tobacconomics
Economic Research Informing Tobacco Control Policy

A Toolkit on

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

INSTITUTE FOR
HEALTH RESEARCH
AND POLICY
UIC

Suggested Citation: John R., Chelwa G., Vulovic V., Chaloupka F., Using Household Expenditure Surveys for Research in the Economics of Tobacco Control. A Tobacconomics Toolkit. Chicago, IL: Tobacconomics, Health Policy Center, Institute for Health Research and Policy, University of Illinois at Chicago, 2019. www.tobacconomics.org

Authors: This Toolkit was written by Rijo John, PhD, Senior Fellow, Centre for Public Policy Research, Kerala, India; Grieve Chelwa, PhD, Senior Lecturer in Economics, University of Cape Town, South Africa; Violeta Vulovic, PhD, Senior Economist, Health Policy Center, University of Illinois at Chicago, and Frank Chaloupka, PhD, Research Professor, University of Illinois at 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 largest preventable cause of death and a main risk factor for several non-communicable diseases resulting in more than 7.2 million annual deaths globally.[1] Worldwide, 12% of all adult deaths (30 years of age and older) are attributed to tobacco (16% among men, 7% 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% of the world's annual gross domestic product (GDP).[5] The global health and economic burden of tobacco use is increasingly borne by LMICs.

"Tobacco – a threat to development" was the theme of the 2017 World No Tobacco Day. It is clear that the unabated consumption of tobacco in various forms has the potential to hinder 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 non-communicable diseases, but also to improve economic growth and reduce poverty. A large body of literature from 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 the 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 policymaking, 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. 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, one could have estimated 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 do several policy-relevant analyses for the purpose of tobacco control policymaking using cross-sectional data from household surveys. Several LMICs conduct household surveys sporadically on a variety of topics which can give interesting insights on consumer behavior with respect to tobacco consumption.

This toolkit will review 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 will demonstrate 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, 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 healthcare expenditures on national poverty head counts. It will briefly discuss the theoretical background and economic rationale of each of these issues, methods of estimation, and the use of the statistical software, Stata®, to implement these methods.

This toolkit is one of several toolkits developed by the World Bank, WHO and Tobacconomics which focus on providing guidance on conducting an economic analysis of tobacco demand, the impact of tobacco consumption on employment, equity, illicit trade, and on 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 which would 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 or economics of tobacco control issues on the part of the reader. However, background in economics and econometrics, with a basic understanding of 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 policymakers, analysts in government agencies, as well as 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 is written to provide technical guidance on three important topics in the area of economics of tobacco control: first, estimating own- and cross-price elasticities (Chapter 3); secondly, estimating the crowding out nature of tobacco spending (Chapter 4); and thirdly, quantifying the impoverishing effect of tobacco use (Chapter 5). All these topics are discussed with the intention of performing analysis with HES data. The discussion in each chapter will start with an introduction and the principles behind the topic along with the rationale for doing the analysis. It will then be followed by a brief technical discussion on the econometric methods used. The discussion of econometric methods, however, are 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 will be followed by a brief discussion of 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 will be presented along with the interpretation of results toward the end of the chapter.

The toolkit will discuss the relevant analysis methods for all tobacco products combined, or smoked and smokeless tobacco products separately, or for individual tobacco products such as cigarettes, bidi and other chewing tobacco products depending on the particular issue being addressed. For example, when estimating own- and cross-price elasticities, it may be useful to present the analysis for each of the tobacco products so that one can estimate 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 bidi 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 better sense to combine all tobacco products into one category and examine the impact by different socioeconomic groups.

The toolkit is organized as follows: Chapter 2 provides an introduction to HES with a focus on surveys in LMICs. It will discuss the contents of HES as it pertains to tobacco. In particular, it will cover various questions pertaining to tobacco consumption and expenditures on different tobacco products of inquiry in HES. The chapter will also briefly discuss 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. The chapter also 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. The primary method discussed will be the one developed by Deaton[7] along with a step-by-step explanation of the Stata commands for estimating price elasticities from HES data. Estimates of price elasticities using local data are often useful and desired for tax policies on tobacco in the respective countries.

Chapter 4 explains the methods to examine the impact of spending on tobacco on intra-household resource allocation. Following an approach of conditional demand systems[8,9] it will show how expenditures on tobacco systematically crowd out expenditures on other commodities within a household. The analysis will discuss ways to estimate the crowding out by different socioeconomic subgroups. The analytical method, as well as the Stata code for executing the model, will also be presented.

Chapter 5 covers the impoverishing effect of spending on tobacco. It will discuss the estimation of the actual amount spent on purchasing tobacco as well as the increased healthcare expenditures attributable to consumption of tobacco and second-hand smoking (SHS). It will then demonstrate how accounting for tobacco spending and associated health expenditures will impact the estimate of national poverty measured by the head count ratio. Step-by-step estimation along with relevant Stata code will be discussed.

As much as possible, these chapters will 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 are italicized. However, the command itself has to 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[10] and has conducted regular and periodic surveys since then every few years. The Living Standard Measurement Surveys (LSMS) were started by the World Bank in 1979 and these multi-topic household surveys have collected household consumption expenditure from about 38 countries around the world,[11] several of them African and Asian countries. There are several 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 decisionmakers, the research community, and other stakeholders,"[12] 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 is available are LMICs. This catalogue 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-series macroeconomic variables, HES provide meaningful cross-sectional data, sometimes for multiple time periods for the same country.

Statistical agencies that usually undertake the HES in most countries only publish summary reports that present only grouped data and 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 that one would like to perform. 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. The microdata, however, is 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, one may receive the data 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[11] 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 assign an equal probability to each household being selected from the frame. Although the sample sizes vary widely depending on the purpose of the survey, given 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 5 million households.[7] 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.[7] Clusters are typically called the first stage units (FSU) or primary sampling units (PSU) as it is the first unit that is 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 that households can be selected based on relevant attributes such as geographical area, ethnic affiliation, level 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.[7] The probability weights for households in each strata 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 significantly vary 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, collected food consumption with a seven-day recall and non-food consumption within both seven-day and 30-day recalls.[13]

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 the 30 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 only 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 levels 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.[7] A brief and conceptual summary of the salient issues follows:

a) **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 formulae 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[7] below gives examples of how to apply survey weights in Stata while computing certain descriptive statistics.

b) **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, i.e., as Deaton[7] 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 Gauss-Markov theorem.[14] 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[7] goes on to say that a weighted regression provides a consistent estimate of the population regression function provided that the assumption about functional form of the regression is correct, i.e., when the regression function itself is the object of interest. If the interest is to estimate behavioural models where behaviour may be different for different subgroups, weighting in the regression is of no use. In conclusion, as Cameron and Trivedi observe,[15] 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.

c) **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 different from those in other clusters which are usually widely separated geographically. In other words, there will be more homogeneity within clusters than between them. To the extent 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.

d) **Heteroskedasticity of OLS residuals:** Distributions of households over different variables of interests such as income and consumption of different goods are usually not normally distributed and, 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. This would leave the OLS estimates inefficient and would invalidate the usual formulas for standard errors and will 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.

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

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

(ii) Omitted explanatory variables—i.e., 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 it is the unobserved variation across individual units of this omitted or unobservable variable; and

(iii) Measurement errors—i.e., 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,[7] 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, one encounters most, if not all, of these different sources of endogeneity. 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) (e.g., two-stage least squares method)[14] 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 several universities and institutions around the world, thereby facilitating exchanges and collaborations between researchers in multiple disciplines and institutions.[16] Below are some useful tips that make working with Stata much easier.

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).

Creating a log file: While a do-file keeps a record of all the commands and allows editing them as necessary, 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 that 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 toward 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*>.

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 the topics of interest within the 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/.)

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 a double quote, 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 which call files (e.g. 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 would vary depending on the operating system. Use of macros is discussed in more detail later.

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 which are referred to in the Stata documentation and accessible online. From Stata's user interface, navigate to "File>Example data sets" 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, and by default, the data sets are obtained from http://www.stata-press.com/data/r15/. 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.

Using logical and relational operators: Stata uses several logical and relational operators to help with manipulating data sets. Some of the commonly used operators and their intended meanings are given here. 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 *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 i.bmi age#sex*> and <*regres bp age##sex*>. Alternatively, to regress *bp* on *age* and *bmi* and the interaction between them, write <*regres bp age##c.bmi*>.

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

Using macros: Macros are abbreviations or *aliases* which have both a name and a value. When its name is dereferenced, it returns its value.[17] 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. We use macros for several purposes including making tasks simpler, making do-files more organized, shortening the length of STATA code, and various other conveniences while programming. Macros can be of two types, local and global, depending on its scope, i.e., 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.[18]

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 (‘’).[18] For example, define a local macro with the name *indvar* as <*local indvar price expenditure hsize*> and issue another command <*summarize ‘indvar’*> it will return 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*> it will return 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.

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 (i.e., 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 has to 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 and are defined to be a natural log of existing variables in the list. We could add multiple lines of commands one below the other and all of which will be repeated over all 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 its 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.

Returning stored results: Stata regularly stores results from commands in local macros which can be called in for various purposes. For example, upon issuing a <*summarize*> command for a variable <*sum varname*> it 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
```

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 which 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 + 3 standard deviations of the variable under consideration. The third line in the loop replaces any values higher than the mean plus 3 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.

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 gives a working example of using some of the Stata tips already covered.

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
```

Using add-on commands: Stata allows people to write third-party commands (called "ado-files") which 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 making 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*>, <*ivreg2*> are some of the popular add-ons. Use the command <*ssc whatshot*> to check out some of the most popular add-on packages available for download.

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 upper case, it will return an error or execute an unintended code.
- Most Stata commands can be abbreviated. For example, *the* <*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. So, it takes 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 with the survey agencies, and other standard practices and customs in different fields. The HES data that is of interest to us will usually be quantitative tabular data. It is usually 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 is 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 which are 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 is archived by the respective survey agencies usually provides 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 is 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 this 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)

```
W155110211307112666212020ll     2    4     33815604     488     573003232    0030251
W155110211307112666212020З1      2    4     33815604000490     547001213    0010211
W1551102113071126662120205l  2  4  33815604  437  460004413  0610251
W15511021130711266621202072 2  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
```

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 for 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 easily be corrected so that 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.

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 5 parts. The first part tells Stata to begin reading the data file from the byte position mentioned in parenthesis. 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 - five column integer variable, %10s - ten column string variable, and %7.2f – 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 (e.g., *dictionary.dct*).

Merging data: Household surveys often come with multiple data files or records for households and individual members within households. Furthermore, for households themselves, there may be multiple records. For example, one file with the basic household characteristics such as household size, SES group they belong to, place of residence etc., and another file for their consumption expenditures. The data on consumption expenditures itself 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 *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 command <*sort hhid*> (where hhid is the identifying variable for households) and then run the command <*drop if hhid==hhid[_n-1]*>. Alternatively, use 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 the 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 to be able to find a set of variables that can uniquely identify every single observation in each of the data to be merged. This needs to be understood from the 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 of one household into the other. 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)[18] 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 1000 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 this data, a unique household id can also be identified using only variables *psu* and *hhold*.

