tobacconomics

Economic Research Informing
Tobacco Control Policy

Conjunto de herramientas actualizado para el

Uso de Encuestas de Gastos de los Hogares para Investigación en Economía del Control del Tabaco

Edición 2023

INSTITUTE FOR
HEALTH RESEARCH
AND POLICY

UIC

Cita sugerida: John R.M., Vulovic V., Chelwa G., Chaloupka F. (2023). Conjunto de herramientas actualizado para el Uso de Encuestas de Gastos de los Hogares para Investigación en Economía del Control del Tabaco. Un Conjunto de herramientas Tobacconomics. Chicago, IL: Tobacconomics, Instituto de Investigación y Políticas de Salud, University of Illinois Chicago.
www.tobacconomics.org

Autores: Este conjunto de herramientas fue escrito por Rijo John, PhD, Profesor Asociado (Adjunto), Facultad de Ciencias Sociales Rajagiri, Kerala, India; Violeta Vulovic, PhD, Economista Senior, Instituto de Investigación y Políticas de Salud, University of Illinois Chicago; Grieve Chelwa, PhD, Director de Investigación, Instituto sobre Raza, Poder y Economía Política, The New School, Ciudad de Nueva York; y Frank Chaloupka, PhD, profesor emérito del Instituto de Investigación y Políticas de Salud de la University of Illinois Chicago. La revisión por pares fue proporcionada por Martín González-Rozada, PhD, Universidad Torcuato Di Tella, Buenos Aires, Argentina; y Guillermo Paraje, PhD, Profesor, Escuela de Negocios, Universidad Adolfo Ibáñez, Santiago, Chile.

Este conjunto de herramientas ha sido financiado por Bloomberg Philanthropies.

Sobre Tobacconomics: Tobacconomics es el resultado de la colaboración de destacados investigadores que desde hace casi treinta años estudian los aspectos económicos de las políticas de lucha contra el tabaco. El equipo se dedica a facilitar a investigadores, defensores y responsables políticos el acceso a los mejores y más recientes trabajos de investigación sobre qué funciona –o no funciona– a la hora de reducir el consumo de tabaco y sus repercusiones en nuestra economía. Como un programa de la University of Illinois at Chicago, Tobacconomics no está vinculado a ningún fabricante de tabaco. Visite **www.tobacconomics.org** o síganos en Twitter **www.twitter.com/tobacconomics**.

Mejorando nuestro conjunto de herramientas: El equipo de Tobacconomics se compromete a hacer que este conjunto de herramientas sea lo más claro y útil posible. Nos gustaría conocer sus comentarios sobre si encontró útil este conjunto de herramientas en su investigación y, de ser así, agradeceríamos conocer su experiencia en cualquier implementación exitosa. También nos gustaría saber si ha encontrado algún problema al aplicar las metodologías presentadas en el conjunto de herramientas y sus opiniones sobre cómo podríamos mejorarlo.

Para cualquier comentario o pregunta sobre el conjunto de herramientas y su contenido, envíenos un correo electrónico a info@tobacconomics.org. Tenemos muchas ganas de escucharlo.

Tabla de contenido

Lista de tablas

1 Introducción

El consumo de tabaco es la causa de muerte prevenible más frecuente y un factor de riesgo principal de varias enfermedades no transmisibles, con más de 7,2 millones de muertos anuales en todo el mundo.[1] A nivel mundial, el 12 por ciento de todas las muertes de adultos (30 años de edad y mayores) se atribuyen al tabaco (16 por ciento entre hombres, 7 por ciento entre mujeres) según la Organización Mundial de la Salud (OMS).[2] Si los patrones actuales de tabaquismo persisten, se espera que el tabaco mate a aproximadamente mil millones de personas en todo el mundo este siglo, principalmente en países de ingresos medianos y bajos (PIMB)[3], donde tanto la prevalencia como el alcance del consumo de tabaco son relativamente altos.[4] El costo económico total del tabaquismo (juntando gastos de salud y pérdidas de productividad) ascendió a US$ 1,4 billones en 2012, o el 1,8% del producto interno bruto (PIB) mundial anual.[5] La carga sanitaria y económica mundial del consumo de tabaco recae cada vez más en los PIMB.

El consumo constante de tabaco en diversas formas obstaculiza el desarrollo y crecimiento económico, especialmente en los PIMB. La morbilidad y la mortalidad resultantes del consumo de tabaco impacta negativamente la productividad, reduce los ingresos disponibles y empuja a las familias a la pobreza. La Agenda 2030 para el Desarrollo Sostenible, adoptada por la Asamblea General de las Naciones Unidas[6] en 2015, reconoce explícitamente la necesidad de fortalecer la implementación del Convenio Marco de la OMS para el Control del Tabaco. La regulación del consumo de tabaco con políticas de salud pública significativas es importante no solo para abordar las crecientes preocupaciones sobre las enfermedades no transmisibles, sino también para mejorar el crecimiento económico y reducir la pobreza. Un sustancial cuerpo de literatura de estudios realizados tanto en países de ingresos altos (PIA) como en PIMB concluye que existen intervenciones de políticas efectivas para reducir la demanda de productos de tabaco y que esas políticas son altamente costo efectivas.[4]

La economía del control del tabaco se ha convertido en una parte integral del discurso del desarrollo, sin embargo, hay escasez de economistas académicos que realicen investigaciones en el área de la economía del control del tabaco, especialmente en los PIMB, donde la necesidad de dicha investigación es relativamente alta. Eso puede deberse a varias razones, incluida la escasez de datos confiables y/o la falta de la experiencia necesaria para llevar a cabo dicha investigación. Aunque la investigación que explora el impacto del control del tabaco en PIMB esté creciendo rápidamente,[4] todavía existe la necesidad de generar más evidencia a nivel local y nacional para respaldar la formulación de políticas de control del tabaco, especialmente en PIMB.

1.1 Propósito de este conjunto de herramientas

El objetivo principal de este conjunto de herramientas es guiar a los investigadores interesados en realizar investigaciones sobre la economía del control del tabaco, especialmente en los PIMB donde existen encuestas de gastos de los hogares (EGH o HES por sus siglas en inglés) sobre el consumo de diferentes productos de tabaco. Es una versión revisada y actualizada de un conjunto de herramientas existente con el mismo nombre publicado por Tobacconomics, con sede en el Institute of Health Research and Policy,

University of Illinois at Chicago.[7] Esta edición del conjunto de herramientas amplía el Capítulo 3 al agregar la estimación del margen extensivo, así como la estimación de las elasticidades por grupos socioeconómicos. Además, los Capítulos 4 y 5 se actualizan con los estudios más recientes. Finalmente, se actualizó la guía técnica en todos los capítulos para incorporar los comentarios y preguntas recibidos de varios usuarios de la primera edición del conjunto de herramientas.

A diferencia de los PIA, los datos de series temporales más largas suelen ser difíciles de obtener en varios PIMB y, como resultado, se vuelve difícil examinar el impacto de ciertas intervenciones de política. Por ejemplo, si se dispusiera de buenos datos de series cronológicas sobre los precios y el consumo de cigarrillos, sería posible estimar cómo las políticas fiscales afectaron los precios y, a su vez, el consumo de cigarrillos. Sin embargo, incluso en ausencia de series cronológicas largas, aún es posible realizar varios análisis relevantes para las políticas sobre los efectos de la formulación de políticas de control del tabaco utilizando datos transversales de encuestas de hogares. La mayoría de los PIMB realizan encuestas de hogares esporádicamente sobre una variedad de temas, y ellas pueden proporcionar información útil sobre el comportamiento del consumidor con respecto al consumo de tabaco.

Este conjunto de herramientas revisa herramientas y técnicas económicas seleccionadas que se pueden usar para analizar datos de EGH con el único propósito de ayudar a la investigación sobre la economía del control del tabaco. El conjunto demuestra el uso de EGH para estimar algunas de las cuestiones importantes en la economía del control del tabaco, incluida la estimación de las elasticidades precio y cruzada, así como la elasticidad del gasto para los productos de tabaco, la elasticidad en los márgenes intensivos y extensivos, el impacto del gasto en tabaco en la asignación de recursos dentro del hogar y el consumo de grupos de productos básicos específicos dentro de un hogar, y el impacto del gasto en tabaco y gastos asociados en atención sanitaria en los recuentos de pobreza nacionales, tanto para la población en general como para las poblaciones en diferentes grupos socioeconómicos. Este conjunto de herramientas analiza brevemente la literatura, los antecedentes teóricos, la justificación económica de cada uno de esos temas, los métodos de estimación y el uso del software estadístico Stata® para implementar esos métodos.

Este es uno de varios conjuntos de herramientas desarrollados por el Banco Mundial, la OMS y Tobacconomics con el objetivo de brindar orientación para realizar un análisis económico de la demanda de tabaco y el impacto del consumo de tabaco en el empleo, la equidad, el comercio ilícito, y los costos económicos. Además, este es el primero de una serie de conjuntos de herramientas de Tobacconomics diseñados para desarrollar capacidades y competencias básicas en el análisis económico de los impuestos al tabaco para apoyar el avance de los argumentos económicos a favor y refutar los argumentos en contra de los aumentos de impuestos al tabaco.

1.2 Quién debería usar este conjunto de herramientas

La discusión en este conjunto de herramientas no presupone conocimiento sobre la tributación del tabaco ni la economía del control del tabaco por parte del lector. Sin embargo, se requiere cierta experiencia en economía y econometría, con una comprensión básica del software econométrico Stata, para hacer un mejor uso de este conjunto de herramientas y realizar estudios independientes en el área de la investigación económica del control del tabaco.

Si bien la discusión de los métodos econométricos y las guías paso a paso con Stata beneficiarían directamente a los investigadores que trabajan en la economía del control del tabaco, con las discusiones de políticas públicas, el razonamiento de los diferentes conceptos económicos del control del tabaco y las interpretaciones de los resultados proporcionados en este conjunto de herramientas también se intenta

beneficiar a los decisores políticos, analistas en agencias gubernamentales y organizaciones de la sociedad civil, para ayudarlos a comprender mejor algunos de los problemas económicos relacionados con el control del tabaco.

1.3 Cómo utilizar este conjunto de herramientas

Este conjunto de herramientas brinda orientación técnica sobre tres temas importantes en el área de la economía del control del tabaco: (i) estimación de las elasticidades precio y precio cruzado (Capítulo 3), (ii) estimación de la naturaleza desplazadora del gasto en tabaco (Capítulo 4), y (iii) cuantificación del efecto empobrecedor del consumo de tabaco (Capítulo 5). Cada tema se discute con la intención de realizar análisis con datos de EGH.

La discusión en cada capítulo comienza con una introducción y con los principios detrás del tema junto con la justificación del análisis. A eso le sigue una breve discusión técnica sobre los métodos econométricos utilizados. La discusión de los métodos econométricos se reduce al mínimo, ya que están disponibles en otros lugares, desde libros de texto de econometría estándar hasta otras fuentes publicadas. Se proporcionan referencias a la lectura necesaria para ayudar a los lectores a obtener conocimientos adicionales sobre los conceptos teóricos presentados.

Una vez que se presentan los métodos, les sigue una breve discusión sobre la preparación de datos para el análisis y luego los diferentes pasos necesarios para realizar el análisis en Stata, junto con el código de Stata necesario. Hacia el final de cada capítulo se presenta un estudio de caso sobre un tema de un país o utilizando un conjunto de datos hipotéticos acompañados con la interpretación de los resultados.

El conjunto de herramientas discute los métodos de análisis relevantes para todos los productos de tabaco combinados, para productos de tabaco con y sin humo por separado, y para productos de tabaco individuales (como cigarrillos, bidi y otros productos de tabaco para mascar) según la cuestión de que se trate. Por ejemplo, al estimar las elasticidades precio y precio cruzada, puede ser útil presentar un análisis para cada uno de los productos de tabaco para facilitar la estimación no solo de la elasticidad precio de diferentes productos de tabaco, sino también la elasticidad precio cruzada, mostrando los patrones de sustitución y complementariedad entre productos de tabaco como bidis y cigarrillos o tabaco con humo y sin humo. Por otro lado, al estimar el impacto del gasto en tabaco en la asignación de recursos dentro del hogar, en lugar de realizar un análisis por diferentes categorías de productos, puede tener más sentido combinar todos los productos de tabaco en una categoría y examinar el impacto en diferentes grupos socioeconómicos.

El conjunto de herramientas está organizado de la siguiente manera. El Capítulo 2 proporciona una introducción a EGH con un enfoque en las encuestas en los PIMB. Discute el contenido de las EGH en lo que respecta al tabaco. En particular, cubre varias preguntas relacionadas con el consumo de tabaco y los gastos en diferentes productos de tabaco de interés en las EGH. El capítulo también analiza brevemente algunos de los problemas econométricos que se deben tener en cuenta al trabajar con las EGH y código de Stata para extraer datos brutos de EGH, entre otros. Además, el capítulo presenta algunos consejos útiles sobre cómo trabajar con el software Stata.

En el Capítulo 3 se analizan los métodos para estimar las elasticidades precio y cruzadas para diferentes productos de tabaco. Presenta métodos para estimar la elasticidad de la prevalencia y de la intensidad de los productos de tabaco. El método principal para estimar la elasticidad de la intensidad (elasticidad en el margen intensivo) es el desarrollado por Deaton,[8] que se presenta junto con una explicación paso a paso de los comandos de Stata para estimar elasticidades precio a partir de datos de EGH. A eso le sigue una

discusión sobre la estimación de la elasticidad de la prevalencia (elasticidad en el margen extensivo) junto con una explicación paso a paso de los comandos de Stata para implementarlos con datos de EGH. En ese capítulo también se puede encontrar una discusión sobre la estimación de las elasticidades precio por grupos de ingreso.

El Capítulo 4 explica los métodos para examinar el impacto del gasto en tabaco en la asignación de recursos dentro del hogar. Siguiendo un enfoque de sistemas de demanda condicional,[9, 10] ese capítulo muestra cómo los gastos en tabaco desplazan sistemáticamente los gastos en otros productos básicos dentro de un hogar. El análisis discute formas de estimar el desplazamiento según diferentes subgrupos socioeconómicos. Se presenta tanto el método analítico como el código de Stata para ejecutar el modelo.

El capítulo 5 cubre el efecto empobrecedor del gasto en tabaco. Se analiza la estimación de la cantidad real gastada en la compra de tabaco, así como el aumento de los gastos de atención sanitaria atribuibles al consumo de tabaco y el humo de segunda mano (HSM). Luego demuestra cómo la contabilización del gasto en tabaco y los gastos en salud asociados impactan la pobreza nacional estimada medida por la tasa de recuento. Se presenta la estimación paso a paso junto con el código Stata pertinente.

En la medida de lo posible, esos capítulos también analizan los resultados empíricos de otros países donde tales estudios se han realizado utilizando EGH.

Los comandos individuales de Stata utilizados en diferentes capítulos se colocan entre corchetes angulares "< >" y en cursiva. Sin embargo, el comando en sí debe usarse sin esos corchetes. Los nombres de variables utilizados en los diferentes ejemplos están todos en cursiva. Los ejemplos específicos que demuestran el uso de algunos códigos de Stata se colocan en cuadros de texto separados en diferentes capítulos. Además, un apéndice de código incluye los códigos de Stata pertinentes para los capítulos respectivos en archivos .do (do-file) separados.

2

Introducción a las encuestas de gastos de los hogares

2.1 Disponibilidad de encuestas de gasto de los hogares

Las encuestas de hogares se han realizado en varios países durante mucho tiempo. La primera encuesta de gastos del consumidor realizada por la Oficina de Estadísticas Laborales (BLS) en los Estados Unidos (EE. UU.), por ejemplo, se realizó en 1888. Aunque es relativamente nueva, la organización National Sample Survey (NSS) en India comenzó sus encuestas de consumo de los hogares ya en la década de 1950[11] y ha realizado encuestas regulares y periódicas cada pocos años desde entonces. El Estudio de Medición de los Niveles de Vida (LSMS, por sus siglas en inglés), el programa insignia del Banco Mundial de encuestas de hogares, existe desde la década de 1980.

Esas encuestas multitemáticas de hogares han recopilado hasta ahora el gasto en consumo de los hogares de más de 40 países de todo el mundo,[12] incluyendo varios países de África y Asia. Hay muchos países, tanto de ingresos altos como de bajos, que realizan encuestas de gastos de los hogares, y muchos de ellos realizan esas encuestas a intervalos regulares.

La Red Internacional de Encuestas de Hogares (IHSN, por sus siglas en inglés), una red informal de agencias internacionales que se esfuerza por "mejorar la disponibilidad, accesibilidad y calidad de los datos de las encuestas en los países en desarrollo, y fomentar el análisis y el uso de esos datos por los decisores de desarrollo nacionales e internacionales, la comunidad investigadora y otras partes interesadas",[13] mantiene un portal para que los investigadores exploren y descarguen documentos y metadatos de censos o encuestas de hasta 201 países; actualmente cuenta con cerca de 7 mil encuestas catalogadas. Alrededor de 137 de los 201 países para los que hay datos disponibles son PIMB. Se puede acceder a ese catálogo en http://catalog.ihsn.org/index.php/catalog e incluye información sobre más de 1000 EGH en su base de datos, de las cuales alrededor de 700 son de PIMB.

En ausencia de variables macroeconómicas de series de tiempo largas, las EGH proporcionan datos transversales significativos, a veces para múltiples períodos de tiempo para el mismo país. Sin embargo, los organismos estadísticos que realizan EGH en la mayoría de los países suelen publicar informes resumidos que presentan solo datos agregados que se difunden libremente al público. Los datos agregados, si bien son útiles para examinar el panorama general, no brindan un tamaño de muestra adecuado para realizar los principales análisis econométricos discutidos en este conjunto de herramientas. Por lo tanto, para realizar análisis econométricos avanzados con los datos de la encuesta, es importante tener acceso a los microdatos (registros de individuos, hogares o unidades) de las encuestas.

Esos microdatos a menudo no están disponibles libremente para el acceso público. Sin embargo, dichos datos generalmente están disponibles directamente de las agencias gubernamentales de estadística a cargo de realizar las encuestas, mediante el pago de una tarifa nominal. Después de pagar la tarifa por el sitio web de la agencia, los datos pueden recibirse en formato digital ya sea descargándolos directamente del sitio

web de la agencia o por correo en un dispositivo de almacenamiento de datos. Algunas agencias permiten la descarga de datos después del registro y una breve descripción del proyecto. Los microdatos de LSMS de diferentes países, por ejemplo, están disponibles gratuitamente para descargar desde el sitio web del Banco Mundial después de registrarse y proporcionar un breve resumen sobre el proyecto.

2.2 Contenido de las encuestas de gasto de los hogares

Las encuestas de hogares más simples recopilan datos sobre una muestra nacional de hogares seleccionados al azar de un "marco" o lista nacional de hogares (a menudo un censo), y se asigna una probabilidad igual a cada hogar seleccionado del marco. Aunque los tamaños de las muestras varían ampliamente dependiendo del propósito de la encuesta, del tamaño de la población del país y de la necesidad de generar estimaciones de submuestras, frecuentemente se encuentran tamaños de muestra de alrededor de 10.000, lo que corresponde a una fracción de muestreo de 1:5000 en una población de cinco millones de hogares.[8]

En la práctica, a menudo se implementa un diseño de dos etapas en la selección de hogares, donde, en la primera etapa, la selección se hace de una lista de "conglomerados" de hogares, generalmente aldeas en áreas rurales o bloques urbanos en centros urbanos, y en la segunda etapa, se seleccionan los hogares de cada conglomerado.[8] Los conglomerados generalmente se denominan unidades primarias de muestreo (UPM), ya que son la primera unidad muestreada en el diseño. Si los conglomerados se seleccionan aleatoriamente con probabilidad proporcional al número de hogares que contienen, y si se selecciona el mismo número de hogares de cada conglomerado, sería como si cada hogar tuviera la misma probabilidad de ser incluido.

Dependiendo de los objetivos de la encuesta, se puede diseñar una muestra para que los hogares puedan ser seleccionados en función de atributos relevantes como el área geográfica, el origen étnico, el nivel de vida, el género o la raza para que los hogares en un determinado grupo puedan tener una cierta probabilidad de ser seleccionados. Tal estratificación convierte efectivamente una muestra de una población en una muestra de muchas poblaciones, garantizando así suficientes observaciones para permitir estimaciones de esos subgrupos.[8]

Las ponderaciones de probabilidad para los hogares en cada estrato pueden diferir. En la mayoría de los casos, puede haber pocas UPM o conglomerados dentro de cada estrato. La NSS en India, por ejemplo, se enfoca en la estratificación por áreas rurales y urbanas dentro de un distrito para sus encuestas de gasto del consumidor. Si bien la estratificación generalmente mejora la precisión de las estimaciones del muestreo, la agrupación de la muestra generalmente reducirá la precisión, ya que los hogares dentro del mismo conglomerado son más similares entre sí y, por lo tanto, reflejan una baja variabilidad.

Las encuestas de hogares, por su propia naturaleza, brindan información sobre los hogares y las personas que los integran. Aunque la definición de hogar utilizada en cada encuesta puede diferir según la estructura de arreglos de vivienda en cada país, en general, aquellos miembros que viven juntos y comen juntos se consideran parte del mismo hogar. Las EGH generalmente brindan datos sobre el consumo, los ingresos o los activos, y las características demográficas de los hogares, incluida la composición del hogar, el tamaño del hogar, la edad y el género de los miembros del hogar, el nivel educativo y la situación laboral de sus miembros, el origen étnico y la raza, entre otros.

Para evaluar el consumo, las EGH miden los gastos incurridos y/o la cantidad consumida por los hogares en diferentes bienes y servicios durante un período de reporte preespecificado, también conocido como período de recuerdo o de referencia. Aunque es poco frecuente, algunas EGH —por ejemplo, la Encuesta de Gastos del

Consumidor (CES por sus siglas en inglés) del BLS en los EE. UU.— también recopilan datos de gastos a nivel individual. En el caso de productos para adultos, como el tabaco, esos datos serían inmensamente útiles.

Según el objetivo de la encuesta y las características de los productos o servicios en cuestión, el período de recuerdo puede variar significativamente para diferentes productos dentro de la misma encuesta y para los mismos productos en diferentes encuestas; puede variar desde tan solo un día hasta un período de un año. Sin embargo, los artículos de consumo comunes en la mayoría de las EGH tienen un período de recuerdo de una semana a un mes. La Encuesta de Ingresos y Gastos de los Hogares (HIES por sus siglas en inglés, 2016) en Liberia, por ejemplo, recopila datos de consumo de alimentos con un recuerdo de siete días y consumo no alimentario dentro de recuerdos de siete y 30 días.[14]

Como parte de la tarea de recopilar datos sobre los gastos incurridos y la cantidad consumida de diferentes bienes, varias EGH recopilan información sobre el consumo de diferentes productos de tabaco de uso común en los respectivos países. El NSS de India, por ejemplo, recopila tanto la cantidad de consumo como los gastos en cigarrillos, bidi y variedades de tabaco sin humo durante los 30 días y los siete días anteriores a la entrevista. Eso proporciona una rica fuente de información para ayudar a examinar varios aspectos económicos relacionados con el consumo de tabaco. Ese nivel de desagregación, sin embargo, puede no estar disponible en todas las EGH.

Dependiendo de los recursos disponibles para las agencias de encuestas, a veces se informan los gastos de productos agregados a grupos más grandes, como el tabaco y los estupefacientes como un solo grupo. Algunas EGH, por otro lado, solo brindan información sobre gastos y no recopilan información sobre cantidades para varios artículos de consumo. Como resultado, puede haber desafíos en el análisis econométrico entre diferentes conjuntos de datos.

Usando otras características específicas de los hogares y la información regional proporcionada en las encuestas de hogares, a menudo es posible clasificar los hogares de una encuesta en diferentes grupos de nivel socioeconómico (NSE) para que el análisis económico se pueda realizar por grupo de NSE. Dicho análisis se puede realizar en función del nivel educativo de los hogares, la situación de los ingresos o los activos, el lugar de residencia, como áreas rurales o urbanas, el origen étnico o el nivel de vida de un hogar, entre otros criterios.

2.3 Problemas econométricos al trabajar con encuestas de hogares

Debido a las características de diseño de las encuestas de hogares discutidas en la sección anterior, existen desafíos específicos para el análisis econométrico. En el Capítulo 2 de *El Análisis de Encuestas de Hogares,* de Deaton, se ofrece una exposición detallada de esos desafíos.[8] A continuación se presenta un breve resumen conceptual de los asuntos más importantes.

2.3.1 Uso de ponderadores de las encuestas para estadísticas descriptivas

Dependiendo del propósito de cada encuesta de hogares, algunos hogares pueden estar sobrerrepresentados o subrepresentados en las encuestas y, como resultado, la media muestral estimada u otras estadísticas muestrales serán estimadores sesgados de sus contrapartes poblacionales. Los ponderadores de la encuesta a menudo se utilizan para volver a ponderar los datos de la muestra y ajustar los elementos de diseño de la encuesta para que las estimaciones sean representativas de la población. La mayoría de las encuestas incluyen los ponderadores de la encuesta junto con los datos publicados y se pueden usar de inmediato, tal cual, mientras se generan las estadísticas necesarias.

Si los ponderadores no se incluyen directamente, la documentación de la encuesta generalmente incluirá instrucciones o fórmulas para calcular esos ponderadores usando variables relevantes incluidas en los datos de la muestra. Es importante aplicar los ponderadores correctos de la encuesta al generar estadísticas descriptivas de los datos de la muestra. La Sección 2 a continuación brinda ejemplos de cómo aplicar ponderadores de encuestas en Stata mientras se calculan ciertas estadísticas descriptivas.

2.3.2 Uso de ponderadores de las encuestas en regresión

A diferencia de las estadísticas descriptivas, no hay acuerdo sobre el uso de ponderadores de encuestas en el contexto de las regresiones. El argumento de la econometría clásica está en contra del uso de ponderadores en la regresión. Según Deaton[8] señala, cuando la población es homogénea de modo que los coeficientes de regresión son idénticos en cada estrato, tanto los estimadores ponderados como los no ponderados serán consistentes, y los mínimos cuadrados ordinarios (MCO) son de hecho más eficientes por medio del teorema de Gauss-Markov.[15] Por otro lado, cuando la población no es homogénea, tanto los estimadores ponderados como los no ponderados son inconsistentes de todos modos y la ponderación no agrega valor.

Sin embargo, Deaton[8] continúa diciendo que una regresión ponderada proporciona una estimación consistente de la función de regresión poblacional, siempre que el supuesto sobre la forma funcional de la regresión sea correcto, es decir, cuando la función de regresión misma es el objeto de interés. Si el interés es estimar modelos de comportamiento donde el comportamiento puede ser diferente para diferentes subgrupos, la ponderación en la regresión no sirve de nada. En conclusión, como observan Cameron y Trivedi,[16] se deben usar ponderadores para la estimación de las medias poblacionales y para predicciones y cálculo de los efectos marginales posteriores a la regresión. No obstante, en la mayoría de los casos, la regresión en sí puede ajustarse sin ponderadores, como es la norma en microeconometría.

2.3.3 Errores estándar inflados debido a los efectos del diseño del conglomerado

Como la mayoría de las encuestas de hogares utiliza un diseño de dos etapas en el que primero se eligen los conglomerados, seguidos de los hogares dentro de cada uno de esos conglomerados, a menudo ocurre que los hogares dentro del mismo conglomerado son bastante similares entre sí —ya que viven cerca uno de otro y son entrevistados más o menos al mismo tiempo— y son diferentes de los de otros grupos que generalmente están muy separados geográficamente. En otras palabras, habrá más homogeneidad dentro de los conglomerados que entre ellos.

En la medida en que las observaciones o los hogares dentro de un conglomerado no sean completamente independientes, las correlaciones positivas entre esas observaciones podrían potencialmente inflar la varianza por encima de lo que sería si fueran independientes. Por lo tanto, es importante corregir los errores estándar estimados en las regresiones basadas en encuestas de hogares para tener en cuenta esos efectos del diseño de conglomerados utilizando técnicas apropiadas.

2.3.4 Heterocedasticidad de los residuos de los MCO

Las distribuciones de los hogares sobre diferentes variables de interés, como el ingreso y el consumo de diferentes bienes, por lo general no se distribuyen normalmente. Como resultado, es bastante común encontrar perturbaciones heterocedásticas en las funciones de regresión estimadas a partir de datos de EGH.

La heterogeneidad entre diferentes grupos también podría resultar en funciones de regresión que arrojan términos de error heterocedásticos. Dichos errores dejarían ineficientes las estimaciones de MCO e invalidarían las fórmulas habituales para los errores estándar. Por lo tanto, tendrían que ser corregidos usando métodos de corrección apropiados.

Combinado con la presencia de efectos de diseño de conglomerados, es importante utilizar fórmulas que corrijan los errores estándar en regresiones basadas en encuestas que tengan en cuenta la presencia de heterocedasticidad y efectos de conglomerados.

2.3.5 Endogeneidad

Eso se refiere a situaciones en una regresión cuando una o más de las variables explicativas se correlacionan con el término de error, lo que da como resultado estimaciones de MCO sesgadas e inconsistentes. La endogeneidad surge principalmente por tres razones:

(i) **Simultaneidad:** X causa Y e Y también causa X. En otras palabras, X e Y se determinan conjuntamente.

(ii) **Variables explicativas omitidas:** cuando una variable omitida afecta a una o más de las variables independientes incluidas y afecta por separado a la variable dependiente. La información omitida contenida en esas variables omitidas también puede denominarse "heterogeneidad no observada" o la variación no observada entre unidades individuales de esa variable omitida o no observable.

(iii) **Errores de medición:** una o más de las variables explicativas se miden incorrectamente. El error de medición en una variable dependiente no sesga el coeficiente de regresión. Los errores de medición en los datos de la encuesta, según Deaton,[8] son un hecho de la vida.

Aunque a menudo se mencionan como fuentes separadas de endogeneidad en la regresión, en realidad no es necesario que sean realmente distintas entre sí. A menudo, en el análisis de regresión que utiliza datos de encuestas, se encuentran la mayoría, si no todas, esas diferentes fuentes de endogeneidad.

En todas las diferentes fuentes de endogeneidad descritas aquí, la función de regresión diferiría del modelo estructural debido a la correlación entre el término de error y las variables explicativas, violando así un supuesto crucial de MCO. El uso de variables instrumentales (VI) (como el método de mínimos cuadrados en dos etapas)[15] es la técnica estándar en tales circunstancias, siempre que sea posible encontrar VI que estén correlacionadas con las variables explicativas pero no correlacionadas con los términos de error para que la regresión arroje estimaciones consistentes.

2.4 Consejos útiles sobre Stata

Stata, un paquete estadístico ampliamente utilizado, es un software de análisis de datos y econométrico preferido por muchas universidades e instituciones de todo el mundo, lo que facilita los intercambios y colaboraciones entre investigadores en múltiples disciplinas e instituciones.[17] A continuación se presentan algunos consejos útiles que hacen que trabajar con Stata sea mucho más fácil.

2.4.1 Creación de un do-file

Stata se puede utilizar por medio de sus menús desplegables desde la interfaz de usuario, emitiendo comandos directamente en una ventana dedicada a comandos o con la ayuda de un do-file, que guarda todos los comandos para ejecutarlos a voluntad. La ejecución por do-file es el método preferido y recomendado, ya que ofrece varias ventajas sobre los otros métodos. Un do-file simplemente registra todos los comandos que se ejecutarán y los guarda en un archivo para uso futuro con la extensión ".do".

La principal ventaja es que el análisis se puede replicar con los comandos guardados en el do-file y el trabajo puede ser compartido y editado por otros colaboradores. Pero, más que nada, un do-file mantiene un registro del trabajo realizado y permite la revisión de los comandos según sea necesario. A diferencia de las

ventanas de comandos o los menús desplegables, en un do-file también se pueden agregar notas y comentarios para otros colaboradores, lo que facilita una colaboración fluida. Puede encontrar información útil sobre cómo crear un do-file en el sitio web de Stata (https://www.stata.com/manuals13/u16.pdf).

2.4.2 Creación de un archivo log

Mientras que un do-file mantiene un registro de todos los comandos y permite editarlos según sea necesario, un archivo de log con la extensión ".log" o ".txt" mantiene un registro de los comandos ejecutados junto con sus resultados durante una sesión de Stata determinada. Es útil crear archivos de log mientras se ejecuta el do-file para que los resultados estén disponibles para futuras referencias o para compartir con colaboradores.

Se crea un archivo de log dentro del do-file mediante un comando *<log using mylog.log, replace>*. Eso creará un archivo con el nombre "mylog.log" en el directorio de trabajo actual de Stata. El argumento opcional *<replace>* se asegurará de que cada vez que se ejecute el do-file, el contenido del archivo de log se reemplace con los nuevos resultados. También se puede usar la opción *<append>* para seguir agregando los resultados de todos los comandos al mismo archivo de log. Antes de cerrar la sección, generalmente hacia el final del do-file, cierre el archivo de log con el comando *<log close>*. El uso del archivo de log también se puede suspender temporalmente y reanudar mediante comandos como *<log off>* y *<log on>*.

2.4.3 Uso de recursos de conocimiento

Todos los manuales de usuario de Stata están integrados en el software. Uno puede simplemente emitir el comando *<help>* seguido del comando particular de Stata para aprender la descripción, sintaxis y ejemplos de cada comando usado en Stata. Por ejemplo, *<help regress>* devolverá la sintaxis, la descripción y los ejemplos necesarios del uso del comando regress. Además, los comandos *<search>* y *<findit>* devuelven información muy útil sobre temas de interés dentro de Stata. Por ejemplo, el comando *<search survey>* devolvería una lista de comandos y módulos que utiliza Stata para analizar datos de encuestas. Stata también tiene un excelente foro de soporte, que es un rico recurso para aprender y familiarizarse con Stata (https://www.statalist.org/forums/).

2.4.4 Configuración de un directorio de trabajo

Mientras trabaja con los datos de la encuesta de hogares, es mejor hacer una copia de los microdatos y moverlos a un directorio dedicado a la encuesta en la computadora. Todos los archivos de programa posteriores de Stata y otros documentos relacionados para el análisis se pueden almacenar en el mismo directorio sin tocar los microdatos originales. El comando *<pwd>* reporta el directorio de trabajo actual de Stata independientemente del sistema operativo. Ese directorio de trabajo se puede cambiar con un comando *<cd "Path">* donde Path, entre comillas dobles, es la ruta del directorio donde se guarda el trabajo; que diferiría dependiendo del sistema operativo.

Una vez que se establece un directorio de trabajo, los comandos subsiguientes que llaman a los archivos (como archivos de datos, do-files, archivos de diccionario, etc.) se pueden ejecutar usando solo el nombre del archivo sin la ruta completa del directorio. Eso también tiene la ventaja de que un colaborador solo necesita cambiar el directorio de trabajo una vez y no necesita cambiar las rutas de archivo mencionadas en diferentes partes del do-file mientras ejecuta un do-file.

Alternativamente, se puede configurar una macro global para asignar un directorio para almacenar los datos y guardar el trabajo. A partir de entonces, simplemente llame al nombre de la macro en lugar de repetir toda la estructura del directorio para usar los datos o guardar algo. Por ejemplo, en Windows, use el comando *<global pathin "C:\Data\HES">*. Más tarde, para importar datos almacenados en ese directorio desde dentro

del do-file, use el comando *<use $pathin\filename.dta>* y Stata buscará automáticamente el archivo de datos en el directorio definido en el pathin macro global. La estructura de la ruta del directorio varía según el sistema operativo. El uso de macros se analiza con mayor detalle más adelante.

2.4.5 Practicar con conjuntos de datos de ejemplo

Stata proporciona dos tipos de conjuntos de datos con fines de demostración y práctica. Estos son: (a) conjuntos de datos de ejemplo instalados con Stata en un equipo local y (b) conjuntos de datos en línea a los que se hace referencia en la documentación de Stata y que son accesibles en línea. Desde la interfaz de usuario de Stata, vaya a "File > Example data sets" y aparecerán listas de datos disponibles. Haga clic en esos conjuntos de datos y ábralos en Stata para practicar.

Alternativamente, use el comando *<sysuse datafile>* donde datafile se refiere al nombre de archivo del conjunto de datos particular en el sistema, si se conocen los nombres de los conjuntos de datos. También se puede usar el comando *<webuse datafile>* para cargar un conjunto de datos específico y obtenerlo en la web. Los conjuntos de datos se obtienen de http://www.stata-press.com/data/r17/. Ese enlace también proporciona una lista detallada de conjuntos de datos organizados por tema, y uno puede navegar por conjuntos de datos disponibles para usar de práctica.

2.4.6 Uso de operadores lógicos y relacionales

Stata utiliza varios operadores lógicos y relacionales para ayudar con el trabajo de conjuntos de datos. Aquí se dan algunos de los operadores de uso común y sus significados.

&	*Y*	*|*	*O*
!	*No*	*~*	*No*
>	*Mayor que*	*<*	*Menor que*
>=	*Mayor o igual*	*<=*	*Menor o igual*
==	*Igual*	*!=*	*No igual*

Aparte de ellos, Stata también tiene operadores para manejar variables categóricas (también conocidas como variables ficticias o dummy; en presencia dos categorías son llamadas también dicotómicas o binarias). Anteponga (i.) a una variable para especificar indicadores para cada categoría de una variable. Eso funciona bien en lugar de crear variables ficticias separadas. El comando *<fvset base>* se puede utilizar para establecer la categoría base. Ingrese (#) entre dos variables ficticias para crear una variable de interacción. Introduzca (##) para especificar tanto los efectos principales de cada variable como sus interacciones. De manera similar, (c.) se puede usar para interactuar una variable continua con una variable categórica anteponiendo la variable continua con (c.).

Por ejemplo, suponga que edad (*age*) y sexo (*sex*) sean variables categóricas y el índice de masa corporal (*bmi*) sea una variable continua. Para realizar una regresión de los efectos de esas variables sobre la presión arterial (*bp*), las siguientes regresiones producen el mismo resultado: *<regres bp i.age age#sex>* y *<regres bp age##sex>*. Alternativamente, para hacer una regresión de la presión arterial sobre la edad y el índice de masa corporal y la interacción entre ellos, escriba *<regres bp age##c.bmi>*.

2.4.7 Uso de macros

Las macros son abreviaturas o "alias", que tienen un nombre y un valor. Cuando se presenta su nombre puntuado en el comando, devuelve su valor.[18] Por lo tanto, una macro tiene un nombre de macro y un contenido de macro. Dondequiera que se use el nombre de la macro con puntuación en el programa, el contenido de la macro se sustituye en su lugar. Las macros se utilizan para varios propósitos, como simplificar las tareas, hacer que los do-file estén más organizados, acortar la longitud del código de Stata y varias otras comodidades durante la programación. Una macro puede ser de dos tipos, local o global, según su alcance, es decir, dónde se reconoce su existencia. Las macros globales, una vez definidas, están disponibles en cualquier parte de Stata, mientras que las macros locales existen únicamente dentro del programa o do-file en el que están definidas.[19]

Para sustituir el contenido de la macro de un nombre de macro global, el nombre de la macro se puntúa con un signo de dólar ($) al frente. De manera similar, para sustituir el contenido de macro de un nombre de macro local, el nombre de macro se puntúa con comillas simples izquierda y derecha (').[19] Por ejemplo, defina una macro local con el nombre *indvar* como <*local indvar price expenditure hsize*> y emita otro comando <*summarize 'indvar'*> para obtener las estadísticas resumidas para cada una de las variables precio (*price*), gasto (*expenditure*) y *hsize* en los resultados.

De manera similar, defina una macro global como <*global xyz age income sex*> y emita el comando <*summarize $xyz*> para obtener el resumen de cada una de esas variables: edad (age), ingresos (income) y sexo (sex). Dado que las macros globales pueden crear conflictos entre los do-file, raramente se utilizan. Por lo general, se prefieren las macros locales al escribir el código en el do-file.

Las macros también se pueden definir como una expresión, y el resultado se convierte en el contenido de la macro. Por ejemplo, defina <*local result = 5+5*> y el comando <*display 'result'*> devolvería 10. Las macros también pueden ofrecer funcionalidades extendidas con funciones extendidas de macro. Use el comando <*help macro*> para obtener más información sobre las macros y sus usos variados y creativos.

2.4.8 Uso de comandos de loop

Los loops son comandos en Stata que ayudan a recorrer una lista arbitraria de cadenas o números. Por ejemplo, un comando de loop puede establecer repetidamente un nombre de macro local para cada elemento de la lista y ejecutar los comandos encerrados entre llaves "{}". Los loops son bastante útiles y convenientes al realizar tareas repetitivas que se realizan de forma secuencial, y se usan mucho durante la programación. Los comandos <*foreach*> y <*forvalues*> de Stata son especialmente útiles para los loops. Esos comandos de loop comienzan y terminan con llaves "{" y "}" en líneas separadas. La llave abierta debe aparecer en la misma línea que <*foreach*> y la llave cerrada debe aparecer sola en una línea al final. Por ejemplo,

```
foreach X in var1 var2 var3 {
replace `X'=. if `X'<=0
generate ln`X'=log(`X')
}
```

La primera línea anterior enumera las diferentes variables sobre las que se debe repetir el comando (var1, var2 y var3) y las siguientes dos líneas dan los comandos reales que se deben repetir. El primer comando le dice a Stata: si una observación para una variable en la lista tiene un valor menor o igual a cero, debe reemplazarse con un punto. El segundo le indica a Stata que genere nuevas variables con un nombre de variable que comience con *ln* seguido de los nombres de las variables en la lista, definido como el logaritmo natural de las variables existentes en la lista.

Se pueden agregar varias líneas de comandos una debajo de la otra, todas las cuales se repetirán sobre todas las variables mencionadas en la primera línea. El código anterior también se puede ejecutar de manera más eficiente utilizando macros locales. Por ejemplo, predefina una macro local *<local varlist var1 var2 var3>* y use el loop:

```
foreach X of local varlist {
replace `X'=. if `X'<=0
generate ln`X'=log(`X')
}
```

Stata también puede ejecutar dichos comandos de loop en diferentes archivos a la vez. De manera similar, el comando *<forvalues>* se puede usar para realizar operaciones similares aplicadas a números. Por ejemplo, suponga que hay 25 estados en una encuesta de hogares, y los gastos de consumo promedio en cada estado están bajo los nombres de variables *" state1, state2, ..., state25 "*. Para convertir todas esas variables a forma logarítmica, use el comando:

```
forvalues i=1/25 {
generate lnstate`i'=ln(state`i')
}
```

La *i* en la primera línea del comando *<forvalues>* se refiere a la macro local dentro del loop.

2.4.9 Devolver resultados almacenados

Stata almacena regularmente los resultados de los comandos en macros locales que se pueden llamar para varios propósitos. Por ejemplo, emitir un comando *<summarize>* para una variable *<sum varname>* devolverá estadísticas descriptivas sobre la variable *<varname>*. Simultáneamente, también almacena esos resultados en macros locales. Por ejemplo, *<summarize mpg>* en los datos de automóviles de Stata arroja los resultados a continuación.

Variable	Obs	Mean	Std. Dev.	Min	Max
mpg	74	21.2973	5.7855	12	41

Ejecute el comando *<return list>* después de esto y obtendrá los resultados que se muestran en la tabla.

```
R(N)      =     74
r(sum_w)  =     74
r(mean)   =     21.2973
r(Var)    =     33.47205
r(sd)     =     5.785503
r(min)    =     12
r(max)    =     41
r(sum)    =     1576
```

Todos los resultados se almacenan en diferentes macros locales. Ellos están disponibles para generar nuevas variables o para usarse en otros comandos inmediatamente después. Similar a *<return list>*, use el comando *<ereturn list>* para mostrar los contenidos almacenados localmente después de los comandos de estimación como *<regress>*. El comando *<help return>* en Stata mostrará otros usos de los comandos *<return>*. A continuación el cuadro 2.1 proporciona un ejemplo práctico utilizando algunos de los consejos de Stata ya cubiertos.

Cuadro 2.1 Consejo de ejemplo de Stata

```
sysuse auto

local items price mpg weight
foreach X of local items {
      quietly sum `X', detail
      local upper = r(mean) + 3 * r(sd)
      replace `X' = r(p50) if `X' > `upper' & `X' <.
}
```

El código demuestra el uso de macros, loops y resultados almacenados, todo en un solo lugar. La primera línea importa los datos "auto" incorporados y la segunda define una macro local llamada items, que consta de tres variables. El tercero abre un comando de loop *<foreach>* y usa la macro local junto con él. Hay tres instrucciones que se ejecutan sucesivamente en las tres variables en las siguientes tres líneas por medio de este loop.

- El primero resume silenciosamente la variable, y con la adición del prefijo "quietly" ejecuta este comando sin mostrar los resultados. La opción *<detail>* después de *<summarize>* exige estadísticas adicionales que normalmente no se calculan, como percentiles, asimetría y curtosis.
- La segunda línea del loop define una nueva macro local *upper*, utilizando los resultados almacenados después de *<summarize>*. Se define como el promedio más tres desviaciones estándar de la variable considerada.
- La tercera línea del loop reemplaza cualquier valor mayor que el promedio más tres desviaciones estándar y menor que los valores faltantes (Stata considera que los valores faltantes son mayores que cualquier valor numérico) con la mediana de esa variable. La llave en la última línea termina el loop.

2.4.10 Uso de delimitadores

El comando *<#delimit ;>* se usa para restablecer el carácter que marca el final de un comando en Stata. Esos se usan solo en do-files y ado-files (definidos en la siguiente sección). Presionar la tecla de retorno le indica a Stata que ejecute el comando. En un do-file, el final de una línea asume la tecla de retorno y esas líneas tienen restricciones de caracteres. Entonces, uno puede indicar a Stata que los comandos son más largos que una línea usando el comando *<#delimit ;>* para dividir libremente las líneas de comando según sea necesario. Stata considerará todas las líneas continuas hasta que vea el carácter delimitador que marca el final del comando como una sola línea lógica.

Alternativamente, se puede usar < /* */ > como delimitador de comentario. Por ejemplo, *<generate X = 3*Y /* esto es un comentario*/ + 5>* es lo mismo que *<gen X = 3*Y + 5>* sin el comentario. También se pueden dividir líneas largas con tres barras diagonales consecutivas (///), en lugar de usar el comando *<#delimit ;>*. Esos son bastante útiles al preparar do-files. Por ejemplo, Stata considera el siguiente comando como una sola línea lógica:

```
regress lnwage educ complete age c.age#c.age    ///
        exp c.exp#c.exp tenure c.tenure#c.tenure  ///
        i.region female
```

2.4.11 Uso de comandos add-on

Stata permite a las personas escribir comandos de terceros (llamados "ado-files") que se pueden almacenar en un archivo de Componentes de Software Estadístico (SSC por sus siglas en inglés), que a menudo se denomina Boston College Archive y es proporcionado por **http://repec.org**. Desde el archivo SSC, los usuarios pueden instalar esos programas add-on usando el comando *<ssc install progname>* donde "progname" es el nombre del ado-file o archivo de programa que debe instalarse. También se puede desinstalar un paquete en particular con el comando *<ssc uninstall progname>*.

