CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS, A.C.

Estimación del Ingreso Promedio por Vivienda en los Municipios del Estado de Sonora

Una Aplicación de la Estimación para Áreas Pequeñas

T E S I S

Que para obtener el grado de

Maestro en Ciencias en Estadística Oficial

Presenta:

Miguel Ángel Suárez Campos

Directora de tesis:

Dra. Graciela González Farías

Guanajuato, Gto. Diciembre de 2010

CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS, A.C.

Estimación del Ingreso Promedio por Vivienda en los Municipios del Estado de Sonora

Una Aplicación de la Estimación para Áreas Pequeñas

T E S I S

Que para obtener el grado de

Maestro en Ciencias en Estadística Oficial

Presenta:

Miguel Ángel Suárez Campos

Comité de Evaluación:

_____

Dr. José Elías Rodríguez Muñoz
(Presidente)

_____

M en C. José Vences Rivera
(Secretario)

_____

Dra. Graciela González Farías
(Vocal y Director de Tesis)

Guanajuato, Gto. Diciembre de 2010

A mis Padres
Rosendo e Irma Socorro
y a David Uquillas
que están con el Creador

A mis hijos
Erika Jazmín,
Ángel Tonalli
y Ulises Tonatiúh

Agradecimientos

Primeramente quiero agradecer a mi esposa Silvia Morales Tintor de quien siempre he tenido su apoyo y comprensión en esos tiempos de arduo trabajo y estudio.

A Gustavo Aguilar Mata, Raúl Mejía González, Alfredo Bustos y de la Tijera, José Vences Rivera, Enrique Navarro Luévano, Virgilio Gómez Rubio, José Elías Rodríguez Muñoz y Rogelio Ramos Quiroga, por la ayuda desinteresada que me brindaron durante la realización de esta tesis.

Agradecimiento especial a mi profesora y asesora de tesis Graciela González Farías por compartir su preciado tiempo y transmitirme un poco de su vasto conocimiento.

A los Profesores del CIMAT que participaron en la primera generación de la Maestría en Ciencias en Estadística Oficial por sus múltiples y valiosas enseñanzas.

Agradezco el apoyo parcial del proyecto de Ciencia Básica del Conacyt 105657.

A los funcionarios y personal del Instituto Nacional de Estadística y Geografía que hicieron posible la realización de la Maestría en Ciencias en Estadística Oficial, por darme esta oportunidad de superación y conocimiento.

Al Dr. Abdón Sánchez Arroyo por el apoyo y consejos brindados a todos los que participamos en la maestría, ya que fueron más allá de su encomienda ofreciendonos una sincera amistad.

A todos mis compañeros y amigos de la maestría por su apoyo y por compartir esos momentos de estudio, estrés y alegría, durante los dos años de formación.

A mis hermanos por el apoyo incondicional que he recibido en cualquier etapa de mi vida.

CONTENIDO

INTRODUCCIÓN

En el muestreo de poblaciones finitas y en particular en el sistema estadístico nacional, existe una demanda creciente de estimaciones precisas sobre promedios o indicadores de interés en áreas cada vez más pequeñas, tales como entidades federativas, municipios o localidades, o bien, en subgrupos o pequeños dominios de la población total, como son las subclases de alguna actividad económica. Estas estimaciones se consideran como un subproducto de los trabajos de muestreo en áreas grandes, en donde se diseña una muestra para obtener estimaciones del total o de la media de una característica de interés con una precisión prefijada. Esto implica que el número de observaciones muestrales en un área pequeña es reducido o incluso nulo, por lo que utilizar estimadores basados en el diseño conduce a grandes errores en las estimaciones e incluso es imposible aplicar dichos estimadores en áreas no muestreadas, haciéndose necesario recurrir por un lado al incremento del tamaño muestral cuando todavía es posible y con su correspondiente costo o por otro lado a los métodos de estimación en áreas pequeñas.

Estos métodos o técnicas hacen uso de información extramuestral que puede provenir de otras encuestas con mayor presencia muestral en las áreas pequeñas de interés, de censos, de registros administrativos e inclusive de tipo geográfico, dicha información está contenida en variables auxiliares, de las que se conoce en unos casos, el valor para cada elemento de la población y en otros solamente el total de cada dominio o área pequeña.

Una de las variables importantes para el cálculo de indicadores económicos estratégicos como la pobreza y el PIB, es el ingreso promedio de los hogares, mismo que por su complejidad de obtención sólo se obtiene en México por medio de una encuesta diseñada conceptualmente para ello, se trata de la Encuesta Nacional de Ingreso y Gasto de los Hogares (ENIGH), cuyo diseño muestral sólo contempla los ámbitos urbano y rural a nivel nacional y en algunas ocasiones por petición y financiamiento extraordinario se amplía para algunos Estados de la República.

En este trabajo se combina la información de la ENIGH del año 2005 y el II Conteo de población y Vivienda levantado también en el año 2005 en particular del Estado de Sonora, mediante la aplicación y comparación de técnicas estadísticas que estiman y/o predicen el valor promedio del ingreso de los hogares en los 72 municipios del estado. Para ello el presente documento se ha dividido en cuatro capítulos.

En el primer capítulo se presenta una breve descripción de la ENIGH haciendo énfasis en los detalles del diseño muestral y el método para el cálculo de sus estimaciones, de manera similar se presenta el II Conteo de Población y Vivienda 2005 (II Conteo), su forma de levantamiento y las variables captadas, y como otro tópico se incluyen variables geográficas que se intuye pueden estar relacionadas con el ingreso de los hogares sobre todo en los municipios rurales.

El segundo capítulo se concentra en presentar la formulación matemática de algunos de los métodos más generales de estimación basada en el diseño utilizados en áreas pequeñas con variables cuantitativas, sin adentrarse ni demostrar los detalles de su obtención, sin embargo, para los métodos de estimación basados en el modelo se proporciona una explicación un poco más amplia.

En el tercer capítulo se muestra la aplicación de las fórmulas del segundo capítulo, con los datos combinados de la ENIGH 2005 y el II Conteo, apoyado con rutinas programadas en el software R, en este capítulo se incluyen los resultados.

El cuarto y último capítulo se reserva para las conclusiones.

CAPÍTULO I. ENIGH Y II CONTEO DE POBLACIÓN Y VIVIENDA
2005

1.1 Descripción de la Encuesta Nacional de Ingresos y Gastos de los Hogares

Dado el carácter oficial de la ENIGH 2005, lo escrito en la sección 1 y 2 de este capítulo fue tomado íntegramente de la *Síntesis Metodológica* publicada en el año 2006 (ver bibliografía).

Antecedentes

La Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) surge en el año de 1984; a partir de 1992 se realiza con una periodicidad de levantamiento de cada dos años, con excepción de 2005, ya que fue un levantamiento extraordinario para tener cifras actualizadas sobre las condiciones de vida de los hogares. Al tiempo que se conservó la comparabilidad del marco conceptual, periodos de referencia, unidades de análisis, cobertura geográfica, instrumentos de captación, diseño muestral y procedimientos operativos utilizados en la generación de datos; también se actualizaron otros aspectos relevantes como la metodología, y se incorporaron nuevos productos con el objetivo de adecuarse a los cambios socio-económicos del país y obtener resultados que reflejan la realidad.

Unidad de observación

La unidad de observación que la ENIGH considera para su estudio es el "Hogar", el cual se define como el conjunto formado por una o más personas que residen habitualmente en la misma vivienda y se sostienen de un gasto común, principalmente para alimentarse y pueden ser parientes o no.

Método de captación

La generación de estadísticas de la ENIGH se basa en la aplicación de un esquema de muestreo probabilístico, a su vez el diseño es bietápico, estratificado y por conglomerados, donde la unidad última de selección es la vivienda y la unidad de observación es el hogar y en consecuencia los resultados obtenidos de la encuesta se generalizan a toda la población.

Marco de muestreo

El marco muestral utilizado es el de propósitos múltiples del INEGI, constituido por la información demográfica y cartográfica obtenida a partir del levantamiento del Censo de Población y Vivienda del 2000.

En este marco de muestreo se excluyen a todas las viviendas colectivas y las de diplomáticos extranjeros, ya que para fines de la encuesta no son objeto de estudio.

Selección y tamaño de la muestra

El procedimiento para conformar la muestra se realiza de la siguiente manera: Se seleccionan las viviendas, distinguiendo entre dos clases; la vivienda particular y la colectiva, siendo objeto de la encuesta sólo la primera; se realiza en forma independiente para cada entidad y estrato; el procedimiento varía dependiendo la zona.

<u>Urbano alto</u>

En la zona urbano alto la selección de la muestra se realizó en forma independiente por cada ciudad y estrato mediante el siguiente procedimiento:

De las n_{ech} UPM que integran el marco de la muestra maestra, se eligieron n^*_{ech} UPM con igual probabilidad para la ENIGH-2005.

En cada UPM se seleccionaron 5 viviendas con igual probabilidad para la ENIGH-2005.

Por lo tanto, la probabilidad de seleccionar una vivienda en la i-ésima UPM, en el h-ésimo estrato, de la c-ésima ciudad, de la e-ésima entidad es:

$$P\{V_{echi}\} = \frac{n_{ech} m_{echi}}{m_{ech}} \frac{n^*_{ech}}{n_{ech}} \frac{5}{m^*_{echi}} = \frac{5\, n^*_{ech} m_{echi}}{m_{ech} m^*_{echi}}$$

Su factor de expansión[1] está dado por:

$$F_{echi} = \frac{m_{ech}\ m^*_{echi}}{5\, n^*_{ech}\ m_{echi}}$$

donde:

n_{ech} = número de UPM seleccionadas, en el h-ésimo estrato, de la c-ésima ciudad, en la e-ésima entidad, para el marco de la muestra maestra.

n^*_{ech} = número de UPM seleccionadas para la ENIGH-2005, en el h-ésimo estrato, de la c-ésima ciudad, en la e-ésima entidad.

m_{ech} = número de viviendas en el h-ésimo estrato, de la c-ésima ciudad, en la e-ésima entidad.

m_{echi} = número de viviendas de la i-ésima UPM, del h-ésimo estrato, de la c-ésima ciudad, en la e-ésima entidad.

m^*_{echi} = número de viviendas en la i-ésima UPM, del h-ésimo estrato, de la c-ésima ciudad, en la e-ésima entidad al momento de la actualización del listado de viviendas, previo al levantamiento de la ENIGH-2005.

Complemento urbano

De las n_{eh} UPM que se seleccionaron para el marco de la muestra maestra, se eligieron n^*_{eh} UPM con igual probabilidad para la ENIGH-2005.

En cada UPM seleccionada, se eligieron 20 viviendas con igual probabilidad.

Por lo tanto, la probabilidad de seleccionar una vivienda en la i-ésima UPM, en el h-ésimo estrato, de la e-ésima entidad es:

$$P\{V_{ehi}\} = \frac{n_{eh}\ m_{ehi}}{m_{eh}} \frac{n^*_{eh}}{n_{eh}} \frac{20}{m^*_{ehi}} = \frac{20\, n^*_{eh}\ m_{ehi}}{m_{eh}\ m^*_{ehi}}$$

Su factor de expansión está dado por:

[1] El factor de expansión se define como el inverso de la probabilidad de selección.

$$F_{ehi} = \frac{m_{eh}\ m^{*}_{ehi}}{20\ n^{*}_{eh}\ m_{ehi}}$$

donde:

n_{eh} = número de UPM seleccionadas para el marco de la muestra maestra de la muestra maestra, en el h-ésimo estrato, de la e-ésima entidad.

m_{ehi} = número de viviendas en la i-ésima UPM, en el h-ésimo estrato, de la e-ésima entidad.

m_{eh} = número de viviendas del h-ésimo estrato, para la e-ésima entidad.

n^{*}_{eh} = número de UPM seleccionadas para la ENIGH-2005, en el h-ésimo estrato, de la e-ésima entidad.

m^{*}_{ehi} = número total de viviendas en la i-ésima UPM, en el h-ésimo estrato, de la e-ésima entidad al momento de levantamiento de la ENIGH-2005.

Rural

De las n_{eh} UPM que se seleccionaron para el marco de la muestra maestra, se eligieron n^{*}_{eh} UPM con igual probabilidad para la ENIGH-2005.

En cada UPM seleccionada, se eligieron 2 segmentos de 10 viviendas aproximadamente, con igual probabilidad.

Por lo tanto, la probabilidad de seleccionar una vivienda de la i-ésima UPM, en el h-ésimo estrato, de la e-ésima entidad es:

$$P\{V_{ehi}\} = \frac{n_{eh}\ m_{ehi}}{m_{eh}} \frac{n^{*}_{eh}}{n_{eh}} \frac{2 \cdot 10}{m^{*}_{ehi}} = \frac{20\ n^{*}_{eh}\ m_{ehi}}{m_{eh}\ m^{*}_{ehi}}$$

En consecuencia, su factor de expansión está dado por:

$$F_{ehi} = \frac{m_{eh}\ m^{*}_{ehi}}{20\ n^{*}_{eh}\ m_{ehi}}$$

donde:

n_{eh} = número de UPM seleccionadas del marco de la muestra maestra, en el h-ésimo estrato, de la e-ésima entidad

m_{ehi} = número de viviendas de la i-ésima UPM, en el h-ésimo estrato, de la e-ésima entidad

m_{eh} = número total de viviendas en el h-ésimo estrato, de la e-ésima entidad.

n^{*}_{eh} = número de UPM seleccionadas para la ENIGH-2005, en el h-ésimo estrato, de la e-ésima entidad.

m^{*}_{ehi} = número total de viviendas en la i-ésima UPM, del h-ésimo estrato, de la e-ésima entidad al momento del levantamiento de la ENIGH-2005.

La selección de la muestra, está calculada para dar estimaciones a los siguientes niveles de desagregación:

- Nivel nacional
- Localidades de 2 500 y más habitantes
- Localidades de menos de 2 500 habitantes

En los estados de Puebla, Sonora, Tabasco y Veracruz se incrementó la muestra para poder dar resultados a nivel entidad.

AJUSTE A LOS FACTORES DE EXPANSIÓN

Los factores de expansión elaborados conforme al procedimiento antes descrito se ajustan para los siguientes conceptos:

Ajuste por no Respuesta

El ajuste por no Respuesta atribuida al informante se realiza a nivel UPM, mediante las siguientes expresiones:

$$F'_{echi} = F_{echi} \frac{nvh_{echi}}{nvhcr_{echi}},$$

donde:

F'_{echi} es el factor de expansión corregido por no Respuesta para las viviendas de la i-ésima UPM, en el h-ésimo estrato, de la c-ésima ciudad, de la e-ésima entidad.

F_{echi} es el factor de expansión de la i-ésima UPM, en el h-ésimo estrato, de la c-ésima ciudad, de la e-ésima entidad.

nvh_{echi} es el número de viviendas en muestra habitadas de la i-ésima UPM, en el h-ésimo estrato, de la c-ésima ciudad, de la e-ésima entidad.

$nvhcr_{echi}$ es el número de viviendas en muestra habitadas con respuesta de la i-ésima UPM, en el h-ésimo estrato, de la c-ésima ciudad , de la e-ésima entidad.

Ajuste por proyección

Los factores ajustados por no-respuesta se corrigen, a fin de asegurar que en cada dominio de interés de la encuesta se obtenga la población total determinada por la proyección de población generada por INEGI referida al punto medio del levantamiento, mediante la siguiente expresión:

$$F''_D = F'_D \frac{PROY_D}{PEXP_D}$$

donde:

F''_D es el factor de expansión corregido por proyección en el dominio *D*.

F'_D es el factor de expansión corregido por no Respuesta en el dominio *D*.

$PROY_D$ es la Población en el dominio *D*, según la proyección.

$PEXP_D$ es la Población total a la que expande la encuesta en el dominio *D*.

1.2 ESTIMADORES UTILIZADOS EN LA ENCUESTA NACIONAL DE INGRESO Y GASTOS DE LOS HOGARES

En esta sección se describen los estimadores que utiliza actualmente el INEGI para proporcionar las cifras oficiales de la ENIGH y de otras encuestas, son estimadores directos derivados del estimador de Horvitz-Thompson, cuyos resultados se mencionan brevemente más adelante en este documento.

ESTIMACIÓN DE TOTALES Y PROPORCIONES

El estimador del total de la característica X, a nivel nacional es:

$$\hat{X} = \sum_e \sum_c \sum_h \sum_i F_{echi}^{UA} \left(\sum_s \sum_l X_{echisl}^{UA} \right) + \sum_e \sum_h \sum_i F_{ehi}^{CU} \left(\sum_s \sum_l X_{ehisl}^{CU} \right) + \sum_e \sum_h \sum_i F_{ehi}^{R} \left(\sum_s \sum_l X_{ehisl}^{R} \right)$$

donde:

F_{echi}^{UA} es el factor de expansión final, de la i-ésima UPM, en el h-ésimo estrato, de la c-ésima ciudad, en la e-ésima entidad en el dominio urbano alto.

X_{echisl}^{UA} es el valor observado de la característica de interés X del l-ésimo hogar, de la s-ésima vivienda, de la i-ésima UPM, en el h-ésimo estrato, de la c-ésima ciudad, en la e-ésima entidad en el dominio urbano alto.

F_{ehi}^{R} es el factor de expansión final de la i-ésima UPM, del h-ésimo estrato, de la e-ésima entidad del dominio rural.

X_{ehisl}^{R} es el valor observado de la característica X en el □-ésimo hogar, de la s-ésima vivienda, en la i-ésima UPM, del h-ésimo estrato, en la e-ésima entidad del dominio rural.

F_{ehi}^{CU} es el factor de expansión final de la i-ésima UPM, del h-ésimo estrato, en la e-ésima entidad, del dominio complemento urbano.

X_{ehisl}^{CU} es el valor observado de la característica X en el □-ésimo hogar, en la s-ésima vivienda, de la i-ésima UPM, del h-ésimo estrato, en la e-ésima entidad del dominio complemento urbano.

Para la estimación de proporciones, tasas y promedios se utiliza el estimador de razón.

$$\hat{R} = \frac{\hat{X}}{\hat{Y}}$$

donde, la variable $\hat{Y}$ es definida en forma análoga a $\hat{X}$.

ESTIMACIÓN DE LAS PRECISIONES

Para la evaluación de los errores de muestreo de las principales estimaciones estatales y nacionales se usó el método de "Conglomerados Últimos",[2] basado en que la mayor contribución a la varianza de un estimador, en un diseño polietápico, es la que se presenta entre las unidades

[2] Vease Hansen, M.H. Horwitz, W.N. y Madow, W.G, *Sample Survey Methods and Therry*, (1953), Vol. 1 página 242.

primarias de muestreo (UPM), el término “Conglomerados Últimos” se utiliza para denotar el total de unidades en muestra de una unidad primaria de muestreo.

Para obtener las precisiones de los estimadores de razón, se aplicó el método de Conglomerados Últimos conjuntamente con el método de series de Taylor, obteniéndose la siguiente fórmula para estimar la precisión de $\hat{R}$.

$$\hat{V}(\hat{R}_{NAL}) = \frac{1}{\hat{Y}^2_{NAL}} \sum_{e=1}^{32} \left\{ \sum_{h=1}^{L_e} \frac{n_{eh}}{n_{eh}-1} \sum_{i=1}^{n_{eh}} \left[\left(\hat{X}_{ehi} - \frac{1}{n_{eh}} \hat{X}_{eh} \right) - \hat{R}_{NAL} \left(\hat{Y}_{ehi} - \frac{1}{n_{eh}} \hat{Y}_{eh} \right) \right]^2 \right\}$$

donde:

$\hat{X}_{ehi}$ es el total ponderado de la variable de estudio X, para la i-ésima UPM, en el h-ésimo estrato, de la e-ésima entidad

$\hat{X}_{eh}$ es el total ponderado de la variable de estudio X, para el h-ésimo estrato, en la e-ésima entidad.

n_{eh} es el número de UPM, en el h-ésimo estrato, para la e-ésima entidad.

Estas definiciones son análogas para la variable de estudio Y.

La estimación de la varianza del estimador de un total, se calcula con la siguiente expresión.

$$\hat{V}\left(\hat{X}_{NAL}\right) = \sum_{e=1}^{32} \sum_{h=1}^{L_e} \frac{n_{eh}}{n_{eh}-1} \sum_{i=1}^{n_{eh}} \left(\hat{X}_{ehi} - \frac{1}{n_{eh}} \hat{X}_{eh} \right)^2 \qquad (1)$$

Las estimaciones del error estándar (E.E.), coeficiente de variación (C.V.) o error relativo del estimador y el efecto de diseño (DEFF) se calculan mediante las siguientes expresiones.

$$E.E. = \sqrt{\hat{V}\left(\hat{\theta}\right)} \qquad C.V. = \frac{\sqrt{\hat{V}\left(\hat{\theta}\right)}}{\hat{\theta}} \qquad DEFF = \frac{\hat{V}\left(\hat{\theta}\right)}{\hat{V}\left(\hat{\theta}\right)_{mas}}$$

donde:

$\hat{\theta}$ es el estimador del parámetro poblacional θ.

$\hat{V}\left(\hat{\theta}\right)_{mas}$ es el estimador de la varianza, bajo muestreo aleatorio simple.

Finalmente, el intervalo de confianza al $(1-\alpha\%)$, se construye de la siguiente forma:

$$I_{1-\alpha} = \left(\hat{\theta} - z_{\alpha/2} \sqrt{\hat{V}\left(\hat{\theta}\right)}, \hat{\theta} + z_{\alpha/2} \sqrt{\hat{V}\left(\hat{\theta}\right)} \right)$$

1.3 Descripción del II Conteo de Población y Vivienda 2005

Para dar continuidad al proceso de actualización de la estadística sociodemográfica, se realizó el II Conteo de Población y Vivienda 2005, con lo cual se contribuye al proceso de planificación general del país, a la formulación de políticas públicas, a la evaluación de programas en aspectos tales

como educación y alfabetización, planificación de la familia, vivienda, desarrollo rural, urbanización y bienestar social, además de que se fortalecen los Sistemas Nacionales Estadístico y de Información Geográfica.

Un censo o un conteo poblacional se concibe como el conjunto integrado de operaciones que permiten organizar y recopilar información sociodemográfica de corte estadístico en forma simultánea y homogénea; procesar, analizar, difundir y evaluar los datos relativos a todos los habitantes de un país, los hogares y las viviendas, en un momento determinado.

Su característica fundamental es la aplicación del método censal, cuya exhaustividad en el empadronamiento permite generar información estadística desagregada para muy pequeñas unidades geográficas y para todos los grupos de población, apegándose estrictamente a las restricciones que marca el principio de confidencialidad.

BASES METODOLÓGICAS

Los componentes metodológicos fundamentales fueron:

El método empleado es censal, porque abarca a todas las viviendas y personas que residen en el territorio nacional en un mismo periodo, con lo que se cumple la característica de universalidad y simultaneidad.

El tipo de levantamiento es de derecho, es decir, la población se captó en su lugar de residencia habitual.

El periodo de levantamiento fue del 4 al 29 de octubre de 2005 y se dispuso de dos semanas adicionales para recuperar la información de aquellas viviendas que por diversos motivos quedaron pendientes.

Dado que el operativo del II Conteo abarcó varias semanas, se definió el 17 de octubre de 2005 como la fecha a la que se refiere la información, que corresponde a la mitad del periodo del levantamiento.

Las unidades de observación fueron:

- Las viviendas y los hogares ubicados en el territorio nacional.
- Los residentes habituales de las viviendas particulares y colectivas.
- Los residentes habituales que no viven en una vivienda.

La información se captó mediante entrevista directa con la aplicación de un cuestionario por hogar; fue proporcionada por el jefe(a) del hogar, su cónyuge o una persona de 15 o más años de edad residente en la vivienda, que conocía la información de ésta y de sus ocupantes.

Para contar con el esquema básico de información que captaría el Conteo de 2005 se efectuó un análisis tanto de la situación como de la dinámica demográfica, social y económica del país; de la demanda de información de interés público para diversos sectores, entre los que destacan las instituciones públicas, privadas y académicas; de la experiencia acumulada y de la evaluación de los resultados de censos anteriores; de la comparabilidad intercensal necesaria; de las recomendaciones internacionales rectoras; de las experiencias exitosas de otros organismos de estadística; y de los desarrollos teóricos y las nuevas perspectivas de análisis sociodemográfico.

Con base en este análisis se identificaron un total de 23 variables para el II Conteo de Población y Vivienda 2005, mismas que se presentan a continuación.

Características de la población
Parentesco
Sexo
Edad
Derchohabiencia a servicios de salud
Lugar de residencia cinco años antes
Condición de habla lengua indígena
Lengua indígena
Condición de habla española
Alfabetismo
Asistencia escolar
Escolaridad
Número de hijos
Características de las viviendas
Clase de vivienda particular
Material en pisos
Número de dormitorios
Total de cuartos
Disponibilidad de energía eléctrica
Disponibilidad de agua
Disponibilidad de excusado o sanitario
Disponibilidad de drenaje
Disponibilidad de bienes
Actividades agropecuarias
Número de hogares

1.4 INFORMACIÓN AUXILIAR GEOGRÁFICA

Además de la información censal, una fuente confiable es la información geográfica, que para algunos temas es posible obtenerla a nivel municipal. Los temas geográficos que se consideran relacionados con el ingreso son; clima, uso del suelo y fisiográficos (topoformas), mismos que fueron elegidos por el CONEVAL en un modelo para la estimación de la pobreza.

De estos temas se construyeron una serie de indicadores dicotómicos, a partir de cartas escala 1:1,000,000 referenciadas y actualizadas en el año 2005, y que están cargadas en la Base de Datos Geográfica del INEGI.

Los indicadores por tema a nivel municipal se presentan como sigue:

Indicador de clima
Seco templado
Seco muy cálido
Muy seco semicálido
Semiseco semicálido
Semiseco semifrío
Semicálido húmedo
Semicálido subhúmedo
Semifrío subhúmedo
Frío
Indicador de uso del suelo
Pastizal
Selva
Área agrícola
Bosque
Indicador fisiográfico de topoforma
Bajada
Depresión
Cañón

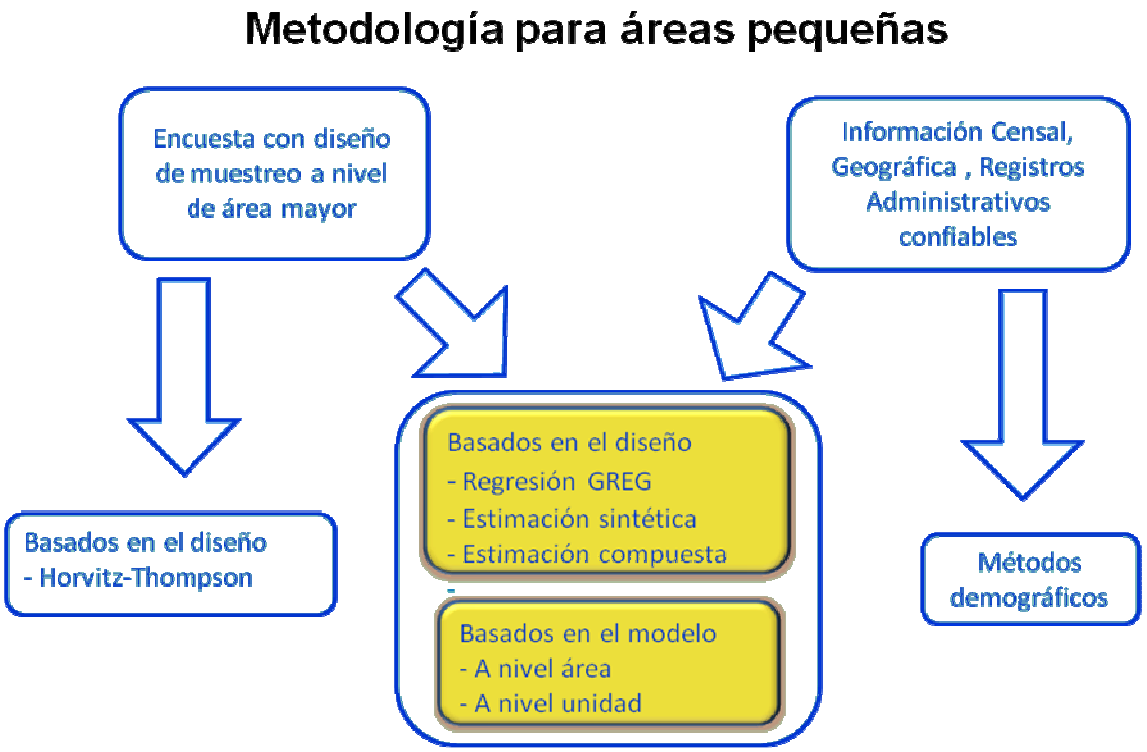
CAPÍTULO II. METODOLOGÍA DE ESTIMACIÓN PARA ÁREAS PEQUEÑAS

2.1 ANTECEDENTES

Tradicionalmente, los procedimientos de inferencia sobre características de poblaciones finitas se han basado en el diseño de la muestra, sin embargo, en la actualidad son cada vez más numerosos los procedimientos de inferencia de aproximación basada en modelos (Cassel, Sarndal y Wretman, 1977). En esta aproximación, se le da más peso al modelo de la distribución de la característica en estudio en la población que al proceso de diseño de la muestra, y se supone, la aceptación de algunos riesgos (Hansen, Madow, 1983) como el de no estar seguros de que el modelo que se adopta es el correcto, pero aún así deben de ser aceptados ya que sólo es posible considerar la inferencia basada en el diseño libre de supuestos cuando la muestra es grande, lo cual no ocurre con la estimación para áreas pequeñas en donde el tamaño de la muestra es pequeño. El recurrir a la inferencia basada en modelos permite solventar problemas no abordables por la vía del diseño como la no respuesta o la estimación en áreas pequeñas (Särndal, 1984).

Diferentes métodos para la estimación en áreas pequeñas fueron recopilados por Harter (1993) para obtener estimaciones de la población en ciudades y municipios en periodos intercensales, más recientemente Rao (2003) hace una compilación de estos métodos en su libro “Small Area Estimation”.

En el presente capítulo, se pretende analizar distintos procedimientos para la estimación en áreas pequeñas, exponiendo brevemente sus alcances y limitaciones, en el diagrama 1 se presentan los procedimientos aquí abordados y su relación con las fuentes de información disponible, siendo esto fundamental para poder proponer los procedimientos posibles para la estimación y elegir el adecuado para tal o cuál área pequeña.



Fuente: Diseño propio

2.2 MÉTODOS DEMOGRÁFICOS TRADICIONALES

Desde mediados del siglo pasado los demógrafos han utilizado una variedad de métodos para la estimación de poblaciones locales y otras características de interés en años intercensales. Los métodos demográficos llamados también "Técnicas con Consideraciones Sintomáticas" (Symptomatic Accouting Techniques) hacen uso de la información captada en los registros administrativos demográficos, como los nacimientos y defunciones cuyas variables son llamadas "sintomáticas" y la ocupan en conjunción con los datos censales.