A unique household id variable (*hhid*) for this 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 stratum, one would have to use all three variables while generating *hhid*. So, any merging of two household level records in this 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 has to 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 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 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.

Therefore, the data from two different data files should be always 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 which matched from both the master and using data files i.e., observations for which the variable (*_merge*) takes the value 3. In order to use only variables with no missing data from both 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 using data file.

Apart from merging different files (e.g., 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 so that there is 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 for each round of HES needs to be prepared first. Once append 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 which 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 inflation adjust any expenditure or price variables so that the values across different rounds of data are in constant terms and are therefore comparable.

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, we will have only 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 is in long format, it 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.

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 while 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 is 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 to do?)	Step (What to do?)	Command (How to do?)
Identify the variables and fix incorrect codes	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, e.g., 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 collaboration	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 design 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 is declared with <*svyset*>, only the prefix <*svy:*> needs to precede each command. The syntax of <*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 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 that data into Stata with the command <*webuse nhanes2*>. The data gives 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. One needs to carefully read the survey documentation to understand the description of variables. Since published reports from the survey also present important point estimates, one can 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 one is using the correct sampling weights and survey design elements 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, one could simply run the command <*svy: mean varname*>. If the mean is computed for a binary variable it would display proportions. One can 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 categorise 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, it is done with the help of weights and robust cluster options. For example, in the above example 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. The price elasticity is one of the most important parameters to be considered when designing tax policy, as it provides an insight to policymakers on the responsiveness of demand to changes in price. Based on the estimated price elasticity, policymakers 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 counter-argument to those claiming that raising taxes would unambiguously result in reduced tax revenues.

Policymakers are interested in responsiveness in tobacco consumption to not only changes in prices of tobacco (i.e., own-price elasticity), but also to changes in prices of other goods, such as their potential complements (e.g., alcohol, coffee, etc.), or their substitutes. Similarly, policymakers may want to know the impact of a change in price of one type of tobacco product (e.g., cigarettes), on other types (e.g., Roll-Your-Own cigarettes), as the impact of their policy may be effectively reduced if, for example, there is a space for downward substitution.

In this chapter these concepts are defined in detail with examples. In the later part of the chapter, Stata code is provided to enable the reader to estimate elasticities. Finally, an example from Uganda is presented.

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 1% 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 5% whenever the good's price rises by 10%. Similarly, a price elasticity of demand of -1.5 implies that the quantity demanded of the good in question declines by 15% whenever its price rises by 10%.

Goods with a price elasticity of demand less than 1 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 1 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, or the addictive/habitual nature of the product. With this in mind, tobacco products having few substitutes and being addictive tend to have relatively price inelastic demand.

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,[4,19,20] meaning that demand for tobacco is inelastic, which is expected given the addictive nature of this product, as well as the availability of a 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.[4,21–23]

In addition to own-price elasticity, we can also define cross-price elasticity. 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 1%, *ceteris paribus*. Unlike the case with own-price elasticity where it 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 that both goods satisfy the same need. An example of substitutes is bottled water and tap water.

Further, 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 1% 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 (rice, maize (corn), etc.) 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. A positive income elasticity of demand, on, for example, cigarettes in a country, implies that tobacco control efforts must be stepped up, especially in periods of rising incomes in that country.

3.2 Econometric issues in demand estimation

There are several theoretical and practical issues to consider in the estimation of price elasticity 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, in reality, things tend to be more complex, because in market interactions, demand influences price as much as price influences demand. One can observe this 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. Further, we know that other factors (e.g., incomes, tastes, the weather, and prices of related goods) can, outside of the influence of price, influence demand of the good.

The issues explained above are in econometric analysis referred to as the *endogeneity problem*, or the *identification problem*, and a failure to adequately address them would lead to obtaining biased estimates; (i.e., the estimates are 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 on unrealistically positive, or negative, impact 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 we are interested in. 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 (i.e., 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,[24–27] because the exact source of the price change that led to a change in quantity demanded could 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.

3.2.2 Angus Deaton's solution to the identification problem

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)[28] 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 non-linear 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[29] and detailed in his book[7] builds on Deaton and Muellbauer (1980).[28] However, it differs slightly from the AIDS in that it allows for zero purchases whereas the original model did not. Allowing for zero purchases is particularly attractive in the case of tobacco given that it is often not consumed by everyone in a population. Moreover, tax policy impacts based on the own-price elasticity estimates are better examined when all households are present in the analysis given that some households which do not consume tobacco now may begin consumption later if and when prices decrease, income increases, etc.

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 cesses in different jurisdictions in the same country. Thus, transportation cost or these other factors affecting price changes across geographical regions implicitly serve as an instrument and is the main factor influencing the price, which in turn influences demand. Therefore, genuine variation in price across clusters is assumed for the identification of price elasticities in this model.

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 3.2.3).

Another significant contribution was that, while households do not report the market price in the survey, it could be inferred from their purchasing decisions by calculating the ratio of household expenditure on a good to the quantity of the good. However, this ratio 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 (i.e., "quality effects"). If not properly dealt with, this might lead to the 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 and/or quantities on goods purchased. Deaton proposes formulae 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.[7,30–32]

3.2.3 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 using HES data. Researchers planning to implement this model are advised to read Chapter 5 from Deaton (1997)[7] to understand 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:

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

Inv_{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 total household expenditure for household *h* in cluster *c* and Inx_{hc} is the log of total household expenditure over the relevant reference period. Z_{hc} is a vector of household-specific characteristics which might include variables on household structure (e.g., household size, proportion of adults, proportion of males, etc.) and household demographics (e.g., age, gender, marital status, schooling and employment status of head of household, etc.). f_c is a cluster-fixed effect and treated as an error in addition to the error term u^0_{hc} in equation 3.2, while u^1_{hc} is the standard regression error term. Both u^0_{hc} and u^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,[7] “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, $In\pi_c$ 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, allows us to check 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 (i.e., the magnitude of β^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 is no price data, prices are proxied by cluster-fixed effects. The relationship between the two errors, u^0_{hc} and u^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:

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

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

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 $\widehat{y^1_c}$ and $\widehat{y^0_c}$ do not have the h subscript because they represent cluster averages. $\widehat{y^1_c}$ and $\widehat{y^0_c}$ 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 $y^1_c = \alpha^1 + \psi ln\pi_c + u^1_c$ *and* $y^0_c = \alpha^0 + \theta ln\pi_c + f_c + u^0_c$, 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, $\widehat{y^0_c}$, on cluster level unit values, $\widehat{y^1_c}$. The coefficient on $\widehat{y^1_c}$ in such a regression can alternatively be obtained by dividing the covariance between $\widehat{y^0_c}$ and $\widehat{y^1_c}$ by the variance of y^1_c . That is $\hat{\phi}$, the estimate of the coefficient on y^1_c, is obtained by:

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

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:

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

where $\overline{w}$ 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:

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

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

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

Finally, Deaton also proposes the following formula for obtaining the estimate of the expenditure elasticity of demand, $\hat{\varepsilon}_1$:

$$\widehat{\varepsilon}_1 = 1 + \left(\frac{\hat{\beta}^0}{\overline{w}}\right) - \hat{\beta}^1 \qquad (3.11)$$

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 India,[33–37] Vietnam,[38] China,[39] Uganda,[40] and Ecuador,[41] 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,[7] for the purpose of tax and price reform, one needs to include all households in the analysis whether they purchase or not. The estimates of own-price elasticity for cigarettes in these studies ranged from -0.1 to -0.6 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[42] 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 tends 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 depends on having 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.[7] 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: (1) the model corrects measurement errors but it cannot guarantee the consistency of parameters with a small number of clusters; and (2) 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,[7] "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 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). Further, other social and demographic data at the individual level should also be present. Whereas such data sets are widely available in high-income countries, they tend to be the exception in LMICs. Researchers having 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)[43] 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)[42] 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 methods for obtaining elasticities.

3.3 Preparing data for analysis

While Chapter 2 provided detailed information on extracting data, cleaning it, merging different data sets, and other necessary data management tips, it is important to provide specific details on the variables necessary for the estimate of price elasticity using Deaton's method discussed above. For any new variables discussed here, it is important to take it through all the processes discussed in Chapter 2. This section discusses how the specific variables 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 as well as the expenditures spent 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 beneficial.

First, unit values for each of the tobacco products for which data is available should be created. 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 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*) using the command <*gen bscig = expcig/exptotal*> where *exptotal* is the total expenditures on all items should be constructed. For those households with no reported expenditures on cigarettes this would generate a missing value. In this case, the command <*replace bscig=0 if bscig==.*> to indicate zero budget share on cigarettes for those households with no spending on cigarettes instead of leaving out all those households using a missing value should be used. This is done because, as Deaton[7] suggests, it is useful to include all households in the analysis for the purpose of tax and price reform whether they purchase or not. 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 (village or primary sampling units in cross-section survey) as in Deaton's original analysis, or it can be a point in time (e.g., survey wave) if combining different rounds of surveys or a combination of both PSU and survey wave.[42]

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 socio-demographic 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 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 (rural/urban areas, province, district, etc.), among others.

3.4 Estimating price elasticity with Stata

This section provides the Stata code for the estimation of own-price elasticity for a single tobacco product (cigarette) using Deaton's method discussed earlier. Deaton provides detailed Stata code for estimating own- and cross-price elasticities for different products and it can be downloaded from http://web.worldbank.org/archive/website00002/WEB/EX5_1-2.HTM. The Code Appendix in Section 7.2 reproduces 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 case in Appendix 7.2 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 elasticites and allows introduction of other theoretical restrictions on the demand system as discussed in Deaton,[7] 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 elasticity.

Testing for spatial variation in unit values

As indicated in the method section, 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,[7] a significant *F-statistic* and R^2 value around 0.5 (i.e., 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 price elasticity.

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*> instead of <*regress*> is used since this is the command used for a linear regression with a large dummy-variable set. The command implicitly includes a dummy variable for each cluster dropping one and yet, 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 are variables after purging off any effects of household-specific characteristics that are the reason for 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. Coefficient for the log expenditure is also stored for later use. The scalar *b1* which is the coefficient of 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 elasticity

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 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.

