1. Introduction

In the light of the increasingly important concerns within the EU regarding the extension of life expectancy, but especially the extension of the years of life lived in healthy conditions, more and more sporting attention is paid to the reduction of avoidable causes of premature death. *„The tobacco epidemic is one of the biggest public health threats the world has ever faced, killing over 8 million people a year around the world. More than 7 million of those deaths are the result of direct tobacco use while around 1.3 million are the result of non-smokers being exposed to second-hand smoke. Around 80% of the 1.3 billion tobacco users worldwide live in low- and middle-income countries , where the burden of tobacco-related illness and death is heaviest. In Romania there are 4.1 million smokers and the annual number of deaths attributable to tobacco smoking is over 36 thousands.*[1] *Tobacco use contributes to poverty by diverting household spending from basic needs such as food and shelter to tobacco. This spending behaviour can be difficult to curb because tobacco is so addictive."*[2]

Last year, Romania adopted the National Plan to Combat and Control Cancer[3], and the regulations for implementing this plan are currently being drawn up. At the same time, within this plan, among the risk factors, smoking occupies the first position as the main cause for the onset of lung cancer. Smoking is also responsible for heart disease, stroke, lung diseases, diabetes, and chronic obstructive pulmonary disease (COPD), which includes emphysema and chronic bronchitis. Smoking also increases risk for tuberculosis, certain eye diseases, and problems of the immune system, including rheumatoid arthritis.[4]

Also in this context, a Horizon Project funded by the European Commission "*Personalized CANcer Primary Prevention Research through Citizen Participation and digitally-enabled social innovation*"[5] was launched this year, in which the effects of smoking on the probability of developing any form of cancer will be analyzed. Reducing tobacco consumption or even preventing the start of consumption will be key elements in reducing cancer risk factors.

Compared to much of the EU, SEE countries are characterized by high levels of tobacco consumption and low prices of cigarettes (Zubović and Vladisavljević, 2020). High tobacco consumption imposes a significant economic burden on households in the region, while at the

[1] WHO. (2021). WHO global report on trends in prevalence of tobacco use 2000-2025, fourth edition (4th ed). World Health Organization. https://apps.who.int/iris/handle/10665/348537

[2] World Health Organization. (2023). WHO report on the global tobacco epidemic, 2023: protect people from tobacco smoke. World Health Organization. https://apps.who.int/iris/handle/10665/372043.

[3] https://ms.ro/media/documents/Planul_Na%C8%9Bional_de_Combatere_%C8%99i_Control_al_Cancerului_RIQiTXG.pdf

[4] Center for Disease Control and Prevention https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm

[5] https://rethink-health.eu/projects/4p-can-project-launch-personalised-cancer-primary-prevention/

same time, the negative effects of tobacco consumption have long-lasting effects on health and well-being in general. Numerous studies such as "*Tobacco taxes as a tobacco control strategy*"[6] (2012) of F.J. Chaloupka, A. Yurekli, G.T. Fong or the Regional Study on *"Impacts of Tobacco Excise Increases on Cigarette Consumption and Government Revenues in Southeastern European Countries"*[7] *(2019) and the WHO "Technical manual on tobacco tax policy and administration"*[8] *(2021)* indicate that tobacco taxation is one of the most important policies to reduce tobacco consumption.

The main objective of this research is to estimate the price elasticity of demand for cigarettes because it plays a crucial role in providing policymakers with essential insights, particularly enabling them to evaluate and model the potential impact of adjusting cigarette taxes on tobacco consumption. Additionally, accurately estimating price elasticity plays a vital role in better forecasting how alterations in cigarette taxes will influence government revenues.

The econometric model for estimating price and income elasticity of demand is based on theoretical framework of the two-part model developed by Manning and Mullahy (2001). This model estimates the overall demand elasticity as a (corrected) sum of two elasticities: prevalence elasticity and conditional demand (in other words, intensity) elasticity. The prevalence elasticity is estimated via a logit model and Generalized Linear Model (GLM) is used for the estimation of conditional demand (intensity).

In order to estimate the price and income elasticity of demand, Household Budget Survey (HBS) data were used. The data were received from the Romanian National Institute of Statistics for the period of 2015 to 2021. Firstly, we estimated the cigarette price and income elasticity of demand on the extensive (prevalence elasticity) and the intensive margin (conditional demand or intensity elasticity). After estimating the demand elasticity for the entire sample, we divided the sample into three income groups (low income, middle income and high income), in order to evaluate if the change in the price of cigarettes affects smokers differently depending on their income. And lastly, a simulation of the impact of an increase in tobacco excise and price on consumption and government revenue was realized.

[6] https://tobaccocontrol.bmj.com/content/tobaccocontrol/21/2/172.full.pdf

[7] https://tobacconomics.org/research/impacts-of-tobacco-excise-increases-on-cigarette-consumption-and-government-revenues-in-southeastern-european-countries/

[8] https://www.who.int/publications/i/item/9789240019188

2. Data description and pre-processing

This chapter describes the data and methodology used in the report. It details the approach for calculating the price participation and intensity (conditional) elasticity of cigarettes in further detail. The mechanism for estimating price elasticity at various income levels is also covered in this chapter. The effects of a price rise on consumption and tax income are then predicted using the estimates. The same two-part econometric models and simulation techniques were used as in the regional study from 2019 on „Impacts of Tobacco Excise Increases on Cigarette Consumption and Government Revenues in Southeastern European Countries" (Zubović J. and Vladisavljević M. 2019). However, there are minor variations in model definition and years of available data due to slight discrepancies in the data that are available and country-specific factors.

To calculate the price and income elasticities of cigarette consumption, microdata from Household Budget Surveys (HBS) data are used in all analyses. The HBS provides the necessary information for the assessment of income, expenses and consumption of the population.

Households from all socio-economic categories are included in the research: employees, employers, self-employed workers in agriculture (farmers) or members of agricultural associations (farmers from commercial agricultural companies), self-employed workers in non-agricultural activities (tradesmen, traders, freelancers, etc.), members of non-agricultural cooperatives (craft, consumer and credit cooperatives), unemployed, pensioners, other categories.

The belonging of the household to one of these socio-economic categories is established based on the declared main occupational status of the head of the household.

The HBS is organized as a continuous quarterly survey over a period of 3 consecutive months, on a sample of 9,504 permanent dwellings, divided into independent monthly sub-samples of 3,168 permanent dwellings. The response rate was 80.5% (76.5% in urban areas and 85.8% in rural areas).

In order to extract the sample a two-stage survey design was used:

- In the first step, 792 research centers (Primary Sampling Units) were selected from the Population and Housing Census - October 2011 (RPL'2011) using the stratified and balanced extraction method of UP within each stratum, constituting the Multifunctional Sample of Territorial Zones (EMZOT'2011 "master" sample) as a survey basis for selective surveys in households, in the intercensal period. The stratification criteria were county and residence environment, by intersecting them, resulting in a number of 88 strata (in the Municipality of Bucharest, the selection was made separately for each of the 6 administrative sectors). EMZOT is a sample of 792 research centers distributed in all the counties of the country and in the sectors of the Municipality of Bucharest (450 in the urban environment and 342 in the rural environment).

- In the second step, 9504 permanent homes were selected per quarter, in 3 monthly waves of 3168 according to a systematic selection algorithm. The homes extracted in the second step are assimilated to the secondary sampling units. From each research center, 12 homes were included in the sample quarterly, respectively 4 homes monthly.

