Market Competition and the Extensive Margin Analysis:

(**MCEMA**) Stata module

Table of Contents

1 Introduction

This User Manual presents the **MCEMA** Stata module, a microsimulation tool that builds on previous efforts to estimate the distributional effects of changes in market structure (e.g., the WELCOM tool which allows to model household level welfare effects of moving from concentrated to competitive markets) due to competition policy or regulatory reforms. However, instead of focusing on the welfare impacts among current users (the traditional approach in the WELCOM tool), the **MCEMA** module expands the analysis to cover the impact among new consumers (the extensive margin or the take up of a good), particularly those users who were previously priced out of the market. **MCEMA** proposes two complementary empirical strategies to estimate both the welfare effects of a price decrease on current users, and those previously priced out of the market. It first focuses on measuring the welfare effects of a price reduction in a specific market using the **WELCOM** (Welfare and Competition) tool. The second approach models consumer behavior by estimating the probability of adoption based on sociodemographic characteristics, and subsequently, estimates the marginal increase of new consumers and their monetary gains (relying on the welfare gains calculated in the first part and the change in price).

It is important to distinguish between the welfare effects among current consumers, and the welfare impacts for new ones. The welfare effects for the current users or consumers is estimated as a function of their observed expenditure in the good (i.e., share of spending in the good relative to total expenditure), their consumption elasticity and the expected change in prices. In contrast, the proposed empirical approach to estimate the new consumer's change in welfare (i.e., the "extensive margin of consumption") can be divided in three steps. First, we estimate a probabilistic model of the likelihood of having positive consumption of the good based on observable characteristics. We use such model to estimate the change in the probability of consumption after a change in prices (or an equivalent change in income) due to increased competition. Second, the expected expenditures of new users/consumers are estimated. The **MCEMA** module allows to estimate the expected expenditure of new consumers using six alternative models. These approaches include simple linear estimators, such as average expenditures by population groups, random imputation techniques, non-linear regression models and random imputation of residuals. Third, we estimate the change in welfare as the product of the change in the probability of consumption and the expected consumption. In sum, MCEMA allows to calculate the aggregate change in welfare among current and new users.

One of the main attributes of **MCEMA** is that it provides a flexible user interface with minimum data requirements, such as the households' consumption of different goods and services from a typical household survey. **MCEMA** was developed by the Global Solutions Group on Markets and Institutions for Poverty Reduction and Shared Prosperity at the Poverty and Equity Global Practice as part of a larger engagement to assess the distributional effects of markets and competition. The first section of this document discusses the

context and main data requirements needed to implement the proposed approach. Then, we show the theoretical model and econometric specifications underlying the estimation of welfare. Finally, we discuss an example of how to apply this approach using a new module of the WELCOM user interphase and relying on real data from Mexico's latest household survey.

We start by assuming that we have a household income or consumption survey with information on households and, potentially, on prices, incomes, and other household characteristics. The main variables needed for the model are (required variables (*)):

- *A dummy variable indicating if the household/individual (denoted by h) has a positive or null consumption on the good or service of interest.
- *Per capita expenditure on the good or service of interest by the household/individual h, denoted by e_h.
- * The welfare aggregate (income or consumption expenditure) before the change in prices, denoted by w_h.
- * The—proportional—price change dp_h or the—proportional—welfare changes dw_h. These variables are calculated using the MCWEL module of WELCOM.
- A set of household characteristics (covariates) that explain the use or consumption of the good of interest (denoted by X_h).

If we denote by π (π_q) the proportion of users at the population or income level (e.g., quintile level), our aim is to assess the change in that proportion implied by the change in price and, indirectly, the change in the welfare aggregate.

Notice that, if prices are available, we will assume that nominal income remains constant. Otherwise (if prices are not directly observable in the data), we keep prices constant, and use instead the income variation that is expected to generate an equivalent impact on consumption. Furthermore, **MCEMA** relies on the assumption that prices of goods and services will decrease after increasing competition.

This manual is organized as follows, in section 2 we introduce the three aforementioned key steps of the proposed empirical approach to estimate the new consumer's change in welfare: 1) Probabilistic model of the likelihood of consumption; 2) estimate the expected expenditures of new entrant; and 3) Estimate the change in welfare. In section three we go over the installation and implementation of the **MCEMA** module, explaining three different examples related to the telecommunications market in Mexico.

