Researching with Secondary Data: A brief overview of possibilities and limitations from the viewpoint of social research

Kumara, Ajantha Sisira

University of Sri Jayewardenepura

April 2022

Researching with Secondary Data: *A brief overview of possibilities and limitations from the viewpoint of social research*

Ajantha Sisira Kumara, *Ph.D.*
Professor in Public Economics, Department of Public Administration, Faculty of Management Studies and Commerce, University of Sri Jayewardenepura, Nugegoda, Sri Lanka
mhasisira@sjp.ac.lk

Abstract

The objective of this paper is to provide information on reliable sources of secondary data available for applied social researchers, feasible studies, and Econometric Modelling techniques that can be applied in those studies. Using secondary data for research projects has now become popular with the availability of nation-wide survey data from reliable sources. The Econometrics methods ranging from multiple regression analysis to dynamic panel analysis can be applied with secondary data to generate nationally representative empirical evidence on the subject of interest. However, the researchers need to be careful of reading secondary data using documentation provided along with datasets. There are limitations and challenges of using secondary data for researching as well.

Keywords: Applied Research; Econometric Modelling; Secondary Data

1. Introduction

High-quality research can also be conducted using secondary data, collected by someone other than the user. The secondary data are from a wide range of sources: censuses, information collected by government departments, organizational records, databases maintained by universities and other research institutions, surveys conducted by universities and research institutions, and so forth. There are advantages of using secondary data for researching. First, any researcher using secondary data can enjoy much information that has been collected in the past, and relevant policy variables can easily be generated by using them. Second, the researcher does not have to wait for a longer time for collecting data, and thereby, the research can be conducted in a timely manner. Accordingly, the researcher may be able to skip the stage of 'data collection' which allows him/her to proceed directly to the stage of 'data analysis'. Third, generally, the secondary data are available for a larger sample size, and the weights or inflation factors are also provided along with datasets. Thus, the researcher may apply statistical techniques to generate weighted-estimates which represent an entire country or an entire sector.

2. Sources of secondary data, Feasible studies, and Research methods

2.1.Department of Census and Statistics, Sri Lanka

The Department of Census and Statistics-Sri Lanka conducts a number of nation-wide surveys. Among them, the household income and expenditure survey-Sri Lanka (The HIES-Sri Lanka) is well-recognized. The HIES-Sri Lanka is conducted in every three-year time period by the Department of Census and Statistics-Sri Lanka to provide information on household income, expenditure, income inequalities, and poverty status. The microdata of the survey are available from 2002 to 2019, and the data can be obtained by researchers by presenting a research proposal. The survey uses two-stage random stratified sampling method for urban, rural, and estate sectors in each district of the country, and therefore, district is used as the main domain for stratification. Initially, 2,500 primary sampling units (census blocks) are selected from the sampling frame by assigning different sample size allocations for each selection domain. The method of selection of primary sampling units is systematic, and the selection probability for each census block is determined by taking the number of housing units available in those census blocks. Secondly, final sampling units (housing units) are selected from 2,500 primary sampling units. Accordingly, from each primary sampling unit, 10 final housing units are systematically selected to create a total sample of 25,000 housing units. The relevant weights (inflation factors) are provided along with datasets, allowing researchers to calculate weighted-estimates, and thus, final results represent the entire Sri Lanka.

The survey questionnaire consists of an array of modules to facilitate a comprehensive data collection process. Initially, it includes a module for demographic characteristics of household members including age, gender, relationship to household head, marital status, ethnicity, and religion. Then, the survey uses two modules for collecting information on household members' education and health. Under education, the survey collects information on the highest educational attainment of members, the type of school or university they are studying or they have studied, the distance from home to the school, travelling mode, and so forth. The health module collects the information on members' health status, the type of treatment they are receiving, communicable or non-communicable illnesses that they are suffering from, the nature of healthcare utilization of members, and their perceived impact of health status on workforce behaviour. The survey provides detailed-data on household-level expenditure on food and non-food categories. Moreover, the survey includes additional modules to provide information at the household-level on waste management, housing conditions, land-ownership, and indebtedness.