Estás técnicas se basan en el supuesto de que existe un comportamiento similar entre el área pequeña y una mayor que la contiene.

MÉTODO DE PROPORCIONES VITALES

Este método utiliza únicamente los nacimientos y defunciones ocurridas en el año t y supone que las variaciones de las proporciones de natalidad y fallecimientos entre el último censo y la actualidad se mantienen en el área pequeña a y un área mayor A que la contiene. Además, es necesario obtener de manera independiente $\hat{t}_{\varsigma,A}$ que es la estimación de la población del área mayor A en el tiempo ς.

La estimación para el total de la población está dada por:

$$\hat{t}_{\varsigma,a}=\frac{1}{2}\left(\frac{n_{\varsigma,a}}{\hat{r}_{1\varsigma,a}}+\frac{d_{\varsigma,a}}{\hat{r}_{2\varsigma,a}}\right) \text{ donde } \hat{r}_{1\varsigma,a}=\frac{n_{\varsigma,A}}{\hat{t}_{\varsigma,A}}\frac{t_{0,A}}{n_{0,A}}\frac{n_{0,a}}{t_{0,a}} \text{ y } \hat{r}_{2\varsigma,a}=\frac{d_{\varsigma,A}}{\hat{t}_{\varsigma,A}}\frac{t_{0,A}}{d_{0,A}}\frac{d_{0,a}}{t_{0,a}}$$

$n_{\varsigma,A}$ son los nacimientos ocurridos en el área o región A en el año ς.
$n_{0,A}$ son los nacimientos ocurridos en el área o región A en el año del último censo.
$n_{\varsigma,a}$ son los nacimientos ocurridos en el área pequeña a en el año ς.
$n_{0,a}$ son los nacimientos ocurridos en el área pequeña a en el año del último censo.
$d_{\varsigma,A}$ son las defunciones ocurridas en el área o región A en el año ς.
$d_{0,A}$ son las defunciones ocurridas en el área o región A en el año del último censo.
$d_{\varsigma,a}$ son las defunciones ocurridas en el área pequeña a en el año ς.
$d_{0,a}$ son las defunciones ocurridas en el área pequeña a en el año del último censo.
$t_{0,A}$ es el total de población en el área o región A en el año del último censo.
$t_{0,a}$ es el total de población en el área pequeña a en el año del último censo.

MÉTODO DE LAS COMPONENTES

Este método utiliza los nacimientos, defunciones y la migración neta ocurridas en el área pequeña en el periodo transcurrido entre el último censo (0) y la actualidad (ς), se basa en los registros administrativos y en el total de población obtenido del último censo ($t_{0,a}$), la población actual puede expresarse como:

$$t_{\varsigma,a}=t_{0,a}+n_{0,\varsigma,a}-d_{0,\varsigma,a}+m_{0,\varsigma,a},$$

MÉTODO DE REGRESIÓN SINTOMÁTICA

Es un método de regresión mediante el cual se quiere explicar a la variable a estimar mediante su relación lineal con variables sintomáticas a través de una regresión lineal:

$$Y_a=\beta_0+\beta_1 x_a+e_a,$$

donde las e_a son errores aleatorios no correlacionados con media 0 y varianza s_e^2.

Tomando el *método de correlación de proporciones* se utilizan las medidas de los cambios en el periodo (0,1) como:

$$Y_a=\frac{t_{1,a}/t_{1,A}}{t_{0,a}/t_{0,A}},\quad x_a=\frac{s_{1,a}/s_{1,A}}{s_{0,a}/s_{0,A}}$$

Y tomando el *método de la correlación por diferencias* las medidas de los cambios se toman como:

$$Y_a=t_{1,a}/t_{1,A}-t_{0,a}/t_{0,A},\quad x_a=s_{1,a}/s_{1,A}-s_{0,a}/s_{0,A}$$

La ecuación se ajusta a los datos x_a y Y_a por mínimos cuadrados ordinarios , así el cambio en el periodo postcensal Y_a está dado por el ajuste de regresión

$$Y_a^*=\hat{\beta}_0+\hat{\beta}_1 x_a$$

Usando los cambios conocidos x_a en las proporciones sintomáticas en el periodo postcensal, el total de población correspondiente al *método de correlación de proporciones* se estima como:

$$\hat{t}_{\varsigma,a}=Y_a^*\left(t_{0,a}/t_{0,A}\right)\hat{t}_{\varsigma,A},$$

El mismo total de población, pero estimado por el *método de la correlación por diferencias* es:

$$\hat{t}_{\varsigma,a}=\left[Y_a^*-\left(t_{0,a}/t_{0,A}\right)\right]\hat{t}_{\varsigma,A}$$

2.3 ESTIMACIÓN BASADA EN EL DISEÑO

Son estimadores insesgados, pero suelen tener grandes varianzas cuando la muestra es pequeña, situación típica en las áreas pequeñas. El diseño de los factores de expansión $w_j(s)$ juega un papel muy importante en la construcción de estos estimadores.

En este tipo de estimación, el diseño de muestreo determina la forma de estimar la variabilidad del muestreo, es decir, la varianza es la desviación cuadrada promedio de la estimación con respecto a su valor esperado, promediado sobre todas las muestras que se podrían obtener mediante un diseño dado (Lohr, 2000).

ESTIMADOR DE HORVITZ-THOMSON

Este estimador es probablemente el más utilizado en los Institutos de estadística y en particular en el INEGI.

Suponiendo que t es el total poblacional del parámetro de interés en el área pequeña a, el estimador de Horvitz-Thompson $\hat{t}_{y,\pi,a}$ del total t de la característica y, se define como:

$$\hat{t}_{y,\pi,a}(s_a) = \sum_{k \in s_a} \frac{y_k}{\pi_k} = \sum_{k \in s_a} w_k y_k \ , \qquad (2)$$

donde S es una muestra $(y_1,\ldots,y_n)$ de tamaño n seleccionada de la subpoblación U_a formada por los elementos $(z_1,\ldots z_{Na})$ con N_a el tamaño de la subpoblación U_a, π_k es la probabilidad de inclusión de primer orden del elemento y_k en la muestra S_a bajo el diseño y $w_k = 1/\pi_k$ es el factor de expansión de y_k.

Un estimador para la media de la variable de interés y está dado por:

$$\hat{\bar{Y}}_{a,\pi} = \frac{1}{\hat{N}_a} \sum_{i \in S_a} w_{ia} y_{ia} \quad \text{donde} \quad \hat{N}_a = \sum_{i \in S_a} w_{ia} \qquad (3)$$

El estimador $E\hat{C}M$ del error cuadrático medio ECM del estimador de la media $\hat{\bar{Y}}_{a,\pi}$, para un tamaño de muestra fijo (Särndal et al, 1992 página 391) , está dado por:

$$E\hat{C}M(\hat{\bar{Y}}_{a,\pi}) = \left(\frac{1}{\hat{N}_a}\right)^2 \sum_{i \in S_a} w_{ia}\left(w_{ia}-1\right)\left(y_{ia} - \hat{\bar{Y}}_{a,\pi}\right)^2, \qquad (4)$$

suponiendo que $\pi_{ia,ja'}=0$, siempre que $a \neq a'$ ó $i \neq j$.

El coeficiente de variación se estima con

$$CV(\hat{\bar{Y}}_{a,\pi}) = \frac{\sqrt{E\hat{C}M(\hat{\bar{Y}}_{a,\pi})}}{\hat{\bar{Y}}_{a,\pi}}, \qquad (5)$$

ESTIMADOR DE REGRESIÓN GENERALIZADO (GREG)

El estimador de regresión generalizado se diferencia del estimador de regresión habitual en que introduce pesos en la estimación de los coeficientes del modelo (normalmente los pesos del muestreo w). Este tipo de estimadores utilizan los modelos de regresión como un medio para conseguir estimadores consistentes desde el punto de vista del diseño, requiere que el muestreo sea aleatorio y que la muestra sea de tamaño grande para proporcionar estimaciones insesgadas, han sido propuestos por Särndal, Swensson y Wrettman (1989). El estimador de regresión generalizado del promedio en el área a viene dado por

$$\hat{\bar{Y}}_{a,GREG} = \hat{\bar{Y}}_{a,\pi} + (\bar{\mathbf{X}}_a - \hat{\bar{X}}_{\pi,a})\hat{\beta}_{GREG} \qquad (6)$$

donde

$\hat{\beta}_{GREG} = \left(\sum_{j=1}^{n} w_j x_j x_j^T\right)^{-1} \sum_{j=1}^{n} w_j x_j y_j$, y $\bar{\mathbf{X}}_a = \left(\bar{X}_{a,1},\ldots,\bar{X}_{a,p}\right)^T$ es un vector de promedios de p variables auxiliares poblacionales conocidas, además se supone que

$$E(e_{ia}) = 0, \quad \text{var}(e_{ia}) = \sigma_e^2$$

Un estimador para el *ECM* del estimador GREG (Särndal et al, 1992, página 401) es:

$$\hat{ECM}(\hat{\bar{Y}}_{a,GREG}) = \left(\frac{1}{\hat{N}_a}\right)^2 \sum_{i \in S_a} w_{ia}(w_{ia} - 1) g_{ia}^2 r_{ia}^2 \qquad (7)$$

donde

$$g_{ia} = 1 + \left(\bar{X}_a + \bar{x}_a\right)^T \left(\sum_{i \in S_a} w_{ia} x_{ia} x_{ia}^T\right)^{-1} x_{ia} \quad \text{y} \quad r_{ia} = y_{ia} - x_{ia}^T \hat{\beta}$$

El estimador del coeficiente de variación del estimador GREG está dado por

$$\hat{CV}(\hat{\bar{Y}}_{a,GREG}) = \frac{\sqrt{\hat{ECM}(\hat{\bar{Y}}_{a,GREG})}}{\hat{\bar{Y}}_{a,GREG}} \qquad (8)$$

ESTIMADOR SINTÉTICO

Se llama estimador sintético a aquel estimador confiable para un área mayor que cubre completamente a las áreas pequeñas. Y se utiliza para derivar un estimador para el área pequeña bajo el supuesto de que ésta tiene las mismas características del área mayor (González 1973), además, utilizan información auxiliar procedente de otros dominios. Estos estimadores tendrán una menor varianza, pero pueden estar sesgados en la medida del incumplimiento del supuesto.

Se dispone de información auxiliar (promedio poblacional por área pequeña) de la variable x, el estimador sintético del promedio de y está dado por:

$$\hat{\bar{Y}}_{a,syn} = \bar{\mathbf{X}}_a \hat{\beta} \text{ donde } \hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i^T\right)^{-1} \sum_{i=1}^{n} x_i y_i \qquad (9)$$

El estimador es aproximadamente insesgado si el vector de coeficientes general β es aproximadamente igual al vector específico de cada área a, esto es:

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i^T\right)^{-1} \sum_{i=1}^{n} x_i y_i \approx \left(\sum_{i=1}^{n_a} x_i x_i^T\right)^{-1} \sum_{i=1}^{n_a} x_i y_i = \hat{\beta}_a$$

La varianza del estimador $\hat{\bar{Y}}_{a,syn}$ es pequeña cuando el tamaño de la muestra global n es grande y está dada por:

$$\hat{\text{var}}(\hat{\bar{Y}}_{a,syn}) = \bar{\mathbf{X}}_a^T \hat{\mathbf{V}}_{\hat{\beta}} \bar{\mathbf{X}}_a \ , \qquad (10)$$

donde $\hat{\mathbf{V}}_{\hat{\beta}}$ es la matriz de varianzas y covarianzas de los coeficientes (betas) estimados.

Debido a que el sesgo en el área pequeña a puede no ser cero y a que el modelo no ajusta bien en el dominio de interés, es deseable estimar el error cuadrático medio (ECM) como medida de precisión del estimador, está dado por

$$ECM(\hat{\bar{Y}}_{a,syn}) = \text{var}(\hat{\bar{Y}}_{a,syn}) + (sesgo_{a,syn})^2$$

A su vez Särndal y Hidiroglou proporcionan una aproximación del sesgo del estimador sintético según la cual $sesgo_{syn} = E(\hat{\bar{Y}}_{a,syn}) - \bar{Y}_{a,syn} \approx -\sum_{j=1}^{n} e_j$, por lo que el ECM se estima mediante la expresión

$$\hat{ECM}(\hat{\bar{Y}}_{a,syn}) = \hat{\text{var}}(\hat{\bar{Y}}_{a,syn}) + (\sum_{j=1}^{n_a} \hat{e}_j)^2 \qquad (11)$$

donde $\hat{e}_j$ son los residuos obtenidos a partir del modelo estimado con todos los datos muestrales, aunque en cada área solamente se suman los específicas de esa área.

El coeficiente de variación viene dado por

$$\hat{CV}(\hat{\bar{Y}}_{a,syn}) = \frac{\sqrt{\hat{ECM}(\hat{\bar{Y}}_{a,syn})}}{\hat{\bar{Y}}_{a,syn}} \qquad (12)$$

ESTIMADOR COMPUESTO

Este tipo de estimador, trata de equilibrar el sesgo potencial del estimador sintético $\hat{\bar{Y}}_{a,syn}$ con la inestabilidad del estimador directo $\hat{\bar{Y}}_{\pi,a}$, como un promedio ponderado de estos estimadores, así el estimador compuesto del total para el área pequeña puede escribirse como:

$$\hat{\bar{Y}}_{a,C} = \phi_a \hat{\bar{Y}}_{a,syn} + (1-\phi_a)\hat{\bar{Y}}_{a,\pi} \text{ ,donde } 0 \le \phi_a \le 1 \qquad (13)$$

El error cuadrático medio de un estimador compuesto está dado por:

$$ECM(\hat{\bar{Y}}_{a,C}) = \phi_a^2 ECM(\hat{\bar{Y}}_{a,syn}) + (1-\phi_a)^2 ECM(\hat{\bar{Y}}_{a,\pi}) + 2\phi_a(1-\phi_a)E(\hat{\bar{Y}}_{a,\pi} - \bar{Y}_a)(\hat{\bar{Y}}_{a,syn} - \bar{Y}_a)$$

Su estimación no es fácil, ya que puede ocurrir que el término de la covarianza no sea pequeña, sin embargo, en proyectos de la Eustat se han utilizado la siguiente aproximación:

$$E(\hat{\bar{Y}}_{a,\pi} - \bar{Y}_a)(\hat{\bar{Y}}_{a,syn} - \bar{Y}_a) \approx ECM(\bar{Y}_{a,\pi}) - \hat{\bar{Y}}_{a,syn}(sesgo_{a,syn})$$

Entonces, el estimador del ECM queda como:

$$\hat{ECM}(\hat{\bar{Y}}_{a,C}) = \phi_a^2 \hat{ECM}(\hat{\bar{Y}}_{a,syn}) + (1-\phi_a)^2 \hat{ECM}(\hat{\bar{Y}}_{a,\pi}) + 2\phi_a(1-\phi_a)\left(\hat{ECM}(\hat{\bar{Y}}_{a,\pi}) - \hat{\bar{Y}}_{a,syn}\left(sesgo_{a,syn}\right)\right) \quad (14)$$

El término $\hat{ECM}(\hat{\bar{Y}}_{a,\pi}) - \hat{\bar{Y}}_{a,syn}(sesgo_{a,syn})$ puede llegar a ser negativo, en este caso puede aproximarse por cero. Cuando n_a = 0 se toma $\hat{ECM}(\hat{\bar{Y}}_{a,C}) = \hat{ECM}(\hat{\bar{Y}}_{a,syn})$ entonces el estimador compuesto es igual al estimador sintético.

Existe un valor de ϕ_a óptimo en el sentido de que se obtiene el menor error cuadrático medio, su valor aproximado es:

$$\phi_a = \frac{ECM(\hat{\bar{Y}}_{a,syn})}{ECM(\hat{\bar{Y}}_{a,syn}) + ECM(\hat{\bar{Y}}_{a,\pi})}, \quad (15)$$

bajo el supuesto de que la covarianza de $E(\hat{\bar{Y}}_{a,\pi} - \bar{Y}_a)(\hat{\bar{Y}}_{a,syn} - \bar{Y}_a)$ es pequeña respecto al $ECM(\hat{\bar{Y}}_{a,syn})$.

Un caso especial del estimador compuesto es el estimador que depende del tamaño de la muestra ,DTM. Drew, Singh y Choudhry (1982) proponen el uso del estimador DTM con los pesos siguientes:

$$\phi_a(T) = \begin{cases} 1 & si\ \hat{N}_a / N_a \geq K_0 \\ \hat{N}_a / K_0 N_a & si\ \hat{N}_a / N_a < K_0 \end{cases}$$

Donde $\hat{N}_a$ es la estimación directa de N_a y K_0 es escogida subjetivamente para controlar la contribución del estimador sintético. La Encuesta de Fuerza de Trabajo de Canadá utiliza K_0 = 2/3.

El estimador del coeficiente de variación viene dado por

$$\hat{CV}(\hat{\bar{Y}}_{a,C}) = \frac{\hat{recm}(\hat{\bar{Y}}_{a,C})}{\hat{\bar{Y}}_{a,C}} \quad \text{donde} \quad \hat{recm}(\hat{\bar{Y}}_{a,C}) = \sqrt{\hat{ECM}(\hat{\bar{Y}}_{a,C})} \quad (16)$$

2.4 ESTIMACIÓN BASADA EN EL MODELO

En la estimación basada en el modelo, el modelo determina la forma de estimar la variabilidad, y el diseño de muestreo no es importante; mientras se conserve el modelo se pueden elegir cualesquiera *n* unidades de la población (Lohr, 2000).

Como ya se ha expuesto, el uso de información auxiliar es necesario para establecer modelos que permitan relacionar a las áreas con escasez o inexistencia de muestra con las áreas próximas con mayor información muestral y poder así incrementar la precisión del estimador. Estos modelos

pueden ser implícitos, como son los estimadores sintéticos o los combinados, o explícitos, como son los basados en modelos. El uso de éstos últimos tiene las siguientes ventajas:

- El diagnóstico del modelo puede usarse para encontrar el modelo adecuado que mejor ajuste a los datos, este tipo de diagnóstico incluye análisis de residuales cuyo resultado verificaría la validez del modelo supuesto, la selección de variables auxiliares para el modelo, y la detección y supresión de observaciones influyentes.
- A diferencia de los estimadores sintéticos y los combinados, los basados en modelos permiten obtener medidas estables de variabilidad de las estimaciones para cada área pequeña.
- Estos métodos pueden utilizarse en casos complejos tanto en casos longitudinales como transversales.
- Pueden utilizarse las metodologías recientemente desarrolladas para modelos de efectos aleatorios, con el fin de lograr inferencias precisas en el área pequeña.

Para comenzar, se presenta un breve repaso sobre los modelos lineales mixtos en general, después se acota a los de estructura de varianza en bloque diagonal, que son los que se aplican en la mayoría de los casos de estimación en áreas pequeñas.

La notación matricial de un modelo lineal mixto está dado por:

$$y = \mathbf{X}\beta + \mathbf{Z}\upsilon + e \ ,$$

donde

y es el vector de respuestas de la variable de interés (n x 1)

$\mathbf{X}$ es la matriz de diseño de los efectos fijos (n x p) de rango completo

β es el vector de coeficientes de los efectos fijos (p x 1)

$\mathbf{Z}$ es la matriz de diseño de los efectos aleatorios (n x q) de rango completo

υ es el vector de coeficientes de los efectos aleatorios (q x 1), las υ_i se piensan como variables aleatorias y no como parámetros

e es el vector de errores residuales del modelo (n x 1)

Se considera que $E(\upsilon)=E(\mathbf{e})=0$ por definición $E(\mathbf{y})=\mathbf{X}\beta$, se supone que υ y e son no correlacionadas. Se denota $\mathbf{R} = \sigma_e^2\mathbf{I}$ a la matriz de varianzas de los errores residuales e y $\mathbf{G} = \sigma_\upsilon^2\mathbf{I}$ a la matriz de varianzas de los efectos aleatorios υ, la matriz de covarianzas del vector de observaciones y es

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathbf{T}} + \mathbf{R}$$

Esto implica que $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \mathbf{V}) = \mathcal{N}(\mathbf{X}\beta, \mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathbf{T}} + \mathbf{R})$

Los componentes de las matrices $\mathbf{G}$ y $\mathbf{R}$ son llamados componentes de la varianza y se aplican a la población en su conjunto, es decir, a las unidades muestreadas y no muestreadas.

Las inferencias de los efectos fijos son llamadas estimaciones, mientras que las de los efectos aleatorios son conocidas como predicciones. Los procedimientos para obtener estas estimaciones

son MELI y MPLI se refieren respectivamente al mejor estimador lineal insesgado y mejor predictor lineal insesgado, *mejor* en el sentido de que minimiza la varianza muestral, *lineal* porque son funciones lineales de las observaciones, e *insesgados* en el sentido de que E[MELI(β)]= β y E[MPLI(υ)] = υ.

Las estimaciones MELI y MPLI para β **y** υ respectivamente son:

$$\hat{\beta} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}y$$

$$\hat{\upsilon} = \mathbf{G}\mathbf{Z}^{\mathbf{T}}\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)$$

La solución de estas ecuaciones requieren la inversa de la matriz de covarianzas $\mathbf{V}$, misma que se puede calcular sin dificultad, sin embargo, cuando y contiene muchos miles de observaciones, como es comúnmente el caso de las encuestas del INEGI, el cálculo de $\mathbf{V}^{-1}$ puede acarrear ciertas dificultades, Henderson (1950, 1963, 1973, 1984) ofrece un método más compacto para obtener conjuntamente $\hat{\beta}$ y $\hat{\upsilon}$ en la forma de las ecuaciones del modelo mixto.

$$\begin{pmatrix} \mathbf{X}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix}\begin{pmatrix} \hat{\beta} \\ \hat{\upsilon} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^{\mathbf{T}}\mathbf{R}^{-1}y \\ \mathbf{Z}^{\mathbf{T}}\mathbf{R}^{-1}y \end{pmatrix}$$

Aunque esta expresión parece más complicada, obtener $\mathbf{R}^{-1}$ y $\mathbf{G}^{-1}$ es trivial ya qe $\mathbf{R}$ y $\mathbf{G}$ son matrices diagonales, las submatrices son mucho más fáciles de calcular que $\mathbf{V}^{-1}$. La matriz que es necesario invertir para encontrar $\hat{\beta}$ y $\hat{\upsilon}$ es de dimensión (p+q) x (p+q), donde p es el número de covariables para los efectos fijos y q es el número de covariables de los efectos aleatorios, la suma de estas es mucho menor que el número de observaciones n, la dimensión de $\mathbf{V}$ es n x n.

Sin embargo, para la solución de las ecuaciones del modelo mixtos suponen que las matrices de covarianzas $\mathbf{R}$ y $\mathbf{G}$ son conocidas. Bajo este supuesto Henderson (1963) demostró que las soluciones $\hat{\beta}$ y $\hat{\upsilon}$ son MELI y MPLI, respectivamente.

Utilizando la inversa de la matriz izquierda de las ecuaciones del modelo mixto, se obtienen los errores estándar de los estimadores para los efectos fijos y aleatorios como sigue:

$$\begin{pmatrix} \mathbf{X}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}$$

Henderson(1975) demostró que la matriz de covarianzas muestral para el MELI de $\boldsymbol{\beta}$ está dado por:

$$\sigma(\hat{\beta}) = \mathbf{C}_{11}$$

La matriz de covarianzas muestrales de los errores de la predicción $\left(\hat{u} - u\right)$ es

$$\sigma\left(\hat{\upsilon} - \upsilon\right) = \mathbf{C}_{22}$$

Y que la covarianza muestral de los efectos estimados y de los errores de predicción está dada por:

$$\sigma\left(\hat{\beta}, \hat{\upsilon} - \upsilon\right) = \mathbf{C}_{12}$$

Se considera $\left(\hat{\upsilon} - \upsilon\right)$ en vez de $\hat{\upsilon}$ ya que ésta incluye tanto la varianza de los errores de predicción como la de los efectos aleatorios de υ en sí mismos. Los errores estándar de los efectos fijos y aleatorios se obtienen respectivamente como la raíz cuadrada de los elementos de la diagonal de $\mathbf{C}_{11}$ y $\mathbf{C}_{22}$.

Es de interés estimar a la media como una combinación lineal de los parámetros de regresión β y de la realización de υ, Henderson (1950) propone el siguiente estimador MELI dado como:

$$\hat{\mu} = l^T \hat{\beta} + m^T \hat{\upsilon}$$

donde l y m son vectores de constantes conocidas $\hat{\beta} = \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{V}^{-1} y$ y $\hat{\upsilon} = \mathbf{G}\mathbf{Z}^T \mathbf{V}^{-1} (y - \mathbf{X}\hat{\beta})$

Ahora, para el caso particular del modelo de covarianza en estructura de bloque diagonal se tiene que

$$y = \left(y_1^T, \ldots, y_m^T\right)^T = y_i, \quad \mathbf{X} = \mathbf{X}_i, \quad \mathbf{Z} = diag_{1 \le i \le m} \mathbf{Z}_i, \quad \upsilon = \upsilon_i, \quad e = e_i,$$

donde m es el número de áreas pequeñas, $\mathbf{X}_i$ es n_i x p, $\mathbf{Z}_i$ es n_i x h_i y y_i es vector n_i x 1, además,

$$\mathbf{R} = diag_{1 \le i \le m}(\mathbf{R}_i), \quad \mathbf{G} = diag_{1 \le i \le m}(\mathbf{G}_i) =, \quad \mathbf{V} = diag_{1 \le i \le m}(\mathbf{V}_i),$$

con

$$\mathbf{V}_i = \mathbf{R}_i + \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T$$

Por lo que el modelo puede ser descompuesto en m submodelos

$$y_i = \mathbf{X}_i \beta + \mathbf{Z}_i \upsilon_i + e_i \qquad (17)$$

El estimador MELI de μ_i se reduce a

$$\hat{\mu}_i = \mathbf{l}_i^T \hat{\beta} + m_i^T \hat{\upsilon}_i \ ,$$

con

$$\hat{\upsilon} = \mathbf{G}_i \mathbf{Z}_i^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \hat{\beta})$$

$$\hat{\beta} = \left(\mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \mathbf{X}_i^T \mathbf{V}_i^{-1} y_i$$

Generalmente los componentes de la varianza son desconocidos en aplicaciones prácticas, por lo que se reemplazan por sus respectivas estimaciones y habrá que utilizar un estimador empírico MELIE.

Henderson (1975) demostró que la sustitución de los valores estimados de los componentes de la varianza en el MELI llevó a predicciones sesgadas. Sin embargo, Kackar y Harville (1981) demostraron que un enfoque en dos fases (estimar primero los componentes de la varianza, y luego utilizar estas para estimar y predecir los parámetros fijos y aleatorios) lleva a predicciones insesgadas, siempre que la distribución del vector de datos sea simétrica respecto al valor esperado y que el valor de los estimadores de los componentes de varianza sean invariantes a transformaciones e incluso sean funciones de los vectores de datos. Ellos demostraron que los estimadores de componentes de la varianza por Máxima Verosimilitud (MV) y Máxima Verosimilitud Restringida (MVR) tienen estas propiedades.

Bajo el supuesto de que la distribución de los componentes aleatorios sea normal, se puede escribir la función de log-verosimilitud del modelo lineal mixto para el vector de observaciones y como:

$$l(\beta,\sigma_e^2,\sigma_v^2 \mid y) = (-(1/2)\left[n\ln\left(2\pi\sigma_e^2\right) + \ln\left|\mathbf{V}\right| + \sigma_e^{-2}\left(y+\mathbf{X}\beta\right)^T \mathbf{V}^{-1}\left(y+\mathbf{X}\beta\right)\right]$$

Diferenciado parcialmente esta función con respecto a β,σ_e^2 y σ_v^2 se obtienen las siguientes funciones score.

$$\frac{\partial l}{\partial \beta} = \sigma_e^{-2}\mathbf{X}\mathbf{V}^{-1}(y-\mathbf{X}\beta)$$

$$\frac{\partial l}{\partial \sigma_e^2} = -(1/2)\left[n\sigma_e^{-2} - \sigma_e^{-4}(y-\mathbf{X}\beta)^T\mathbf{V}^{-1}(y-\mathbf{X}\beta)\right]$$

$$\frac{\partial l}{\partial \sigma_v^2} = -(1/2)\left[tr\left(\mathbf{V}^{-1}Z_i\mathbf{G}_iZ_i^T\right) - \sigma_e^{-2}(y-\mathbf{X}\beta)^T\mathbf{V}^{-1}Z_i\mathbf{G}_iZ_i^T(y-\mathbf{X}\beta)\right] \qquad (18)$$

Igualando estas funciones a cero se tienen las ecuaciones para la estimación de β,σ_e^2 y σ_v^2.

Dadas las estimaciones MV de σ_e^2 y σ_v^2 se obtiene la estimación MV $\hat{\mathbf{V}}$ de $\mathbf{V}$, es claro que la estimación MV de β es justamente su estimación por mínimos cuadrados generalizados con el parámetro $\hat{\mathbf{V}}$, esto es:

$$\hat{\beta}_{MV} = \left(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{V}}^{\mathbf{-1}}\mathbf{X}\right)^{\mathbf{-1}}\mathbf{X}^{\mathbf{T}}\hat{\mathbf{V}}^{\mathbf{-1}}y$$

Sin embargo, las ecuaciones de los componentes de la varianza no tienen una solución analítica, por lo que deben ser resueltos numéricamente.

Un inconveniente del estimador de MV de σ_v^2 es que no toma en cuenta la pérdida de grados de libertad en la estimación de β. El método de MVR toma en cuenta esta pérdida usando una trasformación de datos $y^*=\boldsymbol{A}^T y$ donde $\boldsymbol{A}$ es cualquier matriz ortogonal a la matriz $\boldsymbol{X}$. $y^*=\boldsymbol{A}^T y$ tiene distribución normal (n-p)-variada con media 0 y matriz de covarianza $\mathbf{A}^{\mathrm{T}}\mathbf{V}\mathbf{A}$, por lo que la primera derivada de la función de log-verosimilitud restringida respecto de σ_v^2 es:

$$\frac{\partial l_R}{\partial \sigma_v^2} = -(1/2)\left[tr\left(\mathbf{P}Z_i\mathbf{G}_i Z_i^T\right) - y^T\mathbf{P}Z_i\mathbf{G}_i Z_i^T\mathbf{P}y \right], \qquad (19)$$

donde

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1},$$

Tampoco estas ecuaciones de los componentes de la varianza tienen una solución analítica, por lo que deben ser resueltos numéricamente.

