar ajustes para que cumplan la propiedad "benchmarking".

ESTIMADORES COMPUESTOS

Para equilibrar el sesgo de un estimador sintético y la inestabilidad de un estimador directo para un dominio, tomar:

$$\hat{Y}_d^C = \phi_d \hat{Y}_d + (1 - \phi_d)\hat{Y}_d^{SYN}, \quad 0 \leq \phi_d \leq 1.$$

- **Estimador dependiente de tamaño muestral (SSD):** Para un $\delta > 0$ dado,

$$\phi_d = \begin{cases} 1, & \text{si} \quad \hat{N}_d \geq \delta N_d; \\ \hat{N}_d/(\delta N_d), & \text{si} \quad \hat{N}_d < \delta N_d. \end{cases}$$

✓ *Drew, Singh & Choudhry (1982), SM*

ESTIMADOR DEPENDIENTE DEL TAMAÑO MUESTRAL (SSD)

- Bajo muestreo aleatorio simple (MAS) en la población:

$$\hat{N}_d = \sum_{j \in s_d} d_j = N n_d / n$$

- $\hat{N}_d$ insesgado: $N_d = E_\pi(\hat{N}_d) = N E_\pi(n_d)/n$. Entonces,

$$\hat{N}_d \geq \delta N_d \Leftrightarrow N n_d / n \geq \delta N E_\pi(n_d)/n \Leftrightarrow n_d \geq \delta E_\pi(n_d).$$

- Peso del estimador SSD bajo MAS:

$$\phi_d = \begin{cases} 1 & \text{si} \quad n_d \geq \delta E_d(n_d); \\ n_d / \{\delta E_d(n_d)\} & \text{si} \quad n_d < \delta E_d(n_d) \end{cases}$$

ESTIMADOR DEPENDIENTE DEL TAMAÑO MUESTRAL (SSD)

- Encuesta de Población Activa canadiense: Se producen estimaciones por divisiones censales con $\delta = 2/3$. Para la mayoría de las áreas, $1 - \phi_d = 0$; para otras áreas el peso asignado a $\hat{Y}_d^{SYN}$ estaba en torno a 0.1 pero nunca fue mayor que 0.2.

- Se usa el mismo peso ϕ_d para todas las variables y sin importar las diferencias con respecto a la homogeneidad entre áreas.

ESTIMADOR COMPUESTO ÓPTIMO

- Encontrar ϕ_d que minimice $\text{ECM}_d(\hat{Y}_d^C) \Rightarrow \phi_d^*$
- Peso óptimo depende de los verdaderos ECMs de $\hat{Y}_d^{SYN}$ y $\hat{Y}_d$.
- Peso óptimo estimado:

$$\hat{\phi}_d^* = \text{ECM}_d(\hat{Y}_d^{SYN})/(\hat{Y}_d^{SYN} - \hat{Y}_d)^2$$

- Limitación: $\hat{\phi}_d^*$ es inestable.
- Peso óptimo estimado *común* (promedio sobre áreas):

$$\begin{aligned}
\hat{\phi}^* &= \sum_{\ell=1}^{D} \text{ECM}_d(\hat{\bar{Y}}_\ell^{SYN}) / \sum_{\ell=1}^{D} (\hat{\bar{Y}}_\ell^{SYN} - \hat{\bar{Y}}_\ell)^2 \\
&= 1 - \left\{ \sum_{\ell=1}^{D} \hat{V}_d(\hat{\bar{Y}}_\ell) / \sum_{\ell=1}^{D} (\hat{\bar{Y}}_\ell^{SYN} - \hat{\bar{Y}}_\ell)^2 \right\}
\end{aligned}$$

- $\hat{\phi}^*$ es estable pero es igual para todas las áreas.

BENCHMARKING

- Normalmente se dispone de un estimador directo fiable para una región A que contiene varias áreas, $\hat{Y}_A^{DIR}$.
- Los estimadores indirectos de los totales de las áreas Y_d contenidas en dicha región no tienen por qué sumar $\hat{Y}_A^{DIR}$.
- Ajuste de razón: $\tilde{Y}_d$ estimador indirecto de Y_d con $\sum_{d\in A} \tilde{Y}_d \neq \hat{Y}_A^{DIR}$. Entonces, se toma el estimador

$$\tilde{Y}_d^* = \tilde{Y}_d \frac{\hat{Y}_A^{DIR}}{\sum_{d\in A} \tilde{Y}_d} \Rightarrow \sum_{d\in A} \tilde{Y}_d^* = \hat{Y}_A^{DIR}$$

ESTIMADORES SSD

INDICADORES OBJETIVO:

- Parámetros aditivos.

REQUERIMIENTOS DE DATOS:

- Los mismos que los estimadores directo y sintético que se utilicen.

ESTIMADORES SSD

VENTAJAS:

- Tenderán a tener menor varianza bajo el diseño que el estimador directo y menor sesgo que el sintético.

DESVENTAJAS:

- Si el tamaño muestral del dominio (incluso siendo pequeño) no es inferior al tamaño esperado, no se comparte información.
- El peso del estimador sintético no depende de lo bien explicada que esté la variable de interés por las variables auxiliares.
- No se pueden calcular para dominios no muestreados.
- No se dispone de estimadores del ECM bajo el diseño estables y distintos para cada área.
- Necesitan reajuste para satisfacer la propiedad "benchmarking".

MODELO FAY-HERRIOT

(i) Modelo que **enlaza** las áreas:

$$\delta_d = \mathbf{x}_d'\boldsymbol{\beta} + u_d, \quad d = 1, \ldots, D$$

$$u_d \overset{iid}{\sim} (0, \sigma_u^2), \quad \sigma_u^2 \text{ desconocido}$$

(ii) Modelo del **muestreo**:

$$\hat{\delta}_d^{DIR} = \delta_d + e_d, \quad d = 1, \ldots, D$$

$$e_d \overset{ind}{\sim} (0, \psi_d), \quad \psi_d = V_\pi(\hat{\delta}_d^{DIR}|\delta_d) \text{ conocido } \forall d$$

u_d y e_d independientes

(iii) Modelo **combinado**: Modelo lineal mixto

$$\hat{\delta}_d^{DIR} = \mathbf{x}_d'\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \ldots, D$$

✓ *Fay & Herriot (1979), JASA*

BLUP BAJO EL MODELO FAY-HERRIOT

Best linear unbiased predictor (BLUP)

Bajo el modelo combinado (iii) con $\delta_d = \mathbf{x}_d'\boldsymbol{\beta} + u_d$, el estimador lineal $\tilde{\delta}_d = b + \alpha_1\hat{\delta}_1^{DIR} + \cdots + \alpha_D\hat{\delta}_D^{DIR}$ que es solución al problema:

$$\begin{array}{ll} \min_{(\alpha_1,\ldots,\alpha_D)} & \mathrm{ECM}(\tilde{\delta}_d) = E(\tilde{\delta}_d - \delta_d)^2 \\ \text{s.t.} & E(\tilde{\delta}_d - \delta_d) = 0 \end{array}$$

viene dado por

$$\tilde{\delta}_d^{BLUP} = \mathbf{x}_d'\tilde{\boldsymbol{\beta}} + \tilde{u}_d,$$

donde

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\sigma_u^2) = \left(\sum_{d=1}^{D}\gamma_d\mathbf{x}_d\mathbf{x}_d'\right)^{-1}\sum_{d=1}^{D}\gamma_d\mathbf{x}_d\hat{\delta}_d^{DIR},$$

$$\tilde{u}_d = \tilde{u}_d(\sigma_u^2) = \gamma_d(\hat{\delta}_d^{DIR} - \mathbf{x}_d'\tilde{\boldsymbol{\beta}}), \quad \gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \psi_d}$$