2 Theoretical models and econometric specifications

Two main elements are required to estimate the change in welfare from the extensive margin of consumption (new users or entrants). The first is to identify entrants, or at least, to estimate their proportion *vis-à-vis* current users. The second involves estimating the expected expenditures of the new users or consumers.

The approach we propose loosely follows the literature on two-part models, such as Dow and Norton (2003), Nargis (2013), Cragg (1971), and Heckman (1976).[1]

Two options are available to estimate new entrants, each relying on the relative price change (or the equivalent change in income): 1) Probabilistic model: probability regression (e.g. probit) that estimates the proportion of new entrants in a market based on the aggregate (or sub-group, if applicable) change in probability; 2) Exogenous model: the number of entrants or total users is taken as a given, for instance, as the output of the expected policy (e.g. target of coverage). In subsection 2.1, we discuss how to estimate the proportion of entrants using six alternative probability models using household survey data. Then, in subsection 2.2, we introduce the exogenous model where number of entrants is taken as an input for the model. In subsection 2.3., we propose six different approaches to estimate expenditures of entrants after a change in price due to competition. Next, subsection 2.4 discusses the Taylor approximation approach to estimate change in welfare. Special attention is given to the case of new entrants (i.e., extensive margin of consumption). In subsection 2.5, we discuss the case when prices are not available and finally in subsection 2.6, we address the limitations of the module.

2.1 Probit models and proportion of entrants

MCEMA offers users six alternative probability models to estimate new entrants, giving users a flexible set of tools to best model the decision to of individuals or households to enter a market. Each model is subsequently discussed in this section, with their respective specifications and data requirements.

In the first proposed model (see **M1** below), we exploit the variability of prices or incomes at the household level to estimate a *probability unit* or probit model.[2] Considering the assumption of a normal distribution of the error term in probit models, the probability of use/consumption can be written as

$$Prob(d_h) = f(\log(I_h), \log(p_h), X_h) \dots \boldsymbol{M1}. \quad (1)$$

For the sake of simplicity, we denote all explanatory variables with X and omit the h index in the rest of this section. In the case where the prices' microdata are not available, the equation (1) is reduced to: $Prob(d_h) =$

[1] Note that the literature of Two-Part Heckman models focuses on estimating population level statistics, such as the change in average quantity, the price-elasticity, etc. In contrast, the proposed framework will focus on the distributional impact of price changes on well-being. Using a Taylor approximation approach, the estimation of the impact for current users will not require additional estimations, since the necessary information on their expenditures and price changes is observed. Inversely, to estimate the impact on the extensive margin, the consumption decision is modeled, conditional to positive quantities of the good being consumed.

[2] In Stata, the stepwise prefix can be used to automatically perform the selection of explanatory variables according to their significance levels.

$f(\log(I_h), X_h)$.[3] If X denotes all covariates including the log of income ($\log(I_h) = \log(income_h)$) and potentially prices, we can write

$$Prob(d = 1|X) = \Phi(X\beta). \tag{2}$$

Alternatively,

$$Prob(use = 1) = \Phi\left(\beta_0 + \beta_p \ln(Price) + \cdots + u\right) \tag{3}$$

where $\Phi(.)$ denotes the normal cumulative distribution function (CDF). Then, the probabilistic elasticity is given by

$$\varepsilon_p = \frac{\partial Prob\,(.)}{\partial Price} * \frac{\overline{Price}}{\overline{Prob}} \tag{4}$$

where $\overline{Price}$ and $\overline{Prob}$ refer to the average price and average probability of having positive consumption at the population level (individual of reference), respectively.

