UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS MATEMÁTICAS

FACULTAD CIENCIAS

ESTIMADOR DE REGRESIÓN GENERALIZADO PARA TOTALES EN POBLACIONES FINITAS: PROGRAMACIÓN Y COMPARACIÓN CONTRA LOS ESTIMADORES CLÁSICOS VÍA SIMULACIÓN.

T E S I S

QUE PARA OBTENER EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIAS MATEMÁTICAS

P R E S E N T A

ANNUAR ERVIN AYALA RIVERA

DIRECTOR DE TESIS: DR. IGNACIO MÉNDEZ RAMÍREZ

MÉXICO, D.F. NOVIEMBRE, 2010

Índice general

Índice de figuras

Índice de cuadros

INTRODUCCIÓN

Con cierta frecuencia se plantea la necesidad de efectuar estimaciones en poblaciones finitas, partiendo de muestras relativamente pequeñas. Las limitaciones al tamaño muestral suelen ser debidas, la mayor parte de las veces, a restricciones presupuestarias y a características específicas del universo que en la práctica impiden desarrollar el trabajo de campo en el tiempo deseado. En estos casos, si la única estructura aleatoria que se considera para realizar la inferencia es la que se deriva de la adopción de un diseño muestral aleatorio, donde puede que la precisión de los resultados sea relativamente pequeña.

Cuando se dispone de información auxiliar, una manera de disminuir los errores de estimación sin incrementar el tamaño de la muestra, consiste en incorporar esta información al proceso predictivo, proponiendo para ello la adopción de un modelo de superpoblación.

Las perspectivas que se abren cuando la estructura aleatoria que soporta la inferencia emanan de un modelo, se orientan en varias direcciones. Por un lado, surgen criterios de construcción de estimadores que conducen a soluciones, en ocasiones poco intuitivas, pero con mayor atractivo que los estimadores habitualmente propuestos en la literatura clásica del muestreo en poblaciones finitas. Por otro lado, la no necesidad de recurrir a diseños aleatorios, plantea el interés en buscar diseños intencionados que permitan intervenir sobre la capacidad predictiva de las estrategias, aunque manteniendo su carácter no informativo o ignorable, Murgui (2005).

En las encuestas por muestreo, es frecuente la disponibilidad de alguna o algunas variables correlacionadas con la que es el objeto de estudio, y los datos que proporcionan éstas pueden producir estimaciones más precisas. Esta información auxiliar se utiliza algunas veces para construir el diseño muestral (estratificación de la población, selección de unidades con probabilidades desiguales, etc.) o bien para modificar los estimadores usuales (que sólo utilizan la información muestral de la variable principal) mediante estimadores indirectos.

La literatura de muestreo en poblaciones finitas es abundante en ejemplos en los cuales los métodos indirectos son usados para estimar medias y totales. Ya en el año 1812, Laplace en su libro "Teoría Analítica de las Probabilidades" utilizó un estimador tipo razón para calcular la población de Francia. Estos métodos indirectos han sido ampliamente considerados en los últimos cincuenta años Rueda (1994).

La parte central de este proyecto es considerar el estimador de regresión generalizado (GREG) para poblaciones finitas. Juega un papel muy importante la aleatorización para la selección de la muestra y para evaluar las propiedades estadísticas de las estrategias de estimación en los modelos construidos para tales estimadores. Un buen modelo es crucial para restringir la variabilidad de un estimador de modelo-asistido como el GREG. Si el modelo asumido describe bien las relaciones entre las variables consideradas en la población entonces el GREG traerá una reducción importante de la varianza.

El GREG esencialmente incorpora variables auxiliares relevantes siempre y cuando se conozca su valor en cada una de las unidades de la población; por consiguiente se conoce el total.

Este proyecto consta de tres capítulos, en los cuales se desarrolla una variación del GREG que permite comparar de forma inmediata a los estimadores Horvitz-Thompson, de razón y de regresión; y al mismo tiempo elegir uno mejor que puede o no coincidir con los anteriores.

En el primer capítulo se explica brevemente las características de los estimadores antes señalados, así como la notación que se usa para poder comprender las fórmulas utilizadas para calcular cada uno de ellos.

En el capítulo dos, se plantea de forma intuitiva la relación que existe entre los estimadores de razón y de regresión en términos del GREG y se propone una modificación en la fórmula del GREG que permite expresarlos de una forma más general.

En ese mismo capítulo se presenta el diseño que se realizó de un programa el cual calcula los valores correspondientes para cada uno de los estimadores utilizados y además se complementa con el desarrollo de otro que permite realizar las simulaciones de los mismos, para un ejemplo en particular con datos de población, número de médicos y extensión de territorio por condado de Estados Unidos.

El tercer y último capítulo tiene como objetivo aterrizar la teoría planteada anteriormente con

ejemplos, primero para el caso univariado se toman los datos del INEGI para estimar la población total de México en el 2005 usando como auxiliar el censo del 2000. Para el caso multivariado se separa la variable en hombres y mujeres. También se tiene el mismo análisis para los datos de Estados Unidos, con sus respectivas modificaciones. Por medio de cuadros y gráficas se intenta mostrar la eficacia de los diferentes estimadores.

Al final se plantean y se responden algunas preguntas que surgen de forma natural al desarrollar este tipo de propuestas.

En resumen, el objetivo es encontrar un estimador del tipo GREG que minimice el error estándar apoyado por los estimadores de razón y de regresión y verificar que el beneficio obtenido es mayor que el costo computacional.

Capítulo 1

DEFINICIONES Y CONCEPTOS

En este capítulo se explica la notación a usar durante este proyecto, así como las definiciones y conceptos necesarios para poder comprender de una mejor manera lo que se intenta mostrar.

1.1. Clasificación de estimadores

En muestreo, los estimadores de los parámetros de una población finita son, algunas veces, clasificados en: basados en diseño y en modelo. Otra forma de estudiarlos es en estimadores directos e indirectos, pero pueden haber más criterios. Ya que pueden tener propiedades de muchas clases y los límites entre una y otra no siempre son claros, desde un punto de vista práctico, esto no es un problema ya que la precisión de un estimador no depende de su clasificación.

1.1.1. Estimadores basados en diseño y modelo asistido

En la estimación basada en diseño, la población es considerada finita y fija y sus unidades pueden estar identificadas y etiquetadas. Las variables de estudio son también fijas, por consiguiente la única fuente de aleatoriedad se da en el mecanismo al selecionar la muestra.

Los estimadores basados en diseño usan información acerca de la muestra por medio de los pesos de muestreo llamados factores de expansión, que son el inverso de la probabilidad de seleción.

En la estimación basada en diseño de modelo asistido (o simplemente estimación de modelo asistido), un modelo estadístico es explícitamente usado como una herramienta de ayuda que incorpora información auxiliar en el proceso. Esto requiere tratar la población finita y fija como si hubiera sido generada por el modelo estadístico. Este es usado para hacer predicciones del total de la población,

para estimar las unidades no muestreadas o para los errores esas estimaciones.

No siempre es clara la distinción entre estimadores de modelo asistido de los que no lo son. La definición que propone Myrskylä (2007) servirá de guía, "Un estimador basado en diseño es de modelo asistido si y sólo si, hay un modelo explícito que se usa para hacer predicciones". Con lo anterior se dice que los GREG entran en esa clasificación.

1.1.2. Estimadores directos e indirectos

Como ya se había señalado, los estimadores también pueden ser divididos en directos e indirectos. Esta clasificación es relevante sólo cuando se estudia el total de la población, en la práctica siempre se usan los directos.

Los estimadores directos son definidos como los que usan los valores de las variables de estudio que están en el dominio. Los indirectos, por el contrario, usan información acerca de las variables que no están directamente relacionadas con el dominio bajo estudio.

1.2. Definiciones y notación

Notación importante que se va a usar a lo largo de este trabajo:

y_j : Valor que toma la variable de estudio en la $u_j, j = 1, 2, .., n$;

Y : Total poblacional;

X : Matriz de n filas (observaciones) por p columnas (variables auxiliares);

x_{ij} : Valor que toma la variable auxiliar en la columna i, fila j;

$x_{i.} = \sum_{j=1}^{N} x_{ij}$: Valor total de la columna i;

$X_T = (x_{1.}, x_{2.}, ..., x_{p.})$: Vector de totales para las columnas de X;

N : Tamaño poblacional;

n : Tamaño de la muestra;

π_j : Probabilidad de inclusión de la unidad j;

$w_j = \frac{1}{\pi_j}$: Peso de muestreo factor de expansión de la unidad j;

$\bar{y} = \sum_{j=1}^{n} \frac{y_j}{n}$: Media muestral;

$s_y^2 = \sum_{j=1}^{n} \frac{(y_j - \bar{y})^2}{n-1}$: Varianza muestral;

$\hat{t}_{y.HT} = \frac{N}{n} \sum_{j=1}^{n} y_j = N\bar{y}$: Estimador del total (Horvitz-Thompson);

$\widehat{var}(\hat{t}_{y.HT}) = N^2(1 - \frac{n}{N})\frac{s_y^2}{n}$: Estimador de la varianza del estimador del total;

Conforme aparezca más notación, se especificará qué significa.

Se estudia una población finita de unidades u_i. Entonces se tiene que la población de estudio es:

$$U = \{1, 2, ..., N\}$$

Puede consistir de dominios o estratos, donde $U_h \subseteq U, h = 1, 2, ..., H$ y con la propiedad que $U_h \cap U_i = \emptyset$ de tal forma que $\bigcup_{h=1}^{H} U_h = U$. Algunos ejemplos de estratos podrían ser hombres y mujeres; personas que pertenecen a un cierto grupo de edad, etc. La variable de estudio es el vector y, las caracteríticas auxiliares están dadas por la matriz X. La primera es desconocida antes del muestreo a diferencia de las otras que son conocidas antes del muestreo.

Los parámetros de interés pueden ser las frecuencias o proporciones de éstas, es decir, divididas entre el tamaño del estrato (se supone que se sabe a cuál estrato pertenece cada variable y que el tamaño de cada estrato es conocido).

Una suposición que se hace en este trabajo es $h = 1$, pero todas las fórmulas siguen siendo válidas aunque se tengan varios estratos, ya que son disjuntos.

1.2.1. Estimador de Horvitz-Thompson

El estimador más conocido del total de la población de Y es el Horvitz-Thompson (HT) que pertence a la clasificación de los directos, es lineal e insesgado Tillé (2002). El estimador HT es insesgado pero es ineficiente, ya que no hace uso de la información auxiliar, aunque ésta exista.

Cuando el tamaño muestral es pequeño, no es un estimador adecuado aunque sea insesgado bajo el diseño, ya que es muy inestable y su varianza puede ser muy grande en estos casos EUSTAT (2001). Así que este es usado sólo como referencia y para la elaboración de ciertos conceptos importantes.

El estimador HT del total poblacional Y de la variable y viene dado por

$$\hat{t}_{y.HT} = \sum_{j=1}^{n} w_j y_j$$

donde $w_j = \frac{1}{\pi_j}$ son los pesos muestrales y π_j es la probabilidad de inclusión para el elemento j-ésimo de la muestra. En un muestreo aleatorio simple con n unidades seleccionadas del total N, w_j es igual a N/n para toda $j = 1, ..., n$. Un estimador insesgado de la varianza del estimador HT del total poblacional viene dado por la siguiente expresión:

$$\begin{aligned} \widehat{var}(\hat{t}_{y.HT}) &= N^2(1 - \frac{n}{N})\frac{1}{n}s_y^2 \\ &= N^2(1 - \frac{n}{N})\frac{1}{n}\sum_{j=1}^{n}\frac{(y_j - \bar{y})^2}{n-1} \end{aligned}$$

1.2.2. Estimador de diferencia

Es más elaborado que el anterior, la idea de este estimador como su nombre lo dice, es hacer una resta del total de la variable auxiliar contra su estimado por HT, y el resultado se le suma al HT de la Y, llamado también estimador de traslación Ruiz (1988).

$$\begin{aligned} \hat{t}_{y.Dif} &= \sum_{j=1}^{n} w_j y_j + (\sum_{j=1}^{N} x_j - \sum_{j=1}^{n} w_j x_j) \\ &= \hat{t}_{y.HT} + X_T - \hat{t}_{x.HT} \\ &= \hat{t}_{y.HT} + \bar{A} \end{aligned}$$

Donde $\bar{A} = X_T - \hat{t}_{x.HT}$. Su funcionamiento es bastante lógico ya que $\bar{A}$ es un "ajuste "al HT para estimar Y. La importancia de este estimador es su simplicidad y que usa información auxiliar, además ayuda a comprender un poco al de regresión.

No se necesita ningún cálculo para ver que este estimador es insesgado y su varianza estimada viene dada por

$$\widehat{var}(\hat{t}_{y.Dif}) = N^2(1 - \frac{n}{N})\frac{1}{n}(s_y^2 + s_x^2 - 2s_{xy})$$

1.2.3. Estimador de regresión

Si la relación entre Y_i y X_i, donde la primera es la variable de interés y la segunda es la explicativa, se puede representar aproximadamente como $Y_i = a + bX_i$, entonces esta información se puede incorporar en la construcción de estimadores de regresión.

La idea fundamental de los estimadores de regresión es que si se conoce $\bar{X}$ y el valor de b, se puede conocer $\bar{Y}$.

También de modo aproximado $\bar{y} = a + b\bar{x}$. Es decir los puntos $(\bar{X}, \bar{Y})$ y $(\bar{x}, \bar{y})$ están sobre la recta generada por la ecuación de regresión.