La mayoría de los paquetes add-on brindan funcionalidad adicional en comparación con los comandos integrados de Stata. Por ejemplo, el paquete add-on *<estout>*, que se puede instalar con *<ssc install estout>*, ayuda a crear tablas ordenadas a partir de las estimaciones almacenadas después de los comandos de regresión. Puede crear tablas dignas de publicación con coeficientes de regresión, agregando estrellas para indicar su nivel de significancia, estadísticas resumidas, errores estándar, estadístico *t*, valores *p* e intervalos de confianza para uno o más modelos ajustados anteriormente y almacenados por el comando *<estimates store>*.

Del mismo modo, *<findname>*, *<outreg2>* y *<ivreg2>* son algunos de los add-ons populares. Esos se pueden instalar en Stata usando el comando *< ssc install outreg2 >*, por ejemplo. Outreg2 ayuda a producir tablas listas para publicación a partir de resultados de regresión. Use el comando *<ssc whatshot>* para ver algunos de los paquetes add-ons más populares disponibles para descargar.

2.4.12 Consejos varios

Aquí se incluyen algunos consejos varios no mencionados anteriormente:

- Los comandos de Stata y los nombres de variables distinguen entre mayúsculas y minúsculas. Por ejemplo, si se usa una letra minúscula en lugar de mayúscula, devolverá un error o ejecutará un código no deseado.
- La mayoría de los comandos de Stata se pueden abreviar. Por ejemplo, *<summarize>* se puede abreviar como *<sum>* o *<su>*. En lugar de *<regress>* use *<reg>*, y así sucesivamente.
- El nombre dado a los escalares dentro del do-file debe ser distinto de cualquiera de las otras variables o sus abreviaturas inequívocas presentes en los datos. Si un escalar se define con el mismo nombre que otra variable o su abreviatura inequívoca, Stata priorizará el nombre de la variable o su abreviatura sobre el nombre escalar especificado, lo que generará resultados involuntarios al realizar operaciones que involucran este escalar. Alternativamente, use una pseudo función *<scalar(xyz)>* para deletrear un escalar con el nombre "xyz" cada vez que el escalar se vaya a usar en un cálculo o al definir más escalares.
- Los valores faltantes, indicados por un punto (.) se codifican y tratan como infinito positivo en Stata, lo que significa que toman un valor más alto que todos los demás valores numéricos. Eso es importante al

limpiar los datos. Por ejemplo, *<replace X = 0 if Y>100>* reemplazará X con cero no solo si es mayor que 100, sino también si faltan valores en Y. En su lugar, use *<replace X = 0 if Y>100 & Y<.>*

2.5 Técnicas de extracción de datos usando Stata

Los microdatos de las encuestas de hogares se almacenan en diferentes formatos de archivo según el hardware utilizado para registrar los datos, la disponibilidad de software dentro de las agencias de las encuestas y otras prácticas y costumbres estándar en diferentes campos. Los datos de EGH que son de interés en ese trabajo generalmente serán datos tabulares cuantitativos. A menudo se presentan en archivos de texto delimitados que contienen metainformación, como los que se encuentran en el software estadístico Stata, SPSS y SAS, o en archivos de valores separados por comas simples (.csv), archivos delimitados por tabulaciones (.tab) o en formato fijo ASCII con extensiones de archivo ".ascii", ".dat" o ".txt".

Si los datos están en formato fijo ASCII, que suele ser el caso, habrá un diccionario asociado o un archivo de diseño que describe cada columna de longitudes de registro fijas en el archivo de datos. Por ejemplo, el diccionario diría: la posición de byte 4 en el archivo de datos indica el código para el área rural o urbana, las posiciones de byte 9–10 indican el código para UPM o identificador de grupo, o las posiciones de byte 30–36 indican los gastos en un artículo.

También habrá un archivo, generalmente llamado libro de códigos, que indica el significado de los diferentes códigos utilizados en el archivo de diseño o el archivo de datos. Por ejemplo, indicaría que el valor 1 = rural y 2 = urbano, o 1 = masculino y 2 = femenino.

Los datos finales que archivan las respectivas agencias de encuestas suelen proporcionar toda la documentación necesaria asociada con los datos. El catálogo IHSN[13], por ejemplo, incluye detalles sobre la metodología de la encuesta, los procedimientos de muestreo, los cuestionarios, las instrucciones, los informes de la encuesta, el código utilizado y los libros de códigos de archivos de diseño o diccionario para la mayoría de los datos de la encuesta catalogados allí.

El software que se utiliza para el análisis estadístico debe poder importar microdatos antes de que se puedan realizar diferentes análisis. Se necesita una descripción y documentación detalladas de los datos de la encuesta, la estructura de los archivos de datos y la relación entre los diferentes archivos de datos de la encuesta para tomar una decisión informada sobre qué datos deben extraerse o importarse al software estadístico para un análisis posterior. Para generar cualquier estimación a partir de esos datos, se debe extraer la parte relevante de los datos y agregarla usando los comandos apropiados en el software analítico. Stata utiliza diferentes métodos para importar datos según el tipo de archivo de datos de origen. Ingresar el comando *<help import>* en la ventana de comandos de Stata enumera diferentes opciones y comandos disponibles para importar datos de diferentes formatos.

Dado que los microdatos para la mayoría de las EGH están en formato ASCII fijo, el siguiente ejemplo demuestra una forma sencilla de importar los datos necesarios a Stata. Las siguientes tablas muestran parte de un archivo de datos de formato fijo típico y el archivo de diseño que describe los datos. El archivo de diseño indica qué representa el carácter en cada posición de byte en el archivo de datos ASCII. Para extraer o importar esos datos a un formato legible en Stata, o convertirlos a un conjunto de datos de Stata (.dta), se debe crear un archivo de diccionario de Stata con la extensión de archivo ".dct". En el cuadro 2.2 se proporciona un archivo de diccionario de muestra para extraer partes de la información proporcionada en el archivo de datos ASCII.

Archivo de datos de ejemplo en formato ASCII (formato fijo)

```
W1551102113071266621202011    2   4    33815604    488    573003232    0030251
W1551102113071266621202031    2   4    33815604000490    547001213    0010211
W1551102113071266621202051  2 4  33815604  437  460004413  0610251
W1551102113071266621202072  2 4  33815604  473  554001413  0410251
```

Archivo de diseño de ejemplo

```
-------------------------------------------------------------------------------
item                length          byte-pos.           remarks
-------------------------------------------------------------------------------
work-file-id        2               1-2                  "W1"
round-sch           3               3-5                  "551"
sector              1               6                    -
state region        3               7-9
stratum                             2                    10-11
district                            2                    12-13
sub-rnd                             1                    14
fsu-no              5               16-20
samp. hhno.         2               25-26
hh. size                            3                    58-60
scl-group           1               63
```

Cuadro 2.2 Ejemplo de archivo de diccionario para importar datos de archivos ASCII

```
dictionary using datafile.txt {
_column(1)      str2    ID          %2s     "Work file ID"
_column(6)              sector      %1f     "Rural or Urban"
_column(7)              state       %2f     "States"
_column(9)              region      %1f     "Country regions"
_column(10)             stratum     %2f     "Stratum"
_column(12)             district    %2f     "District"
_column(14)             subround    %1f     "Sample sub Round"
_column(16)             fsu         %5f     "First Stage Unit"
_column(25)             hldno       %2f     "Household number"
_column(58)             hsize       %3f     "Household Size"
_column(63)             socgroup    %1f     "Social group"
}
```

Un archivo de diccionario de Stata comienza con una línea como la siguiente <*dictionary using datafile.txt {*> donde "datafile.txt" es el nombre del archivo de microdatos en el directorio de trabajo de Stata. La definición de las variables individuales sigue a continuación. Cada variable está definida por una línea con cinco partes.

Cuadro 2.2 Ejemplo de archivo de diccionario para importar datos de archivos ASCII (cont.)

La primera parte le dice a Stata que comience a leer el archivo de datos desde la posición del byte mencionado entre paréntesis. El segundo indica el tipo de variable: cadena o numérica. Solo las variables de cadena deben indicarse explícitamente como tales. La tercera parte es el nombre mnemotécnico de la variable. El cuarto es el formato de entrada variable que consta de un signo "%", un número que indica el ancho variable y una letra que indica el formato variable: "f" para números y "s" para cadenas. La quinta parte es una etiqueta opcional dada a la variable. El programa de diccionario termina con una llave de cierre "}".

Algunos ejemplos de formatos de entrada que se pueden usar en las definiciones de variables son:

- *%5f* para una variable entera de cinco columnas,
- *%10s* para una variable de cadena de 10 columnas y
- *%7.2f* para un número de siete columnas con dos decimales implícitos.

Recuerde agregar un carácter de retorno en la última línea, es decir, antes de guardar el archivo, mueva el cursor al comienzo de la siguiente línea debajo de "}". Finalmente, el archivo debe guardarse con la extensión de archivo ".dct" (por ejemplo, "diccionario.dct").

Para ejecutar el programa de diccionario de Stata, abra Stata, establezca el directorio de trabajo y dé el comando: *<infile using dictionary>* donde "dictionary" es el nombre del archivo del diccionario. Si el programa se ejecuta correctamente, el programa aparecerá en la pantalla seguido del mensaje "N observaciones leídas" donde "N" indica el número de observaciones en los datos importados. A continuación, ejecute un comando *<describe>*, que devolverá los resultados con el número de observaciones y variables junto con sus etiquetas.

Una vez que se verifique que todas las variables están en orden, emita el comando *<compress>* para cambiar las variables a su formato más eficiente. Finalmente, los datos importados se pueden guardar en la extensión de formato de datos nativa de Stata (".dta") con el comando *<save mydata>* donde "mydata" es el nombre del archivo de datos de Stata que se guardará en el directorio de trabajo de Stata.

2.6 Preparar y dar formato a los datos para el análisis técnico

Las EGH a menudo proporcionan múltiples conjuntos de datos para registros de individuos, registros de hogares y otras variables. Los gastos para diferentes productos pueden estar en diferentes archivos de datos. Además, los datos pueden estar codificados incorrectamente para ciertas variables, y algunos errores obvios podrían corregirse fácilmente para que esas observaciones no se pierdan durante el análisis final. Además, puede haber algunos valores extremos o faltantes de los que hay que hacerse cargo. Por todas esas razones, es importante limpiar los archivos de datos individualmente y fusionarlos en un solo archivo antes de realizar más análisis. Esta sección proporciona algunos pasos básicos a seguir antes de que se pueda preparar un conjunto de datos final para llevar a cabo un análisis estadístico.

2.6.1 Fusionando datos

Las encuestas de hogares a menudo vienen con múltiples archivos de datos o registros para los hogares y los miembros individuales de los hogares. Además, puede haber múltiples registros para los propios hogares. Por ejemplo, un archivo puede tener características básicas del hogar, como el tamaño del hogar, el NSE al que pertenecen, el lugar de residencia, etc., mientras que otro archivo tiene sus gastos de consumo. Los datos sobre los gastos de consumo en sí mismos podrían distribuirse en diferentes archivos de datos. Por lo tanto, puede ser necesario escribir archivos de diccionario separados para extraer datos de diferentes archivos de datos y fusionarlos después de extraer cada conjunto de datos en archivos de datos de Stata separados.

Debido a que este conjunto de herramientas cubre el análisis a nivel del hogar, la información individual debe agregarse al nivel del hogar. Por ejemplo, el sexo de un individuo no es relevante en un análisis a nivel de hogar. Sin embargo, se puede construir una variable que proporcione la proporción de sexos (relación entre el número de hombres y mujeres en un hogar). De manera similar, el nivel de educación de los miembros individuales de un hogar no es relevante para un análisis a nivel de hogar. Sin embargo, se puede construir el promedio de años de educación recibidos por un hogar, como una variable para el análisis a nivel del hogar para indicar el logro educativo de un hogar.

Una vez que se generan las variables deseables a nivel de hogar a partir de los registros de datos individuales, solo se necesita conservar una sola observación por hogar antes de fusionarla con los datos a nivel de hogar. Por ejemplo, una vez que se genera una variable a nivel del hogar — por ejemplo, la proporción de sexos— a partir de datos a nivel individual, se repetirá el mismo valor para la proporción de sexos para todos los miembros del hogar dentro de un hogar. Para conservar solo una observación por hogar, primero ordene los datos por hogar (o ID de hogar) con el comando <*sort hhid*> (donde *hhid* es la variable de identificación de los hogares) y luego ejecute el comando <*drop if hhid==hhid[_n-1]*>. Como alternativa, utilice el comando <*duplicates drop*> después de organizar los datos según sea necesario.

Fusionar datos a nivel de hogar con datos adicionales, ya sea de los registros individuales o de otros registros específicos del hogar, requerirá el uso del comando <*merge*> en Stata. Ejecute el comando <*help merge*> para ver la sintaxis y las diferentes formas de fusionar archivos de datos en Stata. Stata genera una nueva variable <*_merge*>, después de cada comando de fusión para facilitar la verificación si la fusión se ha realizado correctamente. Es una variable categórica que contiene un código numérico que indica la fuente y el contenido de cada observación en el conjunto de datos fusionados. El comando <*tabulate _merge*> después de la ejecución de un <*merge*> dará la indicación necesaria. Por ejemplo, el código 3 de "_merge" es para observaciones que coinciden correctamente con ambos conjuntos de datos.

El aspecto más importante al fusionar dos archivos diferentes es encontrar un conjunto de variables que puedan identificar de manera única cada observación en cada uno de los conjuntos de datos que se fusionarán. Eso debe comprenderse desde el momento del diseño de la encuesta y extraerse junto con cada extracción de datos utilizando archivos de diccionario. La falta de identificadores únicos o los identificadores definidos incorrectamente puede resultar en la fusión errónea de información de un hogar con la de otro. El Cuadro 2.3 da un ejemplo de cómo identificar esas variables y fusionar archivos correctamente.

Para hacer una fusión uno a uno, tanto los "datos maestros" (master data en inglés) como los "datos de uso" (using data en inglés) deben ser identificables con el mismo conjunto de variables únicas. Solo se pueden realizar análisis adicionales con aquellas observaciones que coincidieron con los archivos de datos "master" y "using", es decir, observaciones para las que la variable (_merge) tome el valor de 3. Para usar solo variables sin datos faltantes tanto del master como del archivo de datos using, es importante descartar las observaciones para las cuales "_merge" no es igual a 3 usando el comando < *drop if _merge!=3*>. Sin

Cuadro 2.3 Un posible desajuste de los hogares durante la fusión

La Encuesta de Ingresos y Gastos de los Hogares de Bangladesh (2010) sigue una técnica de muestreo aleatorio estratificado en dos etapas. La descripción del diseño de la muestra en el informe publicado dice que alrededor de 200 hogares fueron seleccionados de aproximadamente 1000 UPM en todo el país, mientras que las UPM mismas se seleccionaron de aproximadamente 16 estratos diferentes. Está claro que un hogar de esa encuesta debe identificarse de manera única utilizando las variables que representan los estratos, la UPM y el número de hogar. Esas variables son *stratum, psu* y *hhold*, respectivamente, como se indica en la documentación. Dado que los números de UPM en sí mismos son únicos en esos datos, también se puede identificar una identificación de hogar única utilizando solo las variables *psu* y *hhold*.

Se puede generar una variable de identificación de hogar única (*hhid*) para esos datos con el comando <*egen hhid=group(psu hhold)*> donde los valores entre paréntesis corresponden a los nombres de variables requeridos para identificar de manera única el hogar. Por ejemplo, si los números de UPM no fueran únicos y variaran entre estratos, se necesitarían las tres variables para generar *hhid*. Por lo tanto, cualquier fusión de dos registros a nivel de hogar en esos datos utilizará esas variables. Por ejemplo, la HIES tiene un archivo de datos demográficos del hogar (rt001) y un archivo de gastos agregados a nivel del hogar (hhold_exp_hies2010). Si se van a fusionar los archivos, ambos datos deben extraerse por separado y guardarse como archivos de datos de Stata, por ejemplo, con los nombres "hh1.dta" y "hh2.dta". Después de cargar *hh1, hh2* se puede fusionar con él usando el comando <*merge 1:1 psu hhold using hh2*>. Eso fusionaría correctamente los mismos hogares en un archivo de datos con los del otro. El comando <*tab _merge*> mostrará con qué precisión se fusionaron los archivos de datos para que el usuario pueda ver que no hay discrepancias.

Por otro lado, supongamos que primero se generó una variable *hhid* única para cada uno de los archivos de datos por separado, y luego se fusionaron con el comando <*merge 1:1 hhid sing hh2*>, donde la variable de ID única pregenerada (*hhid*) se utilizó para la fusión en lugar de los identificadores de hogar originales (*psu* y *hhold*). Eso también fusionará ambos archivos de datos, y el comando <*tab _merge*> no mostrará discrepancias. Sin embargo, en ese caso, los hogares en ambos conjuntos de datos podrían estar incorrectamente emparejados debido a varias razones:

1. Al generar un *hhid* único en cada archivo de datos individual, Stata asigna ID únicos a cada hogar utilizando el orden de clasificación existente en cada archivo de datos. Si el orden de clasificación de ambos archivos de datos era diferente cuando se generó la variable *hhid*, resultará en familias emparejadas incorrectamente después de la fusión.
2. Suponga que algunos números *psu* o *hhold* fueran diferentes en ambos conjuntos de datos debido a una codificación incorrecta. El <*tab _merge*> después de una fusión correcta usando *psu* y *hhold* mostrará observaciones no emparejadas. Mientras que la fusión con *hhid* pregenerado fusionaría ambos archivos de datos a la perfección, sin poder identificar las discrepancias.
3. Supongamos que el número de observaciones en *hh1* y *hh2* fuera diferente. Una fusión con las variables *psu* y *hhold* coincidiría correctamente con los hogares, mientras que la fusión con *hhid* generado previamente los emparejaría sin darse cuenta.
4. Por lo tanto, los datos de dos archivos de datos diferentes deben fusionarse solamente utilizando todas las variables relevantes que se utilizan para identificar únicamente las observaciones (hogar o persona) en cada archivo de datos. En otras palabras, el comando <*merge*> debe tener todas las variables que identifiquen de forma única una observación presente durante la fusión.

embargo, puede haber situaciones en las que sea necesario conservar en el archivo de datos fusionados aquellas observaciones no emparejadas de los archivos de datos maestros o de datos de uso.

Además de fusionar diferentes archivos (como datos de hogares y datos individuales) de la misma ronda de una EGH determinada, también puede haber situaciones en las que el usuario desee agregar datos de EGH de diferentes años u olas. Obviamente, los hogares en diferentes rondas de EGH pueden ser diferentes entre sí, y lo que se requiere no es una fusión, sino una combinación de diferentes EGH para crear una sección cruzada combinada. En este caso, en lugar de *<merge>*, se debe usar el comando *<append>* en Stata. Para hacer eso, los datos de cada ronda de EGH deben contener el mismo tipo de variables y primero se debe preparar un único conjunto de datos combinados para cada ronda de EGH. Una vez que se realiza la adición, simplemente se agregará al número de observaciones en los datos maestros.

Antes de agregar, es importante crear una variable de año o ciclo y marcarla con números que puedan identificar cada año/ciclo/ronda de la encuesta. Si los datos agrupados finales pertenecen a varios años (generalmente de diferentes oleadas de la encuesta), también es importante ajustar por inflación cualquier variable de gasto o precio para que los valores en diferentes rondas de datos estén en términos constantes y, por lo tanto, sean comparables.

2.6.2 Remodelación de datos

Según el análisis que se realice, puede ser importante remodelar los datos en formato largo o ancho en Stata. Para hacer esto, ejecute el comando *<help reshape>* para comprender cómo se realiza la remodelación de una forma a otra. En un formato ancho, solo habrá tantas observaciones como el número de hogares únicos en un conjunto de datos, mientras que, para un formato de datos largo, los mismos hogares pueden repetirse varias veces, apilados uno debajo del otro. Por ejemplo, suponga que hay información sobre los gastos en cigarrillos y tabaco sin humo. Para los hogares con gastos en ambos productos, habrá dos observaciones para cada hogar bajo un formato largo, mientras que bajo el formato ancho los gastos en cigarrillos y tabaco sin humo aparecerán como variables separadas contra una sola observación del hogar.

Para la mayoría de los análisis, es útil tener los datos remodelados en formato ancho. Por lo tanto, si los datos extraídos están en formato largo, se les debe cambiar la forma a un formato ancho usando el comando *<reshape wide stub, i(i) j(j) >*, después de determinar la observación lógica (i) y la subobservación (j) por el cual organizar los datos.

2.6.3 Limpieza de datos

La limpieza de los datos antes de realizar el análisis estadístico es fundamental, especialmente en el caso de las encuestas de hogares, ya que son datos recopilados por diferentes personas a lo largo del país en diferentes etapas. Por ejemplo, un cero en lugar de un valor faltante podría generar resultados no deseados, como distorsionar la media y las varianzas al realizar un análisis estadístico. Errores similares en los datos son: duplicados, variables categóricas codificadas erróneamente y valores atípicos inaceptablemente altos o bajos para ciertas variables.

De manera similar, si una variable de cadena tiene diferentes ortografías o caracteres de espacio entre las observaciones, Stata consideraría esas entradas como una categoría diferente. Por ejemplo, si masculino bajo la variable sexo se codifica como “Masculino” o “MASCULINO” o “M” o “masculino” u otras variaciones posibles, entonces en lugar de obtener MASCULINO y FEMENINO como dos categorías diferentes, puede haber varias categorías diferentes. Por esas y otras razones, es importante hacer un examen completo de cada una de las variables y asegurarse de que los datos estén codificados de manera consistente.

La tabla 2.1 proporciona una buena secuencia de pasos que se pueden seguir para obtener un conjunto de datos limpio, incluye comandos útiles de Stata que se pueden usar durante estos pasos. Tenga en cuenta

que los pasos mencionados en la tabla no necesitan realizarse estrictamente en el mismo orden que se indican. Usando el comando de ayuda de Stata, seguido de los comandos relevantes de Stata mencionados en esta tabla, el lector puede aprender más sobre cada uno de esos comandos y familiarizarse con diferentes ejemplos.

2.7 Generación de estadísticas descriptivas básicas a partir de encuestas de hogares

Un programa de software estadístico generalmente analiza los datos como si los datos se recopilaran mediante un muestreo aleatorio simple. Sin embargo, como se mencionó anteriormente, la mayoría de las encuestas de hogares utilizan diseños de encuestas más complejos y de varias etapas para recopilar datos,

Tabla 2.1 Estrategia de limpieza de datos

Razón (¿Por qué?)	Paso (¿Qué?)	Comando (¿Cómo?)
Identificar las variables y corregir el código incorrecto	Etiquetar/reetiquetar variables y etiquetar sus valores	label; recode
Identificar observaciones únicas para fusionar correctamente	Comprender los identificadores únicos del diseño de la encuesta y los datos extraídos	egen group(); isid; codebook; inspect; duplicates
Corregir la ortografía; hacer que los datos sean uniformes	Corregir variables de cadena	replace; substr; subinstr; index
Cambiar y transformar variables por necesidad de análisis	Transformar variables	gen; destring; tostring; drop; keep; egen; rename; bysort; encode; recode
Asegurarse de que las conexiones lógicas estén presentes en los datos, como que las madres sean mujeres o que las cantidades tengan las unidades correctas	Verificaciones de consistencia	assert; tabulate; summarize; table; tabstat; count
Crear un solo archivo de datos para trabajar	Combinar o agregar diferentes archivos de datos	merge 1:1; merge m:1; merge 1:m; append
Crear una observación lógica para organizar el archivo de datos	Remodelar los datos al formato ancho o largo apropiado	reshape
Identificar la importancia y la influencia de valores faltantes	Decidir si es necesario eliminar o imputar las observaciones faltantes	sum; mi
Detectar valores atípicos	Eliminar o sustituir los valores atípicos según sea necesario	sum; hist; hilo; stem; graph box; scatter
Mantener un registro de todos los comandos para facilitar la replicación y la colaboración	Documentar cada paso con comentarios y comandos	usar el editor do-file para organizar

y la estratificación y la agrupación en las encuestas por muestreo afectan el cálculo de los errores estándar. Por lo tanto, el análisis estadístico realizado debe poder corregir los elementos de diseño utilizados en la encuesta para obtener estimaciones puntuales y errores estándar más precisos. La documentación que se proporciona junto con los datos de la encuesta generalmente brinda información detallada sobre el diseño de muestreo específico que se utilizó. Esta sección analiza cómo declarar los elementos del diseño de la encuesta y producir estadísticas descriptivas para la muestra completa y por categoría específica. Esta sección también ofrece orientación sobre código de Stata útil para realizar estas acciones.

En Stata, el comando *<svyset>* se utiliza para declarar el diseño de la encuesta de los datos. Designa variables que contienen información sobre el diseño de la encuesta, como las ponderaciones muestrales, la UPM/conglomerado y los estratos, y especifica otras características del diseño de la encuesta, como el número de etapas de muestreo y el método de muestreo. La declaración de diseño, si es necesario, se puede borrar con el comando *<svyset, clear>*. Una vez que los datos se declaran con *<svyset>*, solo el prefijo *<svy:>* debe preceder a cada comando. La sintaxis del comando *<svyset>* para un diseño de encuesta de varias etapas es similar a: *<svyset psu [weight] [, design options] [|| ssu, design options] … [options]>* donde psu es el nombre de una variable que identifica la unidad primaria de muestreo en los datos, *weight* identifica el ponderador muestral, *ssu* identifica las unidades muestrales en la segunda etapa, y así sucesivamente.

Las opciones de diseño declararán los elementos de diseño como los estratos. El sitio web de Stata proporciona un conjunto de datos de encuestas de muestra de la segunda Encuesta Nacional de Examen de Salud y Nutrición (NHANES) en los EE. UU. de 1976 a 1980. Importe esos datos a Stata con el comando *<webuse nhanes2>*. Los datos tienen una variable de ponderación (*finalwgt*), una variable de UPM (*psu*) y una variable de estratos (*strata*). El comando *<svyset>* en este caso se verá así: *<svyset psu [pw=finalwgt], strata(strata)>*, donde “pw” representa los ponderadores de probabilidad.

La mayoría de las encuestas incluyen explícitamente ponderaciones de muestreo, estrato e identificadores de UPM junto con los datos publicados. Es importante leer detenidamente la documentación de la encuesta para comprender la descripción de las variables. Dado que los informes publicados de la encuesta también presentan estimaciones puntuales importantes, es posible comparar los números calculados con los de los informes publicados. Antes de continuar con el análisis, es importante realizar examinaciones cruzadas para asegurarse de que se utilicen las ponderaciones de muestreo y los elementos de diseño de la encuesta correctos como se pretendía originalmente.

Una vez que se declara el diseño de la encuesta por medio de *<svyset>*, la información sobre los estratos y la UPM se puede obtener con el comando *<svydescribe>*. La estimación adicional de estadísticas descriptivas debe tener el prefijo *<svy:>*. Por ejemplo, para estimar la media de una variable, simplemente ejecute el comando *<svy: mean varname>*. Si la media se calcula para una variable binaria, se mostrarían proporciones. Alternativamente, ejecute *<svy: tab binaryvar>* para estimar las proporciones de, por ejemplo, hombres y mujeres, alfabetizados y analfabetos, o variables binarias similares junto con sus errores estándar corregidos para el diseño de la encuesta. De manera similar, *<svy: proportion binaryvar>* produciría una salida con proporciones de la variable de interés junto con sus errores estándar e intervalo de confianza.

Para estimar las mismas estadísticas descriptivas para subgrupos en la encuesta, como grupos de ingresos, género o cualquier otra categoría NSE, el comando *<svy>* se puede ejecutar con opciones adicionales como *<subpop>* u *<over>*. Por ejemplo, el comando *<svy, subpop (female): mean binaryvar>* o *<svy, over(female): mean binaryvar>* proporciona las estimaciones de interés necesarias junto con sus errores estándar. Supongamos que a uno le gustaría encontrar los gastos promedio en cigarrillos por diferentes cuartiles de gasto. Para hacerlo, primero cree una variable para categorizar los hogares en cuatro cuartiles diferentes en función de sus gastos mensuales totales del hogar "exptotal", de la siguiente manera: *<xtile exp_cuartiles =exptotal, n(4)>*. Luego, use el comando *<svy, over(exp_quartiles) : mean exp_cig>* para obtener los gastos mensuales promedio en cigarrillos por diferentes cuartiles de gasto.

Las estimaciones a partir de los datos de la encuesta también se pueden producir sin declarar explícitamente el diseño de la encuesta, pero utilizando las ponderaciones de muestreo correctas y ajustando los errores estándar. En Stata, eso se hace con la ayuda de ponderaciones y opciones de conglomerados robustos. Por ejemplo, en el caso anterior de gastos en cigarrillos por cuartiles de gasto, se pueden obtener los mismos gastos promedio por diferentes cuartiles de gasto con el comando *<mean exp_cig [pw=weightvar], over (exp_quartiles)>*, donde weightvar es el identificador de ponderación de muestreo que se utilizó para declarar el diseño de la encuesta.

Sin embargo, las estadísticas descriptivas que utilizan las ponderaciones muestrales, si bien producen las mismas estimaciones que las que utilizan *<svyset>*, no abordan adecuadamente los problemas de estratificación y, como resultado, podrían producir errores estándar diferentes a los obtenidos con el comando *<svy>*. No obstante, en el contexto de regresión, se podría agregar el argumento opcional *<robust cluster(psuvar)>* después del comando de regresión principal, donde *psuvar* es la variable que identifica el conglomerado o UPM en los datos, y corregiría los efectos del diseño de la encuesta mientras se calculan los errores estándares para las estimaciones de los coeficientes.

3 *Estimación de la elasticidad precio y precio cruzada*

Este capítulo presenta métodos para estimar la elasticidad precio de la demanda usando datos de EGH. La elasticidad de los precios es uno de los parámetros más importantes que se deben tener en cuenta al diseñar la política tributaria, ya que proporciona una idea a los decisores políticos sobre la capacidad de respuesta de la demanda a los cambios en el precio. Con base en la elasticidad de precios estimada, los decisores políticos pueden predecir con cierto grado de confianza el impacto de sus políticas en los objetivos de políticas relevantes, incluidos el consumo de tabaco y los ingresos fiscales. Además, la evidencia empírica sobre la magnitud en la que la demanda de tabaco respondería al precio proporciona un contraargumento muy relevante a los opositores que afirman que aumentar los impuestos resultaría sin ambigüedades en la reducción de los ingresos fiscales.

Los decisores políticos están interesados en la sensibilidad del consumo de tabaco no solo a los cambios en los precios del tabaco (es decir, la elasticidad precio del tabaco), sino también a los cambios en los precios de otros bienes, como sus complementos potenciales (por ejemplo, alcohol o café) o sus sustitutos. De manera similar, los formuladores de políticas públicas pueden querer saber el impacto de un cambio en el precio de un tipo de producto de tabaco (como los cigarrillos) en otros tipos (como los cigarrillos para armar), ya que el impacto de su política puede reducirse efectivamente si, por ejemplo, hay espacio para la sustitución hacia abajo.

También es útil comprender la sensibilidad al precio de los consumidores de tabaco no solo con respecto a la cantidad de producto de tabaco que consumen, sino también con respecto a cómo los cambios de precio afectan su decisión de iniciar y abandonar el consumo de tabaco. Mientras que el primero se mide usando elasticidades en el margen intensivo (también conocido como “elasticidad de cantidad”), el segundo puede evaluarse estimando la elasticidad en el margen extensivo (también conocido como “elasticidad de prevalencia”).

En este capítulo, esos conceptos se definen en detalle junto con ejemplos. A continuación, se realizará una breve discusión teórica sobre la estimación de las elasticidades de cantidad y prevalencia utilizando EGH, en ese orden. En la última parte del capítulo, se proporciona el código de Stata para que el lector pueda estimar las elasticidades. Finalmente, se proporciona un ejemplo utilizando datos de EGH de un país no identificado.

3.1 Definición de conceptos

La elasticidad precio de la demanda se define formalmente como el cambio porcentual en la cantidad demandada de un bien que resulta de un cambio del uno por ciento en el precio de ese bien, manteniendo todo lo demás constante (*ceteris paribus*). Por ejemplo, una elasticidad precio de la demanda de -0,5 implicaría que la cantidad demandada de ese bien en particular disminuye un cinco por ciento cuando el precio del bien aumenta un 10 por ciento. De manera similar, una elasticidad precio de la demanda de -1,5 implica que la cantidad demandada del bien en cuestión disminuye un 15 por ciento cuando su precio aumenta un 10 por ciento. Debido a que mide el cambio porcentual en la cantidad consumida, también se conocen como elasticidades de cantidad o elasticidades en el margen intensivo.

Se dice que los bienes con una elasticidad precio de la demanda inferior a uno en valor absoluto tienen una demanda inelástica porque la respuesta de la demanda es relativamente menor que el cambio de precio. Por otro lado, se dice que los bienes con elasticidad precio de la demanda de más de uno en valor absoluto tienen demanda elástica porque la respuesta de la demanda es relativamente mayor que el cambio de precio. Hay varios factores que afectan la elasticidad precio, como la disponibilidad de sustitutos, si es que el bien es una necesidad, el período de tiempo disponible para encontrar alternativas, qué tan amplia o acotada es la definición del producto y/o la naturaleza adictiva/recurrente del producto. Teniendo en cuenta esos factores, los productos de tabaco, que tienen pocos sustitutos y son adictivos, tienden a tener una elasticidad precio de la demanda relativamente inelástica.

Para la mayoría de los productos ordinarios, son las elasticidades cuantitativas las de interés en lo que se refiere a la política tributaria. Sin embargo, los productos de demérito como el tabaco son consumidos por unos pocos. Eso significa que, para un gran número de personas, su cantidad de consumo está condicionada a su decisión de participar o consumir. Por lo general, esas decisiones siguen un proceso de dos pasos: (i) una decisión sobre si consumir o no, y (ii) una vez decidido consumir, cuánto consumir.

Por ejemplo, la decisión de consumir o "participar" en el proceso de consumo puede tomar la forma de que los no fumadores actuales comiencen a fumar o los fumadores actuales continúen o abandonen el hábito. Juntas, pueden llamarse elasticidad de prevalencia o de participación, o elasticidad en el margen extensivo. Eso se define como el cambio proporcional en la prevalencia del tabaquismo como resultado de un cambio proporcional dado en el precio de los cigarrillos.

Debido a la naturaleza condicional de la cantidad de consumo, las elasticidades de cantidad también se denominan elasticidades condicionales en este contexto. La elasticidad precio de la demanda de cigarrillos es, por lo tanto, una suma de las elasticidades en los márgenes extensivo e intensivo, lo que refleja el impacto de los cambios de precio tanto en la prevalencia como en la intensidad del tabaquismo. La participación respectiva de ambas elasticidades (cantidad y prevalencia) en el coeficiente de elasticidad precio total es algo que debe determinarse empíricamente y dependerá de los datos particulares disponibles.

Que la demanda de un bien sea elástica o inelástica es muy importante para la política tributaria. Se puede esperar que los ingresos tributarios disminuyan cada vez que se aumentan los impuestos a un bien que es elástico en la demanda, ya que la respuesta de la demanda supera el cambio de precio, por lo que los ingresos por ventas y los ingresos tributarios finalmente disminuyen. Por otro lado, se puede esperar que los ingresos tributarios aumenten cada vez que se aumentan los impuestos a un bien que tenga demanda inelástica, ya que su respuesta de la demanda es menor que el cambio de precio, por lo que los ingresos por ventas y los ingresos tributarios finalmente aumentan.

La literatura sobre la estimación de las elasticidades precio de la demanda de cigarrillos tiende a encontrar, en general, elasticidades que oscilan entre 0 y -1,[4, 20, 21] lo que significa que la demanda de tabaco es inelástica, que es lo esperado dada la naturaleza adictiva de ese producto, así como la disponibilidad de muy pocos sustitutos cercanos. La evidencia empírica también confirma que la tributación del tabaco, a través del aumento de precios, es una de las herramientas de política pública más efectivas para disminuir el tabaquismo y sus consecuencias adversas para la salud.[4, 22–24]

Además de la elasticidad precio, también debe definirse la elasticidad precio cruzada. Formalmente, la elasticidad precio cruzada de la demanda entre el bien X y el Y se define como el cambio porcentual en la demanda del bien Y cuando el precio del bien X cambia en un uno por ciento, *ceteris paribus*. A diferencia de la elasticidad precio, que siempre es inequívocamente negativa, la elasticidad precio cruzada puede tener un signo negativo o positivo. Una elasticidad precio cruzada negativa significa que los dos bienes en cuestión

son complementarios. En otras palabras, el consumo conjunto de los dos bienes satisface una necesidad. Un ejemplo sería la gasolina y los automóviles. Por otro lado, una elasticidad precio cruzada positiva significa que los dos bienes son sustitutos. Es decir, un bien puede usarse en lugar del otro bien o ambos satisfacen la misma necesidad. Un ejemplo de sustitutos es el agua embotellada y el agua del grifo.

Además, existe una elasticidad ingreso de la demanda. En este conjunto de herramientas, los términos elasticidad ingreso y elasticidad gasto se usan indistintamente, ya que el gasto total en EGH se usa como un indicador del ingreso. La elasticidad ingreso de la demanda se define formalmente como el cambio porcentual en la cantidad demandada de un bien que surge de un aumento del ingreso del uno por ciento, *ceteris paribus*. Una elasticidad ingreso negativa de la demanda significa que la cantidad demandada del bien disminuye cada vez que aumentan los ingresos. Dichos bienes se denominan bienes “inferiores”. Los alimentos básicos (como el arroz o el maíz) a menudo tienen elasticidades-ingreso negativas de la demanda. Por otro lado, los bienes que tienen elasticidades ingreso de la demanda positivas se denominan bienes "normales".

Conocer la magnitud de la elasticidad ingreso de la demanda es importante para la política de control del tabaco. Por ejemplo, una elasticidad ingreso positiva de la demanda de cigarrillos en un país implica que se deben intensificar los esfuerzos de control del tabaco, especialmente en períodos de aumento de los ingresos en ese país.

Así como la elasticidad ingreso cuantifica cambios en el consumo como resultado de cambios en el ingreso, también se observa que la elasticidad precio de un producto dado puede variar para individuos de diferentes grupos de ingresos o grupos NSE. Las estimaciones de las elasticidades de los precios de los cigarrillos y otros productos de tabaco pueden variar según los grupos NSE. En general, la literatura encuentra que las personas o los hogares en los grupos de ingresos más bajos son mucho más sensibles a los cambios de precio de un producto determinado en comparación con los de los grupos de ingresos más altos. Por ejemplo, la mayoría de las estimaciones de las elasticidades precio de la demanda de cigarrillos de los países de ingresos altos oscilan entre –0,2 y –0,6, agrupados en torno a –0,4, mientras que las estimaciones de los países de ingresos bajos y medianos oscilan entre –0,2 y –0,8, agrupados en torno a –0,5.[4]

Esto sucede incluso dentro de los países de ingresos altos, estudios en el Reino Unido (UK) y Australia[25, 26] muestran una sensibilidad relativamente mayor al precio entre los grupos de nivel socioeconómico más bajo en comparación con los de nivel socioeconómico alto. En los Estados Unidos, mientras que la mayoría de los estudios[27–29] encuentran una respuesta relativamente mayor a los cambios en el precio del tabaco en los grupos de nivel socioeconómico más bajo que en los de nivel socioeconómico alto, algunos estudios[30] ofrecen evidencia no concluyente. En el caso de los PIMB, aunque algunos estudios[31–38] han ofrecido evidencia mixta, un cuerpo de evidencia mucho más grande y creciente muestra una respuesta significativamente mayor en los grupos de ingresos más bajos que entre aquellos con ingresos más altos. Estos incluyen estudios de países como Argentina,[39] Bangladesh,[40] Bosnia y Herzegovina,[41] Brasil,[42] China,[44, 45] El Salvador,[45] India,[46, 47] Indonesia,[48] Irán,[49] Kosovo,[50] México,[51] Montenegro,[52] Pakistán,[53] Perú,[54] Serbia,[55] Tanzania,[56] Tailandia,[57] y Turquía,[58] entre otros.

También se pueden observar diferencias similares en la elasticidad de los precios si el análisis se realiza en diferentes clasificaciones de NSE distintas del ingreso. Este conjunto de herramientas analiza la estimación de las elasticidades de los precios por diferentes grupos de ingresos. Este conjunto de herramientas no repite las instrucciones para la estimación de la elasticidad de precios con otras clasificaciones de NSE, ya que también seguirían un enfoque similar.

3.2 Aspectos econométricos en la estimación de la demanda

Hay varias cuestiones teóricas y prácticas a considerar en la estimación de las elasticidades precio de la demanda. Esta sección cubre algunos de los temas principales.

3.2.1 Problema de identificación en el análisis de la demanda

La ley de la demanda establece que a medida que aumenta el precio de un bien, su demanda disminuye, *ceteris paribus*. Supone que la dirección de la causalidad va del precio a la cantidad demandada. Sin embargo, las interacciones del mercado tienden a ser más complejas, porque la demanda influye en el precio tanto como el precio influye en la demanda. Eso se puede observar en tiempo real en los mercados bursátiles. Es probable que un aumento en el precio de una acción conduzca a una reducción en la cantidad demandada de la acción. Por otro lado, es probable que un aumento en la demanda de la acción conduzca a un aumento en el precio de esa acción. Además, otros factores (como los ingresos, los gustos, el clima y los precios de los bienes relacionados) pueden, además de la influencia del precio, influir en la demanda del bien.

Los problemas explicados anteriormente se conocen en el análisis econométrico como el "problema de endogeneidad" o el "problema de identificación", y no abordarlos adecuadamente conduciría a obtener estimaciones sesgadas; es decir, las estimaciones serían significativamente diferentes del valor real del parámetro que se está estimando. Este es un tema muy relevante para la formulación de políticas públicas, ya que conduciría a una política que puede diseñarse en base a estimaciones irrealmente positivas o negativas, según el signo del sesgo.

Idealmente, el problema de endogeneidad, o el problema de identificación, puede resolverse econométricamente ejecutando un experimento en el que las unidades se asignan aleatoriamente a grupos de tratamiento o de control. Aquí, no hay necesidad de preocuparse por la endogeneidad porque la aleatorización descarta todos los demás factores excepto el factor de interés. Desafortunadamente, con la realidad social, a diferencia de las ciencias físicas, no siempre es fácil o incluso deseable realizar experimentos sociales. Por lo tanto, los economistas y los científicos sociales buscan experimentos “naturales” o cuasiexperimentos que puedan ser explotados para superar el problema de identificación.

Con respecto a la estimación de la elasticidad precio de la demanda de productos de tabaco, los investigadores han buscado casos en los que los gobiernos hayan introducido de forma independiente (es decir, exógenamente) un aumento en los precios del tabaco. Por ejemplo, varios estudios en los EE. UU. en la década de 1990 aprovecharon el aumento del impuesto a los cigarrillos de 25 centavos en California y Massachusetts para estimar la elasticidad precio de la demanda,[59–62] porque la fuente exacta del cambio de precio que condujo a un cambio en la cantidad demandada se puede identificar en esos eventos.

Sin embargo, esos cambios drásticos en los impuestos al tabaco no son muy comunes, especialmente en los países de ingresos medianos y bajos donde, a menos que estén pasando por una reforma tributaria al tabaco, los cambios en los impuestos al tabaco suelen ser graduales y de pequeña magnitud, generalmente para corregir el impacto de la inflación. Esos cambios graduales dificultan aislar el efecto causal del precio sobre la demanda, por lo que el procedimiento de estimación requiere el uso de VI para obtener el efecto causal del precio sobre la demanda (consulte el Capítulo 2 para ver una discusión sobre la endogeneidad y el papel de las VI para resolverlo).

Las VI son difíciles de conseguir en general y en el análisis de la demanda en particular. Afortunadamente, el premio Nobel Angus Deaton ha propuesto una VI adecuada en el contexto de los PIMB que permite la estimación de elasticidades justificables. El método propuesto por Deaton se detalla a continuación.

3.3 Estimación de la elasticidad cantidad con encuestas de gasto de los hogares

Esta sección analiza el marco teórico detrás de la estimación de la elasticidad de la cantidad seguido de su estimación utilizando EGH. Si bien existen algunos modelos diferentes que utilizan un sistema de ecuaciones de demanda, el sistema casi ideal de demanda (AIDS) introducido por Deaton y Muellbauer (1980)[63] ha sido el más popular debido a sus múltiples ventajas. El AIDS tiene una forma flexible y funcional consistente con los datos de gastos de los hogares y diferentes axiomas de elección. No impone restricciones previas sobre las elasticidades, y su especificación mayoritariamente no lineal facilita la estimación, lo que le permite probar explícitamente las restricciones de homogeneidad y simetría. El modelo de Deaton (1988), presentado en este conjunto de herramientas[64] y detallado en su libro,[8] se basa en Deaton y Muellbauer (1980).[63] Sin embargo, se diferencia de AIDS en que corrige tanto por los errores de medición como por el matizado de calidad en valores unitarios, como se analiza a continuación.

El modelo permite que los datos de EGH se utilicen para estimar elasticidades de precios creíbles de la demanda, partiendo del supuesto de que los precios de la mayoría de los bienes en los PIMB varían significativamente a lo largo del espacio geográfico. Esa variación espacial del precio es el resultado de costos de transporte significativos, debido a que los bienes se trasladan de un lugar a otro, o de otros factores, como diferentes impuestos fronterizos o aranceles adicionales en diferentes jurisdicciones en el mismo país. Por lo tanto, los costos de transporte, o esos otros factores que afectan los cambios de precios entre regiones geográficas, sirven implícitamente como un instrumento y son los principales factores que influyen en el precio, que a su vez influye en la demanda. Por lo tanto, se asume una variación genuina en el precio entre los conglomerados para la identificación de las elasticidades precio en ese modelo, resolviendo efectivamente el problema de identificación destacado anteriormente en este capítulo.

El supuesto de que los precios varían espacialmente significa que los hogares que viven cerca unos de otros, como los que se encuentran en el mismo pueblo o bloque urbano, deben afrontar el mismo precio cuando hacen compras en el mismo mercado y al mismo tiempo, si es que se trata de una encuesta transversal. Por otro lado, los hogares que viven muy separados, como los que se encuentran en diferentes pueblos o bloques urbanos, deben enfrentar precios diferentes. En otras palabras, el enfoque requiere que gran parte de la variación observada en el precio se produzca entre conglomerados, como se mencionó en el Capítulo 2, y no dentro de los conglomerados. Desde el punto de vista econométrico, eso requiere que la variación de los precios se explique en gran medida por los “efectos de conglomerados” o las “dummies de conglomerados”. Cualquier variación en el precio dentro de un grupo debe ser el resultado de un error de medición, cuyos patrones se pueden utilizar para corregir las estimaciones finales de dicho error (más información en la Sección 3.3.1).