Taking derivatives and using the chain rule on the Probit specification it is possible to rewrite equation (4) as[4]

$$\begin{aligned} \varepsilon_p &= \frac{\partial \Phi(XB)}{\partial (XB)} \frac{\partial (XB)}{\partial \ln(Price)} \frac{\partial \ln(Price)}{\partial (Price)} \frac{\overline{Price}}{\overline{Prob}} \\ &= \phi(.)\beta_p \left(\frac{1}{\overline{Price}}\right) \frac{\overline{Price}}{\overline{Prob}} \end{aligned} \tag{5}$$

where $\phi(.)$ denotes the probability density function (pdf) of the normal distribution. Thus:

$$\begin{aligned} \varepsilon_p &= \frac{\partial Prob}{\partial Price} \frac{\overline{Price}}{\overline{Prob}} \\ &= \frac{\partial \Phi(XB)}{\partial (XB)} \frac{\partial (XB)}{\partial \ln(Price)} \frac{\partial \ln(Price)}{\partial (Price)} \frac{\overline{Price}}{\overline{Prob}} \\ &= \phi(.) \frac{\beta_p}{\overline{Prob}}. \end{aligned} \tag{6}$$

To estimate the probabilistic price-elasticity, it is not necessary to observe the final prices or final incomes (i.e. the variability of price or incomes between households is exploited). The absolute change in probability is denoted by

[3] Please refer to annex for full derivation.
[4] For more information on price elasticity and the Two-Part/Heckman models, see for instance, Saha et. al (1997).

$$
\begin{aligned}
A_p &= \frac{\partial Prob}{\partial Price} dprice \\
&= \frac{\partial \Phi(XB)}{\partial (XB)} \frac{\partial (XB)}{\partial \ln(Price)} \frac{\partial \ln(Price)}{\partial (Price)} dprice \\
&= \phi(\bar{X}B) * \beta_p * dprice \frac{dprice}{\overline{Price}}.
\end{aligned} \tag{7}
$$

If the price variable is not available, **M1** would take the form

$$
\begin{aligned}
A_I &= \frac{\partial Prob}{\partial Income} dincome \\
&= \frac{\partial \Phi(XB)}{\partial (XB)} \frac{\partial (XB)}{\partial \ln(Income)} \frac{\partial \ln(Income)}{\partial (Income)} dincome \\
&= \phi(\bar{X}\beta) * \beta_I * \frac{dIncome}{\overline{Income}}.
\end{aligned} \tag{8}
$$

In addition, we can also allow for interactions between the price and the quintile or decile dummies to estimate the price coefficients by income partitions, thus resulting in our second suggested model (**M2**).

The distributional effects of competition policies are not necessarily uniform across the income or consumption distribution. As expenditure shares on a specified item might vary across the distribution, it is important to control for the consumer heterogeneity to assess the welfare effects at a granular level. This model allows the user to calculate price elasticities by welfare levels and determine the distributional effects of price variations across the income distribution. For the third proposed model (**M3**), we allow the user to interact price or income with population groups, such as by region. For the model (**M4**), we allow the user to interact price or income with the quintile/decile partition variable and another categorical variable of interest (population groups) such as rural/urban area. In the equation below, we allow for interactions of the log price variable $\ln(Price)$ with quintiles or deciles or any other group variable (rural/urban, for instance).

Then, the model thus takes the form

$$
Prob(use = 1) = \Phi\left(\beta_0 + \sum_{g=1}^{G} \beta_{p,g} \ln(Price) * I_g \ + \cdots + u\right) \ldots\ldots \boldsymbol{M2, M3, M4} \tag{9}
$$

where I_g is an indicator variable that takes the value of one if the household belongs to the group *g* and, otherwise, zero. More precisely, for $\boldsymbol{M2}$ we will have G income partition groups, for instance if we select a quintile income partition, we will have 5 population groups. For $\boldsymbol{M3}$, we will have G depending on the population group variable. For instance, for the *Rural/Urban* variable, $G = 2$. For $\boldsymbol{M4}$, the number of

population groups G is equal to the number of group income partitions times the number of population groups (e.g., we have $G = 10$ when using the quintile income partition and the *Rural/Urban* population group variable).