El valor de $\bar{Y}$ sería exacto si se cumpliera que $\bar{y} = a + b\bar{x}$; sin embargo, esto no se cumple porque tanto $\bar{x}$ como $\bar{y}$ están sujetas a fluctuaciones aleatorias originadas por el proceso de muestreo, de igual manera se debe estimar el coeficiente de regresión, de modo que en realidad se tiene un estimador, Rueda (1994) y Méndez (2009);

$$\hat{\bar{Y}} = \bar{y} - b(\bar{x} - \bar{X})$$

Para estimar el total se tiene:

$$\hat{Y} = N\hat{\bar{Y}} = N(\bar{y} - b(\bar{x} - \bar{X}))$$

Es difícil encontrar expresiones exactas para la varianza o los errores cuadráticos medios de estos estimadores que son sesgados y consistentes. Sin embargo, si n es pequeño, se puede ignorar el factor de corrección por finitud $(1 - n/N)$, entonces se tiene que

$$var(\hat{\bar{Y}}) \simeq \frac{S_y^2(1 - \rho^2))}{n}$$

donde ρ es el coeficiente de correlación entre X y Y.

Además

$$s_y^2 = \frac{1}{N-1}\sum_{j=1}^{N}(Y_j - \bar{Y})^2$$

Si se estiman estos dos parámetros por

$$\hat{\rho} = \frac{\sum_{j=1}^{n}\frac{(x_j - \bar{x})(y_j - \bar{y})}{n-1}}{\sqrt{\sum_{j=1}^{n}\frac{(x_j - \bar{x})^2}{n-1}\sum_{j=1}^{n}\frac{(y_j - \bar{y})^2}{n-1}}}$$

$$\hat{S}_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$$

Se obtiene un estimador de $var(\hat{\bar{Y}})$ y con él se construyen intervalos de confianza para $\bar{Y}$ o para Y.

1.2.4. Estimador de regresión generalizado (GREG)

Un estimador para totales más sofisticado es el de regresión generalizada. Una estimación obtenida por un GREG es una suma de valores observados ponderados mediante el producto de un factor de diseño y uno adicional calculado a través del uso de información extra. Es necesario entonces disponer de una o más variables auxiliares que estén, en lo posible, altamente relacionadas con la variable cuyo total se va a estimar.

El estimador de regresión generalizado utiliza información auxiliar de las variables $x's$ para estimar la variable y. Se diferencia del de regresión habitual en que introduce pesos de los coeficientes del modelo (normalmente el factor de expansión). Los GREG utilizan los modelos de regresión como un medio para conseguir estimadores consistentes desde el punto de vista del diseño. Requieren que el muestreo sea aleatorio. Han sido propuestos fundamentalmente por Särndal et al(1989). El estimador GREG constituye una amplia clase de estimadores que utilizan información auxiliar por modelo.

En la literatura, el término GREG se refiere "algunas veces"a un estimador que tiene un modelo lineal asistido de efectos fijos, aunque ya existen varios estudios acerca del GREG no lineal

Särndal (2007). Las fórmulas y propiedades relacionadas con el GREG se pueden encontrar en Särndal et al(1992), EUSTAT (2001) y Lepik (2007).

El estiomador está dado por:

$$\hat{t}_{y.GREG} = \sum_{j=1}^{N} \hat{y}_j + \sum_{j=1}^{n} w_j (y_j - \hat{y}_j) \tag{1.1}$$

Donde $\hat{y}_j,\ j = 1, ..., N$ son los valores estimados por un modelo dado. El término $\sum_{j=1}^{n} w_j (y_j - \hat{y}_j)$ puede interpretarse como un ajuste de regresión dado el estimador. El efecto es que produce una importante reducción de su varianza, especialmente cuando la relación entre y y x es muy fuerte, Fuller (2002). Si se tiene un modelo de regresión lineal, $y_j = x_j\ \beta + \epsilon_j$, donde β es un parametro desconocido y la $var(\epsilon_j) = \sigma^2$ con $x_j = (1, x_{j1}, ..., x_{jk})'$, entonces $\hat{y}_j = x_j\ \hat{\beta}_{GREG}$, donde

$$\hat{\beta}_{GREG} = (\sum_{j=1}^{n} w_j x_j' x_j c_j)^{-1} \sum_{j=1}^{n} w_j x_j' y_j c_j \tag{1.2}$$

donde c_j son constantes especificadas que sirven para ajustar el estimador según su diseño. La expresión (1.1) usando (1.2)puede escribirse también como:

$$\hat{t}_{y.GREG} = \hat{t}_{y.HT} + (X - \hat{t}_{x.HT})' \hat{\beta}_{GREG} \tag{1.3}$$

Donde $\hat{t}_{y.HT} = \sum_{j=1}^{n} w_j y_j$, es el estimador HT de Y, y $\hat{t}_{x.HT} = \sum_{j=1}^{n} w_j x_j$, es el estimador HT de X_T, donde $X_T = \sum_{j=1}^{N} x_j$, con lo que es un estimador calibrado Sugden and Smith (2007). Efectivamente las expresiones (1.1) y (1.3) coinciden, es fácil verlo sólo sustituyendo.

También se puede expresar el estimador de regresión generalizado como ponderación lineal sobre los y_j de modo que

$$\hat{t}_{y.GREG} = \sum_{j=1}^{n} w_j^*\ y_j = \sum_{j=1}^{n} w_j g_j y_j \tag{1.4}$$

Donde los pesos $w_j^* = w_j g_j$ con $w_j = 1/\pi_j$ y

$$\begin{aligned} g_j &= 1 + (\sum_{i=1}^{N} x_i - \sum_{i=1}^{n} w_i x_i)'\ T^{-1} x_j c_j \\ &= 1 + (X_T - \hat{t}_{x.HT})'\ T^{-1} x_j c_j \end{aligned} \tag{1.5}$$

Y

$$T = \sum_{j=1}^{n} w_j x_j' x_j c_j \tag{1.6}$$

Si se define una nueva variable

$$R = \sum_{j=1}^{n} w_j x_j' y_j c_j \tag{1.7}$$

Entonces usando (1.6) y (1.7) se reescribe a $\hat{\beta}_{GREG}$ como

$$\hat{\beta}_{GREG} = T^{-1}\, R$$

El valor de g_j está próximo a la unidad para la mayoría de los casos. Cuanto mayor es la muestra, debe tener mayor proximidad a la unidad. Es relativamente raro encontrar g_j que sean mayores que 4 o menores que 0. Las w_j* se llaman pesos calibrados ya que aplicados a x_j reproducen exactamente el total poblacional, es decir

$$\sum_{j=1}^{n} w_j^* x_j = \sum_{j=1}^{N} x_j = X_T$$

La varianza del estimador GREG viene dada por

$$var(\hat{t}_{y.GREG}) = \sum_{j=1}^{N}\sum_{k=1}^{N}(\frac{w_j w_k}{w_{jk}} - 1)\epsilon_j \epsilon_k$$

Donde $\epsilon_j = y_j - x_j'\, \beta_{GREG}$ y w_{jk} es la probabilidad de inclusión conjunta de los elementos j-ésimo y k-ésimo. La varianza anterior se estima mediante la expresión:

$$\widehat{var}(\hat{t}_{y.GREG}) = \sum_{j=1}^{n}\sum_{k=1}^{n}(w_j w_k - w_{jk})(g_j \hat{\epsilon}_j)$$

Donde $\hat{\epsilon}_j = y_j - x_j'\, \hat{\beta}_{GREG}$. En el caso del muestreo aleatorio simple, esta expresión toma la forma:

$$\widehat{var}(\hat{t}_{y.GREG}) = N^2(1 - \frac{n}{N})\frac{1}{n}\widehat{var}(g'\hat{\epsilon}) \tag{1.8}$$

Donde $g = (g_1, ..., g_n)'$ y $\hat{\epsilon} = (\hat{\epsilon}_1, ..., \hat{\epsilon}_n)'$.

1.2.5. Estimador directo de razón

Cuando hay una única variable auxiliar, la regresión pasa por el origen y el modelo es heterocedástico, con pesos $c_j = 1/x_j$ entonces el GREG es un estimador directo de razón, Torres (2009). Los valores g_j definidos en la ecuación (1.5) son en este caso constantes para todas las observaciones $j = 1, ..., n$ y vienen dados por

$$g = 1 + \frac{X_T - \hat{t}_{x.HT}}{\hat{t}_{x.HT}}$$

Además, de la ecuación (1.2) se deduce que

$$\hat{\beta}_R = \frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j x_j}$$

El subíndice R significa "directo de razón"y entonces $\hat{t}_{y.R} = X_T \hat{\beta}_R = (\sum_{j=1}^{N} xj)R$. En estadísticas oficiales es frecuente expresar este estimador como

$$\begin{aligned} \hat{t}_{y.R} &= \frac{\sum_{j=1}^{N} x_j}{\sum_{j=1}^{n} w_j x_j} \sum_{j=1}^{n} w_j y_j \\ &= \frac{X_T}{\hat{t}_{x.HT}} \hat{t}_{y.HT} \\ &= (FE)\hat{t}_{y.HT} \end{aligned}$$

Donde FE es el factor de elevación que no depende de la variable a estimar. Obsérvese que coincide con los $g's$ del estimador GREG. Si el dominio fuera pequeño, de modo que haya pocas observaciones de la muestra, se vuelve muy inestable. Se trata de un estimador directo que utiliza solamente información de su propio dominio. Su varianza es de orden $O(1/n)$, por lo tanto, bastante grande. Se obtiene como caso particular de la expresión (1.8), de la que se deduce

$$\widehat{var}(\hat{t}_{y.R}) \approx N^2(1 - \frac{n}{N})\frac{1}{n}(\frac{X_T}{\hat{t}_{x.HT}})^2 \widehat{var}(\hat{\epsilon})$$

Donde $\widehat{var}(\hat{\epsilon})$ es la varianza muestral de los residuos del modelo $y_j = \beta\ x_j + \epsilon_j$, desde $j = 1, ..., n$ con $var(\epsilon_j) = \sigma^2 x_j$. Es decir, se obtienen directamente al calcular $\hat{\epsilon_j} = y_j - \hat{y}_j = y_j - x_j\ \hat{\beta}_R$. Como el sesgo se considera prácticamente nulo EUSTAT (2001), el error cuadrático medio se aproxima por su varianza, es decir $ECM(\hat{t}_{y.R}) \approx var(\hat{t}_{y.R})$. En este caso el estimador de su coeficiente de variación se estima mediante la expresión

$$\widehat{c.v.}(\hat{t}_{y.R}) = \frac{\widehat{e.e.}(\hat{t}_{y.R})}{\hat{t}_{y.R}}$$

Donde el error estándar es $\widehat{e.e.}(\hat{t}_{y.R}) = \sqrt{\widehat{var}(\hat{t}_{y.R})}$.

1.2.6. Error cuadrático medio (ECM)

Para todos los estimadores anteriores, el ECM puede ser calculado con la misma fórmula, para un muestreo aleatorio simple o sistemático teniendo en cuenta las correcciones respectivas para cada uno de ellos. Considerando los cálculos de Ruiz (1991) y Tillé (2002), se deduce sólo para el del GREG pero es similar para los demás.

$$\begin{aligned} ECM(\hat{t}_{y.GREG}) &= E((\hat{t}_{y.GREG} - Y)^2) \\ &= E((\hat{t}_{y.HT} + (X - \hat{t}_{x.HT})'\ \hat{\beta}_{GREG} - Y)^2) \\ &= E(\hat{t}^2_{y.HT} + (X - \hat{t}_{x.HT})'^2\ \hat{\beta}^2_{GREG} + 2\ \hat{t}_{y.HT}(X - \hat{t}_{x.HT})' \\ &\quad \hat{\beta}_{GREG} - 2\ \hat{t}_{y.HT} Y - 2(X - \hat{t}_{x.HT})'\ \hat{\beta}_{GREG}\ Y) \end{aligned}$$

Como la esperanza es lineal se distribuye sobre la suma y haciendo las simplificaciones correspondientes se tiene que

$$\begin{aligned} ECM(\hat{t}_{y.GREG}) &= var(\hat{t}_{y.HT}) + \hat{\beta}^2_{GREG}\ var(\hat{t}_{x.HT}) - 2\ \hat{\beta}_{GREG}\ cov(\hat{t}_{x.HT}, \hat{t}_{y.HT}) \\ &= N^2(1 - \frac{n}{N})\frac{1}{n}(S_y^2 + S_x^2 - 2\ S_{xy}) \end{aligned}$$

Que puede estimarse sin ningún problema por

$$\hat{ECM}(\hat{t}_{y.GREG}) = N^2(1 - \frac{n}{N})\frac{1}{n}(s_y^2 + \hat{\beta}^2_{GREG}\ s_x^2 - 2\ \hat{\beta}_{GREG}\ s_{xy})$$

Para el caso del estimador de razón únicamente se cambia la β_{GREG} por $R = Y/X_T$ que puede estimarse por $\hat{R} = \hat{t}_{y.HT}/\hat{t}_{x.HT}$. Para el de diferencia hacemos $\beta_{GREG} = 1$.

Como el ECM se puede descomponer en

$$ECM(Y) = var(Y) + sesgo^2(Y) \tag{1.9}$$

Entonces el $ECM(\hat{t}_{y.HT}) = var(\hat{t}_{y.HT})$ debido a que es insesgado.

De (1.9) se puede concluir que el estimador de menor ECM coincidirá con el de varianza mínima, por lo que para cálculos posteriores sólo se referire al error estándar.

Capítulo 2

PROPUESTA DE UN NUEVO GREG

En este capítulo se presenta una generalización del estimador GREG para minimizar el error estándar. Y después se realiza la programación del mismo, explicando las restricciones de las variables a ingresar.

2.1. Propuesta de un GREG*

Como fue señalado en el capítulo anterior, el GREG coincide con el estimador de regresión y de razón. La c_j que logra lo anterior se puede reescribir de esta forma $c_j = 1/x_j^z$, donde z pertenece a $[0, \infty)$.