Las encuestas de gastos de los hogares generalmente no informan el precio de mercado en la encuesta. Podría deducirse de las decisiones de compra de los hogares calculando la relación entre el gasto de los hogares en un bien determinado y la cantidad de ese bien. Esa relación, sin embargo, es un valor unitario y no un precio. Los valores unitarios no son lo mismo que los precios debido a los siguientes dos problemas. En primer lugar, los valores unitarios se ven afectados tanto por el precio real como por la elección de la calidad (es decir, los "efectos de la calidad"). Si no se trata adecuadamente, eso podría conducir al llamado "matizado de calidad" o en inglés "quality shading", que se refiere a una situación en la que un cambio de precio no conduce a una reducción en la cantidad demandada, ya que las personas cambian a productos más baratos, pero de menor calidad. En segundo lugar, los valores unitarios no son lo mismo que los precios debido al error de medición, dado que las personas suelen informan mal los gastos y/o las cantidades de bienes comprados. Deaton propone un método para tratar tanto el matizado de calidad como el error de medición. La siguiente sección ofrece una explicación técnica paso a paso del método propuesto originalmente por Deaton en 1988, que desde entonces se ha ampliado en su trabajo posterior.[8, 65–67]

3.3.1 Marco teórico del modelo Deaton

Esta sección describe brevemente los principales pasos involucrados en la derivación del modelo teórico propuesto por Deaton para estimar las elasticidades precio (en el margen intensivo) utilizando datos de EGH. Se recomienda a los investigadores que planeen implementar este modelo que lean el Capítulo 5 de Deaton (1997)[8] para comprender los detalles más finos del modelo. El modelo consta principalmente de seis pasos, desde la obtención de los valores unitarios y pruebas pertinentes hasta la estimación final de las elasticidades precio y gasto.

Paso 1: Obtención de valores unitarios

Primero, los valores unitarios se derivan de los datos de la encuesta a nivel de hogar. Eso se hace dividiendo el gasto total informado en el producto o productos de tabaco en particular sobre los cuales la EGH proporciona datos para su cantidad correspondiente, como:

$$v_{hc} = \frac{x_{hc}}{q_{hc}} \quad (3.1)$$

donde v_{hc}, x_{hc} y q_{hc} son, respectivamente, el valor unitario, el gasto y la cantidad de cigarrillos o cualquier otro producto de tabaco en el hogar *h* ubicado en el conglomerado *c*.

Paso 2: Pruebas de variación espacial en valores unitarios

El segundo paso consiste en comprobar si los valores unitarios obtenidos en el Paso 1 satisfacen el principal supuesto identificador: los valores unitarios varían espacialmente. Eso se hace mediante el análisis de varianza (ANOVA) para dividir la variación total de los valores unitarios en "variaciones dentro de un grupo" y "variaciones entre grupos". Un estadístico *F* significativamente grande para el ejercicio de ANOVA lleva a la conclusión de que los valores unitarios varían a lo largo del espacio geográfico o de los grupos.

Paso 3: Estimación de regresiones dentro de un conglomerado

En un tercer paso, se estiman regresiones dentro de un conglomerado de valores unitarios y participación en el presupuesto utilizando la siguiente especificación:

$$lnv_{hc} = \alpha^1 + \beta^1 lnx_{ic} + \gamma^1 Z_{hc} + \psi ln\pi_c + u^1_{hc} \quad (3.2)$$

$$w_{hc} = \alpha^0 + \beta^0 lnx_{ic} + \gamma^0 Z_{hc} + \theta ln\pi_c + (f_c + u^0_{hc}) \quad (3.3)$$

donde lnv_{hc} es el logaritmo del valor unitario, derivado de acuerdo con la Ecuación 3.1 para el hogar *h* en el conglomerado *c*, mientras w_{hc} representa la participación del gasto en tabaco en el gasto total del hogar para el hogar *h* en el conglomerado *c*. Y lnx_{hc} es el logaritmo del gasto total del hogar durante el período de referencia pertinente. Z_{hc} es un vector de características específicas del hogar que pueden incluir variables sobre la estructura del hogar (como el tamaño del hogar, la proporción de adultos o la proporción de hombres) y la demografía del hogar (como la edad, el sexo, el estado civil o la educación y la situación laboral del jefe de hogar). f_c es un efecto fijo de conglomerado y se trata como un error además del término de error u^0_{hc} en la Ecuación 3.2, mientras que u^1_{hc} es el término de error de regresión estándar. Ambos u^0_{hc} y u^1_{hc}, sin embargo, incorporan cualquier error de medición en las participaciones presupuestarias y los valores unitarios, además de los no observables habituales.

La ecuación de valor unitario no contiene un efecto fijo de aldea porque, como observa Deaton,[8] "condicionados a los precios, los valores unitarios dependen únicamente de los efectos de calidad y los

errores de medición. La introducción de un efecto fijo adicional rompería el vínculo entre precios y valores unitarios, impediría que estos últimos dieran información útil sobre los primeros y, por lo tanto, eliminaría cualquier posibilidad de identificación" de los precios. Finalmente, $ln\pi_c$ son los precios no observados y, en consecuencia, las Ecuaciones 3.2 y 3.3 se estiman sin ellos, pero sus coeficientes se recuperan mediante las fórmulas contenidas en las Ecuaciones 3.8 y 3.9 a continuación. Como se discutió anteriormente, el modelo de Deaton no asume ninguna variación de precios dentro del conglomerado, ya que todos los hogares dentro del mismo conglomerado enfrentan el mismo precio y son encuestados al mismo tiempo. Por lo tanto, incluso si se observaran los precios, se habrían descartado en este paso de la regresión debido a la falta de variación.

La Ecuación 3.2, conocida como la ecuación del "valor unitario", verifica la presencia de efectos de calidad como se analiza en la Sección 3.2.2. Una relación positiva y estadísticamente significativa entre los gastos del hogar y los valores unitarios, después de tener en cuenta las características del hogar, sugeriría la presencia de efectos de calidad. Conocer el patrón de los efectos de la calidad (es decir, la magnitud de β^1) permite corregir las estimaciones finales de la elasticidad del precio para el matizado de la calidad como en el Paso 6. Tenga en cuenta que la Ecuación 3.2, a diferencia de la Ecuación 3.3, se estima sin los efectos fijos de conglomerados. Agregar un efecto fijo a nivel de conglomerado a la Ecuación 3.2 dificultaría la recuperación de los parámetros del modelo.

La Ecuación 3.3, por otro lado, es una ecuación de demanda estándar en la que la participación de los cigarrillos (una representación o proxy de la demanda) se expresa como en función de los ingresos del hogar (con el gasto del hogar como representación o proxy), las características del hogar y los precios. Debido a la suposición de que los precios se fijan dentro de los conglomerados y al hecho de que no hay datos de precios, los precios se representan mediante efectos fijos de conglomerados. La relación entre los dos errores, u^0_{hc} y u^1_{hc}, (capturado, por ejemplo, por la covarianza) es útil para corregir las estimaciones finales de la elasticidad precio por el error de medición, como se explica en el Paso 5.

Paso 4: Obtención de demanda a nivel de conglomerado y valores unitarios

El cuarto paso consiste en separar la demanda a nivel del hogar y los valores unitarios de los efectos del gasto del hogar y las características del hogar y luego promediar entre grupos. La separación y el promedio se realizan porque el interés principal es estimar la elasticidad a nivel de conglomerado utilizando la demanda del conglomerado y el valor unitario del conglomerado separados de todos los demás factores. Este paso requiere las siguientes ecuaciones:

$$\widehat{y_c^1} = \frac{1}{n_c^+}\sum_{h=1}^{n_c^+}(lnv_{hc} - \hat{\beta}^1 lnx_{hc} - \hat{\gamma} Z_{hc}) \quad (3.4)$$

$$\widehat{y_c^0} = \frac{1}{n_c}\sum_{h=1}^{n_c}(w_{hc} - \hat{\beta}^0 lnx_{hc} - \hat{\delta} Z_{hc}) \quad (3.5)$$

donde n_c es el número de hogares en el conglomerado c y n_c^+ es el número de hogares que declaran haber comprado el producto de tabaco para el que se estima la elasticidad. Observe que $\widehat{y_c^1}$ e $\widehat{y_c^0}$ no tienen el subíndice h porque representan promedios de conglomerados. $\widehat{y_c^1}$ e $\widehat{y_c^0}$ son las estimaciones del valor unitario promedio del conglomerado y la demanda promedio del conglomerado, respectivamente, después de eliminar los efectos del gasto del hogar y las características del hogar. En otras palabras, las Ecuaciones 3.4 y 3.5 pueden expresarse alternativamente como $y_c^1 = \alpha^1 + \psi ln\pi_c + u_c^1$ e $y_c^0 = \alpha + \theta ln\pi_c + f_c + u_c^0$, respectivamente.

Paso 5: Regresiones a nivel de conglomerados

Recuerde que el supuesto de identificación es que los precios varían entre conglomerados y no dentro de los conglomerados. Dado esto, las elasticidades precio de la demanda solo se pueden obtener observando

cómo responde la demanda a nivel de conglomerado a los cambios en los precios a nivel de conglomerado. Por lo tanto, el Paso 5 implica hacer una regresión de la demanda a nivel de conglomerado, $\widehat{y_c^0}$, sobre valores unitarios a nivel de conglomerado, $\widehat{y_c^1}$. El coeficiente de $\widehat{y_c^1}$ en tal regresión se puede obtener alternativamente dividiendo la covarianza entre $\widehat{y_c^0}$ e $\widehat{y_c^1}$ por la varianza de $\widehat{y_c^1}$. Eso es $\hat{\phi}$, la estimación del coeficiente de y_c^1, se obtiene por:

$$\hat{\phi} = \frac{Cov\left(\widehat{y_c^0}, \widehat{y_c^1}\right) - \frac{\widehat{\sigma^{10}}}{n_c}}{Var\left(\widehat{y_c^1}\right) - \frac{\widehat{\sigma^{11}}}{n_c^+}} \quad (3.6)$$

donde n_c^+ es el número de hogares en un conglomerado que reportan gastos positivos en tabaco y n_c es el número de hogares en un conglomerado; $\widehat{\sigma^{10}}$ es la estimación de la covarianza de los errores en las Ecuaciones 3.2 y 3.3; $\widehat{\sigma^{11}}$ es la varianza de los errores en la Ecuación 3.2. La ecuación 3.6 es una regresión estándar de errores en variables en la que se utiliza la covarianza y la varianza de los errores para corregir el error de medición. Observe que los factores de corrección para el error de medición se vuelven pequeños a medida que n_c^+ y n_c se agrandan.

Paso 6: Estimación de las elasticidades precio y gasto

El sexto y último paso en el método de Deaton aplica fórmulas de corrección de calidad para obtener la estimación de la elasticidad precio de la demanda, $\widehat{\varepsilon_p}$, como sigue:

$$\widehat{\varepsilon_P} = \left(\frac{\hat{\theta}}{\bar{w}}\right) - \hat{\psi} \quad (3.7)$$

donde $\bar{w}$ es la proporción media del gasto total de los hogares dedicada a los cigarrillos en la muestra. $\hat{\psi}$ y $\hat{\theta}$, las estimaciones de los coeficientes de los términos de precios no observados en las Ecuaciones (3.2) y (3.3) respectivamente, se obtienen de la siguiente manera:

$$\hat{\psi} = 1 - \frac{\hat{\beta}^1(\bar{w} - \hat{\theta})}{\hat{\beta}^0 + \bar{w}} \quad (3.8)$$

$$\hat{\theta} = \frac{\hat{\phi}}{1 + (\bar{w} - \hat{\phi})\hat{\zeta}} \quad (3.9)$$

$$\hat{\zeta} = \frac{\hat{\beta}^1}{\hat{\beta}^0 + \bar{w}(1 - \hat{\beta}^1)} \quad (3.10)$$

Finalmente, Deaton también propone la siguiente fórmula para obtener la estimación de la elasticidad gasto de la demanda, $\widehat{\varepsilon_I}$:

$$\widehat{\varepsilon_I} = 1 + \left(\frac{\hat{\beta}^0}{\bar{w}}\right) - \hat{\beta}^1 \quad (3.11)$$

donde $\hat{\beta}^1$ es la estimación del coeficiente del gasto total del hogar en la Ecuación 3.2, y $\hat{\beta}^0$ es la estimación del coeficiente del gasto total del hogar en la Ecuación 3.3. $\hat{\phi}$ es la estimación del coeficiente de una regresión de la demanda a nivel de conglomerado sobre el valor unitario a nivel de conglomerado (de la Ecuación 3.6). Una vez que se recuperan los parámetros de 3.8 a 3.10, la elasticidad precio de la demanda se puede estimar según la Ecuación 3.7. Por otro lado, la elasticidad gasto de la demanda solo usa coeficientes de primera etapa y puede derivarse usando la Ecuación 3.11. Dado que las fórmulas para la elasticidad precio de la demanda en la Ecuación 3.7 y para la elasticidad gasto de la demanda no son comandos directos de Stata, sus errores estándar deben obtenerse mediante bootstrapping.

Varios estudios han utilizado el método de Deaton para estimar las elasticidades de precio y gasto de la demanda de varios productos de tabaco en diferentes PIMB. Esos incluyen estudios en Albania,[38] Bangladesh,[68] Bosnia y Herzegovina,[81, 82] China,[71] Ecuador,[36] El Salvador,[45] India,[46, 68–71] Montenegro,[79, 80] Pakistán,[78] Serbia,[77, 78] Sudáfrica,[81] Uganda,[82] y Vietnam,[83] entre otros. Algunos estimaron la elasticidad de un solo bien, los cigarrillos, mientras que otros estimaron las elasticidades precio y precio cruzadas de los cigarrillos y algunos otros productos de tabaco.

También se debe tener en cuenta que, si bien algunos de esos estudios consideraron todos los hogares en la regresión de participación presupuestaria para estimar la elasticidad, algunos consideraron solo los hogares con compras positivas en la regresión de participación presupuestaria, estimando así solo una demanda condicional. Sin embargo, como señala Deaton,[8] para el análisis de la reforma tributaria y de precios es necesario incluir a todos los hogares en el análisis, ya sea que compren o no. Por lo tanto, si la elasticidad cantidad-precio se estima solo con hogares con compras positivas (como se muestra en este conjunto de herramientas), complementarla con una estimación de la elasticidad de prevalencia que incluya a todos los hogares proporcionaría una estimación de la elasticidad precio total.

Las estimaciones de la elasticidad precio de los cigarrillos en esos estudios oscilaron entre -0,1 y -1,4, aunque la mayoría de los estudios tenían coeficientes de elasticidad de 0,8 o inferiores en valor absoluto, mientras que las estimaciones de la elasticidad gasto oscilaron entre 0,2 y 2,4. En otras palabras, esos estudios tienden a encontrar estimaciones de elasticidad precio de cigarrillos comparables a las estimadas en la literatura internacional utilizando otros métodos. También tienden a encontrar elasticidades de gasto no negativas de la demanda de cigarrillos, lo que implica que la demanda de cigarrillos no disminuye con un aumento en el gasto.

También es interesante notar que la definición de conglomerado utilizada en esos estudios varía. Mientras que algunos consideraban un pueblo o bloque urbano como el conglomerado predeterminado, otros consideraban un distrito en sí mismo como un conglomerado. También es posible definir un conglomerado sobre variables geográficas y temporales[77, 78, 84] si, por ejemplo, hay EGH de múltiples rondas u olas. Es importante entender que las propiedades de consistencia de los parámetros en el modelo de Deaton dependen de la cantidad de conglomerados (y no de la cantidad de hogares), ya que esos parámetros se derivan de datos promedio a nivel de conglomerados.

Por otro lado, los errores de medición en las Ecuaciones 3.2 y 3.3 tienden a cero solo a medida que aumenta el número de hogares en cada conglomerado. Claramente, hay un trade-off. Por un lado, los conglomerados de tamaños pequeños aumentan la probabilidad de que aumenten los errores de medición, lo que es especialmente cierto en el caso de productos como el tabaco, que solo son consumidos por unos pocos hogares. Con conglomerados más pequeños, también es posible que algunos de ellos no tengan ningún hogar con compras positivas de tabaco. Por otro lado, dado que la regresión de segunda etapa y la estimación de la elasticidad precio dependen de tener un gran número de conglomerados con compras positivas, es importante tener tantos conglomerados como sea posible para obtener estimaciones de parámetros coherentes.

Los propios experimentos de Deaton han demostrado que el estimador funciona adecuadamente incluso cuando hay tan solo dos hogares en cada grupo.[8] Según Deaton, “aumentar el número de pueblos o grupos es mucho más importante que aumentar el número de observaciones en cada uno”. Eso se debe a dos razones: (i) el modelo corrige los errores de medición pero no puede garantizar la consistencia de los parámetros con un pequeño número de conglomerados, y (ii) si los conglomerados se definen o se agregan en áreas geográficas más grandes, entonces los hogares dentro de tales conglomerados puede que no estén en el mismo mercado y, como resultado, puede haber variaciones reales intrarregionales en los valores unitarios dentro de esos conglomerados que, sin darse cuenta, pueden tratarse como errores de medición.

Para que los supuestos del modelo se mantengan, los hogares en un conglomerado determinado deben tener proximidad geográfica y enfrentar las entrevistas más o menos al mismo tiempo. Eso puede volverse aún más difícil a medida que los conglomerados se expanden para incluir regiones geográficas más grandes.

Para la mayoría de las EGH, los conglomerados se dan naturalmente como parte del diseño de la encuesta, como ya se señaló en el Capítulo 2. También es importante señalar que el modelo se basa en la existencia de una variación genuina de los precios entre los conglomerados y requiere que dicha variación sea exógena al proceso que determina la demanda. Como observa Deaton,[8] "si los precios locales están determinados por los precios mundiales, los impuestos fronterizos y los costos de transporte, los supuestos se cumplirán porque la demanda local no tiene efecto sobre los precios". Por otro lado, si los precios del pueblo dependen de la demanda dentro del pueblo, las estimaciones de los parámetros no serán consistentes, por las razones usuales de simultaneidad.

Vale la pena señalar que, aunque la discusión anterior se refiere a los hogares, el análisis también se puede realizar a nivel de individuos. Sin embargo, eso requiere que el investigador tenga acceso a una rica encuesta de gastos con datos recopilados a nivel individual. Por ejemplo, dicha encuesta debe contener información sobre los patrones de gasto (cantidad y monto total gastado) y sobre los productos de tabaco por parte de los individuos (no agregados a nivel de hogar, como suele ser el caso).

Además, también deben estar presentes otros datos sociales y demográficos a nivel individual. Si bien dichos conjuntos de datos están ampliamente disponibles en los PIA, tienden a ser la excepción en los PIMB. Se alienta a los investigadores con acceso a encuestas de gasto recopiladas a nivel individual a utilizar el método de Deaton para la estimación de las elasticidades de la demanda.

El método de Deaton no está exento de críticas. Gibson y Rozelle (2005)[85] muestran que el uso de valores unitarios como sustituto de los precios reales produce estimaciones sesgadas de la elasticidad precio de la demanda incluso después de corregir los efectos de calidad y el error de medición. McKelvey (2011)[84] muestra que el método de Deaton no aborda adecuadamente el problema del matizado de la calidad, que parece prevalecer en muchos entornos. A pesar de esas limitaciones, en ausencia de datos de precios muy detallados, el método de Deaton sigue siendo una de las formas más efectivas de obtener elasticidades.

3.3.2 Preparación de datos para estimar elasticidades de cantidad

Una vez que se extrajeron y limpiaron los datos, se fusionaron diferentes conjuntos de datos y se manejaron los datos según sea necesario, como se detalla en el Capítulo 2, se requieren detalles específicos sobre las variables necesarias para estimar la elasticidad precio utilizando el método de Deaton discutido anteriormente. Toda variable nueva discutida aquí, es importante pasarla por todos los procesos discutidos en el Capítulo 2. Esta sección analiza cómo las variables específicas que se requieren para estimar las elasticidades precio y cruzadas, usando el método de Deaton, pueden generarse usando las variables estándar disponibles de EGH.

Las variables más importantes son la cantidad de consumo y los gastos en diferentes productos de tabaco. Ellas están disponibles directamente en la mayoría de las EGH. Algunas EGH pueden no reportar información de cantidad, como se mencionó anteriormente. En tales casos, la discusión de aquí puede no ser relevante.

Primero, se deben crear valores unitarios para cada uno de los productos de tabaco para los que hay datos disponibles. Eso puede incluir valores unitarios para cigarrillos, bidis y productos sin humo, entre otros. Por ejemplo, la cantidad de cigarrillos (ya sea en cajetillas o en cantidad de cigarrillos) descargada de los datos de EGH tiene el nombre de variable *qcig*, y la variable que representa el gasto gastado en cigarrillos es

expcig. Luego, el valor unitario de los cigarrillos (*uvcig*) se puede generar usando el comando <*gen uvcig=expcig/qcig*>.

El modelo de Deaton utiliza el logaritmo natural de la variable de valor unitario como variable dependiente (*luvcig*). Use el comando <*gen luvcig=ln(uvcig)*> para generarlo. De manera similar, se debe construir una variable para representar participación en el presupuesto de los cigarrillos (*bscig*) usando el comando <*gen bscig = expcig/exptotal*>, donde *exptotal* es el gasto total en todos los artículos. Para aquellos hogares que no reportaron gastos en cigarrillos, eso generaría un valor faltante. Al implementar el modelo de Deaton, *uvcig* y *bscig* serán las variables dependientes en las respectivas regresiones. Se deben generar variables similares de valor unitario y participación en el presupuesto para otros productos de tabaco de EGH que se incluirán en la estimación de la elasticidad del precio.

El precio es definitivamente la variable independiente para usar en un modelo que estima funciones de demanda. Sin embargo, como se señaló anteriormente, el método de Deaton se utiliza en los casos en que no se dispone de información directa sobre precios. En cambio, la variación de precios se captura por medio de las variaciones de precios a nivel de grupo en EGH. Por lo tanto, es crucial contar con una variable que identifique conglomerados (*clust*) o unidades primarias de muestreo. Esa variable suele estar disponible directamente desde la EGH o puede generarse utilizando otras variables disponibles que identifiquen las unidades primarias de muestreo, como se analiza en el Capítulo 2. El conglomerado puede ser una unidad geográfica (como un pueblo o una unidad primaria de muestreo en una encuesta transversal) como en el análisis original de Deaton, o puede ser un punto en el tiempo (como una ola de encuesta) si se combinan diferentes rondas de encuestas o una combinación de UPM y ola de encuesta.[84]

A veces, las encuestas no tienen datos sobre la cantidad de compra, pero hay datos sobre el gasto de un artículo en particular en un hogar y viceversa. Dichos hogares pueden eliminarse del análisis mediante el comando < *drop if [qcig==.&expcig!=.]|[qcig!=.&expcig==.]* >. Después de esto, cualquier variable *bscig* que aún falte se puede reemplazar con ceros.

Dado que es importante que cada conglomerado tenga al menos dos hogares, como se mencionó en la sección anterior, los conglomerados que tienen menos de dos hogares deben eliminarse de la estimación de las elasticidades cuantitativas. Eso se hace creando una variable a nivel de conglomerado que contiene el número de hogares que consumen tabaco (*cigclustsize*).

```
gen dcig=0 if qcig==. | qcig==0
replace dcig=1 if qcig>0 & qcig!=.
bys clust: egen cigclustsize =sum(dcig)
drop if cigclustsize <2
```

Además, es necesario identificar variables específicas a nivel de hogar para usar como variables independientes en el modelo. La literatura ofrece orientación sobre algunas de las variables sociodemográficas comunes a nivel del hogar: logaritmo del tamaño del hogar; proporción de hombres (relación entre el número de hombres y el tamaño del hogar); edad promedio del hogar; educación promedio (educación total recibida por todos los miembros en años dividida por el tamaño del hogar) del hogar; educación máxima (años de educación recibidos por el miembro más educado del hogar); nivel educativo del hogar o del jefe de hogar; variables dummy para caracterizar los hogares en diferentes grupos sociales, étnicos, ocupacionales, religiosos y de ingresos; y variables dummy para indicar la ubicación del hogar (como zonas rurales/urbanas, provincia, región o distrito), entre otras.

3.3.3 Estimación de la elasticidad precio en el margen intensivo con Stata

Esta sección proporciona el código de Stata para la estimación de la elasticidad precio para un solo producto de tabaco (cigarrillos) usando el método de Deaton discutido anteriormente. Deaton proporciona un código de Stata detallado para estimar las elasticidades de precio y precio cruzado para diferentes productos, que se puede descargar de http://web.worldbank.org/archive/website00002/WEB/EX5_1-2.HTM. El Apéndice de Código en la Sección 7.3 proporciona una versión modificada del código de Deaton del sitio web del Banco Mundial, con algunas explicaciones adicionales para que los lectores las sigan.

El código utilizado en esta sección para estimar la elasticidad precio de los cigarrillos produciría parámetros estimados de elasticidad idénticos a los del código de Deaton para casos de múltiples bienes en el Apéndice 7.3 cuando se utiliza para estimar la elasticidad de un solo bien. Mientras que el código para casos de múltiples bienes hace uso de matrices para calcular varios parámetros en el modelo, el código aquí usa solo escalares, ya que es un solo producto. Además, dado que el código para múltiples bienes también estima elasticidades precio cruzadas y permite la introducción de otras restricciones teóricas en el sistema de demanda, como se analiza en Deaton,[8] el código aquí simplemente estima la elasticidad precio de los cigarrillos sin imponer ninguna otra restricción.

El código de esta sección usa las variables *bscig, luvcig, lexp, lhsize, maleratio, meanedu, maxedu, sgp1, sgp2, sgp3* para la estimación de las elasticidades precio, donde *bscig* se refiere a la participación en el presupuesto de cigarrillos en el presupuesto total del hogar, *luvcig* es el logaritmo natural del valor unitario de los cigarrillos, *lexp* es el logaritmo natural de los gastos mensuales del hogar, *lhsize* es el logaritmo natural del tamaño del hogar, *maleratio* es la relación entre el número de hombres y el número de miembros del hogar, *meanedu* es la educación media de los miembros del hogar en años, *maxedu* es la educación máxima recibida en años por cualquiera de los miembros del hogar, y *sgp1* a *sgp3* son los grupos sociales definidos como grupos socioeconómicos o grupos de clase, agrupado según sea adecuado para cada país.

Dado que las elasticidades de cantidad se estiman solo para los hogares que informan sobre el consumo de tabaco, todos los demás hogares se eliminan primero de los datos mediante el comando *< keep if dcig==1 >*.

Pruebas de variación espacial en valores unitarios

Como se indica en la Sección 3.3.1 del método, es conveniente estimar la variación en los valores unitarios entre los conglomerados para evaluar si las variaciones en los valores unitarios son indicativas de la variación en los precios entre los conglomerados. Eso se puede hacer usando el comando *<anova luvcig clust>* o *<regress luvcig i.clust>*. El R^2 y el estadístico *F* de la salida pueden indicar la utilidad de los valores unitarios como informativos de los precios. Según Deaton,[8] un estadístico *F* significativo y un valor de R^2 de alrededor de 0,5 (es decir, las variables ficticias de conglomerados explican aproximadamente la mitad de la variación total de los valores unitarios) significa que los valores unitarios se pueden utilizar con el fin de examinar la variación de precios y estimar las elasticidades de precios.

Estimación de regresiones de primera etapa dentro de un conglomerado y varianzas de errores de medición

A continuación, se estiman las Ecuaciones 3.2 y 3.3 y se almacenan los parámetros relevantes para las etapas posteriores:

```
#delimit;
areg luvcig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust);
scalar sigma11=$S_E_sse / $S_E_tdf;
scalar b1=_coef[lexp];
predict ruvcig, resid;
gen y1cig=luvcig-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio
            -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu
            -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3;

*Repeat for budget shares
areg bscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust);
predict rbscig, resid;
scalar sigma22=$S_E_sse/$S_E_tdf;
scalar b0=_coef[lexp];
gen y0cig=bscig-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio
            -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu
            -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3;

qui areg ruvcig rbscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
scalar sigma12=_coef[rbscig]*sigma22
```

El comando <*areg*> se usa en lugar de <*regress*> ya que se trata de una regresión lineal con un gran conjunto de variables dummy. El comando incluye implícitamente una variable dummy para cada conglomerado menos uno, pero no entrega los coeficientes asociados con esas variables dummy de conglomerado en la salida de la regresión. La opción <*absorb(clust)*> junto con el comando <*areg*> le dice a Stata que use dummies de conglomerados implícitas para la variable de conglomerado clust.

Las variables *y1cig* e *y0cig* después de cada regresión eliminan cualquier efecto de las características específicas del hogar que expliquen la variación de la calidad en los valores unitarios. Esas variables ahora conservan la información de precios contenida en las variables dummy de conglomerados. Los residuos del valor unitario (*ruvcig*) y la regresión de participación presupuestaria (*rbscig*) se generan para usarse en la última regresión de *ruvcig* en *rbscig* para construir el escalar *sigma12*. Este *sigma12* junto con los escalares *sigma11* y *sigma22*, generados después de la regresión de valor unitario y participación presupuestaria, son estimaciones de la varianza y covarianza de los errores de medición que se usarán para la corrección del error de medición en la Ecuación 3.6. El coeficiente para el logaritmo del gasto también se almacena para su uso posterior. El escalar *b1*, que es el coeficiente del logaritmo del gasto en la regresión de valor unitario, es la estimación de la elasticidad calidad. Cuanto menor sea ese número, menor será el matizado de la calidad en valores unitarios.

Estimación de elasticidades de ingresos o gastos

La elasticidad gasto total (o elasticidad ingreso) en la Ecuación 3.11 se puede estimar después de las regresiones de primera etapa usando los resultados guardados. Eso se puede hacer usando el código:

```
qui sum bscig
scalar Wbar=r(mean)
scalar Expel=1-b1+(b0/Wbar)
scalar list Expel
```

El código almacena primero la estimación de la participación presupuestaria promedio en un escalar (*Wbar*) y usa los otros escalares guardados (*b1* y *b0*) de las regresiones de la primera etapa para estimar la elasticidad gasto (*Expel*). La última línea imprimirá la elasticidad gasto en la ventana de resultados de Stata. Para estimar los errores estándar del coeficiente de elasticidad gasto se utiliza un método bootstrap con el siguiente código:

```
cap program drop Expelast
program Expelast, rclass
tempname b1 b0 Wbar
qui areg luvcig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
cap scalar b1=_coef[lexp]
qui areg bscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
cap scalar b0=_coef[lexp]
qui sum bscig
cap scalar Wbar=r(mean)
return scalar Expel=1-b1+(b0/Wbar)
end
Expelast
return list
bootstrap Expel=r(Expel), reps(1000) seed(1): Expelast
```

El código devuelve el coeficiente de elasticidad gasto junto con los errores estándar del bootstrap.

Preparación de datos para la regresión entre conglomerados

El siguiente paso consiste en promediar las variables *y1cig* e *y0cig* por conglomerados para generar *y1c* e *y0c* respectivamente, de modo que puedan usarse para una regresión entre conglomerados de *y0c* sobre *y1c* para obtener la elasticidad precio. Como se mencionó anteriormente, las variables *y1cig* e *y0cig* eliminan de cualquier característica específica del hogar de las regresiones de valor unitario y participación presupuestaria y contienen solo la información de precios en variables dummy de conglomerados, así como los errores de medición.

```
sort clust
egen y0c= mean(y0cig), by(clust)
egen n0c=count(y0cig), by(clust)
egen y1c= mean(y1cig), by(clust)
egen n1c=count(y1cig), by(clust)
sort clust
qui by clust: keep if _n==1
```

Después de generar un valor promedio para todos los hogares en cada conglomerado, solo se debe conservar una observación por conglomerado para el análisis restante. Junto con la generación de las variables a nivel de conglomerado *y0c* e *y1c*, se generan otras dos variables a nivel de conglomerado (*n0c* y *n1c*), que indican el tamaño o la cantidad de todos los hogares en cada conglomerado (*n0c*) y la cantidad de hogares que informan compras positivas en cada conglomerado (*n1c*). Con esos, se estima el tamaño promedio de conglomerado para todos los hogares (*n0*) y el tamaño promedio de conglomerado para

hogares con consumo positivo de cigarrillos (*n1*). Eso se puede hacer usando el siguiente código. Deaton usa la media armónica para estimar los tamaños promedio de conglomerados.

```
ameans n0c
scalar n0=r(mean_h)
ameans n1c
scalar n1=r(mean_h)
drop n0c n1c
```

Regresión entre conglomerados

La regresión entre conglomerados de *y0c* sobre *y1c* produce la estimación del ratio $\phi=\theta/\psi$, cuyo numerador y denominador son los coeficientes de los precios no observados en las Ecuaciones 3.3 y 3.2, respectivamente. En lugar de hacer la regresión real, el parámetro híbrido se puede estimar utilizando un estimador de errores en variable en la Ecuación 3.6, para el cual las estimaciones de *y1* e *y0*, así como las varianzas y covarianzas del error de medición estimadas a partir de las regresiones de primera etapa son utilizados. La Ecuación 3.6 se estima utilizando el siguiente código:

```
qui corr y0c y1c, cov
scalar S=r(Var_2)
scalar R=r(cov_12)
scalar num=scalarI-(sigma12/n0)
scalar den=scalar(S)-(sigma11/n1)
cap scalar phi=num/den
```

Estimación de la elasticidad precio

Una vez que la relación ϕ se estima, como en la Ecuación 3.6, es necesario definir algunos escalares más para estimar la elasticidad precio real. Esto se hace en el siguiente código:

```
cap scalar zeta= b1/((b0 + Wbar*(1-b1)))
cap scalar theta=phi/(1+(Wbar-phi)*zeta)
cap scalar psi=1-((b1*(Wbar-theta))/(b0+Wbar))
return scalar EP=(theta/Wbar)-psi
scalar list EP
```

La última línea del código mostrará la estimación de la elasticidad precio en la pantalla de resultados de Stata. Los otros escalares definidos anteriormente son estimaciones para las Ecuaciones 3.8 a 3.10, no necesariamente en el mismo orden. Para estimar los errores estándar para las estimaciones de la elasticidad precio, las ecuaciones anteriores deben ir a un programa que use el siguiente código:

```
cap program drop elast
program elast, rclass
tempname S R num den phi theta psi
qui corr y0c y1c, cov
scalar S=r(Var_2)
scalar R=r(cov_12)
scalar num=scaI(R)-(sigma12/n0)
scalar den=scalar(S)-(sigma11/n1)
cap scalar phi=num/den
cap scalar zeta= b1/((b0 + Wbar*(1-b1)))
cap scalar theta=phi/(1+(Wbar-phi)*zeta)
cap scalar psi=1-((b1*(Wbar-theta))/(b0+Wbar))
return scalar EP=(theta/Wbar)-psi
end
elast
return list
bootstrap EP=r(EP), reps(1000) seed(1): elast
```

La última línea de código devuelve los errores estándar obtenidos mediante bootstrap para las estimaciones de la elasticidad precio. La sección 7.1 del Apéndice de código incluye un do-file de ejemplo que detalla el código utilizado en esta sección. Los usuarios pueden copiar y pegar ese código en el editor de do-files de Stata y estimar los resultados con los datos/variables correspondientes que se describen allí. Además, la Sección 7.3 reproduce un código detallado de Deaton para estimar las elasticidades precio y precio cruzado utilizando el método de Deaton.

3.3.4 Estudio de caso

Esta sección presenta los resultados de la estimación econométrica de la elasticidad precio para un solo producto básico (cigarrillos) utilizando datos de EGH hipotéticos, aplicando el método de Deaton descrito anteriormente. Los resultados se presentan paso a paso para facilitar la comprensión de la técnica descrita en la sección anterior.

Paso 1: Obtención de valores unitarios y otras variables relevantes

El primer paso en el método de Deaton es obtener los valores unitarios mediante la Ecuación 3.1. En segundo lugar, otras variables utilizadas en el análisis se procesan como se describe en el Capítulo 2. La lista completa de variables que se utilizan para estimar las elasticidades se presenta en la Tabla 3.1 a continuación. Las variables en las líneas 5 a 11 de la tabla 3.1 conforman el vector Z_{ic} de estructura del hogar y variables de control demográfico descrito en las Ecuaciones 3.2 y 3.3.

Paso 2: Pruebas de variación espacial

El segundo paso en el método de Deaton es verificar empíricamente que los valores unitarios satisfacen la hipótesis de variación espacial usando ANOVA. Los resultados del ejercicio ANOVA se encuentran en la Tabla 3.2 a continuación.

Tabla 3.1 Variables utilizadas para la estimación de la elasticidad precio

Variable	Descripción	Obs	Promedio	SD	Mín	Máx
qcig	Número de cigarrillos comprados	9.695	21,58	16,64	1,00	232,50
expcig	Gastos incurridos en cigarrillos	9.695	5.314,34	3.803,90	52,00	45.680,00
luvcig	Logaritmo natural del valor unitario del cigarrillo	9.695	5,53	0,35	3,95	6,49
bscig	Participación del gasto en cigarrillos en el gasto total	9.695	0,09	0,06	0,00	0,51
lexp	Logaritmo del gasto total del hogar	9.695	11,02	0,56	7,79	13,49
lhsize	Logaritmo del tamaño del hogar	9.695	1,07	0,56	0,00	3,00
maleratio	Proporción de hombres en el hogar	9.695	0,51	0,25	0,00	1,00
meanedu	Educación media del hogar	9.694	10,73	2,36	2,00	20,00
maxedu	Educación más alta de cualquiera de los miembros del hogar	9.694	12,01	2,55	2,00	20,00
sgroup	Variable dummy para grupo social	9.695	2,29	0,94	0,00	3,00

El resultado del ejercicio ANOVA muestra que al menos el 87 por ciento (*R*-cuadrado de 0,87) de la variación en los valores unitarios se explica por los efectos entre grupos. El estadístico F está asociado con la hipótesis de que no hay variación espacial en los precios y aquí se rechaza. Como se mencionó anteriormente, un estadístico *F* significativo y un valor de R^2 de alrededor de 0,5 significa que los valores unitarios se pueden usar para examinar la variación de precios y estimar las elasticidades precio.

Tabla 3.2 Prueba de la variación espacial en los logaritmos de los valores unitarios

Estadístico *F*	valor *p*	*R*-cuadrado	*R*-cuadrado ajustado	Observaciones
30,24	0,000	0,871	0,842	9.695

Notas: El estadístico *F* y el valor *p* están asociados con la hipótesis nula de que no hay variación espacial en los valores unitarios. El *R*-cuadrado mide la proporción de variación de precios que tiene lugar entre conglomerados. *n* es el número total de hogares.

Paso 3: Regresiones dentro de un conglomerado en la primera etapa

El siguiente paso es estimar las regresiones dentro del conglomerado, es decir, la regresión de valor unitario y las regresiones de participación presupuestaria, según las Ecuaciones 3.2 y 3.3. Los resultados de esas regresiones se muestran en la Tabla 3.3.

Tabla 3.3 Resultados de la regresión de valor unitario

VARIABLES	Regresión de valor unitario	Regresión de la participación en el presupuesto
lexp	0,0855***	-0,0348***
	(0,00384)	(0,00154)
lhsize	-0,0432***	-0,00848***
	(0,00385)	(0,00154)
maleratio	-0,0159***	0,0216***
	(0,00607)	(0,00243)
meanedu	0,00451***	-0,00148***
	(0,00143)	(0,000573)
maxedu	-0,000244	-0,000719
	(0,00132)	(0,000527)
sgp1	0,00204	0,00850***
	(0,00785)	(0,00314)
sgp2	-0,00808*	-0,00476***
	(0,00447)	(0,00179)
sgp3	-0,00265	-0,000609
	(0,00435)	(0,00174)
Constante	4,602***	0,493***
	(0,0388)	(0,0155)
Observaciones	9.694	9.694
R-cuadrado	0,883	0,377

Nota: Errores estándar entre paréntesis; los coeficientes de efectos fijos de conglomerados se suprimen por razones de espacio, pero en conjunto son estadísticamente significativos. *** p<0,01, ** p<0,05, * p<0,1.

El coeficiente de *lexp* en la regresión de valor unitario es la elasticidad calidad o la elasticidad gasto de la calidad, como se analiza en la Sección 3.3.1. Es estadísticamente significativo al nivel del uno por ciento, lo que implica que el matizado de calidad es significativo, aunque la magnitud en sí misma sea pequeña y quizás insignificante. Los resultados de la regresión de la participación en el presupuesto muestran que la participación en el presupuesto de cigarrillos disminuye con el gasto del hogar. Los hogares con gastos totales más altos suelen ser los de los grupos de ingresos más altos. Tienden a asignar una parte relativamente menor de su presupuesto a la compra de cigarrillos, y el resultado aquí es consistente con eso. Ese resultado es estadísticamente significativo al nivel del uno por ciento. Los coeficientes estimados de las restantes variables específicas del hogar se utilizan para eliminar sus efectos, si los hubiere, sobre las variables de valor unitario y participación en el presupuesto, de modo que se pueda suponer que cualquier variación residual en ellas refleja los errores de medición.

Paso 4 y Paso 5

El paso 4 consiste en obtener el valor unitario a nivel de conglomerado y la demanda a nivel de conglomerado según las Ecuaciones 3.4 y 3.5. El paso 5 es entonces una regresión de la demanda a nivel de conglomerado sobre el valor unitario a nivel de conglomerado según la Ecuación 3.6. Esos resultados no se informan aquí.

Paso 6: Obtención de estimaciones de elasticidad

El paso final aplica las fórmulas de las Ecuaciones 3.7 a 3.11 para obtener estimaciones de la elasticidad precio y gasto. La Tabla 3.6 presenta estimaciones de la elasticidad precio de la demanda de cigarrillos en Uganda. La Tabla 3.7 presenta estimaciones de la elasticidad gasto de la demanda.

Los resultados de la Tabla 3.4 muestran que se espera que la demanda de cigarrillos en este ejemplo aumente alrededor de un 5,2 % cada vez que los ingresos/gastos del hogar aumenten un 10%, como lo indica un coeficiente significativo de elasticidad ingresos/gastos de 0,52. De manera similar, la estimación de la elasticidad del precio de -0,8 indica que por cada aumento del 10 por ciento en el precio de los cigarrillos, se espera que el consumo de cigarrillos en los hogares disminuya en un ocho por ciento. Esas estimaciones están dentro del rango de estimaciones en la literatura que usa el método de Deaton discutido en la Sección 3.3.1.

Tabla 3.4 Estimaciones de la elasticidad ingreso y precio de la demanda de cigarrillos

	Elasticidad gasto	Elasticidad precio total
Coeficiente de elasticidad	0,515***	-0,795***
Intervalo de confianza del 95%	[0,4762283 0,5539125]	[-0,8371711 -0,7528599]
Error estándar de bootstrap	0,0198	0,0215

Nota: Los errores estándar de bootstrap se calcularon haciendo 1000 replicaciones. Asumiendo que las estimaciones siguen una distribución normal, los coeficientes con *** y ** implican niveles de significancia del 1% y 5%, respectivamente.

3.4 Estimación de la elasticidad de prevalencia

A diferencia de las bienes normales, donde la elasticidad precio en el margen intensivo es más importante para fines de modelado tributario, tanto la cantidad de consumo como la prevalencia son importantes cuando se trata de productos de desmerecimiento como el tabaco, que son utilizados solo por una fracción relativamente pequeña de personas. Dado que la prevalencia mundial del tabaquismo en adultos es del 19,6 %,[86] es importante comprender que los resultados como el tabaquismo o el consumo de tabaco (y_i) tienen dos propiedades estadísticas fundamentales[87]:

$$1) \quad y_i \geq 0 \text{ para } i = 1 \ldots . n_1 \quad (3.12)$$

$$2) \quad y_j = 0 \text{ para } j = n_{1+1} \ldots . n_2 \quad (3.13)$$

Es decir, en una distribución de la población, hay n_1 número de personas que fuman cigarrillos en cantidades mayores o iguales a cero y n_{1+1} a n_2 número de personas que no fuman en absoluto. En otras palabras, la distribución acumulada del consumo de cigarrillos se puede caracterizar como una distribución mixta que no es ni discreta ni continua. Si los resultados cero son lo suficientemente grandes, como es obvio a partir de la prevalencia global de tabaquismo del 19,6 por ciento, no se pueden ignorar al modelar empíricamente los

resultados del tabaquismo. Dado que la estimación de la elasticidad de la prevalencia se realiza utilizando los datos de EGH, cabe señalar que las elasticidades de prevalencia estimadas son para los hogares y no para las personas dentro de los hogares.

Así como la elasticidad cantidad se estimó a nivel de hogar en la sección anterior, la elasticidad prevalencia también se estima para el hogar como unidad de análisis. Dado que un hogar consta de personas fumadoras y no fumadoras, las estimaciones de la elasticidad de la prevalencia a nivel del hogar pueden ser menores o iguales a estimaciones similares para individuos. Por ejemplo, si hay dos fumadores en un hogar y solo uno deja de fumar, la elasticidad prevalencia estimada con base en los datos de EGH no captaría eso, mientras que la elasticidad prevalencia estimada con datos a nivel individual sí lo haría.

Ha habido varias estrategias econométricas para modelar variables de resultado con un gran número de ceros y magnitudes positivas en función de un conjunto de covariables exógenas, (x). El enfoque tradicional ha sido utilizar el método de mínimos cuadrados ordinarios (MCO), que trata las cantidades brutas positivas de consumo de cigarrillos como la variable dependiente y_i (resultado). Este método ignora los ceros por completo. Una desventaja importante de ese enfoque es la posibilidad de predecir un consumo negativo a partir del modelo econométrico. Ignorar todos los ceros también significa que la estimación de MCO es ineficiente y, en algunos casos, sesgada.[88] El uso de MCO con una transformación logarítmica de la variable y_i de resultado mitiga parcialmente algunos de esos problemas analíticos. Pero todavía no puede incorporar los ceros, y estos ceros no se pueden transformar en logaritmos. Además, el tabaquismo pronosticado del modelo transformado logarítmicamente aún se vería afectado por lo que se denomina sesgo de transformación.[89, 90]

La presencia de una proporción sustancial de ceros, como en el caso del tabaquismo o el consumo de tabaco, en los datos normalmente se ha manejado usando un modelo de dos partes (2PM), que distingue entre un indicador binario usado para modelar la probabilidad de fumar y un modelo de regresión condicional para el resultado positivo, es decir, en ese caso, la decisión de fumar.[91–93] Si bien existen enfoques econométricos alternativos para manejar una gran cantidad de resultados cero (modelos tobit, modelo de datos de conteo y modelos de Poisson cero inflado, por nombrar algunos), las 2PM se han utilizado ampliamente en el contexto de la modelización de la demanda de cigarrillos o tabaco.[94–99] Una buena exposición de muchos de esos métodos alternativos se puede encontrar en Camerone y Trivedi[16] y así como muchos otros libros de texto econométricos. La primera parte del 2PM utiliza la muestra completa, incluidos los ceros, y estima la probabilidad de observar resultados positivos frente a cero. La segunda parte utiliza una submuestra que se compone únicamente de aquellos que reportan consumos positivos, la cual se estima con un modelo econométrico para variables continuas como MCO o modelo lineal generalizado (GLM).[100] Ambas partes en el 2PM se estiman de forma independiente, lo que permite la independencia entre la decisión de fumar y la decisión de cuánto fumar. Sin embargo, en este conjunto de herramientas, en lugar de estimar 2PM de la manera tradicional, la segunda parte de 2PM se reemplaza con el modelo de Deaton discutido anteriormente, por las razones explicadas en la Sección 3.4.1.