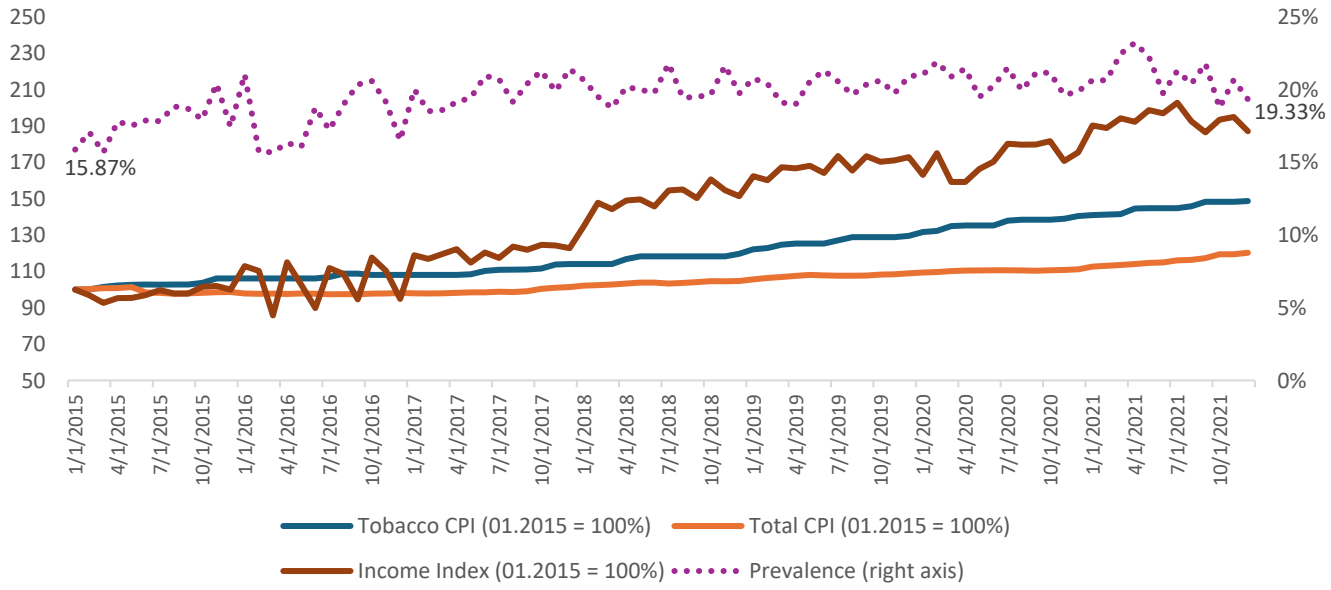
The sample size was calculated to ensure national and regional representativeness for the main survey variables. The survey sample is extracted from the EMZOT-2011 master sample, based on the data recorded at the 2011 Population and Housing Census. EMZOT-2011 is a database composed of approximately 1,500,000 households, selected according to probabilistic criteria, with the aim of serve as a sampling base for all household survey research, for the period 2015 – 2024.

3. Descriptive Statistics and Variable Definitions

Before moving to the econometric model and price and income elasticities estimation, a quick bird-eye view of the HBS data could provide some insights on the Romanian economic context and tobacco consumption. A critical aspect regarding the consumption of cigarettes is the correlation between the increase in income and the increase in cigarette prices.

As can be seen from Figure 1, even though in the analyzed period 2015-2021 tobacco CPI was higher than general CPI, the increase in the average income of households was considerably higher. The cumulative increase in tobacco CPI was somewhere around 50%, while the increase in average income was approximately 90%. Therefore, despite the price increases of tobacco, cigarettes have actually become more affordable. The result of this "cheapening" in relative terms of cigarettes likely contributed to the increase in prevalence rates, though we do not test this claim empirically here. The observed prevalence rate increased in the analyzed period from a little under 16% to over 19%.

Figure 1 Income, Inflation and Tobacco Consumption



Source: National Institute of Statistics and HBS data

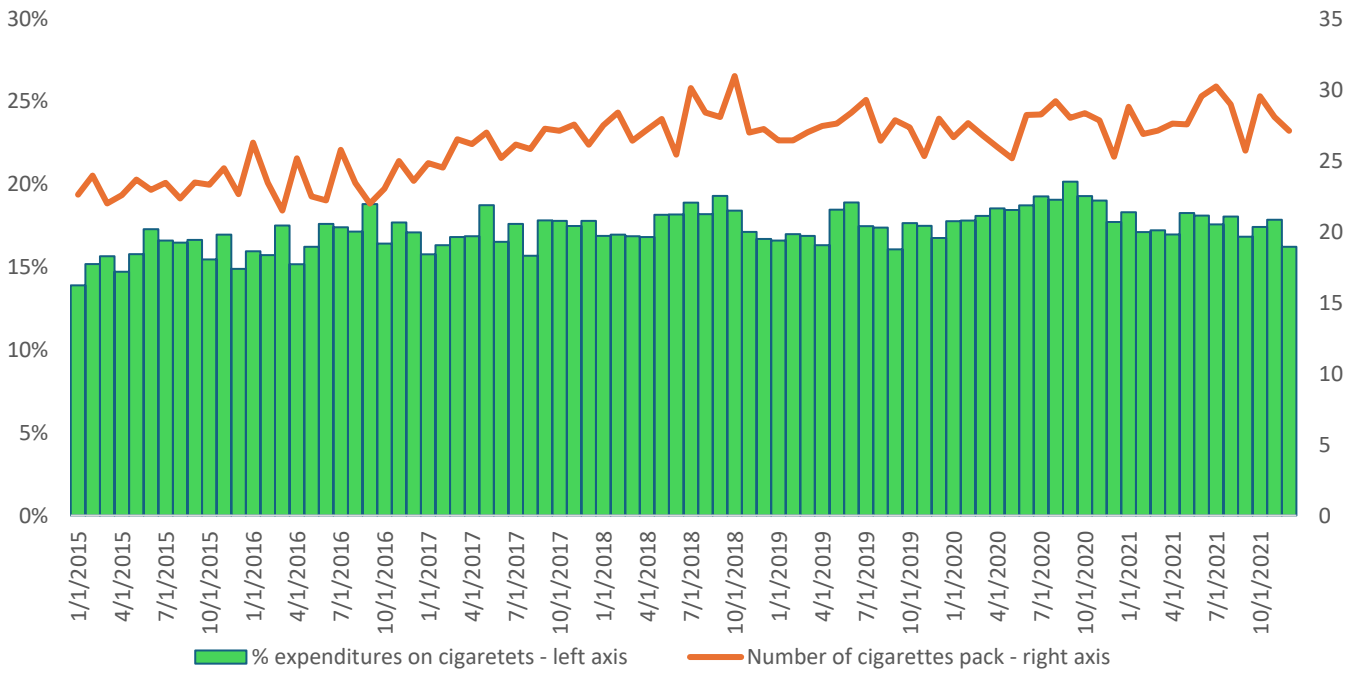
* Note: For Tobacco CPI and Total CPI publicly available data from National Institute of Statistics from Romania (NIS) was used. For income index and prevalence data provided by NIS from HBS data was used. The graph displays the information for income after the winsorizing step that was performed in the data cleaning process (i.e., this is the same data used in the model).

Additionally, during this period the consumption of e- cigarettes, vaping and heated tobacco has increased. This should have been expected to reduce the prevalence rates for cigarette

consumption, because of a switching effect, which did not happen. The reasons why this did not happen can be multiple, among which likely includes the fact that the broader increase in income was considerably higher than the increase in the price of cigarettes. Yet, another fairly common phenomenon is the fact that due to much looser regulations regarding heated tobacco, the same people smoke heated tobacco products in places where cigarettes are prohibited, and when they are in places where smoking is allowed, they smoke manufactured cigarettes therefore showing dual – use of heated tobacco products and traditional cigarettes, thus making them complementary and not substitutes

Moreover, HBS data can inform us regarding the share of smoking expenses in the total expenses of a household (see Figure 2 below). The total expenses of a household are the sum of Food Expenses + Non Food expenses + Services, found in the HBS survey and represent total spending on consumption. The average, over the 7-year analyzed period, cigarettes expenses represented 17% of the total average monthly expenses of a household, which represent an enormous financial burden for the smoking households.

