INEI
INSTITUTO NACIONAL DE ESTADISTICA E INFORMATICA

Estimación de la Pobreza para áreas Menores usando datos Censales y Encuestas de Hogares

Ph. D. Javier Antonio Torres Gómez

Lima, junio 2016

Ph.D. Javier Antonio Torres Gómez

Instituto Nacional de Estadística e Informática
Av. General Garzón N° 658, Jesús María, Lima 11 PERÚ
Teléfonos: (511) 433-8398 431-1340 Fax: 433-3591
Web: www.inei.gob.pe
Junio, 2016

Impreso en los talleres gráficos del:

Instituto Nacional de Estadística e Informática

Tiraje : 100 Ejemplares
Hecho el Depósito Legal en la
Biblioteca Nacional del Perú : 2014-09835

Agradecimiento

Se agradece la colaboración y disponibilidad de los doctores Tarozzi, Elbers & Lanjouw para absolver las dudas relacionadas con sus correspondientes metodologías, así como los códigos de su programación original. Sus aportes hicieron posible la generación de los códigos de Stata ordenados y adaptados para el Perú.

Asimismo, se agradece la valiosa asistencia por parte de Hans Tello Zamudio y Bruno Aranda Vega en el desarrollo de este proyecto. Su laboriosidad y dedicación hicieron posible el término de este trabajo.

Instituto Nacional de Estadística e Informática

Presentación

El Instituto Nacional de Estadística e Informática (INEI), viene impulsando una política orientada al uso intensivo de la información que produce mediante el desarrollo de investigaciones socioeconómicas y estadísticas, en ese contexto, pone a disposición de la comunidad nacional, autoridades, instituciones públicas, privadas, y usuarios en general, el documento **"ESTIMACIÓN DE LA POBREZA PARA ÁREAS MENORES USANDO DATOS CENSALES Y ENCUESTAS DE HOGARES"**

La Investigación tiene como objetivo la estimación de la pobreza para áreas menores o pequeñas y desarrollar estadísticos que muestren su nivel de precisión. Para lograr el objetivo se utilizó el paquete estadístico STATA 13 y se replicaron las metodologías desarrolladas por grupos de autores especialistas en la materia: Elbers, Lanjouw & Lanjow (2003); Tarozzi & Deaton (2009) y Coung (2011).

El estudio ha sido elaborado por el Consultor: Javier Antonio Torres Gómez. Estamos seguros que los resultados de la investigación serán de gran utilidad.

Esta investigación ha sido seleccionada en el concurso nacional de investigaciones que realiza cada año el Instituto Nacional de Estadística e Informática - INEI, a través del Centro de Investigación y Desarrollo (CIDE).
Lima, junio de 2016.

Lima, junio de 2016.

Dr. Aníbal Sánchez Aguilar
Jefe
Instituto Nacional de Estadística e Informática

Resumen

Se desarrollan los algoritmos propuestos por (Elbers, Lanjouw, & Lanjouw, 2003), (Tarozzi & Deaton, 2009) y (Coung, 2011) para el cálculo de la pobreza provincial en el Perú.

En vista de que el objetivo del presente informe es el desarrollo de metodologías alternativas para el cálculo de la pobreza, se realizaron las estimaciones respectivas, solo para cuatro departamentos: Áncash, Arequipa, Ica y Piura, que fueron elegidos debido a que cuentan con el menor error de proyección entre el censo del 2007 y el barrido censal del año 2012-2013, y porque poseen un gran número de habitantes.

Utilizando el método de Elbers et al. (ELL), se ejecutan 500 simulaciones de un error clúster y un error idiosincrático. De este modo, para cada repetición, se estima el logaritmo del gasto de los hogares, el cual se compara con la línea de pobreza para obtener el porcentaje de hogares pobres por provincia en los departamentos escogidos.
Por otro lado, Tarozzi & Deaton intentan simplificar la metodología de ELL. En este sentido, se estima una regresión logística de la línea de pobreza contra una serie de variables socioeconómicas (Ver Anexo N° 9.1) para los cuatro departamentos analizados. Luego, se estima la cantidad de hogares pobres por provincia que no superan esta línea de pobreza. Finalmente, se calcula el porcentaje de hogares pobres para cada provincia de los departamentos escogidos.

Con la metodología de Coung, a diferencia de las metodologías anteriores, se intenta predecir la pobreza de un año a partir del último censo disponible, de modo que se ejecutan 500 simulaciones al igual que ELL para obtener un error clúster y un error idiosincrático. Sin embargo, se estima el logaritmo del gasto de los hogares del año 2009 a partir de la Encuesta Nacional de Hogares (ENAHO) del 2007-2009 y el censo del 2007.

Para ELL, Tarozzi & Deaton y Coung, es fundamental el cumplimiento de dos supuestos: la homogeneidad de área y la medición de los predictores. El primer supuesto indica que el área menor (la provincia) debe estar incluida en un área más grande (el departamento). La segunda condición señala que los regresores que se utilizan para la estimación deben ser similares tanto en la encuesta como en el censo.

Asimismo, se presenta una explicación clara de los comandos utilizados, que puede ser replicada y entendida para otras regiones. De este modo, se aportan diferentes ideas de cómo utilizar la estimación de áreas menores no sólo para el cálculo de la pobreza sino, además, para otras mediciones o como soporte para otros propósitos.

A partir del paquete estadístico *Stata*, se procesa la información de la ENAHO 2007 y 2013, el censo del 2007 y el barrido censal del año 2012-2013. Así, se obtienen dos bases de datos: una para el 2007 y otra para el 2013. Estas bases de datos son de corte transversal y contienen un mismo grupo de variables (Ver Anexo N° 9.1) tanto de la encuesta como del censo de esos años y son las que se utilizan en las metodologías de ELL y Tarozzi & Deaton para calcular la pobreza provincial.

Cuong utiliza un panel de datos para predecir la pobreza de años posteriores. Para replicar esta metodología al caso peruano, se utiliza la muestra el panel 2007-2011 que contiene información para un mismo grupo de hogares encuestados tanto en el 2007 como en el 2009. Este hecho representa un significativo avance en la estimación de índices de bienestar en áreas pequeñas.

Luego de procesar las bases de datos, se elabora la programación pertinente para cada una de las tres metodologías que es explicada a detalle en la sección 6 de este informe. En ellas, se describe cada paso de la programación y el por qué del mismo para el cálculo de la pobreza provincial.

Finalmente, se presentan los resultados obtenidos y se observa que estos no difieren significativamente entre sí (Tarozzi & Deaton y ELL). Sin embargo ambas metodologías difieren de los obtenidos en los mapas de pobreza del INEI 2007, se debe a que el ejercicio utilizó un número limitado de variables (Ver Anexo N° 9.I) en comparación con el número de variables que emplea el INEI. Como se explicó anteriormente, el propósito del presente informe es desarrollar un manual de uso de estas metodologías.

Comparación entre la pobreza de cuatro capitales provinciales calculada por el INEI y las metodologías de Tarozzi & Deaton y ELL para el año 2007

Región	Provincia	Pobreza (Deaton)	Pobreza (ELL)	Pobreza (Mapa INEI)
ÁNCASH	HUARAZ	46%	46%	38.5%
AREQUIPA	AREQUIPA	26%	48%	21.7%
ICA	ICA	34%	37%	15.6%
PIURA	PIURA	52%	54%	37.5%

La pobreza calculada a partir de la metodología de Tarozzi & Deaton y de ELL calculan porcentajes similares. En el resumen sólo se presentan los porcentajes de las capitales de los cuatro departamentos trabajados. Este hecho confirma que la metodología de Tarozzi & Deaton es una simplificación de la de ELL.

ÍNDICE

I. Estimación para áreas pequeñas

1.1. Motivación

La estimación para áreas pequeñas es altamente demandada tanto en el sector público como en el sector privado: en el primero, se utiliza para desarrollar programas sociales; mientras que en el segundo, para decisiones de pequeños negocios que dependen de variables locales (Rao & Molina, 2015).

El gobierno requiere información detallada del gasto de los hogares para el diseño de políticas públicas en aras de combatir la pobreza, la desigualdad, entre otras problemáticas de corte social. No obstante, en el Perú, aquella sólo se encuentra en las encuestas a nivel de hogares (ENAHO) que ejecuta el Instituto Nacional de Estadística e Informática (INEI), el tamaño de la muestra de esta es muy pequeña como para obtener estimadores directos confiables.

Esteban *et. al*. proponen como solución incrementar el tamaño de muestra. Pero, realizarlo es muy costoso (Esteban, Morales, Pérez, & Santamaría, 2012).

La presente investigación desarrolla y explica los algoritmos utilizados por Elbers, Lanjouw y Lanjouw (2003), Tarozzi & Deaton (2009) y Coung (2011); considerando que son de directa aplicación y comúnmente utilizados en la literatura.

En la medida que no se cuente con bases de datos administrativas confiables que permitan inferir la tasa de pobreza para áreas menores (o información censal continua), la necesidad de realizar estimaciones siempre estará presente.

Por consiguiente, surge la metodología de estimación en áreas pequeñas[1/] propuesta por Elbers, Lanjouw y Lanjouw[2/] (2003) a través de la cual se combina la inferencia en una muestra finita (encuesta) con modelos estadísticos.

En concreto, ELL definen una función de gasto para la encuesta y calculan tanto los estimadores como los errores estimados de la ecuación. Luego, seleccionan aleatoriamente de la encuesta un error y un estimador para calcular el gasto para cada área pequeña *(clúster)* en el censo.

Por lo tanto, cada área pequeña estaría dada por un departamento del Perú. Finalmente, calculan el índice de pobreza para cada *clúster* y realizan este procedimiento 500 veces por medio de simulaciones de Monte Carlo (Elbers, Lanjouw, & Lanjouw, Micro–Level Estimation of Poverty and Inequality, 2003). Así, obtienen la media y la varianza de la pobreza para cada área pequeña.

1/ Small Area Estimation (SAE) en inglés.

2/ A partir de aquí el acrónimo ELL hace referencia a estos autores.

En síntesis, el censo o el barrido censal se utilizan en la estimación de áreas menores para incrementar el tamaño de muestra. Esto se debe a que la cantidad de datos presentes en una encuesta para un área pequeña no es lo suficientemente grande como para obtener cálculos consistentes. Aquí, por ejemplo, se utilizó para el cálculo de la pobreza en las provincias de cuatro departamentos del Perú.

La utilización de bases de datos (como el panel de hogares de la ENAHO o, hasta hace poco, el barrido censal 2012-2013) junto a una metodología establecida en la literatura, ofrece la posibilidad de realizar estimados de pobreza para áreas menores y desarrollar estadísticos que muestren su nivel de precisión. De esta manera, se pretende explicar y sistematizar los pasos necesarios para la realización de estos cálculos.

Cabe resaltar el rol que juega en el proceso la unión de dos encuestas de años consecutivos, de este modo, el uso de una mayor cantidad de datos permitiría obtener un cálculo más preciso de la pobreza, lo cual mejoraría los errores de muestreo de este indicador. Esta variante será utilizada para el cálculo de la pobreza bajo la metodología de Tarozzi & Deaton, al unir las encuestas del 2006 y 2007 junto con el censo 2007.

Además, la estimación de áreas menores no sólo es aplicable para el censo nacional, también lo es para otras fuentes de información externa, tal como, el censo agropecuario, el censo escolar, el registro nacional de municipalidades y el censo de infraestructura educativa. En otras palabras, el uso de regresores presentes tanto en la encuesta nacional de hogares como en alguna de estas otras bases de datos mejoraría la matriz de información disponible. De este modo, se podrían calcular otros indicadores con un mayor nivel de precisión que aquellos obtenidos utilizando sólo la encuesta.

El objetivo del presente informe consiste en aplicar la estimación para áreas pequeñas en el Perú. Utilizando el paquete estadístico *STATA 13*, se replican las metodologías desarrolladas por tres grupos de autores especialistas en la materia: Elbers, Lanjouw & Lanjouw (2003); Tarozzi & Deaton (2009); y Cuong (2011).

1.2. Modelos de estimación para áreas pequeñas

Los modelos para estimar pobreza en áreas pequeñas permiten obtener indicadores para todas las áreas objetivo. Esta clase de modelos poseen una serie de ventajas que se detallan a continuación (Rao & Molina, 2015):

a) El diagnóstico del modelo permite hallar uno que se ajuste correctamente a la información disponible (la encuesta y el censo). Estos diagnósticos permiten realizar procedimientos tales como la estimación de residuos y la selección de variables explicativas.
b) Los valores por área poseen una precisión asociable con cada estimador de las áreas pequeñas en comparación con los estimadores sintéticos (indirecto), que utiliza medidas globales (promediadas por el tamaño de cada área pequeña).
c) Se emplea modelos no lineales (logit, probit) y manejar estructuras complejas de datos.
d) Se puede realizar inferencias adecuadas para áreas pequeñas.

Distintos autores señalan la importancia de la elección de las variables explicativas para realizar estimaciones sobre el nivel de bienestar (pobreza) en áreas pequeña. La metodología de ELL emplea una serie de regresores que deben ser comunes tanto en la encuesta como en el censo. Por esta razón, el INEI realiza una selección de variables explicativas para elaborar su mapa de pobreza provincial y distrital el 2009.

Ahora bien, los modelos en cuestión se clasifican en dos: (i) los modelos de área, que relacionan las medidas para áreas pequeñas con regresores para cada área; y (ii) los modelos de unidad o modelos establecidos para unidades individuales (Rao & Molina, 2015). A continuación, se describe los principales modelos de ambos tipos (Rao & Molina, 2015).

1.2.1. Modelo básico por áreas (La enumeración es correcta)

En este modelo se usa una ecuación para describir la relación entre una función de la media de la variable de interés, las variables explicativas seleccionadas y los efectos específicos por área y los efectos aleatorios.

$$\theta_i = z_i^T \beta + b_i v_i$$

Donde $\theta_i = g(\bar{Y}_i)$ es una función de la media de la variable de interés; $z_i^T = \left[z_{1i} \dots z_{pi}\right]$ es un vector de variables explicativas por área; $\beta = \left[\beta_1 \dots \beta_p\right]$son los coeficientes para cada regresor de la estimación; y v_i son los efectos específicos por área y los efectos aleatorios, que se distribuyen iid con media $\mathrm{E}_m(v_i) = 0$ y varianza $V_m(v_i) = \sigma_v^2$. La varianza de los efectos por área y los efectos aleatorios (σ_v^2) mide la homogeneidad de las áreas tomando en cuenta los regresores (z_i).

