tobacconomics

Economic Research Informing
Tobacco Control Policy

Updated Toolkit on

Using Household Expenditure Surveys for Research in the Economics of Tobacco Control

2023 Edition

INSTITUTE FOR
HEALTH RESEARCH
AND POLICY

UIC

Suggested Citation: John R.M., Vulovic V., Chelwa G., Chaloupka F. (2023). Updated Toolkit on Using Household Expenditure Surveys for Research in the Economics of Tobacco Control. A Tobacconomics Toolkit. Chicago, IL: Tobacconomics, Institute for Health Research and Policy, University of Illinois Chicago. **www.tobacconomics.org**

Authors: This Toolkit was written by Rijo John, PhD, Associate Professor (Adjunct), Rajagiri College of Social Sciences, Kerala, India; Violeta Vulovic, PhD, Senior Economist, Institute of Health Research and Policy, University of Illinois Chicago; Grieve Chelwa, PhD, Director of Research, Institute on Race, Power & Political Economy, The New School, New York City; and Frank Chaloupka, PhD, Professor Emeritus at the Institute for Health Research and Policy, University of Illinois Chicago. It was peer-reviewed by Martin Gonzalez-Rozada, PhD, Universidad Torcuato Di Tella, Buenos Aires, Argentina; and Guillermo Paraje, PhD, Professor, Business School, Universidad Adolfo Ibáñez, Santiago, Chile.

This Toolkit is funded by Bloomberg Philanthropies.

About Tobacconomics: Tobacconomics is a collaboration of leading researchers who have been studying the economics of tobacco control policy for nearly 30 years. The team is dedicated to helping researchers, advocates and policymakers access the latest and best research about what's working—or not working—to curb tobacco consumption and the impact it has on our economy. As a program of the University of Illinois at Chicago, Tobacconomics is not affiliated with any tobacco manufacturer. Visit **www.tobacconomics.org** or follow us on Twitter **www.twitter.com/tobacconomics**.

Improving Our Toolkit: The Tobacconomics team is committed to making this toolkit as clear and useful as possible. We would like your feedback on whether you found this toolkit useful in your research and, if so, we would appreciate learning about your experience on any successful implementation. We would also like to hear whether you have encountered any issues in applying the methodologies presented in the toolkit, and your thoughts on how we could improve it.

For any comments or questions about the toolkit and its content, please email us at info@tobacconomics.org. We very much look forward to hearing from you.

Table of Contents

List of tables

Introduction

1

Tobacco use is the most prevalent preventable cause of death and a main risk factor for several noncommunicable diseases, resulting in more than 7.2 million annual deaths globally.[1] Worldwide, 12 percent of all adult deaths (30 years of age and older) are attributed to tobacco (16 percent among men, 7 percent among women) according to the World Health Organization (WHO).[2] If current smoking patterns persist, tobacco is expected to kill approximately one billion people globally this century, mostly in low- and middle-income countries (LMICs)[3] where both the prevalence and extent of tobacco consumption are relatively high.[4] The total economic cost of smoking (from health expenditures and productivity losses together) amounted to US$ 1.4 trillion in 2012, or 1.8 percent of the world's annual gross domestic product (GDP).[5] The global health and economic burden of tobacco use is increasingly borne by LMICs.

Unabated consumption of tobacco in various forms hinders economic development and growth, especially in LMICs. The resulting morbidity and mortality from tobacco use negatively impacts productivity, reduces disposable income, and pushes families into poverty. The 2030 Agenda for Sustainable Development, adopted by the United Nations General Assembly[6] in 2015, explicitly recognizes the need to strengthen the implementation of the WHO Framework Convention on Tobacco Control. Regulating tobacco use with meaningful public health policies is important not only to address growing concerns of noncommunicable diseases, but also to improve economic growth and reduce poverty. A substantial body of literature from studies conducted in both high-income countries (HICs) and LMICs concludes that effective policy interventions are available to reduce demand for tobacco products and that these policies are highly cost-effective.[4]

The economics of tobacco control has become an integral part of the development discourse, and yet there is a paucity of academic economists undertaking research in the area of economics of tobacco control, especially in LMICs where the need for such research is relatively high. This may be due to several reasons including scarcity of reliable data and/or lack of necessary expertise to carry out such research. Although research exploring the impact of tobacco control in LMICs is rapidly growing,[4] there is still a need to generate more local- and country-level evidence to support tobacco control policy making, especially in LMICs.

1.1 Purpose of this toolkit

The primary purpose of this toolkit is to guide researchers interested in carrying out research on the economics of tobacco control, especially in the LMICs where household expenditure surveys (HES) on consumption of different tobacco products exist. It is a revised and updated version of an existing toolkit by the same name published by Tobacconomics, based at the Institute for Health Research and Policy, University of Illinois Chicago.[7] This edition of the toolkit expands Chapter 3 by adding the estimation on extensive margin as well as the estimation of elasticities by socio-

economic groups. In addition, Chapters 4 and 5 are updated with most recent studies. Finally, technical guidance in all chapters was updated to incorporate comments and questions received from various users of the first edition of the toolkit.

Unlike in HICs, longer time-series data are often difficult to obtain in several LMICs and, as a result, it becomes difficult to examine the impact of certain policy interventions. For example, if good time-series data on prices and consumption of cigarettes were available, it would be possible to estimate how tax policies impacted prices and, in turn, consumption of cigarettes. However, even in the absence of long time series, it is still possible to conduct several policy-relevant analyses for the purpose of tobacco control policy making using cross-sectional data from household surveys. Most LMICs conduct household surveys sporadically on a variety of topics, and these can provide useful insights into consumer behavior with respect to tobacco consumption.

This toolkit reviews select economic tools and techniques that can be used to analyze HES data with the sole purpose of aiding research on the economics of tobacco control. It demonstrates the use of HES to estimate some of the important issues in the economics of tobacco control including the estimation of own- and cross-price elasticities as well as expenditure elasticity for tobacco products, elasticity on the extensive and intensive margins, the impact of tobacco spending on intra-household resource allocation and consumption of specific groups of commodities within a household, and the impact of tobacco spending and associated health care expenditures on national poverty head counts all for the general population as well as for populations in different socioeconomic groups. This toolkit briefly discusses the literature, theoretical background, economic rationale of each of these issues, methods of estimation, and the use of the statistical software Stata® to implement these methods.

This is one of several toolkits developed by the World Bank, WHO, and Tobacconomics that provides guidance for conducting an economic analysis of tobacco demand and the impact of tobacco consumption on employment, equity, illicit trade, and economic costs. This is also the first in a series of Tobacconomics toolkits designed to build capacity and core competencies in economic analysis of tobacco taxation to support advancing the economic arguments for—and countering the arguments against—tobacco tax increases.

1.2 Who should use this toolkit

The discussion in this toolkit does not presume knowledge on tobacco taxation nor economics of tobacco control issues on the part of the reader. However, some background in economics and econometrics, with a basic understanding of the econometric software Stata, is required to make better use of this toolkit and carry out independent studies in the area of economics of tobacco control research.

While the discussion of econometric methods and the step-by-step guides with Stata would directly benefit researchers working on the economics of tobacco control, the policy discussions and rationale of different economic concepts in tobacco control and the interpretations of results provided in this toolkit are also intended to benefit policy makers, analysts in government agencies, and those in civil society organizations to help them better understand some of the economic issues around tobacco control.

1.3 How to use this toolkit

This toolkit provides technical guidance on three important topics in the area of economics of tobacco control: (i) estimating own- and cross-price elasticities (Chapter 3), (ii) estimating the crowding-out nature of tobacco spending (Chapter 4), and (iii) quantifying the impoverishing effect of tobacco use (Chapter 5). Each topic is discussed with the intention of performing analysis with HES data.

The discussion in each chapter starts with an introduction and the principles behind the topic along with the rationale for analysis. This is followed by a brief technical discussion on the econometric methods used. The discussion of econometric methods is kept to a minimum, as the same is available elsewhere from standard econometric textbooks and other published sources. References to necessary reading are provided to assist readers in gaining additional knowledge on the theoretical concepts presented.

Once the methods are presented, they are followed by a brief discussion on preparing data for analysis and then the different steps involved in doing the analysis in Stata, along with the necessary Stata code. A case study on the topic from a country or using a hypothetical data set is presented along with the interpretation of results towards the end of each chapter.

The toolkit discusses the relevant analysis methods for all tobacco products combined, for smoked and smokeless tobacco products separately, and for individual tobacco products (such as cigarettes, bidi, and other chewing tobacco products) depending on the issue being addressed. For example, when estimating own- and cross-price elasticities, it may be useful to present analysis for each of the tobacco products to facilitate the estimation of not only the own-price elasticity of different tobacco products but also the cross-price elasticity, showing the substitution and complementary patterns between tobacco products such as bidis and cigarettes or smoked and smokeless tobacco. On the other hand, when estimating the impact of tobacco spending on intra-household resource allocation, rather than conducting an analysis by different product categories it may make more sense to combine all tobacco products into one category and examine the impact across different socioeconomic groups.

The toolkit is organized as follows: Chapter 2 provides an introduction to HES with a focus on surveys in LMICs. It discusses the contents of HES as it pertains to tobacco. In particular, it covers various questions pertaining to tobacco consumption and expenditures on different tobacco products of inquiry in HES. The chapter also briefly discusses some of the econometric issues one needs to be aware of while working with HES and Stata code for extracting data from raw HES, among others. In addition, the chapter presents some useful tips on working with Stata software.

Chapter 3 discusses the methods of estimating own- and cross-price elasticity for different tobacco products. It presents methods to estimate both prevalence and intensity elasticity for tobacco products. The primary method for estimating intensity elasticity (elasticity on the intensive margin) is the one developed by Deaton,[8] which is presented along with a step-by-step explanation of the Stata commands for estimating price elasticities from HES data. This is followed by a discussion of estimating prevalence elasticity (elasticity on the extensive margin) along with a step-by-step explanation of Stata commands for implementing it with HES data. A discussion of estimating price elasticities by income groups can also be found in this chapter.

Chapter 4 explains methods to examine the impact of spending on tobacco on intra-household resource allocation. Following an approach of conditional demand systems,[9, 10] this chapter shows how expenditures on tobacco systematically crowd out expenditures on other commodities within a household. The analysis discusses ways to estimate crowding out according to different socioeconomic subgroups. The analytical method, as well as the Stata code for executing the model, are presented as well.

Chapter 5 covers the impoverishing effect of spending on tobacco. It discusses the estimation of the actual amount spent on purchasing tobacco as well as the increased health care expenditures attributable to consumption of tobacco and secondhand smoking (SHS). It then demonstrates how accounting for tobacco spending and associated health expenditures impacts the estimate of national poverty measured by the head-count ratio. Step-by-step estimation along with relevant Stata code are presented.

As much as possible, these chapters also discuss empirical results from other countries where such studies have been done using HES.

The individual Stata commands used in different chapters are placed in angle brackets "< >" and italicized. However, the command itself must be used without those brackets. The variable names used in different examples are all italicized. Specific examples demonstrating use of certain Stata code are placed in separate text boxes in different chapters. A Code Appendix also includes Stata code relevant to the respective chapters in separate do-files.

An introduction to household expenditure surveys

2

2.1 Availability of household expenditure surveys

Household surveys have been conducted in several countries for a very long time. The first consumer expenditure survey by the Bureau of Labor Statistics (BLS) in the United States (US), for example, was conducted in 1888. Although relatively new, the National Sample Survey (NSS) organization in India started its household consumption surveys as early as the 1950s[11] and has conducted regular and periodic surveys every few years since then. The Living Standards Measurement Study (LSMS), the World Bank's flagship household survey program, has existed since the 1980s. These multi-topic household surveys have collected household consumption expenditure from more than 40 countries around the world so far,[12] including several countries in Africa and Asia. There are many countries—both high- and low-income—that conduct household expenditure surveys, and many of them conduct these surveys at regular intervals.

The International Household Survey Network (IHSN), an informal network of international agencies which strives "to improve the availability, accessibility, and quality of survey data within developing countries, and to encourage the analysis and use of this data by national and international development decision makers, the research community, and other stakeholders,"[13] maintains a portal for researchers to browse and download census or survey documents and metadata from as many as 201 countries; it currently has nearly 7,000 surveys catalogued. About 137 out of the 201 countries for which data are available are LMICs. This catalog is accessible at http://catalog.ihsn.org/index.php/catalog and includes information on more than 1,000 HES in its database, of which about 700 are from LMICs.

In the absence of long time series macroeconomic variables, HES provide meaningful cross-sectional data, sometimes for multiple time periods for the same country. Statistical agencies that undertake the HES in most countries, however, usually publish summary reports presenting only grouped data that are freely disseminated to the public. The grouped data, although helpful in examining the overall picture, does not provide an adequate sample size to undertake the major econometric analyses discussed in this toolkit. Therefore, to conduct advanced econometric analyses with the survey data, it is important to have access to the microdata (individual, household, or unit records) from the surveys.

These microdata are often not freely available for public access. However, such data are usually available directly from the government statistical agencies in charge of conducting the surveys by paying a nominal fee. After paying the fee per the agency's website, the data may be received in digital form either by downloading directly from the agency's website or by mail on a data storage

device. Some agencies allow data download after registration and a brief description of the project. The microdata from LSMS from different countries, for example, are freely available to download from the World Bank website after signing up and providing a brief summary about the project.

2.2 Content of household expenditure surveys

The simplest household surveys collect data on a national sample of households randomly selected from a "frame" or national list of households (often a census), and an equal probability is assigned to each household selected from the frame. Although sample sizes vary widely depending on the purpose of the survey, the population size in the country, and the need for generating subsample estimates, sample sizes of around 10,000 are frequently encountered, corresponding to a sampling fraction of 1:5000 in a population of five million households.[8]

In practice, a two-stage design is often implemented in the selection of households, wherein, at the first stage, selection is made from a list of "clusters" of households—usually villages in rural areas or urban blocks in urban centers—and in the second stage, households are selected from each cluster.[8] Clusters are typically called the first-stage units (FSU) or primary sampling units (PSU), as they are the first unit sampled in the design. If the clusters are randomly selected with probability proportional to the number of households they contain, and if the same number of households is selected from each cluster, it would be as if each household has the same chance of being included.

Depending on the objectives of the survey, a sample may be designed so households can be selected based on relevant attributes such as geographical area, ethnic affiliation, standard of living, gender, or race so that households in a certain group can have a certain probability of being selected. Such stratification effectively converts a sample from one population into a sample from many populations, thus guaranteeing enough observations to permit estimates by these subgroups.[8]

The probability weights for households in each stratum might differ. In most cases, there may be few PSUs or clusters within each stratum. Indian NSS, for example, focuses on stratification by rural and urban areas within a district for its consumer expenditure surveys. While stratification typically enhances the precision of sampling estimates, clustering of the sample will usually reduce the precision, as households within the same cluster are more similar to each other and hence reflect low variability.

Household surveys, by their very nature, provide information on households and the individuals within. Although the definition of household used in each survey can differ depending on the structure of living arrangements in each country, by and large those members who live together and eat together are considered to be part of the same household. The HES typically provide data on consumption, income or assets, and demographic characteristics of households including household composition, household size, age and gender of household members, educational attainment and employment status of household members, ethnicity and race, among others.

To assess consumption, HES measure expenditures incurred and/or quantity consumed by households on different goods and services over a pre-specified reporting period also known as a recall or reference period. Although rare, some HES—for example, the Consumer Expenditure Survey (CES) by the BLS in the US—also collect expenditure data at the individual level. In the case of adult goods like tobacco, such data would be immensely useful.

Depending on the objective of the survey and characteristics of the goods or services in question, the recall period may vary significantly for different goods within the same survey and for the same goods across different surveys; it can range from as low as one day to a period of one year. However, common items of consumption in most HES have a recall period of one week to one month. The Household Income and Expenditure Survey (HIES, 2016) in Liberia, for example, collects food consumption data with a seven-day recall and non-food consumption within both seven-day and 30-day recalls.[14]

As part of the task of collecting data on the expenditures incurred and quantity consumed of different goods, several HES collect information on the consumption of different tobacco products commonly used in the respective countries. The Indian NSS, for example, collects both quantity of consumption and expenditures spent on cigarettes, bidi, and smokeless tobacco varieties over both 30 days and seven days prior to the interview. This provides a rich source of information to aid in examining several economic issues on tobacco consumption. This level of disaggregation, however, may not be available in all HES.

Depending on the resources available to survey agencies, sometimes expenditures are reported for commodities aggregated to larger groups, such as tobacco and intoxicants as a single group. Some HES, on the other hand, provide only expenditure information and do not collect quantity information for several consumption items. As a result, there can be challenges in econometric analysis between different data sets.

Using other household-specific characteristics and regional information given in household surveys, it is often possible to classify the households in a survey into different socioeconomic status (SES) groups so that economic analysis can be performed by SES group. Such analysis may be done based on the educational attainment of households, income or asset status, place of residence like rural or urban areas, ethnic affiliations, or based on the standard of living for a household, among other criteria.

2.3 Econometric issues while working with household surveys

Due to the design characteristics of household surveys discussed in the previous section, there are specific challenges for econometric analysis. A detailed exposition of these challenges is offered in Chapter 2 of *The Analysis of Household Surveys* by Deaton.[8] A brief and conceptual summary of the salient issues follows.

2.3.1 Using survey weights for descriptive statistics

Depending on the purpose of each household survey, some households may be over- or under-represented in surveys and, as a result, the estimated sample mean or other sample statistics will be biased estimators of their population counterparts. Survey weights are often used to re-weight the sample data and adjust for the design elements of the survey to make the estimates representative of the population. Most surveys include the survey weights along with the published data and can be used straight away, as-is, while generating the necessary statistics.

If the weights are not directly given, the survey documentation would usually include instructions or formulas for computing those weights using relevant variables included in the sample data. It is

important to apply the correct survey weights while generating descriptive statistics from sample data. Section 2 below gives examples of how to apply survey weights in Stata while computing certain descriptive statistics.

2.3.2 Using survey weights in regression

Unlike with descriptive statistics, there is no agreement on the use of survey weights in the context of regressions. The classical econometric argument is against the use of weights in regression; as Deaton[8] points out, when the population is homogeneous so that the regression coefficients are identical in each stratum, both weighted and unweighted estimators will be consistent, and ordinary least squares (OLS) is indeed more efficient by way of the Gauss-Markov theorem.[15] On the other hand, when the population is not homogeneous, both weighted and unweighted estimators are inconsistent anyway and weighting adds no value.

Nevertheless, Deaton[8] goes on to say that a weighted regression provides a consistent estimate of the population regression function, provided that the assumption about the functional form of the regression is correct—that is, when the regression function itself is the object of interest. If the interest is to estimate behavioral models where behavior may be different for different subgroups, weighting in the regression is of no use. In conclusion, as Cameron and Trivedi observe,[16] weights should be used for estimation of population means and for post-regression prediction and computation of marginal effects. However, in most cases, the regression itself can be fit without weights, as is the norm in microeconometrics.

2.3.3 Inflated standard errors due to cluster design effects

As most household surveys use a two-stage design in which clusters are chosen first, followed by households from within each of those clusters, it is often the case that households within the same cluster are quite similar to each other—as they live near one another and are interviewed around the same time—and are different from those in other clusters that are usually widely separated geographically. In other words, there will be more homogeneity within clusters than between them.

To the extent that observations or households within a cluster are not fully independent, the positive correlations between these observations could potentially inflate the variance above what it would be if they were independent. Hence, it is important to correct the estimated standard errors in regressions based on household surveys to account for these cluster design effects using appropriate techniques.

2.3.4 Heteroskedasticity of OLS residuals

Distributions of households over different variables of interest, such as income and consumption of different goods, usually are not distributed normally. As a result, it is quite common to find heteroskedastic disturbances in regression functions estimated from HES data.

The heterogeneity between different clusters could also result in regression functions returning heteroskedastic error terms. Such errors would leave the OLS estimates inefficient and would invalidate the usual formulas for standard errors. Thus they would need to be corrected using appropriate correction methods.

Combined with the presence of cluster design effects, it is important to use formulas that correct

standard errors in survey-based regressions that account for the presence of heteroskedasticity as well as cluster effects.

2.3.5 Endogeneity

This refers to situations in a regression when one or more of the explanatory variables is correlated with the error term, resulting in biased and inconsistent OLS estimates. Endogeneity mainly arises due to three reasons:

(i) **Simultaneity**—X causes Y and Y also causes X. In other words, X and Y are jointly determined.

(ii) **Omitted explanatory variables**—when an omitted variable affects one or more of the included independent variables and separately affects the dependent variable. The omitted information contained in those omitted variables may also be referred to as "unobserved heterogeneity," or the unobserved variation across individual units of this omitted or unobservable variable.

(iii) **Measurement errors**—one or more of the explanatory variables are measured incorrectly. Measurement error in a dependent variable does not bias the regression coefficient. Measurement errors in survey data, according to Deaton,[8] are a fact of life.

Although these are often mentioned as separate sources of endogeneity in regression, in reality they need not be truly distinct from each other. Often in regression analysis using survey data, most if not all of these different sources of endogeneity are encountered.

In all the different sources of endogeneity described here, the regression function would differ from the structural model due to the correlation between the error term and explanatory variables, thus violating a crucial OLS assumption. Use of instrumental variables (IVs) (such as two-stage least squares method)[15] is the standard technique in such circumstances, provided it is possible to find IVs that are correlated with the explanatory variables but uncorrelated with the error terms so that the regression yields consistent estimates.

2.4 Useful tips on Stata

Stata, a widely used statistical package, is an econometric and data analysis software preferred by many universities and institutions around the world, thereby facilitating exchanges and collaborations between researchers in multiple disciplines and institutions.[17] Below are some useful tips that make working with Stata much easier.

2.4.1 Creating a do-file

Stata can be used through its pull-down menus from the user interface, by directly issuing commands in a dedicated command window, or with the help of a do-file, which saves all commands for execution at will. Execution by do-file is the preferred and recommended method as it offers several advantages over the other methods. A do-file simply records all the commands to be executed and saves it in a file for future use with the extension ".do".

The main advantage is that the analysis can be replicated with the commands saved in the do-file and the work can be shared and edited by other collaborators. But, more than anything, a do-file keeps a record of work done and enables revision of the commands as needed. Unlike command windows or pull-down menus, in a do-file one can also add notes and comments for other

collaborators, which facilitates seamless collaboration. Useful information on how to create a do-file can be found on the Stata website (https://www.stata.com/manuals13/u16.pdf).

2.4.2 Creating a log file

While a do-file keeps a record of all the commands and allows editing them as needed, a log file with the extension ".log" or ".txt" keeps a record of commands executed along with their results during a given Stata session. It is helpful to create log files while running the do-file so the results are available for future reference or to share with collaborators.

A log file is created within the do-file using a command *<log using mylog.log, replace>*. This will create a file with the name "mylog.log" in the present working directory of Stata. The optional argument *<replace>* will make sure that each time the do-file is executed the contents of the log file are replaced with the new results. One may also use the option *<append>* to keep adding the results of all commands to the same log file. Before closing the section, usually done towards the end of the do-file, close the log file with the command *<log close>*. The use of the log file can also be temporarily suspended and resumed through commands such as *<log off>* and *<log on>*.

2.4.3 Using knowledge resources

All user manuals for Stata are built into the software. One can simply issue the command *<help>* followed by the particular Stata command to learn the description, syntax, and examples of every command used in Stata. For example, *<help regress>* will return the necessary syntax, description, and examples of using the regress command. In addition, *<search>* and *<findit>* commands return very useful information on topics of interest within Stata. For example, the command *<search survey>* would return a list of commands and modules Stata uses to analyze survey data. Stata also has an excellent support forum, which is a rich resource for learning and familiarizing oneself with Stata (https://www.statalist.org/forums/).

2.4.4 Setting a working directory

While working on the household survey data, it is better to make a copy of the microdata and move it to a dedicated directory on the computer. All subsequent Stata program files and other related documents for the analysis can be stored in the same directory while leaving the original microdata untouched. The command *<pwd>* lists the current working directory of Stata irrespective of the operating system. This working directory can be changed with a command *<cd "Path">* where Path, within double quotation marks, is the directory path where work is saved; that would differ depending on the operating system.

Once a working directory is set, the subsequent commands that call files (such as data files, do-files, dictionary files, etc.) can be issued using only the filename without the whole directory path. This also has the advantage that a collaborator only needs to change the working directory once and need not change the file paths mentioned in different parts of the do-file while executing a do-file.

Alternatively, one can set a global macro to assign a directory for storing the data and saving work. Thereafter, simply call the macro name instead of repeating the whole directory structure to use the data or save something. For example, in Windows, use the command *<global pathin "C:\Data\HES">*. Later on, to import data stored in this directory from within the do-file, use the command

<use $pathin\filename.dta> and Stata will automatically look for the data file in the directory defined in the global macro *pathin*. The directory path structure varies depending on the operating system. Use of macros is discussed in greater detail later.

2.4.5 Practicing with example data sets

Stata provides two types of data sets for the purpose of demonstration and practice. They are: (a) example data sets installed with Stata in a local machine and (b) online data sets that are referred to in the Stata documentation and accessible online. From Stata's user interface, navigate to "File > Example data sets", and lists of available data will appear. Click on those data sets and open them inside Stata to practice.

Alternatively, use the command *<sysuse datafile>* where datafile refers to the filename of the particular data set in the system, if the names of the data sets are known. One can also use the command *<webuse datafile>* to load a specified data set, obtaining it over the web. The data sets are obtained from **http://www.stata-press.com/data/r17/**. This link also provides a detailed list of data sets arranged by topic, and one can browse through available data sets to be used for practice.

2.4.6 Using logical and relational operators

Stata uses several logical and relational operators to help work with data sets. Some of the commonly used operators and their intended meanings are given here.

&	*And*	*|*	*Or*
!	*Not*	*~*	*Not*
>	*Greater than*	*<*	*Less than*
>=	*Greater or equal*	*<=*	*Lesser or equal*
==	*Equal*	*!=*	*Not equal*

Apart from these, Stata also has operators to handle categorical variables (also known as factor variables or dummy variables). Prefix a variable with (i.) to specify indicators for each category of a variable. This works well instead of creating separate dummy variables. The command *<fvset base>* can be used to set the base category. Enter (#) between two factor variables to create an interaction variable. Enter (##) to specify both the main effects for each variable and their interactions. Similarly, (c.) can be used to interact a continuous variable with a categorical variable by prefixing the continuous variable with (c.).

For example, assume *age* and *sex* as factor variables and body mass index (*bmi*) as a continuous variable. To regress the effects of these variables on blood pressure (*bp*), the following regressions produce the same result: *<regres bp i.age age#sex>* and *<regres bp age##sex>*. Alternatively, to regress blood pressure on age and body mass index and the interaction between them, write *<regres bp age##c.bmi>*.