Figure 2 % Expenditure on cigarettes of total Household expenditure and Number of Cigarettes[9]

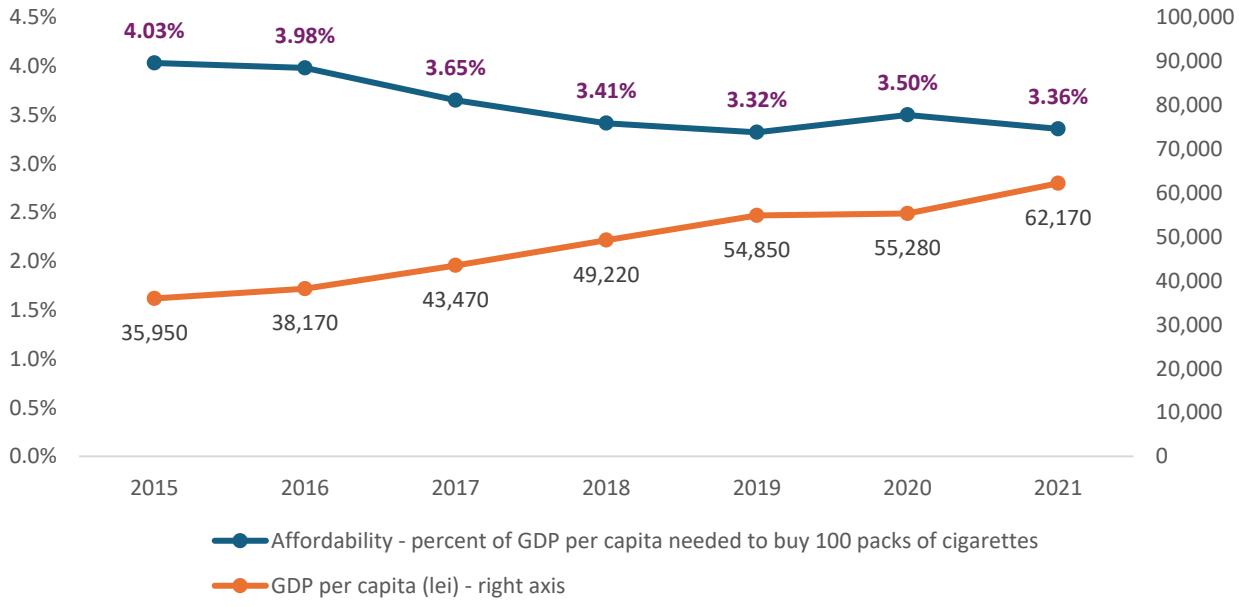


Source: Own processing based on National Institute of Statistics and HBS data

The number of cigarettes consumed on average per household is on a slightly increasing trend. The number of cigarettes packs (of 20 cigarettes sticks, each) consumed in a respective month was determined by dividing the household expenditure on cigarettes by the WAP (TCPI adjusted) for the respective month. This happens because both household income and expenses have increased significantly in this 7-year period, at a much faster rate than the price of cigarettes, and thus with the same allocation of 16% of total expenses at the end of 2021, a household could consume approximately 27 packs of cigarettes compared to 22 packs of cigarettes in March 2015.

Affordability determines how accessible is the consumption of cigarettes. Most commonly is determined by using the Relative Income Price (RIP), here the GDP per Capita was used, and is depicted in figure 3 as what percentage of the GDP per Capita is required to buy 100 packs (of 20 cigarettes each). As it can be observed in 2015, 4.03% of GDP per capita was needed in order to buy 100 packs of cigarettes, while in 2021 only 3.36% of GPD pe capita is needed in order to buy the same amount of cigarettes. This means that in relative terms, cigarettes have become more affordable, they are cheaper in relative terms and this happened because Income (measured here as GBP per Capita) has increase faster than the price of cigarettes. In other words, in the 7 years analyzed period (2015-2021), cigarettes have become 16.7% cheaper in relative terms.

Figure 3 Evolution of affordability and GDP per capita



Source: own processing based on HBS data, National Institute of Statistics

4. Price elasticity of cigarette consumption in Romania

4.1. Model description

Tobacco consumption, specifically cigarettes, is marked by a significant portion of non-smokers. This implies that the distribution has a discrete component and a continuous component. This results in the consumption variable taking a value of zero for these individuals, whereas the rest have strictly positive values. The distribution can be described as:

$$y_i = \begin{cases} 0, & n = 0, 1, \dots, n_i \\ y_i > 0, & n = n_{i+1}, n_{i+2}, \dots n_{i+N} \end{cases} \quad (1)$$

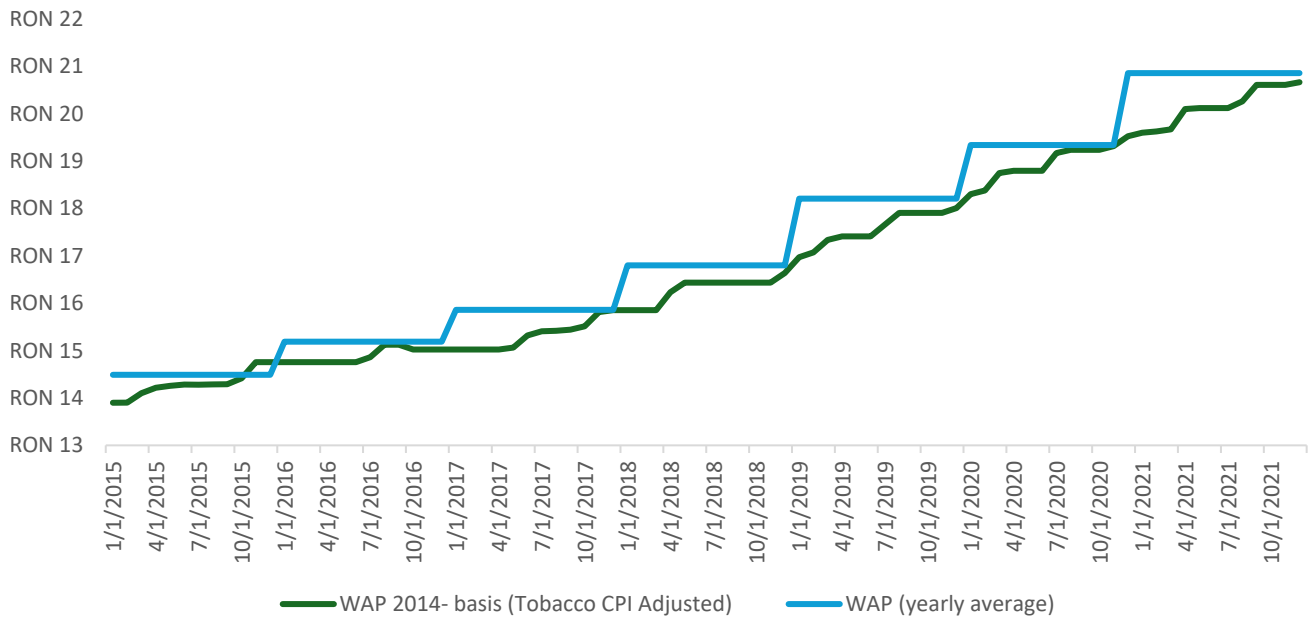
Where:

- y_i - represents the quantity of consumed cigarettes by a household
- n_i - represents the household i

The study examines the distribution of cigarettes. This distribution highlights that individuals, when considering market prices, their financial limits, and the satisfaction they get from smoking, make two primary choices: 1) whether to smoke or not; and 2) if they choose to smoke, how much they consume. Existing literature posits that these two decisions should be analyzed separately in what's termed as the two-part model (Belotti, et al., 2015). This approach is especially relevant when a value of y=0 is frequently observed. This is evident in cigarette consumption, as the global smoking rate stands at about 17.5 percent (WHO, 2021), while the smoking rate in this specific study is situated between 18 and 24 percent for cigarettes.

Price and income are the two primary factors that both models take into account. The computation of the price elasticity, income elasticity, prevalence, and intensity of cigarette smoking is based on these two factors. WAP provided from administrative sources is reported on a yearly basis. In order to generate monthly data, the WAP for 2014 was used as a starting point to compute monthly WAP based on tobacco CPI data. By adjusting the 2014 WAP with monthly tobacco CPI, the values resulted are slightly lower than yearly WAP observed on the Ministry of Finance website.

Figure 4 Evolution of weighted average price (WAP)



Source: own processing based on NIS and Taxation and Customs Union data.

As the models are estimated separately and independently, the total price and income elasticity is calculated as the corrected sum of the prevalence and the conditional demand (intensity) elasticity. Total elasticity cannot be calculated as simple sum of the two elasticities. Instead, this sum needs to be corrected for the fact that a change in the smoking prevalence can attenuate the effect of the conditional demand (intensity) elasticity. Consequently, two models are formulated using the following equations:

$$P\ (y_i > 0) = f(\beta_1 p_{\ j}\ + \beta_2 i_{i,j}\ \ \gamma H_{i,j} + \zeta CF_j) \tag{2}$$

$$E\ (y_i \mid y_i > 0) = \beta_1 p_{\ j}\ + \beta_2 i_{i,j} +\ \ \gamma H_{i,j} + \zeta CF_j \tag{3}$$

Where:

- $p_{\ j}$ denotes the price in period j,
- $i_{i,j}$ represents the income for household i in period j,
- $H_{i,j}$ is a set of household specific variables (detailed description in Table 1)
- CF_j represents a set of control variables (detailed description in Table 1)

In these, equation (2) depicts prevalence, while equation (3) focuses on intensity. Together, these models form a system of two equations detailing the demand for cigarettes.

