tobacconomics

Economic Research Informing Tobacco Control Policy

Conjunto de herramientas para

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

INSTITUTE FOR HEALTH RESEARCH AND POLICY

UIC

Cita sugerida: John R., Chelwa G., Vulovic V., Chaloupka F., *Conjunto de herramientas para: El Uso de Encuestas de Gastos de los Hogares para Investigación en Economía del Control de Tabaco.* Conjunto de herramientas de Tobacconomics. Chicago, IL: Tobacconomics, Health Policy Center, Institute for Health Research and Policy, University of Illinois at Chicago, 2019. www.tobacconomics.org

Autores: Este Conjunto de herramientas fue escrito por Rijo John, PhD, Senior Fellow, Centre for Public Policy Research, Kerala, India; Grieve Chelwa, PhD, Senior Lecturer in Economics, University of Cape Town, South Africa; Violeta Vulovic, PhD, Senior Economist, Health Policy Center, University of Illinois at Chicago, y Frank Chaloupka, PhD, Research Professor, University of Illinois at Chicago. La revisión por pares fue proporcionada por Martin 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 Philantropies.

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 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 está comprometido en lograr que este Conjunto de herramientas sea lo más claro y útil posible. Nos gustaría saber si este Conjunto de herramientas le resultó útil en su investigación y, de ser así, le agradeceríamos nos compartiera su experiencia con cualquier implementación exitosa que haya tenido. También nos gustaría saber si ha encontrado algún problema al aplicar las metodologías que le presentamos en este Conjunto de herramientas, así como sus opiniones sobre cómo podríamos mejorarlo.

Para cualquier comentario o preguntas sobre este Conjunto de herramientas y su contenido, escríbanos un correo electrónico a info@tobacconomics.org. Nos encantaría escucharlo.

Tabla de contenido

Lista de tablas

Introducción

1

El consumo de tabaco es la principal causa evitable de muerte y un factor de riesgo principal de varias enfermedades no transmisibles, lo que provoca más de 7.2 millones de muertes anuales en el mundo.[1] A nivel global, el 12 % de todas las muertes de adultos (de 30 años o mayores) se atribuyen al tabaco (16 % entre los hombres y 7 % entre las mujeres) según la Organización Mundial de la Salud (OMS).[2] Si continúan los patrones actuales de consumo de tabaco, se espera que el tabaco mate a aproximadamente mil millones de personas en todo el mundo en este siglo, principalmente en países de ingresos medianos y bajos (PIMB),[3] donde tanto la prevalencia como la magnitud del consumo de tabaco son relativamente altas.[4] El costo económico total del tabaquismo (considerando el conjunto de los gastos en salud y las pérdidas de productividad) ascendió a $1,4 billones de dólares estadounidenses en 2012, es decir, el 1,8 % del producto interno bruto (PBI) anual de todo el mundo.[5] Los PIMB absorben cada vez más la carga económica y sanitaria del consumo de tabaco a nivel mundial.

El tema del Día Mundial Sin Tabaco de 2017 fue “El tabaco, una amenaza para el desarrollo”. Es evidente que el consumo continuo de tabaco en diversas formas tiene el potencial de obstaculizar el desarrollo y el crecimiento económico, especialmente en los PIMB. La morbilidad y mortalidad resultantes del consumo de tabaco impactan negativamente la productividad, reducen los ingresos disponibles y llevan 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 reforzar la aplicación del Convenio Marco de la OMS para el Control del Tabaco. Regular el consumo de tabaco con políticas de salud pública efectivas es importante no solo para abordar la creciente preocupación de las enfermedades no transmisibles, sino también para mejorar el crecimiento económico y reducir la pobreza. Una gran cantidad de estudios, tanto de países de ingreso alto como de países de ingreso mediano y bajo, llegan a la conclusión de que existen intervenciones de políticas públicas efectivas para reducir la demanda de productos de tabaco y que estas políticas son altamente costo-efectivas.[4]

Los aspectos económicos del control del tabaco se han convertido en una parte integral del discurso sobre el desarrollo y, aun así, hay una 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 dichas investigaciones es relativamente alta. Esto puede deberse a varias razones, entre ellas la escasez de datos confiables y/o la falta de la experiencia necesaria para llevar a cabo estas investigaciones. Si bien la investigación que explora el impacto del control del tabaco en los PIMB está creciendo rápidamente,[4] todavía existe la necesidad de generar más evidencia a nivel local y nacional para apoyar la creación de políticas públicas para el control de tabaco, especialmente en aquellos países.

1.1 Propósito de este Conjunto de herramientas

El propósito principal de este Conjunto de herramientas es guiar a los investigadores interesados en llevar a cabo investigaciones sobre los aspectos económicos del control de tabaco, especialmente en los PIMB donde existen Encuestas de gastos de los hogares (*Household Expenditure Surveys*, HES) sobre el consumo de los diferentes productos de tabaco. A diferencia de los países de ingreso alto, los datos de

series de tiempo más grandes suelen ser difíciles de obtener en varios PIMB y, como consecuencia, resulta difícil examinar el impacto de ciertas intervenciones en políticas públicas. Por ejemplo, si se contara con buenos datos de series de tiempo sobre los precios y el consumo de cigarrillos, se podría haber estimado cómo las políticas fiscales han impactado a los precios y, a su vez, al consumo de cigarrillos. Sin embargo, aún con la ausencia de series de tiempo largas, es posible hacer varios análisis relevantes a la creación de políticas públicas de control del tabaco utilizando los datos transversales de las encuestas de hogares. Varios PIMB realizan encuestas de hogares de forma esporádica acerca de diferentes temas que pueden ofrecer información interesante sobre el comportamiento de los consumidores con respecto al consumo de tabaco.

Este Conjunto de herramientas revisará ciertas herramientas y técnicas económicas que se pueden utilizar para analizar los datos de las HES con el único propósito de ayudar en la investigación sobre los aspectos económicos de control del tabaco. Le mostrará el uso de las HES para calcular algunos de los temas importantes en la economía del control de tabaco, incluyendo la estimación de la elasticidad precio y la elasticidad cruzada, así como la elasticidad del gasto para los productos de tabaco, el impacto del gasto en tabaco en la asignación de recursos dentro del hogar y el consumo de grupos específicos de productos básicos por hogar, el impacto del gasto en tabaco y los gastos asociados en atención médica en los índices de pobreza. Se discutirán brevemente los antecedentes teóricos y la justificación económica de cada uno de estos temas, los métodos de cálculo y el uso del software estadístico Stata®, para implementar estos métodos.

Este Conjunto de herramientas es uno de varios desarrollados por el Banco Mundial, la OMS y Tobacconomics que se centran en proporcionar orientación para realizar un análisis económico de la demanda de tabaco, el impacto del consumo de tabaco en el empleo, en la equidad, en el comercio ilícito y en los costos económicos. Este es también 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, lo que respaldaría el avance de los argumentos económicos a favor de más impuestos y refutaría los argumentos en contra de su aumento.

1.2 Quién debería usar este Conjunto de herramientas

El análisis en este Conjunto de herramientas no presupone que el lector tenga conocimientos sobre los impuestos al tabaco o la economía de los asuntos relacionados al control del tabaco. Sin embargo, se requieren conocimientos previos 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 economía que investiga el control del tabaco. Si bien la discusión de métodos econométricos y las guías paso a paso de Stata beneficiarían directamente a los investigadores que trabajan en esta materia, las discusiones sobre políticas y los fundamentos de diferentes conceptos económicos en el control del tabaco, así como las interpretaciones de los resultados que se ofrecen en este Conjunto de herramientas son también con la intención de beneficiar a los responsables de la formulación de políticas, a los analistas de los organismos gubernamentales, así como a aquellos en las organizaciones de la sociedad civil para ayudarles 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 se ha creado para proporcionar orientación técnica sobre tres temas importantes en el área de la economía del control del tabaco: en primer lugar, estimar la elasticidad precio y la elasticidad cruzada (Capítulo 3); en segundo lugar, estimar la naturaleza de desplazamiento del gasto en tabaco (Capítulo 4); y, en tercer lugar, cuantificar el efecto empobrecedor del consumo de tabaco (Capítulo

5). Todos estos temas se discuten con la intención de realizar análisis con datos de las HES. La discusión en cada capítulo comenzará con una introducción y los principios detrás del tema, junto con la justificación para realizar el análisis. Después, se hará un breve debate técnico sobre los métodos econométricos utilizados. Sin embargo, ese debate se mantiene al mínimo, ya que está disponible en otros lugares desde manuales econométricos hasta otras fuentes publicadas. Se proporcionan referencias a las lecturas necesarias para ayudar a los lectores a adquirir conocimientos adicionales sobre los conceptos teóricos aquí presentados. Una vez que se hayan presentado los métodos, les seguirá una breve discusión sobre la preparación de los datos para el análisis y luego los diferentes pasos necesarios para realizar el análisis en Stata, junto con el código necesario. Hacia el final del capítulo se presentará un caso práctico de algún país relevante al tema junto con la interpretación de los resultados.

El Conjunto de herramientas tratará los métodos de análisis relevantes para todos los productos de tabaco como un conjunto o distinguirá y separará los productos de tabaco con humo y sin humo, o individuales como los cigarrillos, los bidis y otros productos de tabaco para masticar, dependiendo del asunto en particular que se esté abordando. Por ejemplo, al estimar las elasticidades precio y cruzada, puede ser útil presentar el análisis de cada uno de los productos de tabaco para que se pueda estimar no solo la elasticidad precio de los diferentes productos de tabaco, sino también la elasticidad cruzada mostrando la sustitución y los patrones complementarios entre los productos de tabaco como los bidis y los cigarrillos o el 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 sola categoría y examinar el impacto en diferentes grupos socioeconómicos.

El Conjunto de herramientas está organizado de la siguiente manera: en el Capítulo 2 se ofrece una introducción a las HES con un enfoque en las encuestas realizadas en los PIMB. Se discutirá el contenido de las HES en lo que respecta al tabaco. En particular, abarcará varias cuestiones relacionadas con el consumo de tabaco y el gasto en diferentes productos de este que se consultan en las HES. El capítulo también tratará brevemente algunos de los asuntos econométricos que se deben tener en cuenta al trabajar con HES y el código de Stata para extraer datos de dichas encuestas sin procesar, entre otros. El capítulo también presenta algunos consejos útiles para trabajar con el software Stata.

El Capítulo 3 trata los métodos de estimación de la elasticidad precio y cruzada para diferentes productos de tabaco. El principal método que se abarcará será el desarrollado por Deaton[7] junto con una explicación paso a paso de los comandos de Stata para estimar las elasticidades de precios a partir de datos de las HES. Las estimaciones de las elasticidades de precios que utilizan datos locales son a menudo útiles y deseables para su uso en políticas fiscales sobre el tabaco en sus respectivos países.

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 los sistemas de demanda condicional,[8,9] mostrará cómo los gastos en tabaco desplazan sistemáticamente a los gastos en otros productos básicos dentro del hogar. El análisis discutirá las formas de estimar el desplazamiento en diferentes subgrupos socioeconómicos. También se presentará el método analítico, así como el código de Stata para ejecutar el modelo.

El Capítulo 5 abarca el efecto empobrecedor del gasto en tabaco. En este capítulo se discutirá la estimación de la cantidad real que se gasta en la compra de tabaco, así como el aumento de los costos en atención médica atribuibles a su consumo y al SHS (*Second-hand smoking*, Humo de segunda mano). Luego demostrará cómo la contabilidad del gasto en tabaco y los costos de salud asociados impactarán la estimación de la pobreza nacional que se mide con el HCR (*Head Count Ratio*, Índice de recuento de la pobreza). Se presentará la estimación paso a paso junto con el código de Stata pertinente.

En la medida de lo posible, estos capítulos también discutirán los resultados empíricos de otros países donde se han realizado los mismos estudios utilizando las HES.

Los comandos individuales de Stata que se utilizan en los diferentes capítulos se colocan entre corchetes angulares < > y están en cursiva. Sin embargo, el comando en sí se tiene que usar sin esos corchetes. Los nombres de las variables utilizados en diferentes ejemplos también están en cursiva. Los ejemplos específicos que demuestran el uso de ciertos códigos de Stata se colocan en cuadros de texto separados en los diferentes capítulos. Un Apéndice de código también incluye el código de Stata relevante para los respectivos capítulos en archivos .do por separado.

2 Introducción a las encuestas de gastos de los hogares

2.1 Disponibilidad de las encuestas de gastos de los hogares

Las encuestas de hogares se han llevado a cabo en varios países desde hace mucho tiempo. Por ejemplo, la primera encuesta sobre el gasto de los consumidores realizada por la Oficina de Estadísticas Laborales (BLS, por sus siglas en inglés) en los Estados Unidos (EE. UU.), se llevó a cabo en 1888. Aunque es relativamente nueva, la organización que realiza la NSS (*National Sample Survey*, Encuesta Nacional por Muestreo) en la India, comenzó sus encuestas sobre el consumo de los hogares ya a principios de la década de 1950[10] y desde entonces lleva a cabo encuestas regulares y periódicas. Las LSMS (*Living Standard Measurement Surveys*, Encuesta Nacional sobre Calidad de Vida) las inició el Banco Mundial en 1979. Estas encuestas de múltiples temas han recopilado información sobre el gasto de consumo de los hogares de aproximadamente 38 países de todo el mundo,[11] varios de los cuales son países africanos y asiáticos. Hay varios países, tanto de ingreso alto como bajo, que realizan encuestas de gastos de los hogares y muchos de ellos las realizan en intervalos regulares.

La IHSN (*International Household Survey Network*, Red Internacional de Encuestas de Hogares), una red informal de agencias internacionales que procura "mejorar la disponibilidad, accesibilidad y calidad de los datos de encuestas en países en desarrollo, y propiciar el análisis y el uso de estos datos por parte de los responsables nacionales e internacionales de la toma de decisiones en materia de desarrollo, la comunidad investigadora y otras partes interesadas",[12] mantiene un portal para que los investigadores puedan consultar y descargar documentos de censos o encuestas y metadatos de hasta 201 países; actualmente, cuenta con cerca de 7 000 encuestas catalogadas. Aproximadamente 137 de los 201 países de los que se dispone de datos son PIMB. Este catálogo está disponible en http://catalog.ihsn.org/index.php/catalog e incluye información sobre más de 1 000 HES en su base de datos, de las cuales unas 700 son de PIMB. En la ausencia de variables macroeconómicas de series largas, las HES proporcionan datos transversales significativos, en ocasiones para múltiples periodos de tiempo de un mismo país.

Los organismos de estadística que suelen llevar a cabo las HES en la mayoría de los países solo publican informes resumidos que solamente presentan datos agrupados y se difunden gratuitamente al público. Los datos agrupados, aunque son útiles para examinar el panorama general, no proporcionan un tamaño de muestra adecuado para llevar a cabo los principales análisis econométricos que se desearía realizar. Por lo tanto, para realizar análisis econométricos avanzados con los datos de la encuesta, es importante poder tener acceso a los microdatos (registros individuales, de hogares o de unidades) de las encuestas. Los microdatos a menudo no están disponibles gratuitamente para el acceso público. Sin embargo, estos datos usualmente están disponibles de forma directa en los organismos de estadística gubernamentales encargados de realizar las encuestas mediante el pago de una tarifa nominal. Después de pagar la tarifa, según el sitio web del organismo, se puede recibir los datos en formato digital ya sea descargándolos directamente desde el sitio web o por correo regular en un dispositivo de almacenamiento de datos. Algunos organismos permiten la descarga de datos después de registrarse en ellos y dar una breve descripción del proyecto. Por ejemplo, los microdatos de las LSMS[11] de diferentes países se pueden descargar gratuitamente 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 gastos de los hogares

Las encuestas de hogares más sencillas recogen datos sobre una muestra nacional de hogares, seleccionados aleatoriamente de un "marco" o lista nacional de hogares (comúnmente un censo), y le asignan una probabilidad igual a cada hogar del marco. Si bien el tamaño de la muestra varía ampliamente según el propósito de la encuesta, dado el tamaño de la población en el país y la necesidad de generar estimaciones de submuestras, con frecuencia 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 5 millones de hogares.[7] En la práctica, a menudo se implementa un diseño de dos etapas en la selección de hogares en el que, en la primera etapa, la selección se realiza a partir de una lista de "conglomerados" de hogares, generalmente poblaciones en zonas rurales o bloques urbanos en centros urbanos, y en la segunda etapa, los hogares se seleccionan de cada grupo.[7] Los conglomerados se denominan normalmente unidades de primera etapa (FSU, por sus siglas en inglés) o unidades primarias de muestreo (PSU, por sus siglas en inglés), ya que es la primera unidad que se muestrea en el diseño. Si los conglomerados se seleccionan al azar con una 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 posibilidad de ser incluido.

Dependiendo de los objetivos de la encuesta, se puede diseñar una muestra para que los hogares puedan seleccionarse según los atributos relevantes como el área geográfica, la afiliación étnica, el nivel de vida, el género o la raza, de modo que los hogares en un grupo determinado puedan tener una cierta probabilidad de ser seleccionados. Esta estratificación convierte de manera efectiva una muestra de una población en una muestra de muchas poblaciones, garantizando así suficientes observaciones para permitir los cálculos de estos subgrupos.[7] Las ponderaciones de probabilidad para los hogares en cada estrato pueden variar. En la mayoría de los casos, puede haber pocas PSU o conglomerados dentro de cada estrato. Por ejemplo, la NSS de la India se enfoca en la estratificación por zonas rurales y urbanas dentro de un distrito para sus encuestas de gasto de consumo. Si bien la estratificación suele mejorar la precisión de los cálculos de muestreo, la conglomeración de la muestra tiende a reducir 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, proporcionan información sobre los hogares y las personas que los integran. Aunque la definición de hogar que se utiliza en cada encuesta puede variar según la estructura de las disposiciones de vivienda en cada país, en general, los miembros que viven y comen juntos se consideran parte del mismo hogar. Las HES suelen proporcionar datos sobre el consumo, los ingresos o los bienes, y las características demográficas de los hogares, incluyendo la composición del hogar, el tamaño, la edad y el género de los miembros que lo habitan, el nivel de educación, la situación laboral de sus miembros, la etnia y la raza, entre otros.

Para evaluar el consumo, las HES miden los gastos que se hacen y/o la cantidad que se consume en los hogares en diferentes bienes y servicios durante un período de tiempo preestablecido, también conocido como período de referencia o recordatorio. Aunque es raro, algunas HES, por ejemplo, la Encuesta de Gastos del Consumidor (CES, por sus siglas en inglés) realizada por el BLS en los Estados Unidos, también recopila datos de gastos a nivel individual. En el caso de productos para adultos como el tabaco, estos datos serían de gran utilidad. Dependiendo del objetivo de la encuesta y las características de los bienes o servicios en cuestión, el período recordatorio puede variar significativamente para diferentes productos dentro de la misma encuesta y para los mismos productos a través de diferentes encuestas; puede variar desde un día hasta un año. Sin embargo, los artículos típicos de consumo en la mayoría de las HES tienen un período recordatorio de una semana a un mes. La HIES (*Household Income and Expenditure Survey*, Encuesta de ingresos y gasto de los hogares) de 2016 en Liberia, por ejemplo, recopiló el consumo de alimentos con un período recordatorio de 7 días y los demás consumos con periodos recordatorios de 7 y de 30 días.[13]

Como parte de la tarea de recopilar datos sobre los gastos realizados y la cantidad consumida de diferentes productos, varias HES recopilan información sobre el consumo de diferentes productos de tabaco de uso común en los respectivos países. La NSS de la India, por ejemplo, recopila tanto la cantidad de consumo como el gasto realizado en cigarrillos, bidis y diferentes tipos de tabaco sin humo durante los 30 días previos a la entrevista. Esto proporciona una valiosa fuente de información para examinar varios asuntos económicos relacionados con el consumo de tabaco. Sin embargo, este nivel de desagregación puede que no esté disponible en todas las HES. Dependiendo de los recursos disponibles para los organismos encuestadores, a veces solo se informan los gastos en productos agregados a grupos más grandes, como el tabaco en general o los productos intoxicantes, como un solo grupo. Por otro lado, algunas HES solo proporcionan información de gastos y no recopilan información de cantidades en varios artículos de consumo. Como resultado, puede haber desafíos en el análisis econométrico entre diferentes conjuntos de datos.

A menudo es posible clasificar los hogares en una encuesta dentro de diferentes grupos SES (*socioeconomic status*, nivel socioeconómico) usando otras características específicas de los hogares e información regional proporcionada en las encuestas, para que así, el análisis económico se pueda realizar por esos grupos. Dicho análisis se puede hacer en función del nivel educativo de los hogares, de sus ingresos o de su situación patrimonial, del lugar de residencia, como zonas rurales o urbanas, de sus afiliaciones étnicas o en función del nivel de vida de un hogar, entre otros criterios.

2.3 Cuestiones econométricas al trabajar con encuestas de hogares

Debido a las características de diseño de las encuestas de hogares analizados en la sección anterior, existen desafíos específicos para el análisis econométrico. En el Capítulo 2 de *"The analysis of household surveys"* (El análisis de las encuestas de hogares) de Deaton,[7] se ofrece una exposición detallada de estos desafíos. A continuación, se presenta un breve resumen conceptual de los temas más destacados:

a) **Uso de las ponderaciones de encuestas para estadísticas descriptivas:** dependiendo del propósito de cada encuesta, algunos hogares pueden estar sobre o subrepresentados en las encuestas y, como resultado, la media estimada de la muestra u otras estadísticas de esta serían estimadores sesgados de sus homólogos en el resto de la población. Las ponderaciones de la encuesta se utilizan a menudo para volver a preponderar 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 las ponderaciones de esta junto con los datos publicados y pueden utilizarse inmediatamente, tal como están, mientras se generan las estadísticas necesarias. Si las ponderaciones no se dan directamente, la documentación de la encuesta generalmente incluiría instrucciones o fórmulas para calcular esas ponderaciones utilizando las variables relevantes pertinentes incluidas en los datos de la muestra. Es importante aplicar las ponderaciones correctas de la encuesta al generar estadísticas descriptivas a partir de los datos de la muestra. La Sección 2[7] a continuación ofrece ejemplos de cómo aplicar ponderaciones de encuestas en Stata mientras se calculan ciertas estadísticas descriptivas.

b) **Uso de las ponderaciones de la encuesta en regresión:** a diferencia de la estadística descriptiva, no hay un acuerdo en cuanto al uso de ponderaciones de la encuesta en el contexto de regresiones. El argumento econométrico clásico está en contra del uso de ponderaciones en regresión, como lo señala Deaton,[7] cuando la población es homogénea, de manera que los coeficientes de regresión son idénticos en cada estrato, tanto los estimadores ponderados como los no ponderados serán consistentes y los de Mínimos Cuadrados Ordinarios (MCO) serán de

hecho más eficientes a través del teorema de Gauss-Markov.[14] Por otro lado, cuando la población no es homogénea, los estimadores ponderados y no ponderados son inconsistentes de todos modos y la ponderación no agrega ningún valor. Sin embargo, Deaton[7] continúa diciendo que una regresión ponderada proporciona una estimación consistente de la función de regresión de la población siempre que la suposición sobre la forma funcional de la regresión sea correcta, es decir, cuando la función de regresión en sí es el objeto de interés. Si el interés es estimar modelos conductuales donde el comportamiento puede ser diferente para diferentes subgrupos, la ponderación en la regresión no sirve de nada. En conclusión, como lo observan Cameron y Trivedi,[15] se deben usar ponderaciones para la estimación de la media de la población y para la predicción postregresión y el cálculo de efectos marginales. Sin embargo, en la mayoría de los casos, la regresión en sí puede ajustarse sin ponderaciones, como es la norma en microeconometría.

c) **Errores estándar inflados debido a los efectos de diseño del conglomerado:** como la mayoría de las encuestas de hogares utilizan un diseño de dos etapas en el que los conglomerados se eligen primero, seguidos de los hogares dentro de cada uno de esos conglomerados, a menudo se da el caso que los hogares dentro del mismo conglomerado son bastante similares entre sí, ya que viven cerca unos de otros y fueron entrevistados casi al mismo tiempo, y son diferentes de aquellos en otros conglomerados 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 totalmente independientes, las correlaciones positivas entre estas 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 regresiones basadas en encuestas de hogares para tener en cuenta estos efectos de diseño en los conglomerados utilizando las técnicas apropiadas.

d) **Heterocedasticidad de los residuos de MCO:** las distribuciones de los hogares entre diferentes variables de interés, como el ingreso y el consumo de diferentes productos, no suelen estar distribuidas normalmente y, como resultado, es bastante frecuente encontrar alteraciones heteroscedásticas en funciones de regresión estimadas a partir de los datos de las HES. La heterogeneidad entre diferentes conglomerados también podría dar lugar a que las funciones de regresión arrojen términos de error heterocedásticos. Esto dejaría ineficientes los cálculos de MCO e invalidaría las fórmulas habituales para los errores estándar, por lo que sería necesario corregirlos utilizando los métodos de corrección adecuados. En combinación con la presencia de efectos de diseño de conglomerados, es importante usar fórmulas que corrijan los errores estándar en las regresiones basadas en encuestas que consideran la presencia de heterocedasticidad, así como los efectos de conglomerados.

e) **Endogeneidad:** se refiere a situaciones cuando en una regresión, una o más de las variables explicativas se correlaciona con el término de error, lo que resulta en cálculos sesgados e inconsistentes de MCO. La endogeneidad surge principalmente por tres razones:

   (i) simultaneidad, es decir, X causa Y e Y también causa X. En otras palabras, X e Y se determinan conjuntamente;

   (ii) variables explicativas omitidas, es decir, 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 que se encuentra en esas variables omitidas también puede denominarse

heterogeneidad no observada o como la variación no observada entre unidades individuales de esta variable omitida o no observable; y

(iii) errores de medición, es decir, 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 las encuestas, según Deaton,[7] son una realidad innegable.

Aunque a menudo se mencionan como fuentes separadas de endogeneidad en la regresión, en realidad no es necesario que sean realmente distintos entre sí. A menudo, en el análisis de regresión que usa datos de encuestas, uno se encuentra con la mayoría, si no es que todas, de estas diferentes fuentes de endogeneidad. En todas las diferentes fuentes de endogeneidad aquí descritas, 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* (IV, por sus siglas en inglés) (por ejemplo, el método de los mínimos cuadrados en dos etapas)[14] es la técnica estándar en tales circunstancias, siempre que sea posible encontrar IV que estén correlacionadas con las variables explicativas, pero no con los términos de error, de modo que la regresión genere estimaciones consistentes.

2.4 Consejos útiles sobre Stata

Stata, un paquete estadístico ampliamente utilizado, es el software econométrico y de análisis de datos preferido por varias universidades e instituciones alrededor del mundo, lo que facilita el intercambio y la colaboración entre investigadores de múltiples disciplinas e instituciones.[16] A continuación se presentan algunos consejos útiles que hacen que trabajar con Stata sea mucho más fácil.

Creación de un archivo .do: Stata se puede utilizar a través de sus menús desplegables de la interfaz de usuario, a través de comandos directos en una ventana especial para comandos, o con la ayuda de un archivo .do que guarda todos los comandos para usarse como guste. La ejecución de comandos con un archivo .do es el método preferido y recomendado, ya que ofrece varias ventajas sobre los otros métodos. Un archivo de este tipo simplemente registra todos los comandos a ejecutar 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 archivo .do y el trabajo se puede compartir y editar por otros colaboradores. Pero, más que nada, un archivo .do mantiene un registro del trabajo realizado y permite la revisión de los comandos según sea necesario. A diferencia de las ventanas de comando o los menús desplegables, en un archivo .do también se pueden agregar notas y comentarios para otros colaboradores, lo que facilita una colaboración eficiente. Puede encontrar información útil sobre cómo crear un archivo .do en el sitio web de Stata (https://www.stata.com/manuals13/u16.pdf).

Creación de un archivo de registro: Mientras que un archivo .do mantiene un registro de todos los comandos y le permite editarlos según sea necesario, un archivo de registro con la extensión ".log" o ".txt" mantiene un registro de los comandos ejecutados junto con sus resultados durante una sesión dada de Stata. Es útil crear archivos de registro mientras se ejecuta el archivo .do para que los resultados estén disponibles para futuras referencias o para compartirlos con otros colaboradores. Para crear un archivo de registro dentro del archivo .do se utiliza el comando <*log using mylog.log, replace*>. Esto creará un archivo con el nombre *mylog.log* en el directorio del trabajo actual de Stata. El parámetro opcional <*replace*> se encargará de que cada vez que se ejecute el archivo .do, los contenidos del archivo de registro se reemplacen con los nuevos resultados. También se puede usar la opción <*append*> para seguir agregando los resultados de todos los comandos al mismo archivo de registro. Antes de cerrar la sección, que generalmente se hace al final del archivo .do, se debe cerrar el archivo de registro con el comando <*log*

close>. El uso del archivo de registro también se puede suspender temporalmente y reanudar mediante comandos como <*log off*> *y* <*log on*>.

Uso de los recursos de conocimiento: Todos los manuales de uso de Stata están integrados en el software. Puede simplemente ejecutar el comando <*help*> seguido por cualquier comando de Stata para conocer la descripción, la sintaxis y los ejemplos de los comandos que se usan. Por ejemplo, <*help regress*> mostrará la sintaxis, la descripción y los ejemplos necesarios del uso del comando regress. Además, los comandos <*search*> y <*findit*> muestran información muy útil sobre temas de interés dentro de Stata. Por ejemplo, el comando <*search survey*> mostraría una lista de comandos y módulos que Stata utiliza para analizar datos de encuestas. También se cuenta con un excelente foro de soporte que es un recurso invaluable para aprender y familiarizarse más con Stata. (https://www.statalist.org/forums/.)

Configuración de un directorio de trabajo: Mientras se trabaja con los datos de encuestas de hogares, es mejor hacer una copia de los microdatos y moverlos a un directorio específico en la computadora. Todos los archivos subsecuentes del programa de Stata y otros documentos relacionados con el análisis, se pueden almacenar en el mismo directorio dejando los microdatos originales intactos. El comando <*pwd*> muestra el directorio de trabajo actual de Stata independientemente del sistema operativo que use. Puede cambiar este directorio de trabajo con el comando <*cd "Path"*> donde *Path*, dentro de las comillas, es la ruta del directorio donde se guarda el trabajo. Esto podría variar según el sistema operativo que use. Una vez que se ha establecido el directorio de trabajo, los comandos subsecuentes para trabajar con los archivos (archivos de datos, .do, diccionarios, etc.) se pueden utilizar usando solo el nombre de archivo, sin la ruta completa del directorio. Esto tiene la ventaja de que el usuario cambia el directorio de trabajo una sola vez y ya no necesita cambiar las rutas de acceso de los archivos mencionados en diferentes partes del archivo .do mientras trabaja en él. Alternativamente, se puede configurar una macro global para asignar un directorio de almacenamiento de datos y guardar el trabajo. De ahí, solo se invoca el nombre de la macro en vez de repetir toda la estructura del directorio para trabajar con los datos o guardar algo. Por ejemplo, en Windows, use el comando <*global pathin "C:\Data\HES"*>. Más adelante, para importar los datos almacenados en este directorio desde el archivo .do, use el comando <*use $pathin\filename.dta*> y Stata buscará automáticamente el archivo de datos en el directorio definido en la macro global como *pathin*. La estructura de la ruta de directorio puede variar según el sistema operativo que use. Más adelante veremos con más detalle el uso de macros.