As the survey uses different cross-sections of households for different survey periods, the data are not panel in nature but cross-sectional-pooled data. Therefore, the data are more appropriate to conduct cross-sectional studies, and the dynamics can be examined by comparing results across surveys or by including a time-dummy in the pooled-survey data. For instance, Kumara and Samaratunge (2016) examine the patterns and determinants of the likelihood and financial burden of encountering out-of-pocket healthcare expenditure in Sri Lankan households. When considering the household-level out-of-pocket healthcare expenditure data, it can be observed that there are many zeros (0s) under household healthcare expenses category. The reasons might be either those households did not require healthcare facilities within the given reference period or they require such services, but their budget constraint did not allow those households to spend out-of-pocket for healthcare facilities. When modelling this type of a situation, we definitely need to use a two-stage decision making process where at the first stage, the households decide whether to spend out-of-pocket for healthcare (binary choice) and then, they decide how much to spend given the fact that they choose to spend out-of-pocket for healthcare. There are many Econometric models being applied in modelling this type of an environment including, Tobit model and Double-Hurdle model. The paper applies the Tobit model for HIES 2006/2007 and HIES 2009/2010 to examine determinants of household out-of-pocket healthcare expenditure.

$$\Pr obit : y_i = \begin{cases} 0, \ if \ \ y_i^* \leq 0 \\ 1, \ Otherwise \end{cases}$$

$$y_i^* = \beta^{/} x_i + \mu_i \text{ : Latent variable not observed}$$
$$\mu_{\mathrm{i}} \sim N(0, \sigma^2)$$

$$Tobit \text{ model with } \mathrm{y}_{\mathrm{j}} \text{ censored at zero} : y_j = \begin{cases} 0, \ if \ y_i^* \leq 0 \\ \beta^{/} x_i + \mu_i, \ Otherwise \end{cases}$$

The study concludes that household composition, head's level of education, living sector, health status of members, and living standard of households play a key role in determining household-level out-of-pocket healthcare expenditure. Also, the study further reveals that health-related supply-side factors including distance from home to the nearest public and private hospital, bed capacity of public hospitals, number of doctors per 100,000 population, and number of dentists per 100,000 population play a role in determining the same. In this regard, I should note that the HIES does not provide data on healthcare-related supply-side factors. However, the data on healthcare-related supply-side factors including bed capacity of public hospitals, number of specialists, physicians, surgeons, and dentists per 100,000 population are available from the annual reports of Ministry of health-Sri Lanka. Those data are available at the district and divisional levels, and researchers can combine those data collected from other secondary sources with the HIES main dataset using district or division as the 'merging base'.

Another advantage of using HIES data for modelling household behaviour is that the researchers are able to correctly measure the impact of household living standard on household decision making process. Generally, household living standard is measured by household income but, household income data in developing countries are not accurate enough. Therefore, in developing country-context, it is always better for researchers to take per-capita-per-month household expenditure as the living standard measure (Deaton, 2006).

Using nation-wide microdata like HIES-data allows researchers to account for household composition and economies of scale enjoyed by larger households when calculating living standard measure. For instance, researchers are able to calculate Adult-

Equivalent-Scaled (AES) per-capita household expenditure instead of just household per-capita expenditure. The level of consumption of household members are different depending on their age; generally, adults consume more than children. Also, within the same household, there may be many public goods which are non-excludable and non-rival in consumption, and therefore, larger households can enjoy economies of scale. Accordingly, AES-per-capita household expenditure can be calculated by dividing household's total monthly expenditure by the following scale:

$$AES = [1 + \beta(a - 1) + \delta c]^{\theta},$$

where a is total number of adults while c is total number of children living in each household.

The OECD living standard scale can be obtained by setting $\beta = 0.7 \; and \; \delta = 0.5$. Also, if 0<Ɵ<1, economies of scale are not at their maximum but they are admitted up to a certain degree. Higher Ɵ will indicate less economies of scale (as Ɵ =1 means no economies of scale). From an empirical point of view, it is good practice to test the results comparing different values of Ɵ. In applied works, generally Ɵ is set between 0.65 and 0.75.