Equation (9) is equivalent to the following form when price information is not available:

$$Prob(use = 1) = \Phi\left(\beta_0 + \sum_{g=1}^{G} \beta_{I,g} \ln(Income) * I_g \ + \cdots + u\right) \ldots\ldots \boldsymbol{M2, M3, M4} \quad (10)$$

In the rest of this document, we denote $\ln(Price) * I_g$ by $\ln(Price)_g$, such that

$$\frac{\partial Prob}{\partial \ln(Price)_g} \frac{\partial \ln(Price_g)}{\partial Price_g}\Big|_{Price_g = \overline{Price}_g} = \frac{\phi(.)\beta_{p,g}}{\overline{Price}_g} \quad (11)$$

We define the price-elasticity within the group *g* by

$$\varepsilon_{p,g} = \frac{\partial \text{Prob}_g(.)}{\partial Price_g} \frac{\overline{Price}_g}{\overline{Prob}_g}. \quad (82)$$

Then, we can write

$$\begin{aligned} \varepsilon_{p,g} &= \frac{\partial Prob_g}{\partial Prob} \frac{\partial Prob}{\partial \ln(Price_g)} \frac{\partial \ln(Price_g)}{\partial Price_g} \frac{\overline{Price}_g}{\overline{Prob}_g} \\ &= \frac{1}{\varphi_g} \varphi_g \phi(.) \beta_{p,g} \frac{1}{\overline{Prob}_g} \\ &= \phi(.) \frac{\beta_{p,g}}{\overline{Prob}_g} \end{aligned} \quad (13)$$

where φ_g is the population share of group *g*.

In models 2 ,3 and 4, we assume that we run one population model. However, it is also possible to split the data by groups of interests, modelling each group independently. For example, the fifth model (**M5**), we assume that the user would like to run the models by population groups (e.g., Rural and then Urban). Thus, can be written as

$$Prob_g(use = 1) = \Phi_g\left(\beta_{0,g} + \beta_{p,g} \ln(Price) * I_g \ + \cdots + u_g\right). \quad \boldsymbol{M5} \quad (14)$$

The last model, **M6**, is similar to model **M5,** but where price or income variable interacts with the variable *income partition groups*. When the price variable is not available, model M5 would take the following form

$$Prob_g(use = 1) = \Phi_g\left(\beta_{0,g} + \beta_{I,g} \ln(income) * I_g \ + \cdots + u_g\right) \quad \boldsymbol{M5} \quad (15)$$

Based on equation (11) and estimating for each group g, we find that

$$\varepsilon_{p,g} = \phi_g(.) * \frac{\beta_{p,g}}{Prob_g}.$$

After estimating the probit models, we calculate the proportion of entrants in the population, assuming that **the change in the probability of the reference individual**, for whom the determinants have the average values, **represents the proportion of new entrants**. One can also evaluate the change in the probability of the reference individual as

$$A_p = \phi(\bar{X}\beta) * \beta_p * \frac{dprice}{\overline{Price}} \quad (16)$$

where if we do not have the price variable, equation (16) would thus take the following form:

$$A_I = \phi(\bar{X}\beta) * \beta_I * \frac{dIncome}{\overline{Income}}. \quad (17)$$

However, due to nonlinearities, the formula above may be less precise than alternative approached. It is also possible to compute the difference between averages of the predicted probabilities with initial and final prices (which could lead to more accurate estimates), resulting in the estimated proportion of entrants. Formally, the change in the probability of use of the household h is

$$A_{p,h} = Prob_h(use = 1|X'_h) - Prob_h(use = 1|X_h). \quad (18)$$

At population level, the expected change is

$$A_p = E[A_{p,h}]. \quad (19)$$

In other words, the expected change in probability is equal to the expected predicted probability after the increase in price (or income) (X'), minus the expected probability under the initial values of prices (or incomes). The advantage of this second method is that we do not need to evaluate the density for the reference individual, which would provide more accurate results, as the two measurements will capture the main part of the change in probability. Notice that the computation of the proportion of entrants can also be done by population groups (for instance, rural/urban). Equation (19) would take the form $A_I = E[A_{I,h}]$ in the case that the price variable is not available.

2.2 Exogenous total number of users and identification of entrants

In occasions, the user has exogenous information about the number of entrants or on total users, either at the national level or disaggregated by region. This approach is particularly useful for those users that would like to estimate the expected welfare impact of specific goals set out by policies or reforms, such as the rollout of telecommunications infrastructure targeting a given number of new users or the welfare change due to the entry of new consumers due to a competition enhancing reform.