Sustituyendo c, se obtienen los estimadores de regresión y de razón para z igual a 0 y 1 respectivamente. Entonces las nuevas fórmulas dadas por (1.6),(1.7) y (1.5) para T, R y g quedan de esta forma:

$$\hat{T} = \sum_{j=1}^{n} \frac{w_j x_j' x_j}{x_j^z}$$

$$\hat{R} = \sum_{j=1}^{n} \frac{w_j x_j' y_j}{x_j^z}$$

$$\hat{g}_j = 1 + \frac{(X_T - \hat{t}_{x\_HT})' \, T^{-1} x_j}{x_j^z}$$

Debido a que sólo se está haciendo el cambio $c_j = 1/x_j^z$, se conservan todas las propiedades del estimador, es decir, su esperanza, su varianza, etc. No es necesario repetir esos cálculos, además se pueden ver en Särndal et al (1992) capítulo 6, haciendo el respectivo cambio de variable.

Para comprobar que efectivamente se pueden escribir a los estimadores de regresiíon y de razón a partir de esa sustitución, se tiene usando la ecuación (1.4) del GREG que con $z = 1$

$$\begin{aligned}
\hat{t}_{y.GREG} &= \sum_{j=1}^{n} w_j g_j y_j \\
&= \sum_{j=1}^{n} w_j \left(1 + \frac{(X_T - \hat{t}_{x.HT})\, T^{-1} x_j}{x_j^z}\right) y_j \\
&= \sum_{j=1}^{n} w_j \left(1 + \frac{(X_T - \hat{t}_{x.HT})}{\sum_{i=1}^{n} w_i x_i}\right) y_j \\
&= \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i x_i} \sum_{j=1}^{N} x_j \\
&= \frac{\hat{t}_{y.HT}}{\hat{t}_{x.HT}} X_T \\
\hat{t}_{y.GREG} &= \hat{t}_{y.R}
\end{aligned}$$

Obteniendo así el estimador de razón. Para el de regresión no hay nada que hacer, en la fórmula (1.3) se observa que tienen la misma forma, la diferencia radica en el coeficiente β, pero dado que es una proporción, cuando $z = 0$ coincide.

En las fómulas anteriores surge una pregunta natural, ¿Existirá una z en particular que minimice el error estándar? Debido al exponente que tiene la variable x no es posible encontrar de forma analítica la z que minimice al e.e. por lo que ese problema se tendrá que resolver de forma numérica. A este nuevo GREG se le distinguirá con un *, es decir, al estimador con z que minimice el e.e. se le llamará GREG* en este proyecto, con la siguiente fórmula:

$$\hat{t}_{y.GREG^*} = \sum_{j=1}^{n} w_j \left(1 + \frac{(X - \hat{t}_{x.HT})\, T^{-1} x_j}{x_j^{z^*}}\right) y_j$$

donde z^* es aquella que minimiza a (1.8). En un principio $z \in \mathbb{R}^+$ por lo que el error estándar del GREG depende de z

$$e.e.(z) = \sqrt{N^2(1 - \frac{n}{N})\frac{1}{n}\widehat{var}\left(\left(1 + \frac{(X_T - \hat{t}_{x.HT})\, T^{-1} x_j}{x_j^z}\right)\hat{\epsilon}\right)}$$

entonces z^* es aquella que cumpla lo siguiente

$$e.e.(z^*) = \min_z e.e.(z)$$

¿Por qué se puede garantizar que existirá una z que minimice el e.e.? ¿Será ésta única? ¿Qué tan eficiente resultará esa estimación? Son algunas de las preguntas que se plantean debido a que no se puede buscar a z de forma analítica. Pero muy intuitivamente y sin una gran demostración, se puede ver que si z tiende a infinito entonces $\hat{T}$ sería prácticamente cero, lo que indeterminaría a β y no tendría sentido ese estimador por lo que en realidad z no pertenece a $[0, \infty)$, sino más bien a $[0, M]$ pero como M varía dependiendo de los datos, no es posible encontrar ese valor analíticamente. Y por otro lado cuando $z \to \infty$ no podría minimizar el e.e. ya que por la misma razón $\hat{g}_j$ se convierte en una función creciente a partir de un cierto valor para z. Con lo anterior se deduce que el intervalo para valores de z es finito, por lo que se puede garantizar que existe un valor mínimo que toma el e.e. al variar z y este es estrictamente positivo ya que x es positiva.

La tarea ahora es encontrar ese valor y debido a que sólo se puede resolver numéricamente, se tiene que diseñar un programa que calcule esa z que minimiza el e.e.

Otra forma para la c_j, es considerarla como $var(y_j) = \sigma_j^2$, que es un vector, es decir, tiene un valor específico para cada entrada en x, por lo que también se considera el caso cuando se conoce el vector σ^2. Este estimador se denotará por Sig., es decir, el estimador GREG al cual se le proporcione el vector σ^2 se distinguirá con el subíndice Sig.

Cabe aclarar antes de comenzar el programa que la elaboración de éste es específico para un muestreo aleatorio simple o un muestreo sistemático y de un solo estrato, ya que con otro tipo de

diseño se tendrían que hacer las adaptaciones correspondientes a las fórmulas, por lo que hay que tener cuidado.

2.2. Programación del GREG*

Para poder hacer una buena programación, se tiene que diseñar el algoritmo para tenerun programa eficiente, por lo que es de gran ayuda clarificar el objetivo.

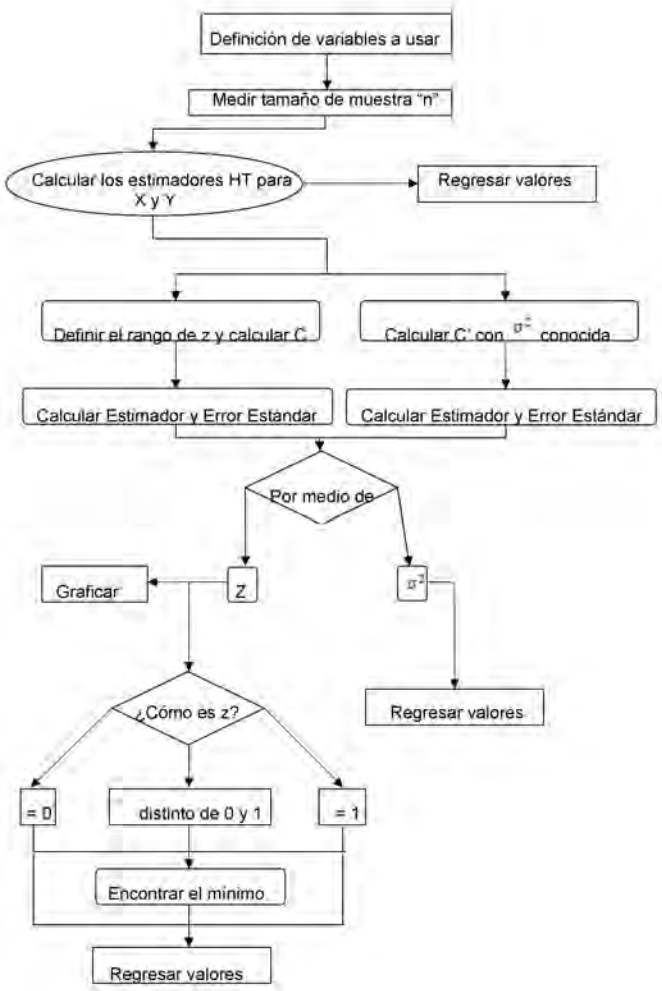
2.2.1. Algoritmo del GREG*

El objetivo es estimar el total de Y a partir de una variable auxiliar llamada X, de la cual se cuenta con el total, por lo que se necesitan las variables Y, X, XT y N donde XT es el total de la variable X y N es el tamaño de la población.

Debido al tipo de diseño (aleatorio simple o sistemático) todos los pesos muestrales son iguales y se calculan dividiendo el número total de elementos en la población entre los de la muestra, por lo que la primera variable que se pide es el total de individuos de la población N.

Ya que los estimadores de HT se pueden calcular con los datos anteriores, entonces no se necesita ningún dato más, ya que la muestra se encuentra en las variables X y Y.

El diagrama del algoritmo queda de la siguiente manera:



Una vez ya teniendo una estructura a seguir del programa, que es como si se tuvieran los planos para construir una casa, se procede a la elaboración del mismo.

2.2.2. Programa en MATLAB para el estimador

El programa se realiza en el paquete de computación científica MATLAB por tener algoritmos eficientes al realizar ciertas operaciones y sus comandos están basados en lenguaje C por lo que

hace más sencillo programar.

Dentro del programa se comienza por definir la función

$function\ [RES, Z] = Ygm(XT, X, Y, N)$

Lo que está escrito después de la palabra function se copia en la ventana de comandos de MATLAB para que corra el programa.

Está diseñado de tal forma que las variables se ingresan como vectores columna, a excepción de X que puede ser una matriz, esto para facilitar el trabajo. Para un mejor resultado se recomienda leer el apéndice referente al programa para poder conocer las restricciones necesarias, ya que si no se ingresan los datos adecuadamente puede que los resultados arrojados no sean correctos.

Se pudo haber desarrollado prácticamente en cualquier leguaje de programación, ya que no requiere de funciones especiales ni de cálculos muy complejos, pero se recurre a un programa porque es realizar muchas veces la misma operación y si en dado caso se quisiera modificar algún dato, esto lo facilitaría.

También se realiza un programa en MATLAB para simular, que se anexa en el apéndice, pero es exclusivamente para el ejemplo de la población de médicos por condado en E.U. En el siguiente capítulo se hacen las aclaraciones correspondientes a los ejemplos.

El objetivo de simular estos datos es conocer la eficacia del nuevo GREG* que se puede medir mediante el número de veces que resulta ser el estimador con menor e.e. por medio de gráficas y tablas de frecuencias, es decir, no se pretende hacer simulaciones para estimar el total de la población, sino para conocer quién es el que obtiene menor variabilidad en la estimación.

Estos programas sirven tanto para el caso univariado como para el multivariado, claro que el caso univariado no tiene ningún problema en expresarse de forma multivariada.

Después de revisar los apéndices para conocer de qué forma trabaja el programa se pueden verificar dos cosas:

- Que el estimador propuesto realmente sea eficiente.
- Que el programa corra como es de esperarse.

Capítulo 3

ESTIMACIÓN NUMÉRICA Y SIMULACIÓN

En este capítulo se hace un ejemplo numérico para verificar el funcionamiento del estimador GREG*, y para comprobar su eficacia se hace una simulación vía Montecarlo.
Una vez que se tiene la teoría el siguiente paso es llevarlo a la práctica.

3.1. Ejemplos con muestreo aleatorio simple (m.a.s.)

Se usa la base de datos counties.dat (tomada del libro de Lohr), en el libro señala que la base es "... una muestra aleatoria simple de 100 de los 3141 condados de Estados Unidos (Oficina de Censos de 1994) ... la población total en 1993 fue estimada en 255,077,536, el área total de Estados Unidos es de 3,536,278 millas cuadradas".

3.2. Ejemplo univariado para una población de médicos por m.a.s.

Para el ejemplo únicamente se usa como variable dependiente a "physicia"que es el número de médicos por condado, y como auxiliar a "totpop"que es el total de la población por condado, tomadas de la base counties.dat.

Se necesita una buena relación de la variable independiente con la dependiente para obtener resultados favorables, entonces hay que verificar este supuesto tan importante antes de realizar la estimación.

En la figura 3.1 se graficaron el total de la población v.s. el número de médicos por condado, a

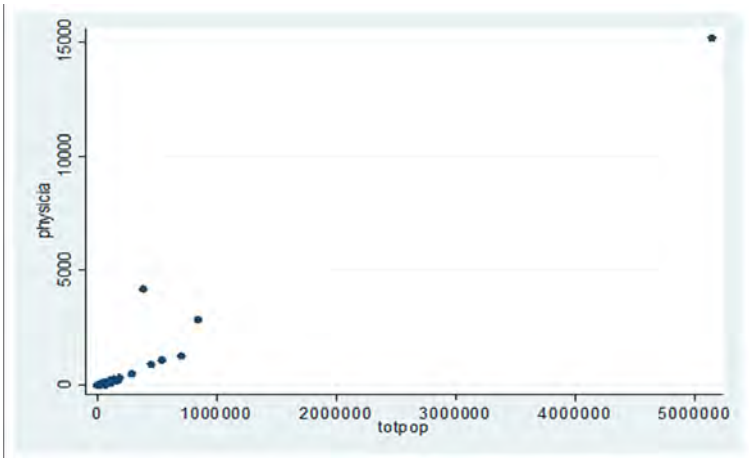


Figura 3.1: Gráfica de número de habitantes v.s. número de médicos, por condado.

simple vista se ve que existe un outlier, se puede pensar que guarda buena proporción con respecto a los demás datos, pero haciendo una observación más detallada esto no es así, ya que si se traza una línea con la dirección que lleva la masa de los datos, se observa en la figura 3.2 que el outlier (en el registro 19, totpop = 5,139,341 queda bastante lejos de esa "buena proporción".

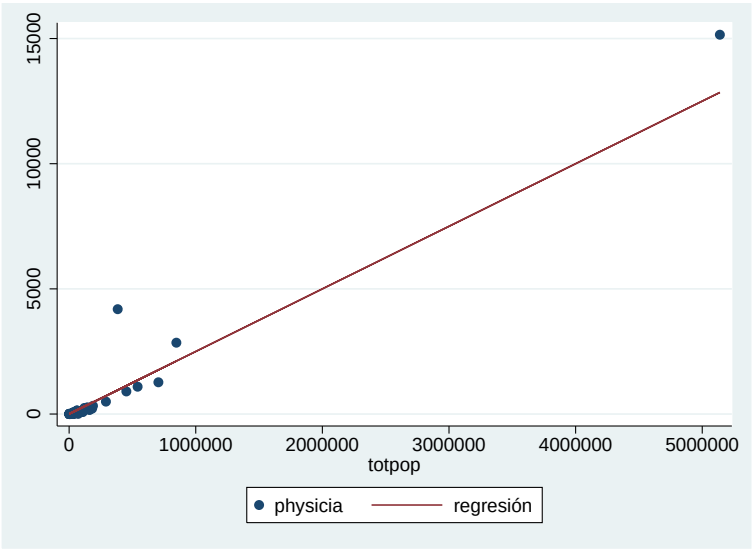


Figura 3.2: Gráfica de número de habitantes v.s. número de médicos, con ajuste de recta

Por lo que se elimina ese dato para garantizar la "buena proporción", y que la estimación sea más confiable, entonces la muestra consiste en eliminar dicha observación de la base original.