Practicar con conjuntos de datos como ejemplo: Stata ofrece dos tipos de conjuntos de datos con fines de demostración y práctica. Estos son: (a) los conjuntos de datos de práctica instalados junto con Stata en alguna computadora local; y (b) los conjuntos de datos en línea a los que se hace referencia en el manual de Stata y que son accesibles en Internet. Desde la interfaz del usuario de Stata, vaya a "*File>Example* data sets" y se enlistarán los datos disponibles. Haga clic en los conjuntos de datos para abrirlos dentro de Stata y practicar con ellos. Como alternativa, si conoce los nombres de los conjuntos de datos, use el comando <*sysuse datafile*> sustituyendo *datafile* por el nombre de un archivo de conjunto de datos en particular dentro del sistema. También se puede usar el comando <*webuse datafile*> para cargar un conjunto de datos especifico, a través de la web y, de forma predeterminada, los conjuntos de datos se cargan desde http://www.stata-press.com/data/r15/. Este enlace también ofrece una lista detallada de conjuntos de datos organizados por temas y se puede navegar a través de los que están disponibles para practicar.

Uso de operadores lógicos y relacionales: Stata utiliza varios operadores lógicos y relacionales para ayudar con la gestión de los conjuntos de datos. Aquí proporcionamos algunos de los operadores más utilizados y sus significados previstos. Además de estos, Stata también tiene operadores para gestionar variables categóricas (también conocidas como variables factoriales, ficticias, dicotómicas o *dummy*). Anteponga una variable con (i.) para especificar indicadores para cada categoría de una variable. Esto funciona mejor en vez de crear variables *dummy* por separado. El comando <*fvset base*> se puede utilizar

para establecer la categoría base. Ingrese (#) entre dos variables factoriales para crear una variable de interacción. Ingrese (##) para especificar los efectos principales para cada variable y sus interacciones. De igual forma, se puede usar (c.) para interactuar una variable continua con una variable categórica anteponiendo la variable continua con (c.). Por ejemplo, supongamos que la edad (*age*) y el sexo (*sex*) son variables factoriales y que el índice de masa corporal (BMI, por sus siglas en inglés) es una variable continua. Para obtener la regresión de los efectos de estas variables en la presión arterial (BP, por sus siglas en inglés), las siguientes regresiones producen el mismo resultado: <*regres bp i.age i.sex age#sex*> *y* <*regres bp age##sex*>. De manera alternativa, para obtener la regresión de *bp* en *age* y *bmi* y la interacción entre ellos, ingrese <*regres bp age##c.bmi*>.

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

Uso de macros: Las macros son abreviaturas o alias que tienen un nombre y un valor. Cuando su nombre no está referenciado, devuelve su valor.[17] Por lo tanto, una macro tiene un nombre y un contenido de macro. En todas partes donde el nombre de la macro se utiliza en el programa con puntuación, los contenidos de esta se sustituyen en su lugar. Las macros se usan para varios propósitos que incluyen simplificar las tareas, hacer que los archivos .do sean más organizados, acortar la longitud del código Stata y otros usos prácticos durante la programación. Las macros pueden ser de dos tipos, locales y globales, dependiendo de su alcance, es decir, donde se reconozca 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 archivo .do en el que están definidas.[18]

Para sustituir el contenido de las macros de un nombre de macro global, el nombre de esta se marca con un signo de dólar ($) al frente. De igual forma, para sustituir el contenido de la macro de un nombre de macro local, el nombre de esta se marca con comillas simples de apertura y de cierre (‘’).[18] Por ejemplo, al definir una macro local con el nombre *indvar* como <*local indvar price expenditure hsize*> y ejecutar el comando <*summarize ‘indvar’*> devolverá las estadísticas de resumen para cada una de las variables *price*, *expenditure* y *hsize* en los resultados. Del mismo modo, al definir una macro global como <*global xyz age income sex*> y ejecutar el comando <*summarize $xyz*> devolverá el resumen de cada una de estas variables: *age, income* y *sex*. Como las macros globales pueden crear conflictos en los archivos .do, rara vez se utilizan. Generalmente se prefieren las macros locales al escribir el código en el archivo .do. Las macros también se pueden definir como una expresión y el resultado se convierte en el contenido de la macro. Por ejemplo, al definir <*local result = 5+5*> junto con el comando <*display ‘result’*> obtendríamos 10. Las macros también pueden ofrecer funcionalidades extendidas con funciones de macro extendidas. Use el comando <*help macro*> para conocer más sobre las macros y su variedad de usos creativos.

Uso de loops: Los bucles o *loops* son comandos de Stata que ayudan a enlazar una lista arbitraria de secuencias o números. Por ejemplo, un comando *loop* puede establecer repetidamente un nombre de macro local para cada elemento de la lista y ejecutar los comandos entre corchetes. Los *loops* son muy útiles y convenientes al realizar tareas repetitivas que se ejecutan de forma secuencial y se utilizan ampliamente durante la programación. Los comandos de Stata <*foreach*> y <*forvalues*> son particularmente útiles para hacer *looping*. Estos comandos *loops* comienzan y terminan con los corchetes “{” y “}” en líneas separadas. El corchete abierto debe aparecer en la misma línea que <*foreach*> y el corchete cerrado debe aparecer en una línea por sí mismo al final. Por ejemplo:

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

La primera línea anterior enlista las diferentes variables sobre las que se tiene que repetir el comando (es decir, *var1, var2* y *var3*) y las siguientes dos líneas dan los comandos que deben repetirse. El primer comando le dice a Stata: si una observación para una variable en la lista tiene un valor menor o igual a cero, entonces debe reemplazarse con un punto. El segundo comando 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 y que se definen como un logaritmo natural de las variables existentes en la lista. Podríamos agregar múltiples líneas de comandos, una debajo de la otra, y todas se repetirían sobre todas las variables mencionadas en la primera línea. El código anterior también se puede ejecutar de una manera más eficiente utilizando macros locales. Por ejemplo, al predefinir una macro local <*local varlist var1 var2 var3*> y usar el *loop*:

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

Stata también puede realizar este tipo de comandos *loop* sobre diferentes archivos a la vez. De igual forma, el comando <*forvalues*> se puede usar para ejecutar operaciones similares aplicadas a números. Por ejemplo, supongamos que hay 25 estados en una encuesta de hogares y los gastos de consumo promedio en cada estado se encuentran bajo los nombres de las variables *state1, state2,…, state25*. Para convertir todas esas variables a su 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*.

Devolución de resultados almacenados: Stata almacena regularmente los resultados de los comandos en macros locales que se pueden ejecutar para varios propósitos. Por ejemplo, al ejecutar 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*> a partir de los datos automáticos en Stata nos devuelve 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 nos dará los resultados como se muestra 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. Estos están disponibles para ser utilizados inmediatamente después para generar nuevas variables o para usarse en otros comandos. Similar a <*return list*>, use el comando <*ereturn list*> para mostrar los contenidos almacenados localmente después de los comandos de cálculo, tales como <*regress*>. El comando <*help return*> en Stata mostrará otros usos de los comandos de retorno. El recuadro 2.1 nos da un ejemplo funcional de cómo utilizar algunas de las sugerencias de Stata que ya hemos visto.

Recuadro 2.1 Sugerencia 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 muestra el uso de macros, *loop* y resultados almacenados, todo en un solo lugar. La primera línea importa los datos de *automóviles* incorporados y la segunda define una macro local llamada *'items'* que consiste en tres variables. El tercero abre un comando *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 a través de este *loop*. La primera resume la variable y con la adición del prefijo *'quietly'* ejecuta este comando sin mostrar los resultados. La opción <*detail*> más <*summarize*> solicita estadísticas adicionales que generalmente no se calculan, como los percentiles, la asimetría y la curtosis.

La segunda línea del *loop* define una nueva macro local *'upper'* utilizando los resultados almacenados después de <*summarize*>. Se define como la media + 3 desviaciones estándar de la variable en consideración. La tercera línea en el loop reemplaza cualquier valor mayor que la media más 3 desviaciones estándar y menor que los valores faltantes (Stata considera que los valores faltantes son mayores que cualquier valor numérico) con el valor de la mediana de esa variable. El corchete en la última línea cierra el loop.

Uso de delimitadores: El comando <*#delimit ;*> se utiliza para restablecer el carácter que marca el final de un comando en Stata. Estos se utilizan solo en archivos .do y archivos .ado (que se definen en la siguiente sección). Al pulsar la tecla de retorno se le indica a Stata que ejecute el comando. En un archivo .do, el final de una línea asume el papel de la tecla de retorno y en sí, estas mismas líneas tienen restricciones de caracteres. Por lo tanto, se le indica a Stata que los comandos pueden ser más largos que una línea al utilizar el comando <*#delimit ;*> para seccionar libremente las líneas de comando según sea necesario. Stata considerará todas las líneas continuas hasta que vea el carácter delimitador que marque el final del comando como una sola línea lógica. De forma alternativa, se puede usar < /* */ > como un delimitador de comentarios. Por ejemplo, <*generate X = 3*Y /* this is a comment*/ + 5*> es lo mismo que <*gen X = 3*Y + 5*> sin el comentario. También se pueden seccionar líneas largas con tres barras diagonales consecutivas (///), en lugar de usar el comando <*#delimit ;*>. Estas son bastante útiles al preparar archivos .do. 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
```

Uso de comandos complementarios: Stata le permite escribir comandos de terceros (conocidos como "archivos .ado") que se pueden almacenar en un archivo SSC (*Statistical Software Components, Componentes* de software estadístico), que comúnmente se le llama el Archivo del Colegio de Boston y es proporcionado por http://repec.org. Los usuarios pueden instalar estos programas complementarios desde el archivo SSC usando el comando <*ssc install progname*> donde *progname* es el nombre del archivo .ado o del archivo del programa que se necesite instalar. También se puede desinstalar algún paquete en particular con el comando <*ssc uninstall progname*>. La mayoría de los paquetes complementarios ofrecen alguna funcionalidad adicional en comparación con los comandos integrados de Stata. Por ejemplo, el paquete complementario <*estout*>, que se puede instalar con <*ssc install estout*>, ayuda a crear tablas ordenadas a partir de cálculos almacenados después de los comandos de regresión. Puede crear tablas dignas de publicación con coeficientes de regresión, usando estrellas para indicar el nivel de importancia, estadísticas de resumen, errores estándar, *t-statistic (estadístico t), valores-p* e intervalos de confianza para uno o más modelos ajustados anteriormente y almacenados con el comando <*estimates store*>. Del mismo modo <*findname*>, <*outreg2*> y <*ivreg2*> son algunos de los complementos populares. Use el comando <*ssc whatshot*> para ver algunos de los paquetes complementarios más populares disponibles para descargar.

Otros consejos: Aquí se incluyen algunos consejos adicionales que no se han mencionado:

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 una 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*> utilice <*reg*> y así sucesivamente.
- El nombre dado a los escalares dentro del archivo .do debe ser distinto de cualquiera de las otras variables o sus abreviaturas únicas dentro de los datos. Si se define un escalar con el mismo nombre que otra variable o su abreviatura única, Stata le dará prioridad al nombre de la variable o su abreviatura antes que al nombre del escalar especificado, llevando a resultados inadvertidos al realizar operaciones que involucren este escalar. Como alternativa, utilice una pseudo función <*scalar(xyz)*> para escribir un escalar con el nombre xyz cada vez que se use el escalar en alguna estimación o al definir más escalares.

- Los valores faltantes, indicados por un punto (.), se codifican y se tratan como un infinito positivo en Stata. Por lo tanto, adquiere un valor más alto que todos los demás valores numéricos. Esto es importante al momento de 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 hay valores faltantes en Y. En su lugar, utilice <*replace X = 0 if Y>100 & y<.*>

2.5 Técnicas para extraer datos utilizando Stata

Los microdatos de las encuestas de hogares se almacenan en diferentes formatos de archivo, dependiendo del hardware que se utiliza para registrar los datos, la disponibilidad del software con los organismos encuestadores y otros usos y costumbres en diferentes campos. Los datos de las HES que nos interesan generalmente serán datos tabulares cuantitativos. Por lo general, se presentan en archivos de texto delimitado que contienen metainformación como la que se encuentra en el software estadístico Stata, SPSS y SAS o en archivos de valores simples separados por comas (.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 describa cada columna del archivo de datos que tenga longitudes de registro fijas. Por ejemplo, el diccionario diría: la posición del byte 4 en el archivo de datos indica el código para el área rural o urbana; las posiciones de los bytes 9 al 10 indican el código para las PSU o el identificador de conglomerados; o, las posiciones de los bytes 30 al 36 indican los gastos en un artículo. También habrá un archivo, generalmente llamado libro de códigos, que indica el significado de los diferentes códigos usados en el archivo de diseño o de datos. Por ejemplo, indicaría que el valor 1 = rural y 2 = urbano, o 1 = masculino y 2 = femenino. La información final que es archivada por los respectivos organismos encuestadores usualmente proporciona toda la documentación necesaria asociada con los datos. El catálogo de la IHSN,[12] 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 del diccionario o del archivo de diseño para la mayoría de los datos de la encuesta catalogados allí.

El software que se utilice para el análisis estadístico debería poder importar microdatos antes de que se lleven a cabo diferentes análisis. Para tomar una decisión informada sobre qué datos deben extraerse o importarse al software estadístico para su análisis posterior, es necesario contar con una descripción detallada y con la documentación 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 generar cualquier estimación a partir de estos datos, se debe extraer la parte relevante de los datos y agregarlos utilizando los comandos apropiados en el software analítico. Stata utiliza diferentes métodos para importar datos según el tipo de archivo de datos de origen. Al ejecutar el comando <*help import*> en la línea de comandos de Stata, se enlistan diferentes opciones y comandos disponibles para importar datos de diferentes formatos.

Ya que los microdatos para la mayoría de las HES están en formato fijo ASCII, el siguiente ejemplo muestra 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 lo que representa el carácter en cada posición de byte en el archivo de datos ASCII. Para extraer o importar estos datos a un formato legible en Stata, o para convertirlos a un conjunto de datos de Stata (.dta), se debe crear un archivo de diccionario Stata con la extensión de archivo ".dct". En el Recuadro 2.2 se muestra un ejemplo de archivo de diccionario para extraer partes de la información proporcionada en el archivo de datos ASCII.

Ejemplo de archivo de datos en formato ASCII (formato fijo)

W155110211307112666212020 11 2 4 33815604 488 573003232 0030251
W155110211307112666212020 31 2 4 33815604000 490 547001213 0010211
W155110211307112666212020 51 2 4 33815604 437 460004413 0610251
W155110211307112666212020722 2 4 33815604 473 554001413 0410251

Ejemplo de archivo de diseño

elemento	*longitud*	*pos. byte*	*comentarios*
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*	

Para ejecutar el programa de diccionario de Stata, abra Stata, establezca el directorio de trabajo y ejecute el comando: <*infile using dictionary*> donde *dictionary* es el nombre del archivo del diccionario. Si el programa se ejecuta correctamente, aparecerá en la pantalla seguido del mensaje *"N observations read"*, donde N indica el número de observaciones en los datos importados. A continuación, ejecute el 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, ejecute el comando <*compress*> para cambiar las variables a su formato más eficiente. Por último, los datos importados se pueden guardar en la extensión nativa de Stata (.dta) con el comando <*save mydata*> donde *mydata* es el nombre del archivo de datos que se guardará en el directorio de trabajo de Stata.

2.6 Preparación y construcción de datos para el análisis técnico

Las HES a menudo proporcionan múltiples conjuntos de datos para registros individuales, de hogares y para otras variables. Los gastos en diferentes productos básicos en sí pueden estar en diferentes archivos de datos. Además, los datos podrían estar codificados incorrectamente en ciertas variables por lo que algunos errores obvios podrían corregirse fácilmente para evitar que se pierdan esas observaciones durante el análisis final. También puede haber algunos valores extremos o faltantes que deban tenerse en cuenta. Por todas estas razones, es importante limpiar los archivos de datos individuales y fusionarlos en un solo archivo antes de llevar a cabo cualquier análisis adicional. Esta sección ofrece algunos pasos básicos a seguir antes de preparar cualquier conjunto de datos finales para realizar el análisis estadístico.

Recuadro 2.2 Ejemplo de archivo de diccionario para importar datos desde archivos ASCII