La aproximación del error cuadrático medio del estimador para la media estimada se obtiene primeramente suponiendo que β y los componentes de la varianza son conocidos y que $\hat{\upsilon}$ es el estimador MELI de υ, de donde se tiene que

$$ECM\left(\hat{\mu}_{MELI}\right) = g_1\left(\sigma_v^2\right) = diag\left(\mathbf{G} - \mathbf{G}\mathbf{V}^{-1}\mathbf{G}\right),$$

después se considera el estimador *mcp* $\hat{\beta}$ de β y se suponen conocidos los componentes de la varianza y se obtiene que

$$\begin{aligned} ECM\left(\hat{\mu}_{MELI}\right) &= \left[l^T - m^T\mathbf{G}\mathbf{V}^{-1}\mathbf{X}\right]\left[\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right]^{-1}\left[l^T - m^T\mathbf{G}\mathbf{V}^{-1}\mathbf{X}\right]^T + diag\left(\mathbf{G} - \mathbf{G}\mathbf{V}^{-1}\mathbf{G}\right), \\ &= g_2(\hat{\sigma}_v^2) + g_1(\hat{\sigma}_v^2) \end{aligned}$$

finalmente, se consideran los estimadores de β y de los componentes de la varianza, obteniendo que

$$\begin{aligned} ECM(\hat{\mu}_{MELIE}) &= E\left(\hat{\mu}_{MELIE} - \hat{\mu}_{MELI} + \hat{\mu}_{MELI} - \mu\right)^2 \\ &= ECM(\hat{\mu}_{MELI}) + E\left(\hat{\mu}_{MELIE} - \hat{\mu}_{MELI}\right)^2 + 2E\left(\hat{\mu}_{MELI} - \mu\right)\left(\hat{\mu}_{MELIE} - \hat{\mu}_{MELI}\right), \end{aligned}$$

Kackard y Harville demostraron que el tercer término es despreciable y para el segundo término presentaron una aproximación, Prasad y Rao (1990) demostraron que había una subestimación del *ECM* e introdujeron una aproximación de segundo orden de la forma

$$\begin{aligned} ECM(\hat{\mu}_{MELIE}) &= ECM(\hat{\mu}_{MELI}) + \sigma_e^2 tr\left[\left(\partial b/\partial\sigma_v^2\right)\mathbf{V}\left(\partial b/\partial\sigma_v^2\right)^T \overline{\mathbf{V}}(\hat{\sigma}_v^2)\right] \qquad (20) \\ &= g_1(\sigma_v^2) + g_2(\sigma_v^2) + g_3(\sigma_v^2) \end{aligned}$$

donde $\overline{\mathbf{V}}(\hat{\sigma}_v^2)$ es la aproximación asintótica de la matriz de varianzas y covarianzas estimando los componentes de la varianza por MV ó MVR.

Así de la ecuación (20), g_1 representa la variabilidad de la estimación MELI sobre la media, g_2 representa la variabilidad de la estimación de β y g_3 la variabilidad de υ.

Retomando a los estimadores basados en modelos, estos se pueden clasificar en dos grandes grupos: los que se basan en modelos de área y los que lo hacen en modelos de unidad.

ESTIMADORES BASADOS EN MODELO DE ÁREA

Se les da este nombre debido a que la información auxiliar disponible sólo se relaciona a nivel de área o subpoblación muestreada o no. La información auxiliar (promedios poblacionales por área) está en el vector $X_a = (x_1, ..., x_p)$ y se supone que están relacionadas con la media en la subpoblación de la variable de interés $\overline{Y}_a$, o una cierta función de ésta, $\theta_a = g(\overline{Y}_a)$, a través del modelo lineal con efectos aleatorios υ_a:

$$\theta_a = X_a\beta + b_a\upsilon_a \quad a = 1, ..., m, \qquad (21)$$

donde β es el vector de parámetros de regresión, b_a son constates conocidas y las υ_a se suponen independientes e idénticamente distribuidas (iid) con media 0 y varianza σ_υ^2; generalmente se supone la normalidad de las υ_a. En caso de que no todas las áreas sean seleccionadas en la muestra, se continúa bajo el supuesto de que las áreas muestreadas obedecen al modelo de población.

Por otro lado, sea $\hat{\overline{Y}}_a$ el estimador directo de la media de la variable y en el área pequeña a-ésima, esto supone que la muestra en el área, n_a, es mayor o igual a 1.

Ahora, tenemos que:

$$\hat{\theta}_a = \theta_a + e_a, \qquad (22)$$

donde $\hat{\theta}_a = g(\hat{\overline{Y}}_a)$ y los errores muestrales e_a son independientes con media 0 y varianza ψ_a conocida, se puede relajar este supuesto reemplazando ψ_a por su estimación $\hat{\psi}_a$ calculada de los datos a nivel unidad.

Sustituyendo en el modelo muestral (21) la θ_a del modelo lineal relacional supuesto (22) obtenemos:

$$\hat{\theta}_a = X_a\beta + b_a\upsilon_a + e_a \qquad (23)$$

Se observa que este modelo involucra tanto errores del diseño muestral e_a, como errores del modelo υ_a. y se supone que e_a y υ_a son independientes.

Comparando el modelo (23) con el modelo (17) tenemos que

$$y_i = \hat{\theta}_a, \quad \mathbf{X}_i = X_a, \quad \mathbf{Z}_i = b_a, \quad \beta = (\beta_1, \ldots, \beta_p), \quad \upsilon_i = \upsilon_a, \quad e_i = e_a,$$

además,

$$\mathbf{G}_i = \sigma_\upsilon^2, \quad \mathbf{R}_i = \psi_a,$$

de donde

$$\mathbf{V}_a = \sigma_\upsilon^2 b_a^2 + \psi_a,$$

también

$$\mu_i = \theta_a = X_a^T \beta + b_i^T \upsilon_i \text{ tomando } l_i = X_a \text{ y } m_i = b_i$$

Entonces, el estimador MELIE para θ_a es

$$\tilde{\theta}_a = \hat{\bar{Y}}_{a,MELIE} = X_a \tilde{\beta} + \hat{\gamma}_a (\hat{\theta}_a - X_a \tilde{\beta}) = \hat{\gamma}_a \hat{\theta}_a + (1 - \hat{\gamma}_a) X_a \tilde{\beta}, \ldots. \quad (24)$$

donde $\tilde{\beta}$ es la estimación de β por aproximación en un algoritmo junto con σ_u^2 utilizando mínimos cuadrados ordinarios, y $\hat{\gamma}_a = \hat{\sigma}_\upsilon^2 b_a^2 / (\hat{\psi}_a + \hat{\sigma}_\upsilon^2 b_a^2)$

Se observa que el estimador MELIE de θ_a se expresa como un promedio ponderado del estimador directo $\hat{\theta}_a$ y el estimador de regresión sintética $X_a \tilde{\beta}$, donde el ponderador $\hat{\gamma}_a$ $(0 \le \hat{\gamma}_a \le 1)$ mide la incertidumbre en la modelización del predictor para cada área pequeña a.

El estimador $\tilde{\theta}_a$ es insesgado si los errores u_a y e_a están simétricamente distribuidos alrededor de 0, en particular $\tilde{\theta}_a$ es insegado si υ_a y e_a son normalmente distribuidas.

Estimación de σ_υ^2

Partiendo de la diferenciación de la función de máxima verosimilitud respecto a σ_υ^2 (ecuación 18) tenemos que

$$s\left(\tilde{\beta}, \sigma_\upsilon^2\right) = \frac{\partial l}{\partial \sigma_\upsilon^2} = -(1/2)\left[tr\left(\mathbf{V}^{-1} Z_i \mathbf{G}_i Z_i^T\right) - \psi_i^{-1} (y - \mathbf{X}\beta)^T \mathbf{V}^{-1} Z_i \mathbf{G}_i Z_i^T (y - \mathbf{X}\beta) \right]$$

$$= -\frac{1}{2} \sum_{i=1}^{m} \frac{b_i^2}{\sigma_\upsilon^2 b_i^2 + \psi_i} + \frac{1}{2} \sum_{i=1}^{m} b_i^2 \frac{(\hat{\theta} - X_a \tilde{\beta})^2}{(\sigma_\upsilon^2 b_i^2 + \psi_i)^2} \quad (25)$$

Ahora, sea $\mathcal{I}\left(\sigma_\upsilon^2\right)$ la segunda derivada de la función $l(\beta, \psi, \sigma_\upsilon^2 \mid y)$ respecto a σ_υ^2, se tiene

$$\mathcal{I}\left(\sigma_v^2\right)=\frac{1}{2}tr\left(\mathbf{V}^{-1}Z_i\mathbf{G}_iZ_i^T\mathbf{V}^{-1}Z_k\mathbf{G}_kZ_k^T\right)=\frac{1}{2}\sum_{i=1}^{m}\frac{b_i^4}{(\sigma_v^2b_i^2+\psi_i)^2} \quad (26)$$

De aquí que para estimar σ_v^2 por máxima verosimilitud, utilizando el algoritmo de Fisher scoring se tiene que:

$$\sigma_v^{2(k+1)}=\sigma_v^{2(k)}+\left[\mathcal{I}\left(\sigma_v^{2(k)}\right)\right]^{-1}s\left(\tilde{\beta}^{(k)},\sigma_v^{2(k)}\right), \quad (27)$$

De igual forma, tomando la diferenciación de la función de máxima verosimilitud restringida respecto a σ_v^2 (ecuación 19) con $Z_i\mathbf{G}_iZ_i^T=diag(b_1^2,....,b_m^2)$ con b_i = 1, se obtiene

$$s_R\left(\sigma_v^2\right)=\frac{\partial l_R}{\partial\sigma_v^2}=-(1/2)\left[tr\left(\mathbf{P}Z_i\mathbf{G}_iZ_i^T\right)-y^T\mathbf{P}Z_i\mathbf{G}_iZ_i^T\mathbf{P}y\right]=-\frac{1}{2}tr\left(\mathbf{P}\right)+\frac{1}{2}\hat{\theta}^T\mathbf{P}\mathbf{P}\hat{\theta}, \quad (28)$$

Ahora, sea $\mathcal{I}_R\left(\sigma_v^2\right)$ la segunda derivada de la función $l_R(\psi,\sigma_v^2\mid y)$ respecto a σ_v^2 se tiene

$$\mathcal{I}_R\left(\sigma_v^2\right)=\frac{1}{2}tr\left[\mathbf{P}\mathbf{P}\right]. \quad (29)$$

De aquí que para estimar σ_v^2 por máxima verosimilitud restringida, utilizando el algoritmo de Fisher scoring, resulta la siguiente expresión:

$$\sigma_v^{2(k+1)}=\sigma_v^{2(k)}+\left[\mathcal{I}_R\left(\sigma_v^{2(k)}\right)\right]^{-1}s_R\left(\sigma_v^{2(k)}\right), \quad (30)$$

Asintóticamente $\mathrm{I}\left(\sigma_v^2\right)/\mathrm{I}_R\left(\sigma_v^2\right)\to 1$ cuando $A\to\infty$.

Estimación del error cuadrático medio ECM

La aproximación del ECM por máxima verosimilitud y máxima verosimilitud restringida se justifica observando las siguientes condiciones de regularidad (Datta y Lahiri 2000):

i) Los elementos de X_i y Z_i son uniformemente acotados
ii) $Sup(n_i)<\infty$ y $Sup(h_i)<\infty$ con i≥1
iii) Las matrices de covarianzas $\mathbf{G}_i$ y $\mathbf{R}_i$ son definidas positivas y sus elementos son uniformemente acotados.

La aproximación por máxima verosimilitud del error cuadrático medio del estimador MELIE de la media se expresa como

$$ECM(\hat{\bar{Y}}_{a,MELIE})\approx g_{1a}(\hat{\sigma}_v^2)-b_{\sigma_u^2}(\hat{\sigma}_v^2)\nabla g_{1a}(\hat{\sigma}_v^2)+g_{2a}(\hat{\sigma}_v^2)+2g_{3a}(\hat{\sigma}_v^2), \quad (31)$$

donde

$$g_{1a}(\hat{\sigma}_v^2) = diag\left(\mathbf{G}_a - \mathbf{G}_a\mathbf{V}_a^{-1}\mathbf{G}_a\right) = \hat{\gamma}_a\hat{\psi}_a$$
$$g_{2a}(\hat{\sigma}_v^2) = \left[X_a^T - m_a^T\mathbf{G}_a\mathbf{V}_a^{-1}\mathbf{X}_a\right]\left[\mathbf{X}_a^T\mathbf{V}_a^{-1}\mathbf{X}_a\right]^{-1}\left[X_a^T - m_a^T\mathbf{G}_a\mathbf{V}_a^{-1}\mathbf{X}_a\right]^T$$
$$g_{3a}(\hat{\sigma}_v^2) = \hat{\psi}_a b_a^4(\hat{\psi}_a + \hat{\sigma}_v^2 b_a^2)^{-3}\bar{V}(\hat{\sigma}_v^2) \qquad (32)$$

$$\nabla g_{1a}(\hat{\sigma}_v^2) = (1-\hat{\gamma})^2$$

$$b_{\sigma_v^2}(\hat{\sigma}_v^2) = -\left[2I(\hat{\sigma}_v^2)\right]^{-1} tr\left[\left\{\mathbf{X}_a^T\mathbf{V}_a^{-1}\mathbf{X}_a\right\}^{-1}\left\{\mathbf{X}_a^T\left(\mathbf{V}_a^{-1}\right)^2\mathbf{X}_a\right\}\right]$$

$b_{\sigma_u^2}$ es el sesgo sobre la estimación MELIE de la media y $\bar{V}$ es la varianza asintótica del estimador $\hat{\sigma}_v^2$ de σ_v^2 estimada por máxima verosimilitud ó máxima verosimilitud restringida y está dada por:

$$\bar{V}(\hat{\sigma}_v^2) = 2\left[\sum_{i=1}^{a} 1/(\sigma_v^2 + \hat{\psi}_a)^2\right]^{-1}$$

Ahora, la estimación del *ECM* por máxima verosimilitud restringida está dada por

$$ECM(\hat{\bar{Y}}_{a,MELIE}) \approx g_{1a}(\hat{\sigma}_v^2) + g_{2a}(\hat{\sigma}_v^2) + 2g_{3a}(\hat{\sigma}_v^2) \qquad (33)$$

Estas estimaciones están basadas en el supuesto de que el número de áreas es grande y de que tanto ψ_a como b_a son uniformemente acotadas.

El *ECM* estimado para las áreas no muestreadas se calcula mediante la siguiente expresión:

$$\hat{ECM}(\hat{\mu}_{.a}) = \hat{\sigma}_v^2 + \bar{\mathbf{X}}_{.a}^T\hat{\mathbf{V}}_{\tilde{\beta}}\bar{\mathbf{X}}_{.a} \qquad (34)$$

donde el punto "." antes del subíndice a se refiere a las áreas no muestreadas, y $\hat{\mathbf{V}}_{\tilde{\beta}}$ es la matriz de varianzas y covarianzas de los coeficientes (betas) estimadas con el algoritmo de scoring junto con σ_v^2.

El intervalo de confianza al 95% está dado por:

$$IC_{a,MELIE} = \hat{\bar{Y}}_{a,MELIE} \pm 2 * \sqrt{\hat{ECM}(\hat{\bar{Y}}_{a,MELIE})}$$

ESTIMADORES BASADOS EN MODELO DE UNIDAD

En este caso, los valores de las unidades poblacionales y_{ia} referida a la unidad i del área a están relacionadas con las variables auxiliares x_{ia}, a través del modelo lineal con efectos mixtos y estructura de covarianza diagonal

$$y_{ai} = x_{ai}^T \beta + \upsilon_a + e_{ai}, \text{ con } i = 1,\ldots,n_a \;\; a = 1,\ldots,m$$

Donde generalmente se supone normalidad para la variables aleatorias υ_a y e_{ai}.

El estimador MELI para la media (para diferenciarlo del modelo de área se añade la letra B al acrónimo MELI quedando como MELIB y su empírico MELIE como MELIEB) es

$$\hat{\bar{y}}_{a,MELIB} = \gamma_a \left[\bar{y}_{ac} + (\bar{X}_a - \bar{x}_{ac})^T \tilde{\beta} \right] + (1-\gamma_a) \bar{X}_a^T \tilde{\beta},$$

donde

$\bar{y}_{ac} = \sum_i c_{ia} y_{ia} / c_a$ y $\bar{x}_{ac} = \sum_i c_{ia} x_{ia} / c_a$ con $c_a = \sum_i c_{ia}$; $C_{ia} = k_{ia}^{-2}$; k_{ia}^2 elementos de la diagonal de la matriz del modelo heteroscedástico de la varianza de los residuales e_{ai} .

$$\gamma_a = \sigma_\upsilon^2 / (\sigma_\upsilon^2 + \sigma_e^2 / c_a)$$

Cuando el modelo es homocedástico y c_{ia} = 1, entonces $c_a = n_a$

Existen diferentes procedimientos para estimar los componentes de varianza σ_υ^2 y σ_e^2 con resultados muy similares, en este trabajo sólo se presenta el método de ajuste de constantes para obtención de las estimaciones.

$\hat{\sigma}_e^2 = \dfrac{1}{n-t-p+1} \sum_{a=1}^{t} \sum_{j=1}^{n_a} \hat{\varepsilon}_{aj}^2$ donde $\hat{\varepsilon}_{aj}$ son los residuos de la regresión lineal ordinaria de $y_{aj} - \bar{y}_a$ sobre $x_{aj} - \bar{x}_a$.

$\hat{\sigma}_\upsilon^2 = \max\left(\dfrac{1}{n^*} \left(\sum_{a=1}^{t} \sum_{j=1}^{n_a} \hat{\upsilon}_{aj}^2 - (n-p)\hat{\sigma}_e^2 \right), 0 \right)$ donde $\hat{\upsilon}_{aj}$ son los residuos de la regresión lineal ordinaria de y_{aj} sobre x_{aj} y la variable z_{aj} que toma el valor 1 para el área pequeña a y 0 para el resto de las áreas y n* es la traza de la matriz $\mathbf{MZZ^T}$, donde $\mathbf{M}$ es la matriz de proyección idempotente $\mathbf{(I-X(X^TX)^{-1}X^T)}$.

La estimación de segundo orden del error cuadrático medio de la estimación de la media está dado como

$$\hat{ECM}(\hat{\bar{y}}_{a,MELIEB}) \approx g_{1a}(\hat{\sigma}) + g_{2a}(\hat{\sigma}) + 2g_{3a}(\hat{\sigma})$$

donde

$$g_{1a}(\hat{\sigma}) = \hat{\gamma}_a(\sigma_e^2 / c_a)$$

$$g_{2a}(\hat{\sigma}) = (\bar{X}_a - \hat{\gamma}_a \bar{x}_{ac})^T \left[\sum_a A_a\right]^{-1} (\bar{X}_a - \hat{\gamma}_a \bar{x}_{ac})$$

$$g_{3a}(\hat{\sigma}) = c_a^{-2}(\sigma_v^2 + \sigma_e^2 / c_a)^{-3}\left(\hat{\sigma}_e^4 \bar{V}_{vv} + \hat{\sigma}_v^4 \bar{V}_{ee} - 2\hat{\sigma}_v^2 \hat{\sigma}_e^2 \bar{V}_{ve}\right)$$

$$A_a = \sigma_e^{-2}\left(\sum_i c_{ai} x_{ai} x_{ai}^T - \gamma_a \bar{x}_{ai} \bar{x}_{ai}^T\right)$$

La varianza y covarianza asintótica del estimador por el método de ajuste de constantes están dadas por

$$\bar{V}_{vv} = 2\eta_1^{-2}\left[\nu_1^{-1}(n-p-\nu_1)(n-p)\sigma_e^4 + \eta_2 \sigma_v^4 + 2\eta_1 \sigma_e^2 \sigma_v^2\right]$$

$$\bar{V}_{ee} = 2\nu_1^{-1}\sigma_e^4$$

$$\bar{V}_{ve} = -2\eta_1^{-1}\nu_1^{-1}(n-p-\nu_1)\sigma_e^4$$

donde

$$\eta_1 = \sum_a c_a \left[1 - c_a \bar{x}_{ac}^T \left(\sum_a \sum_i c_{ai} x_{ai} x_{ai}^T\right)^{-1} \bar{x}_{ac}\right]$$

$$\eta_2 = \sum_a c_a^2 (1 - c_a \bar{x}_{ac}^T \mathbf{A}_1^{-1} \bar{x}_{ac}) + tr\left(\mathbf{A}_1^{-1} \sum_a c_a^2 \bar{x}_{ac} \bar{x}_{ac}^T\right)^2$$

$$\mathbf{A}_1 = \sum_a \sum_i c_{ai} x_{ai} x_{ai}^T$$

Cuando la fracción de muestreo $f_a = n_a / N_a$ no es despreciable, la literatura recomienda utilizar la versión predictiva para obtener la predicción del área pequeña a en lugar de la versión proyectiva, consiste en diferenciar la parte muestreada de la no muestreada. Así, la predicción de la parte muestreada es la misma muestra, mientras que la no muestreada se predice con el predictor de tipo proyectivo.

Se descompone el total $\sum_{j \in N_a} y_{aj} = \sum_{j \in a_r} y_{aj} + \sum_{j \in a_s} y_{aj}$, donde a_s es la muestra en el área pequeña a y a_r el resto de las unidades no pertenecientes a la muestra del área pequeña a. El estimador MELI de la media es:

$$\hat{\bar{Y}}_{a,MELIB} = f_a \bar{y}_a + (1 - f_a)\hat{\bar{y}}_a^* \text{ con}$$

$$\hat{\bar{y}}_a^* = \gamma_a\left[\bar{y}_{ac} + (\bar{x}_a^* - \bar{x}_{ac})^T \tilde{\beta}\right] + (1 - \gamma_a)\bar{x}_a^{*T} \tilde{\beta}$$

$$\bar{x}_a^* = (N_a \bar{X}_a - n_a \bar{x}_a) / (N_a - n_a)$$

El estimador de error cuadrático medio esta dado por

$$\hat{ECM}(\hat{\bar{Y}}_{a,MELIEB}) = (1-f_a)^2 \hat{ECM}(\hat{\bar{y}}_a^*) + (N_a^{-2} k_a^{*T} k_a^*)\hat{\sigma}_e^2 \quad \text{con}$$

$$ECM(\hat{\bar{y}}_a^*) \approx g_{1a}(\hat{\sigma}) + g_{2a}(\hat{\sigma}) + 2g_{3a}(\hat{\sigma})$$

El estimador MELIE del modelo a nivel unidad no considera los pesos del diseño w_{ij} asociados a los elementos seleccionados, como resultado, no es de diseño consistente a menos que el diseño de muestreo, sea auto-ponderado dentro de las áreas, es decir, $w_{ij}=w_j$ para toda j. Por otro lado, el estimador MELIE del modelo a nivel área es consistente con el diseño.

Modelo con pesos de muestreo

Rao 2003 desarrolla el caso especial de estimadores pseudo-MELIE que dependen del diseño de muestreo en donde las varianzas de los errores son iguales, es decir, $k_{ij} = 1$ para toda i,j .

El estimador pseuddo-MELIE ponderado para la media es

$$\hat{\bar{y}}_{aw} = \bar{X}_a^T \hat{\beta}_w + \hat{\gamma}_{aw}(\bar{y}_{aw} - \bar{x}_{aw}^T \hat{\beta}_w)$$

donde

$$\hat{\beta}_w = \left[\sum_a \sum_i w_{ai} x_{ai}(x_{ai} - \gamma_{aw}\bar{x}_{aw})^T\right]^{-1} \times \left[\sum_a \sum_i w_{ai}(x_{ai} - \gamma_{aw}\bar{x}_{aw}) y_{ai}\right]$$

$$\hat{\gamma}_{aw} = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \delta_a \hat{\sigma}_e^2); \quad \bar{y}_{aw} = \bar{x}_{aw}^T \beta + \upsilon_a + \bar{e}_{aw}; \quad \bar{e}_{aw} = \sum_j \tilde{w}_{aj} e_{aj}; \quad \tilde{w}_{aj} = w_{aj} / w_a$$

$$\bar{x}_{aw} = \sum_i \tilde{w}_{ai} x_{ai}$$

con

$$E(\bar{e}_{aw}) = 0 \text{ y } V(\bar{e}_{aw}) = \sigma_e^2 \sum_i w_{ai}^2 = \sigma_e^2 \delta_{aw}$$

Otra forma del estimador pseudo MELIE en términos de los estimadores directos es

$$\sum_a N_a \hat{\bar{y}}_{aw} = \hat{Y}_w + (X - \hat{X}_w)^T \hat{\beta}_w$$

donde

$$\hat{Y}_w = \sum_a w_a \bar{y}_{aw} = \sum_a \sum_i w_{ai} y_{ai}; \quad \hat{X}_w = \sum_a w_a \bar{x}_{aw} = \sum_a \sum_i w_{ai} x_{ai}$$

Bajo condiciones de normalidad de los errores υ_i y e_{ij}, un estimador del error cuadrático medio es

$$\hat{ECM}(\hat{\bar{y}}_{aw}) \approx g_{1aw}(\hat{\sigma}) + g_{2aw}(\hat{\sigma}) + 2g_{3aw}(\hat{\sigma})$$

donde

$$g_{1aw}(\hat{\sigma}) = \hat{\gamma}_{aw}\delta_{aw}\hat{\sigma}_e^2$$
$$g_{2aw}(\hat{\sigma}) = (\overline{X}_a - \hat{\gamma}_{aw}\overline{x}_{aw})^T \Phi_w (\overline{X}_a - \hat{\gamma}_{aw}\overline{x}_{aw})$$
$$g_{3a}(\hat{\sigma}) = \hat{\gamma}_{aw}(1-\hat{\gamma}_{aw})^2 \hat{\sigma}_v^{-2}\hat{\sigma}_e^{-2}\left(\hat{\sigma}_e^4\overline{V}_{vv} + \hat{\sigma}_v^4\overline{V}_{ee} - 2\hat{\sigma}_v^2\hat{\sigma}_e^2\overline{V}_{ve}\right)$$

Además Φ_w es la matriz de covarianzas de $\tilde{\beta}$ y está dada por

$$\Phi_w = \left(\sum_a\sum_i x_{ai}z_{ai}^T\right)^{-1}\left(\sum_a\sum_i z_{ai}z_{ai}^T\right)\left[\left(\sum_a\sum_i x_{ai}z_{ai}^T\right)^{-1}\right]^T\sigma_e^2$$
$$+\left(\sum_a\sum_i x_{ai}z_{ai}^T\right)^{-1}\left[\sum_a\left(\sum_i z_{ai}\right)\left(\sum_i z_{ai}\right)^T\right]\times\left[\left(\sum_a\sum_i x_{ai}z_{ai}^T\right)^{-1}\right]^T\sigma_v^2$$

donde

$$z_{ai} = w_{ai}(x_{ai} - \gamma_{aw}\overline{x}_{aw}).$$

Este estimador es válido para estimaciones de σ_u^2 y σ_e^2 por el método de ajuste de constantes y por máxima verosimilitud restringida.

CAPÍTULO III. APLICACIÓN EN LA ESTIMACIÓN DEL INGRESO PROMEDIO A NIVEL MUNICIPAL

3.1 CONFORMACIÓN DE LOS ARCHIVOS PARA LAS ESTIMACIONES

Las variables captadas por la ENIGH fueron identificadas y homologadas con las captadas en el II Conteo, a partir de ellas, se elaboraron nuevas variables como promedios, máximos, mínimos, proporciones, así como categorizar algunas otras, con el propósito de tener una diversidad de variables que por experiencia e intuición pueden estar relacionadas con el ingreso de la vivienda y, por otro lado, tratar de simplificar la interpretación de los resultados; la categorización se hizo de acuerdo a la distribución de cada variable. Se realizó una depuración de las viviendas tanto de las de la ENIGH como de las del II Conteo, en donde el valor de las variables homologadas era no especificado.

Para el caso de los modelos con efectos aleatorios se incluyeron variables con valores resumidos a nivel municipal, así como variables geográficas que mantuvieran relación con el ingreso y a la vez diferenciaran el comportamiento entre los conglomerados (municipios).

De esta manera, se armaron archivos que contienen 75 variables elaboradas a partir de las variables comunes entre la ENIGH y el II Conteo a nivel registro (ver Anexo 1), 15 variables a nivel municipal y 16 variables geográficas, con el objeto de poder seleccionar las que mejor contribuyan a los modelos a estudiar en este trabajo.

Con la intención de simplificar la carga computacional y la extensión de este trabajo se eligió realizar las estimaciones únicamente para el estado de Sonora, ya que en la encuesta hubo una sobremuestra en el año 2005 con el objeto de proporcionar resultados para algunas entidades federativas, además, su número de municipios (72 en total y 23 en muestra) lo hacen un candidato más o menos adecuado para la construcción de los modelos. Después de la depuración citada, quedaron 1,817 de 1,836 viviendas de la ENIGH y 588,436 de 615,002 del II Conteo para calcular los promedios municipales.

AFIJACIÓN DE LA MUESTRA A NIVEL MUNICIPAL

Con el objeto de asegurar que la expansión de la muestra ajuste al número de viviendas contabilizadas por el II Conteo par cada municipio en muestra, se realizó una corrección al factor de expansión mediante la siguiente expresión:

$$w'_{ai} = w_{ai}\frac{N_{vaconteo}}{\hat{N}_{va}} \quad \text{con} \quad \hat{N}_{va} = \sum_{i=1}^{n_a} w_i \,,$$

donde

w'_{ai} es el factor de expansión ajustado al II Conteo en el municipio a para la vivienda i.
w_{ai} es el factor de expansión en el municipio a para la vivienda i.
$N_{vaconteo}$ es el número de viviendas en el municipio a, según el II Conteo.
N_{va} es el número de viviendas estimada por la encuesta en el municipio a.

3.2 ESTIMACIÓN DIRECTA

Utilizando las expresiones (3), (4) y (5) del Capítulo II para el cálculo del promedio del ingreso total por municipio y su correspondiente error cuadrático medio (ecm), así como su coeficiente de

variación, se obtuvieron las respectivas estimaciones para la entidad de Sonora y a cada uno de los municipios en muestra (ver Tabla 1). Se empleó el software R y las funciones de la librería *"survey"* (ver Anexo 5).

Tabla 1. Municipios en muestra según su estimación Directa

Clave de municipio	Nombre	Tamaño de la muestra	Valores con estimación Directa		
			Ingreso promedio	$\sqrt{\hat{ECM}}$	Coeficiente de variación
	Sonora	1817	39,008	1,605.4	0.072
2	Agua Prieta	38	31,594	3,826.0	0.121
3	Alamos	18	18,928	3,637.6	0.192
12	Bácum	18	20,201	4,803.1	0.238
13	Banámichi	15	30,419	5,773.0	0.190
17	Caborca	19	12,518	4,253.9	0.340
18	Cajeme	233	34,979	4,136.9	0.118
19	Cananea	18	28,288	5,659.3	0.200
25	Empalme	34	23,934	2,938.0	0.123
26	Etchojoa	20	35,005	21,834.5	0.624
29	Guaymas	104	26,105	2,491.7	0.095
30	Hermosillo	737	52,316	3,890.4	0.074
33	Huatabampo	54	13,033	3,547.3	0.272
36	Magdalena	18	23,647	3,978.4	0.168
41	Nacozari de García	18	22,600	4,114.2	0.182
42	Navojoa	113	19,926	4,227.3	0.212
43	Nogales	113	29,402	2,997.1	0.102
48	Puerto Peñasco	36	70,179	9,910.3	0.141
53	San Felipe de Jesús	19	24,294	6,618.4	0.272
55	San Luis Río Colorado	89	34,442	3,640.4	0.106
58	Santa Ana	17	34,832	6,193.6	0.178
66	Ures	14	28,730	3,918.4	0.136
71	Benito Juárez	39	20,014	4,557.6	0.228
72	San Ignacio Río Muerto	32	12,054	2,897.1	0.240

En algunos municipios solamente hay una UPM muestreada por lo que estrictamente no sería posible calcular su ECM, sin embargo, para fines comparativos se calculó suponiendo un muestro aleatorio simple con reemplazo dentro del municipio, este supuesto induce una subestimación del ECM directo real. Una UPM está contenida totalmente en un municipio muestreado.