2.4.7 Using macros

Macros are abbreviations or "aliases," which have both a name and a value. When its name is de-referenced, it returns its value.[18] Hence a macro has a macro name and macro contents. Everywhere the macro name is used in the program with punctuation, the macro contents are substituted in its place. Macros are used for several purposes including making tasks simpler, making do-files more organized, shortening the length of Stata code, and various other conveniences while programming. A macro can be one of two types, local or global, depending on its scope—that is, where its existence is recognized. Global macros, once defined, are available anywhere in Stata, while local macros exist solely within the program or do-file in which they are defined.[19]

To substitute the macro contents of a global macro name, the macro name is punctuated with a dollar sign ($) in front. Similarly, to substitute the macro contents of a local macro name, the macro name is punctuated with surrounding left and right single quotes (‘’).[19] For example, define a local macro with the name indvar as *<local indvar price expenditure hsize>* and issue another command *<summarize 'indvar'>* to obtain the summary statistics for each of the variables' price, expenditure, and hsize in the results.

Similarly, define a global macro as *<global xyz age income sex>* and issue the command *<summarize $xyz>* to obtain the summary of each of those variables: age, income, and sex. As global macros may create conflicts across do-files, they are rarely used. Local macros are usually preferred while writing the code in the do-file.

Macros can also be defined as an expression, and the result becomes the contents of the macro. For example, define *<local result = 5+5>* and the command *<display 'result'>* would return 10. Macros are also able to offer extended functionalities with macro extended functions. Use the command *<help macro>* to learn more about macros and their varied and creative uses.

2.4.8 Using loop commands

Loops are commands in Stata that help to loop over an arbitrary list of strings or numbers. For example, a loop command can repeatedly set a local macro name to each element of the list and execute the commands enclosed in braces "{}". Loops are quite useful and convenient while performing repetitive tasks that are done sequentially, and they are extensively used while programming. Stata's *<foreach>* and *<forvalues>* commands are particularly useful for looping. These loop commands begin and end with braces "{" and "}" in separate lines. The open brace must appear on the same line as *<foreach>* and the close brace must appear on a line by itself in the end. For example,

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

The first line above lists the different variables over which the command has to be repeated (var1, var2, and var3) and the next two lines give the actual commands to be repeated. The first command tells Stata: if an observation for a variable in the list has a value less than or equal to zero, it must be

replaced with a dot. The second one instructs Stata to generate new variables with a variable name starting with *ln* followed by the names of the variables in the list, defined as a natural log of existing variables in the list.

Multiple lines of commands could be added one below the other, all of which will be repeated over all of the variables mentioned in the first line. The code above can also be executed more efficiently using local macros. For example, predefine a local macro *<local varlist var1 var2 var3>* and use the loop:

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

Stata can also perform such loop commands over different files at a time. Similarly, *<forvalues>* command can be used to perform similar operations applied to numbers. For example, suppose there are 25 states in a household survey, and the average consumption expenditures in each state are under variable names "*state1, state2, … state25*". To convert all those variables to logarithmic form, use the command:

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

The *I* in the first line of the *<forvalues>* command refers to the local macro inside the loop.

2.4.9 Returning stored results

Stata regularly stores results from commands in local macros that can be called in for various purposes. For example, issuing a *<summarize>* command for a variable *<sum varname>* will return descriptive statistics on the variable *<varname>*. Simultaneously, it also stores those results in local macros. For example, *<summarize mpg>* from the auto data in Stata returns the results below.

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

Issue the command <return list> after this, and it will give the results as shown in the table.

```
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
```

The results are all stored in different local macros. These are available to be used immediately afterwards to generate new variables or to be used in other commands. Similar to *<return list>*, use the command *<ereturn list>* to show locally stored contents after estimation commands such as *<regress>*. The command *<help return>* in Stata will show other uses of return commands. Box 2.1 above provides a working example using some of the Stata tips already covered.

Box 2.1 Stata example tip

```
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' <.
}
```

The code demonstrates the use of macros, loop, and stored results, all in one place. The first line imports the built-in "auto" data, and the second defines a local macro called *items*, which consists of three variables. The third opens a loop command *<foreach>* and uses the local macro along with it. There are three instructions that are executed successively on all three variables in the next three lines through this loop.

- The first quietly summarizes the variable, and with the addition of the prefix "quietly" it executes this command without displaying the results. The option *<detail>* after *<summarize>* requests additional statistics that are not usually calculated, such as percentiles, skewness, and kurtosis.
- The second line in the loop defines a new local macro *upper*, using the stored results after *<summarize>*. It is defined as the mean plus three standard deviations of the variable under consideration.
- The third line in the loop replaces any values higher than the mean plus three standard deviations and less than missing values—Stata considers missing values to be larger than any numeric value—with the median value of that variable. The brace in the last line ends the loop.

2.4.10 Using delimiters

The command *<#delimit ;>* is used to reset the character that marks the end of a command in Stata. These are used only in do-files and ado-files (defined in the next section). Hitting the return key instructs Stata to execute the command. In a do-file, the end of a line assumes the return key and these lines themselves have character restrictions. So, one can instruct Stata that the commands are longer than one line by using the command *<#delimit ;>* to freely break the command lines as necessary. Stata will consider all lines continuous until it sees the delimiter character that marks the end of the command as a single logical line.

Alternatively one can use < /* */ > as a comment delimiter. For example, *<generate X = 3*Y /* this is a comment*/ + 5>* is the same as *<gen X = 3*Y + 5>* without the comment. One may also break long lines with three consecutive forward slashes (*///*), instead of using the command *<#delimit ;>*. These are quite useful while preparing do-files. For example, Stata considers the following command as a single logical line:

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

2.4.11 Using add-on commands

Stata allows people to write third-party commands (called "ado-files") that can be stored in a Statistical Software Components (SSC) archive, which is often called the Boston College Archive and is provided by **http://repec.org**. From the SSC archive, users can install these add-on programs using the command *<ssc install progname>* where "progname" is the name of the ado-file or program file that needs to be installed. A particular package may also be uninstalled with the command *<ssc uninstall progname>*.

Most add-on packages provide additional functionality compared to built-in Stata commands. For example, the add-on package *<estout>*, which can be installed with *<ssc install estout>*, helps make neat tables from stored estimates after regression commands. It can create publication-worthy tables with coefficients from regression, adding stars to indicate their significance level, summary statistics, standard errors, *t*-statistic, *p*-values, and confidence intervals for one or more models fitted earlier and stored by the command *<estimates store>*.

Similarly, *<findname>*, *<outreg2>*, and *<ivreg2>* are some of the popular add-ons. These can be installed onto Stata using the command *<ssc install outreg2>*, for example. Outreg2 helps produce publication-ready tables from regression output. Use the command *<ssc whatshot>* to check out some of the most popular add-on packages available for download.

2.4.12 Miscellaneous tips

Some miscellaneous tips not mentioned above are included here:

- Stata commands and variable names are case sensitive. For example, if a lowercase letter is used in place of uppercase, it will return an error or execute an unintended code.
- Most Stata commands can be abbreviated. For example, *<summarize>* can be abbreviated as *<sum>* or *<su>*. Instead of *<regress>* use *<reg>*, and so on.
- The name given to scalars within the do-file should be distinct from any of the other variables or their unambiguous abbreviations present in the data. If a scalar is defined with the same name as another variable or its unambiguous abbreviation, Stata will prioritize the variable name or its abbreviation over the scalar name specified, leading to inadvertent results while doing operations involving this scalar. Alternatively, use a pseudo function *<scalar(xyz)>* to spell out a scalar with the name "xyz" every time the scalar is to be used in a calculation or while defining more scalars.

- Missing values, denoted by a dot (.) are coded and treated as positive infinity in Stata, meaning they take a value higher than all other numeric values. This is important while cleaning the data. For example, *<replace X = 0 if Y>100>* will replace X with zero not only if it is greater than 100, but also if there are any missing values in Y. Instead, use *<replace X = 0 if Y>100 & y<.>*

2.5 Techniques for extracting data using Stata

The microdata from household surveys are stored in different file formats depending on the hardware used to record the data, availability of software within the survey agencies, and other standard practices and customs in different fields. The HES data that are of interest in this work will usually be quantitative tabular data. They are often presented in delimited text files containing meta information such as those found in statistical software Stata, SPSS, and SAS, or in simple comma-separated values files (.csv), tab delimited files (.tab), or in fixed ASCII format with either ".ascii", ".dat", or ".txt" file extensions.

If the data are in fixed ASCII format, which is often the case, there will be an associated dictionary or layout file that describes each column in the data file of fixed record lengths. For example, the dictionary would say: byte position 4 in the data file indicates the code for rural or urban area, byte positions 9–10 indicate the code for PSU or cluster identifier, or byte positions 30–36 indicate the expenditures on an item.

There will also be a file, usually called a codebook, which indicates the meaning of different code used in the layout file or data file. For example, it would indicate that value 1 = rural and 2 = urban, or 1 = male and 2 = female.

The final data that are archived by the respective survey agencies usually provide all the necessary documentation associated with the data. The IHSN catalogue, 12 for example, includes details on survey methodology, sampling procedures, questionnaires, instructions, survey reports, code used, and dictionary or layout file codebooks for most of the survey data catalogued there.

The software that is used for statistical analysis should be able to import microdata before different analyses can be performed. A detailed description and documentation of the survey data, the structure of data files, and the relationship between different data files in the survey are necessary to make an informed decision on what data should be extracted or imported into the statistical software for further analysis. For generating any estimate from these data, one must extract the relevant portion of the data and aggregate it using appropriate commands in the analytical software. Stata uses different methods to import data depending on the source data file type. Entering the command *<help import>* in Stata's command prompt lists different options and commands available to import data of different formats.

Since the microdata for most HES are in fixed ASCII format, the example below demonstrates a simple way to import the necessary data into Stata. The tables below show part of a typical fixed format data file and the layout file describing the data. The layout file tells what the character in each byte position in the ASCII data file represents. In order to extract or import these data into a readable format in Stata, or convert it to a Stata data set (.dta), a Stata dictionary file with the file extension ".dct" must be created. A sample dictionary file to extract parts of the information given in the ASCII data file is given in Box 2.2.

Example data file in ASCII format (fixed format)

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

Example layout file

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

Box 2.2 Example dictionary file to import data from ASCII files

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

A Stata dictionary file begins with a line that looks like <*dictionary using datafile.txt {*> where "datafile.txt" is the name of the microdata file in the Stata working directory. The definition of individual variables follows next. Each variable is defined by a line with five parts.

Box 2.2 Example dictionary file to import data from ASCII files (cont'd)

The first part tells Stata to begin reading the data file from the byte position mentioned in parentheses. The second indicates the variable type: string or numeric. Only the string variables need to be explicitly indicated as such. The third part is the mnemonic name of the variable. The fourth is the variable input format which consists of a "%" sign, a number indicating the variable width, and a letter indicating the variable format: "f" for numbers and "s" for strings. The fifth part is an optional label given to the variable. The dictionary program ends with a closing brace "}".

Some examples of input formats that may be used in the variable definitions are:

- %5f for a five-column integer variable,
- %10s for a 10-column string variable, and
- %7.2f for a seven-column number with two implied decimal places.

Remember to add a return character at the last line—that is, before saving the file move the cursor to the beginning of the next line below the "}". Finally, the file must be saved with the file extension ".dct" (for example, "dictionary.dct").

To execute the Stata dictionary program, open Stata, set the working directory, and give the command: *<infile using dictionary>* where dictionary is the filename of the dictionary file. If the program runs correctly, the program will appear on the screen followed by the message "N observations read" where "N" indicates the number of observations in the imported data. Next, run a command *<describe>*, which will return the results with the number of observations and variables along with their labels.

Once it is verified that the variables are all in order, issue the command *<compress>* to change variables to their most efficient format. Finally, the imported data can be saved in Stata's native data format extension (".dta") with the command *<save mydata>* where "mydata" is the name of the Stata data file that will be saved in the Stata working directory.

2.6 Preparing and building data for technical analysis

HES often provides multiple data sets for individual records, household records, and other variables. The expenditures for different commodities themselves may be in different data files. Moreover, data may be incorrectly coded for certain variables, and some obvious errors could be corrected easily so those observations are not lost during the final analysis. In addition, there may be some extreme or missing values that need to be accounted for. For all these reasons, it is important to clean individual data files and merge them all into a single file before carrying out further analysis. This section provides some basic steps to undertake before a final data set can be prepared to carry out statistical analysis.

2.6.1 Merging data

Household surveys often come with multiple data files or records for households and individual members within households. Furthermore, there may be multiple records for households themselves. For example, one file may have basic household characteristics such as household size, SES group they belong to, place of residence etc., while another file has their consumption expenditures. The data on consumption expenditures themselves could be distributed across different data files. Therefore, it may be necessary to write separate dictionary files for extracting data from different data files and merge them together after each data set is extracted into separate Stata data files.

Because this toolkit covers household-level analysis, the individual information needs to be aggregated to household level. For example, the sex of an individual is not relevant in a household-level analysis. However, a variable that gives sex ratio (ratio of number of males to females in a household) can be constructed. Similarly, education level of individual members in a household is not relevant for a household-level analysis. However, average years of education received by a household, as a variable for household-level analysis to indicate a household's educational attainment, can be constructed.

Once desirable household-level variables are generated from the individual data records, only a single observation needs to be retained per household before it is merged with household-level data. For example, once a household-level variable—say, sex ratio—is generated from individual-level data, the same value for the sex ratio will be repeated for all household members within a household. To retain only one observation per household, first sort the data by household (or by household IDs) with the command *<sort hhid>* (where *hhid* is the identifying variable for households) and then run the command *<drop if hhid==hhid[_n-1]>*. Alternatively, use the command *<duplicates drop>* after arranging the data as necessary.

Merging household-level data with additional data either from the individual records or from other household-specific records will require use of the *<merge>* command in Stata. Run the command *<help merge>* to see the syntax as well as different ways of merging data files in Stata. Stata generates a new variable *<_merge>*, after each merge command to facilitate checking if merging has been done correctly. It is a categorical variable containing a numeric code indicating the source and contents of each observation in the merged data set. The command *<tabulate _merge>* after execution of a *<merge>* will give the necessary indication. For example, code 3 for "_merge" is for observations correctly matching both data sets.

The most important aspect of merging two different files is finding a set of variables that can uniquely identify every single observation in each of the data sets to be merged. This needs to be understood from the time of survey design and extracted along with every single data extraction using dictionary files. A lack of unique identifiers or incorrectly defined identifiers can result in inadvertently combining information from one household with that of another. Box 2.3 gives an example of identifying these variables and merging files correctly.

Box 2.3 A potential mismatch of households while merging

The Bangladesh Household Income and Expenditure Survey (2010) follows a two-stage stratified random sampling technique. The description of the sample design in the published report says around 200 households each were selected from about 1,000 PSUs across the country, while the PSUs themselves were selected from about 16 different strata. It is clear that a household from this survey should be uniquely identified using the variables representing strata, PSU, and household number. These variables are *stratum, psu*, and *hhold*, respectively, as given in the documentation. Since the PSU numbers themselves are unique in these data, a unique household ID can also be identified using only the variables *psu* and *hhold*.

A unique household ID variable (*hhid*) for these data can be generated with the command <*egen hhid=group(psu hhold)*> where the values in parentheses correspond to the variable names required to uniquely identify the household. For example, if PSU numbers were not unique and varied across strata, all three variables would be needed to generate *hhid*. So, any merging of two household-level records in these data will use these variables. For example, HIES has a household demographics data file (rt001) and a household-level aggregate expenditures file (hhold_exp_hies2010). If the files are to be merged, both data must be extracted separately and saved as Stata data files, for example, with the names, "hh1.dta" and "hh2.dta." After loading *hh1, hh2* can be merged with it using the command <*merge 1:1 psu hhold using hh2*>. This would correctly merge the same households in one data file with those in the other. The command <*tab _merge*> will show how accurately the data files were merged so the user can see there are no mismatches.

On the other hand, suppose a unique *hhid* variable was generated first for each of the data files separately, and they were merged afterwards with the command <*merge 1:1 hhid using hh2*>, where the pre-generated unique ID variable (*hhid*) was used for merging instead of the original household identifiers (*psu* and *hhold*). This will also merge both the data files, and the command <*tab _merge*> will show no mismatches. However, in this case, the households in both data sets could be incorrectly matched due to several reasons:

1. While generating a unique *hhid* in each individual data file, Stata assigns unique IDs to each household using the existing sort order in each data file. If the sort order of both data files were different when the *hhid* variable was generated, it will result in incorrectly matching households after the merge.
2. Suppose some psu or hhold numbers were different in both the data sets due to incorrect coding. The <*tab _merge*> after a correct merge using both *psu* and *hhold* will show mismatched observations. Whereas merging with pre-generated *hhid* would merge both data files perfectly, failing to identify mismatches.
3. Suppose the number of observations in *hh1* and *hh2* were different. A merge with both *psu* and *hhold* variables would correctly match the households, whereas merging with pre-generated *hhid* would match them inadvertently.
4. Therefore, the data from two different data files should be merged only using all relevant variables that are used to identify the unique observations (household or person) in each data file. In other words, the <*merge*> command should have all variables that uniquely identify an observation present while merging.

To do a one-to-one merge, both the "master data" as well as the "using data" should be identifiable with the same set of unique variables. Further analysis can only be performed with those observations that matched from both the "master" and "using" data files—that is, observations for which the variable (_merge) takes the value of 3. In order to use only variables with no missing data from both the master and using data files, it is important to drop observations for which "_merge" is not equal to 3 using the command <*drop if _merge!=3*>. However, there may be situations in which it is necessary to retain in the merged data file those unmatched observations from either the master data or the using data file.

Apart from merging different files (such as household data and individual data) from the same round of a given HES, there may also be situations where the user wants to pool HES data from different years or waves. Obviously, the households in different rounds of HES may be different from each other, and what is required is not a merging but pooling of different HES to create a pooled cross section. In this case, instead of <*merge*> one should use the <*append*> command in Stata. To do this, data from each round of HES need to contain the same type of variables and a single merged data set for each round of HES needs to be prepared first. Once the appending is done, it will simply add to the number of observations in the master data.

Before appending, it is important to create a year or wave variable and mark it with numbers that can identify each year/wave/round of the survey. If the final pooled data belong to multiple years (usually from different waves of the survey), it is also important to adjust for inflation any expenditure or price variables so the values across different rounds of data are in constant terms and are therefore comparable.

2.6.2 Reshaping data

Depending on the analysis one undertakes, it may be important to reshape the data into long format or wide format in Stata. To do this, run the command <*help reshape*> to understand how reshaping from one form to the other is done. In a wide format, there will only be as many observations as the number of unique households in a data set, whereas, for a long data format the same households may be repeated multiple times, stacked under one another. For example, assume there is information on the expenditures on cigarettes as well as smokeless tobacco. For households with expenditures on both products, there will be two observations for each household under a long format, whereas under the wide format expenditures on cigarettes and smokeless tobacco will appear as separate variables against a single observation of the household.

For most analyses, it is useful to have the data reshaped in wide format. So, if the extracted data are in long format, they should be reshaped to a wide format using the command <*reshape wide stub, i(i) j(j)* >, after determining the logical observation (i) and the sub-observation (j) by which to organize the data.

2.6.3 Cleaning data

Cleaning data before performing statistical analysis is essential, especially in the case of household surveys, as these are data collected by different people across the country in different stages. For example, a zero in the place of a missing value could result in generating undesirable results, such as distorting the mean and variances when doing statistical analysis. Similar errors in data are: duplicates, erroneously coded categorical variables, and unacceptably high or low outlier values for certain variables.

Similarly, if a string variable has different spellings or space characters between observations, Stata would consider these entries as a different category. For example, if male under the variable sex is coded as "Male" or "MALE" or "M" or "male" or other possible variations, then instead of getting MALE and FEMALE as two different categories, there may be several different categories. For these and other reasons, it is important to do a thorough examination of each of the variables and make sure the data are consistently coded.

Table 2.1 provides a good sequence of steps that can be taken to obtain a clean data set, including useful Stata commands that can be used during these steps. Please note that the steps mentioned in the table need not be performed strictly in the same order as given. Using Stata's help command, followed by the relevant Stata commands mentioned in this table, the reader can learn more about each of those commands and become familiar with different examples.

Table 2.1 Data-cleaning strategy

Reason (Why?)	Step (What?)	Command (How?)
Identify the variables and fix incorrect code	Label/re-label variables and label their values	label; recode
Identify unique observations to correctly merge	Understand unique identifiers from survey design and extracted data	egen group(); isid; codebook; inspect; duplicates
Correct spellings; make the data uniform	Correct string variables	replace; substr; subinstr; index
Change and transform variables per need for analysis	Transforming variables	gen; destring; tostring; drop; keep; egen; rename; bysort; encode; recode
Ensure logical connections are present in data, such as mothers are females or quantities have correct units	Consistency checks	assert; tabulate; summarize; table; tabstat; count
Create a single data file to work with	Merge or append different data files	merge 1:1; merge m:1; merge 1:m; append
Create a logical observation to organize the data file	Reshape data to appropriate wide or long format	reshape
Identify the importance and influence of missing values	Decide if missing observations need to be removed or imputed	sum; mi
Detect outliers	Remove or substitute outliers as necessary	sum; hist; hilo; stem; graph box; scatter
Keep a record of all commands to facilitate replication and	Document every step with comments and commands	use do-file editor to organize

2.7 Generating basic descriptive statistics from household surveys

A statistical software program usually analyzes data as if the data were collected using simple random sampling. However, as previously mentioned, most household surveys use more complex and multi-stage survey designs to collect data, and stratification and clustering in sample surveys affect the calculation of the standard errors. Therefore, the performed statistical analysis should be able to correct for the design elements used in the survey in order to obtain more accurate point estimates and standard errors. The documentation provided along with the survey data usually gives detailed information on the specific sampling design that was used. This section discusses how to declare the survey design elements and produce descriptive statistics for the full sample and by specific category. This section also offers guidance on useful Stata code to perform these actions.

In Stata, the command <*svyset*> is used to declare the survey design of the data. It designates variables that contain information about the survey design, such as the sampling weights, PSU/cluster, and strata and specifies other design characteristics of the survey, such as the number of sampling stages and the sampling method. The design declaration, if need be, can be cleared with the command <*svyset, clear*>. Once the data are declared with <*svyset*>, only the prefix <*svy:*> needs to precede each command. The syntax of the <*svyset*> command for a multi-stage survey design looks like: <*svyset psu [weight] [, design options] [|| ssu, design options] … [options]*> where psu is the name of a variable identifying the primary sampling unit in the data, *weight* identifies the sampling weight, ssu identifies the sampling units in the second stage, and so on.

Design options will declare the design elements like strata. The Stata website, for example, provides a sample survey data set from the second National Health and Nutrition Examination Survey (NHANES) in the US from 1976–80. Import those data into Stata with the command <*webuse nhanes2*>. The data give a weight variable (*finalwgt*), a PSU variable (*psu*), and a strata variable (*strata*). The <*svyset*> command in this case will look like: <*svyset psu [pw=finalwgt], strata(strata)*>, where “pw” stands for probability weights.

Most surveys explicitly include sampling weights, stratum, and PSU identifiers along with the published data. It is important to carefully read the survey documentation to understand the description of variables. Since published reports from the survey also present important point estimates, it is possible to compare the calculated numbers with those in the published reports. Before proceeding with further analysis, it is important to perform such cross-examinations to make sure the correct sampling weights and survey design elements are being used as originally intended.

Once the survey design is declared through <*svyset*>, information on strata and PSU can be obtained with the command <*svydescribe*>. Further estimation of descriptive statistics should be prefixed with <*svy:*>. For example, to estimate the mean of a variable, simply run the command <*svy: mean varname*>. If the mean is computed for a binary variable, it would display proportions. Alternatively, run <*svy: tab binaryvar*> to estimate the proportions of, say, males and females, literate and illiterate, or similar binary variables along with their standard errors corrected for the survey design. Similarly, <*svy: proportion binaryvar*> would provide an output with proportions of the variable of interest along with their standard errors and confidence interval.

To estimate the same descriptive statistics for subgroups in the survey, such as income groups, gender, or any other SES categories, the <*svy*> command can be executed with additional options

like <*subpop*> or <*over*>. For example, the command <*svy, subpop (female): mean binaryvar*> or <*svy, over(female): mean binaryvar*> gives the necessary estimates of interest along with their standard errors. Suppose one would like to find the average expenditures on cigarettes by different expenditure quartiles. To do so, first create a variable to categorize households into four different quartiles based on their total monthly household expenditures "exptotal," as follows: <*xtile exp_quartiles =exptotal, n(4)*>. Then, use the command <*svy, over(exp_quartiles) : mean exp_cig*> to obtain the average monthly expenditures on cigarettes by different expenditure quartiles.

The estimates from the survey data can also be produced without explicitly declaring the survey design but using the correct sampling weights and adjusting for the standard errors. In Stata, this is done with the help of weights and robust cluster options. For example, in the above case of cigarette expenditures by expenditure quartiles, the same average expenditures by different expenditure quartiles may be obtained with the command <*mean exp_cig [pw=weightvar], over (exp_quartiles)*>, where weightvar is the sampling weight identifier that was used to declare the survey design.

However, descriptive statistics using the sampling weights—while producing the same estimates as those using <*svyset*>—do not adequately address the stratification issues and, as a result, could produce standard errors different than those obtained using the <*svy*> command. In the regression context, however, one could add the optional argument <*robust cluster(psuvar)*> after the main regress command where *psuvar* is the variable that identifies cluster or PSU in the data, and it would correct for survey design effects while computing standards errors for the coefficient estimates.

3 Estimating own- and cross-price elasticities

This chapter presents methods of estimating the price elasticity of demand using HES data. Price elasticity is one of the most important parameters to be considered when designing tax policy, as it provides an insight to policy makers on the responsiveness of demand to changes in price. Based on the estimated price elasticity, policy makers can predict with some degree of confidence the impact of their policies on relevant policy objectives, including tobacco consumption and tax revenues. Moreover, empirical evidence on the magnitude at which tobacco demand would respond to price provides a very relevant counterargument to opponents claiming that raising taxes would unambiguously result in reduced tax revenues.

Policy makers are interested in the responsiveness of tobacco consumption to changes not only in prices of tobacco (that is, own-price elasticity), but also to changes in prices of other goods, such as their potential complements (for example, alcohol or coffee) or their substitutes. Similarly, policy makers may want to know the impact of a change in price of one type of tobacco product (such as cigarettes) on other types (such as roll-your-own cigarettes), as the impact of their policy may be effectively reduced if, for example, there is room for downward substitution.

It is also useful to understand the price responsiveness of tobacco users not only with respect to the quantity of tobacco product they consume, but also with respect to how price changes impact their decision to initiate and quit tobacco use. While the former is measured using elasticities on the intensive margin (also known as a "quantity elasticity"), the latter can be assessed by estimating elasticity on the extensive margin (also known as "prevalence elasticity").