BLUP BAJO EL MODELO FAY-HERRIOT

Demostración: Se demuestra un resultado más general. Se expresa el modelo (iii) en notación matricial

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \mathbf{e},$$

donde

$$\mathbf{y} = \begin{pmatrix} \hat{\delta}_1^{DIR} \\ \vdots \\ \hat{\delta}_D^{DIR} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_D' \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_D \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_D \end{pmatrix}.$$

Matrices de covarianzas: $V(\mathbf{u}) = \sigma_u^2 \mathbf{I}_D$, $V(\mathbf{e}) = \text{diag}(\psi_d)$.
Demostramos que el BLUP de un efectos mixto

$$\mu = \boldsymbol{\ell}'\boldsymbol{\beta} + \mathbf{m}'\mathbf{u},$$

para vectores $\boldsymbol{\ell}$ and $\mathbf{m}$ dados de dimensiones $p \times 1$ y $D \times 1$, es

$$\tilde{\mu} = \boldsymbol{\ell}'\tilde{\boldsymbol{\beta}} + \mathbf{m}'\tilde{\mathbf{u}}, \quad \tilde{\mathbf{u}} = (\tilde{u}_1, \ldots, \tilde{u}_D)'.$$

BLUP BAJO EL MODELO FAY-HERRIOT

- Predictor **Lineal** de $\mu = \boldsymbol{\ell}'\boldsymbol{\beta} + \mathbf{m}'\mathbf{u}$:

$$\tilde{\mu} = \boldsymbol{\alpha}'\mathbf{y} + b,$$

para un vector dado $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)'$ y escalar b.

- Error de predicción:

$$\tilde{\mu} - \mu = \boldsymbol{\alpha}'\mathbf{y} + b - \boldsymbol{\ell}'\boldsymbol{\beta} - \mathbf{m}'\mathbf{u} = \boldsymbol{\alpha}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\alpha}'\mathbf{u} + \boldsymbol{\alpha}'\mathbf{e} + b - \boldsymbol{\ell}'\boldsymbol{\beta} - \mathbf{m}'\mathbf{u}.$$

- $\tilde{\mu}$ **insesgado bajo el modelo** para μ si y solo si $E(\tilde{\mu} - \mu) = 0$.
- Tomando esperanza del error de predicción,

$$E(\tilde{\mu} - \mu) = (\boldsymbol{\alpha}'\mathbf{X} - \boldsymbol{\ell}')\boldsymbol{\beta} + b = 0 \ \forall \boldsymbol{\beta} \Leftrightarrow \boldsymbol{\alpha}'\mathbf{X} = \boldsymbol{\ell}', \ b = 0.$$

- Si $\tilde{\mu}$ es insesgado para μ, entonces

$$\text{ECM}(\tilde{\mu}) = V(\tilde{\mu} - \mu) = V(\boldsymbol{\alpha}'\mathbf{y} - \mathbf{m}'\mathbf{u}) = \boldsymbol{\alpha}'\mathbf{V}\boldsymbol{\alpha} + \sigma_u^2\mathbf{m}'\mathbf{m} - 2\sigma_u^2\boldsymbol{\alpha}'\mathbf{m},$$

donde $\mathbf{V} = V(\mathbf{y}) = \sigma_u^2\mathbf{I}_D + \text{diag}(\psi_d)$.

BLUP BAJO EL MODELO FAY-HERRIOT

- Problema de minimización:

$$\begin{array}{ll} \min_{\boldsymbol{\alpha}} & \text{ECM}(\tilde{\mu}) = \boldsymbol{\alpha}'\mathbf{V}\boldsymbol{\alpha} + \sigma_u^2\mathbf{m}'\mathbf{m} - 2\sigma_u^2\boldsymbol{\alpha}'\mathbf{m} \\ \text{s.t.} & \boldsymbol{\alpha}'\mathbf{X} = \boldsymbol{\ell}' \end{array}$$

- Mediante el método de **multiplicadores de Lagrange**, se obtiene:

$$\boldsymbol{\alpha}' = \boldsymbol{\ell}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} + \sigma_u^2\mathbf{m}'\mathbf{V}^{-1}\left[\mathbf{I}_D - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\right].$$

- Entonces, el BLUP de μ es

$$\tilde{\mu}^{BLUP} = \boldsymbol{\alpha}'\mathbf{y} = \boldsymbol{\ell}'\tilde{\boldsymbol{\beta}} + \mathbf{m}'\underbrace{\sigma_u^2\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})}_{\tilde{\mathbf{u}}} = \boldsymbol{\ell}'\tilde{\boldsymbol{\beta}} + \mathbf{m}'\tilde{\mathbf{u}}.$$

- Para $\boldsymbol{\ell} = \mathbf{x}_d$ and $\mathbf{m} = (\mathbf{0}'_{d-1}, 1, \mathbf{0}'_{D-d})'$, obtenemos

$$\tilde{\delta}_d^{BLUP} = \mathbf{x}'_d\tilde{\boldsymbol{\beta}} + \tilde{u}_d.$$

BUENA PROPIEDAD DEL BLUP

- El BLUP se puede expresar como

$$\tilde{\delta}_d^{BLUP} = \gamma_d \, \hat{\delta}_d^{DIR} + (1 - \gamma_d)\mathbf{x}_d'\tilde{\boldsymbol{\beta}}, \quad \gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \psi_d}.$$

- Composición del estimador directo $\hat{\delta}_d^{DIR}$ y el estimador "sintético de regresión" $\mathbf{x}_d'\tilde{\boldsymbol{\beta}}$.
- Da **mayor peso** a $\hat{\delta}_d^{DIR}$ cuando la varianza muestral ψ_d es pequeña small ($\hat{\delta}_d^{DIR}$ **fiable**).
- Da **más peso** al estimador **sintético** $\mathbf{x}_d'\tilde{\boldsymbol{\beta}}$ cuando ψ_d es grande ($\hat{\delta}_d^{DIR}$ no fiable) o σ_u^2 pequeño ($\mathbf{x}_d'\tilde{\boldsymbol{\beta}}$ **fiable**).

BLUP EMPÍRICO (EBLUP)

- $\tilde{\delta}_d^{BLUP}$ depende de σ_u^2 a través de $\tilde{\boldsymbol{\beta}}$ y γ_d:

$$\tilde{\delta}_d^{BLUP} = \tilde{\delta}_d^{BLUP}(\sigma_u^2)$$

- BLUP **empírico** (EBLUP) de δ_d: $\hat{\sigma}_u^2$ estimador de σ_u^2,

$$\hat{\delta}_d^{EBLUP} = \tilde{\delta}_d^{BLUP}(\hat{\sigma}_{\mathbf{u}}^{\mathbf{2}}), \quad d = 1, \ldots, D$$

- El EBLUP se mantiene **insesgado bajo el modelo**, si:
  - ✓ La distribución de u_d es simétrica.
  - ✓ $\hat{\sigma}_u^2$ par: $\hat{\sigma}_u^2(\mathbf{y}) = \hat{\sigma}_u^2(-\mathbf{y})$.
  - ✓ $\hat{\sigma}_u^2$ invariante por traslaciones: $\hat{\sigma}_u^2(\mathbf{y} + \mathbf{X}\boldsymbol{\gamma}) = \hat{\sigma}_u^2(\mathbf{y})$ para todo $\mathbf{y}$ y $\boldsymbol{\gamma}$.