It's crucial to consider the potential reciprocal relationship between prices and demand indicators when evaluating price elasticities. Prior research assessing the independence of tobacco prices has deduced that such prices can be considered exogenous (Karki et al., 2003; Kyaing, 2003; NCI, 2016; Kostova & Dave, 2015), even when derived from a similar aggregation level (Huang, et al., 2018)). Lastly, it's worth noting that prices aren't solely driven by market dynamics. For one, state-determined excise taxes significantly shape them. Moreover, in the SEE region, price alignment with the EU heavily sways them, meaning a dip in demand wouldn't necessarily modify cigarette prices. However, to assess possible endogeneity problems a Hausman test was performed on the model. The test suggested that there are possible endogeneity issues. To address this problem an instrumental variable approach was used. First the price was estimated using the same regressors as in the model and an instrumental variable, in this case excise was chosen as instrumental variable for price as it clearly has an impact on the price, but the level of the excise should have no impact on smoking decisions. The estimated level of price was used as a regressor instead of the original price variable.

4.2. Estimation of prevalence elasticity

The initial segment of the model examines how tobacco prices influence a household's choice to smoke, given the set of independent variables. This decision is commonly represented using a binary choice model. The distinguishing factor between a binary choice and the standard linear regression model is the character of the dependent variable. In binary choice models, rather than modeling a continuous variable, it focuses on the likelihood that the dependent variable y_i equals one — indicating households that spend on or consume cigarettes — as opposed to zero, which signifies households that do not consume cigarettes. As a result, the model employs a (nonlinear) function of the linear combination of independent variables to articulate the probability of a household incurring positive tobacco expenses.

For the first part of the model that estimates prevalence, a logit model was used to estimate Equation (2). County-level, month, and year cluster corrected standard errors are applied to account for any differences that may manifest locally or that have appeared over the course of the studied interval, as well as heteroscedasticity-robust standard errors to control for potential heteroscedasticity in both parts of the model. For the purpose of this report the variables were used both in level as well as logarithm form to test the robustness of the results. Also, further tests have been performed to assess the performance of each of the estimated models in order to select the best functional form for the model. These tests include Akaike and Bayesian Information Criteria, collinearity diagnostics, link test, and goodness of fit tests (the test results for each model is presented in the annex 2 to 5).

Given the structure of HBS data, where households change each year, a typical panel regression is not applicable in its standard form as it typically relies on observing the same units across different time periods to capture unobserved individual heterogeneity. To address this, the current model is a pooled regression which combines all data into a single model without accounting for the individual effects specific to each unit or time interval. If cross-sectional units change each

period, pooled regression can still be applied as it combines all data into a single model without accounting for the individual effects specific to each unit. We considered this approach to be better suited given our data structure since panel regression is not applicable in its standard form, as it typically relies on observing the same units across different time periods to capture unobserved individual heterogeneity.

4.3. Estimation of intensity elasticity

The dependent variable in intensity models, Equation (3), is typically represented in log form as it helps to stabilize non-constant error variance; however, similarly to Equation (2) both level and log form were estimated and tested for performance. A standard practice in health economics in this case is to use the Generalized Linear Model (GLM) with gamma family and a log link function. This method has been proposed as a more robust alternative to a log regression specification (Manning, et al., 2005). In this situation, GLM is the preferred model as the Ordinary Least Squares (OLS) estimator requires retransformation which can cause a prediction bias.

In order to implement a two-part model each component was estimated and tested separately and then aggregated into a two-part model. The best functional form of the model was identified as model (5) in log form which has the lowest AIC and BIC scores for the second part of the model, passes the link test and does not suffer from multicollinearity. However, model 3, has lowest AIC and BIC scores for first part and evethou it does not pass the link test, it is right on the limit with a score of 0.049. but because it includes variables that show the occupational status within the households in comparison with model 5, we selected model 3. After the selection of model (3) the goodness of fit was assessed for deciles indicating suggesting that there's no evidence of systematic bias in the residuals across the deciles of predicted values, indicating a good model fit across different levels of the predictors.[10] Also, after the identification of the best form of the two-part model and using the same log specifications the model was re-run, conditional on each separate income group to assess if there are significant differences of elasticity according to the income group. The whole sample of observations, approximatively 105 thousands was divided into three equal number of observations groups, representing low income, middle income and high income groups.

4.4. Data description

Before estimating the models some of the variables needed some transformations in order to increase the quality of the modelling process. This section provides a description of all the

[10] The test was perform both for the original model with F-test was F(10, 20637) = 0.94, and Prob > F = 0.4921, as well as for the model updated to accommodate endogeneity problems with a F(10, 20637) = 0.78, and Prob > F = 0.6475

variables used and the transformations that were performed on each of them before the modelling process.

To perform all the relevant modelling of the data, some transformation had to be performed on the raw data. These transformations are described below, while the descriptive statistics are presented at large in Annex 1 Descriptive statistics

Variable	Obs	Mean	Std. dev.	Min	Max
year	105,010	2018	2.000	2015	2021
month	105,010	6.5	3.5	1.0	12.0
Wap (weighted average price)	105,010	16.8	2.1	13.9	20.7
Nrgl (Order number of the household within the dwelling)	105,010	1.0	0.1	1.0	4.0
sex	105,010	1.5	0.5	1.0	2.0
Lunn (month)	105,010	6.4	3.3	1.0	12.0
Ann (year)	105,010	1962.0	18.0	1916.0	2006.0
Nat (nationality)	105,010	1.1	0.4	1.0	5.0
Nive (The last level of education of the highest degree graduated)	105,010	5.9	2.3	1.0	14.0
Stocup (Occupational status in the reference month)	105,010	6.4	4.0	1.0	14.0
Stocupan (Main occupational status in the last 12 months)	105,010	6.3	4.0	1.0	14.0
Idhh (household ID)	105,010	8075.0	4686.0	1.0	16449.0
reg nuts2 (region – nuts2)	105,010	4.2	2.3	1.0	8.0
Mediu (environment)	105,010	2.0	1.0	1.0	3.0
Weight	105,010	262.5	245.2	35.9	3375.0
district	105,010	21.9	13.4	1.0	52.0
age	105,010	56.1	18.0	15.0	100.0
Hsize (Household size)	105,010	4.0	2.8	1.0	27.0
Adultratio	105,010	1.0	0.1	0.1	1.0
Maleratio	105,010	0.4	0.3	0.0	1.0
Maxedu (highest level of education in household)	104,988	3.6	1.4	1.0	6.0
educ avg years (average years of education in an Household**)**	105,010	10.4	2.7	0.0	20.0
Htype (The occupational status of the household's member with the highest occupational status)	105,010	3.3	0.9	1.0	4.0
hhd avg activity	105,010	3.0	1.0	1.0	5.0
urban	105,010	0.5	0.5	0.0	1.0
tot	104,999	1175.0	866.2	0.0	12920.0
Ecig (Household expenditure on cigarettes)	20,657	447.7	332.5	4.0	3464.0
tot1	84,015	1287.0	1369.0	0.3	79698.0
tot2	104,932	1028.0	971.8	0.0	36733.0
mce30 (The income at household level)	105,010	0.6	0.5	0.0	3.4
total cons (total consumption)	105,010	3232.0	2653.0	0.0	83477.0
insd ban (indoor smoking ban)	105,010	0.8	0.4	0.0	1.0
Ncig (The number of cigarettes consumed by the household)	105,010	103.6	268.1	0.0	3442.0
Idcig (A dummy variable that highlights if a household has any expenditure on cigarettes)	105,010	0.2	0.4	0.0	1.0
Disid	105,010	1775.0	1083.0	1.0	3947.0
income grp	105,010	2.0	0.8	1.0	3.0
htdum1 (A set of dummy variables corresponding to each level of occupation)	105,010	0.0	0.1	0.0	1.0
htdum2	105,010	0.3	0.5	0.0	1.0
htdum3	105,010	0.1	0.2	0.0	1.0
htdum4	105,010	0.6	0.5	0.0	1.0
hsize2 (Household size squared)	105,010	24.0	35.5	1.0	729.0
lnp (The natural logarithm transformation of price)	105,010	2.8	0.1	2.6	3.0
lny (The natural logarithm transformation of income)	104,941	-0.9	0.9	-5.5	1.2
lnp2 (The square of log transformation of price)	105,010	7.9	0.7	6.9	9.2
lny2 (The square of the log transformation of income)	104,941	1.7	2.3	0.0	30.0
wap2 (The square of price (wap))	105,010	286.2	71.0	193.3	427.3
mce2 (The square of income (mce30))	105,010	0.6	1.2	0.0	11.5
maleratio2 (The square of male ratio)	105,010	0.3	0.2	0.0	1.0
adultratio2 (The square of adult ratio)	105,010	0.9	0.2	0.0	1.0

.