In this chapter, these concepts are defined in detail along with examples. A brief theoretical discussion on the estimation of both quantity and prevalence elasticities using HES, in that order, will follow. In the later part of the chapter, Stata code is provided to enable the reader to estimate elasticities. Finally, an example is provided using HES data from an unidentified country.

3.1 Definition of concepts

The own-price elasticity of demand is formally defined as the percentage change in the quantity demanded of a good that results from a one-percent change in the price of that good, keeping everything else constant (*ceteris paribus*). For example, a price elasticity of demand of -0.5 would imply that the quantity demanded of that particular good declines by five percent when the good's price rises by 10 percent. Similarly, a price elasticity of demand of -1.5 implies that the quantity demanded of the good in question declines by 15 percent when its price rises by 10 percent. Because it measures the percentage change in quantity consumed, these are also known as quantity elasticities or elasticities on the intensive margin.

Goods with a price elasticity of demand less than one in absolute value are said to have inelastic demand because the demand response is relatively less than the price change. On the other hand, goods with price elasticity of demand more than one in absolute value are said to have elastic demand because the demand response is relatively greater than the price change. There are various factors impacting price elasticity, such as the availability of substitutes, whether a good is a necessity, the period of time available to find alternatives, how broadly or narrowly the commodity is defined, and/or the addictive/habitual nature of the product. Considering these factors, tobacco products—having few substitutes and being addictive—tend to have relatively price-inelastic demand.

For most ordinary commodities, it is the quantity elasticities that are of interest as far as tax policy is concerned. However, demerit products such as tobacco are consumed by only a few. This means for a large number of people, their quantity of consumption is conditional on their decision to participate or consume. Typically, these decisions follow a two-step process: (i) a decision on whether to consume at all, and (ii) having decided to consume, how much to consume.

For example, the decision to consume or "participate" in the consumption process can be either in the form of current non-smokers initiating smoking or current smokers continuing or quitting the habit. Together, they may be called prevalence or participation elasticity, or elasticity on the extensive margin. This is defined as the proportionate change in prevalence of smoking as a result of a given proportionate change in price of cigarettes.

Because of the conditional nature of the quantity of consumption, the quantity elasticities are also referred to as conditional elasticities in this context. The price elasticity of demand for cigarettes is thus a sum of the elasticities on the extensive and intensive margins, reflecting the impact of price changes on both prevalence as well as intensity of smoking. The respective share of both elasticities—quantity and prevalence—in the total price elasticity coefficient is something to be empirically determined and will depend on the particular data available.

Whether a good's demand is elastic or inelastic matters a great deal for tax policy. Tax revenues can be expected to decline whenever taxes are raised on a good that is demand-elastic, as the demand response outstrips the price change so that sales revenues and tax revenues ultimately decline. On the other hand, tax revenues can be expected to increase whenever taxes are raised on a good that is demand-inelastic, as their demand response is smaller than the price change so that sales revenues and tax revenues ultimately increase.

The literature on the estimation of own-price elasticities of demand for cigarettes tends to find, in general, elasticities ranging between 0 and -1,[4, 20, 21] meaning that demand for tobacco is inelastic, which is expected given the addictive nature of this product as well as the availability of very few close substitutes. The empirical evidence also confirms that tobacco taxation, through higher prices of tobacco, is one of the most effective policy tools for decreasing smoking and its adverse health consequences.[4, 22–24]

In addition to own-price elasticity, cross-price elasticity should also be defined. Formally, the cross-price elasticity of demand between good X and Y is defined as the percent change in the demand for good Y when the price of good X changes by one percent, *ceteris paribus*. Unlike with own-price elasticity, which is always unambiguously negative, the cross-price elasticity can have a negative or positive sign. A negative cross-price elasticity means the two goods in question are complements.

In other words, the joint consumption of the two goods satisfies a need. An example would be gasoline and cars. On the other hand, a positive cross-price elasticity means the two goods are substitutes. That is, one good can be used in place of the other good or both goods satisfy the same need. An example of substitutes is bottled water and tap water.

Furthermore, there is an income elasticity of demand. In this toolkit, the terms income elasticity and expenditure elasticity are used interchangeably, as total expenditure in HES is used as a proxy for income. The income elasticity of demand is formally defined as the percent change in the quantity demanded of a good arising from a one-percent increase in income, *ceteris paribus*. A negative income elasticity of demand means that the quantity demanded of the good declines whenever incomes rise. Such goods are referred to as "inferior" goods. Staple foods (such as rice or maize/corn) often have negative income elasticities of demand. On the other hand, goods having positive income elasticities of demand are referred to as "normal" goods.

Knowing the magnitude of the income elasticity of demand is important for tobacco control policy. For example, a positive income elasticity of demand on cigarettes in a country implies that tobacco control efforts must be stepped up, especially in periods of rising incomes in that country.

Just as income elasticity quantifies changes in consumption as a result of changes in income, it is also observed that the price elasticity for a given product can vary for individuals from different income groups or SES groups. Estimates of price elasticities for cigarettes and other tobacco products could vary by SES groups. In general, the literature finds that people or households in lower-income groups are far more sensitive to price changes of a given product compared to those in higher-income groups. For example, most estimates of price elasticities of demand for cigarettes from HICs range from −0.2 to −0.6, clustering around −0.4, whereas estimates from LMICs range from −0.2 to −0.8, clustering around −0.5.[4]

Even within HICs, studies in the United Kingdom (UK) and Australia[25, 26] show relatively greater responsiveness to price among lower- compared to higher-SES groups. In the United States, while most studies[27–29] find a relatively greater responsiveness to tobacco price changes in lower- than in higher-SES groups, a few studies[30] offer inconclusive evidence. In the case of LMICs, although some studies[31–38] have offered mixed evidence, a much larger and growing body of evidence shows a significantly greater response in the lower-income groups than among those with higher income. These include studies from countries such as Argentina,[39] Bangladesh,[40] Bosnia and Herzegovina,[41] Brazil,[42] China,[44, 45] El Salvador,[45] India,[46, 47] Indonesia,[48] Iran,[49] Kosovo,[50] Mexico,[51] Montenegro,[52] Pakistan,[53] Peru,[54] Serbia,[55] Tanzania,[56] Thailand,[57] and Turkey,[58] among others.

Similar differences in price elasticity can also be observed if the analysis is done across different SES groups other than income. This toolkit discusses the estimation of price elasticities by different income groups. This toolkit does not repeat instructions for price elasticity estimation with other SES classifications since those would also follow a similar approach.

3.2 Econometric issues in demand estimation

There are several theoretical and practical issues to consider in the estimation of price elasticities of demand. This section covers some of the main issues.

3.2.1 Identification problem in demand analysis

The law of demand states that as the price of a good increases, its demand decreases, *ceteris paribus*. It assumes that the direction of causation runs from price to quantity demanded. However, market interactions tend to be more complex, because demand influences price as much as price influences demand. This can be observed in real time in the stock markets. An increase in the price of a stock is likely to lead to a reduction in the quantity demanded of the stock. On the other hand, an increase in demand for the stock is likely to lead to an increase in the price of that stock. Furthermore, other factors (such as incomes, tastes, the weather, and prices of related goods) can—outside of the influence of price—influence demand for the good.

The issues explained above are referred to in econometric analysis as the "endogeneity problem," or the "identification problem," and a failure to address them adequately would lead to obtaining biased estimates; that is, the estimates would be significantly different from the true value of the parameter being estimated. This is a very relevant issue for policy formulation, as it would lead to a policy that may be designed based on unrealistically positive or negative estimates, depending on the sign of the bias.

Ideally, the endogeneity problem, or the identification problem, can be econometrically resolved by running an experiment where units are randomly assigned into treatment or control groups. Here, there is no need to worry about endogeneity because randomization rules out all other factors except the factor of interest. Unfortunately, with social reality, unlike in the physical sciences, it is not always easy or even desirable to run social experiments. Therefore, economists and social scientists search for "natural" experiments or quasi-experiments that can be exploited in overcoming the identification problem.

In regard to estimating the price elasticity of demand for tobacco products, researchers have searched for instances where governments have independently (that is, exogenously) introduced an increase in tobacco prices. For instance, several studies in the US in the 1990s took advantage of the 25-cent cigarette tax increase in California and Massachusetts to estimate the price elasticity of demand,[59–62] because the exact source of the price change that led to a change in quantity demanded can be pinpointed in these events.

However, such dramatic changes in tobacco taxes are not very common, especially in LMICs, where, unless they are undergoing a tobacco tax reform, the changes in tobacco taxes are most commonly gradual and small in magnitude, usually to correct for the impact of inflation. These gradual changes make it difficult to isolate the causal effect of price on demand, so the estimation procedure requires using the IVs in obtaining the causal effect of price on demand (see Chapter 2 for a discussion on endogeneity and the role of IVs in resolving it).

IVs are difficult to come by in general and in demand analysis in particular. Fortunately, Nobel Laureate Angus Deaton has proposed a suitable IV within the context of LMICs that allows for the estimation of defensible elasticities. The method proposed by Deaton is detailed below.

3.3 Estimation of quantity elasticity with household expenditure surveys

This section discusses the theoretical framework behind the estimation of quantity elasticity followed by its estimation using HES. While there are a few different models using a system of

demand equations, the almost ideal demand system (AIDS) introduced by Deaton and Muellbauer (1980)[63] has been the most popular due to its many advantages. AIDS has a flexible, functional form consistent with household expenditure data and different axioms of choice. It does not impose any prior restrictions on elasticities, and its mostly nonlinear specification makes it easy to estimate, allowing it to explicitly test the restrictions of homogeneity and symmetry. Deaton's (1988) model, presented in this toolkit[64] and detailed in his book,[8] builds on Deaton and Muellbauer (1980).[63] However, it differs from AIDS in that it corrects for both measurement errors and quality shading in unit values, as discussed below.

The model allows for data from HES to be utilized to estimate credible price elasticities of demand, starting from the assumption that prices of most goods in LMICs vary significantly across geographical space. This spatial variation of price is the result of either significant transportation costs, due to goods moving from one place to the next, or other factors such as different border taxes or additional duties in different jurisdictions in the same country. Thus, transportation costs, or these other factors affecting price changes across geographical regions, implicitly serve as an instrument and are the main factors influencing price, which in turn influences demand. Therefore, genuine variation in price across clusters is assumed for the identification of price elasticities in this model, effectively solving the identification problem highlighted earlier in this chapter.

The assumption about spatially varying prices means that households living close to one another, such as those in the same village or urban block, should face the same price as they make purchases in the same market and at the same time, if it is a cross-sectional survey. On the other hand, households living far apart, such as those in different villages or urban blocks, should face different prices. In other words, the approach requires that much of the observed variation in price should take place between clusters as mentioned in Chapter 2, as opposed to within clusters. Econometrically, this requires that price variation should largely be explained by "cluster effects" or "cluster dummies." Any within-cluster variation in price should be a result of measurement error, patterns of which can be utilized in correcting final estimates for such error (more in Section 3.3.1).

Household expenditure surveys usually do not report the market price in the survey. It could be inferred from the purchasing decisions of households by calculating the ratio of household expenditure on a given good to the quantity of that good. This ratio, however, is a unit value and not price. Unit values are not the same thing as prices because of the following two problems. First, unit values are affected by both the actual price and the choice of quality (that is, "quality effects"). If not properly dealt with, this might lead to so-called "quality shading," which refers to a situation where a price change does not lead to a reduction in quantity demanded as people trade down to cheaper but lower quality products. Second, unit values are not the same thing as prices because of measurement error, given that people often misreport expenditure on and/or quantities of goods purchased. Deaton proposes a method to deal with both quality shading and measurement error. The next section gives a technical step-by-step explanation of the method originally proposed by Deaton in 1988, which has since been extended in his later work.[8, 65–67]

3.3.1 Theoretical framework of the Deaton model

This section briefly describes the main steps involved in deriving the theoretical model proposed by Deaton to estimate price elasticities (on the intensive margin) using HES data. Researchers planning to implement this model are advised to read Chapter 5 from Deaton (1997)[8] to understand

the finer details of the model. The model mainly consists of six steps, from deriving the unit values through relevant tests to finally estimating the price and expenditure elasticities.

Step 1: Deriving unit values

First, the unit values are derived from the survey data at the household level. This is done by dividing reported total expenditure on the particular tobacco product or products on which HES provides data by their corresponding quantity, as:

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

where υ_{hc}, x_{hc} and q_{hc} are, respectively, the unit value, expenditure, and quantity of cigarettes or any other tobacco product in household *h* located in cluster *c*.

Step 2: Testing for spatial variation in unit values

The second step consists of checking whether obtained unit values in Step 1 satisfy the main identifying assumption: unit values vary spatially. This is done by using analysis of variance (ANOVA) to divide the total variation in unit values into "within-cluster variations" and "between-cluster variations." A significantly large *F*-statistic for the ANOVA exercise leads to the conclusion that unit values vary across geographical space or clusters.

Step 3: Estimating within-cluster regressions

In a third step, one estimates within-cluster regressions of unit values and budget shares using the following specification:

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

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

where lnv_{hc} is the log of the unit value, derived according to Equation 3.1 for household *h* in cluster *c*, while w_{hc} represents the share of tobacco expenditure in the total household expenditure for household *h* in cluster *c*. And lnx_{hc} is the log of total household expenditure over the relevant reference period. Z_{hc} is a vector of household-specific characteristics that might include variables on household structure (such as household size, proportion of adults, or proportion of males) and household demographics (such as age, gender, marital status, or schooling and employment status of head of household). f_c is a cluster-fixed effect and treated as an error in addition to the error term u^1_{hc} is the standard regression error term. Both u^0_{hc} in Equation 3.2, while u^1_{hc} is the standard regression error term. Both u^0_{hc} and $^1_{hc}$, however, incorporate any measurement errors in budget shares and unit values, apart from the usual unobservables.

The unit value equation contains no village-fixed effect because, as Deaton observes,[8] "conditional on prices, unit values depend only on quality effects and measurement errors. The introduction of an additional fixed effect would break the link between prices and unit values, would prevent the latter giving any useful information about the former, and would thus remove any possibility of identification" of prices. Finally, $ln\pi_c$ are the unobserved prices and consequently, Equations 3.2 and 3.3 are estimated without them but their coefficients are recovered through the formulas contained

in Equations 3.8 and 3.9 below. As discussed above, Deaton's model assumes no within-cluster variation in prices, as all households within the same cluster face the same price and are surveyed at the same time. Therefore, even if the prices were observed, they would have been dropped in this step from the regression due to a lack of variation.

Equation 3.2, referred to as the "unit value" equation, checks for the presence of quality effects as discussed in Section 3.2.2. A positive and statistically significant relationship between household expenditures and unit values, after accounting for household characteristics, would suggest the presence of quality effects. Knowing the pattern of the quality effects (that is, the magnitude of $\beta 1$), allows correction of the final price elasticity estimates for quality shading as in Step 6. Note that Equation 3.2, unlike Equation 3.3, is estimated without the cluster-fixed effects. Adding a cluster-level fixed effect to Equation 3.2 would make it difficult to recover the model's parameters.

Equation 3.3, on the other hand, is a standard demand equation where the cigarette share (a proxy for demand) is expressed as a function of household income (proxied by household expenditure), household characteristics, and prices. Because of the assumption that prices are fixed within clusters and the fact that there are no price data, prices are proxied by cluster-fixed effects. The relationship between the two errors, u^0_{hc} and $^1_{hc}$, (as captured by, say, the covariance) is useful in correcting the final price elasticity estimates for measurement error as explained in Step 5.

Step 4: Obtaining cluster-level demand and unit values

The fourth step involves stripping the household-level demand and unit values of the effects of household expenditure and household characteristics and then averaging across clusters. The stripping and averaging are done because the primary interest is to estimate elasticity at the cluster level using cluster demand and cluster unit value stripped of all other factors. This step requires the following equations:

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

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

where n_c is the number of households in cluster c and n_c^+ is the number of households reporting purchase of the tobacco product for which elasticity is estimated. Notice that $\widehat{y_c^1}$ and $\widehat{y_c^0}$ do not have the h subscript because they represent cluster averages. $\widehat{y_c^1}$ and $\widehat{y_c^0}$ are the estimates of, respectively, cluster average unit value and cluster average demand after removing the effects of household expenditure and household characteristics. In other words, equations 3.4 and 3.5 can alternatively be expressed as $y_c^1 = \alpha^1 + \psi ln\pi_c + u_c^1$ *and* $y_c^0 = \alpha + \theta ln\pi_c + f_c + u_c^0$, respectively.

Step 5: Cluster-level regressions

Recall that the identifying assumption is that prices vary between clusters and not within clusters. Given this, price elasticity of demand can only be obtained by seeing how cluster level demand responds to changes in cluster level prices. Thus, Step 5 involves regressing cluster level demand, $\widehat{y_c^0}$, on cluster level unit values, $\widehat{y_c^1}$. The coefficient on $\widehat{y_c^1}$ in such a regression can alternatively be obtained by dividing the covariance between $\widehat{y_c^0}$ and $\widehat{y_c^1}$ by the variance of $\widehat{y_c^1}$. That is $\hat{\phi}$, the estimate of the coefficient on $\widehat{y_c^1}$, is obtained by:

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

where n_c^+ is the number of households in a cluster reporting positive expenditures on tobacco and n_c is the number of households in a cluster; $\widehat{\sigma^{10}}$ is the estimate of the covariance of the errors in Equations 3.2 and 3.3; $\widehat{\sigma^{11}}$ is the variance of the errors in Equation 3.2. Equation 3.6 is a standard errors-in-variables regression where the covariance and variance of errors is used to correct for measurement error. Notice that the correction factors for measurement error become small as n_c^+ and n_c become large.

Step 6: Estimating price and expenditure elasticities

The sixth and final step in Deaton's method applies quality correction formulas in obtaining the estimate of the price elasticity of demand, $\widehat{\varepsilon_P}$, as follows:

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

where $\bar{w}$ is the average share of total household expenditure dedicated to cigarettes in the sample. $\hat{\psi}$ and $\hat{\theta}$, the estimates of the coefficients on the unobserved price terms in Equations (3.2) and (3.3) respectively, are recovered as follows:

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

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

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

Finally, Deaton also proposes the following formula for obtaining the estimate of the expenditure elasticity of demand, $\widehat{\varepsilon_I}$:

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

where $\hat{\beta}^1$ is the estimate of the coefficient on total household expenditure in Equation 3.2, and $\hat{\beta}^0$ is the estimate of the coefficient on total household expenditure in Equation 3.3. $\hat{\phi}$ is the estimate of the coefficient of a regression of cluster-level demand on cluster-level unit value (from Equation 3.6). Once the parameters in 3.8 to 3.10 are recovered, the price elasticity of demand can be estimated as per Equation 3.7. On the other hand, the expenditure elasticity of demand only uses first stage coefficients and can be derived using Equation 3.11. Given that the formulas for the price elasticity of demand in Equation 3.7 and for the expenditure elasticity of demand are not direct Stata commands, their standard errors have to be obtained by bootstrapping.

A number of studies have used Deaton's method to estimate price and expenditure elasticities of demand for various tobacco products in different LMICs. These include studies in Albania,[38] Bangladesh,[68] Bosnia and Herzegovina,[81, 82] China,[71] Ecuador,[36] El Salvador,[45] India,[46, 68–71] Montenegro,[79, 80] Pakistan,[78] Serbia,[77, 78] South Africa,[81] Uganda,[82] and Vietnam,[83] among others.

Some estimated elasticity for a single good—cigarettes—while others estimated own- and cross-price elasticities for cigarettes and a few other tobacco products.

It should also be noted that, while some of these studies considered all households in the budget share regression for estimating elasticity, some considered only households with positive purchases in the budget share regression, thus estimating only a conditional demand. However, as Deaton points out,[8] for the purpose of tax and price reform, one needs to include all households in the analysis whether they purchase or not. Hence, if the quantity price elasticity is estimated with only households with positive purchases (as shown in this toolkit), supplementing it with an estimate of prevalence elasticity that includes all households would provide an estimate of total price elasticity.

The estimates of own-price elasticity for cigarettes in these studies ranged from -0.1 to -1.4, although most studies had elasticity coefficients 0.8 or lower in absolute value, while expenditure elasticity estimates ranged from 0.2 to 2.4. In other words, these studies tend to find price elasticity estimates for cigarettes comparable to the ones estimated in the international literature using other methods. They also tend to find non-negative expenditure elasticities of demand for cigarettes, implying that cigarette demand does not decline with an increase in expenditure.

It is also interesting to note that the definition of cluster used in these studies varies. While some considered a village or urban block as the default cluster, others considered a district itself as a cluster. It is also possible to define a cluster over both geographical and time variables[77, 78, 84] if, for example, there are HES from multiple rounds or waves. It is important to understand that the consistency properties of parameters in Deaton's model depend on the number of clusters (and not on the number of households) as these parameters are derived from average cluster-level data.

On the other hand, the measurement errors in Equations 3.2 and 3.3 tend to zero only as the number of households in each cluster increases. Clearly, there is a trade-off. On the one hand, small cluster sizes increase the probability of increased measurement errors which is especially true in the case of products like tobacco which are consumed by only a few households. With smaller clusters, it is also possible that some of them do not have any households with positive tobacco purchases at all. On the other hand, since the second stage regression and the estimation of price elasticity depend on having a large number of clusters with positive purchases, it is important to have as many clusters as possible in order to derive consistent parameter estimates.

Deaton's own experiments have shown that the estimator performs adequately even when there are as few as two households in each cluster.[8] According to Deaton, "increasing the number of villages or clusters is much more important than increasing the number of observations in each." This is due to two reasons: (i) the model corrects measurement errors but it cannot guarantee the consistency of parameters with a small number of clusters, and (ii) if clusters are defined or aggregated over larger geographical areas, then the households within such clusters may not be facing the same market and, as a result, there may be true intra-regional variations in unit values within those clusters that may inadvertently be treated as measurement errors. For the model assumptions to hold, the households in a given cluster should have geographical proximity and face interviews at more or less the same time. This may become all the more difficult as clusters are expanded to include larger geographical regions.

For most HES, the clusters are naturally given as part of the survey design, as already noted in Chapter 2. It is also important to note that the model relies on the existence of genuine variation in

prices across clusters and requires that such variation be exogenous to the process that determines demand. As Deaton observes,[8] "if local prices are determined by world prices, border taxes, and transport costs, the assumptions will be satisfied because local demand has no effect on prices." On the other hand, if village prices depend on demand within the village, the parameter estimates will not be consistent, for the usual simultaneity reasons.

It is worth noting that even though the above discussion refers to households, the analysis can also be conducted at the level of individuals. However, this requires that the researcher have access to a rich expenditure survey with data collected at the individual level. For example, such a survey should contain information on the expenditure patterns (quantity and total amount spent) and on tobacco products by individuals (not aggregated at the household level, as is often the case).

Furthermore, other social and demographic data at the individual level should also be present. Whereas such data sets are widely available in HICs, they tend to be the exception in LMICs. Researchers with access to expenditure surveys collected at the individual level are encouraged to use Deaton's method for the estimation of demand elasticities.

Deaton's method is not without its critics. Gibson and Rozelle (2005)[85] show that using unit values as a substitute for actual prices yields biased estimates for the price elasticity of demand even after correcting for quality effects and measurement error. McKelvey (2011)[84] shows that Deaton's method does not adequately deal with the issue of quality shading that appears to be prevalent in many settings. These limitations notwithstanding, in the absence of very detailed price data, Deaton's method remains one of the most effective ways to obtain elasticities.

3.3.2 Preparing data for estimating quantity elasticities

Once data have been extracted and cleaned, different data sets merged, and data otherwise managed as needed—as detailed in Chapter 2—specific details are required about the variables necessary for estimating price elasticity using Deaton's method discussed above. For any new variable discussed here, it is important to take it through all the processes discussed in Chapter 2. This section discusses how the specific variables that are required for the estimation of own- and cross-price elasticities, using Deaton's method, can be generated using the standard variables available from HES.

The most important variables are the quantity of consumption and the expenditures on different tobacco products. These are directly available from most HES. Some HES may not report quantity information, as mentioned earlier. In such cases, the discussion here may not be relevant.

First, unit values should be created for each of the tobacco products for which data are available. This may include unit values for cigarettes, bidis, and smokeless products, among others. For example, the quantity of cigarettes (either in packs or number of sticks) as downloaded from the HES data has the variable name *qcig*, and the variable representing the expenditure spent on cigarettes is *expcig*. Then, the unit value of cigarettes (*uvcig*) can be generated using the command *<gen uvcig=expcig/qcig>*.

Deaton's model uses the natural log of the unit value variable as the dependent variable (*luvcig*). Use the command *<gen luvcig=ln(uvcig)>* to generate this. Similarly, a variable to represent the budget shares devoted to cigarettes (*bscig*) should be constructed using the command *<gen bscig =*

expcig/exptotal>, where *exptotal* is the total expenditures on all items. For those households with no reported expenditures on cigarettes, this would generate a missing value. While implementing Deaton's model, *uvcig* and *bscig* will be the dependent variables in the respective regressions. Similar unit value and budget share variables should be generated for other tobacco products from HES that will be included in the estimation of price elasticity.

Price is definitely one independent variable to use in a model estimating demand functions. However, as noted earlier, Deaton's method is used in cases where direct price information is not available. The price variation is instead captured through the cluster-level variations of prices in HES. It is, therefore, crucial to have a variable that identifies clusters (*clust*) or primary sampling units. This variable is usually directly available from the HES or may be generated using other available variables identifying primary sampling units as discussed in Chapter 2. The cluster can be a geographical unit (such as a village or primary sampling unit in a cross-section survey) as in Deaton's original analysis, or it can be a point in time (such as survey wave) if combining different rounds of surveys or a combination of both PSU and survey wave.[84]

Sometimes surveys have no data on quantity of purchase, but there are data on expenditure for a particular item against a household and vice versa. Such households may be dropped from the analysis using the command *< drop if [qcig==.&expcig!=.]|[qcig!=.&expcig==.] >*. After this, any *bscig* variables that are still missing may be replaced with zeros.

Since it is important that each cluster has at least two households, as mentioned in the previous section, clusters having fewer than two households should be dropped from the estimation of quantity elasticities. This is done by creating a cluster-level variable containing the number of households consuming tobacco (*cigclustsize*).

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

In addition, it is necessary to identify specific household-level variables to use as independent variables in the model. The literature offers guidance on some of the common household-level sociodemographic variables: log of household size; male ratio (ratio of number of males to household size); average age of household; average education (total education received by all the members in years divided by the household size) of the household; max education (years of education received by the most educated member in the household); educational attainment of household or the household head; dummy variables to characterize households into different social, ethnic, occupational, religious, and income groups; and dummy variables to indicate the location of the household (such as rural/urban areas, province, region, or district), among others.