MÉTODOS DE AJUSTE

- ✓ Método de ajuste FH;
- ✓ Máxima Verosimilitud (ML);
- ✓ Máxima Verosimilitud Restringida/Residual (REML);
- ✓ Método de momentos de Prasad-Rao.

MÉTODO DE AJUSTE FH

- Se verifica

$$\hat{\delta}_d^{DIR} \overset{ind}{\sim} N(\mathbf{x}_d'\boldsymbol{\beta}, \sigma_u^2 + \psi_d) \Rightarrow \sum_{d=1}^{D} \frac{\left\{\hat{\delta}_d^{DIR} - \mathbf{x}_d'\tilde{\boldsymbol{\beta}}(\sigma_u^2)\right\}^2}{\sigma_u^2 + \psi_d} \sim \mathcal{X}_{D-p}^2$$

- **Método Fay-Herriot:** Resolver iterativamente para σ_u^2 la ecuación

$$h(\sigma_u^2) = \sum_{d=1}^{D} \frac{\left(\hat{\delta}_d^{DIR} - \mathbf{x}_d'\tilde{\boldsymbol{\beta}}(\sigma_u^2)\right)^2}{\sigma_u^2 + \psi_d} = D - p.$$

  Parar cuando las iteraciones convergen a una solución $\tilde{\sigma}_u^2$

  Tomar $\hat{\sigma}_u^2 = \max(\tilde{\sigma}_u^2, 0)$ y $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_u^2)$.

  No se requiere normalidad.

OTROS MÉTODOS DE AJUSTE

- **Máxima verosimilitud:** Habitualmente bajo normalidad

$$\hat{\delta}_d^{DIR} \sim N(\mathbf{x}_d'\boldsymbol{\beta}, \sigma_u^2 + \psi_d)$$

  Los estimadores ML son consistentes en ausencia de normalidad (bajo ciertas condiciones de regularidad).

- **Máxima verosimilitud restringida (REML):** Reduce el sesgo de los estimadores ML para tamaño muestral pequeño n en comparación con p.

- **Método de Prasad-Rao:** Método de momentos. Proporciona buenos valores iniciales para algoritmos iterativos de ajuste.

(✓ Prasad & Rao, 1990)

ERROR CUADRÁTICO MEDIO

- Bajo normalidad de u_d y e_d, cuando $D \to \infty$,

$$\text{ECM}(\hat{\delta}_d^{EBLUP}) = g_{1d}(\sigma_u^2) + g_{2d}(\sigma_u^2) + g_{3d}(\sigma_u^2) + o(D^{-1}),$$

donde

$$\begin{aligned} g_{1d}(\sigma_u^2) &= \gamma_d \psi_d = O(1), \\ g_{2d}(\sigma_u^2) &= (1-\gamma_d)^2 \mathbf{x}_d' \left(\sum_{d=1}^{D} \gamma_d \mathbf{x}_d \mathbf{x}_d' \right)^{-1} \mathbf{x}_d = O(D^{-1}), \\ g_{3d}(\sigma_u^2) &= (1-\gamma_d)^2 \gamma_d \sigma_u^{-2} \bar{V}(\hat{\sigma}_u^2) = O(D^{-1}), \end{aligned}$$

- $\bar{V}(\hat{\sigma}_u^2)$ varianza asintótica de $\hat{\sigma}_u^2$: Depende del método de estimación usado para σ_u^2.

✓ *Prasad & Rao (1990), JASA*

ERROR CUADRÁTICO MEDIO

Esquema de la demostración (estimación ML): Hemos obtenido $\mu = \alpha'\mathbf{u}$, donde

$$\alpha' = \ell'\mathbf{Q}\mathbf{X}'\mathbf{V}^{-1} + \sigma_u^2\mathbf{m}'\mathbf{P}, \quad \mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\mathbf{Q}\mathbf{X}'\mathbf{V}^{-1},$$

para $\mathbf{Q} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = (\sum_d \gamma_d \mathbf{x}_d\mathbf{x}_d')^{-1}$.

Reemplazando α' y $\mathbf{m}' = (\mathbf{0}_{d-1}', 1, \mathbf{0}_{D-d}')$ en ECM$(\tilde{\mu})$, y observando que

$$\mathbf{PVP} = \mathbf{P}, \ \mathbf{PX} = \mathbf{0}_D,$$

obtenemos que

$$\text{ECM}(\tilde{\delta}_d^{BLUP}) = g_{1d}(\sigma_u^2) + g_{2d}(\sigma_u^2),$$

donde

$$\begin{aligned} g_{1d}(\sigma_u^2) &= \gamma_d\psi_d, \\ g_{2d}(\sigma_u^2) &= (1-\gamma_d)^2\mathbf{x}_d'(\sum_d \gamma_d\mathbf{x}_d\mathbf{x}_d')^{-1}\mathbf{x}_d. \end{aligned}$$

ERROR CUADRÁTICO MEDIO

- Descomposición del ECM:

$$\begin{aligned} \text{ECM}(\hat{\delta}_d^{EBLUP}) &= E(\hat{\delta}_d^{EBLUP} - \delta_d)^2 \\ &= E(\hat{\delta}_d^{EBLUP} - \tilde{\delta}_d^{BLUP} + \tilde{\delta}_d^{BLUP} - \delta_d)^2 \\ &= \text{ECM}(\tilde{\delta}_d^{BLUP}) + E(\hat{\delta}_d^{EBLUP} - \tilde{\delta}_d^{BLUP})^2 \\ &+ 2E(\hat{\delta}_d^{EBLUP} - \tilde{\delta}_d^{BLUP})(\tilde{\delta}_d^{BLUP} - \delta_d). \end{aligned}$$

- Si $\hat{\sigma}_u^2$ es par e invariante por translaciones, entonces bajo normalidad

$$E(\hat{\delta}_d^{EBLUP} - \tilde{\delta}_d^{BLUP})(\tilde{\delta}_d^{BLUP} - \delta_d) = 0.$$

- Por tanto,

$$\text{ECM}(\hat{\delta}_d^{EBLUP}) = \text{ECM}(\tilde{\delta}_d^{BLUP}) + E(\hat{\delta}_d^{EBLUP} - \tilde{\delta}_d^{BLUP})^2.$$

✓ *Kackar & Harville (1984), JASA*

ERROR CUADRÁTICO MEDIO

- Expansión de Taylor de primer orden de $\tilde{\delta}_d^{BLUP}(\hat{\sigma}_u^2)$ en torno a σ_u^2:

$$\hat{\delta}_d^{EBLUP} \approx \tilde{\delta}_d^{BLUP} + \frac{\partial \tilde{\delta}_d^{BLUP}}{\partial \sigma_u^2}(\hat{\sigma}_u^2 - \sigma_u^2).$$

- Entonces,

$$E(\hat{\delta}_d^{EBLUP} - \tilde{\delta}_d^{BLUP})^2 \approx E\left[\left(\frac{\partial \tilde{\delta}_d^{BLUP}}{\partial \sigma_u^2}\right)^2 (\hat{\sigma}_u^2 - \sigma_u^2)^2\right].$$