Table 1 Variable names and descriptions

Variable name in the model	Description
hsize	The total household size computed for each combination of year and household (including children).
nadults	The number of adults in each household was determined by counting all persons above 14 years of age.
adultratio	The ratio between the number of adults and household size
namales	The number of males in each household
maleratio	The ratio between the number of males and household size
educc	The level of education of the household's member with the highest education. In the HBS data there are 12 categories of education. For the purpose of this study, the number of categories was reduced to 6: 1 "less than primary", 2 "primary", 3 "secondary 4 years" , 4 "post-secondary ", 5 "tertiary BA", 6 "tertiary MA, PhD"
deduc1 to deduc6	A set of dummy variables corresponding to each level of education (some of the specifications and tests needed this type of data structure). In all the specifications of the models, one of these variables is dropped.
htype	The occupational status of the household's member with the highest occupational status. In the HBS data there are 14 categories of occupational status. For the purpose of this study the number of categories was reduced to 4: 1 "Unemployed or other", 2 "Pensioner", 3 "Farmers", 4 "Employed"
htdum1 to htdum4	A set of dummy variables corresponding to each level of occupation. In all the specifications of the models, one of these variables is dropped.
ecig	Household expenditure on cigarettes
insdban	A dummy variable that highlights when there was a change in Law that prohibited smoking in indoor places like restaurants, bars, office buildings etc.
wap	Given that the quantities of cigarettes are not collected in HBS in Romania, in order to determine the price, we had to rely on administrative information regarding the weighted average price per pack (which is available at yearly level from the Ministry of Finance – MoF). In order to infer monthly data, we used the WAP data for 2014 as basis and computed the monthly values for Jan-2015 to Dec-2021 using tobacco CPI provided by NIS. This approach lead to slightly different average yearly WAP compared to the Ministry of Finance values for WAP.
ncig	The number of cigarettes consumed by the household. Since quantity data is not readily available in HBS, this was determined by using expenditure data

	and WAP data. The number of cigarettes consumed was estimated as the ratio between the expenditure on cigarettes and the WAP multiplied by 20 (the number of cigarettes in a pack).
idcig	A dummy variable that highlights if a household has any expenditure on cigarettes.
wap2	The square of price (wap)
lnp	The natural logarithm transformation of price
lnp2	The square of log transformation of price
mce30	The income at household level extracted from HBS data. The income was divided by 10000, to reduce the possibility that the models will be affected by the difference in scales between variables. Furthermore, to mitigate the impact of outliers on the model, income above the 99th percentile was replaced by the income corresponding to the 99th percentile.
mce2	The square of income (mce30)
lny	The natural logarithm transformation of income
lny2	The square of the log transformation of income
income_grp	A new variable containing information about the income group was created. For each year the households were divided into terciles based on their income: the lowest income households, middle income households, the highest income households.

5. Results

5.1. Prevalence elasticity

The results for prevalence elasticity (extensive margin), using variables in logarithm, are presented in 2, indicating a price elasticity of -0,111 and an income elasticity of 0.240. The detailed results of all the model specifications are presented in Annex 2 and Annex 3 (the selected model is presented in Annex 3, model (3)).

Table 2 Prevalence elasticity

	ey/dx	std. err.	z	P>|z|	95% conf. interval	
Price elasticity	-0.111***	0.0711	-1.58	0.115	-0.251	0.027
Income elasticity	0.240***	0.0219	10.94	0	0.197	0.283

Source: Own calculation

All the estimated models performed modestly for the first part. However, adding more interaction terms or higher order terms is not supported in literature and in majority of the cases it also

induces severe multicollinearity into the model. The current form that was kept was the one that (i) didn't present multicollinearity problems, and (ii) had for the second part of the model the values very close to lowest values for the Information criteria (AIC and BIC) and included variables on occupational status within the household.

One possible cause for these results is the structure of the data: it contains information at household level when the decision to smoke is rather a personal one, thus being very hard to add household information data that will help classify actions (i.e., decision to smoke or not) that are taken by individual from those household (e.g., some individuals may decide to smoke or quit smoking, but if some other person from that household already smokes and keeps smoking, the information about the household doesn't change, even though at individual level some changes happened).

5.2. Intensity elasticity

The results for intensity elasticity (intensive margin), using variables in logarithm, are presented in Table Table 2, indicating an price elasticity of -0,432 and an income elasticity of 0.508. The detailed results of all the model specifications are presented in Annex 4 and Annex 5 (the selected model is presented in Annex 5, model (3)).

Table 3 Intensity elasticity

	ey/dx	std. err.	z	P>|z|	95% conf. interval	
Price elasticity	-0.432***	0.0471	-9.18	0	-0.525	-0.340
Income elasticity	0.508***	0.0146	34.85	0	0.480	0.537

Source: Own calculation

Compared to prevalence intensity for the second part, the model performs better in the link test, eventhou the value (0.049) is still under, but not significant, the threshold of 0.05. It also passes the specification as well as the Collin test for multicollinearity. The same caveats as to the prevalence model still apply, even though they are mitigated by the fact that the decision on how many cigarettes to buy (being continuous not discrete) is highlighted even at household level as previously argued by Deaton and Ng in *"Parametric and Nonparametric Approaches to Price and Tax Reform"*. Given that the actual consumption is derived from average prices and the fact that the real quantity is not collected in the surveys and computed from average prices, one of the drawbacks of this approach is that it doesn't capture quality switching. However, at national level should capture the phenomenon accurately.

5.3. Two-part model

The two-part approach essentially takes both Equation (2) and (3) and estimates together,. Although the literature suggests that these two decisions can be modelled independently, total elasticity cannot be calculated as simple sum of the two elasticities. Instead, this sum needs to be corrected for the fact that a change in the smoking prevalence can attenuate the effect of the conditional demand (intensity) elasticity. The total price elasticity and income elasticity are the sum of prevalence, and intensity elasticity, **resulting in a price elasticity of -0.545 and income elasticity of 0.749**. Both price and income elasticities of demand are within the invervals observed in the literature for medium-high and high income countries as Romania.

Table 4 Two-part model estimations

	Prevalence model		Intensity model			
VARIABLES	Coef.	std. err.	Coef.	std. err.	Estimated elasticities	std. err.
Linear prediction	-0.139	(0.088)	-0.433***	(0.047)	-0.545***	(0.085)
lny	0.299***	(0.027)	0.508***	(0.015)	0.749***	(0.026)
lny2	-0.050***	(0.010)	0.036***	(0.006)		
hsize	-0.030***	(0.005)	0.016***	(0.002)		
maleratio	1.248***	(0.037)	0.307***	(0.019)		
adultratio	-0.390***	(0.078)	0.160***	(0.042)		
educc==less than primary	-0.451***	(0.076)	-0.110**	(0.051)		
educc==primary	-0.075**	(0.033)	-0.047***	(0.017)		
educc==post secondary	-0.091***	(0.027)	-0.005	(0.013)		
educc==tertiary BA	-0.099**	(0.040)	-0.043**	(0.020)		
educc==tertiary MA, PhD	-0.138***	(0.024)	-0.031***	(0.012)		
htype== 1.0000	-0.266**	(0.116)	0.010	(0.070)		
htype== 2.0000	-0.573***	(0.028)	-0.008	(0.014)		
htype== 3.0000	-0.178***	(0.041)	-0.053**	(0.025)		
Constant	-0.584**	(0.279)	7.323***	(0.152)		
Observations	104,919		104,919		104,919	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Own calculation

5.4. Income group estimation

Using the same log specification as above the model was re-run, conditional on each separate income group, the elasticities are presented in Table 55 while the detailed results of estimations are presented in Annex 6 .