En la primera parte, el objetivo es estimar la elasticidad precio de la participación en el tabaquismo o la prevalencia del tabaquismo. Estima la probabilidad de que una persona fume utilizando un modelo de probabilidad binario paramétrico, como logit o probit.[101] En otras palabras, este modelo estima si el precio del tabaco impacta o no la decisión de un individuo de consumir tabaco, condicionado a un conjunto de variables independientes. Basado en Cameron y Trivedi,[16] a continuación se proporciona una breve introducción a los modelos probit y logit.

Dado que son modelos de elección binaria, la variable dependiente puede tomar el valor de 1 si un individuo tiene un consumo de tabaco positivo y cero en caso contrario, cada uno tiene sus respectivas probabilidades. Eso es,

$$y = \begin{cases} 1 & con\ probabilidad \quad p \\ 0 & con\ probabilidad \quad 1-p \end{cases} \quad (3.14)$$

La función de masa de probabilidad para el resultado observado *y* es $p^y\ (1\text{-}p)^{1-y}$ with $E(y)=p$ y $Var(y)=p(1-p)$. Se puede formar un modelo de regresión a partir de esto parametrizando *p* para depender de una función índice $X'\beta$, donde *X* es un vector de variables que incluyen el precio del tabaco, los gastos de consumo total de una persona o un hogar (que es un indicador de sus ingresos) y otras covariables utilizadas como variables de control. β es un vector de parámetros desconocidos a estimar. La probabilidad condicional toma la forma $p_i=Pr(y_i=1|X)=F(X_i'\beta)$, donde $F(\cdot)$ es una función paramétrica especificada de $X_i'\beta$ asegurando los límites $0\leq p\leq 1$. Es la elección de la función de enlace $F(\cdot)$ que distingue entre modelos logit y probit, como se muestra en la Tabla 3.5.

Para los modelos probit y logit, $F(z)\rightarrow 0$ cuando $z\rightarrow -\infty$ y $F(z)\rightarrow 1$ cuando $z\rightarrow +\infty$. También, $f(z)=\partial F(z)/\partial z$ es positiva, ya que $\partial F(z)$ es estrictamente creciente. Si $y_i^*= x_i'\beta+u_i$ es una función no observable, el modelo probit es el indicado cuando u_i tiene una distribución normal, y el modelo logit lo es cuando tiene una distribución logística.

Tabla 3.5 Modelos de resultados binarios

Modelo	Probabilidad $Pr(y=1\|X)$	Efecto Marginal $\partial p/(\partial x_j$
Logit	$\Lambda(X'\beta) = e^{X'\beta}/(1+e^{X'\beta})$	$\Lambda(X'\beta)\{1-\Lambda(X'\beta)\}\beta_j$
Probit	$\Phi(X'\beta) = \int_{-\infty}^{X'\beta} \phi(z)dz$	$\phi(X'\beta)\beta_j$

Nota: Tabla adaptada de Cameron & Trivedi (2010)[16].

Los investigadores deben tener cuidado con la interpretación de los resultados, ya que los coeficientes estimados no representan los efectos marginales y no tienen una interpretación clara. En los modelos de elección binaria, los efectos marginales no son constantes sino una función de todas las variables explicativas utilizadas en el modelo. Los efectos marginales para los modelos logit y probit también se muestran en la Tabla 3.5. El efecto marginal para la variable precio, por ejemplo, será:

$$ME_p = \partial P\,(Y=1)/\partial p_i = f(z) * \beta_1 \quad (3.15)$$

Los efectos marginales se interpretan como un aumento en la probabilidad de que un hogar *h* tendría un gasto positivo en productos de tabaco por un aumento de una unidad en el precio p_i. A partir de la Ecuación 3.15, la elasticidad precio se calcula como:

$$\varepsilon_{pp} = ME_p * (\overline{p_i}/Y) = \frac{\partial P(Y=1)}{\partial p_i} * \frac{\overline{p_i}}{Y} \quad (3.16)$$

donde $\overline{p_i}$ e *Y* son el precio promedio de los productos del tabaco y la prevalencia del tabaquismo antes del aumento de precio, respectivamente. Dado que la prevalencia en sí misma es una tasa y la elasticidad se interpreta como el cambio porcentual en la prevalencia con respecto a un cambio del uno por ciento en el precio, una forma más intuitiva de expresar los efectos marginales puede ser en términos de un cambio porcentual en la prevalencia como resultado de un aumento del uno por ciento en el precio. Eso se puede estimar de la siguiente manera:

$$\varepsilon_{pp*} = ME_p * (p_i) = \frac{dP(Y=1)}{dp_i} * p_i \quad (3.17)$$

Para la segunda parte del 2PM, serían suficientes las elasticidades cantidad obtenidas en la Sección 3.3. Una vez que se estiman las elasticidades cantidad de la Sección 3.3 y la elasticidad prevalencia de la Sección 3.4, la elasticidad precio total se convierte en una suma directa de ambas elasticidades. Por supuesto, el 2PM estima elasticidades para el margen extensivo e intensivo, de forma independiente. Ya se han publicado algunos estudios que estiman la elasticidad de la demanda de productos de tabaco utilizando el enfoque descrito en esta sección. Estos son estudios de Bosnia y Herzegovina,[70] Montenegro[77] y Serbia.[80] Todos ellos utilizan una regresión logit para estimar elasticidades prevalencia y el método de Deaton para estimar elasticidades cantidad.

3.4.1 Preparación de datos para estimar elasticidades prevalencia

La preparación de los datos sigue los mismos pasos que ya se describieron en la Sección 3.3.2. Las estimaciones de la elasticidad prevalencia utilizarían el mismo conjunto de variables que se utilizan en el caso de la elasticidad cantidad. Se necesita una variable indicadora binaria para el consumo de tabaco. Sin embargo, esa variable también se definió en la Sección 3.3.2 con el nombre de variable *dcig*. Para los precios se utilizarán los valores unitarios definidos anteriormente.

A diferencia del modelo de Deaton, que estima la elasticidad cantidad, como se analiza en la Sección 3.3, la estimación logit o probit no corrige la variación de la calidad ni los errores de medición en los valores unitarios. Dado que los datos de precios a nivel individual no están disponibles, sino solo los valores unitarios a nivel de hogar de la EGH, también existe una posible endogeneidad involucrada cuando los valores unitarios se utilizan como sustitutos del precio.

Una forma de mitigar parcialmente ese problema es usar una variable de precio a nivel de conglomerado en lugar de usar valores unitarios a nivel de hogar individual. Por lo tanto, los valores unitarios promedio a nivel de conglomerado se utilizan bajo el supuesto de que todos los hogares en un conglomerado determinado enfrentan el mismo valor unitario promedio. Eso se ocuparía de la endogeneidad, así como de las posibles variaciones de calidad en los valores unitarios entre los hogares hasta cierto punto (porque todas las variables a nivel del hogar se utilizan como controles en la regresión de la elasticidad prevalencia). Cuanto mayor sea el número de hogares que consumen tabaco en un conglomerado, mejor será la mitigación de la endogeneidad asociada y los problemas de matizado de calidad.

Sin embargo, eso aún no corregiría los posibles errores de medición en valores unitarios y, por desgracia, actualmente no existe una solución fácil para eso al estimar modelos logit o probit. Por esa razón, este conjunto de herramientas continúa confiando en el modelo de Deaton para estimar la elasticidad cantidad en lugar del 2PM convencional, que estima tanto la elasticidad prevalencia como la de la cantidad, las que no corrigen ni el matizado de la calidad ni el error de medición en ambas partes. Mediante el uso de un modelo logit/probit para estimar la primera parte y el modelo de Deaton para estimar la segunda parte, este conjunto de herramientas argumenta que las estimaciones presentadas serían mejores que las estimadas con un 2PM convencional.

3.4.2 Estimación de la elasticidad precio en el margen extensivo con Stata

Paso 1: Generación de una variable adicional para la estimación de la elasticidad prevalencia

La variable binaria que indica el estado de fumador (*dcig*) y las variables del lado derecho de la regresión para estimar la elasticidad precio ya se generaron en la Sección 3.3. La única variable adicional requerida es un valor unitario promedio del conglomerado que se puede asignar a todos los hogares en un conglomerado determinado. En caso de que no haya hogares disponibles en un conglomerado que reporten consumo de

tabaco, es posible que no haya un valor unitario para asignar. En ese caso, un valor unitario promedio debe definirse en un agregado geográfico más alto, como rural/urbano o distrito o región. El modelo tendrá al menos tantos valores unitarios promedio como el número de conglomerados con hogares fumadores. La variable *pcig* se utiliza como proxy de los precios en las regresiones logit/probit.

```
egen pcig=mean(uvcig), by(clust)
egen pcig2=mean(uvcig), by(region)
replace pcig=pcig2 if pcig==.
```

Paso 2: Ejecutando la regresión logit

Los siguientes comandos primero definen una macro global para las variables independientes, ejecutan la regresión logit con la dicotómica *dcig* como variable de resultado y estiman las probabilidades pronosticadas de resultados positivos (fumar) en una nueva variable *yhat_p*.

```
global $xvar lexp lhsize maleratio meanedu maxedu sgp1-sgp3
logit dcig $xvar
predict yhat_p, pr
```

Paso 3: Estimación de la elasticidad prevalencia

Los coeficientes obtenidos de la regresión logit no son elasticidades. Para obtener elasticidades (cambio porcentual en la probabilidad de que un hogar consuma cigarrillos con respecto a un cambio porcentual en el precio del cigarrillo), es necesario utilizar el comando *<margins>*. La sintaxis varía dependiendo de si las variables anteriores estaban en niveles o en forma logarítmica. Dado que la variable dependiente es binaria, está en niveles. Si la variable independiente precio también está en niveles, la elasticidad precio se obtiene con el siguiente comando:

```
margins, eyex(pcig)
```

Entonces, *eyex* obtiene un cambio porcentual en y para un cambio porcentual en *x*. En ese caso, usar *eyex* como en *<margins, eydx(pcig)>* produciría la llamada semielasticidad, que representa el cambio porcentual en *y* para una unidad de cambio en *x*. Sin embargo, si la variable precio *pcig* está en forma logarítmica, el código para obtener la elasticidad precio sería

```
margins, eyex(pcig)
```

En ese caso, *eydx* obtiene el cambio porcentual en *y* para un cambio unitario en *x*, que es un logaritmo del precio. El mismo procedimiento se puede utilizar para estimar la elasticidad gasto de la prevalencia. Las fórmulas utilizadas por el comando *margins* para estimar las elasticidades respectivas y sus errores estándar están disponibles en el manual de referencia base de Stata para cada versión de Stata.

Paso 4: Diagnóstico de regresión

Antes de que los resultados del modelo puedan usarse para hacer alguna inferencia estadística, y para que el análisis sea válido, es necesario verificar si el modelo cumple con los supuestos de los modelos de elección binaria. Las posibles comprobaciones incluyen probar si el modelo está especificado correctamente (prueba

de error de especificación), si el modelo general es estadísticamente significativo (prueba de bondad de ajuste) y si los regresores son ortogonales (prueba de multicolinealidad).

La prueba de error de especificación se realiza para confirmar que la función de probabilidad está correctamente especificada. Se hace con el comando *<linktest>* justo después del comando de regresión logit. El comando *<linktest>* reconstruye el modelo utilizando el valor predicho lineal (*_hat*) y el valor predicho lineal al cuadrado (*_hatsq*) como predictores. Si *_hat* resulta ser significativo, significa que se seleccionaron predictores significativos para el modelo y el modelo se especificó correctamente, por lo que cualquier significación estadística de la variable *_hatsq* debería ser puramente casual. Un *_hatsq* estadísticamente significativo puede ser indicativo de un error de especificación debido a la omisión de algunas variables importantes o a la omisión de algunos efectos de interacción de las variables incluidas en el modelo. La especificación del modelo debe cambiarse hasta que pueda pasar la prueba de error de especificación. Eso se puede hacer de muchas maneras, incluida la adición de nuevas variables, la adición de polinomios de orden superior de una o más de las variables existentes y la adición de nuevos efectos de interacción entre las variables existentes.

Las **pruebas de bondad de ajuste** son la prueba de razón de verosimilitud (LR) y la prueba de bondad de ajuste de Hosmer y Lemeshow (HL). Las estadísticas de LR se informan de forma predeterminada cuando se estima el modelo: el modelo se ajusta bien si las estadísticas de LR son estadísticamente significativas. La prueba HL comprueba si la frecuencia prevista y observada coinciden estrechamente, y se puede obtener con el comando *<lfit, group (10) table>* inmediatamente después de ejecutar el comando de regresión logit. Un valor p no significativo para las estadísticas de chi-cuadrado de HL indicaría que el modelo se ajusta bien a los datos. Alternativamente, el comando *<fitstat>* de Stata después de la regresión logit devolvería una serie de estadísticas de aptitud, incluido el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC).

La **prueba de multicolinealidad** se realiza para verificar si las variables son ortogonales entre sí (es decir, no están correlacionadas en absoluto). Se puede usar el programa *<collin>* escrito por usuarios de Stata que prueba la multicolinealidad con el comando *<collin var1 var2>*. Cuanto más se acercan a uno la tolerancia (que es igual a $1 - R^2$) y el factor de inflación variable (VIF = 1 / tolerancia), menos grave es el problema de multicolinealidad en el modelo. Como regla general, una tolerancia de 0,1 o menos (y VIF de 10 o más) debería ser preocupante. El programa *collin* se puede instalar en Stata usando el comando *<findit collin>*. Vale la pena señalar que el "problema" de la multicolinealidad, sin embargo, es solo una cuestión de grado, y apenas hay alguna solución práctica para ello. Descartar una variable relevante daría como resultado un sesgo de variable omitida. Por lo tanto, siempre que los coeficientes estimados de las variables incluidas tengan errores estándar lo suficientemente bajos e intervalos de confianza estrechos, la multicolinealidad puede ignorarse por completo.

La sección 7.1 del Apéndice de código incluye un do-file de ejemplo que detalla el código utilizado en esta sección.

3.4.3 Estudio de caso

Los mismos datos utilizados en la Sección 3.3.4 se utilizan para estimar la elasticidad precio y gasto de la prevalencia del tabaquismo. La Tabla 3.6 muestra las estimaciones de los coeficientes de la regresión logística junto con las elasticidades precio y gasto de la prevalencia estimadas usando el comando *<margin>* en Stata. La elasticidad precio de la prevalencia del tabaquismo es negativa y significativa, pero de muy baja magnitud. La elasticidad gasto, sin embargo, tiene signo positivo y un coeficiente de mayor magnitud.

Dado que la elasticidad precio de la prevalencia del tabaquismo (elasticidad en el margen extensivo) es -0,0528 y la elasticidad precio de la cantidad de tabaco (elasticidad en el margen intensivo) es -0,795, la elasticidad precio total de fumar se estima en (-0,0528) + (-0,795) = -0,8478.

Es ideal estimar primero las elasticidades de prevalencia, ya que eso utiliza la muestra completa. Posteriormente, mantenga solo los hogares que reportan consumo positivo y proceda a la estimación de la elasticidad cantidad. Sin embargo, la explicación en este conjunto de herramientas está en orden inverso, ya que es importante comprender la razón detrás del uso del método de Deaton para la estimación de la elasticidad cuando se utilizan datos de EGH y por qué se elige ese método para la estimación de las elasticidades cuantitativas en la segunda parte de 2PM en lugar de GLM, que se utiliza convencionalmente para estimar la segunda parte en 2PM.

Tabla 3.6 Resultados de regresiones logísticas y elasticidades

VARIABLES	Coeficiente	Errores estándar
pcig	-0,000317**	(0,000151)
lexp	0,956***	(0,0328)
lhsize	0,0760**	(0,0347)
maleratio	0,582***	(0,0530)
meanedu	-0,0315**	(0,0129)
maxedu	-0,0298**	(0,0122)
sgp1	0,0641	(0,0686)
sgp2	-0,506***	(0,0395)
sgp3	-0,107***	(0,0412)
Constante	-10,24***	(0,328)
Elasticidad precio	-0,0528**	(0,0251944)
Elasticidad gasto	0,5883***	(0,0203771)
Observaciones	25.188	

Nota: *** p<0,01, ** p<0,05, * p<0,1.

3.5 Estimación de elasticidades por grupos de ingreso

Tanto la elasticidad prevalencia como la cantidad pueden estimarse por grupos de ingresos de los hogares o por otras categorías socioeconómicas. Este conjunto de herramientas presenta solo la estimación de elasticidades por grupo de ingresos, que es la más utilizada. La primera decisión al estimar las elasticidades precio por grupo de ingreso es el número de cuantiles en los que se divide la muestra total de hogares en función del ingreso. Que el análisis se haga por terciles, quintiles o deciles depende de su objetivo y del contexto del país, pero también del tamaño de la muestra de la EGH.

La elasticidad precio estimada con el método de Deaton depende del número de conglomerados (es decir, la propiedad de consistencia), ya que se deriva de los promedios a nivel de conglomerados, pero no depende del número de hogares en cada conglomerado. Sin embargo, cuanto menor sea el número de hogares por

conglomerado, mayor será el error de medición. Como es probable que la variable de precio sea endógena, Deaton aborda ese problema estimando la elasticidad utilizando el precio promedio del grupo. Sin embargo, cuanto menor sea el número de hogares por conglomerado, es menos probable que se aborde el problema de la endogeneidad.

Entonces, una vez que la muestra de hogares de la EGH se divide en tres, cinco o 10 cuantiles de hogares en función de sus ingresos, el tamaño del conglomerado en cada cuantil se vuelve aún más pequeño. En otras palabras, un conglomerado en la muestra original ahora puede dividirse en dos o más subconglomerados dependiendo del ingreso de los hogares en ese conglomerado. Ese menor tamaño del conglomerado exacerba aún más el problema de la endogeneidad en la variable precio y el problema del error de medición. Ese es especialmente el caso en países con una población más pequeña y una muestra de EGH más pequeña.

Por lo tanto, se recomienda utilizar un número más pequeño en lugar de un número mayor de grupos de ingresos. Incluso en los casos en que la muestra EGH es lo suficientemente grande como para tener incluso 10 grupos de ingresos, la diferencia en las elasticidades estimadas entre los grupos de ingresos más bajos puede ser tan pequeña como para sugerir que pueden considerarse como un grupo de ingresos. Lo mismo puede observarse para los grupos de ingresos medianos y altos. Es posible realizar una prueba estadística para determinar si los coeficientes de elasticidad entre diferentes subgrupos son estadísticamente diferentes entre sí. Ese también puede ser un principio rector para decidir el número de subgrupos que se utilizarán para el análisis.

La segunda decisión al estimar las elasticidades precio por grupo de ingresos es si se debe usar el mismo precio promedio para todos los hogares en subconglomerados que se derivan del mismo conglomerado en la muestra original, independientemente de sus ingresos, o si el precio promedio por subconglomerado debe ser diferente. Los estudios que han estimado las elasticidades precio por grupo socioeconómico usando el método de Deaton hasta ahora han asumido diferentes precios promedio por grupo de ingreso. Sin embargo, como se mencionó anteriormente, para abordar la endogeneidad potencial en la variable de precio, el método de Deaton deriva un precio promedio a nivel de conglomerado y lo aplica a todos los hogares en ese conglomerado.

Eso se basa en el supuesto de que no hay variación de precios dentro de un conglomerado, ya que todos los hogares de un conglomerado viven lo suficientemente cerca unos de otros y, por lo tanto, realizan compras en el mismo mercado y son encuestados al mismo tiempo. Además, al utilizar el precio promedio a nivel de conglomerado, se minimiza, si no se elimina, el problema potencial de la autoselección de los hogares en función de su nivel de ingresos y sus preferencias. Por lo tanto, tanto intuitivamente como de acuerdo con los supuestos del método de Deaton, se debe aplicar el mismo precio promedio a nivel de conglomerado a todos los hogares en el mismo conglomerado, independientemente de su nivel de ingresos. Como el método de Deaton está diseñado para estimar la elasticidad a nivel de población, pero no dentro de esa misma población por grupos NSE, la aplicación del método requiere ajustes en el código, que se presenta en este conjunto de herramientas.

3.5.1 Preparación de datos para estimar la elasticidad por grupos de ingreso

La categorización de los hogares por grupo de ingreso es el primer paso para estimar las elasticidades por grupo de ingreso. Dado que la información sobre ingresos generalmente no se proporciona en las EGH, se representa mediante la suma de todos los gastos del hogar informados durante el período de informe, como en

$$x_{hc} = \sum_{i=1}^{N} x_{ihc} \quad (3.18)$$

donde x_{hc} es el gasto total informado de un hogar *h* en el conglomerado *c*, que es la suma del gasto en artículos *i* reportado por ese hogar.

Dado que los datos de la EGH son a nivel de hogar, los hogares deben dividirse en grupos de ingresos no según el gasto total del hogar, sino según el gasto por miembro del hogar:

$$x_{phc} = \frac{1}{s_{hc}} \sum_{i=1}^{N} x_{ihc} \quad (3.19)$$

donde x_{phc} se informa el gasto total por miembro del hogar (*p* significa per cápita) de un hogar *h* en conglomerado *c* y s_{hc} es el número de miembros del hogar. La variable de gasto total del hogar (*exptotal*), así como la variable del tamaño del hogar (*hsize*), ya se definieron en la Sección 3.3. Luego, el gasto per cápita del hogar se genera usando *<gen exppc=exptotal/hsize>*. Los cuantiles se pueden crear con el comando *<xtile inc = exppc [w=weights], nq(3)>* donde la opción *nq (.)* especifica el número de cuantiles y *weights* son la variable para los ponderadores de los hogares en la encuesta.

Se eligen tres grupos de ingresos (bajo, mediano, alto) para los fines de este análisis, ya que esos proporcionarían una variación adecuada entre los grupos de ingresos, al tiempo que garantizan la cantidad máxima de subconglomerados en cada grupo junto con tamaños de subconglomerados razonables. El comando crea una nueva variable *inc*, que distribuye los hogares en tres terciles utilizando las ponderaciones adecuadas de la encuesta. Es importante notar que, debido a que se usan los ponderadores, los tres grupos pueden no tener el mismo número de observaciones. También se crea una variable de subconglomerado (*subcluster*) usando la variable de conglomerado existente (*clust*) y la nueva variable de grupo de ingresos (*inc*) usando el comando *<egen subclust=group(clust inc)>*.

Al igual que en la Sección 3.3, los subconglomerados con menos de dos hogares que reportan consumo de tabaco se eliminan del análisis. Eso puede hacerse de la siguiente manera:

```
bys subclust: egen cigsubclust =sum(dcig)
drop if cigsubclust <2
```

Estas, junto con las demás variables específicas del hogar utilizadas en la Sección 3.3, serían suficientes para estimar las elasticidades cantidad por grupo de ingresos utilizando el método de Deaton. Cabe señalar que muchos conjuntos de datos de EGH vienen con grupos que ya son bastante pequeños. Cuando se realiza un análisis por subgrupos y se generan subconglomerados para obtener porcentajes de presupuesto promedio a nivel de subconglomerados, es muy posible terminar con varios subconglomerados que tienen menos de dos hogares con fumadores. Si eso sucede, muchas observaciones se eliminarán del análisis y, como resultado, las estimaciones de elasticidad que utilizan la muestra general no necesariamente estarán dentro del rango de elasticidades estimadas para diferentes subgrupos. Eso se debe a que las muestras fusionadas utilizadas para un análisis de subconglomerados no tendrán las mismas observaciones que la muestra de todos los hogares antes de crear los subgrupos.

Para la estimación de las elasticidades de prevalencia, se crea una variable de precio como el promedio de los valores unitarios de los subconglomerados recién creados, de la siguiente manera:

```
egen pcig=mean(uvcig), by(subclust)
egen pcig2=mean(uvcig), by(clust)
replace pcig=pcig2 if pcig==.
```

La prueba de la significancia de la diferencia de los coeficientes de elasticidad entre los grupos de ingresos se puede implementar con la ayuda de una regresión aparentemente no relacionada (SUR). Para ello, simultáneamente se estiman modelos por subgrupos (ingresos), y se almacenan primero sus resultados. Posteriormente, se puede realizar una prueba de significación de chi-cuadrado (χ^2) con los resultados almacenados para probar la significación estadística de diferentes combinaciones lineales o igualdad de coeficientes.[102, 103] El comando *<suest test>* de Stata se puede usar para ese propósito. Más detalles se proporcionan a continuación.

3.5.2 Estimación de la elasticidad de la prevalencia por grupo de ingresos con Stata

Dado que las variables necesarias ya están definidas y se eliminaron los subconglomerados con menos de dos hogares que reportan un consumo positivo, los códigos de la sección 3.4.2 se pueden usar extendiendo los mismos códigos con un comando de loop simple para el grupo de ingresos, como se muestra a continuación.

```
global xvar "pcig lexp lhsize maleratio meanedu maxedu sgp1 sgp2 sgp3"
local append "replace"
forvalues i=1/3 {
    logit dcig $xvar if inc== `i'
    outreg2 using PrevalenceElastInc.doc, ctitle (Income group: `i') `append'
    predict yhat_p `i', pr
    local append "append"
}
```

La elasticidad precio se puede obtener con el comando *<margins, eyex(pcig)>* y la elasticidad gasto con el comando *<margins, eydx(lexp)>*, como se indicó anteriormente.

La Tabla 3.7 muestra los resultados de la regresión logística junto con la elasticidad precio e ingreso de la prevalencia de tabaquismo obtenidos con los comandos de margen utilizando los mismos datos que en las secciones anteriores. Se puede ver que la elasticidad precio de la prevalencia, aunque de signo negativo, no es estadísticamente significativa para ninguno de los grupos de ingresos. Las elasticidades gasto son positivas y estadísticamente significativas.

La prueba de significancia de la diferencia de los coeficientes de elasticidad entre los grupos de ingresos se puede realizar utilizando el siguiente código:

```
Global xvar “pcig lexp lhsize maleratio meanedu maxedu sgp1 sgp2 sgp3”
local append “replace”
forvalues i=1/3 {
    logit dcig $xvar if inc== `i’
    outreg2 using PrevalenceElastInc.doc, ctitle (Income group: `i’) `append’
    predict yhat_p `i’, pr
    margins, eyex(pcig)
    estimates store inc`i’
    local append “append”
}
suest inc*
test [inc1_dcig]pcig-[inc2_dcig]pcig=0
```

```
test [inc1_dcig]pcig-[inc3_dcig]pcig=0
test [inc2_dcig]pcig-[inc3_dcig]pcig=0
```

Después de almacenar cada estimación con el comando *<estimates store name>*, el comando *<suest inc*>* devolverá los resultados de una regresión SUR. Después de eso, las diferencias entre cada coeficiente se pueden probar con un comando de prueba simple que, por ejemplo, prueba si la diferencia entre el coeficiente de elasticidad precio entre el grupo de ingresos 1 y el coeficiente de elasticidad precio entre el grupo de ingresos 2 es igual a 0. La prueba devolvería un estadístico de chi-cuadrado (χ^2). Un estadístico significativo aquí implicaría que la diferencia en los coeficientes de elasticidad entre los grupos de ingreso 1 y 2 es estadísticamente significativa.

Tabla 3.7 Resultados de regresiones logísticas y elasticidades de prevalencia por grupo de ingreso

VARIABLES	Ingresos bajos	Ingresos medianos	Ingresos altos
pcig	-7,76e-05 (0,000338)	0,000104 (0,000297)	-0,000117 (0,000273)
lexp	1,178*** (0,107)	0,793*** (0,220)	0,438*** (0,0885)
lhsize	0,247** (0,106)	0,246 (0,222)	0,182* (0,0933)
maleratio	0,711*** (0,138)	0,929*** (0,112)	0,278*** (0,0839)
meanedu	-0,0376 (0,0255)	-0,0596** (0,0252)	0,00293 (0,0278)
maxedu	-0,0279 (0,0239)	0,00471 (0,0236)	-0,0570** (0,0268)
sgp1	0,254** (0,117)	0,102 (0,160)	0,0787 (0,170)
sgp2	-0,319*** (0,0876)	-0,563*** (0,0800)	-0,638*** (0,0708)
sgp3	-0,0342 (0,0798)	-0,165** (0,0795)	-0,118 (0,0861)
Constante	-12,66*** (1,024)	-8,758*** (2,179)	-4,213*** (0,927)
Elasticidad precio	-0,011926 (0,0519689)	0,0151632 (0,0432479)	-0,016525 (0,038618)
Elasticidad gasto	0,706927*** (0,0646166)	0,4342673** (0,12084)	0,2208638*** (0,0446629)
Observaciones	5.835	6.291	6.135

Nota: Errores estándar entre paréntesis; *** p<0,01, ** p<0,05, * p<0,1.

3.5.3 Estimación de la elasticidad cantidad por grupo de ingresos con Stata

En el caso del método de Deaton, hay algunos cambios significativos en el código presentado en la Sección 3.3.3 para estimar las elasticidades de la demanda de un solo producto. El siguiente código estima las elasticidades precio e ingreso de la demanda de cigarrillos junto con sus errores estándar de bootstrap para los tres grupos de ingresos simultáneamente. En primer lugar, los hogares de no fumadores se eliminan del análisis con el comando *<keep if dcig==1>*, ya que solo se estiman las elasticidades condicionales. Las variaciones espaciales en los valores unitarios se pueden probar de la misma manera que en la Sección 3.3.3 usando el comando *<anova luvcig clust>*. Tenga en cuenta que eso no se hace a nivel de subconglomerado, sino a nivel de conglomerado. Eso se debe a que se supone que todos los hogares en todos los subconglomerados dentro de un conglomerado enfrentan los mismos precios promedio de mercado.

Estimación de regresiones de primera etapa dentro de un conglomerado y varianzas de errores de medición

La regresión del valor unitario de la primera etapa será la misma para todos los grupos de ingresos, mientras que la regresión de la participación en el presupuesto de la primera etapa se realiza por separado para cada grupo de ingresos. Los parámetros de esas regresiones se guardan para la estimación de la elasticidad en la segunda etapa.

```
#delimit;
areg luvcig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust);
predict ruvcig, resid;
scalar sigma11=$S_E_sse / $S_E_tdf;
scalar b1=_coef[lexp];
gen y1cig=luvcig-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio
          -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu
          -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3;

#delimit;
forvalues i=1/3 {;
    areg bscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3 if inc== `i', absorb(clust);
    predict rbscig `i', resid;
    scalar sigma22 `i'=$S_E_sse/$S_E_tdf ;
    scalar b0 `i'=_coef[lexp];
    gen y0cig `i'=bscig-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio
          -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu
          -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3 if inc== `i' ;
    qui areg ruvcig rbscig `i' lexp lhsize maleratio meanedu maxedu sgp1-sgp3 if inc== `i',
absorb(clust) ;
    scalar sigma12 `i'=_coef[rbscig `i']*sigma22 `i' ;
};
```

A diferencia de la Sección 3.3.3, la covarianza de u_0 (*sigma22*) y la de u_1, el coeficiente b_0 y la variable *y0cig* para la segunda etapa son todos diferentes para cada grupo de ingreso.

Estimación de elasticidades de ingresos o gastos

El siguiente código estima el coeficiente de elasticidad gasto junto con los errores estándar obtenidos mediante bootstrap y los almacena en un archivo separado como “deatonExpElast.doc” en un formato listo para usar.

```
cap program drop Expelast
program define Expelast, rclass
args i
qui areg bscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3 if inc== `i', absorb(clust)
local a0 =_coef[lexp]
qui sum bscig if inc== `i'
local vbar =r(mean)
return scalar Expel`i' =1-b1+(`a0'/`vbar')
end
forvalues i=1/3 {
bootstrap Expel`i'=r(Expel`i'), reps(1000) seed(1): Expelast  `i'
estimates store exp_el`i'
}
```

Preparación de datos para la regresión entre conglomerados

El siguiente paso consiste en promediar las variables *y1cig* e *y0cig* por conglomerados para generar *y1c* e *y0c* respectivamente, de modo que puedan usarse para una regresión entre conglomerados de *y0c* sobre *y1c* para derivar la elasticidad precio. Pero solo se genera *y1c* a continuación, ya que se hace para todos los grupos de ingresos juntos. *y0c* se genera junto con la estimación de la elasticidad en la siguiente etapa a medida que se estiman los errores estándar mediante bootstrap para cada elasticidad del grupo de ingresos, uno a la vez.

```
sort clust
egen y1c= mean(y1cig), by(clust)
egen n1c=count(y1cig), by(clust)
ameans n1c
scalar n1=r(mean_h)
drop n1c
mean y1c
```

Estimación de la elasticidad precio

El siguiente código estima la elasticidad del precio, así como los errores estándar mediante bootstrap por separado para cada grupo de ingresos y los almacena en un archivo como “DeatonPriceElast.doc” en un formato listo para usar.

```
cap program drop elast
program define elast, rclass
      qui sum inc
      local a=r(mean)
      global j= `a'
      tempname S R num den phi theta psi
      qui corr y0c$j y1c, cov
      scalar S=r(Var_2)
      scalar R$j=r(cov_12)
      scalar num$j = scalar(R$j) - (sigma12$j / n0$j)
      scalar den=scalar(S)-(sigma11/n1)
      cap scalar phi$j = num$j / den
      cap scalar zeta$j = b1/((b0$j + Wbar$j * (1-b1)))
      cap scalar theta$j =phi$j /(1+(Wbar$j - phi$j) * zeta$j)
      cap scalar psi$j = 1-((b1*(Wbar$j - theta$j ))/(b0$j + Wbar$j))
      return scalar EP$j = (theta$j / Wbar$j )- psi$j
end

local append "replace"
forvalues i=1/3 {
      preserve
      egen y1c= mean(y1cig), by(clust)
      keep if inc== `i'
      sort clust
      egen y0c `i'= mean(y0cig `i'), by(clust)
      egen n0c `i'=count(y0cig `i'), by(clust)
      egen n1c=count(y1cig), by(clust)
      qui sum bscig if inc== `i'
      scalar Wbar `i' = r(mean)
      sort clust
      qui by clust: keep if _n==1
      qui ameans n0c `i'
      scalar n0 `i'=r(mean_h)
      qui ameans n1c
      scalar n1=r(mean_h)
      drop n0c `i' n1c
      elast
      return list
      bootstrap EP `i'=r(EP `i'), reps(1000) seed(1): elast
      outreg2 using DeatonPriceElast.doc, dec(4) ctitle (Income group:  `i') `append'
      local append "append"
      restore
}
```

La Sección 7.2 del Apéndice de Código reproduce el código anterior junto con el de la estimación de las elasticidades de prevalencia por grupos de ingresos. La Tabla 3.8 muestra los resultados de la elasticidad precio y gasto de cigarrillos en el margen intensivo utilizando los mismos datos que en la sección anterior. Todos los coeficientes de elasticidad son altamente significativos y tienen los signos esperados. La elasticidad total puede obtenerse sumando las elasticidades de prevalencia y cantidad para cada grupo de ingresos, respectivamente.

Si también se quiere comparar si las elasticidades son significativamente diferentes desde el punto de vista estadístico entre los grupos de ingresos, no es posible utilizar *suest* como antes. En su lugar, se pueden utilizar los siguientes códigos para verificar las diferencias estadísticas en las elasticidades de los precios entre los grupos de ingresos. El comando test al final del código muestra si las elasticidades estimadas difieren significativamente entre sí.

Tabla 3.8 Elasticidad precio y gasto de la demanda de cigarrillos por grupo de ingreso

	Elasticidad gasto	Elasticidad precio
Ingresos bajos		
Coeficiente de elasticidad	0,649***	-0,808***
Intervalo de confianza del 95%	[0,506 0,791]	[-0,8941 -0,7216]
Error estándar de bootstrap	(0,0714)	(0,0440)
Observaciones	2 333	728
Ingresos medianos		
Coeficiente de elasticidad	0,605***	-0,826***
Intervalo de confianza del 95%	[0,369 0,841]	[-0,9014 -0,7507]
Error estándar de bootstrap	(0,1218)	(0,0385)
Observaciones	2 844	883
Ingresos altos		
Coeficiente de elasticidad	0,422***	- 0,816***
Intervalo de confianza del 95%	[0,321 0,523]	[-0,8729 -0,7586]
Error estándar de bootstrap	(0,0503)	(0,0292)
Observaciones	3 044	920

Nota: Los errores estándar obtenidos mediante bootstrap se calcularon haciendo 1000 replicaciones. Asumiendo que las estimaciones siguen una distribución normal, los coeficientes con *** y ** implican niveles de significancia del 1% y 5%, respectivamente.

```
cap program drop elast
program define elast, rclass
forvalues i=1/3 {
    preserve
    egen y1c `i'= mean(y1cig), by(clust)
    keep if inc== `i'
    sort clust
    egen y0c `i'= mean(y0cig `i'), by(clust)
    egen n0c `i'= count(y0cig `i'), by(clust)
```

```
        egen n1c `i'= count(y1cig), by(clust)
        qui sum bscig
        scalar Wbar `i' = r(mean)
        sort clust
        qui by clust: keep if _n==1
        qui ameans n0c `i'
        scalar n0 `i'=r(mean_h)
        qui ameans n1c `i'
        scalar n1 `i'=r(mean_h)
        drop n0c `i' n1c `i'
        tempname S `i' R `i' num `i' den `i' phi `i' theta `i' psi `i'
        qui corr y0c `i' y1c `i', cov
        scalar S `i'=r(Var_2)
        scalar R `i'=r(cov_12)
        scalar num `i' = scalar(R `i') - (sigma12 `i' / n0 `i')
        scalar den `i'=scalar(S `i')-(sigma11/n1 `i')
        cap scalar phi `i' = num `i' / den `i'
        cap scalar zeta `i' = b1/((b0 `i' + Wbar `i' * (1-b1)))
        cap scalar theta `i' =phi `i' /(1+(Wbar `i' - phi `i') * zeta `i')
        cap scalar psi `i' = 1-((b1*(Wbar `i' - theta `i'))/(b0 `i' + Wbar `i'))
        return scalar Elast_ `i' = (theta `i' / Wbar `i' )- psi `i'
        restore
    }
    end
    elast
    bootstrap elast1=r(Elast_1)elast2=r(Elast_2)elast3=r(Elast_3), reps(1000) seed(1):elast

    test _b[elast1]=_b[elast2]
    test _b[elast2]=_b[elast3]
    test _b[elast1]=_b[elast3]
```

3.6 Estimación de elasticidades cuando los valores unitarios no están disponibles en la EGH

El enfoque de Deaton permite la estimación de la demanda y el cálculo de las elasticidades precio y precio cruzadas utilizando cantidades y valores unitarios obtenidos de los datos de la EGH. Sin embargo, a veces los datos de la EGH recopilan información solo sobre los gastos en los que incurren los hogares para diferentes grupos de productos básicos. No proporciona información sobre las cantidades compradas y, como resultado, no es posible construir valores unitarios cuya variación espacial pueda utilizarse para informar la variabilidad de los precios a nivel de hogar. En ese caso, no se puede aplicar el enfoque de Deaton que se analiza en este capítulo. Dado que las EGH proporcionan información valiosa sobre el consumo de los hogares junto con el de los productos de tabaco, no sería prudente ignorar dichos datos simplemente porque no se dispone de información sobre la cantidad. Afortunadamente, existen métodos

para recuperar valores unitarios (o pseudovalores unitarios) de modo que puedan usarse para la estimación de funciones de demanda y para derivar la elasticidad precio.

Tradicionalmente, cuando la información sobre cantidades no está disponible en la EGH, las fuentes externas de variabilidad de precios obtenidas de los índices de precios nacionales agregados, como los índices de precios al consumidor (IPC) y los datos de precios promedio ponderados o no ponderados disponibles en los niveles administrativos locales, a menudo se fusionan con el gasto de los hogares para obtener estimaciones de las elasticidades precio.[104] Los sistemas de demanda populares como AIDS o el sistema casi ideal de demanda cuadrático (QAIDS) se emplean a menudo al usar dichos índices de precios para estimar las funciones de demanda. Sin embargo, ese enfoque es criticado por no tener en cuenta la variabilidad espacial y de los hogares, lo que da como resultado estimaciones distorsionadas de los parámetros de la demanda y no es coherente con la teoría.[105–108] Además, los índices de precios agregados suelen estar muy correlacionados y pueden sufrir problemas de endogeneidad.[109]

Literatura reciente,[110] sin embargo, sugiere que la construcción de índices de precios a nivel de hogares (precios de Stone-Lewbel (SL) [111]) para grupos de productos básicos puede mitigar los problemas relacionados con el uso exclusivo de índices de precios agregados en situaciones en las que la información sobre cantidades no está disponible en la encuesta. Los índices de precios SL para grupos de productos básicos se construyen utilizando información sobre las participaciones en el presupuesto de los subgrupos, las características demográficas de los hogares y los índices de precios nacionales agregados, y permiten recuperar los precios a nivel de los hogares o los valores unitarios.[110] Se ha encontrado que el uso de precios SL específicos de los hogares da como resultado parámetros de demanda que son más precisos y económicamente plausibles que los que se obtienen usando solo índices de precios agregados.[108] El programa escrito por usuarios en Stata, *<pseudounit>*[104] ayuda a estimar dichos valores unitarios (pseudovalores unitarios) usando este método para las EGH sin información de cantidad.

Un sistema de demanda marshalliano implícito del Índice Exacto Afín de Stone (EASI) recientemente propuesto hace uso de esos métodos para estimar la elasticidad precio[110, 112] y tiene varias ventajas sobre los sistemas de demanda tradicionales como el AIDS. En la literatura también se encuentran disponibles diferentes métodos empíricos para el cálculo del índice de precios SL para agregados de productos.[113] Sin embargo, este conjunto de herramientas no aborda esos temas y los desarrollos que los rodean, ya que, en la mayoría de los casos, los datos de las EGH brindan tanto la cantidad como los gastos para diferentes productos básicos de interés. Sin embargo, los lectores que tengan datos de EGH sin información sobre cantidades deben familiarizarse con la literatura de esta sección antes de intentar estimar la elasticidad precio a partir de dichos datos.

4 Estimación del efecto desplazamiento del gasto en tabaco

4.1 Cómo el gasto en tabaco desplaza el gasto en otros bienes y servicios

Si bien la prevalencia mundial del consumo de tabaco ha disminuido del 26,7% en 2010 al 22,3% en 2020, gran parte de esa disminución se ha producido en los PIMB.[114] La mayoría (alrededor del 80 por ciento) de los aproximadamente 1300 millones de consumidores actuales de tabaco en el mundo viven en PIMB.[115] También se encuentra que la prevalencia del consumo de tabaco sin humo es mucho mayor en los países de ingresos medianos y bajos. De los aproximadamente 335 millones de adultos que consumen tabaco sin humo en el mundo, 266 millones se encuentran en el Sudeste Asiático.[114] Varios estudios también han demostrado que el consumo de tabaco es desproporcionadamente mayor entre las personas relativamente pobres. Un metaanálisis de 201 estudios realizado por la OMS encontró una asociación estadísticamente significativa entre una mayor prevalencia de tabaquismo actual entre adultos y menores ingresos, tanto para hombres como para mujeres.[116]

El gasto en tabaco representa una parte importante del presupuesto familiar en muchos países, desde el uno por ciento en países como México y Hong Kong hasta el 11 por ciento en países como Zimbabue y China.[117] Los hogares operan sobre la base de ingresos disponibles limitados y, como resultado, cuando gastan sus presupuestos limitados en tabaco, tienen un costo de oportunidad enorme. Inevitablemente eso significa que tienen que reducir los gastos en otros bienes y servicios, algunos de los cuales pueden ser necesidades como alimentos, ropa y vivienda. La idea de que los hogares que gastan dinero en el consumo de tabaco desvían fondos del consumo de otros productos básicos se denomina efecto de "desplazamiento" del gasto en tabaco.

Hubo algunos intentos iniciales de explicar el problema del desplazamiento con análisis descriptivos de datos de Bangladesh[118] y China[119] en los años 2001 y 2002, respectivamente. Un examen empírico formal de la idea de desplazamiento debido al gasto en tabaco utilizando métodos econométricos vino más tarde de los EE. UU.[120] y China[121] en los años 2004 y 2006. Esos estudios, sin embargo, no pudieron modelar explícitamente la cuestión de la endogeneidad presente en dicho análisis.

La generación actual de métodos econométricos que estiman el efecto de desplazamiento del gasto en tabaco comenzó en 2008 utilizando datos de gastos de los hogares de la India.[117] Utiliza técnicas de VI para dar cuenta de la posible endogeneidad en el sistema de demanda al tiempo que trata el gasto en tabaco como un regresor y encuentra que el gasto en tabaco desplaza a los alimentos, la educación y el entretenimiento mientras que atrae los gastos en salud, ropa y combustibles. Se han realizado estudios que utilizan métodos econométricos similares y datos de gastos de los hogares en otros países como (cronológicamente) Taiwán,[122] Sudáfrica,[123] Camboya,[124] Zambia,[125] Turquía,[126] Bangladesh,[127] Mauricio,[128] Chile,[129] Vietnam,[130] Ghana,[131] y Kenia.[132] También hay estudios que utilizan diferentes métodos para examinar el desplazamiento en Indonesia,[133] Sudáfrica,[134] Corea del Sur,[135] y otros PIMB.[136]

Tabla 4.1 Estudios econométricos sobre el efecto desplazamiento del gasto en tabaco

Año	Autores	País	Método	Datos de encuesta utilizados	Artículos desplazados
2004	Busch et al.[120]	EE. UU.	Regresiones MCO separadas	Encuesta de Gastos del Consumidor	Ropa, vivienda
2006	Wang et al.[121]	China	Modelo logit fraccional	Encuesta primaria	Educación, mantenimiento de equipos agrícolas, ahorros.
2008	John, RM[117]	India	Variables instrumentales (VI)	Encuesta Nacional por Muestreo	Comida, educación, entretenimiento
2008	Pu et al.[122]	Taiwán	VI	Encuesta de Ingresos y Gastos Familiares	Ropa, atención médica, transporte
2008	Koch & Tshiswaka-Kashalala[123]	Sudáfrica	VI	Encuesta de Ingresos y Gastos de Sudáfrica	Educación, combustible, ropa, salud, transporte
2009	Block & Webb[133]	Indonesia	Ecuaciones de forma reducida	Datos del sistema de vigilancia nutricional	Alimento
2012	John et al.[124]	Camboya	VI	Encuesta Socioeconómica de Camboya	Alimentación, educación, ropa
2014	Chelwa & Walbeek[125]	Zambia	VI	Encuesta de Seguimiento de las Condiciones de Vida	Alimentos, educación, ropa, transporte, mantenimiento de equipos
2015	San & Chaloupka[126]	Turquía	VI	Encuesta de Presupuesto de los Hogares Turcos	Alimentación, vivienda, educación, bienes duraderos/no duraderos
2015	Do & Bautista[136]	40 PIMB	Modelos de pendiente aleatoria	Encuesta Mundial de Salud	Educación, cuidado de la salud
2018	Husain et al.[127]	Bangladesh	VI	Encuesta de Ingresos y Gastos de los Hogares	Ropa, vivienda, educación, energía, transporte, comunicación
2018	Paraje & Araya[129]	Chile	Modelo de QAIDS	Encuesta de Presupuestos Familiares de Chile	Salud, educación, vivienda

Tabla 4.1 Estudios econométricos sobre el efecto desplazamiento del gasto en tabaco (cont.)