- Reemplazamos $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v}$ en $\tilde{\delta}_d^{BLUP} = \boldsymbol{\alpha}'\mathbf{y}$:

$$\tilde{\delta}_d^{BLUP} = \boldsymbol{\ell}'\boldsymbol{\beta} + \mathbf{b}'\mathbf{v}, \quad \mathbf{v} = \mathbf{u} + \mathbf{e} \sim N(\mathbf{0}_D, \mathbf{V}).$$

Entonces,

$$\frac{\partial \tilde{\delta}_d^{BLUP}}{\partial \sigma_u^2} = \frac{\partial \mathbf{b}'}{\partial \sigma_u^2}\mathbf{v}.$$

ERROR CUADRÁTICO MEDIO

- Por tanto,

$$E(\hat{\delta}_d^{EBLUP} - \tilde{\delta}_d^{BLUP})^2 \approx E\left[\frac{\partial \mathbf{b}'}{\partial \sigma_u^2}\mathbf{v}\mathbf{v}'\left(\frac{\partial \mathbf{b}'}{\partial \sigma_u^2}\right)'(\hat{\sigma}_u^2 - \sigma_u^2)^2\right].$$

- $\hat{\sigma}_u^2$ estimador ML de σ_u^2.
- Por la expansión de primer orden de Taylor de $s(\hat{\sigma}_u^2) = \partial \log L(\hat{\sigma}_u^2)/\partial \hat{\sigma}_u^2$ en el valor σ_u^2, y sabiendo que $\partial s(\sigma_u^2)/\partial \sigma_u^2 \xrightarrow{P} -\mathcal{I}(\sigma_u^2)$, donde $\mathcal{I}(\sigma_u^2)$ es la información de Fisher,

$$\hat{\sigma}_u^2 \approx \sigma_u^2 + \mathcal{I}(\sigma_u^2)s(\sigma_u^2)$$

ERROR CUADRÁTICO MEDIO

- Función de log-verosimilitud:

$$\log L(\sigma_u^2) = -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).$$

- Score (gradiente de la log-verosimilitud):

$$s(\sigma_u^2) = -\frac{1}{2}\text{tr}(\mathbf{V}^{-1}) - \underbrace{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'}_{\mathbf{v}'}\mathbf{V}^{-3}\underbrace{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}_{\mathbf{v}}.$$

- Información de Fisher:

$$\mathcal{I}(\sigma_u^2) = -\frac{1}{2}\text{tr}(\mathbf{V}^{-2}).$$

- Finalmente, usando las expresiones del score y la inf. de Fisher, se calcula la siguiente esperanza teniendo en cuenta la normalidad de $\mathbf{v}$:

$$E(\hat{\delta}_d^{EBLUP} - \tilde{\delta}_d^{BLUP})^2 \approx E\left[\frac{\partial \mathbf{b}'}{\partial \sigma_u^2}\mathbf{v}\mathbf{v}'\left(\frac{\partial \mathbf{b}'}{\partial \sigma_u^2}\right)'\mathcal{I}^2(\sigma_u^2)s^2(\sigma_u^2)\right].$$

ERROR CUADRÁTICO MEDIO

- Se cumple que

$$E[g_{1d}(\hat{\sigma}_u^2)] \approx g_{1d}(\sigma_u^2) - g_{3d}(\sigma_u^2),$$
$$E[g_{2d}(\hat{\sigma}_u^2)] \approx g_{2d}(\sigma_u^2), \quad E[g_{3d}(\hat{\sigma}_u^2)] \approx g_{3d}(\sigma_u^2).$$

- El estimador del ECM cuando $\hat{\sigma}_u^2$ se obtiene mediante REML:

$$\text{mse}(\hat{\delta}_d^{EBLUP}) = g_{1d}(\hat{\sigma}_u^2) + g_{2d}(\hat{\sigma}_u^2) + 2g_{3d}(\hat{\sigma}_u^2)$$

- Es casi insesgado:

$$E\left[\text{mse}(\hat{\delta}_d^{EBLUP})\right] = \text{ECM}(\hat{\delta}_d^{EBLUP}) + o(D^{-1})$$

- Cuando $\hat{\sigma}_u^2$ se obtiene por FH o ML, se debe añadir un término debido al sesgo en $\hat{\sigma}_u^2$.

✓ *Prasad & Rao (1990), JASA*

EBLUP BAJO EL MODELO FH

INDICADORES OBJETIVO:

- Indicadores generales.

REQUERIMIENTOS DE DATOS:

- Valores agregados de p variables auxiliares A nivel de dominio.
- Tamaños poblacionales de los dominios.

EBLUP BAJO EL MODELO FH

VENTAJAS:

- Requiere solo información auxiliar a **nivel de área**, que está disponible **fácilmente** y evita problemas de confidencialidad.
- Hace uso de los **pesos muestrales** mientras $\gamma_d \neq 0$. Es consistente bajo el diseño cuando $n_d \to \infty$. Por tanto, se verá menos afectado por muestreo informativo.
- Asigna automáticamente **mayor peso** al estimador sintético de regresión cuando el tamaño muestral del área es **pequeño**.
- A menudo es más eficiente que el estimador directo.
- Tiene en cuenta la heterogeneidad no explicada entre áreas si $\gamma_d \neq 0$.

EBLUP BAJO EL MODELO FH

VENTAJAS:

- Tiende al estimador directo cuando aumenta el tamaño muestral del dominio (ψ_d decrece).
- Para estimadores directos **lineales**, el T. Central del Límite garantiza una **mínima bondad de ajuste** en las áreas de tamaños muestrales no demasiado pequeños. Atípicos aislados tienen efecto pequeño debido a la agregación.
- El estimador del ECM de Prasad-Rao es un estimador estable del ECM bajo el diseño y es **insesgado bajo el diseño** cuando se **promedia** a lo largo de un número grande de áreas.
- Para estimar en dominios no muestreados, se puede usar el componente sintético ($\gamma_d = 0$).

EBLUP BAJO EL MODELO FH

DESVENTAJAS:

- **Pérdida** de información en el proceso de agregación de las variables auxiliares.
- Solo D (típicamente $<< n$) observaciones para ajustar el modelo. En nuestros ejemplos, ganacias **pequeñas** respecto a los estimadores directos.
- Es necesario diagnosticar del modelo. Probleams potenciales de **linealidad** para parámetros no lineales.
- Se requiere estimación **preliminar** de las varianzas muestrales ψ_d. ¡El mismo problema de áreas pequeñas!

EBLUP BAJO EL MODELO FH

DESVENTAJAS:

- Si queremos estimar varios indicadores definidos en términos de la misma variable objetivo, se requiere encontrar un buen modelo para cada indicador.
- Los estimadores no se pueden desagregar para subdominios.
- La fórmula del estimador de ECM de Prasad-Rao es correcta **bajo el modelo** con normalidad, pero no es insesgado bajo el diseño para el ECM en un área concreta).
- Se requiere reajuste para satisfacer la propiedad "benchmarking".