Moreover, the HIES-Sri Lanka enables researchers to establish complex categorical dependent variable models or limited-dependent variable models. For instance, Kumara and Pallegedara (2020) use HIES 2006/2007, HIES 2012/2013, and HIES 2016 for a multinomial logit regression model to examine patterns and determinants of household waste disposal mechanisms in Sri Lanka. The HIES consists of a separate module for household-level waste management, and accordingly, there are five mechanisms being used for disposing household waste: local government collection, burning, dumping within the premises, dumping outside the premises, and composting. Thus, the dependent variable of the regression model is categorical with five categories, and multinomial logit model is recognized to be the most appropriate way of modelling the situation as specified below:

$$j = 1,2,3,4,5$$
$$\textit{Taking 5 as the base} - \textit{category}$$

$$\Pr(y_i = j \mid X_i) = \exp(X_i^{/}\beta_j)/[1 + \sum_{j=1}^{4}\exp(X_i^{/}\beta_j)]$$
$$\Pr(y_i = 5 \mid X_i) = 1/[1 + \sum_{j=1}^{4}\exp(X_i^{/}\beta_j)]$$
$$\textit{such that}$$
$$\sum_{j=1}^{5}\Pr(y_i = j \mid X_i) = 1$$

The comprehensiveness of HIES-dataset allows the researchers to evaluate the role of an array of covariates including, household head's characteristics, household characteristics, housing characteristics, and living standard in determining the probability of choosing each mechanism of waste disposal. The larger sample of HIES enables researchers to provide provincial-level results with regard to the subject.

Working with secondary datasets may sometimes be problematic due to frequently occurring 'endogeneity biases' in regression models. As you already know, in regression analysis, the endogeneity biases can be possible due to three reasons including, omitted variables, measurement errors, and reverse-causality. However, luckily, the solutions for the issue are also feasible from the same dataset. The best remedy to address the issue of endogeneity is using an instrumental variable or a set of such instrumental variables for the problematic independent variable. Then, the problem arises for researchers as to what variable to be used as a qualified instrument. In order to be a qualified instrument, the proposed instrumental variable should be strongly correlated with the problematic endogenous regressor and it should not be correlated with the error term of the secondly estimated regression model. However, nation-wide household datasets like HIES provide ample opportunities for researchers to find out an appropriate instrumental variable. For instance, Pallegedara and Kumara (2021) examine the nexus between firewood burning for cooking at the household-level and members' respiratory health and healthcare utilization.

$$\Pr(y_i = 1 \mid d_i, X_i) = f(\beta_1 d_i + \beta_2 X_i + \varepsilon_i)$$

The dummy variable of interest (di) indicating whether a household uses firewood for cooking or not may be endogenous to respiratory health (asthma prevalence) and health care utilisation of households. In this research design, endogeneity may be a possibility first due to omitted variables or unobserved heterogeneity. Second, endogeneity may arise in this setup due to simultaneity which means that health variables (dependent variables) and household energy choice (independent variable of interest) are codetermined with each influencing the other or reverse causality. As an instrumental variable, the study tries to use the variable of 'distance from household to firewood sources'. It is reasonable to believe that the distance to firewood sources from households is correlated with using firewood for cooking but not with the dependent variables. However, in the paper, the appropriateness of the instrumental variable has been econometrically tested. The instrumental variable regression finally concludes that firewood burning increases household members' probability of asthma prevalence by 10.9 percentage points, out-patient care utilisation by 33.1 percentage points, and in-patient care utilisation by 17.5 percentage points, on average.