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* on *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 ϕ=θ/ψ 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, one can simply estimate this hybrid parameter 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=scalar(R)-(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=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
```

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 7.2 reproduces detailed code from Deaton to estimate own- and cross-price elasticities using Deaton's method.

3.5 Case study from Uganda

This section presents results from a study in Uganda with a step-by-step process that eventually leads to estimates of elasticities. The study used data from the 2005 and 2009 editions of the Uganda National Panel Survey (UNPS) and used only households reporting positive consumption of cigarettes in the analysis. The UNPS is conducted by the Uganda Bureau of Statistics with assistance from the World Bank. The data are easily downloadable from the World Bank's Living Standards Measurement Survey website (http://microdata.worldbank.org/index.php/catalog/lsms). Results are presented below in a step-by-step manner to aid understanding of the technique. Further, the 2005 and 2009 editions of the UNPS are treated as separate cross-sections.

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 were processed as described in Chapter 2. The full list of variables that are used to estimate elasticities in Uganda 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.

Table 3.1 Variables used for own-price elasticity estimation from the 2005 and 2009 UNPS

Variable	
1	Average cigarette share in total household expenditure
2	Natural logarithm of unit value
3	Natural logarithm of household expenditure
5	Natural logarithm of household size
6	Natural logarithm of years of schooling of household head
7	Natural logarithm of age of household head
8	Proportion of males in the household
9	Proportion of adults in the household
10	Dummy variable for whether adult head works
11	Dummy variable for whether household head is male

Notes: Relevant variables from the 2005 and 2009 editions of the UNPS

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 3.2 Testing spatial variation in log unit values

2005 sample				2009 sample			
F-statistic	p-value	R-squared	n	F-statistic	p-value	R-squared	n
1.29	0.08	0.70	274	1.12	0.33	0.72	173

Notes: The F-statistic and the p-value are associated with the null hypothesis of no spatial variation in unit values. The hypothesis is rejected in the 2005 but not in the 2009 sample. The R-squared measures the proportion of variation in prices taking place between clusters. n is the total number of households.

The outcome of the ANOVA exercise shows that at least 70% (R-squared of 0.70) of the variation in unit values is explained by between-cluster effects. The *F-statistic* is associated with the hypothesis of no spatial variation in prices—which is rejected in the 2005 sample and not rejected in the 2009 sample.

Step 3: Within-cluster regressions

The next step is to estimate the within-cluster regressions, i.e., the unit value regression and budget share regressions, as per equations 3.2 and 3.3 above. The results of these regressions are contained in Table 3.3 and Table 3.4.

The results of the unit value regression in Table 3.3 show that reported unit values are positively correlated with household expenditure. This result is statistically significant at the 5% level for both years of the survey. This is indicative of the presence of quality effects in the data as per the discussion in Section 3.2.3. The results of the budget share regression in Table 3.4 show that the cigarette budget share declines with household expenditure. This result is statistically significant at the 1% level for both survey years.

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.

Table 3.3 Results from the unit value regression

Variables	2005 lnv	2009 lnv
Lnx	0.234***	0.115**
	(0.051)	(0.048)
Size	-0.042	-0.010
	(0.124)	(0.119)
Adults	-0.203	0.159
	(0.295)	(0.300)
Males	0.261	0.131
	(0.216)	(0.223)
Education	-0.143*	0.108
	(0.080)	(0.074)
Age	-0.015	-0.409**
	(0.153)	(0.166)
Gender	0.217	0.218
	(0.163)	(0.183)
Work	-0.144	0.101
	(0.141)	(0.118)
Constant	4.957***	6.602***
	(0.692)	(0.739)
No. of households	233	147
R-squared	0.115	0.126

Notes: Results of the regression of the log of unit value (lnv) on the log of household expenditure (lnx) and other household characteristics. Household size (Size), education of household head (Education) and age of household head (Age) are in natural logarithms. Adults refers to the proportion of adults in a household and adults are defined as aged 18 years or older. Males is the proportion of males in a household. Gender is a dummy variable which takes on the value of 1 if the household head is male and zero if they are female. Work is a dummy variable which takes on the value of 1 if the household head is employed and zero otherwise. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 3.4 Results from the budget share regression

Variables	2005 w	2009 w
Lnx	-0.056***	-0.065***
	(0.017)	(0.023)
Size	0.002	0.039
	(0.031)	(0.043)
Adults	0.008	0.092
	(0.072)	(0.103)
Males	0.013	0.010
	(0.059)	(0.068)
Education	-0.001	-0.012
	(0.020)	(0.025)
Age	0.028	-0.077
	(0.044)	(0.072)
Gender	-0.038	-0.108*
	(0.037)	(0.056)
Work	0.037	0.058
	(0.037)	(0.039)
Constant	0.533***	0.963***
	(0.193)	(0.292)
No. of households	233	147
R-squared	0.866	0.909

Notes: Results of the regression of the cigarette budget share (w) on the log of household expenditure (lnx) and other household characteristics. Household size (Size), education of household head (Education) and age of household head (Age) are in natural logarithms. Adults refers to the proportion of adults in a household and adults are defined as aged 18 years or older. Males is the proportion of males in a household. Gender is a dummy variable which takes on the value of 1 if the household head is male and zero if they are female. Work is a dummy variable which takes on the value of 1 if the household head is employed and zero otherwise. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Cluster-fixed effects are suppressed for space reasons but are jointly statistically significant at the 1% level for the 2005 and pooled samples and at 10% for the 2009 sample.

Table 3.5 Estimates of the own-price elasticity of demand for cigarettes in Uganda

	2005	2009
$\widehat{\varepsilon_P}$	-0.326*** [0.021]	-0.258*** [0.011]
	(-0.368 , -0.284)	(-0.280 , -0.235)
No. of households	233	147
No. of clusters	184	130

Notes: Estimates of the price elasticity of demand for cigarettes in Uganda. Bootstrapped standard errors are in square brackets. 95% confidence intervals are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 3.6 Estimates of expenditure elasticity of demand for cigarettes in Uganda

	2005	2009
$\widehat{\varepsilon_I}$	0.132 [0.338]	0.043 [0.539]
	(-0.531 , 0.796)	(-1.014 , 1.100)
No. of households	233	147

Notes: Estimates of the expenditure elasticity of demand for cigarettes in Uganda for the 2005 and 2009 samples. Bootstrapped standard errors are in square brackets. 95% confidence intervals are in parentheses. Since the expenditure elasticity of demand is estimated at the household level (see equation 3.11), I only report the number of households.

The results in Table 3.5 show that cigarette demand in Uganda is expected to decline by about 0.3% every time cigarette prices rise by 1%. These estimates are statistically significant at the 1% level of significance. These estimates are within the range of estimates in the literature that uses Deaton's method discussed in Section 3.2.3. Table 3.6 presents results of the expenditure elasticity of demand for 2005 and 2009. Given that the expenditure elasticity estimates are not precisely estimated (i.e., standard errors are large), it is difficult to draw strong inferences. At the very least, the results in Table 3.6 suggest that cigarette demand does not decline with an increase in household expenditure.

3.6 Estimating elasticity when unit values are not available from HES

Deaton's approach allows us to estimate demand and compute 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, we cannot construct unit values whose spatial variation can be used as informative of 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 the 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) were often merged with household expenditure to obtain estimates of price elasticities.[44] Popular demand systems such as AIDS or Quadratic Almost Ideal Demand System (QAIDS) were 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.[45–48] Moreover, aggregate price indices are often highly correlated and may suffer from endogeneity problems.[49]

Recent literature,[50] however, suggests that construction of household level price indices (Stone-Lewbel (SL) prices[51]) 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 it allows for household level prices or unit values to be recovered.[50] It was 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.[48] The user-written program in Stata, <*pseudounit*>,[44] 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[50,52] 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.[53] This toolkit, however, does not go into these issues and the developments around it as, more often than not, HES data provides 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 does tobacco spending crowd out spending on other goods and services?

While global smoking prevalence has declined from 23.5% in 2007 to 20.7% in 2015, much of that decline has occurred in HICs while the decline has been the lowest in LICs.[54] The majority (around 77%) of the world's approximately 1.1 billion current smokers live in LMICs.[21] The prevalence of smokeless tobacco use is also found to be much higher in lower middle-income countries (14.6%) and LICs (11.2%) compared to the global prevalence (6.5%).[4] 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.[55]

Expenditure on tobacco accounts for a significant portion of the household budget in many countries, ranging from 1% in countries such as Mexico and Hong Kong to 10% in countries such as Zimbabwe and China.[56] Households operate based on limited disposable income and, as a result, when they spent their limited budgets on tobacco, it has a huge opportunity cost. It would inevitably mean they have to cut down expenditures on certain other goods and services, some of which may be necessary items of consumption such as food, clothing, and housing. The idea that households which 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[57] and China[58] 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 on from the US[59] and China[60] 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.[56] It used IV techniques to account for the possible endogeneity in the demand system while treating tobacco spending as a regressor and found that spending on tobacco crowded out food, education, and entertainment while crowding in expenditures on health, clothing, and fuels. Studies using similar econometric methods and household expenditure data were done in other countries such as Taiwan,[61] South Africa,[62] Cambodia,[63] Zambia,[64] Turkey,[65] and Bangladesh.[66] There were also other studies which examined crowding out in Indonesia[67] and other LMICs,[68] but with slightly different methods.

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.*[59]	US	Separate OLS regressions	Consumer Expenditure Survey	Clothing, housing
2006	Wang *et al.*[60]	China	Fractional Logit model	Primary Survey	Education, agriculture equipment maintenance, savings
2008	John, RM[56]	India	Instrumental variables	National Sample Survey	Food, education, entertainment
2008	Pu *et al.*[61]	Taiwan	Instrumental variables	Survey of Family Income & Expenditure	Clothing, medical care, transportation
2008	Koch & Tshiswaka-Kashalala[62]	South Africa	Instrumental variables	The South African Income and Expenditure Survey	Education, fuel, clothing, healthcare and transportation
2009	Block & Webb[67]	Indonesia	Reduced form equations	Nutrition surveillance system data	Food
2012	John *et al.*[63]	Cambodia	Instrumental variables	Cambodia Socio-Economic Survey	Food, education, clothing
2014	Chelwa & Walbeek[64]	Zambia	Instrumental variables	Living Conditions Monitoring Survey	Food, schooling, clothing, transportation, equipment maintenance
2015	San & Chaloupka[65]	Turkey	Instrumental variables	Turkish Household Budget Survey	Food, housing, education, durable/non-durable goods
2015	Do & Bautista[68]	40 LMICs	Random-slope models	World Health Survey	Education, healthcare
2018	Husain *et al.*[66]	Bangladesh	Instrumental variables	Household Income and Expenditure Survey	Clothing, housing, education, energy, transportation and communication
2018	Paraje & Araya	Chile	Quadratic AIDS model	Chilean Household Budget Survey (EPF)	Healthcare, education, housing

Table 4.1 above provides a summary of the different econometric studies that were 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 10 years. Most of these 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 inter-generational 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, 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—how it may impact social welfare is uncertain. As Deaton points out,[7] if women systematically get less than men, or if children and old people are systematically worse-off than other members of the households, 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,[54] if the allocation decisions are made by the male-heads in a household, it could potentially be unfavourable to women and/or children within a household. In fact, some of the findings from the crowding out literature described above underscore this. When school or 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 inter-generational impacts on society. The literature from India,[56] for example, showed tobacco spending households systematically allocate less money on clean cooking fuels and allocate more money on 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 the households pre-allocate a certain portion of the budget for purchase of tobacco. It 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. If it is found that compromises are made in the case of necessary commodities like food, education, and clothing, which may directly impact health and development of all members of a household, tobacco control policies should be able to address those.

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-spending households provides a preliminary indication on 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 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 labelled with the following commands:

```
gen tob=0
replace tob=1 if 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, healthcare, 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 makes further analysis simpler. For comparing the mean budget shares dedicated to these products between tobacco spenders and non-spenders a budget share variable is defined, one for each of this commodity group. Given the total expenditures on all items together as *exptotal*, the budget share on each of the variables 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 all these products.