3.2.1. Estimación numérica

Ya que se tienen todos los elementos para realizar la estimación, se introducen los datos. Las estimaciones se presentan en el cuadro 3.1

Estimador	**Pob. Tot. Estimada**	**e. e.**
HT	462,076	165,899
Regresión	582,763	143,030
Razón	553,332	138,039
GREG*	539,800	137,118
Sig.	546,956	117,614

Cuadro 3.1: Estimación del total de médicos en E.U. ejemplo univariado

El valor de z para el GREG* en este ejemplo fue de 1.24. (Poco arriba del estimador de razón, el cual es $z = 1$)

Con los resultados mostrados en el cuadro 3.1 se deduce que el estimador de menor e.e. fue el Sig., se usó $c_j = v\hat{a}r(y_j)$ para calcular el estimador. Como en todo experimento controlado, se conocen los resultados verdaderos, en este ejemplo el número total de médicos que había en E.U. en 1993 era de 532,638 médicos, con lo que el GREG* es el más cercano. Por lo que se recurre al uso de la simulación para verificar si en verdad el Sig. tiene siempre menor e.e. que el GREG*. En la figura 3.3 se muestra el comportamiento del estimador conforme fue variando la z.

Para este ejemplo en particular se observa que conforme la z se fue incrementando, el valor del estimador GREG* coincide con el HT, no sólo en valor de la población, sino también en e.e., esto no necesariamente debe de suceder, pero muestra la gran generalización que puede tener cuando existe una buena relación entre las variables a analizar.

3.2.2. Simulación

Para este ejemplo es muy útil la simulación, ya que como se tomó la muestra del libro de Lohr (2000), no se pueden tomar más muestras para verificar que en general el GREG* obtendrá menor

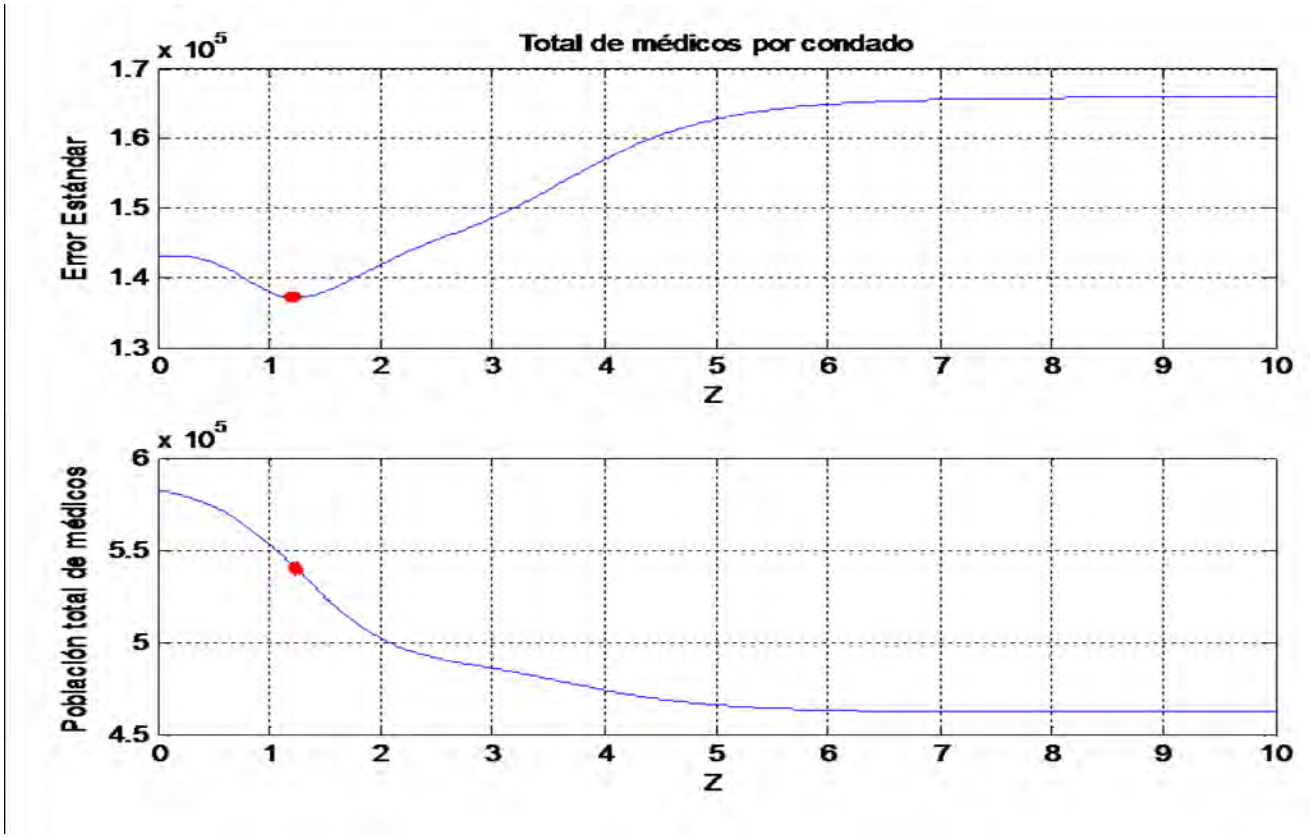


Figura 3.3: Comportamiento del e.e. y del GREG* al variar z, ejemplo univariado

e.e. en lugar del Sig. como lo muestra el ejemplo anterior.

Se simulan el número de médicos por condado con respecto a un dato proporcionado por la oficina de censos de los Estados Unidos, el cual señala que een el año de 1993, los porcentajes de médicos por número de habitantes por condado fueron, 0.91 % para el concdado con mayor porcentaje y cero para el menor.

Como no se sabe de alguna distribución que proporcione el comportamiento de los médicos con respecto al número de habitantes, se usa una uniforme [0 , 0.0041 * X] donde X es la variable auxiliar, número de habitantes por candado.

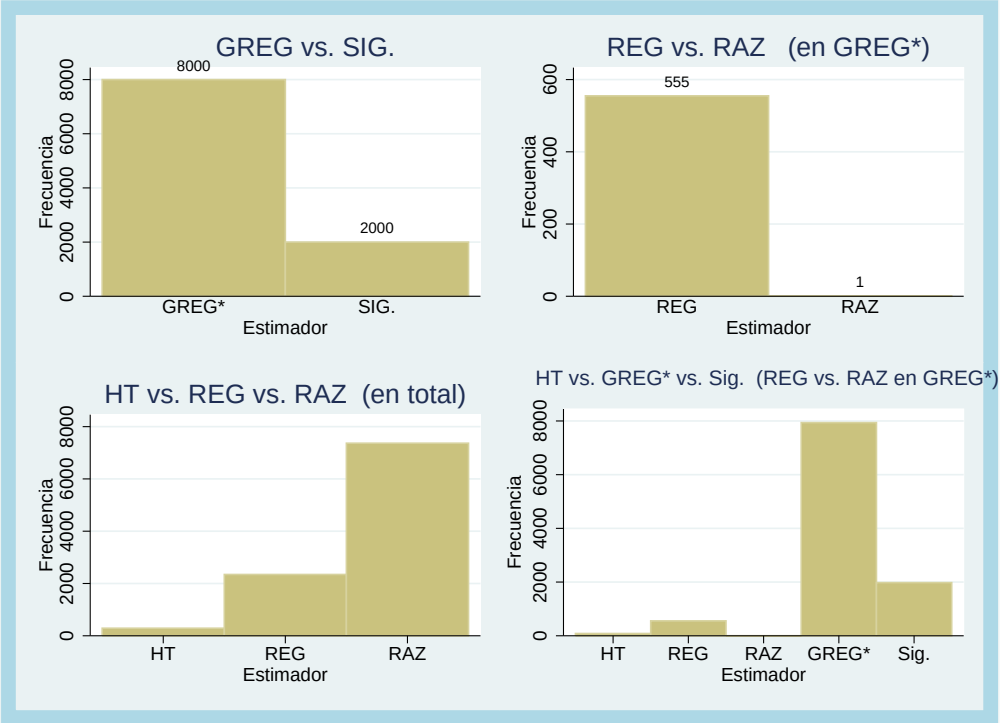


Figura 3.4: Frecuencias del menor e.e. para la simulación de cada estimador, ejemplo univariado

En la figura 3.4 se muestra el número de veces que el estimador obtuvo el menor e.e.

En la primera gráfica (GREG vs. Sig.) que es la de mayor interés, se observa que de las 10,000 iteraciones, en 8,000 el GREG* obtuvo menor e.e. que el Sig.

La gráfica REG vs. RAZ (en GREG*) muestra que del total de las 8,000 veces en las que el estimador GREG* obtuvo el menor e.e. 555 veces coincidió con el de regresión y sólo una vez con el de razón, con lo anterior se podría pensar que en las 10,000 iteraciones, el de regresión fue mejor

que el de razón pero para eso está la siguiente gráfica HT vs. REG vs. RAZ (en total) la cual indica quien obtuvo un menor e.e. en las 10,000 simulaciones, sin tomar en cuenta al GREG* ni al Sig.; y se observa que el de razón obtuvo más veces un menor e.e. con respecto a los otros dos estimadores.

La última gráfica muestra la proporción que tienen todos en total, pero los valores para el de regresión y para el de razón son cuando el GREG* coincide con ellos.

El resumen de la información de la figura 3.4 aparece en el cuadro 3.2.

Estimador	**Simulaciones con el menor e.e.**	**Porcentaje**
HT	84	0.84
Regresión	555	5.55
Razón	1	0.01
GREG*	7,940	79.4
Sig.	1,980	19.8

Cuadro 3.2: Porcentaje del número de veces que obtuvo el menor e.e. cada estimador en la simulación, ejemplo univariado

Existe un excedente de 560 simulaciones debido al diseño del GREG* el cual puede coincidir con otros estimadores, por lo que se duplicó el conteo.

Se debe de tener en cuenta que cuando se hace simulación, no siempre van a dar los mismos resultados, es decir, puede variar la tabla anterior, pero no cabe ninguna duda al decir que el mejor estimador es el GREG*.

3.3. Ejemplo multivariado para una población de médicos por m.a.s.

Para este ejemplo se usan las mismas variables que en el caso univariado pero la parte auxiliar no sólo dependerá de la población sino que también considerará el área de territorio para cada condado, ya que se cuenta también con su total, el cual sería el área de los Estados Unidos (3,536,278 millas cuadradas), en la base counties.dat ésta variable se llama “landarea”.

Como en el ejemplo anterior se verificó el supuesto de "buena relación" entonces se realiza directamente la estimación.

3.3.1. Estimación numérica

Algo que se tiene que modificar al ingresar los datos es que ahora la auxiliar es una matriz y la variable de los totales XT se tiene que introducir como vector columna, aunque estas especificaciones se hicieron en el capítulo anterior.

Una vez ingresando los datos, se obtuvieron los resultados mostrados en el cuadro 3.3

Estimador	**Pob. Tot. Estimada**	**e. e.**
HT	462,076	165,899
Regresión	609,506	168,687
Razón	554,164	138,489
GREG*	533,172	127,522
Sig.	586,906	149,677

Cuadro 3.3: Estimación del total de médicos en E.U. ejemplo multiivariado

El valor de z para el GREG* en este ejemplo fue de 3.33.

En el cuadro 3.3 se observan varias cosas muy interesantes, una de ellas es que en este ejemplo multivariado, resultó ser precisamente el GREG* el que tuvo el menor e.e., algo más, es que el total de médicos estimado por el GREG* usando como variables auxiliares, el número de habitantes y el área total por condado, únicamente tiene un error de 535 médicos por encima del valor verdadero, obviamente en la realidad no se podría saber esto, pero como es un ejemplo controlado se tiene esa información.

Otro dato importante a destacar es que como era de esperarse el HT tomó el mismo valor que en el ejemplo univariado, pero los otros estimadores no mejoraron, al contrario, esto hace resaltar aun más la eficacia que tiene el GREG*.

Aún falta hacer la simulación para garantizar que efectivamente el GREG* siempre tenga menor

e.e. En la figura 3.5 se muestra la gráfica que resultó al variar la z.