Año	Autores	País	Método	Datos de encuesta utilizados	Artículos desplazados
2018	Ross et al.[128]	Mauricio	VI	Encuestas de Presupuesto Familiar	Transporte, comunicación, salud, educación
2019	Chelwa & Koch[134]	Sudáfrica	Coincidencia genética/ no paramétrica	Encuestas de Ingresos y Gastos	Artículos de alimentación seleccionado
2020	Masa-ud et al.[131]	Ghana	GMM-3SLS / VI	Encuesta de Niveles de Vida en Ghana	Alimentación, vivienda, atención médica
2020	Nguyen & Nguyen[130]	Vietnam	GMM-3SLS / VI	Encuesta sobre el Nivel de Vida de los Hogares de Vietnam	Educación
2020	Nyagwachi et al.[132]	Kenia	Diferencias en Diferencias Emparejadas	Encuesta Integrada de Hogares y Presupuesto de Kenia	Educación, comunicación,, y algunos alimentos
2021	Jin & Cho[135]	Corea del Sur	Diferencias en Diferencias Emparejadas	Encuesta de Ingresos y Gastos de los Gogares Coreanos	Artículos de alimentación seleccionados
2021	Djutaharta et al.[137]	Indonesia	Modelo AIDS	SUSENAS, PODS, y RISKESDAS	Varios artículos de alimentación
2022	Vladisavljevic et al.[138]	Serbia	VI	Encuesta de Presupuesto Familiar de Serbia	Alimentación, ropa, educación, recreación
2022	Wisana et al.[139]	Indonesia	GMM-3SLS / VI	SUSENAS	Alimentos, ropa, vivienda, servicios públicos, educación, atención médica
2022	Mugoša et al.[140]	Montenegro	VI	Encuesta de Presupuestos Familiares	Ropa, vivienda y educación
2022	Gómez et al.[141]	México	GMM-3SLS	Encuesta Nacional de Ingresos y Gastos de los Hogares	Alimentos, alcohol, transporte, bienes duraderos

La Tabla 4.1 anterior proporciona un resumen de los diferentes estudios econométricos que se han realizado para examinar el efecto de desplazamiento del gasto en tabaco. Como se puede ver, la técnica de VI es el método preferido adoptado por la mayoría de los estudios de los últimos 15 años. Sin embargo, algunos de los estudios más recientes han sido críticos con la técnica de VI y, en cambio, han propuesto el uso de

métodos de Diferencias en Diferencias Emparejadas (MDID) como alternativa. Sin embargo, el requisito de datos para el método MDID puede ser más restrictivo, ya que idealmente requeriría datos de panel de hogares. La mayoría de los estudios sobre desplazamiento encuentran que el gasto en tabaco desplaza los gastos en artículos necesarios de consumo doméstico, como alimentos, ropa, vivienda y educación, entre otros, lo que implica que el gasto en tabaco puede tener impactos intergeneracionales y en el desarrollo.

4.2 Importancia de la asignación de recursos dentro del hogar

Los hogares a menudo agrupan los recursos de los miembros individuales de la familia y toman decisiones sobre el gasto o la asignación de presupuestos entre los bienes de consumo alternativos que requiere cada miembro individual. En la mayoría, si no en todas las EGH, el hogar es la unidad para la cual se informa el consumo. Sin embargo, no se informa cómo se produce la distribución del consumo entre los miembros de la familia. Si las decisiones de asignación las toman determinados miembros adultos en un hogar (a menudo hombres en varios PIMB[142,143]), su impacto en el bienestar social es incierto. Como señala Deaton,[8] si las mujeres reciben sistemáticamente menos que los hombres, o si los niños y los ancianos están sistemáticamente en peores condiciones que otros miembros del hogar, el bienestar social se exagerará cuando se utilicen medidas que supongan que todos los miembros del hogar reciben el mismo trato.

Las decisiones de asignación de recursos dentro del hogar se vuelven aún más importantes cuando los ingresos disponibles se reducen una vez que el dinero se asigna a gastos improductivos, como el gasto en tabaco. Dado que el consumo de tabaco es más frecuente entre los hombres que entre las mujeres en la mayoría de los países,[144] si las decisiones de asignación las toman los jefes varones de un hogar, podrían ser potencialmente desfavorables para las mujeres y/o los niños dentro de esos hogares. De hecho, algunos de los hallazgos de la literatura sobre desplazamiento descritos anteriormente subrayan eso.

Por ejemplo, cuando los gastos educativos se ven comprometidos como resultado de una mayor asignación para el consumo de tabaco, afecta directamente a los niños de un hogar y su potencial de ingresos futuros, al tiempo que impone impactos intergeneracionales a largo plazo en la sociedad. La literatura de la India[117] muestra que los hogares que gastan en tabaco asignan sistemáticamente menos dinero a combustibles limpios para cocinar y asignan más dinero a fuentes de combustible no limpias, como la leña, que puede ser más peligrosa para las mujeres que se dedican a recolectarla y quemarla mientras cocinan.

Dado que el consumo de tabaco es en gran medida adictivo, es muy posible que los hogares preasignen una determinada parte del presupuesto para la compra de tabaco. Eso significa que el hogar tiene que maximizar su utilidad asignando de manera óptima el presupuesto restante (total menos el presupuesto preasignado para tabaco) entre bienes alternativos. Ciertamente, dado que el presupuesto disponible se reduce después de la preasignación, se deben hacer algunas concesiones. Los estudios de desplazamiento han encontrado que se hacen concesiones en el caso de productos básicos como alimentos, educación y ropa, que pueden afectar directamente la salud y el desarrollo de todos los miembros de un hogar. Por lo tanto, es importante que las políticas de control del tabaco aborden esos desafíos.

4.3 Comparación de la participación promedio en el presupuesto

Verificar las diferencias en la participación promedio en el presupuesto o el gasto promedio en diferentes grupos de productos básicos entre los hogares que gastan en tabaco y los hogares que no gastan en tabaco proporciona una indicación preliminar de las posibles concesiones, si las hay, como resultado del gasto en tabaco. Esta sección examina esas diferencias dividiendo los hogares en diferentes grupos en base de sus hábitos de consumo de tabaco y comparando la parte del presupuesto que cada grupo asigna a la compra de diferentes grupos de productos básicos.

Paso 1: Creación de participación en el presupuesto promedio por tipo de hogar

Como primer paso, cree una variable categórica *tob* que tome el valor de 1 si los hogares gastan dinero en tabaco y 0 en caso contrario. Como ejemplo, *exptobac* es la variable que representa la cantidad gastada en tabaco por un hogar extraída de la EGH. Luego, se puede generar la variable indicadora de tabaco, y se pueden etiquetar sus valores como "Tobacco spenders" (Consumidores de tabaco) y "Tobacco non-spenders" (No consumidores de tabaco) con los siguientes comandos:

```
gen tob= exptobac >0 & exptobac <.
label define tob 1 "Tobacco spenders" 0 "Tobacco non-spenders"
label values tob tob
```

En términos generales, hay 10 grupos de productos básicos (*tabaco, alimentos, atención médica, educación, vivienda, ropa, entretenimiento, transporte, bienes duraderos* y *otros*) que agotan el presupuesto familiar. La mayoría de los estudios en la literatura sobre desplazamiento han considerado algunos o todos esos para su análisis. Las variables que representan los gastos en esos productos básicos son *exptobac, expfood, exphealth, expeducn, exphousing, expcloths, expentertmnt, exptransport, expdurable* y *expother*, respectivamente, tal como se extraen de los datos de las EGH.

Tenga en cuenta que todas las variables tienen el mismo prefijo *exp*. Esa forma de nombrar las variables simplifica el análisis posterior. Para comparar la participación promedio en el presupuesto dedicada a esos productos entre los consumidores de tabaco y los no consumidores de tabaco, se define una variable de participación en el presupuesto para cada grupo de productos básicos. Dados los gastos totales en todos los artículos juntos como *exptotal*, la participación en el presupuesto en cada grupo de productos se puede generar con el siguiente comando de loop:

```
#delimit;
local items "tobac food health educn housing cloths entertmnt transport durable other";
foreach X of local items{ ;
gen bs_ `X'=(exp `X'/exptotal) ;
} ;
```

Se definirán nuevas variables para la participación en el presupuesto con el prefijo (*bs_*) para cada una de esas categorías de productos.

Paso 2: Probar si la diferencia en la participación promedio en el presupuesto es estadísticamente significativa

Una prueba estadística de la igualdad de la participación promedio en el presupuesto entre dos grupos (consumidores de tabaco y no consumidores de tabaco) es una prueba *t* de Student de dos muestras para la igualdad de la media. La prueba *t* se puede realizar en Stata con el comando *<ttest bs_food, by(tob) unequal>*, donde *tob* es la variable binaria que indica el estado del gasto en tabaco definido en el Paso 1. Eso comparará la parte del presupuesto dedicada a alimentos por parte de los hogares que gastan en tabaco y los hogares que no gastan en tabaco y probará si la diferencia es estadísticamente significativa. La hipótesis nula es que la diferencia en la participación promedio en el presupuesto es igual a 0. También se reporta el estadístico *t* para la diferencia de medias. Como regla general, si el valor absoluto de *t* es mayor que 2, se rechaza la hipótesis nula y se puede concluir que la diferencia en la participación promedio en el presupuesto observada es estadísticamente significativa.

La prueba *t*, sin embargo, no permite el uso de ponderadores de encuesta. Tampoco permite el uso del comando <*svy*> de Stata. Como resultado, las participaciones promedio en el presupuesto calculadas para los consumidores y no consumidores de tabaco con el comando <*ttest*> pueden estar sesgadas. Sería ideal calcular las participaciones en el presupuesto para ambos grupos después de ponderarlos con ponderaciones de encuesta apropiadas o usar el prefijo "svy" después de declarar el diseño de la encuesta de los datos con el comando <*svyset*> como se explica en el Capítulo 2. La prueba t anterior en este caso se puede hacer de la siguiente manera:

```
mean bs_food [pw=weight], over(tob)
lincom _b[c.bs_food@0.tob] - _b[c.bs_food@1.tob]
```

Aquí, *weight* es la variable para el ponderador de la encuesta. El comando <*lincom*> informa la diferencia en la participación promedio ponderada en el presupuesto entre los dos grupos y muestra la prueba *t*, así como el valor *p* para la hipótesis nula de que la diferencia en la media es igual a 0. Ese método producirá estimaciones idénticas a las de la prueba *t* si no se usaran ponderadores.

En lugar de usar ponderadores en el comando anterior, el comando <*svy: mean bs_food, over(tob)*> también se puede usar después de declarar el diseño de la encuesta. Alternativamente, se puede usar el comando <*test (_b[c.bs_food@0.tob] = _b[c.bs_food@1.tob])*>, que realiza una prueba de Wald en lugar de la prueba t realizada por <*lincom*>. Dado que se están estimando las participaciones promedio en el presupuesto de la EGH, se debe usar una opción de la prueba que permita usar el ponderador o usar el prefijo "svy" en lugar de usar una prueba t directa que no permita en absoluto usar ponderadores

Paso 3: Informe de los resultados de las pruebas

A efectos de la presentación de informes, solo es necesario conocer la participación promedio en el presupuesto para los grupos de productos básicos determinados, la diferencia en las participaciones promedio en el presupuesto y la significancia estadística de la diferencia según lo indicado por el valor de estadístico *t*. A continuación, se proporciona un programa para los diez grupos de productos básicos:

```
#delimit;
local items tobac food health educn housing cloths entertmnt transport durable other;
local nvar: word count `items';
matrix B = J(`nvar', 4, .);
forvalues I = 1/`nva" {;
local X: word `" of `item";
qui mean bs_ `" [pw=weight], over(tob);
matrix tmp=r(table);
matrix B[`", 1] = tmp[1,1];
matrix B[`", 2] = tmp[1,2];
qui lincom _b[c.bs_ `"@0.tob] - _b[c.bs_ `"@1.tob];
matrix B[`", 3] = r(estimate);
matrix B[`", 4] = r(t);
};
matrix rownames B = `item";
matrix colnames B = non-spenders spenders Difference t-stat;
matrix list B;
```

El código anterior presentará una tabla con las participaciones en el presupuesto para gastos no relacionados con el tabaco, el gasto, la diferencia en las participaciones en el presupuesto y el estadístico t para la prueba de igualdad de las participaciones medias en el presupuesto entre consumidores y no consumidores de tabaco para cada uno de los grupos de bienes en la macro local *items*.

4.4 Un marco para examinar empíricamente el desplazamiento

La prueba *t* simple de igualdad de medias, como se discutió en la sección anterior, no controla otras características específicas del hogar que pueden influir en las decisiones de asignación presupuestaria. Al no controlarlos, es posible atribuir inadvertidamente las decisiones de asignación a los hábitos de consumo de tabaco de un hogar. Por esa razón, existe la necesidad de un modelo econométrico formal que pueda explicar si los hogares que gastan en tabaco reducen sistemáticamente sus gastos en otros grupos de productos básicos y, de ser así, cuáles. Esta sección describe el enfoque conceptual y econométrico que se sigue en la mayor parte de la literatura actual para estimar el grado de desplazamiento debido al gasto en tabaco. Además, la sección discute algunas mejoras metodológicas en la literatura existente sobre ese tema.

4.4.1 Un marco teórico para examinar el desplazamiento

La teoría microeconómica enseña que la solución a la maximización de la utilidad de un individuo sujeto a una restricción presupuestaria devuelve un conjunto de funciones de demanda marshallianas de la forma:

$$q_i = f^i(p_1, \dots, p_n, Y; h) \quad \forall\, i = 1\ to\ n \quad (4.1)$$

donde q_i es la cantidad consumida del bien *i*, *Y* es el gasto total, *h* es un vector de características y $p_1,\dots,p_n$ son los precios de n productos básicos en la función utilidad de un individuo. Dado que los gastos del hogar se informan para todo el hogar como una sola unidad, se usa una función de demanda a nivel del hogar y se necesita el supuesto de que el hogar busca maximizar una sola función de utilidad. Si la demanda de un hogar por uno de los bienes, por ejemplo de tabaco, está predeterminada, son funciones de demanda condicional. El marco teórico para eso se detalla en Pollak (1969).[9] La idea es que el hogar maximizaría la siguiente función utilidad:

$$Max\ U = U(q_1, \dots, q_n; a) \qquad s.a. \sum_{i=1}^{n-1} p_i q_i = M \quad \& \quad q_n = \bar{q}_n \quad (4.2)$$

donde $\bar{q}_n$ denota la demanda de tabaco de un hogar y $M=Y-(p_n*\bar{q}_n)$. Resolviendo eso para $n-1$ bienes se obtiene la siguiente función de demanda condicional, que está condicionada al consumo del bien *n* (tabaco, en este caso):

$$q_i = g^i(p_1, \dots, p_{n-1}, M; \bar{q}_n; h) \quad \forall\, i \neq n \quad (4.3)$$

La función de demanda de cualquier bien dado (q_i) aquí está condicionada a los precios de todas las mercancías excepto el bien condicionante (q_n), gasto restante total (*M*) después de deducir los gastos en el bien condicionante, cantidad del bien condicionante ($\bar{q}_n$), y un vector de características del hogar (*h*). Cuando se trata de bienes que no son consumidos por muchos hogares (como el tabaco), es conveniente utilizar funciones de demanda condicional, como señalan Browning y Meghir.[145]

4.4.2 El modelo econométrico para examinar el desplazamiento

Esta sección analiza una ecuación econométrica específica que se estima para examinar el impacto del desplazamiento y una breve descripción de los posibles métodos de estimación que se utilizan en la literatura hasta el momento, junto con sus deficiencias. Luego propone un método de estimación alternativo que es más eficiente y teóricamente preferido.

Especificación del modelo econométrico

La implementación empírica del modelo requiere el uso de una forma funcional específica. La literatura sobre el desplazamiento ha utilizado en gran medida QAIDS[146] para estimar el impacto del desplazamiento. Dado que a menudo no se dispone de información directa sobre los precios de los diferentes grupos de productos de las encuestas de hogares, para la especificación econométrica se utilizan las curvas de Engel, que permiten trabajar con los gastos. QAIDS, con la presencia de un término de ingreso cuadrático, si bien es consistente con la teoría de la utilidad, permite que los bienes sean de lujo en algunos niveles de ingreso y necesarios en otros.[117] La curva de Engel condicional toma la siguiente forma para el bien *i* y el hogar *j*:

$$w_{ij} = \alpha_{1i} + \alpha_{2i} p_{nj} \bar{q}_{nj} + \delta_i' h_j + \beta_{1i} lnM_j + \beta_{2i} (lnM_j)^2 + u_{ij} \quad (4.4)$$

donde $w_{ij}=p_{ij}q_{ij}/M_j$ es la participación en el presupuesto asignada para el hogar *j* para el grupo *i* de productos básicos del presupuesto restante (M_j) después de deducir los gastos en tabaco, $p_{nj}\bar{q}_{nj}$ es el gasto en tabaco, h_j es un vector de características del hogar que permite que las preferencias sean heterogéneas,[147] lnM_j y lnM_j^2 son los logaritmos naturales de M_j y M_j^2, que son el gasto después de deducir los gastos con tabaco, y u_{ij} es el término de error aleatorio.

Método de estimación 1: Estimación de variables instrumentales ecuación por ecuación (2SLS)

El modelo, según lo especificado en la Ecuación 4.4, no puede ser estimado con el método MCO, ya que las variables $p_n \bar{q}_n$ y *lnM* son probablemente endógenas debido a la simultaneidad involucrada. Si ese es el caso, esas variables se correlacionarán con el término de error u_{ij} y podrían resultar en estimaciones de MCO sesgadas e inconsistentes. En otras palabras, la suposición fundamental de MCO de que el término de error del modelo no está correlacionado con los regresores (es decir, $E(u/x)=0$) se viola y las estimaciones de MCO no dan una interpretación causal. En tales casos, si se pueden encontrar variables exógenas que están correlacionadas con esos regresores endógenos pero que no están correlacionadas con el término de error (VI), se podría usar el método VI para estimar los parámetros de manera más consistente. Eso también se denomina a veces estimación de mínimos cuadrados en dos etapas (2SLS).

Sin embargo, el estimador de VI es menos eficiente que MCO y debe usarse solo si hay variables endógenas presentes en el modelo. Eso se puede probar con la prueba de exogeneidad de Durbin-Wu-Hausman (DWH),[148] si los errores son homocedásticos. Si los errores son heterocedásticos, se suelen utilizar diferentes pruebas, como la prueba de Wooldridge, una prueba basada en una regresión auxiliar o el estadístico *C*, según el tipo de heterocedasticidad asumida.[149] Todos los estudios en la generación actual de literatura sobre desplazamiento muestran que esas variables son, de hecho, endógenas.

La estimación de VI proporciona un estimador consistente bajo la fuerte suposición de que existe un instrumento *z* válido que satisface dos condiciones: (1) el instrumento *z* está parcialmente correlacionado con los regresores endógenos *x* (es decir, $Cov(x_i,z_j)\neq 0$) y (2) el instrumento *z* afecta la variable dependiente w_i solo a través de los regresores o *z* en sí mismo no causa w_i (es decir, $E(u/z)=0$). La primera condición a veces se denomina restricción de inclusión, mientras que la segunda condición se conoce popularmente como restricción de exclusión. Si bien la restricción de inclusión se puede probar estadísticamente al verificar la

asociación entre un instrumento (z) y las variables endógenas (x) con una regresión de forma reducida (cuanto más fuerte sea la asociación, más fuerte será la identificación del modelo), probar la restricción de exclusión es imposible, especialmente en el caso recién identificado (es decir, cuando el número de instrumentos es igual al número de regresores endógenos).

En el caso de sobreidentificación (es decir, cuando hay más instrumentos que el número de regresores endógenos), se puede realizar una prueba de restricciones de sobreidentificación para probar la exogeneidad de los instrumentos, siempre que los parámetros del modelo se estimen utilizando el método de momentos generalizado (GMM).[16] Esa prueba nuevamente difiere dependiendo de si los errores son homocedásticos. Si los errores son homocedásticos, se debe realizar una prueba de Sargan. Si no, se utiliza el estadístico J de Hansen o el estadístico de Hansen-Sargan. Si el estadístico de prueba es estadísticamente significativo, indica que los instrumentos pueden no ser válidos; eso puede suceder si los instrumentos no son verdaderamente exógenos o porque son excluidos incorrectamente de la regresión.[149]

Incluso si hay instrumentos válidos y coeficientes consistentes en la estimación, su matriz de covarianza puede ser inconsistente si los errores son heterocedásticos.[149] El estadístico de Pagan-Hall se puede utilizar para probar la presencia de heterocedasticidad en la regresión de VI. Bajo la hipótesis nula de homocedasticidad, el estadístico de Pagan-Hall se distribuye como χ^2, independientemente de la presencia de heterocedasticidad en otras partes del sistema.[149] Un estadístico significativo implicará la presencia de heterocedasticidad. Si ese es el caso, se tendrá que usar un error estándar consistente con la heterocedasticidad mientras se emplea una estimación VI ecuación por ecuación. Las estimaciones de los coeficientes, así como sus errores estándar, serán consistentes. Eso se puede hacer por medio de una estimación 2SLS o GMM, a la que Wooldridge[15] se refiere como un "estimador de un sistema 2SLS", y que es más eficiente que el estimador de VI simple[149] en presencia de heterocedasticidad.

Método de estimación 2: Sistema de estimación de variables instrumentales (3SLS)

Para estimar un sistema de curvas de Engel (una para cada grupo de productos básicos, para encontrar dónde y cómo se produce el desplazamiento), es necesario estimar una ecuación para cada grupo de productos básicos que se va a considerar. Cada una de esas ecuaciones tendría el gasto en tabaco como un bien condicionante junto con *M* y otras características específicas del hogar, como se muestra en la Ecuación 4.4.

Dado que los regresores en cada ecuación son los mismos, el sistema de ecuaciones es muy parecido a una regresión aparentemente no relacionada (SUR) con la adición del método VI, que es efectivamente un método de mínimos cuadrados de tres etapas (3SLS).[150] Bajo el supuesto de que los errores son homocedásticos, 3SLS proporciona una estimación más eficiente en comparación con 2SLS+VI al explotar la correlación cruzada de errores.[16] La literatura ha usado consistentemente ese método, en oposición al uso de VI en SUR. Una buena descripción de la estimación del sistema 3SLS, que también se denomina 3SLS tradicional, se puede encontrar en Wooldridge[15] Capítulo 8.

Método de estimación 3: Estimación GMM 3SLS

El estimador 3SLS tradicional, según Wooldridge,[15] es menos eficiente y su estimador de varianza es inapropiado si los errores son heterocedásticos. En las encuestas transversales descritas en el Capítulo 2, la heterocedasticidad es la norma y no la excepción. Un estimador de sistema que es consistente y más eficiente que el estimador 3SLS tradicional en presencia de heterocedasticidad es un estimador GMM, y Wooldridge[15] lo llama el estimador "GMM 3SLS". Eso extiende el estimador 3SLS tradicional al permitir la heterocedasticidad y diferentes instrumentos para diferentes ecuaciones.[151] La estimación GMM permite la selección de diferentes matrices de ponderación con las que obtener estimadores que pueden tolerar

heterocedasticidad , conglomeración, autocorrelación y otras violaciones clásicas del término de error *u*. El 3SLS tradicional, por ejemplo, es un estimador GMM que utiliza una matriz de ponderación particular, que asume que los errores son independientes e idénticamente distribuidos (i.i.d.).[15] Sin embargo, al igual que los estimadores VI/3SLS, el estimador GMM también puede tener malas propiedades en muestra finita.[149]

Según Wooldridge,[15] el estimador GMM 3SLS que usa la matriz de ponderación consistente de heterocedasticidad nunca es peor, asintóticamente, que el 3SLS tradicional; y en algunos casos importantes es estrictamente mejor. Sin embargo, la literatura previa sobre desplazamiento parece haber ignorado una prueba de heterocedasticidad en el modelo 3SLS que se utilizó y estimó el modelo 3SLS tradicional asumiendo que los errores son i.i.d. Eso puede haber producido estimaciones de parámetros menos eficientes si la heterocedasticidad estaba realmente presente en esos modelos. Literatura más reciente,[130, 131] sin embargo, ha hecho uso de los métodos GMM-3SLS para estimar el impacto de desplazamiento del gasto en tabaco.

Prueba de heterogeneidad en las preferencias entre consumidores y no consumidores de tabaco

Típicamente, en los datos de EGH, hay una gran cantidad de ceros o valores faltantes contra los gastos en tabaco. Eso puede deberse a que los precios del tabaco son actualmente inaccesibles para algunos hogares debido a restricciones en su presupuesto (también conocida como *solución de esquina*), o debido a la abstención (es decir, el tabaco no está en la función de utilidad de un hogar o en su canasta de consumo, no importa cuál sea el precio). Si se trata del último caso, los consumidores y no consumidores de tabaco tienen preferencias fundamentalmente heterogéneas. Teóricamente, no hay razón *a priori* por la que uno deba asumir cualquiera de los dos casos. Sin embargo, junto con la estimación del desplazamiento, para permitir también la heterogeneidad en las preferencias entre los hogares que gastan en tabaco y los que no gastan en tabaco, la Ecuación 4.4 puede complementarse con la adición de una variable binaria que indique el estado de consumo de tabaco, como en alguna literatura,[117, 126, 148] como sigue:

$$w_{ij} = \left(\alpha_{1i} + \alpha_{2i} d_j + \alpha_{3ij} p_{nj} \bar{q}_{nj} + \delta_i' h_j\right) + \left(\beta_{1i} + \beta_{2i} d_j\right) lnM_j + \left(\gamma_{1i} + \gamma_{2i} d_j\right)(lnM_j)^2 + u_{ij} \qquad (4.5),$$

donde *d* es un indicador binario que toma el valor 1 si un hogar gasta en tabaco y 0 en caso contrario.

Si los parámetros asociados con la variable binaria *d* son conjuntamente significativos —es decir, si la hipótesis nula $H_0: \alpha_{2i} = \beta_{2i} = \gamma_{2i} = 0$ es rechazada— se puede concluir que el tabaco no está en la función utilidad de aquellos hogares que actualmente reportan gastos cero en tabaco. En otras palabras, tanto los consumidores como los no consumidores de tabaco tienen funciones utilidad o preferencias que son diferentes entre sí, y la Ecuación 4.5 se usa para estimar el desplazamiento en tal caso.

En la Ecuación 4.5, la variable binaria que indica el consumo de tabaco (*d*), su interacción con los gastos del hogar (*dlnM*) y su término cuadrático ($dlnM^2$) juntos sirven para distinguir entre los hogares que gastan en tabaco y los que no gastan. No obstante, si no se rechaza la hipótesis nula, significa que los coeficientes asociados a la variable binaria de tabaco, y los de las variables de gasto con las que interactúa la binaria de tabaco, no son significativos y que las preferencias no son diferentes para usuarios y no usuarios de tabaco. En ese caso, solo se necesita la especificación de la Ecuación 4.4 para estimar el desplazamiento. La literatura al respecto utiliza una prueba de Wald para probar la significancia conjunta de los tres parámetros después de la regresión.

Si un investigador tiene interés en probar esa hipótesis, la Ecuación 4.5, en lugar de la Ecuación 4.4, debe especificarse en primer lugar. Si la hipótesis $H_0: \alpha_{2i} = \beta_{2i} = \gamma_{2i} = 0$ es rechazada, entonces se debe usar la

especificación de la Ecuación 4.5 para estimar el desplazamiento. En ese caso, los coeficientes asociados a las variables serán diferentes tanto para los consumidores como para los no consumidores de tabaco. En otras palabras, las preferencias son efectivamente heterogéneas entre los hogares que gastan y los que no gastan en tabaco y que los que no gastan en tabaco no tienen tabaco en su función de utilidad, sin importar cuál sea su precio.

Si, por el contrario, no se rechaza la hipótesis, se puede proceder con la especificación de la Ecuación 4.4, en cuyo caso tanto los hogares que gastan tabaco como los que no gastan tendrán las mismas estimaciones de parámetros. En otras palabras, no hay razón para indagar sobre el desplazamiento del gasto en tabaco en el caso de aquellos hogares para los cuales el tabaco no forma parte de su función utilidad o canasta de consumo, sin importar cuál sea su precio.

4.4.3 Limitaciones del modelo y desarrollos recientes

La discusión de diferentes métodos para estimar el desplazamiento en la Sección 4.4.2 asume la disponibilidad de VI adecuadas para abordar la endogeneidad presente en la especificación del modelo. Sin embargo, encontrar una VI adecuada que cumpla con los requisitos econométricos necesarios a menudo puede ser desafiante y, a veces, es posible que uno no pueda encontrarlas en absoluto. De hecho, existe literatura que estima el desplazamiento ignorando tal endogeneidad,[120, 121, 129, 133] a menudo debido a la falta de disponibilidad de VI adecuadas. Sin embargo, las regresiones que ignoran la presencia de variables endógenas pueden dar como resultado estimaciones de parámetros que conducen a una inferencia incorrecta. En tales casos, se pueden adoptar métodos menos sofisticados. Uno de esos métodos es una comparación simple de las proporciones del presupuesto entre los consumidores y no consumidores de tabaco en varios artículos de compra utilizando una prueba *t*, como ya se describió en la Sección 4.3. También es posible comparar los gastos absolutos asignados a diferentes artículos

Literatura reciente sobre desplazamiento,[132, 134, 135] sin embargo, ha hecho uso de otros métodos que pueden no necesitar el uso explícito de variables instrumentales para examinar el desplazamiento. Esos incluyen métodos tales como un modelo de emparejamiento genético no paramétrico,[134, 149] y modelo de Diferencias en Diferencias Emparejadas (MDID)[132, 135, 150–152] como alternativas. Una de las principales críticas de esos estudios hacia la literatura existente sobre el desplazamiento es que las VI utilizadas en la literatura actual (como la proporción de sexos en adultos o la prevalencia regional del tabaquismo) son imperfectas y, a menudo, no es posible probar la restricción de exclusión necesaria para esas VI. Sin embargo, como ya se señaló en la Sección 4.4.2, cuando hay más instrumentos que el número de regresores endógenos, se puede realizar una prueba de sobreidentificación de restricciones para probar la exogeneidad de los instrumentos, siempre que los parámetros del modelo se estimen utilizando el GMM óptimo.[16] El modelo GMM-3SLS propuesto en este conjunto de herramientas también lo permite. Sin embargo, encontrar más instrumentos que el número de regresores endógenos a menudo puede ser un desafío.

En los casos de los modelos recién identificados, sería imposible probar la restricción de exclusión. Idealmente, los métodos que no dependen de la obtención de instrumentos adecuados serían preferibles en tales situaciones, y la literatura reciente aborda esa preocupación. Sin embargo, se debe tener en cuenta que la implementación eficiente de MDID requiere datos de panel o secciones transversales repetidas que se puedan convertir en *pseudopaneles*. Esa puede ser una limitación importante para los países con una sola ronda de EGH transversales. Dado que MDID imita el diseño de investigación experimental utilizando datos observacionales, debería ser posible agrupar de manera efectiva a los hogares en los datos en grupos de tratamiento y control, ambos con características sociodemográficas idénticas distintas al estado de tratamiento, que, en este caso, sería el estado de consumo de tabaco.

Dado que el análisis de desplazamiento explicado anteriormente compara las participaciones en el presupuesto de diferentes productos básicos solo por hogares que gastan y que no gastan en tabaco, no arroja mucha luz sobre las asignaciones dentro del hogar como resultado del desplazamiento. Esa es otra limitación de este análisis. Por ejemplo, el análisis puede mostrar que el gasto en salud o en educación se ve desplazado como resultado del gasto en tabaco. Pero es difícil determinar qué miembro del hogar se ve afectado debido a ese desplazamiento. El hecho de que el análisis solo considere grupos más agregados de productos básicos hace que tales consideraciones dentro del hogar sean aún más difíciles de examinar.

4.5 Preparación de datos para el análisis

Si bien el Capítulo 2 proporcionó información detallada sobre la extracción de datos, su limpieza, la fusión de variables que provienen de diferentes conjuntos de datos y otros consejos necesarios para la gestión de datos, es importante brindar detalles específicos sobre las variables necesarias para el análisis en este capítulo. Para cualquier variable nueva que se discuta aquí, es importante llevarla por todos los procesos discutidos en el Capítulo 2. Esta sección explica cómo se pueden generar las variables específicas requeridas para el análisis de desplazamiento usando las variables estándar disponibles de las EGH. También muestra formas de clasificar los hogares para satisfacer las necesidades analíticas específicas de este capítulo.

Las variables más importantes requeridas son los gastos en tabaco, así como otros grupos de productos básicos mencionados anteriormente, que deben probarse para determinar si se produce un desplazamiento. Esos están disponibles directamente desde cualquier EGH. A continuación, deben construirse las participaciones en el presupuesto de cada uno de los grupos de productos básicos restantes, después de restar el gasto en tabaco. Por ejemplo, se puede crear una variable para la participación en el presupuesto de alimentos en Stata usando el código *<generate bsfood = expfood/exp_less>* donde *bsfood* es la variable de participación en el presupuesto de alimentos, que se usará como variable dependiente en la regresión, *expfood* son los gastos en alimentos que se extraen de la EGH, y *exp_less* son los gastos totales en todos los artículos (*exptotal*) menos los gastos en tabaco (*exptobac*). Para todos los grupos de productos básicos juntos, se puede usar un loop para generar la participación en el presupuesto de la siguiente manera:

```
#delimit;
gen exp_less = exptotal – exptobac ;
local items "food health educn housing cloths entertmnt transport durable other";
foreach X of local items{ ;
    gen bs`X'=(exp`X'/exp_less) ;
    } ;
```

Esas son las variables que entrarían en la regresión (VI, 3SLS o GMM 3SLS) como variables dependientes. Eso es diferente de las variables de participación presupuestaria creadas en la Sección 4.3 para la prueba *t*, ya que tenían el gasto total como denominador. Aunque los gastos en diferentes productos están disponibles directamente de la EGH, es posible que la EGH no informe esos datos al nivel de agregación requerido. Por ejemplo, los gastos en alimentos pueden registrarse en la EGH como gastos en muchos otros artículos alimentarios. Si la información agregada no está disponible, es posible que deba agregar gastos en artículos más pequeños para crear grupos agregados como los que se presentan aquí. Después de todo, tener demasiados productos básicos desagregados puede no servir de mucho, desde el punto de vista de las políticas públicas, al analizar el impacto de desplazamiento del gasto en tabaco. Sin embargo, dependiendo de las circunstancias socioeconómicas de cada país, la selección de grupos de productos podría variar.

Es necesario generar los logaritmos naturales y los cuadrados de las variables *exptotal* y *exp_less* que se utilizarán en la regresión. Hay que identificar las variables específicas a nivel del hogar para usar como controles y las variables que normalmente pueden funcionar como instrumentos para las variables endógenas en el modelo 3SLS. La literatura ofrece alguna orientación. Algunas de las variables sociodemográficas comunes a nivel del hogar utilizadas en esta literatura incluyen logaritmo del tamaño del hogar; proporción de adultos (ratio entre el número de adultos y el tamaño del hogar); edad promedio del hogar; educación promedio (educación total recibida por todos los miembros en años dividida por el tamaño del hogar) del hogar; educación máxima (años de educación recibidos por el miembro más educado del hogar); variables dummy para caracterizar los hogares en diferentes grupos sociales, étnicos, ocupacionales, religiosos y de ingresos; y una variable dummy para indicar el área de residencia de un hogar, como zonas rurales o urbanas, entre otras.

Elegir las variables adecuadas para que sirvan como instrumentos es uno de los aspectos clave en la preparación de la lista de variables para el análisis. Una vez más, la literatura ofrece alguna orientación. Gran parte de la literatura reciente sobre el desplazamiento[117, 122, 125–127] utiliza los gastos totales del hogar o el valor total de los activos del hogar como instrumento para el gasto del grupo *M* (*exp_less*) y la proporción de hombres adultos o mujeres adultas en el número total de adultos en el hogar (proporción de adultos por sexo) o la proporción de hombres adultos a las mujeres adultas como instrumento para el gasto en tabaco.

Se cree que la proporción de adultos por sexo es un instrumento sensato para el gasto en tabaco, ya que el consumo de tabaco suele ser mucho más frecuente entre los hombres que entre las mujeres en la mayoría de esos países. Por lo tanto, se espera que un aumento en la proporción masculina (proporción de hombres adultos a mujeres adultas) se relacione positivamente con el gasto en tabaco, y no es algo que pueda afectar directamente la participación en el presupuesto en otros grupos de productos básicos para los cuales se estima el impacto de desplazamiento. Pero en países donde la tasa de tabaquismo no es significativamente diferente entre géneros, la proporción de sexos puede no ser un instrumento apropiado para usar. En tales casos, se han adoptado enfoques alternativos. Algunos estudios[123, 138] han utilizado una medida compuesta de prevalencia e intensidad del tabaquismo como instrumento para el gasto en tabaco. Cualquier variable exógena que aparezca en el lado derecho de las otras ecuaciones en el modelo puede servir potencialmente como un instrumento para estimar la variable endógena en el lado derecho de la ecuación. Independientemente de la variable que se utilice como instrumento, es importante verificar que los instrumentos seleccionados estén correlacionados con la variable endógena del lado derecho y que no tengan un efecto directo sobre la variable dependiente.

4.6 Estimando el desplazamiento con Stata

Esta sección demuestra los diferentes métodos de estimación (3SLS tradicional, 3SLS GMM y VI ecuación por ecuación) discutidos en la Sección 4.4 para estimar los efectos de desplazamiento. Primero, analiza la configuración general de las variables que se pueden usar en todos los métodos. Después de una discusión sobre la implementación de los tres métodos de estimación, se analizan las pruebas de varios requisitos del modelo, incluida la validez de los instrumentos y la heterocedasticidad , entre otros. Los resultados de esas pruebas guiarán la decisión sobre el tipo de método de estimación a utilizar.

Como se detalló anteriormente, según las propiedades de los datos, existen diferentes estrategias de modelado. A continuación, se presentan algunas variables que son necesarias para estimar la Ecuación 4.4:

```
gen pq=exptobac
gen lnM=log(exp_less)
gen lnX=log(exptotal)
gen lnM2=lnM*lnM
gen lnX2=lnX*lnX
```

Además, para simplificar el modelo de regresión para estimar las estimaciones de 3SLS tradicional o GMM 3SLS o VI, es útil crear ciertas macros globales que indiquen la lista de variables dependientes, variables endógenas, variables exógenas e instrumentos en el modelo. Por ejemplo, para estimar el impacto del desplazamiento entre ocho grupos de productos básicos (alimentos, salud, educación, vivienda, vestimenta, entretenimiento, transporte y bienes duraderos), dejando fuera el grupo de "otros", como se hace comúnmente en la literatura, las siguientes macros se definen:

```
global ylist bsfood bshealth bseducn bshousing bscloths bsentertmnt bstransport bsdurable
global x1list pq lnM lnM2
global x2list hsize meanedu maxedu sd1-sd3
global zlist asexratio lnX lnX2
```

La macro *ylist* incluye las variables dependientes que entran en la regresión, *x1list* incluye las variables endógenas del lado derecho como se explica en la Ecuación 4.4 (esas son variables que se sospecha que sean endógenas), *x2list* incluye las variables exógenas (tamaño del hogar, educación media, máx. educación, tres variables dummy para representar el estado socioeconómico de los hogares), y *zlist* incluye las VI para corregir la endogeneidad en el modelo (proporción de adultos por sexo, logaritmo del gasto total y logaritmo del gasto total al cuadrado, en este caso). En el modelo, sin embargo, cada variable exógena puede ser un instrumento por su cuenta. El número de variables en *zlist* debe ser al menos tan grande como el de *x1list* para que el modelo sea identificado. Las variables utilizadas en las macros globales aquí son solo para fines de demostración. En el análisis real puede haber un número mayor o menor de variables en cualquiera de las listas anteriores. Por ejemplo, el *x2list* puede contener otras características específicas del hogar además de las presentadas aquí.

4.6.1 Estimación de 3SLS

Una vez que se crean esas macros globales, la estimación del modelo 3SLS en Stata se puede hacer simplemente usando el comando <*reg3*>. La ayuda de Stata sobre reg3: <*help reg3*> proporciona una sintaxis detallada y ejemplos útiles para usar ese comando. No obstante, para ese propósito, una vez que se definen las macros globales como se indicó anteriormente, solo se necesita usar el siguiente comando para obtener las estimaciones de 3SLS:

```
reg3 ($ylist = $x1list $x2list), exog($zlist) endog($x1list) 3sls
```

donde las opciones *exog* y *endog* especifican la lista de regresores exógenos y endógenos en el lado derecho de cada una de las ecuaciones. Sin el uso de macros globales, ese comando también podría escribirse como:

```
reg3 (bsfood bshealth bseducn bshousing bscloths bsentertmnt bstransport bsdurable =
exptobac lnexp_less lnexp_less2 hsize meanedu maxedu sd1-sd3), exog(asexratio
lnexptotal lnexptotal2) endog(exptobac lnexp_less lnexp_less2) 3sls
```

Es esencial que el código esté en una sola línea en el do-file o debe dividirse con delimitadores apropiados para marcar el final del comando y aceptables para Stata. El uso de macros hace que el código sea mucho más ordenado. Además, no hay razón para usar un comando de regresión separado para cada ecuación, incluso si los instrumentos varían para algunas de las ecuaciones. Todos los instrumentos se pueden juntar en la lista *exog* mientras se usa el comando *reg3*.

Como se señaló anteriormente, 3SLS es un estimador GMM que utiliza una matriz de ponderación particular que asume errores i.i.d. Por lo tanto, los resultados de 3SLS anteriores del comando *<reg3>* se pueden reproducir con una estimación de GMM con una matriz de ponderación pertinente. Esto se hace en el siguiente código:

```
gmm (eq1: bsfood - {food: $x1list $x2list _cons}) ///
(eq2: bshealth - {health: $x1list $x2list _cons}) ///
    (eq3: bseducn - {educn: $x1list $x2list _cons}) ///
    (eq4: bshousing - {housing: $x1list $x2list _cons}) ///
    (eq5: bscloths - {cloths: $x1list $x2list _cons}) ///
    (eq6: bsentertmnt - {entertmnt: $x1list $x2list _cons}) ///
    (eq7: bstransport  - {transport: $x1list $x2list _cons}) ///
    (eq8: bsdurable - {durable: $x1list $x2list _cons}) ///
    , instruments($zlist $x2list) ///
    winitial(unadjusted, independent)  wmatrix(unadjusted) twostep
```

La opción *<winitial()>* especifica la matriz de ponderación que se usará para obtener las estimaciones de parámetros del primer paso. La subopción *<independent>* le dice a GMM que suponga que los residuos son independientes en las condiciones de los momentos. La opción *<wmatrix()>* controla cómo se calcula la matriz de ponderación sobre la base de las estimaciones del primer paso antes del segundo paso de la estimación. Al especificar *<wmatrix(unadjusted)>*, se exige una matriz de ponderación que asume homocedasticidad condicional, pero no impone la independencia de las ecuaciones cruzadas como la matriz de ponderación inicial.[151] Tenga en cuenta que el código *<gmm>* anterior podría tardar mucho más de lo que tardaría *<reg3>* en converger en una solución, incluso varias horas, dependiendo de la capacidad física de la computadora. Eso se debe a que GMM, a diferencia de 3SLS, es un estimador muy general y no lineal, y busca numéricamente una solución.

4.6.2 Estimación de GMM 3SLS

Si los errores son heterocedásticas, eso significa que las estimaciones 3SLS tradicionales son menos eficientes y sus errores estándar son inconsistentes. En ese caso, se debe utilizar una matriz de ponderación consistente con la heterocedasticidad para obtener estimaciones de parámetros consistentes. Eso es posible con GMM usando la opción *<wmatrix(robust)>* como se implementa en el siguiente código:

```
gmm (eq1: bsfood - {food: $x1list $x2list _cons}) ///
    (eq2: bshealth - {health: $x1list $x2list _cons}) ///
    (eq3: bseducn - {educn: $x1list $x2list _cons}) ///
    (eq4: bshousing - {housing: $x1list $x2list _cons}) ///
    (eq5: bscloths - {cloths: $x1list $x2list _cons}) ///
    (eq6: bsentertmnt - {entertmnt: $x1list $x2list _cons}) ///
    (eq7: bstransport  - {transport: $x1list $x2list _cons}) ///
    (eq8: bsdurable - {durable: $x1list $x2list _cons}) ///
    , instruments($zlist $x2list) ///
    winitial(unadjusted, independent)  wmatrix(robust) twostep
```

La opción *wmatrix(robust)* exige una matriz de ponderación apropiada para errores que son independientes, pero no necesariamente idénticamente distribuidos. También es posible exigir una matriz de ponderación que tenga en cuenta la correlación arbitraria entre las observaciones dentro de los conglomerados, como suele observarse en los datos de las encuestas. Para ese propósito, la opción se puede modificar a *<wmatrix(cluster clustvar)>*, donde clustvar es el nombre de la variable que identifica los conglomerados en los datos.

En lugar de los errores estándar robustos en *<gmm>*, también se pueden obtener errores estándar mediante bootstrap usando *<reg3>* con un prefijo de bootstrap. Por ejemplo, *<bootstrap, reps(1000) seed(1010):reg3 ($ylist = $x1list $x2list), exog($zlist) endog($x1list) 3sls>*. Eso es mejor que estimar un 3SLS *<reg3>* ignorando la posible heterocedasticidad. Sin embargo, *<reg3>* con 1000 replicaciones de bootstrap puede tardar tanto como *<gmm>* en lograr la convergencia. El uso de *<gmm>*, por otro lado, tiene la ventaja adicional de especificar una matriz de ponderación que da cuenta de la heterocedasticidad del conglomerado y la autocorrelación.