```
dictionary using datafile.txt {
_column(1) str2     ID          %2s   "ID del archivo de trabajo"
_column(6)          sector      %1f   "Rural o Urbano"
_column(7)          state       %2f   "Estados"
_column(9)          region      %1f   "Regiones del país"
_column(10)         stratum     %2f   "Estrato"
_column(12)         district    %2f   "Distrito"
_column(14)         subround    %1f   "Muestra sub Ronda"
_column(16)         fsu         %5f   "Unidades de primera etapa"
_column(25)         hldno       %2f   "Número de hogar"
_column(58)         hsize       %3f   "Tamaño del hogar"
_column(63)         socgroup    %1f   "Grupo social"
}
```

Un archivo de diccionario de Stata comienza con una línea que se parece a esta <*dictionary using datafile.txt {*> donde *datafile.txt* es el nombre del archivo de microdatos en el directorio de trabajo de Stata. La definición de variables individuales sigue a continuación. Cada variable está definida por una línea con 5 partes. La primera parte le dice a Stata que comience a leer el archivo de datos desde la posición de byte mencionada entre paréntesis. La segunda indica el tipo de variable: de cadena o numérica. Solo las variables de cadena deben indicarse explícitamente como tales. La tercera parte es el nombre mnemotécnico de la variable. La cuarta es el formato de entrada de la variable que consiste en un signo "%", un número que indica el ancho de la variable y una letra que indica el formato de la variable: f para los números y s para las cadenas. La quinta parte es una etiqueta opcional dada a la variable. El programa del diccionario termina con un corchete de cierre, "}".

Algunos ejemplos de formatos de entrada que se pueden usar en las definiciones de variables son: %5f — variable numérica de cinco columnas, %10s — variable de cadena de diez columnas y %7.2f — 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 principio de la siguiente línea debajo de "}". Finalmente, el archivo debe guardarse con la extensión de archivo .dct (por ejemplo, *dictionary.dct*).

Fusión de datos: Las encuestas de hogares a menudo vienen con múltiples archivos o registros de datos de hogares y de los miembros que los conforman. Asimismo, para los propios hogares, puede haber múltiples registros. Por ejemplo, un archivo con las características básicas del hogar, como su tamaño, el grupo SES al que pertenece, el lugar de residencia, etc., y otro archivo para sus gastos de consumo. Los datos sobre los gastos de consumo en sí podrían distribuirse en diferentes archivos de datos. Por lo tanto, podría ser necesario crear archivos de diccionario por separado para extraer los datos de diferentes archivos y fusionarlos después de que cada conjunto de datos se haya extraído en archivos de Stata separados.

Ya que este Conjunto de herramientas cubre el análisis a nivel de los hogares, la información individual necesita agregarse a nivel de los hogares. 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 dé la proporción de sexo (proporción del número de hombres y mujeres en un hogar). De igual forma, el nivel de educación de los miembros individuales en un hogar no es relevante para un análisis a nivel de hogar. No obstante, se puede crear una variable para el análisis a nivel de hogar con el promedio de años de educación recibidos por hogar para conocer el nivel educativo de los mismos.

Una vez que las variables deseables a nivel de hogar se hayan generado a partir de los registros de datos individuales, solo es necesario conservar una observación por hogar antes de fusionarla con los datos a nivel de hogar. Por ejemplo, una vez que se genera una variable a nivel de hogar a partir de datos de nivel individual, por ejemplo, *sex ratio* (la proporción de sexos), se repetirá el mismo valor para *sex ratio* para todos los miembros dentro de un hogar. Para conservar solo una observación por hogar, primero se debe ordenar los datos por hogar (o por ID [identificación] de hogar) con el comando <*sort hhid*> (donde *hhid* es la variable de identificación para los hogares) y luego ejecutar el comando <*drop if hhid==hhid[_n-1]*>. Alternativamente, se puede usar el comando <*duplicates drop*> después de organizar los datos según sea necesario.

La fusión de datos a nivel de hogar con datos adicionales ya sea de los registros individuales o de otros registros específicos de hogar, requerirá el uso del comando <*merge*> en Stata. Ejecute el comando <*help merge*> para ver la sintaxis, así como las diferentes formas de fusionar archivos de datos en Stata. Para facilitar la comprobación de si la fusión se ha realizado correctamente, Stata genera una nueva variable, <*_merge*>, después de cada comando de fusión. 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*> dará la indicación necesaria después de la ejecución de <*merge*>. Por ejemplo, el código 3 para *_merge*, es para observaciones que coinciden correctamente con ambos conjuntos de datos.

El aspecto más importante de la fusión de dos archivos diferentes es poder encontrar un conjunto de variables que puedan identificar de forma única todas las observaciones individuales en cada uno de los datos que se fusionarán. Esto debe entenderse a partir del diseño de la encuesta y extraerse junto con cada extracción de datos mediante archivos de diccionario. La falta de identificadores únicos o la existencia de identificadores definidos incorrectamente, puede resultar en la combinación errónea de información de un hogar con otro. El Recuadro 2.3 da un ejemplo de cómo identificar estas variables y fusionar archivos correctamente.

Recuadro 2.3 Posible discrepancia al fusionar los hogares

La Encuesta sobre ingresos y gasto de los hogares de Bangladesh (2010)[18] sigue una técnica de muestreo aleatorio estratificado en dos etapas. La descripción del diseño de la muestra en el informe publicado dice que se seleccionaron alrededor de 200 hogares de entre unas 1 000 PSU en todo el país, mientras que las propias PSU se seleccionaron de entre unos 16 estratos diferentes. Está claro que un hogar de esta encuesta debe identificarse de manera única utilizando las variables que representan los estratos, la PSU y el número del hogar. Estas variables son *stratum, psu* y *hhold*, respectivamente, como se indica en la documentación. Ya que los números de la PSU son únicos en estos datos, también se puede identificar una ID de hogar única usando solo las variables *psu* y *hhold*.

Con el comando <*egen hhid=group(psu hhold)*> se puede generar una variable única de ID de hogar (*hhid*) para estos datos donde los valores entre paréntesis corresponden a los

nombres de variables requeridos para identificar de manera única al hogar. Por ejemplo, si los números de la PSU no fueran únicos y variaran entre estratos, se tendrían que usar las tres variables al generar el *hhid*. Por lo tanto, cualquier fusión de dos registros a nivel de hogar en estos datos utilizará estas variables. Por ejemplo, la HIES tiene un archivo de datos demográficos de hogar (rt001) y un archivo de gastos agregados a nivel de 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*>. Esto 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 que luego se fusionaron usando el comando <*merge 1:1 hhid using hh2*> donde se utilizó la variable de identificación única (*hhid*) generada previamente para fusionar en vez de los identificadores de hogar originales (*psu* y *hhold*). Esto también fusionará los archivos de datos y el comando <*tab _merge*> no mostrará discrepancias. Sin embargo, en este caso, los hogares en ambos datos podrían estar fusionados de forma incorrecta debido a varias razones:

1. Al generar un *hhid* único en cada archivo de datos individual, Stata asigna identificadores únicos a cada hogar usando 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*, se obtendrá una concordancia incorrecta entre los hogares después de la fusión.

2. Supongamos que algunos números *psu* o *hhold* fueran diferentes en ambos conjuntos de datos debido a una codificación incorrecta. El comando <*tab _merge*> después de una fusión correcta usando tanto *psu* como *hhold* mostrará observaciones con discrepancias. Mientras que al usar *hhid*, generado previamente, para fusionar ambos archivos de datos, los fusionaría perfectamente fallando al no identificar discrepancias.

3. Supongamos que el número de observaciones en *hh1* y *hh2* fuera diferente. Una fusión con las variables *psu* y *hhold* combinaría los hogares de forma correcta, mientras que la fusión con *hhid* generado previamente los combinaría inadvertidamente.

Por lo tanto, los datos de dos archivos de datos diferentes deben fusionarse siempre utilizando únicamente todas las variables relevantes que se utilizan para identificar las observaciones únicas (hogar o persona) en cada archivo de datos. En otras palabras, el comando <*merge*> debería tener todas las variables que identifican de forma única una observación presente durante la fusión.

Para realizar una fusión uno a uno, tanto los datos *master (datos maestros, están en la memoria)* como los datos *using (datos en uso)* deben ser identificables con el mismo conjunto de variables únicas. Solo se puede realizar un análisis posterior con aquellas observaciones que coinciden con los archivos de *datos master* y con los de datos *using*, es decir, observaciones para las cuales la variable *(_merge)* toma el valor 3. Para solo usar las variables sin datos faltantes, tanto de los archivos de datos *master* como de los de datos *using*, es importante omitir las observaciones para las que *_merge* no es igual a 3 usando el

comando <*drop if _merge!=3*>. Sin embargo, puede haber situaciones en las que sea necesario mantener aquellas observaciones no coincidentes, ya sea del archivo de datos *master* o del de datos *using*, en el archivo de datos fusionados.

Además de fusionar diferentes archivos (por ejemplo, datos de hogares y datos individuales) de la misma ronda de una HES dada, también puede haber situaciones en las que el usuario quiera agrupar datos de las HES de diferentes años u olas. Obviamente, los hogares en diferentes rondas de HES pueden ser diferentes entre sí y lo que se requiere no es una fusión, sino la agrupación de diferentes HES para que haya una sección transversal agrupada. En este caso, en lugar de <*merge*> se debe usar el comando <*append*> en Stata. Para hacer esto, los datos de cada ronda de HES deben contener el mismo tipo de variables y primero se debe preparar una sola fusión de datos para cada ronda de HES. Una vez que se haya anexado, simplemente se agregará al número de observaciones en los datos maestros. Antes de anexar, es importante crear una variable de año u ola y marcarla con números que puedan identificar a cada año/ola/ronda de la encuesta. Si los datos agrupados finales pertenecen a varios años (generalmente de diferentes olas de la encuesta), también es importante ajustar la inflación en cualquier variable de gasto o precio para que los valores de las diferentes rondas de datos estén en términos constantes y, por lo tanto, sean comparables.

Remodelación de datos: Dependiendo del análisis que se realice, puede ser importante para remodelar datos a formato largo o ancho en Stata. Para hacer esto, ejecute el comando <*help reshape*> para entender cómo se hace la remodelación de una forma a otra. En un formato ancho, solo tendremos 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, supongamos que hay información sobre el gasto en cigarrillos y en tabaco sin humo. Para los hogares con gasto en ambos productos, habrá dos observaciones para cada hogar en un formato largo, mientras que, en el formato ancho, el gasto en cigarrillos y en tabaco sin humo aparecerá como variables separadas frente a una sola observación del hogar. Para la mayoría de los análisis, es útil remodelar los datos en formato ancho. Por lo tanto, si los datos extraídos están en formato largo, deben remodelarse a formato ancho con el comando <*reshape wide stub, i(i) j(j)* >, después de determinar la observación lógica (i) y la subobservación (j) mediante las cuales se organizarán los datos.

Limpieza de datos: Limpiar los datos antes de realizar el análisis estadístico es esencial, especialmente en el caso de las encuestas de hogares, ya que se trata de datos recopilados por diferentes personas en todo el país en diferentes etapas. Por ejemplo, un cero en lugar de un valor faltante podría generar resultados indeseables, como la distorsión de la media y las varianzas al realizar análisis estadísticos. Los 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 formas de escribirse o espacios entre observaciones, Stata consideraría estas entradas como una categoría diferente. Por ejemplo, si el sexo 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, podría haber varias categorías diferentes. Por estas y otras razones, es importante hacer un examen minucioso 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, incluyendo comandos útiles de Stata que se pueden usar durante estos pasos. Tenga en cuenta que los pasos mencionados en la tabla no tienen que realizarse estrictamente en el mismo orden que se indican. Usando el comando de ayuda de Stata, seguido por los comandos pertinentes mencionados en esta tabla, el lector puede aprender más sobre cada uno de esos comandos y familiarizarse con diferentes ejemplos.

Tabla 2.1 Estrategia de limpieza de datos

Razón (¿Por qué hacerlo?)	Paso (¿Qué hacer?)	Comando (¿Cómo hacerlo?)
Identificar las variables y corregir códigos incorrectos	Etiquetar/volver a etiquetar las variables y etiquetar sus valores	label; recode
Identificar observaciones únicas para fusionarlas correctamente	Comprender los identificadores únicos del diseño de la encuesta y de 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	Transformación de variables	gen; destring; tostring; drop; keep; egen; rename; bysort; encode; recode;
Asegurarse de que las conexiones lógicas estén presentes en los datos, por ejemplo, que las madres estén como femenino o las cantidades tengan unidades correctas	Comprobación de consistencia	assert; tabulate; summarize; table; tabstat; count;
Crear un único archivo de datos en cual trabajar	Fusionar o anexar 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 largo o ancho apropiado	Reshape
Identificar la importancia e influencia de los valores faltantes	Decidir si las observaciones faltantes deben ser eliminadas o imputadas	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	Utilice el editor de archivos .do para organizarlo

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

Un programa de software estadístico generalmente analiza los datos como si se hubieran recopilado mediante un muestreo aleatorio simple. Sin embargo, como se mencionó anteriormente, la mayoría de las encuestas de hogares utilizan un diseño de encuesta más complejo y de multietapa para recopilar datos, y la estratificación y la conglomeración en encuestas por muestreo afectan el cálculo de los errores estándar. Por lo tanto, el análisis estadístico realizado debe ser capaz de corregir los elementos de diseño utilizados

en la encuesta para obtener estimaciones puntuales más precisas y errores estándar. La documentación proporcionada junto con los datos de la encuesta suele proporcionar información detallada sobre el diseño de muestreo específico que se utilizó. En esta sección se 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 el código útil de Stata para realizar estas acciones.

En Stata, el comando <*svyset*> se utiliza para declarar el diseño de los datos de la encuesta. Designa variables que contienen información sobre el diseño de la encuesta, como las ponderaciones de muestreo, la PSU o el conglomerado y los estratos, y especifica otras características de diseño de la encuesta, como el número de etapas y el método de muestreo. La declaración de diseño, de ser necesaria, se puede borrar con el comando <*svyset, clear*>. Una vez declarados los datos con <*svyset*>, solo el prefijo <*svy:*> tiene que preceder a cada comando. La sintaxis del comando <*svyset*> para el diseño de una encuesta multietapa es la siguiente: <*svyset psu [weight] [, design options] [|| ssu, design options] … [options]*> donde *psu* es el nombre de una variable que identifica la unidad de muestreo primaria en los datos, *weight* identifica la ponderación de muestreo, *ssu* identifica las unidades de muestreo en la segunda etapa, y así sucesivamente. Las opciones de diseño declararán los elementos de diseño como estratos. Por ejemplo, el sitio web de Stata proporciona un conjunto de datos de la encuesta de muestra de la segunda Encuesta Nacional de Examen de Salud y Nutrición (NHANES, por sus siglas en inglés) en los Estados Unidos entre 1976 y 1980. Importe esos datos a Stata con el comando <*webuse nhanes2*>. Los datos proporcionan una variable de ponderación *(finalwgt)*, una variable de la PSU *(psu)* y una variable de estratos *(strata)*. El comando <*svyset*> en este caso se verá como: <*svyset psu [pw=finalwgt], strata(strata)*> donde *pw* representa ponderaciones probabilísticas.

La mayoría de las encuestas explícitamente incluyen ponderaciones de muestreo, estratos e identificadores de PSU junto con los datos publicados. Se necesita leer cuidadosamente la documentación de la encuesta para entender la descripción de las variables. Dado que los informes publicados de la encuesta también presentan estimaciones puntuales importantes, se pueden comparar las cifras calculadas con las de los informes publicados. Antes de continuar con análisis adicionales, es importante realizar dicha comparación cruzada para asegurarse de que se están utilizando las ponderaciones de muestreo y los elementos de diseño de la encuesta correctos, tal como se pretendía originalmente.

Una vez que se ha declarado el diseño de la encuesta a través de <*svyset*>, la información sobre los estratos y la PSU se puede obtener con el comando <*svydescribe*>. La estimación adicional de las estadísticas descriptivas debe ir precedida de <*svy:*>. Por ejemplo, para estimar la media de una variable, simplemente se podría ejecutar el comando <*svy: mean varname*>. Si se calcula la media para una variable binaria, mostrará las proporciones. Alternativamente se puede ejecutar <*svy: tab binaryvar*> para estimar las proporciones de, digamos, hombres y mujeres, analfabetas y alfabetizados o variables binarias similares junto con sus errores estándar corregidos para el diseño de la encuesta. Del mismo modo, <*svy: proportion binaryvar*> proporcionaría un resultado con proporciones de la variable de interés junto con sus errores estándar y su intervalo de confianza.

Para estimar las mismas estadísticas descriptivas para los subgrupos de la encuesta, tales como grupos de ingresos, género o cualquier otra categoría SES, 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 se quiera encontrar el gasto promedio en cigarrillos por diferentes cuartiles de gasto. Para ello, primero cree una variable para categorizar los hogares en cuatro cuartiles diferentes basados en su gasto familiar total mensual *(exptotal),* como se indica a continuación: <*xtile exp_quartiles =exptotal, n(4)*>. A continuación, utilice el comando <*svy, over(exp_quartiles) : mean exp_cig*> para obtener el gasto medio mensual en cigarrillos por diferentes cuartiles de gasto.

Las estimaciones 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, se hace con la ayuda de ponderaciones y opciones de conglomerados robustos. Por ejemplo, en el ejemplo anterior de gastos en cigarrillos por cuartiles de gasto, se puede obtener el mismo gasto promedio por diferentes cuartiles de gasto utilizando el comando <*mean exp_cig [pw=weightvar], over (exp_quartiles)*>, donde *weightvar* es el identificador de la ponderación de muestreo que se utilizó para declarar el diseño de la encuesta. Sin embargo, aunque las estadísticas descriptivas que utilizan las ponderaciones de muestreo producen las mismas estimaciones que las que utilizan <*svyset*>, no abordan adecuadamente los problemas de estratificación y, como resultado, podrían producir errores estándar diferentes de los que se obtienen con el comando <*svy*>. En el contexto de la regresión, sin embargo, se podría añadir el argumento opcional <*robust cluster(psuvar)*> después del comando de regresión principal donde *psuvar* es la variable que identifica el conglomerado o la PSU en los datos y corrige los efectos del diseño de la encuesta mientras calcula los errores estándar para las estimaciones de los coeficientes.

3 Estimación de la elasticidad precio y elasticidad cruzada

En este capítulo se presentan métodos para estimar la elasticidad precio de la demanda utilizando HES. La elasticidad precio es uno de los parámetros más importantes que deben tenerse en cuenta al diseñar políticas fiscales ya que proporciona a los responsables de la creación de políticas públicas un entendimiento sobre la capacidad de respuesta de la demanda a los cambios en los precios. Basándose en la elasticidad precio estimada, los responsables de la creación de políticas pueden predecir con cierto grado de confianza el impacto de sus políticas en sus objetivos relevantes, incluyendo el consumo de tabaco y los ingresos fiscales. Además, la evidencia empírica sobre la magnitud a la cual la demanda de tabaco respondería a los precios proporciona un contraargumento muy relevante para los que afirman que el aumento de los impuestos resultaría, indudablemente, en una reducción de los ingresos fiscales.

Los responsables de crear políticas públicas están interesados en la capacidad de respuesta del consumo de tabaco, no solo en cuanto a los cambios en los precios del tabaco (es decir, la elasticidad de los precios), sino también en los cambios en los precios de otros productos, como sus posibles complementos (por ejemplo, el alcohol, el café, etc.), o sus sustitutos. De manera similar, los creadores de políticas públicas tal vez quieran conocer el impacto de un cambio en el precio de algún tipo de producto de tabaco (por ejemplo, los cigarrillos), en otros tipos (como los cigarrillos de tabaco a granel), ya que el impacto de su política puede reducirse efectivamente si, por ejemplo, hay margen para la sustitución a la baja.

En este capítulo estos conceptos se definen en detalle con algunos ejemplos. En la última parte del capítulo, se proporciona el código de Stata para que el lector pueda estimar las elasticidades. Por último, se presenta un ejemplo de Uganda.

3.1 Definición de conceptos

La elasticidad precio de la demanda se define formalmente como el cambio porcentual en la cantidad demandada de un producto que resulta de un cambio del 1 % en el precio de ese producto, manteniendo todo lo demás sin cambios (*ceteris paribus*). Por ejemplo, una elasticidad de precio de la demanda de -0.5 implicaría que la cantidad demandada de ese producto en particular disminuye en un 5 % cuando el precio del producto sube en un 10 %. Del mismo modo, una elasticidad precio de la demanda de -1.5 implica que la cantidad demandada del producto en cuestión disminuye en un 15 % cada vez que su precio aumenta en un 10 %.

Se dice que los productos con una elasticidad precio de la demanda con valor absoluto inferior a 1 tienen una demanda inelástica porque la respuesta de la demanda es relativamente menor que la variación de precios. Por otro lado, se dice que los productos con una elasticidad de la demanda con un valor superior a 1 tienen una demanda elástica porque la respuesta de la demanda es relativamente mayor que el cambio de precios. Existen varios factores que influyen en la elasticidad de los precios; la disponibilidad de sustitutos, si un producto es de necesidad básica, el período de tiempo disponible para encontrar alternativas, qué tan amplia o estrechamente está definido, o su naturaleza adictiva/usual. Teniendo esto en

cuenta, los productos de tabaco que tienen pocos sustitutos y que son adictivos tienden a tener una demanda de precio relativamente inelástica.

Que la demanda de un producto sea elástica o inelástica es muy importante para la política fiscal. Se puede esperar que los ingresos fiscales disminuyan cuando se aumentan los impuestos en algún producto que es de demanda elástica ya que la respuesta de la demanda supera al cambio de precios, de modo que los ingresos por ventas y los ingresos fiscales al final disminuyen. Por otra parte, se puede esperar que los ingresos fiscales aumenten cuando se recaudan impuestos de algún producto que es de demanda inelástica ya que su respuesta de la demanda es menor que el cambio de precios, de modo que los ingresos por ventas y los ingresos fiscales al final aumentan.

Los estudios sobre la estimación de las elasticidades de la demanda de cigarrillos tienden a encontrar, en general, elasticidades que oscilan entre 0 y -1,[4,19,20] lo que significa que la demanda de tabaco es inelástica, lo que es de esperarse dada la naturaleza adictiva de este producto, así como la disponibilidad de muy pocos sustitutos cercanos. La evidencia empírica también confirma que los impuestos al tabaco, a través de precios más altos de tabaco, son una de las herramientas políticas más efectivas para reducir el tabaquismo y sus consecuencias adversas para la salud.[4,21–23]

Además de la elasticidad precio, también podemos definir la elasticidad cruzada. Formalmente, la elasticidad cruzada de la demanda entre los productos X e Y se define como el cambio porcentual en la demanda del producto Y cuando el precio del producto X cambia en un 1 %, *ceteris paribus*. A diferencia del caso con elasticidad de precio, donde siempre es indudablemente negativa, la elasticidad cruzada puede tener un signo negativo o positivo. Una elasticidad cruzada negativa significa que los dos productos en cuestión son complementos. En otras palabras, el consumo conjunto de los dos productos satisface una necesidad. Un ejemplo sería la gasolina y los automóviles. Por otro lado, una elasticidad cruzada positiva significa que los dos productos son sustitutos. Es decir, se puede usar un producto en lugar del otro o ambos productos 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 utilizan indistintamente, ya que el gasto total en las HES se utiliza como sustituto de los ingresos. La elasticidad ingreso de la demanda se define formalmente como el cambio porcentual en la cantidad demandada de un producto que surge de un aumento del 1 % del ingreso, *ceteris paribus*. Una elasticidad ingreso negativa de la demanda significa que la cantidad demandada del producto disminuye cuando el ingreso aumenta. Estos productos se denominan productos “inferiores”. Los alimentos básicos (arroz, maíz, etc.) a menudo tienen elasticidades ingreso negativas de la demanda. Por otra parte, los productos con elasticidades ingreso positivas de la demanda se denominan productos “normales”. Conocer la magnitud de la elasticidad ingreso de la demanda es importante para las políticas de control de tabaco. Una elasticidad ingreso positiva de la demanda, por ejemplo, en los cigarrillos de un país, implica que los esfuerzos de control de tabaco deben intensificarse, especialmente en períodos de aumento de los ingresos en ese país.

3.2 Cuestiones econométricas en la estimación de la demanda

Hay varias cuestiones teóricas y prácticas que se deben considerar en la estimación de las elasticidades precio de la demanda. Esta sección cubre algunas de las cuestiones 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 producto, su demanda disminuye, *ceteris paribus*. Asume que la dirección de la causalidad va del precio a la cantidad demandada.

Sin embargo, en realidad, las cosas tienden a ser más complejas porque en las interacciones de mercado la demanda influye en el precio tanto como el precio influye en la demanda. Esto se puede observar en tiempo real en los mercados de valores. Es probable que un aumento en el precio de una acción lleve a una reducción en la cantidad demandada de la acción. Por otra parte, es probable que un aumento en la demanda de la acción lleve a un aumento en su precio. Además, sabemos que otros factores (por ejemplo, los ingresos, los gustos, el clima y los precios de los productos relacionados) pueden, fuera de la influencia del precio, influir en la demanda del producto.

Las cuestiones que se explicaron anteriormente son conocidas en el análisis econométrico como *endogeneity problem (problema de endogeneidad)*, o *identification problem (problema de identificación)*, y el no abordarlas adecuadamente llevaría a obtener estimaciones sesgadas (es decir, las estimaciones son significativamente diferentes al valor real del parámetro que se está estimando). Esta es una cuestión muy relevante en la formulación de políticas ya que conduciría a una política que podría diseñarse con un impacto positivo o negativo de manera poco realista, dependiendo del sentido del sesgo.

Idealmente, el problema de endogeneidad o de identificación, se puede resolver econométricamente al realizar un experimento donde las unidades sean asignadas aleatoriamente en grupos de tratamiento o grupos de control. Aquí, no hay necesidad de preocuparse por la endogeneidad porque la aleatoriedad descarta todos los demás factores excepto el factor que nos interesa. Desafortunadamente, con la realidad social, a diferencia de las ciencias físicas, no siempre es fácil, ni siquiera deseable, el realizar experimentos sociales. Por lo tanto, los economistas y los científicos sociales buscan experimentos “naturales” o cuasi experimentos que puedan aprovecharse 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, exógena) un aumento de los precios del tabaco. Por ejemplo, varios estudios realizados en Estados Unidos en la década de 1990 aprovecharon el aumento de 25 centavos del impuesto a los cigarrillos en California y Massachusetts para estimar la elasticidad precio de la demanda,[24-27] ya que la fuente exacta del cambio de precio que propició un cambio en la cantidad demandada podía ser señalada en estos eventos.

Sin embargo, estos cambios drásticos en los impuestos al tabaco no son muy comunes, especialmente en los PIMB, donde, a menos que se estén reformando, los cambios en estos impuestos son más comúnmente graduales y de pequeña magnitud, por lo general para corregir el impacto de la inflación. Estos cambios graduales hacen que sea difícil aislar el efecto causal del precio en la demanda, por lo que el procedimiento de estimación requiere el uso de las IV para obtener el efecto causal del precio en la demanda (consulte el Capítulo 2 para una discusión sobre la endogeneidad y el papel de las IV para resolverlo).

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

3.2.2 La solución de Angus Deaton al problema de la identificación

Si bien existen algunos modelos diferentes que utilizan un sistema de ecuaciones de demanda, el Sistema de Demanda Casi Ideal (AIDS, por sus siglas en inglés) introducido por Deaton y Muellbauer (1980)[28] ha sido el más popular gracias a sus múltiples ventajas. El AIDS tiene una forma funcional flexible consistente con los datos de gasto de los hogares y los diferentes axiomas de elección. No impone ninguna restricción previa sobre las elasticidades y su especificación, en general no lineal, es fácil de estimar, lo que le permite probar explícitamente las restricciones de homogeneidad y simetría. El modelo de Deaton (1988) que se presenta en este Conjunto de herramientas[29] y explicado en su libro,[7] se basa en Deaton y Muellbauer

(1980).[28] Sin embargo, difiere ligeramente del AIDS ya que permite compras cero, mientras que el modelo original no lo permitía. Permitir compras cero es especialmente atractivo en el caso del tabaco, ya que a menudo no lo consumen todas las personas de una población. Además, los efectos de la política fiscal basados en las estimaciones de la elasticidad precio se examinan mejor cuando todos los hogares están presentes en el análisis, dado que algunos hogares que no consumen tabaco ahora pueden empezar a consumirlo después si disminuyen los precios, si aumentan sus ingresos, etc.

El modelo permite utilizar los datos de las HES para estimar elasticidades precio creíbles de la demanda, partiendo del supuesto de que los precios de la mayoría de los productos en los PIMB varían significativamente en el espacio geográfico. Esta variación espacial del precio es el resultado ya sea de los costos considerables de transporte debido al traslado de los productos de un lugar a otro, u otros factores tales como diferentes impuestos fronterizos o aranceles en diferentes jurisdicciones del mismo país. Por lo tanto, el costo de transportación u otros factores que afectan los cambios de precios entre regiones geográficas sirven implícitamente como un instrumento y es el principal factor que influye en el precio, lo que a su vez influye en la demanda. Por lo tanto, para identificar las elasticidades precio en este modelo se supone que existe una variación genuina en los precios entre conglomerados.

El supuesto de que los precios varían espacialmente significa que los hogares que viven cerca unos de otros, como los que viven en la misma "localidad" o "bloque urbano", deberían enfrentar el mismo precio, ya que realizan compras en el mismo mercado y al mismo tiempo si se trata de una encuesta transversal. Por otro lado, los hogares que viven lejos, como los que viven en diferentes localidades o bloques urbanos, deberían enfrentar precios diferentes. En otras palabras, el enfoque requiere que gran parte de la variación observada en el precio ocurra entre conglomerados, como se menciona en el Capítulo 2, y no dentro de los conglomerados. Desde el punto de vista econométrico, esto requiere que la variación de precios se explique en gran medida por los "efectos de conglomerado" o "variables *dummy* por conglomerado". Cualquier variación en el precio dentro del conglomerado debe ser el resultado de un error de medición, cuyos patrones pueden ser utilizados para corregir las estimaciones finales de dicho error (más en la Sección 3.2.3).

Otra contribución significativa fue que, si bien los hogares no informan sobre el precio de mercado en la encuesta, puede deducirse de sus decisiones de compra calculando la relación entre el gasto del hogar en un producto y la cantidad del producto. Sin embargo, esta relación es un valor unitario y no un precio. Los valores unitarios no son lo mismo que los precios debido a los dos problemas siguientes. 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 calidad"). Si no se trata adecuadamente, esto podría dar lugar a lo que se conoce como "dilución de la calidad", que se refiere a una situación en la que un cambio de precio no lleva a una reducción en la cantidad de la demanda, ya que la gente cambia a productos más baratos, pero de menor calidad. En segundo lugar, los valores unitarios no son lo mismo que los precios debido a un error de medición, dado que a menudo la gente da información errónea sobre los gastos y/o las cantidades de los productos comprados. Deaton propone fórmulas para tratar tanto con la dilución de calidad como con el error de medición. La siguiente sección ofrece una explicación técnica paso a paso del método propuesto originalmente por Deaton en 1988, que se ha ampliado en sus trabajos posteriores.[7,30–32]

3.2.3 Marco teórico del Modelo Deaton

Esta sección describe brevemente los principales pasos para derivar el modelo teórico propuesto por Deaton para estimar las elasticidades de precios utilizando datos de las HES. Se aconseja a los investigadores que planean implementar este modelo que lean el Capítulo 5 de Deaton (1997)[7] para entender los detalles más sutiles del modelo. El modelo consiste principalmente de seis pasos, desde la derivación de los valores unitarios, pasando por las pruebas pertinentes, hasta la estimación de las elasticidades del precio y del gasto.

Paso 1: Derivación de valores unitarios

En primer lugar, los valores unitarios se derivan de los datos de la encuesta a nivel de hogares. Esto se hace dividiendo el gasto total reportado en el producto o productos de tabaco en particular sobre los cuales la HES proporciona datos por su cantidad correspondiente, como:

$$\upsilon_{hc} = \frac{x_{hc}}{q_{hc}} \qquad (3.1)$$

dónde υ_{hc}, x_{hc} y q_{hc} son, respectivamente, el valor unitario, el gasto y la cantidad de cigarrillos o de cualquier otro producto de tabaco en el hogar *h* localizado 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. Esto se hace utilizando el Análisis de Varianza (ANOVA, por su acrónimo en inglés) para dividir la variación total en valores unitarios en "variaciones dentro de un mismo conglomerado" y "variaciones entre conglomerados". Un *F-statistic (estadístico-F)* significativamente grande para el ejercicio ANOVA lleva a la conclusión de que los valores unitarios varían según el espacio geográfico o los conglomerados.

Paso 3: Estimación de regresiones dentro del conglomerado

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

$$lnv_{hc} = \alpha^{1} + \beta^{1} lnx_{ic} + \gamma^{1} Z_{hc} + \psi ln\pi_{c} + u^{1}_{hc} \qquad (3.2)$$
$$w_{hc} = \alpha^{o} + \beta^{o} lnx_{ic} + \gamma^{o} Z_{hc} + \theta ln\pi_{c} + (f_{c} + u^{o}_{hc}) \qquad (3.3)$$

lnv_{hc} es el logaritmo del valor unitario, derivado según la ecuación 3.1 para el hogar *h* en el conglomerado *c*, mientras que w_{hc} representa la parte del gasto en tabaco en el gasto familiar total de los hogares *h* en el conglomerado *c* y lnx_{hc} es el logaritmo del gasto familiar total de los hogares durante el período de referencia correspondiente. Z_{hc} es un vector de características específicas del hogar que puede incluir variables sobre la estructura del hogar (por ejemplo, el tamaño del hogar, proporción de adultos, proporción de hombres, etc.) y la demografía del hogar (por ejemplo, edad, sexo, estado civil, escolaridad y situación laboral de la o del jefe de del hogar, etc.). f_c es un efecto fijo de conglomerado y tratado 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. Ambas u^0_{hc} y u^1_{hc}, incorporan cualquier error de medición en las participaciones en el presupuesto y los valores unitarios, aparte de los inobservables habituales. La ecuación del valor unitario no contiene ningún efecto fijo de la localidad porque, como observa Deaton,[7] "condicionados a los precios, los valores unitarios dependen únicamente de los efectos de la calidad y de los errores de medición. La introducción de un efecto fijo adicional rompería el vínculo entre los precios y los valores unitarios, impediría que los últimos proporcionaran 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, 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 los precios fueran observables, se habrían reducido en esta etapa de la regresión debido a la falta de variación.

La ecuación 3.2, denominada “valor unitario”, nos permite comprobar la presencia de los efectos de calidad, como se explica en la Sección 3.2.2. Una relación positiva y estadísticamente significativa entre el gasto del hogar y los valores unitarios, después de tener en cuenta las características de los hogares, sugeriría la presencia de los efectos de calidad. Conocer el patrón de los efectos de calidad (es decir, la magnitud de β^1), permite corregir las estimaciones finales de elasticidad precio para la dilución de calidad en el Paso 6. Note que la ecuación 3.2, a diferencia de la ecuación 3.3, se estima sin los efectos fijos del conglomerado. Agregar un efecto fijo a nivel de conglomerado a la ecuación 3.2 dificultaría la recuperación de los parámetros del modelo.

Por otra parte, la ecuación 3.3 es una ecuación de demanda estándar en la que la participación de cigarrillos (como sustituta de la demanda) se expresa en función de los ingresos de los hogares (sustituidos por el gasto del hogar), sus características 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, estos se sustituyen por los efectos fijos del conglomerado. La relación entre los dos errores, u^0_{hc} y u^1_{hc}, (como lo refleja, por ejemplo, la covarianza) es útil para corregir las estimaciones finales de la elasticidad precio para el error de medición, como se explica en el Paso 5.

Paso 4: Obtención de la demanda y los valores unitarios a nivel de conglomerados

El cuarto paso consiste en despojar la demanda a nivel de hogares y a los valores unitarios de los efectos del gasto del hogar y de las características de los hogares y, después, promediar entre conglomerados. El despojo y la premediación se realizan porque el interés principal es estimar la elasticidad a nivel de conglomerados utilizando la demanda de estos y su valor unitario despojados de todos los demás factores. Este paso requiere las siguientes ecuaciones:

$$\widehat{y^1_c} = \frac{1}{n^+_c} \sum_{h=1}^{n^+_c} (lnv_{hc} - \hat{\beta}^1 lnx_{hc} - \hat{\gamma} Z_{hc}) \qquad (3.4)$$

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

donde n_c es el número de hogares en el conglomerado c y n^+_c es el número de hogares que informan de la compra del producto de tabaco para el que estamos estimando la elasticidad. Observe que $\widehat{y^1_c}$ y $\widehat{y^0_c}$ no tienen el subíndice h porque representan promedios de conglomerados. $\widehat{y^1_c}$ y $\widehat{y^0_c}$ son las estimaciones, respectivamente, del valor unitario medio de los conglomerados y de la demanda media de los conglomerados después de eliminar los efectos del gasto del hogar y las características de los hogares. En otras palabras, las ecuaciones 3.4 y 3.5 pueden expresarse alternativamente como $y^1_c = \alpha^1 + \psi ln\pi_c + u^1_c$ y $y^0_c = \alpha^0 + \theta ln\pi_c + f_c + u^0_c$, respectivamente.

Paso 5: Regresiones a nivel de conglomerados

Recuerde que nuestra suposición identificadora es que los precios varían entre conglomerados y no dentro de ellos. Dada esta situación, las elasticidades de precios de la demanda solo se pueden obtener al ver cómo la demanda a nivel de conglomerado responde a los cambios en los precios a nivel de conglomerado. Por lo tanto, el Paso 5 implica la regresión de la demanda a nivel de conglomerado, $\widehat{y^0_c}$, en valores unitarios también a nivel de conglomerado, $\widehat{y^1_c}$. El coeficiente en $\widehat{y^1_c}$ en tal regresión puede obtenerse alternativamente dividiendo la covarianza entre $\widehat{y^0_c}$ y $\widehat{y^1_c}$ por la variación de $\widehat{y^1_c}$. Es decir $\hat{\phi}$, la estimación del coeficiente de y^1_c, se obtiene por:

$$\hat{\phi} = \frac{Cov(\widehat{y^0_c}, \widehat{y^1_c}) - \frac{\widehat{\sigma^{10}}}{n_c}}{Var(\widehat{y^1_c}) - \frac{\widehat{\sigma^{11}}}{n^+_c}} \qquad (3.6)$$

donde n_c^+ es el número de hogares en un conglomerado que informan de 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 de errores estándar en variables donde la covarianza y la varianza de errores se utiliza para corregir el error de medición. Note que los factores de corrección para el error de medición se vuelven pequeños a medida que n_c^+ y n_c se vuelven grandes.

Paso 6: Estimación de las elasticidades precio y gasto

El sexto y último paso del método de Deaton aplica fórmulas de corrección de calidad para obtener la estimación de la elasticidad precio de la demanda, $\widehat{\varepsilon_p}$, según se indica a continuación:

$$\widehat{\varepsilon_p} = \left(\frac{\hat{\theta}}{\overline{w}}\right) - \hat{\psi} \qquad (3.7)$$

donde $\overline{w}$ es la proporción media del gasto del hogar total dedicado a cigarrillos en la muestra. $\hat{\psi}$ y $\hat{\theta}$, las estimaciones de los coeficientes sobre los términos de precios no observados en las ecuaciones (3.2) y (3.3), respectivamente, se recuperan de la siguiente manera:

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

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

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

Finalmente, Deaton también propone la siguiente fórmula para obtener la estimación de la elasticidad del gasto de la demanda, $\widehat{\varepsilon_1}$:

$$\widehat{\varepsilon_1} = 1 + \left(\frac{\hat{\beta}^0}{\overline{w}}\right) - \hat{\beta}^1 \qquad (3.11)$$

dónde $\hat{\beta}^1$ es la estimación del coeficiente del gasto del hogar total en la ecuación 3.2, y $\hat{\beta}^0$ es la estimación del coeficiente sobre el gasto del hogar total en la ecuación 3.3. $\hat{\phi}$ es la estimación del coeficiente de una regresión de la demanda sobre el valor unitario, ambas a nivel de conglomerado (de la ecuación 3.6). Una vez recuperados los parámetros del 3.8 al 3.10, se puede estimar la elasticidad precio de la demanda en función de la ecuación 3.7. Por otra parte, la elasticidad del gasto de la demanda solo utiliza coeficientes de la primera etapa y puede obtenerse utilizando 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 tienen que obtenerse con *bootstrapping.*

Varios estudios han utilizado el método de Deaton para estimar las elasticidades precio y gasto de la demanda de diversos productos de tabaco en diferentes PIMB. Estos incluyen estudios en la India,[33-37] Vietnam,[38] China,[39] Uganda[40] y Ecuador,[41] entre otros. Algunos estimaron la elasticidad para un solo producto, los cigarrillos, mientras que otros estimaron las elasticidades precio propias y cruzadas para los cigarrillos y algunos otros productos de tabaco. Cabe señalar que, si bien algunos de estos estudios consideraron todos los hogares en la regresión de la participación en el presupuesto para estimar la elasticidad, otros consideraron solo los hogares con compras positivas, por lo que estimaron solo una

demanda condicional. Sin embargo, como lo señala Deaton,[7] para fines de reformas fiscales y de precios, es necesario incluir todos los hogares en el análisis, independientemente de que compren o no. Las estimaciones de la elasticidad de los precios propios de los cigarrillos en estos estudios oscilaron entre -0.1 y -0.6, mientras que las estimaciones de la elasticidad de los gastos oscilaron entre 0.2 y 2.4. En otras palabras, estos estudios tienden a encontrar estimaciones de elasticidad de precios para los cigarrillos comparables a las estimadas en los estudios internacionales que usan otros métodos. También tienden a encontrar elasticidades no negativas del gasto de la demanda de cigarrillos, lo que implica que la demanda de cigarrillos no disminuye con un aumento en el gasto.

También es interesante observar que en estos estudios la definición utilizada de conglomerado varía. Mientras que algunos consideraban una localidad o un 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 de tiempo[42] si, por ejemplo, hay HES de múltiples rondas u olas. Es importante entender que las propiedades de consistencia de los parámetros en el modelo de Deaton dependen del número de conglomerados (y no del número de hogares), ya que estos parámetros se derivan de los datos promedio a nivel de conglomerados. Por otro lado, los errores de medición en las ecuaciones 3.2 y 3.3 tienden a ser cero solo a medida que aumenta el número de hogares en cada conglomerado. Claramente, hay una compensación. Por un lado, los conglomerados 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 los consumen unos cuantos hogares. Con conglomerados más pequeños, también es posible que algunos de ellos no tengan ningún hogar con compras de tabaco positivas en lo absoluto. Por otra parte, dado que la segunda etapa de regresión y estimación de la elasticidad precio depende de que haya un gran número de conglomerados con compras positivas, es importante contar con el mayor número posible de conglomerados a fin de obtener estimaciones de parámetros consistentes.

Los experimentos de Deaton han demostrado que el estimador funciona adecuadamente incluso cuando hay tan solo dos hogares en cada conglomerado.[7] Según Deaton, “aumentar el número de localidades o conglomerados es mucho más importante que aumentar el número de observaciones en cada uno”. Esto se debe a dos razones: (1) el modelo corrige los errores de medición, pero no puede garantizar la consistencia de los parámetros con un número pequeño de conglomerados; y (2) si los conglomerados se definen o agregan en áreas geográficas más extensas, es posible que los hogares dentro de dichos conglomerados no enfrenten el mismo mercado y, como resultado, puede haber verdaderas variaciones intrarregionales en los valores unitarios dentro de dichos conglomerados que pueden ser tratados inadvertidamente como errores de medición. Para que los supuestos del modelo se mantengan, los hogares en un conglomerado dado deben tener proximidad geográfica y tener las entrevistas más o menos al mismo tiempo. Esto puede resultar aún más difícil a medida que se amplíen los conglomerados para incluir regiones geográficas más extensas.

Para la mayoría de las HES, los conglomerados se presentan naturalmente como parte del diseño de la encuesta, como ya se indicó en el Capítulo 2. También es importante señalar que el modelo se basa en la existencia de una variación genuina de los precios entre conglomerados y requiere que dicha variación sea exógena al proceso que determina la demanda. Como observa Deaton,[7] “si los precios locales se determinan por los precios mundiales, los impuestos fronterizos y los costos de transporte, las suposiciones se cumplirán pues la demanda local no tiene efecto sobre los precios”. Por otro lado, si los precios de la localidad dependen de la demanda dentro de la misma localidad, las estimaciones de los parámetros no serán consistentes, por las razones de simultaneidad habitual.

Cabe señalar que, aunque la discusión anterior se refiere a los hogares, el análisis también puede llevarse a cabo a nivel individual. Sin embargo, esto requiere que el investigador tenga acceso a una rica encuesta de gastos recopilada a nivel individual. Por ejemplo, una encuesta de este tipo debería contener

información sobre los patrones de gasto (cantidad e importe total gastado) y sobre los productos de tabaco por parte de los individuos (no agregados a nivel de los hogares, como suele ser el caso). Además, también deberían estar presentes otros datos sociales y demográficos a nivel individual. Si bien estos conjuntos de datos están ampliamente disponibles en los países de ingresos altos, tienden a ser la excepción en los PIMB. Se alienta a los investigadores que tienen acceso a las encuestas de gastos recopiladas a nivel individual a que utilicen el método de Deaton para estimar las elasticidades de la demanda.

El método de Deaton no está exento de críticas. Gibson y Rozelle (2005)[43] muestran que la utilización de valores unitarios como sustituto de los precios reales da lugar a estimaciones sesgadas de la elasticidad de la demanda en función de los precios, incluso después de haber corregido los efectos de la calidad y los errores de medición. Mckelvey (2011)[42] muestra que el método de Deaton no trata adecuadamente el tema de la dilución de calidad que parece prevalecer en muchos entornos. A pesar de estas limitaciones, y a falta de datos de precios a gran detalle, el método de Deaton sigue siendo uno de los métodos más efectivos para obtener elasticidades.

3.3 Preparación de datos para el análisis

Si bien el Capítulo 2 proporcionó información detallada sobre la extracción de datos, su limpieza, la fusión de diferentes conjuntos de datos y otros consejos necesarios para la gestión de datos, es importante proporcionar detalles específicos sobre las variables necesarias para la estimación de la elasticidad del precio utilizando el método de Deaton discutido anteriormente. Para cualquier nueva variable que se discuta aquí, es importante analizarla a través de todos los procesos discutidos en el Capítulo 2. Esta sección discute cómo las variables específicas requeridas para la estimación de la elasticidad precio y la elasticidad cruzada usando el método de Deaton pueden generarse usando las variables estándar disponibles de las HES.

Las variables más importantes son la cantidad de consumo, así como el gasto en diferentes productos de tabaco. Estos están disponibles directamente en la mayoría de las HES. Es posible que algunas HES no reporten información de cantidad como se mencionó anteriormente. En tales casos, la discusión aquí puede no ser beneficiosa.

En primer lugar, deben crearse valores unitarios para cada uno de los productos de tabaco de los que se dispone de datos. Esto puede incluir valores unitarios para cigarrillos, bidis y productos sin humo, entre otros. Por ejemplo, la cantidad de cigarrillos (en paquetes o individuales) tal como se descarga de los datos de la HES, tiene el nombre de variable *qcig* y la variable que representa el gasto en cigarrillos es *expcig*. Entonces, el valor unitario de los cigarrillos (*uvcig*) se puede generar utilizando el comando <*gen uvcig=expcig/qcig*>. El modelo de Deaton utiliza el logaritmo natural de la variable de valor unitario como la variable dependiente (*luvcig*). Utilice el comando <*gen luvcig=ln(uvcig)*> para generarlo. Del mismo modo, una variable para representar las participaciones presupuestarias dedicadas a los cigarrillos (*bscig*) utilizando el comando <*gen bscig = expcig/exptotal*> donde *exptotal* es el gasto total en todos los elementos y debe construirse. Para aquellos hogares sin gasto reportado en cigarrillos, esto generaría un valor faltante. En este caso, se debería usar el comando <*replace bscig=0 if bscig==.*> para indicar una participación cero en el presupuesto de cigarrillos para los hogares que no gastan en cigarrillos, en lugar de dejar fuera a todos los hogares al utilizar un valor faltante. Esto se debe a que, como lo sugiere Deaton,[7] es útil incluir a todos los hogares en el análisis para los efectos de la reforma fiscal y de precios, independientemente de que compren cigarrillos o no. Al implementar el modelo de Deaton, *uvcig* y *bscig* serían las variables dependientes en las regresiones respectivas. Se deben generar variables similares de valor unitario y participación en el presupuesto para otros productos de tabaco de las HES que se incluirán en la estimación de la elasticidad precio.

El precio es definitivamente una variable independiente en un modelo que estima las 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 los precios. En cambio, la variación de precios se captura a través de las variaciones de precios a nivel de conglomerados en la HES. Por lo tanto, es crucial tener una variable que identifique a los conglomerados (*clust*) o a las unidades primarias de muestreo. Esta variable suele estar disponible directamente desde la HES o se puede generar utilizando otras variables disponibles que identifiquen a las unidades primarias de muestreo, como se discutió en el Capítulo 2. El conglomerado puede ser una unidad geográfica (localidad o unidades primarias de muestreo en encuesta transversal) como en el análisis original de Deaton, o puede ser un punto en el tiempo (por ejemplo, alguna ola de la encuesta) si se combinan diferentes rondas de encuestas o una combinación de PSU y olas de la encuesta.[42]

Además, es necesario identificar variables específicas a nivel de los hogares para utilizarlas como variables independientes en el modelo. Los estudios ofrecen orientación sobre algunas de las variables sociodemográficas comunes a nivel de los hogares: logaritmo del tamaño del hogar; proporción de hombres (relación entre el número de hombres y el tamaño del hogar); edad media del hogar; educación media del hogar (educación total recibida por todos los miembros en años divididos por el tamaño del hogar); educación máxima (años de educación recibida por el miembro con mayor educación del hogar); nivel educativo del jefe del hogar; variables *dummy* para caracterizar hogares en diferentes grupos sociales, étnicos, ocupacionales, religiosos y de ingresos; y variables *dummy* para indicar la ubicación del hogar (áreas rurales/urbanas, provincia, distrito, etc.), entre otras.

3.4 Estimación de la elasticidad precio 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 que se discutió anteriormente. Deaton proporciona un código detallado de Stata para estimar elasticidades cruzadas para diferentes productos y se puede descargar de http://web.worldbank.org/archive/website00002/WEB/EX5_1-2.HTM. El Apéndice de código en la Sección 7.2 reproduce el código de Deaton del sitio web del Banco Mundial con algunas explicaciones adicionales para que los lectores puedan seguirlo. El código que se utiliza en esta sección para estimar la elasticidad precio de los cigarrillos producirá estimaciones idénticas a la elasticidad obtenida en el código de Deaton para el caso de múltiples productos del Apéndice 7.2, que se utilizó para estimar la elasticidad de un solo producto. Mientras que el código para casos de múltiples productos hace uso de matrices para calcular varios parámetros en el modelo, el código aquí usa solo escalares ya que se trata de un solo producto. Además, como el código para múltiples productos también estima las elasticidades cruzadas y permite la introducción de otras restricciones teóricas sobre el sistema de demanda, como se discutió en Deaton,[7] el código simplemente estima la elasticidad precio de los cigarrillos sin imponer ninguna otra restricción. El código de esta sección utiliza las variables *bscig, luvcig, lexp, lhsize, maleratio, meanedu, maxedu, sgp1, sgp2, sgp3* para la estimación de las elasticidades precio.

Pruebas de variación espacial en valores unitarios

Como se indica en la sección del método, es útil estimar la variación de los valores unitarios entre conglomerados para evaluar si estas variaciones son indicativas de la variación de los precios entre conglomerados. Esto se puede hacer usando el comando <*anova luvcig clust*> o <*regress luvcig i.clust*>. El R^2 y el *estadístico-F* del resultado pueden indicar la utilidad de los valores unitarios como información de los precios. Según Deaton,[7] un valor significativo del *estadístico-F* y un R^2 en torno a 0.5 (es decir, las variables *dummy* 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 los precios y estimar las elasticidades de los precios.

Estimación de las regresiones de primera etapa dentro del conglomerado y de las varianzas de los errores de medición

Aquí abajo se estiman las ecuaciones 3.2 y 3.3 y se almacenan los parámetros relevantes para las etapas siguientes:

```
#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;

* Repetir para participaciones en el presupuesto:
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
```

Se utiliza el comando <*areg*> en lugar de <*regress*> ya que se usa para una regresión lineal con un conjunto de variables *dummy* extenso. El comando incluye implícitamente una variable *dummy* para cada conglomerado menos uno y, sin embargo, no enlista el coeficiente asociado con estas variables dummy por conglomerado en el resultado de la regresión. La opción <*absorb(clust)*> junto con el comando <*areg*> le dice a Stata que use las variables *dummy* por conglomerado implícitas para la variable del conglomerado *clust*. Las variables *y1cig* y *y0cig*, después de cada regresión, son variables depuradas de cualquier efecto de las características específicas de los hogares que sean la razón de la variación de calidad de los valores unitarios. Estas variables conservan ahora la información de precios contenida en las variables *dummy* por conglomerado. Los residuos del valor unitario (*ruvcig*) y de la regresión de la participación en el presupuesto (*rbscig*) se generan para utilizarse 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 las regresiones del valor unitario y de la participación en el presupuesto, son estimaciones de la varianza y covarianza de los errores de medición que se usarán para la corrección de errores de medición en la ecuación 3.6. También se almacena el coeficiente para el logaritmo del gasto para usarse más tarde. El escalar *b1*, que es el coeficiente del logaritmo del gasto en la regresión del valor unitario, es el estimador de la elasticidad de la calidad. Cuanto menor sea este número, menor será la dilución de calidad en valores unitarios.

Estimación de las elasticidades de los ingresos y el gasto

La elasticidad total del gasto (o elasticidad ingreso) en la ecuación 3.11 puede estimarse después de estas regresiones de primera etapa utilizando los resultados guardados. Esto 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 media del presupuesto en un escalar (*Wbar*) y utiliza los otros escalares guardados (*b1 y b0*) de las regresiones de primera etapa para estimar la elasticidad del gasto (*Expel*). La última línea imprimirá la elasticidad del gasto en la ventana de resultados de Stata.

Preparación de datos para la regresión entre conglomerados

El siguiente paso consiste en promediar las variables *y1cig* y *y0cig* por conglomerados para generar *y1c* y *y0c* respectivamente, de modo que se puedan utilizar para una regresión entre conglomerados de *y0c* sobre *y1c* para derivar la elasticidad precio. Como se mencionó anteriormente, las variables *y1cig* y *y0cig* se depuran de cualquier característica específica de los hogares de las regresiones de valor unitario y de la participación en el presupuesto y contienen solo la información de precios en las variables *dummy* por conglomerado, 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 medio para todos los hogares en cada conglomerado, solo se necesita mantener una observación por conglomerado para el análisis restante. Junto con la generación de las variables de nivel de conglomerado y0c en y1c, se generan otras dos variables de nivel de conglomerado n0c y n1c que indican el tamaño o el número de todos los hogares en cada conglomerado (n0c) y el número de hogares que reportan compras positivas en cada conglomerado (n1c). Utilizando estos datos, se estima el tamaño medio de los conglomerados para todos los hogares (n0) y para aquellos hogares con un consumo positivo de cigarrillo (n1). Esto se puede hacer usando el siguiente código. Deaton utiliza una media armónica para estimar los tamaños promedio de estos 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 de la relación $\phi=\frac{\theta}{\psi}$ cuyo numerador y denominador son los coeficientes de precios no observados en las ecuaciones 3.3 y 3.2,

respectivamente. En lugar de hacer la regresión real, se puede simplemente estimar este parámetro híbrido usando un estimador de errores en la variable en la ecuación 3.6 para la cual se usan las estimaciones para *y1* y *y0* así como las varianzas y covarianzas del error de medición estimadas a partir de las regresiones de primera etapa. 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=scalar(R)-(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 ha estimado, como en la ecuación 3.6, unos cuantos escalares más se deben definir 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 la 3.10, no necesariamente en el mismo orden. Con el fin de estimar los errores estándar para las estimaciones de elasticidad precio, las ecuaciones anteriores deben entrar en un programa que utilice 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=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
```

La última línea de código devuelve los errores estándar *bootstrapped* para las estimaciones de elasticidad precio. La Sección 7.1 del Apéndice de código incluye un archivo .do de ejemplo que detalla el código utilizado en esta sección. Los usuarios pueden copiar y pegar ese código en el editor de archivos .do de Stata y estimar los resultados con los datos/variables correspondientes que se describen ahí mismo. Además, la Sección 7.2 reproduce el código detallado de Deaton para estimar las elasticidades cruzadas utilizando el método de Deaton.

3.5 Caso práctico de Uganda

Esta sección presenta resultados de un estudio realizado en Uganda con un proceso paso a paso que al final conduce a estimaciones de elasticidades. El estudio utilizó datos de las ediciones de 2005 y 2009 de la UNPS (*Uganda National Panel Survey,* Encuesta de Panel Nacional de Uganda) y utilizó solamente hogares que reportaron consumo positivo de cigarrillos en el análisis. La Oficina de Estadística de Uganda dirige la UNPS con la asistencia del Banco Mundial. Los datos se pueden descargar fácilmente del sitio web de la Encuesta de Medición de los Niveles de Vida del Banco Mundial (http://microdata.worldbank.org/index.php/catalog/lsms). A continuación, se presentan los resultados paso a paso para facilitar la comprensión de la técnica. Además, las ediciones de 2005 y 2009 de la UNPS se tratan como secciones transversales separadas.

Paso 1: Derivación de valores unitarios y otras variables relevantes

El primer paso en el método de Deaton es derivar los valores unitarios según la ecuación 3.1. Segundo, otras variables utilizadas en el análisis fueron procesadas como se describe en el Capítulo 2. La lista completa de variables que se utilizan para estimar las elasticidades en Uganda se presenta en la Tabla 3.1 a continuación. Las variables de las líneas 5 a la 11 de la Tabla 3.1 constituyen el vector de Z_{ic} de las variables de estructura del hogar y de control demográfico descritas en las ecuaciones anteriores 3.2 y 3.3.

Tabla 3.1 Variables utilizadas para el cálculo de la elasticidad precio de 2005 y 2009 en la UNPS

	Variable
1	Participación promedio del gasto en cigarrillos en el gasto total del hogar.
2	Logaritmo natural del valor unitario.
3	Logaritmo natural del gasto del hogar.
5	Logaritmo natural del tamaño del hogar.
6	Logaritmo natural de años de escolaridad del jefe de hogar.
7	Logaritmo natural de la edad del jefe de hogar.
8	Proporción de hombres en el hogar.
9	Proporción de adultos en el hogar.
10	Variable *dummy* para determinar si el jefe de hogar trabaja.
11	Variable *dummy* para determinar si el jefe de hogar es un hombre.

Notas: Variables relevantes de las ediciones de 2005 y 2009 de la UNPS

Paso 2: Hipótesis 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 se presentan en la Tabla 3.2. a continuación.

Tabla 3.2 Prueba de variación espacial del logaritmo de los valores unitarios

Muestra de 2005				Muestra de 2009			
Estadístico-F	valor-p	R cuadrado	n	Estadístico-F	valor-p	R cuadrado	n
1.29	0.08	0.70	274	1.12	0.33	0.72	173

Notas: El *estadístico-F* y el *valor-p* se asocian con la hipótesis nula de que no hay variación espacial en los valores unitarios. La hipótesis se rechaza en la muestra de 2005, pero no en la de 2009. El *R cuadrado* mide la proporción de variación en los precios que tiene lugar entre conglomerados. *n* es el número total de hogares.

El resultado del ANOVA muestra que al menos el 70 % (*R cuadrado* de 0.70) de la variación de los valores unitarios, se explica por los efectos entre conglomerados. El *estadístico-F* está asociado con la hipótesis de que no hay variación espacial en los precios, lo que se rechaza en la muestra de 2005, pero no se rechaza en la muestra de 2009.

Paso 3: Regresiones dentro del conglomerado

El siguiente paso es estimar las regresiones dentro del conglomerado, es decir, la regresión del valor unitario y las de la participación en el presupuesto, según las ecuaciones anteriores 3.2 y 3.3. Los resultados de estas regresiones se presentan en las Tablas 3.3 y 3.4.

Los resultados de la regresión del valor unitario de la Tabla 3.3 muestran que los valores unitarios reportados están correlacionados positivamente con el gasto del hogar. Este resultado es estadísticamente significativo al nivel del 5 % para ambos años de la encuesta. Esto es indicativo de la presencia de efectos de calidad en los datos según el análisis de la Sección 3.2.3. Los resultados de la regresión de la participación en el presupuesto en la Tabla 3.4 muestran que la participación de los cigarrillos en el presupuesto disminuye con el gasto del hogar. Este resultado es estadísticamente significativo al nivel del 1 % para ambos años de la encuesta.

Paso 4 y Paso 5:

El Paso 4 implica la obtención del 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 al mismo nivel según la ecuación 3.6. Estos resultados no se reportan aquí.

Paso 6: Obtención de estimaciones de elasticidad

El paso final es aplicar las fórmulas de las ecuaciones 3.7 a la 3.11 para obtener estimaciones de elasticidad precio y gasto. La Tabla 3.5 presenta estimaciones de la elasticidad precio de la demanda de cigarrillos en Uganda. La Tabla 3.6 presenta estimaciones de la elasticidad gasto de la demanda.

Tabla 3.3 Resultados de la regresión del valor unitario

Variables	2005 lnv	2009 lnv
Lnx	0.234***	0.115**
	(0.051)	(0.048)
Tamaño	-0.042	-0.010
	(0.124)	(0.119)
Adultos	-0.203	0.159
	(0.295)	(0.300)
Hombres	0.261	0.131
	(0.216)	(0.223)
Educación	-0.143*	0.108
	(0.080)	(0.074)
Edad	-0.015	-0.409**
	(0.153)	(0.166)
Género	0.217	0.218
	(0.163)	(0.183)
Trabajo	-0.144	0.101
	(0.141)	(0.118)
Constante	4.957***	6.602***
	(0.692)	(0.739)
Número de hogares	233	147
R cuadrado	0.115	0.126

Notas: Resultados de la regresión del logaritmo del valor unitario (*lnv*) en el logaritmo del gasto del hogar (lnx) y otras características de los hogares. El tamaño del hogar (Tamaño), la educación del jefe de hogar (Educación) y la edad del jefe de hogar (Edad) están en logaritmos naturales. Por Adultos se refiere a la proporción de adultos en un hogar y se refiere a aquellos de 18 años de edad o más. Hombres es la proporción de hombres en el hogar. Género es una variable *dummy* que toma el valor de 1 si el jefe de hogar es hombre y cero si es mujer. Trabajo es una variable *dummy* que toma el valor de 1 si el jefe de hogar está empleado y cero en caso contrario. Errores estándar entre paréntesis. *** p<0.01, ** p<0.05, * p<0.1.

Tabla 3.4 Resultados de la regresión de la participación en el presupuesto

Variables	2005 w	2009 w
Lnx	-0.056***	-0.065***
	(0.017)	(0.023)
Tamaño	0.002	0.039
	(0.031)	(0.043)
Adultos	0.008	0.092
	(0.072)	(0.103)
Hombres	0.013	0.010
	(0.059)	(0.068)
Educación	-0.001	-0.012
	(0.020)	(0.025)
Edad	0.028	-0.077
	(0.044)	(0.072)
Género	-0.038	-0.108*
	(0.037)	(0.056)
Trabajo	0.037	0.058
	(0.037)	(0.039)
Constante	0.533***	0.963***
	(0.193)	(0.292)
Número de hogares	233	147
R cuadrado	0.866	0.909

Notas: Resultados de la regresión de la participación del gasto en cigarrillos en el presupuesto (*w*) en el logaritmo del gasto del hogar (lnx) y otras características de los hogares. El tamaño del hogar (Tamaño), la educación del jefe de hogar (Educación) y la edad del jefe de hogar (Edad) están en logaritmos naturales. Por Adultos se refiere a la proporción de adultos en un hogar y se refiere a aquellos de 18 años de edad o más. Hombres es la proporción de hombres en el hogar. Género es una variable *dummy* que toma el valor de 1 si el jefe de hogar es hombre y cero si es mujer. Trabajo es una variable *dummy* que toma el valor de 1 si el jefe de hogar está empleado y cero en caso contrario. Errores estándar entre paréntesis. *** p<0.01, ** p<0.05, * p<0.1. Los efectos fijos del conglomerado se suprimen por razones de espacio, pero son estadísticamente significativos en conjunto al 1 % del nivel para las muestras de 2005 y las muestras agrupadas y al 10 % para la muestra de 2009.

Tabla 3.5 Estimaciones de la elasticidad precio de la demanda de cigarrillos en Uganda

	2005	2009
$\widehat{\varepsilon}_P$	-0.326*** [0.021]	-0.258*** [0.011]
	(-0.368 , -0.284)	(-0.280 , -0.235)
Número de hogares	233	147
Número de conglomerados	184	130

Notas: Estimaciones de la elasticidad precio de la demanda de cigarrillos en Uganda. Los errores estándar *bootstrapped* están entre corchetes. Los intervalos de confianza de 95 % están entre paréntesis. *** p<0.01, ** p<0.05, * p<0.1.

Tabla 3.6 Estimaciones de la elasticidad del gasto de la demanda en cigarrillos en Uganda

	2005	2009
$\widehat{\varepsilon}_I$	0.132 [0.338]	0.043 [0.539]
	(-0.531 , 0.796)	(-1.014 , 1.100)
Número de hogares	233	147

Notas: Estimaciones de la elasticidad del gasto de la demanda de cigarrillos en Uganda para las muestras de 2005 y 2009. Los errores estándar *bootstrapped* están entre corchetes. Los intervalos de confianza del 95 % están entre paréntesis. Dado que la elasticidad gasto de la demanda se estima a nivel de los hogares (véase la ecuación 3.11), solo se informa sobre el número de estos.

Los resultados de la Tabla 3.5 muestran que se espera que la demanda de cigarrillos en Uganda disminuya alrededor de un 0.3 % cada vez que los precios de los cigarrillos aumenten un 1 %. Estas estimaciones son estadísticamente significativas en el nivel de relevancia del 1 % y están dentro del rango de estimaciones obtenidas en los estudios que utilizan el método de Deaton discutido en la Sección 3.2.3. La Tabla 3.6 presenta resultados de la elasticidad gasto de la demanda para 2005 y 2009. Dado que las estimaciones de la elasticidad gasto no se estiman con precisión (es decir, los errores estándar son grandes), es difícil extraer conclusiones sólidas. Como mínimo, los resultados de la Tabla 3.6 sugieren que la demanda de cigarrillos no disminuye con un aumento del gasto del hogar.

3.6 Estimación de elasticidades cuando los valores unitarios no están disponibles en las HES

El enfoque de Deaton nos permite estimar la demanda y calcular la elasticidad precio y cruzada utilizando cantidades y valores unitarios obtenidos a partir de datos de las HES. Sin embargo, a veces los datos de las HES solo recogen información sobre los gastos en que incurren los hogares para diferentes grupos de productos. No proporcionan información sobre las cantidades compradas y, por lo tanto, no podemos construir valores unitarios cuya variación espacial pueda utilizarse como información sobre la variabilidad de los precios a nivel de los hogares. En este caso, el enfoque de Deaton, tal como se discute en este capítulo, no se puede aplicar. Dado que las HES proporcionan información rica sobre el consumo de los

hogares junto con consumo de productos de tabaco, no sería prudente ignorar tales datos simplemente porque no hay información disponible sobre la cantidad. Afortunadamente, existen métodos para recuperar valores unitarios (o pseudovalores unitarios) de modo que los mismos puedan utilizarse para la estimación de las funciones de demanda y para derivar la elasticidad precio.

Tradicionalmente, cuando la información cuantitativa no está disponible en las HES, las fuentes externas de variabilidad de precios obtenidas a partir de los índices de precios nacionales agregados, como los Índices de Precios al Consumidor (IPC), a menudo se fusionan con el gasto del hogar para obtener estimaciones de las elasticidades de los precios.[44] Con frecuencia se emplean sistemas de demanda conocidos como el AIDS o el Sistema Cuadrático de Demanda Casi Ideal (QAIDS, por sus siglas en inglés) cuando se utilizan dichos índices de precios para estimar las funciones de demanda. Sin embargo, este enfoque es criticado por no tener en cuenta la variabilidad espacial y de los hogares, lo que da lugar a estimaciones distorsionadas de los parámetros de la demanda y no ser coherentes con la teoría.[45-48] Además, los índices de precios agregados suelen estar muy correlacionados y pueden presentar problemas de endogeneidad.[49]

Sin embargo, estudios recientes[50] sugieren que la construcción de índices de precios a nivel de los hogares (precios Stone-Lewbel [SL][51]) para grupos de productos puede mitigar los problemas relacionados con utilizar únicamente índices de precios agregados en situaciones en las que no se dispone de información cuantitativa en la encuesta. Los índices de precios SL para grupos de productos se construyen utilizando información sobre la participación en el presupuesto del subgrupo, las características demográficas de los hogares y los índices de precios nacionales agregados, y permite recuperar los precios o valores unitarios a nivel de los hogares.[50] Se comprobó que la utilización de los precios SL específicos de los hogares resulta en parámetros de demanda más precisos y económicamente viables que los obtenidos al utilizar solamente índices de precios agregados.[48] El programa escrito por usuarios en Stata, *<pseudounit>*,[44] ayuda a estimar tales valores unitarios (valores pseudounitarios) utilizando este método para las HES sin información cuantitativa.

Un sistema implícito de demanda *marshalliana*, propuesto recientemente, el EASI (*Exact Affine Stone Index*, Índice exacto afín de Stone) utiliza estos métodos para estimar la elasticidad precio[50,52] y tiene varias ventajas sobre los sistemas de demanda tradicionales, como el AIDS. En los estudios también se dispone de diferentes métodos empíricos para el cálculo del índice de precios SL para productos agregados.[53] Sin embargo, este Conjunto de herramientas no aborda estas cuestiones ni los acontecimientos que las rodean, ya que, en la mayoría de los casos, los datos de las HES proporcionan tanto la cantidad como el gasto de los diferentes productos de interés. Sin embargo, los lectores que tienen datos de las HES sin información cuantitativa deberían familiarizarse con los estudios en esta sección antes de intentar estimar la elasticidad precio a partir de dichos datos.

4 Estimación del efecto de 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 tabaquismo ha disminuido del 23.5 % en 2007 al 20.7 % en 2015, gran parte de esa disminución se ha producido en los países de ingresos altos, mientras que la disminución ha sido menor en los países de ingresos bajos.[54] La mayoría (alrededor del 77 %) de los aproximadamente 1 100 millones de fumadores actuales del mundo, viven en PIMB.[21] La prevalencia del consumo de tabaco sin humo también es mucho mayor en los PIMB (14.6 %) y en los países de ingresos bajos (11.2 %) en comparación con la prevalencia mundial (6.5 %).[4] Varios estudios también han demostrado que el consumo de tabaco es desproporcionadamente mayor entre las personas relativamente pobres. Un metaanálisis de 201 estudios de la OMS encontró una asociación estadísticamente significativa de una mayor prevalencia del tabaquismo actual entre adultos y un menor ingreso, tanto en hombres como mujeres.[55]

El gasto en tabaco representa una parte significativa del presupuesto de los hogares en muchos países, desde el 1 % en países como México y Hong Kong, hasta el 10 % en países como Zimbabue y China.[56] Los hogares funcionan basados en ingresos disponibles limitados y, como resultado, cuando gastan su limitado presupuesto en tabaco, tiene un enorme costo de oportunidad. Esto significaría inevitablemente que tendrían que reducir los gastos en otros bienes y servicios, algunos de los cuales podrían ser artículos de consumo necesarios, 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 tempranos de explicar la cuestión del desplazamiento con un análisis descriptivo de los datos de Bangladesh[57] y China[58] en los años 2001 y 2002, respectivamente. Un examen empírico formal de la idea del desplazamiento debido al gasto en tabaco mediante métodos econométricos surgió más tarde de los EE. UU.[59] y China[60] en los años 2004 y 2006. Sin embargo, estos estudios no pudieron modelar explícitamente el tema de la endogeneidad presente en dicho análisis. La generación actual de métodos econométricos que estiman el impacto del desplazamiento del gasto en tabaco comenzó en 2008 utilizando datos sobre el gasto del hogar de la India.[56] Utilizó técnicas de IV para tener en cuenta la posible endogeneidad del sistema de demanda, al tiempo que trataba el gasto en tabaco como un regresor, y se descubrió que el gasto en tabaco desplazaba a los alimentos, la educación y el entretenimiento, a la vez que desplazaba el gasto en salud, ropa y combustibles. En otros países como Taiwán,[61] Sudáfrica,[62] Camboya,[63] Zambia,[64] Turquía[65] y Bangladesh,[56] se realizaron estudios similares usando métodos econométricos y datos sobre el gasto del hogar parecidos. También hubo otros estudios que examinaron el desplazamiento en Indonesia[67] y otros PIMB,[68] pero con métodos ligeramente diferentes.

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

Año	Autores	País	Método	Datos de la Encuesta utilizada	Artículos desplazados
2004	Busch *et al.*[59]	Estados Unidos	Regresiones MCO separadas	Encuesta sobre el gasto de los consumidores	Ropa, vivienda
2006	Wang *et al.*[60]	China	Modelo Logit fraccionado	Encuesta principal	Educación, mantenimiento de equipos agrícolas, ahorros
2008	John, RM[56]	India	Variables instrumentales	Encuesta nacional por muestreo	Comida, educación, entretenimiento.
2008	Pu *et al.*[61]	Taiwán	Variables instrumentales	Encuesta de Ingresos y gastos de los hogares	Ropa, atención médica, transporte
2008	Koch & Tshiswaka-Kashalala[62]	Sudáfrica	Variables instrumentales	Encuesta de Ingresos y Gastos de Sudáfrica	Educación, combustible, ropa, atención médica y transporte
2009	Block & Webb[67]	Indonesia	Ecuaciones de forma reducida	Datos del sistema de vigilancia nutricional.	Comida
2012	John *et al.*[63]	Camboya	Variables instrumentales	Encuesta Socio-Económica de Camboya	Comida, educación, ropa
2014	Chelwa & Walbeek[64]	Zambia	Variables instrumentales	Condiciones de vida Encuesta de seguimiento	Alimentación, escolarización, ropa, transporte, mantenimiento del equipo
2015	San & Chaloupka[65]	Turquía	Variables instrumentales	Encuesta de presupuesto de Hogares Turcos	Alimentos, vivienda, educación, bienes duraderos/no duraderos
2015	Do & Bautista[68]	40 PIMB	Modelos de pendiente aleatoria	Encuesta Mundial de Salud	Educación, atención médica
2018	Husain *et al.*[66]	Bangladesh	Variables instrumentales	Encuesta sobre ingresos y gastos de los hogares	Ropa, vivienda, educación, energía, transporte y comunicación
2018	Paraje & Araya	Chile	Modelo cuadrático AIDS	Encuesta de Presupuestos Familiares de Chile (EPF)	Atención médica, educación, vivienda

En la Tabla 4.1 se resumen 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 IV es el método preferido que es adoptado por la mayoría de los estudios de los últimos 10 años. La mayoría de estos estudios encuentran que el gasto en tabaco desplaza los gastos en artículos necesarios para el consumo de los hogares, como alimentos, ropa, vivienda y educación, entre otros, lo que implica que el gasto en tabaco puede tener impactos en el desarrollo y entre las generaciones.

4.2 Importancia de la asignación de recursos dentro del hogar

Los hogares a menudo juntan los recursos de 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 individualmente. En la mayoría, si no es que en todas las HES, el hogar es la unidad por la cual se reporta el consumo. Sin embargo, no se informa cómo se distribuye el consumo entre los miembros de la familia. Si las decisiones de asignación son tomadas por ciertos miembros adultos en un hogar, a menudo hombres en varios PIMB, la forma en que pueden impactar el bienestar social es incierta. Como señala Deaton,[7] si las mujeres reciben sistemáticamente menos que los hombres, o si los niños y los ancianos están sistemáticamente en peor situación que otros miembros del hogar, se sobreestimará el bienestar social al utilizar medidas que supongan que todos los miembros del hogar reciben el mismo trato.

Las decisiones de asignación de recursos dentro de los hogares adquieren mayor importancia cuando se reducen los ingresos disponibles una vez que se asigna dinero para 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,[54] si las decisiones de asignación las toman los jefes de hogar, podría ser potencialmente desfavorable para las mujeres y/o los niños de un hogar. De hecho, algunos de los hallazgos de los estudios de desplazamiento descritos anteriormente lo subrayan. Cuando los gastos escolares o educativos se ven comprometidos como resultado de una mayor asignación al consumo de tabaco, esto tiene un impacto directo en los niños del hogar y en su potencial de ingresos futuros, a la vez que impone impactos intergeneracionales a largo plazo en la sociedad. Los estudios de la India,[56] por ejemplo, muestran que los hogares que gastan en tabaco asignan sistemáticamente menos dinero en combustibles limpios para cocinar y más dinero en fuentes de combustible contaminantes, 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 cierta parte del presupuesto para la compra de tabaco. Significa que el hogar tiene que maximizar su utilidad asignando de manera óptima el presupuesto restante (el total menos el presupuesto preasignado para el tabaco) entre los productos alternativos. Ciertamente, como el presupuesto disponible se reduce después de la preasignación, hay que llegar a algunas concesiones. Si se llega a la conclusión de que se hacen concesiones en el caso de productos básicos necesarios como alimentos, educación y ropa, que pueden afectar directamente a la salud y el desarrollo de todos los miembros de un hogar, las políticas de control del tabaco deben ser capaces de abordarlas.

4.3 Comparación de la participación media en el presupuesto

La comprobación de las diferencias en la media del presupuesto o en la media del gasto dedicado a los diferentes grupos de productos entre los hogares que gastan en tabaco y los que no lo hacen proporciona un indicio preliminar de las posibles concesiones, si las hubiere, derivadas del gasto en tabaco. En esta sección se examinan estas diferencias dividiendo a los hogares en diferentes grupos en función de sus hábitos de gasto en tabaco y comparando la participación del presupuesto que cada grupo asigna a la compra de diferentes grupos de productos.