Para realizar las estimaciones correspondientes, se asume el método de James-Stein, quien plantea que el estimado de la función de la media estimada de la variable de interés $\hat{\theta}_i = g(\bar{\hat{Y}}_i)$, es igual a esta función θ_i más una perturbación e_i independiente de θ_i, con media $\mathrm{E}_p(e_i|\theta_i) = 0$ y con una varianza conocida $V_p(e_i|\theta_i) = \psi_i$:

$$\hat{\theta}_i = g\left(\bar{\hat{Y}}_i\right) = \theta_i + e_i$$

Por consiguiente, se puede presentar el estimador $\hat{\theta}_i$ como:

$$\hat{\theta}_i = z_i^T \beta + b_i v_i + e_i$$

Donde se asume que v_i y e_i son independientes entre sí.

1.2.2. Extensiones: Modelo Fay-Harriot

Fay-Harriot plantean una extensión del modelo básico por área, donde estiman un vector de características $rx1$ por cada área pequeña i ($\theta_{ij}=[\theta_{i1}...\theta_{ir}]^T$), donde $\theta_{ij} = g(\bar{Y}_{ij})$ o $\bar{Y}_{ij}$ es la media de la i-ésima área pequeña, (i=1,...,m) para la j-ésima característica (j=1,...,r). La estimación de estos parámetros será representada por $\hat{\theta}_{ij} = \left[\hat{\theta}_{i1} \dots \hat{\theta}_{ir}\right]^T$.

Para realizar las inferencias se plantea el modelo de error de muestreo James-Stein y se propone un modelo de enlace similar al modelo básico por área:

$$\hat{\theta}_{ij} = g\left(\bar{\hat{Y}}_{ij}\right) = \theta_{ij} + e_{ij}$$
$$\theta_{ij} = z_{ij}^T\beta + b_{ij}v_{ij}$$

De este modo, se presenta el estimador $\hat{\theta}_{ij}$, como:

$$\hat{\theta}_{ij} = z_{ij}^T\beta + v_{ij} + e_{ij}$$

Donde b_{ij}=1, de este manera se toma ventaja de las correlaciones entre los componentes de $\hat{\theta}_i$ y se obtienen estimadores más eficientes.

1.2.3. Extensiones: Series de tiempo y modelos de corte transversal

Algunas encuestas registran a los mismos hogares a través de múltiples periodos. Por ejemplo, en la encuesta mensual de fuerza laboral canadiense (LFS), un hogar puede permanecer en la muestra por seis meses consecutivos. De esta forma, con muestras repetidas se puede ganar eficiencia a través de múltiples periodos. Rao y Yu proponen una extensión al modelo básico a nivel de área para lidiar con series de tiempo y datos de corte transversal.

Específicamente, los autores proponen un modelo de error de muestreo y un modelo de enlace:

$$\hat{\theta}_{it} = \theta_{it} + e_{it}$$
$$\theta_{it} = z_{it}^T\beta + v_i + u_{it}$$

Donde $\theta_{it} = g(\bar{Y}_{it})$ es una función de la media del área pequeña i en el tiempo t ($\bar{Y}_{it}$). También $\hat{\theta}_{it}$ será el estimador directo de la encuesta y el error e_{it} tendrá una distribución normal con media cero.

Para la estimación de θ_{it}, z_{it} es un vector de variables específicas al área que pueden cambiar a lo largo del tiempo. Asimismo, se asume que el error v_i tendrá una distribución normal con media cero y varianza homocedástica, y que el error u_{it} sigue un proceso autorregresivo de primer orden común para cada área i.

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, \quad |\rho| < 1$$

Donde se asume que el error ε_{it} conserva una distribución normal con media cero, varianza homocedástica y que es independiente de los otros errores, e_{it} y v_i.

Por otro lado, el modelo anterior también puede ser expresado, como uno con rezagos, donde se relaciona θ_{it} con el parámetro del periodo anterior $\theta_{i,t-1}$, los valores de las variables auxiliares para el periodo t y $t-1$, los efectos de área v_i y el efecto en el área por periodo ε_{it} :

$$\theta_{it} = \rho\theta_{i,t-1} + (z_{it} - \rho z_{i,t-1})^T\beta + (1-\rho)v_i + \varepsilon_{it}$$

1.2.4. Modelo básico por unidades

El modelo básico por unidades asume que las variables explicativas a nivel de unidad (individuo) x_{ij} se encuentran disponibles para cada elemento de la población *j* en cada área pequeña *i*. Además, se propone que la variable de interés y_{ij} está relacionada con x_{ij} a través de un modelo de regresión lineal de error anidado:

$$y_{ij} = x_{ij}^T\beta + v_i + e_{ij}$$

Donde el error específico al área es v_i, y el error específico a la unidad es e_{ij}. Se asume que ambos son variables aleatorias independientes e idénticamente distribuidas, con media cero y varianza homocedástica[3].

Se toma una muestra s_i de tamaño n_i de las variables explicativas (x_{ij}) para la i-ésima área (que tiene un tamaño N_i). Se considera que la muestra sigue la distribución del modelo básico antes planteado (si el muestreo fue realizado de forma aleatoria). De esta manera se construye una forma matricial con P número de regresores.

$$y_i^P = X_i^P\beta + v_i 1_i^P + e_i^P$$

La matriz X_i^P tiene orden de $N_i \times p$, y_i^P y e_i^P son vectores de N_i filas y 1_i^P es un vector de unos. A continuación se expresa la forma anterior en una de matrices particionadas, donde una parte es aquella que fue muestreada y la otra no lo fue:

$$y_i^P = \begin{bmatrix} y_i \\ y_{ir} \end{bmatrix} = \begin{bmatrix} X_i \\ X_{ir} \end{bmatrix}\beta + v_i \begin{bmatrix} 1_i \\ 1_{ir} \end{bmatrix} + \begin{bmatrix} e_i \\ e_{ir} \end{bmatrix}$$

Donde el subíndice r denota las unidades no muestreadas. Si el modelo se mantiene para la muestra, entonces es posible estimar los valores de y_{ir} dado los estimadores $\beta, \sigma_v^2, \sigma_e^2$, y los valores de X_{ir}.

Finalmente, la media del área pequeña i ($\bar{Y}_i$) será,

$$\bar{Y}_i = f_i\bar{y}_i + (1-f_i)\hat{\bar{y}}_{ir}$$

Con $f_i = {}^{n_i}/_{N_i}$, donde $\bar{y}_i$ e $\bar{y}_{ir}$ a la media de los elementos muestreados y no muestreados, respectivamente.

3/ Otro supuesto usualmente empleado es la distribución normal de ambos errores.

1.2.5. Extensiones: Modelo a dos niveles

El modelo a dos niveles permite la variación del intercepto y las pendientes entre áreas. Por lo tanto, cada coeficiente se expresa como $\beta_{ip} = \beta_p + v_i$, donde el p-ésimo coeficiente (β_p) para el área i está formado por un componente fijo y otro que es aleatorio al área. De esta manera, el conjunto de coeficientes $\beta_i = (\beta_{i1}, \dots, \beta_{ip})$ será modelado en términos de un conjunto de variables a nivel de área $\tilde{Z}_i$.

El modelo a dos niveles está compuesto por dos ecuaciones:

$$y_{ij} = x_{ij}^T \beta_i + e_{ij}, \qquad j = 1, \dots, N_i, \qquad i = 1, \dots, m$$

$$\beta_i = \tilde{Z}_i \alpha + \mathbb{v}_i$$

Donde $\tilde{Z}_i$ es una matriz $p \times q$, α es un vector de los parámetros de la regresión con q filas, $\mathbb{v}_i$ es un término de perturbación de media cero y varianza heterocedástica, y $e_{ij} = k_{ij}\tilde{e}_{ij}$, donde $\tilde{e}_{ij}$ posee media 0 y varianza homocedástica.

La primera ecuación puede ser expresada de forma matricial,

$$y_i^P = X_i^P \beta_i + e_i^P$$

Si se une a la segunda ecuación, se tiene un modelo que usa variables a nivel de área y unidad:

$$y_i^P = X_i^P \tilde{Z}_i \alpha + X_i^P \mathbb{v}_i + e_i^P$$

Puesto que se asume que no existe sesgo por una selección muestral, si N_i es grande se puede expresar la media $\bar{Y}_i$ como.

$$\bar{Y}_i = \bar{X}_i^T \tilde{Z}_i \alpha + \bar{X}_i^T \mathbb{v}_i$$

Entonces la estimación de $\bar{Y}_i$ es aproximadamente equivalente a la estimación de una combinación lineal de α y la realización de un vector aleatorio $\mathbb{v}_i$ con una matriz de covarianzas desconocida.

2. Resultados Previos Relevantes

Algunas mejoras a la metodología inicial de ELL se encuentran en el análisis realizado por Nepal. Por otra parte, los informes de mapas de pobreza de Paraguay, Sri Lanka, Bulgaria y Georgia reportan conclusiones relevantes a tomar en cuenta al elaborar la estimación de áreas pequeñas. A continuación presentamos un detalle de la experiencia internacional:

2.1. Resultados internacionales

2.1.1. Nepal

En adición a la estimación tradicional de áreas pequeñas, el gobierno de Nepal junto con instituciones tales como el Banco Mundial, las Naciones Unidas y el Centro de Estadística de Nepal estiman no sólo la pobreza de dicho país, sino también la ingesta calórica y la malnutrición. Para medir la ingesta de calorías directa se realiza un ajuste a la edad y sexo de cada miembro del hogar, de modo que sea un equivalente a la ingesta adulta y luego se toma el promedio de la ingesta calórica de un hogar como el total de kilocalorías consumidas por día dividido entre el número de adultos que habitan allí. Para el cálculo de la malnutrición, se toman en cuenta: la talla por edad, el peso y el peso por edad.

Se realiza una estimación con la encuesta demográfica y de salud del 2001, y se extrapola al país utilizando datos censales. Se complementa la estimación con información espacial adicional tal como acceso a infraestructura y servicios públicos; condiciones agroecológicas; gasto público, entre otras (Central Bureau of Statistics; Government of Nepal; United Nations World Food Programme & The World Bank, 2006).

2.1.2. Paraguay

El mapa de pobreza de Paraguay muestra que un mayor nivel de desigualdad no necesariamente implica un incremento en la incidencia de pobreza. En contraposición a lo que generalmente se supone, el informe de Robles y Santander encuentra que *"para algunos distritos del país la elevada incidencia de la pobreza no se halla acompañada por mayores niveles de desigualdad, ocurriendo en otros distritos la situación inversa"* (Robles & Santander, 2004).

2.1.3. Sri Lanka

Sri Lanka emplea los mapas de pobreza como soporte para identificar las zonas más vulnerables en caso de tsunami. Esta aplicación brinda una dimensión adicional al cálculo de pobreza. No como medida de bienestar actual de la población sino como probabilidad de vulnerabilidad ante un desastre natural[4].

Cabe resaltar, sin embargo, que el documento señala la importancia de usar estos mapas sólo como un primer paso en la aplicación de políticas contra la pobreza y no como sustitutos del desarrollo de estas (Poverty Reduction & Economic Management, South Asia Region, World Bank, 2005).

[4]Esta visión podría ser incorporada y adaptada para el Perú.

2.1.4. Bulgaria

Oleksiy Ivaschenko utiliza la estimación de áreas pequeñas para comparar los niveles de vida inter e intra zonas geográficas. Encuentra que a pesar de la gran desigualdad existente entre áreas urbanas y rurales, existe también mucha diferencia en los niveles de pobreza dentro de las mismas áreas. Del mismo modo señala que los estándares de vida varían mucho, incluso entre municipalidades del mismo distrito. Estos resultados son de importancia y utilidad para el adecuado enfoque de programas de reducción de pobreza y el análisis de los factores regionales asociados.

2.1.5. Georgia

En su análisis para Georgia, Labbate et al. proponen el uso de mapas de pobreza como una herramienta adicional para el adecuado diseño de proyectos públicos y de infraestructura. Por ejemplo, una concesión que pretende rehabilitar carreteras para una mejor conexión con socios comerciales, y acceso a los distritos más pobres de Georgia, necesita de más de una herramienta de análisis. La superposición del mapa de pobreza con la red de carreteras permite distinguir las zonas donde el impacto sería mayor.

2.1.6. Chile

Desde 1987, el Ministerio de Desarrollo Social levanta la Encuesta de Caracterización Socioeconómica Nacional (Casen) como un reflejo de la realidad social y económica del país. Con ella se han estimado y publicado tasas de pobreza a nivel nacional, regional y comunal. En esa tarea, la metodología estándar generaba estimadores insesgados y consistentes, pero solo a nivel nacional y regional. En consecuencia, en el año 2011 se realizó un proyecto de investigación para el desarrollo de mejores metodologías.

Como resultado, el Ministerio de Desarrollo Social emplea la metodología para la estimación de pobreza a nivel local que el U.S. Census Bureau toma como referencia en la distribución de fondos públicos (método Fay-Harriot). El método se basa en un promedio entre una tasa de pobreza directa y una tasa de pobreza sintética. La primera es una estimación que solo emplea los datos de la encuesta Caen. La segunda toma además información auxiliar de datos censales para la generación de un modelo de regresión lineal. Finalmente, el promedio ponderado se realiza en función inversa del nivel de varianza que cada estimación genere.

Con la nueva metodología, el ministerio obtiene mayores ventajas sobre la estimación de la incidencia de pobreza a nivel comunal. Los resultados de un ejercicio aplicado al 2009 arrojaron: tasas de pobreza consistentes con metodologías previas, y mayores niveles de precisión reflejados en intervalos de confianza significativamente más reducidos. Adicionalmente, la mayor precisión antes mencionada permite realizar mejores comparaciones entre comunas o entre periodos (Ministerio de Desarrollo Social, 2013).