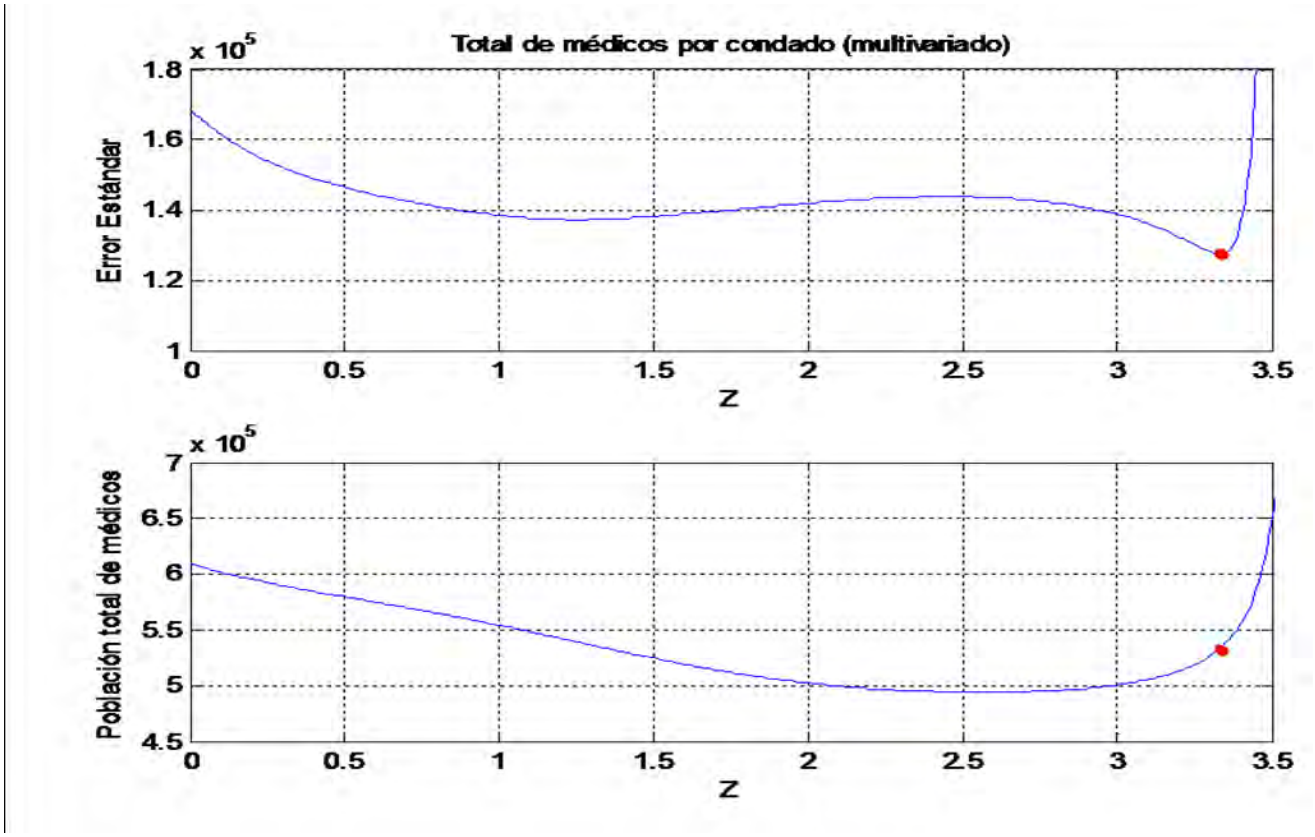


Figura 3.5: Comportamiento del e.e. y del GREG* al variar z, ejemplo multivariado

3.3.2. Simulación

Aunque en el ejemplo univariado fue contundente que el GREG* obtenía el menor e.e. más veces que los otros estimadores, se espera que para este ejemplo, la diferencia sea aún más notoria.

Al usar el mismo razonamiento que en el ejemplo anterior para simular el número de médicos por condado los resultados fueron

Las gráficas de la figura **??** muestran el número de veces que el estimador obtuvo el menor e.e.

Como se había previsto, en la primera gráfica se ve que es muy clara la mejoría que tiene el GREG* con respecto al Sig., este último de las 10,000 simulaciones sólo dos veces obtuvo un menor e.e..

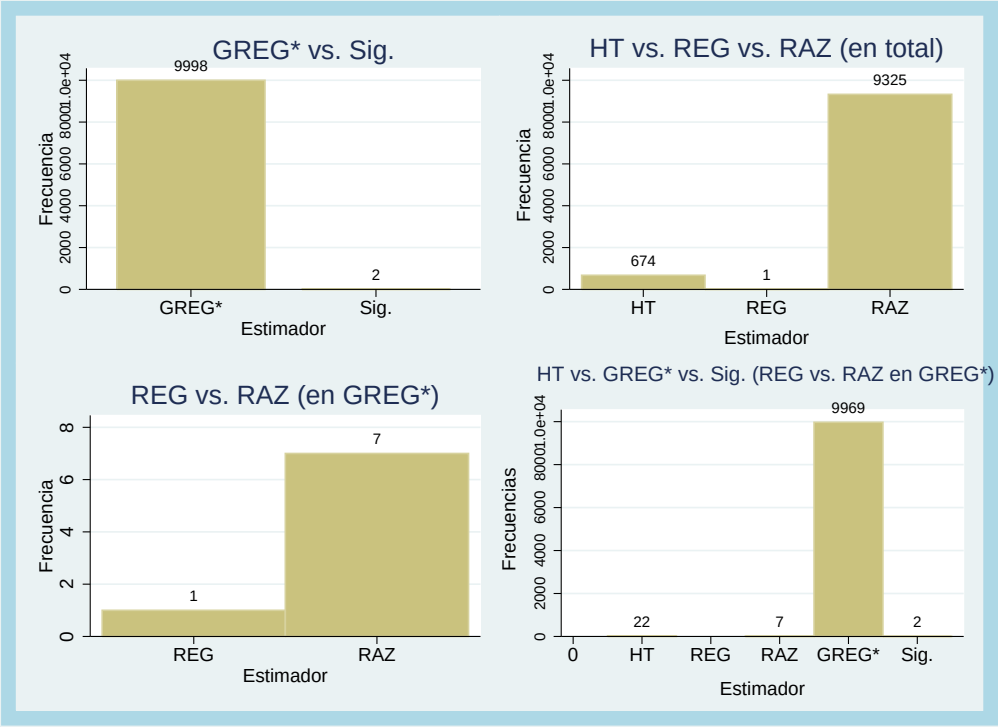


Figura 3.6: Frecuencias del menor e.e. para la simulación de cada estimador, ejemplo multivariado

En la gráfica de REG vs. RAZ se observa que en esta ocasión fue al revés del caso univariado, pero el número de veces que el de razón coincidió con el GREG* es muy insignificante, solo cuenta son 7 veces de las 10,000 simulaciones.

En la tercera gráfica, HT vs. REG vs. RAZ, fue el estimador de razón el que obtuvo mayor número de veces el menor e.e., el HT sólo lo tuvo 674 de las 10,000 simulaciones.

Por último se ve en la cuarta gráfica que prácticamente el GREG* obtuvo el menor e.e. en las 10,000 simulaciones salvo que el HT obtuvo 22 veces, el Sig. fue menor en 2 ocasiones, y sólo en 7 veces coincidió el de razón con el GREG*. Se puede notar que esta vez el de regresión en ningún momento obtuvo un menor e.e. que el resto de los estimadores, de hecho, cuando se estimó el total de médicos fue precisamente el que obtuvo el mayor e.e.

Aunque es clara la diferencia, se muestran los resultados en el cuadro 3.4, igual que en el ejemplo univariado.

Estimador	Simulaciones con el menor e.e.	Porcentaje
HT	22	0.22
Regresión	0	0
Razón	7	0.07
GREG*	9,987	99.87
SIg.	2	0.02

Cuadro 3.4: Porcentaje del número de veces que obtuvo el menor e.e. cada estimador en la simulación, ejemplo multivariado

3.4. Estimación por simulación

En los ejemplos pasados se desarrolló la simulación para determinar qué estimador obtenía más veces el menor e.e., no para comparar los estimadores del total de médicos. Como no era el objetivo, no se entra mucho en detalle pero se presentan en el cuadro 3.5 los valores que se obtuvieron para cada estimador vía simulación pero únicamente para el caso univariado.

Estimador	Total de médicos vía simulación
Regresión	542,615
Razón	535,500
GREG*	532,934
Sig.	517,759

Cuadro 3.5: Valor promedio de las estimaciones del total de médicos en E.U. vía simulación

El estimador GREG* en el caso univariado estima el total de médicos de una forma bastante buena, parecida al caso multivariado, lo cual era de esperarse.

Esta estimación vía simulación, sólo es el número promedio de médicos que se obtuvieron de las 10,000 iteraciones.

3.5. Ejemplos con muestreo sistemático (m.s.)

Aunque la muestra obtenida para hacer los ejemplos anteriores fue obtenida de datos reales, no deja de ser un ejemplo de libro, por lo que ahora se realizarán otros dos ejemplos similares a los anteriores pero ahora con datos del Censo del 2000 y Conteo 2005 realizados por el INEGI.

El muestreo sistemático consiste en dividir la población de N unidades en n subgrupos de k elementos , y tomar de esos subgrupos al azar uno de ellos, como la muestra; Méndez (2009).

Se calcula k como el cociente $k = [N/n]$

Se obtiene un número aleatorio entre 1 y k. Se seleccionan en la muestra los elementos $i, i + k, i + 2k, i + 3k, ...i + (n - 1)k$. Esto equivale a partir la población en n conjuntos de k elementos; y seleccionar uno de ellos.

El muestreo sistemático procede al tomar al azar un número entre 1 y k, sea i entonces los elementos de la muestra son: $u_i, u_{i+k}, u_{i+2k}, ..., u_{i+(n-1)k}$.

Para fines de estimación se tiene que la probabilidad de seleccionar en la muestra cualquier elemento o unidad es la misma que para el m.a.s..

No hay expresiones válidas para estimar la varianza del estimador cuando se usa el muestreo sistemático en poblaciones que no tienen orden aleatorio.

Si la población tiene un orden que se refleja en cambios periódicos de los valores de y_i, el muestreo sistemático puede producir varianzas mayores. En este caso el problema es que la muestra puede coincidir con valores todos bajos (o altos) de y_i, siendo de esta manera poco representativa y con fluctuaciones fuertes de muestra a muestra Esto puede suceder cuando la población consiste en los volúmenes de ventas de una tienda en un periodo de tiempo. Otro ejemplo es en el muestreo de plantas cultivadas (maíz, trigo, etcétera) en donde ciertas áreas del terreno se riegan en un día determinado y otras áreas otro día.

En resumen el muestreo sistemático es una herramienta delicada que puede ser mejor, igual o peor que el m.a.s..

Por lo que la base se ordenará con respecto a la tasa de crecimiento de cada municipio, ya que

a final de cuentas es precisamente lo que se quiere estimar. Se elaboró un programa para tomar dicha muestra, el cual se anexa en el apéndice junto con la muestra resultante.

3.6. Ejemplo univariado para la población de México en el 2005 por m.s.

Para este caso se va a estimar la población total de México para el año 2005, usando como variable dependiente "pop05"que es el número de habitantes por municipio en el 2005 y como auxiliar "pop00"que es del censo del 2000. El total de pobladores en el 2000 era de 95,753,396 personas, en 2442 municipios.

La relación que tienen las dos variables es bastante buena, se puede constatar en la figura 3.7.

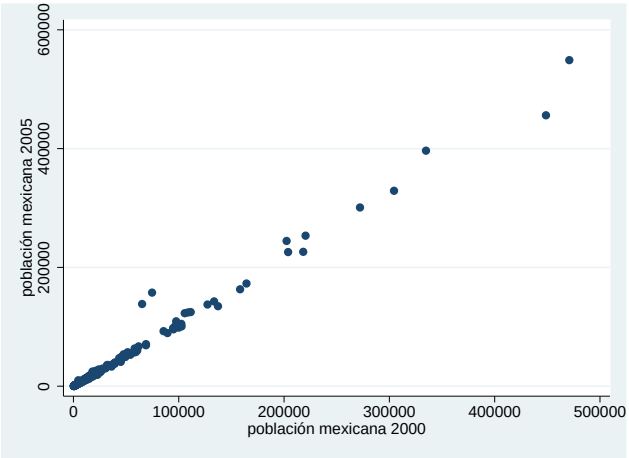


Figura 3.7: Numéro de habitantes por municipio, año 2000 v.s. año 2005

3.6.1. Estimación numérica

Ingresando la muestra (por muestreo sistemático) del número de habitantes por municipio para el año 2000 en México, en el programa se obtienen los resultados del cuadro 3.6

En la figura 3.8 se muestra en las gráficas el comportamiento de las estimaciones y del e.e. al variar z. Para $z = 0.49$ se minimiza el e.e.

Estimador	Pob. Tot. Estimada	e. e.
HT	101,738,771	16,094,915
Regresión	103,297,604	1,751,325
Razón	103,242,257	1,766,149
GREG*	103,274,251	1,729,195
Sig.	103,054,421	2,994,175

Cuadro 3.6: Estimación del número total de habitantes en México para el año 2005, ejemplo univariado