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 1 si los hogares gastan cualquier cantidad positiva de dinero en tabaco y 0 si no es así. Como ejemplo, *exptobac* es la variable que representa la cantidad gastada en tabaco por un hogar según se extrajo de las HES. Después, se puede generar la variable indicadora *tobacco*, y sus valores se pueden etiquetar con los siguientes comandos:

```
gen tob=0
replace tob=1 if 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, *tabaco, alimentos, atención médica, educación, vivienda, ropa, entretenimiento, transporte, bienes duraderos* y *otros*, que agotan el presupuesto del hogar. La mayoría de los estudios sobre el desplazamiento han considerado algunos o todos ellos para su análisis. Las variables que representan los gastos en estos productos son *exptobac, expfood, exphealth, expeducn, exphousing, expcloths, expentertmnt, exptransport, expdurable* y *expother,* respectivamente, según se extrajo de los datos de las HES. Tenga en cuenta que todas las variables tienen el mismo prefijo *exp*. Esta forma de nombrarlas hace que el análisis sea más sencillo. Para comparar la media de las participaciones del presupuesto dedicadas a estos productos entre los consumidores de tabaco y los no consumidores, se define una variable de participación del presupuesto; una para cada uno de este grupo de productos. Dados los gastos totales en todos los artículos juntos como *exptotal*, la participación del presupuesto en cada una de las variables 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 las cuotas de presupuesto con el prefijo (bs_) para todos estos productos.

Paso 2: Comprobar si la diferencia en la media de la participación del presupuesto es estadísticamente significativa

Una prueba estadística de la igualdad de la media de la participación presupuestaría entre dos grupos (consumidores y no consumidores de tabaco) es una *prueba t* de Student para la media de igualdad de dos muestras. La *prueba t* que 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. Esto comparará la cuota del presupuesto que dedican a alimentos los hogares que gastan y los que no gastan en tabaco y comprobará si la diferencia es estadísticamente relevante. La hipótesis nula es que la diferencia en la media de la cuota de presupuesto es = 0. También se informa el *estadístico t* para la diferencia en la media. Por regla general, si el valor absoluto de t es mayor que 2, se rechaza el valor nulo y se puede concluir que la diferencia en la media de la cuota del presupuesto observada es estadísticamente relevante.

La *prueba t*, sin embargo, no permite el uso de encuestas ponderadas. Tampoco permite el uso del comando <*svy*> de Stata. Como resultado, el promedio de las cuotas de presupuesto calculadas para los consumidores y no consumidores de tabaco bajo el comando <*ttest*> pueden llegar a ser sesgadas. Lo ideal sería calcular la participación del presupuesto para ambos grupos después de equilibrarlo con las 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. En este caso, la prueba t se puede realizar de la siguiente manera:

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

Aquí, *weight* es la variable para la ponderación de la encuesta. El comando <*lincom*> informa de la diferencia en la media de la participación de presupuesto ponderadas 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 = 0. Este método producirá estimaciones idénticas a las de la *prueba t* si no se utilizara la ponderación. En lugar de usar la ponderación en el comando anterior, también se puede usar el comando <*svy: mean bs_food, over(tob)*> después de declarar el diseño de la encuesta. También se puede usar el comando <*test [bs_food]0 - [bs_food]1*> que realiza una *prueba Wald*, en lugar de la *prueba t* que se realiza con <*lincom*>. Dado que se están calculando la media de las participaciones de presupuesto de las HES, se debería utilizar una opción de la prueba que permita utilizar la ponderación o el prefijo svy en lugar de utilizar una *prueba t* directa que no permita las ponderaciones en absoluto.

Paso 3: Informe de resultados de la prueba

Para la elaboración de informes, solo es necesario conocer la media de la participación en el presupuesto para los grupos de productos dados, la diferencia en la media de estas y la relevancia estadística de la diferencia, tal y como se indica en el valor del *estadístico t*. A continuación, se ofrece un programa para los diez grupos de productos:

```
#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/`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 [bs_ `X']0 - [bs_ `X']1;
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;
```

El código anterior enlistará una tabla con la participación en el presupuesto para los que no gastan, los que si gastan, la diferencia y el *estadístico t* para la prueba de igualdad de la media de la participación en el presupuesto entre los que gastan y los que no gastan en tabaco para cada uno de los grupos de productos en la macro local *items.*

4.4 Marco para el análisis empírico del desplazamiento

La simple *prueba t* de igualdad de la media, como se discutió en la sección anterior, no controla otras características específicas del hogar que pueden influir en las decisiones de asignación del presupuesto y, al no controlarlas, se puede estar atribuyendo inadvertidamente las decisiones de asignación a los hábitos de gasto en tabaco de un hogar. Por esta razón, es necesario un modelo econométrico formal que pueda explicar si los hogares que gastaron en tabaco redujeron sistemáticamente sus gastos en otros grupos de productos y, en caso afirmativo, en cuáles. Esta sección describe el enfoque conceptual y econométrico que se sigue en la mayor parte de los estudios actuales para estimar el grado de desplazamiento debido al gasto en tabaco. Además, la sección discute algunas mejoras metodológicas en los estudios existentes sobre este tema.

4.4.1 Marco teórico para analizar el desplazamiento

La teoría microeconómica nos enseña que la solución a la maximización de la utilidad de un individuo sujeta a una restricción presupuestaria devuelve un conjunto de funciones de demanda *marshalliana* incondicionales de la forma:

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

donde q_i es la cantidad de productos consumidos, Y es el gasto total, h es un vector de características y $p_1,\ldots,p_n$ son los precios de n productos en la función de utilidad de un individuo. Dado que los gastos del hogar se reportan para todo el hogar como una sola unidad, se utiliza una función de demanda a nivel de hogar y se necesita la suposición de que el hogar busca maximizar una función de utilidad única. Si la demanda de un hogar de uno de los productos, por ejemplo, el tabaco, está predeterminada, entonces existen funciones de demanda condicional. El marco teórico para ello se detalla en Pollak (1969).[8] La idea es que el hogar maximice la siguiente función de utilidad:

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

Donde $\bar{q}_n$ denota la demanda de tabaco de un hogar y $M = Y - p_n * \bar{q}_n$. Resolviendo esto para los productos *n-1* se obtiene la siguiente función de demanda condicional, que es condicional al consumo del *n-ésimo* producto (tabaco en nuestro caso):

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

La función de demanda de cualquier producto determinado (q_i) aquí es condicional a los precios de todos los productos excepto el producto condicionante (q_n), el total de los gastos restantes (M) después de deducir los gastos en el producto condicional, la cantidad de dicho producto ($\bar{q}_n$), y un vector de las características de los hogares ($\boldsymbol{h}$).

Cuando se trata de productos que no se consumen por muchos hogares (por ejemplo, el tabaco), es ventajoso utilizar funciones de demanda condicional, tal como lo señalan Browning y Meghir.[69]

4.4.2 Modelo econométrico para analizar el desplazamiento

Esta sección se analiza una ecuación econométrica específica que se estima para examinar el impacto del desplazamiento y un breve resumen de los posibles métodos de estimación que se han utilizado hasta ahora en los estudios, junto con sus deficiencias. Luego propone un método de estimación alternativo que es más eficiente y teóricamente preferible.

4.4.2.1 Especificación del modelo econométrico

La implementación empírica del modelo requiere el uso de una forma funcional específica. Los estudios sobre el desplazamiento han utilizado en gran medida el QAIDS[70] para estimar el impacto del desplazamiento. Dado que las encuestas de hogares no suelen disponer de información directa sobre los precios de los distintos grupos de productos, se utilizan curvas de Engel, que permiten trabajar con los gastos, para la especificación econométrica. Con la presencia de un término cuadrático de ingresos, el QAIDS, además de ser consistente con la teoría de la utilidad, permite que los productos sean un lujo en algunos niveles de ingresos y necesarios en otros.[56] La curva condicional de Engel toma la siguiente forma para el producto *i* y el hogar *j:*

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

donde $w_{ij} = p_{ij} q_{ij}/M_j$ es la participación en el presupuesto asignada por el hogar *j* al grupo de productos *i* del presupuesto restante después de deducir los gastos en tabaco (M_j), $p_{nj}\bar{q}_{nj}$ es el gasto en tabaco, $\mathbf{h}_j$ es un vector de las características de los hogares que permite que las preferencias sean heterogéneas,[71] *lnM* y lnM^2 son los logaritmos naturales de *M* y M^2 que representa al gasto después de deducir el gasto en tabaco, y u_{ij} es el término de error aleatorio.

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

El modelo que se especifica en la ecuación 4.4 no puede estimarse con el método MCO ya que las variables $p_n\bar{q}_n$ y *lnM* son probablemente endógenas debido a la simultaneidad involucrada. Si este es el caso, estas variables se correlacionarán con el término de error u_{ij} y podrían resultar en estimaciones sesgadas e inconsistentes de MCO. En otras palabras, una suposición fundamental de MCO es que el término de error del modelo no está correlacionado con los regresores, es decir, $E(u|\mathbf{x}) = 0$, y las estimaciones MCO de la encuesta fallan al dar una interpretación causal. En tales casos, si se pueden encontrar variables exógenas que estén correlacionadas con estos regresores endógenos, pero que no están correlacionadas con el término de error (IV), se podría utilizar el método IV para estimar los parámetros de manera más consistente. Esto también se conoce como una estimación de mínimos cuadrados de dos etapas (2SLS).

El estimador IV, sin embargo, es menos eficiente que el MCO y debe usarse solo si hay variables endógenas presentes en el modelo. Si los errores son homocedásticos, esto se puede probar con la prueba de *exogeneidad* de *Durbin-Wu-Hausman* (DWH)[72]. Si los errores son heterocedásticos, se suelen utilizar diferentes pruebas, como la *prueba de puntuación de Wooldridge*, una prueba basada en regresión auxiliar o una prueba de *C-statistic (estadístico C)*, dependiendo del tipo de heterocedasticidad que se asuma.[73] Todos los estudios de la generación actual de literatura sobre el desplazamiento muestran que estas variables son realmente endógenas.

La estimación por IV proporciona un estimador consistente bajo la fuerte suposición de que existe un instrumento válido ***z*** que satisface dos condiciones: (1) El instrumento ***z*** está parcialmente correlacionado con los regresores endógenos ***x***, es decir, $Cov(\boldsymbol{x}_i,\boldsymbol{z}_i)\neq 0$; y (2) El instrumento ***z*** afecta a la variable dependiente w_i únicamente a través de los regresores o el ***z*** en sí mismo no causa $\boldsymbol{w}_i$, es decir, $E(u/\boldsymbol{z})=0$. La primera condición a veces se llama restricción de inclusión, mientras que la segunda condición se conoce popularmente como restricción de exclusión. Mientras que la restricción de inclusión pueda ser probada estadísticamente comprobando 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 sobreidentificado (es decir, cuando hay más instrumentos que el número de regresores endógenos), se puede hacer una prueba de sobreidentificación de las restricciones para probar la exogeneidad de los instrumentos, siempre y cuando los parámetros del modelo se estimen usando de manera óptima el Método de Momentos Generalizados (GMM, por sus siglas en inglés).[15] Esta prueba también difiere dependiendo de si los errores son homocedásticos o no. Si los errores son homocedásticos, se debe realizar una *prueba de Sargan*. Si no, se usa *el estadístico J de Hansen* o el *estadístico Hansen-Sargan*. Si el estadístico de la prueba es estadísticamente significativo, indica que los instrumentos pueden no ser válidos; esto puede suceder si los instrumentos no son verdaderamente exógenos, o porque están siendo excluidos incorrectamente de la regresión.[73]

Incluso si existen instrumentos válidos y coeficientes consistentes con las estimaciones, su matriz de covarianza puede ser inconsistente si los errores son heterocedásticos.[73] El *estadístico Pagan-Hall* puede usarse para probar la presencia de heterocedasticidad en la regresión de IV. Bajo la hipótesis nula de la homocedasticidad, el *estadístico Pagan-Hall* se distribuye como χ^2, independientemente de la presencia de heterocedasticidad en otras partes del sistema.[73] Un estadístico significativo implicará la presencia de heterocedasticidad. Si este es el caso, se deberá utilizar un error estándar consistente a heterocedasticidad mientras se emplea una estimación ecuación por ecuación con IV. Las estimaciones de los coeficientes, así como sus errores estándar, entonces serán consistentes. Esto se puede hacer a través de una estimación 2SLS o GMM, a la que Wooldridge[14] se refiere como un "estimador de sistema 2SLS" y que es más eficiente que el simple estimador IV[73] en presencia de heterocedasticidad.

4.4.2.3 Método de estimación 2: Estimación de la variable instrumental del sistema (3SLS)

Para estimar un sistema de curvas de Engel, una para cada grupo de productos, para determinar dónde y cómo se produce el desplazamiento, deberían haber tantas ecuaciones como el número de grupos de productos considerados. Cada una de estas ecuaciones tendría el gasto en tabaco como un producto 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, por sus siglas en inglés) con la adición del método de IV, que es efectivamente un método de mínimos cuadrados en tres etapas (3SLS, por sus siglas en inglés).[74] Bajo el supuesto de que los errores son homocedásticos, el 3SLS proporciona una estimación más eficiente en comparación con el 2SLS+IV al explotar la correlación de ecuaciones cruzadas de errores.[15] Los estudios han usado consistentemente este método en oposición al uso de IV en el SUR. Una buena descripción de la estimación del sistema 3SLS, que también se llama 3SLS tradicional, se puede encontrar en Wooldridge,[14] Capítulo 8.

4.4.2.4 Método de estimación 3: Estimación de 3SLS por GMM

El estimador tradicional 3SLS, según Wooldridge,[14] es menos eficiente y su estimador de varianza es inapropiado si los errores son heterocedásticos. En las encuestas transversales, en el Capítulo 2, la

heterocedasticidad es la norma y no la excepción. Un estimador del sistema que es consistente y más eficiente que el estimador tradicional 3SLS en presencia de heterocedasticidad es un estimador de GMM, y Wooldridge[14] lo llama el estimador "GMM 3SLS". Este extiende el estimador tradicional 3SLS al permitir heterocedasticidad y diferentes instrumentos para diferentes ecuaciones.[75] La estimación de GMM permite seleccionar diferentes matrices de ponderación para obtener estimadores que pueden tolerar la heterocedasticidad, la conglomeración, la autocorrelación y otras violaciones clásicas del término de error *u*. El 3SLS tradicional, por ejemplo, es un estimador de GMM que utiliza una matriz de ponderación particular, que asume errores i.i.d.[14] Sin embargo, al igual que los estimadores IV/3SLS, el estimador GMM también puede tener propiedades de muestra limitada deficientes.[73]

Según Wooldridge,[14] el estimador GMM 3SLS utilizando la matriz de ponderación consistente de heterocedasticidad nunca es peor, asintóticamente, que el 3SLS tradicional, y en algunos casos importantes es estrictamente mejor. Los estudios anteriores sobre el desplazamiento, sin embargo, parecen haber ignorado una prueba de heterocedasticidad en el modelo 3SLS que han usado y estimado el modelo 3SLS tradicional asumiendo que los errores son i.i.d. Esto puede haber producido estimaciones de parámetros menos eficientes si la heterocedasticidad estaba realmente presente en esos modelos.

4.4.2.5 Comprobación de la heterogeneidad de las preferencias entre los consumidores y los no consumidores de tabaco

Típicamente, en los datos de las HES, se ve un gran número de ceros o valores faltantes contra el gasto en tabaco. Esto puede deberse a que los precios del tabaco son actualmente inasequibles para algunos de los hogares debido a las limitaciones de su presupuesto (también conocido como *solución de esquina*), o a la abstención (es decir, el tabaco no está en la función de utilidad de un hogar o en su canasta básica, sin importar cuál sea el precio). En este último caso, los consumidores y no consumidores de tabaco tienen preferencias fundamentalmente heterogéneas. Teóricamente no hay ninguna razón *a priori* por la que se deba asumir cualquiera de los dos casos. Sin embargo, junto con la estimación del desplazamiento, si también se desea tener en cuenta la heterogeneidad de las preferencias entre los hogares que gastan y los que no gastan, la ecuación 4.4 puede aumentarse con la adición de una variable binaria que indique la situación del consumo de tabaco, como se indica en algunos estudios[56,65,76] de la siguiente manera:

$$w_{ij} = (\alpha_{1i} + \alpha_{2i} d_j + \alpha_{3ij} p_{nj} \overline{q}_{nj} + \delta_i' \boldsymbol{h}_j) + (\beta_{1i} + \beta_{2i} d_j) ln M_j + (\gamma_{1i} + \gamma_{2i} d_j)(ln\, M_j)^2 + u_{ij} \qquad (4.5)$$

donde *d* es un indicador binario que toma el valor 1 si un hogar gasta en tabaco y 0 si no.

Si los parámetros asociados a la variable binaria *d* no son conjuntamente significativos, es decir, si la hipótesis nula H_0:$\alpha_{2i} = \beta_{2i} = \gamma_{2i} = 0$ no se rechaza, se puede concluir que los hogares contra los que actualmente se informa de un gasto cero en tabaco no están gastando en tabaco, probablemente porque actualmente no es asequible para ellos. En otras palabras, tanto los que gastan en tabaco como los que no lo hacen tienen funciones de utilidad similares y los que no lo hacen actualmente no gastan en tabaco solo porque su precio es inasequible. Pero, si se rechaza la nulidad, significa que los coeficientes asociados a la variable *dummy* de tabaco y a las variables de gasto con las que interactúa la esta variable *dummy* de tabaco, son significativas y que las preferencias son realmente diferentes para los consumidores de tabaco y los no consumidores. Los estudios sobre esto, después de la regresión, usan una *prueba de Wald* para probar la significancia conjunta de los tres parámetros.

Si el investigador tiene interés en probar esta hipótesis, se debe especificar en primer lugar la ecuación 4.5, en vez de la ecuación 4.4. Si la hipótesis H_0:$\alpha_{2i} = \beta_{2i} = \gamma_{2i} = 0$ se rechaza, entonces la especificación de la ecuación 4.5 debería usarse para estimar el desplazamiento. En ese caso, los coeficientes asociados a las variables serán diferentes tanto para los consumidores de tabaco como para los no consumidores. En otras palabras, las preferencias son realmente heterogéneas entre los hogares que gastan y los que no gastan y que los que no gastan en tabaco no lo tienen en su función de utilidad, independientemente de cuál sea su precio. Si, por otro lado, la hipótesis no se rechaza, se puede proceder con la especificación de la ecuación 4.4, en cuyo caso tanto los hogares que gastan en tabaco como los que no lo hacen tendrán las mismas estimaciones de parámetros. No hay razón para hablar de desplazamiento del gasto en tabaco en el caso de aquellos hogares para los que el tabaco no forma parte de su función de utilidad ni de su canasta básica, independientemente de cuál sea su precio.

4.4.3 Limitaciones del modelo

La discusión de los diferentes métodos para estimar el desplazamiento en la Sección 4.4.2 supone la disponibilidad de IV adecuadas para abordar la endogeneidad presente en la especificación del modelo. Sin embargo, encontrar una IV adecuada que cumpla con los requisitos econométricos necesarios a menudo puede ser difícil y, a veces, es posible que no se pueda encontrar en absoluto. En efecto, existen estudios que estiman el desplazamiento ignorando tal endogeneidad,[59,60,67,77] a menudo debido a la no disponibilidad de IV adecuadas. Sin embargo, las regresiones que ignoran la presencia de variables endógenas pueden dar lugar a estimaciones de parámetros que llevan a una inferencia errónea. En tales casos, se pueden adoptar métodos menos sofisticados. Uno de estos métodos es una simple comparación de las cuotas del presupuesto entre los que gastan y los que no gastan en distintos artículos de compra utilizando una prueba t, como ya se ha descrito en la Sección 4.3. También se puede comparar el gasto absoluto asignado a diferentes artículos entre ambos grupos de hogares. En lugar de una prueba t, también se podrían realizar otras herramientas de comparación descriptivas o gráficas para comparar los promedios.

Dado que el análisis de desplazamiento explicado anteriormente compara la participación del presupuesto en diferentes productos únicamente consumidores de tabaco y los hogares no consumidores, no arroja mucha luz sobre las asignaciones intrahogares como resultado del desplazamiento. Esta es otra limitación de este análisis. Por ejemplo, el análisis puede mostrar que el gasto en salud o educación se ve desplazado como resultado del gasto en tabaco. Sin embargo, es difícil determinar qué miembro de la familia se ve afectado a causa de este desplazamiento. El hecho de que en el análisis sólo se tengan en cuenta grupos de productos agregados más grandes dificulta aún más el examen de esas consideraciones dentro de los hogares. Por otra parte, herramientas menos sofisticadas como una *prueba t,* discutida en la Sección 4.3, permite la comparación directa de las partes del presupuesto o de los gastos entre los que gastan y los que no gastan para cualquier artículo desagregado. De hecho, se pueden recoger solamente artículos de interés y comparar los patrones de gasto entre ambos grupos. En un estudio realizado en la India,[56] se utilizó una *prueba t* sencilla para comparar la participación en el presupuesto asignadas a los gastos de los autobuses escolares y la participación en el presupuesto sobre los diferentes tipos de combustible para cocinar entre personas que gastan en tabaco y personas que no. Se encontró que las personas que gastan en tabaco gastaban menos en el autobús escolar (lo que implica que los niños en el hogar se ven directamente afectados). También se encontró que los hogares que gastaban en tabaco gastaban menos en combustible limpio para cocinar y más en combustible impuro como leña (lo que implica que la salud de las mujeres en estos hogares probablemente se vea afectada).

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 vienen de diferentes conjuntos de datos y otros consejos necesarios para la gestión de datos, es importante proporcionar detalles específicos sobre las variables necesarias para el análisis en este capítulo. Es importante analizar cualquier nueva variable que se discuta aquí, a través de todos los procesos discutidos en el Capítulo 2. Esta sección discute cómo las variables específicas requeridas para el análisis del desplazamiento se pueden generar usando las variables estándar disponibles de las HES. También muestra formas de clasificar los hogares para que se ajusten a las necesidades analíticas específicas de este capítulo.

Las variables más importantes que se requieren son los gastos en tabaco, así como otros grupos de productos mencionados anteriormente, que necesitan probarse para determinar si se produce un desplazamiento. Estas están disponibles directamente en la mayoría de las HES. A continuación, debe calcularse la participación de cada uno de los grupos de productos del presupuesto restante después de haber deducido los gastos en tabaco. Por ejemplo, se puede crear una variable para la participación de alimentos en el presupuesto en Stata usando el código <*generate bsfood = expfood/exp_less*> donde *bsfood* es la variable de la participación de los alimentos que se utilizará como variable dependiente en la regresión, *expfood* es el gasto en alimentos que se extrae de las HES y *exp_less* es el gasto total en todos los artículos (*exptotal*) menos el gasto en tabaco (*exptobac*). Para todos los grupos de productos juntos, se puede utilizar un bucle para generar las cuotas de presupuesto como se indica a continuación:

```
#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) ;
    } ;
```

Estas son las variables que entrarían en la regresión (IV, 3SLS o GMM 3SLS) como variables dependientes. Esto es diferente de las variables de la participación del presupuesto creadas en la Sección 4.3 para la *prueba t*, ya que esta tenía el gasto total como denominador. Aunque los gastos en diferentes productos están disponibles directamente de las HES, es posible que los datos de las HES no reporten estos datos al nivel de agregación requerido. Por ejemplo, los gastos en alimentos pueden registrarse en las HES como gastos en varios otros artículos alimenticios. Si no hay información agregada disponible, es posible que haya que agregar los gastos en artículos más pequeños para crear grupos agregados como los que se enlistan a continuación. Al analizar el efecto de desplazamiento del gasto en tabaco, tener demasiados productos desagregados puede no servir de mucho después de todo, desde el punto de vista de las políticas públicas. Sin embargo, dependiendo de las circunstancias socioeconómicas de cada país, la selección de los grupos de productos podría variar.

Es necesario generar logaritmos naturales y cuadrados de las variables *exptotal* y *exp_less* para utilizarse en la regresión. También es necesario identificar las variables específicas a nivel de hogar que se utilizarán como controles y las variables que típicamente pueden funcionar como instrumentos para las variables endógenas en el modelo 3SLS. Los estudios realizados ofrecen orientación. Algunas de las variables sociodemográficas comunes a nivel de los hogares que se utilizan en este estudio incluyen un logaritmo del tamaño del hogar; proporción de adultos (relación entre el número de adultos y el tamaño del hogar); edad media del hogar; educación media del hogar (educación total recibida por todos los miembros en años

divididos por el tamaño del hogar); educación máxima (años de educación recibida por el miembro con mayor educación del hogar); variables *dummy* para caracterizar hogares en diferentes grupos sociales, étnicos, ocupacionales, religiosos y de ingresos; y variables *dummy* para indicar la ubicación del hogar ya sea como áreas rurales o urbanas, entre otras.

La elección de las variables correctas para que sirvan como instrumentos es uno de los aspectos clave de la preparación de la lista de variables para el análisis. De nuevo, los estudios ofrecen orientación. Gran parte de los estudios recientes sobre el desplazamiento[56,61,64–66] utilizan el gasto total del hogar o el valor total de los bienes de los hogares como un instrumento para grupo del gasto *M (exp_less)* y la proporción de hombres o mujeres adultos en el número total de adultos en el hogar (proporción del sexo de los adultos) o la proporción de hombres adultos a mujeres adultas como instrumento para el gasto en tabaco. Se considera que la proporción del sexo de los adultos 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 estos países. Por lo tanto, se espera que un aumento en la proporción de hombres (proporción de hombres adultos con respecto a las mujeres adultas) esté positivamente relacionado con el gasto en tabaco, y no es algo que pueda tener un impacto directo en la cuota del presupuesto de otros grupos de productos para los que se estima el impacto del desplazamiento. Un estudio[62] utiliza una medida compuesta de prevalencia del tabaquismo como instrumento para el gasto en tabaco. De hecho, cualquier variable exógena que aparezca en el RHS (*righ-hand side*, lado derecho) de las otras ecuaciones en el modelo puede servir potencialmente como un instrumento para la variable endógena del lado derecho en la ecuación que se estimará. Independientemente de la variable que se utilice como instrumento, es importante comprobar que los instrumentos seleccionados están correlacionados con la variable del RHS endógena y no tienen un efecto directo sobre la variable dependiente.

4.6 Estimando el desplazamiento con Stata

Esta sección mostrará los diferentes métodos de estimación (3SLS tradicional, GMM 3SLS y una IV de ecuación por ecuación) que se discutieron en la Sección 4.4 para estimar el desplazamiento. En primer lugar, discutirá la configuración general de las variables que pueden utilizarse bajo todos los métodos. Después de una discusión de la implementación de los tres métodos de estimación, se discutirá la prueba de varios requerimientos del modelo incluyendo la validez de los instrumentos y la heterocedasticidad, entre otros. Los resultados de estas pruebas guiarán la decisión sobre el tipo de método de estimación que ha de utilizar.

Como se detalló anteriormente, dependiendo de las propiedades de los datos, existen diferentes estrategias de modelación. A continuación, se muestran 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 calcular las estimaciones tradicionales 3SLS o GMM 3SLS o IV, es útil crear ciertas macros globales que indiquen la lista de variables dependientes, endógenas, exógenas e instrumentos en el modelo. Por ejemplo, se definen las siguientes macros para estimar el impacto del desplazamiento entre ocho grupos de productos de alimentos, salud, educación, vivienda,

ropa, entretenimiento, transporte y bienes duraderos, dejando fuera al grupo de productos “otros” como se hace comúnmente en los estudios:

```
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 RHS como se explica en la ecuación 4.4 (estas son variables que se sospecha son endógenas), *x2list* incluye las variables exógenas (tamaño del hogar, educación media, educación máxima, tres variables *dummy* para representar el nivel SES de los hogares), y *zlist* incluye las IV para corregir endogeneidad en el modelo (proporción del sexo de los adultos, logaritmo de los gastos totales y logaritmo de los gastos totales al cuadrado en este caso). Sin embargo, en el modelo cada variable exógena puede ser un instrumento propio. El número de variables en *zlist* debe ser al menos tan grande como las de *x1list* para que el modelo sea identificado. Las variables utilizadas en las macros globales aquí solo están para fines demostrativos. En el análisis real puede haber un número menor o mayor de variables en cualquiera de las listas anteriores. Por ejemplo, la *x2list* puede contener varias otras características específicas del hogar que las que se enlistan aquí.

4.6.1 Estimación de 3SLS

Una vez creadas estas macros globales, la estimación del modelo 3SLS en Stata puede hacerse simplemente usando el comando <*reg3*>. Con <*help reg3*> se obtiene ayuda de Stata sobre el comando reg3, proporciona la sintaxis detallada y ejemplos útiles para usarlo. Pero, para este propósito, una vez que las macros globales están definidas como aquí arriba, solo se necesita usar el siguiente comando para obtener las estimaciones de 3SLS:

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

Las opciones *exog* y *endog* especifican la lista de regresores exógenos y endógenos en el RHS de cada una de las ecuaciones. Sin el uso de macros globales, este comando también se podría escribir 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
```

Recuerde, el código debe estar en una sola línea en el archivo .do o debe ser seccionado con los delimitadores apropiados de Stata para que se marque el final del comando. Sin embargo, el uso de macros hace que el código sea mucho más limpio. Como se señaló anteriormente, 3SLS es un estimador GMM que utiliza una matriz de ponderación particular que asume errores de i.i.d. Por lo tanto, los resultados anteriores del 3SLS con el comando <*reg3*> se pueden reproducir con una estimación GMM que tenga una matriz de ponderación adecuada. 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 a utilizar para obtener las estimaciones de los parámetros del primer paso. La subopción <*independent*> le dice a *gmm* que asuma que los residuos son independientes en todas las condiciones del momento. 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 indica una matriz de ponderación que asume la homocedasticidad condicional, pero que no impone la independencia de la ecuación cruzada como se solicita en la matriz de ponderación inicial.[75] Tenga en cuenta que el código anterior <*gmm*> podría tomar mucho más tiempo, a veces hasta varias horas dependiendo de la capacidad física de la computadora, que el código <*reg3*> para converger en una solución. Esto se debe a que el GMM, a diferencia del 3SLS, es un estimador muy general y no lineal y busca una solución de forma numérica.

4.6.2 Estimación de GMM 3SLS

Si los errores son heterocedásticos, sabemos que las estimaciones tradicionales de 3SLS son menos eficientes y sus errores estándar inconsistentes. Se debe utilizar una matriz de ponderación consistente a la heterocedasticidad para obtener estimaciones de parámetros consistentes en este caso. Esto es posible con GMM utilizando la opción <*wmatrix(robust)*> tal como se aplica en el código a continuación:

```
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)* solicita una matriz de ponderación apropiada para los errores que son independientes, pero no necesariamente distribuidos idénticamente. Si se prefiere solicitar una matriz de ponderación que también tenga en cuenta la correlación arbitraria entre las observaciones dentro de los conglomerados, como se suele observar en los datos de la encuesta, 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 podrían obtener errores estándar *bootstrapped* si se utilizara <*reg3*> con un prefijo *bootstrap*. Por ejemplo, <*bootstrap, reps(1000)*

seed(1010):reg3 ($ylist = $x1list $x2list), exog($zlist) endog($x1list) 3sls>. Esto es mejor que estimar un 3SLS con <*reg3*> ignorando la posible heterocedasticidad. Sin embargo, <*reg3*> con 1 000 réplicas de *bootstrap* puede tomar tanto tiempo como <*gmm*> para lograr la convergencia. Por otro lado, <*gmm*> tiene la ventaja adicional de especificar una matriz de ponderación que tiene en cuenta la heterocedasticidad de la conglomeración y la autocorrelación.

Los modelos implementados anteriormente son modelos identificados, ya que el número de instrumentos es igual al número de variables del RHS endógenas. Si hay un modelo sobreidentificado en su lugar, la implementación del código de Stata sería la misma, excepto que los nombres de esos instrumentos adicionales serían agregados a la lista de IV en la macro global *zlist*.