Table 5 Elasticities estimation – national level and income group level

	Group 1 - low income		Group 2 – middle income		Group 3 - high income		Total	
VARIABLES	Prevalence	Intensity	Prevalence	Intensity	Prevalence	Intensity	Prevalence	Intensity
Price elasticity	-0.349**	-0.386***	-0.114	-0.520***	-0.107	-0.463***	-0.112	-0.433***
	(0.158)	(0.0973)	(0.127)	(0.0800)	(0.111)	(0.0738)	(0.071)	(0.047)
Income elasticity	0.566***	0.800***	0.130	0.534***	0.232***	0.469***	0.240***	0.508***
	(0.175)	(0.120)	(0.137)	(0.0846)	(0.0362)	(0.0251)	(0.022)	(0.015)
Observations	34,937	4,146	34,987	7,538	34,995	8,963	104,919	20,647

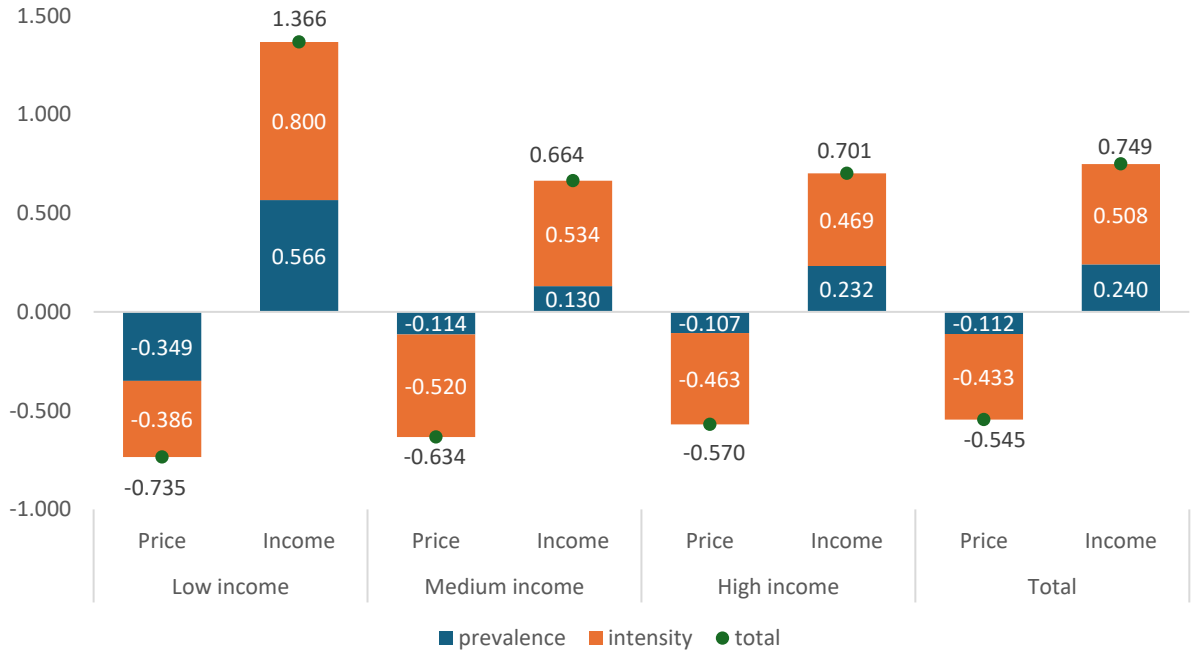
Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Own calculation

Based on the estimates of prevalence and conditional demand elasticity from the previous sections, total demand elasticity is calculated and presented below by income group in Figure 5

Figure 5 Elasticities – national level and by income group



Source: Authors' calculation based on the estimated elasticities Table 6

Total price elasticity is the highest for low-income households at -0.735, which means that a 10 percent price increase leads to a decrease in consumption by 7.35% percent. Elasticity is the lowest in the high-income group at -0.570, the effect of price on prevalence being almost insignificant, meaning that most of the increase in price will lead to a reduction in consumption.

It is important to note that a similar percentual increase in income as in prices will offset the increase of cigarettes prices, especially for the low-income category where income elasticity is highest, standing at almost 1.4. And, as shown in the descriptive statistics chapter, in the last decade income increases were substantial.

Annex 1 Descriptive statistics

Variable	Obs	Mean	Std. dev.	Min	Max
year	105,010	2018	2.000	2015	2021
month	105,010	6.5	3.5	1.0	12.0
Wap (weighted average price)	105,010	16.8	2.1	13.9	20.7
Nrgl (Order number of the household within the dwelling)	105,010	1.0	0.1	1.0	4.0
sex	105,010	1.5	0.5	1.0	2.0
Lunn (month)	105,010	6.4	3.3	1.0	12.0
Ann (year)	105,010	1962.0	18.0	1916.0	2006.0
Nat (nationality)	105,010	1.1	0.4	1.0	5.0
Nive (The last level of education of the highest degree graduated)	105,010	5.9	2.3	1.0	14.0
Stocup (Occupational status in the reference month)	105,010	6.4	4.0	1.0	14.0
Stocupan (Main occupational status in the last 12 months)	105,010	6.3	4.0	1.0	14.0
Idhh (household ID)	105,010	8075.0	4686.0	1.0	16449.0
reg nuts2 (region – nuts2)	105,010	4.2	2.3	1.0	8.0
Mediu (environment)	105,010	2.0	1.0	1.0	3.0
Weight	105,010	262.5	245.2	35.9	3375.0
district	105,010	21.9	13.4	1.0	52.0
age	105,010	56.1	18.0	15.0	100.0
Hsize (Household size)	105,010	4.0	2.8	1.0	27.0
Adultratio	105,010	1.0	0.1	0.1	1.0
Maleratio	105,010	0.4	0.3	0.0	1.0
Maxedu (highest level of education in household)	104,988	3.6	1.4	1.0	6.0
educ avg years (average years of education in an Household**)**	105,010	10.4	2.7	0.0	20.0
Htype (The occupational status of the household's member with the highest occupational status)	105,010	3.3	0.9	1.0	4.0
hhd avg activity	105,010	3.0	1.0	1.0	5.0
urban	105,010	0.5	0.5	0.0	1.0
tot	104,999	1175.0	866.2	0.0	12920.0
Ecig (Household expenditure on cigarettes)	20,657	447.7	332.5	4.0	3464.0
tot1	84,015	1287.0	1369.0	0.3	79698.0
tot2	104,932	1028.0	971.8	0.0	36733.0
mce30 (The income at household level)	105,010	0.6	0.5	0.0	3.4
total cons (total consumption)	105,010	3232.0	2653.0	0.0	83477.0
insd ban (indoor smoking ban)	105,010	0.8	0.4	0.0	1.0
Ncig (The number of cigarettes consumed by the household)	105,010	103.6	268.1	0.0	3442.0
Idcig (A dummy variable that highlights if a household has any expenditure on cigarettes)	105,010	0.2	0.4	0.0	1.0
Disid	105,010	1775.0	1083.0	1.0	3947.0
income grp	105,010	2.0	0.8	1.0	3.0
htdum1 (A set of dummy variables corresponding to each level of occupation)	105,010	0.0	0.1	0.0	1.0

htdum2	105,010	0.3	0.5	0.0	1.0
htdum3	105,010	0.1	0.2	0.0	1.0
htdum4	105,010	0.6	0.5	0.0	1.0
hsize2 (Household size squared)	105,010	24.0	35.5	1.0	729.0
lnp (The natural logarithm transformation of price)	105,010	2.8	0.1	2.6	3.0
lny (The natural logarithm transformation of income)	104,941	-0.9	0.9	-5.5	1.2
lnp2 (The square of log transformation of price)	105,010	7.9	0.7	6.9	9.2
lny2 (The square of the log transformation of income)	104,941	1.7	2.3	0.0	30.0
wap2 (The square of price (wap))	105,010	286.2	71.0	193.3	427.3
mce2 (The square of income (mce30))	105,010	0.6	1.2	0.0	11.5
maleratio2 (The square of male ratio)	105,010	0.3	0.2	0.0	1.0
adultratio2 (The square of adult ratio)	105,010	0.9	0.2	0.0	1.0