Los modelos implementados anteriormente son modelos identificados justamente, ya que el número de instrumentos es igual al número de variables de lado derecho endógenas. Si, en cambio, hay un modelo sobreidentificado, la implementación del código de Stata sería la misma, excepto que los nombres de esos instrumentos adicionales se agregarían a la lista de VI en la macro global *zlist*.

4.6.3 VI ecuación por ecuación

Como se indicó en la Sección 4.4, una alternativa a hacer una estimación del sistema, como en el 3SLS tradicional, es hacer la estimación para cada ecuación, una por una, usando 2SLS. Eso se puede implementar con la ayuda del comando *<ivregress>* de Stata de la siguiente manera:

```
#delimit;
local depvar "food health educn housing cloths entertmnt transport durable";
foreach X of local depvar{;
    ivregress 2sls bs `X' $x2list ($x1list = $zlist);
};
```

Stata también tiene un excelente comando escrito por usuarios *<ivreg2>*[157] que se puede usar en lugar de *<ivregress>* y ofrece una funcionalidad adicional en comparación con *<ivregress>*. Se puede instalar usando el comando *<ssc install ivreg2>*. La implementación de *<ivreg2>* es bastante similar a la de *<ivregress>*. Por ejemplo, *<ivregress 2sls bsfood $x2list ($x1list = $zlist)>* y *<ivreg2 bsfood $x2list ($x1list = $zlist)>* darían estimaciones idénticas.

VI ecuación por ecuación, que Wooldridge[15] refiere como un "estimador 2SLS de sistema" se puede implementar omitiendo la opción *<twostep> y <wmatrix()>* de la implementación tradicional 3SLS en un comando *<gmm>* como se muestra a continuación. Eso debería generar resultados similares a los obtenidos de *<ivregress>* o *<ivreg2>*, pero con errores estándar robustos.

```
gmm (eq1: bsfood - {food: $x1list $x2list _cons}) ///
(eq2: bshealth - {health: $x1list $x2list _cons}) ///
(eq3: bseducn - {educn: $x1list $x2list _cons}) ///
(eq4: bshousing - {housing: $x1list $x2list _cons}) ///
(eq5: bscloths - {cloths: $x1list $x2list _cons}) ///
(eq6: bsentertmnt - {entertmnt: $x1list $x2list _cons}) ///
(eq7: bstransport  - {transport: $x1list $x2list _cons}) ///
(eq8: bsdurable - {durable: $x1list $x2list _cons}) ///
, instruments($zlist $x2list) ///
winitial(unadjusted, independent)
```

Para ver también errores estándar idénticos a los del comando *<ivregress>*, agregue la opción *<vce(unadjusted) onestep>* después de *<winitial(unadjusted, independent)>*. Si una prueba de heterocedasticidad después de VI ecuación por ecuación indica que los errores no son homocedásticos, entonces se puede usar el estimador 2SLS de sistema con *<gmm>*, dado anteriormente, que arroja errores estándar robustos, o se puede modificar el comando *<ivregress>* con el comando opcional *<vce(robust)>*. Por ejemplo, se puede implementar para la ecuación bsfood como *<ivregress 2sls bsfood $x2list ($x1list = $zlist), vce(robust)>*. El comando *<ivregress>* también permite especificar una matriz de ponderación con el uso del estimador GMM como *<ivregress gmm bsfood $x2list ($x1list = $zlist), wmatrix(robust)>* o con otras especificaciones de la matriz de ponderación, como *<wmatrix(cluster clustvar)>*. Las estimaciones de los coeficientes, así como sus errores estándar, serán consistentes, como se indica en la Sección 4.4.

4.6.4 Realización de diferentes pruebas para decidir el método de estimación

Antes de decidir qué método de estimación en particular se debe utilizar, es importante realizar varias pruebas. Eso incluye una prueba de endogeneidad de variables, una prueba de validez de los instrumentos utilizados y una prueba de homocedasticidad de errores, entre otras. Esas pruebas se implementan más fácilmente después de una estimación VI ecuación por ecuación.

1) Prueba de endogeneidad de los regresores: Como se señaló en la Sección 4.4, no es necesario utilizar un estimador VI a menos que las variables endógenas sean realmente endógenas. La endogeneidad se puede probar con la ayuda de la prueba de exogeneidad DWH[148] en caso de errores i.i.d., o la prueba de Wooldridge o una prueba basada en una regresión auxiliar en el caso de errores no i.i.d.,[149] como se discutió anteriormente. Después del comando *<ivregress>*, se puede usar el comando *<estat endogenous>* para hacer esto. Reportará sobre el estadístico de DWH o cualquiera de los otros estadísticos consistentes con heterocedasticidad discutidos anteriormente, dependiendo de la matriz de ponderación opcional utilizada con el comando *<ivregress>*. En cualquier caso, la hipótesis nula es que las variables son exógenas y un estadístico de prueba significativo indicaría que la variable debe tratarse como endógena.

Del mismo modo, si se usa *<ivreg2>*, el comando *<ivendog>* se puede usar después de *<ivreg2>* e informará el estadístico de DWH. Alternativamente, la opción *<endog(varname)>* se puede usar junto con el comando *<ivreg2>* para probar si un instrumento es endógeno. Por ejemplo, *<ivreg2 bsfood $x2list ($x1list = $zlist), gmm2s robust endogtest($x1list)>* prueba la endogeneidad de las tres variables endógenas, además de

mostrar los resultados de la regresión. Esa opción es particularmente útil para probar la endogeneidad cuando hay heterocedasticidad.

2) Probar la validez de los instrumentos: Como se señaló anteriormente, los estimadores VI son consistentes solo bajo el supuesto muy fuerte de que existe un instrumento válido (*z*) que satisface las restricciones de inclusión y exclusión. Probar la restricción de inclusión es sencillo. Comprueba si los instrumentos son débiles o fuertes. Con el comando <*ivreg2*>, simplemente se necesita agregar la opción <*first*>; por ejemplo, <*ivreg2 bsfood $x2list ($x1list = $zlist), first*>. Eso informaría los resultados de la regresión de la primera etapa, uno para cada regresor endógeno. En este caso, dado que hay tres variables endógenas del lado derecho (*pq, lnM, lnM2*), informaría tres resultados de regresión de primera etapa con cada una de esas variables endógenas como variable dependiente y todos los regresores exógenos restantes y las VI como variables del lado derecho. El R^2 y el estadístico *F* de esas regresiones de primera etapa indican qué tan fuertes o débiles son los instrumentos.

Una regla general común sugiere que un estadístico *F* de menos de 10, en el caso de un único regresor endógeno, es indicativa de un instrumento débil.[16, 154] Si hay un único instrumento y un único regresor endógeno, eso se traduce en un valor *t* de 3,2 o superior y el correspondiente valor *p* de 0,0016 o inferior para el instrumento. Los resultados de esta prueba *F* deben informarse al informar las estimaciones de VI. Esa regla empírica, sin embargo, es *ad hoc* y puede no ser lo suficientemente conservadora si el modelo está sobreidentificado. Para ecuaciones con más de un regresor endógeno, se puede usar una estadístico llamado R^2 parcial de Shea en lugar del valor *F* crítico.[16] Sin embargo, no hay consenso sobre qué tan bajo el valor de R^2 indica un problema.[16] Usar la opción <*first*> después de <*ivreg2*>, así como el comando <*estat firststage*> después de ejecutar <*ivregress*> informa el R^2 parcial de Shea. Ver Cameron y Trivedi[16] Capítulo 6 para una exposición detallada de esas estadísticas. Alternativamente, consulte el manual de referencia de Stata[151] en las notas técnicas posteriores a la estimación de *ivregress* en las páginas 1212–1213.

En general, no es posible probar la restricción de exclusión o probar la exogeneidad de los instrumentos, especialmente en el caso anterior. Sin embargo, en el caso sobreidentificado, se puede realizar una prueba de restricciones sobreidentificadas con el comando <*estat overid*> después de <*ivregress*> o con el comando <*overid*> después de <*ivreg2*>. Informaría los resultados de una prueba de Sargan en el caso de homocedasticidad. Si <*ivregress*> se hubiera usado con la opción <*gmm*> junto con una matriz de ponderación consistente con la heterocedasticidad, entonces <*estat overid*> informaría un estadístico *J* de Hansen o estadístico Hansen-Sargan, que explica las perturbaciones heterocedásticas. Un estadístico de prueba estadísticamente significativo indica que los instrumentos pueden no ser válidos. Eso puede suceder si los instrumentos no son verdaderamente exógenos, o porque se excluyeron incorrectamente de la regresión,[149] como se mencionó anteriormente.

3) Prueba de heterocedasticidad : Como se señaló en la Sección 4.4, si los errores son heterocedásticos, la regresión VI produce errores estándar inconsistentes y las estimaciones 3SLS tradicionales son menos eficientes y los errores estándar son inconsistentes. El estadístico de Pagan-Hall se puede utilizar para probar la presencia de heterocedasticidad en la regresión VI. Eso se puede implementar con el comando <*ivhettest*>. Por ejemplo, después de <*ivreg2 bsfood $x2list ($x1list = $zlist)*>, aplique el comando <*ivhettest*>, y reportaría el estadístico Pagan-Hall con la hipótesis nula de perturbaciones homocedásticas. Un estadístico significativo implicará un rechazo de la hipótesis nula, indicativo de la presencia de heterocedasticidad. Desafortunadamente, el <*ivhettest*> no funciona después del <*ivregress*> hasta ahora.

También hay un programa escrito por usuarios *<lmhreg3>*[159] que se puede instalar con el comando *<ssc install lmhreg3>*, que realiza las pruebas tanto de ecuaciones individuales como de heterocedasticidad general del sistema después del comando *<reg3>*. Entonces, si se usó *<reg3>* para hacer una estimación 3SLS, se puede aplicar el comando *<lmhreg3>* inmediatamente después para verificar si cada una de las ecuaciones individuales, así como el sistema en su conjunto, satisface el supuesto de homocedasticidad. La hipótesis nula es que los errores son homocedásticos y, como de costumbre, un estadístico de prueba significativo (Pagan-Hall u otras pruebas del multiplicador de Lagrange utilizadas en *lmhreg3*) es indicativo de heterocedasticidad.

4) Prueba de heterogeneidad en las preferencias entre consumidores y no consumidores de tabaco: Para examinar si las preferencias son heterogéneas entre los hogares que gastan en tabaco y los que no, se puede estimar la Ecuación 4.5 en lugar de la Ecuación 4.4 para probar la significancia conjunta de los parámetros asociados con el indicador binario para el consumo de tabaco y las interacciones con él. Se traduce en probar la hipótesis nula H_0:$\alpha_{2i}=\beta_{2i}=\gamma_{2i}=0$ en la Ecuación 4.5. Para eso, primero estime el modelo en la Ecuación 4.5 usando *<ivregress>* de la siguiente manera:

```
#delimit;
local depvar "food health educn housing cloths entertmnt transport durable";
foreach X of local depvar{;
    ivregress 2sls bs `X' $x2list tob tob#c.lnM tob#c.lnM2 ($x1list = $zlist);
    test (tob=0) (1.tob#c.lnM=0) (1.tob#c.lnM2=0);
};
```

El comando *<test>* después de cada ecuación sucesiva realiza una prueba de Wald para probar una hipótesis lineal compuesta de que los tres coeficientes asociados con la variable dicotómica *tob* son cero de manera conjunta. Rechazar (es decir, un estadístico de prueba significativo) sugiere que la Ecuación 4.5 puede ser una especificación más apropiada, mientras que no rechazar implicaría que la Ecuación 4.4 puede ser la especificación correcta. Si la prueba concluye que la Ecuación 4.5 es la especificación de elección, todas las pruebas de (1) a (3) anteriores deben realizarse nuevamente en la nueva especificación. Y si la heterocedasticidad está presente, se debe utilizar un método de estimación GMM 3SLS para obtener los parámetros finales.

Resumen de las pruebas y decisión sobre el método de estimación: Para repasar, antes de decidir qué método de estimación usar —ya sea el 3SLS tradicional *<reg3>* o GMM 3SLS *<gmm>* o VI ecuación por ecuación (ya sea con *ivregress* o *ivreg2*)— se recomienda primero estimar VI ecuación por ecuación. Eso permitiría determinar si existe endogeneidad en el modelo y si los instrumentos utilizados son válidos. A continuación, se debe realizar la prueba de heterocedasticidad. Si la prueba de heterocedasticidad indica que los errores son i.i.d., entonces se podría optar por un *<reg3>* para hacer la estimación 3SLS tradicional. De lo contrario, se debe usar un método de estimación GMM 3SLS que use el comando *<gmm>* en Stata para producir estimaciones de parámetros eficientes. Según Wooldridge,[15] el estimador GMM 3SLS que usa la matriz de ponderación consistente a heterocedasticidad nunca es peor, asintóticamente, que el 3SLS tradicional; y en algunos casos importantes es estrictamente mejor. Por lo tanto, sería más seguro utilizar un método de estimación GMM 3SLS para estimar el desplazamiento en cualquiera de los casos.

Finalmente, probar la significancia conjunta de los parámetros asociados con la variable dicotómica para el gasto en tabaco junto con sus variables de interacción indicará si es apropiado usar una forma funcional que trate completamente diferente a los consumidores y los no consumidores de tabaco. Si concluye que deben tratarse de manera diferente, entonces se debe especificar la Ecuación 4.5 y todas las pruebas sugeridas desde (1) hasta (3) deben repetirse en la nueva especificación.

4.6.5 Estimación del desplazamiento por subgrupos

Dado que el consumo de tabaco está más concentrado en las comunidades de bajos ingresos, o que se sabe que las comunidades de bajos ingresos gastan una parte desproporcionadamente mayor de su presupuesto en la compra de productos de tabaco, es posible que el impacto del desplazamiento sea mayor entre comunidades de bajos ingresos. De manera similar, los hogares también pueden clasificarse en términos de la intensidad de su gasto en tabaco en consumidores moderados, medianos y altos. Es posible que el desplazamiento sea mucho mayor entre los que gastan mucho en comparación con los que gastan moderadamente. Por esas y otras razones, los investigadores pueden querer examinar el impacto de desplazamiento por diferentes subgrupos definidos, ya sea por ingresos o por otras características. La literatura ha utilizado diferentes subgrupos para examinar el impacto, incluidos los grupos de ingresos,[117, 127, 134] intensidad del gasto en tabaco,[124] y diferentes tipos de tabaco.[127]

Además de los detalles discutidos hasta ahora, estimar el impacto del desplazamiento por subgrupos requiere solo dos pasos adicionales:

(1) definir una variable categórica que indique el subgrupo; y

(2) agregar la opción de subgrupo al comando de Stata pertinente.

A continuación, se muestran ejemplos.

Paso 1: Definición de variables categóricas para indicar subgrupo

Ese paso ya se discutió en la Sección 3.5.1. Para reiterar, primero es necesario crear una variable de gasto per cápita (*pcexp*) para diferentes hogares usando el comando <*gen exppc=exptotal/hsize*>. Los grupos/cuantiles de gastos domésticos per cápita (como representación de los ingresos) se pueden generar con el comando:

```
<xtile incgrp = exppc [w=weights], nq(3)>
```

donde la opción *nq (.)* especifica el número de cuantiles.

De manera similar, los hogares también pueden clasificarse según la distribución de la participación en el presupuesto del gasto en tabaco en gastadores bajos o altos, y así sucesivamente.

Paso 2: Adición de opciones de subgrupos a los comandos pertinentes de Stata

Una vez que se genera la variable categórica, digamos *incgrp*, la estimación se puede realizar agregando una opción <*by(incgrp)*> o <*over(incgrp)*> o el prefijo <*bysort incgrp:*> a los comandos de Stata, según el comando particular. Por ejemplo, el <*ivregress*> se puede estimar con el prefijo de la siguiente manera:

```
#delimit;
local depvar "food health educn housing cloths entertmnt transport durable"
foreach X of local depvar{
bysort incgrp: ivregress 2sls bs `X' $x2list ($x1list = $zlist)
}
```

También para GMM 3SLS, el prefijo *<bysort incgrp:>* se puede agregar antes del comando *<gmm>*.

La Sección 7.4 en el Apéndice de código proporciona un do-file de ejemplo que detalla el código utilizado en ese capítulo. Los usuarios podrán copiar y pegar eso en el editor de do-files de Stata y podrán estimar los resultados con los datos/variables correspondientes que se describen allí.

4.7 Estudio de caso de Turquía

Los hogares turcos, a pesar de vivir en un país de ingresos medianos-altos, gastaron más del ocho por ciento de su presupuesto familiar en la compra de tabaco en 2011. Mientras que las personas ricas en Turquía gastaron alrededor del 6,2 por ciento de su presupuesto familiar en tabaco, las personas pobres gastaron hasta el 10,7 por ciento.[126] Dado que una gran parte del presupuesto de los hogares se está desviando hacia el gasto en tabaco, es posible que sea compensado con los gastos en otras necesidades del hogar. En ese contexto, San & Chaloupka[126] examinan el desplazamiento del gasto en tabaco en una variedad de grupos de productos básicos en Turquía. El estudio estima el modelo QAIDS con una variante de la Ecuación 4.5 para estimar los efectos del desplazamiento. El modelo econométrico utilizado es el método 3SLS discutido en la Sección 4.4. El estudio utiliza el gasto total como instrumento para el gasto neto de tabaco, y el ratio de mujeres —relación entre mujeres adultas y el total de adultos en los hogares— como instrumento para el gasto en tabaco. La Tabla 4.2 muestra una panorámica de los resultados que encontraron para 2011 y presenta los resultados de solo un subconjunto de los grupos de productos básicos que analizaron los autores. La primera columna debajo del grupo de productos básicos muestra las estimaciones de los parámetros y la segunda columna presenta los errores estándar.

La variable binaria (*d*), que indica el gasto en tabaco, es significativa en todos los productos, excepto la educación. El signo negativo indica que el gasto en tabaco tiene un impacto negativo en el gasto en el grupo de productos básicos correspondiente. Sin embargo, la magnitud del coeficiente de la variable binaria d no proporciona una interpretación directa. Debe interpretarse solo junto con su interacción con otras variables (*dlnM* y $dlnM^2$) y su significado conjunto, como se muestra en la Ecuación 4.5. De hecho, en ese estudio se rechaza una prueba de Wald de significancia conjunta de los tres coeficientes, lo que implica que los hogares que gastan y los que no gastan en tabaco tienen funciones utilidad fundamentalmente diferentes.

Tabla 4.2 Efecto de desplazamiento del gasto en tabaco en Turquía, 2011

	Alimento		Vivienda		Ropa		Transporte		Educación	
	Coef.	**S.E**	**Coef.**	**S.E**	**Coef.**	**S.E**	**Coef.**	**S.E**	**Coef.**	**S.E**
d	0,7616*	-0,196	−0,7572*	-0,365	−0,3641*	-0,098	2,273*	-0,302	−0,0542	-0,094
p.q	−0,0002	0,000	−0,0022*	0,000	−0,0003*	0,000	0,0021*	0,000	−0,0003*	0,000
lnM	0,1045*	-0,003	0,1352*	-0,006	0,0041*	-0,002	−0,0373*	-0,005	−0,0189*	-0,002
lnM^2	−0,0121*	0,000	−0,0135*	-0,001	0,0005*	0,000	0,0092*	-0,001	0,0025*	0,000
dlnM	−0,2004*	-0,055	0,2316*	-0,102	0,0955	-0,027	−0,6456*	-0,084	0,0228	-0,026
$dlnM^2$	0,0122*	0,003	−0,0105*	0,006	−0,0056	-0,002	0,0410*	0,005	−0,0012	-0,002

Notas: Resultados de la especificación en la Ecuación 4.5. Los valores de las variables dependientes van de 0 a 1.
*Esos resultados son significativos al nivel del 5%. Fuente: San & Chaloupka (2016)[65]

La variable *p.q* es el gasto total preasignado en tabaco, y su coeficiente proporciona una indicación del grado de desplazamiento. Por ejemplo, por cada aumento de liras en la cantidad preasignada para el tabaco, hay una reducción en la participación presupuestaria asignada a la vivienda en el presupuesto restante del hogar de 0,0022 puntos porcentuales o 0,0022 x *M* liras, donde *M* es el presupuesto restante después de gastar en tabaco.

Suponga que los gastos mensuales después de gastar en tabaco son de aproximadamente 1200 liras (ya que 106 liras gastadas en tabaco constituyen aproximadamente el 8,17 por ciento del presupuesto). Luego, utilizando las estimaciones de parámetros presentadas por los autores, se puede calcular que un aumento de 10 liras en la cantidad preasignada para el tabaco conduce a una disminución de 26,4 liras en los gastos de vivienda, al tiempo que redistribuye los gastos en todos los productos básicos restantes, aumentando unos y disminuyendo otros. Por ejemplo, un aumento de 10 liras en la cantidad preasignada para el tabaco reduciría los gastos en alimentos, servicios públicos, bienes duraderos, vestimenta, salud y educación en alrededor de 2,4, 1,2, 9,6, 3,6, 2,4 y 3,6 liras, respectivamente, y aumentaría los gastos en transporte, entretenimiento, alcohol y otros productos básicos en 25,2, 20,4, 2,4 y 1,2 liras, respectivamente.

Lo que es importante ver es que un aumento en el gasto en tabaco redistribuye claramente los gastos, beneficiando algunos artículos, pero perjudicando a varios otros. En ese caso particular, los artículos con consumo reducido son en su mayoría de primera necesidad, y eso amerita una intervención de política pública para regular el consumo de tabaco. Las variables restantes de la tabla 4.2, incluidas las utilizadas en la regresión pero que no se muestran en la tabla, sirven como variables de control en la regresión.

5 Cuantificación del efecto empobrecedor del consumo de tabaco

5.1 Introducción

Las estimaciones nacionales de pobreza son una variable política importante en la mayoría de los países. La estimación del porcentaje de personas pobres en una población determina el curso de los debates sobre políticas de desarrollo en muchos países. La reducción de la pobreza es un objetivo declarado en numerosos países del mundo, y la erradicación de la pobreza en todas sus formas es el primer objetivo de los Objetivos de Desarrollo Sostenible de las Naciones Unidas.[6] Sin embargo, el consumo de tabaco es un factor importante entre los que obstaculizan la capacidad de una nación para alcanzar los objetivos de reducción de la pobreza.

El consumo de tabaco y la pobreza son componentes de un círculo vicioso.[4] A medida que se gasta más dinero en tabaco, los hogares se ven privados de ciertas necesidades, como la alimentación y la nutrición, como se explica en el Capítulo 4, lo que crea un enorme costo de oportunidad y exacerba aún más la pobreza. Dado que el dinero gastado en tabaco es altamente improductivo y aumenta las enfermedades relacionadas con el tabaco, el aumento de los costos de atención de la salud resultante y la pérdida de ingresos debido a las muertes prematuras y la morbilidad también pueden aumentar la carga de la pobreza. En todo el mundo, alrededor del 80% de los fumadores viven en PIMB, y en la mayoría de esos países el consumo de tabaco se concentra en poblaciones de bajos ingresos.[4] Las desigualdades relacionadas con la riqueza y con la educación en el consumo de tabaco entre hombres y mujeres son mayores entre los PIMB en comparación con los países de ingresos medianos altos.[160]

El Capítulo 4 explica cómo el gasto en tabaco desplaza los gastos en diferentes grupos de productos, ofreciendo una cierta dimensión del costo de oportunidad del gasto en tabaco. Este capítulo muestra cómo cuantificar el impacto directo del gasto en tabaco sobre la pobreza medido por recuento de la pobreza, analiza cómo el gasto en tabaco contribuye al empobrecimiento y presenta métodos para cuantificar esos conceptos. También demuestra cómo se puede hacer eso con la ayuda de EGH usando Stata.

5.2 Recuento de la pobreza y su relevancia

Las definiciones de pobreza varían de un país a otro según las circunstancias sociales y económicas específicas que prevalecen en cada país. Sin embargo, "casi todas las líneas nacionales de pobreza (NPL, por sus siglas en inglés) están ancladas al costo de una canasta de alimentos, lo que las personas pobres de ese país comerían habitualmente, que proporciona una nutrición adecuada para una buena salud y una actividad normal, más una asignación para gastos no alimentarios."[161] A medida que cambian las canastas de alimentos o los gustos y preferencias, las naciones suelen redefinir la línea de pobreza en consecuencia. En esencia, la línea de pobreza contempla cierta privación de recursos y define una cantidad que es necesaria para sostener una noción percibida localmente de lo que se necesita para no vivir en pobreza.

Por lo general, eso se traduce a una unidad monetaria local. Por ejemplo, Estadísticas de Sudáfrica[162] define una línea de pobreza alimentaria, la cantidad de dinero que una persona necesitará para pagar la ingesta de energía diaria mínima requerida, también conocida como la línea de pobreza "extrema", como 547 rand por persona al mes. También define otras líneas de pobreza que tienen en cuenta ciertos gastos mínimos en artículos no alimentarios. De manera similar, la Oficina del Censo de los Estados Unidos (USCB) utiliza un conjunto de umbrales de ingresos en dólares que varían según el tamaño y la composición de la familia para determinar quién se encuentra en la pobreza.[163] La definición de 2022 de la USCB muestra que una persona soltera menor de 65 años que gana menos de $13.590 dólares estadounidenses por año se considera que vive por debajo del umbral de la pobreza.[164]

Si bien existen varios métodos para medir la pobreza, el índice de recuento (HCR), que es una medida absoluta de la pobreza, es uno de los indicadores de pobreza más utilizados, especialmente en los PIMB.[165] El HCR, una medida de conteo, se define como la fracción de la población que vive por debajo de la NPL, y permite una interpretación muy intuitiva y sencilla. Esa fracción a menudo se calcula utilizando las EGH, ya que permite calcular los gastos promedio de cada hogar, o los gastos de consumo per cápita de los individuos, y compararlos con la línea de pobreza definida. El HCR, sin embargo, no tiene en cuenta el grado de pobreza. En otras palabras, la tasa de pobreza medida por HCR permanecería igual incluso si las personas por debajo de esa línea de pobreza se empobrecieran aún más.

Las NPL entre países a menudo no son comparables, ya que la noción de vivir en pobreza puede variar significativamente entre países y culturas. Aunque no sean comparables entre países, las líneas de pobreza son bastante útiles en el contexto de las políticas de desarrollo interno de un país. Se pueden utilizar como referencia para facilitar ciertos programas de bienestar social, por ejemplo, para desarrollar intervenciones dirigidas específicamente a personas en situación de pobreza.

5.3 ¿Cómo contribuye el consumo de tabaco al empobrecimiento?

El objetivo de este capítulo es cuantificar el impacto del consumo de tabaco en la estimación del HCR. Para entender eso, ayuda distinguir dos tipos de pobreza, como lo explica el sociólogo británico B. Seebohm Rowntree[166] y está reproducido en la monografía de la OMS/NCI sobre *La economía del tabaco y el control del tabaco*.[4] La primera es la pobreza primaria, que se refiere a una situación en la que los ingresos u otros recursos son insuficientes para cubrir las necesidades básicas como alimentos, agua o ropa. Esencialmente, los hogares que caen por debajo de la NPL en un país pueden clasificarse como aquellos que sufren de pobreza primaria.

La segunda es la pobreza secundaria, que se refiere a una situación en la que los hogares cuentan con recursos suficientes para satisfacer sus necesidades básicas, pero esos recursos no son utilizados de manera eficiente. En consecuencia, a pesar de poseer una mayor cantidad de recursos, esos hogares pueden estar viviendo en condiciones similares o inferiores a las de la pobreza primaria. Por ejemplo, una cantidad significativa de los ingresos se gasta en el consumo improductivo y dañino de bienes como el tabaco o el alcohol por parte de un hogar que, por lo demás, se encuentra por encima del umbral de la pobreza. Debido a un efecto de desplazamiento, el hogar no puede satisfacer sus necesidades básicas, al igual que los hogares en pobreza primaria.

Pero las estimaciones de HCR solo captarían a aquellos que se encuentran en la pobreza primaria, aunque muchos hogares en el país pueden estar realmente en la pobreza secundaria y, por lo tanto, no cubren sus necesidades básicas debido al consumo excesivo de tabaco. Sería ideal incluir dichos hogares en el cálculo del HCR para que las políticas y los programas puedan enfocarse de manera más efectiva. Alternativamente, se deberán adoptar políticas para sacar a los hogares de la pobreza secundaria ayudándolos a reducir o detener el consumo derrochador y dañino para que sus recursos disponibles totales puedan satisfacer sus necesidades básicas.

Dado que los presupuestos familiares son limitados, el consumo de cualquier cosa, incluido el tabaco, implica necesariamente concesiones. La literatura discutida en el Capítulo 4 muestra que la concesión ocurre en forma de desplazamiento de ciertas necesidades. Hay tres canales principales por los cuales un mayor consumo de tabaco puede disminuir efectivamente los ingresos de un hogar y empujarlo a un estado de pobreza, como se explica a continuación:

1) **Canal 1: Ingresos perdidos por la compra de tabaco**
   El ingreso disponible directo para satisfacer las necesidades básicas se reduce en la misma cantidad que se gastó en la compra de tabaco.

2) **Canal 2: Pérdida de ingresos por el tratamiento de morbilidad relacionada con el tabaco**
   Dado que el consumo de tabaco y la exposición al humo de segunda mano conducen inevitablemente a la aparición de varias enfermedades y la morbilidad asociada, los costos del tratamiento de estas afecciones médicas reducen aún más los ingresos disponibles para satisfacer las necesidades básicas. Si bien el aumento de los gastos médicos afecta directamente los ingresos disponibles, también puede afectar la productividad y el potencial de generación de ingresos.

3) **Canal 3: Pérdida de ingresos por el tratamiento de la mortalidad relacionada con el tabaco**
   El consumo de tabaco y las enfermedades relacionadas con el humo de segunda mano a menudo provocan una muerte prematura. Eso se traduce en la pérdida de ingresos futuros, lo que afecta el bienestar de los otros miembros del hogar.

Todos esos canales tienen el efecto final de empobrecer aún más a un hogar pobre. Dado que las personas pobres suelen destinar una mayor parte de su presupuesto al tabaco en comparación con las personas ricas,[4] el impacto empobrecedor del gasto en tabaco es relativamente mayor en personas pobres que en personas ricas. Las políticas de control del tabaco que reducen el consumo de tabaco tienen el efecto contrario, especialmente si los consumidores de tabaco son más sensibles a los precios.[167] Como resultado de la disminución del gasto en tabaco y, en consecuencia, la reducción del gasto en atención médica, esos hogares tendrán más ingresos disponibles para gastar en necesidades esenciales (como alimentos, ropa y educación).

Aunque la literatura que examina las desigualdades socioeconómicas en el tabaquismo y el consumo de tabaco es bastante sustancial,[4] la literatura que cuantifica el efecto empobrecedor del gasto en tabaco en términos de su impacto en medidas cuantificables de pobreza es limitada. Uno de los primeros estudios se hizo en Vietnam,[168] y cuantifica el efecto empobrecedor de los pagos de bolsillo para la atención de la salud. El primer estudio para estimar el efecto empobrecedor del gasto doméstico directo en el tabaquismo y el gasto médico excesivo atribuible al tabaquismo se realizó en China.[169] Encuentran que esos dos efectos combinados son responsables del empobrecimiento de 30,5 millones de residentes urbanos y 23,7 millones de residentes rurales en China. Un estudio de la India[170] también encuentra que el efecto combinado de esos dos factores resulta en el empobrecimiento de 15 millones de personas en la India.

Un estudio en el Reino Unido[171] resta solo los gastos en tabaco de los ingresos del hogar para estimar su impacto en la pobreza y encuentra que se puede considerar que más de 432 mil niños han sido arrastrados a la pobreza por el tabaquismo de sus padres. Otro estudio del Reino Unido[172] muestra que, cuando se tiene en cuenta el gasto en tabaco, alrededor de 500 mil hogares adicionales, que comprenden más de 850 mil adultos y casi 400 mil niños, se clasifican como pobres en el Reino Unido en comparación con las cifras oficiales de *Hogares por Debajo del Ingreso Promedio*. Un estudio más reciente[173] del Reino Unido encuentra que 230 mil hogares (400 mil adultos y 180 mil niños) viven en la pobreza cuando se tiene en cuenta el gasto semanal en tabaco. Otro estudio reciente de Vietnam[174] estima que los gastos relacionados con el tabaco empobrecieron a 0,31 millones adicionales (que corresponden al 3,77% de la estimación oficial de la OSG) de personas en Vietnam en 2018.

Esos estudios concluyen que muchas personas que de otro modo estarían por encima de la NPL en esos países (es decir, en pobreza secundaria) están efectivamente en la pobreza porque su ingreso disponible, después de los gastos en tabaco y los gastos de salud asociados, es menor que el de las personas que están clasificadas oficialmente por debajo de la NPL. En otras palabras, esas personas son etiquetadas inadvertidamente como si estuvieran por encima de la línea de pobreza, cuando en realidad no lo están.

Ninguno de los estudios hasta el momento ha estimado el impacto empobrecedor de los ingresos perdidos por las muertes prematuras relacionadas con el tabaco (Canal 3) y los ingresos perdidos por la morbilidad relacionada con el humo de segunda mano (parte del Canal 2). Dado que la pobreza o HCR se mide para un momento determinado, es insostenible restar los ingresos perdidos debido a la mortalidad prematura o la pérdida futura de ingresos de los ingresos familiares actuales. Sin embargo, los costos médicos directos atribuibles a humo de segunda mano (parte del Canal 2) son claramente candidatos a ser restados como ingresos perdidos de los ingresos disponibles actuales al evaluar el impacto empobrecedor del consumo de tabaco. Pero eso tampoco se ha incorporado en ninguno de los estudios hasta el momento.

5.4 Marco conceptual para estimar el impacto en HCR

Para estimar el cambio en HCR, es necesario restar dos tipos diferentes de ingresos perdidos de los ingresos del hogar: (i) ingresos perdidos debido a la compra de tabaco, e (ii) ingresos perdidos debido a los costes sanitarios directamente atribuibles al consumo de tabaco y al humo de segunda mano. Antes de poder restar esos diferentes componentes de ingresos perdidos de los ingresos totales del hogar, es importante identificar la NPL según la forma en que se calcula el HCR. La NPL puede ser un número único para todo el país o números diferentes para áreas rurales y urbanas y para cada subregión o estado dentro del país. Suele estar disponible en los organismos de estadística u otras fuentes gubernamentales de cada país. La variable de ingresos contra la cual se suele calcular el HCR se toma de las EGH representativas a nivel nacional. Dado que los estimados de consumo o gasto reportados son mucho más confiables que los ingresos reportados para representar el ingreso real,[8] los gastos estimados a partir de las EGH se utilizan como sustitutos de los ingresos para estimar la proporción de personas por debajo del umbral de pobreza.

Lo que también es importante es el hecho de que la mayoría de las EGH son encuestas de hogares que tratan a los hogares como una sola unidad y los gastos de consumo se informan para el hogar como un todo. Sin embargo, la pobreza la experimentan los individuos, no los hogares *per se*, y por lo tanto es la pobreza entre las personas la que debe medirse. Aunque sea posible que no se sepa nada sobre la distribución dentro de los hogares, muchos estudios asumen una distribución uniforme dentro de los hogares al construir la distribución estimada del consumo individual.[175] Además, si bien es más aceptable suponer una distribución uniforme del consumo dentro de los hogares al construir la distribución estimada del consumo individual, puede no ser tan aceptable suponer una distribución uniforme dentro de un hogar en el caso de productos conocidos para adultos, como el tabaco. Una solución propuesta por Deaton[8] es "un sistema de ponderadores, por el cual los niños cuentan como una fracción de un adulto, y la fracción depende de la edad, de modo que el tamaño efectivo del hogar es la suma de esas fracciones, y no se mide en número de personas, sino en número de adultos equivalentes." Ha habido tales estudios que enfatizan la importancia de estimar la pobreza usando el gasto por adulto equivalente (PAE) que controla las economías de escala y las necesidades reducidas de los niños.[176] Una revisión reciente de esa literatura,[176] sin embargo, concluye que "el uso de escalas de equivalencia, si bien no carece de importancia, no es convincente en la práctica" ya que los estudios no están de acuerdo en cuanto a si las estimaciones de pobreza son sensibles al uso de escalas de equivalencia que se ajustan a la composición del hogar y las economías de escala. Además, dado que el hogar es una sola unidad a todos los efectos prácticos y el dinero gastado en tabaco necesariamente reduce los ingresos disponibles para todo el hogar, incluidos los niños, el impacto empobrecedor podría muy bien ser soportado por igual por niños y adultos. Por lo tanto, la discusión de HCR

y el impacto empobrecedor del consumo de tabaco en este capítulo no ha considerado el uso de dichos *adultos equivalentes*.

Al estimar el HCR, es importante utilizar ponderaciones de encuestas que puedan generar estadísticas a nivel de población para individuos y no para hogares. Esa estimación se puede obtener multiplicando el tamaño del hogar por las ponderaciones de la encuesta dadas para generar estadísticas a nivel de hogar en las EGH. El HCR total y la pobreza se calculan antes de restar los ingresos perdidos relacionados con el tabaco. Sea *z* la variable o escalar que representa la NPL. El HCR simplemente cuenta el número de personas cuyos ingresos están por debajo de la línea de pobreza *z* y divide ese número por el número total de personas en el país o región. Sea *x* la medida de bienestar (es decir, los gastos de consumo per cápita, que son los gastos totales de consumo del hogar divididos por el tamaño del hogar), entonces el HCR denotado como (P_0) se calcula de la siguiente manera:[165]

$$P_0 = \frac{1}{N}\sum_{i=1}^{N} I(x_i \leq z) \qquad (5.1)$$

donde *I (.)* es una función indicadora que toma el valor 1 si su argumento es verdadero y 0 en caso contrario. Si bien se calcula usando EGH, se deben usar ponderaciones de encuesta apropiadas. $P_0 \times N$ da el número total de personas en pobreza en el país.

5.4.1 Exceso de pobreza atribuido a la pérdida de ingresos por la compra de tabaco

Los gastos en tabaco por hogar suelen estar disponibles en las mismas encuestas de hogares a partir de las cuales se calcula el HCR (P_0). Sean *t* los gastos de consumo per cápita en la compra de tabaco en el mismo período de tiempo para el cual se captura la medida de bienestar (*x*). En otras palabras, ese es el ingreso perdido por la compra de tabaco. Entonces, el HCR, después de deducir el gasto en tabaco o los ingresos perdidos por la compra de tabaco, denotados por (P_1), puede calcularse como:

$$P_1 = \frac{1}{N}\sum_{i=1}^{N} I([x_i - t_i] \leq z) \qquad (5.2)$$

donde, nuevamente, *I (.)* es una función indicadora que toma el valor de 1 si su argumento es verdadero y 0 en caso contrario. $x_i - t_i$ es el ingreso disponible per cápita después de restar el ingreso perdido por las compras de tabaco. $(P_1 - P_0) \times N$ es el exceso de personas empobrecidas debido al gasto en tabaco. En otras palabras, ese es el exceso de pobreza atribuido a los gastos directos de compra de tabaco.

5.4.2 Exceso de pobreza atribuido a la pérdida de ingresos por la compra de tabaco y el tratamiento de la morbilidad relacionada con el tabaco

La morbilidad relacionada con el tabaco puede ocurrir entre quienes consumen tabaco, así como entre quienes están expuestos al humo de segunda mano. Sean *t* y *h* el gasto en tabaco per cápita y los gastos en salud per cápita atribuibles al consumo de tabaco y el humo de segunda mano, respectivamente, en el mismo período de tiempo para el que se mide el bienestar (*x*). Entonces, el HCR, después de deducir ese ingreso perdido de las compras de tabaco y los gastos en tratamientos de salud atribuibles al tabaco, denotados por (P_2), puede calcularse como:

$$P_2 = \frac{1}{N}\sum_{i=1}^{N} I([x_i - t_i - h_i] \leq z) \qquad (5.3)$$

donde *I (.)* es una función indicadora que toma el valor 1 si su argumento es verdadero y 0 en caso contrario.

$x_i - t_i - h_i$ es el ingreso disponible per cápita después de restar los gastos en tabaco y los gastos sanitarios atribuibles al consumo de tabaco y al humo de segunda mano. $(P_2 - P_1) \times N$ es el número adicional de personas empobrecidas debido al gasto en salud atribuible al consumo de tabaco y al humo de segunda mano. $(P_2 - P_0) \times N$ será el exceso total de personas empobrecidas después de tener en cuenta los ingresos perdidos tanto por el gasto en tabaco como por gastos atribuibles en atención de salud.

Si bien las EGH brindan información sobre los gastos de atención médica, no distinguen la cantidad de atención médica que se puede atribuir al consumo de tabaco o la exposición al humo de segunda mano. Eso debe estimarse por separado, y la resta debe ser solo para los gastos en atención de la salud que pueden atribuirse al consumo de tabaco o a la exposición al humo de segunda mano. Los costos atribuibles se pueden estimar utilizando un enfoque específico de la enfermedad o un enfoque inclusivo o de todas las causas.[177] Dado que las EGH a menudo proporcionan gastos agregados de atención de la salud, es más apropiado utilizar el enfoque inclusivo. Este descompone la parte de los costos médicos totales atribuibles al consumo de tabaco o la exposición al humo de segunda mano multiplicando los costos totales de atención médica por la fracción atribuible al consumo de tabaco, o fracción atribuible al humo de segunda mano, comúnmente conocida como fracción atribuible al tabaquismo (SAF). SAF es la proporción del uso total de la atención médica que se atribuye al tabaquismo por parte de los fumadores actuales y ex fumadores.[177] De manera similar, SAF para el humo de segunda mano sería la fracción de los gastos de atención médica que se pueden atribuir al humo de segunda mano.

Por lo tanto, los gastos sanitarios atribuibles al consumo de tabaco y al humo de segunda mano (es decir, *h* en la Ecuación 5.3) se pueden calcular de la siguiente manera:

$$h_i = (exphealth_i / hsize_i) * (SAF_{tob} + SAF_{HSM}) \quad (5.4)$$

donde *exphealth* y *hsize* son los gastos del hogar en salud y el tamaño del hogar, respectivamente. Ambas variables se obtienen directamente de las EGH. SAF_{tob} y SAF_{HSM} son fracciones de los gastos de atención médica atribuibles al consumo de tabaco y al humo de segunda mano, respectivamente. El SAF debe estimarse externamente utilizando datos de varias fuentes diferentes. También puede tomarse de estudios disponibles en otras partes del país.

El SAF se puede estimar utilizando el enfoque epidemiológico o un enfoque econométrico.[178] El enfoque econométrico requiere "datos exhaustivos representativos a nivel nacional que contengan información detallada sobre el historial de tabaquismo de cada encuestado, características sociodemográficas, situación laboral, otros comportamientos de riesgo para la salud, estado de salud, condiciones médicas, gastos anuales de atención médica por tipo de servicios de atención médica (como hospitalización y visitas ambulatorias), y días anuales de baja laboral o incapacidad".[178] Por otro lado, el enfoque epidemiológico requiere menos datos y "se puede hacer con datos agregados y, por lo tanto, se puede usar cuando no se dispone de datos detallados de encuestas de salud".[178]

Por esas razones, en muchos PIMB se prefiere el enfoque epidemiológico para estimar SAF. La OMS proporciona un conjunto de herramientas[179] para estimar los costos económicos del tabaquismo, que incluye explicaciones y métodos detallados tanto para los métodos epidemiológicos como econométricos para estimar SAF. Por lo tanto, este conjunto de herramientas no aborda ese tema.

A diferencia de los datos requeridos para estimar SAF para el consumo de tabaco, los datos requeridos para estimar SAF para el humo de segunda mano pueden ser más difíciles de obtener. Quizás esa sea la razón por la cual los estudios previos que cuantificaron el efecto empobrecedor del consumo de tabaco en la pobreza ignoraron esa fuente particular de pérdida de ingresos en el cálculo.

5.5 Preparación de datos para estimar el efecto empobrecedor

Como se detalla en el Capítulo 2, los datos primero deben limpiarse y prepararse para el análisis. Dado que el objetivo es cuantificar el efecto empobrecedor del tabaco, las variables más importantes son los gastos en tabaco (*exptobac*), así como los gastos en todos los productos básicos juntos como proxy del ingreso familiar (*exptotal*). Además, se requieren los gastos en atención médica (*exphealth*) para calcular los costos de atención médica atribuibles al tabaco y al humo de segunda mano, según la disponibilidad de SAF. Las otras variables necesarias de las EGH para el análisis incluyen el tamaño del hogar, las ponderaciones de la encuesta y las variables para declarar el diseño de la encuesta. Es necesaria una variable o escalar para representar la NPL. Si la NPL es una variable que varía entre regiones, o por áreas rurales o urbanas, o estados dentro del país, entonces la variable debe fusionarse con los datos de la encuesta de hogares antes de poder realizar el análisis. Para ello, debe estar presente una variable de identificación común tanto en los datos de gasto de los hogares como en los datos de la línea de pobreza.

Por ejemplo, si la NPL en un país varía según el estado y el área de residencia (rural o urbana), los datos de pobreza deben tener tres variables, una variable que indique la NPL (*npl*), generalmente en unidades monetarias locales, una variable con los nombres o código numérico para diferentes estados (*stateid*), y una variable de residencia que indica si la NPL pertenece a áreas rurales o urbanas (*residence*). Del mismo modo, los datos de EGH también deben tener variables de *stateid* y de *residencia*. Luego, ambos conjuntos de datos se pueden fusionar con el comando *<merge>* en Stata.

Para hacer esto, primero prepare un conjunto de datos de Stata con la variable *npl* y otras variables de identificación según sea necesario y guárdelo con el nombre "poverty.dta". Luego, abra los datos maestros de la EGH con la información de gastos de cada hogar y asegúrese de que tenga las mismas variables *stateid* y *residence* que en el archivo de poverty.dta. Luego use el comando *<merge m:1 stateid residence using poverty.dta>*. Aquí se usa una fusión de muchos a 1 (*m:1*), ya que el conjunto de datos maestros tiene varios hogares con el mismo *stateid* y *residence*. Después del comando de fusión, utilice el comando *<tabulate _merge>* para comprobar si la fusión se ha realizado con correctamente.

Si bien la EGH considera a los hogares como una sola unidad e informa todos los gastos a nivel del hogar, la NPL generalmente es para un individuo, por lo que es importante convertir los datos de gastos para que sean comparables con los datos de la línea de pobreza. También es importante verificar la duración de la declaración de los gastos (por ejemplo, por mes, por semana o cualquier otro intervalo) y asegurarse de que la línea de pobreza también tenga la misma duración.