3.3.3 Estimating price elasticity on the intensive margin with Stata

This section provides the Stata code for the estimation of own-price elasticity for a single tobacco product (cigarettes) using Deaton's method discussed earlier. Deaton provides detailed Stata code for estimating own- and cross-price elasticities for different products, which can be downloaded from **http://web.worldbank.org/archive/website00002/WEB/EX5_1-2.HTM**. The Code Appendix in Section 7.3 provides a modified version of Deaton's code from the World Bank website, with some added explanations for readers to follow.

The code used in this section for estimating price elasticity for cigarettes would produce identical parameter estimates for elasticity as Deaton's code for multi-good cases in Appendix 7.3 used to estimate elasticity for a single good. While the code for multi-good cases makes use of matrices for computing several parameters in the model, the code here uses only scalars, as it is a single commodity. Moreover, as the code for multi-goods also estimates cross-price elasticities and allows introduction of other theoretical restrictions on the demand system, as discussed in Deaton,[8] the code here simply estimates own-price elasticity for cigarettes without imposing any other restrictions.

The code for this section uses the variables *bscig, luvcig, lexp, lhsize, maleratio, meanedu, maxedu, sgp1, sgp2, sgp3* for the estimation of own-price elasticities, where *bscig* refers to budget share for cigarettes in total household budget, *luvcig* is the natural log of the unit value of cigarettes, *lexp* is the natural log of the monthly household expenditures, *lhsize* is the natural log of the household size, *maleratio* is the ratio of the number of males to number of household members, *meanedu* is mean education of household members in years, *maxedu* is the maximum education received by any of the household members in years, and *sgp1* to *sgp3* are the social groups defined as socioeconomic groups or caste groups any such grouping as appropriate for each country.

Since the quantity elasticities are estimated only for households reporting consumption of tobacco, all other households are first dropped from the data using the command < *keep if dcig==1* >.

Testing for spatial variation in unit values

As indicated in the method Section 3.3.1, it is useful to estimate the variation in unit values across clusters to assess if variations in unit values are indicative of variation in prices across clusters. This can be done using the command <*anova luvcig clust*> or <*regress luvcig i.clust*>. The R^2 and *F*-statistic from the output can indicate the usefulness of unit values as informative of prices. According to Deaton,[8] a significant *F*-statistic and R^2 value around 0.5 (that is, cluster dummies explains about half of the total variation in unit values) means the unit values can be used for the purpose of examining price variation and to estimate the price elasticities.

Estimating within-cluster first-stage regressions and measurement error variances

Below, Equations 3.2 and 3.3 are estimated and relevant parameters are stored for the subsequent stages:

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

*Repeat for budget shares
```

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

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

The command <*areg*> is used instead of <*regress*> since this is a linear regression with a large dummy-variable set. The command implicitly includes a dummy variable for each cluster dropping one but does not list the coefficient associated with these cluster dummies in the regression output. The option <*absorb(clust)*> along with the command <*areg*> tells Stata to use implicit cluster dummies for the cluster variable clust.

The variables *y1cig* and *y0cig* after each regression purge off any effects of household-specific characteristics that explain quality variation in unit values. These variables now preserve the price information contained in cluster dummies. The residuals from the unit value (*ruvcig*) and budget share regression (*rbscig*) are generated to be used in the last regression of *ruvcig* on *rbscig* to construct the scalar *sigma12*. This *sigma12* along with the scalars *sigma11* and *sigma22*, generated after unit value and budget share regression, are estimates of the variance and covariance of measurement errors to be used for the measurement error correction in Equation 3.6. The coefficient for the log expenditure is also stored for later use. The scalar *b1*, which is the coefficient of the log expenditure in the unit value regression, is the estimate of quality elasticity. The lower this number, the lower the quality shading in unit values.

Estimating income or expenditure elasticities

The total expenditure elasticity (or income elasticity) in Equation 3.11 can be estimated after these first stage regressions using the saved results. This can be done using the code:

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

The code stores the estimate of the average budget share into a scalar (*Wbar*) first and uses the other saved scalars (*b1* and *b0*) from the first-stage regressions to estimate the expenditure elasticity (*Expel*). The last line will print the expenditure elasticity on Stata's result window. In order to estimate the standard errors for the expenditure elasticity coefficient, a bootstrap method is used with the following code:

```
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
```

The code returns the expenditure elasticity coefficient along with the bootstrapped standard errors.

Preparing data for between-cluster regression

The next step involves averaging the variables *y1cig* and *y0cig* by clusters to generate *y1c* and *y0c* respectively, so that they can be used for a between-cluster regression of *y0c* on *y1c* to derive the own-price elasticity. As mentioned earlier, the variables *y1cig* and *y0cig* are purged of any household-specific characteristics from unit value and budget share regressions and contain only the price information in cluster dummies as well as the measurement errors.

```
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
```

After generating an average value for all households in each cluster, only one observation per cluster needs to be kept for the remaining analysis. Along with generating the cluster-level variables *y0c* and *y1c*, two other cluster-level variables are generated (*n0c* and *n1c*), indicating the size or the number of all households in each cluster (*n0c*) and the number of households reporting positive purchases in each cluster (*n1c*). Using these, the average cluster size for all households (*n0*) and the average cluster size for households with positive consumption of cigarettes (*n1*) are estimated. This can be done using the following code. Deaton uses harmonic mean to estimate these average cluster sizes.

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

Between-cluster regression

The between cluster regression of *y0c* on *y1c* yields the estimate of the ratio $\phi = \theta/\psi$, the numerator and denominator of which are the coefficients of unobserved prices in Equations 3.3 and 3.2, respectively. Instead of doing the actual regression, the hybrid parameter can be estimated using an errors-in-variable estimator in Equation 3.6, for which the estimates for *y1* and *y0* as well as the measurement error variances and covariance estimated from the first-stage regressions are used. The Equation 3.6 is estimated using the following code:

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

Estimating own-price elasticity

Once the ratio ϕ is estimated, as in Equation 3.6, a few more scalars need to be defined to estimate the actual own-price elasticity. This is done in the code below:

```
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
```

The last line of the code will display the estimate of own-price elasticity on the Stata result screen. The other scalars defined above are estimates for Equations 3.8 to 3.10, not necessarily in the same order. In order to estimate the standard errors for the price elasticity estimates, the above equations should go into a program using the following code:

```
cap program drop elast
program elast, rclass
tempname S R num den phi theta psi
qui corr y0c y1c, cov
scalar S=r(Var_2)
```

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

The last line of code returns the bootstrapped standard errors for the own-price elasticity estimates. Section 7.1 in the Code Appendix includes an example do-file that details the code used in this section. Users can copy and paste that code into Stata's do-file editor and estimate the results with appropriate accompanying data/variables described therein. In addition, Section 7.3 reproduces detailed code from Deaton to estimate own- and cross-price elasticities using Deaton's method.

3.3.4 Case study

This section presents the results of econometric estimation of own-price elasticity for a single commodity (cigarettes) using hypothetical HES data applying Deaton's method described above. Results are presented in a step-by-step manner to aid understanding of the technique described in the previous section.

Step 1: Derivation of unit values and other relevant variables

The first step in Deaton's method is to derive unit values as per Equation 3.1 above. Second, other variables used in the analysis are processed as described in Chapter 2. The full list of variables that are used to estimate elasticities are reported in Table 3.1 below. Variables in lines 5–11 in Table 3.1 make up the Z_{ic} vector of household structure and demographic control variables described in Equations 3.2 and 3.3 above.

Step 2: Spatial variation hypothesis

The second step in Deaton's method is to empirically verify that the unit values satisfy the spatial variation hypothesis using ANOVA. The results of the ANOVA exercise are contained in Table 3.2 below.

Table 3.1 Variables used for own-price elasticity estimation

Variable	Description	Obs	Mean	Std. dev.	Min	Max
qcig	Number of cigarettes purchased	9,695	21.58	16.64	1.00	232.50
expcig	Expenditure incurred on cigarettes	9,695	5,314.34	3,803.90	52.00	45,680.00
luvcig	Natural log of unit value of cigarette	9,695	5.53	0.35	3.95	6.49
bscig	Budget share spent on cigarettes	9,695	0.09	0.06	0.00	0.51
lexp	Logarithm of total household expenditures	9,695	11.02	0.56	7.79	13.49
lhsize	Logarithm of household size	9,695	1.07	0.56	0.00	3.00
maleratio	Proportion of males in the household	9,695	0.51	0.25	0.00	1.00
meanedu	Mean education of household	9,694	10.73	2.36	2.00	20.00
maxedu	Highest education by any of the household member	9,694	12.01	2.55	2.00	20.00
sgroup	Dummy variable for social group	9,695	2.29	0.94	0.00	3.00

The outcome of the ANOVA exercise shows that at least 87 percent (*R*-squared of 0.87) of the variation in unit values is explained by between-cluster effects. The *F*-statistic is associated with the hypothesis that there is no spatial variation in prices and it is rejected here. As mentioned earlier, a significant *F*-statistic and R^2 value around 0.5 means the unit values can be used for the purpose of examining price variation and to estimate the price elasticities.

Table 3.2 Testing spatial variation in log unit values

F-statistic	*p*-value	*R*-squared	Adjusted *R*-squared	Observations
30.24	0.000	0.871	0.842	9,695

Notes: The *F*-statistic and the *p*-value are associated with the null hypothesis of no spatial variation in unit values. The *R*-squared measures the proportion of variation in prices taking place between clusters. *n* is the total number of households.

Step 3: Within-cluster regressions in the first stage

The next step is to estimate the within-cluster regressions—that is, the unit value regression and budget share regressions, as per Equations 3.2 and 3.3 above. The results of these regressions are shown in Table 3.3.

Table 3.3 Results from the unit value regression

VARIABLES	Unit value regression	Budget share regression
lexp	0.0855***	-0.0348***
	(0.00384)	(0.00154)
lhsize	-0.0432***	-0.00848***
	(0.00385)	(0.00154)
maleratio	-0.0159***	0.0216***
	(0.00607)	(0.00243)
meanedu	0.00451***	-0.00148***
	(0.00143)	(0.000573)
maxedu	-0.000244	-0.000719
	(0.00132)	(0.000527)
sgp1	0.00204	0.00850***
	(0.00785)	(0.00314)
sgp2	-0.00808*	-0.00476***
	(0.00447)	(0.00179)
sgp3	-0.00265	-0.000609
	(0.00435)	(0.00174)
Constant	4.602***	0.493***
	(0.0388)	(0.0155)
Observations	9,694	9,694
R-squared	0.883	0.377

Note: Standard errors in parentheses; Coefficients of cluster-fixed effects are suppressed for space reasons but are jointly statistically significant. *** p<0.01, ** p<0.05, * p<0.1.

The coefficient of *lexp* in the unit value regression is the quality elasticity or the expenditure elasticity of quality as discussed in Section 3.3.1. It is statistically significant at the one-percent level, implying that quality shading is significant although the magnitude itself is small and perhaps negligible. The results of the budget share regression show that the cigarette budget share declines with household expenditure. Households with higher total expenditures are usually the ones in higher income groups. They tend to allocate a relatively lower share of their budget towards cigarette purchases, and the result here is consistent with that. This result is statistically significant at the one-percent level. The estimated coefficients of the remaining household-specific variables are used to remove their effects, if any, from unit value and budget share variables so that any residual variation in them can be assumed to reflect the measurement errors.

Step 4 and Step 5

Step 4 involves obtaining cluster-level unit value and the cluster-level demand as per Equations 3.4 and 3.5. Step 5 is then a regression of cluster-level demand on cluster-level unit value as per Equation 3.6. These results are not reported here.

Step 6: Obtaining elasticity estimates

The final step applies the formulas in Equations 3.7 to 3.11 to obtain price and expenditure elasticity estimates. Table 3.5 presents estimates of the own-price elasticity of demand for cigarettes in Uganda. Table 3.6 presents estimates of the expenditure elasticity of demand.

The results in Table 3.4 show that cigarette demand in this example is expected to increase by about 5.2 percent every time household income/expenditures increase by 10 percent, as indicated by a significant income/expenditure elasticity coefficient of 0.52. Similarly, the own-price elasticity estimate of -0.8 indicate means that for every 10-percent increase in the price of cigarettes, it is expected that the household consumption of cigarettes would decrease by eight percent. These estimates are within the range of estimates in the literature that uses Deaton's method discussed in Section 3.3.1.

Table 3.4 Estimates of income and own-price elasticity of demand for cigarettes

	Expenditure elasticity	Own-price elasticity
Elasticity coefficient	0.515***	-0.795***
95% confidence interval	[.4762283 .5539125]	[-.8371711 -.7528599]
Bootstrapped standard error	.0198	.0215

Note: Bootstrapped standard errors were calculated by making 1,000 draws. Assuming the estimates follow a normal distribution, the coefficients with *** and ** imply levels of significance at 1% and 5%, respectively.

3.4 Estimation of prevalence elasticity

Unlike normal commodities where price elasticity in the intensive margin is most important for tax modelling purposes, both quantity of consumption as well as prevalence are important when it comes to demerit products such as tobacco, which are used only by a relatively smaller fraction of people. Given that global adult smoking prevalence is 19.6 percent,[86] it is important to understand that outcomes such as smoking or tobacco use (y_i) have two fundamental statistical properties[87]:

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

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

That is, in a distribution of population, there is n_1 number of people who smoke cigarettes in quantities greater than or equal to zero and n_{1+1} to n_2 number of people who do not smoke at all. In other words, the cumulative distribution of cigarette consumption can be characterized as a mixed distribution that is neither discrete nor continuous. If the zero outcomes are sufficiently large, as is

obvious from the 19.6 percent global prevalence of smoking, they cannot be ignored while empirically modeling smoking outcomes. Since the estimation of prevalence elasticity is done using the HES data, it should be noted that the estimated prevalence elasticities are for households and not for individuals within the households.

Just as quantity elasticity was estimated at the household level in the previous section, the prevalence elasticity is also estimated for the household as the unit of analysis. Since a household consists of both smoking and non-smoking individuals, the estimates of prevalence elasticity at the household level may be smaller than or equal to similar estimates for individuals. For example, if there are two smokers in a household and only one quits smoking, the prevalence elasticity estimated based on the HES data would not capture that, while prevalence elasticity estimated using individual-level data would.

There have been several econometric strategies to model outcome variables with large number of zeros and positive magnitudes as a function of a set of exogenous covariates, (x). The traditional approach has been to use the ordinary least squares (OLS) method that treats positive raw quantities of cigarette consumption as the dependent (outcome) variable y_i. This method ignores the zeros completely. One major downside of this approach is the possibility of predicting negative consumption from the econometric model. Ignoring all the zeros also means the OLS estimation is inefficient and, in some cases, biased.[88] Use of OLS with a log transformation of outcome variable y_i partially mitigates some of these analytical issues. But this still could not incorporate the zeros, and those cannot be log-transformed. Moreover, the predicted smoking from the log-transformed model would still be affected by what is called a re-transformation bias.[89, 90]

The presence of a substantial proportion of zeros, as in the case of smoking or tobacco use, in the data typically has been handled by using a two-part model (2PM), which distinguishes between a binary indicator used to model the probability of smoking and a conditional regression model for the positive outcome—that is, in this case, the decision to smoke.[91–93] Although there are alternative econometric approaches to deal with a large number of zero outcomes—tobit models, count data model, and zero inflatfed poison models, to name a few—2PM has been used widely in the context of modeling cigarette or tobacco demand.[94–99] A good exposition of many of these alternative methods can be found in Camerone and Trivedi[16] and as well as many other econometric text books. The first part of the 2PM uses the full sample, including zeros, and estimates the probability of observing positive versus zero outcomes. The second part uses a subsample that consists of only those reporting positive consumption, which is estimated with an econometric model for continuous variables such as OLS or generalized linear model (GLM).[100] Both parts in the 2PM are estimated independently, which allows for independence between the decision to smoke and the decision on how much to smoke. In this toolkit, however, instead of estimating 2PM the traditional way the second part of the 2PM is replaced with Deaton's model discussed earlier, for reasons explained in Section 3.4.1.

In the first part, the goal is to estimate the price elasticity of smoking participation or smoking prevalence. It estimates the probability that an individual would smoke by using a parametric binary probability model, such as logit or probit.[101] In other words, this model estimates whether the price of tobacco impacts an individual's decision to consume tobacco or not, conditional on a set of independent variables. Based on Cameron & Trivedi,[16] a brief introduction to probit and logit models is provided below.

Since they are binary choice models, the dependent variable can take the value of 1 if an individual has positive tobacco consumption and zero otherwise, each having their respective probabilities. That is,

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

The probability mass function for the observed outcome y is $p^y\ (1\text{-}p)^{1\text{-}y}$ with $E(y)=p$ and $Var(y)=p(1\text{-}p)$. A regression model can be formed from this by parameterizing p to depend on an index function $X'\beta$, where X is a vector of variables including the price of tobacco, an individual or household's total consumption expenditures (which is a proxy for their income), and other covariates used as control variables. β is a vector of unknown parameters to be estimated. The conditional probability takes the form $p_i=Pr(y_i=1|X)=F(X_i'\beta)$, where $F(\cdot)$ is a specified parametric function of $X_i'\beta$ ensuring the bounds $0\leq p\leq 1$. It is the choice of the link function $F(\cdot)$ that distinguishes between logit and probit models, as shown in Table 3.5.

Table 3.5 Binary outcome models

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

Note: Table adapted from Cameron & Trivedi (2010)[16].

For both probit and logit models, $F(z)\rightarrow 0$ as $z\rightarrow -\infty$ and $F(z)\rightarrow 1$ as $z\rightarrow +\infty$. Also, $f(z)=^{\partial F(z)}/_{\partial z}$ is positive, since $\partial F(z)$ is strictly increasing. If $y_i^*= x_i'\beta+u_i$ is an unobservable function, the probit model is the case where u_i has a normal distribution, and the logit model is when it has a logistic distribution.

Researchers need to be careful about interpretation of the results, as estimated coefficients do not represent the marginal effects and have no clear interpretation. In binary choice models, marginal effects are not constant but are rather a function of all explanatory variables used in the model. The marginal effects for both logit and probit models are also shown in Table 3.5. The marginal effect for the price variable, for example, will be:

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

Marginal effects are interpreted as an increase in the probability that a household h would have positive spending on tobacco products for an increase in price p_i by one unit. From Equation 3.15, price elasticity is calculated as:

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

where $\overline{pi}$ and *Y* are the average price of tobacco products and smoking prevalence before the price increase, respectively. Since prevalence itself is a rate and elasticity is interpreted as the percentage change in prevalence with respect to a one-percent change in price, a more intuitive way of expressing marginal effects may be in terms of a percentage point change in prevalence as a result of a one-percent increase in price. This can be estimated as follows:

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

For the second part of the 2PM, the quantity elasticities derived in Section 3.3 would suffice. Once the quantity elasticities from Section 3.3 and prevalence elasticity from Section 3.4 are estimated, the total price elasticity becomes a straightforward summation of both elasticities. By assumption, the 2PM estimates elasticities for the extensive and intensive margin, independently. A few studies estimating the elasticity of demand for tobacco products have already been published using the approach described in this section. Those were studies from Bosnia and Herzegovina,[70] Montenegro[77] and Serbia.[80] All of them use a logit regression to estimate prevalence elasticities and Deaton's method to estimate the quantity elasticities.

3.4.1 Preparing data for estimating prevalence elasticities

The preparation of data follows the same steps as already described in Section 3.3.2. The prevalence elasticity estimates would use the same set of variables as used in the case of quantity elasticity. A binary indicator variable is needed for tobacco consumption. However, this variable was also defined in Section 3.3.2 with the variable name *dcig*. For the prices, the unit values will be used as defined earlier.

Unlike the Deaton model estimating quantity elasticity as discussed in Section 3.3, the logit or probit estimation does not correct for quality variation or measurement errors in unit values. Since the individual-level data on prices are not available but only household-level unit values from the HES, there is also possible endogeneity involved when the unit values are used as proxies for price.

One way to partially mitigate this issue is to use a cluster-level price variable instead of using unit values at the individual household level. So, the average unit values at the cluster level are used under an assumption that every household in a given cluster faces the same average unit value. This would take care of endogeneity as well as possible quality variations in unit values across households to an extent (because all household-level variables are used as controls in the prevalence elasticity regression). The larger the number of households consuming tobacco in a cluster, the better the mitigation of associated endogeneity and quality shading issues.

However, this still would not correct for the possible measurement errors in unit values, and unfortunately there is currently no easy solution for this while estimating logit or probit models. For this reason this toolkit continues relying on Deaton's model to estimate the quantity elasticity rather than the conventional 2PM estimating both prevalence and quantity elasticity which does not correct for either quality shading or measurement error in both parts. By using a logit/probit model to estimate the first part and Deaton's model to estimate the second part, this toolkit argues that the presented estimates would be better than the ones estimated using a conventional 2PM.

3.4.2 Estimating price elasticity on the extensive margin with Stata

Step 1: Generating an additional variable for prevalence elasticity estimation

The binary variable indicating smoking status (*dcig*) and the right-hand side variables for the regression to estimate price elasticity were already generated in Section 3.3. The only additional variable required is a cluster average unit value that can be assigned to all households in a given cluster. In the event there are no households available in a cluster that report consumption of tobacco, there may not be a unit value to assign. In that case, an average unit value should be defined at a higher geographical aggregate such as rural/urban or district or region. The model will have at least as many average unit values as the number of clusters with smoking households. The variable *pcig* is used as a proxy for prices in the logit/probit regressions.

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

Step 2: Running the logit regression

The following commands first define a global macro for the independent variables, run the logit regression with the dichotomous *dcig* as outcome variable, and estimate the predicted probabilities of positive outcomes (smoking) in a new variable *yhat_p*.

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

Step 3: Estimating prevalence elasticity

The coefficients obtained from the logit regression are not elasticities. To obtain elasticities (percentage change in the probability that a household consumes cigarettes with respect to a percentage change in cigarette price), it is necessary to use the command *<margins>*. The syntax varies depending on whether the variables above were in levels or in logarithmic form. Since the dependent variable is a binary, it is in levels. If the independent variable, price, is also in levels, the price elasticity is obtained with the following command:

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

So, eyex obtains a percentage change in *y* for a percentage change in *x*. In this case, using *eydx* as in *<margins, eydx(pcig)>* would produce the so-called semi-elasticity, which represents the percentage change in *y* for a unit change in *x*. However, if the price variable *pcig* is in logarithmic form, the code for obtaining price elasticity would be