Annex 2 Prevalence elasticity estimations – variables used in level

	(1)	(2)	(3)	(4)	(5)
VARIABLES	idcig	idcig	idcig	idcig	idcig
weigheted averge price	0.00713	0.115	-0.00690	0.196*	-0.0264***
	(0.00475)	(0.111)	(0.00548)	(0.111)	(0.00479)
weigheted averge price^2		-0.00347		-0.00552*	
		(0.00314)		(0.00314)	
income	0.311***	0.826***	0.833***	0.301***	1.529***
	(0.0246)	(0.0622)	(0.0621)	(0.0249)	(0.0558)
income^2		-0.190***	-0.192***		-0.389***
		(0.0209)	(0.0209)		(0.0200)
household size	-0.0147***	-0.0237***	-0.0242***	-0.0135***	-0.0296***
	(0.00431)	(0.00449)	(0.00447)	(0.00433)	(0.00447)
males in household	1.193***	1.209***	1.209***	1.193***	1.260***
	(0.0346)	(0.0351)	(0.0351)	(0.0346)	(0.0345)
adults in household	-0.307***	-0.357***	-0.358***	-0.306***	-0.701***
	(0.0779)	(0.0780)	(0.0780)	(0.0779)	(0.0781)
insd_ban		0.0242	0.0466*	0.0288	
		(0.0345)	(0.0277)	(0.0344)	
education level 1	-0.805***	-0.742***	-0.741***	-0.803***	-0.965***
	(0.0724)	(0.0727)	(0.0728)	(0.0724)	(0.0721)
education level 2	-0.253***	-0.204***	-0.203***	-0.252***	-0.295***
	(0.0318)	(0.0323)	(0.0323)	(0.0318)	(0.0318)
education level 4	-0.0430	-0.0740***	-0.0749***	-0.0407	-0.114***
	(0.0265)	(0.0267)	(0.0267)	(0.0265)	(0.0265)
education level 5	-0.0531	-0.0990**	-0.0986**	-0.0570	-0.159***
	(0.0401)	(0.0402)	(0.0402)	(0.0401)	(0.0399)
education level 6	-0.0975***	-0.134***	-0.136***	-0.0927***	-0.182***
	(0.0236)	(0.0239)	(0.0239)	(0.0237)	(0.0239)
household ocupation 1	-0.757***	-0.639***	-0.638***	-0.757***	
	(0.111)	(0.111)	(0.112)	(0.111)	
household ocupation 2	-0.755***	-0.672***	-0.670***	-0.757***	
	(0.0253)	(0.0273)	(0.0273)	(0.0253)	
household ocupation 3	-0.398***	-0.315***	-0.314***	-0.400***	
	(0.0393)	(0.0405)	(0.0405)	(0.0393)	
Constant	-1.647***	-2.632***	-1.591***	-3.258***	-1.357***
	(0.111)	(0.947)	(0.113)	(0.946)	(0.112)
Obs.	104,988	104,988	104,988	104,988	104,988
Robust standard errors in parentheses					
*** p<0.01, ** p<0.05, * p<0.1					
		Information criteria			
AIC	99,484	99,388	99,387	99,477	100,090
BIC	99,617	99,550	99,540	99,630	100,205
		Link test			

_hatsq z score	-13.04	-13.33	-13.33	-13.02	-14.82
_hatsq p value	0.0000	0.0000	0.0000	0.0000	0.0000
		LM test prob > chi2			
10 groups	0.0000	0.0000	0.0000	0.0000	0.0000
20 groups	0.0000	0.0000	0.0000	0.0000	0.0000
50 groups	0.0000	0.0000	0.0000	0.0000	0.0000
		Collin			
Mean VIF	1.51	91.27	3.19	95.16	3.27
Det(correlation matrix)	0.1037	0.0000	0.0068	0.0001	0.0201

Annex 3 Prevalence elasticity estimations – variables used in log

	(1)	(2)	(3)	(4)	(5)
VARIABLES	idcig	idcig	idcig	idcig	idcig
ln(weighter average price)	-0.295***	5.438	-0.211**	10.51*	-0.540***
	(0.0943)	(5.376)	(0.0956)	(5.365)	(0.0817)
ln(weighter average price)^2		-0.989		-1.808*	
		(0.941)		(0.940)	
ln(income)	0.398***	0.297***	0.299***		0.465***
	(0.0192)	(0.0271)	(0.0270)		(0.0259)
ln(income)^2		-0.0503***	-0.0500***	-0.127***	-0.0606***
		(0.00959)	(0.00958)	(0.00731)	(0.0101)
household size	-0.0366***	-0.0293***	-0.0296***	0.00148	-0.0366***
	(0.00436)	(0.00454)	(0.00453)	(0.00356)	(0.00451)
males in household	1.235***	1.247***	1.247***	1.254***	1.303***
	(0.0361)	(0.0365)	(0.0365)	(0.0364)	(0.0365)
adults in household	-0.415***	-0.389***	-0.390***	-0.263***	-0.699***
	(0.0780)	(0.0780)	(0.0780)	(0.0775)	(0.0778)
insd_ban	0.0339	0.0116	0.0351	0.0110	
	(0.0282)	(0.0361)	(0.0282)	(0.0361)	
education level 1	-0.536***	-0.452***	-0.451***	-0.427***	-0.543***
	(0.0744)	(0.0764)	(0.0764)	(0.0763)	(0.0762)
education level 2	-0.106***	-0.0751**	-0.0748**	-0.0865***	-0.106***
	(0.0327)	(0.0332)	(0.0332)	(0.0331)	(0.0330)
education level 4	-0.0974***	-0.0904***	-0.0911***	-0.0461*	-0.128***
	(0.0267)	(0.0266)	(0.0266)	(0.0265)	(0.0264)
education level 5	-0.121***	-0.101**	-0.100**	-0.0182	-0.150***
	(0.0398)	(0.0399)	(0.0399)	(0.0391)	(0.0396)
education level 6	-0.171***	-0.137***	-0.138***	-0.0132	-0.183***
	(0.0229)	(0.0237)	(0.0237)	(0.0207)	(0.0236)
household ocupation 1	-0.357***	-0.268**	-0.266**	-0.293**	
	(0.115)	(0.116)	(0.116)	(0.116)	
household ocupation 2	-0.580***	-0.574***	-0.573***	-0.649***	
	(0.0276)	(0.0277)	(0.0277)	(0.0268)	
household ocupation 3	-0.194***	-0.179***	-0.178***	-0.258***	
	(0.0411)	(0.0410)	(0.0410)	(0.0400)	
Constant	-0.0910	-8.450	-0.414	-16.63**	0.858***
	(0.278)	(7.651)	(0.285)	(7.627)	(0.258)
Obs.	104,919	104,919	104,919	104,919	104,919
Robust standard errors in parentheses					
*** p<0.01, ** p<0.05, * p<0.1					
		Information criteria			
AIC	99,124	99,100	99,099	99,216	99,597
BIC	99,267	99,262	99,252	99,369	99,712
		Link test			
_hatsq z score	-13.1	-11.94	-11.94	-11.18	-9.42

_hatsq p value	0.0000	0.0000	0.0000	0.0000	0.0000
		LM test prob > chi2			
10 groups	0.0000	0.0000	0.0000	0.0000	0.0000
20 groups	0.0000	0.0000	0.0000	0.0000	0.0000
50 groups	0.0000	0.0000	0.0000	0.0000	0.0000
		Collin			
Mean VIF	1.66	724.68	2.42	767.58	2.56
Det(correlation matrix)	0.0506	0.0000	0	0	0.029