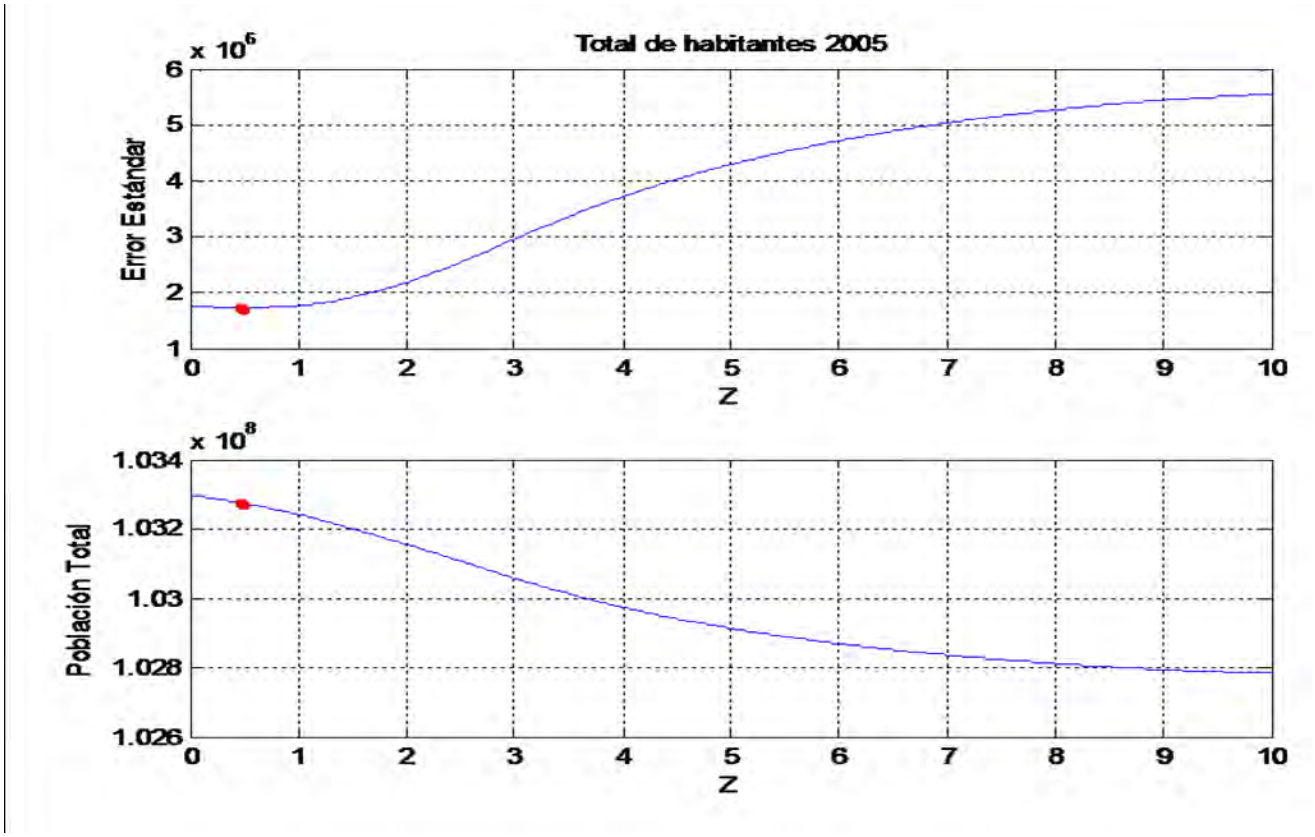


Figura 3.8: Comportamiento del e.e. y del GREG* al variar z, ejemplo univariado

En el cuadro 3.6 se observa que el de menor e.e. fue el GREG*, cabe señalar que el INEGI estimó la población en 103,263,388 personas, la diferencia que existe entre el valor de estas dos estimaciones, puede deberse al muestreo realizado, ya que lo más probable es que el INEGI haya usado otra técnica más efectiva y que las calibraciones sean más exactas, a diferencia del realizado aquí, el cual tiene el mismo factor de expansión para cada observación. Pero esas son buenas noticias ya que si con un muestreo más simple se obtienen esos resultados, eso significa que si se mejoran los factores de expansión entonces la estimación será mucho mejor. No se intenta corregir ese problema debido a que no es el tema de la tesis.

3.7. Ejemplo multivariado para la población de México en el 2005 por m.s.

El problema que se presenta para este ejemplo es que no se cuenta con otras variables que puedan ser de gran ayuda como lo sería el área territorial por municipio, comercio, etc. Por lo que se tuvo que usar la clasificación por sexo. No se intenta recopilar el área de cada municipio para el estudio ya que no entra en el objetivo de la tesis.

3.7.1. Estimación numérica

Ahora se tienen las variables "pop00", "poh00", y "pom00", que son la población total, población de hombres y población de mujeres por condado respectivamente, como matriz auxiliar; los resultados se muestran en el cuadro 3.7.

Estimador	**Pob. Tot. Estimada**	**e. e.**
HT	101,738,771	16,094,915
Regresión	104,022,587	7,416,858
Razón	103,889,077	6,206,860
GREG*	103,233,851	1,927,584
Sig.	103,608,926	3,196,345

Cuadro 3.7: Estimación del número total de habitantes en México para el año 2005, ejemplo multivariado

Para $z = 4.66$ se minimiza el e.e.

Se puede observar que haciendo la comparación con el ejemplo univariado, ningún estimador mejoró, al contrario, eso debido a que las variables anexadas no brindan mucha ayuda, pero aun así se ve que el GREG se comporta de una forma muy consistente, lo cual habla muy bien de esa nueva forma de calcularlo.

En la figura 3.9 se muestra el comportamiento que se tuvo al variar z.

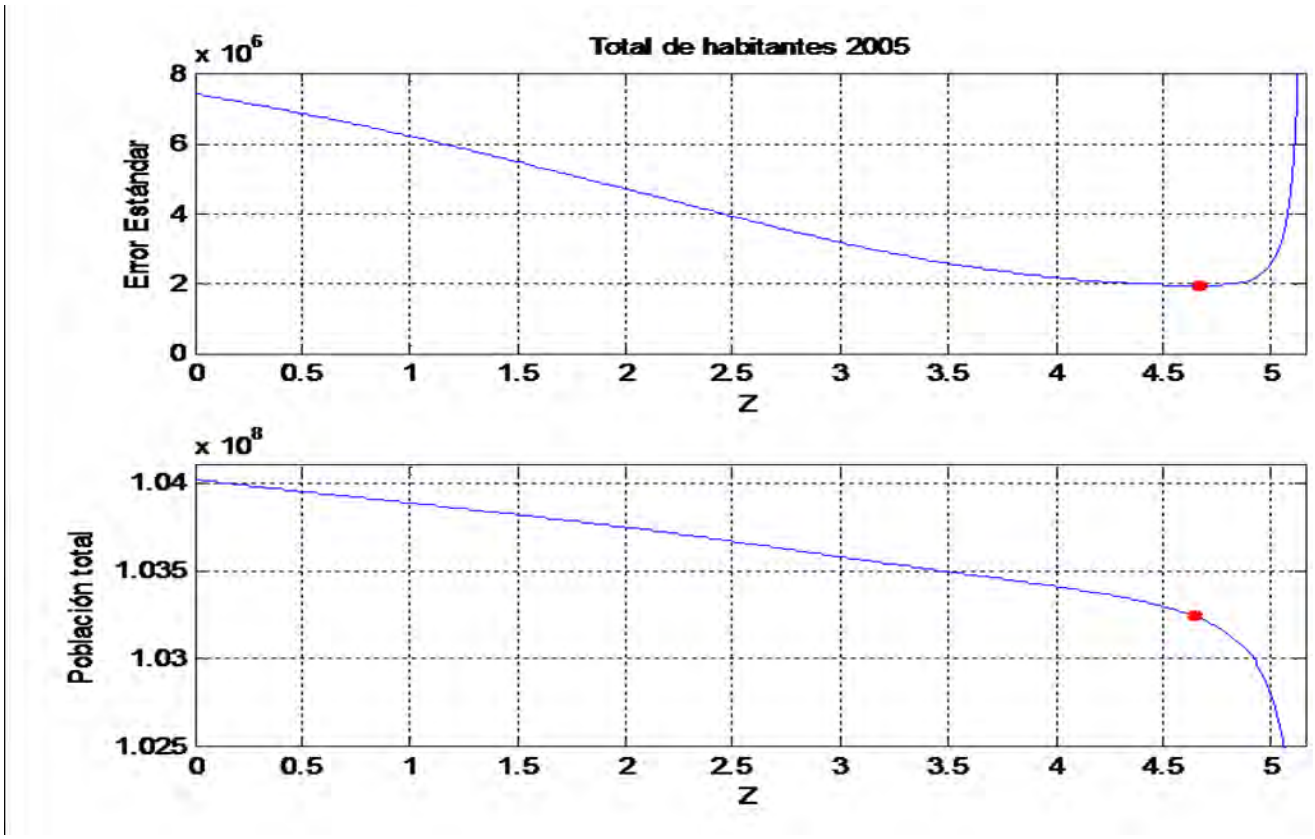


Figura 3.9: Comportamiento del e.e. y del GREG* al variar z, ejemplo multivariado

Para estos ejemplos no se recurre a la simulación debido a que el programa creado para simular únicamente funciona cuando no se tiene una distribución asociada al comportamiento de las variables, por lo que en su lugar como se cuenta con el número de habitantes para los 2442 municipios, entonces se van a tomar las 10 muestras diferentes que se pueden obtener bajo el muestreo sistemático, sólo son 10 debido a que el tamaño de muestra es de 244, por lo que se obtiene $k = 10$.

El cuadro 3.8 representa el resumen de los e.e. para las 10 muestras, en él se puede observar

K	HT	REG	RAZ	GREG	SIG
1	13,558,306	1,714,179	1,854,471	1,631,828	**1,625,724**
2	10,833,423	1,860,106	1,955,569	1,737,684	**1,620,464**
3	21,276,329	**1,540,476**	1,828,390	**1,540,233**	2,096,737
4	18,834,288	**1,717,031**	1,972,483	**1,717,031**	1,888,245
5	16,747,919	1,927,387	1,848,438	**1,727,777**	1,928,090
6	19,244,926	1,753,326	1,741,900	**1,653,966**	1,825,263
7	23,069,370	1,460,469	1,467,543	**1,443,972**	1,855,497
8	16,771,314	**1,753,045**	1,791,950	**1,753,045**	2,110,558
9	19,101,724	2,008,508	2,282,564	**1,954,527**	2,365,508
10	16,094,915	1,751,325	1,766,149	**1,729,195**	2,994,175
total	175,532,512	17,186,882	18,509,457	16,889,259	20,310,261
promedio	17,553,251	1,718,688	1,850,946	**1,688,926**	2,031,026
Min	1,443,972				
Max	23,069,370				

Cuadro 3.8: Errores estándar para las diferentes muestras por estimador

que el GREG* fue quién obtuvo el menor e.e. tanto en promedio 1,688,926 como en general para todas las muestras 1,443,972.

CONCLUSIONES

El estimador GREG es un estimador bastante útil cuando los supuestos que necesita son satisfechos, como lo son la buena relación lineal entre la variable dependiente con las auxiliares, cuando son conocidos los totales, entre otros.

Tal vez en la práctica muchas veces no se puedan garantizar esos supuestos debido a la naturaleza de la información con la que se cuenta, pero la mayoría de las veces se puede replantear el problema y utilizar este tipo de estimadores, por señalar un ejemplo bastante sencillo, supóngase que en el ejemplo del capítulo anterior se contara con una variable auxiliar que fuera dicotómica la cual midiera si existe alguna universidad en ese condado, y otra variable que mostrara si en aquella imparten medicina, de forma muy intuitiva se puede pensar que habrá un mayor número de médicos si existe una universidad con la carrera de medicina, pero para obtener los totales de esas variables tal vez no sea tan fácil como se pudiera pensar ya que entonces se necesitaría otra base que dijera cuantas universidades existen en E.U. y además en cuantas de esas imparten medicina. En el supuesto anterior no suena tan difícil conseguir esa información si se piensa que ese registro existe y está prácticamente al alcance de cualquier persona, pero habrá casos en los que esa información no se puede conseguir y si no la se tiene entonces no se puede usar este método, una opción sería no contemplarla pero se sabe que será bastante útil para hacer una mejor estimación, entonces se tiene que recurrir a otro tipo de estimadores donde no se requiera contar con los totales.

De cualquier forma, cuando se pueda usar el estimador GREG será de muchísima ayuda.

A lo largo del proyecto se vio que siempre será mejor usar el estimador GREG que los estimadores de regresión o el de razón, la pregunta que se plantea es:

¿El beneficio que se consigue en la estimación justifica el costo computacional que se genera para calcular el GREG?

Para calcular el GREG no se tiene un costo computacional elevado, por lo que siempre será mejor

calcular el GREG para hacer la estimación si se cuenta con la información necesaria.

El GREG* que se planteó suena bastante intuitivo desde el punto de vista que se pueden escribir a los estimadores de regresión y de razón en términos del GREG, y se llega que ingresando una variable z se puede calcular cualquiera de los dos estimadores sólo modificando el valor de z, (0 para el de regresión y 1 para el de razón) y también se puede escribir al HT pero la z no será la misma para cualquier conjunto de datos como si lo es con los otros dos estimadores. Pero teniendo eso en mente, entonces se llega de forma muy natural a la pregunta:

¿Qué pasa con otros valores de z?

Al generalizar y usar un rango de valores para z se encontraron mejores estimaciones para el problema, por lo que se llegó a la conclusión que el GREG con la z que minimice el e.e. será el GREG*.

Pero el GREG*, si tiene un costo computacional bastante elevado debido a que para cada valor del rango de z se tiene que calcular un estimador, entonces se vuelve a preguntar

¿El beneficio que se consigue en la estimación justifica el costo computacional que se genera para calcular el GREG*?

Esa pregunta a simple vista pudiera pensarse que "no vale la pena", ya que en el ejemplo anterior se vio que comparando al GREG* con el de razón la diferencia aproximada es de 11,000 médicos y pensando que el total es de más de medio millón pues la diferencia es del 2 % del total lo cual hace pensar que es insignificante la mejoría que se consigue, pero gracias a la simulación se puede observar que en un porcentaje bastante alto el estimador GREG* tiene una buena aproximación, por lo que en lugar de simular el estimador de razón para encontrar una mejor estimación la cual llevaría más tiempo que calcular el GREG*, basta con calcular una única vez el GREG* y será un resultado muy confiable.

Además se puede optimizar el programa de tal forma que al restringir el rango de z no se tengan que hacer iteraciones innecesarias para encontrar la de menor e.e. y ese valor será una aproximación muy eficiente al verdadero valor. Y ya teniendo el programa optimizado entonces sin duda es de mayor beneficio calcular el GREG*.

Otra respuesta a la misma pregunta es qué día con día las computadoras cada vez son más rápidas

y los paquetes científicos tienen mejores algoritmos para poder reducir tiempo y costo a la hora de hacer un cálculo de esa índole, por lo que la respuesta a esa pregunta en este momento no será la misma que dentro de unos meses ya que el avance tecnológico crece de una forma impresionante.

El programa que se diseñó en este proyecto es bastante eficiente, salvo por el caso que aun se puede optimizar la parte del cálculo del rango de z.