MODELO CON ERRORES ANIDADOS

- y_{dj} valor de la variable objetivo para la unidad j dentro del área d
- u_d efecto aleatorio del área d
- Modelo de regresión lineal con **errores anidados**:

$$y_{dj} = \mathbf{x}'_{dj}\boldsymbol{\beta} + u_d + e_{dj}, \quad j = 1, \ldots, N_d, \; d = 1, \ldots, D$$
$$u_d \overset{iid}{\sim} N(0, \sigma_u^2), \quad e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2)$$

- Modelo en notación matricial:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

- Esperanza y varianza marginales:

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, \quad V(\mathbf{y}) = \sigma_u^2 \mathbf{Z}\mathbf{Z}' + \sigma_e^2 \mathbf{I}_N$$

✓ *Battese, Harter & Fuller (1988), JASA*

BLUP: MODELO LINEAL GENERAL

Modelo lineal más general:

- $\mathbf{y} = (y_1, \ldots, y_N)'$ vector poblacional (**aleatorio**)
- Modelo lineal:

$$E(\mathbf{y}) = \mathbf{X}\beta, \quad V(\mathbf{y}) = \mathbf{V}$$

- Descomposición en partes muestreada y no muestreada

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{pmatrix}$$

- Parámetro lineal objetivo:

$$\delta = \mathbf{a}'\mathbf{y} = \mathbf{a}_s'\mathbf{y}_s + \mathbf{a}_r'\mathbf{y}_r$$

BLUP: MODELO LINEAL GENERAL

Mejor predictor lineal insesgado (BLUP): V conocido

El predictor lineal $\tilde{\delta} = \boldsymbol{\alpha}' \mathbf{y}_s$ que es solución del problema:

$$\begin{array}{ll} \text{mín}_{\boldsymbol{\alpha} \in \boldsymbol{R}^n} & \text{ECM}(\tilde{\delta}) = E(\tilde{\delta} - \delta)^2 \\ \text{s.a.} & E(\tilde{\delta} - \delta) = 0 \end{array}$$

viene dado por

$$\tilde{\delta}^{BLUP} = \mathbf{a}_s' \mathbf{y}_s + \mathbf{a}_r' \tilde{\mathbf{y}}_r^{BLUP},$$

donde

$$\begin{aligned} \tilde{\mathbf{y}}_r^{BLUP} &= \mathbf{X}_r \tilde{\boldsymbol{\beta}} + \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} (\mathbf{y}_s - \mathbf{X}_s \tilde{\boldsymbol{\beta}}), \\ \tilde{\boldsymbol{\beta}} &= (\mathbf{X}_s' \mathbf{V}_{ss}^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s' \mathbf{V}_{ss}^{-1} \mathbf{y}_s \end{aligned}$$

✓ *Royall (1970), Biometrika*

BLUP BAJO MOD. ERRORES ANIDADOS

- BLUP de $\delta = \bar{Y}_d$ Bajo el **modelo con errores anidados**:

$$\tilde{\bar{Y}}_d^{BLUP} = \frac{1}{N_d}\left(\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \tilde{y}_{dj}^{BLUP}\right),$$

donde

$$\tilde{y}_{dj}^{BLUP} = \mathbf{x}'_{dj}\tilde{\boldsymbol{\beta}} + \tilde{u}_d,\ \tilde{\boldsymbol{\beta}} \text{ estimador WLS de } \boldsymbol{\beta},$$
$$\tilde{u}_d = \gamma_d(\bar{y}_d - \bar{\mathbf{x}}'_d\tilde{\boldsymbol{\beta}}),\ \gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/n_d).$$

- Cuando $n_d/N_d \approx 0$,

$$\tilde{\bar{Y}}_d^{BLUP} \approx \gamma_d\left\{\bar{y}_d + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_d)'\tilde{\boldsymbol{\beta}}\right\} + (1-\gamma_d)\bar{\mathbf{X}}'_d\tilde{\boldsymbol{\beta}}$$

- Composición entre estimadores **"survey regression"** $\bar{y}_d + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_d)'\tilde{\boldsymbol{\beta}}$ y **sintético de regresión** $\bar{\mathbf{X}}'_d\tilde{\boldsymbol{\beta}}$.

BLUP EMPÍRICO (EBLUP)

- BLUP depende de $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)'$ desconocido:

$$\tilde{\delta}^{BLUP} = \tilde{\delta}^{BLUP}(\boldsymbol{\theta}).$$

- EBLUP de δ: $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ estimador de $\boldsymbol{\theta}$

$$\hat{\delta}^{EBLUP} = \tilde{\delta}^{BLUP}(\hat{\boldsymbol{\theta}}),$$

- Estimadores de σ_u^2 y σ_e^2:
  - ✓ Método de Henderson III (método de momentos);
  - ✓ ML;
  - ✓ REML.

EBLUP BAJO MODELO A NIVEL UNIDAD

INDICADORES OBJETIVO:

- Medias de totales de la variable de interés.

REQUERIMIENTOS DE DATOS:

- Microdatos para las p variables auxiliares en la encuesta.
- Indicador del dominio en la encuesta.
- Medias poblacionales de las p variables auxiliares para los dominios.

EBLUP BAJO MODELO A NIVEL UNIDAD

VENTAJAS:

- Usa información auxiliar a nivel unidad, que es más **detallada** que la información a nivel de área.
- El tamaño total muestral es típicamente **grande** ($n >> D$), así que se comparte mucha información.
- Incorpora heterogeneidad no explicada entre áreas.
- Es necesario diagnosticar el modelo.
- No requiere disponer de las varianzas muestrales de los estimadores directos.
- Automáticamente comparte información entre áreas ("borrows strength") cuando el tamaño muestral del dominio es pequeño y tiende al estimador "survey-regression" cuando el tamaño muestral del dominio aumenta.

EBLUP BAJO MODELO A NIVEL UNIDAD

VENTAJAS:

- Los estimadores se pueden desagregar para subáreas (sin efecto de sub-área) o incluso para individuos.
- Estimadores **insesgados** bajo el modelo (no es necesaria normalidad pero sí simetría).
- Estimadores del ECM con sesgo **despreciable** bajo el modelo con **normalidad**.
- Estimador del ECM bajo el modelo estable para el ECM bajo el diseño e insesgado bajo el diseño cuando se promedia para muchos dominios.
- Para estimar en áreas no muestreadas, se puede usasr la parte sintética.

EBLUP BAJO MODELO A NIVEL UNIDAD

DESVENTAJAS:

- Información auxiliar a nivel unidad **de difícil acceso** por temas de confidencialidad.
- Solo se aplica a parámetros **lineales**.
- **No se usan los pesos muestrales**, de modo que puede ser sesgado bajo el diseño, especialmente bajo muestreo **informativo**.
- Se puede ver afectado por datos anómalos y/o falta de normalidad.

EBLUP BAJO MODELO A NIVEL UNIDAD

DESVENTAJAS:

- **Sensible** a desviaciones del modelo. **Diagnóstico del modelo** muy importante.
- Estimador del ECM por la fórmula de Prasad-Rao correcto **bajo el modelo** con normalidad. (no insesgado bajo el diseño para el MSE bajo el diseño en un área concreta).
- Es necesario un reajuste para satisfacer la propiedad "benchmarking".