3.3 Estimación GREG

Los modelos tipo GREG consideran los pesos de muestreo y para realizar los análisis y estimaciones se utilizaron nuevamente las funciones de la librería "survey" de R (ver Anexo 5), mismas que consideran el diseño de muestreo, sin embargo, es pertinente mencionar que la

selección de las unidades muestrales se realizaron mediante un muestreo sistemático, con este tipo de muestreo las probabilidades de segundo orden no están bien definidas por la dependencia que guardan la selección del primer elemento con los subsiguientes elementos. Debido a que no hay o no se encontró en la literatura ni en el software estas consideraciones, se continuará con el supuesto de un *mas* en la elección de las viviendas dentro de una UPM en lugar del muestreo sistemático.

El modelo propuesto para analizar el ingreso promedio de las viviendas es:

$$\log(y_{ai}) = X_{ai}\beta + e_{ai}$$

La transformación logarítmica se basa en que $\log(y)$ normaliza el comportamiento de la distribución de y, además de la conveniencia de utilizar una escala mejor relacionada con la escala de las variables auxiliares o independiaentes.

Selección de variables

Para tener una idea de la relación de las variables independientes respecto a la variable dependiente, se obtuvo la correlación de cada una con respecto al log-ingreso total de la vivienda. A continuación se presentan las variables ordenadas por su correlación de mayor a menor.

numcua	**escmax**	**compu**	**escpromh**	**peprom5n**	**pemedn**	**escmenj**	**escpronj**
0.30	0.29	0.29	0.28	0.27	0.27	0.27	0.27
anosescj	**cuadorc**	**cuador**	**nivedj**	**matpiso3**	**numcuac**	**matpiso**	**esccatj**
0.27	0.26	0.25	0.25	0.23	0.23	0.23	0.18
matpiso2	**banoagua1**	**banoagua**	**lava**	**pp15pinc**	**banoagua3**	**rezinegi**	**trabdomi**
−0.17	0.16	−0.16	0.16	−0.14	−0.13	−0.13	0.12
rezaeduj	**hacinai**	**hacina**	**pp15pri**	**matpiso1**	**rezahaci**	**refri**	**agua**
−0.12	−0.12	−0.12	−0.11	−0.10	−0.10	0.10	−0.10
taseschc0	**taseschc**	**panalf15**	**nas1524c**	**nas1524**	**e25a49c**	**banoagua2**	**pmuj12**
−0.09	0.09	−0.08	0.08	0.08	0.08	−0.07	0.07
muj12c	**imigeu**	**tele**	**analfc**	**e50a64c0**	**tasesch**	**phimue12**	**muj12**
0.07	0.07	0.07	−0.07	−0.06	0.06	−0.06	0.06
hombresc	**hombres**	**sinbien**	**agua2**	**migeuj**	**pe65mas**	**e50a64c**	**e50a64**
0.06	0.06	−0.06	−0.06	0.06	−0.06	0.06	0.06
banoagua4	**hombresc0**	**e50a64c2**	**taseschc3**	**taseschc2**	**taseschc1**	**phog**	**phinv12**
−0.05	−0.05	0.05	0.05	0.05	0.05	0.05	−0.05
imigra	**hijosc**	**electri**	**pe25a49**	**e50a64c1**	**hijosmc**	**edadmin**	**e65mas**
0.05	0.05	0.05	0.05	0.04	0.04	−0.04	−0.04
hombresc3	**hombresc1**	**phijo12**	**ninosc**	**hinv15**	**hijo12**	**hijasc**	**pe50a64**
0.03	−0.03	−0.03	−0.03	−0.03	−0.03	0.03	0.03
e15a24c	**e15a24**	**nase614c1**	**hombresc4**	**factorM**	**edadprom**	**agro**	**asisescj**
0.03	0.03	0.02	0.02	−0.02	−0.02	−0.02	0.02
nase614c2	**nase614c0**	**hombresc5**	**hombresc2**	**taseschc4**	**numhog**	**nase614**	**edadmax**
−0.01	−0.01	0.01	0.01	0.01	−0.01	0.01	−0.01
edadj2	**edadj**	**pe15a24**	**nase614c**	**e6a14c**			
−0.01	0.01	0.01	0.00	0.00			

Se ejecutaron los modelos de regresión mediante la función *svyglm* de *R* considerando el diseño de muestreo, se incluyeron una a una las variables comenzando por la de mayor correlación y así hasta la que presentó menor correlación, se compararon los modelos mediante un análisis de deviance que es una generalización del análisis de varianza, la deviance D_p (p es el número de

parámetros en el modelo) se puede usar como una medida de discrepancia y la diferencia de deviance entre modelos $D_{p_i} - D_{p_j}, p_j > p_i$ es interpretada como una medida de la variación de los datos explicadas por el modelo M_{pj} y que no están en M_{pi}, cuando esta diferencia es > $\chi^2_{p_j - p_i, \alpha}$ indica que los efectos de los términos que están en M_{pj} y no están en M_{pi} son significativos. Tomando este criterio, observamos la columna "`P(>|Chi|)`" de la "Tabla para la selección de variables del modelo GREG" presentada en el Anexo 2, que contiene el resultado de la comparación de los 76 modelos, en donde se eliminan aquellos con valor mayor a 0.95. En total son 55 variables cuyo aporte no es explicativo para la variable dependiente.

A los modelos restantes se vuelve aplicar el análisis de deviance para eliminar aquellos que no son explicativos, también se eliminan aquellas variables que presentan alta correlación (>0.7) y su deviance haya sido el menor, además se eliminan aquellas cuyo coeficiente β en el modelo no es significativo, de esta forma, se llega a un modelo que incluye a 19 variables, el resultado al aplicar la función svyglm de R es el siguiente:

```
Call:
svyglm(log(sum_ingtot) ~ escmax + numcua + compu + matpiso1 +
    matpiso2 + banoagua2 + banoagua3 + lava + trabdomi + pp15pri +
    refri + taseschc0 + imigeu + hombresc0 + hombresc1 + hombresc2 +
    hombresc3 + e50a64c0 + e50a64c1, design = dstrat, family = gaussian(link = identity))

Survey design:
svydesign(id = ~upm, strata = ~est, weights = ~factorM, data = datos,
    nest = TRUE)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.041615   0.131368  68.827  < 2e-16 ***
escmax       0.068165   0.005274  12.926  < 2e-16 ***
numcua       0.085092   0.012575   6.767 9.99e-11 ***
compu        0.252779   0.043552   5.804 2.04e-08 ***
matpiso1    -0.182887   0.084398  -2.167 0.031222 *
matpiso2    -0.207726   0.043051  -4.825 2.49e-06 ***
banoagua2   -0.157522   0.060431  -2.607 0.009716 **
banoagua3   -0.181782   0.050489  -3.600 0.000386 ***
lava         0.142316   0.042402   3.356 0.000918 ***
trabdomi     0.896530   0.113409   7.905 9.70e-14 ***
pp15pri     -0.151735   0.064003  -2.371 0.018544 *
refri        0.234590   0.057197   4.101 5.63e-05 ***
taseschc0    0.116104   0.050828   2.284 0.023232 *
imigeu       0.662327   0.142213   4.657 5.31e-06 ***
hombresc0   -0.447456   0.064673  -6.919 4.12e-11 ***
hombresc1   -0.259018   0.063363  -4.088 5.94e-05 ***
hombresc2   -0.239056   0.054513  -4.385 1.73e-05 ***
hombresc3   -0.137878   0.069922  -1.972 0.049770 *
e50a64c0    -0.134949   0.049345  -2.735 0.006707 **
e50a64c1    -0.103459   0.048878  -2.117 0.035318 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.3216295)
Number of Fisher Scoring iterations: 2
```

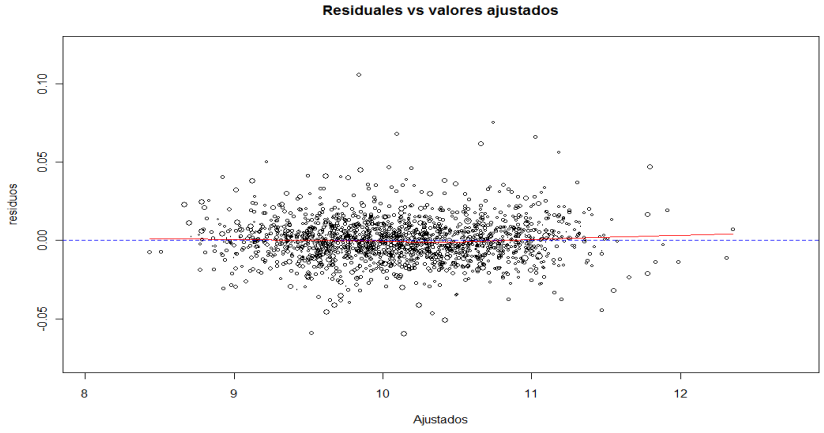
En donde se puede observar que todos los coeficientes son significativos, además, debido a que los valores de las variables son positivas, en el modelo se puede interpretar que la media tiene un valor aproximado de 9.04, que las variables *escmax* (escolaridad máxima en la vivienda), *numcua (*número de cuartos en la vivienda), *compu* (existencia de computadora en la vivienda), *lava* (existencia de lavadora en la vivienda), *refri* (existencia de refrigerador en la vivienda), *taseschc0* (ningún habitante de la vivienda, mayor de 15 años asiste a la escuela), *trabdomi* (existencia de sirvientes en la vivienda), *imigeu* (existencia de habitantes que emigran a los Estados Unidos) contribuyen incrementando el ingreso, por el contrario, las variables *matpiso1* (existencia de piso de tierra en la vivienda), *matpiso2* (existencia de piso de cemento en la vivienda), *banoagua2 (le echan agua con cubeta al baño)*, *banoagua3* (no le pueden echar agua al baño), *pp15pri* (proporción de personas de más de 15 años que sólo completó la primaria), *hombresc0* (ningún habitante masculino en la vivienda), *hombresc1* (existencia de un habitante masculino en la vivienda), *hombresc2* (existencia de dos habitantes masculinos en la vivienda), *hombresc3* (existencia de tres habitantes masculinos en la vivienda) y *e50a64c01* (ningún ó un habitante entre 50 y 64 años en la vivienda) contribuyen disminuyendo el ingreso.

La matriz de correlaciones de las variables seleccionadas se presenta en el Anexo 4 como "Matriz de correlaciones para el modelo GREG", en el Anexo 3 se presenta la "Tabla de colinealidad de las variables del modelo GREG", en donde se observa que las variables seleccionadas guardan una baja correlación y no presentan problemas de colinealidad, para esto último se analiza el resultado de la aplicación de la función colldiag de R que utiliza el método de descomposición de proporciones, en donde se debe observar que el índice de condición no se incremente desproporcionalmente de una componente a otra o que para las misma componente las proporciones de variables diferentes sean mayores a 0.5, situaciones que por lo general se presentan simultáneamente (Rawlings et.al. 1998).

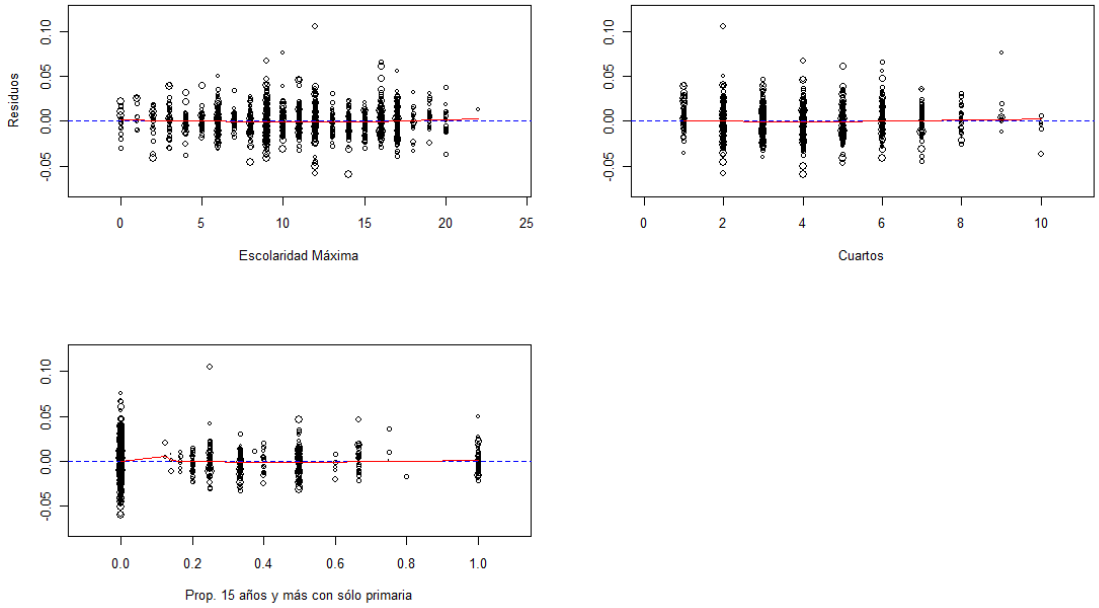
Análisis de residuales

En la Gráfica 1 podemos observar que el comportamiento de los residuales respecto de los valores ajustados no presenta ninguna tendencia y su dispersión acotada entre 0.05 y -0.05 indica homocedasticidad de los residuales. En las gráficas 2 a 5 se presentan los residuales contra las variables cuantitativas del modelo, en donde se observa que su comportamiento es lineal y sin tendencia clara, por lo que para estas variables no es necesaria ninguna transformación. En estas gráficas se considera el peso del diseño muestral, esto es, el tamaño del círculo representa el peso muestral dado a cada vivienda muestreada.

Gráfica 1. Residuos del modelo GREG ponderados contra valores ajustados

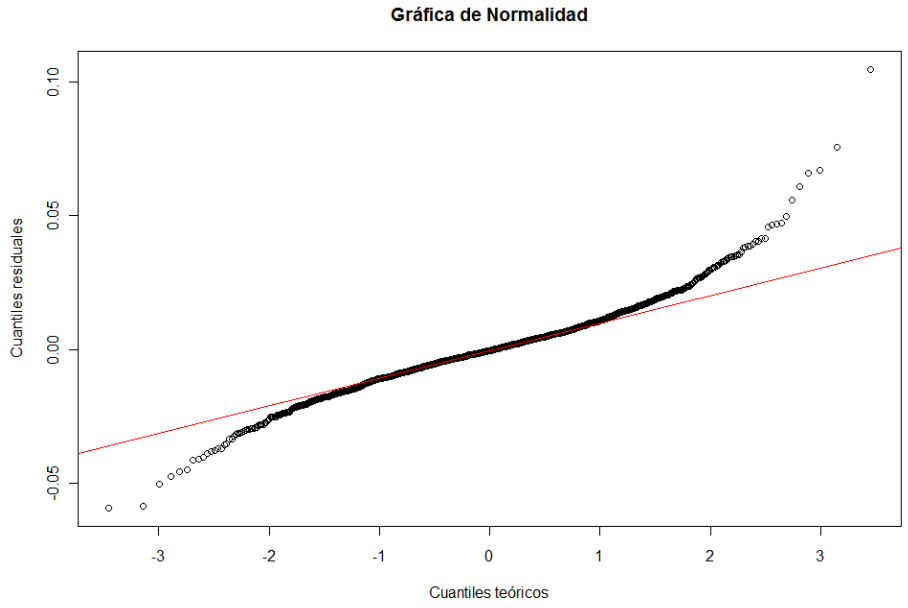


Gráficas 2 a 5. Residuos contra variables seleccionadas de tipo cuantitativo.



En la Gráfica 6 se observa que existen valores que salen del comportamiento de normalidad, como ejercicio se realizó un análisis de puntos extremos e influyentes, identificando 20 puntos, mismos que fueron eliminados del archivo, se aplicaron al modelo final, logrando un mejor ajuste y en el análisis de residuales cumplió la condición de normalidad, sin embargo, se decidió no eliminar estos casos ya que según comentaron las personas responsables del levantamiento, que los datos fueron comprobados y calificados como datos reales. Por lo que se continuó este trabajo con el archivo completo.

Grafica 6. Normalidad de residuales del modelo GREG.



Ahora que se cuenta con un modelo razonablemente correcto, se calculan las estimaciones del ingreso promedio, su error cuadrático medio y su coeficiente variación utilizando las expresiones (6), (7) y (8) descritas en el capítulo 2 para la estimación GREG, (ver "Código en R"), las estimaciones para los municipios en muestra se presentan en la Tabla 2.

Tabla 2. Municipios en muestra según su estimación GREG

Clave de municipio	Nombre	Tamaño de la muestra	Valores con estimación GREG		
			Ingreso promedio	$\sqrt{\hat{ECM}}$	Coeficiente de variación
	Sonora	1817	38,002	630.4	0.017
2	Agua Prieta	38	27,082	2,507.5	0.093
3	Alamos	18	14,775	1,452.1	0.098
12	Bácum	18	16,186	2,244.2	0.139
13	Banámichi	15	25,307	3,222.9	0.127
17	Caborca	19	22,591	2,459.7	0.109
18	Cajeme	233	28,915	1,292.7	0.045
19	Cananea	18	28,497	3,576.5	0.126
25	Empalme	34	21,657	2,100.4	0.097
26	Etchojoa	20	31,857	7,350.6	0.231
29	Guaymas	104	23,087	1,261.8	0.055
30	Hermosillo	737	41,540	988.3	0.024
33	Huatabampo	54	16,591	1,205.9	0.073
36	Magdalena	18	27,964	2,543.5	0.091
41	Nacozari de García	18	19,686	2,537.0	0.129
42	Navojoa	113	19,377	948.1	0.049
43	Nogales	113	27,408	1,433.1	0.052
48	Puerto Peñasco	36	51,365	5,489.1	0.107
53	San Felipe de Jesús	19	20,919	3,189.6	0.152
55	San Luis Río Colorado	89	30,050	1,729.2	0.058
58	Santa Ana	17	28,634	3,861.3	0.135
66	Ures	14	22,088	3,453.8	0.156
71	Benito Juárez	39	12,014	1,614.1	0.134
72	San Ignacio Río Muerto	32	11,956	1,703.9	0.143

De los resultados de la Tabla 2 podemos observar respecto a la estimación directa (Tabla 1) que son mas centralizadas respecto a la estimación estatal, y que prácticamente para todos los municipios con excepción de Ures el coeficiente de variación disminuye alrededor de 50%.

3.4 ESTIMACIÓN SINTÉTICA

El modelo sintético aplicado es de la forma:

$$\hat{\bar{y}}_{syn,a} = \bar{X}_a \hat{\beta} + e_a$$

En esta ocasión las variables auxiliares se seleccionaron con el método *forward,* debido a que sólo se cuenta con veintitrés promedios de la variable dependiente y por otro lado se tienen 189 variables a escoger, por lo que se compararon los modelos lineales con cada una de las variables mediante la función anova y se eligió la que menor suma de residuales cuadráticos presento en el modelo, después se fijó esta variable en el modelo y se probó una segunda variable significativa, verificando que la correlación entre ellas no fuera alta (<|0.7|) y que mantuvieran baja colinealidad, y así sucesivamente, se finalizó la selección de variables cuando ya ninguna aportaba una mejora significativa al modelo. Las variables seleccionadas fueron:

- Promedio de escolaridad masculina en la vivienda.
- Proporción de viviendas cuyo jefe del hogar migra a EU.
- Proporción de viviendas con un hijo menor de 12 años.
- Proporción de viviendas con tres habitantes de 15 a 24 años.
- Proporción de viviendas en donde no se le puede echar agua al baño.
- Proporción de viviendas con un habitante de 50 a 64 años.

El modelo lineal obtenido con la función lm de R es el siguiente:

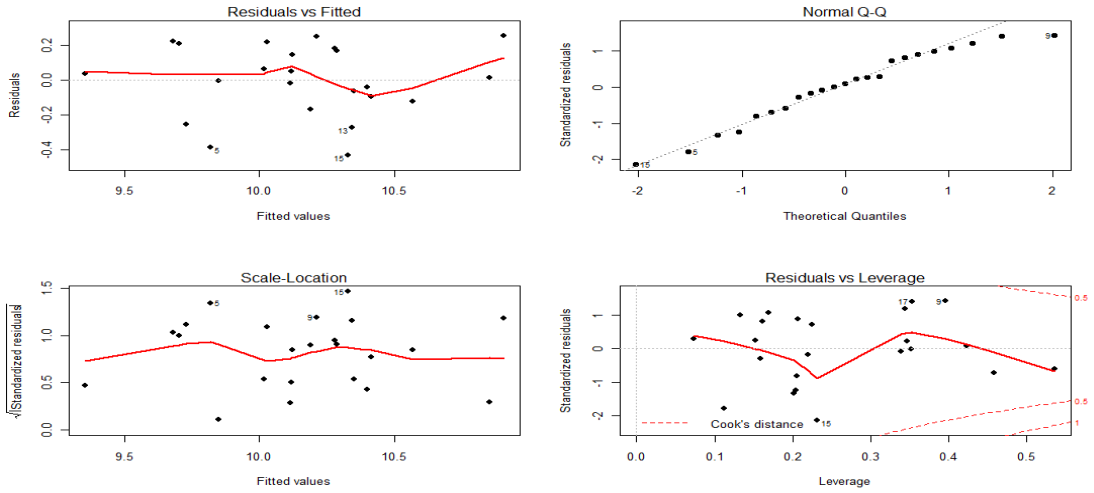
```
Call:
lm(formula = yHT ~ escpromh + migeuj_1 + hijo12c1 + e15a24_3 +
    banoagua3 + e50a64_1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39449 -0.09218  0.07120  0.11164  0.25237

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5812     1.2400   5.307 7.08e-05 ***
escpromh      0.4142     0.1206   3.434 0.003406 **
migeuj_1     67.2048    16.2774   4.129 0.000788 ***
hijo12c1    -10.1229     3.0765  -3.290 0.004612 **
e15a24_3     45.6350    12.1238   3.764 0.001697 **
banoagua3    -1.3072     0.5850  -2.234 0.040073 *
e50a64_1      5.5678     2.8857   1.929 0.071606 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2123 on 16 degrees of freedom
Multiple R-squared: 0.813,      Adjusted R-squared: 0.7428
F-statistic: 11.59 on 6 and 16 DF,  p-value: 4.686e-05
```

La tabla de colinealidad y la matriz de correlación entre las variables del modelo sintético se presentan en los Anexos 3 y 4 respectivamente , en ellas se puede observar que la correlación no es alta (>0.7) y que del análisis de colienalidad solamente en la componentes principal 7 la proporción de la varianza es mayor al 0.5 en dos coeficientes *intercept* y *e50a64_1,* sin embargo,

el índice de condición no es grande 80.938, con lo que se concluye que no hay problemas severos de colinealidad.

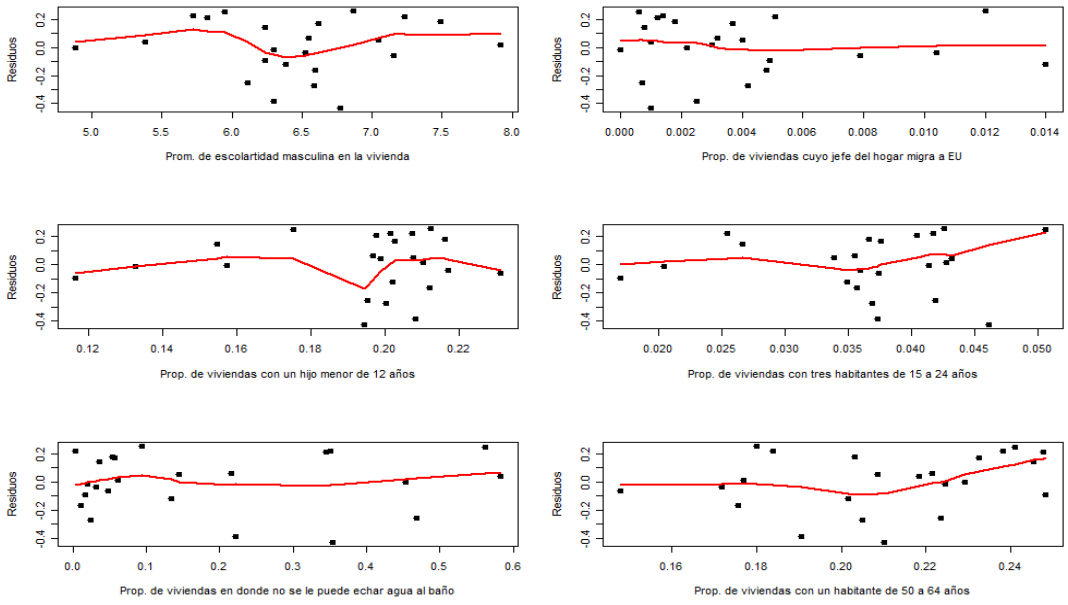
Para el análisis de residuales, R proporciona cuatro gráficas que dan información acerca del comportamiento de los residuos, mismas que se presentan a continuación.

Gráficas 7 a 10.- Gráficas para análisis de residuales.



En las gráficas de la izquierda, podemos observar que el comportamiento de los residuales respecto a los valores ajustados no presenta tendencia, representado por la línea que presenta trayectoria casi lineal, lo que indica homocedasticidad de los residuales, en la gráfica superior derecha se puede observar que los residuos guardan una relación lineal al compararlos con los cuantiles teóricos de una distribución normal, lo que indica normalidad de los residuos, la gráfica inferior derecha nos revela que no hay puntos influyentes.

Gráficas 11 a 16. Residuales contra variables cuantitativas modelo sintético



En las gráficas 11 a16 se observa el comportamiento de cada una de las variables auxiliares del modelo respecto a los residuales del modelo, en todos los casos se observa un comportamiento casi lineal representado por la línea, por lo que no es necesaria ninguna transformación en alguna de las variables.

Una vez que el modelo cumple con las condiciones de normalidad de residuos, homocedasticidad, variables auxiliares incorrelacionadas y significativas y sin problemas de colinealidad, se procede a utilizar las expresiones (9), (10), (11) y (12) del Capítulo 2 y se obtienen las estimaciones del ingreso promedio total, el ECM y el coeficiente de variación, para cada municipio en muestra, mismos que se presentan en la Tabla 3.

Tabla 3. Municipios en muestra según su estimación Sintética

Clave de municipio	Nombre	Tamaño de la muestra	Valores con estimación Sintética		
			Ingreso promedio	$\sqrt{\hat{ECM}}$	Coeficiente de variación
	Sonora				
2	Agua Prieta	38	32,778	3,713.0	0.113
3	Alamos	18	18,967	2,571.4	0.136
12	Bácum	18	16,367	3,705.1	0.226
13	Banámichi	15	33,381	6,388.1	0.191
17	Caborca	19	18,408	7,238.4	0.393
18	Cajeme	233	29,159	6,110.9	0.210
19	Cananea	18	22,690	5,856.8	0.258
25	Empalme	34	22,449	2,000.4	0.089
26	Etchojoa	20	27,225	7,886.0	0.290
29	Guaymas	104	24,762	2,569.0	0.104
30	Hermosillo	737	51,527	7,699.6	0.149
33	Huatabampo	54	16,809	4,615.2	0.275
36	Magdalena	18	31,015	8,993.7	0.290
41	Nacozari de García	18	26,632	5,172.8	0.194
42	Navojoa	113	30,577	13,519.5	0.442
43	Nogales	113	31,261	3,431.5	0.110
48	Puerto Peñasco	36	54,285	15,772.2	0.291
53	San Felipe de Jesús	19	24,658	3,306.2	0.134
55	San Luis Río Colorado	89	38,811	7,584.8	0.195
58	Santa Ana	17	29,382	5,681.3	0.193
66	Ures	14	24,850	4,500.1	0.181
71	Benito Juárez	39	16,026	3,867.8	0.241
72	San Ignacio Río Muerto	32	11,578	1,626.1	0.140

De los resultados de la Tabla 3, se puede observar que las estimaciones son más centralizadas que la estimación con el modelo GREG (Tabla 2) y la estimación directa (Tabla 1), sin embargo, sus coeficientes de variación son mayores que el modelo GREG y muy inestables. Esto da evidencia de que el comportamiento estatal no es parecido al de los municipios, aportando sesgos importantes al ECM.

3.5 Estimación Compuesta

Tomando la expresión (15) del Capítulo 2 referente a la estimación de ϕ_a y con los ahora conocidos errores cuadráticas medios de la estimación directa y la sintética presentadas en las tablas 1 y 3 respectivamente, los valores de $\hat{\phi}_a$ para los municipios en muestra se presentan en la Tabla 4.

Tabla 4. Valor de $\hat{\phi}_a$ para los municipios en muestra

Clave	Nombre del municipio	Valor de $\hat{\phi}_a$
2	Agua Prieta	0.4850093
3	Alamos	0.3332025
12	Bácum	0.3730679
13	Banámichi	0.5504475
17	Caborca	0.7432860
18	Cajeme	0.6857322
19	Cananea	0.5171451
25	Empalme	0.3167517
26	Etchojoa	0.1153923
29	Guaymas	0.5152607
30	Hermosillo	0.7966212
33	Huatabampo	0.6286225
36	Magdalena	0.8363486
41	Nacozari de García	0.6125280
42	Navojoa	0.9109357
43	Nogales	0.5672542
48	Puerto Peñasco	0.7169443
53	San Felipe de Jesús	0.1997037
55	San Luis Río Colorado	0.8127725
58	Santa Ana	0.4569397
66	Ures	0.5687633
71	Benito Juárez	0.4186665
72	San Ignacio Río Muerto	0.2395681

Para los municipios no muestreados ϕ_a toma el valor cero, es decir, las estimaciones son las mismas que la del estimador sintético, ahora tomando las expresiones (13), (14) y (16), se calculan las estimaciones compuestas para la media, el ecm y el coeficiente de variación del ingreso municipal de las viviendas para cada municipio, mismos que se muestran en la Tabla 5.