In this scenario, l_h denotes the total of eligible members of sampled household h that consumes or can consume the good of interest. Also, sw_h denotes the sampling weight of household h. The sum of the product $s_h = sw_h . l_h$ is thus an estimate of the total eligible population. If we denote the total eligible population in the region g by L_g, we can write $L_g = \sum_h s_{h,g}$. In addition, one can estimate the total current users in the region g as

$$O_g = \sum_h s_{h,g} I[e_h > 0], \quad (20)$$

where $I[True]$= 1 and zero otherwise, and component e_h denotes the expenditures on the good of interest of the household h. If we denote the total entrants in the region g by E_g, we can write the following condition

$$T_g = O_g + E_g \leq L_g. \quad (21)$$

This means that total users must be equal or higher than the final users including entrants. The eligibility of the user depend on different factors, such as coverage in the region, age, gender, etc. Assume that we dispose the total number of entrants or that of total final users by region. When using household surveys, it is possible to estimate the component O_g. In addition, if there is exogenous information available on T_g instead of that on E_g, one can estimate the E_g as: $E_g = \max(0;\ T_g - E_g)$. In this latter formula, we impose the non-negativeness of the number of entrants.

After estimating the total number of entrants, we need to identify them among the sampled households. In the proposed approach, we sequentially include those with highest probabilities. The identification can be done with the following steps:

a- Computing the proportion of entrants among the non-users within the group g, and this based on the exogenous information (ρ_g);

b- Modelling and estimating the probability of use ($Pr_{g,h}$: probability of use of household h) (see subsection 2.1 on the different suggested Probit models);

c- For non-users, generating the normalized probability times the proportion of increase as: $\kappa_{g,h} = \rho_g\left(Pr_{g,h}/\overline{\overline{Pr_g}}\right)$ where $\overline{\overline{Pr_g}}$ is the average probabilities of non-users in group g, and $\kappa_{g,h}$ is equal to zero for current users.

Let $sw_{g,h}$ be the sampling weight of household h. We can check that: $\sum_h \kappa_{g,h}\, sw_{g,h} = \rho_g \sum_h sw_{g,h}$. Thus, for a given group g, entrant subgroup is composed of the non-users of that group when they with weights $\kappa_{g,h}\, sw_{g,h}$.

2.3 Estimation of expenditures of entrants

In addition to estimating the proportion of entrants, we also need to estimate expenditures of the relevant product among new consumers or users. **MCEMA** offers six alternative ways to model the expenditures of new entrants, offering practitioners a diverse toolkit that can be tailored to overcome any set of constraints that could arise from household surveys.

The estimation of consumption among new entrants is based on the probability or likelihood that they consume, thus consumption is modeled as $E[e \mid e > 0]$ where e denotes expenditure. The underlying premise is that, at the margin, those who start to consume the good are the better-off among the non-consumers and, thus, they will have a similar expenditure model to that of the current consumer group.[5] The non-randomness of the decision to consume is addressed assuming that, at the margin, the new consumers of the good share similar observable characteristics with current consumers. For the estimation of the expenditures, we will rely on imputations.

Imputation techniques can be grouped into two main categories: Single Value Imputation (SI) and Multiple Value Imputation (MI), each approach has advantages and disadvantages. One of the main advantages of SI is that the sample preserves its original size. In addition, the MI is more demanding computationally.

The extent of the bias related to the prediction of expenditures of entrants will depend on many factors including: (i) the randomness (or not) of decision to consume[6]; (ii) the imputation method and the information available in the data set to predict expenditures; (iii) the proportion of missing values.