4.6.3 Variables instrumentales 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. Esto se puede implementar con la ayuda del comando de Stata <*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' $x2list ($x1list = $zlist);
};
```

Stata también tiene un excelente comando escrito por usuarios <*ivreg2*>[78] que puede utilizarse 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.

La IV ecuación por ecuación, a la que Wooldridge[14] se refiere como un “estimador del sistema 2SLS”, se puede implementar omitiendo la opción <*twostep*> y <*wmatrix()*> desde la implementación tradicional del 3SLS en un comando <*gmm*> como se muestra a continuación. Esto debería dar un resultado similar a los obtenidos con <*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 los 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 la IV ecuación por ecuación indica que los errores no son homocedásticos, entonces se puede usar el estimador del sistema 2SLS con <*gmm*> como se indica anteriormente, ya que devuelve errores estándar robustos, o el comando <*ivregress*> se puede modificar 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, por ejemplo, <*wmatrix(cluster clustvar)*>. Las estimaciones de los coeficientes, así como sus errores estándar, entonces serán consistentes como se indica en la Sección 4.4.

4.6.4 Ejecución 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. Estos incluyen una prueba de endogeneidad de las variables, una prueba de validez de los instrumentos utilizados y una prueba de homocedasticidad de los errores, entre otros. Estas pruebas se implementan más fácilmente después de una estimación de IV ecuación por ecuación.

1) Prueba de endogeneidad de los regresores: como se señaló en la Sección 4.4, no es necesario utilizar un estimador IV a menos que las variables endógenas sean efectivamente endógenas. La endogeneidad puede probarse con la ayuda de la *prueba DWH* de exogeneidad[72] en caso de errores i.i.d., o la *prueba de puntuación de Wooldridge*, o una prueba auxiliar basada en regresión en caso de errores no i.i.d.,[73] como se discutió anteriormente. Después del comando <*ivregress*>, el comando <*estat endogenous*> se puede utilizar para hacer esto. Informará el estadístico DWH o cualquiera de las otras estadísticas consistentes con la heterocedasticidad discutidas anteriormente, dependiendo de la matriz de ponderación opcional usada con el comando <*ivregress*>. En cualquier caso, la hipótesis nula es que las variables son exógenas y una estadística de prueba significativa indicaría que la variable debe tratarse como endógena.

De forma similar, si se usa <*ivreg2*>, el comando <*ivendog*> se puede usar después de <*ivreg2*> e informará el *estadístico DWH*. Alternativamente, la opción <*endog(varname)*> se puede utilizar junto con el comando <*ivreg2*> para comprobar si un instrumento es endógeno. Por ejemplo, <*ivreg2 bsfood $x2list ($x1list = $zlist), gmm2s robust endogtest($x1list)*> prueba la endogeneidad de las tres variables endógenas junto con la visualización de los resultados de la regresión. Esta opción es particularmente útil para probar la endogeneidad cuando la heterocedasticidad está presente.

2) Prueba de validez de los instrumentos: como se señaló anteriormente, los estimadores IV son consistentes solo bajo la fuerte suposición de que existe un instrumento válido *z* que satisface tanto las restricciones de inclusión como las de exclusión. La prueba de la restricción de inclusión es sencilla. Esta comprueba si los instrumentos son débiles o fuertes. Con el comando <*ivreg2*> simplemente hay que añadir la opción <*first*>; por ejemplo, <*ivreg2 bsfood $x2list ($x1list = $zlist), first*>. Esto reportaría los resultados de la primera etapa de regresión, uno para cada regresor endógeno. Por ejemplo, en este caso, dado que hay tres variables endógenas de RHS (*pq, lnM, lnM2*), reportaría tres resultados de regresión de primera etapa con cada una de estas variables endógenas como la variable dependiente y todos los regresores exógenos restantes y las IV como variables del RHS. El R^2 y el *estadístico-F* de estas regresiones de la primera etapa indican qué tan fuertes o débiles son los instrumentos.

Una regla empírica común sugiere que un *estadístico-F* de menos de 10, en el caso de un solo regresor endógeno, es indicativo de un instrumento débil.[15,79] Si hay un solo instrumento y regresor endógeno, esto se traduce en un *valor-t* de 3.2 o mayor y el *valor-p* de 0.0016 o menor para el instrumento. Los resultados de esta *prueba F* deben reportarse cuando se reportan las estimaciones IV. Esta regla empírica, sin embargo, es *ad hoc* y puede no ser suficientemente conservadora si el modelo está sobreidentificado. Para ecuaciones con más de un regresor endógeno, se puede usar un estadístico llamado R^2 parcial *de Shea* en lugar del valor *crítico de F*.[15] Sin embargo, no hay consenso sobre qué tan bajo debe ser el valor del R^2 para indicar un problema.[15] La opción <*first*> después de <*ivreg2*>, así como el comando <*estat firststage*> después de ejecutar <*ivregress*>, reporta el R^2 *parcial de Shea*. Ver Cameron y Trivedi[15], Capítulo 6 para una exposición detallada de estas estadísticas. Alternativamente, consulte el manual de referencia de Stata[75] sobre notas técnicas postestimación de ivregress en la página 1212-13.

En general, no es posible probar la restricción de exclusión o la exogeneidad de los instrumentos, especialmente en el caso anterior. En el caso sobreidentificado, sin embargo, se puede realizar una prueba de sobreidentificación de restricciones 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 la homocedasticidad. Si <*ivregress*> hubiera utilizado la opción <*gmm*> junto con una matriz de ponderación consistente con la heterocedasticidad, entonces el comando <*estat overid*> reportaría un *estadístico de Hansen's J* o un *estadístico de Hansen-Sargan* que explica las alteraciones heteroscedásticas. Un estadístico de prueba estáticamente significativo indica que los instrumentos pueden no ser válidos. Esto puede ocurrir si los instrumentos no son verdaderamente exógenos, o si están siendo excluidos incorrectamente de la regresión,[73] como se señaló antes.

3) Prueba de heteroscedasticidad: como se señaló en la Sección 4.4, si los errores son heterocedásticos, la regresión IV produce errores estándar inconsistentes y las estimaciones tradicionales de 3SLS son menos eficientes y los errores estándar inconsistentes. El *estadístico Pagan-Hall* se puede utilizar para probar la presencia de heterocedasticidad en la regresión IV. Esto se puede implementar con el comando <*ivhettest*>. Por ejemplo, después de <*ivreg2 bsfood $x2list ($x1list = $zlist)*>, ejecute el comando <*ivhettest*> y reportará el *estadístico Pagan-Hall* con la hipótesis nula de alteraciones homocedásticas. Un estadístico significativo implicará un rechazo de la nula, indicativo de la presencia de heterocedasticidad. Desafortunadamente, el <*ivhettest*> no funciona después del comando <*ivregress*> por ahora. También hay un programa escrito por usuarios, <*lmhreg3*>[80], que puede instalarse con el comando <*ssc install lmhreg3*> y realiza las pruebas tanto de la ecuación única como de la heterocedasticidad general del sistema después del comando <*reg3*>. Así que, si <*reg3*> se usara para hacer una estimación 3SLS, se puede aplicar el comando <*lmhreg3*> inmediatamente después para comprobar si cada una de las ecuaciones individuales, así como el sistema en su conjunto, satisfacen 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) Comprobación de heterogeneidad de las preferencias entre los consumidores y los no consumidores de tabaco: si se quiere examinar si las preferencias son heterogéneas entre los hogares que gastan y los que no gastan en tabaco, se puede estimar la ecuación 4.5 en lugar de la ecuación 4.4 para comprobar la importancia 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 esto, primero estime el modelo en la ecuación 4.5 usando <*ivregress*> como se muestra a continuación:

```
#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 comprobar una hipótesis lineal compuesta de que los tres coeficientes asociados con la variable *dummy tob* son conjuntamente cero. Un rechazo (es decir, un estadístico de prueba significativo) sugiere que la ecuación 4.5 puede ser una especificación más apropiada, mientras que ningún rechazo 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 anteriores de la 1) a la 3) deben realizarse 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.

Resumen de las pruebas y decisión sobre el método de estimación: Para revisar, antes de decidir qué método de estimación utilizará, ya sea el tradicional 3SLS <*reg3*>, o GMM 3SLS <*gmm*>, o el de IV ecuación por ecuación (ya sea con *ivregress* o *ivreg2*), se recomienda primero estimar con IV ecuación por ecuación. Esto permitiría determinar si existe endogeneidad en el modelo y si los instrumentos utilizados son válidos. Después, se debe realizar la prueba de heterocedasticidad. Si la prueba de heterocedasticidad indica que los errores son i.i.d., entonces se puede optar por un <*reg3*> para hacer la estimación tradicional 3SLS. Si no es así, se debe usar un método de estimación GMM 3SLS usando el comando <*gmm*> en Stata para producir estimaciones de parámetros eficientes. Según Wooldridge,[14] nunca es peor el estimador GMM 3SLS utilizando la matriz de ponderación consistente de heterocedasticidad, asintóticamente, que el tradicional 3SLS 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 cualquier caso. Por último, la comprobación de la importancia conjunta de los parámetros asociados a la variable indicadora para el gasto en tabaco, junto con sus variables de interacción, indicará si es apropiado utilizar una forma funcional que trate a los consumidores de tabaco y a los no consumidores como algo totalmente diferente. Si llega a la conclusión de que se les trate de forma diferente, deberá especificarse la ecuación 4.5 y deberán repetirse en la nueva especificación todas las pruebas sugeridas anteriormente de la 1) a la 3).

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 estas comunidades. De manera similar, también podemos clasificar a los hogares en términos de la gravedad de su gasto en tabaco como consumidores moderados, medios y altos. Es posible que el desplazamiento sea mucho mayor entre los altos consumidores en comparación con los consumidores moderados. Por estas y otras razones, el investigador puede querer examinar el impacto del desplazamiento por diferentes subgrupos definidos, ya sea por los ingresos o por otras características. Los estudios han utilizado diferentes subgrupos para examinar el impacto, incluyendo grupos de ingresos,[56,66] la gravedad del gasto en tabaco[63] y diferentes tipos de tabaco.[66]

Aparte 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 Stata relevante.

A continuación, se muestran algunos ejemplos.

Paso 1: Definición de variables categóricas para indicar el subgrupo

El siguiente código de Stata clasifica a los hogares en tres grupos de ingresos; bajo, medio y alto basados en la distribución de los gastos mensuales per cápita de cada hogar. Esto puede hacerse primero creando una variable de gasto per cápita (*pcexp*) por diferentes hogares.

```
#delimit;
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;
```

Como se indica aquí arriba, el código clasifica como de ingresos altos (*High income*) a aquellos hogares que superan el percentil 70 de la distribución del gasto per cápita y como de ingresos bajos (*Low income*) a aquellos que se encuentran por debajo del percentil 30 de la distribución, mientras que los que se encuentran entre ambos se clasifican como de ingresos medios (*Middle income*). El código también asigna etiquetas para cada uno de los valores que toma la nueva variable *incgrp*. Del mismo modo, también se puede clasificar a los hogares en función de la distribución de la parte del presupuesto que se gasta en tabaco entre los que gastan poco o mucho, y así sucesivamente.

Paso 2: Agregar opciones de subgrupos a los comandos relevantes de Stata

Una vez que se genera la variable categórica, digamos *incgrp*, la estimación se puede hacer ya sea agregando una opción <*by(incgrp)*> o <*over(incgrp)*> o el prefijo <*bysort incgrp:*> a los comandos de Stata, dependiendo del comando en particular. Por ejemplo, el comando <*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)
}
```

Para el GMM 3SLS también se puede añadir el prefijo <*bysortincgrp:*> antes del comando <*gmm*>.

La Sección 7.3 del Apéndice de código provee un archivo .do de ejemplo que detalla el código utilizado en este capítulo. Los usuarios podrán copiar y pegar ese código en el editor de archivos .do de Stata y estimar los resultados con los datos/variables correspondientes que se describen ahí mismo.

4.7 Caso práctico de Turquía

Los hogares turcos, a pesar de vivir en un país de ingresos medios altos, gastaron más del 8 % de su presupuesto familiar en la compra de tabaco en 2011. Mientras que los ricos en Turquía gastaron alrededor del 6.2 % del presupuesto familiar en tabaco, los pobres gastaron hasta un 10.7 %.[65] Dado que una gran parte del presupuesto familiar se está desviando hacia el gasto en tabaco, es posible que los gastos en otras necesidades básicas se estén sustituyendo. En este contexto, San & Chaloupka[65] examinaron el desplazamiento del gasto en tabaco de una serie de grupos de productos en Turquía. El estudio estimó el modelo QAIDS con una variante de la ecuación 4.5 para estimar los efectos del desplazamiento. El modelo econométrico utilizado fue el método 3SLS que se discute en la Sección 4.4.2.3. El estudio utilizó el gasto total para instrumentalizar el gasto neto en tabaco y la proporción de mujeres para instrumentalizar el gasto en tabaco. La Tabla 4.2 muestra un resumen de los resultados que encontraron en 2011.

Como se indica a continuación, los autores estimaron el modelo especificado en la ecuación 4.5 de este capítulo con alguna variación en las variables de control y utilizaron la técnica tradicional de estimación 3SLS. La Tabla 4.2 enlista solo un subconjunto de los grupos de productos analizados por los autores. La primera columna bajo el grupo de productos muestra las estimaciones de los parámetros y la segunda columna presenta los errores estándar. La variable binaria que indica el gasto en tabaco es significativa en el caso de todos los productos básicos excepto la educación. Su signo negativo indica que el gasto en tabaco tiene un impacto negativo en el gasto destinado al grupo de productos correspondiente. El *p.q.* muestra el total de los gastos preasignados en tabaco y da una indicación del grado de exclusión. Por ejemplo, por cada incremento en liras en la cantidad preasignada para tabaco, hay una reducción en la parte del presupuesto asignada a vivienda en el presupuesto restante del hogar de 0.0022 puntos porcentuales o 0.0022 x M lira, donde M es el presupuesto restante después del gasto en tabaco.

Tabla 4.2 Impacto del desplazamiento del gasto en tabaco en Turquía, 2011

	Comida		Vivienda		Ropa		Transporte		Educación	
	Coeff.	S.E	Coeff.	S.E	Coeff.	S.E	Coeff.	S.E	Coeff.	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
lnM²	−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
dlnM²	0.0122*	0.003	−0.0105*	0.006	−0.0056	-0.002	0.0410*	0.005	−0.0012	-0.002

Resultados de la especificación de la ecuación 4.5. Los valores de las variables dependientes van de 0 a 1. *Estos resultados son significativos en el nivel del 5 %. Origen: San & Chaloupka (2016).[65]

Supongamos que los gastos mensuales después del gasto en tabaco son de unas 1 200 liras (ya que 106 liras gastadas en tabaco constituyen alrededor del 8.17 % del presupuesto). Luego, utilizando las estimaciones de los parámetros presentados por los autores, se puede calcular que un aumento de 100 liras en la cantidad preasignada para tabaco conduce a una disminución de 264 liras en los gastos de vivienda, a la vez que se redistribuyen los gastos en todos los demás productos, aumentando algunos y disminuyendo otros. Por ejemplo, un aumento de 100 liras en la cantidad preasignada para tabaco reduciría los gastos en alimentos, servicios públicos, bienes duraderos, ropa, salud y educación en unas 24, 12, 96, 36, 24 y 36 liras, respectivamente, y aumentaría los gastos en transporte, entretenimiento, alcohol y otros productos básicos en 252, 204, 24 y 12 liras, respectivamente. Lo que es importante ver es que un aumento en el gasto en tabaco redistribuye claramente los gastos, beneficiando a algunos artículos, pero perjudicando a otros. En este caso en particular, los artículos de consumo reducido son en su mayoría necesidades básicas y eso justifica la intervención de políticas públicas para regular el consumo de tabaco.

5 Medición del efecto empobrecedor del consumo de tabaco

5.1 Introducción

Las estimaciones nacionales de la pobreza son una variable política importante en muchos países. La estimación del porcentaje de pobres determina el curso de los debates sobre políticas de desarrollo en varios países. La reducción de la pobreza es un objetivo declarado en muchos países y la erradicación de la pobreza en todas sus formas es el primero 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. Esto se debe a que el consumo de tabaco y la pobreza forman parte de un círculo vicioso.[4] A medida que se gasta más dinero en tabaco, los hogares se ven privados de ciertas necesidades, entre ellas la alimentación y la nutrición, como se explica en el Capítulo 4, lo que genera un enorme costo de oportunidad y exacerba la pobreza. Dado que el dinero que se gasta en tabaco es muy improductivo y aumenta las enfermedades relacionadas con el tabaco, el aumento consecuente de los costos de atención médica y de la pérdida de ingresos debido a las muertes prematuras y a la morbilidad, también puede agravar el problema de la pobreza. En 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 las poblaciones de bajos ingresos.[4] Las desigualdades relacionadas con la riqueza y la educación en el consumo de tabaco entre hombres y mujeres son mayores en PIMB en comparación con los países de ingresos medios altos.[81]

En el Capítulo 4 se explica cómo el gasto en tabaco sustituye o desplaza los gastos en diferentes grupos de productos, ofreciendo una cierta dimensión del costo de oportunidad del gasto en tabaco. Este capítulo mostrará cómo cuantificar el impacto directo del gasto en tabaco sobre la pobreza, medido por el número de personas que viven en pobreza; cómo el gasto en tabaco contribuye al empobrecimiento; y los métodos actuales para cuantificar el mismo. También demostrará cómo se puede hacer esto con la ayuda de las HES usando Stata.

5.2 Medición de la pobreza y su importancia

Las definiciones de pobreza varían de un país a otro dependiendo de las circunstancias sociales y económicas específicas que prevalecen en cada país. Sin embargo, “casi todas las NPL (*National Poverty Line*, Línea o umbral nacional de pobreza) están ancladas al costo de una canasta básica de alimentos, lo que los pobres en ese país comen habitualmente, que proporciona una nutrición adecuada para una buena salud y actividad normal, más una asignación para gastos no alimentarios”.[82] A medida que cambian las canastas básicas de alimentos o los gustos y preferencias, las naciones típicamente redefinen la línea nacional de pobreza. En esencia, la línea de pobreza toma en cuenta una cierta privación de recursos y define una cantidad que es necesaria para sostener una noción percibida localmente de lo que se requiere para no ser pobre. Normalmente, esto se convierte a una unidad de moneda local. Por ejemplo, la Oficina de Estadísticas de Sudáfrica[83] define una línea de pobreza alimentaria, la cantidad de dinero que un

individuo necesitará para poder costear la ingesta mínima de energía diaria requerida, también conocida como la línea de pobreza "extrema", en 547 rands sudafricanos por persona al mes. También define otras líneas de pobreza que tienen en consideración ciertos gastos mínimos en artículos no alimentarios. De manera similar, la Oficina del Censo de los Estados Unidos (USCB, por sus siglas en inglés) utiliza un conjunto de umbrales de ingresos en dólares que varían según el tamaño y la composición del hogar para determinar quién está en estado de pobreza.[84] La definición de 2017 de la USCB muestra que una persona menor de 65 años que gana menos de $12 752 dólares al año, se considera que vive por debajo de la línea de pobreza.

Aunque existen varios métodos para medir la pobreza, el índice de recuento de la pobreza (HCR, por sus siglas en inglés) es una medida absoluta y uno de los indicadores más utilizados, especialmente en los PIMB.[85] El HCR, una medida de recuento que se define como la fracción de la población que vive por debajo de la NPL y permite una interpretación muy intuitiva y sencilla. Esta fracción se calcula comúnmente utilizando las HES, ya que permite calcular el gasto promedio de cada hogar, o los gastos de consumo per cápita y compararlos con la línea de pobreza definida. Sin embargo, el HCR no toma en consideración el grado de pobreza. En otras palabras, la tasa de pobreza que se mide con el HCR seguiría siendo la misma incluso si los pobres 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 pobreza puede variar significativamente de un país a otro y de una cultura a otra. Aunque no son comparables entre países, las líneas de pobreza son bastante útiles en el contexto de las políticas nacionales de desarrollo de un país. Se pueden utilizar como referencia para facilitar ciertos programas de bienestar social, por ejemplo, para desarrollar intervenciones dirigidas específicamente a los pobres.

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 esto, ayuda distinguir dos tipos de pobreza, tal como lo explica el sociólogo británico B. Seebohm Rowntree[86] y se reproduce en la Monografía de la OMS/NCI.[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 se encuentran 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 tienen recursos suficientes para satisfacer sus necesidades básicas, pero esos recursos no se utilizan de manera eficiente. Como resultado, a pesar de poseer una mayor cantidad de recursos, estos hogares pueden estar viviendo en condiciones similares o inferiores a las de la pobreza primaria. Por ejemplo, una cantidad importante de ingresos se gasta en el consumo improductivo y nocivo de productos como el tabaco o el alcohol por parte de un hogar que, por todo lo demás, se encuentra por encima de la línea de pobreza. Debido a un efecto de desplazamiento, los hogares no pueden satisfacer sus necesidades básicas, al igual que los hogares en situación de pobreza primaria. Pero las estimaciones del HCR solo reflejarían a quienes se encuentran en la pobreza primaria, aunque muchos hogares del país podrían estar en la pobreza secundaria y, por lo tanto, no satisfacen sus necesidades básicas debido al derroche en el consumo de tabaco. Sería ideal incluir a estos hogares en el cálculo del HCR para que las políticas públicas y los programas puedan ser más efectivos. Alternativamente, habrá que adoptar políticas para que los hogares puedan salir de la pobreza secundaria ayudándoles a reducir o detener el consumo derrochador y nocivo, de modo que el total de sus recursos disponibles pueda satisfacer sus necesidades básicas.

Dado que el presupuesto del hogar es limitado, el consumo de cualquier cosa, incluyendo el tabaco, implica necesariamente sustituciones. Los estudios sobre el desplazamiento que se discuten en el Capítulo 4

muestran que la sustitución ocurre en la forma de desplazamiento de ciertas necesidades básicas. Existen tres canales principales a través de los cuales el aumento del 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: Pérdida de ingresos por la compra de tabaco**
El ingreso directo disponible 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 la morbilidad relacionada con el tabaco**
Dado que el consumo de tabaco y la exposición al SHS conducen inevitablemente a la aparición de varias enfermedades y a 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 a los ingresos disponibles, también puede afectar a la productividad y al potencial de ganar 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 SHS a menudo causan la muerte prematura. Esto resulta en la pérdida de ingresos futuros, lo que repercute en el bienestar de otros miembros del hogar.

Todos estos canales tienen el efecto final de empobrecer aún más a un hogar pobre. Como los pobres suelen asignar una mayor proporción de su presupuesto al tabaco en comparación con los ricos,[4] el impacto empobrecedor del gasto en tabaco es relativamente mayor en los primeros que en los segundos. 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.[87] Como resultado de la disminución del gasto en tabaco y, en consecuencia, de la reducción del gasto en atención médica, estos hogares tendrán más ingresos disponibles para gastar en necesidades esenciales (por ejemplo, alimentos, ropa y educación).

Aunque los estudios que examinan las desigualdades socioeconómicas en el tabaquismo y el consumo de tabaco son bastante sustanciales,[4] los estudios que cuantifican el efecto empobrecedor del gasto en tabaco en términos de su impacto sobre las medidas cuantificables de la pobreza son limitados. Uno de los primeros estudios se realizó en Vietnam[88] y cuantificó el efecto empobrecedor de los pagos directos en la atención médica. Sin embargo, el primer estudio que estimó el efecto empobrecedor del gasto directo de los hogares en tabaco y el gasto médico excesivo atribuible al tabaquismo se realizó en China.[89] Se encontró que estos dos efectos combinados fueron responsables del empobrecimiento de 30.5 millones de residentes urbanos y 23.7 millones de residentes rurales en China. Otro estudio en la India[90] también encontró que el efecto combinado de estos dos factores resultó en el empobrecimiento de 15 millones de personas en la India. Un estudio más reciente en el Reino Unido[91] restó solo los gastos en tabaco de los ingresos de los hogares para estimar su impacto en la pobreza y encontró que más de 432 000 niños podían ser vistos como arrastrados a la pobreza a causa del tabaquismo de sus padres. Otro estudio en el Reino Unido[92] mostró que, si se tiene en cuenta el gasto en tabaco, alrededor de 500 000 hogares adicionales, que comprenden más de 850 000 adultos y casi 400 000 niños, están clasificados dentro de la pobreza en el Reino Unido en comparación con las cifras oficiales de los *Hogares con ingresos por debajo del promedio.*

Estos estudios concluyeron que muchas personas que de otra manera estaban por encima de la NPL en estos países (es decir, en la pobreza secundaria) eran efectivamente pobres porque sus ingresos disponibles después de gastar en tabaco y en gastos de salud asociados eran más bajos que los de las personas que estaban oficialmente clasificadas como por debajo de la NPL. En otras palabras, estas personas son etiquetadas inadvertidamente como que están por encima de la línea de pobreza mientras que, en realidad, no lo están.

Ninguno de los estudios realizados hasta ahora ha estimado el impacto empobrecedor de la pérdida de ingresos por las muertes prematuras relacionadas con el tabaco (Canal 3) y la pérdida de ingresos por la morbilidad relacionada con el SHS (parte del Canal 2). Dado que la pobreza o el HCR se mide en un momento dado, es insostenible restar la pérdida de ingresos por la mortalidad prematura o a la pérdida futura de ingresos de los ingresos actuales de los hogares. Sin embargo, los costos médicos directos atribuibles al SHS (parte del Canal 2) son claramente un candidato para que los ingresos no percibidos se resten del ingreso actual disponible mientras se evalúa el impacto empobrecedor del consumo de tabaco. Pero esto tampoco se ha incorporado en ninguno de los estudios hasta ahora.

5.4 Marco conceptual para estimar el impacto en el HCR

Para estimar el cambio en el HCR, es necesario restar dos tipos diferentes de pérdida de ingresos de los ingresos totales de los hogares para estimar dicho cambio: (1) pérdida de ingresos por la compra de tabaco; y (2) pérdida de ingresos debido al consumo de tabaco y a los costos directos de atención médica atribuibles al SHS. Antes de poder restar estos diferentes componentes de la pérdida de ingresos del ingreso total del hogar, es importante identificar la NPL con base en la forma en que se calcula el HCR. La NPL es un número único para todo el país, o un número diferente para las zonas rurales y urbanas y para cada subregión o estado dentro del país. Por lo general, se puede obtener de los organismos de estadística u otras fuentes gubernamentales de cada país. La variable de ingreso contra la cual se calcula generalmente el HCR se toma de las HES representativas a nivel nacional. Dado que las estimaciones de consumo o de gastos declarados son mucho más fiables que los ingresos declarados al representar los ingresos reales,[7] los gastos estimados a partir de las HES se utilizan como indicador sustituto de los ingresos para estimar la proporción de personas que se encuentran por debajo de la línea de pobreza.

Lo que también es importante es el hecho de que la mayoría de las HES son encuestas de hogares que tratan a los hogares como una sola unidad y los gastos de consumo se reportan para el hogar como un todo. Sin embargo, la pobreza la experimentan las personas, no los hogares per se, y por lo tanto es la pobreza entre las personas la que se debe medir. Aunque no se sabe nada sobre la distribución dentro de los hogares, es una práctica común asumir una distribución uniforme dentro de los hogares cuando se decide la distribución estimada de los consumos individuales.[93] Por lo tanto, al estimar el HCR, es importante utilizar las ponderaciones de la encuesta que puedan generar estadísticas a nivel de población para las personas y no para los hogares. Esta estimación se puede obtener multiplicando el tamaño de los hogares por las ponderaciones dadas de la encuesta para generar estadísticas a nivel de hogares en las HES.

En primer lugar, el HCR total y la pobreza se calculan antes de restar la pérdida de ingresos relacionada 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. Si x es la medida de bienestar (es decir, el gasto en consumo per cápita, que es el gasto total en consumo de los hogares dividido por el tamaño de los hogares), entonces el HCR denotado como (P_0) se calcula de la siguiente manera:[85]

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

Donde $I(.)$ es una función indicadora que toma valor de 1 si su argumento es verdadero y 0 en caso contrario. Ya que se calcula usando las HES, se deben usar las ponderaciones apropiadas de la encuesta. $P_0 \times N$ da el número total de pobres en el país.

5.4.1 *Exceso de pobreza a causa de 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). Sea *t* el gasto en consumo per cápita en la compra de tabaco en el mismo período de tiempo para el que se ha capturado la medida de bienestar (*x*). En otras palabras, es la pérdida de ingresos por la compra de tabaco. Entonces, el HCR, después de deducir el gasto en tabaco o la pérdida de ingresos por la compra de este, denotados por (P_1), se puede calcular de la siguiente manera:

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

donde, de nuevo, *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 en compras de tabaco. $(P_1 - P_0) \times N$ es el número excesivo de personas que se empobrecen a causa del gasto en tabaco. En otras palabras, esto es del exceso de pobreza atribuido al gasto directo de compra de tabaco.

Si bien es más aceptable suponer una distribución uniforme del consumo dentro de los hogares al decidir la distribución estimada del consumo individual,[93] puede no ser tan aceptable suponer una distribución uniforme dentro de un hogar en el caso de productos comúnmente consumidos por los adultos como el tabaco. Una de las soluciones propuestas por Deaton[7] es "un sistema de ponderaciones, en el que los niños cuentan como una fracción de un adulto, con la fracción dependiente de la edad, de modo que el tamaño efectivo del hogar es la suma de estas fracciones y se mide no por el número de personas, sino por el número de *equivalentes adultos*". Sin embargo, dado que el hogar es una sola unidad para todos los fines prácticos y que el dinero gastado en tabaco reduce necesariamente los ingresos disponibles para todo el hogar, incluidos los niños, el impacto empobrecedor podría ser soportado por igual tanto por los niños como por los adultos. Por lo tanto, la consideración de dicha *equivalencia adulta* al examinar el efecto empobrecedor del gasto en tabaco puede no dar los resultados deseados.

5.4.2 *El exceso de pobreza atribuido a la pérdida de ingresos por la compra de tabaco y el tratamiento de la morbilidad asociada al tabaco*

La morbilidad relacionada con el tabaco puede ocurrir entre aquellos que consumen tabaco, así como entre aquellos que están expuestos al SHS. Sean *t* y *h* el gasto per cápita en tabaco y su consumo total y los gastos en salud per cápita atribuibles al SHS, respectivamente, en el mismo período de tiempo para el que se mide el bienestar (*x*). Entonces, el HCR después de deducir la pérdida de ingresos de las compras de tabaco y el tratamiento de los gastos médicos atribuibles al tabaco, denotados por (P_2), se puede calcular 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 valor de 1 si su argumento es verdadero y 0 en caso contrario. $x_i - t_i - h_i$ es el ingreso per cápita disponible después de restar tanto los gastos en tabaco como los gastos en atención médica atribuibles a su consumo y al SHS. $(P_2 - P_1) \times N$ es el número adicional de personas que se empobrecen debido al consumo de tabaco y al gasto médico atribuible al SHS. $(P_2 - P_0) \times N$ será el número total de personas empobrecidas después de contabilizar la pérdida de ingresos procedente tanto del gasto en tabaco como de los gastos atribuibles a la atención médica.

Aunque las HES proporcionan información sobre los gastos de atención médica, no distinguen la cantidad de atención médica que se puede atribuir al consumo de tabaco o a la exposición al SHS. Esto debe

estimarse por separado y la sustracción debe ser solo para los gastos en atención médica que puedan ser atribuidos al consumo de tabaco o a la exposición al SHS. Los costos atribuibles pueden estimarse utilizando un enfoque específico de la enfermedad o un enfoque inclusivo o de causa múltiple.[94] Dado que las HES a menudo proporcionan gastos en atención médica agregados, el enfoque inclusivo es más apropiado para su uso. Descompone la parte de los costos médicos totales atribuibles al consumo de tabaco o exposición al SHS al multiplicar los costos totales de atención médica por la fracción atribuible al uso de tabaco, o fracción atribuible del SHS, comúnmente conocida como Fracción Atribuible al Tabaquismo (FAT en español y SAF, por sus siglas en inglés). FAT es la porción de la utilización total de la atención médica que se atribuye al tabaquismo por parte de exfumadores y fumadores.[94] De manera similar, la FAT para el SHS sería la fracción de los gastos de atención médica que pueden atribuirse al SHS.

Por lo tanto, los gastos de atención médica atribuibles al consumo de tabaco y al SHS (es decir, a *h* en la ecuación 5.3) se pueden calcular de la siguiente manera:

$$h_i = (exphealth_i / hsize_i)^* (FAT_{tob} + FAT_{SHS}) \qquad (5.4)$$

donde *exphealth* y *hsize* son gastos del hogar en salud y el tamaño de los hogares, respectivamente. Ambas variables se obtienen directamente de las HES. FAT_{tob} y FAT_{SHS} son fracciones de los gastos de atención médica atribuibles al consumo de tabaco y al SHS, respectivamente. La FAT debe estimarse externamente utilizando datos de varias fuentes diferentes. También se puede obtener de estudios disponibles en otras partes del país.

La FAT se puede estimar utilizando el enfoque epidemiológico o el enfoque econométrico. El enfoque econométrico requiere “amplios datos representativos a nivel nacional que contengan información detallada sobre el historial de tabaquismo de cada encuestado, sus características sociodemográficas, su situación laboral, otras conductas de riesgo para la salud, su estado de salud, sus afecciones médicas, los gastos anuales de atención médica por tipo de servicios de atención (como hospitalizaciones y visitas ambulatorias), y los días de incapacidad o pérdida de trabajo anuales”.[95] Por otra parte, el enfoque epidemiológico es menos intensivo en datos y “puede hacerse con datos agregados y, por lo tanto, puede utilizarse cuando no se dispone de datos detallados de las encuestas de salud”.[95] Por estas razones, en muchos PIMB se prefiere el enfoque epidemiológico para estimar la FAT. La OMS proporciona un “Conjunto de herramientas” para estimar los costos económicos del tabaquismo y proporciona una explicación detallada y una metodología para ambos métodos de estimación de la FAT, el epidemiológico y el econométrico. Por lo tanto, en este Conjunto de herramientas no se tratará esta cuestión. A diferencia de los datos requeridos para estimar la FAT para el consumo de tabaco, los datos requeridos para estimar la FAT para el SHS pueden ser más difíciles de obtener. Tal vez esta sea la razón por la que estudios anteriores que cuantificaban el efecto empobrecedor del consumo de tabaco sobre la pobreza ignoraron en sus cálculos esta fuente particular de pérdida de ingresos.

5.4.3 Brecha de pobreza a causa del tabaco

Los cambios incrementales en el número de pobres debido a la sustracción sucesiva del gasto en tabaco, y el gasto en atención médica atribuido al tabaquismo y a la exposición al SHS, pueden no ser significativos, ya que muchos caen por debajo de la línea de pobreza debido únicamente al gasto en tabaco, y se vuelven aún más pobres debido al gasto en atención médica atribuible. Este es un motivo de preocupación y es exactamente el principal defecto de una medida como el HCR. En otras palabras, el HCR no tiene en cuenta el grado de pobreza y no se vería afectado si los pobres se empobrecieran aún más. Una manera de abordar este problema es utilizar una medida llamada “brecha de pobreza” que le da un mayor peso a la

persona en pobreza agregada cuanto más pobre es. La brecha de pobreza se puede calcular utilizando la siguiente fórmula:[7]

$$P_G = \frac{1}{N} \sum_{(i=1)}^{N} \left(1 - \frac{x_i}{z}\right) I(x_i \leq z) \qquad (5.5)$$

Deaton[7] señala que la PG (*Poverty Gap*, Brecha de pobreza) se puede interpretar como una medida per cápita del déficit total de bienestar individual por debajo de la línea de pobreza, ya que es la suma de todos los déficits divididos por la población y expresados como una relación de la propia línea de pobreza. $P_G \times z \times N$ da la cantidad total por la cual los pobres están por debajo de la línea de pobreza. Comparando la brecha de pobreza antes y después de restar el gasto en tabaco y otros gastos atribuibles a la atención médica, se puede estimar el grado en que el tabaco está empobreciendo a las personas en la pobreza secundaria. Sin embargo, esto tampoco se ha hecho en estudios anteriores sobre la cuantificación del impacto empobrecedor del consumo de tabaco.

5.5 Preparación de datos para estimar el efecto empobrecedor

Como se detalla en el Capítulo 2, los datos deben limpiarse primero 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 en conjunto, como una aproximación a los ingresos de los hogares (*exptotal*). Además, los gastos en atención médica (*exphealth*) son necesarios para calcular los costos de atención médica atribuibles al tabaco y al SHS en función de la disponibilidad de la FAT. Las otras variables que se necesitan de las HES para el análisis incluyen el tamaño del hogar, las ponderaciones de la encuesta y las variables para declarar el diseño de esta. Se necesita una variable o escalar para representar la NPL. Si la NPL es una variable que varía de una región a otra, o por áreas rurales o urbanas, o estados dentro del país, entonces la variable tendrá que fusionarse con los datos de la encuesta de hogares antes de que se pueda hacer el análisis. Para ello, una variable de identificación común deberá estar presente tanto en los datos de gasto de los hogares como en los datos de la línea de pobreza.

Por ejemplo, si la NPL de un país varía según el estado y la residencia (rural o urbana), entonces los datos de pobreza deben tener tres variables, una variable que indique la NPL *(npl),* generalmente en unidades de moneda local, una variable con los nombres o el código numérico de los diferentes estados *(stateid)* y una variable de residencia que indique si la *npl* pertenece a áreas rurales o urbanas *(residence).* Del mismo modo, los datos de las HES también deben tener variables de *stateid* y *residence*. 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 *npl* y otras variables de identificación según sea necesario y guárdelo con el nombre *poverty.dta*. Luego, abra los datos maestros de las HES con la información de gasto de cada hogar y asegúrese de que tenga las mismas variables de *stateid* y *residence* que en el *poverty.dta.* A continuación, utilice el comando <*merge m:1 stateid residence using poverty.dta*>. Aquí se utiliza una fusión de varios a uno *(m:1)*, ya que el conjunto de datos maestros tiene varios hogares con el mismo estado y residencia. Después del comando de fusión, utilice el comando <*tabulate _merge*> para comprobar si la fusión se ha llevado a cabo con precisión.

Mientras que las HES consideran a los hogares como una sola unidad y reportan todos los gastos a ese nivel, la NPL es generalmente para un individuo, por lo que es importante convertir los datos de gasto para que sean comparables a los de la línea de pobreza. También es importante verificar si la duración de la presentación de informes sobre los gastos (por ejemplo, por mes, por semana o cualquier otro intervalo) es igual y asegurarse de que la línea de pobreza también lo sea durante el mismo período de tiempo. Por

ejemplo, tanto el gasto en consumo o el gasto en atención médica atribuible al consumo de tabaco como la línea de pobreza deberían ser por persona por mes. Para ello, en Stata, cree nuevas variables para generar gastos per cápita para ser comparados con la línea de pobreza utilizando la variable de tamaño del hogar *(hsize)*. Por ejemplo, los gastos per cápita pueden generarse como <*gen pce =exptotal/hsize*>. De manera similar, las variables sobre el gasto per cápita en tabaco *(pcetob)* y en salud *(pcehealth)* deben generarse dividiendo el gasto total correspondiente entre la variable de tamaño del hogar. Además, el uso del valor de la FAT y *pcehealth* crean la variable *pcehealthtob* que representa el consumo de tabaco per cápita y los gastos en atención médica atribuibles al tabaquismo. Por ejemplo, si la FAT para el consumo de tabaco es 0.2, entonces se puede generar una nueva variable *pcehealthtob* con el comando <*gen pcehealthtob=pcehealth*0.2*> y si la FAT para la exposición al SHS 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 SHS.

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

(1) *pcet* (pce, gasto per cápita después de que se compensan los gastos en tabaco): <*gen pcet=pce-pcetob*>, y
(2) *pceh* (pce, gasto per cápita después de que se compensan los gastos de tabaco y el consumo de tabaco y los gastos atribuibles al SHS): <*pceh=pcet-pcehealthtob- pcehealthshs*>. En caso de que no se disponga de estimaciones de la FAT para la exposición al SHS, la fórmula para pceh puede reducirse a <*gen pceh=pcet-pcehealthtob*>.

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

5.6 Estimación del impacto empobrecedor del consumo de tabaco

En Stata, la estimación del HCR es bastante sencilla, el software ofrece varios módulos escritos por usuarios para esto. Por ejemplo, <*povdeco*>[96] es un módulo que estima el HCR y varias otras medidas de pobreza con un solo comando. Para ello, instale el módulo con <*ssc install povdeco*> y ejecute el comando <*povdeco pce [fw=pweight], varpline(npl)*> donde *pce* es la variable para los gastos mensuales per cápita, *npl* es la variable para la NPL y *pweight* es la ponderación de la encuesta ajustada al tamaño del hogar. *povdeco* informará sobre el 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 el HCR solamente, un simple comando de proporción 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]
```