Tabla 5. Municipios en muestra según su estimación Compuesta

Clave de municipio	Nombre	Tamaño de la muestra	Valores con estimación Compuesta		
			Ingreso promedio	$\sqrt{E\hat{C}M}$	Coeficiente de variación
	Sonora				
2	Agua Prieta	38	32,203	3,796.4	0.118
3	Alamos	18	18,954	3,207.7	0.169
12	Bácum	18	17,797	4,404.3	0.247
13	Banámichi	15	31,751	5,902.5	0.186
17	Caborca	19	14,030	4,511.8	0.322
18	Cajeme	233	33,150	4,371.7	0.132
19	Cananea	18	25,585	5,705.9	0.223
25	Empalme	34	22,919	2,543.7	0.111
26	Etchojoa	20	28,123	12,342.7	0.439
29	Guaymas	104	25,454	2,510.1	0.099
30	Hermosillo	737	52,156	4,118.4	0.079
33	Huatabampo	54	14,435	3,712.9	0.257
36	Magdalena	18	24,853	4,191.6	0.169
41	Nacozari de García	18	24,162	4,289.8	0.178
42	Navojoa	113	20,875	4,379.3	0.210
43	Nogales	113	30,207	3,083.1	0.102
48	Puerto Peñasco	36	65,680	10,501.2	0.160
53	San Felipe de Jesús	19	24,585	4,769.6	0.194
55	San Luis Río Colorado	89	35,260	3,847.6	0.109
58	Santa Ana	17	31,872	6,047.1	0.190
66	Ures	14	27,057	4,033.0	0.149
71	Benito Juárez	39	17,696	4,336.8	0.245
72	San Ignacio Río Muerto	32	11,692	2,251.4	0.193

Como se observa en los resultados de la Tabla 5 las estimaciones en los municipios en muestra están entre la estimación directa HT (Tabla 1) y la estimación sintética (Tabla 3), en esta composición predomina la estimación en la que el error cuadrático medio es menor.

3.6 ESTIMACIÓN BASADA EN MODELO DE ÁREA

Partiendo de que la información referente a cada vivienda en la ENIGH no es posible asociarla a la misma vivienda del II Conteo, más aún, tampoco es posible asociar la UPM o el estrato del diseño

de muestreo a conglomerados de viviendas en el II Conteo, pero sí es posible la asociación a nivel de municipio, ya que al tomar esta conglomeración de viviendas, las UPM del diseño de la ENIGH están totalmente contenidas en su municipio correspondiente, es decir, en cada municipio de la muestra existe por lo menos una UPM.

Entonces, la forma de estimación que se utilizará es la de "Modelo de Área", el área pequeña considerada es el municipio, por lo que se propone el siguiente modelo lineal mixto:

$$\log(\bar{y}_a) = \bar{X}_a\beta + \upsilon_a + e_a$$

Donde $\bar{y}_a$ es el estimador directo del promedio municipal del ingreso para cada vivienda con errores de muestreo e_a, $\bar{X}_a$ es un vector de variables auxiliares que contienen promedios municipales asociados con el ingreso promedio, β es el vector de parámetros de la regresión, υ_a es el error del modelo (efectos aleatorios de área pequeña) para el área a. Suponemos $e_a \sim N\left(0, \psi_a^2\right)$ y $\upsilon_a \sim N\left(0, \sigma_\upsilon^2\right)$ son independientes.

Para el caso que nos ocupa, tomamos a $\mathrm{I}\psi_a^2$ como la matriz de varianzas estimadas de forma directa para cada municipio, la estimación MELIE para $log(y_a)$ es

$$\log(\tilde{\bar{y}}_a) = \bar{X}_a\tilde{\beta} + \hat{\gamma}_a(\hat{\theta}_a - \bar{X}_a\tilde{\beta}) \text{ donde } \hat{\gamma}_a = \hat{\sigma}_\upsilon^2 / (\hat{\psi}_a^2 + \hat{\sigma}_\upsilon^2) \quad (35)$$

Los parámetros $\tilde{\beta}$ y los componentes de la varianza $\hat{\sigma}_\upsilon^2$ se estiman de forma iterativa utilizando las expresiones (25), (26) y (27) para MV y las expresiones (28), (29) y (30) para MVR tomadas del Capítulo 2, esto una vez que se eligieron las variables auxiliares que entran en la regresión.

Debido a que tal y como está planteado el modelo sintético para los municipios se puede ver que es el mismo que el de "Modelo de Área", excepto por los efectos aleatorios de municipio, los que a su vez, son efectos aleatorios en el intercepto, entonces, se consideran las mismas variables auxiliares del modelo sintético para el "Modelo de Área".

Los valores obtenidos para los coeficientes $\tilde{\beta}$ y su error estándar se presentan en la Tabla 6, en donde es fácil observar que ajustan sin ningún problema al modelo propuesto.

Tabla 6. Variables seleccionadas para el Modelo de Área con valores de $\tilde{\beta}$ ajustados y su error estándar

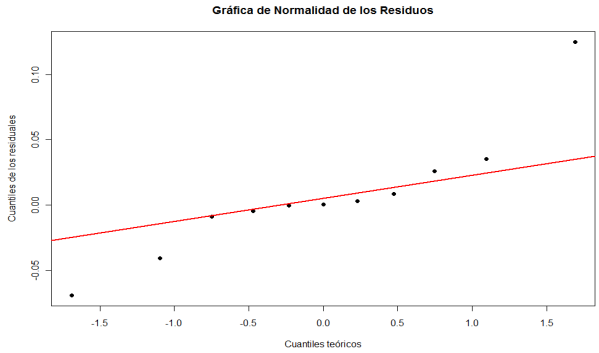
Variable		Valor estimado MELIE	Error estándar
Descripción	Mnemónico		
Intercepto		8.230	0.552
Promedio de escolaridad masculina en la vivienda	Escpromh	0.443	0.101
Proporción de viviendas cuyo jefe del hogar migra a EU.	migeuj_1	58.200	13.255
Proporción de viviendas con un	hijo12c1	-14.254	2.292

Variable		Valor estimado MELIE	Error estándar
Descripción	Mnemónico		
hijo menor de 12 años.			
Proporción de viviendas con tres habitantes de 15 a 24 años.	e15a24_3	49.044	10.027
Proporción de viviendas en donde no se le puede echar agua al baño.	banoagua3	-1.009	0.482

NORMALIDAD DE RESIDUALES Y DE EFECTOS ALEATORIOS DEL MODELO DE ÁREA

Los valores de las varianzas en la estimación directa son indefinidas para 11 de los 23 municipios en muestra, debido a que sólo existe una UPM en cada uno de esos municipios; la estimación de los promedios del ingreso obtenida para ellos con el modelo de área son idénticas a la estimación directa, entonces, solamente se verifica la normalidad de la distribución de los residuos diferentes de cero. En la Gráfica 17 se comparan los cuantilies teóricos con los cuantiles de los residuales del modelo y muestra de forma visual la normalidad de los residuales.

Gráfica 17. Normalidad de los residuos del Modelo de Área



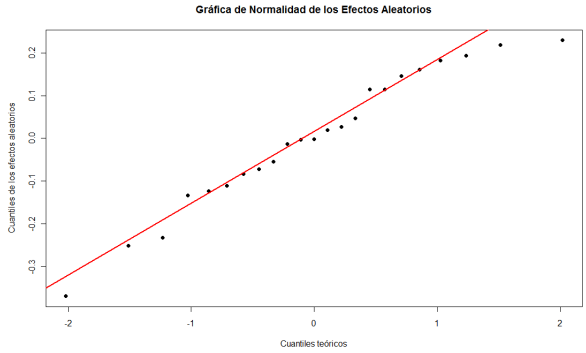
Para comprobar de forma numérica, se aplica la prueba de normalidad de Shapiro-Wlik con el siguiente resultado.

```
        Shapiro-Wilk normality test
data:  resE[resE != 0]
W = 0.8754, p-value = 0.09086
```

En donde el p-valor está en la zona de aceptación al 95% de confianza.

La normalidad de los efectos aleatorios se muestra visualmente en la gráfica 18.

Gráfica 18. Normalidad de los efectos aleatorios del Modelo de Área



La comprobación numérica, se hace también mediante la prueba de normalidad de Shapiro-Wlik con el siguiente resultado.

```
        Shapiro-Wilk normality test
data:  eblup1$randeff
W = 0.9614, p-value = 0.4917
```

En donde el p-valor está en la zona de aceptación al 95% de confianza.

Una vez obtenidas las estimaciones de $\tilde{\beta}$ y $\hat{\sigma}_v^2$ así como de los efectos aleatorios, se estiman las componentes de ECM g_1, g_2 y g_3 tomando las expresiones (31) y (32) para MV y la expresión (33) para MVR. Los ECM para los municipios con una sola UPM son cero, esto aunado a que las estimaciones del promedio del ingreso son idénticas a las estimaciones directas, motiva a realizar las estimaciones tomando el modelo de área obtenido y aplicarlo a estos municipios, como si no hubieran tenido muestra, así, aplicando las expresiones (35) y (34) se obtiene el valor del promedio del ingreso, el ECM y su coeficiente de variación.

En la Tabla 7 se presentan los resultados para todos los municipios en muestra.

Tabla 7. Municipios en muestra estimados para Áreas Pequeñas con Modelo de Área

Clave de municipio	Nombre	Tamaño de la muestra	Valores con estimación AP con MA		
			Ingreso promedio	$\sqrt{\hat{ECM}}$	Coeficiente de variación
	Sonora				
2	Agua Prieta	38	31,609	2,155.7	0.068
3	Alamos	18	18,963	3,735.7	0.197
12	Bácum	18	16,226	2,921.0	0.145
13	Banámichi	15	34,011	7,113.2	0.234
17	Caborca	19	18,122	3,226.3	0.258
18	Cajeme	233	33,765	2,882.7	0.085
19	Cananea	18	22,487	4,418.1	0.156
25	Empalme	34	23,928	306.2	0.013
26	Etchojoa	20	30,269	6,170.5	0.176

Clave de municipio	Nombre	Tamaño de la muestra	Valores con estimación AP con MA		
			Ingreso promedio	$\sqrt{\hat{ECM}}$	Coeficiente de variación
29	Guaymas	104	25,888	2,499.3	0.097
30	Hermosillo	737	52,788	3,051.2	0.058
33	Huatabampo	54	13,968	1,175.0	0.084
36	Magdalena	18	30,424	5,618.5	0.238
41	Nacozari de García	18	25,580	4,739.1	0.210
42	Navojoa	113	29,290	4,779.5	0.163
43	Nogales	113	29,540	1,405.9	0.048
48	Puerto Peñasco	36	68,384	4,081.0	0.060
53	San Felipe de Jesús	19	24,617	4,786.8	0.197
55	San Luis Río Colorado	89	35,872	3,844.5	0.107
58	Santa Ana	17	29,017	5,266.3	0.151
66	Ures	14	24,449	4,548.7	0.158
71	Benito Juárez	39	17,662	2,306.2	0.131
72	San Ignacio Río Muerto	32	12,017	673.0	0.056

Validación cruzada

Para cuantificar la confiabilidad de las predicciones del modelo se utiliza la validación cruzada, misma que fue introducida por Stone (1974), la idea es dividir la muestra en k submuestras aproximadamente iguales, en cada caso una de las partes se convierte en una muestra de prueba que sirve para validar el modelo y las k-1 muestras restantes constituyen lo que es una muestra de entrenamiento que sirve para construir el modelo (en particular si la submuestra es de un elemento se tiene una validación similar al Jackknife, se le conoce como *"leave-on-out")*, luego se cuantifica el error de predicción de la que se dejó afuera, se repite el proceso para cada una de las observaciones y se promedia los errores cuadráticos, el resultado es llamado el $PRESS$ (Allen 1974).

Para su aplicación en modelos de regresión se debe satisfacer que los errores ϵ_i deben ser iid normal estándar, entonces,

$$PRESS = \frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2 \qquad (36)$$

Evaluando (21) con las predicciones obtenidas a partir del modelo de área ajustado con los n-1 municipios muestreados en cada caso, se obtiene que $PRESS$ = 0.0780

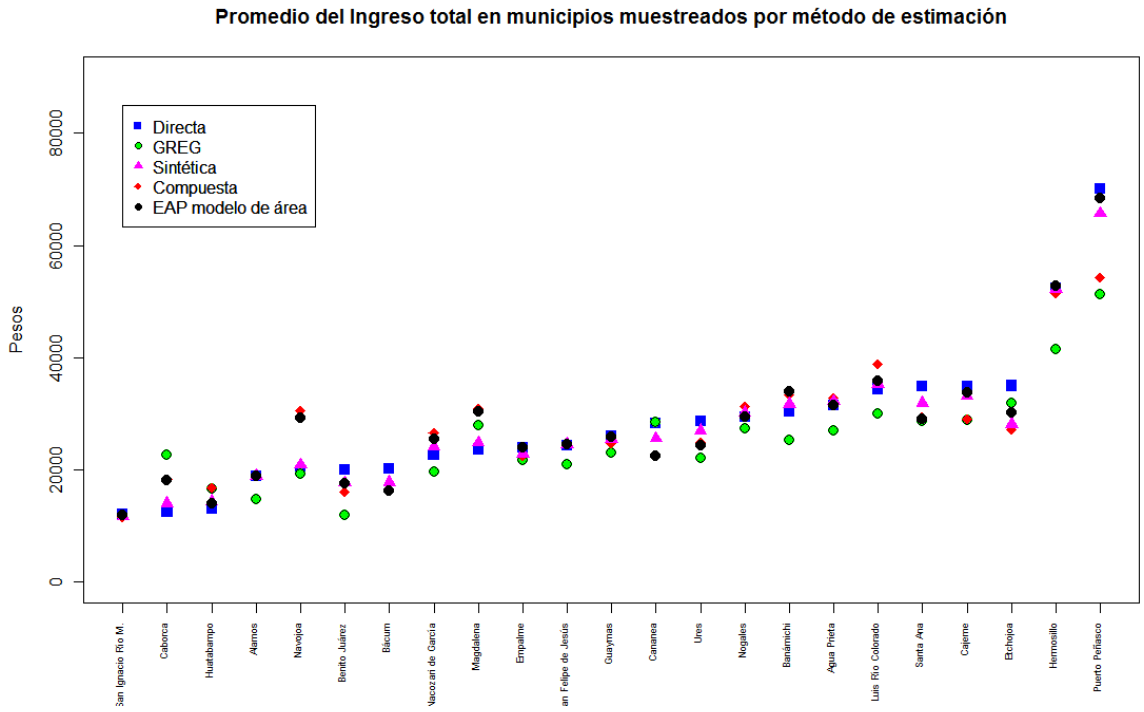
Por otro lado, tenemos que el error cuadrático residual del modelo de área propuesto con los n municipios en muestra es 0.0979, es decir, 25% mayor que el *PRESS*, y aunque no hay reglas firmes al respecto, se considera que el modelo es fiable cuando el error cuadrático residual y el PRESS es semejante.

3.7 COMPARACIÓN DE RESULTADOS

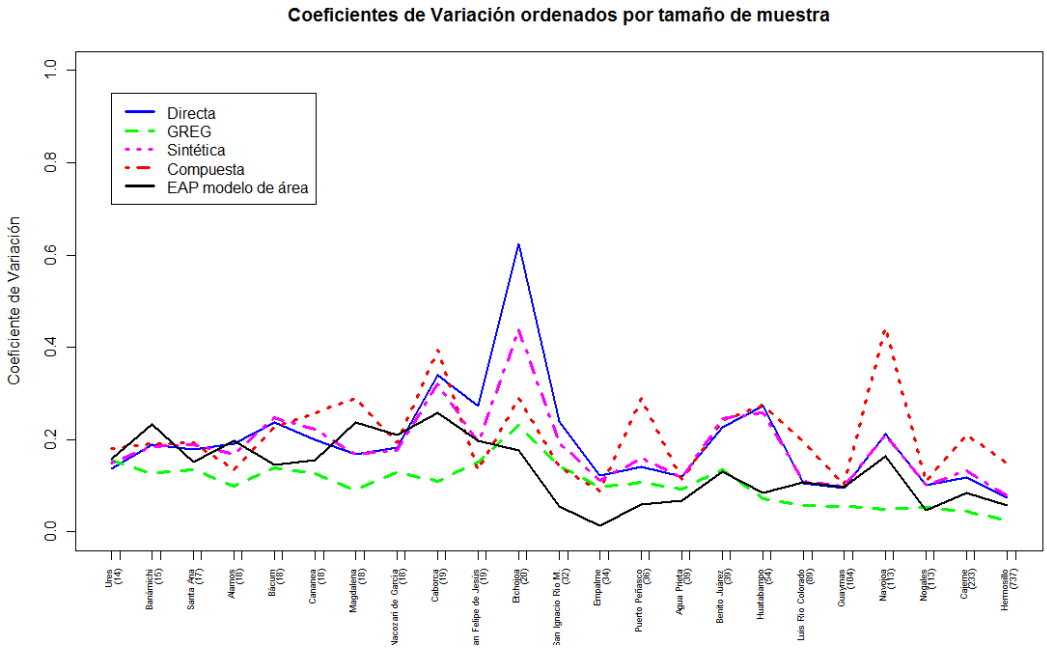
En las siguientes gráficas se comparan los resultados obtenidos por los métodos de estimación abarcados en este documento, en la Gráfica 19 se presentan las estimaciones del ingreso total promedio por vivienda para los municipios muestreados por la ENIGH, en donde se observa que la estimación GREG proporciona por lo general los valores más bajos y, por el contrario, la estimación directa los más altos, los valores de la estimación GREG son los más centralizados debido a que provienen de un modelo general para el estado, obtenido con todas las unidades de la muestra, en cuanto a centralización le siguen los de la estimación sintética, que también provienen de un modelo estatal pero obtenido con los promedios municipales de los 23 municipios muestreados, después están los obtenidos mediante el modelo compuesto, que como ya se dijo, son un promedio ponderado entre la estimación sintética y la directa, los valores de la estimación para áreas pequeñas son más dispersos debido a que aunque provienen de un modelo obtenido con los promedios de los 23 municipios en muestra, éste considera las diferencias municipales observadas en la muestra incluidas como efectos aleatorios de área.

Para tener una idea del nivel de precisión de las estimaciones, en la Gráfica 20 se muestran los coeficientes de variación estimados ordenados de menor a mayor por el tamaño de la muestra, se observa que para las estimación directa, sintética y compuesta los valores superan en la mayoría de los casos a 0.15, punto en donde generalmente se considera que la estimación comienza a ser imprecisa, además se observa una gran inestabilidad en las precisiones estimadas, para los modelos GREG y el de áreas pequeñas con modelo a nivel área se observa que las estimaciones tienden a ser más precisas conforme aumenta el tamaño de la muestra, es decir, su coeficiente de variación es menor para áreas con mayor muestra, el modelo de GREG muestra por lo general el menor coeficiente de variación y mayor estabilidad, sin embargo, el inconveniente del GREG es que no es posible hacer estimaciones en municipios sin muestra.

Gráfica 19. Ingreso total promedio municipal por vivienda según método de estimación

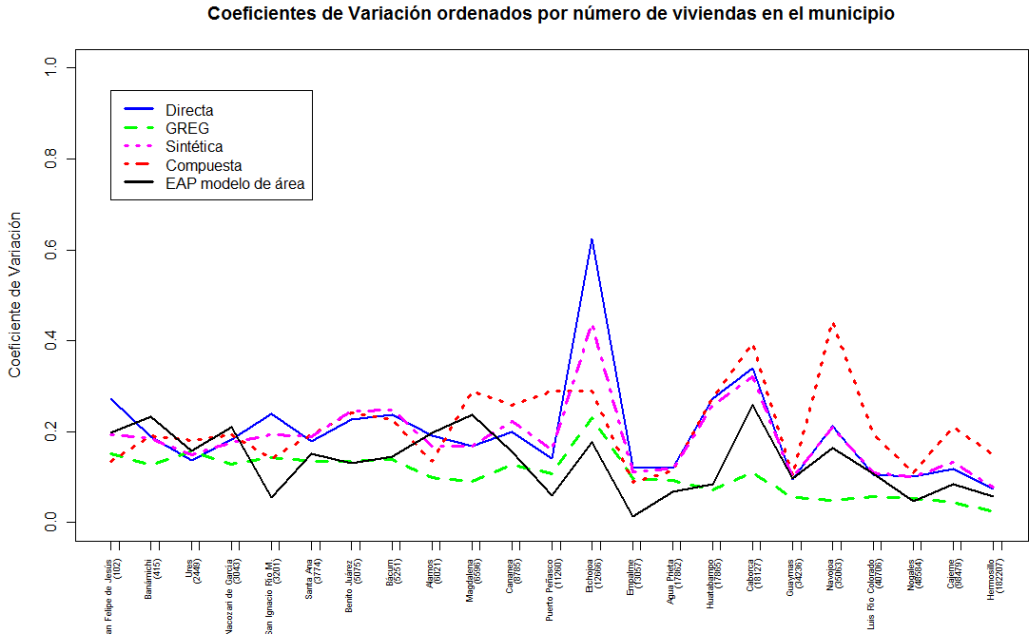


Gráfica 20. Coeficientes de Variación municipal según método de estimación ordenados por tamaño de muestra



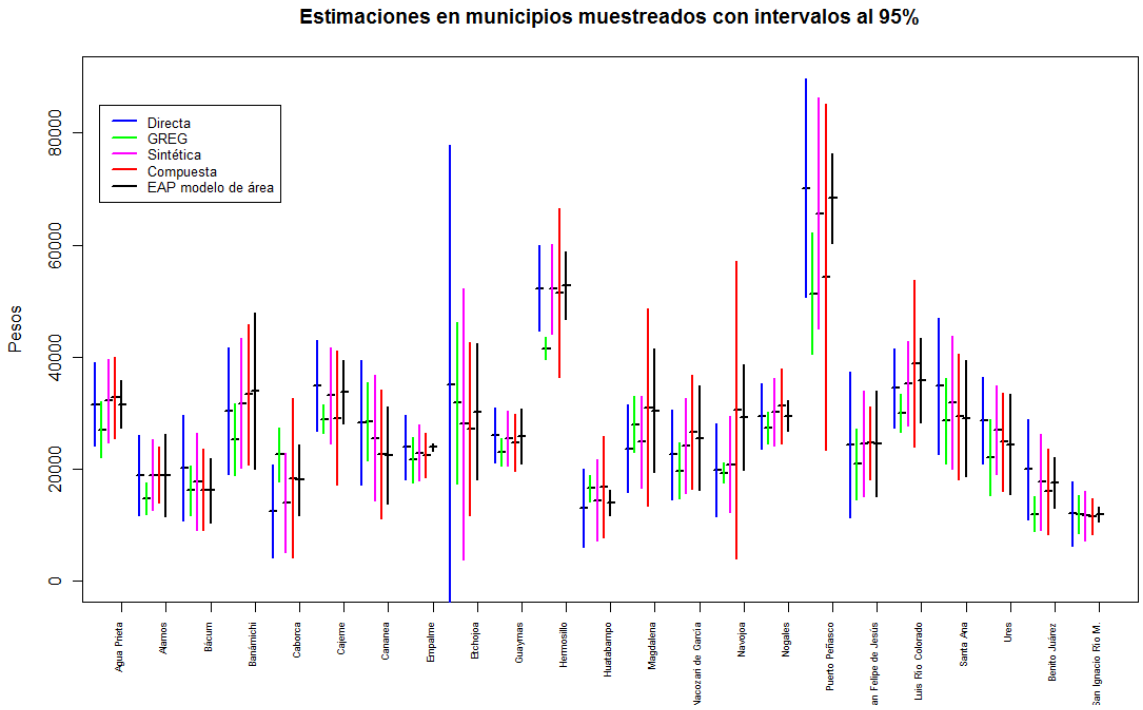
En la Gráfica 21 se muestran los coeficientes de variación estimados ordenados de menor a mayor por el número de viviendas en el municipio para el año 2005, las observaciones son prácticamente las mismas que las de la Gráfica 20, lo cual confirma que en el diseño de la muestra se incluyó el tamaño de la localidad y que está muy relacionada para la entidad considerada con el tamaño del municipio.

Grafica 21. Coeficientes de variación municipal según método de estimación ordenados por número de viviendas en el municipio



A manera de resumen, en la Gráfica 22 se presenta para cada municipio en muestra, las estimaciones del ingreso promedio con intervalos de confianza al 95% para los métodos de estimación aquí abordados.

Gráfica 22. Estimación del ingreso promedio municipal por vivienda por método de estimación con intervalos de confianza al 95%



RESULTADOS DE LA ESTIMACIÓN EN ÁREAS NO MUESTREADAS

En este trabajo, exclusivamente los métodos de estimación sintética y el de áreas pequeñas con "Modelo de Área" proporcionan estimaciones para los municipios no muestreados.

Para el modelo sintético se utilizan las expresiones (9) y (11) tomando a $\boldsymbol{X}_a$ como $\boldsymbol{X}_{a.}$, esto es, los valores de las variables auxliares (promedios municipales censales) en los municipios no muestreados.

En el caso del "Modelo de Área", para calcular el promedio del ingreso total en la vivienda se recurre a la expresión (24) en donde la se indefine debido a que la estimación directa también lo está, entonces, (24) se reduce a $\log(\tilde{\bar{y}}_{a.}) = \bar{X}_{a.}\tilde{\boldsymbol{\beta}}$, el ECM se estimada mediante la expresión (34).

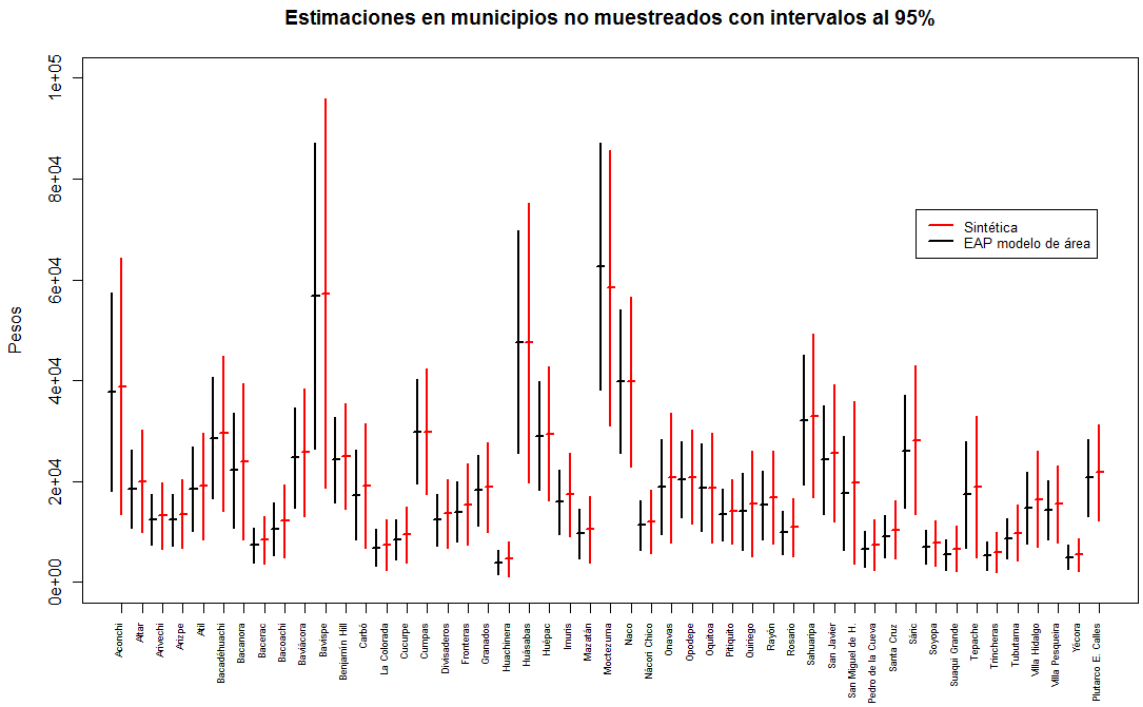
La Tabla 8 contiene los valores numéricos usados para la elaboración de la Gráfica 23, que muestra las estimaciones con sus respectivos intervalos de confianza; se observa que las estimaciones por ambos métodos son similares ya que las variables utilizadas en cada modelo son las mismas, la diferencia está en los efectos aleatorios por municipio y a la variabilidad de la variable dependiente, ambos considerados en el "Modelo de Área".