Another type of studies that one can perform using nation-wide survey data is quasi-experimental studies. Accordingly, the applied researchers perform propensity score matching and regression discontinuity design-related studies using household survey data. Availability of data for an array of covariates at the household- and individual-level for a larger sample allows researchers to perform such quasi-experimental studies with household survey data. The biases emerging from unobserved-heterogeneity in traditional regression models can be mitigated using quasi-experimental studies. In such exercises, depending on the objective (s) of the study, treatment variable, outcome variable(s), and balancing covariates need to be determined. Generally, the treatment variable is a dummy variable indicating whether a particular household or individual receives the defined-treatment, and accordingly, the study identifies separate treatment and control groups. Then, using a propensity score (generally, this is the predicted probability from a Probit or a Logit model) and one of matching algorithm, a matched- and an un-matched samples need to be identified. Finally, the researchers are interested in calculating average treatment effect on treated group (ATT) as follows:

$$\tau ATT = E(\tau \mid T=1) = E[Y(1) \mid T=1] - E[Y(0) \mid T=1]$$

Samaratunge et al. (2020) conducted a quasi-experimental study using HIES-2016 to uncover the impact of internal and international private remittances on household expenditure behaviour. This study defines the treatment as whether a household receives private internal and/or international remittances. The outcome variables consist of different household expenditure categories including, food, education, healthcare, durables, housing, and so forth. In order to minimize unobserved-heterogeneity, the households are balanced using an array of balancing covariates including, household size, dependency ratio, household composition, health status of members, level of education of members, and living sector and province. The study concludes that private remittances have significantly increased household per-capita expenditure and initiated positive behavioural changes via increased-allocations for basic needs, human and physical capital investment. Further, the study finds that compared with internal remittances, the impact of international remittances shows a strong potential for reducing poverty incidence of Sri Lankan households.

Even though the Department of Census and Statistics does not have panel household surveys which can be used to estimate Panel regression models accounting for fixed-and random-effects, some researchers have been tactful enough to examine social dynamics by converting original household survey data into a Pseudo-panel dataset. The Pseudo panel method can be regarded as an alternative to using panel data for estimating fixed-effects models when the researchers have only independent repeated-cross-sectional data. In constructing Pseudo-panel datasets, we need to observe cohorts, stable groups of individuals or households, rather than individuals or households overtime. For instance, one can observe households representing each division or each province overtime. Alternatively, the individuals from each district can be observed overtime in terms of their health status. Finally, a Pseudo-panel dataset can be constructed by replacing individual entry with cohort mean values. Now, one can account for social dynamics by estimating fixed-effect models, and so forth. Pallegedara and Grimm (2018) use the HIES data collected in 1990/1991, 1995/1996, 2006/2007, 2009/2010, and 2012/2013 to create a Pseudo-panel dataset when they examine why out-of-pocket healthcare payments at the household-level have risen under the free healthcare policy in

Sri Lanka. In this study, they take the aggregated-mean values of variables of interest at the district-level. Their random-effect model concludes that when households get richer, they spend an increasing amount on private healthcare facilities, implying a dissatisfaction with the quality offered by the public healthcare sector.

Apart from the HIES, the Department of Census and Statistics conducts the nation-wide surveys on agriculture, health, trade, poverty, computer literacy, industry, gender, public employment, education, labour force, and population. The relevant microdata files are also available for researchers to use in their studies.

2.2. The Variety of Democracy Institute (V-Dem Institute)

This is an independent research institute based at the Department of Political Science at the university of Gothenburg, Sweden. The institute collects data with the help of country experts, arranges them, and makes the data available to the users. The available data are mainly on democracy and democratization of countries, and they are available over a long period of time. Accordingly, the data are long-country-panels in nature which are useful in research projects in Economics, Sociology, History, and Political Science. The datasets include more than 50 democracy-related indicators annually from 1789 to the present for all countries of the world.

The V-Dem data have been arranged by the themes of elections, political parties, direct democracy, the executive, the legislature, deliberation, the judiciary, civil liberty, media, political equality, exclusion, legitimization, civic and academic space, and COVID-19. Under each theme, there are measures for an array of constructs. For instance, under media, they have the measures for government censorship efforts, Internet censorship efforts, print/broadcast media freedom, female journalists, media biases, harassment of journalists, and media-related corruption.