Esto también se puede hacer después de declarar el diseño de la medición usando el comando *svyset* como se explica en el Capítulo 2. En este caso el comando puede escribirse como <*svy: proportion povdum*>.

Dado que el cambio en el HCR debe determinarse después de la sustracción incremental de las diferentes pérdidas de ingresos, como se discutió anteriormente, esto puede implementarse mejor con el siguiente código. El código a continuación 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 usadas son: *pce, pcet, pceh, npl*, y *pweight.* Si los datos se han preparado con estos nombres de variables, al ejecutar el código se generaría una matriz 3X2 en la ventana de resultados de Stata mostrando *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 cualquier ingreso perdido), *pcet* (HCR después de restar el ingreso perdido por la compra directa de tabaco), y *pceh* (HCR después de restar la pérdida de ingresos tanto de la compra de tabaco como del uso de tabaco y los gastos atribuibles de salud del SHS). Los valores correspondientes bajo la columna “Poor” (“Pobre”) muestran el número estimado de personas pobres en cada paso sucesivo. Comparar dos filas sucesivas permite ver el cambio tanto en el HCR como en el número de pobres después de la sustracción sucesiva de cada componente de la pérdida de ingresos. El número de pobres en el código se calcula multiplicando HCR por la población total, como se estima en la encuesta de hogares, lo que es posible utilizando la variable de peso específico de la persona. El escalar *r(sum)* es un resultado guardado después del comando *summarize* y muestra el resultado de multiplicar la media por el tamaño de la población. Alternativamente, se puede multiplicar el HCR por los datos de población disponibles a nivel nacional de otras fuentes para llegar al cambio en el número de pobres.

El análisis anterior se puede hacer con diferentes subgrupos, así como utilizando cualquiera de los métodos discutidos anteriormente. Sin embargo, es necesario modificar los datos y generar nuevas variables para poder realizar el análisis a nivel de subgrupo. La Sección 7.4 del Apéndice de código incluye un archivo .do de ejemplo que detalla el código utilizado en esta sección. Los usuarios podrán copiar y pegar ese código en el editor de archivos .do de Stata y estimar los resultados con los datos/variables correspondientes que se describen ahí mismo.

5.7 Caso práctico de la India

En la India, durante 2004 y 2005, fuentes oficiales del gobierno consideraron que alrededor del 28.3 % de la población rural y el 25.6 % de la urbana estaban por debajo de la NPL. Las estadísticas oficiales de pobreza se presentan por separado para las zonas rurales y urbanas del país y también se presentan por estado. La línea de pobreza también está disponible a esos niveles de agregación. La India también tiene el segundo mayor número de consumidores de tabaco del mundo.[33] 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. En este contexto, John *et al.*,[90] examinaron el impacto empobrecedor del gasto en tabaco, así como el del gasto en atención médica relacionado con el consumo de tabaco en dicho país. La Tabla 5.1 muestra los resultados de su análisis.

En la tabla se presentan en primer lugar las estimaciones oficiales del HCR y el número de pobres en la India por zonas rurales y urbanas. Luego muestra el efecto separado de restar el gasto en tabaco y el de atención médica atribuible al consumo de tabaco de los gastos per cápita para las áreas rurales y urbanas de la India y luego el efecto combinado de restar ambos gastos de los gastos totales per cápita. Los resultados muestran que la tasa de pobreza o HCR aumentó en 1.6 y 0.8 puntos porcentuales en las zonas rurales y urbanas de la India, respectivamente, tras restar los ingresos no percibidos de la compra de tabaco y los gastos sanitarios relacionados con este. En otras palabras, el gasto en tabaco y el gasto en atención médica asociado al consumo de tabaco empobreció a unas 15 millones de personas adicionales en la India. Es decir, 15 millones de personas en la India que están por encima de la línea de pobreza oficial se encuentran en pobreza secundaria, pero disfrutan de un menor nivel de vida en términos de su capacidad para gastar en las necesidades diarias porque su dinero se está desviando hacia gastos innecesarios en tabaco.

Esto también tiene serias implicancias políticas. Si las medidas de bienestar social (por ejemplo, un subsidio alimentario) se dirigen a los que 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 social y seguirían viviendo en la pobreza.

Tabla 5.1 Cambios en el HCR y el número de pobres después de considerar el consumo de tabaco en la India

	Rural	Urban	Total
(1) Estimaciones Oficiales			
Población total (millones)	780.2	315.5	1095.7
Población DLP (%)	28.3	25.6	
Población DLP (millones)	220.7	80.8	301.6
(2) Contabilidad de las compras de tabaco			
Población DLP (%)	29.8	26.3	
Población DLP (millones)	232.5	83.1	315.6
(3) Contabilidad de los gastos médicos relacionados con el tabaco			
Población DLP (%)	28.4	25.7	
Población DLP (millones)	221.4	81.1	302.5
(4) Efecto combinado de (2) y (3)			
Población DLP (%)	29.8	26.4	
Población DLP (millones)	232.9	83.3	316.2

DLP = Por debajo de la línea de pobreza. Origen: John *et al.* (2011)[90]

6 Bibliografía

1. Organización Mundial de la Salud. *Tobacco Control for Sustainable Development*. Nueva Delhi, India: Organización Mundial de la Salud, Oficina Regional para Asia Sudoriental; 2017. http://apps.who.int/iris/handle/10665/255509. (Consultado el 5 de octubre de 2018).
2. Organización Mundial de la Salud. *WHO Global Report: Mortality Attributable to Tobacco*. Ginebra, Suiza; 2012. *http://apps.who.int/iris/bitstream/10665/44815/1/9789241564434_eng.pdf*. (Consultado el 4 de septiembre de 2018).
3. Jha P., Peto R. *Global Effects of Smoking, of Quitting, and of Taxing Tobacco. N Engl J Med*. 2014; 370 (1): 60-68. doi:10.1056/NEJMra1308383
4. Instituto Nacional del Cáncer de Estados Unidos y la Organización Mundial de la Salud. *The Economics of Tobacco and Tobacco Control.* Bethesda, MD: Departamento de Salud y Servicios Humanos de los Estados Unidos (HHS, por sus siglas en inglés), Institutos Nacionales de la Salud (NIH, por sus siglas en inglés), Instituto Nacional del Cáncer (NCI, por sus siglas en inglés); y Ginebra, CH: Organización Mundial de la Salud; 2016. http://cancercontrol.cancer.gov/brp/tcrb/monographs/21/index.html. (Consultado el 19 de febrero de 2017 y el 19 de septiembre de 2018).
5. Goodchild M., Nargis N., d'Espaignet, E. T. *Global economic cost of smoking-attributable diseases. Tob. Control.* Enero de 2017: tobaccocontrol-2016-053305. doi:10.1136/tobaccocontrol-2016-053305
6. ONU. *Transforming Our World: The 2030 Agenda for Sustainable Development.* Nueva York, EE. UU.: Asamblea General Nacional de las Naciones Unidas; 2015. https://sustainabledevelopment.un.org/post2015/transformingourworld. (Consultado el 19 de septiembre de 2018).
7. Deaton, A. S. *The Analysis of Household Surveys.* Baltimore: Johns Hopkins University Press para el Banco Mundial; 1997.
8. *Pollak, R. A. Conditional Demand Functions and Consumption Theory. Q J Econ.* 1969; 83 (1): 60-78.
9. Pollak, R. A. *Conditional Demand Functions and the Implications of Separable Utility.* South Econ J. 1971; 37 (4): 423-433.
10. Indian Statistical Institute (Instituto Indio de Estadística). *The National Sample Survey: General Report No. 1. First Round: October 1950 - March 1951. Sankhy Indian J Stat* 1933-1960. 1953; 13 (1/2): 47-214.
11. Banco Mundial. Encuestas de Medición del Nivel de Vida (LSMS). http://microdata.worldbank.org/index.php/catalog/lsms. Publicado en 2018. (Consultado el 2 de septiembre de 2018).

12. Red Internacional de Encuestas de Hogares (IHSN). Catálogo de encuestas de la IHSN. http://catalog.ihsn.org/index.php/catalog/central. Publicado en 2018. (Consultado el 23 de septiembre de 2018).

13. *LISGIS. Household Income and Expenditure Survey 2016*. (Encuesta de ingresos y gasto de los hogares, 2016). Monrovia, Liberia: *Liberia Institute for Statistics and Geo-Information Services* (Instituto de Estadística y Servicios de Geo-Información de Liberia) - Gobierno de Liberia; 2017. http://catalog.ihsn.org/index.php/catalog/7279. (Consultado el 11 de septiembre de 2018).

14. Wooldridge, J. M. *Econometric Analysis of Cross Section and Panel Data.* 2ª ed. Cambridge, Massachusetts: The MIT Press; 2010. https://mitpress.mit.edu/books/econometric-analysis-cross-section-and-panel-data-second-edition.

15. Cameron, A. C., Trivedi, P. K. *Microeconometrics Using Stata*, Edición Revisada. 2ª ed. Texas, EE. UU.: Stata Press; 2010. https://www.stata.com/bookstore/microeconometrics-stata/. (Consultado el 14 de octubre de 2018).

16. StataCorp. *Stata Statistical Software: Release 15*. (Software estadístico de Stata: versión 15). College Station, TX: StataCorp LP; 2018. http://www.stata.com/.

17. Baum, C. F. *A Little Bit of Stata Programming Goes a Long Way*. Boston, MA: Departamento de Economía del Boston College; 2005. http://ideas.repec.org/e/pba1.html. (Consultado el 10 de junio de 2018).

18. StataCorp. *Stata programming reference manual: Release 15*. (Manual de referencia de programación de Stata: versión 15). 2017.

19. Chaloupka, F. J., Warner K. E. *The Economics of Smoking.* En: *The Handbook of Health Economics;* 2000:1539-1627.

20. IARC. *IARC Handbooks of Cancer Prevention in Tobacco Control, Volume 14: Effectiveness of Tax and Price Policies for Tobacco Control.* Lyon, Francia; 2011. http://www.iarc.fr/en/publications/pdfs-online/prev/handbook14/handbook14.pdf. (Consultado el 2 de junio de 2015).

21. Organización Mundial de la Salud. *WHO Report on the Global Tobacco Epidemic, 2015: Raising Taxes on Tobacco.* Ginebra, Suiza; 2015. http://www.who.int/tobacco/global_report/2015/report/en/. (Consultado el 12 de septiembre de 2018).

22. Departamento de Salud y Servicios Humanos de los Estados Unidos (HHS). *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*, 2014. Rockville, MD: Departamento de Salud y Servicios Humanos de los Estados Unidos, Centros para el Control y Prevención de Enfermedades (CDC, por sus siglas en inglés), Centro Nacional para la Prevención de Enfermedades Crónicas y la Promoción de la Salud, Oficina de Tabaquismo y Salud (OHS, por sus siglas en inglés); 2014. http://www.surgeongeneral.gov/library/reports/50-years-of-progress. . (Consultado el 2 de septiembre de 2018).

23. Jha, P., Chaloupka, F. J. *Tobacco Control in Developing Countries*. Oxford, Nueva York: Oxford University Press; 2000.

24. 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.* 1993; 12 (1): 1-18. doi:10.1016/0167-6296(93)90037-F

25. 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.* 1994; 15 (1): 26-36.

26. 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.* 1995; 85 (9): 1218-1222.

27. Sung, H. Y., Hu, T. W., Keeler, T. E. *Cigarette Taxation and Demand: An Empirical Model. Contemp Econ Policy.* 1994; 12 (3): 91-100. doi:10.1111/j.1465-7287.1994.tb00437.x

28. Deaton, A., Muellbauer, J. *An Almost Ideal Demand System. Am Econ Rev.* 1980; 70 (3): 312-326.

29. Deaton, A. *Quality, Quantity, and Spatial Variation of Price. Am Econ Rev. 1988;* 78 (3): 418-430.

30. Deaton, A. *Household survey data and pricing policies in developing countries. World Bank Econ Rev.* 1989; 3 (2 [mayo de 1989]): 183-210.

31. Deaton, A. *Price elasticities from survey data: Extensions and Indonesian results. J Econom.* 1990; 44 (3): 281-309. doi:10.1016/0304-4076(90)90060-7

32. Deaton, A., Grimard F. Demand Analysis and Tax Reform in Pakistan. Banco Mundial; 1992. http://www.worldbank.org/html/prdph/lsms/research/wp/a81_100.html#wp85. (Consultado el 12 de septiembre de 2018).

33. John, R. M., Rao, R. K., Rao, M. G., et al. *The Economics of Tobacco and Tobacco Taxation in India*. París: Unión Internacional Contra la Tuberculosis y Enfermedades Respiratorias; 2010.

34. John, R. M. *Consumption of Tobacco in India: An Economic Analysis.* 2007.

35. John, R. M. *Price Elasticity Estimates for Tobacco in India. Health Policy Plan.* 2008; 23 (3): 200-209.

36. 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.* 2011; No. 17580. http://www.nber.org/papers/w17580. (Consultado el 10 de septiembre de 2018).

37. Selvaraj, S., Srivastava, S., Karan, A. Price elasticity of tobacco products among economic classes in India, 2011–2012. BMJ Open. 2015; 5 (12): e008180. doi:10.1136/bmjopen-2015-008180

38. Eozenou P, Fishburn B. *Price Elasticity Estimates for Cigarette Demand in Vietnam.* Centro de Investigación sobre Desarrollo y Políticas (DEPOCEN, por sus siglas en inglés), Vietnam; 2009. https://ideas.repec.org/p/dpc/wpaper/0509.html. (Consultado el 3 de diciembre de 2018).

39. Chen, Y., Xing, W. *Quantity, quality, and regional price variation of cigarettes: Demand analysis based on a household survey in China. China Econ Rev.* 2011; 22 (2): 221-232. doi:10.1016/j.chieco.2011.01.004

40. Chelwa, G. T*he economics of tobacco control in some African countries.* 2015. https://open.uct.ac.za/handle/11427/16529. (Consultado el 12 de marzo de 2018).

41. Chávez, R. *Price elasticity of demand for cigarettes and alcohol in Ecuador*, con base en datos de hogares. *Rev Panam Salud Publica Pan Am J Public Health*. 2016; 40 (4): 222-228.

42. McKelvey, C. *Price, unit value, and quality demanded. J Dev Econ.* 2011; 95 (2): 157-169. doi:10.1016/j.jdeveco.2010.05.004

43. Gibson, J., Rozelle, S. *Prices and Unit Values in Poverty Measurement and Tax Reform Analysis. World Bank Econ Rev.* 2005; 19 (1): 69-97.

44. Menon, M., Perali, F., Tommasi, N. *Estimation of unit values in household expenditure surveys without quantity information.* Stata J. 2017; 17 (1): 222-239.

45. Atella V, Menon M, Perali F. *Estimation of Unit Values in Cross Sections Without Quantity Information and Implications for Demand and Welfare Analysis.* Rochester, NY: Social Science Research Network; 2003. https://papers.ssrn.com/abstract=391481. (Consultado el 29 de noviembre de 2018).

46. Coondoo, D., Majumder, A., Ray, R. *Method of Calculating Regional Consumer Price Differentials with Illustrative Evidence from India. Rev Income Wealth.* 2004; 50 (1): 51-68. doi:10.1111/j.0034-6586.2004.00111.x

47. Slesnick, D. T. *Prices and demand: New evidence from micro data. Econ Lett.* 2005; 89 (3): 269-227. doi:10.1016/j.econlet.2005.05.034

48. Hoderlein, S., Mihaleva, S. *Increasing the price variation in a repeated cross section. J Econom.* 2008; 147 (2): 316-325. doi:10.1016/j.jeconom.2008.09.022

49. Lecocq, S., Robin, J. M. *Estimating almost-ideal demand systems with endogenous regressors. Stata J.* 2015; 15 (2): 554-573.

50. 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.* 2015; 47 (6): 553-568. doi:10.1080/00036846.2014.975332

51. Lewbel, A. *Identification and Estimation of Equivalence Scales under Weak Separability. Rev Econ Stud.* 1989; 56 (2): 311-316. doi:10.2307/2297464

52. Lewbel, A., Pendakur, K. Tricks with Hicks: The EASI Demand System. Am Econ Rev. 2009; 99 (3): 827-863. doi:10.1257/aer.99.3.827

53. 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.* 2018; 25 (8): 557-561. doi:10.1080/13504851.2017.1346353

54. Organización Mundial de la Salud. *WHO Report on the Global Tobacco Epidemic, 2017: Monitoring Tobacco Use and Prevention Policies.* Ginebra, Suiza; 2017. http://apps.who.int/iris/bitstream/10665/255874/1/9789241512824-eng.pdf?ua=1. (Consultado el 4 de agosto de 2017).

55. Organización Mundial de la Salud. *Systematic Review of the Link between Tobacco and Poverty.* Ginebra, Suiza: Organización Mundial de la Salud; 2014. http://www.who.int/tobacco/publications/syst_rev_tobacco_poverty/en/index.html. (Consultado el 20 de junio de 2018).

56. John, R. M. *Crowding out effect of tobacco expenditure and its implications on household resource allocation in India. Soc Sci Med.* 2008; 66 (6): 1356-1367. doi:10.1016/j.socscimed.2007.11.020

57. Efroymson, D., Ahmed, S., Townsend, J., et al. *Hungry for tobacco: An analysis of the economic impact of tobacco consumption on the poor in Bangladesh. Tob Control.* 2001; 10: 212-217. doi:10.1136/tc.10.3.212

58. Thomson, G. W., Wilson, N. A., O'Dea D., Reid, P. J., Chapman, P. H. T*obacco spending and children in low income households. Tob Control.* 2002; 11 (4): 372-375.

59. 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.* 2004; 3 (4): 263-272.

60. Wang, H., Sindelar, J. L., Busch, S. H. *The impact of tobacco expenditure on household consumption patterns in rural China. Soc Sci Med.* 2006; 62 (6): 1414-1426. doi:10.1016/j.socscimed.2005.07.032

61. 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.* 2008; 66 (9): 1979-1989. doi:10.1016/j.socscimed.2008.01.007

62. Koch, S. F., Tshiswaka-Kashalala, G. *Tobacco Substitution and the Poor.* Sudáfrica: Departamento de Economía, Universidad de Pretoria; 2008. https://www.up.ac.za/media/shared/Legacy/UserFiles/wp_2008_32.pdf. (Consultado el 14 de octubre de 2018).

63. John, R. M., Ross, H., Blecher, E. *Tobacco expenditures and its implications for household resource allocation in Cambodia. Tob Control.* 2011. doi:10.1136/tc.2010.042598

64. Chelwa, G., Walbeek, C. van. *Assessing the Causal Impact of Tobacco Expenditure on Household Spending Patterns in Zambia.* Sudáfrica: Economic Research Southern Africa; 2014. https://econrsa.org/2017/wp-content/uploads/working_paper_453.pdf. (Consultado el 14 de octubre de 2018).

65. San, S., Chaloupka, F. J. *The impact of tobacco expenditures on spending within Turkish households. Tob Control.* 2016; 25 (5): 558-563. doi:10.1136/tobaccocontrol-2014-052000

66. 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.* 2018; 13 (10): e0205120. doi:10.1371/journal.pone.0205120

67. Bloque, S., Webb, P. *Up in Smoke: Tobacco Use, Expenditure on Food, and Child Malnutrition in Developing Countries. Econ Dev Cult Change.* 2009; 58 (1): 1-23. doi:10.1086/605207

68. 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.* 2015; 15. doi:10.1186/s12889-015-2423-9

69. Browning, M., Meghir, C. *The Effects of Male and Female Labor Supply on Commodity Demands. Econometrica.* 1991; 59 (4): 925-951.

70. Banks, J., Blundell R., Lewbel, *A. Quadratic Engel Curves and Consumer Demand. Rev Econ Stat.* 1997; 79 (4): 527-539.

71. Pollak, R. A., Gales, T. J. Demand System Specification and Estimation. Oxford, Nueva York: Oxford University Press; 1995.

72. Davidson, R., MacKinnon, J. G. *Estimation and Inference in Econometrics.* Nueva York; 1993.

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

74. Zellner, A., Theil, H. *Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations. Econometrica.* 1962; 30 (1): 54-78. doi:10.2307/1911287

75. StataCorp. *Stata base reference manual: Release 15.* (Manual de referencia de base de Stata: versión 15). 2017.

76. Vermeulen, F. *Do Smokers Behave Differently? A Tale of Zero Expenditures and Separability Concepts. Econ Bull.* 2003; 4 (6): 1-7.

77. Paraje, G., Araya, D. *Relationship between smoking and health and education spending in Chile. Tob Control.* 2018; 27 (5): 560-567. doi:10.1136/tobaccocontrol-2017-053857

78. Baum, C. F., Schaffer, M. E., Stillman, S. IVREG2: Stata Module for Extended Instrumental Variables/2SLS and GMM Estimation. Departamento de Economía de Boston College; 2007. https://ideas.repec.org/c/boc/bocode/s425401.html. (Consultado el 30 de octubre de 2018).

79. Staiger, D., Stock, J. H. *Instrumental Variables Regression with Weak Instruments. Econometrica.* 1997; 65 (3): 557-586. doi:10.2307/2171753

80. Shehata, E. A. E. LMHREG3: *Stata Module to Compute Overall System Heteroscedasticity Tests after (3SLS-SURE) Regressions.* Departamento de Economía de Boston College; 2011. https://ideas.repec.org/c/boc/bocode/s457381.html. (Consultado el 14 de noviembre de 2018).

81. 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.* 2018; 27 (1): 26-34. doi:10.1136/tobaccocontrol-2016-053266

82. Banco Mundial. WDI - Poverty and Inequality. http://datatopics.worldbank.org/world-development-indicators/themes/poverty-and-inequality.html#national-poverty-lines. Publicado en 2018. (Consultado el 1 de noviembre de 2018).

83. Statistics South Africa. *National Poverty Lines.* Pretoria, Sudáfrica; 2018. http://www.statssa.gov.za/publications/P03101/P031012018.pdf. (Consultado el 11 de enero de 2018).

84. Oficina del Censo de los Estados Unidos. *Poverty*. https://www.census.gov/topics/income-poverty/poverty.html. Publicado en 2018. (Consultado el 1 de noviembre de 2018).

85. Foster, J., Seth, S., Lokshin, M., Sajaia, Z. *Unified Approach to Measuring Poverty and Inequality: Theory and Practice.* Washington, D. C.: Banco Mundial; 2013. http://documents.worldbank.org/curated/en/281001468323965733/A-unified-approach-to-measuring-poverty-and-inequality-theory-and-practice. (Consultado el 2 de noviembre de 2018).

86. Rowntree, B. Seebohm. *Poverty: A Study of Town Life.* Londres, Reino Unido: MacMillan; 1901.

87. Fuchs Tarlovsky, A., Del Carmen, G., Mukong, A. K. *Long-Run Impacts of Increasing Tobacco Taxes: Evidence from South Africa.* Banco Mundial; 2018: 1-39. http://documents.worldbank.org/curated/en/122081521480061194/Long-run-impacts-of-increasing-tobacco-taxes-evidence-from-South-Africa. (Consultado el 2 de noviembre de 2018).

88. Wagstaff, A., Doorslaer, E. van. *Paying for Health Care: Quantifying Fairness, Catastrophe, and Impoverishment, with Applications to Vietnam, 1993-98.* Banco Mundial; 2001. http://ideas.repec.org/p/wbk/wbrwps/2715.html. (Consultado el 2 de noviembre de 2018).

89. Liu, Y., Rao, K., Hu, T., Sun, Q., Mao, Z. *Cigarette smoking and poverty in China. Soc Sci Med.* 2006; 63 (11): 2784-2790.