```
margins, eydx(pcig)
```

In this case, *eydx* obtains the percentage change in *y* for a unit change in *x*, which is a logarithm of price. The same procedure can be used for estimating expenditure elasticity of prevalence. The formulas used by the margins command to estimate the respective elasticities and their standard errors are all available from the Stata base reference manual for each version of Stata.

Step 3: Regression diagnostics

Before the results of the model can be used to make any statistical inference, and for the analysis to be valid, it is necessary to check whether the model meets the assumptions of the binary choice models. Possible checks include testing whether the model is correctly specified (the specification error test), whether the overall model is statistically significant (the goodness-of-fit test), and whether the regressors are orthogonal (the multicollinearity test).

The **specification error test** is conducted to confirm that the probability function is correctly specified. It is done with the command *<linktest>* right after the logit regression command. The command *<linktest>* rebuilds the model using the linear predicted value *(_hat)* and the linear predicted value squared *(_hatsq)* as the predictors. If *_hat* turns out to be significant, it means that meaningful predictors are selected for the model and the model is correctly specified, so any statistical significance of the variable *_hatsq* should be purely by chance. A statistically significant *_hatsq* may be indicative of specification error because of omitting some important variables or omitting some interaction effects of included variables in the model. The model specification needs to be changed until it is able to pass the specification error test. This can be done many ways including adding new variables, adding higher order polynomials of one or more of the existing variables, and adding new interaction effects between existing variables.

The **goodness-of-fit tests** are the likelihood ratio (LR) test and Hosmer and Lemeshow's goodness-of-fit (HL) test. The LR statistics are reported by default when the model is estimated: the model fits well if the LR statistics are statistically significant. The HL test tests whether the predicted and observed frequency closely match, and it can be obtained with the command *<lfit, group (10) table>* right after running the logit regression command. An insignificant p-value for the HL chi-square statistics would indicate that the model fits the data well. Alternatively, Stata's *<fitstat>* command after the logit regression would return a number of fitness statistics including Akaike information criterion (AIC) and Bayesian information criterion (BIC).

The multicollinearity test is conducted to verify whether variables are orthogonal to each other (that is, completely uncorrelated). A Stata user-written program *<collin>* test for multicollinearity can be used with the command *<collin var1 var2>*. The closer the tolerance (which equals $1 - R^2$) and the variable inflation factor (VIF = 1 / tolerance) to one, the less severe the multicollinearity problem in the model. As a rule of thumb, a tolerance of 0.1 or less (and VIG of 10 or higher) should be concerning. The *collin* program can be installed from within Stata using the command *<findit collin>*. It is worth noting that the "problem" of multicollinearity, however, is just a matter of degree, and there is hardly any practical solution to it. Dropping a relevant variable would result in omitted variable bias. So, as long as the estimated coefficients of included variables have sufficiently low standard errors and narrow confidence intervals, multicollinearity may be ignored altogether.

Section 7.1 in the Code Appendix includes an example do-file that details the code used in this section.

3.4.3 Case study

The same data used in Section 3.3.4 are used to estimate the smoking prevalence elasticity of price and expenditure. Table 3.6 shows the coefficient estimates from the logistic regression along with the price and expenditure elasticities of prevalence estimated using the <*margin*> command in Stata. The price elasticity of prevalence for smoking is negative and significant, yet very low in magnitude. The expenditure elasticity, however, has a positive sign and a coefficient that is higher in magnitude.

Since the price elasticity of prevalence for smoking (elasticity in the extensive margin) is -0.0528 and the price elasticity of quantity of smoking (elasticity in the intensive margin) is -0.795, the total price elasticity of smoking is estimated to be (-0.0528) + (-0.795) = -0.8478.

It is ideal to estimate the prevalence elasticities first, as this uses the full sample. Later, keep only the households reporting positive consumption and proceed to the quantity elasticity estimation. However, the explanation in this toolkit is in reverse order since it is important to understand the reason behind using Deaton's method for elasticity estimation when using HES data and why that method is chosen for the estimation of quantity elasticities in the second part of 2PM instead of GLM, which is used conventionally for estimating the second part in 2PM.

Table 3.6 Results of logistic regressions and elasticities

VARIABLES	Coefficient	Standard errors
pcig	-0.000317**	(0.000151)
lexp	0.956***	(0.0328)
lhsize	0.0760**	(0.0347)
maleratio	0.582***	(0.0530)
meanedu	-0.0315**	(0.0129)
maxedu	-0.0298**	(0.0122)
sgp1	0.0641	(0.0686)
sgp2	-0.506***	(0.0395)
sgp3	-0.107***	(0.0412)
Constant	-10.24***	(0.328)
Price elasticity	-0.0528**	(0.0251944)
Expenditure elasticity	0.5883***	(0.0203771)
Observations	25,188	

Note: *** p<0.01, ** p<0.05, * p<0.1.

3.5 Estimating elasticities by income groups

Both prevalence and quantity elasticities can be estimated by household income groups or by other socioeconomic categories. This toolkit presents only the estimation of elasticities by income group, which is the most commonly used. The first decision in estimating price elasticities by income group is on the number of quantiles to which the total sample of households is split based on income. Whether the analysis is done by terciles, quintiles, or deciles depends on its objective and the country context, but also on the size of the HES sample.

Price elasticity estimated using Deaton's method depends on the number of clusters (that is, the consistency property), since it is derived based on the cluster-level averages, but it does not depend on the number of households in each cluster. However, the smaller the number of households per cluster, the higher the measurement error. As the price variable is likely endogenous, Deaton addresses this issue by estimating elasticity using the cluster-average price. However, the smaller the number of households per cluster, the less likely the problem of endogeneity is addressed.

So, once the HES sample of households is split into three, five, or 10 quantiles of households based on their income, the cluster size in each quantile becomes even smaller. In other words, a cluster in the original sample may now split into two or more subclusters depending on the income of households in that cluster. This smaller size of the cluster further exacerbates the problem of endogeneity in the price variable and the measurement error problem. This is especially the case in countries with a smaller population and smaller HES sample.

It is therefore advised to use a smaller rather than larger number of income groups. Even in cases where the HES sample is large enough to have even 10 income groups, the difference in estimated elasticities between the bottom income groups may be so small to suggest that they may be considered as one income group. The same may be observed for the middle and the top income groups. It is possible to conduct a statistical test to determine whether the elasticity coefficients across different subgroups are statistically different from each other. This may also be a guiding principle on deciding the number of subgroups to be used for the analysis.

The second decision in estimating price elasticities by income group is whether the same average price shall be used for all households in subclusters that are derived from the same cluster in the original sample, regardless of their income, or should the average price per subcluster be different. Studies that have estimated price elasticities by socioeconomic group using Deaton's method have so far assumed different average prices by income group. However, as mentioned above, to address potential endogeneity in the price variable, Deaton's method derives a cluster-level average price and applies it to all households in that cluster.

This is based on the assumption that there is no price variation within a cluster, as all households in a cluster live close enough to each other and therefore make purchases in the same market and are surveyed at the same time. Moreover, by using cluster-level average price, the potential problem of households' self-selection based on their income level and preferences is minimized if not eliminated. So, both intuitively and in line with the assumptions of Deaton's method, the same cluster-level average price should be applied to all households in the same cluster, regardless of their level of income. As Deaton's method is designed for estimating elasticity at the population level, but not within that same population by SES groups, the application of the method requires an adjustments in the code, which is presented in this toolkit.

3.5.1 Preparing data for estimating elasticity by income groups

Categorizing households by income group is the first step to estimating elasticities by income group. As information on income is usually not provided in HES, it is proxied by the sum of all reported household expenditure during the reporting period, as in

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

where x_{hc} is the total reported spending of a household *h* in cluster *c*, which is the sum of spending on items i reported by that household.

As HES data are at the household level, households should be divided to income groups not based on total household spending, but based on spending per member of household:

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

where x_{phc} is reported total spending per household member (*p* stands for per capita) of a household *h* in cluster *c* and s_{hc} is the number of household members. The total household expenditure variable (*exptotal*) as well as household size variable (*hsize*) were already defined in Section 3.3. Then, the per capita household expenditure is generated using *<gen exppc=exptotal/hsize>*. The quantiles can be created with the command *<xtile inc = exppc [w=weights], nq(3)>* where option *nq (.)* specifies the number of quantiles and weights is the variable for household weights in the survey.

Three income groups (low, middle, high) are chosen for the purpose of this analysis, as these would provide adequate variation across income groups while ensuring the maximum number of subclusters in each group along with reasonable subcluster sizes. The command creates a new variable inc, which distributes households into three terciles using appropriate weights from the survey. It is important to note that—because the weights are used—the three groups may not have an equal number of observations. A subcluster variable (*subclust*) is also created using the existing cluster variable (*clust*) and the new income group variable (*inc*) using the command *<egen subclust=group(clust inc)>*.

Just as in Section 3.3, subclusters with fewer than two households reporting tobacco consumption are dropped from the analysis. This can be done as follows:

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

These along with the other household specific variables used in the Section 3.3 would be sufficient to estimate the quantity elasticities by income group using Deaton's method. It should be noted that many HES data sets come with clusters that are already quite small. When an analysis is done by subgroups, and subclusters are generated to derive average budget shares at the subcluster level, it is quite possible to end up with several subclusters having fewer than two households with smokers. If that happens, a lot of observations will be dropped from the analysis and, as a result, the elasticity estimates using the overall sample will not necessarily fall within the range of elasticities estimated for different subgroups. This is because the combined samples used for an analysis for subclusters will not have the same observations as the sample of all households before creating subgroups.

For the estimation of prevalence elasticities, a price variable is created as the average of unit values by the newly created subclusters, as follows:

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

Testing the significance of the difference of elasticity coefficients across income groups can be implemented with the help of a seemingly unrelated regression (SUR). For this, models by subgroups (income) are simultaneously estimated, and their results stored first. Later, a chi-squared ($\chi 2$) test of significance can be done with the stored results to test the statistical significance of different linear combinations or equality of coefficients.[102, 103] Stata's *<suest test>* command can be used for this purpose. More details are provided below.

3.5.2 Estimating prevalence elasticity by income group with Stata

Since the necessary variables are already defined and subclusters with fewer than two households reporting positive consumption are dropped, the codes from section 3.4.2 can be used by extending the same codes with a simple loop command for the income group, as shown below.

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

The price elasticity can be obtained with the command *<margins, eyex(pcig)>* and expenditure elasticity with the command *<margins, eydx(lexp)>*, as indicated earlier.

Table 3.7 shows the results of the logistic regression along with both price and income elasticity of smoking prevalence obtained with the margin commands using the same data as in previous sections. One can see that price elasticity of prevalence, although negative in sign, is not statistically significant for any of the income groups. The expenditure elasticities are positive and statistically significant.

Testing the significance of the difference of elasticity coefficients across income groups can be done using the following code:

```
Global xvar “pcig lexp lhsize maleratio meanedu maxedu sgp1 sgp2 sgp3”
local append “replace”
forvalues i=1/3 {
    logit dcig $xvar if inc== `i’
    outreg2 using PrevalenceElastInc.doc, ctitle (Income group: `i’) `append’
```

```
        predict yhat_p `i', pr
        margins, eyex(pcig)
        estimates store inc `i'
        local append "append"
}
suest inc*
test [inc1_dcig]pcig-[inc2_dcig]pcig=0
test [inc1_dcig]pcig-[inc3_dcig]pcig=0
test [inc2_dcig]pcig-[inc3_dcig]pcig=0
```

Table 3.7 Results of logistic regressions and prevalence elasticities by income group

VARIABLES	Low-income	Middle-income	High-income
pcig	-7.76e-05 (0.000338)	0.000104 (0.000297)	-0.000117 (0.000273)
lexp	1.178*** (0.107)	0.793*** (0.220)	0.438*** (0.0885)
lhsize	0.247** (0.106)	0.246 (0.222)	0.182* (0.0933)
maleratio	0.711*** (0.138)	0.929*** (0.112)	0.278*** (0.0839)
meanedu	-0.0376 (0.0255)	-0.0596** (0.0252)	0.00293 (0.0278)
maxedu	-0.0279 (0.0239)	0.00471 (0.0236)	-0.0570** (0.0268)
sgp1	0.254** (0.117)	0.102 (0.160)	0.0787 (0.170)
sgp2	-0.319*** (0.0876)	-0.563*** (0.0800)	-0.638*** (0.0708)
sgp3	-0.0342 (0.0798)	-0.165** (0.0795)	-0.118 (0.0861)
Constant	-12.66*** (1.024)	-8.758*** (2.179)	-4.213*** (0.927)
Price elasticity	-0.011926 (0.0519689)	0.0151632 (0.0432479)	-0.016525 (0.038618)
Expenditure elasticity	0.706927*** (0.0646166)	0.4342673** (0.12084)	0.2208638*** (0.0446629)
Observations	5,835	6,291	6,135

Note: Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1.

After each estimate is stored with <*estimates store name*> command, the command <*suest inc**> will return the results of a SUR regression. After this, differences between each coefficient can be tested with a simple test command, which, for example, tests the price elasticity coefficient among income group 1– price elasticity coefficient among income group 2 = 0. The test would return a chi-squared ($\chi 2$) statistic. A significant statistic here would imply that the difference in elasticity coefficients between income groups 1 and 2 is statistically significant.

3.5.3 Estimating quantity elasticity by income group with Stata

In the case of Deaton's method, there are some significant changes in the code presented in Section 3.3.3 to estimate the elasticities of demand for a single commodity. The code below estimates both price and income elasticities of demand for cigarettes along with their bootstrapped standard errors for all three income groups simultaneously. First of all, the non-smoking households are dropped from the analysis with the command <*keep if dcig==1*>, as only the conditional elasticities are estimated. The spatial variations in unit values can be tested the same way as done in Section 3.3.3 using the command <*anova luvcig clust*>. Note that this is not done at the subcluster level, but at the cluster level. This is because it is assumed that all households in all subclusters within a cluster are facing the same average market prices.

Estimating within-cluster first-stage regressions and measurement error variances

The first stage unit value regression will be the same for all income groups, while the first stage budget share regression is done separately for each income group. The parameters from these regressions are saved for the elasticity estimation in the second stage.

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

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

Unlike in Section 3.3.3, the covariance of u_0 *(sigma22)* and that of u_1, the coefficient b_0, and the purged *y0cig* for the second stage are all different for each income group.

Estimating income or expenditure elasticities

The following code estimates the expenditure elasticity coefficient along with the bootstrapped standard errors and stores them in a separate file such as "deatonExpElast.doc" in a ready-to-use format.

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

Preparing data for between-cluster regression

The next step involves averaging the variables *y1cig* and *y0cig* by clusters to generate *y1c* and *y0c* respectively, so they can be used for a between-cluster regression of *y0c* on *y1c* to derive the own-price elasticity. But only *y1c* is generated below as that is done for all income groups together. The *y0c* is generated along with the elasticity estimation in the next stage as bootstrap standard errors are estimated for each income group elasticity, one at a time.

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

Estimating own-price elasticity

The code below estimates the own-price elasticity as well as the bootstrapped standard errors separately for each income group and stores them in a file such as "DeatonPriceElast.doc" in a ready-to-use format.

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

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

The Code Appendix Section 7.2 reproduces the code above along with that of estimating prevalence elasticities by income groups. Table 3.8 shows the results of own-price and expenditure elasticity for cigarettes on the intensive margin using the same data as in the previous section. All elasticity coefficients are highly significant and have the expected signs. The total elasticity can be obtained by summing up the prevalence and quantity elasticities for each income group, respectively.

If one wants to also compare whether the elasticities are statistically significantly different across the income groups it is not possible to use suest as before. Instead, the following codes can be used to check the statistical differences in price elasticities across income groups. The test command at the end of the code shows whether the estimated elasticities differ significantly from each other.

Table 3.8 Price and expenditure elasticity of demand for cigarettes by income group

	Expenditure elasticity	Own-price elasticity
Low-income		
Elasticity coefficient	0.649***	-0.808***
95% confidence interval	[0.506 - 0.791]	[-0.8941 -0.7216]
Bootstrapped standard error	(0.0714)	(0.0440)
Observations	2,333	728
Middle-income		
Elasticity coefficient	0.605***	-0.826***
95% confidence interval	[0.369 - 0.841]	[-0.9014 -0.7507]
Bootstrapped standard error	(0.1218)	(0.0385)
Observations	2,844	883
High-income		
Elasticity coefficient	0.422***	- 0.816***
95% confidence interval	[0.321 - 0.523]	[-0.8729 -0.7586]
Bootstrapped standard error	(0.0503)	(0.0292)
Observations	3,044	920

Note: Bootstrapped standard errors were calculated by making 1,000 draws. Assuming the estimates follow a normal distribution, the coefficients with *** and ** imply levels of significance at 1% and 5%, respectively.