Por ejemplo, tanto el gasto de consumo o gasto en salud atribuible al consumo de tabaco como la línea de pobreza deben ser por persona, por mes. Para hacerlo en Stata, cree nuevas variables para generar gastos per cápita para compararlos con la línea de pobreza utilizando la variable tamaño del hogar (*hsize*). Por ejemplo, los gastos per cápita se pueden generar como *<gen pce =exptotal/hsize>*. De manera similar, las variables sobre el gasto en tabaco per cápita (*pcetob*) y sobre los gastos en salud per cápita (*pcehealth*) deben generarse dividiendo los gastos totales correspondientes por la variable del tamaño del hogar. Además, al usar el valor de SAF y *pcehealth*, se crea la variable pcehealthtob que representa los gastos de atención médica atribuibles al consumo de tabaco y al humo de segunda mano per cápita. Por ejemplo, si el SAF para el consumo de tabaco es 0,2, se puede generar una nueva variable *pcehealthtob* con el comando *<gen pcehealthtob =pcehealth*0.2>*. Y si el SAF para la exposición al humo de segunda mano es 0,1, se debe crear una nueva variable *pcehealthshs* con el comando *<gen pcehealthshs=pcehealth*0.1>* para representar los gastos de atención médica per cápita atribuibles al humo de segunda mano.

Con el fin de calcular el cambio en HCR después de la resta incremental de diferentes variables de interés, se deben crear las siguientes variables adicionales:

(1) *pcet* (*pce* después de sustraer los gastos en tabaco): *<gen pcet=pce-pcetob>*, y

(2) *pceh* (*pce* después de sustraer los gastos en tabaco y los gastos en salud atribuibles al consumo de tabaco y al humo de segunda mano): *<gen pceh=pcet-pcehealthtob- pcehealthshs>*. En caso de que no se disponga de estimaciones de SAF para la exposición al humo de segunda mano, la fórmula para *pceh* puede reducirse a *<gen pceh=pcet-pcehealthtob>*.

Por último, la variable de ponderación de la encuesta proporcionada en los datos de gastos de los hogares (como *hweight*) debe ajustarse para tener en cuenta la estimación de la pobreza a nivel individual. Eso se puede hacer multiplicando la variable por el tamaño del hogar, es decir, *<gen pweight=hweight*hsize>*. Una vez generadas todas las variables anteriores, se puede estimar en Stata el efecto empobrecedor del tabaco.

5.6 Estimación del efecto empobrecedor del consumo de tabaco

En Stata, la estimación de HCR es bastante sencilla y Stata ofrece varios módulos escritos por usuarios para ello. Por ejemplo, *<povdeco>*[180] es un módulo que estima HCR y varias otras medidas de pobreza con un solo comando. Para hacerlo, instale el módulo con *<ssc install povdeco>* y ejecute el comando *<povdeco pce [fw=pweight], varpline(npl)>* donde *pce* es la variable de gasto per cápita mensual, *npl* es la variable de NPL, y *pweight* es el ponderador de la encuesta ajustado por el tamaño del hogar. *Povdeco* reportará HCR junto con una brecha de pobreza y una brecha de pobreza al cuadrado, por defecto. También permite estimar la pobreza por diferentes subgrupos utilizando la opción *<bygroup(groupvar)>*.

Sin embargo, para estimar solo el HCR, un comando de proporción simple en Stata funcionará. Por ejemplo, con el siguiente comando, se puede estimar el HCR:

```
gen povdum = 0
replace povdum = 1 if pce <= npl
proportion povdum [fw = pweight]
```

Eso también se puede hacer después de declarar el diseño de la encuesta usando el comando *<svyset>* como se explica en el Capítulo 2. En ese caso, el comando se puede escribir como *<svy: proportion povdum>*.

Dado que el cambio en HCR debe determinarse después de la resta incremental de diferentes ingresos perdidos como se explicó anteriormente, eso se puede implementar mejor con el siguiente código. El siguiente código asume que las variables han sido generadas como se discutió en la Sección 5.5.

```
#delimit;
local subtr pce pcet pceh;
local nvar: word count `subtr';
matrix M = J(`nvar', 2, .);
forvalues i = 1/`nvar' {;
   local X: word `i' of `subtr';
   qui gen ind = (`X'<=npl);
   qui sum ind [fw=pweight];
   matrix M[`i', 1] = r(mean);
   matrix M[`i', 2] = r(sum);
   drop ind;
};
```

```
matrix rownames M = `subtr';
matrix colnames M = HCR Poor;
matlist M, cspec(& %12s | %5.4f & %9.0f &) rspec(--&&-);
```

Como muestra el código, las únicas variables de los datos utilizados en el código anterior son: *pce, pcet, pceh, npl* y *pweight*. Si los datos se prepararon con esos nombres de variables, ejecutar el código generaría una matriz de 3x2 en la ventana de resultados de Stata que mostraría *pce, pcet* y *pceh* como encabezados de fila y "HCR" y "Poor" como encabezados de columna. La primera columna muestra el HCR estimado (valor de 0 a 1) para *pce* (antes de restar los ingresos perdidos), *pcet* (HCR después de restar los ingresos perdidos de la compra directa de tabaco) y *pceh* (HCR después de restar los ingresos perdidos de la compra de tabaco y gastos sanitarios atribuibles al consumo de tabaco y al humo de segunda mano). Los valores correspondientes bajo la columna "Poor" muestran el número estimado de personas en pobreza en cada paso sucesivo. Si desea derivar también los errores estándar para cada estimación, el módulo de Stata *<povdeco>* daría resultados idénticos para HCR junto con otras medidas de pobreza. El siguiente código se puede utilizar para ese propósito. Las variables *pce* y *npl* son las mismas que en el código anterior.

```
ssc install povdeco, replace
povdeco pce [fw=pweight], varpline(npl)
```

La comparación de dos filas sucesivas ilustra el cambio tanto en HCR como en el número de personas en pobreza después de la sustracción sucesiva de cada componente de ingreso perdido. En el código se estima el número de personas pobres multiplicando HCR por la población total estimada a partir de la propia encuesta de hogares, lo que es posible utilizando la variable de ponderación específica de la persona. El escalar *r(sum)* es un resultado guardado después del comando *<summary>*, y muestra el resultado de multiplicar la media por el tamaño de la población. Alternativamente, uno puede multiplicar HCR por los datos de población disponibles a nivel nacional de otras fuentes para llegar al cambio en el número de personas pobres.

El análisis anterior también se puede hacer con diferentes subgrupos usando cualquiera de los métodos discutidos anteriormente. Sin embargo, es necesario modificar los datos y es posible que se deban generar nuevas variables para poder realizar el análisis a nivel de subgrupo. La sección 7.5 del Apéndice de código incluye un do-file de ejemplo que detalla el código utilizado en esta sección. Los usuarios podrán copiar y pegar eso en el editor de do-files de Stata y estimar los resultados con los datos/variables correspondientes que se describen allí.

5.7 Estudio de caso de la India

En India durante 2004–2005, fuentes oficiales del gobierno consideraron que aproximadamente el 28,3% de la población rural y el 25,6% de la población urbana se encontraban por debajo de la NPL. Las estadísticas oficiales de pobreza se reportan por separado para las áreas rurales y urbanas del país y también se reportan por estado. La línea de pobreza también está disponible por separado para cada estado y por áreas rurales y urbanas. India también tiene el segundo mayor número de consumidores de tabaco en el mundo.[72] La tasa de pobreza y las tendencias a lo largo del tiempo siempre han ocupado un lugar central en el discurso de la política de desarrollo de la India. En ese contexto, John et al.[170] examinan el impacto empobrecedor del gasto en tabaco, así como el gasto en atención médica relacionado con el consumo de tabaco en la India. La Tabla 5.1 muestra los resultados de su análisis.

La tabla primero informa las estimaciones oficiales de HCR y el número de personas pobres en la India por áreas rurales y urbanas. Luego muestra el efecto separado de restar el gasto en tabaco y el gasto en atención médica atribuible al consumo de tabaco de los gastos per cápita para áreas rurales y urbanas en India y luego el efecto combinado de restar ambos gastos de los gastos per cápita. Los resultados muestran que la tasa de pobreza o HCR aumentó en 1,6 y 0,8 puntos porcentuales en la India rural y urbana, respectivamente, después de restar los ingresos perdidos por la compra de tabaco y los gastos de atención médica relacionados con el tabaco. El gasto en tabaco y el gasto asociado en atención de la salud empobrecieron a unos 15 millones de personas más en la India. En otras palabras, 15 millones de personas en la India que están por encima del umbral oficial de pobreza se encuentran en pobreza secundaria, experimentando niveles de vida más bajos en términos de su capacidad para gastar en las necesidades diarias porque su dinero se desvía hacia gastos innecesarios relacionados con el tabaco.

Eso también tiene serias implicancias en términos de políticas públicas. Si las medidas de bienestar social (un subsidio alimentario, por ejemplo) se dirigen a quienes están oficialmente por debajo de la NPL, los que se encuentran en pobreza secundaria no podrán disfrutar de los beneficios derivados de dichas medidas de bienestar y seguirán viviendo en la pobreza.

Tabla 5.1 Cambios en HCR y número de personas pobres después de tener en cuenta el consumo de tabaco en India

	Rural	Urbano	Total
(1)) Estimaciones oficiales			
Población total (millones)	780,2	315,5	1095,7
Población BPL (%)	28,3	25,6	
Población BPL (millones)	220,7	80,8	301,6
(2) Considerando las compras de tabaco			
Población BPL (%)	29,8	26,3	
Población BPL (millones)	232,5	83,1	315,6
(3) Considerando los gastos médicos relacionados con el tabaco			
Población BPL (%)	28,4	25,7	
Población BPL (millones)	221,4	81,1	302,5
(4) Efecto combinado de (2) y (3)			
Población BPL (%)	29,8	26,4	
Población BPL (millones)	232,9	83,3	316,2

Nota: BPL= Por debajo de la línea de pobreza. Fuente: John et al. (2011).[170]

6 Bibliografía

1. World Health Organization. *Tobacco control for sustainable development.* http://apps.who.int/iris/handle/10665/255509 (2017).
2. World Health Organization. *WHO global report: mortality attributable to tobacco.* http://apps.who.int/iris/bitstream/10665/44815/1/9789241564434_eng.pdf (2012).
3. Jha, P. & Peto, R. Global Effects of Smoking, of Quitting, and of Taxing Tobacco. *N. Engl. J. Med. 370*, 60–68 (2014).
4. NCI & WHO. *The Economics of Tobacco and Tobacco Control.* https://cancercontrol.cancer.gov/sites/default/files/2020-06/m21_complete.pdf (2016).
5. Goodchild, M., Nargis, N. & d'Espaignet, E. T. Global economic cost of smoking-attributable diseases *Tob. Control 27*, 58–64 (2018).
6. UN. *Transforming our world: the 2030 Agenda for Sustainable Development.* https://sustainabledevelopment.un.org/post2015/transformingourworld (2015).
7. John, R. M., Grieve Chelwa, Violeta Vulovic, & Frank J Chaloupka. *Using Household Expenditure Surveys for Research in the Economics of Tobacco Control. A Tobacconomics Toolkit.* https://tobacconomics.org/research/a-toolkit-on-using-household-expenditure-surveys-for-research-in-the-economics-of-tobacco-control/ (2019).
8. Deaton, A. S. *The Analysis of Household Surveys.* (Johns Hopkins University Press for the World Bank, 1997).
9. Pollak, R. A. Conditional Demand Functions and Consumption Theory. *Q. J. Econ. 83*, 60–78 (1969).
10. Pollak, R. A. Conditional Demand Functions and the Implications of Separable Utility. *South. Econ. J. 37*, 423–433 (1971).
11. Indian Statistical Institute. The National Sample Survey: General Report No. 1. First Round: October 1950 - March 1951. *Sankhyā Indian J. Stat. 1933-1960 13*, 47–214 (1953).
12. World Bank. Living Standards Measurement Study (LSMS). https://web.worldbank.org/archive/website00002/WEB/INDEX-5.HTM (2022).
13. International Household Survey Network. IHSN Survey Catalog. http://catalog.ihsn.org/index.php/catalog/central (2018).
14. LISGIS. *Household Income and Expenditure Survey* 2016. http://catalog.ihsn.org/index.php/catalog/7279 (2017).
15. Wooldridge, J. M. *Econometric Analysis of Cross Section and Panel Data.* (The MIT Press, 2010).
16. Cameron, A. C. & Trivedi, P. K. *Microeconometrics Using Stata, Revised Edition.* (Stata Press, 2010).

17. StataCorp. Stata Statistical Software v.15. (2018).

18. Baum, C. F. *A little bit of Stata programming goes a long way.* http://ideas.repec.org/e/pba1.html (2005).

19. StataCorp. *Stata programming reference mannual Release 15.* (2017).

20. Chaloupka, F. J. & Warner, K. E. The Economics of Smoking. in *The Handbook of Health Economics* 1539–1627 (2000).

21. IARC. *IARC Handbooks of Cancer Prevention in Tobacco Control, Volume 14: Effectiveness of Tax and Price Policies for Tobacco Control.* (2011).

22. World Health Organization. *WHO report on the global tobacco epidemic, 2015: raising taxes on tobacco.* http://www.who.int/tobacco/global_report/2015/report/en/ (2015).

23. U.S. Department of Health and Human Services. *The health consequences of smoking—50 years of progress: a report of the Surgeon General, 2014*. http://www.surgeongeneral.gov/library/reports/50-years-of-progress (2014).

24. Jha, P. & Chaloupka, F. J. *Tobacco Control in Developing Countries.* (Oxford University Press, 2000).

25. Townsend, J., Roderick, P. & Cooper, J. Cigarette smoking by socioeconomic group, sex, and age: effects of price, income, and health publicity. *BMJ 309*, 923–927 (1994).

26. Siahpush, M., Wakefield, M. A., Spittal, M. J., Durkin, S. J. & Scollo, M. M. Taxation Reduces Social Disparities in Adult Smoking Prevalence. *Am. J. Prev. Med. 36*, 285–291 (2009).

27. Chaloupka, F. J. Rational Addictive Behavior and Cigarette Smoking. *J. Polit. Econ. 99*, 722–742 (1991).

28. Farrelly, M. C., Bray, J. W., Pechacek, T. & Woollery, T. Response by Adults to Increases in Cigarette Prices by Sociodemographic Characteristics. *South. Econ. J. 68*, 156–165 (2001).

29. Colman, G. J. & Remler, D. K. Vertical Equity Consequences of Very High Cigarette Tax Increases: If the Poor Are the Ones Smoking, How Could Cigarette Tax Increases Be Progressive? *J. Policy Anal. Manage. 27*, 376–400 (2008).

30. Franks, P. *et al.* Cigarette Prices, Smoking, and the Poor: Implications of Recent Trends. *Am. J. Public Health 97*, 1873–1877 (2007).

31. Onder, Z. The economics of tobacco in Turkey: new evidence and demand estimates. (2002).

32. Karki, Y. B., Pant, K. D. & Pande, B. R. *A Study on the Economics of Tobacco in Nepal.* (World Bank, Washington, DC, 2003).

33. Sarntisart, I. An economic analysis of tobacco control in Thailand. (2003).

34. Levy, D. T., Chaloupka, F. J. & Gitchell, J. The effects of tobacco control policies on smoking rates: a tobacco control scorecard. *J. Public Health Manag. Pract. 10*, 338–353 (2004).

35. Chaloupka, F. J., Yurekli, A. & Fong, G. T. Tobacco taxes as a tobacco control strategy. *Tob. Control 21*, 172 (2012).

36. Chávez, R. Price elasticity of demand for cigarettes and alcohol in Ecuador, based on household data. *Rev. Panam. Salud Publica Pan Am. J. Public Health 40*, 222–228 (2016).

37. Gonzalez-Rozada, M. & Ramos-Carbajales, A. Implications of raising cigarette excise taxes in Peru. Rev. *Panam. Salud Publica Pan Am. J. Public Health 40*, 250–255 (2016).

38. Gjika, A., Zhllima, E., Rama, K. & Imami, D. Analysis of Tobacco Price Elasticity in Albania Using Household Level Data. *Int. J. Environ. Res. Public. Health 17*, E432 (2020).

39. Cruces, G., Falcone, G. & Puig, J. *Tobacco taxes in Argentina: Toward a comprehensive cost-benefit analysis*. https://tobacconomics.org/files/research/592/CEDLAS_FinalReport_EN.pdf (2020).

40. Nargis, N. *et al.* The price sensitivity of cigarette consumption in Bangladesh: evidence from the International Tobacco Control (ITC) Bangladesh Wave 1 (2009) and Wave 2 (2010) Surveys. *Tob. Control 23*, i39–i47 (2014).

41. Gligorić, D., Preradović Kulovac, D., Mićić, L. & Pepić, A. Price and income elasticity of cigarette demand in Bosnia and Herzegovina by different socioeconomic groups. *Tob. Control* tobaccocontrol-2021-056881 (2022) doi:10.1136/tobaccocontrol-2021-056881.

42. Divino, J. A., Ehrl, P., Candido, O. & Valadao, M. A. P. Extended cost-benefit analysis of tobacco taxation in Brazil. *Tob. Control* tobaccocontrol-2021-056806 (2021) doi:10.1136/tobaccocontrol-2021-056806.

43. Huang, J., Zheng, R., Chaloupka, F. J., Fong, G. T. & Jiang, Y. Differential responsiveness to cigarette price by education and income among adult urban Chinese smokers: findings from the ITC China Survey. *Tob. Control 24*, iii76–iii82 (2015).

44. Verguet, S. *et al.* The consequences of tobacco tax on household health and finances in rich and poor smokers in China: an extended cost-effectiveness analysis. *Lancet Glob. Health 3*, e206–e216 (2015).

45. Paraje, G., Araya, D., De Paz, A. & Nargis, N. Price and expenditure elasticity of cigarette demand in El Salvador: a household-level analysis and simulation of a tax increase. *Tob. Control* tobaccocontrol-2019-055568 (2020) doi:10.1136/tobaccocontrol-2019-055568.

46. Selvaraj, S., Srivastava, S. & Karan, A. Price elasticity of tobacco products among economic classes in India, 2011–2012. *BMJ Open 5*, (2015).

47. Dauchy, E. P. & John, R. M. The Effect of Price and Tax Policies on the Decision to Smoke or Use Smokeless Tobacco in India. *Prev. Sci. Off. J. Soc. Prev. Res.* (2022) doi:10.1007/s11121-022-01360-w.

48. Adioetomo, S. M. & Djutaharta, T. Cigarette consumption, taxation, and household income: Indonesia case study. (2005).

49. Raei, B. *et al.* Distributional health and financial consequences of increased cigarette tax in Iran: extended cost-effectiveness analysis. *Health Econ. Rev. 11*, 30 (2021).

50. Kosovo. in *Impacts of Tobacco Excise Increases on Cigarette Consumption and Government Revenues in Southeastern European Countries* (Institute for Health Research and Policy, University of Illinois Chicago).

51. Macías Sánchez, A., Villarreal Páez, H. J., Méndez Méndez, J. S. & García Góme, A. *Extended Cost-Benefit Analysis of Tobacco Consumption in Mexico.* https://tobacconomics.org/files/research/605/extended-cost-benefit-analysis-tobacco-ciep-en.pdf (2020).

52. Cizmovic, M., Mugosa, A., Kovacevic, M. & Lakovic, T. Effectiveness of tax policy changes in Montenegro: smoking behaviour by socio-economic status. *Tob. Control* tobaccocontrol-2021-056876 (2022) doi:10.1136/tobaccocontrol-2021-056876.

53. Nayab, D., Nasir, M., Memon, J. A., Khalid, M. & Hussain, A. Estimating the price elasticity for cigarette and chewed tobacco in Pakistan: evidence from microlevel data. *Tob. Control 29*, s319 (2020).

54. de los Rios, C., Medina, D. & Aguilar, J. Cost-benefit analysis of tobacco consumption in Peru. *Inst. Estud. Peru. Doc. Trab. No 270* (2020).

55. Vladisavljević, M., Zubović, J., Đukić, M. & Jovanović, O. Inequality-Reducing Effects of Tobacco Tax Increase: Accounting for Behavioral Response of Low-, Middle-, and High-Income Households in Serbia. *Int. J. Environ. Res. Public. Health 18*, (2021).

56. Kidane, A., Mduma, J., Naho, A. & Hu, T.-W. Impact of Smoking on Food Expenditure among Tanzanian Households. *Afr. Stat. J. J. Stat. Afr. 18*, 69–78 (2015).

57. Jankhotkaew, J., Pitayarangsarit, S., Chaiyasong, S. & Markchang, K. Price elasticity of demand for manufactured cigarettes and roll-your-own cigarettes across socioeconomic status groups in Thailand. *Tob. Control 30*, 542–547 (2021).

58. Onder, Z. & Yurekli, A. A. Who pays the most cigarette tax in Turkey. *Tob. Control 25*, 39 (2016).

59. Keeler, T. E., Hu, T.-W., Barnett, P. G. & Manning, W. G. Taxation, regulation, and addiction: A demand function for cigarettes based on time-series evidence. *J. Health Econ. 12*, 1–18 (1993).

60. Hu, T. W., Bai, J., Keeler, T. E., Barnett, P. G. & Sung, H. Y. The impact of California Proposition 99, a major anti-smoking law, on cigarette consumption. *J. Public Health Policy 15*, 26–36 (1994).

61. Hu, T. W., Sung, H. Y. & Keeler, T. E. Reducing cigarette consumption in California: tobacco taxes vs an anti-smoking media campaign. *Am. J. Public Health 85*, 1218–1222 (1995).

62. Sung, H.-Y., Hu, T.-W. & Keeler, T. E. Cigarette Taxation and Demand: An Empirical Model. *Contemp. Econ. Policy 12*, 91–100 (1994).

63. Deaton, A. & Muellbauer, J. An Almost Ideal Demand System. *Am. Econ. Rev. 70*, 312–326 (1980).

64. Deaton, A. S. Quality, Quantity, and Spatial Variation of Price. *Am. Econ. Rev*. 78, 418–430 (1988).

65. Deaton, A. Household survey data and pricing policies in developing countries. *World Bank Econ. Rev. 3*, 183–210 (1989).

66. Deaton, A. Price elasticities from survey data: Extensions and Indonesian results. *J. Econom. 44*, 281–309 (1990).

67. Deaton, A. & Grimard, F. *Demand Analysis and Tax Reform in Pakistan.* http://www.worldbank.org/html/prdph/lsms/research/wp/a81_100.html#wp85 (1992).

68. Ahmed, N., Mozumder, T. A., Hassan, M. T. & Huque, R. Demand for tobacco products in Bangladesh. *Tob. Control* tobaccocontrol-2020-056297 (2021) doi:10.1136/tobaccocontrol-2020-056297.

69. Gligorić, D., Pepić, A., Petković, S., Ateljević, J. & Vukojević, B. Price elasticity of demand for cigarettes in Bosnia and Herzegovina: microdata analysis. *Tob. Control 29*, s304–s309 (2020).

70. Gligorić, D., Kulovac, D. P., Mićić, L. & Pepić, A. Price and income elasticity of cigarette demand in Bosnia and Herzegovina by different socioeconomic groups. *Tob. Control* (2022) doi:10.1136/tobaccocontrol-2021-056881.

71. Chen, Y. & Xing, W. Quantity, quality, and regional price variation of cigarettes: Demand analysis based on a household survey in China. *China Econ. Rev. 22*, 221–232 (2011).

72. John, R. M. *et al. The Economics of Tobacco and Tobacco Taxation in India.* (2010).

73. John, R. M. Consumption of Tobacco in India: An Economic Analysis. (Indira Gandhi Institute of Development Research, 2007).

74. John, R. M. Price Elasticity Estimates for Tobacco in India. *Health Policy Plan. 23*, 200–209 (2008).

75. Guindon, G. E., Nandi, A., Chaloupka, F. J. & Jha, P. Socioeconomic Differences in the Impact of Smoking Tobacco and Alcohol Prices on Smoking in India. *Natl. Bur. Econ. Res. Work. Pap. Ser. No. 17580*, (2011).

76. Mugosa, A., Cizmovic, M., Lakovic, T. & Popovic, M. Accelerating progress on effective tobacco tax policies in Montenegro. *Tob. Control 29*, s293–s299 (2020).

77. Cizmovic, M., Mugosa, A., Kovacevic, M. & Lakovic, T. Effectiveness of tax policy changes in Montenegro: smoking behaviour by socio-economic status. *Tob. Control* (2022) doi:10.1136/tobaccocontrol-2021-056876.

78. Nayab, D., Nasir, M., Memon, J. A., Khalid, M. & Hussain, A. Estimating the price elasticity for cigarette and chewed tobacco in Pakistan: evidence from microlevel data. *Tob. Control 29*, s319–s325 (2020).

79. Vladisavljevic, M., Zubović, J., Đukić, M. & Jovanović, O. Tobacco price elasticity in Serbia: evidence from a middle-income country with high prevalence and low tobacco prices. *Tob. Control 29*, s331–s336 (2020).

80. Vladisavljević, M., Zubović, J., Đukić, M. & Jovanović, O. Inequality-Reducing Effects of Tobacco Tax Increase: Accounting for Behavioral Response of Low-, Middle-, and High-Income Households in Serbia. *Int. J. Environ. Res. Public. Health 18*, 9494 (2021).

81. Dare, C., Boachie, M. K., Tingum, E. N., Abdullah, S. M. & van Walbeek, C. Estimating the price elasticity of demand for cigarettes in South Africa using the Deaton approach. *BMJ Open 11*, e046279 (2021).

82. Chelwa, G. & van Walbeek, C. Does cigarette demand respond to price increases in Uganda? Price elasticity estimates using the Uganda National Panel Survey and Deaton's method. *BMJ Open 9*, e026150 (2019).

83. Eozenou, P. & Fishburn, B. *Price Elasticity Estimates for Cigarette Demand in Vietnam.* **https://ideas.repec.org/p/dpc/wpaper/0509.html** (2009).

84. McKelvey, C. Price, unit value, and quality demanded. *J. Dev. Econ. 95*, 157–169 (2011).

85. Gibson, J. & Rozelle, S. Prices and Unit Values in Poverty Measurement and Tax Reform Analysis. *World Bank Econ. Rev. 19*, 69–97 (2005).

86. Drope, J. *et al. The Tobacco Atlas.* **https://tobaccoatlas.org/** (2022).

87. Mullahy, J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *J. Health Econ. 17*, 247–281 (1998).

88. Manning, W. Dealing with skewed data on costs and expenditures. in *The Elgar Companion to Health Economics* 439–446 (Edward Elgar, 2006).

89. Manning, W. G. The logged dependent variable, heteroscedasticity, and the retransformation problem. *J. Health Econ. 17*, 283–295 (1998).

90. Duan, N. Smearing Estimate: A Nonparametric Retransformation Method. *J. Am. Stat. Assoc. 78*, 605–610 (1983).

91. Jones, A. M. *Models For Health Care.* **https://www.york.ac.uk/media/economics/documents/herc/wp/10_01.pdf** (2010).

92. Cragg, J. G. Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica 39*, 829–844 (1971).

93. Belotti, F., Deb, P., Manning, W. G. & Norton, E. C. Twopm: Two-Part Models. *Stata J. 15*, 3–20 (2015).

94. Tauras, J. A. An Empirical Analysis of Adult Cigarette Demand. East. *Econ. J. 31*, 361–375 (2005).

95. Kostova, D., Ross, H., Blecher, E. & Markowitz, S. *Prices and Cigarette Demand: Evidence from Youth Tobacco Use in Developing Countries.* http://www.nber.org/papers/w15781 (2010) doi:10.3386/w15781.

96. Ross, H. & Chaloupka, F. J. *The effect of cigarette prices on youth smoking*. *Health Econ. 12*, 217–230 (2003).

97. Nikaj, S. & Chaloupka, F. J. The effect of prices on cigarette use among youths in the global youth tobacco survey. *Nicotine Tob. Res. Off. J. Soc. Res. Nicotine Tob. 16 Suppl 1*, S16-23 (2014).

98. Joseph, R. A. & Chaloupka, F. J. The Influence of Prices on Youth Tobacco Use in India. *Nicotine Tob. Res. 16*, S24–S29 (2014).

99. Kostova, D. *et al.* Exploring the relationship between cigarette prices and smoking among adults: a cross-country study of low- and middle-income nations. *Nicotine Tob. Res. Off. J. Soc. Res. Nicotine Tob. 16 Suppl 1*, S10-15 (2014).

100. Manning, W. G. & Mullahy, J. Estimating log models: to transform or not to transform? *J. Health Econ. 20*, 461–494 (2001).

101. William H. Greene. *Econometric Analysis.* (Prentice Hall, 2002).

102. Zellner, A. An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *J. Am. Stat. Assoc. 57*, 348–368 (1962).

103. Zellner, A. Estimators for Seemingly Unrelated Regression Equations: Some Exact Finite Sample Results. *J. Am. Stat. Assoc. 58*, 977–992 (1963).

104. Menon, M., Perali, F. & Tommasi, N. Estimation of unit values in household expenditure surveys without quantity information. *Stata J. 17*, 222–239 (2017).

105. Atella, V., Menon, M. & Perali, F. *Estimation of Unit Values in Cross Sections Without Quantity Information and Implications for Demand and Welfare Analysis.* https://papers.ssrn.com/abstract=391481 (2003).

106. Coondoo, D., Majumder, A. & Ray, R. A Method of Calculating Regional Consumer Price Differentials with Illustrative Evidence from India. R*ev. Income Wealth 50*, 51–68 (2004).

107. Slesnick, D. T. Prices and demand: New evidence from micro data. *Econ. Lett. 89*, 269–274 (2005).

108. Hoderlein, S. & Mihaleva, S. Increasing the price variation in a repeated cross section. *J. Econom. 147*, 316–325 (2008).

109. Lecocq, S. & Robin, J.-M. Estimating almost-ideal demand systems with endogenous regressors. *Stata J. 15*, 554–573 (2015).

110. Castellón, C. E., Boonsaeng, T. & Carpio, C. E. Demand system estimation in the absence of price data: an application of Stone-Lewbel price indices. *Appl. Econ. 47*, 553–568 (2015).

111. Lewbel, A. Identification and Estimation of Equivalence Scales under Weak Separability. *Rev. Econ. Stud. 56*, 311–316 (1989).

112. Lewbel, A. & Pendakur, K. Tricks with Hicks: The EASI Demand System. *Am. Econ. Rev. 99*, 827–863 (2009).

113. Moro, D., Castellari, E. & Sckokai, P. Empirical issues in the computation of Stone–Lewbel price indexes in censored micro-level demand systems. *Appl. Econ. Lett. 25*, 557–561 (2018).

114. WHO. *WHO global report on trends in prevalence of tobacco use 2000-2025, fourth edition.* https://www.who.int/publications/i/item/9789240039322 (2021).

115. World Health Organization. *Tobacco: Key Facts*. https://www.who.int/news-room/fact-sheets/detail/tobacco (2022).

116. World Health Organization. *Systematic review of the link between tobacco and poverty.* http://www.who.int/tobacco/publications/syst_rev_tobacco_poverty/en/index.html (2014).

117. John, R. M. Crowding out effect of tobacco expenditure and its implications on household resource allocation in India. *Soc. Sci. Med. 66*, 1356–1367 (2008).

118. Efroymson, D. Hungry for tobacco: an analysis of the economic impact of tobacco consumption on the poor in Bangladesh. *Tob. Control 10*, 212–217 (2001).

119. Thomson, G. W., Wilson, N. A., D Dea, Reid, P. J. & Chapman, P. H. Tobacco spending and children in low income households. *Tob. Control 11*, 372–375 (2002).

120. Busch, S. H., Jofre-Bonet, M., Falba, T. A. & Sindelar, J. L. Burning a Hole in the Budget: Tobacco Spending and its Crowd-Out of Other Goods. *Appl. Health Econ. Health Policy. 3*, 263–272 (2004).

121. Wang, H., Sindelar, J. L. & Busch, S. H. The impact of tobacco expenditure on household consumption patterns in rural china. *Soc. Sci. Med. 62*, 1414–1426 (2006).

122. Pu, C., Lan, V., Chou, Y.-J. & Lan, C. The crowding-out effects of tobacco and alcohol where expenditure shares are low: Analyzing expenditure data for Taiwan. *Soc. Sci. Med. 66*, 1979–1989 (2008).

123. Koch, S. F. & Tshiswaka-Kashalala, G. *Tobacco Substitution and the Poor.* https://www.up.ac.za/media/shared/Legacy/UserFiles/wp_2008_32.pdf (2008).

124. John, R. M., Ross, H. & Blecher, E. Tobacco expenditures and its implications for household resource allocation in Cambodia. *Tob. Control 21*, 341–346 (2012).

125. Chelwa, G. & Walbeek, C. van. *Assessing the Causal Impact of Tobacco Expenditure on Household Spending Patterns in Zambia.* https://econrsa.org/2017/wp-content/uploads/working_paper_453.pdf (2014).

126. San, S. & Chaloupka, F. J. The impact of tobacco expenditures on spending within Turkish households. *Tob. Control 25*, 558–563 (2016).

127. Husain, M. J., Datta, B. K., Virk-Baker, M. K., Parascandola, M. & Khondker, B. H. The crowding-out effect of tobacco expenditure on household spending patterns in Bangladesh. *PLOS ONE 13*, e0205120 (2018).

128. Ross, H., Moussa, L., Harris, T. & Ajodhea, R. The heterogeneous impact of a successful tobacco control campaign: a case study of Mauritius. *Tob. Control 27*, 83–89 (2018).

129. Paraje, G. & Araya, D. Relationship between smoking and health and education spending in Chile. *Tob. Control 27*, 560–567 (2018).

130. Nguyen, N.-M. & Nguyen, A. Crowding-out effect of tobacco expenditure in Vietnam. *Tob. Control 29*, s326–s330 (2020).

131. Masa-ud, A. G. A., Chelwa, G. & van Walbeek, C. Does tobacco expenditure influence household spending patterns in Ghana?: Evidence from the Ghana 2012/2013 Living Standards Survey. *Tob. Induc. Dis. 18*, 48 (2020).

132. Nyagwachi, A. O., Chelwa, G. & van Walbeek, C. The effect of tobacco- and alcohol-control policies on household spending patterns in Kenya: An approach using matched difference in differences. *Soc. Sci. Med. 256*, 113029 (2020).

133. Block, S. & Webb, P. Up in Smoke: Tobacco Use, Expenditure on Food, and Child Malnutrition in Developing Countries. *Econ. Dev. Cult. Change 58*, 1–23 (2009).

134. Chelwa, G. & Koch, S. F. The effect of tobacco expenditure on expenditure shares in South African households: A genetic matching approach. *PLOS ONE 14*, e0222000 (2019).

135. Jin, H. J. & Cho, S. M. Effects of cigarette price increase on fresh food expenditures of low-income South Korean households that spend relatively more on cigarettes. *Health Policy Amst. Neth. 125*, 75–82 (2021).

136. Do, Y. K. & Bautista, M. A. Tobacco use and household expenditures on food, education, and healthcare in low- and middle-income countries: a multilevel analysis. *BMC Public Health 15*, (2015).

137. Djutaharta, T., Nachrowi, N. D., Ananta, A. & Martianto, D. Impact of price and non-price policies on household cigarette consumption and nutrient intake in smoking-tolerant Indonesia. B*MJ Open 11*, e039211 (2021).

138. Vladisavljevic, M., Zubovic, J., Đukić, M. & Jovanović, O. Crowding-out effect of tobacco consumption in Serbia. *Tob. Prev. Cessat. 8*, (2022).

139. Wisana, I. D. G. K., Swarnata, A., Kamilah, F. Z., Meilissa, Y. & Kusnadi, G. *The Crowding-out Effect of Tobacco Consumption in Indonesia.* (2022).

140. Mugoša, A., Čizmović, M. & Vulović, V. *Impact of tobacco spending on intra-household resource allocation in Montenegro.* **https://tobacconomics.org/impact-of-tobacco-spending-on-intra-householdresource-allocation-in-montenegro-working-paper-series/** (2022).

141. Gómez, A. G., Macías, A. & Páez, H. J. V. *Crowding-Out and Impoverishing Effect of Tobacco in Mexico.* **https://www.tobacconomics.org/research/crowding-out-and-impoverishing-effect-of-tobacco-in-mexico/** (2022).

142. Lassi, Z. S., Ali, A. & Meherali, S. Women's Participation in Household Decision Making and Justification of Wife Beating: A Secondary Data Analysis from Pakistan's Demographic and Health Survey. *Int. J. Environ. Res. Public. Health 18*, 10011 (2021).

143. Seidu, A.-A., Dzantor, S., Sambah, F., Ahinkorah, B. O. & Ameyaw, E. K. Participation in household decision making and justification of wife beating: evidence from the 2018 Mali Demographic and Health Survey. *Int. Health 14*, 74–83 (2022).

144. World Health Organization. *WHO report on the global tobacco epidemic, 2017: Monitoring tobacco use and prevention policies.* **http://apps.who.int/iris/bitstream/10665/255874/1/9789241512824-eng.pdf?ua=1** (2017).

145. Browning, M. & Meghir, C. The Effects of Male and Female Labor Supply on Commodity Demands. *Econometrica 59*, 925–951 (1991).

146. Banks, J., Blundell, R. & Lewbel, A. Quadratic Engel Curves and Consumer Demand. *Rev. Econ. Stat. 79*, 527–539 (1997).

147. Pollak, R. A. & Wales, T. J. *Demand System Specification and Estimation.* (Oxford University Press, 1995).

148. Davidson, R. & MacKinnon, J. G. *Estimation and Inference in Econometrics.* (1993).

149. Baum, C., Schaffer, M. & Stillman, S. Instrumental variables and GMM: Estimation and testing. *Stata J. 3*, 1–31 (2003).

150. Zellner, A. & Theil, H. Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations. *Econometrica 30*, 54–78 (1962).

151. StataCorp. Stata base reference mannual Release 15. (2017).

152. Vermeulen, F. Do Smokers Behave Differently? A Tale of Zero Expenditures and Separability Concepts. *Econ. Bull. 4*, 1–7 (2003).

153. Diamond, A. & Sekhon, J. S. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Rev. Econ. Stat. 95*, 932–945 (2013).

154. Abadie, A. Semiparametric Difference-in-Differences Estimators. *Rev. Econ. Stud. 72*, 1–19 (2005).

155. Bertrand, M., Duflo, E. & Mullainathan, S. How Much Should We Trust Differences-In-Differences Estimates?*. *Q. J. Econ. 119*, 249–275 (2004).

156. Stuart, E. A. *et al.* Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Serv. Outcomes Res. Methodol. 14*, 166–182 (2014).

157. Baum, C. F., Schaffer, M. E. & Stillman, S. IVREG2: Stata module for extended instrumental variables/2SLS and GMM estimation. (2007).

158. Staiger, D. & Stock, J. H. Instrumental Variables Regression with Weak Instruments. *Econometrica 65*, 557–586 (1997).

159. Shehata, E. A. E. LMHREG3: Stata module to compute Overall System Heteroscedasticity Tests after (3SLS-SURE) Regressions. (2011).

160. Sreeramareddy, C. T., Harper, S. & Ernstsen, L. Educational and wealth inequalities in tobacco use among men and women in 54 low-income and middle-income countries. *Tob. Control 27*, 26–34 (2018).

161. World Bank. WDI - Poverty and Inequality. http://datatopics.worldbank.org/world-development-indicators/themes/poverty-and-inequality.html#national-poverty-lines (2018).

162. Statistics South Africa. *National Poverty Lines.* http://www.statssa.gov.za/publications/P03101/P031012018.pdf (2018).

163. US Census Bureau. Poverty. https://www.census.gov/topics/income-poverty/poverty.html (2018).

164. Department of Health and Human Services, Office of the Secretary. *Annual update of the HHS poverty guidelines.* https://www.govinfo.gov/content/pkg/FR-2022-01-21/pdf/2022-01166.pdf (2022).

165. Foster, J., Seth, S., Lokshin, M. & Sajaia, Z. *A unified approach to measuring poverty and inequality : theory and practice.* (The World Bank, 2013).

166. B. Seebohm Rowntree. *Poverty: a study of town life.* (MacMillan, 1901).

167. Fuchs Tarlovsky, A., Del Carmen, G. & Mukong, A. K. *Long-run impacts of increasing tobacco taxes : evidence from South Africa.* 1–39 http://documents.worldbank.org/curated/en/122081521480061194/Long-run-impacts-of-increasing-tobacco-taxes-evidence-from-South-Africa (2018).

168. Wagstaff, A. & Doorslaer, E. van. *Paying for health care : quantifying fairness, catastrophe, and impoverishment, with applications to Vietnam, 1993-98.* http://ideas.repec.org/p/wbk/wbrwps/2715.html (2001).

169. Liu, Y., Rao, K., Hu, T., Sun, Q. & Mao, Z. Cigarette smoking and poverty in China. *Soc. Sci. Med. 63*, 2784–2790 (2006).

170. John, R. M., Sung, H.-Y., Max, W. B. & Ross, H. Counting 15 million more poor in India, thanks to tobacco. *Tob. Control 20*, 349–352 (2011).

171. Belvin, C., Britton, J., Holmes, J. & Langley, T. Parental smoking and child poverty in the UK: an analysis of national survey data. *BMC Public Health 15*, 507 (2015).

172. Howard Reed. *Estimates of poverty in the UK adjusted for expenditure on tobacco.* http://ash.org.uk/information-and-resources/health-inequalities/health-inequalities-resources/estimates-of-poverty-in-the-uk-adjusted-for-expenditure-on-tobacco/ (2015).

173. Nyakutsikwa, B., Britton, J. & Langley, T. The effect of tobacco and alcohol consumption on poverty in the United Kingdom. *Addict. Abingdon Engl. 116*, 150–158 (2021).

174. Nguyen, A. N., Nguyen, N.-M. & Bui, T. H. *The Impoverishing Effect of Tobacco Use in Viet Nam (Report)*. https://tobacconomics.org/files/research/730/dpc-rp-poverty-final.pdf (2021).

175. Ravallion, M. *Poverty comparisons : a guide to concepts and methods.* 1 http://documents.worldbank.org/curated/en/290531468766493135/Poverty-comparisons-a-guide-to-concepts-and-methods (1992).

176. Regier, G., Zereyesus, Y. A., Dalton, T. J. & Amanor-Boadu, V. Do Adult Equivalence Scales Matter in Poverty Estimates? A Northern Ghana Case Study and Simulation. *J. Int. Dev. 31*, 80–100 (2019).

177. Cutler, D. M. et al. How Good a Deal Was the Tobacco Settlement?: Assessing Payments to Massachusetts. *J. Risk Uncertain. 21*, 235–261 (2000).

178. World Health Organization. *Assessment of the economic costs of smoking. World Health Organization economicsof tobacco toolkit.* http://whqlibdoc.who.int/publications/2011/9789241501576_eng.pdf (2011).