3. Algoritmos de Metodologías Propuestas

3.1. Tarozzi y Deaton

El estimador propuesto por Tarozzi & Deaton, requiere de dos supuestos centrales:

Medición de los predictores (MP): Sea X_h el valor de los regresores para el hogar h (si h está incluido en la muestra de la encuesta), y sea $\tilde{X}_h$ la correspondiente medida en el censo. Entonces, $X_h = \tilde{X}_h$ para todo h.

Homogeneidad de área (AH) (o independencia condicional):

$$f\left(y_h \middle| X_h, h \in H(A)\right) = f\left(y_h \middle| X_h, h \in H(R)\right)$$

Donde A es el área pequeña en el censo y R es la región a la que pertenece dicha área pequeña. En consecuencia, $A \subset R$.

Ambos autores señalan que el parámetro W_A (índice de pobreza para el área pequeña A en el censo) para una línea de pobreza z está determinado por[6] :

$$W_A = \frac{1}{N_A} \sum_{h \in H(A)} 1\,(y_h \leq z)$$

Así, bajo los dos supuestos antes mencionados, el parámetro de interés puede ser estimado en dos etapas: primero, los parámetros γ que describen la probabilidad condicional $P(y_h \leq z|X_h;\hat{\gamma})$ se estiman con un modelo de variable dependiente paramétrica binaria tal como un logit o un probit a partir de la información de la encuesta para la región R. Luego, el contador de pobreza se estima como:

$$\widehat{W}_A = \frac{1}{N_A} \sum_{h \in H(A)} P(y_h \leq z|X_h;\hat{\gamma})$$

[5] Nuevamente agradezco a los autores por su aporte. Sus metodologías son ordenadas, detalladas y codificadas para el caso peruano.

[6] Sin tomar en cuenta las diferencia entre el nivel de hogares y el de personas.

Esto es la media de las probabilidades imputadas para cada región perteneciente a un área A del censo. Ahora bien, Tarozzi y Deaton argumentan que como la población del censo se mantiene fija y, por tanto, no hay aleatoriedad, la única fuente de error muestral en el estimador $\widehat{W}_A$ es la estimación de γ, de modo que el error estándar puede ser calculado utilizando el método delta. Si se adopta un modelo logit en la primera etapa de estimación, el método delta conduce a que:

$$\widehat{Var}\left(\widehat{W}_A\right)_{\mathbf{1x1}} = \hat{G}_{1\,\mathbf{1xk}}\widehat{Var}(\hat{\gamma})_{\mathbf{kxk}}\hat{G}'_{\mathbf{kx1}}$$

Dónde:

$$\hat{G}_{\mathbf{1xk}} = \frac{1}{N_A}\sum_{h\in H(A)}\frac{e^{\widetilde{X_h}'\hat{\gamma}}}{1+e^{\widetilde{X_h}'\hat{\gamma}}}\widetilde{X_h}'_{\mathbf{1xk}}$$

$\tilde{X}_{h\,\mathbf{kx1}}$: regresores del censo empleados para estimar $\widehat{W}_A$

$\widehat{Var}(\hat{\gamma})$: matriz de varianzas y covarianzas de la estimación de $\widehat{W}_A$ en la primera etapa
∧ k: número de regresores

Sin embargo, el error cuadrático medio del estimador debe tomar en cuenta no sólo la varianza de $\widehat{W}_A$ sino también la diferencia entre W_A y la media de $P(y_h \le z|X_h;\ \gamma)$. Esta diferencia, a la que Tarozzi & Deaton se refieren como sesgo, puede presentarse en el estimador de W_A incluso si γ fuese conocido.

Los autores en cuestión muestran que la contribución de este sesgo al error cuadrático medio puede ser aproximado por:

$$\hat{b}^2\left(\widehat{W}_A\right) = \frac{1}{N_A}\sum_{h\in H(A)}[(p_h - 1(y_h < z))^2] + \frac{N_A - 1}{N_A}\sum_{h\in H(A)}[(p_h - 1(y_h < z))(p_{h'} - 1(y_{h'} < z))]$$

Dónde: $p_h = P(y_h \le z|X_h;\gamma)$ segunda expresión está dada por el producto de diferentes hogares en una misma área pequeña A. Ambas sumatorias pueden ser estimadas utilizando los respectivos análogos muestrales.

En síntesis, el intervalo de confianza para $\widehat{W}_A$ con una cobertura nominal de $(1-\tau)$ se construye como:

$$\widehat{W}_A \pm \Phi^{-1}(1-\tau/2)\left(\widehat{Var}\left(\widehat{W}_A\right) + \hat{b}^2\left(\widehat{W}_A\right)\right)$$

Donde $\Phi^{-1}(\ .\ )$ es la inversión de la distribución acumulada de una normal estándar. Cabe resaltar que mientras el primer término del sesgo de $\widehat{W}_A$ tiende a 0 conforme el tamaño de muestra de un área pequeña aumenta, el segundo término no lo hace. Por consiguiente, el componente del sesgo puede ser relativamente mayor cuando los residuos de la primera etapa de estimación exhiben una gran correlación intra clúster debido a que esta correlación puede conducir a un gran segundo término.

3.2. Elbers, Lanjouw & Lanjouw

Al igual que Deaton, Elbers et al. requieren de ambos supuestos: medición de los predictores (MP) y homogeneidad de área (AH). Los autores estiman el gasto para áreas pequeñas. Este último, al ser un estimado basado en la simulación, depende de supuestos paramétricos sobre la distribución de los residuos de la regresión. Sin embargo, estos mismos supuestos son los que proveen esta metodología la habilidad para generar índices de pobreza a partir del proceso de simulación pertinente.

En primer lugar, se desarrolla un modelo empírico donde la variable de interés es el gasto *per cápita* del hogar h en el *clúster* $c\,(y_{ch})$. Entonces, se toma una aproximación lineal para la distribución condicional y_{ch} respecto a las características del hogar (x_{ch}).

$$\ln y_{ch} = E\left[\ln y_{ch} \mid x_{ch}^{T}\right] + u_{ch} = x_{ch}^{T}\beta + u_{ch}$$

Se separa la perturbación u_{ch} en 2 componentes: el componente idiosincrático (ε_{ch}) y el componente *clúster* (η_c), el cual permite una correlación *intra-clúster* de la perturbación, ambos componentes son independientes entre sí, y también con las variables explicativas x_{ch}.

$$u_{ch} = \eta_c + \varepsilon_{ch}$$

En segundo lugar, y con ambas ecuaciones planteadas, se estima $\hat{\beta}$ a partir de mínimos cuadrados ordinarios y se toma el residuo de la regresión $(\hat{u}_{ch})$. Dado el pequeño número de *clústers*, se asume la homocedasticidad de su componente en la perturbación. Sin embargo, la varianza del componente idiosincrático, sí toma una forma flexible. En consecuencia, cada perturbación clúster muestral es definida como la media de $\hat{u}_{ch}$ para el conjunto de observaciones procedentes de un clúster común $(\hat{u}_{c.})$, y el resto es el estimado de la perturbación idiosincrática (e_{ch}).

$$\hat{u}_{ch} = \hat{u}_{c.} + (\hat{u}_{ch} - \hat{u}_{c.}) = \hat{\eta}_c + e_{ch}$$

En tercer lugar, para modelar la heterocedasticidad de la perturbación idiosincrática se eligen de diez a veinte variables (z_{ch}) que expliquen mejor la variación de e_{ch}^2. Se estima $\hat{\alpha}$ por mínimos cuadrados ordinarios y con los resultados se estima un modelo logarítmico de la varianza de ε_{ch} condicionado z_{ch}.

$$\ln\left[\frac{e_{ch}^2}{A^*-e_{ch}^2}\right] = z_{ch}^T\alpha + r_{ch}$$

$$\sigma^2(z_{ch},\alpha,A,B) = \left[\frac{A^* e^{z_{ch}^T\alpha} + B}{1+e^{z_{ch}^T\alpha}}\right]$$

Donde $A^* = (1.05)\max\{e_{ch}^2\}$ y B=0. De esta forma, se evitan varianzas estimadas extremadamente altas o negativas. Entonces, con las varianzas estimadas se genera una distribución de errores idiosincráticos estandarizados en la que H es el número de observaciones.

$$e_{ch}^* = \frac{e_{ch}}{\hat{\sigma}_{\varepsilon,ch}} - \left[\frac{1}{H}\sum_{ch}\frac{e_{ch}}{\hat{\sigma}_{\varepsilon,ch}}\right]$$

Para la estimación del índice de pobreza (W_0) y su error estándar, se realizará una simulación Monte Carlo que toma como insumos las estimaciones puntuales de $\hat{\beta}$, las distribuciones empíricas del error *clúster* y las del error idiosincrático. La estructura de una simulación r es como sigue. En primer lugar, se toma un $\hat{\beta}^r$ de la distribución muestral. Luego, a cada *clúster* en el censo se le es asignado un error *clúster* $\hat{\eta}^r$ tomado de la distribución empírica de $\hat{\eta}$. A continuación, para cada observación censal es tomado un error idiosincrático $e_{ch}^{*\,r}$. Con este último es que se calcula el error heterocedástico $e_{ch}^r = \hat{\sigma}_{\varepsilon,ch}(e_{ch}^{*\,r})$. Finalmente, los valores simulados del logaritmo del gasto *per cápita* son generados con la ecuación:

$$\ln y^r = x_{ch}^T\hat{\beta}^r + \hat{\eta}_c^r + e_{ch}^r$$

Con los datos estimados del gasto se genera el índice de pobreza W^r para la simulación r. Es así que el valor estimado del índice es la media sobre las R simulaciones.

$$\widehat{W} = \frac{1}{R}\sum_{r=1}^{R} W^r$$

3.3. Cuong

3.3.1. Método de estimación en áreas pequeñas

El tercer método es el de Coung, el autor proyecta un mapa de pobreza para el área rural en Vietnam para el año 2008 utilizando el 50% de la muestra del Censo Rural de Agricultura y Pesca y el panel de la encuesta nivel de hogares de Vietnam del 2004 al 2006. Como la encuesta del 2008 se encuentra disponible, el autor también construye un mapa de pobreza del 2008 utilizando estimaciones del panel de la encuesta del 2006 al 2008 y proyectándola con la información del 2006.

En concreto, el autor basa su estimación en una versión simplificada del método de Elbers et. al que se describe en dos etapas: primero, construye una relación funcional entre el gasto y las características del hogar en el panel data. Así, por medio de mínimos cuadrados generalizados (MCG), Cuong permite la presencia de heterocedasticidad en los errores y correlación entre departamentos.

$\ln(y_h) = X_h\beta + \varepsilon_h$ para cada departamento

Donde y_h es el gasto per cápita del hogar h en la encuesta, x_h es un vector de variables explicativas del hogar disponible en ambas bases de datos (censo y encuesta). Cabe resaltar que Cuong, al igual que ELL, descompone los errores en componente idiosincrático (u_{ch}) y un componente clúster (por departamento) (η_h). Sin embargo, para fines prácticos, se simplifica la nomenclatura de los errores (ε_h).

Segundo, a partir de los estimadores obtenidos en el paso anterior, se seleccionan los betas $(\hat{\beta})$ y errores $(\hat{\varepsilon}_h)$ estimados de forma aleatoria. De este modo, se aplica el modelo funcional descrito en el paso anterior junto a los estimadores escogidos anteriormente.

$\hat{y}_h^s = \exp(X_h\hat{\beta}_h^s + \hat{\varepsilon}_h^s)$

Obteniéndose un gasto estimado $(\hat{y}_h)$ con el cuál se calcula el índice de pobreza Foster-Greer-Thorbecke (FTG) para cada área pequeña (departamento) en el censo.

$\hat{P}_\alpha = \frac{1}{n}\sum_{h-1}^{n}[\frac{z-\hat{y}_h^s}{z}]^\alpha \, 1(\hat{y}_h^s < z)$

Este procedimiento se repite 500 veces a través de una simulación de Monte Carlo, por la cual se obtiene la media y la varianza del índice FTG estimado.

3.3.2. Proyección de índices de pobreza

Por intermedio de una encuesta panel a nivel de hogares y la data censal, se busca proyectar el índice de pobreza para un periodo posterior al disponible en el censo. Así, Cuong supone que existe una relación funcional entre el gasto de los hogares en el periodo actual (t_2) y las características de los hogares en el periodo anterior (t_1):

$$\ln(y_{h2}) = X_{h1}\beta_{12} + \varepsilon_{h2}$$

Donde el subíndice 2 hace referencia al periodo 2, mientras que el subíndice 12 hace alusión a que el coeficiente de la regresión indica una relación entre el gasto del segundo periodo y las variables independientes del primer periodo.

También, se puede asumir que una relación funcional entre la variable dependiente en un periodo posterior (t_3) y las variables independientes del periodo actual (t_2) :

$$\ln(y_{h3}) = X_{h2}\beta_{23} + \varepsilon_{h3}$$

La predicción del nivel de pobreza en el periodo (t_3) se fundamenta en los siguientes supuestos:

a) $\beta_{23} = \beta_{12}$: significa que la correlación entre el gasto en periodo actual (t_2) y las características del periodo anterior (t_1) no presentan variaciones a lo largo del tiempo.

b) $Var(\varepsilon_{h3}) = Var(\varepsilon_{h2})$: esta condición es fundamental para que el error simulado en el periodo posterior (t_3) pueda ser tomado de la distribución del error del periodo actual (t_2).

4. Bases de Datos

4.1. Censos Nacionales de Población y Vivienda de 2007

Los Censos Nacionales 2007 estuvieron a cargo del Instituto Nacional de Estadística e Informática (INEI), en su condición de ente rector del Sistema Estadístico Nacional. En esta ocasión, los censos realizados fueron el XI de Población y el VI de Vivienda. El primero reveló un total de 28'220,000 habitantes; mientras que el segundo censo mostró 7'566,000 viviendas particulares en el territorio peruano.