Si se miran las clasificaciones que existen para los tipos de estimadores, se puede crear un estimador para cada tipo de problema que se presente, pero obviamente es bastante ineficiente estar haciendo uno particular, y en la práctica habrá ocasiones en las que difícilmente se puede decidir qué tipo será mejor, por lo que sólo puedo decir:

"*No creo en un estimador, porque ...*

$Es - timador$"

Apéndice A

Programa para calcular los estimadores

```
    function [RES,Z]=Ygm(XT,X,Y,N)

[n,m]= size(X);
w = N/n;
Z2 = 1;
Z = 1;
G = zeros(n,1);
T = zeros(m);
t = zeros(m,1);
Yb = norm(Y,1)/n;
Ty = norm(w.*Y,1);
Tx = zeros(m,1);
P = (N ∧ 2)*((1/n)-(1/N));
.   for j=1:m
.       Tx(j) = norm(w.*X(:,j),1);
.   end
.   while (Z2 == Z)
.       Z2 = Z2*10;
.       z1=linspace(0,Z2,1001);
.           for i=1:1001
.               if (i==1)
.                   C = 1/(Y-Yb).∧ 2;
```

```
.                 for j=1:m
.                     for k=1:n
.                         t(j) = t(j)+w*(X(k,j)*Y(k))*C(k);
.                     end
.                 end
.                 for i1=1:m
.                     for j=i1:m
.                         for k=1:n
.                             T(j,i1) = T(j,i1)+w*(X(k,j)*X(k,j))*C(k);
.                             T(i1,j) = T(j,i1);
.                         end
.                     end
.                 end
.             B = (T^ -1)*t;
.             Tsig=Ty+(XT-Tx)'*B;
.                 for k=1:n
.                     G(k) = 1+(XT-Tx)'*(T^ -1)*X(k,:)'*C(k);
.                 end
.              E = Y-X*B;
.              GE = G.*E;
.              GEp = norm(GE,1)/n;
.             VARsig = P*(norm((GE-GEp).^ 2,1))/(n-1);
.             EEsig = VARsig^ 5;
.         end
.     T = zeros(m);
.     t = zeros(m,1);
.     C = 1/X.^ (z1(i));
.         for j=1:m
.             for k=1:n
.                 t(j) = t(j)+w*(X(k,j)*Y(k))*C(k);
.             end
.         end
.         for i1=1:m
.             for j=i1:m
```

```
.                   for k=1:n
.                       T(j,i1) = T(j,i1)+w*(X(k,j)*X(k,j))*C(k);
.                       T(i1,j) = T(j,i1);
.                    end
.                end
.            end
.        B = (T∧ -1)*t;
.         YT(i) = Ty+(XT-Tx)'*B;
.            for k=1:n
.                G(k) = 1+(XT-Tx)'*(T∧ -1)*X(k,:)'/C(k);
.            end
.         E = Y-X*B;
.         GE = G.*E;
.         GEp = norm(GE,1)/n;
.         VAR(i) = P*(norm((GE-GEp).∧ 2,1))/(n-1);
.         EE1(i) = VAR(i)∧ 5;
.            if (z1(i)==1)
.                Traz = YT(i);
.                EEraz = EE1(i);
.            end
.        end
.     [EEgreg,I]= min(EE1);
.     Z = z1(I);
.   end
EEHT = (P*norm((Y-norm(Y,1)/n).∧ 2,1)/(n-1))∧ 5;
TOT = [Ty; YT(1); Traz; YT(I); Tsig];
EE = [EEHT; EE1(1); EEraz; EEgreg; EEsig];
EST = ['HT.. '; 'REG. '; 'RAZ. '; 'GREG '; 'SIG. '];
RE = [TOT,EE];
RES = [EST,num2str(RE)];
EE1 = EE1';
YT = YT';
subplot(2,1,1),plot(z1,EE1),grid,...
```

subplot(2,1,2),plot(z1,YT'),grid,...

Apéndice B

Programa para realizar la simulación

```
function [sh]=Ygms(XT,X,Y,N)

[n,m]= size(X);
w = N/n;
sh = 0;
sreg = 0;
sgreg = 0;
sraz = 0;
ssig = 0;
s1h = 0;
s1reg = 0;
s1raz = 0;
for s=1:10000
.    if s ≥ 2
.        MS = rand(n);
.            for j=1:n
.            Y(j) = MS(j)*(0.0041*X(j,1));
.        end
.    end
Z2 = 1;
Z = 1;
G = zeros(n,1);
T = zeros(m);
t = zeros(m,1);
```

```
Yb = norm(Y,1)/n;
Ty = norm(w.*Y,1);
Tx = zeros(m,1);
P = (N^ 2)*((1/n)-(1/N));
.   for j=1:m
.       Tx(j) = norm(w.*X(:,j),1);
.   end
.   while (Z2 == Z)
.       Z2 = Z2*10;
.       z1=linspace(0,Z2,1001);
.           for i=1:1001
.               if (i==1)
.                   C = 1/(Y-Yb).^ 2;
.                    for j=1:m
.                        for k=1:n
.                            t(j) = t(j)+w*(X(k,j)*Y(k))*C(k);
.                        end
.                   end
.                   for i1=1:m
.                       for j=i1:m
.                           for k=1:n
.                               T(j,i1) = T(j,i1)+w*(X(k,j)*X(k,j))*C(k);
.                               T(i1,j) = T(j,i1);
.                           end
.                       end
.                   end
.               B = (T^ -1)*t;
.               Tsig=Ty+(XT-Tx)'*B;
.                    for k=1:n
.                        G(k) = 1+(XT-Tx)'*(T^ -1)*X(k,:)'*C(k);
.                    end
.                E = Y-X*B;
.                GE = G.*E;
```

```
.             GEp = norm(GE,1)/n;
.             VARsig = P*(norm((GE-GEp).^ 2,1))/(n-1);
.             EEsig = VARsig^ .5;
.          end
.       T = zeros(m);
.       t = zeros(m,1);
.       C = 1/X.^ (z1(i));
.          for j=1:m
.             for k=1:n
.                t(j) = t(j)+w*(X(k,j)*Y(k))*C(k);
.             end
.          end
.          for i1=1:m
.             for j=i1:m
.                for k=1:n
.                   T(j,i1) = T(j,i1)+w*(X(k,j)*X(k,j))*C(k);
.                   T(i1,j) = T(j,i1);
.                end
.             end
.          end
.       B = (T^ -1)*t;
.        YT(i) = Ty+(XT-Tx)'*B;
.           for k=1:n
.              G(k) = 1+(XT-Tx)'*(T^ -1)*X(k,:)'/C(k);
.           end
.        E = Y-X*B;
.        GE = G.*E;
.        GEp = norm(GE,1)/n;
.        VAR(i) = P*(norm((GE-GEp).^ 2,1))/(n-1);
.        EE1(i) = VAR(i)^ .5;
.           if (z1(i)==1)
.              Traz = YT(i);
.              EEraz = EE1(i);
```

```
.             end
.         end
.     [EEgreg,I]= min(EE1);
.     Z = z1(I);
.   end
EEHT = (P*norm((Y-Yb).^2,1)/(n-1))^.5;
EE = [EEHT;EE1(1);EEraz;EEgreg;EEsig];
[v p]= min(EE);
.     if p == 1
.         sh = sh+1;
.     elseif p == 2
.         sreg = sreg+1;
.         sgreg = sgreg+1;
.     elseif p == 3
.         sraz = sraz+1;
.         sgreg = sgreg+1;
.     elseif p == 4
.         sgreg = sgreg+1;
.     elseif p == 5
.         ssig = ssig+1;
.     end
[v q]= min([EEHT;EE1(1);EEraz]);
.     if q == 1
.         s1h = s1h+1;
.     elseif q == 2
.         s1reg = s1reg+1;
.     elseif q == 3
.         s1raz = s1raz+1;
.     end
s
end

    pri = [sgreg ssig];
sec = [sreg sraz];
```

```
ter = [s1h s1reg s1raz];
cua = [sh sreg sraz sgreg ssig];
subplot(2,2,1),bar(pri)
subplot(2,2,3),bar(sec)
subplot(2,2,2),bar(ter)
subplot(2,2,4),bar(cua)
```

Apéndice C

Bases de datos

C.1. Médicos en Estados Unidos

Muestra proporcionada por el libro de Lohr.

RN	STATE	COUNTY	TOTPOP	Medicos
27	AL	Escambia	36023	24
48	AL	Marshall	73524	44
85	AK	Prince of Wales	6408	7
126	AR	Cross	19261	11
158	AR	Newton	7649	3
186	CA	Butte	188377	327
254	CO	Custer	2140	1
286	CO	Ouray	2497	3
305	CT	Hartford	847009	2851
340	FL	Hardee	20084	11
350	FL	Lake	161228	167
371	FL	St. Lucie	161106	176
422	GA	Crisp	20377	20
432	GA	Echols	2291	0
527	GA	Walton	40750	29
559	ID	Camas	755	0
586	ID	Shoshone	13644	9
617	IL	Ford	13914	11
630	IL	Jasper	10519	3
639	IL	Lake	541047	1093
698	IN	Boone	38381	81

RN	STATE	COUNTY	TOTPOP	Medicos
702	IN	Clark	89658	109
703	IN	Clay	25078	11
743	IN	Martin	10510	4
780	IN	Washington	24398	9
895	KS	Cheyenne	3189	2
917	KS	Grant	7625	4
932	KS	Kiowa	3582	2
943	KS	Meade	4230	3
1040	KY	Henry	13486	5
1137	LA	Lafourche	86723	70
1145	LA	Ouachita	144910	268
1149	LA	Red River	9257	6
1156	LA	St. John the Baptist	41179	20
1191	MD	Baltimore	705138	1269
1219	MA	Hampden	452140	904
1269	MI	Lake	9029	2
1303	MI	Schoolcraft	8478	6
1363	MN	Norman	7709	2
1381	MN	Sibley	14274	7
1411	MS	Copiah	27831	12
1485	MO	Bates	15047	6
1494	MO	Cape Girardeau	63318	137
1544	MO	Miller	21267	3
1548	MO	Montgomery	11264	2
1554	MO	Osage	12117	1
1572	MO	Ste. Genevieve	16163	7
1593	MO	St. Louis City	383733	4189
1640	MT	Silver Bow	34128	67
1751	NV	Lander	6691	1
1818	NM	San Miguel	26486	27
1854	NY	Montgomery	52065	66
1904	NC	Caswell	20698	7
1956	NC	Pamlico	11676	6
1973	NC	Surry	62972	50
1988	ND	Adams	3007	13
1992	ND	Bottineau	7604	4
2025	ND	Renville	2944	3
2031	ND	Slope	897	0
2057	OH	Crawford	47660	39
2062	OH	Erie	77512	107
2063	OH	Fairfield	109318	86
2099	OH	Morrow	28577	6
2125	OH	Wayne	103908	84
2142	OK	Cleveland	181388	217

RN	STATE	COUNTY	TOTPOP	Medicos
2211	OR	Coos	61894	98
2225	OR	Lane	290866	497
2234	OR	Tillamook	22307	23
2324	SC	Cherokee	45602	34
2333	SC	Fairfield	22412	10
2380	SD	Douglas	3651	4
2398	SD	Lake	10585	6
2454	TN	Grainger	17766	4
2484	TN	Marshall	22974	18
2505	TN	Trousdale	5949	3
2547	TX	Burnet	23080	17
2548	TX	Caldwell	26819	12
2554	TX	Cass	29981	14
2592	TX	Erath	28426	22
2610	TX	Gray	23436	30
2614	TX	Guadalupe	66613	47
2662	TX	La Salle	5431	2
2682	TX	Maverick	40647	24
2752	TX	Uvaide	24192	16
2763	TX	Wichita	120386	243
2840	VA	Craig	4496	1
2847	VA	Fauquier	50686	48
2899	VA	Shenandoah	32282	23
2902	VA	Spotsylvania	61435	144
2903	VA	Stafford	70900	7
2925	VA	Falls Church City	9324	0
2990	WA	Whatcom	137913	203
3009	WV	Jackson	26108	15
3044	WV	Wirt	5298	0
3071	WI	Green Lake	18822	16
3095	WI	Pierce	33442	22
3107	WI	Sheboygan	105039	114
3132	WY	Natrona	62565	130
3141	WY	Weston	6600	3

Cuadro C.1: Muestra de población de médicos en E.U.

C.2. INEGI 2000 y 2005, muestra y programa

Muestra utilizada para hacer la estimadción de la población total de México para el año 2005.