```
cap program drop elast
program define elast, rclass
forvalues i=1/3 {
    preserve
    egen y1c `i'= mean(y1cig), by(clust)
    keep if inc== `i'
    sort clust
    egen y0c `i'= mean(y0cig `i'), by(clust)
    egen n0c `i'= count(y0cig `i'), by(clust)
    egen n1c `i'= count(y1cig), by(clust)
```

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

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

3.6 Estimating elasticities when unit values are not available from HES

Deaton's approach allows for the estimation of demand and computation of own- and cross-price elasticities using quantities and unit values obtained from HES data. However, sometimes HES data collects information only about the expenditures households incur for different commodity groups. It does not provide any information on quantities purchased and, as a result, it is not possible to construct unit values whose spatial variation can be used to inform variability in prices at the household level. In this case, Deaton's approach as discussed in this chapter cannot be applied. Given that HES otherwise provide rich information on household consumption along with that of tobacco products, it would be unwise to ignore such data simply because quantity information is not available. Fortunately, there are methods to recover unit values (or pseudo unit values) so that the same can be used for the estimation of demand functions and to derive price elasticity.

Traditionally, when quantity information is not available in HES, the external sources of price variability obtained from aggregate national price indices such as consumer price indices (CPI) and weighted or unweighted average price data available at the local administrative levels are often merged with household expenditure to obtain estimates of price elasticities.[104] Popular demand systems such as AIDS or quadratic almost ideal demand system (QAIDS) are often employed while using such price indices to estimate demand functions. However, this approach is criticized for not accounting for spatial and household variability, thus resulting in distorted estimates of demand parameters and not being coherent with the theory.[105–108] Moreover, aggregate price indices are often highly correlated and may suffer from endogeneity problems.[109]

Recent literature,[110] however, suggests that construction of household-level price indices (Stone-Lewbel (SL) prices [111]) for commodity groups can mitigate the issues around using only aggregate price indices in situations where quantity information is not available in the survey. SL price indices for commodity groups are constructed using information on the subgroup budget shares, household demographic characteristics, and the aggregate national price indices, and allow for household-level prices or unit values to be recovered.[110] It has been found that the use of household-specific SL prices results in demand parameters that are more precise and economically plausible than the ones obtained by using only aggregate price indices.[108] The user-written program in Stata, *<pseudounit>*[104] helps to estimate such unit values (pseudo unit values) using this method for HES with no quantity information.

A recently proposed Exact Affine Stone Index (EASI) implicit Marshallian demand system makes use of these methods to estimate price elasticity[110, 112] and has several advantages over traditional demand systems such as the AIDS. Different empirical methods for the computation of the SL price index for product aggregates are also available in literature.[113] This toolkit, however, does not go into these issues and the developments around them as, more often than not, HES data provide both quantity and expenditures for different commodities of interest. However, readers having HES data without quantity information should familiarize themselves with the literature in this section before attempting to estimate price elasticity from such data.

4 Estimating the crowding-out effect of tobacco spending

4.1 How tobacco spending crowds out spending on other goods and services

While the global prevalence of tobacco use has declined from 26.7 percent in 2010 to 22.3 percent in 2020, much of that decline has occurred in LMICs.[114] The majority (around 80 percent) of the world's approximately 1.3 billion current tobacco users live in LMICs.[115] The prevalence of smokeless tobacco use is also found to be much higher in lower middle-income countries. Of the estimated 335 million adults using smokeless tobacco in the world, 266 million are in southeast Asia.[114] Several studies have also shown that tobacco use is disproportionately higher among relatively poor people. A meta-analysis of 201 studies by WHO found a statistically significant association between higher prevalence of current smoking among adults and lower income, for both men and women.[116]

Expenditure on tobacco accounts for a significant portion of the household budget in many countries, ranging from one percent in countries such as Mexico and Hong Kong to 11 percent in countries such as Zimbabwe and China.[117] Households operate based on limited disposable income and, as a result, when they spend their limited budgets on tobacco it has a huge opportunity cost. It inevitably means they have to reduce expenditures on certain other goods and services, some of which may be necessities such as food, clothing, and housing. The idea that households that spend money on consuming tobacco divert funds from the consumption of other commodities is called the "crowding out" effect of tobacco spending.

There were some early attempts to explain the issue of crowding out with descriptive analysis of data from Bangladesh[118] and China[119] in the years 2001 and 2002, respectively. A formal empirical examination of the idea of crowding out due to tobacco spending using econometric methods came later from the US[120] and China[121] in the years 2004 and 2006. These studies, however, could not explicitly model the issue of endogeneity present in such analysis.

The current generation of econometric methods estimating the crowding-out impact of tobacco spending started in 2008 using household expenditure data from India.[117] It uses IV techniques to account for the possible endogeneity in the demand system while treating tobacco spending as a regressor and finds that spending on tobacco crowds out food, education, and entertainment while crowding in expenditures on health, clothing, and fuels. Studies using similar econometric methods and household expenditure data have been done in other countries such as (chronologically) Taiwan,[122] South Africa,[123] Cambodia,[124] Zambia,[125] Turkey,[126] Bangladesh,[127] Mauritius,[128] Chile,[129] Vietnam,[130] Ghana,[131] and Kenya.[132] There are also studies using different methods to examine crowding out in Indonesia,[133] South Africa,[134] South Korea,[135] and other LMICs.[136]

Table 4.1 Econometric studies on the crowding-out effect of tobacco spending

Year	Authors	Country	Method	Survey data used	Items crowded out
2004	Busch et al.[120]	USA	Separate OLS regressions	Consumer Expenditure Survey	Clothing, housing
2006	Wang et al.[121]	China	Fractional logit model	Primary survey	Education, agriculture equipment maintenance, savings
2008	John, RM[117]	India	Instrumental variables (IV)	National Sample Survey	Food, education, entertainment
2008	Pu et al.[122]	Taiwan	(IV)	Survey of Family Income & Expenditure	Clothing, medical care, transportation
2008	Koch & Tshiswaka-Kashalala[123]	South Africa	(IV)	The South African Income and Expenditure Survey	Education, fuel, clothing, health care, transportation
2009	Block & Webb[133]	Indonesia	Reduced-form equations	Nutrition surveillance system data	Food
2012	John et al.[124]	Cambodia	IV	Cambodia Socio-Economic Survey	Food, education, clothing
2014	Chelwa & Walbeek[125]	Zambia	IV	Living Conditions Monitoring Survey	Food, schooling, clothing, transportation, equipment maintenance
2015	San & Chaloupka[126]	Turkey	IV	Turkish Household Budget Survey	Food, housing, education, durable/non-durable goods
2015	Do & Bautista[136]	40 LMICs	Random-slope models	World Health Survey	Education, health care
2018	Husain et al.[127]	Bangladesh	IV	Household Income and Expenditure Survey	Clothing, housing, education, energy, transportation, communication
2018	Paraje & Araya[129]	Chile	Quadratic AIDS model	Chilean Household Budget Survey	Health care, education, housing
2018	Ross et al.[128]	Mauritius	IV	Household Budget Surveys	Transportation, communication, health, education
2019	Chelwa & Koch[134]	South Africa	Genetic matching/ nonparametric	Income and Expenditure Surveys	Select food items

Table 4.1 Econometric studies on the crowding-out effect of tobacco spending (cont'd)

Year	Authors	Country	Method	Survey data used	Items crowded out
2020	Masa-ud et al.[131]	Ghana	GMM-3SLS / IV	Ghana Living Standards Survey	Food, housing, health care
2020	Nguyen & Nguyen[130]	Vietnam	GMM-3SLS / IV	Vietnam Household Living Standard Survey	Education
2020	Nyagwachi et al.[132]	Kenya	Matched difference in differences	Kenya Integrated Household and Budget Survey	Education, communication, and some food items
2021	Jin & Cho[135]	South Korea	Matched difference in differences	Korean Household Income and Expenditure Survey	Select food items
2021	Djutaharta et al.[137]	Indonesia	AIDS Model	SUSENAS, PODS, and RISKESDAS	Various food items
2022	Vladisavljevic et al.[138]	Serbia	IV	Serbian Household Budget Survey	Food, clothing, education, recreation
2022	Wisana et al.[139]	Indonesia	GMM-3SLS / IV	SUSENAS	Food, clothing, housing, utilities, education, health care
2022	Mugoša et al.[140]	Montenegro	IV	Household Budget Survey	Clothing, housing, and education
2022	Gómez et al.[141]	Mexico	GMM-3SLS	National Survey of Household Income and Expenditure	Food, alcohol, transport, durables

Table 4.1 above provides a summary of the different econometric studies that have been done to examine the crowding-out impact of tobacco spending. As one can see, the IV technique is the preferred method adopted by most of the studies from the past 15 years. Some of the more recent studies, however, have been critical of the IV technique and have instead proposed the use of matched difference in differences (MDID) methods as an alternative. However, the data requirement for the MDID method can be more restrictive, as it would ideally require household panel data. Most crowding-out studies find that spending on tobacco crowds out expenditures on necessary items of household consumption such as food, clothing, housing, and education, among others, implying that tobacco spending can have developmental and intergenerational impacts.

4.2 Importance of intra-household resource allocation

Households often pool resources from individual family members and make decisions on spending or allocating budgets among alternative consumption goods that are required by each individual member. In most if not all HES, the household is the unit for which consumption is reported. However, how the distribution of consumption occurs among family members is not reported. If the allocative decisions are made by certain adult members in a household—often males in several LMICs[142,143]—their impact on social welfare is uncertain. As Deaton points out,[8] if women systematically get less than men, or if children and old people are systematically worse off than other members of the household, social welfare will be overstated when using measures that assume everyone in the household is equally treated.

The intra-household resource allocation decisions become all the more important when disposable incomes are reduced once money is allotted for unproductive spending such as spending on tobacco. Given that consumption of tobacco is more prevalent among males than females in most countries,[144] if the allocation decisions are made by the male heads in a household, they could potentially be unfavorable to women and/or children within those households. In fact, some of the findings from the crowding-out literature described above underscore this.

For example, when educational expenses are compromised as a result of increased allocation on tobacco consumption, it directly impacts children in a household and their future earning potential while imposing long-run intergenerational impacts on society. The literature from India[117] shows tobacco-spending households systematically allocate less money on clean cooking fuels and allocate more money to unclean fuel sources such as firewood, which may be more hazardous to the women who engage in collecting it and burning it while cooking.

Since tobacco consumption is largely addictive, it is quite possible that households pre-allocate a certain portion of the budget for the purchase of tobacco. This means the household has to maximize its utility by optimally allocating the remaining budget (total minus the pre-allocated budget on tobacco) among alternative goods. Certainly, as the disposable budget is reduced after the pre-allocation, some compromises have to be made. Crowding-out studies have found that compromises are made in the case of necessary commodities like food, education, and clothing, which may directly impact the health and development of all members of a household. Therefore, it is important that tobacco control policies address these challenges.

4.3 Comparison of mean budget shares

Checking the differences in mean budget share or mean expenditure spent on different commodity groups between tobacco-spending households and non-tobacco-spending households provides a preliminary indication of potential compromises, if any, made as a result of tobacco spending. This section examines these differences by dividing households into different groups on the basis of their tobacco spending habits and comparing the share of budget each group allocates to the purchase of different commodity groups.

Step 1: Creating average budget shares by type of household

As a first step, create a categorical variable *tob* which takes the value of 1 if households spend any money on tobacco and 0 otherwise. As an example, *exptobac* is the variable representing the amount spent on tobacco by a household as extracted from HES. Then, the indicator variable tobacco can be generated, and their values can be labeled with the following commands:

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

Generally speaking, there are 10 commodity groups—*tobacco, food, health care, education, housing, clothing, entertainment, transportation, durables*, and *other*—that exhaust the household budget. Most studies in the literature on crowding out have considered some or all of these for their analysis. The variables representing the expenditures on these commodities are *exptobac, expfood, exphealth, expeducn, exphousing, expcloths, expentertmnt, exptransport, expdurable*, and *expother*, respectively, as extracted from the HES data.

Note that all variables have the same prefix *exp*. This way of naming variables makes further analysis simpler. For comparing the mean budget shares dedicated to these products between tobacco spenders and non-tobacco spenders a budget share variable is defined for each commodity group. Given the total expenditures on all items together as *exptotal*, the budget share on each of the commodity group can be generated with the following loop command:

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

New variables for budget shares with the prefix (*bs_*) will be defined for each of these product categories.

Step 2: Testing if the difference in mean budget share is statistically significant

A statistical test of the equality of mean budget shares between two groups (tobacco spenders and non-tobacco-spenders) is a two-sample student’s *t*-test for the equality of mean. The *t*-test can be performed in Stata with the command <*ttest bs_food, by(tob) unequal*>, where *tob* is the binary variable indicating the status of tobacco spending defined in Step 1. This will compare the budget share dedicated to food by tobacco-spending households and non-tobacco-spending households and test if the difference is statistically significant. The null hypothesis is that the difference in mean budget share = 0. The *t*-statistic for the difference in mean is also reported. As a rule of thumb, if the absolute *t* value is greater than 2, the null is rejected, and it may be concluded that the difference in mean budget share observed is statistically significant.

The *t*-test, however, does not allow the use of survey weights. It does not allow the use of Stata's <*svy*> command either. As a result, the average budget shares computed for tobacco users and non-users under the <*ttest*> command can be biased. It would be ideal to compute the budget shares for both groups after weighting them with appropriate survey weights or to use the "svy" prefix after declaring the survey design of the data with the <*svyset*> command as explained in Chapter 2. The above *t*-test in this case can be done as follows:

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

Here, *weight* is the variable for survey weight. The command <*lincom*> reports the difference in the weighted mean budget shares between the two groups and shows the *t*-test as well as the *p*-value for the null hypothesis that the difference in mean = 0. This method will produce identical estimates as the ones from *t*-test if weights were not used.

Instead of using weights in the command above, the command <*svy: mean bs_food, over(tob)*> can also be used after declaring the survey design. Alternatively, the command <*test (_b[c.bs_food@0.tob] = _b[c.bs_food@1.tob])*> can be used, which performs a Wald test instead of the *t*-test performed by <*lincom*>. Since mean budget shares from HES are being estimated, an option of the test that allows either using the weight or using the "svy" prefix instead of using a direct *t*-test that does not allow using weights at all should be used.

Step 3: Reporting test results

For the purpose of reporting, one only needs to know the mean budget shares for the given commodity groups, the difference in mean budget shares, and the statistical significance of the difference as indicated by the value of the *t*-statistic. A program is provided below for all ten commodity groups:

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

The code above will list a table with the budget shares for non-tobacco-spending, spending, difference in the budget shares, and t-statistic for the test of equality of mean budget shares between tobacco spenders and non-spenders for each of the commodity groups in the local macro *items*.

4.4 A framework for the empirical examination of crowding out

The simple *t*-test of equality of mean, as discussed in the previous section, does not control for other household-specific characteristics that may influence budget allocation decisions. By not controlling for these, it is possible to inadvertently attribute allocation decisions to a household's tobacco-spending habits. For this reason, there is a need for a formal econometric model that can explain whether households that spend on tobacco systematically reduce their expenditures on other commodity groups and, if so, which ones. This section describes the conceptual and econometric approach that is followed in most of the current literature to estimate the extent of crowding out due to tobacco spending. In addition, the section discusses some methodological improvements on the existing literature on this subject.

4.4.1 A theoretical framework to examine crowding out

Microeconomic theory teaches that the solution to an individual's utility maximization subject to a budget constraint returns a set of unconditional Marshallian demand functions of the form:

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

where q_i is the quantity of i^{th} good consumed, Y is total expenditures, h is a vector of characteristics, and $p_1,\dots,p_n$ are the prices of n commodities in an individual's utility function. Given that household expenditures are reported for the whole household as a single unit, a household-level demand function is used and needs the assumption that the household seeks to maximize a single utility function. If a household's demand for one of the goods—say, tobacco—is predetermined, there are conditional demand functions. The theoretical framework for this is detailed in Pollak (1969).[9] The idea is that the household would maximize the following utility function:

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

Where $\bar{q}_n$ denotes a household's demand for tobacco and $M=Y-(p_n*\bar{q}_n)$. Solving this for $n-1$ goods yields the following conditional demand function, conditional on the consumption of the n^{th} good (tobacco in this case):

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

The demand function of any given good (q_i) here is conditional on the prices of all commodities except the conditioning good (q_n), total remaining expenditure (M) after deducting expenditures on conditional good, quantity of the conditioning good ($\bar{q}_n$), and a vector of household characteristics (h). When dealing with goods that are not consumed by many households (such as tobacco) it is advantageous to use conditional demand functions as noted by Browning & Meghir.[145]

4.4.2 The econometric model to examine crowding out

This section discusses a specific econometric equation that is estimated for examining the crowding-out impact and a brief overview of possible estimation methods that are used in the literature so far, along with their shortcomings. It then proposes an alternative estimation method that is more efficient and theoretically preferred.

Specification of the econometric model

The empirical implementation of the model requires the use of a specific functional form. The literature on crowding out has largely used the QAIDS[146] to estimate the impact of crowding out. Since direct price information is often not available for different commodity groups from household surveys, Engel curves, which allow work with expenditures, are used for the econometric specification. QAIDS, with the presence of a quadratic income term, while being consistent with the utility theory, permits goods to be luxuries at some income levels and necessities at others.[117] The conditional Engel curve takes the following form for the good *i* and household *j*:

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

where $w_{ij}=p_{ij}q_{ij}/M_j$ is the budget share allocated by the j^{th} household to the i^{th} commodity group out of the remaining budget (M_j) after deducting the expenditures on tobacco, $p_{nj}\bar{q}_{nj}$ is the expenditures on tobacco, h_j is a vector of household characteristics allowing for the preferences to be heterogeneous,[147] *lnM* and lnM^2 are the natural logs of *M* and M^2, which is the expenditure after deducting the expenditure on tobacco, and u_{ij} is the random error term.

Estimation method 1: Equation-by-equation instrumental variables estimation (2SLS)

The model as specified in Equation 4.4 cannot be estimated with the OLS method as the variables $p_n \bar{q}_n$ and *lnM* are likely endogenous because of the simultaneity involved. If this is indeed the case, these variables will be correlated with the error term u_{ij} and could result in biased and inconsistent OLS estimates. In other words, a fundamental OLS assumption that the model error term is uncorrelated with the regressors (that is, $E(u/x)=0$) is violated and the OLS estimates fail to give causal interpretation. In such cases, if exogenous variables that are correlated with these endogenous regressors but are not correlated with the error term (IVs) can be found, the IV method could be used to estimate the parameters more consistently. This is also sometimes referred to as a two-stage least-squares (2SLS) estimation.

The IV estimator, however, is less efficient than OLS and should be used only if there are endogenous variables present in the model. This can be tested with the Durbin-Wu-Hausman (DWH) test of exogeneity,[148] if the errors are homoskedastic. If the errors are heteroskedastic, different tests such as Wooldridge's score test, an auxiliary regression-based test, or *C*-statistic are usually used depending on the type of heteroskedasticity assumed.[149] All studies in the current generation of crowding-out literature show that these variables are indeed endogenous.

The IV estimation provides a consistent estimator under the very strong assumption that a valid instrument *z* exists that satisfies two conditions: (1) instrument *z* is partially correlated with the endogenous regressors *x* (that is, $Cov(x_i,z_i)\neq 0$) and (2) instrument *z* affects the dependent variable w_i only through the regressors or z itself does not cause wi (that is, $E(u/z)=0$). The first condition is

sometimes called the inclusion restriction, while the second condition is popularly known as the exclusion restriction. While the inclusion restriction can be tested statistically by checking the association between an instrument (z) and endogenous variables (x) with a reduced-form regression—the stronger the association, the stronger the identification of the model—testing the exclusion restriction is impossible, especially in the just-identified case (that is, when the number of instruments equals the number of endogenous regressors).

In the over-identified case (that is, when there are more instruments than the number of endogenous regressors), a test of over-identifying restrictions can be conducted to test the exogeneity of instruments, provided the parameters of the model are estimated using the optimal generalized method of moment (GMM).[16] This test again differs depending on whether the errors are homoskedastic. If the errors are homoskedastic, a Sargan or score test should be performed. If not, Hansen's J-statistic or Hansen-Sargan statistic is used. If the test statistic is statistically significant, it indicates that the instruments may not be valid; this can happen if the instruments are not truly exogenous or because they are being incorrectly excluded from the regression.[149]

Even if there are valid instruments and estimate-consistent coefficients, its covariance matrix can be inconsistent if the errors are heteroskedastic.[149] The Pagan-Hall statistic can be used to test for the presence of heteroskedasticity in the IV regression. Under the null hypothesis of homoskedasticity, the Pagan-Hall statistic is distributed as χ^2, irrespective of the presence of heteroskedasticity elsewhere in the system.[149] A significant statistic will imply the presence of heteroskedasticity. If this is the case, a heteroskedasticity-consistent standard error will have to be used while employing an equation-by-equation IV estimation. The coefficient estimates, as well as their standard errors, will then be consistent. This can be done through either a 2SLS or GMM estimation, which Wooldridge[15] refers to as a "system 2SLS estimator," and which is more efficient than the simple IV estimator[149] in the presence of heteroskedasticity.

Estimation method 2: System instrumental variable estimation (3SLS)

In order to estimate a system of Engel curves—one for each commodity group, to find where and how the crowding out is occurring—it is necessary to estimate an equation for each commodity group to be considered. Each of these equations would have tobacco spending as a conditioning commodity along with *M* and other household-specific characteristics as shown in Equation 4.4.

Since the regressors in each equation are the same, the system of equations is much like a seemingly unrelated regression (SUR) with the addition of the IV method, which is effectively a three-stage least squares (3SLS) method.[150] Under the assumption that the errors are homoskedastic, 3SLS provides a more efficient estimation compared to 2SLS+IV by exploiting cross-equation correlation of errors.[16] The literature has consistently used this method as opposed to the use of IVs in SUR. A good description of the 3SLS system estimation, which is also called the traditional 3SLS, can be found in Wooldridge[15] Chapter 8.

Estimation method 3: GMM 3SLS estimation

The traditional 3SLS estimator, according to Wooldridge,[15] is less efficient and its variance estimator is inappropriate if errors are heteroskedastic. In the cross-sectional surveys described in Chapter 2, heteroskedasticity is the norm rather than the exception. A system estimator that is consistent and more efficient than the traditional 3SLS estimator in the presence of heteroskedasticity is a GMM

estimator, and Wooldridge[15] calls it the "GMM 3SLS" estimator. It extends the traditional 3SLS estimator by allowing for heteroskedasticity and different instruments for different equations.[151] The GMM estimation allows selection of different weight matrices with which to obtain estimators that can tolerate heteroskedasticity, clustering, autocorrelation, and other classical violations of the error term *u*. The traditional 3SLS, for example, is a GMM estimator that uses a particular weighting matrix, which assumes that errors are independent and identically distributed (i.i.d.).[15] However, just like the IV/3SLS estimators, the GMM estimator, too, may have poor finite sample properties.[149]

According to Wooldridge,[15] the GMM 3SLS estimator using the heteroskedasticity consistent weighting matrix is never worse, asymptotically, than traditional 3SLS; and in some important cases it is strictly better. The previous literature on crowding out, however, seems to have ignored a test of heteroskedasticity in the 3SLS model that was used and estimated the traditional 3SLS model assuming the errors are i.i.d. This may have produced less efficient parameter estimates if heteroskedasticity was indeed present in those models. More recent literature,[130, 131] however, has made use of GMM-3SLS methods to estimate the crowding-out impact of tobacco spending.

Testing heterogeneity in preferences between tobacco users and non-users

Typically, in HES data, there are a large number of zeros or missing values against the expenditures on tobacco. This can be either because tobacco prices are currently unaffordable to some of the households due to constraints in their budget (also known as a *corner solution*), or because of abstention (that is, tobacco is not in a household's utility function or its consumption basket, no matter what the price is). If it is the latter case, tobacco users and non-users have fundamentally heterogeneous preferences. Theoretically there is no *a priori* reason why one should assume either case. However, along with the estimation of crowding out, in order to also allow for heterogeneity in preferences between tobacco-spending and non-tobacco-spending households, Equation 4.4 can be augmented with the addition of a binary variable indicating tobacco consumption status, as in some literature,[117, 126, 148] as follows:

where *d* is a binary indicator taking the value of 1 if a household spends on tobacco and 0 otherwise.

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

If the parameters associated with the binary variable *d* are jointly significant—that is, if the null hypothesis H_0:$\alpha_{2i}=\beta_{2i}=\gamma_{2i}=0$ is rejected—it can be concluded that tobacco is not in the utility function of those households for which zero expenditures on tobacco are currently reported. In other words, both tobacco spenders and non-spenders have utility functions or preferences that are different from each other, and Equation 4.5 is used to estimate the crowding out in such a case.

In Equation 4.5, the binary variable indicating the consumption of tobacco (*d*), its interaction with household expenditures (*dlnM*), and its squared term ($dlnM^2$) together serve the purpose of distinguishing between households who spend on tobacco and those who do not. But, if the null hypothesis is not rejected, it means the coefficients associated with the tobacco dummy variable, and those of the expenditure variables with which the tobacco dummy interacts, are not significant and that the preferences are not different for tobacco users and non-users. In that case, only the specification in Equation 4.4 is needed to estimate the crowding out. The literature on this uses a Wald test to test the joint significance of the three parameters after the regression.

If a researcher has an interest in testing this hypothesis, Equation 4.5, instead of Equation 4.4, should be specified in the first place. If the hypothesis H_0:$\alpha_{2i}=\beta_{2i}=\gamma_{2i}=0$ is rejected, then the specification in Equation 4.5 should be used for estimating crowding out. In that case, the coefficients associated with the variables will be different for both tobacco spenders and non-spenders. In other words, the preferences are indeed heterogeneous between tobacco spending and non-spending households and that tobacco non-spenders do not have tobacco in their utility function, no matter what its price is.

If, on the other hand, the hypothesis fails to be rejected, one may proceed with the specification in Equation 4.4, in which case both tobacco spending and non-spending households will have the same parameter estimates. In other words, there is no reason to inquire about crowding out of tobacco expenditures in the case of those households for which tobacco is not part of their utility function or consumption basket, no matter what its price is.

4.4.3 Limitations of the model and recent developments

The discussion of different methods of estimating crowding out in Section 4.4.2 assumes the availability of suitable IVs to address the endogeneity present in the model specification. However, finding a suitable IV that meets the necessary econometric requirements can often be challenging and, sometimes, one may not be able to find them at all. There is indeed literature that estimates crowding out ignoring such endogeneity,[120, 121, 129, 133] often due to the unavailability of suitable IVs. Regressions ignoring the presence of endogenous variables, however, could result in parameter estimates that lead to a wrong inference. In such cases, less sophisticated methods may be adopted. One such method is a simple comparison of budget shares between tobacco spenders and non-spenders on various items of purchase using a *t*-test, as already described in Section 4.3. It is also possible to compare absolute expenditures allotted to different items between both groups of households. Instead of a *t*-test, other descriptive or graphical comparison tools could also be performed to compare the averages.

Recent literature on crowding out,[132, 134, 135] however, has made use of other methods that may not need the explicit use of instrumental variables to examine crowding out. These include methods such as a non-parametric genetic mapping model,[134, 149] and matched difference in differences (MDID) model[132, 135, 150–152] as alternatives. One of the major criticisms in these studies against the existing literature on crowding out is that the IVs used in the present literature (such as adult sex ratio or regional smoking prevalence) are imperfect, and it is often not possible to test the exclusion restriction needed for these IVs. However, as already noted in the Section 4.4.2, when there are more instruments than the number of endogenous regressors, a test of over-identifying restrictions can be done to test the exogeneity of instruments, provided the parameters of the model are estimated using optimal GMM.[16] The GMM-3SLS model proposed in this toolkit allows this, too. However, finding more instruments than the number of endogenous regressors can often be challenging.

In the cases of the just-identified models, it would be impossible to test the exclusion restriction. Ideally, methods that do not rely on obtaining adequate instruments would be preferred in such situations, and the recent literature addresses this concern. However, it is to be noted that efficient implementation of MDID requires panel data or repeated cross sections that can be converted to *pseudo* panels. This may be a major limitation for countries with a single round of cross-sectional HES. Since MDID mimics experimental research design using observational data, it should be

possible to effectively group households in the data into treatment and control groups, both having identical sociodemographic characteristics other than the treatment status, which, in this case, would be tobacco spending status.

Since the crowding-out analysis explained above compares the budget shares on different commodities by tobacco spending and non-spending households only, it does not shed much light on intra-household allocations as a result of crowding out. This is another limitation of this analysis. For example, the analysis may show that health expenditure or education expenditure is crowded out as a result of tobacco spending. But which household member is impacted due to this crowding out is difficult to ascertain. The fact that the analysis only considers larger aggregated groups of commodities makes such intra-household considerations all the more difficult to examine.

4.5 Preparing data for analysis

While Chapter 2 provided detailed information on extracting data, cleaning it, merging variables that come from different data sets, and other necessary data management tips, it is important to provide specific details on the variables necessary for the analysis in this chapter. For any new variables that are discussed here, it is important to take them through all of the processes discussed in Chapter 2. This section discusses how the specific variables required for the crowding-out analysis can be generated using the standard variables available from HES. It also shows ways of classifying households to suit the specific analytic needs of this chapter.

The most important variables required are the expenditures spent on tobacco as well as other commodity groups mentioned earlier on, which need to be tested to determine if crowding out occurs. These are directly available from any HES. Next, the shares devoted to each of the commodity groups from the remaining budget after subtracting expenditure on tobacco should be constructed. For example, a variable for budget share on food can be created in Stata using the code *<generate bsfood = expfood/exp_less>* where *bsfood* is the budget share variable on food to be used as a dependent variable in the regression, *expfood* is expenditures on food that is extracted from HES and *exp_less* is the total expenditures on all items (*exptotal*) minus the expenditure on tobacco (*exptobac*). For all commodity groups together, a loop can be used to generate the budget shares as follows:

```
#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) ;
    } ;
```

These are the variables that would go into the regression (IV, 3SLS, or GMM 3SLS) as dependent variables. This is different from the budget share variables created in Section 4.3 for the *t*-test, since that had total expenditure as the denominator. Although expenditures on different commodities are available directly from HES, it is possible that the HES does not report these data at the level of aggregation required. For example, expenditures on food may be recorded in HES as expenditures on several other food items. If aggregate information is not available, one may have to aggregate

expenditures on smaller items to create aggregate groups like the ones listed here. Having too many disaggregated commodities may not serve much purpose after all, from a policy point of view, while analyzing the crowding-out impact of tobacco spending. However, depending on the socioeconomic circumstances in each country, the selection of commodity groups could vary.

Natural logs and squares of variables *exptotal* and *exp_less* to be used in the regression need to be generated. Specific household-level variables to use as controls and the variables, which can typically work as instruments for the endogenous variables in the 3SLS model, need to be identified. The literature offers some guidance. Some of the common household-level sociodemographic variables used in this literature include log of household size; adult ratio (ratio of number of adults to household size); average age of household; average education (total education received by all the members in years divided by the household size) of the household; max education (years of education received by the most educated member in the household); dummy variables to characterize households into different social, ethnic, occupational, religious, and income groups; and a dummy variable to indicate a household's residence, such as rural or urban areas, among others.

Choosing the right variables to serve as instruments is one of the key aspects of preparing the list of variables for the analysis. Again, the literature offers some guidance. Much of the recent literature on crowding out[117, 122, 125–127] uses total household expenditures or total value of household assets as an instrument for the group expenditure *M (exp_less)* and the ratio of adult males or adult females in the total number of adults in the household (adult sex ratio) or the ratio of adult males to adult females as the instrument for tobacco expenditure.

The adult sex ratio is thought to be a sensible instrument for tobacco spending as tobacco consumption is usually much more prevalent among males than females in most of these countries. Therefore, an increase in male ratio (ratio of adult males to adult females) is expected to be positively related to tobacco spending, and it is not something that may directly impact the budget share on other commodity groups for which the crowding out impact is estimated. But in countries where the smoking rate is not significantly different across genders, sex ratio may not be an appropriate instrument to use. Alternative approaches have been taken in such instances. Some studies[123, 138] have used a composite smoking prevalence and intensity measure as an instrument for tobacco spending. Any exogenous variables that appear on the right-hand side (RHS) of the other equations in the model can potentially serve as an instrument for the endogenous RHS variable in the equation to be estimated. No matter which variable is used as an instrument, it is important to check that the selected instruments are correlated with the endogenous RHS variable and that they do not have a direct effect on the dependent variable.

4.6 Estimating crowding out with Stata

This section demonstrates the different estimation methods (traditional 3SLS, GMM 3SLS and an equation-by-equation IV) discussed in Section 4.4 to estimate crowding-out effects. First, it discusses the general setup of variables that can be used under all the methods. After a discussion of the implementation of all three estimation methods, the testing of various requirements of the model including validity of instruments and heteroskedasticity, among others, are discussed. The results of these tests will guide the decision on the type of estimation method to be used.

As detailed earlier, depending on the properties of data there are different modeling strategies. Below are a few variables that are necessary for estimating Equation 4.4:

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

In addition, to simplify the regression model for estimating the traditional 3SLS or GMM 3SLS or IV estimations, it is useful to create certain global macros indicating the list of dependent variables, endogenous variables, exogenous variables, and instruments in the model. For example, for estimating the impact of crowding out among eight commodity groups—food, health, education, housing, clothing, entertainment, transportation, and durable goods—leaving out the commodity group "other" as commonly done in the literature, the following macros are defined:

```
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
```

The macro *ylist* includes the dependent variables that go into the regression, *x1list* includes the RHS endogenous variables as explained in Equation 4.4 (these are variables that are suspected to be endogenous), *x2list* includes the exogenous variables (household size, mean education, max education, three dummy variables to represent the SES status of households), and *zlist* includes the IVs to correct for endogeneity in the model (adult sex ratio, log of total expenditures, and log of total expenditure squared, in this case). In the model, however, every exogenous variable can be an instrument of its own. The number of variables in *zlist* must be at least as large as those in the *x1list* for the model to be identified. The variables used in the global macros here are only for the purpose of demonstration. In the actual analysis there can be a smaller or greater number of variables in any of the lists above. For example, the *x2list* may contain several other household-specific characteristics than those listed here.

4.6.1 Estimation of 3SLS

Once these global macros are created, estimation of the 3SLS model in Stata can simply be done by using the command <*reg3*>. Stata help with reg3: <*help reg3*> provides detailed syntax and useful examples for using this command. But, for this purpose, once the global macros are defined as above, only the following command needs to be used to obtain the 3SLS estimates:

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

where the *exog* and *endog* options specify the list of exogenous and endogenous regressors on the RHS of each of the equations. Without the use of global macros, this command could also be written as:

```
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
```

It is essential that the code must be either in one single line in the do-file or it should be broken with appropriate delimiters acceptable to Stata to mark the end of the command. The use of macros makes the code much neater. Moreover, there is no reason to use a separate regression command for each equation, even if the instruments vary for some of the equations. All instruments can be put together into the *exog* list while using the *reg3* command.

As previously noted, 3SLS is a GMM estimator that uses a particular weighting matrix that assumes i.i.d. errors. So, the above 3SLS results from the <*reg3*> command can be reproduced with a GMM estimation with an appropriate weighting matrix. This is done in the code below:

```
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 option <*winitial()*> specifies the weight matrix to use to obtain the first-step parameter estimates. The <*independent*> sub-option tells the GMM to assume that the residuals are independent across momentary conditions. The option <*wmatrix()*> controls how the weight matrix is computed on the basis of the first-step estimates before the second step of estimation. By specifying <*wmatrix(unadjusted)*>, a weight matrix is requested that assumes conditional homoskedasticity, but does not impose the cross-equation independence like the initial weight matrix.[151] Please note that the <*gmm*> code above could take much longer—sometimes several hours, depending on the physical capacity of the computer—than <*reg3*> would to converge on a solution. This is because GMM, unlike 3SLS, is a very general and nonlinear estimator, and it searches numerically for a solution.

4.6.2 Estimation of GMM 3SLS

If the errors are heteroskedastic that means traditional 3SLS estimates are less efficient and their standard errors inconsistent. A heteroskedasticity consistent weighting matrix should be used to obtain consistent parameter estimates in this case. This is possible with GMM using the option <*wmatrix(robust)*> as implemented in the code below:

```
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
```

The option wmatrix(robust) requests a weight matrix appropriate for errors that are independent but not necessarily identically distributed. It is also possible to request a weight matrix that accounts for arbitrary correlation among observations within clusters, as is usually observed in survey data. For that purpose, the option can be modified to *<wmatrix(cluster clustvar)>*, where *clustvar* is the name of the variable that identifies clusters in the data.

Instead of the robust standard errors in *<gmm>*, bootstrapped standard errors may also be obtained by using *<reg3>* with a bootstrap prefix. For example, *<bootstrap, reps(1000) seed(1010):reg3 ($ylist = $x1list $x2list), exog($zlist) endog($x1list) 3sls>*. This is better than estimating a 3SLS *<reg3>* ignoring possible heteroskedasticity. However, *<reg3>* with 1,000 bootstrap replications may take as much time as *<gmm>* to achieve convergence. Using *<gmm>*, on the other hand, has the added advantage of specifying a weighting matrix that accounts for heteroskedasticity from clustering and autocorrelation.

The models as implemented above are just-identified models, as the number of instruments is equal to the number of endogenous RHS variables. If there is an over-identified model instead, the implementation of the Stata code would be the same except that the names of those additional instruments would be added to the list of IVs in the global macro *zlist*.

4.6.3 Equation-by-equation IV

As noted in Section 4.4, an alternative to doing a system estimation, as in traditional 3SLS, is to do the estimate for each equation, one by one, using 2SLS. This can be implemented with the help of Stata's *<ivregress>* command as follows:

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

Stata also has an excellent user-written command *<ivreg2>*[157] that can be used instead of <ivregress>, and it offers additional functionality compared to *<ivregress>*. It can be installed using the command *<ssc install ivreg2>*. The implementation of *<ivreg2>* is quite similar to that of *<ivregress>*. For example, *<ivregress 2sls bsfood $x2list ($x1list = $zlist)>* and *<ivreg2 bsfood $x2list ($x1list = $zlist)>* would give identical estimates.

The equation-by-equation IV, which Wooldridge[15] refers to as a "system 2SLS estimator" can be implemented by omitting the option *<twostep>* and *<wmatrix()>* from the traditional 3SLS implementation in a *<gmm>* command as below. This should give output similar to the ones obtained from *<ivregress>* or *<ivreg2>*, but with robust standard errors.

```
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)
```

To also see the standard errors identical to the ones in the *<ivregress>* command, add the option *<vce(unadjusted) onestep>* after *<winitial(unadjusted, independent)>*. If a test for heteroskedasticity after equation-by-equation IV indicates errors are not homoskedastic, then either the system 2SLS estimator with *<gmm>* can be used, as given above, which returns robust standard errors, or the <ivregress> command can be modified with the optional command *<vce(robust)>*. For example, it can be implemented for the bsfood equation as *<ivregress 2sls bsfood $x2list ($x1list = $zlist), vce(robust)>*. The *<ivregress>* command also allows specifying a weighting matrix with the use of the GMM estimator as *<ivregress gmm bsfood $x2list ($x1list = $zlist), wmatrix(robust)>* or with other specifications of the weighting matrix, such as *<wmatrix(cluster clustvar)>*. The coefficient estimates as well as their standard errors will then be consistent, as noted in Section 4.4.

4.6.4 Performing different tests to decide on the estimation method

Before deciding which particular estimation method should be used, it is important to perform several tests. These include a test for endogeneity of variables, a test of validity of used instruments, and a test for homoskedasticity of errors, among others. These tests are more easily implemented after an equation-by-equation IV estimation.

1) Testing endogeneity of regressors: As noted in Section 4.4, it is not necessary to use an IV estimator unless the endogenous variables are indeed endogenous. Endogeneity can be tested with the help of the DWH test of exogeneity[148] in case of i.i.d errors, or Wooldridge's score test, or an auxiliary regression-based test in the case of non-i.i.d. errors,[149] as discussed earlier. After the *<ivregress>* command, the command *<estat endogenous>* can be used to do this. It will report either the DWH-statistic or any of the other heteroskedasticity-consistent statistics discussed above, depending on the optional weighting matrix used with the *<ivregress>* command. In either case, the null hypothesis is that the variables are exogenous, and a significant test statistic would indicate that the variable should be treated as endogenous.

Similarly, if *<ivreg2>* is used, the command *<ivendog>* can be used after *<ivreg2>* and will report the DWH-statistic. Alternatively, the option *<endog(varname)>* can be used along with the *<ivreg2>* command to test if an instrument is endogenous. For example, *<ivreg2 bsfood $x2list ($x1list =*

$zlist), gmm2s robust endogtest($x1list)> tests for endogeneity of all three endogenous variables, along with displaying the regression results. This option is particularly useful to test for endogeneity when heteroskedasticity is present.

2) Testing the validity of instruments: As previously noted, IV estimators are consistent only under the very strong assumption that a valid instrument (*z*) exists that satisfies both inclusion and exclusion restrictions. Testing the inclusion restriction is straightforward. It checks if the instruments are weak or strong. With the *<ivreg2>* command, one simply needs to add the option *<first>*; for example, *<ivreg2 bsfood $x2list ($x1list = $zlist), first>*. This would report the first-stage regression results, one for each endogenous regressor. In this case, since there are three endogenous RHS variables (*pq, lnM, lnM2*), it would report three first-stage regression results with each of these endogenous variables as the dependent variable and all the remaining exogenous regressors and the IVs as RHS variables. The R^2 and *F*-statistic from these first-stage regressions indicate how strong or weak the instruments are.