Contó con 53 preguntas que dieron a conocer las características sociodemográficas más relevantes de la población. Asimismo, su estructura permite su comparación con las otras bases de datos empleadas en el presente trabajo.

4.2. Encuesta Nacional de Hogares (ENAHO)

La Encuesta Nacional de Hogares es una investigación estadística continua que genera indicadores trimestrales, que permiten conocer la evolución de la pobreza, del bienestar y de las condiciones de vida de los hogares. Esta es una encuesta por muestreo, donde la unidad de análisis está constituida por el hogar y las personas, se realiza a nivel nacional, en el área urbana y rural en los 24 departamentos.

4.3. Encuesta Nacional de Hogares Panel (ENAHO PANEL)

La ENAHO Panel es una base de datos muestral que es elaborada por el Instituto Nacional de Estadística e Informática (INEI). En el periodo 2007-2011, con el objetivo de medir cambios en el comportamiento de algunas características de la población peruana, se implementaron muestras de panel de viviendas en la cual viviendas encuestadas son visitadas e investigadas anualmente. Esta muestra permite obtener estimaciones de las características socio-demográficas de la población peruana a nivel nacional.

De la misma manera que la ENAHO no panel, ésta encuesta posee información relevante respecto de las características de la vivienda y del hogar, características de los miembros del hogar, educación, empleo, salud, ingresos y gastos.

4.4. Barrido Censal 2012-2013

El Barrido Censal fue realizado por el Instituto Nacional de Estadística e Informática (INEI) a inicios del año 2012 y fue culminado a finales del 2013. (el propósito fue aplicar una evaluación socioeconómica de los hogares del país).

Esta base de datos recoge datos de características de la vivienda, así como ciertas variables a nivel de individuos, como educación, empleo y salud. Sin embargo, a diferencia de un censo, el barrido censal presenta proyecciones de dichas variables en el Perú.

5. Selección de Departamentos a Estudiar

A partir de las proyecciones de población del INEI, se obtuvo la población por departamento hacia 2013; mientras que, sobre la base del Barrido Censal del 2012-2013 se calcularon las proyecciones de la población en estos años.

Se trató de escoger aquellos departamentos cuya diferencia entre la población encuestada por el barrido censal y la población proyectada sea la mínima posible. Es decir, aquellos departamentos donde el barrido censal haya conseguido la mayor cobertura posible.

Tabla N° 1:

Departamento	Cobertura	Proyecciones INEI 2013
Ica	90%	771 507
Piura	87%	1 800 000
Tumbes	84%	231 48
Arequipa	83%	1 300 000
Áncash	83%	1 100 000
Amazonas	83%	419 404
Moquegua	83%	176 736
Lambayeque	82%	1 200 000
Cajamarca	82%	1 500 000
Ucayali	82%	483 708
San Martín	81%	818 061
Callao	81%	982 8
La Libertas	79%	1 800 000
Ayacucho	78%	673 609
Apurimac	77%	454 324
Lima	77%	9 500 000
Loreto	77%	1 000 000
Junin	77%	1 300 000
Tacna	77%	333 276
Huanuco	74%	847 714
Madre de Dios	74%	130 876
Cusco	72%	1 300 000
Huancavelica	72%	487 472
Puno	70%	1 400 000
Pasco	69%	299 807

Así, se encontró que los cinco (5) departamentos con la mayor cobertura son: Ica, Piura, Tumbes, Arequipa y Áncash. De estos, se decidió analizar solo cuatro, descartando Tumbes por tener un tamaño relativamente pequeño. A estos cuatro (4) departamentos se aplicarán las metodologías indicadas.

6. Aplicación de Metodologías de Estimación de Áreas Pequeñas

A continuación se presentan los algoritmos de estimación de las metodologías desarrolladas por (Elbers, Lanjouw, & Lanjouw, Micro–Level Estimation of Poverty and Inequality, 2003); y (Deaton & Tarozzi, 2009).

6.1. Deaton y Tarozzi

Pobreza

Tal como se explicó en el apartado correspondiente a metodologías, Tarozzi & Deaton (2009) calculan la probabilidad de que un hogar sea pobre o no en los cuatro departamentos escogidos $P(y_h \leq z | X_h; \gamma)$.

Para ello, emplean un modelo de variable dependiente bivariada (modelo logístico), donde la variable línea toma el valor de 1 si los ingresos del hogar no superan la línea de pobreza establecida por el INEI y 0 de otro modo.

quietly logit linea cable internet telefono pob1599 pet1564 edusup2 ratsup4 piso2 pared2 ltamhog costa urbano if region==`departamento' & base_datos==0 [pweight=factor07], vce(robust)

Luego, a través del comando predict, el programa estima la probabilidad de que el ingreso de los hogares no supere la línea de pobreza en el censo.

quietly predict lineaestimada_`departamento'_`prov'

Finalmente, para obtener la media de dicha probabilidad para cada provincia en los cuatro departamentos escogidos (Áncash, Arequipa, Ica y Tacna), simplemente habría que sumar las probabilidades obtenidas en cada provincia y dividirlas entre el número de hogares existentes en cada provincia.

El comando summarize ejecuta esta operación para cada provincia en los cuatro departamentos elegidos. De este modo, se obtiene la pobreza por provincia en el Perú $(\widehat{W}_A)$. En consecuencia, para cada departamento, se crea una matriz cuyo orden es $N_A x 1$ donde N_A es el número de provincias que posee. Ahí se almacena la pobreza para cada provincia en ese departamento.

quietly summarize lineaestimada_`departamento'_`prov' if region==`departamento' & provincia==`prov' & base_datos==1
matrix pobreza_`departamento'[`prov',1]=r(mean)

Varianza

La varianza del estimador de la pobreza se descompone de la siguiente manera:

$$\widehat{Var}\left(\widehat{W}_A\right)_{\mathbf{1x1}} = \hat{G}_{1\,\mathbf{1xk}}\widehat{Var}(\hat{\gamma})_{\mathbf{kxk}}\hat{G}'_{\mathbf{kx1}}$$

La matriz de varianzas y covarianzas $\widehat{Var}(\hat{\gamma})_{\mathbf{kxk}}$ para cada departamento es calculada por Stata en la regresión logística realizada inicialmente:

matrix varbeta_`departamento'_`prov'=e(V)

Mientras que, para obtener la matriz G, primero se debe calcular la distribución marginal de la función logística $f(xb)=F(xb)(1-F(xb))$.

quietly generate Fprima_`departamento'_`prov'=lineaestimada_`departamento'_`prov' (1-lineaestimada_`departamento'_`prov')*

Para luego multiplicarla por las variables explicativas en el censo ($f(xb)*X$) y obtener la matriz G de medias ($f(xb)*X)/N_A$, operación realizada en Stata por medio de la función *accum:*

quietly foreach var of varlist cable internet telefono pob1599 pet1564 edusup2 ratsup4 piso2 pared2 ltamhog costa urbano

{generate _`var'_`departamento'_`prov'=Fprima_`departamento'_`prov'*`var'}

*quietly matrix accum G_`departamento'_`prov'= _cable_`departamento'_`prov'_ internet_`departamento'_`prov' /**

/ _telefono_`departamento'_`prov' _pob1599_`departamento'_`prov'_ pet1564_`departamento'_`prov' _edusup2_`departamento'_`prov' /

/ _ratsup4_`departamento'_`prov' _piso2_`departamento'_`prov' _pared2_`departamento'_`prov' _ltamhog_`departamento'_`prov'/

**/ _costa_`departamento'_`prov' _urbano_`departamento'_`prov' Fprima_`departamento'_`prov' if region==`departamento' & provincia==`prov' & base_datos==1, means(M) nocons*

Con ambos componentes obtenidos, simplemente se obtiene el producto matricial de la siguiente manera:

*matrix varianza_`departamento'[`prov',1]=vecdiag(M*varbeta_`departameto'_`prov'*M')'*

Sesgo

En cuanto al sesgo del error cuadrático medio, Tarozzi & Deaton (2009) señalan que este no debe tomar en cuenta sólo la varianza de la pobreza sino también la diferencia de la pobreza calculada en la encuesta y como se obtiene en el censo. Así, este se define de la siguiente manera:

$$\hat{b}^2(\widehat{W}_A) = \frac{1}{N_A} \sum_{h \in H(A)} [(p_h - 1(y_h < z))^2] + \frac{N_A - 1}{N_A} \sum_{h \in H(A)} [(p_h - 1(y_h < z))(p_{h'} - 1(y_{h'} < z))]$$

El primer componente de este sesgo es:

$$\sum_{h \in H(A)} (P(y_h \leq z | X_h;\ \gamma) - 1(y_h < z))^2)$$

Para su cálculo, primero se debe restar la probabilidad de que el hogar sea pobre en un área menor ($P(y_h \leq z \mid X\_h;\ \gamma)$) con la línea de pobreza calculada por el INEI ($1(y_h < z)$) de la siguiente manera:

```
quietly generate desv_`departamento'_`prov'=(linea-lineaestimada_`departamento'_`prov')
if region==`departamento' & provincia==`prov' & base_datos==0
```

Posteriormente se obtiene dicha diferencia al cuadrado y, finalmente, se obtiene su media por medio del comando summarize (los resultados se almacenan en la matriz sesgovarianza_`departamento' de orden igual al número de provincias (`prov') por uno):

```
quietly generate desv2_`departamento'_`prov'=desv_`departamento'_`prov'^2
quietly summarize desv2_`departamento'_`prov' if region==`departamento' & provincia==`prov'
matrix sesgovarianza_`departamento'[`prov',1]=r(mean)
```

El segundo componente está dado por:

$$\sum_{h \in H(A)} (P(y_h \leq z | X_h;\ \gamma) - 1(y_h < z))(P(y_{h'} \leq z | X_{h'};\ \gamma) - 1(y_{h'} < z))]$$

Para calcular el segundo componente para cada área menor, primero se generaron matrices W de orden igual al número de hogares en dicha área menor por el número de hogares en dicha área menor $(N_A x N_A)$. En ella se resta una matriz de unos con una matriz identidad, de modo que sólo los componentes distintos a la diagonal sean unos.

quietly summarize unos if region==`departamento' & provincia==`prov' & base_datos==0 & desv_`departamento'_`prov'!=.

scalar cantidadhogares_`departamento'_`prov'=r(N)

matrix W_`departamento'_`prov'=J(cantidadhogares_`departamento'_`prov',cantidadhogares_ `departamento'_`prov',1)-I(cantidadhogares_`departamento'_`prov')

Luego, se generaron matrices denominadas **desv** de orden $N_A x1$ que contenían las desviaciones $(P(y_h \leq z|X_h;\ \gamma) - 1(y_h < z))$ para cada área menor. Así, por medio de la multiplicación de la transpuesta de la matriz desv por la matriz W por la matriz desv, se obtiene una matriz de orden 1x1 denominada A que contiene el segundo componente del sesgo.

preserve

drop if region!=`departamento' & provincia!=`prov' & base_datos!=0

drop if desv_`departamento'_`prov'==.

sort region provincia

mkmat desv_`departamento'_`prov' if region==`departamento' & provincia==`prov' & base_ datos==0, matrix(desviacion_`departamento'_`prov')

*matrix A_`departamento'_`prov'=desviacion_`departamento'_`prov''*W_`departameto'_ `prov'*desviacion_`departamento'_`prov'*

count
matrix sesgocovarianza_`departamento'[`prov',1]=A_`departamento'_`prov'[1,1]/(`r(N)')
restore

Finalmente, el sesgo se calcula a partir de la fórmula expresada por Tarozzi & Deaton (2009):

$$\hat{b}^2(\widehat{W}_A) = \frac{1}{N_A} \sum_{h \in H(A)} [(p_h - 1(y_h < z))^2] + \frac{N_A - 1}{N_A} \sum_{h \in H(A)} [(p_h - 1(y_h < z))(p_{h'} - 1(y_{h'} < z))]$$

scalar ssesgo_`departamento'_`prov'=sesgovarianza_`departamento'[`prov',1]/ (cantidadhogares_`departamento'_`prov')+(max(0,sesgocovarianza_`departamento'[`prov',1])) (cantidadhogares_`departamento'_`prov'-1)/(cantidadhogares_`departamento'_`prov')*

Intervalo de confianza

Los autores definen el intervalo de confianza de la siguiente forma:

$$\widehat{W}_A \pm \Phi^{-1}(1 - \tau/2) x (\widehat{Var}(\widehat{W}_A) + \widehat{b^2}(\widehat{W}_A)$$

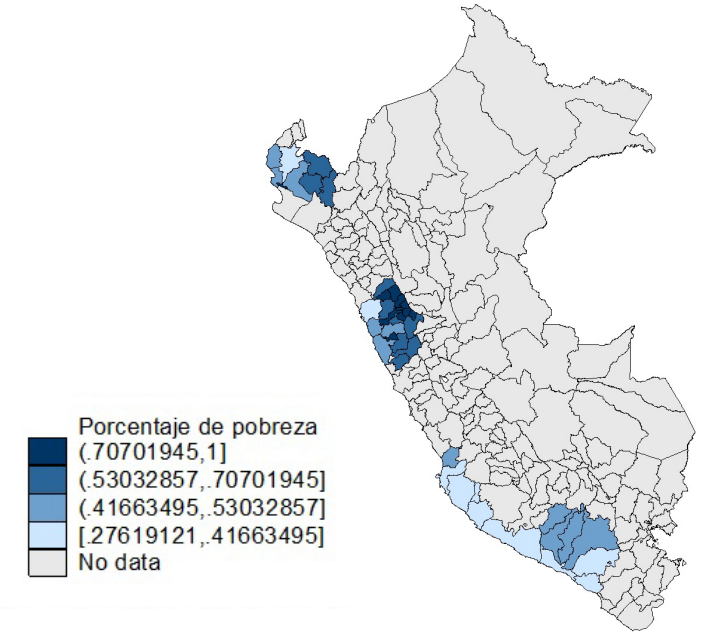
Tabla N° 2: Resultados del ejercicio Pobreza provincial (2007)**