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

A statistical test of the equality of mean budget shares between two groups (tobacco spenders and non-spenders) is a two-sample *Student’s t-test* for the equality of mean. The *t-test* which 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-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 the groups after weighting it 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 [bs_food]0 - [bs_food]1
```

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

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 *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/`nvar' {;
local X: word `i' of `items';
qui mean bs_`X' [pw=weight], over(tob);
matrix tmp=r(table);
matrix B[`i', 1] = tmp[1,1];
matrix B[`i', 2] = tmp[1,2];
qui lincom [bs_`X']0 - [bs_`X']1;
matrix B[`i', 3] = r(estimate);
matrix B[`i', 4] = r(t);
};
matrix rownames B = `items';
matrix colnames B = non-spenders spenders Difference t-stat;
matrix list B;
```

The code above will list a table with the budget shares for non-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 and by not controlling for the same, one may be inadvertently attributing allocation decisions to a household's tobacco spending habits. For this reason, there is a need for a formal econometric model which can explain whether households that spent on tobacco systematically cut down 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:

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

where q_i is the quantity of i^{th} good consumed, Y is total expenditures, h is a vector of characteristics and $p_1,\ldots,p_n$ are the prices of n commodities in an individual's utility function. Given that the 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).[8] The idea is that the household would maximize the following utility function:

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

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):

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

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 (e.g., tobacco) it is advantageous to use conditional demand functions as noted by Browning and Meghir.[69]

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.

4.4.2.1 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[70] 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.[56] The conditional Engel curve takes the following form for the good *i* and household *j:*

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

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, $\boldsymbol{h}_j$ is a vector of household characteristics allowing for the preferences to be heterogeneous,[71] *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.

4.4.2.2 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, i.e., $E(u/\boldsymbol{x}) = 0$, is violated and the OLS estimates fail to give causal interpretation. In such cases, if one can find exogenous variables which are correlated with these endogenous regressors, but are not correlated with the error term (IVs), one could use the IV method 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,[72] if the errors are homoskedastic. If 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.[73] 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 $\boldsymbol{x}$, i.e., $Cov(\boldsymbol{x}_i, \boldsymbol{z}_i) \neq 0$; and (2) Instrument $\boldsymbol{z}$ affects the dependent variable *wi* only through the regressors or $\boldsymbol{z}$ itself does not cause $\boldsymbol{w}_i$, i.e., $E(u/\boldsymbol{z})=0$. The first condition is sometimes called inclusion restriction, while the second condition is popularly known as exclusion restriction. While the inclusion restriction can be tested statistically by checking the association between an instrument ($\boldsymbol{z}$) and endogenous variables ($\boldsymbol{x}$) 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 (i.e., when the number of instruments equals the number of endogenous regressors). In the over-identified case (i.e., 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 Generalized Method of Moment (GMM).[15] This test again differs depending on whether the errors are homoskedastic. If the errors are homoskedastic, perform 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.[73]

Even if there are valid instruments and estimate-consistent coefficients, its covariance matrix can be inconsistent if the errors are heteroskedastic.[73] 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.[73] 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[14] refers to as a "system 2SLS estimator" and which is more efficient than the simple IV estimator[73] in the presence of heteroskedasticity.

4.4.2.3 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, there should be as many equations estimated as the number of considered commodity groups. 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.[74] 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.[15] 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[14] Chapter 8.

4.4.2.4 Estimation method 3: GMM 3SLS estimation

The traditional 3SLS estimator, according to Wooldridge,[14] is less efficient and its variance estimator is inappropriate if errors are heteroskedastic. In the cross-sectional surveys, 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[14] calls it the "GMM 3SLS" estimator. It extends the traditional 3SLS estimator by allowing for heteroskedasticity and different instruments for different equations.[75] 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 i.i.d. errors.[14] However, just like the IV/3SLS estimators, the GMM estimator, too, may have poor finite sample properties.[73]

According to Wooldridge,[14] the GMM 3SLS estimator using the heteroskedasticity consistent weighting matrix is never worse, asymptotically, than traditional 3SLS, and in some important cases is strictly better. The previous literature on crowding out, however, seems to have ignored a test of heteroskedasticity in the 3SLS model they have 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.

4.4.2.5 Testing heterogeneity in preferences between tobacco users and non-users

Typically, in the HES data, one would see 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 the constraints in their budget (also known as a *corner solution*), or because of abstention (i.e., 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, if one would also like to allow for heterogeneity in preferences between tobacco spending and non-spending households, the equation 4.4 can be augmented with the addition of a binary variable indicating tobacco consumption status as in some literature[56,65,76] as follows:

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

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

If the parameters associated with the binary variable *d* are not jointly significant i.e., if the null hypothesis $H_0{:}\alpha_{2i} = \beta_{2i} = \gamma_{2i} = 0$ fails to be rejected, one may conclude that those households, against whom zero expenditures on tobacco are currently reported, are not spending on tobacco probably because it is currently not affordable to them. In other words, both tobacco spenders and non-spenders have similar utility functions and tobacco non-spenders currently do not spend on tobacco only because its price is unaffordable. But, if the null is rejected, it means that the coefficients associated with tobacco dummy and that of expenditure variables where tobacco dummy is interacted with, are significant and that the preferences are indeed different for tobacco users and non-users. The literature on this uses a *Wald test* to test the joint significance of the three parameters after the regression.

If the 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. There is no reason to talk 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

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 which estimates crowding out ignoring such endogeneity,[59,60,67,77] 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. One may also compare absolute expenditures allotted to different items between both groups of households. Instead of a *t-test*, one could also perform other descriptive or graphical comparison tools to compare the averages.

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. On the other hand, less sophisticated tools like a *t-test*, discussed in Section 4.3, allows direct comparison of budget shares or expenditures between spenders and non-spenders for any disaggregated item. In fact, one can simply pick up items of interest only and compare the spending patterns between both groups. A study in India,[56] for example, used a simple *t-test* to compare budget shares allotted to school bus expenses and budget shares on different types of cooking fuel between tobacco spenders and non-spenders. It was found that tobacco spenders spent less on school bus expenses (implying that the young kids in the household are directly impacted). It also found tobacco spending households spent less on clean cooking fuel and spent more on unclean fuel like firewood (implying that the health of women in these households is likely affected).

4.5 Preparing data for analysis

While Chapter 2 provided detailed information on extracting data, cleaning it, merging variables that are 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 it through all 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 *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 data does not report this 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 socio-economic 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 socio-demographic 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 characterise 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[56,61,64–66] 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. One study[62] uses a composite smoking prevalence measure as an instrument for tobacco spending. In fact, 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 they do not have a direct effect on the dependent variable.

4.6 Estimating crowding out with Stata

This section will demonstrate the different estimation methods (traditional 3SLS, GMM 3SLS and an equation-by-equation IV) discussed in Section 4.4 to estimate the crowding out. First, it will discuss the general set-up 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 will be 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. First, 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 which go into the regression, *x1list* includes the RHS endogenous variables as explained in equation 4.4 (these are variables which are suspected as 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 square 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 less or a greater number of variables in any of the lists above. For example, the *x2list* may contain several other household-specific characteristics than are 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 on *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, one only need to use the following command to obtain the 3SLS estimates:

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

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 would 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
```

Remember, the code either has to be 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. However, the use of macros makes the code much neater. As previously noted, 3SLS is a GMM estimator that uses a particular weighting matrix which assumes i.i.d. errors. So, the above 3SLS results from <*reg3*> command can be reproduced with a GMM estimation with 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 *gmm* to assume that the residuals are independent across moment 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 that assumes conditional homoskedasticity, but that does not impose the cross-equation independence like the initial weight matrix is requested.[75] Please note that the <*gmm*> code above could take much longer—sometimes several hours depending on the physical capacity of the computer—than <*reg3*> to converge on a solution. This is because GMM, unlike 3SLS, is a very general and non-linear estimator and it searches numerically for a solution.

4.6.2 Estimation of GMM 3SLS

If the errors are heteroskedastic we know that 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 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. If one prefers to request a weight matrix that also accounts for arbitrary correlation among observations within clusters, as is usually observed in survey data, 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*>, one could also obtain bootstrapped standard errors if one were to use <*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 1000 bootstrap replications may take as much time as <*gmm*> to achieve convergence. <*gmm*>, on the other hand, has the added advantage of specifying a weighting matrix that accounts for hetroskedasticity 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*>[78] which 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[14] 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 heteroskedasicity after equation-by-equation IV indicates errors are not homoskedastic, then one can either use the system 2SLS estimator with <*gmm*> 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, for example, <*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 one does not need 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[72] 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,[73] 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 hetroskedasticity-consistent statistic 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 $\boldsymbol{z}$ 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. For example, 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.[15,79] 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.[15] However, there is no consensus on how low of a value of R^2 indicates a problem.[15] 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[15] Chapter 6 for a detailed exposition of these statistics. Alternatively, refer to the Stata reference manual[75] on *ivregress* post-estimation technical notes on page 1212-13.

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 overidentifying 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,[73] 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, 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*>[80] which can be installed with the command <*ssc install lmhreg3*> and which performs the tests of both single equation and overall system heteroskedasticy 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, satisfy 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: If one wants 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 (i.e., 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 determining 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, one must use a GMM 3SLS estimation method using the <*gmm*> command in Stata to produce efficient parameter estimates. According to Wooldridge,[14] the GMM 3SLS estimator using the heteroskedasticity-consistent weighting matrix is never worse, asymptotically, than traditional 3SLS, and in some important cases 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 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, we can also classify households 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, the researcher 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,[56,66] severity of tobacco spending,[63] and different types of tobacco.[66]

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 steps are shown below.

Step 1: Defining categorical variables to indicate subgroup

The following Stata code categorizes households into three income groups—low-, middle- and high-income—based on the distribution of per capita monthly expenditures for each household. This can be done by first creating a per capita expenditure variable (*pcexp*) by different households.