Tabla 8. Municipios no muestreados según su estimación Sintética y de AP con Modelo de Área

Municipio		Estimación sintética		Estimación AP con Modelo de Área	
Clave	Nombre	Ingreso promedio	$\sqrt{\hat{ECM}}$	Ingreso promedio	$\sqrt{\hat{ECM}}$
1	Aconchi	38,817	12,955.8	37,845	10,030.2
4	Altar	20,037	5,183.0	18,560	3,931.3
5	Arivechi	13,241	3,366.4	12,399	2,583.5
6	Arizpe	13,576	3,479.5	12,365	2,610.3
7	Atil	19,076	5,405.0	18,517	4,230.9
8	Bacadéhuachi	29,486	7,795.8	28,626	6,120.3
9	Bacanora	23,927	7,896.1	22,210	5,804.5
10	Bacerac	8,381	2,419.2	7,411	1,759.2
11	Bacoachi	12,167	3,685.0	10,522	2,641.5
14	Baviácora	25,774	6,456.6	24,688	5,053.8
15	Bavispe	57,270	19,692.4	56,799	15,488.9
16	Benjamín Hill	25,028	5,309.0	24,296	4,329.1
20	Carbó	19,108	6,278.5	17,257	4,535.1
21	La Colorada	7,459	2,538.9	6,881	1,870.7
22	Cucurpe	9,444	2,815.9	8,475	2,070.7
23	Cumpas	29,858	6,335.8	29,855	5,305.6
24	Divisaderos	13,616	3,488.2	12,415	2,618.0
27	Fronteras	15,425	4,134.0	13,991	3,070.4
28	Granados	18,862	4,507.4	18,249	3,580.2
31	Huachinera	4,600	1,804.8	3,931	1,242.8
32	Huásabas	47,512	14,121.0	47,674	11,297.5
34	Huépac	29,431	6,753.3	29,000	5,482.7
35	Imuris	17,365	4,226.9	15,894	3,242.9
37	Mazatán	10,488	3,385.7	9,710	2,509.0
38	Moctezuma	58,384	13,952.7	62,633	12,474.2
39	Naco	39,791	8,578.4	39,905	7,241.2
40	Nácori Chico	11,935	3,194.6	11,321	2,475.7
44	Onavas	20,700	6,543.8	18,892	4,781.8
45	Opodepe	20,899	4,774.8	20,396	3,862.6
46	Oquitoa	18,716	5,524.3	18,768	4,405.3
47	Pitiquito	14,009	3,264.4	13,395	2,599.1
49	Quiriego	15,579	5,379.3	14,028	3,864.6
50	Rayón	16,868	4,725.1	15,284	3,476.3

Municipio		Estimación sintética		Estimación AP con Modelo de Área	
Clave	Nombre	Ingreso promedio	$\sqrt{E\hat{C}M}$	Ingreso promedio	$\sqrt{E\hat{C}M}$
51	Rosario	10,881	2,943.0	9,831	2,189.5
52	Sahuaripa	32,994	8,299.5	32,200	6,608.4
54	San Javier	25,551	6,950.7	24,263	5,529.6
56	San Miguel de Horcasitas	19,819	8,198.9	17,676	5,750.3
57	San Pedro de la Cueva	7,445	2,550.7	6,538	1,800.3
59	Santa Cruz	10,446	2,984.0	9,082	2,143.2
60	Sáric	28,218	7,508.4	25,943	5,738.6
61	Soyopa	7,765	2,308.6	6,959	1,688.5
62	Suaqui Grande	6,634	2,312.1	5,456	1,550.3
63	Tepache	18,990	7,140.1	17,367	5,348.0
64	Trincheras	6,018	2,029.6	5,268	1,444.5
65	Tubutama	9,808	2,808.7	8,586	2,020.5
67	Villa Hidalgo	16,450	4,857.4	14,814	3,630.1
68	Villa Pesqueira	15,528	3,896.4	14,324	2,956.2
69	Yécora	5,466	1,684.6	4,955	1,251.7
70	Gral. Plutarco Elías Calles	21,797	4,851.2	20,759	3,885.4

Gráfica 23. Estimación del Ingreso promedio municipal de la vivienda en municipios no muestreados por método de estimación con intervalos de confianza al 95%



COMPROBACIÓN CONTRA LA ESTIMACIÓN ESTATAL DIRECTA

Para verificar que la estimación municipal del promedio del ingreso total vía el modelo de área sea consistente con la estimación del promedio estatal vía estimación directa, se procede a calcular el promedio ponderado del ingreso municipal, con la siguiente expresión:

$$\hat{\bar{y}}_{EMELIE} = \sum_{a=1}^{72} \frac{N_a}{N} \hat{\bar{y}}_{a,EMELIE} \quad \text{donde } N = \sum_{a=1}^{72} N_a$$

Así, $\hat{\bar{y}}_{EMELIE} = 36,026$

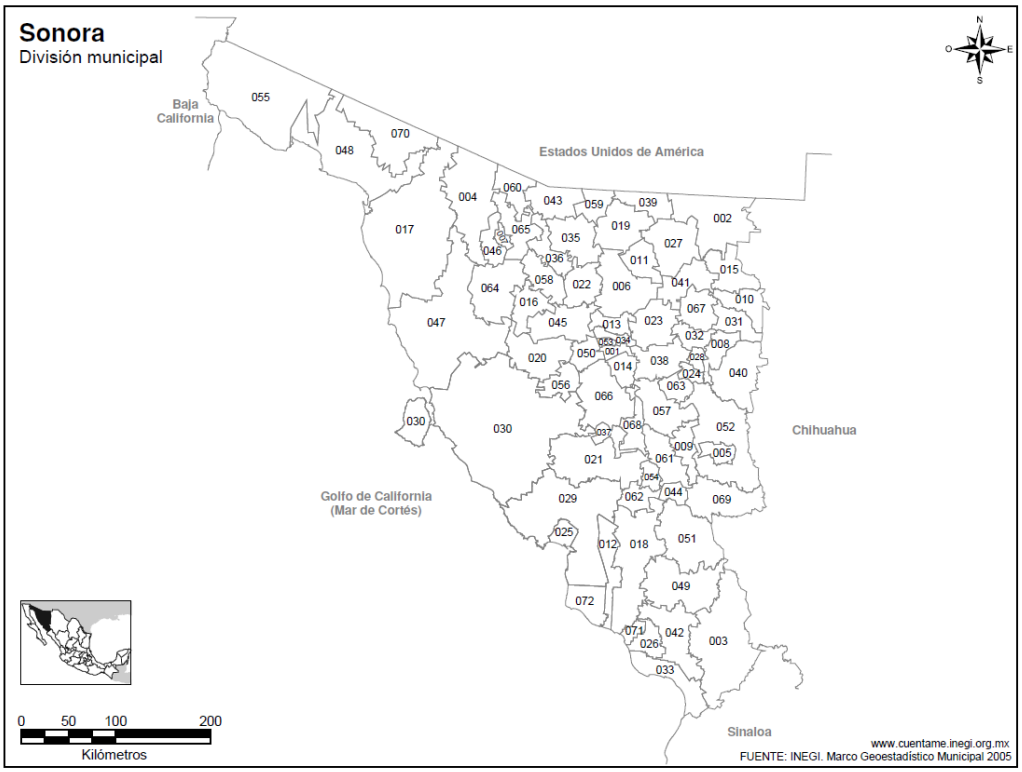
Por otra parte, el valor estimado directo es de $\hat{\bar{y}}_{HT} = 39,008$ con un error estándar $e\hat{s} = 2,797.6$, de donde se construye el intervalo de confianza al 95% para $\hat{\bar{y}}_{HT} = 39,008$ como:

$$I = \hat{\bar{y}}_{HT} \pm 1.95(s\hat{e}) = [33524, 44491]$$

Esto último nos dice que, con una confianza del 95% este intervalo capta al promedio real (desconocido).

Debido a que también en dicho intervalo está contenido $\hat{\bar{y}}_{EMELIE}$, se puede considerar que es consistente con el promedio estimado directo $\hat{\bar{y}}_{HT}$.

Mapa del estado de Sonora con división municipal y clave de municipio



CAPÍTULO IV.- CONCLUSIONES Y COMENTARIOS

Para la aplicación de modelos en la estimación de parámetros de una variable dada, es necesario contar con información auxiliar que tenga alguna relación con la variable a estimar, se puede decir que esta información se tiene con cierta abundancia en el INEGI, sin embargo, encontrar las variables que tengan correspondencia para modelar el comportamiento de la variable objetivo, es una labor ardua, en donde quizás el contar con el conocimiento empírico del fenómeno a analizar, o bien, acudir a una persona muy familiarizada con el tema puede aligerar dicha labor.

En este ejercicio, a pesar de la limitante en cuanto al número y sencillez de variables recolectadas en el II Conteo, se pudo encontrar un modelo razonablemente bueno, en donde las variables incluidas fueron seleccionadas con argumentos puramente estadísticos, quedaron incluidos los temas de escolaridad, migración, estructura de edad de los integrantes de la vivienda, sexo, y condiciones sanitarias de la vivienda, que conceptualmente se puede argumentar son los que sobresalen para relacionarlos con los ingresos en las viviendas del estado de Sonora en el año 2005. Las variables de tipo geográfico fueron incluidas en el análisis, aunque no resultaron significativas, sin embargo, no se deben descartar al analizar modelos para otros estados u otros cortes.

La estimación GREG fue la que presentó menor coeficiente de variación (CV) en los municipios con muestra, sin embargo, al contar con más parámetros lo convierte en un modelo menos eficaz para la predicción en los municipios con poca muestra y sin posibilidad en los municipios sin muestra. La EAP con modelo de área es la siguiente con menor CV, y además, considera en los efectos aleatorios de área, las características particulares del fenómeno vinculadas con la demarcación municipal y a su vez con su ubicación geográfica. Por lo que se puede considerar, que la EAP es la metodología que mejor se ajusta para la estimación del Ingreso promedio por vivienda para los municipios con y sin muestra en la ENIGH 2005 para el estado de Sonora.

Aunque en los valores mostrados en esta tesis muestran que la estimación basada en el modelo proporciona mejores resultados, sobre todo en las áreas con poco ó ninguna unidad muestral, no se debe ver como una metodología superior a la estimación directa, sino como un complemento de esta, por lo que sería prudente tomar en cuenta a los modelos de estimación en las etapas de diseño de las encuestas y los censos.

Es claro que este trabajo abarca solamente una pequeña porción de la estimación basada en el modelo, ya que es necesario todavía asimilar la estimación introduciendo la componente espacial y temporal al modelo de área, así como el modelo de áreas pequeñas a nivel unidad, encontrar el tamaño de muestra adecuado para la estimación por modelo, la estimación para áreas pequeñas con modelos logísticos, con diseño de panel, con más de dos etapas, etc.

La modelación estadística de los fenómenos socioeconómicos aprovechando los datos que recolecta el INEGI, daría como resultado un incremento radical en la información que se proporciona al Sistema Nacional de Información Estadística y Geográfica, además, de incrementar su calidad, validando la relación intrínseca entre variables de las diferentes fuentes de información.

RELACIÓN DE ARCHIVOS

Nombre	Descripción
Viv26.csv	Archivo en formato texto delimitado por comas que contiene las 1817 viviendas de la ENIGH 2005 del estado de Sonora con sus repectivos valores para las variables relacionadas en el Anexo 1.
Vivconteo.csv	Archivo en formato texto delimitado por comas que contiene el número de viviendas por muncipio del estado de Sonora según el II Conteo 2005.
Avgmun26.csv	Archivo en formato texto delimitado por comas que contiene el promedio por municipio del estado de Sonora de las variables referidas en el Anexo 1.
Tesis-MASC.R	Código R utilizado para generar los resultados y las gráficas.
R	Carpeta con las librerías y la versión de R utilizadas en este trabajo.
Estimación del ingreso promedio por vivienda en los municipios de Sonora.pdf	Versión en formato PDF de este documento

BIBLIOGRAFÍA

ACUÑA, E. **"Métodos para Estimar el Error de Predicción en Regresión: Validación Cruzada y Bootstraping"**. *Universidad de Puerto Rico*, 2003.

CASSEL, C. M., SÄRNDAL, C. E. and WRETMAN, J. H. **"Foundations of Inference in Survey Sampling"**. *John Wiley & Sons*, 1977.

CASSEL, C. M., SÄRNDAL, C. E. and WRETMAN, J. H. **"Some results on Generalized Difference Estimation and Generalized Regression Estimation for finite populations"..** *Biometrika*, Vol. 63, pp. 615-620, 1976.

DREW, D., SINGH, M. P. and CHOUDHRY, H. **"Evaluation of Small Area Estimation Techniques"**. *Survey Methodology*, Vol. 8, pp. 1747, 1982.

ELBERS, C., LANJOUW, J. O. and LANJOUW, P. **"Micro-level Estimation of Poverty and Inequality"**. *Econometrica*, Vol. 71, pp. 355-364, 2003.

EURAREA CONSORSIUM, "**Enhancing Small Area Estimation Techniques to meet European Needs",** Project Reference Vol. 2 Explanatory Appendices, 2004.

FAY, R. E. and HERRIOT, R. A. **"Estimates of income for Small Places: An Application of James-Stein procedures to census data".** *Journal of the American Statistical Association*, Vol. 74, pp. 269-277, 1979.

FOX, J. **"Linear Mixed Models, Appendix to an R and S-PLUS companion to Applied Regression"**. *Cran.R-Project docs* (Internet), disponible en: http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-mixed-models.pdf , 2002.

GHOSH, M. and RAO, J. N. K. **"Small Area Estimation: An Appraisal"**. *Statistical Science*, Vol. 9, pp. 55-93, 1994.

GÓMEZ RUBIO VIRGILIO, "**Introduction to Small Area Estimation",** Imperial College, London, UK, 2007

GONZÁLEZ, M. E. **"Use and Evaluation of Synthetic Estimates"**. *American Statistical Association: Proceedings of Social Statistics Section*, 1973.

HANSEN, M. H., HORWITZ, W. N. and MADOW, W. G. **"Sample Survey. Methods and Theory".** Vol. 1, *John Wiley & Sons*, 1953.

HANSEN, M. H., MADOW, W. G. and TEPPING, B. J. **"An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys"**. *Journal of the American Statistical Association*, Vol. 78, pp. 776-793, 1983.

HARTER, R. M. **"Small Area Estimation using nested-error and for sugar models and auxiliary data"**. *Ph. D. Theses*, Iowa State University (U.S.A), 1983.

HENDERSON, C. R **"Equivalent Linear Models to reduce computations"..** *Journal of Dairy Science*, Vol. 68, pp.2267–2277, 1985.

HENDERSON, C. R. **"Application of Linear Models in animal breeding".** *University of Guelph* (Ontario, Canada), 1984.

HENDERSON, C. R. **"Best Linear Unbiased Estimation and prediction under a Selection Model".** *Biometrics*, Vol. 31, pp. 423-449, 1975.

HENDERSON, C. R. **"Estimation of Variance and Covariance Components".** *Biometrics*, Vol. 9, pp. 226-252, 1953.

HENDERSON, C. R. **"Selection Index and Expected Genetic Advance".** Hanson, W. D. and Robinson, H. F. (Eds.), *Statistical Genetics and Plant Breeding*, No. 992, pp.141-163, 1963.

HENDERSON, C. R. **"Sire Evaluation and Genetic Trends".** *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush*, American Society of Animal Science and the American Dairy Science Association, pp.10-41, 1973.

HENDERSON, C. R. **"Specific and General Combining Ability".** Gowen, J. W (Ed.) *Heterosis*, Iowa State College Press (U.S.A.), pp. 352-370, 1950.

INEGI. **"Diccionario de datos climáticos escala 1:250 000 y 1:1 000 000 (vectorial)"**. 2000.

INEGI. **"II Conteo de Población y Vivienda 2005. Características Metodológicas y Conceptuales"**. 2006.

INEGI. **"Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH 2005). Síntesis Metodológica"**. 2006.

INSTITUTO VASCO DE ESTADÍSTICA (EUSTAT) **"Estimación de Áreas Pequeñas en la Encuesta Industrial de la C.A. de Euskadi"**. *Documentos metodológicos*, 2006.

INSTITUTO VASCO DE ESTADÍSTICA (EUSTAT) **"Cálculo de coeficientes de variación para diferentes estimadores directos e indirectos utilizados en las encuestas económicas de la Eustat"**. *Documentos metodológicos*, 2004.

LOHR S. L. **"Muestreo: Diseño y Análisis"**. *Thomson International*, 2000.

MARTÍNEZ, I. L. **"Muestreo de Áreas: Diseño de muestras y estimación en Pequeñas Áreas"**. *Tesis Doctoral*, Universidad Politécnica de Madrid (España), 1998.

OFFICE FOR NATIONAL STATISTICS **"Project EURAREA"** *Documentos metodológicos*, Unión Europea, 2005

PURCELL, N. J. and KISH, L. **"Estimation for Small Domains"**. *Biometrics*, Vol. 35, pp. 365-384, 1979.

RAMOS, Q. R. **"Apuntes de Modelos Jerárquicos"**. *Centro de Investigación en Matemáticas (CIMAT)*, 2008.

RAO, J.N.K. **"Small Area Estimation"**. *Wiley Interscience*, 2003.

SÄRNDAL, C. E. **"Design-Consistent versus Model-Dependent Estimators for Small Domains"**. *Journal of the American Statistical Association*, Vol. 79, pp. 624-631, 1984.

SÄRNDAL, C. E. and HIDIROGLOU, M. A.. **"Small Domain Estimation: A Conditional Analysis"**. *Journal of the American Statistical Association*, Vol. 84, pp. 266-275, 1989.

SÄRNDAL, C. E., SWENSSON, B. and WRETMAN, J. H. **"The Weighted Technique for estimating the variance of the General Regression Estimator of the finite population total"**. *Biometrika*, Vol. 76, pp. 527-537, 1989.

WALSH B. **"Course of Mixed Models"**. *University of Arizona*, 2003.

ANEXOS

ANEXO 1. VARIABLES EMPLEADAS EN LOS MODELOS

Mnemónico	Descripción	Valor[3]
agro	Actividades agropecuarias o forestales en la vivienda	0 No tiene 1 Si tiene
agua	Procedencia del agua en el hogar	1 Red pública dentro de la vivienda 2 Red pública fuera de la vivienda dentro del terreno 3 Llave pública o hidrante 4 Otra vivienda
agua2	Disponibilidad de agua entubada en la vivienda	0 No tiene 1 Si tiene
analfc	Analfabetas (categórico)	0 No tiene 1 Uno 2 Dos o más
anosescj	Años de escolaridad del jefe o jefa del hogar principal	De 0 al 24 años
asisescj	Asistencia escolar del jefe o jefa del hogar principal	0 No asiste actualmente a la escuela 1 Asiste actualmente a la escuela
banoagua	Servicio sanitario con agua	1 Con conexión de agua 2 Le echan agua con cubeta 3 No se le puede echar agua 4 No tiene servicio sanitario
banoagua1	Indicadora de servicio sanitario con conexión de agua	0 No tiene 1 Si tiene
banoagua2	Indicadora de servicio sanitario en la vivienda que le echan agua con cubeta	0 No tiene 1 Si tiene
banoagua3	Indicadora de servicio sanitario que no le pueden echar agua con cubeta	0 No tiene 1 Si tiene
banoagua4	Indicadora de servicio sanitario que procede de otra vivienda	0 No tiene 1 Si tiene
compu	Existe al menos una computadora en la vivienda	0 No tiene 1 Si tiene
cuador	Número cuartos para dormir en la vivienda	De 1 al 25 cuartos

[3] Estos valores se emplearon para la selección de variables en el modelo GREG con información de la ENIGH 2005 (muestra), para los modelos Sintético, Compuesto y de Áreas Pequeñas, las variables son las mismas pero calculadas como promedios ó proporciones municipales con la información del II Conteo 2005.

Mnemónico	Descripción	Valor[3]
cuadorc	Cuartos para dormir en la vivienda (categórico)	1 Cuarto 2 Dos cuartos 3 Tres cuartos 4 Cuatro cuartos 5 Cinco cuartos o más
e15a24	Número de personas de 15 a 24 años en la vivienda	De 0 a 14 personas
e15a24c	Número de personas de 15 a 24 años en la vivienda (categórico)	0 No tiene 1 Una persona 2 Dos personas 3 Tres personas 4 Cuatro o más personas
e25a49c	Número de personas de 25 a 49 años en la vivienda (categórico)	0 No tiene 1 Una persona 2 Dos personas 3 Tres o más personas
e50a64	Número de personas de 50 a 64 años en la vivienda	De 0 a 6 personas
e50a64c	Número de personas de 50 a 64 años en la vivienda (categórico)	0 No tiene 1 Una persona 2 Dos o más personas
e50a64c0	Indicadora de ninguna persona de 50 a 64 años en la vivienda	0 No tiene 1 Si tiene
e50a64c1	Indicadora de una persona de 50 a 64 años en la vivienda	0 No tiene 1 Si tiene
e50a64c2	Indicadora de dos personas de 50 a 64 años en la vivienda	0 No tiene 1 Si tiene
e65mas	Número de personas de 65 y más años en la vivienda	De 0 a 8 personas
e6a14c	Número de personas de 6 a 14 años en la vivienda (categórico)	0 No tiene 1 Una persona 2 Dos personas 3 Tres personas 4 Cuatro o más personas
edadj	Edad del jefe del hogar principal	De 10 a 97 años o más
edadj2	Edad del jefe del hogar principal al cuadrado	De 100 a 9409 años o más
edadmax	Edad máxima en la vivienda	De 0 a 97 años o más
edadmin	Edad mínima en la vivienda	De 0 a 97 años o más
edadprom	Edad promedio de los residentes de la	De 0 a 97 años

Mnemónico	Descripción	Valor[3]
	vivienda	
electri	Electricidad en la vivienda	0 No tiene 1 Si tiene
esccatj	Escolaridad categórica del jefe o jefa del hogar principal	0 Sin instrucción 1 Primaria incompleta 2 Primaria completa 3 Secundaria incompleta 4 Educación básica y más
escmax	Escolaridad máxima de los residentes de la vivienda	De 0 a 24
escmenj	Escolaridad del jefe o jefa del hogar principal relativa mediana para mayores de 5años	De -2 a 22
escpromh	Promedio de escolaridad de los residentes de la vivienda	De 0 a 22
escpronj	Escolaridad del jefe o jefa del hogar principal relativa al promedio para mayores de 5 años	De -1 a 6.23
factorM	Factor de expansión con afijación municipal	
hacina	Índice de hacinamiento en la vivienda	De 0.04 a 29
hacinai	Indicadora de hacinamiento	0 No 1 Si
hijasc	Número de hijas en la vivienda	0 Sin hijas 1 Una hija 2 Dos hijas 3 Tres hijas o más
hijo12	Número de hijos menores de 12 años en la vivienda	De 0 a 14
hijosc	Número de hijos (hombres - mujeres) en la vivienda	De 0 a 6
hijosmc	Número de hijos varones en la vivienda	0 Sin hijos 1 Un hijo 2 Dos hijos 3 Tres hijos 4 Cuatro hijos o más
hinv15	Número de hijos nacidos vivos de mujeres mayores de 15 a 45 años en la vivienda	De 0 a 30
hombres	Número de hombres en la vivienda	De 0 a 12
hombresc	Número de hombres en la vivienda (categórico)	0 No tiene 1 Uno 2 Dos

Mnemónico	Descripción	Valor[3]
		3 Tres 4 Cuatro 5 Cinco o más
hombresc0	Indicadora de ningún hombre en la vivienda	0 No tiene 1 Si tiene
hombresc1	Indicadora de un hombres en la vivienda	0 No tiene 1 Si tiene
hombresc2	Indicadora de dos hombres en la vivienda	0 No tiene 1 Si tiene
hombresc3	Indicadora de tres hombres en la vivienda	0 No tiene 1 Si tiene
hombresc4	Indicadora de cuatro hombres en la vivienda	0 No tiene 1 Si tiene
hombresc5	Indicadora de cinco hombres en la vivienda	0 No tiene 1 Si tiene
imigeu	Indicadora de cambio de residencia de algún miembro de la vivienda a Estados Unidos en el año 2000	0 No tiene 1 Si tiene
imigra	Indicadora de cambio de residencia de algún miembro de la vivienda en el año 2000	0 No tiene 1 Si tiene
lava	Lavadora en la vivienda	0 No tiene 1 Si tiene
matpiso	Material de piso en la vivienda	1 Tierra 2 Cemento o firme 3 Madera, mosaico u otro material
matpiso1	Material de piso de tierra en la vivienda	0 No tiene 1 Si tiene
matpiso2	Material de piso en la vivienda de Cemento o firme	0 No tiene 1 Si tiene
matpiso3	Material de piso en la vivienda de Madera, mosaico u otro	0 No tiene 1 Si tiene
migeuj	Indicadora de cambio de residencia del jefe o jefa del hogar a Estados Unidos en el año 2000	0 No tiene 1 Si tiene
muj12	Número de Mujeres de 12 años o más en la vivienda	De 0 a 22
muj12c	Número de Mujeres de 12 años o más en la vivienda (categórico)	0 No tiene 1 Uno 2 Dos

Mnemónico	Descripción	Valor[3]
		3 Tres 4 Cuatro 5 Cinco o más
nas1524	Número de personas en el hogar de 15 a 24 años que asisten a la escuela	De 0 a 11
nas1524c	Número de personas en el hogar de 15 a 24 años que asisten a la escuela (categórico)	0 No tiene 1 Uno 2 Dos 3 Tres
nase614	Número de personas en la vivienda de 6 a 14 años que asisten a la escuela	De 0 a 11
nase614c	Número de personas en la vivienda de 6 a 14 años que asisten a la escuela (categórico)	0 No tiene 1 Uno 2 Dos
nase614c0	Indicadora de ninguna persona en la vivienda de 6 a 14 años que asiste a la escuela	0 No tiene 1 Si tiene
nase614c1	Indicadora de Una persona en la vivienda de 6 a 14 años que asiste a la escuela	0 No tiene 1 Si tiene
nase614c2	Indicadora de Dos personas en la vivienda de 6 a 14 años que asisten a la escuela	0 No tiene 1 Si tiene
ninosc	Número de niños en la vivienda	0 No tiene 1 Uno 2 Dos 3 Tres 4 Cuatro 5 Cinco o más
nivedj	Nivel de instrucción del jefe o jefa del hogar principal	0 Sin instrucción 1 Primaria 2 Secundaria o equivalente 3 Preparatoria o equivalente 4 Profesional o equivalente
numcua	Número de cuartos en la vivienda	De 0 a 25
numcuac	Número de cuartos en la vivienda (categórico)	1 Uno 2 Dos 3 Tres 4 Cuatro 5 Cinco o más
numhog	Número de hogares en la vivienda	De 1 a 5

Mnemónico	Descripción	Valor[3]
panalf15	Proporción de analfabetas mayores de 15 años en la vivienda	De 0 a 1
pe15a24	Proporción de personas de 15 a 24 años en la vivienda	De 0 a 1
pe25a49	Proporción de personas de 25 a 49 años en la vivienda	De 0 a 1
pe50a64	Proporción de personas de 50 a 64 años en la vivienda	De 0 a 1
pe65mas	Proporción de personas de 65 años o más en la vivienda	De 0 a 1
pemedn	Promedio escolar relativo a la mediana quinquenal nacional	De -2 a 22
peprom5n	Promedio escolar relativo al promedio quinquenal nacional	De 0 a 7.23
phijo12	Proporción de hijos menores de 12 años en la vivienda	De 0 a 1
phimue12	Promedio de hijos muertos de mujeres mayores de 12 años en la vivienda	De 0 a 20
phinv12	Promedio de hijos vivos de mujeres mayores de 12 años en la vivienda	De 0 a 25
phog	Población total en la vivienda	De 0 a 37
pmuj12	Proporción de mujeres de 12 años y más en la vivienda	De 0 a 1
pp15pinc	Proporción de personas de 15 años o más con primaria incompleta	De 0 a 1
pp15pri	Proporción de personas de 15 años o más que solamente completó la primaria	De 0 a 1
refri	Refrigerador en la vivienda	0 No tiene 1 Si tiene
rezaeduj	Rezago educativo del jefe o jefa del hogar principal	De 0 a 1
rezahaci	Rezago en disponibilidad de espacio en la vivienda (definición FAIS)	De 0 a 1
rezinegi	Indicadora de rezago educativo en personas de 15 a 49 años (definición INEGI)	De 0 a 1
sinbien	No dispone de televisión, refrigerador, lavadora ni computadora	0 con al menos un bien 1 Sin bien
tasesch	Total de personas en la vivienda que asisten a la escuela	De 0 a 17
taseschc	Total de personas de 15 años o más que	0 No tiene

Mnemónico	Descripción	Valor[3]
	asisten a la escuela (categórica)	1 Uno 2 Dos 3 Tres 4 Cuatro o más
taseschc0	Indicadora de ninguna persona de 15 años o más en la vivienda que asiste a la escuela	0 No tiene 1 Si tiene
taseschc1	Indicadora de una persona en la vivienda que asiste a la escuela	0 No tiene 1 Si tiene
taseschc2	Indicadora de dos personas en la vivienda que asisten a la escuela	0 No tiene 1 Si tiene
taseschc3	Indicadora de tres personas en la vivienda que asisten a la escuela	0 No tiene 1 Si tiene
taseschc4	Indicadora de cuatro personas en la vivienda que asisten a la escuela	0 No tiene 1 Si tiene
tele	Televisión en la vivienda	0 No tiene 1 Si tiene
trabdomi	Trabajadores domésticos en la vivienda	0 No tiene 1 Si tiene

ANEXO 2. TABLA PARA SELECCIÓN DE VARIABLES DEL MODELO GREG

modelo	Deviance	prueba P(>|Chi|)	modelo	Deviance	prueba P(>|Chi|)
1	**0.19010**	**0.54323**	39	0.00034	0.97359
2	**0.13479**	**0.55119**	40	0.00006	0.98848
3	**0.01112**	**0.86209**	**41**	**0.00304**	**0.92055**
4	**0.00377**	**0.91900**	42	0.00023	0.97800
5	0.00104	0.95732	43	0.00002	0.99348
6	0.00001	0.99546	44	0.00001	0.99553
7	0.00060	0.96756	45	0.00062	0.96398
8	0.00000	0.99819	46	0.00005	0.98993
9	0.00009	0.98722	**47**	**0.00302**	**0.92027**
10	0.00282	0.92945	48	0.00000	0.99860
11	0.00303	0.92655	49	0.00000	0.99794
12	0.00086	0.96075	50	0.00011	0.98458
13	0.00007	0.98846	51	0.00045	0.96923
14	**0.00859**	**0.87508**	52	0.00000	0.99782
15	0.00022	0.97974	53	0.00034	0.97312
16	**0.00205**	**0.93854**	54	0.00018	0.98061
17	**0.00491**	**0.90441**	55	0.00029	0.97540
18	0.00000	0.99772	56	0.00030	0.97491
19	0.00094	0.95798	57	0.00020	0.97923
20	**0.00174**	**0.94275**	58	0.00012	0.98398
21	**0.00364**	**0.91684**	59	0.00001	0.99565
22	**0.00254**	**0.93027**	60	0.00001	0.99564
23	**0.00472**	**0.90430**	61	0.00005	0.98957
24	**0.00245**	**0.93070**	62	0.00001	0.99564
25	0.00094	0.95696	63	0.00013	0.98369
26	**0.00171**	**0.94195**	64	0.00098	0.95427
27	0.00029	0.97596	65	0.00019	0.97965
28	**0.00544**	**0.89562**	66	0.00059	0.96450
29	0.00010	0.98546	67	0.00063	0.96325
30	0.00011	0.98533	68	0.00000	0.99803
31	0.00010	0.98571	69	0.00004	0.99104
32	**0.00143**	**0.94623**	70	0.00006	0.98882
33	0.00001	0.99549	71	0.00077	0.95936
34	0.00079	0.96005	72	0.00064	0.96285
35	**0.00333**	**0.91740**	73	0.00044	0.96919
36	0.00048	0.96863	74	0.00002	0.99277
37	0.00043	0.97015	**75**	**0.00120**	**0.94897**
38	0.00018	0.98050	76	0.00000	0.99727