Región	Provincia	Pobreza	Varianza	Sesgo	Intervalo superior	Intervalo inferior
ÁNCASH	HUARAZ	47.59%	0.07%	0.10%	47.60%	47.57%
ÁNCASH	AIJA	0.00%	0.00%	0.00%	0.00%	0.00%
ÁNCASH	ANTONIO RAYMONDI	75.63%	0.12%	18.95%	76.83%	74.44%
ÁNCASH	ASUNCION	71.20%	0.14%	1.06%	71.28%	71.13%
ÁNCASH	BOLOGNESI	62.04%	0.08%	19.02%	63.23%	60.84%
ÁNCASH	CARHUAZ	69.16%	0.11%	25.20%	70.74%	67.57%
ÁNCASH	CARLOS FERMIN FITZCARRALD	75.14%	0.15%	0.75%	75.19%	75.08%
ÁNCASH	CASMA	50.93%	0.11%	0.32%	50.95%	50.90%
ÁNCASH	CORONGO	0.00%	0.00%	0.00%	0.00%	0.00%
ÁNCASH	HUARI	70.70%	0.10%	58.35%	74.37%	67.04%
ÁNCASH	HUARMEY	46.15%	0.11%	0.54%	46.19%	46.11%
ÁNCASH	HUAYLAS	67.70%	0.11%	4.51%	67.98%	67.41%
ÁNCASH	MARISCAL LUZURIAGA	77.78%	0.15%	11.63%	78.52%	77.05%
ÁNCASH	OCROS	64.11%	0.07%	3.14%	64.31%	63.91%
ÁNCASH	PALLASCA	68.12%	0.09%	0.69%	68.17%	68.07%
ÁNCASH	POMABAMBA	73.31%	0.13%	18.63%	74.48%	72.13%
ÁNCASH	RECUAY	61.13%	0.07%	127.12%	69.10%	53.15%
ÁNCASH	SANTA	34.17%	0.10%	0.04%	34.17%	34.16%
ÁNCASH	SIHUAS	73.97%	0.13%	117.34%	81.34%	66.60%
ÁNCASH	YUNGAY	73.01%	0.13%	0.50%	73.05%	72.97%
AREQUIPA	AREQUIPA	27.62%	0.08%	1.51%	27.72%	27.52%
AREQUIPA	CAMANA	35.75%	0.10%	0.30%	35.77%	35.72%
AREQUIPA	CARAVELI	38.20%	0.12%	1.31%	38.29%	38.11%
AREQUIPA	CASTILLA	41.66%	0.15%	0.76%	41.72%	41.61%
AREQUIPA	CAYLLOMA	47.45%	0.18%	0.27%	47.48%	47.42%
AREQUIPA	CONDESUYOS	49.48%	0.20%	101.84%	55.87%	43.08%
AREQUIPA	ISLAY	29.39%	0.07%	0.21%	29.41%	29.38%
AREQUIPA	LA UNION	53.03%	0.24%	78.80%	57.99%	48.08%
ICA	ICA	35.36%	0.11%	0.05%	35.37%	35.35%
ICA	CHINCHA	42.12%	0.16%	0.07%	42.14%	42.11%
ICA	NAZCA	31.27%	0.11%	0.14%	31.29%	31.26%
ICA	PALPA	33.12%	0.15%	21.02%	34.45%	31.79%
ICA	PISCO	39.96%	0.14%	0.13%	39.97%	39.94%
PIURA	PIURA	72.69%	0.20%	185.09%	84.30%	61.07%
PIURA	AYABACA	52.66%	0.07%	136.76%	61.24%	44.08%
PIURA	HUANCABAMBA	69.50%	0.21%	0.15%	69.52%	69.47%
PIURA	MORROPON	61.64%	0.11%	33.79%	63.76%	59.51%
PIURA	PAITA	55.30%	0.11%	7.48%	55.78%	54.83%
PIURA	SULLANA	52.35%	0.09%	30.80%	54.29%	50.42%
PIURA	TALARA	37.63%	0.08%	0.16%	37.65%	37.62%
PIURA	SECHURA	52.55%	0.14%	2.25%	52.70%	52.40%

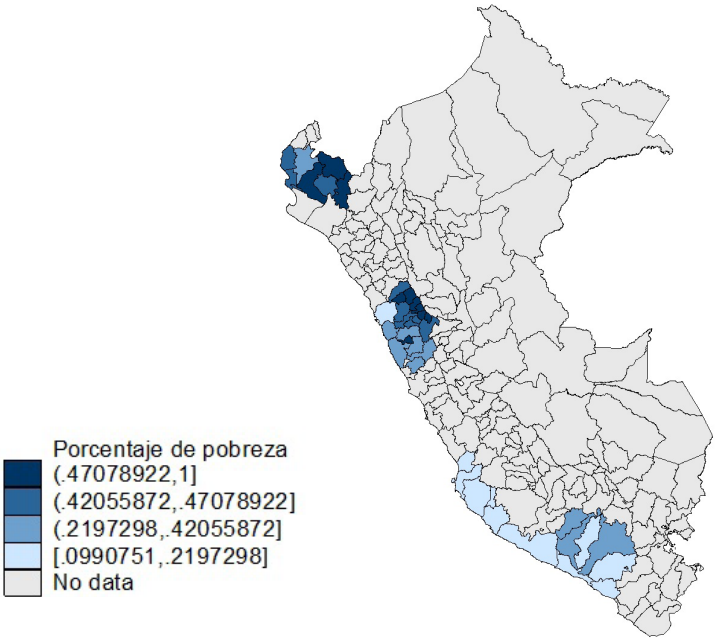
**Aquellas provincias con pobreza igual a 0% carecían de datos en la encuesta

Gráficos de la pobreza

Mapa N° 1: Estimación de pobreza en áreas menores Metodología de Tarozzi & Deaton (2007)



Mapa N° 2: Estimación de pobreza en áreas menores Metodología de Tarozzi & Deaton (2013)



6.2. Elbers, Lanjouw y Lanjouw

Pobreza

Elbers, Lanjouw, & Lanjouw (2003) calculan el nivel de pobreza departamental estimando el logaritmo del nivel de gasto per cápita por hogar y comparando si este cruza, o no, el logaritmo de la línea de pobreza correspondiente.

Para ello, emplean un modelo lineal que relaciona las características propias de la vivienda y el hogar, con el nivel de gasto per cápita dentro del mismo. De esta manera, se corre un MCO ajustado por clúster que genere los estimadores indicados. En el siguiente algoritmo, la expresión *$dependiente<.* se asegura que solamente se usen las observaciones de la ENAHO en la regresión.

reg $dependiente $independientes if $dependiente<. [pweight= factor07] , cluster(provincia)

Donde *$independientes* llama a las variables explicativas de la regresión, y *$dependiente* hace lo mismo para la variable endógena, y la variable *factor07* incorpora el factor de expansión de la encuesta dentro de la regresión.

Adicionalmente, la metodología incorpora la simulación de los errores. En este caso, existen dos: un error clúster y un error idiosincrático. Para el primero, se tomará la media por clúster (provincias) de los errores

predict error if $dependiente<. , residual
egen errorclus = mean(error) if $dependiente<. , by (provincia)

Para el segundo, se tomará el resto del error.

generate errorid = error - errorclus if $dependiente<.

Además, considerando la heterocedasticidad del error, se aproximan las varianzas del error idiosincrático a través de su valor predicho elevado al cuadrado. En consecuencia, se emplea un modelo no lineal que toma, en este ejercicio, al valor ajustado del logaritmo del gasto per cápita y su cuadrado como regresores.

*nl (errorid2 = exp({b1}*varh+{b2}*varh2)/(1+exp({b1}*varh+{b2}*varh2))) if $dependiente<. [pweight=factor07] , variables(errorid2 varh varh2)*

Con los estimadores de ambos modelos, y el conjunto de errores (clúster e idiosincrático), se predice el logaritmo del gasto per cápita para cada hogar.

Entonces, se emplea una simulación de Monte Carlo de 500 repeticiones. En cada una de las repeticiones se toma una muestra de los estimadores (asumiendo que conservan una distribución normal) y se imputan ambos errores de forma aleatoria.

Al comparar los valores simulados del logaritmo del gasto per cápita por hogar con las líneas de pobreza, se define si un hogar es pobre.

```
generate sufrepobreza=0
replace sufrepobreza=1 if simulado<=lnlinea
```

Luego, se consigue el porcentaje de pobreza por provincia.

```
generate pobpobre=sufrepobreza*pnum
collapse (rawsum) pobpobre (rawsum) pnum , by (provincia)
```

Empleando estas simulaciones de Monte Carlo, se obtiene una muestra con niveles de pobreza para cada provincia. Finalmente, con el comando *summarize* se obtiene cada una de las medias y desviaciones estándar para la generación de los intervalos de confianza.

Tabla N° 3: Resultados del ejercicio: "Elbers, Lanjouw y Lanjouw" Pobreza provincial (2007)

Región	Provincia	Pobreza	Intervalo superior	Intervalo inferior
ÁNCASH	AIJA	58%	91,58%	23,58%
ÁNCASH	ANTONIO RAYMONDI	62%	96,06%	28,90%
ÁNCASH	ASUNCION	59%	92,39%	26,34%
ÁNCASH	BOLOGNESI	52%	86,05%	18,38%
ÁNCASH	CARHUAZ	57%	92,94%	20,72%
ÁNCASH	CARLOS FERMIN FITZCARRALD	63%	95,13%	30,76%
ÁNCASH	CASMA	44%	80,66%	7,89%
ÁNCASH	CORONGO	60%	94,39%	25,72%
ÁNCASH	HUARAZ	46%	78,93%	13,71%
ÁNCASH	HUARI	58%	92,10%	24,30%
ÁNCASH	HUARMEY	43%	78,41%	7,43%
ÁNCASH	HUAYLAS	58%	91,35%	24,50%
ÁNCASH	MARISCAL LUZURIAGA	63%	98,85%	27,04%
ÁNCASH	OCROS	53%	88,09%	17,05%
ÁNCASH	PALLASCA	60%	92,24%	27,18%
ÁNCASH	POMABAMBA	63%	96,41%	30,35%
ÁNCASH	RECUAY	54%	88,55%	19,34%
ÁNCASH	SANTA	39%	73,27%	4,38%
ÁNCASH	SIHUAS	62%	96,35%	27,82%
ÁNCASH	YUNGAY	59%	94,49%	24,50%
AREQUIPA	AREQUIPA	48%	76,60%	19,40%
AREQUIPA	CAMANA	43%	71,80%	13,79%
AREQUIPA	CARAVELI	51%	79,04%	23,30%
AREQUIPA	CASTILLA	53%	82,05%	23,10%
AREQUIPA	CAYLLOMA	58%	86,59%	28,70%
AREQUIPA	CONDESUYOS	50%	76,12%	23,07%
AREQUIPA	ISLAY	49%	77,29%	20,05%
AREQUIPA	LA UNION	59%	87,20%	30,96%
ICA	CHINCHA	43%	67,12%	19,42%
ICA	ICA	37%	59,56%	14,29%
ICA	NAZCA	37%	59,10%	15,59%
ICA	PALPA	42%	66,70%	16,76%
ICA	PISCO	42%	65,35%	18,28%
PIURA	AYABACA	75%	103,17%	45,94%
PIURA	HUANCABAMBA	77%	102,94%	50,54%
PIURA	MORROPON	57%	97,28%	17,58%
PIURA	PAITA	56%	97,08%	15,24%
PIURA	PIURA	54%	91,69%	15,51%
PIURA	SECHURA	56%	66,07%	46,33%
PIURA	SULLANA	54%	93,39%	14,57%
PIURA	TALARA	48%	87,36%	8,20%

6.2.1. Estimación con dos encuestas consecutivas

Para mejorar la precisión del cálculo de la pobreza en áreas menores, se puede tomar más de una encuesta en la realización del ejercicio. En este caso, el uso de dos encuestas consecutivas permite incrementar el número de observaciones sobre la cual se estimará la pobreza. Sin embargo, los nuevos resultados tendrían sentido si se presenta una buena estimación y la distribución de la incidencia de pobreza ajustada, se comporte como una normal. De lo contrario, esta nueva estimación no tendría mayor efecto y sería mejor conservar los resultados del ejercicio con una sola encuesta. A continuación, se proyectan los índices de pobreza para el 2007 con data censal 2007 y dos encuestas consecutivas (2006 y 2007) para las provincias del departamento de Áncash.

Tabla N° 4: Resultados del ejercicio con dos encuestas (ELLPobreza provincial (2007)

Región	Provincia	Pobreza	Intervalo superior	Intervalo inferior
ÁNCASH	AIJA	57%	91,82%	23,03%
ÁNCASH	ANTONIO RAYMONDI	61%	95,10%	27,64%
ÁNCASH	ASUNCION	58%	93,10%	23,11%
ÁNCASH	BOLOGNESI	52%	85,45%	19,05%
ÁNCASH	CARHUAZ	56%	91,85%	19,71%
ÁNCASH	CARLOS FERMIN FITZCARRALD	63%	95,73%	29,52%
ÁNCASH	CASMA	41%	74,83%	6,40%
ÁNCASH	CORONGO	58%	92,99%	22,46%
ÁNCASH	HUARAZ	46%	79,88%	12,77%
ÁNCASH	HUARI	57%	90,54%	24,34%
ÁNCASH	HUARMEY	40%	74,12%	5,56%
ÁNCASH	HUAYLAS	57%	90,73%	23,98%
ÁNCASH	MARISCAL LUZURIAGA	61%	96,72%	25,68%
ÁNCASH	OCROS	53%	88,56%	17,75%
ÁNCASH	PALLASCA	58%	92,65%	24,00%
ÁNCASH	POMABAMBA	63%	96,31%	29,36%
ÁNCASH	RECUAY	54%	89,75%	18,37%
ÁNCASH	SANTA	36%	69,65%	2,66%
ÁNCASH	SIHUAS	61%	95,99%	26,63%
ÁNCASH	YUNGAY	57%	92,59%	22,23%

A continuación, se presentan los resultados obtenidos para el cálculo de la pobreza con dos encuestas consecutivas (2012 y 2013) y el barrido censal del 2012-2013 a partir de la metodología de Tarozzi & Deaton (2009).