MEJOR PREDICTOR

Mejor predictor (BP)

Consideramos $\delta = h(\mathbf{y})$, **no necesariamente lineal**. El predictor $\tilde{\delta} = g(\mathbf{y}_s)$ que minimiza $\text{ECM}(\tilde{\delta}) = E(\tilde{\delta} - \delta)^2$ es

$$\tilde{\delta}^{BP} = E_{\mathbf{y}_r}(\delta|\mathbf{y}_s).$$

Demostración: Definimos $\delta^0 = E_{\mathbf{y}_r}(\delta|\mathbf{y}_s)$. Observemos que

$$\text{ECM}(\tilde{\delta}) = E_{\mathbf{y}}\{(\tilde{\delta} - \delta^0)^2\} + 2\,E_{\mathbf{y}}\{(\tilde{\delta} - \delta^0)(\delta^0 - \delta)\} + E_{\mathbf{y}}\{(\delta^0 - \delta)^2\}.$$

El último término no depende de $\tilde{\delta}$. Para el segundo término,

$$\begin{aligned} E_{\mathbf{y}}\{(\tilde{\delta} - \delta^0)(\delta^0 - \delta)\} &= E_{\mathbf{y}_s}\left[E_{\mathbf{y}_r}\left\{(\tilde{\delta} - \delta^0)(\delta^0 - \delta)|\mathbf{y}_s\right\}\right] \\ &= E_{\mathbf{y}_s}\left[(\tilde{\delta} - \delta^0)\left\{\delta^0 - E_{\mathbf{y}_r}(\delta|\mathbf{y}_s)\right\}\right] = 0. \end{aligned}$$

El mínimo de $E_{\mathbf{y}}\{(\tilde{\delta} - \delta^0)^2\}$ se alcanza para $\tilde{\delta}^{BP} = \delta^0 = E_{\mathbf{y}_r}(\delta|\mathbf{y}_s)$.

MEJOR ESTIMADOR EMPÍRICO

- El mejor predictor es insesgado:

$$E_{\mathbf{y}_s}(\tilde{\delta}^{BP}) = E_{\mathbf{y}_s}\{E_{\mathbf{y}_r}(\delta|\mathbf{y}_s)\} = E_{\mathbf{y}}(\delta).$$

- Para un modelo lineal con $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ y $V(\mathbf{y}) = \mathbf{V}(\boldsymbol{\theta})$ con $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$ desconocidos, el BP depende de $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$:

$$\tilde{\delta}^{BP} = \tilde{\delta}^{BP}(\boldsymbol{\beta}, \boldsymbol{\theta}).$$

- Mejor predictor **empírico** (EBP): $\hat{\boldsymbol{\theta}}$ estimador de $\boldsymbol{\theta}$. Entonces

$$\hat{\delta}^{EBP} = \tilde{\delta}^{BP}(\tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}).$$

MEJOR PREDICTOR: PARÁMETRO LINEAL

- Caso particular: Consideramos un parámetro lineal

$$\delta = \mathbf{a}'\mathbf{y} = \mathbf{a}_s'\mathbf{y}_s + \mathbf{a}_r'\mathbf{y}_r$$

Si $\mathbf{y}$ se distribuye como una normal, entonces el BP es

$$\tilde{\delta}^{BP} = \mathbf{a}_s'\mathbf{y}_s + \mathbf{a}_r'\tilde{\mathbf{y}}_r^{BP},$$

donde

$$\tilde{\mathbf{y}}_r^{BP} = \mathbf{X}_r\boldsymbol{\beta} + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}).$$

- Usando $\tilde{\boldsymbol{\beta}}$ para estimar $\boldsymbol{\beta}$, el EBP coincide con el EBLUP.

MÉTODO EB: INDICADORES DE POBREZA

- Indicador de pobreza para el dominio d:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left(\frac{z - E_{dj}}{z} \right)^{\alpha} I\left(E_{dj} < z\right), \quad d = 1, \ldots, D.$$

- La distribución de la renta E_{dj} es marcádamente asimétrica por la derecha.
- Seleccionamos una transformación $T()$ tal que la distribución de $y_{dj} = T(E_{dj})$ sea aprox. normal.
- **Hipótesis:** $y_{dj} = T(E_{dj})$ satisface el modelo con errores anidados

$$y_{dj} = \mathbf{x}'_{dj}\boldsymbol{\beta} + u_d + e_{dj}, \quad u_d \overset{iid}{\sim} N(0, \sigma_u^2), \quad e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2).$$

MÉTODO EB: INDICADORES DE POBREZA

- Vector para área d: $\mathbf{y}_d = (y_{d1}, \ldots, y_{dN_d})'$.
- Indicador de pobreza en términos de $\mathbf{y}_d$:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left\{ \frac{z - T^{-1}(y_{dj})}{z} \right\}^{\alpha} I\left\{ T^{-1}(y_{dj}) < z \right\} = h_\alpha(\mathbf{y}_d).$$

- Partición de $\mathbf{y}_d$ en muestra y no-muestra: $\mathbf{y}_d = (\mathbf{y}'_{ds}, \mathbf{y}'_{dr})'$
- **Mejor predictor:**

$$\tilde{F}^{BP}_{\alpha d} = E_{\mathbf{y}_{dr}}\left[F_{\alpha d} | \mathbf{y}_{ds} \right].$$

✓ *Molina and Rao (2010), CJS*

MÉTODO EB

- Distribución de $\mathbf{y}_{dr}$ dado $\mathbf{y}_{ds}$ bajo el modelo con errores anidados:

$$\mathbf{y}_{dr}|\mathbf{y}_{ds} \sim N(\boldsymbol{\mu}_{dr|s}, \mathbf{V}_{dr|s}),$$

donde

$$\boldsymbol{\mu}_{dr|s} = \mathbf{X}_{dr}\boldsymbol{\beta} + \gamma_d(\bar{y}_{ds} - \bar{\mathbf{x}}'_{ds}\boldsymbol{\beta})\mathbf{1}_{N_d-n_d},$$
$$\mathbf{V}_{dr|s} = \sigma_u^2(1-\gamma_d)\mathbf{1}_{N_d-n_d}\mathbf{1}'_{N_d-n_d} + \sigma_e^2\mathbf{I}_{N_d-n_d},$$

y

$$\gamma_d = \sigma_u^2(\sigma_u^2 + \sigma_e^2/n_d)^{-1}.$$

- La distribución condicionada depende de $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$.
- **Mejor predictor empírico (EB):** Reemplazamos un estimador consistente $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$

$$\hat{F}_{\alpha d}^{EBP} = \tilde{F}_{\alpha d}^{BP}(\hat{\boldsymbol{\theta}}).$$

✓ *Molina & Rao (2010), CJS*

APROXIMACIÓN MONTE CARLO

(a) Generar L vectores fuera de muestra $\mathbf{y}_{dr}^{(\ell)}$, $\ell = 1, \ldots, L$ de la distribución condicionada (estimada) de $\mathbf{y}_{dr}|\mathbf{y}_{ds}$.

(b) Unir los elementos de la muestra para formar un vector censal $\mathbf{y}_{d}^{(\ell)} = (\mathbf{y}_{ds}, \mathbf{y}_{dr}^{(\ell)})$, $\ell = 1, \ldots, L$.

(c) Calcular el indicador de interés con cada vector poblacional $F_{\alpha d}^{(\ell)} = h_\alpha(\mathbf{y}_{d}^{(\ell)})$, $\ell = 1, \ldots, L$. Después tomar el promedio para las L simulaciones Monte Carlo:

$$\hat{F}_{\alpha d}^{EBP} = \frac{1}{L}\sum_{\ell=1}^{L} F_{\alpha d}^{(\ell)}.$$

(d) El ECM se puede estimar mediante bootstrap paramétrico.