ANEXO 3. TABLAS DE COLINEALIDAD

TABLA DE COLINEALIDAD DE LAS VARIABLES DEL MODELO GREG

```
> colldiag(fitfplm)
Condition
Index                    Variance Decomposition Proportions
           intercept escmax numcua compu matpiso1 matpiso2 banoagua2 banoagua3 lava  trabdomi
1      1.000 0.000     0.001  0.001  0.002 0.001    0.003     0.001     0.001     0.002 0.000
2      2.314 0.000     0.001  0.002  0.059 0.097    0.004     0.023     0.108     0.006 0.010
3      2.679 0.000     0.001  0.000  0.020 0.033    0.008     0.011     0.021     0.000 0.031
4      2.745 0.000     0.000  0.000  0.009 0.028    0.023     0.202     0.046     0.000 0.058
5      2.836 0.000     0.000  0.000  0.004 0.010    0.005     0.005     0.005     0.000 0.015
6      2.841 0.000     0.000  0.000  0.000 0.011    0.001     0.081     0.002     0.000 0.153
7      2.876 0.000     0.000  0.000  0.001 0.008    0.003     0.010     0.002     0.000 0.375
8      2.942 0.000     0.000  0.000  0.001 0.011    0.004     0.077     0.014     0.000 0.242
9      3.025 0.000     0.000  0.000  0.009 0.021    0.011     0.288     0.036     0.000 0.055
10     3.070 0.000     0.000  0.000  0.040 0.106    0.080     0.006     0.008     0.000 0.044
11     3.485 0.000     0.000  0.000  0.011 0.012    0.051     0.015     0.076     0.000 0.001
12     3.978 0.000     0.000  0.000  0.373 0.046    0.020     0.132     0.288     0.007 0.001
13     5.255 0.000     0.001  0.003  0.162 0.332    0.521     0.128     0.336     0.051 0.000
14     5.891 0.000     0.003  0.003  0.165 0.088    0.058     0.000     0.001     0.186 0.001
15     6.900 0.002     0.024  0.012  0.064 0.038    0.049     0.002     0.010     0.581 0.002
16     8.499 0.001     0.030  0.221  0.034 0.021    0.028     0.003     0.000     0.093 0.002
17     9.516 0.000     0.061  0.067  0.019 0.001    0.001     0.002     0.000     0.020 0.005
18    10.700 0.001     0.401  0.556  0.003 0.000    0.002     0.001     0.010     0.006 0.003
19    10.768 0.000     0.271  0.001  0.020 0.005    0.014     0.000     0.004     0.046 0.000
20    24.226 0.997     0.204  0.133  0.004 0.132    0.115     0.016     0.032     0.002 0.003

                         Variance Decomposition Proportions
   pp15pri refri taseschc0 imigeu hombresc0 hombresc1 hombresc2 hombresc3 e50a64c0 e50a64c1
1  0.003   0.001 0.003     0.000  0.001     0.001     0.001     0.001     0.001     0.001
2  0.030   0.000 0.003     0.002  0.000     0.000     0.000     0.001     0.000     0.000
3  0.036   0.000 0.003     0.048  0.047     0.032     0.023     0.036     0.003     0.027
4  0.000   0.000 0.000     0.041  0.049     0.003     0.001     0.000     0.008     0.068
5  0.007   0.000 0.000     0.403  0.143     0.032     0.001     0.004     0.000     0.010
6  0.001   0.000 0.000     0.025  0.009     0.000     0.052     0.114     0.003     0.025
7  0.011   0.000 0.000     0.098  0.035     0.010     0.039     0.047     0.002     0.011
8  0.001   0.000 0.001     0.096  0.147     0.004     0.003     0.000     0.009     0.058
9  0.003   0.000 0.002     0.096  0.004     0.012     0.000     0.053     0.008     0.061
10 0.043   0.000 0.000     0.182  0.008     0.035     0.026     0.008     0.001     0.002
11 0.761   0.000 0.002     0.001  0.010     0.002     0.002     0.000     0.000     0.013
12 0.048   0.001 0.044     0.000  0.008     0.000     0.008     0.013     0.000     0.005
13 0.002   0.002 0.000     0.001  0.000     0.000     0.001     0.000     0.000     0.002
14 0.000   0.003 0.595     0.000  0.001     0.000     0.010     0.018     0.001     0.000
15 0.001   0.002 0.209     0.003  0.012     0.031     0.032     0.025     0.092     0.061
16 0.001   0.009 0.001     0.001  0.004     0.000     0.000     0.000     0.545     0.471
17 0.001   0.000 0.068     0.000  0.389     0.622     0.617     0.530     0.113     0.066
18 0.011   0.110 0.002     0.000  0.000     0.003     0.015     0.009     0.087     0.033
19 0.012   0.695 0.007     0.001  0.037     0.037     0.025     0.019     0.000     0.000
20 0.028   0.177 0.061     0.001  0.095     0.175     0.144     0.122     0.126     0.085
```

TABLA DE COLINEALIDAD DE LAS VARIABLES DEL MODELO SINTÉTICO

```
Condition
Index    Variance Decomposition Proportions
           intercept escpromh migeuj_1 hijo12c1 e15a24_3 banoagua3 e50a64_1
1      1.000 0.000     0.000    0.004    0.000    0.000    0.001     0.000
2      2.893 0.000     0.000    0.167    0.000    0.000    0.057     0.000
3      5.237 0.000     0.001    0.475    0.001    0.000    0.131     0.002
4     13.538 0.002     0.001    0.138    0.033    0.079    0.020     0.072
5     31.540 0.005     0.053    0.003    0.174    0.856    0.547     0.087
6     42.735 0.001     0.481    0.034    0.648    0.024    0.144     0.145
7     80.938 0.992     0.464    0.178    0.145    0.041    0.099     0.693
```

ANEXO 4. MATRICES DE CORRELACIÓN

MATRIZ DE CORRELACIONES DE LAS VARIABLES DEL MODELO GREG

	escmax	numcua	compu	matpiso1	matpiso2	banoagua2	banoagua3	lava	trabdomi	pp15pri
escmax	1.000	0.408	0.485	-0.238	-0.272	-0.152	-0.249	0.324	0.077	-0.269
numcua	0.408	1.000	0.412	-0.306	-0.271	-0.180	-0.339	0.371	0.117	-0.102
compu	0.485	0.412	1.000	-0.176	-0.273	-0.126	-0.209	0.275	0.066	-0.176
matpiso1	-0.238	-0.306	-0.176	1.000	-0.272	0.117	0.371	-0.293	-0.016	0.060
matpiso2	-0.272	-0.271	-0.273	-0.272	1.000	0.122	0.134	-0.154	-0.046	0.110
banoagua2	-0.152	-0.180	-0.126	0.117	0.122	1.000	-0.104	-0.125	-0.014	0.039
banoagua3	-0.249	-0.339	-0.209	0.371	0.134	-0.104	1.000	-0.313	-0.020	0.091
lava	0.324	0.371	0.275	-0.293	-0.154	-0.125	-0.313	1.000	0.010	-0.104
trabdomi	0.077	0.117	0.066	-0.016	-0.046	-0.014	-0.020	0.010	1.000	-0.021
pp15pri	-0.269	-0.102	-0.176	0.060	0.110	0.039	0.091	-0.104	-0.021	1.000
refri	0.179	0.240	0.152	-0.331	-0.014	-0.127	-0.183	0.332	0.016	-0.054
taseschc0	-0.308	-0.203	-0.325	0.084	0.100	0.057	0.057	-0.146	-0.013	0.099
imigeu	0.044	0.054	0.032	-0.011	-0.031	-0.027	-0.006	0.051	-0.005	0.002
hombresc0	-0.095	0.015	-0.034	-0.026	-0.047	-0.039	-0.053	-0.022	-0.015	0.032
hombresc1	-0.123	-0.012	-0.068	-0.042	-0.003	0.025	-0.042	-0.064	-0.015	0.039
hombresc2	0.108	-0.001	0.050	0.012	-0.002	0.015	-0.005	0.036	0.011	-0.075
hombresc3	0.039	-0.024	0.008	0.036	0.019	-0.013	0.057	0.028	0.003	0.003
e50a64c0	-0.009	-0.153	-0.024	0.016	0.047	0.037	-0.016	-0.016	-0.009	-0.071
e50a64c1	-0.006	0.098	0.009	-0.014	-0.023	-0.023	0.014	-0.003	0.025	0.054

	refri	taseschc0	imigeu	hombresc0	hombresc1	hombresc2	hombresc3	e50a64c0	e50a64c1
escmax	0.179	-0.308	0.044	-0.095	-0.123	0.108	0.039	-0.009	-0.006
numcua	0.240	-0.203	0.054	0.015	-0.012	-0.001	-0.024	-0.153	0.098
compu	0.152	-0.325	0.032	-0.034	-0.068	0.050	0.008	-0.024	0.009
matpiso1	-0.331	0.084	-0.011	-0.026	-0.042	0.012	0.036	0.016	-0.014
matpiso2	-0.014	0.100	-0.031	-0.047	-0.003	-0.002	0.019	0.047	-0.023
banoagua2	-0.127	0.057	-0.027	-0.039	0.025	0.015	-0.013	0.037	-0.023
banoagua3	-0.183	0.057	-0.006	-0.053	-0.042	-0.005	0.057	-0.016	0.014
lava	0.332	-0.146	0.051	-0.022	-0.064	0.036	0.028	-0.016	-0.003
trabdomi	0.016	-0.013	-0.005	-0.015	-0.015	0.011	0.003	-0.009	0.025
pp15pri	-0.054	0.099	0.002	0.032	0.039	-0.075	0.003	-0.071	0.054
refri	1.000	-0.073	0.011	0.025	-0.027	0.019	-0.013	0.004	-0.017
taseschc0	-0.073	1.000	0.040	0.099	0.150	-0.053	-0.080	0.022	-0.006
imigeu	0.011	0.040	1.000	0.012	0.057	-0.030	-0.032	-0.014	0.004
hombresc0	0.025	0.099	0.012	1.000	-0.210	-0.195	-0.138	0.011	0.067
hombresc1	-0.027	0.150	0.057	-0.210	1.000	-0.477	-0.336	-0.029	0.021
hombresc2	0.019	-0.053	-0.030	-0.195	-0.477	1.000	-0.312	0.007	-0.032
hombresc3	-0.013	-0.080	-0.032	-0.138	-0.336	-0.312	1.000	0.055	-0.044
e50a64c0	0.004	0.022	-0.014	0.011	-0.029	0.007	0.055	1.000	-0.736
e50a64c1	-0.017	-0.006	0.004	0.067	0.021	-0.032	-0.044	-0.736	1.000

MATRIZ DE CORRELACIONES DE LAS VARIABLES DEL MODELO SINTÉTICO

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	1.0000000	0.28182085	0.45987688	-0.13714634	-0.66701017	-0.61339356
[2,]	0.2818208	1.00000000	0.35966504	-0.08577914	-0.44183800	-0.57377013
[3,]	0.4598769	0.35966504	1.00000000	0.53154450	-0.05827526	-0.68312437
[4,]	-0.1371463	-0.08577914	0.53154450	1.00000000	0.66199841	-0.09643757
[5,]	-0.6670102	-0.44183800	-0.05827526	0.66199841	1.00000000	0.43914334
[6,]	-0.6133936	-0.57377013	-0.68312437	-0.09643757	0.43914334	1.00000000

