Nota metodológica de la estimación del ingreso a nivel municipal

La técnica de imputación de ingresos empleada en México sigue a Elbers et al. (2002). Los pasos son los siguientes:

Etapa Uno: Selección de las variables

Se seleccionaron características tanto de los individuos (edad, sexo, escolaridad, etc.) y los hogares (tamaño, sexo del jefe del hogar, escolaridad promedio, etc.) como de las viviendas (luz eléctrica, agua potable, drenaje, etc.) que estuvieran disponibles en la ENIGH (2000 y 2005) y el Censo 2000 y la muestra del Conteo 2005. Para asegurar la comparabilidad entre ambas fuentes, se realizaron pruebas estándares de todas las variables, a fin de seleccionar solamente aquellas cuya distribución es similar.

Etapa Dos: Modelación del ingreso

La estimación de la primera etapa requiere la modelación del ingreso per cápita del hogar. Dado que la base de datos se encuentra a nivel de individuo, el ingreso correspondiente a cada individuo se repite para todos los miembros del hogar.

A fin de no forzar los parámetros a sólo un modelo de imputación al nivel nacional, se dividió el país en cinco grupos de estados haciendo una estratificación según los niveles de marginalidad, sin utilizar ningún criterio de proximidad geográfica. La justificación es puramente metodológica: a fin de minimizar la heterogeneidad, es importante crear agrupamientos estadísticos, antes que geográficos ya que los resultados tienen mayor precisión.

Así pues, para cada región geográfica, la primera etapa se inicia con un modelo de asociación del ingreso per cápita para un hogar estimado a partir de la ENIGH llevándose a cabo una transformación logarítmica del mismo. Se estima un modelo de regresión por mínimos cuadrados generalizados (MCG) donde la variable dependiente es el logaritmo del ingreso per cápita y las variables explicativas son un grupo de características observables que sean comunes entre la encuesta y el censo, como se mencionó con anterioridad. La selección de las variables explicativas se toma en cuenta el nivel de significancia de las variables y mediante un proceso secuencial se van eliminando aquellas que no cuentan con impacto significativo o que contribuyen poco a la R^2 ajustada.

En resumen, se estimaron cinco variantes del modelo. Para poder realizar las estimaciones, se obtiene la matriz de varianzas y covarianzas.

Así consideramos una estimación lineal de la distribución condicional de y_h,

$$\ln y_{ch} = E\left[\ln y_{ch} \middle| x_{ch}\right] + u_{ch} = x_{ch}^T \beta + u_{ch} \qquad (1)$$

en donde el vector de errores se distribuye como $u \sim \Gamma(0, \Sigma)$. El vector β carece de toda interpretación económica.

El error puede desagregarse en

$$u_{ch} = \eta_c + \varepsilon_{ch} \qquad (2)$$

Donde η_c corresponde al error de la comunidad *c* y ε_{ch} corresponde al error intrínseco del hogar *h* que vive en la comunidad *c*.

Para minimizar la proporción del error que corresponde al factor de localidad (características intrínsecas de la localidad), se agrega a la estimación de ingresos, variables que correspondan a ésta y que no estén relacionadas con otras localidades, es decir, variables que no solamente expliquen la condición de cierto nivel de ingreso por el hecho mismo de pertenecer a dicha localidad, sino también que logren capturar la heterogeneidad entre las localidades, para lo cual se crea una base proveniente de las mismas u otras fuentes de información, por ejemplo: información geográfica. Dicha base se encuentra a nivel estatal y municipal.

Una vez que se permite la presencia de heteroscedastidad, procedemos a modelar e_{ch}^2. Esto se hace seleccionando un vector de variables z_{ch} (las cuales pueden ser transformaciones cuadráticas de las variables independientes del modelo de ingreso, el ingreso predicho y sus interacciones) que explican mejor la variación de e_{ch}^2. Así se estima la matriz de varianza- covarianza y se procede a estimar el modelo original mediante MCG.

Etapa Tres: Simulación

En esta etapa se combinaron los parámetros estimados de la primera etapa con las características observables para cada individuo del Censo a fin de estimar un ingreso y simular los errores (véase a López Calva et al., 2005), para cada simulación se empleó un grupo de parámetros de la etapa uno. Dado que se requieren estimaciones por género, la base del Censo se divide en dos: hombres y mujeres De esta forma se obtuvo como resultado de la simulación, un ingreso para cada individuo por sexo, que es utilizado para estimar medidas de bienestar para cada región: pobreza y desigualdad.

Preparación de datos

Para México, el proceso anterior tiene lugar de la siguiente manera:

a) Preparación de la base de datos a partir de la ENIGH: Se crean, en principio, los diferentes rubros de ingreso y gasto para obtener el logaritmo natural del ingreso per cápita, mismo que servirá como la variable dependiente en los modelos, a su vez se crean las variables compuestas a nivel de hogar que serán utilizadas como regresores. Una vez creada la base de datos que contenga dicha información se procede a separarla en las cinco regiones. El resultado final son cinco bases de datos a partir de la ENIGH.

b) Preparación de la base de datos a partir del Censo 2000: Se genera una base de datos que contenga las variables compuestas del Censo (las mismas variables que contienen las bases creadas a partir de la ENIGH). Sin embargo en este caso, como ya se mencionó con anterioridad, dado que se requieren estimaciones por género, la base se divide en dos: hombres y mujeres. Una vez hecho lo anterior, de nuevo se separa por regiones.

c) Preparación de la base de datos de efectos fijos: Se procede a elaborar una base que contiene variables que se encuentran en el censo o en diferentes fuentes de datos, y que pueden definir características propias de cada comunidad (efecto localidad o cluster). Por ejemplo, la mortalidad infantil, grado de marginación, esperanza de vida, etc. Como se mencionó con anterioridad, esta base está disponible a nivel estatal y municipal. Estas bases son anexadas a las bases anteriores.

Variable de Identificación

Para llevar a cabo la imputación del ingreso de la encuesta al Censo, es necesario construir un identificador geográfico (ID) del nivel de datos, que se debe encontrar tanto en la encuesta como en el Censo. El ID está compuesto por un valor que representa desde la región hasta la localidad. Hay que señalar que solamente un identificador es permitido en la base que genera el software para las diversas estimaciones, por lo que ambas bases "de insumo" (ENIGH y Censo) deben contener el mismo identificador. En este caso, el identificador fue construido de la siguiente forma para ambas bases:

Región(1 dígitos)+Entidad(2 dígitos)+Municipio(3 dígitos)