✓ *Molina & Rao (2010), CJS*

ECM POR BOOTSTRAP PARAMÉTRICO

(i) Generar B vectores poblacionales (censos) bootstrap a partir del modelo ajustado

$$\mathbf{y}^{*(b)} = (\mathbf{y}_1^{*(b)}, \ldots, \mathbf{y}_D^{*(b)}), \quad b = 1, \ldots, B.$$

(ii) Calcular los verdaderos parámetros bootstrap

$$\delta_d^{*(b)} = h(\mathbf{y}_d^{*(b)}), \quad b = 1, \ldots, B.$$

(iii) Con la parte de la muestra $\mathbf{y}_s^{*(b)} = (\mathbf{y}_{1s}^{*(b)}, \ldots, \mathbf{y}_{Ds}^{*(b)})'$ del vector poblacional $\mathbf{y}^{*(b)}$, calcular los estimadores EB:

$$\hat{\delta}_d^{EBP*(b)}, \quad b = 1, \ldots, B.$$

(iv) **Estimador naïve del ECM por bootstrap paramétrico:**

$$mse_*(\hat{\delta}_d^{EBP}) = \frac{1}{B} \sum_{b=1}^{BP} \left(\hat{\delta}_d^{EBP*(b)} - \delta_d^{*(b)} \right)^2$$

✓ *González-Manteiga et al. (2008), JSCS*

EB BAJO EL MODELO A NIVEL UNIDAD

INDICADORES OBJETIVO:

- Indicadores generales definidos en términos de una variable continua (ej. renta) que será modelizada.

REQUERIMIENTOS DE DATOS:

- Microdatos de las p variables auxiliares en la encuesta.
- Indicador de dominio en la encuesta.
- Microdatos de las p variables auxiliares para todas las unidades de la población (censo o registro administrativo).

EB BAJO EL MODELO A NIVEL UNIDAD

VENTAJAS:

- Se usa información auxiliar a nivel de unidad, que es más **detallada** que la información a nivel de área.
- El tamaño muestral total es habitualmente muy **grande** ($n >> D$), por lo que se comparte mucha información.
- Incorpora heterogeneidad no explicada entre áreas.
- Permite estimar parámetros no lineales **generales** $h(\mathbf{y})$, donde $\mathbf{y}$ se distribuye según una normal.

EB BAJO EL MODELO A NIVEL UNIDAD

DESVENTAJAS:

- Se generan censos completos. Por tanto, se pueden estimar **varios indicadores** a partir del mismo modelo.
- Estimadores aprox. **insesgados y óptimos** bajo el modelo con normalidad.
- Las estimaciones se pueden desagregar en cualquier subdominio (sin efecto de subdominio), incluso a nivel de unidad.
- Estimadores del ECM bajo el modelo **con sesgo despreciable** bajo **normalidad**, para parámetros lineales.
- Estimador del ECM bajo el modelo es estable para el ECM bajo el diseño cuando se promedia para muchos dominios, para parámetros lineales.

EB BAJO EL MODELO A NIVEL UNIDAD

DESVENTAJAS:

- Información auxiliar para cada unidad de la población (censo/registro) **no es fácilmente accesible**.
- Computacionalmente **intensivo**.
- **No utiliza los pesos de muestreo**, por lo que puede ser sesgado bajo el diseño, especialmente bajo muestreo **informativo**.
- **Sensible** a desviaciones del modelo. Es muy importante encontrar la transformación correcta de la variable y **la diagnosis del modelo**.
- Los estimadores del ECM por bootstrap son computacionalmente intensivos.

MODELOS LINEALES GENERALIZADOS

- $y_{dj} \in \{0, 1\}$, donde 1=presencia de la característica de interés, 0=ausencia.
- Parámetros objetivo: proporciones de individuos con dicha característica,

$$P_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}, \quad d = 1, \ldots, D.$$

- Modelo logístico mixto:

$$y_{dj}|u_d \overset{ind.}{\sim} \text{Bern}(p_{dj}), \quad j = 1, \ldots, N_d, \; d = 1, \ldots, D,$$
$$p_{dj} = \frac{\exp(\mathbf{x}'_{dj}\boldsymbol{\beta} + u_d)}{1 + \exp(\mathbf{x}'_{dj}\boldsymbol{\beta} + u_d)}, \quad u_d \overset{iid}{\sim} N(0, \sigma_u^2).$$

✓ *González-Manteiga et al. (2007), CSDA*

ESTIMADORES DE ÁREAS PEQUEÑAS

- Mejor predictor:

$$\hat{P}_d^{BP} = \frac{1}{N_d}\left\{\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} E(y_{dj}|\mathbf{y}_{ds})\right\}, \quad d = 1,\ldots,D.$$

- La esperanza de $E(y_{dj}|\mathbf{y}_{ds})$ no se puede calcular analíticamente: es necesario utilizar métodos de simulación para las aproximaciones (ej. Laplace o Monte Carlo).
- Estimadores simples tipo *plug-in*:

$$\hat{P}_d^{Plug} = \frac{1}{N_d}\left(\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \hat{p}_{dj}^{Plug}\right), \quad d = 1,\ldots,D.$$

- $\hat{p}_{dj}^{Plug} = \exp(\mathbf{x}'_{dj}\hat{\boldsymbol{\beta}} + \hat{u}_d)/\{1+\exp(\mathbf{x}'_{dj}\hat{\boldsymbol{\beta}} + \hat{u}_d)\}$ predicción de las probabilidades mediante el ajuste del GLMM.

MÉTODOS DE AJUSTE

- Verosimilitud muestral:

$$f(\mathbf{y}_s) = \int_{R^D} f(\mathbf{y}_s, \mathbf{u}) d\mathbf{u} = \int_{R^D} f_1(\mathbf{y}_s|\mathbf{u}) f_2(\mathbf{u}) d\mathbf{u}$$

- No se puede obtener una expresión analítica para la verosimilitud.

- ML: Se requieren aproximaciones (ej. Laplace) o métodos numéricos para maximizar la verosimilitud.

CUASI-VEROSIMILITUD PENALIZADA (PQL) +ML APROXIMADA

(A) σ_u^2 conocida: algoritmo PQL (✓ Breslow & Clayton, 1993):

$$(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}) = \text{argmax}_{(\boldsymbol{\beta},\mathbf{u})} f(\mathbf{y}_s, \mathbf{u})$$

(B) $\boldsymbol{\beta}$ y $\mathbf{u}$ conocido: ML aproximado

$$\hat{\sigma}_u^2 = \text{argmax}_{\sigma_u^2} f_L(\mathbf{y}_s)$$

f_L verosimitud normal multivariante de un modelo lineal mixto que aproxima el GLMM.

✓ *Schall (1991)* ✓ *Saei & Chambers (2003)*

CUASI-VEROSIMILITUD PENALIZADA (PQL) +ML APROXIMADA

- Proporciona estimadores que pueden ser inconsistentes.
- El ECM se puede estimar por bootstrap paramétrico.
- Ajuste GLMM + EBP + ECM bootstrap: altamente intensivo a nivel computacional. Inviable para poblaciones grandes.
- Ajuste GLMM + Estimador *plug-in* + ECM bootstrap: más viable pero no óptimo.