Municipio	**pop00**	**poh00**	**pom00**	**pop05**
Tejupilco	88244	42851	45393	60946
Cuauhtémoc	504759	236002	268757	488677
Jerécuaro	55011	26436	28575	45956
Tecuala	42117	21253	20864	37172
Huetamo	44969	21547	23422	41015
Coeneo	22933	10234	12699	19466
Tempoal	36207	18242	17965	33049
La Huerta	22547	11297	11250	19708
Zitácuaro	137278	65902	71376	134681
San Dimas	21695	11071	10624	19250
Ecuandureo	14719	6509	8210	12393
Acatlán de Pérez Figueroa	44503	22153	22350	42314
Tecalitlán	17839	8700	9139	15778
Temapache	102662	50884	51778	100670
Tepalcatepec	23903	11848	12055	21997
San José Tenango	19953	9831	10122	18115
Río Grande	59218	27805	31413	57473
Cuerámaro	25478	11924	13554	23778
Tlalchapa	12870	6187	6683	11258
Mecayapan	15178	7602	7576	13625
Teocuitatlán de Corona	11641	5540	6101	10160
Actopan	39182	19274	19908	37746
Taxco de Alarcón	100009	48457	51552	98645
Jesús Carranza	25340	12449	12891	24030
Copala	13028	6382	6646	11784
Huanímaro	19637	9068	10569	18442
Balancán	54101	27077	27024	52946
Tlahuiltepa	10385	5160	5225	9264
Allende	20787	10294	10493	19684
San Miguel del Puerto	8568	4316	4252	7486
Atoyac	22527	10985	11542	21487
Juchipila	12601	5902	6699	11583
San Miguel Panixtlahuaca	6701	3248	3453	5724
Tepeojuma	8375	3846	4529	7417
Alaquines	8745	4390	4355	7802

Municipio	pop00	poh00	pom00	pop05
Ocampo	10068	5104	4964	9151
Santa María Apazco	2519	1209	1310	1629
Juan Aldama	19335	9300	10035	18477
San Francisco Tlapancingo	2060	968	1092	1235
Santiago Apóstol	4624	2098	2526	3825
Santa Lucía del Camino	44188	21486	22702	43411
Santo Domingo Tonalá	7284	3425	3859	6520
Mexticacán	6798	3001	3797	6054
Mascota	13633	6666	6967	12924
Sahuaripa	6308	3277	3031	5622
Nombre de Dios	17919	8757	9162	17262
Huandacareo	11652	5291	6361	11022
Acatlán	18519	8972	9547	17914
Misantla	60539	29969	30570	59956
Acatic	19018	9169	9849	18457
Villagrán	6961	3605	3356	6417
Emiliano Zapata	26827	13219	13608	26302
Chilchota	30507	14491	16016	29999
Juchitlán	5731	2687	3044	5240
Temascalcingo	57238	27651	29587	56756
Jáltipan	37600	17900	19700	37128
La Reforma	3544	1693	1851	3086
Cañadas de Obregón	4239	1948	2291	3797
Landa de Matamoros	19309	9447	9862	18879
Santiago Tapextla	3230	1649	1581	2810
Cuautla	2429	1147	1282	2024
Tlapacoya	6422	3198	3224	6034
Casas	4497	2430	2067	4123
Olintla	12449	6094	6355	12084
Amacueca	5418	2610	2808	5065
Jiménez del Teul	5191	2538	2653	4845
San Marcial Ozolotepec	1737	852	885	1399
Carbó	4964	2625	2339	4636
Tepalcingo	23445	11418	12027	23124
Coxcatlán	17336	8692	8644	17028
Santa Ana Tlapacoyan	1986	887	1099	1687
Guelatao de Juárez	754	414	340	468
Fresnillo de Trujano	1122	519	603	848
San Miguel Tlacotepec	3517	1626	1891	3252
Tetecala	6673	3300	3373	6417
Yogana	1409	650	759	1160
San Pedro Mártir Yucuxaco	1551	734	817	1309
San Miguel Chicahua	2268	983	1285	2035
Citlaltépetl	11232	5556	5676	11005
Tuxcacuesco	3980	2007	1973	3758
Tepatlaxco	7836	4101	3735	7618
Mazatán	1568	820	748	1363

Municipio	pop00	poh00	pom00	pop05
San Melchor Betaza	1118	529	589	919
Parás	1142	604	538	950
Santa María Sola	1675	835	840	1490
San Pablo Macuiltianguis	1131	540	591	956
Tepeapulco	49327	23843	25484	49157
Ixhuatán	8852	4428	4424	8689
Tlalixtaquilla de Maldonado	6675	3173	3502	6515
Etzatlán	17122	8383	8739	16968
San Francisco Cahuacuá	3320	1643	1677	3170
San Juan Bautista Lo de Soto	2282	1152	1130	2140
Rojas de Cuauhtémoc	1061	512	549	925
San Andrés Yaá	509	231	278	378
Ixtlán de Juárez	7279	3611	3668	7151
Xicotlán	1353	648	705	1229
Isla Mujeres	10965	5756	5209	10850
San Lorenzo Albarradas	2583	1273	1310	2477
San Baltazar Loxicha	2853	1379	1474	2751
San Felipe Usila	11672	5641	6031	11573
San Francisco Cajonos	460	207	253	368
Jilotzingo	13890	6874	7016	13800
Yutanduchi de Guerrero	1255	585	670	1171
San Andrés Nuxiño	2063	1012	1051	1983
Bacadéhuachi	1336	702	634	1260
Santiago Nejapilla	266	142	124	195
Trincheras	1692	912	780	1626
San Francisco Teopan	452	224	228	390
Agua Blanca de Iturbide	8483	4159	4324	8423
San Miguel Ixitlán	619	301	318	568
Santa María Yavesía	456	215	241	409
Santa María Temaxcalapa	958	453	505	920
Tepakán	2114	1079	1035	2080
Quintana Roo	989	500	489	957
Santa Cruz de Bravo	406	194	212	379
Teotitlán del Valle	5542	2662	2880	5517
San Mateo Etlatongo	1104	524	580	1085
San Antonio Huitepec	4303	1958	2345	4288
Magdalena Zahuatlán	430	206	224	422
Chigmecatitlán	1145	517	628	1143
San Cristóbal Amoltepec	1176	569	607	1179
Galeana	38863	19766	19097	38873
San Joaquín	7613	3573	4040	7629
Santa Catarina Yosonotú	1840	809	1031	1860
Escobedo	2744	1324	1420	2774
Santa Catarina Tlaltempan	759	379	380	795

Municipio	pop00	poh00	pom00	pop05
Dzitás	3401	1747	1654	3443
Yécora	5993	3137	2856	6043
San Juan Comaltepec	2334	1147	1187	2389
Villa Unión	6067	3134	2933	6130
San José Ayuquila	1267	600	667	1338
Sanahcat	1448	721	727	1526
San Jerónimo Tecuanipan	5135	2456	2679	5221
Coyuca de Benítez	68871	33528	35343	68968
Hidalgo	1401	733	668	1505
San Andrés Huaxpaltepec	5630	2730	2900	5746
Dzemul	3130	1615	1515	3255
Concepción del Oro	11692	5902	5790	11821
San Nicolás de los Ranchos	9561	4655	4906	9704
Ucú	2901	1486	1415	3057
San Juan Petlapa	2551	1244	1307	2717
San Diego la Mesa Tochimiltzingo	1076	531	545	1252
Teteles de Avila Castillo	5336	2322	3014	5519
Valle de Guadalupe	5862	2818	3044	6052
El Fuerte	89283	45333	43950	89486
San Pedro Yólox	2542	1206	1336	2758
Benito Juárez	16205	7956	8249	16436
Xoxocotla	4397	2258	2139	4641
Santo Domingo Zanatepec	10449	5296	5153	10708
Nautla	9750	4762	4988	10019
San Lucas	5638	2868	2770	5918
Miahuatlán	3791	1845	1946	4083
Los Ramones	5917	2974	2943	6221
Magdalena	2323	1166	1157	2639
San Martín Toxpalan	3250	1623	1627	3586
Otatitlán	5220	2477	2743	5562
Soltepec	10756	5284	5472	11115
San Diego de la Unión	33980	16079	17901	34361
Zaragoza	13474	6526	6948	13873
Cuzamá	4379	2226	2153	4787
Xochitlán Todos Santos	4957	2312	2645	5387
Alpatláhuac	8545	4200	4345	8988
EL Arenal	14383	6923	7460	14845
Huehuetlán	14269	7095	7174	14754
Zaragoza	21870	10766	11104	22366
La Concordia	39580	20336	19244	40100
San José Teacalco	4575	2198	2377	5118
San Agustín de las Juntas	4958	2398	2560	5527
Ixtacomitán	9108	4496	4612	9696

Municipio	pop00	poh00	pom00	pop05
Guadalupe Victoria	14377	6902	7475	15003
Huiramba	6659	3156	3503	7303
Carichí	7704	3933	3771	8365
Tzimol	11870	5911	5959	12569
San Felipe	95031	45087	49944	95748
Batopilas	12497	6401	6096	13239
Texcalyacac	3685	1810	1875	4441
Marín	4627	2384	2243	5398
Arteaga	18730	9605	9125	19530
Reyes Etla	2419	1172	1247	3247
San Matías Tlalancaleca	15981	7838	8143	16835
Quechultenango	32461	15800	16661	33333
Nogales	30865	14754	16111	31786
Ziracuaretiro	12819	6183	6636	13768
Juan Rodríguez Clara	33391	16582	16809	34373
Tlayacapan	13411	6673	6738	14409
Atizapán	7692	3769	3923	8740
Cosío	12575	6111	6464	13658
La Perla	17804	8992	8812	18930
Tochtepec	16943	8134	8809	18101
Tetela del Volcán	15984	7797	8187	17189
Ocotepec	9251	4616	4635	10538
Santa María Chimalapa	7102	3605	3497	8439
Lázaro Cárdenas	20207	10435	9772	21611
San Jacinto Amilpas	8307	3945	4362	9763
Mártir de Cuilapan	13749	6533	7216	15248
Tlalixtac de Cabrera	6765	3226	3539	8307
Altepexi	15579	7550	8029	17177
Tlachichuca	25042	12399	12643	26721
Espita	12614	6258	6356	14355
Yehualtepec	19008	9284	9724	20815
Xicotepec	68900	33415	35485	70815
San Mateo del Mar	10653	5440	5213	12657
Hecelchakán	24853	12373	12480	26909
Allende	27293	13841	13452	29462
Acatlán de Juárez	20100	10468	9632	22400
Chilapa de Alvarez	102489	48542	53947	104911
Huatusco	46341	22480	23861	48872
General Felipe Ángeles	14789	6810	7979	17432
Los Reyes de Juárez	20413	9885	10528	23149
Soledad Atzompa	16368	8076	8292	19183
Puente de Ixtla	52957	25820	27137	55943
Tantoyuca	94625	47366	47259	97756
Zacatelco	31775	15231	16544	35128

Municipio	**pop00**	**poh00**	**pom00**	**pop05**
Siltepec	32352	16594	15758	35841
Guadalupe y Calvo	48183	24421	23762	51774
Amecameca	43499	21018	22481	47237
Villagrán	45689	21768	23921	49609
Tumbalá	23976	11810	12166	28145
San Luis de la Paz	96481	45933	50548	100927
Chignautla	21207	10462	10745	25983
Huimanguillo	158261	78817	79444	163146
Tenango del Valle	61903	30078	31825	67057
Tecamachalco	58121	28116	30005	63565
Cadereyta de Montes	51450	24153	27297	57138
Ocoyoacac	47623	23350	24273	53662
Tezoyuca	18088	8992	9096	24734
Villaflores	85592	42741	42851	92654
Atizapán de Zaragoza	448706	219016	229690	456112
La Magdalena Contreras	218210	104549	113661	226264
Comalcalco	164325	81143	83182	172906
San Juan Bautista Tuxtepec	133565	64457	69108	142783
Lagos de Moreno	127322	60997	66325	137528
San Pedro Cholula	97582	46874	50708	109264
Ixtlahuaca	111349	53420	57929	124781
Acuña	108767	55707	53060	124175
Almoloya de Juárez	105695	52439	53256	122891
La Paz	203866	100330	103536	225836
Tepic	304432	147173	157259	329015
Gómez Palacio	272119	134049	138070	300920
Tehuacán	220410	105254	115156	253451
Chalco	202412	99720	102692	244541
Tonalá	334801	168831	165970	396555
Tlaquepaque	470882	232536	238346	548935
Tijuana	1103532	557107	546425	1289969
Juárez	65205	33024	32181	138412
Chicoloapan	74663	36643	38020	157561
Santo Domingo Ixcatlán	878	409	469	1582
Sitalá	4602	2291	2311	9966

Cuadro C.2: Muestra de la población de México, 2000 y 2005

El programa con el cual se seleccionó la muestra anterior es una macro en excel

```
Sub muestra()
'
' muestra Macro
'
For I = 1 To 244
.       Range(Cells(10 + (10 * (I - 1)), 3), Cells(10 + (10 * (I - 1)), 7)).Select
.       Selection.Copy
.       Cells(4 + I, 10).Select
.       ActiveSheet.Paste
Next I
End Sub
```

Bibliografía

[1] EUSTAT (2001). Cálculo de coeficientes de variación para diferentes estimadores directos e indirectos utilizados en las encuestas económicas de Eustat. Instituto Vasco de Estadística.

[2] FULLER WAYNE A. (2002). Regression Estimation for Survey Samples. Survey Methodology, Statistics Canada

[3] LEPIK NATALJA (2007). On the Bias of the Generalized Regression Estimator in Survey Sampling. Springer Science + Business Media B.V.

[4] LOHR SHARON L. (2000). Muestreo: Diseño y Analisis. Arizona State University, International Thomson Editores.

[5] MÉNDEZ IGNACIO (2009). Notas del curso de Análisis y Diseño de Muestras. IIMAS, UNAM.

[6] MURGUI, S.; COLOM, M.C. Y MOLÉS, M.C. (2005b): Alternativas al estimador de regresión en poblaciones finitas. Aplicación a un colectivo de empresas. XIX Reunión Anual de ASEPELT, Badajoz.

[7] MYRSKYLÄ MIKKO (2007). Generalised Regression Estimation for Domain Class Frequencies. Statistics Finland.

[8] RUEDA MARIA y ARCOS ANTONIO (1994). Sobre un estimador de razón múltiple. ESTADISTICA ESPAÑOLA Vol. 36, Núm. 137, págs. 459 a 471.

[9] RUIZ ESPEJO M. (1988). Un ensayo sobre criterios de inferencia en poblaciones finitas. Qüestiió, vol. 12, n°3. Universidad Complutense de Madrid

[10] RUIZ ESPEJO M. (1991). Comparación de estimadores óptimos de razón, producto y regresión. Trabajos de Estadística vol.6, págs 81 a 87. Universidad Complutense de Madrid

[11] SÄRNDAL CARL-ERIK, SWENSSON BENGT, AND WRETMAN JAN (1989). The Weighted Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total. Biometrika, 76, 3, 527-537.

[12] SÄRNDAL CARL-ERIK, SWENSSON BENGT, AND WRETMAN JAN (1992). Model Assisted Survey Sampling (Springer Series in Statistics). Springer.

[13] SÄRNDAL CARL-ERIK (2007). The calibration approach in survey theory and practice. Survey Methodology Vol. 33, No. 2, pp. 99-119

[14] SUGDEN ROGER AND SMITH FRED (2007). Design-Based Properties of Linear Calibrated Estimators of a Finite Population Total. International Statistical Review.

[15] TILLÉ YVES (2002). Sampling Algorithms. Institut de Statistique, Université de Neuchâtel.

[16] TORRES LINDA (2009). Estrategia de muestreo usando estimadores de regresión generalizada para la estimación de tasas de favoritismo en elecciones presidenciales en Colombia. Comunicaciones en Estadística, Universidad de Santo Tomas.

[17] http://www.wikipedia.com