A common rule of thumb suggests an *F*-statistic of less than 10, in the case of a single endogenous regressor, to be indicative of a weak instrument.[16, 154] If there is a single instrument and a single endogenous regressor, this translates to a *t*-value of 3.2 or higher and the corresponding *p*-value of 0.0016 or lower for the instrument. The results of this *F*-test should be reported when reporting IV estimates. This rule of thumb, however, is ad hoc and may not be sufficiently conservative if the model is over-identified. For equations with more than one endogenous regressor, a statistic called Shea's partial R^2 can be used instead of the *F*-critical value.[16] However, there is no consensus on how low of a value for R^2 indicates a problem.[16] Using the option *<first>* after *<ivreg2>*, as well as the command *<estat firststage>* after executing the <ivregress> reports Shea's partial R^2. See Cameron and Trivedi[16] Chapter 6 for a detailed exposition of these statistics. Alternatively, refer to the Stata reference manual[151] on *ivregress* post-estimation technical notes on page 1212–1213.

Testing the exclusion restriction or testing the exogeneity of the instruments is, in general, not possible, especially in the previous case. In the over-identified case, however, a test of over-identifying restrictions can be performed with the command *<estat overid>* after *<ivregress>* or with command *<overid>* after *<ivreg2>*. It would report the results of a Sargan test in the case of homoskedasticity. If *<ivregress>* had used the option *<gmm>* along with a heteroskedasticity-consistent weighting matrix, then the *<estat overid>* would report a Hansen's *J*-statistic or Hansen-Sargan statistic, which account for heteroskedastic disturbances. A statically significant test statistic indicates that the instruments may not be valid. This can happen if the instruments are not truly exogenous, or because they are being incorrectly excluded from the regression,[149] as noted earlier.

3) Testing for heteroskedasticity: As noted in Section 4.4, if the errors are heteroskedastic, IV regression produces inconsistent standard errors and the traditional 3SLS estimates are less efficient and standard errors inconsistent. The Pagan-Hall statistic can be used to test the presence of heteroskedasticity in the IV regression. This can be implemented with the command *<ivhettest>*. For example, after *<ivreg2 bsfood $x2list ($x1list = $zlist)>*, apply the command *<ivhettest>*, and it would report the Pagan-Hall statistic with the null hypothesis of homoskedastic disturbances. A significant statistic will imply a rejection of the null hypothesis, indicative of the presence of heteroskedasticity. Unfortunately, the *<ivhettest>* does not work after the *<ivregress>* as of now.

There is also a user-written program *<lmhreg3>*[159] that can be installed with the command *<ssc install lmhreg3>*, which performs the tests of both single equations and overall system heteroskedasticity after the *<reg3>* command. So, if *<reg3>* were used to do a 3SLS estimation, one can apply the command *<lmhreg3>* immediately afterwards to check whether each of the individual equations, as well as the system as a whole, satisfies the homoskedasticity assumption. The null hypothesis is that the errors are homoskedastic, and, as usual, a significant test statistic (Pagan-Hall or other Lagrange multiplier tests used in lmhreg3) is indicative of heteroskedasticity.

4) Testing heterogeneity in preferences between tobacco users and non-users: To examine whether the preferences are heterogeneous between tobacco spending and non-spending households, Equation 4.5 can be estimated instead of Equation 4.4 to test for the joint significance of parameters associated with the binary indicator for tobacco use and the interactions with it. It translates to testing the null hypothesis H_0:$\alpha_{2i}=\beta_{2i}=\gamma_{2i}=0$ in Equation 4.5. For this, first estimate the model in Equation 4.5 using *<ivregress>* as follows:

```
#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);
};
```

The *<test>* command after each successive equation performs a Wald test to test a composite linear hypothesis that all three coefficients associated with the dummy variable *tob* are jointly zero. A rejection (that is, significant test statistic) suggests that Equation 4.5 may be a more appropriate specification, whereas no rejection would imply Equation 4.4 may be the right specification. If the test concludes that Equation 4.5 is the specification of choice, all tests from (1) to (3) above need 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.

Summary of tests and decision on the estimation method: To review, before deciding on which method of estimation to use—either the traditional 3SLS *<reg3>* or GMM 3SLS *<gmm>* or equation-by-equation IV (either with *ivregress* or *ivreg2*)—it is recommended to first estimate equation-by-equation IV. This would allow one to determine whether there is endogeneity in the model, and if the used instruments are valid. Next, the heteroskedasticity test must be performed. Should the heteroskedasticity test indicate that the errors are i.i.d., then one could opt for a *<reg3>* to do the traditional 3SLS estimation. If not, a GMM 3SLS estimation method using the *<gmm>* command in Stata must be used to produce efficient parameter estimates. According to Wooldridge,[15] the GMM 3SLS estimator using the heteroskedasticity-consistent weighting matrix is never worse, asymptotically, than traditional 3SLS, and in some important cases it is strictly better. So, it would be safer to use a GMM 3SLS estimation method to estimate the crowding out in any case.

Finally, testing the joint significance of parameters associated with the indicator variable for tobacco spending along with their interaction variables will indicate whether it is appropriate to use a functional form that treats tobacco spenders and non-spenders as entirely different. If it

concludes that they should be treated differently, then Equation 4.5 must be specified and all suggested tests from (1) to (3) above need to be repeated on the new specification.

4.6.5 Estimation of crowding out by subgroups

Since tobacco use is more concentrated in low-income communities, or low-income communities are known to spend a disproportionately larger share of their budget on purchasing tobacco products, it is possible that the impact of crowding out may be larger among these low-income communities. Similarly, households can also be classified in terms of the severity of their spending on tobacco into moderate, medium, and high spenders. It is possible that the crowding out could be much higher among high spenders compared to moderate spenders. For these and other reasons, researchers may want to examine the crowding-out impact by different subgroups defined either by income or by other characteristics. The literature has used different subgroups for examining the impact including income groups,[117, 127, 134] severity of tobacco spending,[124] and different types of tobacco.[127]

Apart from the details discussed so far, estimating crowding-out impact by subgroups requires only two additional steps:

(1) defining a categorical variable indicating the subgroup; and

(2) adding the subgroup option to the relevant Stata command.

Examples of these are shown below.

Step 1: Defining categorical variables to indicate subgroup

This step was already discussed in Section 3.5.1. To reiterate, first it is necessary to create a per capita expenditure variable (*pcexp*) by different households using the command *<gen exppc=exptotal/hsize>*. The per capita household expenditure groups/quantiles (as proxy for income) can be generated using the command:

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

where *option nq (.)* specifies the number of quantiles.

Similarly, households can also be classified based on the distribution of budget share spent on tobacco into low or high spenders, and so on.

Step 2: Adding subgroup options to relevant Stata commands

Once the categorical variable is generated, say incgrp, the estimation can be done by either adding a *<by(incgrp)>* or *<over(incgrp)>* option or *<bysort incgrp:>* prefix to the Stata commands, depending on the particular command. For example, the *<ivregress>* can be estimated with the prefix as follows:

```
#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)
}
```

For the GMM 3SLS, too, the prefix <*bysort incgrp:*> can be added before the command <*gmm*>.

Section 7.4 in the Code Appendix provides an example do-file that details the code used in this chapter. Users will be able to copy and paste that into Stata's do-file editor and will be able to estimate the results with appropriate accompanying data/variables described therein.

4.7 Case study from Turkey

Turkish households, despite living in an upper-middle-income country, spent more than eight percent of their household budget on tobacco purchases in 2011. While the rich in Turkey spent about 6.2 percent of their household budget on tobacco, the poor spent as high as 10.7 percent.[126] Given that a large portion of household budgets is being diverted to tobacco spending, it is possible that expenditures on other household necessities are traded off. In this context, San & Chaloupka[126] examine the crowding out of tobacco spending on a variety of commodity groups in Turkey. The study estimates the QAIDS model with a variant of Equation 4.5 to estimate the effects of crowding out. The econometric model used is the 3SLS method discussed in Section 4.4. The study uses total expenditure as an instrument for the expenditure net of tobacco and a female ratio—ratio of adult females to total adults in the households—as an instrument for tobacco spending. Table 4.2 shows a snapshot from the results they found for 2011 and lists the results of only a subset of the commodity groups the authors analyzed. The first column under the commodity group shows the parameter estimates and the second column presents the standard errors.

Table 4.2 Crowding-out effect of tobacco spending in Turkey, 2011

	Food		Housing		Clothing		Transportation		Education	
	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
lnM^2	−0.0121*	0.000	−0.0135*	-0.001	0.0005*	0.000	0.0092*	-0.001	0.0025*	0.000
dlnM	−0.2004*	-0.055	0.2316*	-0.102	0.0955	-0.027	−0.6456*	-0.084	0.0228	-0.026
$dlnM^2$	0.0122*	0.003	−0.0105*	0.006	−0.0056	-0.002	0.0410*	0.005	−0.0012	-0.002

Notes: Results from the specification in Equation 4.5. The values of dependent variables run from 0 to 1. *These results are significant at the 5% level. Source: San & Chaloupka (2016)[65]

The binary variable (*d*), indicating tobacco spending, is significant in the case of all commodities except education. Its negative sign indicates that spending on tobacco has a negative impact on spending on the corresponding commodity group. The magnitude of the coefficient for the dummy variable, *d*, however, does not provide a straightforward interpretation. It should be interpreted only in conjunction with its interaction with other variables ($dlnM$ & $dlnM^2$) and their joint significance as shown in Equation 4.5. A Wald test of the joint significance of all three coefficients is indeed rejected in this study, implying that households that spend and do not spend on tobacco have fundamentally different utility functions.

The variable *p.q* is the total pre-allocated expenditures on tobacco, and its coefficient provides an indication of the extent of crowding out. For example, for every lira increase in the pre-allocated amount on tobacco, there is a reduction in the budget share allotted to housing in the remaining budget of the household by 0.0022 percentage points or 0.0022 *x M* lira, where *M* is the remaining budget after spending on tobacco.

Assume the monthly expenditures after spending on tobacco are about 1200 lira (since 106 lira spent on tobacco constitutes about 8.17 percent of the budget). Then, using the parameter estimates presented by the authors, one can compute that a 10-lira increase in the pre-allocated amount on tobacco leads to a 26.4-lira decrease in housing expenses, while also redistributing expenditures on all the remaining commodities, increasing some and decreasing others. For example, a 10-lira increase in the pre-allocated amount on tobacco would decrease expenditures on food, utilities, durables, clothing, health, and education by about 2.4, 1.2, 9.6, 3.6, 2.4, and 3.6 lira, respectively, and increase expenditures on transport, entertainment, alcohol, and other commodities by 25.2, 20.4, 2.4, and 1.2 lira, respectively.

What is important to see is that an increase in tobacco spending clearly redistributes the expenditures, benefiting some items but hurting several others. In this particular case, the items with reduced consumption are mostly necessities, and that warrants public policy intervention to regulate tobacco use. The remaining variables in Table 4.2, including the ones used in the regression but not shown on the table, serve the purpose of control variables in the regression.

5 Quantifying the impoverishing effect of tobacco use

5.1 Introduction

National poverty estimates are an important political variable in most countries. The estimate of the percentage of poor people in a population determines the course of development policy debates in many countries. Poverty reduction is a stated objective in numerous countries around the world, and the eradication of poverty in all its forms is the very first goal of the Sustainable Development Goals of the United Nations.[6] However, tobacco use is a major factor among those that hinder a nation's ability to achieve poverty-reduction goals.

Tobacco use and poverty are components of a vicious cycle.[4] As more money is spent on tobacco, households are deprived of certain necessities including food and nutrition, as explained in Chapter 4, thus creating a huge opportunity cost and further exacerbating poverty. As the money spent on tobacco is highly unproductive and increases tobacco-related diseases, the resulting increased health care costs and loss of income due to premature deaths and morbidity can also add to the burden of poverty. Worldwide, around 80 percent of smokers live in LMICs, and in most of those countries tobacco use is concentrated in low-income populations.4 The wealth-related and education-related inequalities in tobacco use among men and women are higher among LMICs compared to upper middle-income countries.[160]

Chapter 4 explains how spending on tobacco displaces or crowds out expenditures on different commodity groups, offering a certain dimension of the opportunity cost of spending on tobacco. This chapter shows how to quantify the direct impact of tobacco spending on poverty measured by poverty head counts, discusses how tobacco spending contributes to impoverishment, and presents methods for quantifying these concepts. It also demonstrates how this can be done with the help of HES using Stata.

5.2 Poverty head counts and their relevance

Definitions of poverty vary from country to country depending on the specific social and economic circumstances prevailing in each country. However, "almost all national poverty lines (NPL) are anchored to the cost of a food basket—what the poor in that country would customarily eat—that provides adequate nutrition for good health and normal activity, plus an allowance for non-food spending."[161] As the food baskets or the tastes and preferences change, nations typically redefine the poverty line accordingly. In essence, the poverty line takes a certain resource deprivation into account and defines an amount that is necessary to sustain a locally perceived notion of what it takes not to be poor.

Usually this is translated into a local currency unit. For example, Statistics South Africa[162] defines a food poverty line—the amount of money that an individual will need to afford the minimum required daily energy intake, also known as the "extreme" poverty line—as 547 rand per person per month. It also defines other poverty lines that take into account certain minimum expenditures on non-food items. Similarly, the United States Census Bureau (USCB) uses a set of dollar income thresholds that vary by family size and composition to determine who is in poverty.[163] The USCB's 2022 definition shows that a single person under the age of 65 earning less than $13,590 per annum is considered to be living below the poverty line.[164]

Although there are several methods to measure poverty, the head-count ratio (HCR), which is an absolute measure of poverty, is one of the most commonly used poverty indicators, especially in LMICs.[165] The HCR, a counting measure, is defined as the fraction of the population living below the NPL, and it allows for a highly intuitive and simple interpretation. This fraction is often computed using HES as it allows one to compute the average expenditures by each household, or per capita consumption expenditures by individuals, and to compare that against the defined poverty line. The HCR, however, does not take into account the degree of poverty. In other words, the rate of poverty measured by HCR would remain the same even if the poor below that poverty line became even poorer.

The NPLs across countries are often not comparable, as the notion of being poor can vary significantly across countries and cultures. Although not comparable across countries, poverty lines are quite useful in the context of a country's domestic development policies. They can be used as yardsticks for facilitating certain social welfare programs, for example, to develop interventions to specifically target the poor.

5.3 How does tobacco consumption contribute to impoverishment?

The objective of this chapter is to quantify the impact of tobacco consumption on the estimate of HCR. To understand this, it helps to distinguish two types of poverty as explained by the British sociologist B. Seebohm Rowntree[166] and reproduced in the WHO/NCI monograph on *The Economics of Tobacco and Tobacco Control*.[4] The first one is primary poverty, which refers to a situation in which income or other resources are insufficient to afford the basic necessities like food, water, or clothing. Essentially, households that fall below the NPL in a country can be classified as those suffering from primary poverty.

The second one is secondary poverty, which refers to a situation in which households have sufficient resources to meet their basic needs, but those resources are not used efficiently. As a result, despite possessing a higher amount of resources, these households may be living in conditions similar or inferior to those in primary poverty. For example, a significant amount of income is spent on unproductive and harmful consumption of goods such as tobacco or alcohol by a household that is otherwise above the poverty line. Due to a crowding-out effect, the household is consequently unable to meet their basic needs, just as those households in primary poverty.

But the estimates of HCR would only capture those who are in primary poverty, although many households in the country may actually be in secondary poverty and hence not meeting their basic needs due to wasteful consumption on tobacco. It would be ideal to include such households in the calculation of HCR so that policies and programs can be more effectively targeted. Alternatively, policies will have to be adopted for households to be lifted out of secondary poverty by helping

them to reduce or stop wasteful and harmful consumption so that their total available resources can meet their basic needs.

As household budgets are limited, consumption of anything—including tobacco—necessarily involves trade-offs. The literature discussed in Chapter 4 shows that the trade-off happens in the form of crowding out of certain necessities. There are three major channels through which increased consumption of tobacco can effectively diminish a household's income and push it into a state of poverty, as explained below:

1) **Channel 1: Forgone income from tobacco purchase**
   The direct disposable income to meet basic needs is reduced by the same amount that was spent on the purchase of tobacco.

2) **Channel 2: Forgone income from treating tobacco-related morbidity**
   As tobacco consumption and exposure to SHS inevitably lead to the onset of several diseases and the associated morbidity, the costs of treatment of these medical conditions further reduce the disposable income available to meet basic needs. While the increased medical expenditure directly impacts disposable income, it can also impact productivity and income-earning potential.

3) **Channel 3: Forgone income from treating tobacco-related mortality**
   Tobacco consumption and SHS-related diseases often result in premature death. This results in the loss of future earnings, impacting the welfare of other members of the household.

All these channels have the ultimate effect of impoverishing a poor household even further. As the poor usually allocate a larger share of their budget to tobacco compared to the rich,[4] the impoverishing impact of tobacco spending is relatively larger on the poor than on the rich. Tobacco control policies that reduce consumption of tobacco have the opposite effect, especially if the tobacco users are more price-sensitive.[167] As a result of decreased spending on tobacco and, consequently, reduced health care spending, these households will have more disposable income to spend on essential needs (such as food, clothing, and education).

Although the literature examining the socioeconomic inequalities in smoking and tobacco use is quite substantial,[4] the literature quantifying the impoverishing effect of tobacco spending in terms of its impact on quantifiable measures of poverty is limited. One of the first studies was done in Vietnam,[168] and it quantifies the impoverishing effect of out-of-pocket payments for health care. The first study to estimate the impoverishing effect of direct household spending on smoking and excess medical spending attributable to smoking was done in China.[169] It finds that these two effects combined are responsible for impoverishing 30.5 million urban residents and 23.7 million rural residents in China. A study from India[170] also finds that the combined effect of these two factors result in the impoverishment of 15 million people in India.

A study from the UK[171] subtracts only tobacco expenditures from household income to estimate its impact on poverty and finds that more than 432,000 children may be viewed as having been drawn into poverty by parental smoking. Yet another study from the UK[172] shows that, when expenditure on tobacco is taken into account, around 500,000 extra households, comprising more than 850,000 adults and almost 400,000 children, are classified as being in poverty in the UK compared to the official *Households Below Average Income* figures. A more recent study[173] from the UK finds that 230,000 households (400,000 adults and 180,000 children) are living in poverty when weekly

tobacco expenditure is taken into account. Another recent study from Vietnam[174] estimates that tobacco-related expenditures impoverished an additional 0.31 million (corresponding to 3.77 percent of GSO's official estimate) people in Vietnam in 2018.

These studies conclude that so many people who otherwise are above the NPL in these countries (that is, in secondary poverty) are effectively in poverty because their disposable income, after spending on tobacco and associated health expenditures, is lower than that of people who are officially classified as being under the NPL. In other words, these people are inadvertently labeled as being above the poverty line, while in reality they are not.

None of the studies so far have estimated the impoverishing impact of the income forgone from tobacco-related premature deaths (Channel 3) and income forgone from SHS-related morbidity (part of Channel 2). Since poverty or HCR is measured for a given point in time, subtracting the forgone income due to premature mortality or that of future loss of income from present household incomes is untenable. However, the direct medical costs attributable to SHS (part of Channel 2) are clearly a candidate for forgone income to be subtracted from the present disposable income while assessing the impoverishing impact of tobacco use. But this has not been incorporated in any of the studies so far either.

5.4 Conceptual framework to estimate the impact on HCR

To estimate the change in HCR, subtraction of two different types of forgone incomes from household incomes to estimate the change in HCR is necessary: (i) income forgone on account of the purchase of tobacco, and (ii) income forgone due to tobacco use and SHS-attributable direct health care costs. Before being able to subtract these different components of forgone income from total household income, it is important to identify the NPL based on which way the HCR is computed. The NPL is either a single number for the whole country, or different numbers for rural and urban areas and for each subregion or state within the country. It is usually available from the statistical agencies or other government sources in each country. The income variable against which the HCR is usually computed is taken from nationally representative HES. Since the reported consumption or expenditure estimates are far more reliable than reported income in representing the true income,[8] the expenditures estimated from HES are used as a proxy for income to estimate the proportion of people below the poverty line.

What is also important is the fact that most HES are household surveys that treat households as a single unit and the consumption expenditures are reported for the household as a whole. Poverty, however, is experienced by individuals, not by households per se, and therefore it is poverty among persons that must be measured. Although one may not know anything about the distribution within households, many studies assume a uniform distribution within households when constructing the estimated distribution of individual consumption.[175] Additionally, while it is more acceptable to assume a uniform distribution of consumption within households when constructing the estimated distribution of individual consumption, it may not be as acceptable to assume a uniform distribution within a household in the case of known adult goods like tobacco. One solution proposed by Deaton[8] is "a system of weights, whereby children count as some fraction of an adult, with the fraction dependent on age, so that effective household size is the sum of these fractions, and is measured not in numbers of persons, but in numbers of *adult equivalents*." There have been such studies that emphasize the importance of estimating poverty using per adult equivalent (PAE) expenditures that controls for economies of scale and the reduced needs of children.[176] A recent

review of this literature,[176] however, concludes that "the use of equivalence scales, while not unimportant, is not compelling in practice" since studies are not in agreement as to whether the poverty estimates are sensitive to the use of equivalence scales that adjust for household composition and economies of scale. Moreover, since the household is a single unit for all practical purposes and the money spent on tobacco necessarily reduces the disposable income available to the whole household including children, the impoverishing impact could very well be equally borne by children as well as adults. Therefore, the discussion of HCR and the impoverishing impact of tobacco use in this chapter has not considered the use of such *adult equivalence*.

While estimating the HCR, it is important to use survey weights that can generate population-level statistics for individuals and not for households. This estimate can be obtained by multiplying household size by the survey weights given to generate household-level statistics in HES. Total HCR and poverty are calculated before subtracting the tobacco-related forgone incomes. Let *z* be the variable or scalar that represents the NPL. The HCR simply counts the number of people whose incomes are below the poverty line *z* and divides that number by the total number of people in the country or region. Let *x* be the welfare measure (that is, per capita consumption expenditures, which is total household consumption expenditures divided by household size), then the HCR denoted as (P_0) is calculated as follows:[165]

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

where *I (.)* is an indicator function that takes the value of 1 if its argument is true and 0 otherwise. While it is computed using HES, appropriate survey weights are to be used. $P_0 \times N$ gives the total number of poor in the country.

5.4.1 Excess poverty attributed to forgone income from tobacco purchase

Tobacco expenditures by household are usually available from the same household surveys from which the HCR (P_0) is computed. Let *t* be the per capita consumption expenditures on purchasing tobacco in the same time period for which the welfare measure (*x*) is captured. In other words, this is the forgone income from tobacco purchase. Then, the HCR, after deducting tobacco spending or the forgone income from tobacco purchase, denoted by (P_1), can be calculated as:

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

where, again, *I (.)* is an indicator function that takes the value of 1 if its argument is true and 0 otherwise. $x_i - t_i$ is the per capita disposable income after subtracting the forgone income from tobacco purchases. $(P_1 - P_0) \times N$ is the excess number of people who are impoverished because of spending on tobacco. In other words, this is the excess poverty attributed to direct tobacco purchase expenditures.

Excess poverty attributed to forgone income from tobacco purchase and treating tobacco-related morbidity

Tobacco-related morbidity can occur among those who consume tobacco as well as those who are exposed to SHS. Let *t* and *h* be the per capita tobacco expenditure and total tobacco use– and SHS-attributable per capita health expenditures, respectively, in the same time period for which the

welfare (x) is measured. Then, the HCR, after deducting this forgone income from tobacco purchases and treating tobacco-attributable health expenditures, denoted by (P_2), can be calculated as:

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

where *I (.)* is an indicator function that takes the value of 1 if its argument is true and 0 otherwise.

$x_i - t_i - h_i$ is the per capita disposable income after subtracting both expenditures on tobacco and the attributable health care expenditures due to tobacco consumption and SHS. $(P_2 - P_1) \times N$ is the additional number of people who are impoverished due to tobacco use and SHS attributable health care spending. $(P_2 - P_0) \times N$ will be the total excess number of people impoverished after accounting for forgone income from both tobacco spending and attributable health care expenditures.

While HES provides information on health care expenditures, they do not distinguish the amount of health care that can be attributed to tobacco use or SHS exposure. This must be estimated separately, and the subtraction should be only for the expenditures on health care that can be attributed to either tobacco use or SHS exposure. The attributable costs can be estimated using either a disease-specific approach or an inclusive or all-cause approach.177 Since the HES often provide aggregate health care expenditures, the inclusive approach is more appropriate to use. It decomposes the share of total medical costs attributable to tobacco use or SHS exposure by multiplying total health care costs by the tobacco use attributable fraction, or SHS attributable fraction, commonly known as the smoking-attributable fraction (SAF). SAF is the portion of total medical care utilization that is attributable to smoking by current and former smokers.[177] Similarly, SAF for SHS would be the fraction of health care expenditures that can be attributed to SHS.

Therefore, the attributable health care expenditures due to tobacco consumption and SHS, (that is, *h* in the Equation 5.3) can be computed as follows:

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

where *exphealth* and *hsize* are household expenditures on health and household size, respectively. Both these variables are directly obtained from HES. SAF_{tob} and SAF_{SHS} are fractions of health care expenditures attributable to tobacco use and SHS, respectively. The SAF must be externally estimated using data from several different sources. It also may be taken from available studies elsewhere in the country.

The SAF can be estimated either by using the epidemiological approach or an econometric approach.[178] The econometric approach requires "extensive nationally representative data that contain detailed information on each respondent's smoking history, sociodemographic characteristics, employment status, other health risk behaviours, health status, medical conditions, annual health care expenditures by type of health care services (such as inpatient hospitalizations and outpatient visits), and annual work-loss or disability days."[178] On the other hand, the epidemiological approach is less data intensive and "can be done with aggregate data and therefore can be used when detailed health survey data are not available."[178]

For these reasons, the epidemiological approach to estimating SAF is preferred in many LMICs. WHO provides a toolkit[179] for estimating the economic costs of smoking, which includes detailed explanations and methods for both the epidemiological and econometric methods of estimating SAF. Therefore, this toolkit does not discuss this issue.

Unlike the data required for estimating SAF for tobacco use, the data required to estimate the SAF for SHS may be more difficult to obtain. Perhaps this is the reason why previous studies quantifying the impoverishing effect of tobacco use on poverty ignored this particular source of forgone income from calculation.