90. John, R. M., Sung, H. Y., Max, W. B., Ross, H. *Counting 15 million more poor in India, thanks to tobacco. Tob Control.* 2011; 20 (5): 349-352. doi:10.1136/tc.2010.040089

91. 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.* 2015; 15 (1): 507. doi:10.1186/s12889-015-1797-z

92. Reed, H. *Estimates of Poverty in the UK Adjusted for Expenditure on Tobacco.* Londres, Reino Unido: Action on Smoking and Health; 2015. http://ash.org.uk/information-and-resources/health-inequalities/health-inequalities-resources/estimates-of-poverty-in-the-uk-adjusted-for-expenditure-on-tobacco/. (Consultado el 11 de marzo de 2018).

93. Ravallion, M. *Poverty Comparisons: A Guide to Concepts and Methods.* Washington, D. C.: Banco Mundial; 1992: 1. http://documents.worldbank.org/curated/en/290531468766493135/Poverty-

comparisons-a-guide-to-concepts-and-methods. (Consultado el 2 de noviembre de 2018).

94. Cutler, D. M., Epstein, A. M., Frank, R. G., et al. *How Good a Deal Was the Tobacco Settlement?: Assessing Payments to Massachusetts. J Risk Uncertain.* 2000; 21 (2): 235-261. doi:10.1023/A:1007863408004

95. Organización Mundial de la Salud. *Assessment of the Economic Costs of Smoking. World Health Organization Economics of Tobacco Toolkit.* Ginebra, Suiza: Organización Mundial de la Salud; 2011. http://whqlibdoc.who.int/publications/2011/9789241501576_eng.pdf. (Consultado el 4 de octubre de 2018).

96. Jenkins, S.P. *POVDECO: Stata Module to Calculate Poverty Indices with Decomposition by Subgroup.* Departamento de Economía de Boston College; 2008. https://ideas.repec.org/c/boc/bocode/s366004.html. (Consultado el 6 de noviembre de 2018).

Anexos de Código

7

7.1 Archivo .do de Stata para calcular la elasticidad precio usando el método Deaton para un solo producto

```
*=====================================================================
* Date : November 2018
* Topic: Stata do-file made as part of the toolkit on Using Household
* Expenditure Surveys for Economics of Tobacco Control Research
* This do-file estimates the own price elasticity and expenditure
* elasticity for a single commodity, for example, cigarette.
* Data base used: hbs_data.dta
* Key variables:
* - exptotal - total household expenditures in local currency units (LCU)
* - expcig - total household cigarette expenditures in LCU
* - qcig - number of sticks or packs of cigarettes purchased
* - hsize - household size
* - meanedu - mean education of household in years
* - maxedu - maximum education of household in years
* - sgroup - factor variable represeting household social groups
* - maleratio - ratio of number of males to household size
* - clust - variable identifying the primary samping unit or cluster
*=====================================================================

clear
version 15
set mem 1000m
set more off
*change the directory paths below to inform stata where the data are
*stored and where output is to be stored
global pathin  "C:\Data\"
global pathout "C:Data\Demand"

capture log close
log using $pathout\Demand.log, replace
use $pathin\hbs_data.dta
```

```
*generating additional variables for the model
gen uvcig=expcig/qcig
gen luvcig=ln(uvcig)
gen bscig=expcig/exptotal
replace bscig=0 if bscig==.
gen lhsize=ln(hsize)
gen lexp=ln(exptotal)
tab sgroup, gen(sgp)

*Testing for spatial variation in unit values
*Any of the following two commands may be used. Both give identical estimates
anova luvcig clust
*regress luvcig i.clust

*Estimating within-cluster first stage regressions
*Here we run two equations
*Running unit value regression and storing the results
areg luvcig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
scalar sigma11=$S_E_sse / $S_E_tdf
scalar b1=_coef[lexp] //*Expenditure elasticity of quality

predict ruvcig, resid  // residuals from the unit value regression
*These residuals still have cluster effects in

*Purged y's for next stage
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 // residuals from the budget share regression

scalar sigma22=$S_E_sse/$S_E_tdf // var-covar matrix of u0 (e0e0)
scalar b0=_coef[lexp]       // Coefficients of lnexp in BS regression
*Purged y's for next stage
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

*This next regression is necessary to get covariance of residuals
qui areg ruvcig rbscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
scalar sigma12=_coef[rbscig]*sigma22  // covar matrix of u1 (e1e0)
```

```
*expenditure elasticity of quantity
qui sum bscig
scalar Wbar=r(mean)
scalar Expel=1-b1+(b0/Wbar)
di Expel


/*To estimate the bootstrap standard errors for expenditure elasticity
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
*/

*Next, equations (3.4) and (3.5) are derived via the following sets of commands:
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
*keeping one obs per cluster
qui by clust: keep if _n==1

*Deaton uses harmonic mean to estimate average cluster size
ameans n0c
scalar n0=r(mean_h)
ameans n1c
scalar n1=r(mean_h)
drop n0c n1c
```

```
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)  //Var of y1
scalar R=r(cov_12) //Covariance y0c and y1c
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
```

7.2 Archivo do de Stata para calcular la elasticidad precio y elasticidad cruzada usando el Método Deaton para múltiples bienes

```
*=====================================================================
* Topic: Stata do-file reproduced from Deaton and modified
* for the toolkit on Using Household Expenditure Surveys for
* Economics of Tobacco Control Research
*
* It provides the code for calculating the system of demand
* equations, including the own and cross-price elasticities,
* for completing the system, and for calculating the
* symmetry-constrained estimates. There are four
* separate programs: the first, allindia.do, is for estimating
* the demand system. Appended to it is a program mkmats.do, that
* calculates the commutation and selection matrices required for
* the symmetry-constrained estimates, as well as procedures for
* making the "vec" of a matrix, and for reversing the operation.
* The code bootall.do bootstraps the procedure in order to obtain
* measures of sampling variability.
* please make three separate do-files namely, allindia.do mkmats.do and
* bootstrap.do and save them all in the same directory before the elasticity
* estimates are done as in the do-file allindia.do
*
* Note: This code was written as part of "The Analysis of Household Surveys:
* A Microeconometric Approach to Development Policy", by Angus Deaton.
* This book, published for the World Bank by The Johns Hopkins University
* Press and scheduled for release in 1997. The original coda is available
* from http://web.worldbank.org/archive/website00002/WEB/EX5_1-2.HTM
```

```
*
* Data base used: hbs_data.dta
* Key variables needed to execute this code:
* - The log unit values begine with luv, e.g., luvcig luvbeer
* - The budget shares begine with bs, e.g., bscig bsbeer
* - lexp - natural log of total household expenditures
* - lhsize - matiral log of household size
* - Additional household-specific variables as available to be added by the user
* - The following are added here
* - meanedu - mean education of household in years
* - maxedu - maximum education of household in years
* - sgp1 to sgp3 - factor variable represeting household social groups
* - maleratio - ratio of number of males to household size
* - clust - variable identifying the primary samping unit or cluster
*============================================================
clear all
set matsize 10000
cd "C:\Users\Rijo\Documents\Dropbox\Work\Frank\TA-Dhaka"
global pathin  "C:\Data\"
global pathout "C:Data\poverty"
capture log close
log using $pathout\Elasticity.log, replace
use $pathin\hbs_data.dta
*#########################################################################
*allindia.do (with modifications of variable names, number of goods.
*We also add comments at various places for the ease of understanding
*Equation numbers added at various places refers to the correspnding equations
*in Deaton's book Analayis of household Surveys referred above
*Executing the program part by part may return errors.
*#########################################################################
version 7.0
*These are the commodity identifiers to be added by the user
global goods "cig beer"
*number of goods in the system to be declared by the user
global ngds=2

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
```

```
* Average Budget shares
cap program drop mkwbar    // creating average budget shares
program def mkwbar
        local ig=1
        while "`1'" ~= ""{
        qui summ bs`1'
        matrix wbar[`ig',1]=_result(3)
        local ig=`ig'+1
        mac shift}
end

mkwbar $goods

/*************************************************************
FIRST STAGE REGRESSIONS: WITHIN - CLUSTER
*************************************************************/
cap program drop st1reg  // stage 1 within village regression
program def st1reg
        local ig=1
        while "`1'" ~= ""{
*Cluster-fixed effect regression
*areg, instead of reg, is used for linear regression with a large dummy-variable set
areg luv`1' lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)

*Measurement error variance
*Summ of squares of errors / total degree of freedom for error;
matrix ome[`ig',1]=$S_E_sse/$S_E_tdf  //var-covar matrix of u1 (e1e1)
matrix b1[`ig',1]=_coef[lexp] //*Expenditure elasticity of quality
*These residuals still have cluster effects in
predict ruv`1', resid  // residuals from the unit value regression

*Purged y's for next stage
gen y1`1'=luv`1'-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio ///
                -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu ///
                -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3

drop luv`1'

*Repeat for budget shares
areg bs`1' lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
predict rbs`1', resid // residuals from the budget share regression

matrix sig[`ig',1]=$S_E_sse/$S_E_tdf // var-covar matrix of u0 (e0e0)
matrix b0[`ig',1]=_coef[lexp]       // Coefficients of lnexp in BS regression
gen y0`1'=bs`1'-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio ///
                -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu ///
                -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3
```

```
*This next regression is necessary to get covariance of residuals
qui areg ruv`1' rbs`1' lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)

matrix lam[`ig',1]=_coef[rbs`1']*sig[`ig',1]  // covar matrix of u1 (e1e0)
drop bs`1' rbs`1' ruv`1'
local ig=`ig'+1
mac shift}
end

st1reg $goods
matrix list sig          // var-covar matrix of u0 (e0e0)
matrix list ome          // var-covar matrix of u1 (e1e1)
matrix list lam          // covar matrix of u1 (e1e0)
matrix list b0           // Coefficients of lnexp in BS regression
matrix list b1           // elasticity of quality w.r.t exp
matrix list wbar // average budget shares

*this completes the first stage regression and estimation of all necessary
*parameters from it. Saving so far as a protection
save tempa.dta, replace
drop _all
use tempa.dta

*************************************************************************
*SECOND STAGE REGRESSIONS - BETWEEN CLUSTER
*************************************************************************
*Averaging by cluster
*Counting numbers of obs in each cluster for n and n+
cap program drop clustit
program def clustit
local ig=1
while "`1'" ~= ""{
egen y0c`ig'=mean(y0`1'), by(clust)
egen n0c`ig'=count(y0`1'), by(clust)
egen y1c`ig'=mean(y1`1'), by(clust)
egen n1c`ig'=count(y1`1'), by(clust)
drop y0`1' y1`1'
local ig=`ig'+1
mac shift }
end
clustit $goods
sort clust
*keeping only one observation per cluster
qui by clust: keep if _n==1
*Saving cluster level information
*Use this for shortcut bootstrapping
save tempclus.dta, replace
```

```
/*Removing region (province) effects
* This is optional and may or may not be used depending on the data
* This assumes the availability of the categorical variable region in the data
tab region, gen(regiond)
cap program drop purge
program define purge
local ig=1
while `ig' <= $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
local ig=`ig'+1
}
end
purge
drop regiond*
*/

matrix define n0=J($ngds,1,0)
matrix define n1=J($ngds,1,0)
*Estimating average cluster sizes using harmonic mean
cap program drop mkns
program define mkns
local ig=1
while `ig' <= $ngds {
replace n0c`ig'=1/n0c`ig'
replace n1c`ig'=1/n1c`ig'
qui summ n0c`ig'
matrix n0[`ig',1]=(_result(3))^(-1)
qui summ n1c`ig'
 matrix n1[`ig',1]=(_result(3))^(-1)
drop n0c`ig' n1c`ig'
local ig=`ig'+1
}
end
mkns

*Making the intercluster variance and covariance matrices (eqn. 5.83)
*This is done in pairs because of the missing values
matrix s=J($ngds,$ngds,0)  // between-cluster var-covar matrix of y1 [cov(y1Gc,y1Hc)]
matrix r=J($ngds,$ngds,0)  // between-cluster covar matrix of y1 [cov(y1Gc,y0Hc)]
```

```
cap program drop mkcov
program def mkcov
local ir=1
while `ir' <= $ngds {
local ic=1
while `ic' <= $ngds {
qui corr y1c`ir' y1c`ic', cov
matrix s[`ir',`ic']=_result(4)
qui corr y1c`ir' y0c`ic', cov
matrix r[`ir',`ic']=_result(4)
local ic=`ic'+1
}
local ir=`ir'+1
}
end
mkcov
*We don't need the data any more
drop _all
matrix list s  // between-cluster var-covar matrix of y1 [cov(y1Gc,y1Hc)]
matrix list r  // between-cluster covar matrix of y1 [cov(y1Gc,y0Hc)]
*Making OLS estimates
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")
*Corrections for measurement error
cap program drop fixmat
program def fixmat
matrix def sf=s
matrix def rf=r
local ig=1
while `ig' <= $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]
local ig=`ig'+1
}
end
fixmat
matrix invs=syminv(sf)
matrix bhat=invs*rf   // The errors-in-variable estimator with ME correction Eqn 5.85
*Estimated B matrix without restrictions
matrix list bhat // The errors-in-variable estimator with ME correction).
*The ratio Phi from which Psi and Theta matrices has to be disentangled.
*Housekeeping matrices, including elasticities
cap program drop mormat
program def mormat
```

```
matrix def xi=J($ngds,1,0) // Xi vector in Eqn 5.92
matrix def el=J($ngds,1,0) // Expenditure elasticity matrix in Eqn 5.89 or 5.50
local ig=1
while `ig' <= $ngds {
matrix xi[`ig',1]=b1[`ig',1]/(b0[`ig',1]+ ///
((1-b1[`ig',1])*wbar[`ig',1]))
matrix el[`ig',1]=1-b1[`ig',1]+b0[`ig',1]/wbar[`ig',1]
local ig=`ig'+1
}
end
mormat
global ng1=$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)
display("Average budget shares")
matrix tm=wbar'
matrix list tm   // Average budget shares
display("Expenditure elasticities")
matrix tm=el'    // Expenditure elasticities (dlnq/dlnx)
matrix list tm
display("Quality elasticities")
matrix tm=b1'
matrix list tm   // Expenditure elasticity of quality (dlnuv/dlnx)

*This all has to go in a program to use it again later
*Basically uses the b from eqn 5.85 matrix to form price elasticity matrix
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 // Theta matrix in Eqn 5.90
display("Theta matrix")
matrix list theta  // Theta matrix in Eqn 5.90
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
*****************************************************************
*If program is executed only up to this point and with a single commoditiy
*by specifying ngds=1 and retain only one good in global macro this will return
*the same estimate of price elasticity derived from code in chapter 3 of this
*tool kit. The code below completes the system of demand equation by filling out
*the matrices. This essentially adds a single composite commodity to the
*equation to complete the system using homogeneity and adding-up restrictions.
*****************************************************************
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'
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 // this command can be dropped if we are only interested in
*symmetry constrained estimates as given below. If it is only the unconstrainted
*estimates that we are intereted in there is no need to run rest of the code too
*****************************************************************
**Calculating symmetry restricted estimators
**These are only approximately valid & assume no quality effects
*the do-file mkmats.do should be executed for this
run mkmats.do
vecmx bhat vbhat
** R matrix for restrictions
lmx $ngds llx
commx $ngds k
global ng2=$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
mkels
complet
*The program will display the own and cross-price elasticities for the two
*googs cigarette and beer along with that of the composite commodity used for
*completing the system
log close

*###########################################################################
*mkmats.do (there is nothign the user has to add in this particular do-file
*They simply have to save this do-file in their working directory
**It calculates two matrices, the commutation matrix and the lower diagonal
**selection matrix that are needed in the main calculations; these are
**valid only for square matrices also a routine for taking the vec of a matrix
**and a matching unvec routine for calculating the commutation matrix k
**the matrix is defined by K*vec(A)=vec(A')
*###########################################################################
cap program drop commx
program define commx
local n2=`1'^2
matrix `2'=J(`n2',`n2',0)
local i=1
local ik=0
while `i' <= `1'{
local j=1
local ij=`i'
while `j' <= `1'{
local ir=`j'+`ik'
matrix `2'[`ir',`ij']=1
local ij=`ij'+`1'
local j=`j'+1
}
local i=`i'+1
local ik=`ik'+`1'
}
end
**for vecing a matrix, i.e., stacking it into a column vector
```

```
cap program drop vecmx
program def vecmx
local n=rowsof(`1')
local n2=`n'^2
matrix def `2'=J(`n2',1,0)
local j=1
while `j' <= `n' {
local i=1
while `i' <= `n' {
local vcel=(`j'-1)*`n'+`i'
matrix `2'[`vcel',1]=`1'[`i',`j']
local i=`i'+1
}
local j=`j'+1
}
end


*program for calculating the matrix that extracts
*from vec(A) the lower left triangle of the matrix 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

**program for unvecing the vec of a square matrix
cap program drop unvecmx
program def unvecmx
local n2=rowsof(`1')
local n=sqrt(`n2')
matrix def `2'=J(`n',`n',0)
```

```
local i=1
while `i' <= `n' {
local j=1
while `j' <= `n' {
local vcel=(`j'-1)*`n'+`i'
matrix `2'[`i',`j']=`1'[`vcel',1]
local j=`j'+1
}
local i=`i'+1
}
end

*##############################################################################
*boostrap.do for bootstrapping demand estimates to derive standard errors
*##############################################################################
version 7.0

capture log close
set more 1
drop _all
do allindia.do
run mkmats.do
log using bstrapDemand.log, replace
drop _all
vecmx xep vxep
set obs 1
gen reps=0
global nels=$ng1*$ng1
global nmc=1000 // the simulation is repeated 1000 times
cap program drop vtodat
program define vtodat
local ic=1
while `ic' <= $nels {
gen e`ic'=vxep[`ic',1]
local ic=`ic'+1
}
end
vtodat
save bootstrap.dta, replace
drop _all

/* This should be used only if the region dummies in allindia.do were used
cap program drop purge
program define purge
local ig=1
while `ig' <= $ngds {
qui regress y0c`ig' regiond*
```

```
predict tm, resid
replace y0c`ig'=tm
drop tm
qui regress y1c`ig' regiond*
predict tm, resid
replace y1c`ig'=tm
drop tm
local ig=`ig'+1
}
end
*/

cap program drop mkns
program define mkns
local ig=1
while `ig' <= $ngds {
replace n0c`ig'=1/n0c`ig'
replace n1c`ig'=1/n1c`ig'
qui summ n0c`ig'
matrix n0[`ig',1]=(_result(3))^(-1)
qui summ n1c`ig'
matrix n1[`ig',1]=(_result(3))^(-1)
drop n0c`ig' n1c`ig'
local ig=`ig'+1
}
end
cap program drop mkcov
program def mkcov
local ir=1
while `ir' <= $ngds {
local ic=1
while `ic' <= $ngds {
qui corr y1c`ir' y1c`ic', cov
matrix s[`ir',`ic']=_result(4)
qui corr y1c`ir' y0c`ic', cov
matrix r[`ir',`ic']=_result(4)
local ic=`ic'+1
}
local ir=`ir'+1
}
end
cap program drop fixmat
program def fixmat
matrix def sf=s
matrix def rf=r
local ig=1
while `ig' <= $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]
local ig=`ig'+1
}
end
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
end
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 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
end

cap program drop bootindi
program define bootindi
local expno=1
while `expno' <= $nmc {
display("Simulation Number `expno'")
quietly {
use tempclus.dta
bsample _N

/*
qui tab region, gen(regiond)
*qui tab subrnd, gen(quard)
purge
drop regiond*
*/

matrix define n0=J($ngds,1,0)
matrix define n1=J($ngds,1,0)
*Averaging (harmonically) numbers of obs over clusters
mkns
*Making the intercluster variance and covariance matrices
*This is done in pairs because of the missing values
matrix s=J($ngds,$ngds,0)
matrix r=J($ngds,$ngds,0)
```

```
mkcov
*We don't need the data any more
drop _all
*Making OLS estimates
matrix bols=syminv(s)
matrix bols=bols*r
*Corrections for measurement error
fixmat
matrix invs=syminv(sf)
matrix bhat=invs*rf
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)
mkels
**Completing the system by filling out the matrices
** Gives standard errors for elasticities without symmetry restrictions
complet //Drop this command if the intend is to estimate symmetry constrained standard errors
*If it is only the unconstrainted estimates that we are intereted in there is
*no need to run the code from this point till the next command complet
**Calculating symmetry restricted estimators
vecmx bhat vbhat
** R matrix for restrictions
lmx $ngds llx
commx $ngds k
global ng2=$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
**going back to get elasticities and complete sytem
matrix bhat=btild
mkels
** Gives standard errors for elasticity with symmetry restrictions
complet

vecmx xep vxep
set obs 1
gen reps=`expno'
vtodat
append using bootstrap.dta
save bootstrap.dta, replace
drop _all
local expno=`expno'+1
  }
sleep 900
}
end
bootindi
use bootstrap.dta
display("Monte Carlo results")
summ
log close
*Note on reading the standard errors:
*The standard errors are displayed in a single column. The final elasticiy matrix
*derived from allindia.do should be stacked (vec of the elasticiy matrix) into a
*single column and the standard errors in the single column diplayed after
*bootstrap will correspond to the vec of elasticity matrix
```

7.3 Archivo .do de Stata para calcular el efecto de desplazamiento del gasto en tabaco

```
*=====================================================================
* Date: November 2018
* Topic: Stata do-file made as part of the toolkit on Using Household
*        Expenditure Surveys for Economics of Tobacco Control Research
* This do-file estimates the crowding out impact of tobacco spending
* Data base used: DataQAIDS.dta
* Key variables:
* - exptotal - total household expenditures in local currency units (LCU)
* - exptobac - total household tobacco expenditures in LCU
* - exphealth - total household healthcare expenditures in LCU
```

```
* - expfood - total household food expenditure in LCU
* - expeducn - total household education expenditure in LCU
* - exphousing - total household housing expenditure in LCU
* - expcloths - total household clothing expenditure in LCU
* - expentertmnt - total household entertainment expenditure in LCU
* - exptransport - total household transportation expenditure in LCU
* - expdurable - total household durable goods expenditure in LCU
* - expother - total household other items expenditure in LCU
* - hsize - household size
* - meanedu - mean education of household in years
* - maxedu - maximum education of household in years
* - sgroup - factor variable represeting household social groups
* - asexratio - adult sex ratio (ratio of adult males to adult females)
* - weight - survey weights
*=================================================================================
clear
version 15
set mem 1000m
set more off

*change the directory paths below to inform Stata where data are
*stored and where output is to be stored
global pathin  "C:\Data\"
global pathout "C:\Data\QAIDS"

capture log close
log using $pathout\Crowdout.log, replace
use $pathin\DataQAIDS.dta

cd "C:\Users\Rijo\Documents\Dropbox\Work\Frank"
use DataQAIDS.dta

*********************************************************************************
*T-test for comparing mean budget shares
*********************************************************************************
*Generate a binary variable for tobacco spending
gen tob=0
replace tob=1 if exptobac >0 & exptobac <.
label define tob 1 "Tobacco spenders" 0 "Tobacco non-spenders", replace

*generating budget share variables for t-test of comparison
*here the denominator is the total expenditures on all goods combined
local items "tobac food health educn housing cloths entertmnt transport durable other"
foreach X of local items{
        gen bs_`X'=(exp`X'/exptotal)
}
*t-test using survey weights
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 [bs_`X']0 - [bs_`X']1
         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
*dropping this budget share variables
drop bs_*

*****************************************************************************
*Preparing variables for estimating crowding out
*****************************************************************************
*generate dummies social groups
tab sgroup, gen(sd)

*creating budget shares for crowding out analysis. Here the denominator is the
*total expendituer minus the expenditures on tobacco
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

*Estimating Crowding out with different models
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

*********************************************
*Traditional 3SLS estimation
```

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

*Traditional 3SLS using 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

*The above two implimentations (reg3 and gmm) should give identical results
*and are traditional 3SLS estimation. But, converging gmm can take much longer
*than reg3 above. Be preapred to wait few hours depending on the machine.
*One possible alternative is to save the reg3 results first using the command
*<matrix b = e(b)> and use these as the starting value for gmm so that
*convergence may be faster. This is done by adding the option
*<center twostep from(b)> to the last line in gmm instead of using only <twostep>

**********************************************
*GMM 3SLS estimation (wooldridge): adjusts for heteroskedasticity
**********************************************
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

*One could also use option <wmatrix(cluster clustvar)> where clustvar is
*the name of the variable that identifies clusters

**********************************************
* Equation-by-equation IV or 2SLS using ivregress:
```

```
**********************************************
*Using Stata's built-in iv regression command
local depvar "food health educn housing cloths entertmnt transport durable"
foreach X of local depvar{
        ivregress 2sls bs`X' $x2list ($x1list = $zlist)
}

*Using user-written program <ivreg2>
*Source: Baum CF, Schaffer ME, Stillman S. IVREG2: Stata Module for
*Extended Instrumental Variables/2SLS and GMM Estimation. Boston College
*Department of Economics; 2007.
*https://ideas.repec.org/c/boc/bocode/s425401.html. Accessed October 30, 2018

local depvar "food health educn housing cloths entertmnt transport durable"
foreach X of local depvar{
        ivreg2 bs`X' $x2list ($x1list = $zlist)
}

*both of the above sets of commands should return identical results.
*But ivreg2, by default, also displays few test statistics of interest

*Using System 2SLS estimator (equation by equation IV)
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)

*This gives parameter estimates similar to the ivregress above, but with
*Robust standard errors. To have the same standard errors
*as in ivregress instead add the option <vce(unadjusted) onestep>
*after winitial(unadjusted, independent)

*if there is heteroskedasticity present, one can perform either the system 2SLS
*using gmm as given above, which returns robust standard errors, or modify the
*ivregress with the option vce(robust) or use the gmm estimator in ivregress
*command to specify additional options like <wmatrix(robust)> or
*<wmatrix(cluster clustvar)>. This is done below.

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)
}

*Where clustvar is the name of cluster variable in the data
*This would return heteroskedasticity consistent standard errors which also
*accounts for arbitrary correlation among observations within clusters

*******************************************************************************
* Performing different tests to decide on the estimation method
*******************************************************************************
*The tests are all shown for equation-by-equation IV and for a single equation
* i.e., for bsfood. One can simply construct a loop around to do this in one
*shot for all equations

*****************************************
*(1)Testing Endogeneity of regressors:
*****************************************
*dpending on whether or not the vce(robust) option is used the output of the
*test results will differ. In either case, a significant statistic implies
*rejecting the null Ho: variables are exogenous.

ivregress 2sls bsfood $x2list ($x1list = $zlist)
estat endogenous

ivregress 2sls bsfood $x2list ($x1list = $zlist), vce(robust)
estat endogenous

*These tests can also be done in a loop for all commodities together as follows:
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
}
*with ivreg2, however, do the tests along with the regression itself
*with the option endogtest() as follows
ivreg2 bsfood $x2list ($x1list = $zlist), endogtest($x1list)

*********************************************
*(2) Testing the validity of instruments
```

```
*********************************************
**Testing inclusion restriction. Checks if instruments are strong or weak

ivregress 2sls bsfood $x2list ($x1list = $zlist)
estat firststage, all
*This will show as many first stage regression results are the number of
*endogenous variables. Since we've three here it will report three first stage
*results. Rule of thumb- suggests an F-statistic of less than 10, in case of a
*a single endogenous regressor, to be indicative of a weak instrument
*Since we have three here, a statistic called Shea's partial R2 can used
*instead of the F-critical value. These are also listed after the command.
*Plese note there is no consensus on how low of a value of R2 indicates a
*problem. See Cameron & Trivedi25 (Chapter 6.4.2) for a detailed exposition of
*these statistics

*with ivreg2, however, do the tests along with the regression itself
*with the option endogtest() as follows:

ivreg2 bsfood $x2list ($x1list = $zlist), first

**Testing exclusion restriction. (instrument exogeneity)

*It is not possible to test the exclusion restriction when the model is just
*identified as we have in the specifications above. If there are more instruments
*than the number of endogenous variables, we can perform a test of
*over identifying restrictions. This is done as

ivregress 2sls bsfood $x2list ($x1list = $zlist)
estat overid
*In just identified case, it will simply return an error
*"no overidentifying restrictions".
* For the purpose of demonstration, suppose we specifiy the following:
* it returns the results of Sargan statistic. But, remember, this is just
* an arbitrary specification in which we keep the number of instruments higher
* The results are not to be taken anyways.
ivregress 2sls bsfood $x2list (pq lnM = $zlist)
estat overid
*if the heteroskedasticity consistent standard errors are used, estata overid
* will return Score chi2 or Hansen's J chi2-statistic. A significant
*test statistic indicates that the instruments may not be valid.
ivregress 2sls bsfood $x2list (pq lnM = $zlist), vce(robust)
estat overid

*********************************************
*(3) Testing for heteroskedasticity
```

```
*********************************************
*The test is more easily done with ivreg2 as follows:
ivreg2 bsfood $x2list ($x1list = $zlist)
ivhettest
*It reports the Pagan-Hall statistic with the Ho: Disturbance is homoskedastic

*********************************************
*(4) Testing heterogeneity in preferences between tobacco users and non-users
*********************************************
*Testing this would need an alternative specification of the model
*Equation 5 in the chapter 4. The addition of dummy variables can be added to
*the model using the factor notations.

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)
}

*A rejection (i.e., significant test statistic) suggests that equation 5 may
*be a more appropriate specification whereas no rejection imply equation 4
*may be used as the right specification. If the test concludes that equation 5
*is the specification of choice, all tests from 1 to 3 above needs to be
*performed again on the new specification. And if heteroskedasticity is present
* a GMM 3SLS estimation method must be used to obtain the final parameters.

*********************************************************************
*Analysis by different sub group
*********************************************************************
*generate indicator variable for different income groups
*First generate percapita expenditues and then generate the 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

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

```
*for GMM 3SLS estimation too, one can add the prefix <bysort incgrp:> before
*the command gmm and obtain results by each income group.
log close
```

7.4 Archivo .do de Stata para calcular el efecto de empobrecimiento del consumo de tabaco

```
*=====================================================================
* Date : November 2018
* Topic: Stata do-file made as part of the toolkit on Using Household
*        Expenditure Surveys for Economics of Tobacco Control Research
* This do-file estimates the impoverishing impact of tobacco use
* Data base used: DataHH.dta
* Key variables:
* - exptotal - total household expenditures in local currency units (LCU)
* - exptobac - total household tobacco expenditures in LCU
* - exphealth - total household healthcare expenditures  in LCU
* - hsize - household size
* - hweight - survey weights
* - npl - National poverty line in local currency units
*=====================================================================

clear
version 15
set mem 1000m
set more off

*change the directory paths below to inform stata where the data are
*stored and where output is to be stored
global pathin  "C:\Data\"
global pathout "C:Data\poverty"

capture log close
log using $pathout\poverty.log, replace
use $pathin\DataHH.dta

*following loop generate per capita expendituers and label them
foreach X in total tobac health{
          gen pce`X'=exp`X'/hsize
          label var pce`X' "percapita expenditure of `X'"
          }

*SAF is Smoking (tobacco use) attributable fraction estimated externally
scalar SAF=0.2
replace pcehealth=pcehealth*SAF
*If SAF for SHS exposure is available, instead multiply the pcehealth
*variable with the sum of both SAFs
```

```
*preparing variables for analysis
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


*generating an indicator variable for poverty
gen povdum = 0
replace povdum = 1 if pce <= npl
proportion povdum [fw = pweight]

*the following user written module also gives identical result for HCR
*along with other poverty measures. To use this, first apply the following
*command without the star.
*ssc install povdeco, replace
povdeco pce [fw=pweight], varpline(npl)


*Code for computing changes in HCR and number of poor in one shot
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
*the following lists the results with special formating options
matlist M, cspec(& %12s | %5.4f & %9.0f &) rspec(--&&-)

log close
```

tobacconomics

Economic Research Informing Tobacco Control Policy

INSTITUTE FOR HEALTH RESEARCH AND POLICY

UIC

www.tobacconomics.org
@tobacconomics