Tabla N° 5: Resultados del ejercicio con dos encuestas (Tarozzi & Deaton)

Pobreza provincial (2012-2013)

Región	Provincia	Pobreza	Varianza	Sesgo	Intervalo superior	Intervalo inferior
ÁNCASH	HUARAZ	32.35%	0.03%	92.70%	38.17%	26.54%
ÁNCASH	AIJA	49.32%	0.08%	3.14%	49.53%	49.12%
ÁNCASH	ANTONIO RAYMONDI	51.30%	0.11%	9.98%	51.93%	50.67%
ÁNCASH	ASUNCION	48.15%	0.10%	113.82%	55.30%	41.01%
ÁNCASH	BOLOGNESI	42.82%	0.05%	0.24%	42.84%	42.80%
ÁNCASH	CARHUAZ	45.17%	0.08%	95.37%	51.15%	39.18%
ÁNCASH	CARLOS FERMIN FITZCARRALD	51.87%	0.13%	109.54%	58.75%	45.00%
ÁNCASH	CASMA	29.37%	0.05%	0.11%	29.38%	29.36%
ÁNCASH	CORONGO	0.00%	0.00%	0.00%	0.00%	0.00%
ÁNCASH	HUARI	47.42%	0.09%	75.53%	52.16%	42.68%
ÁNCASH	HUARMEY	26.16%	0.05%	0.15%	26.18%	26.15%
ÁNCASH	HUAYLAS	45.31%	0.09%	9.60%	45.92%	44.70%
ÁNCASH	MARISCAL LUZURIAGA	53.22%	0.13%	304.88%	72.35%	34.10%
ÁNCASH	OCROS	43.32%	0.05%	0.49%	43.35%	43.29%
ÁNCASH	PALLASCA	47.79%	0.08%	29.28%	49.63%	45.94%
ÁNCASH	POMABAMBA	50.09%	0.13%	139.54%	58.84%	41.33%
ÁNCASH	RECUAY	42.93%	0.04%	0.35%	42.95%	42.90%
ÁNCASH	SANTA	21.98%	0.04%	0.01%	21.99%	21.98%
ÁNCASH	SIHUAS	49.34%	0.11%	0.33%	49.37%	49.31%
ÁNCASH	YUNGAY	47.16%	0.10%	0.19%	47.18%	47.14%
AREQUIPA	AREQUIPA	13.92%	0.02%	8.84%	14.48%	13.37%
AREQUIPA	CAMANA	19.91%	0.03%	0.05%	19.92%	19.91%
AREQUIPA	CARAVELI	21.35%	0.04%	54.22%	24.75%	17.95%
AREQUIPA	CASTILLA	25.93%	0.05%	0.13%	25.94%	25.91%
AREQUIPA	CAYLLOMA	30.96%	0.07%	43.74%	33.71%	28.21%
AREQUIPA	CONDESUYOS	35.62%	0.09%	0.14%	35.64%	35.61%
AREQUIPA	ISLAY	15.40%	0.02%	0.05%	15.40%	15.39%
AREQUIPA	LA UNION	42.74%	0.12%	151.28%	52.24%	33.25%
ICA	ICA	10.01%	0.04%	0.00%	10.01%	10.00%
ICA	CHINCHA	12.55%	0.06%	0.01%	12.56%	12.55%
ICA	NAZCA	8.56%	0.03%	9.36%	9.15%	7.97%
ICA	PALPA	8.24%	0.03%	0.04%	8.24%	8.23%
ICA	PISCO	11.07%	0.05%	10.86%	11.75%	10.38%
PIURA	PIURA	40.24%	0.04%	89.63%	45.86%	34.61%
PIURA	AYABACA	59.63%	0.16%	169.01%	70.23%	49.02%
PIURA	HUANCABAMBA	57.39%	0.15%	0.10%	57.40%	57.37%
PIURA	MORROPON	48.15%	0.06%	126.66%	56.10%	40.21%
PIURA	PAITA	37.24%	0.05%	101.68%	43.62%	30.86%
PIURA	SULLANA	37.57%	0.04%	0.03%	37.58%	37.57%
PIURA	TALARA	26.06%	0.03%	2.87%	26.24%	25.88%
PIURA	SECHURA	37.27%	0.06%	0.20%	37.28%	37.25%

Intervalo de confianza

Los autores definen el intervalo de confianza de la siguiente forma:

$$\widehat{W}_A \pm \Phi^{-1}(1-\tau/2)x(\widehat{Var}(\widehat{W}_A) + \widehat{b^2}(\widehat{W}_A)$$

Donde, $\widehat{b^2}(\widehat{W}_A)$, es el sesgo del error cuadrático medio.

6.3. Coung (2011)

Pobreza

A continuación, se desarrolla el ejercicio propuesto en la metodología de Coung replicado al caso peruano. Además, se proyectan los índices de pobreza para el 2007 con data censal del 2007. Así se obtiene la tabla con la pobreza provincial del 2009 para Áncash, Arequipa, Ica y Piura la cual se presenta a continuación:

Tabla N° 6: Pobreza provincial (2009)

Región	Provincia	Pobreza
ÁNCASH	HUARAZ	27%
ÁNCASH	AIJA	34%
ÁNCASH	ANTONIO RAYMONDI	38%
ÁNCASH	ASUNCION	34%
ÁNCASH	BOLOGNESI	32%
ÁNCASH	CARHUAZ	31%
ÁNCASH	CARLOS FERMIN FITZCARRALD	38%
ÁNCASH	CASMA	38%
ÁNCASH	CORONGO	36%
ÁNCASH	HUARI	33%
ÁNCASH	HUARMEY	36%
ÁNCASH	HUAYLAS	36%
ÁNCASH	MARISCAL LUZURIAGA	37%
ÁNCASH	OCROS	31%
ÁNCASH	PALLASCA	35%
ÁNCASH	POMABAMBA	40%
ÁNCASH	RECUAY	31%
ÁNCASH	SANTA	36%
ÁNCASH	SIHUAS	37%
ÁNCASH	YUNGAY	38%
AREQUIPA	AREQUIPA	37%
AREQUIPA	CAMANA	17%
AREQUIPA	CARAVELI	35%
AREQUIPA	CASTILLA	32%
AREQUIPA	CAYLLOMA	40%
AREQUIPA	CONDESUYOS	28%
AREQUIPA	ISLAY	35%
AREQUIPA	LA UNION	39%
ICA	ICA	31%
ICA	CHINCHA	36%
ICA	NAZCA	33%
ICA	PALPA	36%
ICA	PISCO	34%
PIURA	PIURA	47%
PIURA	AYABACA	67%
PIURA	HUANCABAMBA	69%
PIURA	MORROPON	52%
PIURA	PAITA	48%
PIURA	SULLANA	47%
PIURA	TALARA	44%
PIURA	SECHURA	50%

7. Conclusiones y Recomendaciones

El presente estudio desarrolla los algoritmos necesarios para estimar la pobreza para áreas menores a través de las metodologías (Elbers, Lanjouw, & Lanjouw, Micro–Level Estimation of Poverty and Inequality, 2003), (Deaton & Tarozzi, 2009) y (Coung, 2011).

Las metodologías fueron aplicadas para 4 departamentos específicos: Ica, Piura, Arequipa y Áncash; debido a que mostraban el mayor grado de cobertura en el barrido censal de 2012 y 2013. Asimismo, el cálculo se realizó para un subconjunto de las variables empleadas por el INEI en la generación del mapa de pobreza del 2013.

Los resultados permiten tener información sobre el nivel de bienestar a nivel provincial y distrital con un acertado nivel de confianza.

La pobreza calculada a partir de la metodología de Tarozzi & Deaton y aquella obtenida a partir del ELL arrojan porcentajes similares. Esto último confirma que la metodología de Tarozzi & Deaton es una simplificación de la de ELL.

En el caso específico de Coung, este utiliza un panel de datos para predecir la pobreza de años posteriores. Este hecho representa un significativo avance en la estimación de índices de bienestar en áreas menores.

La aplicación y publicación de estas metodologías, ofrecería una mayor replicabilidad de los resultados estimados. Así, se recomienda que estos algoritmos sean utilizados en otras regiones a fin de cotejar los resultados obtenidos hasta ahora.

8. Bibliografía

Central Bureau of Statistics; Government of Nepal; United Nations World Food Programme & The World Bank. (2006). *Small Area Estimation of Poverty, Caloric Intake and Malnutrition in Nepal. Katmandú: Government of Nepal.*

Coung, N. V. (2011). Poverty projection using a small area estimation method: Evidence from Vietnam. *Journal of Comparative Economics, 368-382.*

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003, January). Micro-level estimation of poverty and inequality. *Econometrica, 71(1), 355-364.*

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro–Level Estimation of Poverty and Inequality. *Econometrica, 355-364.*

Esteban, M., Morales, D., Pérez, A., & Santamaría, L. (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis, 2840-2855.*

Ministerio de Desarrollo Social. (2013, Febrero 6). Ministerio de Desarrollo Social. Retrieved from Ministerio de Desarrollo Social: http://observatorio.ministeriodesarrollosocial.gob.cl/indicadores/docs/Incidencia_de_la_Pobreza_Comunal_Chile_2009y2011_SAE_11feb13_5284f2200bd3e.pdf

Poverty Reduction & Economic Management, South Asia Region, World Bank. (2005). *A Poverty Map for Sri Lanka—Findings and Lessons.* Washington DC: Department of Census and Statistics, Colombo, Sri Lanka, and The World Bank.

Rao, J. N., & Molina, I. (2015). *Small Area Estimation.* Nueva Jersey: John Wiley & Sons.

Robles, M., & Santander, H. (2004). Paraguay: Pobreza y desigualdad de ingresos a nivel distrital. 49.

Tarozzi, A., & Deaton, A. (2009). Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas. *The Review of Economics and Statistics, 773-792.*

9. Anexos

9.1. VARIABLES UTILIZADAS

Linea: línea de pobreza de acuerdo al INEI

Cable: *dummy* de valor igual a 1 si el hogar posee cable y 0 de otro modo

Internet: *dummy* de valor igual a 1 si el hogar posee internet y 0 de otro modo

Pob1599: Miembros del hogar de 15 a más años de edad

Pet1564: Número de miembros con edades entre 16 y 64 años

Edusup2: Número de miembros del hogar de 18 a más años con educacion superior universitaria completa

Ratsup4: Ratio de la población de 18 a 24 que asiste a la universidad entre la poblacion total de 18 a 24 años

Piso2: *dummy* de valor igual a 1 si el hogar posee piso de vinílicos o losetas, madera o cemento y 0 de otro modo

Pared2: *dummy* de valor igual a 1 si el hogar posee pared exterior de piedra o sillar con cal o cemento, adobe o tapia y 0 de otro modo

Ltamhog: logaritmo del número total de miembros del hogar

Costa: *dummy* de valor igual a 1 si el hogar se encuentra en la costa y 0 de otro modo

Urbano: *dummy* de valor igual a 1 si el hogar se encuentra en una zona urbana y 0 de otro modo

9.2. DEATON Y TAROZZI (2009)

Tabla N° A1: Pobreza provincial (2013)

Región	Provincia	Pobreza	Varianza	Sesgo	Intervalo superior	Intervalo inferior
ÁNCASH	HUARAZ	30.66%	0.07%	57.70%	40.92%	20.41%
ÁNCASH	AIJA	48.07%	0.17%	3.06%	48.64%	47.50%
ÁNCASH	ANTONIO RAYMONDI	49.88%	0.23%	2.68%	50.40%	49.36%
ÁNCASH	ASUNCION	47.08%	0.19%	44.56%	55.03%	39.13%
ÁNCASH	BOLOGNESI	40.89%	0.11%	0.46%	40.99%	40.79%
ÁNCASH	CARHUAZ	43.93%	0.15%	28.00%	48.92%	38.93%
ÁNCASH	CARLOS FERMIN FITZCARRALD	50.74%	0.26%	141.14%	75.85%	25.63%
ÁNCASH	CASMA	28.88%	0.11%	0.24%	28.94%	28.82%
ÁNCASH	CORONGO	0.00%	0.00%	0.00%	0.00%	0.00%
ÁNCASH	HUARI	45.89%	0.18%	44.72%	53.86%	37.92%
ÁNCASH	HUARMEY	25.63%	0.11%	0.37%	25.72%	25.55%
ÁNCASH	HUAYLAS	44.05%	0.18%	62.62%	55.20%	32.90%
ÁNCASH	MARISCAL LUZURIAGA	52.08%	0.28%	214.27%	90.17%	13.98%
ÁNCASH	OCROS	42.06%	0.10%	32.43%	47.83%	36.28%
ÁNCASH	PALLASCA	45.78%	0.18%	0.67%	45.93%	45.63%
ÁNCASH	POMABAMBA	48.95%	0.26%	8.01%	50.42%	47.49%
ÁNCASH	RECUAY	41.28%	0.09%	0.53%	41.39%	41.17%
ÁNCASH	SANTA	20.99%	0.09%	0.02%	21.01%	20.97%
ÁNCASH	SIHUAS	48.24%	0.23%	0.65%	48.40%	48.08%
ÁNCASH	YUNGAY	45.91%	0.20%	0.34%	46.00%	45.81%
AREQUIPA	AREQUIPA	10.97%	0.03%	0.92%	11.14%	10.81%
AREQUIPA	CAMANA	17.16%	0.06%	5.61%	18.17%	16.16%
AREQUIPA	CARAVELI	18.31%	0.07%	16.08%	21.18%	15.44%
AREQUIPA	CASTILLA	21.97%	0.09%	0.16%	22.02%	21.93%
AREQUIPA	CAYLLOMA	26.91%	0.12%	108.35%	46.17%	7.65%
AREQUIPA	CONDESUYOS	31.78%	0.17%	0.27%	31.86%	31.71%
AREQUIPA	ISLAY	13.20%	0.04%	0.14%	13.24%	13.17%
AREQUIPA	LA UNION	38.06%	0.26%	78.50%	52.04%	24.07%
ICA	ICA	14.13%	0.15%	0.01%	14.16%	14.10%
ICA	CHINCHA	16.46%	0.20%	8.05%	17.93%	14.99%
ICA	NAZCA	11.47%	0.10%	0.02%	11.50%	11.45%
ICA	PALPA	9.91%	0.09%	0.09%	9.94%	9.88%
ICA	PISCO	14.76%	0.18%	8.71%	16.34%	13.19%
PIURA	PIURA	63.99%	0.26%	55.96%	73.97%	54.01%
PIURA	AYABACA	45.31%	0.10%	95.40%	62.26%	28.35%
PIURA	HUANCABAMBA	61.31%	0.26%	0.19%	61.39%	61.23%
PIURA	MORROPON	52.98%	0.13%	0.13%	53.03%	52.94%
PIURA	PAITA	43.30%	0.13%	81.89%	57.86%	28.73%
PIURA	SULLANA	43.07%	0.10%	0.05%	43.09%	43.04%
PIURA	TALARA	31.32%	0.11%	0.10%	31.36%	31.28%
PIURA	SECHURA	43.47%	0.16%	0.32%	43.55%	43.39%

Pasos para correr el programa

1. En un directorio determinado, crear las siguientes tres carpetas: Bases de datos, Datos del mapa e Imágenes.

2. En la carpeta Bases de datos, almacenar el archivo zip ENAHO2007 y extraerlo en esa dirección. Asimismo, almacenar las dos bases de datos correspondientes al censo: *base concatenada vivienda 2007 (02 de agosto).dta y base_concatenada_poblacion_2007_02_de_agosto_stata.dta*

3. En la carpeta Datos del mapa, almacenar y extraer los archivos zip *Limite_provincial y Limite_distrital*

4. Abrir el archivo do Base de datos y modificar la línea 10 por la dirección en la que colocó la carpeta Bases de datos. Una vez que corre el do en cuestión obtiene una base de datos a nivel de hogares con información de la ENAHO 2007 y del censo del mismo periodo.