5.5 Preparing data for estimating the impoverishing effect

As detailed in Chapter 2, the data must first be cleaned and prepared for analysis. Since the objective is to quantify the impoverishing effect of tobacco, the most important variables are expenditures spent on tobacco (*exptobac*) as well as expenditures on all commodities together as a proxy for household income (*exptotal*). In addition, expenditures on health care (exphealth) are required in order to compute health care costs attributable to tobacco and SHS depending on the availability of SAF. The other variables needed from HES for the analysis include household size, survey weights, and variables to declare survey design. A variable or scalar to represent the NPL is necessary. If the NPL is a variable that varies across regions, or by rural or urban areas, or states within the country, then the variable has to be merged with the household survey data before the analysis can be done. To do so, a common identifying variable has to be present in both the household expenditure data as well as in the poverty line data.

For example, if the NPL in a country varies by state and residence (rural or urban), then the poverty data should have three variables, a variable indicating the NPL (*npl*), usually in local currency units, a variable with either the names or numeric code for different states (*stateid*), and a residence variable indicating whether the NPL belongs to rural or urban areas (*residence*). Similarly, the HES data must also have *stateid* and *residence* variables. Then, both data sets can be merged with the <merge> command in Stata.

To do this, first prepare a Stata data set with the *npl* and other identifying variables as necessary and save it with the name "poverty.dta." Then, open the HES master data with the expenditure information for each household, and make sure it has the same *stateid* and *residence* variables as in the poverty.dta file. Then use the command *<merge m:1 stateid residence using poverty.dta>*. A many-to-1 (*m:1*) merge is used here since the master data set has several households with the same *stateid* and *residence*. After the merge command, use the command *<tabulate _merge>* to check if the merging has taken place accurately.

While the HES considers households as a single unit and reports all expenditures at the household level, the NPL is usually for an individual, so it is important to convert the expenditure data to be comparable to the poverty line data. It is also important to check if the duration of reporting the expenditures (such as by month, by week, or any other interval) is equal and to make sure that the poverty line is also for the same time duration.

For example, both the consumption expenditure or tobacco use-attributable health care expenditures and the poverty line should be per person, per month. To do so in Stata, create new variables to generate per capita expenditures to be compared with the poverty line using the

household size variable (*hsize*). For example, per capita expenditures can be generated as *<gen pce =exptotal/hsize>*. Similarly, variables on per capita tobacco spending (*pcetob*) and on per capita health expenditures (*pcehealth*) should be generated by dividing the corresponding total expenditures by the household size variable. Furthermore, using the SAF value and pcehealth create the *pcehealthtob* variable that represents the per capita tobacco use– and SHS-attributable health care expenses. For example, if the SAF for tobacco use is 0.2, then a new variable *pcehealthtob* with the command *<gen pcehealthtob =pcehealth*0.2>* can be generated. And if the SAF for SHS exposure is 0.1, a new variable *pcehealthshs* with the command *<gen pcehealthshs=pcehealth*0.1>* to represent SHS-attributable per capita health care expenditures should be created.

For the purpose of computing the change in HCR after the incremental subtraction of different variables of interest, the following additional variables should be created:

(1) *pcet* (*pce* after tobacco expenditures are netted out): *<gen pcet=pce-pcetob>*, and

(2) *pceh* (*pce* after tobacco expenditures and tobacco use– and SHS-attributable health care expenditures are netted out): *<gen pceh=pcet-pcehealthtob- pcehealthshs>*. In case estimates of SAF for SHS exposure are not available, the formula for *pceh* may be reduced to *<gen pceh=pcet-pcehealthtob>*.

Lastly, the survey weight variable provided in the household expenditure data (such as *hweight*) should be adjusted to account for individual-level estimation of poverty. This can be done by multiplying this variable with the household size, that is, *<gen pweight=hweight*hsize>*. Once all of the above variables are generated, the impoverishing effect of tobacco can be estimated in Stata.

5.6 Estimating impoverishing impact of tobacco use

In Stata, estimation of HCR is quite straightforward, and Stata offers several user-written modules for this. For example, *<povdeco>*[180] is a module that estimates HCR and several other poverty measures with a single command. To do so, install the module with *<ssc install povdeco>* and run the command *<povdeco pce [fw=pweight], varpline(npl)>* where *pce* is the variable for monthly per capita expenditures, *npl* is the variable for the NPL, and *pweight* is the survey weight adjusted for household size. *Povdeco* will report HCR along with a poverty gap and squared poverty gap, by default. It also allows estimation of poverty by different subgroups using the option *<bygroup(groupvar)>*.

To estimate the HCR alone, however, a simple proportion command in Stata will work. For example, with the command below, the HCR can be estimated:

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

This can also be done after declaring the survey design using *<svyset>* command as explained in Chapter 2. In this case the command can be written as *<svy: proportion povdum>*.

Since the change in HCR must be determined after incremental subtraction of different forgone incomes as discussed earlier, this can be better implemented with the following code. The code below assumes that the variables have been generated as discussed in Section 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(--&&-);
```

As the code shows, the only variables from the data used in the code above are: *pce, pcet, pceh, npl,* and *pweight*. If the data have been prepared with these variable names, running the code would generate a 3x2 matrix in the Stata result window showing *pce, pcet,* and *pceh* as row headings and "HCR" and "Poor" as column headings. The first column shows the estimated HCR (value from 0 to 1) for *pce* (before subtracting any forgone income), *pce*t (HCR after subtracting forgone income from direct tobacco purchase), and *pceh* (HCR after subtracting forgone incomes from both tobacco purchase and tobacco use– and SHS-attributable health care expenditures). The corresponding values under the column "Poor" show the estimated number of poor persons in each successive step. If you like to also derive standard errors for each estimate, the stata module <*povdeco*> would give identical results for HCR along with other poverty measures. The following code can be used for this purpose. Variables pce and npl are the same as in the previous code.

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

Comparing two successive rows illustrates the change in both HCR and the number of poor people after the successive subtraction of each forgone income component. The number of poor people in the code is estimated by multiplying HCR by the total population as estimated from the household survey itself, which is possible using the person-specific weight variable. The scalar *r(sum)* is a saved result after the <*summarize*> command, and it shows the result of multiplying the mean by the population size. Alternatively, one can multiply HCR by the nationally available population data from other sources to arrive at the change in the number of poor people.

The analysis above can be done with different subgroups as well using any of the methods discussed above. However, the data need to be modified and new variables may have to be generated in order to do the analysis at the subgroup level. Section 7.5 in the Code Appendix includes an example do-file that details the code used in this section. Users will be able to copy and paste that into Stata's do-file editor and estimate the results with the appropriate accompanying data/variables described therein.

5.7 Case study from India

In India during 2004–2005, about 28.3 percent of the rural and 25.6 percent of the urban population were considered to be below the NPL by official government sources. The official poverty statistics are reported separately for rural and urban areas in the country and are also reported by state. The poverty line is also available separately for each state and by rural and urban areas. India also has the second largest number of tobacco users in the world.[72] The poverty rate and trends over time have always taken center stage in Indian development policy discourse. In this context, John et al.[170] examine the impoverishing impact of tobacco spending as well as that of tobacco use-related health care spending in India. Table 5.1 shows the results from their analysis.

The table first reports official estimates for HCR and the number of poor people in India by rural and urban areas. It then shows the separate effect of subtracting tobacco spending and tobacco use–attributable health care spending from per capita expenditures for rural and urban areas in India and then the combined effect of subtracting both the expenditures from per capita expenditures. The results show that the rate of poverty or HCR increased by 1.6 and 0.8 percentage points in rural and urban India, respectively, after subtracting forgone incomes from tobacco purchase and tobacco-related health care expenditures. Spending on tobacco and the associated health care spending impoverished about 15 million additional people in India. In other words, 15 million people in India who are above the official poverty line are in secondary poverty, experiencing lower standards of living in terms of their ability to spend on daily necessities because their money is being diverted to wasteful expenditures related to tobacco.

This has serious policy implications, too. If social welfare measures (a food subsidy, for example) are targeted to those who are officially below the NPL, those in secondary poverty will not be able to enjoy the benefits arising from such welfare measures and will continue to live in poverty.

Table 5.1 Changes in HCR and number of poor after accounting for tobacco use in India

	Rural	Urban	Total
(1) Official estimates			
Total population (million)	780.2	315.5	1095.7
Population BPL (%)	28.3	25.6	
Population BPL (million)	220.7	80.8	301.6
(2) Accounting for tobacco purchases			
Population BPL (%)	29.8	26.3	
Population BPL (million)	232.5	83.1	315.6
(3) Accounting for tobacco-related medical expense			
Population BPL (%)	28.4	25.7	
Population BPL (million)	221.4	81.1	302.5
(4) Combined effect of (2) and (3)			
Population BPL (%)	29.8	26.4	
Population BPL (million)	232.9	83.3	316.2

Note: BPL= Below poverty line. Source: John et al. (2011).[170]

6 Bibliography

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

16. Cameron, A. C. & Trivedi, P. K. *Microeconometrics Using Stata, Revised Edition.* (Stata Press, 2010).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Code appendices

7

7.1 Stata do-file to estimate prevalence and quantity elasticity for a single commodity

```
*====================================================================
* Date : January 20 2023
* 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
cd "Directory Path"
capture log close
log using Elasticity.log, replace
use hbs_data, clear
drop if [qcig==.&expcig!=.]|[qcig!=.&expcig==.]
gen uvcig=expcig/qcig
gen luvcig=ln(uvcig)
gen bscig=expcig/exptotal
gen lhsize=ln(hsize)
gen lexp=ln(exptotal)
```

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

*************************************************************************
*Estimating prevalence elasticity using logit model
*************************************************************************
egen pcig=mean(uvcig), by(clust)
egen pcig2=mean(uvcig), by(region)
replace pcig=pcig2 if pcig==.
global xvar "pcig lexp lhsize maleratio meanedu maxedu sgp1 sgp2 sgp3"
logit dcig $xvar
outreg2 using PrevalenceElast.doc, replace sideway
predict yhat_p, pr
margins, eyex(pcig)
margins, eydx(lexp)

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

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

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

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

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

*To estimate the bootstrap standard errors for price elasticity
cap program drop elast
program elast, rclass
tempname S R num den phi theta psi
qui corr y0c y1c, cov
scalar S=r(Var_2)
scalar R=r(cov_12)
scalar num=scalar(R)-(sigma12/n0)
scalar den=scalar(S)-(sigma11/n1)
cap scalar phi=num/den
```

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

7.2 Stata do-file to estimate prevalence and quantity elasticity for a single commodity by income groups

```
*=================================================================
* Date : January 20 2023
* 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 (both prevalence and quantity elasticities) for a single commodity,
* for example, cigarette, by different income groups.
* 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
*Add the directory path in close quotes below
cd "Directory Path"
capture log close
log using Elasticity.log, replace
use hbs_data, clear
drop if [qcig==.&expcig!=.]|[qcig!=.&expcig==.]
gen uvcig=expcig/qcig
gen luvcig=ln(uvcig)
```

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

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

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

```
*****************************************************************************
* Estimating quantity elasticity using Deaton's model
*****************************************************************************
keep if dcig==1
*anova luvcig clust
areg luvcig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
predict ruvcig, resid
scalar sigma11=$S_E_sse / $S_E_tdf
scalar b1=_coef[lexp]
gen y1cig=luvcig-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio ///
                -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu ///
                -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3
* Purging and averaging of unit values for the second stage is done for all income group combined
* so that all households in the same cluster faces same average unit values.
forvalues i=1/3 {
        areg bscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3 if inc==`i', absorb(clust)
        predict rbscig`i', resid
        scalar sigma22`i'=$S_E_sse/$S_E_tdf
        scalar b0`i'=_coef[lexp]
        gen y0cig`i'=bscig-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio ///
                -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu ///
                -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3 if inc==`i'
        qui areg ruvcig rbscig`i' lexp lhsize maleratio meanedu maxedu sgp1-sgp3 if inc==`i',
absorb(clust)
        scalar sigma12`i'=_coef[rbscig`i']*sigma22`i'
}

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

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

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

```
***********************************
*Use the code below if you like to test whether elasticites across income groups
*are statistically different
***********************************
cap program drop elast
program define elast, rclass
forvalues i=1/3 {
	preserve
	egen y1c`i'= mean(y1cig), by(clust)
	keep if inc==`i'
	sort clust
	egen y0c`i'= mean(y0cig`i'), by(clust)
	egen n0c`i'= count(y0cig`i'), by(clust)
	egen n1c`i'= count(y1cig), by(clust)
	qui sum bscig
	scalar Wbar`i' = r(mean)
	sort clust
	qui by clust: keep if _n==1
	qui ameans n0c`i'
	scalar n0`i'=r(mean_h)
	qui ameans n1c`i'
	scalar n1`i'=r(mean_h)
	drop n0c`i' n1c`i'
	tempname S`i' R`i' num`i' den`i' phi`i' theta`i' psi`i'
	qui corr y0c`i' y1c`i', cov
	scalar S`i'=r(Var_2)
	scalar R`i'=r(cov_12)
	scalar num`i' = scalar(R`i') - (sigma12`i' / n0`i')
	scalar den`i'=scalar(S`i')-(sigma11/n1`i')
	cap scalar phi`i' = num`i' / den`i'
	cap scalar zeta`i' = b1/((b0`i' + Wbar`i' * (1-b1)))
	cap scalar theta`i' =phi`i' /(1+(Wbar`i' - phi`i') * zeta`i')
	cap scalar psi`i' = 1-((b1*(Wbar`i' - theta`i'))/(b0`i' + Wbar`i'))
	return scalar Elast_`i' = (theta`i' / Wbar`i' )- psi`i'
	restore
}
end
elast
bootstrap elast1=r(Elast_1)elast2=r(Elast_2)elast3=r(Elast_3), reps(1000) seed(1):elast

*Testing the difference in elasticity.
test _b[elast1]=_b[elast2]
test _b[elast2]=_b[elast3]
test _b[elast1]=_b[elast3]
```

7.3 Stata do-file for estimating own- and cross-price elasticities for multiple goods using Deaton method

```
/*============================================================
Date: June 30 2022
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.

Note: These codes were written as part of "The Analysis of Household Surveys: A Microeconometric Approach to Development Policy", by Angus Deaton.
The original codes are 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, such as luvcig, luvbiri, etc.
* - The budget shares begine with bs, such as bscig, bsbiri
* - lnexp - 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 only for expository purposes
                -Education year of the household head - edu_head
                -Protion of adult members in the family - adultratio
                -Portion of male members in the family - maleratio
* - clust - variable identifying the primary samping unit or cluster
==============================================================*/

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

generate cluster=psu
*These are the commodity identifiers
gl goods "cig  biri slt"

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

*generate log of household size, expenditures
```

```
gen lhsize=ln(hsize)
gen lnexp=ln(dce)
gen lnyexp=ln(y_exp)

foreach X in $goods{
  gen luv`X'=ln(uv`X')
}

*To find the number of PSUs with no purchases of items below
gen anytobac=0
foreach X in $goods{
  recode anytobac 0=1 if uv`X'!=.
  egen psu_`X'=mean(uv`X'), by(psuid)
  egen tag=tag(psuid) if psu_`X'==.
  egen `X'_none=total(tag)
  drop tag
}

foreach X of global goods{
drop psu_`X' `X'_none
}
*Dropping clusters with less than 2 households reporting any type of tobacco consumption
bys psuid: egen consumingHH_psu=sum(anytobac)
drop if consumingHH_psu<2
drop anytobac consumingHH_psu

save "EditedData.dta", replace
* Defining a program data_matrix to use in bootstrap
cap program drop data_matrix
program define data_matrix
        *Number of goods in the system. It will be automatically taken.
        gl ngds : word count $goods

        matrix define sig=J($ngds,1,0) // var-covar matrix of u0 (e0e0)
        matrix define ome=J($ngds,1,0) // var-covar matrix of u1 (e1e1)
        matrix define lam=J($ngds,1,0) // covar matrix of u1 (e1e0)
        matrix define wbar=J($ngds,1,0) // average budget shares
        matrix define b1=J($ngds,1,0) // elasticity of quality w.r.t exp
        matrix define b0=J($ngds,1,0) // Coefficients of lnexp in BS regression

        *Average budget shares
        local ig=1
        foreach X in $goods{
                qui summ bs`X'
                matrix wbar[`ig',1]=r(mean)
                local ig=`ig'+1
        }
```

```
/*============================================================
First Stage Regression: Within-cluster
============================================================*/
/* creating a global for the variables we are controlling for. for example log of
expenditure, religion, education etc. We will have tho input these variables only ones in here.
*/
gl controls "lnexp lhsize lnyexp edu_head adultratio maleratio"
local ig=1
foreach X in $goods{
        *Cluster fixed effect regression
        *areg, instead of reg, is used for linear regression with a large dummy-variable set
        areg luv`X' $controls , absorb(cluster)

        *Measurement error variance
        *Summ of squares of errors / total degree of freedom for error
        *calculating var-covar matrix of u1 (e1e1)
        matrix ome[`ig',1]=$S_E_sse/$S_E_tdf
        *calculating Expenditure elasticity of quality
        matrix b1[`ig',1]=_coef[lnexp]

        *These residuals still have cluster effects in
        predict ruv`X', resid

        *Purged y's for next stage
        predict dresidual, dr
        gen y1`X'=_b[_cons]+dresidual
        drop dresidual luv`X'

        **Repeat for budget shares
        areg bs`X' $controls , absorb(cluster)
        *calculating residuals from the budget share regression
        predict rbs`X', resid

        *calculating var-covar matrix of u0 (e0e0)
        matrix sig[`ig',1]=$S_E_sse/$S_E_tdf
        *Calculating Coefficients of lnexp in BS regression
        matrix b0[`ig',1]=_coef[lnexp]
        predict dresidual, dr
        gen y0`X'=_b[_cons]+dresidual

        *This next regression is necessary to get covariance of residuals
        qui areg ruv`X' rbs`X' $controls , absorb(cluster)
        *Calculating covar matrix of u1 (e1e0)
        matrix lam[`ig',1]=_coef[rbs`X']*sig[`ig',1]
        drop bs`X' rbs`X' ruv`X' dresidual
        local ig=`ig'+1
}
```

```
        matrix list sig         // 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
        *drop lnexp lhsize adultratio meanedu maxedu res* hhtyp*

        *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 Regression: Between-cluster
        ============================================================*/
        *Averaging by cluster
        *Counting numbers of obs in each cluster
        local ig=1
        foreach X in $goods{
                egen y0c`ig'=mean(y0`X'), by(cluster)
                egen n0c`ig'=count(y0`X'), by(cluster)
                egen y1c`ig'=mean(y1`X'), by(cluster)
                egen n1c`ig'=count(y1`X'), by(cluster)
                drop y0`X' y1`X'
                local ig=`ig'+1
        }

        sort clust
        *keeping one obs per cluster
        *NB subround and region are constant within cluster
        qui by clust: keep if _n==1
        *Saving cluster level information
end
data_matrix

/*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)

foreach ig of numlist 1/$ngds{
        regress y0c`ig' regiond2 regiond3 regiond4
        predict tm, resid
        replace y0c`ig'=tm
```

```
        drop tm
        qui regress y1c`ig' regiond2 regiond3 regiond4
        predict tm, resid
        replace y1c`ig'=tm
        drop tm
}
drop regiond*
*/
cap program drop var_covar
program define var_covar
        matrix define n0=J($ngds,1,0)
        matrix define n1=J($ngds,1,0)

        *Averaging (harmonically) numbers of obs over clusters

        foreach ig of numlist 1/$ngds{
                *replace n0c`ig'=1/n0c`ig'
                *replace n1c`ig'=1/n1c`ig'
                qui ameans n0c`ig'
                matrix n0[`ig',1]=(r(mean_h))
                qui ameans n1c`ig'
                matrix n1[`ig',1]=(r(mean_h))
                drop n0c`ig' n1c`ig'
        }

        *Making the intercluster variance and covariance matrices
        *This is done in pairs because of the missing values
        matrix s=J($ngds,$ngds,0) // between-cluster var-covar matrix of y1
        matrix r=J($ngds,$ngds,0) // between-cluster covar matrix of y1
        local ir=1
        foreach ir of numlist 1/$ngds{
                local ic=1
                foreach ic of numlist 1/$ngds{
                        qui corr y1c`ir' y1c`ic', cov
                        matrix s[`ir',`ic']=r(cov_12)
                        qui corr y1c`ir' y0c`ic', cov
                        matrix r[`ir',`ic']=r(cov_12)
                        }
        }

        *We don't need the data any more
        drop _all
        matrix list s // between-cluster var-covar matrix of y1
        matrix list r // between-cluster covar matrix of y1
        *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
        matrix def sf=s
        matrix def rf=r
        foreach ig of numlist 1/$ngds{
                matrix sf[`ig',`ig']=sf[`ig',`ig']-ome[`ig',1]/n1[`ig',1]
                matrix rf[`ig',`ig']=rf[`ig',`ig']-lam[`ig',1]/n0[`ig',1]
        }

        matrix invs=syminv(sf)
        matrix bhat=invs*rf // The errors-in-variable estimator with ME correction

        *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
        matrix def xi=J($ngds,1,0)
        matrix def el=J($ngds,1,0)
        foreach ig of numlist 1/$ngds{
                matrix xi[`ig',1]=b1[`ig',1]/(b0[`ig',1]+((1-b1[`ig',1])*wbar[`ig',1]))
                matrix el[`ig',1]=1-b1[`ig',1]+b0[`ig',1]/wbar[`ig',1]
        }

        global ng1=$ngds+1
        matrix iden=I($ngds)
        matrix iden1=I($ng1)
        matrix itm=J($ngds,1,1)
        matrix itm1=J($ng1,1,1)
        matrix dxi=diag(xi)
        matrix dwbar=diag(wbar)
        matrix idwbar=syminv(dwbar)
end
var_covar
display("Average budget shares")
matrix tm=wbar'
matrix list tm // 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
        display("Theta matrix")
        matrix list theta
        matrix ep=bhat'
        matrix ep=idwbar*ep
        matrix ep=ep-iden
        matrix ep=ep*psi
        display("Matrix of price elasticities")
        matrix list ep // price elasticity of demand without symmetry restrictions)
end
mkels

/*================================================================
**Completing the system 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'
        matrix exp_elas=xe'
        display("extended outlay elasticities (or total expenditure elasticities)")
        matrix list tm  // expenditure elasticities from the complete system
        matrix xxi=itm1-xb1
        matrix xxi=dxwbar*xxi
        matrix xxi=xxi+xb0
        matrix tm=diag(xb1)
        matrix tm=syminv(tm)
        matrix xxi=tm*xxi
        matrix dxxi=diag(xxi)
        *Extending psi
        matrix xpsi=dxxi*xtheta
        matrix xpsi=xpsi+iden1
        matrix atm=dxxi*dxwbar
        matrix atm=atm+iden1
        matrix atm=syminv(atm)
        matrix xpsi=atm*xpsi
        matrix ixpsi=inv(xpsi)
        *Extending bhat & elasticity matrix
        matrix xbhatp=xtheta*ixpsi
        matrix xep=idxwbar*xbhatp
        matrix xep=xep-iden1
        matrix xep=xep*xpsi
        display("extended matrix of elasticities")
        matrix list xep   // price elasticities from the complete system without symmetry
end
complet // this command can be dropped if we are only interested in
*symmetry constrained estimates as given below
*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
*Calculates two matrices, the commutation matrix and the lower diagonal
*selection matrix that are needed in the main calculations;
cap program drop commx
program define commx
        mata: st_matrix("`2'", kmatrix(`1',`1'))
end

**for vecing a matrix, that is, stacking it into a column vector
cap program drop vecmx
program def vecmx
        mata: st_matrix("`2'", vec(st_matrix("`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
        mata: st_matrix("`2'", colshape(st_matrix("`1'"),$ngds)')
end

vecmx bhat vbhat
** R matrix for restrictions
lmx $ngds llx
commx $ngds k
cap program drop matrices
program def matrices
        global ng2=$ngds*$ngds
        matrix bigi=I($ng2)
        matrix k=bigi-k
        matrix r=llx*k
        matrix drop k
        matrix drop bigi
        matrix drop llx
        ** r vector for restrictions, called rh
        matrix rh=b0#wbar
        matrix rh=r*rh
        matrix rh=-1*rh
        **doing the constrained estimation
```

```
        matrix iss=iden#invs
        matrix rp=r'
        matrix iss=iss*rp
        matrix inn=r*iss
        matrix inn=syminv(inn)
        matrix inn=iss*inn
        matrix dis=r*vbhat
        matrix dis=rh-dis
        matrix dis=inn*dis
        matrix vbtild=vbhat+dis
        unvecmx vbtild btild
        **the following matrix should be symmetric
        matrix atm=b0'
        matrix atm=wbar*atm  // Eqn. 5.98
        matrix atm=btild+atm
        matrix list atm
        **going back to get elasticities and complete sytem
        matrix bhat=btild
end
matrices
mkels
complet

/*To estimate the bootstrap standard errors for price and expenditure elasticities */

drop _all
vecmx xep vxep
vecmx exp_elas vexp_elas
matrix obs=vxep\vexp_elas
matrix observe=obs'
global nels=$ng1*$ng1
drop _all

use "EditedData.dta", replace
        capture program drop bootstrap
        program define booststrap, rclass
                preserve
                bsample _N
                data_matrix
                var_covar
                mkels
                vecmx bhat vbhat
                lmx $ngds llx
                commx $ngds k
                matrices
                mkels
                complet
```

```
                vecmx xep vxep
                vecmx exp_elas vexp_elas
                foreach ic of numlist 1/$nels{
                        return scalar e`ic'=vxep[`ic',1]
                }
                foreach iex of numlist 1/$ng1{
                        return scalar exp`iex'=vexp_elas[`iex',1]
                }
                restore
        end

local x ""
local x1 ""
foreach X of numlist 1/$nels {
        local y "e`X'=r(e`X')"
        local z "`x' `y'"
        local x "`z'"
        }
foreach X of numlist 1/$ng1{
        local y1 "exp`X'=r(exp`X')"
        local z1 "`x1' `y1'"
        local x1 "`z1'"
        }
simulate `x' `x1', reps(1000) seed(122002): booststrap

bstat, stat(observe)
log close
```

7.4 Stata do-file for estimating crowding-out effect of tobacco spending

```
*============================================================
* 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 health care 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

use "$pathin/DataQAIDS.dta", clear
**********************************************************************
*T-test for comparing mean budget shares
**********************************************************************
*Generate a binary variable for tobacco spending
gen tob= 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 _b[c.bs_`X'@0.tob] - _b[c.bs_`X'@1.tob]          matrix B[`i', 3] = r(estimate)
   matrix B[`i', 4] = r(t)
}
matrix rownames B =`items'
matrix colnames B = non-spenders spenders Difference t-stat
matrix list B
*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
* that is, 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 (that is, 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 subgroup
*********************************************************************
*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
```

7.5 Stata do-file for estimating impoverishing effect of tobacco use

```
*============================================================
* 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 health care 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 expenditures 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
```

INSTITUTE FOR HEALTH RESEARCH AND POLICY

uic

tobacconomics

Economic Research Informing Tobacco Control Policy

www.tobacconomics.org
@tobacconomics