```
#delimit;
gen pcexp=exptotal/hsize;
_pctile pcexp, p(30, 70) ;
Local lower = `r(r1)’;
local upper = `r(r2)';
gen incgrp=0 ;
replace incgrp=1 if pcexp<= `lower';
replace incgrp=2 if pcexp> `lower' & pcexp< `upper';
replace incgrp=3 if pcexp>= `upper’;
label define incgrp 1 "Low income" 2 "Middle income" 3 "High income" ;
label values incgrp incgrp;
```

As indicated above, the code classifies those household above the 70th percentile of the distribution of per capita expenditures as high-income and those below the 30th percentile of the distribution as low-income while the ones in between are classified as middle-income. The code also assigns labels for each of the values the new variable *incgrp* takes. Similarly, one can also classify households 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, one can add the prefix <*bysort incgrp:*> before the command <gmm>.

Section 7.3 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 8% of their household budget on purchasing tobacco in 2011. While the rich in Turkey spent about 6.2% of the household budget on tobacco, the poor spent as high as 10.7%.[65] Given that a large portion of household budget is being diverted to tobacco spending, it is possible that expenditures on other household necessities are traded off. In this context, San & Chaloupka[65] examined the crowding out of tobacco spending on a variety of commodity groups in Turkey. The study estimated the QAIDS model with a variant of equation 4.5 to estimate the effects of crowding out. The econometric model used was the 3SLS method discussed in Section 4.4.2.3. The study used total expenditure to instrument for the expenditure net of tobacco and a women ratio to instrument for tobacco spending. Table 4.2 shows a snapshot from the results they found for 2011.

As indicated below, the authors estimated the model specified in equation 4.5 in this chapter with some variation in the control variables and used the traditional 3SLS estimation technique. Table 4.2 lists 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. The binary variable 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 *p.q* shows the total pre-allocated expenditures on tobacco and it gives 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.

Table 4.2 Crowding out impact 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

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]

Assume the monthly expenditures after spending on tobacco are about 1200 Lira (since 106 Lira spent on tobacco constituted about 8.17% of the budget). Then, using the parameter estimates presented by the authors, one can compute that a 100 Lira increase in the pre-allocated amount on tobacco leads to a 264 Lira decrease in housing expenses, while also redistributing expenditures on all the remaining commodities, increasing some and decreasing others. For example, a 100 Lira increase in the pre-allocated amount on tobacco would decrease expenditures on food, utilities, durables, clothing, health, and education by about 24, 12, 96, 36, 24 and 36 Lira, respectively, and increase expenditures on transport, entertainment, alcohol, and other commodities by 252, 204, 24 and 12 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.

Quantifying the impoverishing effect of tobacco use

5

5.1 Introduction

National poverty estimates are an important political variable in many countries. The estimate of the percentage of poor determines the course of development policy debates in several countries. Poverty reduction is a stated objective in many 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 the factors that hinder a nation's ability to achieve poverty reduction goals. This is because tobacco use and poverty are parts of a vicious circle.[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 exacerbating poverty. As the money spent on tobacco is highly unproductive and increases tobacco-related diseases, the resulting increased healthcare costs and loss of income due to premature deaths and morbidity can also add to the burden of poverty. Worldwide, around 80% 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.[81]

Chapter 4 explained 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 will show how to quantify the direct impact of tobacco spending on poverty measured by poverty head counts; discuss how tobacco spending contributes to impoverishment; and present methods to quantify the same. It will also demonstrate 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."[82] 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, the Statistics South Africa[83] 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.[84] The USCB’s 2017 definition shows that a single person under the age of 65 earning less than $12,752 per annum is considered to be living below the poverty line.

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.[85] The HCR, a counting measure, is defined as the fraction of the population living below the NPL and allows 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[86] and reproduced in the WHO/NCI Monograph.[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 on crowding out 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 purchase of tobacco.

2) **Channel 2: Forgone income from treating tobacco-related morbidity**
As tobacco consumption and exposure to SHS inevitably leads 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.[87] As a result of decreased spending on tobacco and, consequently, reduced healthcare spending, these households will have more disposable income to spend on essential needs (e.g., food, clothing, and education).

Although the literature examining the socio-economic 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[88] which quantified the impoverishing effect of out-of-pocket payments for healthcare. However, the first study which estimated the impoverishing effect of direct household spending on smoking and excess medical spending attributable to smoking was done in China.[89] It found that these two effects combined were responsible for impoverishing 30.5 million urban residents and 23.7 million rural residents in China. Another study from India[90] also found that the combined effect of these two factors resulted in impoverishment of 15 million people in India. A more recent study from the UK[91] subtracted only tobacco expenditures from household income to estimate its impact on poverty and found that over 432,000 children may be viewed as having been drawn into poverty by parental smoking. Yet another study from the UK[92] showed that when expenditure on tobacco is taken into account, around 500,000 extra households, comprising over 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.

These studies concluded that so many people who otherwise were above the NPL in these countries (i.e., in secondary poverty) were effectively in poverty because their disposable income after spending on tobacco and associated health expenditures was lower than that of people who were 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: (1) income forgone on account of the purchase of tobacco; and (2) income forgone due to tobacco use and SHS-attributable direct healthcare 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,[7] 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, it is a common practice to assume a uniform distribution within households when constructing the estimated distribution of individual consumptions.[93] Therefore, while estimating the HCR, it is important to use the 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.

First, 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 (i.e., 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:[85]

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

Where *I (.)* is an indicator function that takes value 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 value 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.

While it is more acceptable to assume a uniform distribution of consumption within households when constructing the estimated distribution of individual consumption,[93] 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[7] 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*." However, 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, consideration of such *adult equivalence* while examining the impoverishing effect of tobacco spending may not give the desired results.

5.4.2 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 expenditure, 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 value 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 healthcare 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 healthcare 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 healthcare expenditures.

While HES provides information on healthcare expenditures, they do not distinguish the amount of healthcare 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 healthcare that can be attributed to either tobacco use or SHS exposure. The attributable costs can be estimated either using a disease-specific approach or an inclusive or all-cause approach.[94] Since the HES often provide aggregate healthcare 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 healthcare 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 which is attributable to smoking by current and former smokers.[94] Similarly, SAF for SHS would be the fraction of healthcare expenditures that can be attributed to SHS.

Therefore, the attributable healthcare expenditures due to tobacco consumption and SHS, (i.e., *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 healthcare expenditures attributable to tobacco use and SHS, respectively. The SAF must be externally estimated using data from several different sources. It may be also be taken from available studies elsewhere in the country.

The SAF can be estimated either by using the epidemiological approach or an econometric approach.[95] 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 healthcare expenditures by type of healthcare services (such as inpatient hospitalizations and outpatient visits), and annual work-loss or disability days.”[95] 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.”[95] For these reasons, the epidemiological approach to estimating SAF is preferred in many LMICs. WHO provides a “toolkit” for estimating the economic costs of smoking and it provides a detailed explanation and methods for both the epidemiological and econometric methods of estimating SAF. Therefore, this toolkit will 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.

5.4.3 Poverty gap due to tobacco use

The incremental changes to the number of poor due to the successive subtraction of tobacco spending, and healthcare spending attributed to smoking and SHS exposure, may not be significant as many fall below the poverty line due to only tobacco spending, and become even poorer due to attributable healthcare spending. This is a matter of concern and is exactly the major flaw of a measure like HCR. In other words, HCR takes no account of the degree of poverty and would be unaffected if the poor became even poorer. One way to address this is using a measure called the "poverty gap" which assigns a larger weight to an individual in the aggregate poverty the poorer he or she is. The poverty gap can be computed using the formula:[7]

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

Deaton[7] notes that P_G can be interpreted as a per capita measure of the total shortfall of individual welfare below the poverty line as it is the sum of all the shortfalls divided by the population and expressed as a ratio of the poverty line itself. $P_G \times z \times N$ gives the total amount by which the poor are below the poverty line. Comparing the poverty gap before and after subtracting tobacco spending and other attributable healthcare expenditures, one can estimate the degree to which tobacco is impoverishing people in secondary poverty. However, this has also not been done in the previous literature on quantifying the impoverishing impact of tobacco use on poverty.

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 healthcare (*exphealth*) are required in order to compute healthcare 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 will have to be merged with the household survey data before the analysis can be done. To do so, a common identifying variable will have 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*. 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 (e.g., 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 healthcare 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 healthcare 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 healthcare 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 & SHS attributable healthcare 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 (e.g., *hweight*) should be adjusted to account for individual level estimation of poverty. This can be done by multiplying this variable with the household size, i.e., <*gen pweight=hweight*hsize*>. Once all 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*>[96] 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 captia 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 has 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), *pcet* (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 healthcare expenditures). The corresponding values under the column “Poor” show the estimated number of poor persons in each successive step. Comparing two successive rows enables one to see the change in both HCR and number of poor after the successive subtraction of each forgone income component. The number of poor 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 the HCR by the nationally available population data from other sources to arrive at the change in the number of poor.

The analysis above can be done with different subgroups as well using any of the methods discussed above. However, the data needs to be modified and new variables may have to be generated in order to do the analysis at the subgroup level. Section 7.4 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 will be able to estimate the results with the appropriate accompanying data/variables described therein.

5.7 Case study from India

In India during 2004-05, about 28.3% of the rural and 25.6% 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.[33] The poverty rate and trends over time have always taken center stage in Indian development policy discourse. In this context, John *et al*,[90] examined the impoverishing impact of tobacco spending as well as that of tobacco use-related healthcare 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 in India by rural and urban areas. It then shows the separate effect of subtracting tobacco spending and tobacco use-attributable healthcare 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 healthcare expenditures. In other words, spending on tobacco and the associated healthcare 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, yet enjoying lesser standards of living in terms of their ability to spend on daily necessities because their money is being diverted to wasteful expenditures on 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

BPL= Below Poverty Line. Source: John *et al.* (2011)[90]

6 Bibliography

1. World Health Organization. *Tobacco Control for Sustainable Development*. New Delhi, India: World Health Organization, Regional Office for South-East Asia; 2017. http://apps.who.int/iris/handle/10665/255509. Accessed October 05, 2018

2. World Health Organization. *WHO Global Report: Mortality Attributable to Tobacco*. Geneva, Switzerland; 2012. *http://apps.who.int/iris/bitstream/10665/44815/1/9789241564434_eng.pdf*. Accessed September 4, 2018.

3. Jha P, Peto R. Global Effects of Smoking, of Quitting, and of Taxing Tobacco. N Engl J Med. 2014;370(1):60-68. doi:10.1056/NEJMra1308383

4. U.S. National Cancer Institute and World Health Organization. *The Economics of Tobacco and Tobacco Control.* Bethesda, MD: U.S. Department of Health and Human Services, National Institutes of Health, National Cancer Institute; and Geneva, CH: World Health Organization; 2016. http://cancercontrol.cancer.gov/brp/tcrb/monographs/21/index.html. Accessed February 19, 2017. Accessed September 19, 2018

5. Goodchild M, Nargis N, d'Espaignet ET. Global economic cost of smoking-attributable diseases. *Tob Control.* January 2017:tobaccocontrol-2016-053305. doi:10.1136/tobaccocontrol-2016-053305

6. UN. *Transforming Our World: The 2030 Agenda for Sustainable Development.* New York, US: United National General Assembly; 2015. https://sustainabledevelopment.un.org/post2015/transformingourworld. Accessed September 19, 2018

7. Deaton AS. *The Analysis of Household Surveys.* Baltimore: Johns Hopkins University Press for the World Bank; 1997.

8. Pollak RA. Conditional Demand Functions and Consumption Theory. *Q J Econ.* 1969;83(1):60-78.

9. Pollak RA. Conditional Demand Functions and the Implications of Separable Utility. *South Econ J.* 1971;37(4):423-433.

10. Indian Statistical Institute. The National Sample Survey: General Report No. 1. First Round: October 1950 - March 1951. *Sankhy Indian J Stat 1933-1960.* 1953;13(1/2):47-214.

11. World Bank. Living Standards Measurement Study (LSMS). http://microdata.worldbank.org/index.php/catalog/lsms. Published 2018. Accessed September 2, 2018.

12. International Household Survey Network. IHSN Survey Catalog. http://catalog.ihsn.org/index.php/catalog/central. Published 2018. Accessed September 23, 2018.

13. LISGIS. *Household Income and Expenditure Survey 2016.* Monrovia, Liberia: Liberia Institute for Statistics and Geo-Information Services - Government of Liberia; 2017. http://catalog.ihsn.org/index.php/catalog/7279. Accessed September 11, 2018.

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

15. Cameron AC, Trivedi PK. *Microeconometrics Using Stata, Revised Edition.* 2nd ed. Texas, US: Stata Press; 2010. https://www.stata.com/bookstore/microeconometrics-stata/. Accessed October 14, 2018.

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

17. Baum CF. *A Little Bit of Stata Programming Goes a Long Way.* Boston, MA: Boston College Department of Economics; 2005. http://ideas.repec.org/e/pba1.html. Accessed June 10, 2018.

18. StataCorp. Stata programming reference manual. Release 15. 2017.

19. Chaloupka FJ, Warner KE. The Economics of Smoking. In: *The Handbook of Health Economics.* ; 2000:1539-1627.

20. IARC. *IARC Handbooks of Cancer Prevention in Tobacco Control, Volume 14: Effectiveness of Tax and Price Policies for Tobacco Control.* Lyon, France; 2011. http://www.iarc.fr/en/publications/pdfs-online/prev/handbook14/handbook14.pdf. Accessed June 2, 2015.

21. World Health Organization. *WHO Report on the Global Tobacco Epidemic, 2015: Raising Taxes on Tobacco.* Geneva, Switzerland; 2015. http://www.who.int/tobacco/global_report/2015/report/en/. Accessed Accessed September 12, 2018.

22. U.S. Department of Health and Human Services. *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General, 2014.* Rockville, Md: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2014. http://www.surgeongeneral.gov/library/reports/50-years-of-progress. Accessed September 12, 2018

23. Jha P, Chaloupka FJ. *Tobacco Control in Developing Countries.* Oxford, New York: Oxford University Press; 2000.

24. Keeler TE, Hu T-W, Barnett PG, Manning WG. Taxation, regulation, and addiction: A demand function for cigarettes based on time-series evidence. *J Health Econ.* 1993;12(1):1-18. doi:10.1016/0167-6296(93)90037-F

25. Hu TW, Bai J, Keeler TE, Barnett PG, Sung HY. The impact of California Proposition 99, a major anti-smoking law, on cigarette consumption. *J Public Health Policy*. 1994;15(1):26-36.

26. Hu TW, Sung HY, Keeler TE. Reducing cigarette consumption in California: tobacco taxes vs. an anti-smoking media campaign. *Am J Public Health.* 1995;85(9):1218-1222.

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

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

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

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

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

32. Deaton A, Grimard F. *Demand Analysis and Tax Reform in Pakistan*. World Bank; 1992. http://www.worldbank.org/html/prdph/lsms/research/wp/a81_100.html#wp85. Accessed September 12, 2018

33. John RM, Rao RK, Rao MG, et al. *The Economics of Tobacco and Tobacco Taxation in India.* Paris: International Union Against Tuberculosis and Lung Disease; 2010.

34. John RM. Consumption of Tobacco in India: An Economic Analysis. 2007.

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

36. Guindon GE, Nandi A, Chaloupka FJ, Jha P. Socioeconomic Differences in the Impact of Smoking Tobacco and Alcohol Prices on Smoking in India. *Natl Bur Econ Res Work Pap Ser.* 2011;No. 17580. http://www.nber.org/papers/w17580. Accessed September 10, 2018

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

38. Eozenou P, Fishburn B. *Price Elasticity Estimates for Cigarette Demand in Vietnam.* Development and Policies Research Center (DEPOCEN), Vietnam; 2009. https://ideas.repec.org/p/dpc/wpaper/0509.html. Accessed December 3, 2018.

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

40. Chelwa G. The economics of tobacco control in some African countries. 2015. https://open.uct.ac.za/handle/11427/16529. Accessed March 12, 2018.

41. 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.* 2016;40(4):222-228.

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

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

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

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

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

47. Slesnick DT. Prices and demand: New evidence from micro data. *Econ Lett.* 2005;89(3):269-274. doi:10.1016/j.econlet.2005.05.034

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

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

50. Castellón CE, Boonsaeng T, Carpio CE. Demand system estimation in the absence of price data: an application of Stone-Lewbel price indices. *Appl Econ.* 2015;47(6):553-568. doi:10.1080/00036846.2014.975332

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

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

53. Moro D, Castellari E, Sckokai P. Empirical issues in the computation of Stone–Lewbel price indexes in censored micro-level demand systems. *Appl Econ Lett.* 2018;25(8):557-561. doi:10.1080/13504851.2017.1346353

54. World Health Organization. *WHO Report on the Global Tobacco Epidemic, 2017: Monitoring Tobacco Use and Prevention Policies.* Geneva, Switzerland; 2017. http://apps.who.int/iris/bitstream/10665/255874/1/9789241512824-eng.pdf?ua=1. Accessed August 4, 2017.

55. World Health Organization. *Systematic Review of the Link between Tobacco and Poverty.* Geneva, Switzerland: World Health Organization; 2014. http://www.who.int/tobacco/publications/syst_rev_tobacco_poverty/en/index.html. Accessed June 20, 2018.

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

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

58. Thomson GW, Wilson NA, D Dea, Reid PJ, Chapman PH. Tobacco spending and children in low income households. *Tob Control.* 2002;11(4):372-375.

59. Busch SH, Jofre-Bonet M, Falba TA, Sindelar JL. Burning a Hole in the Budget: Tobacco Spending and its Crowd-Out of Other Goods. *Appl Health Econ Health Policy.* 2004;3(4):263-272.

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

61. Pu C, Lan V, Chou Y-J, Lan C. The crowding-out effects of tobacco and alcohol where expenditure shares are low: Analyzing expenditure data for Taiwan. *Soc Sci Med.* 2008;66(9):1979-1989. doi:10.1016/j.socscimed.2008.01.007

62. Koch SF, Tshiswaka-Kashalala G. *Tobacco Substitution and the Poor.* South Africa: Department of Economics, University of Pretoria; 2008. https://www.up.ac.za/media/shared/Legacy/UserFiles/wp_2008_32.pdf. Accessed October 14, 2018.

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

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

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

66. Husain MJ, Datta BK, Virk-Baker MK, Parascandola M, Khondker BH. The crowding-out effect of tobacco expenditure on household spending patterns in Bangladesh. *PLOS ONE*. 2018;13(10):e0205120. doi:10.1371/journal.pone.0205120

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

68. Do YK, Bautista MA. Tobacco use and household expenditures on food, education, and healthcare in low- and middle-income countries: a multilevel analysis. *BMC Public Health.* 2015;15. doi:10.1186/s12889-015-2423-9

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

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

71. Pollak RA, Wales TJ. *Demand System Specification and Estimation.* Oxford, New York: Oxford University Press; 1995.

72. Davidson R, MacKinnon JG. *Estimation and Inference in Econometrics.* New York; 1993.

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

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

75. StataCorp. Stata base reference manual. Release 15. 2017.

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

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

78. Baum 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.

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

80. Shehata EAE. LMHREG3: *Stata Module to Compute Overall System Heteroscedasticity Tests after (3SLS-SURE) Regressions.* Boston College Department of Economics; 2011. https://ideas.repec.org/c/boc/bocode/s457381.html. Accessed November 14, 2018.

81. Sreeramareddy CT, Harper S, Ernstsen L. Educational and wealth inequalities in tobacco use among men and women in 54 low-income and middle-income countries. *Tob Control.* 2018;27(1):26-34. doi:10.1136/tobaccocontrol-2016-053266

82. World Bank. WDI - Poverty and Inequality. http://datatopics.worldbank.org/world-development-indicators/themes/poverty-and-inequality.html#national-poverty-lines. Published 2018. Accessed November 1, 2018.

83. Statistics South Africa. *National Poverty Lines.* Pretoria, South Africa; 2018. http://www.statssa.gov.za/publications/P03101/P031012018.pdf. Accessed January 11, 2018.

84. US Census Bureau. Poverty. https://www.census.gov/topics/income-poverty/poverty.html. Published 2018. Accessed November 1, 2018.

85. Foster J, Seth S, Lokshin M, Sajaia Z. *A Unified Approach to Measuring Poverty and Inequality : Theory and Practice.* Washington, D.C.: The World Bank; 2013. http://documents.worldbank.org/curated/en/281001468323965733/A-unified-approach-to-measuring-poverty-and-inequality-theory-and-practice. Accessed November 2, 2018.

86. B. Seebohm Rowntree. *Poverty: A Study of Town Life.* London, UK: MacMillan; 1901.

87. Fuchs Tarlovsky A, Del Carmen G, Mukong AK. *Long-Run Impacts of Increasing Tobacco Taxes : Evidence from South Africa.* The World Bank; 2018:1-39. http://documents.worldbank.org/curated/en/122081521480061194/Long-run-impacts-of-increasing-tobacco-taxes-evidence-from-South-Africa. Accessed November 2, 2018.

88. Wagstaff A, Doorslaer E van. *Paying for Health Care : Quantifying Fairness, Catastrophe, and Impoverishment, with Applications to Vietnam, 1993-98.* The World Bank; 2001. http://ideas.repec.org/p/wbk/wbrwps/2715.html. Accessed October 21, 2018.

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

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

91. Belvin C, Britton J, Holmes J, Langley T. Parental smoking and child poverty in the UK: an analysis of national survey data. *BMC Public Health.* 2015;15(1):507. doi:10.1186/s12889-015-1797-z

92. Howard Reed. *Estimates of Poverty in the UK Adjusted for Expenditure on Tobacco.* London, UK: Action on Smoking and Health; 2015. http://ash.org.uk/information-and-resources/health-inequalities/health-inequalities-resources/estimates-of-poverty-in-the-uk-adjusted-for-expenditure-on-tobacco/. Accessed March 11, 2018.

93. Ravallion M. *Poverty Comparisons : A Guide to Concepts and Methods.* Washington DC: The World Bank; 1992:1. http://documents.worldbank.org/curated/en/290531468766493135/Poverty-comparisons-a-guide-to-concepts-and-methods. Accessed November 2, 2018.

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

95. World Health Organization. *Assessment of the Economic Costs of Smoking. World Health Organization Economics of Tobacco Toolkit.* Geneva, Switzerland: World Health Organization; 2011. http://whqlibdoc.who.int/publications/2011/9789241501576_eng.pdf. Accessed October 4, 2018.

96. Jenkins SP. *POVDECO: Stata Module to Calculate Poverty Indices with Decomposition by Subgroup.* Boston College Department of Economics; 2008. https://ideas.repec.org/c/boc/bocode/s366004.html. Accessed November 6, 2018.

Code Appendices

7

7.1 Stata do-file for estimating own-price elasticity using Deaton method for a single commodity

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

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

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

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

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

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

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

*Purged y's for next stage
gen y1cig=luvcig-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio ///
              -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu ///
              -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3

*Repeat for budget shares
areg bscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
predict rbscig, resid // residuals from the budget share regression

scalar sigma22=$S_E_sse/$S_E_tdf // var-covar matrix of u0 (e0e0)
scalar b0=_coef[lexp]       // Coefficients of lnexp in BS regression
*Purged y's for next stage
gen y0cig=bscig-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio ///
              -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu ///
              -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3

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

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


/*To estimate the bootstrap standard errors for expenditure elasticity
cap program drop Expelast
program Expelast, rclass
tempname b1 b0 Wbar
qui areg luvcig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
cap scalar b1=_coef[lexp]
qui areg bscig lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)
cap scalar b0=_coef[lexp]
qui sum bscig
cap scalar Wbar=r(mean)
return scalar Expel=1-b1+(b0/wbar)
end
expelast
return list

bootstrap Expel=r(Expel), reps(1000) seed(1): Expelast
*/

*Next, equations (3.4) and (3.5) are derived via the following sets of commands:
sort clust
egen y0c= mean(y0cig), by(clust)
egen n0c=count(y0cig), by(clust)
egen y1c= mean(y1cig), by(clust)
egen n1c=count(y1cig), by(clust)
sort clust
*keeping one obs per cluster
qui by clust: keep if _n==1

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