Annex 4 Intensity elasticity estimations – variables used in level

	(1)	(2)	(3)	(4)	(5)
VARIABLES	ncig	ncig	ncig	ncig	ncig
weigheted averge price	-0.0122***	-0.0667	-0.0255***	0.0253	-0.0259***
	(0.00262)	(0.0586)	(0.00288)	(0.0585)	(0.00259)
weigheted averge price^2		0.00117		-0.00118	
		(0.00166)		(0.00165)	
income	0.465***	1.044***	1.041***	0.460***	1.127***
	(0.0135)	(0.0306)	(0.0304)	(0.0136)	(0.0273)
income^2		-0.208***	-0.207***		-0.229***
		(0.00969)	(0.00965)		(0.00903)
household size	0.0291***	0.0182***	0.0184***	0.0296***	0.0159***
	(0.00222)	(0.00226)	(0.00225)	(0.00224)	(0.00221)
males in household	0.240***	0.267***	0.267***	0.240***	0.262***
	(0.0198)	(0.0194)	(0.0194)	(0.0198)	(0.0193)
adults in household	0.237***	0.180***	0.181***	0.236***	0.158***
	(0.0436)	(0.0430)	(0.0430)	(0.0437)	(0.0421)
insd_ban		0.0387**	0.0311**	0.0417**	
		(0.0185)	(0.0152)	(0.0187)	
education level 1	-0.289***	-0.222***	-0.222***	-0.287***	-0.256***
	(0.0561)	(0.0538)	(0.0539)	(0.0565)	(0.0518)
education level 2	-0.168***	-0.111***	-0.111***	-0.167***	-0.129***
	(0.0168)	(0.0165)	(0.0165)	(0.0168)	(0.0164)
education level 4	0.0308**	0.00221	0.00247	0.0316**	-0.00281
	(0.0130)	(0.0131)	(0.0131)	(0.0131)	(0.0130)
education level 5	-0.00275	-0.0434**	-0.0435**	-0.00521	-0.0508**
	(0.0199)	(0.0200)	(0.0200)	(0.0199)	(0.0199)
education level 6	0.0106	-0.0282**	-0.0276**	0.0133	-0.0353***
	(0.0115)	(0.0117)	(0.0117)	(0.0116)	(0.0116)
household ocupation 1	-0.262***	-0.146*	-0.147*	-0.264***	
	(0.0814)	(0.0766)	(0.0766)	(0.0812)	
household ocupation 2	-0.150***	-0.0600***	-0.0606***	-0.151***	
	(0.0127)	(0.0133)	(0.0133)	(0.0126)	
household ocupation 3	-0.220***	-0.131***	-0.131***	-0.222***	
	(0.0244)	(0.0244)	(0.0243)	(0.0245)	
Constant	5.630***	6.009***	5.658***	5.305***	5.670***
	(0.0636)	(0.503)	(0.0628)	(0.503)	(0.0621)
Obs.	20,655	20,655	20,655	20,655	20,655

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Information criteria					
AIC	297,363	297,212	297,210	297,362	297,221
BIC	297,475	297,347	297,337	297,489	297,221
Link test					
_hatsq z score	-1.62	-0.69	-0.69	-1.38	-2.5
_hatsq p value	0.1050	0.4920	0.4910	0.1660	0.0120
Collin					
Mean VIF	1.51	91.27	3.19	95.16	3.27
Det(correlation matrix)	0.1037	0.0000	0.0068	0.0001	0.0201

Annex 5 Intensity elasticity estimations – variables used in log

	(1)	(2)	(3)	(4)	(5)
VARIABLES	ncig	ncig	ncig	ncig	ncig
ln(weighter average price)	-0.432***	-1.040	-0.507***	8.930***	-0.471***
	(0.0506)	(2.822)	(0.0498)	(2.928)	(0.0440)
ln(weighter average price)^2		0.0931		-1.516***	
		(0.494)		(0.512)	
ln(income)	0.436***	0.512***	0.512***		0.519***
	(0.0125)	(0.0145)	(0.0144)		(0.0141)
ln(income)^2		0.0375***	0.0374***	-0.0832***	0.0366***
		(0.00647)	(0.00647)	(0.00711)	(0.00632)
household size	0.0217***	0.0154***	0.0154***	0.0735***	0.0145***
	(0.00230)	(0.00226)	(0.00225)	(0.00179)	(0.00221)
males in household	0.323***	0.307***	0.307***	0.313***	0.305***
	(0.0191)	(0.0193)	(0.0193)	(0.0206)	(0.0193)
adults in household	0.182***	0.158***	0.158***	0.406***	0.156***
	(0.0429)	(0.0425)	(0.0425)	(0.0448)	(0.0418)
insd_ban	0.0294*	0.0326*	0.0304**	0.0171	
	(0.0152)	(0.0190)	(0.0151)	(0.0199)	
education level 1	-0.0732	-0.110**	-0.110**	-0.126**	-0.111**
	(0.0478)	(0.0503)	(0.0503)	(0.0539)	(0.0496)
education level 2	-0.0295*	-0.0466***	-0.0466***	-0.0923***	-0.0512***
	(0.0169)	(0.0170)	(0.0170)	(0.0199)	(0.0171)
education level 4	7.37e-05	-0.00550	-0.00545	0.0730***	-0.00659
	(0.0131)	(0.0130)	(0.0130)	(0.0135)	(0.0130)
education level 5	-0.0299	-0.0452**	-0.0452**	0.0978***	-0.0451**
	(0.0198)	(0.0199)	(0.0199)	(0.0199)	(0.0199)
education level 6	-0.00806	-0.0319***	-0.0318***	0.177***	-0.0340***
	(0.0116)	(0.0117)	(0.0116)	(0.0107)	(0.0116)
household ocupation 1	0.0765	0.0101	0.00974	-0.0628	
	(0.0694)	(0.0699)	(0.0699)	(0.0798)	
household ocupation 2	-0.00581	-0.00702	-0.00715	-0.165***	
	(0.0141)	(0.0137)	(0.0137)	(0.0149)	
household ocupation 3	-0.0368	-0.0524**	-0.0525**	-0.203***	
	(0.0248)	(0.0246)	(0.0246)	(0.0254)	
Constant	7.225***	8.270**	7.512***	-7.708*	7.444***
	(0.155)	(4.022)	(0.152)	(4.168)	(0.142)
Observations	20,647	20,647	20,647	20,647	20,647
Robust standard errors in parentheses					
*** p<0.01, ** p<0.05, * p<0.1					
		Information criteria			
AIC	297,026	297,010	297,008	297,478	297,003
BIC	297,145	297,145	297,135	297,605	297,099
		Link test			
_hatsq z score	5.61	1.98	1.97	5.93	1.88
_hatsq p value	0.0000	0.0480	0.0490	0.0000	0.0610
		Collin			
Mean VIF	1.66	724.68	2.42	767.58	2.56
Det(correlation matrix)	0.0506	0.0000	0.0087	0.0000	0.029

Annex 6 Detailed model estimation of the two-part model at income group level

	Group 1		Group 2		Group 3	
VARIABLES	idcig	ncig	idcig	ncig	idcig	ncig
lnp	-0.942***	-0.387***	-0.697***	-0.613***	-0.123	-0.551***
	(0.168)	(0.0886)	(0.155)	(0.0769)	(0.146)	(0.0735)
lny	1.228***	0.796***	0.478***	0.556***	0.327***	0.503***
	(0.183)	(0.120)	(0.172)	(0.0834)	(0.0488)	(0.0257)
lny2	0.131***	0.0993***	-0.114	0.00232	0.0503	0.0499*
	(0.0420)	(0.0297)	(0.101)	(0.0499)	(0.0586)	(0.0282)
hsize	0.0358*	-0.0172	-0.104***	-0.0133***	-0.0150***	0.0229***
	(0.0187)	(0.0106)	(0.00966)	(0.00480)	(0.00535)	(0.00264)
maleratio	1.641***	0.238***	0.960***	0.304***	0.909***	0.434***
	(0.0525)	(0.0292)	(0.0655)	(0.0310)	(0.0980)	(0.0494)
adultratio	-1.021***	-0.0497	-0.804***	-0.0288	-0.185	0.236***
	(0.202)	(0.112)	(0.123)	(0.0647)	(0.123)	(0.0630)
deduc1	-0.505***	-0.133***	0.161	0.350		
	(0.0800)	(0.0434)	(0.305)	(0.268)		
deduc2	-0.103**	-0.0402*	-0.0569	-0.0274	0.318*	-0.0459
	(0.0414)	(0.0214)	(0.0588)	(0.0287)	(0.164)	(0.0930)
deduc4	-0.0295	0.0582	-0.147***	-0.0150	-0.131***	-0.0263
	(0.0808)	(0.0404)	(0.0407)	(0.0203)	(0.0387)	(0.0192)
deduc5	0.0266	0.147**	-0.307***	-0.00730	-0.0897*	-0.0880***
	(0.145)	(0.0746)	(0.0747)	(0.0368)	(0.0508)	(0.0254)
deduc6	-0.0540	0.168***	-0.207***	-0.0595***	-0.152***	-0.0483***
	(0.119)	(0.0544)	(0.0424)	(0.0205)	(0.0313)	(0.0154)
o.deduc1					-	-
Constant	2.643***	7.787***	1.891***	8.167***	-0.800*	7.494***
	(0.644)	(0.349)	(0.503)	(0.250)	(0.449)	(0.230)
Observations	34,937	4,146	34,987	7,538	34,995	8,963

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1