ANEXO 5. CÓDIGO R

```
################################################################################
#
#    ESTIMACION DEL PROMEDIO DEL INGRESO TOTAL DE LAS VIVIENDAS POR MUNICIPIO
#                          PARA EL ESTADO DE SONORA
#    UTILIZANDO LAS SIGUIENTES METODOLOGÍAS:
#    - ESTIMACIÓN DIRECTA
#    - MODELO GREG A NIVEL ESTATAL CON UNIDAES MUESTRALES
#    - MODELO SINTÉTICO CON PROMEDIOS MUNICIPALES
#    - MODELO COMPUESTO
#    - MODELO DE ÁREA PEQUEÑA CON INFORMACIÓN AUXILIAR A NIVEL ÁREA
#
################################################################################
#
#         R versión 2.9.2
#
################################################################################
#
#     LIBRERÍAS:  SURVEY_3.6-13
#                 PERTURB_2.03
################################################################################

rm(list=ls())
library(survey)
# LECTURA DE DATOS
datos<-read.csv("D:\\pruebastesis\\viv26.csv",head=T)
attach(datos)
dim(datos)

# ESTATAL SONORA

# ESTIMACION DIRECTA
dstrat<-svydesign(id=~upm,strata=~est, weights=~factor, data=datos, nest=TRUE)
medtHT<-svymean(sum_ingtot,design=dstrat,deff=T,se=T)
coef(medtHT)
SE(medtHT)
cvt<-cv(svymean(sum_ingtot,design=dstrat))
cvt
lmedtHT<-svymean(log(sum_ingtot),design=dstrat)
coef(lmedtHT)
SE(lmedtHT)

# MUNICIPAL SONORA
# AFIJACIÓN DE LA MUESTRA MUNICIPAL
Na<-tapply(factor,municipio,sum)
na<-as.vector(table(municipio))
munson<-as.integer(names(table(municipio)))
Nconteo<-read.csv("D:\\pruebastesis\\vivconteo.csv",head=T)
Nf<-Nconteo[munson,1]/Na
factorMun<-rep(0,dim(datos)[2])
j<-0
for(i in munson){
   j=j+1
   factorMun[municipio==i]<-factor[municipio==i]*Nf[j]
}
datos$factorM<-factorMun
dstrat<-svydesign(id=~upm,strata=~est, weights=~factorM, data=datos, nest=TRUE)
detach(datos)
attach(datos)
mmHT<-svyby(sum_ingtot,municipio,dstrat,svymean,deff=T,se=T)
medmHT<-coef(mmHT)
Na<-tapply(factorM,municipio,sum)

#  CALCULO DEL ERROR CUADRÁTICO MEDIO ESTATAL
Ymed<-rep(medmHT,na)
sqrt(sum(factorM*(factorM-1)*(sum_ingtot-Ymed)^2)/sum(Na)^2)

#  CALCULO DEL ERROR CUADRÁTICO MEDIO MUNICIPAL
```

```
ecmHT<-tapply(factorM*(factorM-1)*(sum_ingtot-Ymed)^2,municipio,sum)/Na^2
round(sqrt(ecmHT),1)
CVHT<-sqrt(ecmHT)/medmHT
round(CVHT,3)
minHT<-round(as.vector(medmHT)-1.96*sqrt(ecmHT),0)
maxHT<-round(as.vector(medmHT)+1.96*sqrt(ecmHT),0)

# PROMEDIOS LOGARITMICOS
medmHTl<-coef(svyby(log(sum_ingtot),municipio,dstrat,svymean))
Vad<-(SE(svyby(log(sum_ingtot),municipio,dstrat,svymean)))^2

# DENSISDAD DEL INGRESO POR VIVIENDA PARA SONORA
hist(sum_ingtot[sum_ingtot<200000],freq=F,breaks=20,
     xlab='Ingreso por vivenda',ylab='densidad',
     main='Densidad del Ingreso para Sonora ENIGH 2005*')
curve(dgamma(x,shape=2.6,scale=10300),add=T,col='red')

##############################################################################
# OBTENCION DEL MODELO PARA LA ESTIMACIÓN GREG
##############################################################################
# SELECCIÓN DE VARIABLES
#
c1<-round(cor(cbind(sum_ingtot,(datos[,-c(1:7,56,74,86,87)]))),2)[1,]
or<-order(abs(round(cor(cbind(sum_ingtot,(datos[,-c(1:7,56,74,86,87)]))),2)[1,]))
nva<-(c1[or])[(length(c1)-1):1]
nva
dev<-rep(0,length(nva))
prueba<-rep(0,length(nva))
fit1<-svyglm(log(sum_ingtot)~1,design=dstrat,family=gaussian(link=identity))
fit2<-svyglm(log(sum_ingtot)~numcua, design=dstrat, family=gaussian(link=identity))
AA<-anova(fit1,fit2,test='Chisq')
dev[1]<-AA$Deviance[2]
prueba[1]<-AA$P[2]
fit1<-svyglm(log(sum_ingtot)~numcua, design=dstrat, family=gaussian(link=identity))
fit2<-svyglm(log(sum_ingtot)~numcua+escmax, design=dstrat, family=gaussian(link=identity))
AA<-anova(fit1,fit2,test='Chisq')
dev[2]<-AA$Deviance[2]
prueba[2]<-AA$P[2]
base<-paste("log(sum_ingtot)~",names(nva)[1])
for(i in 3:length(nva)){
    fit1<-svyglm(as.formula(paste(base,names(nva)[i-
1],sep="+")),design=dstrat,family=gaussian(link=identity))
    base<-paste(base,names(nva)[i-1],sep="+")
    fit2<-
svyglm(as.formula(paste(base,names(nva)[i],sep="+")),design=dstrat,family=gaussian(link=iden
tity))
    AA<-anova(fit1,fit2,test='Chisq')
    dev[i]<-AA$Deviance[2]
    prueba[i]<-AA$P[2]}
res1<-round(cbind(1:length(nva),dev,prueba),5)
res1
sel1<-res1[prueba<.95,][,1]
dev1<-rep(0,length(sel1))
prueba1<-rep(0,length(sel1))
sel1b<-sel1[-c(1,2)]
dev1[c(1,2)]<-dev[c(1,2)]
prueba1[c(1,2)]<-prueba[c(1,2)]
base<-paste("log(sum_ingtot)~",names(nva)[1])
j<-2
for(i in sel1b){
    fit1<-svyglm(as.formula(paste(base,names(nva)[i-
1],sep="+")),design=dstrat,family=gaussian(link=identity))
    base<-paste(base,names(nva)[i-1],sep="+")
    fit2<-
svyglm(as.formula(paste(base,names(nva)[i],sep="+")),design=dstrat,family=gaussian(link=iden
tity))
    AA<-anova(fit1,fit2,test='Chisq')
    j<-j+1
    dev1[j]<-AA$Deviance[2]
    prueba1[j]<-AA$P[2]}
res2<-round(cbind(sel1,dev1,prueba1),5)
```

```
res2
sel2<-res2[prueba1<.95,][,1]
sel2
names(nva)[sel2]
varsel<-cbind(numcua,escmax,compu,escpromh,cuadorc,cuador,matpiso,banoagua,
        lava,trabdomi,hacina,pp15pri,refri,taseschc,imigeu,hombresc,e50a64c)
cor(varsel)

# SE ELIMINAN cuador y cuadorc POR CORRELACIÓN CON numcua
# TAMBIÉN SE ELIMINA escpromh POR CORRELACIÓN CON escmax

fit<-svyglm(log(sum_ingtot)~numcua+escmax+compu+matpiso+banoagua+lava+trabdomi+
            hacina+pp15pri+refri+taseschc+imigeu+hombresc+e50a64c+nase614c,
            design=dstrat,family=gaussian(link=identity))
summary(fit)

#########################
#   CREACIÓN DUMMYS     #
#########################
matpiso1<-matpiso
matpiso1[matpiso==2]<-0
matpiso1[matpiso==3]<-0
matpiso2<-matpiso
matpiso2[matpiso==1]<-0
matpiso2[matpiso==2]<-1
matpiso2[matpiso==3]<-0
matpiso3<-matpiso
matpiso3[matpiso==1]<-0
matpiso3[matpiso==2]<-0
matpiso3[matpiso==3]<-1
datos$matpiso1<-matpiso1
datos$matpiso2<-matpiso2
datos$matpiso3<-matpiso3
taseschc0<-taseschc
taseschc0[taseschc==0]<-1
taseschc0[taseschc==1]<-0
taseschc0[taseschc==2]<-0
taseschc0[taseschc==3]<-0
taseschc0[taseschc==4]<-0
taseschc1<-taseschc
taseschc1[taseschc==0]<-0
taseschc1[taseschc==2]<-0
taseschc1[taseschc==3]<-0
taseschc1[taseschc==4]<-0
taseschc2<-taseschc
taseschc2[taseschc==0]<-0
taseschc2[taseschc==1]<-0
taseschc2[taseschc==2]<-1
taseschc2[taseschc==3]<-0
taseschc2[taseschc==4]<-0
taseschc3<-taseschc
taseschc3[taseschc==0]<-0
taseschc3[taseschc==1]<-0
taseschc3[taseschc==2]<-0
taseschc3[taseschc==3]<-1
taseschc3[taseschc==4]<-0
taseschc4<-taseschc
taseschc4[taseschc==0]<-0
taseschc4[taseschc==1]<-0
taseschc4[taseschc==2]<-0
taseschc4[taseschc==3]<-0
taseschc4[taseschc==4]<-1
datos$taseschc0<-taseschc0
datos$taseschc1<-taseschc1
datos$taseschc2<-taseschc2
datos$taseschc3<-taseschc3
datos$taseschc4<-taseschc4
e50a64c0<-e50a64c
e50a64c0[e50a64c==0]<-1
e50a64c0[e50a64c==1]<-0
e50a64c0[e50a64c==2]<-0
```

```
e50a64c1<-e50a64c
e50a64c1[e50a64c==0]<-0
e50a64c1[e50a64c==2]<-0
e50a64c2<-e50a64c
e50a64c2[e50a64c==0]<-0
e50a64c2[e50a64c==1]<-0
e50a64c2[e50a64c==2]<-1
e50a64c01<-e50a64c
e50a64c01[e50a64c==0]<-1
e50a64c01[e50a64c==1]<-1
e50a64c01[e50a64c==2]<-0
datos$e50a64c0<-e50a64c0
datos$e50a64c1<-e50a64c1
datos$e50a64c2<-e50a64c2
datos$e50a64c01<-e50a64c01
hombresc0<-hombresc
hombresc0[hombresc==0]<-1
hombresc0[hombresc==1]<-0
hombresc0[hombresc==2]<-0
hombresc0[hombresc==3]<-0
hombresc0[hombresc==4]<-0
hombresc0[hombresc==5]<-0
hombresc1<-hombresc
hombresc1[hombresc==0]<-0
hombresc1[hombresc==2]<-0
hombresc1[hombresc==3]<-0
hombresc1[hombresc==4]<-0
hombresc1[hombresc==5]<-0
hombresc2<-hombresc
hombresc2[hombresc==0]<-0
hombresc2[hombresc==1]<-0
hombresc2[hombresc==2]<-1
hombresc2[hombresc==3]<-0
hombresc2[hombresc==4]<-0
hombresc2[hombresc==5]<-0
hombresc3<-hombresc
hombresc3[hombresc==0]<-0
hombresc3[hombresc==1]<-0
hombresc3[hombresc==2]<-0
hombresc3[hombresc==3]<-1
hombresc3[hombresc==4]<-0
hombresc3[hombresc==5]<-0
hombresc4<-hombresc
hombresc4[hombresc==0]<-0
hombresc4[hombresc==1]<-0
hombresc4[hombresc==2]<-0
hombresc4[hombresc==3]<-0
hombresc4[hombresc==4]<-1
hombresc4[hombresc==5]<-0
hombresc5<-hombresc
hombresc5[hombresc==0]<-0
hombresc5[hombresc==1]<-0
hombresc5[hombresc==2]<-0
hombresc5[hombresc==3]<-0
hombresc5[hombresc==4]<-0
hombresc5[hombresc==5]<-1
datos$hombresc0<-hombresc0
datos$hombresc1<-hombresc1
datos$hombresc2<-hombresc2
datos$hombresc3<-hombresc3
datos$hombresc4<-hombresc4
datos$hombresc5<-hombresc5
banoagua1<-banoagua
banoagua1[banoagua==2]<-0
banoagua1[banoagua==3]<-0
banoagua1[banoagua==4]<-0
banoagua2<-banoagua
banoagua2[banoagua==1]<-0
banoagua2[banoagua==2]<-1
banoagua2[banoagua==3]<-0
banoagua2[banoagua==4]<-0
```

```
banoagua3<-banoagua
banoagua3[banoagua==1]<-0
banoagua3[banoagua==2]<-0
banoagua3[banoagua==3]<-1
banoagua3[banoagua==4]<-0
banoagua4<-banoagua
banoagua4[banoagua==1]<-0
banoagua4[banoagua==2]<-0
banoagua4[banoagua==3]<-0
banoagua4[banoagua==4]<-1
datos$banoagua1<-banoagua1
datos$banoagua2<-banoagua2
datos$banoagua3<-banoagua3
datos$banoagua4<-banoagua4
nase614c0<-nase614c
nase614c0[nase614c==0]<-1
nase614c0[nase614c==1]<-0
nase614c0[nase614c==2]<-0
nase614c1<-nase614c
nase614c1[nase614c==0]<-0
nase614c1[nase614c==2]<-0
nase614c2<-nase614c
nase614c2[nase614c==0]<-0
nase614c2[nase614c==1]<-0
nase614c2[nase614c==2]<-1
datos$nase614c0<-nase614c0
datos$nase614c1<-nase614c1
datos$nase614c2<-nase614c2
detach(datos)
attach(datos)
dstrat<-svydesign(id=~upm,strata=~est, weights=~factorM, data=datos, nest=TRUE)

######################
# MODELO GREG FINAL #
######################

#  SE ELIMINAN AQUELLAS VARIABLES QUE NO AJUSTA AL MODELO, COMENZANDO POR
#  LA DE MENOS AJUSTE HASTA QUE SE LLEGO AL SIGUIENTE MODELO

fitf<-svyglm(log(sum_ingtot)~escmax+numcua+compu+matpiso1+matpiso2+
   banoagua2+banoagua3+lava+trabdomi+pp15pri+refri+taseschc0+imigeu+
   hombresc0+hombresc1+hombresc2+hombresc3+e50a64c0+e50a64c1
   ,design=dstrat,family=gaussian(link=identity))
summary(fitf)

# MATRIZ DE CORRELACIONES
round(cor(cbind(escmax,numcua,compu,matpiso1,matpiso2,banoagua2,banoagua3,
     lava,trabdomi,pp15pri,refri,taseschc0,imigeu,hombresc0,hombresc1,
     hombresc2,hombresc3,e50a64c0,e50a64c1)),3)

# ANALISIS DE COLINELIDAD
library(perturb)
fitflm<-lm(log(sum_ingtot)~escmax+numcua+compu+matpiso1+matpiso2+
   banoagua2+banoagua3+lava+trabdomi+pp15pri+refri+taseschc0+imigeu+
   hombresc0+hombresc1+hombresc2+hombresc3+e50a64c0+e50a64c1,weights=factorM)
summary(fitflm)
colldiag(fitflm)

# ANALISIS DE RESIDUALES
svyhist(~residuals(fitf),design=dstrat,main="Residuos ponderados",col="yellow")
qqnorm(residuals(fitf),main="Gráfica de Normalidad",xlab="Cuantiles teóricos",
       ylab="Cuantiles residuales"  )
qqline(residuals(fitf),col="red")
shapiro.test(residuals(fitf))

svyplot(residuals(fitf)~fitted.values(fitf),design=dstrat,xlab='Ajustados',
        ylab='residuos',main="Residuales vs valores ajustados")
abline(0,0,col='blue',lty=2,)
lines(lowess(residuals(fitf)~fitted.values(fitf)),col="red")
par(mfrow=c(2,2))
svyplot(residuals(fitf)~escmax,design=dstrat,xlab='Escolaridad Máxima',ylab='Residuos')
```

```
abline(0,0,col='blue',lty=2)
lines(lowess(residuals(fitf)~escmax),col="red")
svyplot(residuals(fitf)~numcua,design=dstrat,xlab='Cuartos',ylab='')
abline(0,0,col='blue',lty=2)
lines(lowess(residuals(fitf)~numcua),col="red")
svyplot(residuals(fitf)~pp15pri,design=dstrat,xlab='Prop.   15   años   y   más   con   sólo
primaria',ylab='')
abline(0,0,col='blue',lty=2)
lines(lowess(residuals(fitf)~pp15pri),col="red")

# ESTIMACIÓN CON EL MODELO GREG
Datxmunt<-read.csv("D:\\pruebastesis\\avgmun26.csv",head=T)
Nommun<-Datxmunt[,2]
Datxmun<-Datxmunt[munson,]
pp1<-predict(fitf,data.frame(INTERCEP = 1,escmax=Datxmun[,93],
 numcua=Datxmun[,55],compu=Datxmun[,79],matpiso1=Datxmun[,46],matpiso2=Datxmun[,47],
 banoagua2=Datxmun[,83],banoagua3=Datxmun[,84],lava=Datxmun[,77],trabdomi=Datxmun[,166],
 pp15pri=Datxmun[,154],refri=Datxmun[,75],taseschc0=Datxmun[,160],imigeu=Datxmun[,120],
 hombresc0=Datxmun[,113],hombresc1=Datxmun[,114],hombresc2=Datxmun[,115],
 hombresc3=Datxmun[,116],e50a64c0=Datxmun[,21],e50a64c1=Datxmun[,22]))
p1<-as.vector(pp1)
p2<-as.vector(fitted(fitf))
p2f1<-exp(p2)*factorM
p2f<-as.vector(by(p2f1,municipio,sum))/Na
medGREG<-medmHT+(exp(p1)-p2f)
round(medGREG,0)

# CÁLCULO DEL ECM GREG
xa<-cbind(escmax,numcua,compu,matpiso1,matpiso2,banoagua2,banoagua3,lava,trabdomi,
    pp15pri,refri,taseschc0,imigeu,hombresc0,hombresc1,hombresc2,hombresc3,e50a64c0,
    e50a64c1)
Xabarra<-cbind(Datxmun[,93],Datxmun[,55],Datxmun[,79],Datxmun[,46],Datxmun[,47],
    Datxmun[,83],Datxmun[,84],Datxmun[,77],Datxmun[,166],Datxmun[,154],Datxmun[,75],
    Datxmun[,160],Datxmun[,120],Datxmun[,113],Datxmun[,114],Datxmun[,115],
    Datxmun[,116],Datxmun[,21],Datxmun[,22])
xabarra<-
cbind(coef(svyby(escmax,municipio,dstrat,svymean)),coef(svyby(numcua,municipio,dstrat,svymea
n)),
 coef(svyby(compu,municipio,dstrat,svymean)),coef(svyby(matpiso1,municipio,dstrat,svymean)),

coef(svyby(matpiso2,municipio,dstrat,svymean)),coef(svyby(banoagua2,municipio,dstrat,svymean
)),
 coef(svyby(banoagua3,municipio,dstrat,svymean)),coef(svyby(lava,municipio,dstrat,svymean)),

coef(svyby(trabdomi,municipio,dstrat,svymean)),coef(svyby(pp15pri,municipio,dstrat,svymean))
,

coef(svyby(refri,municipio,dstrat,svymean)),coef(svyby(taseschc0,municipio,dstrat,svymean)),

coef(svyby(imigeu,municipio,dstrat,svymean)),coef(svyby(hombresc0,municipio,dstrat,svymean))
,

coef(svyby(hombresc1,municipio,dstrat,svymean)),coef(svyby(hombresc2,municipio,dstrat,svymea
n)),

coef(svyby(hombresc3,municipio,dstrat,svymean)),coef(svyby(e50a64c0,municipio,dstrat,svymean
)),
 coef(svyby(e50a64c1,municipio,dstrat,svymean)))
rmseGREG<-rep(0,length(na))
j<-0
for(i in munson){
  xa1<-xa[municipio==i,]
  fac1<-factorM[municipio==i]
  j=j+1
  Xabarra1<-Xabarra[j,]
  xabarra1<-xabarra[j,]
  denom<-rep(0,length(xabarra1))
  for(k in 1:length(fac1)){
      denom<-denom+diag(fac1[k]*(xa1[k,]%*%t(xa1[k,]))) }
  g<-rep(0,length(fac1))
  for(k in 1:length(fac1)){
```

```
      gii<-xa1[k,]/denom
      gii[is.na(gii)]<-0
      g[k]<-1+t(Xabarra1-xabarra1)%*%gii }
  rmseGREG[j]<-sqrt((sum(fac1*(fac1-
1)*fitf$resid[municipio==i]^2*g^2)/Na[j]^2)*medGREG[j]^2)
}
round(rmseGREG,1)
CVGREG<-rmseGREG/medGREG
round(CVGREG,3)      # coeficiente de variación GREG
minGREG<-round(as.vector(medGREG)-1.96*rmseGREG,0)
maxGREG<-round(as.vector(medGREG)+1.96*rmseGREG,0)


############################
# ESTIMACION SINTÉTICA     #
############################
yHT<-log(as.vector(medmHT))
Datxmunt<-read.csv("D:\\pruebastesis\\avgmun26.csv",head=T)
Datxmun<-Datxmunt[munson,]
dim(Datxmun)
detach(datos)
attach(Datxmun)
Datxmun[1,]

# SELECCION DE VARIABLES
c2<-round(cor(cbind(yHT,(Datxmun[,-c(1:2)]))),2)[1,]
or<-order(abs(round(cor(cbind(yHT,(Datxmun[,-c(1:2)]))),2)[1,]))
nv<-(c2[or])[(length(c2)-1):1]

rm(matpiso1,hombresc1,matpiso2,hombresc0,hombresc2,taseschc0)
rm(banoagua3)
rm(banoagua1)
rm(matpiso3)
rm(banoagua4)
rm(nase614c2)
rm(nase614c0)
rm(hombresc3)
rm(hombresc4)
rm(hombresc5)
rm(nase614c1)
rm(taseschc4)
rm(taseschc3)
rm(taseschc2)
rm(taseschc1)
rm(banoagua2)

# PRIMERA VARIABLE
lm1<-lm(as.formula(paste("yHT~",names(nv)[1])))
sqm<-rep(0,length(nv))
for(i in 2:length(nv)) {
   lm2<-lm(as.formula(paste("yHT~",names(nv)[i])))
   AA<-anova(lm1,lm2)
   sqm[i]<-AA$Sum[2] }
round(sqm,2)
plot(1:189,sqm)

# SEGUNDA VARIABLE
lm1<-lm(as.formula(paste("yHT~escpromh")))
base<-paste("yHT~",names(nv)[1])
sqm<-rep(0,length(nv))
for(i in 1:length(nv)) {
   lm2<-lm(as.formula(paste(base,names(nv)[i],sep="+")))
   AA<-anova(lm1,lm2)
   sqm[i]<-AA$Sum[2] }
round(sqm,2)
plot(1:189,sqm)  # seleccionar linea de corte para las var. más significativas

seg<-names(nv)[sqm>.3]
coli<-rep(0,length(seg))
for(i in 1:length(seg)) {
   out<-lm(as.formula(paste(base,seg[i],sep="+")))
```

```
   coll<-colldiag(out)
   coli[i]<-coll$condindx[3]}
round(coli,2)
plot(1:length(seg),coli) # seleccionar las variables con menor indice de cond.
                         # y que ademas tengan baja correlación
seg[coli<50]
#[1] "p_p_f_oe"   "p_p_eu2000" "imigra1"    "imigeu1"    "migeuj_1"   "muj12c0"    "e6a14_4"


# TERCERA VARIABLE
lm1<-lm(as.formula(paste("yHT~migeuj_1+",names(nv)[1])))
base<-paste("yHT~migeuj_1+",names(nv)[1])
sqm<-rep(0,length(nv))
for(i in 1:length(nv)) {
   lm2<-lm(as.formula(paste(base,names(nv)[i],sep="+")))
   AA<-anova(lm1,lm2)
   sqm[i]<-AA$Sum[2] }
round(sqm,2)
plot(1:189,sqm)

seg<-names(nv)[sqm>.3]
coli<-rep(0,length(seg))
for(i in 1:length(seg)) {
   out<-lm(as.formula(paste(base,seg[i],sep="+")))
   coll<-colldiag(out)
   coli[i]<-coll$condindx[4]}
round(coli,2)
plot(1:length(seg),coli)
seg[coli<50]
#  [1] "nivedj_3"  "e50a64_2"  "e50a64"     "agro1"      "agro0"      "e65mas"     "hijo12c0"
"hijo12c1"
# [9] "ninosc2"  "ient_12"

# CUARTA VARIABLE
nv<-(c2[or])[(length(c2)-1):1]
lm1<-lm(as.formula(paste("yHT~hijo12c1+migeuj_1+",names(nv)[1])))
base<-paste("yHT~hijo12c1+migeuj_1+",names(nv)[1])
sqm<-rep(0,length(nv))
for(i in 1:length(nv)) {
   lm2<-lm(as.formula(paste(base,names(nv)[i],sep="+")))
   AA<-anova(lm1,lm2)
   sqm[i]<-AA$Sum[2] }
round(sqm,2)
plot(1:189,sqm)

seg<-names(nv)[sqm>.25]
coli<-rep(0,length(seg))
for(i in 1:length(seg)) {
   out<-lm(as.formula(paste(base,seg[i],sep="+")))
   coll<-colldiag(out)
   coli[i]<-coll$condindx[5]}
round(coli,2)
plot(1:length(seg),coli)
seg[coli<50]
# [1] "lava1"     "lava0"     "rezahaci"  "cuadorc1"  "refri1"    "refri0"    "p_vph_3y"
# [8] "numcuac1"  "hombresc4" "hombresc5" "hijosmc3"  "taseschc4" "taseschc3" "muj12c4"
#[15] "e15a24_3"  "nas1524c3" "e15a24_4"  "e15a24"    "e15a24_2"

# QUINTA VARIABLE
nv<-(c2[or])[(length(c2)-1):1]
lm1<-lm(as.formula(paste("yHT~e15a24_3+hijo12c1+migeuj_1+",names(nv)[1])))
base<-paste("yHT~e15a24_3+hijo12c1+migeuj_1+",names(nv)[1])
sqm<-rep(0,length(nv))
for(i in 1:length(nv)) {
   lm2<-lm(as.formula(paste(base,names(nv)[i],sep="+")))
   AA<-anova(lm1,lm2)
   sqm[i]<-AA$Sum[2] }
round(sqm,2)
plot(1:189,sqm)

seg<-names(nv)[sqm>.15]
```

```
coli<-rep(0,length(seg))
for(i in 1:length(seg)) {
   out<-lm(as.formula(paste(base,seg[i],sep="+")))
   coll<-colldiag(out)
   coli[i]<-coll$condindx[6]}
round(coli,2)
plot(1:length(seg),coli)
seg[coli<60]
#[1] "banoagua3" "muj12c3"   "p_hog_ind" "ninosc2"

# SEXTA VARIABLE
nv<-(c2[or])[(length(c2)-1):1]
lm1<-lm(as.formula(paste("yHT~banoagua3+migeuj_1+hijo12c1+e15a24_3+",names(nv)[1])))
base<-paste("yHT~banoagua3+migeuj_1+hijo12c1+e15a24_3+",names(nv)[1])
sqm<-rep(0,length(nv))
for(i in 1:length(nv)) {
   lm2<-lm(as.formula(paste(base,names(nv)[i],sep="+")))
   AA<-anova(lm1,lm2)
   sqm[i]<-AA$Sum[2] }
round(sqm,2)
plot(1:189,sqm)

seg<-names(nv)[sqm>.1]
coli<-rep(0,length(seg))
for(i in 1:length(seg)) {
   out<-lm(as.formula(paste(base,seg[i],sep="+")))
   coll<-colldiag(out)
   coli[i]<-coll$condindx[7]}
round(coli,2)
plot(1:length(seg),coli)
seg[coli<60]
# [1] "hijo12c2"

# LA UNICA VARIABLE SELECIONADA ESTÁ CORRELACIONADA, POR LO TANTO EL MODELO
# SE QUEDA CON CINCO VARIABLES

# MODELO FINAL Y ANÁLISIS DE RESIDUALES
out1<-lm(yHT~escpromh+migeuj_1+hijo12c1+e15a24_3+banoagua3)
summary(out1)
par(mfrow=c(2,2))
plot(out1)
#plot(out1$fitted,residuals(out1))
#lines(lowess(residuals(out1)~out1$fitted),col="red")
par(mfrow=c(3,2))
plot(escpromh,residuals(out1),xlab="Prom. de escolartidad masculina en la vivienda",
    ylab="Residuos")
lines(lowess(residuals(out1)~escpromh),col="red")
plot(migeuj_1,residuals(out1),xlab="Prop. de viviendas cuyo jefe del hogar migra a EU",
    ylab="Residuos")
lines(lowess(residuals(out1)~migeuj_1),col="red")
plot(hijo12c1,residuals(out1),xlab="Prop. de viviendas con un hijo menor de 12 años",
    ylab="Residuos")
lines(lowess(residuals(out1)~hijo12c1),col="red")
plot(e15a24_3,residuals(out1),xlab="Prop. de viviendas con tres habitantes de 15 a 24 años",
    ylab="Residuos")
lines(lowess(residuals(out1)~e15a24_3),col="red")
plot(banoagua3,residuals(out1),xlab="Prop. de viviendas en donde no se le puede echar agua
al baño",
    ylab="Residuos")
lines(lowess(residuals(out1)~banoagua3),col="red")
plot(e50a64_1,residuals(out1),xlab="Prop. de viviendas con un habitante de 50 a 64 años",
    ylab="Residuos")
lines(lowess(residuals(out1)~e50a64_1),col="red")

#  CALCULO DEL PROMEDIO SINTÉTICO
ySYN<-fitted(out1)
round(exp(ySYN),0)
XpoS<-cbind(rep(1,23),Datxmun[,94],Datxmun[,45],Datxmun[,107],Datxmun[,12],Datxmun[,84])

# CÁLCULO DEL ECM SINTÉTICA
sesSYN2<-residuals(out1)^2
```

```
ecmSY<-diag(XpoS%*%vcov(out1)%*%t(XpoS))+sesSYN2
CVSYN<-sqrt(ecmSY)
round(sqrt(ecmSY),3)                                    #coeficiente de variación
ecmSYN<-ecmSY*exp(ySYN)^2
as.vector(round(sqrt(ecmSYN),1))
minSYNm<-round(as.vector(exp(ySYN))-1.96*sqrt(ecmSYN),0)
maxSYNm<-round(as.vector(exp(ySYN))+1.96*sqrt(ecmSYN),0)

# municipios que no están en muestra

Datxmun<-Datxmunt[-munson,]
XpoS1<-cbind(rep(1,49),Datxmun[,94],Datxmun[,45],Datxmun[,107],Datxmun[,12],
       Datxmun[,84])
ySYNn<-XpoS1%*%out1$coeff
round(as.vector(exp(ySYNn)),0)
ecmSYn<-diag(XpoS1%*%vcov(out1)%*%t(XpoS1))+sum(sesSYN2)/length(na)
round(sqrt(ecmSYn),3)                                    #coeficiente de variación
ecmSYNn<-ecmSYn*exp(ySYNn)^2
as.vector(round(sqrt(ecmSYNn),1))
minSYN<-round(as.vector(exp(ySYNn))-1.96*sqrt(ecmSYNn),0)
maxSYN<-round(as.vector(exp(ySYNn))+1.96*sqrt(ecmSYNn),0)

############################
# ESTIMACION COMPUESTA     #
############################
gam<-ecmSYN/(ecmSYN+ecmHT)
gam
medmC<-round(gam*medmHT+(1-gam)*exp(ySYN),0)
round(medmC,0)
ecmC<-gam^2*ecmHT+(1-gam)^2*ecmSYN+2*gam*(1-gam)*ifelse(ecmHT-sqrt(sesSYN2)*ySYN^2>0,ecmHT-
sqrt(sesSYN2)*ySYN^2,0)
recmC<-round(sqrt(ecmC),1)
CVCOM<-recmC/medmC
round(CVCOM,3)
minC<-round(as.vector(medmC)-1.96*recmC,0)
maxC<-round(as.vector(medmC)+1.96*recmC,0)


##################
# SMALL AREAS     #
##################

#########################################################################
#  función EBLUP.area     AUTOR:   DR. VIRGILIO GOMEZ RUBIO
#  LE HICE UNAS ADECUACUIONES EN LA CONVERGENCIA DEL ALGORITMO
#########################################################################

EBLUP.area<-function(ydir, Xpop, vardir, m, tol=10e-6, maxiter=50, method="ML")
{

        res<-switch(method,
                ML = EBLUP.area.ML(ydir, Xpop, vardir, m, tol, maxiter),
                REML = EBLUP.area.REML(ydir, Xpop, vardir, m, tol, maxiter)
        )

        if(is.null(res))
                print("Method should be ML or REML\n")

        return(res)
}


#traditional EBLUB ML procedure, scoring algorithm.
EBLUP.area.ML<-function(ydir, Xpop, vardir, m, tol, maxiter)
{
        sigma2.u.stim<-0  #variance components
        sigma2.u.stim[1]<-5000000 #starting value
        k<-0

        diff<-tol+1
        while ((diff>tol) & (k<maxiter))
```

```
        {
                k<-k+1

                V<-diag(1,m) #variances-covariances matrix
                for (i in 1:m)
                {
                        V[i,i]<-(sigma2.u.stim[k]+vardir[i,1])
                } #vardir is the mX1 vector of sampling variance

                Vinv<-solve(V)

                BETA<-solve(t(Xpop)%*%Vinv%*%Xpop)%*%t(Xpop)%*%Vinv%*%ydir[,1]  #ydir  is  the
mX1 vector of direct (sampling) estimates
                sdev<-(-0.5)*sum(diag(Vinv))-(0.5)*t(ydir[,1]-(Xpop)%*%BETA)%*%((-
1)*Vinv%*%Vinv)%*%(ydir[,1]-(Xpop)%*%BETA) #Xpop is the mXp matrix of p auxiliary variables
                Idev<-((0.5)*(sum(diag(Vinv%*%Vinv))))^(-1)   #Idev   is   the   inverse   of
information matrix
                sigma2.u.stim[k+1]<-sigma2.u.stim[k]+Idev*sdev #scoring algorithm
                diff<-abs(sigma2.u.stim[k+1]-sigma2.u.stim[k])
        }

        if (sigma2.u.stim[k+1]<0)
                sigma2u<-0
        else
                sigma2u<-sigma2.u.stim[k+1]

        V<-diag(1,m)
        G<-diag(1,m)*sigma2u
        for (i in 1:m) {V[i,i]<-(sigma2u+vardir[i,1])}

        Vinv<-solve(V)

        Bstim<-solve(t(Xpop)%*%Vinv%*%Xpop)%*%t(Xpop)%*%Vinv%*%ydir[,1]
        m1<-diag(1,m)
        thetaEBLUP<-Xpop%*%Bstim+m1%*%G%*%Vinv%*%(ydir[,1]-(Xpop%*%Bstim)) #EBLUP estimator
        varbeta<-solve(t(Xpop)%*%solve(V)%*%Xpop)
        randeff<-m1%*%G%*%Vinv%*%(ydir[,1]-(Xpop%*%Bstim))#Estimate of the random effects

        sig<-Bstim/sqrt(diag(solve(t(Xpop)%*%Vinv%*%Xpop))) #significance of Beta estimates

        #to estimate MSE for each small area
        #g1
        g1<-diag((G-G%*%Vinv%*%G))
        #g2
        g2<-matrix(0,m,1)
        for (i in 1:m) {m1<-matrix(0,m,1)
                  m1[i]<-1
        g2[i]<-(Xpop[i,]-
t(m1)%*%G%*%Vinv%*%Xpop)%*%solve(t(Xpop)%*%Vinv%*%Xpop)%*%t(Xpop[i,]-
t(m1)%*%G%*%Vinv%*%Xpop)}

        Idev<-((0.5)*(sum(diag(Vinv%*%Vinv))))^(-1)
        #g3
        g3<-matrix(0,m,1)
        for (i in 1:m) {m1<-matrix(0,m,1)
                  m1[i]<-1
        g3[i]<-(t(m1)%*%(Vinv+G%*%((-1)*Vinv%*%Vinv))%*%V%*%t((t(m1)%*%(Vinv+G%*%((-
1)*Vinv%*%Vinv)))))*Idev}

        #the bias
        bdist<-0
        btr<-0
        gradg1<-0
        I<-diag(1,m)
        distorsione<-matrix(0,m,1)
        for (i in 1:m)
        {
                m1<-matrix(0,m,1)
                m1[i]<-1
                gradg1<-t(m1)%*%(I-((I%*%Vinv%*%G)+(G%*%((-
1)*Vinv%*%I%*%Vinv)%*%G)+(G%*%Vinv%*%I)))%*%m1
```

```
                btr<-sum(diag(solve(t(Xpop)%*%Vinv%*%Xpop)%*%t(Xpop)%*%((-
1)*Vinv%*%I%*%Vinv)%*%Xpop))
                bdist<-(1/(m*2))*Idev*btr
                distorsione[i,1]<-bdist*gradg1
        }

        #estimated MSE
        msestim<-g1-distorsione+g2+2*g3


#       list(beta=Bstim, betasig=sig, EBLUP=thetaEBLUP,g1=g1, g2=g2, g3=g3,
#               MSE=msestim)
        list(EBLUP=thetaEBLUP, beta=Bstim, sigma2u=sigma2u, iterations=k,
  g1=g1, g2=g2, g3=g3, mse=msestim, randeff=randeff, varbeta=varbeta)
}


EBLUP.area.REML<-function(ydir, Xpop, vardir, m, tol, maxiter)
{
        sigma2.u.stim<-0
        sigma2.u.stim[1]<-10
        k<-0
        diff<-tol+1

        while ( (diff>tol) & (k<maxiter) )
        {
                k<-k+1

                V<-diag(1,m)
                for (i in 1:m) {V[i,i]<-(sigma2.u.stim[k]+vardir[i,1])}

                Vinv<-solve(V)

                BETA<-solve(t(Xpop)%*%Vinv%*%Xpop)%*%t(Xpop)%*%Vinv%*%ydir[,1]
                P<-Vinv-(Vinv%*%Xpop%*%solve(t(Xpop)%*%Vinv%*%Xpop)%*%t(Xpop)%*%Vinv)       #P
matrix
                sdev<-(-0.5)*sum(diag(P))+((0.5)*t(ydir[,1])%*%P%*%P%*%ydir[,1])
                Idev<-((0.5)*(sum(diag(P%*%P))))^(-1)
                sigma2.u.stim[k+1]<-sigma2.u.stim[k]+Idev*sdev
                diff<-abs(sigma2.u.stim[k+1]-sigma2.u.stim[k])
                if (k>100) {diff.S=0.00001}
        }

        if (sigma2.u.stim[k+1]<0)
                sigma2uREML<-0
        else
                sigma2uREML<-sigma2.u.stim[k+1]

#       sigma.sim.REML[w]<-sigma2uREML
        V<-diag(1,m)
        G<-diag(1,m)*sigma2uREML
        for (i in 1:m)
                V[i,i]<-(sigma2uREML+vardir[i,1])

        Vinv<-solve(V)

        Bstim<-solve(t(Xpop)%*%Vinv%*%Xpop)%*%t(Xpop)%*%Vinv%*%ydir[,1]
        m1<-diag(1,m)
        thetaEBLUPREML<-Xpop%*%Bstim+m1%*%G%*%Vinv%*%(ydir[,1]-(Xpop%*%Bstim))
        randeff<-m1%*%G%*%Vinv%*%(ydir[,1]-(Xpop%*%Bstim))#Estimate of the random effects
        varbeta<-solve(t(Xpop)%*%solve(V)%*%Xpop)

        #to estimate MSE

        g1REML<-diag((G-G%*%Vinv%*%G))


        g2REML<-matrix(0,m,1)
        for (i in 1:m)
        {
                m1<-matrix(0,m,1)
```

```
                m1[i]<-1
                g2REML[i]<-(Xpop[i,]-
t(m1)%*%G%*%Vinv%*%Xpop)%*%solve(t(Xpop)%*%Vinv%*%Xpop)%*%t(Xpop[i,]-
t(m1)%*%G%*%Vinv%*%Xpop)
        }

        Idev<-((0.5)*(sum(diag(P%*%P))))^(-1)
        g3REML<-matrix(0,m,1)
        for (i in 1:m)
        {
                m1<-matrix(0,m,1)
                m1[i]<-1
                g3REML[i]<-(t(m1)%*%(Vinv+G%*%((-
1)*Vinv%*%Vinv))%*%V%*%t((t(m1)%*%(Vinv+G%*%((-1)*Vinv%*%Vinv)))))*Idev
        }


        msestimREML<-g1REML+g2REML+2*g3REML



#       list(beta=Bstim, betasig=NULL, EBLUP=thetaEBLUPREML,g1=g1REML,
#               g2=g2REML, g3=g3REML, MSE=msestimREML)
        list(EBLUP=thetaEBLUPREML, beta=Bstim,
  sigma2u=sigma2uREML, iterations=k, g1=g1REML, g2=g2REML,
  g3=g3REML, mse=msestimREML, randeff=randeff, varbeta=varbeta)
}

#####################################################################
#
#   ESTIMACIÓN EBLUP MODELO DE AREA
#
#####################################################################
Datxmun<-Datxmunt[munson,]
Xpo<-cbind(rep(1,23),Datxmun[,94],Datxmun[,45],Datxmun[,107],Datxmun[,12],
     Datxmun[,84])
#eblup1<-EBLUP.area(as.matrix(yHT,ncol=1),        Xpo[,c(1:7)],        as.matrix(Vad,ncol=1),
23,method="ML")
eblup1<-EBLUP.area(as.matrix(yHT,ncol=1), Xpo, as.matrix(Vad,ncol=1), 23,method="ML")
round(sqrt(diag(eblup1$varbeta)),3)
round(as.vector(eblup1$beta),3)
eblup1a<-EBLUP.area(as.matrix(yHT,ncol=1),        Xpo[,c(1:7)],        as.matrix(Vad,ncol=1),
23,method="REML")
round(sqrt(diag(eblup1a$varbeta)),3)
round(as.vector(eblup1a$beta),3)
round(as.vector(exp(eblup1$EBLUP)),0)
round(as.vector(sqrt(eblup1$mse*exp(eblup1$EBLUP)^2)/exp(eblup1$EBLUP)),3)
mse<-as.vector(eblup1$mse*exp(eblup1$EBLUP)^2)
eer<-sqrt(sum((eblup1$EBLUP-yHT)^2)/(23-5))
eer
# ANALISIS DE RESIDUALES MODELO DE ÁREA
resE<-yHT-eblup1$EBLUP
qqnorm(resE[resE!=0][-7],main="Gráfica de Normalidad",xlab="Cuantiles teóricos",
       ylab="Cuantiles de los residuales")
qqline(resE[resE!=0][-7],col="red")
shapiro.test(resE[resE!=0][-7])
cor(Xpo[,-1])

# NORMALIDAD DE EFECTOS ALEATORIOS
qqnorm(eblup1$randeff,main="Gráfica de Normalidad",xlab="Cuantiles teóricos",
       ylab="Cuantiles de los efectos aleatorios")
qqline(eblup1$randeff,col="red")
shapiro.test(eblup1$randeff)

#  VALIDACION CRUZADA
muVC<-rep(0,23)
ecmVC<-rep(0,23)
for (i in 1:23){
   XpoVC<-Xpo[-i,]
   eblupVC<-EBLUP.area(as.matrix(yHT[-i],ncol=1), XpoVC[,c(1:6)], as.matrix(Vad[-i],ncol=1),
            22,method="ML")
```

```
   muVC[i]<-Xpo[i,]%*%eblupVC$beta+eblup1$randeff[i]
#    ecmVC[i]<-eblupVC$sigma2u+t(as.matrix(Xpo[i,]))%*%eblupVC$varbeta%*%as.matrix(Xpo[i,])
}
round(as.vector(exp(muVC)),0)
eVC<-eblup1$EBLUP-muVC
PRESS<-sum(eVC^2)/(23-5)
sqrt(PRESS)
Mmunson<-munson
munson<-munson[resE!=0]

# ESTIMACIÓN PARA MUNICIPIOS QUE NO ESTÁN EN MUESTRA Y VARIANZA INDEFINIDA
Datxmun<-Datxmunt[-munson,]
Xpo1<-cbind(rep(1,dim(Datxmun)[1]),Datxmun[,94],Datxmun[,45],Datxmun[,107],Datxmun[,12],
     Datxmun[,84])
eblupfs1<-Xpo1%*%eblup1$beta
round(as.vector(exp(eblupfs1)),0)
eblupfs1a<-Xpo1%*%eblup1a$beta
round(as.vector(exp(eblupfs1a)),0)
yEBLUP<-rep(0,72)
yEBLUP[munson]<-round(as.vector(exp(eblup1$EBLUP)),0)[resE!=0]
yEBLUP[-munson]<-round(as.vector(exp(eblupfs1)),0)
yEBLUP
ecmynEBLUP<-eblup1$sigma2u+diag(Xpo1%*%eblup1$varbeta%*%t(Xpo1))
ECMynEBLUP<-as.vector(round(ecmynEBLUP*exp(eblupfs1)^2,3))
round(sqrt(ecmynEBLUP),3)       # coeficiente de variación
ECMEBLUP<-rep(0,72)
ECMEBLUP[munson]<-mse[resE!=0]
ECMEBLUP[-munson]<-ECMynEBLUP
CVmEBLUP<-sqrt(ECMEBLUP[Mmunson])/exp(eblup1$EBLUP) # coeficiente de variación muestra
muS<-1:72
cbind(muS[-Mmunson],as.character(Nommun[-
Mmunson]),round(exp(ySYNn),0),round(sqrt(ecmSYNn),1),
     round(yEBLUP[-Mmunson],0),round(sqrt(ECMEBLUP[-Mmunson]),1))

#  INTERVALOS DE CONFIANZA
minEBLUPm<-yEBLUP[Mmunson]-1.96*sqrt(ECMEBLUP[Mmunson])
maxEBLUPm<-yEBLUP[Mmunson]+1.96*sqrt(ECMEBLUP[Mmunson])
minEBLUP<-yEBLUP[-Mmunson]-1.96*sqrt(ECMEBLUP[-Mmunson])
maxEBLUP<-yEBLUP[-Mmunson]+1.96*sqrt(ECMEBLUP[-Mmunson])

# GRAFICA DE PROMEDIOS ESTIMADOS
o<-order(medmHT)
plot(1:23,medmHT[o],ylim=c(0,90000),main="Promedio   del   Ingreso   total   en   municipios
muestreados por método de estimación",
     xlab=" ",ylab="Pesos",pch=15,xaxt = "n",cex=1.4,col="blue")
axis(1, 1:23, labels = Nommun[Mmunson][o],lwd=.4,padj=0,cex.axis=.6,las=2)
points(1:23,medGREG[o],pch=16,cex=1.4,col="orange")
points(1:23,exp(ySYN)[o],pch=18,cex=1.4,col="red")
points(1:23,medmC[o],pch=17,cex=1.4,col="magenta")
#points(1:23,medGREG[o],pch=1,cex=1)
points(1:23,yEBLUP[Mmunson][o],pch=19,cex=1.4)
legend(1,85000,legend=c("Directa","GREG","Sintética","Compuesta",
       "EAP modelo de área"),pch=c(15,16,17,18,19),col=c(4,"orange",6,2,1),cex=1)

# GRAFICA DE CV ESTIMADOS
oo<-order(na)
plot(1:23,CVHT[oo],ylim=c(0,1),main="Coeficientes  de  Variación  ordenados  por  tamaño  de
muestra",
     xlab=" ",ylab="Coeficiente de Variación",xaxt = "n",lwd=2,col="blue",lty=1,type="l")
axis(1, 1:23, labels = Nommun[Mmunson][oo],lwd=.4,padj=0,cex.axis=.6,las=2)
axis(1, 1:23+.2, labels = paste("(",na,")",sep='')[oo],lwd=.4,padj=0,cex.axis=.6,las=2)
lines(1:23,CVGREG[oo],lty=2,lwd=2,col="orange")
lines(1:23,CVSYN[oo],lty=3,lwd=2,col="red")
lines(1:23,CVCOM[oo],lty=4,lwd=2,col="magenta")
lines(1:23,CVmEBLUP[oo],lty=1,lwd=2)
legend(1,.95,legend=c("Directa","GREG","Sintética","Compuesta",
       "EAP modelo de área"),lty=c(1,2,3,4,1),col=c(4,"orange",6,2,1),cex=1,lwd=2)

oo2<-order(Na)
plot(1:23,CVHT[oo2],ylim=c(0,1),main="Coeficientes  de  Variación  ordenados  por  número  de
viviendas en el municipio",
```

```
     xlab=" ",ylab="Coeficiente de Variación",xaxt = "n",lwd=2,col="blue",lty=1,type="l")
axis(1, 1:23, labels = Nommun[Mmunson][oo2],lwd=.4,padj=0,cex.axis=.6,las=2)
axis(1, 1:23+.2, labels = paste("(",Na,")",sep='')[oo2],lwd=.4,padj=0,cex.axis=.6,las=2)
lines(1:23,CVGREG[oo2],lty=2,lwd=2,col="orange")
lines(1:23,CVSYN[oo2],lty=3,lwd=2,col="red")
lines(1:23,CVCOM[oo2],lty=4,lwd=2,col="magenta")
lines(1:23,CVmEBLUP[oo2],lty=1,lwd=2)
legend(1,.95,legend=c("Directa","GREG","Sintética","Compuesta",
       "EAP modelo de área"),lty=c(1,2,3,4,1),col=c(4,"orange",6,2,1),cex=1,lwd=2)


# GRAFICA CON INTERVALOS DE CONFIAZA
plot(1:23,yEBLUP[Mmunson],ylim=c(0,90000),main="Estimaciones en municipios muestreados con
intervalos al 95%",
     xlab=" ",ylab="Pesos",pch=3,xaxt = "n",cex=0.25)
segments(1:23,minEBLUPm,1:23,maxEBLUPm)
axis(1, 1:23, labels = Nommun[Mmunson],lwd=.4,padj=0,cex.axis=.6,las=2)
points(1:23-.15,exp(ySYN),pch=3,cex=0.25)
segments(1:23-.15,minSYNm,1:23-.15,maxSYNm,col="red")
points(1:23-.3,medmC,pch=3,cex=0.25)
segments(1:23-.3,minC,1:23-.3,maxC,col="magenta")
points(1:23-.45,medGREG,pch=3,cex=0.25)
segments(1:23-.45,minGREG,1:23-.45,maxGREG,"orange")
points(1:23-.6,medmHT,pch=3,cex=0.25)
segments(1:23-.6,minHT,1:23-.6,maxHT,col="blue")
legend(.5,85000,legend=c("Directa","GREG","Sintética","Compuesta",
       "EAP modelo de área"),lty=1,col=c(4,"orange",6,2,1),cex=.8)

plot(1:49,exp(ySYNn),ylim=c(0,100000),main="Estimaciones en municipios no muestreados con
intervalos al 95%",
     xlab=" ",ylab="Pesos",pch=3,col="red",xaxt = "n",cex=0.25)
segments(1:49,minSYN,1:49,maxSYN,col="red")
points(1:49-.5,yEBLUP[-Mmunson],pch=3,cex=0.25)
segments(1:49-.5,minEBLUP,1:49-.5,maxEBLUP)
legend(40,74000,legend=c("Sintética","EAP modelo de área"),lty=1,col=c(2,1),cex=.8)
axis(1, 1:49, labels = Nommun[-Mmunson],lwd=.4,padj=0,cex.axis=.6,las=2)

# VERIFICACIÓN DEL PROMEDIO ESTATAL
PEmin<-coef(medtHT)-1.96*SE(medtHT)
PEmax<-coef(medtHT)+1.96*SE(medtHT)
pondmun<-as.vector(Nconteo[,1])/sum(as.vector(Nconteo[,1]))
prompond<-sum(yEBLUP*pondmun)
prompond>PEmin&prompond<PEmax

# [1] TRUE
# EL PROMEDIO PONDERADO ESTÁ EN EL INTERVALO DE CONFIANZA DEL PROMEDIO ESTATAL DIRECTO
```