```
cap program drop elast
program elast, rclass
tempname S R num den phi theta psi
qui corr y0c y1c, cov
scalar S=r(Var_2)  //Var of y1
scalar R=r(cov_12) //Covariance y0c and y1c
scalar num=scalar(R)-(sigma12/n0)
scalar den=scalar(S)-(sigma11/n1)
cap scalar phi=num/den
cap scalar zeta= b1/((b0 + Wbar*(1-b1)))
cap scalar theta=phi/(1+(Wbar-phi)*zeta)
cap scalar psi=1-((b1*(Wbar-theta))/(b0+Wbar))
return scalar EP=(theta/Wbar)-psi
end
elast
return list
bootstrap EP=r(EP), reps(1000) seed(1): elast
log close
```

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

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

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

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

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

mkwbar $goods

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

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

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

drop luv`1'

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

matrix sig[`ig',1]=$S_E_sse/$S_E_tdf // var-covar matrix of u0 (e0e0)
matrix b0[`ig',1]=_coef[lexp]       // Coefficients of lnexp in BS regression
gen y0`1'=bs`1'-_coef[lexp]*lexp-_coef[lhsize]*lhsize-_coef[maleratio]*maleratio ///
                    -_coef[meanedu]*meanedu-_coef[maxedu]*maxedu ///
                    -_coef[sgp1]*sgp1-_coef[sgp2]*sgp2-_coef[sgp3]*sgp3
*This next regression is necessary to get covariance of residuals
qui areg ruv`1' rbs`1' lexp lhsize maleratio meanedu maxedu sgp1-sgp3, absorb(clust)

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

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

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

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