5. Abrir el archivo do *Tarozzi & Deaton (2009)*. Allí se calcula la pobreza a nivel regional, provincial y departamental.

   a.En las líneas 1, 93 y 254 modificar la dirección donde se colocó la carpeta Base de datos.
   b.En las líneas 198, 201, 378 y 381 modificar la dirección donde se colocó la carpeta Datos del mapa.
   c.En las líneas 199 y 379 modificar la dirección donde se colocó la carpeta Imágenes.
   d.Luego de correr el do file, se obtiene la pobreza a nivel regional, provincial y departamental.

Código de STATA

```
*-------------------------------------------------------------------------------------------------------------------
*2.1 A NIVEL PROVINCIAL
*-------------------------------------------------------------------------------------------------------------------
global b "C:\Users\Inspiron\Desktop\ "
use "$b\basededatos.dta", clear
sort region provincia
set matsize 10000

*DEFINIMOS LAS MATRICES QUE ALMACENAN LAS VARIABLES DE INTERÉS

forvalues departamento=1(1)4 {
summarize provincia if region==`departamento' & base_datos==1

        forvalues prov=1(1)`r(max)' {
        matrix pobreza_`departamento'=J(`prov',1,1)
        matrix varianza_`departamento'=J(`prov',1,1)
        matrix sesgovarianza_`departamento'=J(`prov',1,1)
        matrix sesgocovarianza_`departamento'=J(`prov',1,1)
        matrix sesgo_`departamento'=J(`prov',1,1)
        matrix IS_`departamento'=J(`prov',1,1)
        matrix II_`departamento'=J(`prov',1,1)        } }
```

```
*APLICAMOS LA METODOLOGÍA PROPUESTA POR DEATON A NIVEL PROVINCIAL
forvalues departamento=1(1)4 {
summarize provincia if region==`departamento' & base_datos==1

forvalues prov=1(1)`r(max)' {

*Si existen datos para la encuesta en esa provincia, continuar con los cálculos

count if region==`departamento' & provincia==`prov' & base_datos==0
if `r(N)'!=0 {

quietly logit linea cable internet telefono pob1599 pet1564 edusup2 piso2 pared2 ltamhog costa urbano if region==`departamento' /*
*/ & base_datos==0 [pweight=factor07], vce(robust)
matrix varbeta_`departamento'_`prov'=e(V)


*ESTIMADOR DE POBREZA

quietly predict lineaestimada_`departamento'_`prov'
quietly summarize lineaestimada_`departamento'_`prov' if region==`departamento' & provincia==`prov' & base_datos==1
scalar spobreza_`departamento'_`prov'=r(mean)
quietly summarize lineaestimada_`departamento'_`prov' if region==`departamento' & provincia==`prov' & base_datos==1
matrix pobreza_`departamento'[`prov',1]=r(mean)


*VARIANZA DEL ESTIMADOR DE POBREZA

*1°: Se genera F'(X'b) que es igual a F(1-F) en el logit

quietly generate Fprima_`departamento'_`prov'=lineaestimada_`departamento'_`prov'*(1-lineaestimada_`departamento'_`prov')

*2°: Para aplicar el método Delta, se debe calcular F'(X'b)*X para cada observación
quietly foreach var of varlist cable internet telefono pob1599 pet1564 edusup2 piso2 pared2 ltamhog costa urbano {
generate _`var'_`departamento'_`prov'=Fprima_`departamento'_`prov'*`var'
                                        }
```

*3°: Se calcula la media de F'(X'b) para el CENSO

quietly matrix accum G_`departamento'_`prov'= _cable_`departamento'_`prov' _internet_`departamento'_`prov' /*

/ _telefono_`departamento'_`prov' _pob1599_`departamento'_`prov' _pet1564_`departamento'_`prov' _ edusup2_`departamento'_`prov' /

/ _piso2_`departamento'_`prov' _pared2_`departamento'_`prov' _ltamhog_`departamento'_`prov'/

*/ _costa_`departamento'_`prov' _urbano_`departamento'_`prov' Fprima_`departamento'_`prov' if region==`departamento' & provincia==`prov' & base_datos==1, means(M) nocons

*4°: Se obtiene la matriz de la varianza

matrix varianza_`departamento'[`prov',1]=vecdiag(M*varbeta_`departamento'_`prov'*M')'
scalar svarianza_`departamento'_`prov'=varianza_`departamento'[`prov',1]
matrix drop M

*SESGO EN EL ERROR CUADRÁTICO MEDIO (MSE)

*PRIMER COMPONENTE DEL SESGO (varianza)
quietly generate desv_`departamento'_`prov'=(linea-lineaestimada_`departamento'_`prov') if region==`departamento' & provincia==`prov' & base_datos==0
quietly generate desv2_`departamento'_`prov'=desv_`departamento'_`prov'^2
quietly summarize desv2_`departamento'_`prov' if region==`departamento' & provincia==`prov'
matrix sesgovarianza_`departamento'[`prov',1]=r(mean)

*SEGUNDO COMPONENTE DEL SESGO (covarianza)

*1°: Se crea una variable que arroja una secuencia numérica (1,2,3...) para una área pequeña en una región determinada
quietly capture egen hhseq=seq() if region==`departamento' & provincia==`prov' & base_datos==0

*2°: Se crea una matriz de 1´s para todos los valores distintos a la matriz diagonal para obtener las covarianzas y no las varianzas

quietly summarize unos if region==`departamento' & provincia==`prov' & base_datos==0 & desv_`departamento'_`prov'!=.

scalar cantidadhogares_`departamento'_`prov'=r(N)
matrix W_`departamento'_`prov'=J(cantidadhogares_`departamento'_`prov',cantidadhogares_`departamento'_`prov',1)-I(cantidadhogares_`departamento'_`prov')

```
*3°: Se calcula el producto de las desviaciones para hogares diferentes
preserve
drop if region!=`departamento' & provincia!=`prov' & base_datos!=0
drop if desv_`departamento'_`prov'==.
sort region provincia
mkmat desv_`departamento'_`prov' if region==`departamento' & provincia==`prov' & base_datos==0, matrix
(desviacion_`departamento'_`prov')
matrix  A_`departamento'_`prov'=desviacion_`departamento'_`prov''*W_`departamento'_`prov'*desviacion_
departamento'_`prov'
count
matrix sesgocovarianza_`departamento'[`prov',1]=A_`departamento'_`prov'[1,1]/(`r(N)')
restore

scalar ssesgo_`departamento'_`prov'=sesgovarianza_`departamento'[`prov',1]/(cantidadhogares_`departamento'_
`prov')+(max(0,sesgocovarianza_`departamento'[`prov',1]))*(cantidadhogares_`departamento'_`prov'-1)/
(cantidadhogares_`departamento'_`prov')
matrix sesgo_`departamento'[`prov',1]=ssesgo_`departamento'_`prov'

*INTERVALOS DE CONFIANZA (AL 95%)

*Intervalo superior
scalar sintervalo_superior_`departamento'_`prov'=spobreza_`departamento'_`prov'+(1- invnomal(0.95)/2)*
(svarianza_`departamento'_`prov'+ssesgo_`departamento'_`prov'))
matrix IS_`departamento'[`prov',1]=sintervalo_superior_`departamento'_`prov'

*Intervalo inferior
scalar sintervalo_inferior_`departamento'_`prov'=spobreza_`departamento'_`prov'(1- invnormal(0.95)/2)*
(svarianza_`departamento'_`prov'+ssesgo_`departamento'_`prov'))
matrix II_`departamento'[`prov',1]=sintervalo_inferior_`departamento'_`prov'
}
else {

matrix pobreza_`departamento'[`prov',1]=0
matrix varianza_`departamento'[`prov',1]=0
matrix sesgovarianza_`departamento'[`prov',1]=0
matrix sesgocovarianza_`departamento'[`prov',1]=0
matrix sesgo_`departamento'[`prov',1]=0
matrix IS_`departamento'[`prov',1]=0
matrix II_`departamento'[`prov',1]=0            } } }

*GRAFICANDO LA POBREZA A NIVEL PROVINCIAL EN LOS CUATRO DEPARTAMENTOS ESCOGIDOS

global c "C:\Users\Inspiron\Desktop\Hans Tello\UP\Vida Profesional\Investigación\Datos del mapa"
global i "C:\Users\Inspiron\Desktop\ \Imágenes"
```

```
cd “C:\Users\Inspiron\Desktop\Datos del mapa “
shp2dta using “$c\BAS_LIM_PROVINCIA.shp”, data(mapaperu) coor(perucoor) genid(id) replace

use “$c\mapaperu.dta”, clear

generate region=substr(FIRST_IDPR,1,2)
generate provincia=substr(FIRST_IDPR,3,2)

generate orden=5
replace orden=1 if region==”02”
replace orden=2 if region==”04”
replace orden=3 if region==”11”
replace orden=4 if region==”20”

sort orden region provincia
matrix tabla_pobreza=pobreza_1\pobreza_2\pobreza_3\pobreza_4
svmat double tabla_pobreza, name(pobreza)

matrix tabla_varianza=varianza_1\varianza_2\varianza_3\varianza_4
svmat double tabla_varianza, name(varianza)

matrix tabla_IS=IS_1\IS_2\IS_3\IS_4
svmat double tabla_IS, name(IS)

matrix tabla_II=II_1\II_2\II_3\II_4
svmat double tabla_II, name(II)

matrix tabla_sesgo=sesgo_1\sesgo_2\sesgo_3\sesgo_4
svmat double tabla_sesgo, name(sesgo)

sort id
save “$c\mapafinalperu.dta”, replace


use “$c\mapafinalperu.dta”, clear

*Gráfico de la pobreza en el Perú para las cuatro provincias estudiadas

spmap pobreza using “$c\perucoor.dta”,id (id) clmethod(quantile) fcolor(Blues2) ndfcolor(dimgray)
legtitle(“Porcentaje de pobreza”)/*
*/ title(“{bf:Estimación de pobreza en áreas menores}” “{bf: Metodología de Tarozzi & Deaton (2009)}”)
note(“Fuente: INEI. Elaboración propia”)
graph export “$i\PobrezaProvincias_Peru.png”, replace

save “$c\finalperu.dta”, replace
```

```
*Para obtener la tabla en Excel
use "$c\mapafinalperu.dta", clear

drop if FIRST_NOMB!="ANCASH" & FIRST_NOMB!="AREQUIPA" & FIRST_NOMB!="ICA" & FIRST_
NOMB!="PIURA"

rename FIRST_IDPR codigoINEI

sort codigoINEI

preserve
use "$b\basededatos.dta", clear
sort region provincia distrito
keep if base_datos==1
collapse (sum) unos (sum) lavadora, by (region provincia)
generate porcentaje_lavadora=lavadora/unos
mkmat porcentaje_lavadora, matrix(lavadoras)
restore

svmat double lavadoras, name(porcentaje_lavadora)
bro FIRST_NOMB NOMBPROV pobreza varianza sesgo IS II porcentaje_lavadora
```

9.3. ELBERS, LANJOUW Y LANJOUW (2003)

Pasos para correr el programa

1. Elegir un directorio determinado y crear una carpeta nueva.

2. En la carpeta nueva, almacenar el archivo zip ENAHO2007 y extraerlo en esa dirección. Asimismo, almacenar las dos bases de datos correspondientes al censo: *base concatenada vivienda 2007 (02 de agosto).dta y base_concatenada_poblacion_2007_02_de_agosto_stata.dta.* También se debe almacenar los dos do-files: *generador_censo.do, generador_enaho.do y Elbers_Lanjouw_Lanjouw.do.*

3. Abrir el archivo do *generador_enaho* y modificar las líneas 9 y 22 con el número ccdd que represente al departamento seleccionado para la estimación. También modificar la línea 159 con el valor correspondiente al departamento. Una vez que corre el do en cuestión se obtiene una base de datos a nivel de hogares con información de la ENAHO 2007.