It should be noted that the V-Dem data are being used extensively by the researchers around the world. To highlight one of many, Kratou and Laakso (2021) examine the nexus between academic freedom and democracy in African countries using the V-Dem datasets. In particular, they use the indices of academic freedom and quality of elections and estimate a dynamic panel model using generalized method of moment (GMM) technique:

$$Quality\_elections_{i,t} = \alpha_i + \beta_1 Quality\_elections_{i,t-1} + \beta_2 Academic\_freedom_{i,t-10} \\ + \delta School\_enrollment_{i,t} + \varepsilon_{i,t}$$

They conclude that there is a positive impact of academic freedom on quality of elections in Africa after the cold-war. The study provides an array of robustness checks using V-Dem data.

2.3.The World Bank

The World Bank conducts a number of surveys and makes data available for researchers to use. Among many of such surveys, the World Bank Enterprise Survey (WBES) has been recognized widely by researchers and policy makers. The WBES is a firm-level survey of a representative sample of an economy's private sector. Accordingly, it provides an expansive array of economic data from 174,000 firms in 151 countries. The themes of data cover business environment and performance indicators which are created by taking weighted averages of businesses' responses to the questions in the WBES. The indices are available at the country-level as well as at the firm-level from 1990 to the present.

The raw individual country datasets, aggregated datasets, panel datasets along with survey documentation are publicly available on the World Bank website. During the survey, private sector business owners and top managers are interviewed by private survey contractors on behalf of the World Bank. In case of larger economies, around 1,800 interviews are conducted. However, for medium-sized economies, around 360 interviews whereas for smaller economies, 150 interviews are conducted. The WBES takes into account formal (registered) companies with 5 or more employees for obtaining data.

The sampled-companies represent manufacturing and services sectors. The services sector covers an array of sub-sectors including, construction, retail, wholesale, hotels, restaurants, transport, storage, communication, and IT. The WBES uses random stratified sampling technique to create a sample of firms from each country. The strata include firm size (small, medium, and large), business sector, and geographic region (located cities) within a country. The standard WBES topics cover firm characteristics, gender participation, access to finance, annual sales, cost of inputs/labour, workforce composition, bribery, trade, crime, competition, capacity utilization, land and permits, taxation, informality, business-government-relations, innovation and technology, creativity, and performance measures.

The WBES has been a popular source of firm-level data for researchers in Economics, Entrepreneurship, Business Development, and Business-Government-Relations. Moving onto a recently conducted study based on the WBES, Kumanayake (2021) provides empirical evidence from Asian firms on whether custom and other trade regulatory constraints lead firms to bribe. The study uses firm-level data from 11 Asian countries for the Endogenous Treatment Model to examine the impact of custom and other trade barriers on the firms' decision to bribe. The study reveals that the Asian firms facing custom and other trade barriers are likely to pay more bribes amounting to 6.4% of their annual total sales. Also, the study confirms that the firms trusting the fair functioning of their judiciary systems are less likely to bribe government officials.

2.4.GEM: Global Entrepreneurship Monitor

GEM is a global research source that collects data on entrepreneurship from individual entrepreneurs in countries. GEM conducts around 200,000 interviews annually with entrepreneurs, and the data are available for 115 countries across a period of 22 years. Thus, the data allow researchers to conduct longitudinal analyses across regions on multiple-levels.

The data have been arranged under two themes: entrepreneurial behaviour and attitudes and entrepreneurial framework conditions. Under the theme of entrepreneurial behaviour and attitudes, the data are available on perceived opportunities, perceived capabilities, fear of failure rate, entrepreneurial intensions, business ownership, motivational index, female/male participation, innovation, and high job creation expectations. Under the theme of entrepreneurial framework, GEM provides the firm-level data on financing, government support and policies, taxes and bureaucracy, governmental programs, basic school entrepreneurial education and training, post school entrepreneurial education and training, research and development transfer, commercial and professional infrastructure, internal market dynamics, internal market openness, physical and service infrastructure, and cultural and social norms.