```
/*Removing region (province) effects
* This is optional and may or may not be used depending on the data
* This assumes the availability of the categorical variable region in the data
tab region, gen(regiond)
cap program drop purge
program define purge
local ig=1
while `ig' <= $ngds {
regress y0c`ig' regiond2 regiond3 regiond4
predict tm, resid
replace y0c`ig'=tm
drop tm
qui regress y1c`ig' regiond2 regiond3 regiond4
predict tm, resid
replace y1c`ig'=tm
drop tm
local ig=`ig'+1
}
end
purge
drop regiond*
*/

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

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

```
cap program drop mkcov
program def mkcov
local ir=1
while `ir' <= $ngds {
local ic=1
while `ic' <= $ngds {
qui corr y1c`ir' y1c`ic', cov
matrix s[`ir',`ic']=_result(4)
qui corr y1c`ir' y0c`ic', cov
matrix r[`ir',`ic']=_result(4)
local ic=`ic'+1
}
local ir=`ir'+1
}
end
mkcov
*We don't need the data any more
drop _all
matrix list s  // between-cluster var-covar matrix of y1 [cov(y1Gc,y1Hc)]
matrix list r  // between-cluster covar matrix of y1 [cov(y1Gc,y0Hc)]
*Making OLS estimates
matrix bols=syminv(s)
matrix bols=bols*r
display("Second-stage OLS estimates: B-matrix") // eqn 5.84
matrix list bols
display("Column 1 is coefficients from 1st regression, etc")
*Corrections for measurement error
cap program drop fixmat
program def fixmat
matrix def sf=s
matrix def rf=r
local ig=1
while `ig' <= $ngds {
matrix sf[`ig',`ig']=sf[`ig',`ig']-ome[`ig',1]/n1[`ig',1]
matrix rf[`ig',`ig']=rf[`ig',`ig']-lam[`ig',1]/n0[`ig',1]
local ig=`ig'+1
}
end
fixmat
matrix invs=syminv(sf)
matrix bhat=invs*rf   // The errors-in-variable estimator with ME correction Eqn 5.85
*Estimated B matrix without restrictions
matrix list bhat // The errors-in-variable estimator with ME correction).
*The ratio Phi from which Psi and Theta matrices has to be disentangled.
*Housekeeping matrices, including elasticities
cap program drop mormat
program def mormat
```

```
matrix def xi=J($ngds,1,0) // Xi vector in Eqn 5.92
matrix def el=J($ngds,1,0) // Expenditure elasticity matrix in Eqn 5.89 or 5.50
local ig=1
while `ig' <= $ngds {
matrix xi[`ig',1]=b1[`ig',1]/(b0[`ig',1]+ ///
((1-b1[`ig',1])*wbar[`ig',1]))
matrix el[`ig',1]=1-b1[`ig',1]+b0[`ig',1]/wbar[`ig',1]
local ig=`ig'+1
}
end
mormat
global ng1=$ngds+1
matrix iden=I($ngds)
matrix iden1=I($ng1)
matrix itm=J($ngds,1,1)
matrix itm1=J($ng1,1,1)
matrix dxi=diag(xi)
matrix dwbar=diag(wbar)
matrix idwbar=syminv(dwbar)
display("Average budget shares")
matrix tm=wbar'
matrix list tm   // Average budget shares
display("Expenditure elasticities")
matrix tm=el'    // Expenditure elasticities (dlnq/dlnx)
matrix list tm
display("Quality elasticities")
matrix tm=b1'
matrix list tm   // Expenditure elasticity of quality (dlnuv/dlnx)

*This all has to go in a program to use it again later
*Basically uses the b from eqn 5.85 matrix to form price elasticity matrix
cap program drop mkels
program define mkels
matrix cmx=bhat'
matrix cmx=dxi*cmx
matrix cmx1=dxi*dwbar
matrix cmx=iden-cmx
matrix cmx=cmx+cmx1
matrix psi=inv(cmx)
matrix theta=bhat'*psi // Theta matrix in Eqn 5.90
display("Theta matrix")
matrix list theta  // Theta matrix in Eqn 5.90
matrix ep=bhat'
matrix ep=idwbar*ep
matrix ep=ep-iden
matrix ep=ep*psi
display("Matrix of price elasticities")
```

```
matrix list ep  // price elasticity of demand without symmetry restrictions)
end
mkels
**************************************************************
*If program is executed only up to this point and with a single commoditiy
*by specifying ngds=1 and retain only one good in global macro this will return
*the same estimate of price elasticity derived from code in chapter 3 of this
*tool kit. The code below completes the system of demand equation by filling out
*the matrices. This essentially adds a single composite commodity to the
*equation to complete the system using homogeneity and adding-up restrictions.
***************************************************************
cap program drop complet
program define complet
*First extending theta
matrix atm=theta*itm
matrix atm=-1*atm
matrix atm=atm-b0
matrix xtheta=theta,atm
matrix atm=xtheta'
matrix atm=atm*itm
matrix atm=atm'
matrix xtheta=xtheta\atm
*Extending the diagonal matrices
matrix wlast=wbar'*itm
matrix won=(1)
matrix wlast=won-wlast
matrix xwbar=wbar\wlast
matrix dxwbar=diag(xwbar)
matrix idxwbar=syminv(dxwbar)
matrix b1last=(0.25)
matrix xb1=b1\b1last
matrix b0last=b0'*itm
matrix b0last=-1*b0last
matrix xb0=b0\b0last
matrix xe=itm1-xb1
matrix tm=idxwbar*xb0
matrix xe=xe+tm
matrix tm=xe'
display("extended outlay elasticities (or total expenditure elasticities)")
matrix list tm  // expenditure elasticities from the complete system
matrix xxi=itm1-xb1
matrix xxi=dxwbar*xxi
matrix xxi=xxi+xb0
matrix tm=diag(xb1)
matrix tm=syminv(tm)
matrix xxi=tm*xxi
matrix dxxi=diag(xxi)
```

```
*Extending psi
matrix xpsi=dxxi*xtheta
matrix xpsi=xpsi+iden1
matrix atm=dxxi*dxwbar
matrix atm=atm+iden1
matrix atm=syminv(atm)
matrix xpsi=atm*xpsi
matrix ixpsi=inv(xpsi)
*Extending bhat & elasticity matrix
matrix xbhatp=xtheta*ixpsi
matrix xep=idxwbar*xbhatp
matrix xep=xep-iden1
matrix xep=xep*xpsi
display("extended matrix of elasticities")
matrix list xep   // price elasticities from the complete system without symmetry
end
complet // this command can be dropped if we are only interested in
*symmetry constrained estimates as given below. If it is only the unconstrainted
*estimates that we are intereted in there is no need to run rest of the code too
************************************************************
**Calculating symmetry restricted estimators
**These are only approximately valid & assume no quality effects
*the do-file mkmats.do should be executed for this
run mkmats.do
vecmx bhat vbhat
** R matrix for restrictions
lmx $ngds llx
commx $ngds k
global ng2=$ngds*$ngds
matrix bigi=I($ng2)
matrix k=bigi-k
matrix r=llx*k
matrix drop k
matrix drop bigi
matrix drop llx
** r vector for restrictions, called rh
matrix rh=b0#wbar
matrix rh=r*rh
matrix rh=-1*rh
**doing the constrained estimation
matrix iss=iden#invs
matrix rp=r'
matrix iss=iss*rp
matrix inn=r*iss
matrix inn=syminv(inn)
matrix inn=iss*inn
matrix dis=r*vbhat
matrix dis=rh-dis
matrix dis=inn*dis
matrix vbtild=vbhat+dis
unvecmx vbtild btild
**the following matrix should be symmetric
matrix atm=b0'
matrix atm=wbar*atm  // Eqn. 5.98
matrix atm=btild+atm
matrix list atm
**going back to get elasticities and complete sytem
matrix bhat=btild
mkels
complet
*The program will display the own and cross-price elasticities for the two
*googs cigarette and beer along with that of the composite commodity used for
*completing the system
log close

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

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


*program for calculating the matrix that extracts
*from vec(A) the lower left triangle of the matrix A
cap program drop lmx
program define lmx
local ng2=`1'^2
local nr=0.5*`1'*(`1'-1)
matrix def `2'=J(`nr',`ng2',0)
local ia=2
local ij=1
while `ij' <= `nr'{
local ik=0
local klim=`1'-`ia'
while `ik' <= `klim' {
local ip=`ia'+(`ia'-2)*`1'+`ik'
matrix `2'[`ij',`ip']=1
local ij=`ij'+1
local ik=`ik'+1
}
local ia=`ia'+1
}
end

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

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

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

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

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

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

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