4. Abrir el archivo do *generador_censo* y modificar la línea 6 con la dirección de la carpeta nueva. También modificar la línea 131 con el departamento que desee estimar. Además, entre las líneas 139 y 158 validar los comandos respectivos al departamento. El comando de la línea 175 solo aplica para la estimación del departamento de Ica. Una vez que corre el do en cuestión se obtiene una base de datos a nivel de hogares con información del censo 2007 que es agregada a base de datos generada en el paso anterior.

5. Abrir el archivo do *Elbers_Lanjouw_Lanjouw.* Allí se calcula la pobreza a nivel provincial.

   a.En la línea 4 modificar la dirección por la correspondiente a la carpeta nueva.
   b.Entre las líneas 236 y 249 validar los comandos respectivos al departamento.
   c.Luego de correr el do file, se obtiene la pobreza a nivel provincial.

Código de STATA

```
* Método Elbers, Lanjouw y Lanjouw (ELL).
* Este do file aplica el método ELL para el caso específico de la base de datos adjunta.
cd " C:\Users\Inspiron\Desktop\Investigación\Bases de datos "
*(Se espera una base de datos con los resultados del censo sobre el departamento elegido y los resultados correspondientes
* para la Enaho)
* En primer lugar se estimarán los parámetros de la encuesta de hogares. Se almacenarán los estimados
* puntuales y las correspondientes matrices de varianzas y covarianzas estimadas para ser usadas
* en el procedimiento ELL. Además, guarda el error clúster y el error idiosincrático en dos bases de datos.
* Los parámetros heterocedásticos son estimados usando mínimos cuadrados no linealizados (NLLS).
* Siguiendo la versión de Deaton, los regresores para el modelo heterocedástico serán el valor ajustado
```

```
* y el mismo al cuadrado. Sin embargo, puede ser reemplazado por los regresores que se consideren pertinentes.

* Se abre la base de datos
use “basededatos.dta” , clear

*******************************************************************************
* Variable dependiente: gasto (habrá que volverlo logaritmo neperiano)
global dependiente “lngasto”

* Variables independientes:
 global independientes “ cable internet telefono pob1599 pet1564 edusup2 piso2 pared2 ltamhog ratsup4 region_natural area “

* Variables independientes modelo heterocedástico:
* global independientesh “ “

* Variable cluster ( serán las provicias del departamento)

*******************************************************************************
1° Se estiman los parámetros de la regresión lineal, corrigiendo por clúster los errores estándar

replace factor07=factor07*mieperho

reg $dependiente $independientes if $dependiente<. [pweight=factor07], cluster(provincia)
* 2° SE GUARDA EL VECTOR DE BETAS ESTIMADOS Y LA MATRIZ DE VARIANZAS Y COVARIANZAS*
matrix ell_varcovbeta = e(V)
matrix ell_beta = e(b)’

* 3° Se generan los residuos de la regresión lineal
predict error if $dependiente<., residual

* 4° SE ESTIMAN LOS ERRORES CLÚSTER
egen errorclus = mean(error) if $dependiente<., by (provincia)

* errorclus: es la variable que conserva los errores por clúster(por provincia).

* 5° Se estiman los errores idiosincráticos y el cuadrado de los mismos.
gen errorid = error - errorclus if $dependiente<.
gen errorid2 = errorid^2 if $dependiente<.

*******************************************************************************
* ESTIMACIÓN DEL MODELO HETEROCEDÁSTICO USANDO MÍNIMOS CUADRADOS NO LINEALES (NLLS)
```

```
* 6° Ahora, para la generación del error idiosincrático estandarizado, se regresionará un modelo no lineal.

* Para este caso se tomarán como regresores: la variable dependiente ajustada y su cuadrado.

* (REPITO, de todas formas se pueden usar los regresores que crean pertinentes)

* varh es el valor del ln del gasto ajustado y varh2 es el ln del gasto ajustado al cuadrado
predict varh if $dependiente<.
gen varh2=varh^2 if $dependiente<.

* Página 357 de ELL, ecuación 2 (A=1 y B=0)
nl ( errorid2 = exp({b1}*varh+{b2}*varh2)/(1+exp({b1}*varh+{b2}*varh2)))if $dependiente<. [pweight=factor07],
variables(errorid2 varh varh2)

* 7° SE GUARDA EL VECTOR DE BETAS ESTIMADOS Y LA MATRIZ DE VARIANZAS Y COVARIANZAS
EN EL MODELO HETEROCEDÁSTICO*
matrix ell_varcovbetahet = e(V)
matrix ell_betahet = e(b)'

* 8° El error al cuadrado se vuelve la aproximación a la varianza heterocedástica; entonces, es estimado.
predict varianza_est if $dependiente<.

* varianza_est: varianza estimada del error heterocedástico
*********************************************************************************
* 9° Con los estimados de la varianza ahora se pueden obtener los "residuos estandarizados"
gen errorides = errorid/sqrt(varianza_est) if $dependiente<.
egen mediaest = mean(errorides) if $dependiente <., by(provincia)
replace errorides= errorides - mediaest if $dependiente<.

*errorides: Es la variable del error idiosincrático estandarizado

* 10° Se genera una nueva identificación para los clústers, que sea por número
egen ell_psuid = group(provincia) if $dependiente<.
* ell_pusid es la nueva identificaicón clúster (provincia) por número

* 11° Se despeja solo las variables condicionadas "<."(solo se tomará la encuesta)
keep if $dependiente<.
keep ell_psuid errorclus errorides
save temporal , replace

* 12° SE GENERA EL DATAFILE CON LOS ERRORES CLUSTER (por provincia)
* Además, se conserva un escalar con el número de errores cluster
keep errorclus ell_psuid
duplicates drop ell_psuid, force
sort ell_psuid
```

```
save erclus, replace
scalar n_erclus = _N
* erclus: es un file que guarda los erroes clúster
* n_erclus: es un escalar que guarda el número de errores clúster

* 13° SE GENERA EL DATAFILE CON LOS ERRORES IDIOSINCRÁTICOS ESTANDARIZADOS
* Además, se conserva un escalar con el número de errores idiosincráticos
use errorides using temporal
gen ell_n = _n
sort ell_n
save erid, replace
scalar n_erid = _N
* erid: es un file que guarda los errores idiosincráticos estandarizados
* n_erid: es un escalar que guarda el número de errores idiosincráticos estandarizados

* 14° Se modifican los nombres de las columnas de los vectores de estimadores
mat colnames ell_beta = beta
mat colnames ell_betahet = betahet

* 15° SE SALVAN LOS VECTORES DE LOS PARÁMETROS ESTIMADOS Y LAS MATRICES DE VARIANZAS
Y COVARIANZAS
matrix varcovbeta = ell_varcovbeta
matrix beta = ell_beta
matrix varcovbetahet = ell_varcovbetahet
matrix betahet = ell_betahet

* Incluso, se conserva un vector con las desviaciones estándar de los estimadores del modelo lineal.
matrix varianzas_lineal=vecdiag(varcovbeta)'
local  reg= rowsof(varianzas_lineal)
forvalues f=1/`reg'{ matrix varianzas_lineal[`f',1]=sqrt(varianzas_lineal[`f',1]) }
* AHORA SE GENERARÁ UNA SIMULACIÓN DEL PROCEDIMINETO ELL PROPUESTO POR
TAROZZI(2009)

* CON LA AYUDA DE LOS COMPONENETES ANTES ESTIMADOS(LA ESTIMACIÓN ES REALIZADA AL
MISMO TIEMPO PARA TODAS LAS PROVINCIAS ELEGIDAS):

* 15.5° Se genenar un bucle para la simulación de montecarlo
forvalues h=1/500 {

* 16°Primero se vuelve a abrir la base de datos original que conserva los resultados del censo para el
departamento seleccionado
use "basededatos.dta" , clear

* 17° Se conserva solo las observaciones pertenecientes al censo
keep if $dependiente==.
```

```
* 18° Se generará un id para relacionar los clúster (provincias) en el censo con los errores clúster guardados en erclus
gen ell_psuid = floor(n_erclus*runiform() + 1)

* 19° ahora cada cluster (provincia) conservará un solo id que fue elegido de forma aleatoria en el paso anterior
bysort provincia: replace ell_psuid = ell_psuid[1]

* 20° Se asignará ahora los errores clúster
merge m:1 ell_psuid using erclus
drop _merge

* 21° Se generará un id para relacionar los errores idiosincráticos guardados en erid con cada observación del censo.

* En este caso, y siguiendo con la independencia entre errores, el error idiosincrático será tomado sin importar

* el error clúster que tome la observación.
gen ell_n = floor(n_erid*runiform() + 1)

* 22° Se asignarán ahora los errores idiosincráticos
merge m:1 ell_n using erid
drop _merge

* 23° Se guarda la base de datos que tiene las variables explicativas y los errores asignados.
save censo_errores , replace

* Se generará ahora una muestra para el conjunto de parámetros estimados, para el modelo lineal y el heterocedástico(Estilo Tarozzi)

* CASO DEL MODELO LINEAL

* 24° Se genera una nueva combinación de parámetros estimados, asumiendo un distribución normal para cada estimador.

* Se genera un vector con el número de parámetros estimados
local  reg= rowsof(varianzas_lineal)
gen muestra =0 in 1/`reg'
mkmat muestra in 1/`reg'
drop muestra

* 25° Se genera la muestra
forvalues s=1/`reg'{
matrix muestra[`s',1]=rnormal(beta[`s',1],varianzas_lineal[`s',1]) }
mat colnames muestra = muestra
```

```
* CASO DEL MODELO HETEROCEDÁSTICO

* 26° Se realiza una factorización de cholesky con la matriz de varianzas y covarianzas de los estimadores.

* Se genera un local que capture el número de estimadores
mat cholhet=cholesky(varcovbetahet)
local  reghet= rowsof(betahet)

* 27° Se genera una variable llamada norm, de valores aleatorios procedentes de una distribución normal,

* que tendrá un número de observaciones igual al número de estimadores.
gen normhet = invnorm(uniform()) in 1/`reghet'

* 28° la variable antes especificada será transformada en una matriz
mkmat normhet in 1/`reghet'
drop normhet

* 29° Se tomarán los estimadores y la matriz de cholesky para generar una nueva combinación de parámetros estimados
mat muestrahet = betahet + cholhet*normhet
mat colnames muestrahet = muestrahet

* SE GENERARÁN LOS VALORES AJUSTADOS PARA EL CENSO, EN FUNCIÓN DE LA MUESTRA DE ESTIMADORES DEL MODELO LINEAL.

* 30° Se genera un local llamado contador
local contador=1

* 31° Se toma el global $independientes, con los regresores del modelo lineal, y se cuenta el número de regresores

* (incluso la constante).
foreach var of varlist $independientes {        local contador = `contador' + 1    }

* 32° Se generará una variable que capture el valor ajustado
gen depend_ajus=0

* 33° Se genera un bucle cerrado que genere los valores ajustados y al final se agrega la constante
local i=1

foreach variable of varlist $independientes{replace depend_ajus= depend_ajus + muestra[`i',1]*`variable'
local i=`i' + 1}

replace depend_ajus = depend_ajus + muestra[`contador',1]
```

```
*(Tomando el valor ajustado y su cuadrado como regresores del modelo heterocedástico)
* 34° Se genera una variable (variablehetero) que tomará los valores ajustados para la aproximación del error heterocedástico

gen variablehetero = depend_ajus

gen simulado= depend_ajus + errorclus /*
*/ + sqrt( /*
*/ exp(muestrahet[1,1]*variablehetero+muestrahet[2,1]*variablehetero^2) /*
*/              /(1 + exp(muestrahet[1,1]*variablehetero+muestrahet[2,1]*variablehetero^2)) /*
*/              )*errorides

* 35° Se define que hogares son pobres
generate sufrepobreza=0
replace sufrepobreza=1 if simulado<=lnlinea

* 36° Se define el número de personas pobres
 gen pobpobre=sufrepobreza*pnum

* 37° Se colapsa la base de datos para solo tener el número de personas, y el número de personas pobre, de cada provincia
collapse (rawsum) pobpobre (rawsum) pnum , by (provincia)

* 38° Se genera la variable porcentajepobre, que define el porcentaje de pobreza por provincia
generate porcentajepobre`h'=pobpobre/pnum
drop pobpobre pnum

* 39° Se salva la primera simulación
if `h'==1 {
save "montecarlo.dta" , replace}

* 40° Se salva la simulación
else { save "provicional.dta" , replace

* 41° Se fusiona cada una de las simulaciones de montecarlo
use "montecarlo.dta" , clear
merge 1:1 provincia using "provicional.dta"
drop _merge

* 42° Se salva la nueva base de datos con todas las simulaciones
save "montecarlo.dta" , replace

* 43° Se borra el file innecesario
*erase "provicional.dta" } }
```

```
* 44° Se generan los labels para cada provincia

*PARA AREQUIPA
*label define provincia_nombres 1 Arequipa 2 Camaná 3 Caravelí 4 Castilla 5 Caylloma 6 Condesuyos 7 Islay 8 La_Union
*label values provincia provincia_nombres

*PARA ANCASH
*label define provincia_nombres 1 Huaraz 2 Aija 3 Antonio_Raymondi 4 Asunción 5 Bolognesi 6 Carhuaz 7 Carlos_Fermín 8 Casma /*
* */ 9 Corongo 10 Huari 11 Huarmey 12 Huaylas 13 Mariscal_Luzu 14 Ocros 15 Pallasca 16 Pomabamba 17 Recuay 18 Santa 19 Sihuas 20 Yungay
*label values provincia provincia_nombres

*PARA ICA
label define provincia_nombres 1 Ica 2 Chincha 3 Nazca 4 Palpa 5 Pisco
label values provincia provincia_nombres

*PARA PIURA
*label define provincia_nombres 1 Piura 2 Ayabaca 3 Huancabamba 4 Morropón 5 Paita 6 Sullana 7 Talara 8 Sechura
*label values provincia provincia_nombres

* 45° Se crea una variable string a partir de los label values
decode provincia, generate(_varname)
drop provincia

save “montecarlo.dta”, replace
drop if porcentajepobre1==.

* 46° se transpone las variables con las observaciones
 xpose, clear

* 48° Se obtienen las desviaciones estándar y medias de cada provincia
summarize
browse
```