EXTENSIÓN: VARIAS CATEGORÍAS

- Y_{d1} total de desempleados en área d;
- Y_{d2} total de empleados en área d;
- R_d tasa de desempleados en área d;

$$R_d = \frac{Y_{d1}}{Y_{d1} + Y_{d2}} \times 100.$$

MODELO LOGÍSTICO MIXTO MULTINOMIAL

- Tres categorías excluyentes:
  y_{dj1} 1=desempleados, 0=otros
  y_{dj2} 1=empleados, 0=otros
  y_{dj3} 1=inactivo, 0=otros

- **Modelo Multivariante:**

$$(y_{dj1}, y_{dj2}, y_{dj3}) \sim \text{Multin}(m_{dj}; p_{dj1}, p_{dj2}, p_{dj3})$$

$$\begin{aligned} \text{Desempleado}: &\quad \log(p_{dj1}/p_{dj3}) = \mathbf{x}'_{dj1}\boldsymbol{\beta}_1 + u_{d1} \\ \text{Empleado}: &\quad \log(p_{dj2}/p_{dj3}) = \mathbf{x}'_{dj2}\boldsymbol{\beta}_2 + u_{d2} \end{aligned}$$

- Efectos aleatorios específicos para las categorías:
  $\mathbf{u} = (u_{d1}, u_{d2})' \sim N_2(0, \Sigma_u)$.

MODELO LOGÍSTICO MIXTO MULTINOMIAL

- Estimadores *plug-in* de totales de desempleados/empleados:

$$\hat{Y}_{dk}^{Plug} = \sum_{j \in s_d} y_{djk} + \sum_{j \in r_d} \hat{p}_{djk}^{Plug}, \quad k = 1, 2.$$

- Estimadores *plug-in* de tasas de desempleados:

$$R_d^{Plug} = \frac{\hat{Y}_{d1}^{Plug}}{\hat{Y}_{d1}^{Plug} + \hat{Y}_{d2}^{Plug}} \times 100.$$

✓ *Molina, Saei & Lombardía (2007), JRSSA*

MODELOS PARA DATOS BINARIOS

INDICADORES OBJETIVO:

- Proporciones o totales de una variable binaria (ej. acceso o no a cierto servicio o comodidad).

REQUERIMIENTOS DE DATOS:

- Microdatos para las p variables auxiliares en la encuesta.
- Indicador del dominio en la encuesta.
- Microdatos de las p variables auxiliares para todas las unidades poblacionales (censo o registro administrativo).

MODELOS PARA DATOS BINARIOS

VENTAJAS:

- Utiliza información auxiliar de nivel de unidad, que es **más detallada** que la información de nivel de área.
- El tamaño total de la muestra es típicamente muy **grande** ($n >> D$), por lo que se comparte mucha información.
- Incorpora heterogeneidad no explicada entre áreas.
- EB es aprox. **insesgado y óptimo** bajo el modelo.
- Las estimaciones se pueden desagregar para cualquier subdominio (sin efecto de subdominio), incluso a nivel de unidad.
- Para estimar en áreas no muestreadas, se puede usar la parte sintética.

MODELOS PARA DATOS BINARIOS

DESVENTAJAS:

- Información auxiliar para cada unidad de población (censo/registro) **no es fácilmente accesible**.
- Computacionalmente **intensivo**.
- **No utiliza los pesos del muestreo**, por lo que puede ser sergado bajo el diseño, especialmente bajo el muestreo **informativo**.
- **Sensible** a desviaciones del modelo. Encontrar la transformación correcta de la variable y la **diagnosis de modelo** muy importante.
- Estimador EB (al contrario que el tipo *plug-in*) es computacionalmente intensivo.
- Estimadores del ECM por bootstrap son computacionalmente intensivos (aún más para estimadores EB).
- Es necesario un reajuste para que verifiquen la propiedad "benchmarking".

SOFTWARE

El paquete R `sae` contiene funciones:

- Estimadores directos: `direct`.
- Estimadores tradicionales indirectos: `pssynt`, `ssd`.
- Modelo FH: `eblupFH`, `mseFH`.
- Modelo FH espacial: `eblupSFH`, `mseSFH`, `pbmseSFH`, `npbmseSFH`.
- Modelo FH espacio-temporal: `eblupSTFH`, `pbmseSTFH`.
- Modelo con errores anidados: `eblupBHF`, `pbmseBHF`.
- Método EB bajo modelo con errores anidados: `ebBHF`, `pbmseebBHF`.
- Conjuntos de datos y ejemplos.

✓ *Molina & Marhuenda (2015), The R Journal*

RESUMEN

(a) Medidas preventivas sobre el diseño muestral pueden reducir significativamente la necesidad de estimaciones indirectas.

(b) Información auxiliar de calidad relacionada con la variable de interés juega un papel crucial en la estimación basada en modelos. Es necesario facilitar el acceso a la información auxiliar mediante coordinación y cooperación entre instituciones.

(c) La validación del modelo es crucial. También son deseables estudios de evaluación externa.

(d) Los modelos de nivel de área tienen mayor alcance que los modelos de nivel de unidad debido a que la información auxiliar a nivel de área es más fácilmente accesible. Pero la necesidad de conocer las varianzas muestrales de los estimadores directos es restrictiva. Se necesita más investigación para obtener buenas aproximaciones a dichas varianzas muestrales. Los modelos a nivel de unidad pueden ganar mucha más eficiencia si se dispone de los datos necesarios.

REFERENCIAS

- Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data, *J. Amer. Statist. Assoc.*, **83**, 28–36.
- Breslow, N.E., Clayton, D.G. (1993), Approximate inference in generalized linear mixed models. J. Amer. Statist. Assoc. **88**, 9–25.
- Deville, J.C. and Särndal, C.E. (1992). Calibration estimation in Survey Sampling, *J. Amer. Statist. Assoc.*, **87**, 376–382.
- Drew, D., Singh, M.P. and Choudhry, G.H. (1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey, *Survey Methodology*, **8**, 17–47.
- Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures, *Econometrica*, **52**, 761–766.

REFERENCIAS

- Fay, R.E. and Herriot, R.A. (1979). Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data, *J. Amer. Statist. Assoc.*, **74**, 269–277.
- González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. and Santamaría, L. (2007), Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model, *Computational Statistics and Data Analysis*, **51**, 2720–2733.
- Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). Sample Survey Methods and Theory I, New York: Wiley.
- Kackar, R.N. and Harville, D.A. (1984). Approximations for Standard Errors of Estimators of Fixed Random Effects in Mixed Linear Models, *J. Amer. Statist. Assoc.*, **79**, 853–862.
- Prasad, N.G.N. and Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators, *J. Amer. Statist. Assoc.*, **85**, 163–171.

REFERENCIAS

- Molina, I. and Marhuenda, Y. (2015), sae: An R Package for Small Area Estimation, *The R Journal*, **7**.
- Molina, I., Saei, A. and Lombardía, M.J. (2007), Small area estimates of labour force participation under a multinomial logit mixed model, *Journal of the Royal Statistical Society, Series A*, **170**, 975–1000.
- Molina, I. and Rao (2010). Small Area Estimation of Poverty Indicators. *Canadian Journal of Statistics*, **38**, 369–385.
- Rao, J.N.K. and Molina, I. (2015). Small Area Estimation, Second Edition. Wiley: Hoboken, NJ.
- Royall, R.M. (1970). On Finite Population Sampling Theory Under Certain Linear Regression, *Biometrika*, **57**, 377–387.

REFERENCIAS

- Saei, A., Chambers, R., 2003. Small area estimation under linear and generalized linear mixed models with time and area effects. S3RI Methodology Working Paper M03/15. Southampton Statistical Sciences Research Institute, University of Southampton.
- Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719—727.
- Särndal, C.E., Swenson, B. and Wretman, J.H. (1992). Model Assisted Survey Sampling, New York: Springer-Verlag.