179. WHO, W. H. O. Asssessment of the Economic Cost of Smoking. (2011).

180. Jenkins, S. P. POVDECO: Stata module to calculate poverty indices with decomposition by subgroup. (2008).

7 Apéndice de código

7.1 Do-file de Stata para estimar las elasticidades prevalencia y cantidad de un solo producto

```
*=====================================================================
* Fecha: 20 de enero de 2023
* Tema: Do-file de Stata creado como parte del conjunto de herramientas para el
* Uso de Encuestas de Gastos de los Hogares para la Investigación en
* Economía del Control del Tabaco
* Este do-file estima la elasticidad precio y gasto
* elasticidad para un solo producto, por ejemplo, cigarrillos.
* Base de datos utilizada: hbs_data.dta
* Variables clave:
* - exptotal - gasto total del hogar en unidades monetarias locales (UML)
* - expcig - gasto total en cigarrillos del hogar en UML
* - qcig - número de cigarrillos o paquetes de cigarrillos comprados
* - hsize - tamaño del hogar
* - meanedu - educación media del hogar en años
* - maxedu - educación máxima del hogar en años
* - sgroup - variable categórica que representa los grupos sociales del hogar
* - maleratio – ratio entre el número de hombres y el tamaño del hogar
* - clust - variable que identifica la unidad primaria de muestreo o conglomerado

*=====================================================================
clear
versión 15
set mem 1000m
set more off
cd "Directory Path"
capture log close
log using Elasticity.log, replace
use hbs_data, clear
drop if [qcig==.&expcig!=.]|[qcig!=.&expcig==.]
gen uvcig=expcig/qcig
gen luvcig=ln(uvcig)
gen bscig=expcig/exptotal
gen lhsize=ln(hsize)
```

```
gen lexp=ln(exptotal)
tab sgroup, gen(sgp)
drop if [qcig==.&expcig!=.]|[qcig!=.&expcig==.]
gen dcig=0 if qcig==. | qcig==0
replace dcig=1 if qcig>0 & qcig!=.
by clust, sort: egen cigclustsize =sum(dcig)
drop if cigclustsize <2

*************************************************************************
*Estimación de la elasticidad de la prevalencia utilizando el modelo logit
*************************************************************************
egen pcig=mean(uvcig), by(clust)
egen pcig2=mean(uvcig), by(region)
replace pcig=pcig2 if pcig==.
global xvar "pcig lexp lhsize maleratio meanedu maxedu sgp1 sgp2 sgp3"
logit dcig $xvar
outreg2 using PrevalenceElast.doc, replace sideway
predict yhat_p, pr
margins, eyex(pcig)
margins, eydx(lexp)

*Regression diagnostics
*logit dcig $xvar
*linktest
*logit dcig $xvar
*lfit, group (10) table

*************************************************************************
*Estimación de la elasticidad cantidad utilizando el modelo de Deaton
*************************************************************************
keep if dcig==1
*anova luvcig clust
areg luvcig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
outreg2 using FirstStagereg.doc, replace ctitle("Unit value Regression")
scalar b1=_coef[lexp]
predict ruvcig, resid
scalar sigma11=$S_E_sse / $S_E_tdf
gen y1cig=luvcig-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio ///
                -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu ///
                -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3
areg bscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
outreg2 using FirstStagereg.doc, append ctitle("Budget share Regression")
predict rbscig, resid
scalar sigma22=$S_E_sse/$S_E_tdf
scalar b0=_coef[lexp]
gen y0cig=bscig-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio ///
                -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu ///
                -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3
```

```
qui areg ruvcig rbscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
scalar sigma12=_coef[rbscig]*sigma22

qui sum bscig
scalar Wbar=r(mean)
scalar Expel=1-b1+(b0/Wbar)
*Elasticidad gasto de cantidad
di Expel
*Para estimar los errores estándar de la elasticidad gasto mediante bootstrap
cap program drop Expelast
program Expelast, rclass
tempname b1 b0 Wbar
qui areg luvcig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
cap scalar b1=_coef[lexp]
qui areg bscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
cap scalar b0=_coef[lexp]
qui sum bscig
cap scalar Wbar=r(mean)
return scalar Expel=1-b1+(b0/Wbar)
end
Expelast
return list
bootstrap Expel=r(Expel), reps(1000) seed(1): Expelast

sort clust
egen y0c= mean(y0cig), by(clust)
egen n0c=count(y0cig), by(clust)
egen y1c= mean(y1cig), by(clust)
egen n1c=count(y1cig), by(clust)
sort clust
qui by clust: keep if _n==1
ameans n0c
scalar n0=r(mean_h)
ameans n1c
scalar n1=r(mean_h)
drop n0c n1c

*Para estimar los errores estándar de la elasticidad precio mediante bootstrap
cap program drop elast
program elast, rclass
tempname S R num den phi theta psi
qui corr y0c y1c, cov
scalar S=r(Var_2)
scalar R=r(cov_12)
scalar num=scalar(R)-(sigma12/n0)
scalar den=scalar(S)-(sigma11/n1)
cap scalar phi=num/den
```

```
cap scalar zeta= b1/((b0 + Wbar*(1-b1)))
cap scalar theta=phi/(1+(Wbar-phi)*zeta)
cap scalar psi=1-((b1*(Wbar-theta))/(b0+Wbar))
return scalar EP=(theta/Wbar)-psi
end
elast
return list
bootstrap EP=r(EP), reps(1000) seed(1): elast
log close
clear all
```

7.2 Do-file de Stata para estimar las elasticidades de prevalencia y de cantidad de un solo producto por grupos de ingresos

```
*================================================================
* Fecha: 20 de enero de 2023
* Tema: Do-file de Stata creado como parte del conjunto de herramientas sobre el Uso
* de Encuestas de Gastos de los Hogares para Investigación en Economía del
* Control del Tabaco. Este do-file estima la elasticidad precio y
* elasticidad gasto (tanto elasticidades de prevalencia como de cantidad) para un solo
* producto, por ejemplo, cigarrillo, por diferentes grupos de ingresos.
* Base de datos utilizada: hbs_data.dta

* Variables clave:
* - exptotal - gasto total del hogar en unidades monetarias locales (UML)
* - expcig - gasto total en cigarrillos del hogar en UML
* - qcig - número de cigarrillos o paquetes de cigarrillos comprados
* - hsize - tamaño del hogar
* - meanedu - educación media del hogar en años
* - maxedu - educación máxima del hogar en años
* - sgroup - variable categórica que representa los grupos sociales del hogar
* - maleratio - ratio entre el número de hombres y el tamaño del hogar
* - clust - variable que identifica la unidad primaria de muestreo o conglomerado

*================================================================
clear
versión 15
set mem 1000m
set more off
*Agregue la ruta del directorio entre comillas
cd "Directory Path"
capture log close
log using Elasticity.log, replace
use hbs_data, clear
drop if [qcig==.&expcig!=.]|[qcig!=.&expcig==.]
gen uvcig=expcig/qcig
```

```
gen luvcig=ln(uvcig)
gen bscig=expcig/exptotal
gen lhsize=ln(hsize)
gen lexp=ln(exptotal)
tab sgroup, gen(sgp)
gen exppc=exptotal/hsize
xtile inc = exppc [w=weights], nq(3)
tab inc
xtile inc_temp = exppc, nq(3)
egen subclust=group(clust inc)
gen dcig=0 if qcig==. | qcig==0
replace dcig=1 if qcig>0 & qcig!=.
bys subclust: egen cigsubclust =sum(dcig)
drop if cigsubclust <2

*****************************************************************************
* Estimación de la elasticidad de la prevalencia utilizando el modelo logit
*****************************************************************************
egen pcig=mean(uvcig), by(clust)
egen pcig2=mean(uvcig), by(region)
replace pcig=pcig2 if pcig==.
global xvar "pcig lexp lhsize maleratio meanedu maxedu sgp1 sgp2 sgp3"
local append "replace"
forvalues i=1/3 {
        logit dcig $xvar if inc==`i'
        outreg2 using PrevalenceElastInc.doc, ctitle (Income group: `i') `append'
        predict yhat_p`i', pr
        *margins, eyex(pcig) coeflegend post
        margins, eyex(pcig)
        estimates store inc`i'
        *margins, eydx(lexp)
        local append "append"
}
suest inc*
test [inc1_dcig]pcig-[inc2_dcig]pcig=0
test [inc1_dcig]pcig-[inc3_dcig]pcig=0
test [inc2_dcig]pcig-[inc3_dcig]pcig=0

*Regression diagnostics
logit dtob pcig $xvar if inc==1
linktest
logit dtob pcig $xvar if inc==1
lfit, group (10) table
```

```
*************************************************************************
* Estimación de la elasticidad cantidad utilizando el modelo de Deaton
*************************************************************************
keep if dcig==1
*anova luvcig clust
areg luvcig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
predict ruvcig, resid
scalar sigma11=$S_E_sse / $S_E_tdf
scalar b1=_coef[lexp]
gen y1cig=luvcig-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio ///
                -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu ///
                -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3
* La depuración y el promedio de valores unitarios para la segunda etapa se realizan
* para todos los grupos de ingresos combinados para que todos los hogares en el
* mismo conglomerado enfrenten los mismos valores unitarios promedio.
forvalues i=1/3 {
        areg bscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3 if inc==`i', absorb(clust)
        predict rbscig`i', resid
        scalar sigma22`i'=$S_E_sse/$S_E_tdf
        scalar b0`i'=_coef[lexp]
        gen y0cig`i'=bscig-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio ///
                -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu ///
                -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3 if inc==`i'
        qui areg ruvcig rbscig`i' lexp lhsize maleratio meanedu maxedu sgp1-sgp3 if inc==`i',
absorb(clust)
        scalar sigma12`i'=_coef[rbscig`i']*sigma22`i'
}

*Elasticidad del gasto por grupos de ingresos con errores estándar de bootstrap
cap program drop Expelast
program define Expelast, rclass
args i
qui areg bscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3 if inc==`i', absorb(clust)
local a0 =_coef[lexp]
qui sum bscig if inc==`i'
local vbar =r(mean)
return scalar Expel`i' =1-b1+(`a0'/`vbar')
end
forvalues i=1/3 {
bootstrap Expel`i'=r(Expel`i'), reps(1000) seed(1): Expelast `i'
estimates store exp_el`i'
}
*Para estimar la elasticidad precio y los errores estándar mediante bootstrap
cap program drop elast
program define elast, rclass
        qui sum inc
        local a=r(mean)
```

```
        global j=`a'
        tempname S R num den phi theta psi
        qui corr y0c$j y1c, cov
        scalar S=r(Var_2)
        scalar R$j=r(cov_12)
        scalar num$j = scalar(R$j) - (sigma12$j / n0$j)
        scalar den=scalar(S)-(sigma11/n1)
        cap scalar phi$j = num$j / den
        cap scalar zeta$j = b1/((b0$j + Wbar$j * (1-b1)))
        cap scalar theta$j =phi$j /(1+(Wbar$j - phi$j) * zeta$j)
        cap scalar psi$j = 1-((b1*(Wbar$j - theta$j ))/(b0$j + Wbar$j))
        return scalar EP$j = (theta$j / Wbar$j )- psi$j
end

local append "replace"
forvalues i=1/3 {
        preserve
        egen y1c= mean(y1cig), by(clust)
        keep if inc==`i'
        sort clust
        egen y0c`i'= mean(y0cig`i'), by(clust)
        egen n0c`i'=count(y0cig`i'), by(clust)
        egen n1c=count(y1cig), by(clust)
        qui sum bscig if inc==`i'
        scalar Wbar`i' = r(mean)
        sort clust
        qui by clust: keep if _n==1
        qui ameans n0c`i'
        scalar n0`i'=r(mean_h)
        qui ameans n1c
        scalar n1=r(mean_h)
        drop n0c`i' n1c
        elast
        return list
        bootstrap EP`i'=r(EP`i'), reps(1000) seed(1): elast
        outreg2 using DeatonPriceElast.doc, dec(4) ctitle (Income group: `i') `append'
        local append "append"
        restore
}
```

```
**********************************
*Utilice el código a continuación si desea probar si las elasticidades entre grupos de
* ingresos son estadísticamente diferentes
**********************************
cap program drop elast
program define elast, rclass
forvalues i=1/3 {
	preserve
	egen y1c`i'= mean(y1cig), by(clust)
	keep if inc==`i'
	sort clust
	egen y0c`i'= mean(y0cig`i'), by(clust)
	egen n0c`i'= count(y0cig`i'), by(clust)
	egen n1c`i'= count(y1cig), by(clust)
	qui sum bscig
	scalar Wbar`i' = r(mean)
	sort clust
	qui by clust: keep if _n==1
	qui ameans n0c`i'
	scalar n0`i'=r(mean_h)
	qui ameans n1c`i'
	scalar n1`i'=r(mean_h)
	drop n0c`i' n1c`i'
	tempname S`i' R`i' num`i' den`i' phi`i' theta`i' psi`i'
	qui corr y0c`i' y1c`i', cov
	scalar S`i'=r(Var_2)
	scalar R`i'=r(cov_12)
	scalar num`i' = scalar(R`i') - (sigma12`i' / n0`i')
	scalar den`i'=scalar(S`i')-(sigma11/n1`i')
	cap scalar phi`i' = num`i' / den`i'
	cap scalar zeta`i' = b1/((b0`i' + Wbar`i' * (1-b1)))
	cap scalar theta`i' =phi`i' /(1+(Wbar`i' - phi`i') * zeta`i')
	cap scalar psi`i' = 1-((b1*(Wbar`i' - theta`i'))/(b0`i' + Wbar`i'))
	return scalar Elast_`i' = (theta`i' / Wbar`i' )- psi`i'
	restore
}
end
elast
bootstrap elast1=r(Elast_1)elast2=r(Elast_2)elast3=r(Elast_3), reps(1000) seed(1):elast

* Prueba de diferencias de elasticidad.
test _b[elast1]=_b[elast2]
test _b[elast2]=_b[elast3]
test _b[elast1]=_b[elast3]
```

7.3 Do-file de Stata para estimar las elasticidades precio y precio cruzadas para múltiples bienes utilizando el método Deaton

```
/*============================================================
Fecha: 30 de junio de 2022
Tema: Do-file Stata reproducido de Deaton y modificado para el conjunto de herramientas para el Uso de Encuestas de Gastos de los Hogares para Investigación en Economía del control del tabaco. Proporciona el código para calcular el sistema de ecuaciones de demanda, incluidas las elasticidades precio y precio cruzadas, para completar el sistema y para calcular estimaciones restringidas por simetría.

Nota: Estos códigos fueron escritos como parte de "El análisis de las encuestas de hogares: Un enfoque microeconométrico de las políticas de desarrollo", por Angus Deaton.
Los códigos originales están disponibles en
http://web.worldbank.org/archive/website00002/WEB/EX5_1-2.HTM

*Base de datos utilizada: hbs_data.dta
* Variables clave necesarias para ejecutar este código:
* - Los logaritmos de los valores unitarios comienzan con luv, como luvcig, luvbiri, etc.
* - Las participaciones en el presupuesto comienzan con bs, como bscig, bsbiri
* - lnexp - logaritmo natural de los gastos totales del hogar
* - lhsize – logaritmo natural del tamaño del hogar
* - Variables adicionales específicas del hogar disponibles para ser agregadas por el usuario
Lo siguiente se agrega aquí solo con fines expositivos
                -Año de educación del jefe de hogar - edu_head
                -Porción de miembros adultos en la familia - adultratio
                -Porción de miembros masculinos en la familia - maleratio
* - clust - variable que identifica la unidad primaria de muestreo o conglomerado
============================================================*/

clear all
versión 15
set mem 1000m
set more off
cd "Directory Path"
capture log close
log using "Elasticity.log", replace
use hbs_data.dta

generate cluster=psu
*Estos son los identificadores de productos básicos
gl goods "cig  biri slt"

rename psu psuid
label var hhold "Unique ID for the HH"
```

```
label var psuid "Unique ID for the PSU"


*generar el logaritmo de el tamaño del hogar, gastos
gen lhsize=ln(hsize)
gen lnexp=ln(dce)
gen lnyexp=ln(y_exp)


foreach X in $goods{
  gen luv`X'=ln(uv`X')
}

*Para encontrar la cantidad de UPM sin compras de los artículos a continuación
gen anytobac=0
foreach X in $goods{
  recode anytobac 0=1 if uv`X'!=.
  egen psu_`X'=mean(uv`X'), by(psuid)
  egen tag=tag(psuid) if psu_`X'==.
  egen `X'_none=total(tag)
  drop tag
}

foreach X of global goods{
drop psu_`X' `X'_none
}
*Eliminación de conglomerados con menos de 2 hogares que reportan algún tipo de
*consumo de tabaco
bys psuid: egen consumingHH_psu=sum(anytobac)
drop if consumingHH_psu<2
drop anytobac consumingHH_psu

save "EditedData.dta", replace
* Definición de un programa data_matrix para usar en bootstrap
cap program drop data_matrix
program define data_matrix
        *Número de bienes en el sistema. Se tomará automáticamente.
        gl ngds : word count $goods

        matrix define sig=J($ngds,1,0) // var-covar matrix of u0 (e0e0)
        matrix define ome=J($ngds,1,0) // var-covar matrix of u1 (e1e1)
        matrix define lam=J($ngds,1,0) // covar matrix of u1 (e1e0)
        matrix define wbar=J($ngds,1,0) // average budget shares
        matrix define b1=J($ngds,1,0) // elasticity of quality w.r.t exp
        matrix define b0=J($ngds,1,0) // Coefficients of lnexp in BS regression
```

```
*Participación promedio en el presupuesto
local ig=1
foreach X in $goods{
                    qui summ bs`X'
                    matrix wbar[`ig',1]=r(mean)
                    local ig=`ig'+1
}

/*============================================================
Regresión de la primera etapa: Dentro del conglomerado
============================================================*/
/* creación de una macro global para las variables que estamos controlando. por
ejemplo log de gasto, religión, educación, etc. Tendremos que ingresar esas variables
solo una vez aquí.
*/
gl controls "lnexp lhsize lnyexp edu_head adultratio maleratio"
local ig=1
foreach X in $goods{
        *Regresión de efectos fijos de conglomerados
        *areg, en lugar de reg, se usa para la regresión lineal con un gran conjunto de
        *variables dummy
        areg luv`X' $controls , absorb(cluster)

        * Varianza del error de medición
        *Suma de los cuadrados de los errores/grado de libertad total del error
        *calculando la matriz var-covar de u1 (e1e1)
        matrix ome[`ig',1]=$S_E_sse/$S_E_tdf
        *calculando la elasticidad gasto de la calidad
        matrix b1[`ig',1]=_coef[lnexp]

        *Esos residuos todavía tienen efectos de conglomerados
        predict ruv`X', resid

        *Limpiando y's para el próximo paso
        predict dresidual, dr
        gen y1`X'=_b[_cons]+dresidual
        drop dresidual luv`X'

        **Repetir para participación presupuestaría
        areg bs`X' $controls , absorb(cluster)
        *calculando lo residuos de la regresión de participación presupuestaría
        predict rbs`X', resid

        *calculando la matriz var-covar de u0 (e0e0)
        matrix sig[`ig',1]=$S_E_sse/$S_E_tdf
        *Cálculo de coeficientes de lnexp en regresión BS
        matrix b0[`ig',1]=_coef[lnexp]
```

```
		predict dresidual, dr
		gen y0`X'=_b[_cons]+dresidual

		*Esta próxima regresión es necesaria para obtener la covarianza de los residuos
		qui areg ruv`X' rbs`X' $controls , absorb(cluster)
		*Cálculo de la matriz covar de u1 (e1e0)
		matrix lam[`ig',1]=_coef[rbs`X']*sig[`ig',1]
		drop bs`X' rbs`X' ruv`X' dresidual
		local ig=`ig'+1
	}

	Matrix list sig		// matriz var-covar de u0 (e0e0)
	matrix list ome		// matriz var-covar de u1 (e1e1)
	matrix list lam		// matriz covar de u1 (e1e0)
	matrix list b0		// Coeficeintes de lnexp en regresión BS
	matrix list b1		// elasticidad de calidad w.r.t exp
	matrix list wbar // Participación presupuestaria promedio
	*drop lnexp lhsize adultratio meanedu maxedu res* hhtyp*


	*esto completa la regresión de la primera etapa y la estimación de todos los
	*parámetros necesarios a partir de ella
	* Guardar hasta ahora como precaución
	save "tempa.dta", replace
	drop _all
	use "tempa.dta"

	/*=============================================================
	Regresión de segunda etapa: Entre conglomerados
	=============================================================*/
	*Promediando por conglomerado
	*Contando número de observaciones por cada conglomerado
	local ig=1
	foreach X in $goods{
		egen y0c`ig'=mean(y0`X'), by(cluster)
		egen n0c`ig'=count(y0`X'), by(cluster)
		egen y1c`ig'=mean(y1`X'), by(cluster)
		egen n1c`ig'=count(y1`X'), by(cluster)
		drop y0`X' y1`X'
		local ig=`ig'+1
	}

	sort clust
	*se mantiene una observación por conglomerado
	*NB subround and region son constante dentro del conglomerado
	qui by clust: keep if _n==1
	*Guardando información a nivel conglomerado
end
data_matrix
```

```
/*Eliminando los efectos de región o provincia
* Esto es opcional y puede usarse o no dependiendo de los datos
* Eso supone la disponibilidad de la región variable categórica en los datos
tab region, gen(regiond)

foreach ig of numlist 1/$ngds{
	regress y0c`ig' regiond2 regiond3 regiond4
	predict tm, resid
	replace y0c`ig'=tm
	drop tm
	qui regress y1c`ig' regiond2 regiond3 regiond4
	predict tm, resid
	replace y1c`ig'=tm
	drop tm
}
drop regiond*
*/
cap program drop var_covar
program define var_covar
	matrix define n0=J($ngds,1,0)
	matrix define n1=J($ngds,1,0)

	*Promedio (armónico) del número de observaciones en los conglomerados

	foreach ig of numlist 1/$ngds{
		*replace n0c`ig'=1/n0c`ig'
		*replace n1c`ig'=1/n1c`ig'
		qui ameans n0c`ig'
		matrix n0[`ig',1]=(r(mean_h))
		qui ameans n1c`ig'
		matrix n1[`ig',1]=(r(mean_h))
		drop n0c`ig' n1c`ig'
	}

	*Elaboración de las matrices de varianza y covarianza entre conglomerados
	*Esto se hace en parejas debido a los valores faltantes
	matrix s=J($ngds,$ngds,0) // between-cluster var-covar matrix of y1
	matrix r=J($ngds,$ngds,0) // between-cluster covar matrix of y1
	local ir=1
	foreach ir of numlist 1/$ngds{
		local ic=1
		foreach ic of numlist 1/$ngds{
			qui corr y1c`ir' y1c`ic', cov
			matrix s[`ir',`ic']=r(cov_12)
			qui corr y1c`ir' y0c`ic', cov
			matrix r[`ir',`ic']=r(cov_12)
			}
```

```
        }

        *Ya no necesitamos los datos
        drop _all
        matrix list s // between-cluster var-covar matrix of y1
        matrix list r // between-cluster covar matrix of y1
        *Estimando MCO
        matrix bols=syminv(s)
        matrix bols=bols*r
        display("Second-stage OLS estimates: B-matrix") // eqn 5.84
        matrix list bols
        display("Column 1 is coefficients from 1st regression, etc")

        *Correcciones por errores de medición
        matrix def sf=s
        matrix def rf=r
        foreach ig of numlist 1/$ngds{
                matrix sf[`ig',`ig']=sf[`ig',`ig']-ome[`ig',1]/n1[`ig',1]
                matrix rf[`ig',`ig']=rf[`ig',`ig']-lam[`ig',1]/n0[`ig',1]
        }

        matrix invs=syminv(sf)
        matrix bhat=invs*rf // Estimador de errores en la variable con corrección de errores de
        *medición

        *Matriz B estimada sin restricciones
        matrix list bhat // Estimador de errores en la variable con corrección de errores de medición
        *El ratio Phi del cual las matrices Psi y Theta se deben separar
        *Matrices de hogar, incluyendo elasticidades
        matrix def xi=J($ngds,1,0)
        matrix def el=J($ngds,1,0)
        foreach ig of numlist 1/$ngds{
                matrix xi[`ig',1]=b1[`ig',1]/(b0[`ig',1]+((1-b1[`ig',1])*wbar[`ig',1]))
                matrix el[`ig',1]=1-b1[`ig',1]+b0[`ig',1]/wbar[`ig',1]
        }

        global ng1=$ngds+1
        matrix iden=I($ngds)
        matrix iden1=I($ng1)
        matrix itm=J($ngds,1,1)
        matrix itm1=J($ng1,1,1)
        matrix dxi=diag(xi)
        matrix dwbar=diag(wbar)
        matrix idwbar=syminv(dwbar)
end
var_covar
display("Average budget shares")
matrix tm=wbar'
```

```
matrix list tm // Participación en el gasto promedio
display("Expenditure elasticities")
matrix tm=el' // Elasticidades gasto (dlnq/dlnx)
matrix list tm
display("Quality elasticities")
matrix tm=b1'
matrix list tm // Elasticidad gasto de la calidad (dlnuv/dlnx)

*Todo esto tiene que ir en un programa para volver a usarlo más tarde
*Básicamente usa la matriz b de la ecuación 5.85 para formar la matriz de elasticidad precio
cap program drop mkels
program define mkels
        matrix cmx=bhat'
        matrix cmx=dxi*cmx
        matrix cmx1=dxi*dwbar
        matrix cmx=iden-cmx
        matrix cmx=cmx+cmx1
        matrix psi=inv(cmx)
        matrix theta=bhat'*psi
        display("Theta matrix")
        matrix list theta
        matrix ep=bhat'
        matrix ep=idwbar*ep
        matrix ep=ep-iden
        matrix ep=ep*psi
        display("Matrix of price elasticities")
        matrix list ep // price elasticity of demand without symmetry restrictions)
end
mkels

/*================================================================
**Completar el sistema llenando las matrices
* Esto esencialmente agrega un solo producto compuesto a la ecuación para completar
* el sistema utilizando homogeneidad y restricciones de suma.
================================================================*/
cap program drop complet
program define complet
        *First extending theta
        matrix atm=theta*itm
        matrix atm=-1*atm
        matrix atm=atm-b0
        matrix xtheta=theta,atm
        matrix atm=xtheta'
        matrix atm=atm*itm
        matrix atm=atm'
        matrix xtheta=xtheta\atm
        *Extending the diagonal matrices
        matrix wlast=wbar'*itm
```

```
        matrix won=(1)
        matrix wlast=won-wlast
        matrix xwbar=wbar\wlast
        matrix dxwbar=diag(xwbar)
        matrix idxwbar=syminv(dxwbar)
        matrix b1last=(0.25)
        matrix xb1=b1\b1last
        matrix b0last=b0'*itm
        matrix b0last=-1*b0last
        matrix xb0=b0\b0last
        matrix xe=itm1-xb1
        matrix tm=idxwbar*xb0
        matrix xe=xe+tm
        matrix tm=xe'
        matrix exp_elas=xe'
        display("extended outlay elasticities (or total expenditure elasticities)")
        matrix list tm  // expenditure elasticities from the complete system
        matrix xxi=itm1-xb1
        matrix xxi=dxwbar*xxi
        matrix xxi=xxi+xb0
        matrix tm=diag(xb1)
        matrix tm=syminv(tm)
        matrix xxi=tm*xxi
        matrix dxxi=diag(xxi)
        *Extending psi
        matrix xpsi=dxxi*xtheta
        matrix xpsi=xpsi+iden1
        matrix atm=dxxi*dxwbar
        matrix atm=atm+iden1
        matrix atm=syminv(atm)
        matrix xpsi=atm*xpsi
        matrix ixpsi=inv(xpsi)
        *Extending bhat & elasticity matrix
        matrix xbhatp=xtheta*ixpsi
        matrix xep=idxwbar*xbhatp
        matrix xep=xep-iden1
        matrix xep=xep*xpsi
        display("extended matrix of elasticities")
        matrix list xep   // price elasticities from the complete system without symmetry
end
complet // este comando se puede eliminar si solo estamos interesados en
*estimaciones con restricciones de simetría como se indica a continuación
* estimadores en los que estamos interesados, no es necesario ejecutar el resto del código también
************************************************************
**Cálculo de estimadores con restricciones de simetría
**Estos son solo aproximadamente válidos y no asumen efectos de calidad
*Calcula dos matrices, la matriz de conmutación y la diagonal inferior
*matriz de selección que se necesita en los cálculos principales;
```

```
cap program drop commx
program define commx
	mata: st_matrix("`2'", kmatrix(`1',`1'))
end

** para vectorizar una matriz, es decir, apilarla en un vector columna
cap program drop vecmx
program def vecmx
	mata: st_matrix("`2'", vec(st_matrix("`1'")))
end

*programa para calcular la matriz que extrae
*de vec(A) el triangúlo inferior de la matriz A
cap program drop lmx
program define lmx
	local ng2=`1'^2
	local nr=0.5*`1'*(`1'-1)
	matrix def `2'=J(`nr',`ng2',0)
	local ia=2
	local ij=1
	while `ij' <= `nr'{
		local ik=0
		local klim=`1'-`ia'
		while `ik' <= `klim' {
			local ip=`ia'+(`ia'-2)*`1'+`ik'
			matrix `2'[`ij',`ip']=1
			local ij=`ij'+1
			local ik=`ik'+1
			}
			local ia=`ia'+1
	}
end

**programa para reacomodar el vector en una matrix cuadrada
cap program drop unvecmx
program def unvecmx
	mata: st_matrix("`2'", colshape(st_matrix("`1'"),$ngds)')
end

vecmx bhat vbhat
** Matrix R para restricciones
lmx $ngds llx
commx $ngds k
cap program drop matrices
program def matrices
	global ng2=$ngds*$ngds
	matrix bigi=I($ng2)
```

```
        matrix k=bigi-k
        matrix r=llx*k
        matrix drop k
        matrix drop bigi
        matrix drop llx
        ** r vector for restrictions, called rh
        matrix rh=b0#wbar
        matrix rh=r*rh
        matrix rh=-1*rh
        **doing the constrained estimation
        matrix iss=iden#invs
        matrix rp=r'
        matrix iss=iss*rp
        matrix inn=r*iss
        matrix inn=syminv(inn)
        matrix inn=iss*inn
        matrix dis=r*vbhat
        matrix dis=rh-dis
        matrix dis=inn*dis
        matrix vbtild=vbhat+dis
        unvecmx vbtild btild
        **the following matrix should be symmetric
        matrix atm=b0'
        matrix atm=wbar*atm  // Eqn. 5.98
        matrix atm=btild+atm
        matrix list atm
        **going back to get elasticities and complete sytem
        matrix bhat=btild
end
matrices
mkels
complet

/*Para estimar los errores estándar bootstrap de las elasticidades precio y gasto */

drop _all
vecmx xep vxep
vecmx exp_elas vexp_elas
matrix obs=vxep\vexp_elas
matrix observe=obs'
global nels=$ng1*$ng1
drop _all

use "EditedData.dta", replace
        capture program drop bootstrap
        program define booststrap, rclass
                preserve
                bsample _N
```

```
                data_matrix
                var_covar
                mkels
                vecmx bhat vbhat
                lmx $ngds llx
                commx $ngds k
                matrices
                mkels
                complet
                vecmx xep vxep
                vecmx exp_elas vexp_elas
                foreach ic of numlist 1/$nels{
                        return scalar e`ic'=vxep[`ic',1]
                }
                foreach iex of numlist 1/$ng1{
                        return scalar exp`iex'=vexp_elas[`iex',1]
                }
                restore
        end

local x ""
local x1 ""
foreach X of numlist 1/$nels {
        local y "e`X'=r(e`X')"
        local z "`x' `y'"
        local x "`z'"
        }
foreach X of numlist 1/$ng1{
        local y1 "exp`X'=r(exp`X')"
        local z1 "`x1' `y1'"
        local x1 "`z1'"
        }
simulate `x' `x1', reps(1000) seed(122002): booststrap

bstat, stat(observe)
log close
```

7.4 Do-file de Stata para estimar el efecto de desplazamiento del gasto en tabaco

```
*============================================================
* Fecha: Noviembre de 2018
* Tema: Do-file de Stata creado como parte del conjunto de herramientas para el Uso de
* Encuestas de Gastos de los Hogares para Investigación en Economía del Control del Tabaco
* Este do-file estima el impacto de desplazamiento del gasto en tabaco
* Base de datos utilizada: DataQAIDS.dta
* Variables clave:
* - exptotal - gasto total del hogar en unidades monetarias locales (UML)
* - exptobac - gasto total en tabaco del hogar en UML
* - exphealth - gastos totales de atención médica del hogar en UML
* - expfood - gasto total en alimentos del hogar en UML
* - expeducn - gasto total en educación del hogar en UML
* - exphousing - gasto total de vivienda del hogar en UML
* - expcloths - gasto total en ropa para el hogar en UML
* - expentertmnt - gasto total de entretenimiento del hogar en UML
* - exptransport - gasto total de transporte del hogar en UML
* - expdurable - gasto total en bienes duraderos de los hogares en UML
* - exother - gasto total en otros artículos del hogar en UML
* - hsize - tamaño del hogar
* - meanedu - educación media del hogar en años
* - maxedu - educación máxima del hogar en años
* - sgroup - variable categórica que representa los grupos sociales del hogar
* - asexratio - proporción de sexos de adultos (proporción de hombres adultos a mujeres adultas)
* - weight - ponderadores de la encuesta
*============================================================================
clear
versión 15
set mem 1000m
set more off

* cambie las rutas de directorio a continuación para informar a Stata dónde están los datos
*almacenados y donde se almacenará la salida
global pathin  "C:\Data\"
global pathout "C:\Data\QAIDS"

capture log close
log using $pathout\Crowdout.log, replace
use $pathin\DataQAIDS.dta
use "$pathin/DataQAIDS.dta", clear

*********************************************************************
*Prueba T para comparar participación promedio en el presupuesto
*********************************************************************
*Generar una variable binaria para el gasto en tabaco
gen tob= exptobac >0 & exptobac <.
label define tob 1 "Tobacco spenders" 0 "Tobacco non-spenders", replace
```

```
*generación de variables de participación en el presupuesto para la prueba t de comparación
*aquí el denominador es el gasto total en todos los bienes combinados
local items "tobac food health educn housing cloths entertmnt transport durable other"
foreach X of local items{
	gen bs_`X'=(exp`X'/exptotal)
}
*prueba t usando ponderaciones de encuesta
local items tobac food health educn housing cloths entertmnt transport durable other
local nvar: word count `items'
matrix B = J(`nvar', 4, .)
forvalues i = 1/`nvar' {
   local X: word `i' of `items'
   qui mean bs_`X' [pw=weight], over(tob)
	matrix tmp=r(table)
	matrix B[`i', 1] = tmp[1,1]
	matrix B[`i', 2] = tmp[1,2]
	qui lincom _b[c.bs_`X'@0.tob] - _b[c.bs_`X'@1.tob]	matrix B[`i', 3] = r(estimate)
   matrix B[`i', 4] = r(t)
}
matrix rownames B =`items'
matrix colnames B = non-spenders spenders Difference t-stat
matrix list B
*eliminando las variables de participación en el presupuesto
drop bs_*

**********************************************************************
*Preparación de variables para la estimación del desplazamiento
**********************************************************************
*generar variables dummy de grupos sociales
tab sgroup, gen(sd)

*crear participación en el presupuesto para el análisis de desplazamiento. Aquí el denominador es
*el gasto total menos los gastos en tabaco
gen exp_less=exptotal-exptobac
local items "food health educn housing cloths entertmnt transport durable other"
foreach X of local items{
	gen bs`X'=(exp`X'/exp_less)
}

gen lnM=log(exp_less)
gen lnX=log(exptotal)
gen lnM2=lnM*lnM
gen lnX2=lnX*lnX
gen pq=exptobac
*Estimación de desplazamiento con diferentes modelos
global ylist bsfood bshealth bseducn bshousing bscloths bsentertmnt bstransport bsdurable
global x1list pq lnM lnM2
```

```
global x2list hsize meanedu maxedu sd1-sd3
global zlist asexratio lnX lnX2

*********************************************
*Estimación 3SLS tradicional
*********************************************
**3SLS using reg3
reg3 ($ylist = $x1list $x2list), exog($zlist) endog($x1list) 3sls

*3SLS tradicional usando GMM
gmm (eq1: bsfood - {food: $x1list $x2list _cons}) ///
        (eq2: bshealth - {health: $x1list $x2list _cons}) ///
        (eq3: bseducn - {educn: $x1list $x2list _cons}) ///
        (eq4: bshousing - {housing: $x1list $x2list _cons}) ///
        (eq5: bscloths - {cloths: $x1list $x2list _cons}) ///
        (eq6: bsentertmnt - {entertmnt: $x1list $x2list _cons}) ///
        (eq7: bstransport  - {transport: $x1list $x2list _cons}) ///
        (eq8: bsdurable - {durable: $x1list $x2list _cons}) ///
        , instruments($zlist $x2list) ///
        winitial(unadjusted, independent)  wmatrix(unadjusted) twostep

*Las dos implementaciones anteriores (reg3 y gmm) deberían dar resultados idénticos
*y son estimaciones 3SLS tradicionales. Pero, converger gmm puede llevar mucho más tiempo
*que reg3. Esté preparado para esperar algunas horas dependiendo de la máquina.
*Una alternativa es guardar primero los resultados de reg3 usando el comando
*<matrix b = e(b)> y utilícelos como el valor inicial de gmm para que
*la convergencia pueda ser más rápida. Eso se hace agregando la opción
*<center twostep from(b)> a la última línea en gmm en lugar de usar solo <twostep>

*********************************************
*Estimación GMM 3SLS (wooldridge): ajustada por heterocedasticidad
*********************************************
gmm (eq1: bsfood - {food: $x1list $x2list _cons}) ///
        (eq2: bshealth - {health: $x1list $x2list _cons}) ///
        (eq3: bseducn - {educn: $x1list $x2list _cons}) ///
        (eq4: bshousing - {housing: $x1list $x2list _cons}) ///
        (eq5: bscloths - {cloths: $x1list $x2list _cons}) ///
        (eq6: bsentertmnt - {entertmnt: $x1list $x2list _cons}) ///
        (eq7: bstransport  - {transport: $x1list $x2list _cons}) ///
        (eq8: bsdurable - {durable: $x1list $x2list _cons}) ///
        , instruments($zlist $x2list) ///
        winitial(unadjusted, independent)  wmatrix(robust) twostep

*También se podría usar la opción <wmatrix(cluster clustvar)> donde clustvar es
*el nombre de la variable que identifica los conglomerados
```

```
*********************************************
* VI ecuación por ecuación o 2SLS usando ivregress:
*********************************************
*Usando el comando de regresión VI integrado de Stata
local depvar "food health educn housing cloths entertmnt transport durable"
foreach X of local depvar{
        ivregress 2sls bs`X' $x2list ($x1list = $zlist)
}

*Usando el programa escrito por usuarios <ivreg2>
*Fuente: Baum CF, Schaffer ME, Stillman S. IVREG2: Módulo Stata para
*Variables Instrumentales Extendidas/Estimación 2SLS y GMM. Boston College
*Departamento de Economía; 2007.
*https://ideas.repec.org/c/boc/bocode/s425401.html. Consultado el 30 de octubre de 2018

local depvar "food health educn housing cloths entertmnt transport durable"
foreach X of local depvar{
        ivreg2 bs`X' $x2list ($x1list = $zlist)
}

*los dos conjuntos de comandos anteriores deberían arrojar resultados idénticos.
*Pero ivreg2, por defecto, también muestra algunas estadísticas de prueba de interés

*Usando el estimador System 2SLS (VI ecuación por ecuación)
gmm (eq1: bsfood - {food: $x1list $x2list _cons}) ///
        (eq2: bshealth - {health: $x1list $x2list _cons}) ///
        (eq3: bseducn - {educn: $x1list $x2list _cons}) ///
        (eq4: bshousing - {housing: $x1list $x2list _cons}) ///
        (eq5: bscloths - {cloths: $x1list $x2list _cons}) ///
        (eq6: bsentertmnt - {entertmnt: $x1list $x2list _cons}) ///
        (eq7: bstransport  - {transport: $x1list $x2list _cons}) ///
        (eq8: bsdurable - {durable: $x1list $x2list _cons}) ///
        , instruments($zlist $x2list) ///
        winitial(unadjusted, independent)

*Eso proporciona estimaciones de parámetros similares al ivregress anterior, pero con
*Errores estándar robustos. Para tener los mismos errores estándar
*como en ivregress en su lugar agregue la opción <vce(unadjusted) onestep>
*después de winitial(unadjusted, independent)

*si hay heterocedasticidad presente, uno puede realizar el sistema 2SLS
*Usando gmm como se indicó anteriormente, que devuelve errores estándar robustos, o modifica el
*ivregress con la opción vce(robust) o usar el estimador gmm en ivregress
*comando para especificar opciones adicionales como <wmatrix(robust)> o
*<wmatrix(cluster clustvar)>. Eso se hace a continuación.
```

```
local depvar "food health educn housing cloths entertmnt transport durable"
foreach X of local depvar{
	ivregress gmm bs`X' $x2list ($x1list = $zlist), wmatrix(cluster clustvar)
}

*Donde clusvar es el nombre de la variable de conglomerado en los datos
*Eso devolvería errores estándar consistentes con la heterocedasticidad que también
*considera la correlación arbitraria entre las observaciones dentro de los conglomerados

*******************************************************************************
* Realización de diferentes pruebas para decidir el método de estimación
*******************************************************************************
*Todas las pruebas se muestran para VI ecuación por ecuación y para una sola ecuación
* es decir, para bsfood. Uno puede simplemente construir un loop para hacer eso de
*una vez para todas las ecuaciones

*****************************************
*(1) Prueba de endogeneidad de los regresores:
*****************************************
* dependiendo de si se usa o no la opción vce(robust), la salida de
*los resultados de las pruebas serán diferentes. En cualquier caso, una estadística significativa
*implica rechazar la hipótesis nula Ho: las variables son exógenas.

ivregress 2sls bsfood $x2list ($x1list = $zlist)
estat endogenous

ivregress 2sls bsfood $x2list ($x1list = $zlist), vce(robust)
estat endogenous

*Esas pruebas también se pueden realizar en un loop para todos los productos juntos de la
*siguiente manera:
local depvar "food health educn housing cloths entertmnt transport durable"
foreach X of local depvar{
	ivregress 2sls bs`X' $x2list ($x1list = $zlist)
	estat endogenous
	ivregress 2sls bs`X' $x2list ($x1list = $zlist), vce(robust)
	estat endogenous
}
*con ivreg2, sin embargo, haga las pruebas en conjunto con la regresión misma
*con la opción endogtest() de la siguiente manera
ivreg2 bsfood $x2list ($x1list = $zlist), endogtest($x1list)
```

```
*********************************************
*(2) Comprobación de la validez de los instrumentos
*********************************************
**Restricción de inclusión de pruebas. Comprueba si los instrumentos son fuertes o débiles

ivregress 2sls bsfood $x2list ($x1list = $zlist)
estat firststage, all
*Eso mostrará tantos resultados de regresión de primera etapa como el número de
*variables endógenas. Como tenemos tres aquí, informará resultados tres de primera etapa.
*Regla empírica: sugiere un estadístico F de menos de 10, en caso de
*un solo regresor endógeno, para ser indicativo de un instrumento débil
*Dado que tenemos tres aquí, se puede usar un estadístico llamada R2 parcial de Shea
*en lugar del valor F-crítico. Esos también se reportan después del comando.
*Tenga en cuenta que no hay consenso sobre qué tan bajo de un valor de R2 indica un
*problema. Ver Cameron & Trivedi25 (Capítulo 6.4.2) para una exposición detallada de
*esas estadísticas

*con ivreg2, sin embargo, haga las pruebas en conjunto con la regresión misma
*con la opción endogtest() de la siguiente manera:

ivreg2 bsfood $x2list ($x1list = $zlist), first

**Prueba de restricción de exclusión. (exogeneidad del instrumento)

*No es posible probar la restricción de exclusión cuando el modelo es solo
*identificado como lo tenemos en las especificaciones anteriores. Si hay más instrumentos
*que el número de variables endógenas, podemos realizar una prueba de
*restricciones de sobreidentificación. Esto se hace como

ivregress 2sls bsfood $x2list ($x1list = $zlist)
estat overid
*En el caso identificado justo, simplemente devolverá un error
*"no overidentifying restrictions".
* Con fines de demostración, supongamos que especificamos lo siguiente:
* devuelve los resultados de la estadística Sargan. Pero, recuerde, esto es sólo
* una especificación arbitraria en la que mantenemos el número de instrumentos más alto
* Los resultados no deben tomarse de todos modos.
ivregress 2sls bsfood $x2list (pq lnM = $zlist)
estat overid
*si se utilizan los errores estándar consistentes con la heterocedasticidad , estat overid
* devolverá un Score chi2 o el estadístico chi2 J de Hansen. Un estadístico
* de prueba significativo indica que los instrumentos pueden no ser válidos.
ivregress 2sls bsfood $x2list (pq lnM = $zlist), vce(robust)
estat overid
```

```
*******************************************
*(3) Prueba de heterocedasticidad
*******************************************
*La prueba se realiza más fácilmente con ivreg2 de la siguiente manera:
ivreg2 bsfood $x2list ($x1list = $zlist)
ivhettest
*Reporta el estadístico Pagan-Hall con la Ho: La perturbación es homocedástica.

*******************************************
*(4) Prueba de heterogeneidad en las preferencias entre consumidores y no consumidores de
tabaco
*******************************************
*Probar esto necesitaría una especificación alternativa del modelo de la
*Ecuación 5 en el capítulo 4. La adición de variables dummy se puede agregar a
*el modelo usando las notaciones factoriales.

local depvar "food health educn housing cloths entertmnt transport durable"
foreach X of local depvar{
        ivregress 2sls bs`X' $x2list tob tob#c.lnM tob#c.lnM2 ($x1list = $zlist)
        test (tob=0) (1.tob#c.lnM=0) (1.tob#c.lnM2=0)
}

*Rechazar (es decir, un estadístico de prueba significativo) sugiere que la Ecuación 5 puede
*ser una especificación más apropiada mientras que no rechazar implica la Ecuación 4
*puede usarse como la especificación correcta. Si la prueba concluye que la Ecuación 5
* es la especificación de elección, todas las pruebas de 1 a 3 deben ser
* realizadas de nuevo en la nueva especificación. Y si la heterocedasticidad está presente
* Se debe utilizar un método de estimación GMM 3SLS para obtener los parámetros finales.

*********************************************************************
*Análisis por diferentes subgrupos
*********************************************************************
*generar variable indicadora para diferentes grupos de ingresos
*Primero generar gasto per cápita y luego generar la variable
gen pcexp=exptotal/hsize
_pctile pcexp, p(30, 70)
local lower = `r(r1)'
local upper = `r(r2)'
gen incgrp=0
replace incgrp=1 if pcexp<=`lower'
replace incgrp=2 if pcexp>`lower' & pcexp<`upper'
replace incgrp=3 if pcexp>=`upper'
label define incgrp 1 "Low income" 2 "Middle income" 3 "High income"
label values incgrp incgrp
```

```
*VI ecuación por ecuación
local depvar "food health educn housing cloths entertmnt transport durable"
foreach X of local depvar{
        bysort incgrp: ivregress 2sls bs`X' $x2list ($x1list = $zlist)
}
*También para la estimación GMM 3SLS, se puede agregar el prefijo <bysort incgrp:> antes del
*comando gmm y obtener resultados por cada grupo de ingreso.
log close
```

7.5 Do-file de Stata para estimar el efecto empobrecedor del consumo de tabaco

```
*============================================================
* Fecha: Noviembre de 2018
* Tema: Do-file de Stata creado como parte del conjunto de herramientas para el Uso de
* Encuestas de Gastos de los Hogares para Investigación en Economía del Control del Tabaco
* Este do-file estima el impacto empobrecedor del consumo de tabaco
* Base de datos utilizada: DataHH.dta
* Variables clave:
* - exptotal - gasto total del hogar en unidades monetarias locales (UML)
* - exptobac - gasto total en tabaco del hogar en UML
* - exphealth - gastos totales de atención médica del hogar en UML
* - hsize - tamaño del hogar
* - hweight – ponderadores de la encuesta
* - npl – Línea de pobreza nacional en unidades monetarias locales
*============================================================

clear
version 15
set mem 1000m
set more off

* cambie las rutas de directorio a continuación para informar a Stata dónde están los datos
*almacenados y donde se almacenará la salida
global pathin  “C:\Data\”
global pathout “C:Data\poverty”

capture log close
log using $pathout\poverty.log, replace
use $pathin\DataHH.dta

*el siguiente loop genera los gastos per cápita y los etiqueta
foreach X in total tobac health{
        gen pce`X'=exp`X'/hsize
        label var pce`X' "percapita expenditure of `X'"
        }
```

```
*SAF es la fracción atribuible al tabaquismo (uso de tabaco) estimada externamente
scalar SAF=0.2
replace pcehealth=pcehealth*SAF
*Si SAF para la exposición al humo de segunda mano está disponible, en su lugar
*multiplique la variable pcehealth con la suma de ambos SAF

* preparación de variables para el análisis
ren pcetotal pce
gen pcet=pce-pcetobac
label var pcet "pce-expenditure on tobacco"
gen pceh=pcet-pcehealth
label var pceh "pct-tobacco attributable health care exp."
gen pweight=hweight*hsize

*generación de una variable indicadora de pobreza
gen povdum = 0
replace povdum = 1 if pce <= npl
proportion povdum [fw = pweight]

*el siguiente módulo escrito por usuarios también da un resultado idéntico para HCR
*junto con otras medidas de pobreza. Para usar esto, primero aplique el siguiente
*comando sin la estrella.
*ssc install povdeco, replace
povdeco pce [fw=pweight], varpline(npl)

*Código para calcular cambios en HCR y número de pobres de una sola vez
local subtr pce pcet pceh
local nvar: word count `subtr'
matrix M = J(`nvar', 2, .)
forvalues i = 1/`nvar' {
   local X: word `i' of `subtr'
   qui gen ind = (`X'<=npl)
   qui sum ind [fw=pweight]
   matrix M[`i', 1] = r(mean)
   matrix M[`i', 2] = r(sum)
   drop ind
}
matrix rownames M = `subtr'
matrix colnames M = HCR Poor
*a continuación se enumeran los resultados con opciones de formato especiales
matlist M, cspec(& %12s | %5.4f & %9.0f &) rspec(--&&-)

log close
```

INSTITUTE FOR HEALTH RESEARCH AND POLICY

UIC

tobacconomics

Economic Research Informing Tobacco Control Policy

www.tobacconomics.org
@tobacconomics