```
matrix sf[`ig',`ig']=sf[`ig',`ig']-ome[`ig',1]/n1[`ig',1]
matrix rf[`ig',`ig']=rf[`ig',`ig']-lam[`ig',1]/n0[`ig',1]
local ig=`ig'+1
}
end
cap program drop mkels
program define mkels
matrix cmx=bhat'
matrix cmx=dxi*cmx
matrix cmx1=dxi*dwbar
matrix cmx=iden-cmx
matrix cmx=cmx+cmx1
matrix psi=inv(cmx)
matrix theta=bhat'*psi
display("Theta matrix")
matrix list theta
matrix ep=bhat'
matrix ep=idwbar*ep
matrix ep=ep-iden
matrix ep=ep*psi
end
cap program drop complet
program define complet
*First extending theta
matrix atm=theta*itm
matrix atm=-1*atm
matrix atm=atm-b0
matrix xtheta=theta,atm
matrix atm=xtheta'
matrix atm=atm*itm
matrix atm=atm'
matrix xtheta=xtheta\atm
*Extending the diagonal matrices
matrix wlast=wbar'*itm
matrix won=(1)
matrix wlast=won-wlast
matrix xwbar=wbar\wlast
matrix dxwbar=diag(xwbar)
matrix idxwbar=syminv(dxwbar)
matrix b1last=(0.25)
matrix xb1=b1\b1last
matrix b0last=b0'*itm
matrix b0last=-1*b0last
matrix xb0=b0\b0last
matrix xe=itm1-xb1
matrix tm=idxwbar*xb0
matrix xe=xe+tm
matrix tm=xe'
matrix xxi=itm1-xb1
matrix xxi=dxwbar*xxi
matrix xxi=xxi+xb0
matrix tm=diag(xb1)
matrix tm=syminv(tm)
matrix xxi=tm*xxi
matrix dxxi=diag(xxi)
*Extending psi
matrix xpsi=dxxi*xtheta
matrix xpsi=xpsi+iden1
matrix atm=dxxi*dxwbar
matrix atm=atm+iden1
matrix atm=syminv(atm)
matrix xpsi=atm*xpsi
matrix ixpsi=inv(xpsi)
*Extending bhat & elasticity matrix
matrix xbhatp=xtheta*ixpsi
matrix xep=idxwbar*xbhatp
matrix xep=xep-iden1
matrix xep=xep*xpsi
end

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

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

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

```
mkcov
*We don't need the data any more
drop _all
*Making OLS estimates
matrix bols=syminv(s)
matrix bols=bols*r
*Corrections for measurement error
fixmat
matrix invs=syminv(sf)
matrix bhat=invs*rf
global ng1=$ngds+1
matrix iden=I($ngds)
matrix iden1=I($ng1)
matrix itm=J($ngds,1,1)
matrix itm1=J($ng1,1,1)
matrix dxi=diag(xi)
matrix dwbar=diag(wbar)
matrix idwbar=syminv(dwbar)
mkels
**Completing the system by filling out the matrices
** Gives standard errors for elasticities without symmetry restrictions
complet //Drop this command if the intend is to estimate symmetry constrained standard errors
*If it is only the unconstrainted estimates that we are intereted in there is
*no need to run the code from this point till the next command complet
**Calculating symmetry restricted estimators
vecmx bhat vbhat
** R matrix for restrictions
lmx $ngds llx
commx $ngds k
global ng2=$ngds*$ngds
matrix bigi=I($ng2)
matrix k=bigi-k
matrix r=llx*k
matrix drop k
matrix drop bigi
matrix drop llx
** r vector for restrictions, called rh
matrix rh=b0#wbar
matrix rh=r*rh
matrix rh=-1*rh
**doing the constrained estimation
matrix iss=iden#invs
matrix rp=r'
matrix iss=iss*rp
matrix inn=r*iss
matrix inn=syminv(inn)
matrix inn=iss*inn
matrix dis=r*vbhat
matrix dis=rh-dis
matrix dis=inn*dis
matrix vbtild=vbhat+dis
unvecmx vbtild btild
**going back to get elasticities and complete sytem
matrix bhat=btild
mkels
** Gives standard errors for elasticity with symmetry restrictions
complet

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

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

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

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

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

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

*generating budget share variables for t-test of comparison
*here the denominator is the total expenditures on all goods combined
local items "tobac food health educn housing cloths entertmnt transport durable other"
foreach X of local items{
	gen bs_`X'=(exp`X'/exptotal)
}
*t-test using survey weights
local items tobac food health educn housing cloths entertmnt transport durable other
local nvar: word count `items'
matrix B = J(`nvar', 4, .)
forvalues i = 1/`nvar' {
   local X: word `i' of `items'
   qui mean bs_`X' [pw=weight], over(tob)
         matrix tmp=r(table)
         matrix B[`i', 1] = tmp[1,1]
         matrix B[`i', 2] = tmp[1,2]
         qui lincom [bs_`X']0 - [bs_`X']1
         matrix B[`i', 3] = r(estimate)
   matrix B[`i', 4] = r(t)
}
matrix rownames B =`items'
matrix colnames B = non-spenders spenders Difference t-stat
matrix list B
*dropping this budget share variables
drop bs_*

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

*creating budget shares for crowding out analysis. Here the denominator is the
*total expendituer minus the expenditures on tobacco
gen exp_less=exptotal-exptobac
local items "food health educn housing cloths entertmnt transport durable other"
foreach X of local items{
         gen bs`X'=(exp`X'/exp_less)
}

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

*Estimating Crowding out with different models
global ylist bsfood bshealth bseducn bshousing bscloths bsentertmnt bstransport bsdurable
global x1list pq lnM lnM2
global x2list hsize meanedu maxedu sd1-sd3
global zlist asexratio lnX lnX2

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

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

*Traditional 3SLS using GMM
gmm (eq1: bsfood - {food: $x1list $x2list _cons}) ///
        (eq2: bshealth - {health: $x1list $x2list _cons}) ///
        (eq3: bseducn - {educn: $x1list $x2list _cons}) ///
        (eq4: bshousing - {housing: $x1list $x2list _cons}) ///
        (eq5: bscloths - {cloths: $x1list $x2list _cons}) ///
        (eq6: bsentertmnt - {entertmnt: $x1list $x2list _cons}) ///
        (eq7: bstransport  - {transport: $x1list $x2list _cons}) ///
        (eq8: bsdurable - {durable: $x1list $x2list _cons}) ///
        , instruments($zlist $x2list) ///
        winitial(unadjusted, independent)  wmatrix(unadjusted) twostep

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

**********************************************
*GMM 3SLS estimation (wooldridge): adjusts for heteroskedasticity
**********************************************
gmm (eq1: bsfood - {food: $x1list $x2list _cons}) ///
        (eq2: bshealth - {health: $x1list $x2list _cons}) ///
        (eq3: bseducn - {educn: $x1list $x2list _cons}) ///
        (eq4: bshousing - {housing: $x1list $x2list _cons}) ///
        (eq5: bscloths - {cloths: $x1list $x2list _cons}) ///
        (eq6: bsentertmnt - {entertmnt: $x1list $x2list _cons}) ///
        (eq7: bstransport  - {transport: $x1list $x2list _cons}) ///
        (eq8: bsdurable - {durable: $x1list $x2list _cons}) ///
        , instruments($zlist $x2list) ///
        winitial(unadjusted, independent)  wmatrix(robust) twostep

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

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

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

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

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

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

*Using System 2SLS estimator (equation by equation IV)
gmm (eq1: bsfood - {food: $x1list $x2list _cons}) ///
	(eq2: bshealth - {health: $x1list $x2list _cons}) ///
	(eq3: bseducn - {educn: $x1list $x2list _cons}) ///
	(eq4: bshousing - {housing: $x1list $x2list _cons}) ///
	(eq5: bscloths - {cloths: $x1list $x2list _cons}) ///
	(eq6: bsentertmnt - {entertmnt: $x1list $x2list _cons}) ///
	(eq7: bstransport  - {transport: $x1list $x2list _cons}) ///
	(eq8: bsdurable - {durable: $x1list $x2list _cons}) ///
	, instruments($zlist $x2list) ///
	winitial(unadjusted, independent)

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

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

local depvar "food health educn housing cloths entertmnt transport durable"
foreach X of local depvar{
        ivregress gmm bs`X' $x2list ($x1list = $zlist), wmatrix(cluster clustvar)
}

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

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

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

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

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

*These tests can also be done in a loop for all commodities together as follows:
local depvar "food health educn housing cloths entertmnt transport durable"
foreach X of local depvar{
        ivregress 2sls bs`X' $x2list ($x1list = $zlist)
        estat endogenous
        ivregress 2sls bs`X' $x2list ($x1list = $zlist), vce(robust)
        estat endogenous
}
*with ivreg2, however, do the tests along with the regression itself
*with the option endogtest() as follows
ivreg2 bsfood $x2list ($x1list = $zlist), endogtest($x1list)

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

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

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

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

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

**Testing exclusion restriction. (instrument exogeneity)

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

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

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

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

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

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

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

*******************************************************************
*Analysis by different sub group
******************************************************************
*generate indicator variable for different income groups
*First generate percapita expenditues and then generate the variable
gen pcexp=exptotal/hsize
_pctile pcexp, p(30, 70)
local lower = `r(r1)'
local upper = `r(r2)'
gen incgrp=0
replace incgrp=1 if pcexp<=`lower'
replace incgrp=2 if pcexp>`lower' & pcexp<`upper'
replace incgrp=3 if pcexp>=`upper'
label define incgrp 1 "Low income" 2 "Middle income" 3 "High income"
label values incgrp incgrp

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

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

7.4 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 healthcare expenditures  in LCU
* - hsize - household size
* - hweight - survey weights
* - npl - National poverty line in local currency units
*===============================================================================

clear
version 15
set mem 1000m
set more off

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

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

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

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

```
*preparing variables for analysis
ren pcetotal pce
gen pcet=pce-pcetobac
label var pcet "pce-expenditure on tobacco"
gen pceh=pcet-pcehealth
label var pceh "pct-tobacco attributable health care exp."
gen pweight=hweight*hsize


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

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


*Code for computing changes in HCR and number of poor in one shot
local subtr pce pcet pceh
local nvar: word count `subtr'
matrix M = J(`nvar', 2, .)
forvalues i = 1/`nvar' {
   local X: word `i' of `subtr'
   qui gen ind = (`X'<=npl)
   qui sum ind [fw=pweight]
   matrix M[`i', 1] = r(mean)
   matrix M[`i', 2] = r(sum)
   drop ind
}
matrix rownames M = `subtr'
matrix colnames M = HCR Poor
*the following lists the results with special formating options
matlist M, cspec(& %12s | %5.4f & %9.0f &) rspec(--&&-)

log close
```

tobacconomics
Economic Research Informing Tobacco Control Policy

INSTITUTE FOR
HEALTH RESEARCH
AND POLICY
UIC

www.tobacconomics.org
@tobacconomics