The data collection process uses two complementary tools: adult population survey and the national expert survey. The adult population survey focuses on evolution of individuals in the lifecycle of entrepreneurial process. In addition to collecting data on business characteristics, it focuses on motivation for starting a business, actions taken to run a business, and entrepreneurial attitudes. Moreover, national expert survey focuses on the conditions that enhance or hinder new venture creation.

3. Conclusion

3.1. Validity and Reliability of Secondary Data

Generally, we do not conduct traditional 'validity' and 'reliability' tests when we use secondary data for studies. Traditional validity and reliability tests are conducted only for primary data to check the appropriateness of measurement instrument or questionnaires. In case of secondary data, we would rather conduct pre-diagnostic tests to ensure validity of data with regard to a particular policy model. For instance, in multiple regression models, the researcher(s) may test for normality, serial correlation, homogeneity, and outliers before estimating the model. In panel data analysis, one can test for panel unit root and panel cointegration. We may also conduct post-estimation tests to ensure the goodness-of-fit of policy models estimated using secondary data. With regard to reliability of secondary data, usually, researchers compare their data structures with national-level parameters to see whether the results are generalizable to the whole country.

3.2. Limitations

Researching with secondary data has limitations as well. The researchers do not have a control over the data collection process. Thus, researchers have less opportunity to correct for errors if any. In case of primary data, researchers have a full control over data collection process, and error-correction is direct via many mechanisms like conducting pilot surveys. Accordingly, it is advisable that researchers need to evaluate the data collection process wherever possible via procedural documentation of secondary datasets (Pederson et al., 2020).

Moreover, the secondary datasets generally suffer from having a large number of missing values for certain variables (DeCarlo, 2018). Hence, the researchers should be able to effectively deal with missing data. In this regard, researchers need to check whether those data are missing at random or not. Accordingly, they may use several standard methods of treating for missing values.

Finally, secondary datasets may not provide all of the information that researchers are interested in. Using primary data allows researchers to thoroughly study about a variable that the study is focused on. When using secondary data, researchers may have to depend on very limited amount of information about key variables that surveys provide. It frequently happens as researchers' key variable of interest may not be a central variable of surveys.

References

Deaton, A. (2006). 'Measuring Poverty'. Understanding Poverty. Oxford: Oxford University Press.

DeCarlo, M. (2018). 'Scientific Inquiry in Social Work'. Open Social Work Education. Link: https://scientificinquiryinsocialwork.pressbooks.com/

Kratou, H. and Laakso, L. (2021). The Impact of Academic Freedom on Democracy in Africa, *The Journal of Development Studies*, DOI: 10.1080/00220388.2021.1988080.

Kumanayake, N.S. (2021). Do customs and other trade regulatory barriers lead firms to bribe? Evidence from Asia, *The Journal of International Trade & Economic Development*, DOI: 10.1080/09638199.2021.1962391.

Kumara, A.S. and Pallegedara, A. (2020). Household waste disposal mechanisms in Sri Lanka: Nation-wide survey evidence for their trends and determinants, *Waste Management* 114, 62-71.

Kumara, A.S. and Samaratunge, R, (2016). Patterns and determinants of out-of-pocket health care expenditure in Sri Lanka: evidence from household surveys, *Health Policy and Planning* 31(8), 970-983.

Pallegedara, A. and Kumara, A.S. (2021). Impacts of firewood burning for cooking on respiratory health and healthcare utilisation: Empirical evidence from Sri Lankan micro-data, *International Journal of Health Planning and Management* 37(1),465-485.

Pallegedara, A., Grimm M. (2018). Have out-of-pocket health care payments risen under free health care policy? The case of Sri Lanka, *International Journal of Health Planning and Management* 33(3), e781–e97.

Pederson, L.L., Vingilis, E., Wickens, C.M., Koval, J., Mann, R.E. (2020). Use of secondary data analyses in research: Pros and Cons, *Journal of Addiction Medicine and Therapeutic Science* 6(1), 58-60.

Samaratunge, R., Kumara, A.S., and Abeysekera, L. (2020). Where do Remittances Go in Household Consumption? Empirical Evidence from Sri Lanka-Wide Micro-data, *International Migration* 58(5), 194-219.