The first element (i) can arise when current users' characteristics are significantly different from those of non-users. The proposed approach focuses on the well-being of entrants, and not on that of all non-users. It follows that, at the margin, if the size of entrant group is moderate, they could be expected to have similar characteristics to the non-users. Concerning the second factor (ii), the nature and characteristics of the data will partially determine the convenience of the methods. For the third factor (iii), regardless of the imputation method used, the higher the proportion of estimated entrants the higher the bias in the prediction of expenditures. For

[5] We assume that the unobserved expenditures of entrants can be treated similarly to that of missing values, where these missing values are the unobserved expenditures of entrants.
[6] This refers to the extent to which use or consumption is selective or non-random.

instance, if the number of users is only two, but the number of entrants is ten, the prediction based on only the two non-missing values from the two users is likely to be biased. Thus, in presence of large proportion of entrants, the user must be aware of the nature of the assumptions underlying the results (for instance, see Jakobsen et. al (2017)). In the following sub-sections, each of the six model is discussed in detail, particularly juxtaposing the advantages and disadvantages of each.

2.3.1 Imputation using average expenditures (Unconditional Mean Imputation)

In the first method we attribute the average of current users' expenditures to new entrants by population groups (PSU, percentile, etc). However, this method does not consider the variability of expenditures across households. Formally, we denote the average expected expenditures on good i of entrants of group g (e.g., decile 2) by $e_{g,i}^{entrant}$ and those of the current consumer of the same group by $e_{g,i}^{old}$. Thus, we have that

$$e_{g,i}^{entrant} = e_{g,i}^{old} = \frac{\sum_{i=1}^{N} I[e_{g,i,h}>0] e_{g,i,h}}{\sum_{i=1}^{N} I[e_{g,i,h}>0]}. \quad (22)$$

The main limitation of this approach is the low variability it produces within each group of entrants (Enders, 2010; Eekhout et al, 2013). Empirically, this approach can be useful, for instance, if we do not have other explanatory variables to estimate the expenditures model, or if the variables available to model expenditure have an overall weak predicting power, and it also offers the advantage that it is easy to communicate.

2.3.2 Random imputation using expenditures of current users

Alternatively, we can use the hot deck imputation technique to assign the expenditure value of a random selected current user to the (new) entrant from a subsample of households with similar characteristics (for instance, belonging to the same population group).

The **MCEMA** module uses the hotdeck command for Stata developed by Mander and Clayton (2000) (see also Siddique and Belin (2008)). Operationally, the Stata user can indicate the desired number of non-missing values to replace each missing value. As a first step, the hotdeck command makes different data files from the non-missing values. If, for instance, we decide to replace each missing value with two non-missing values, the command draws randomly two sub-samples—without replacement—from the no missing data.[7]

[7] Further, if our aim is to estimate coefficients of a given linear model, the **hotdeck** command run the regressions with the different the generated data (no missing data plus a given random drawn sub-sample). After that, the averages of coefficients of the different regressions are computed. For the estimation of the standard error of coefficients, see Rubin (1987). In our study case, and for the second proposed **mcema** method, we will be limited to the first stage of the hot deck imputation for the one random imputation.

An advantage of the hotdeck approach is that it constrains the imputed values to those of current users by population groups. Another advantage is that it adds variability, due to its random component, which is crucial to accurately estimate standard errors (see for instance, Mander and Clayton (2000) and Andridge (2010), and the help of the Stata command **hotdeck**.). As indicated in subsection 2.2.1, imputation based on a fixed predicted value, such as the average per PSU, tends to reduce the overall variability of the estimations.

Formally, expenditures of current users *vis-à-vis* new entrants are imputed such that:

$$e_{g,i,h}^{entrant} = e_{g,i,Random}^{old} \tag{23}$$

Equation (23) indicates that the expected expenditures of entrant h in group g are simply the expenditures of a random selected current consumer, when the latter is selected from the same group g. A disadvantage of this second method is that it does not consider the need of households according to their characteristics, and the random process can attribute higher levels of expenditures, even by PSUs, to a poor entrant.

2.3.3 Imputation based on expenditure models and linear regressions (regression/conditional mean imputation)

Single-demand models are characterized by using a single equation model to predict expenditures. Starting from the approach proposed by Working-Leser model (see Working (1943) and Leser (1963), Muellbauer (1999)), and where the expenditure share model was

$$s_i = \beta_0 + \beta_1 \log(I) + \sum_j \beta_j \log(p_j) + \sum_k \beta_k X_k + \varepsilon_i \tag{24}$$

where s_i denotes the expenditures share of good i. An updated Working-Leser model that enables to easily compute the income elasticity is:

$$\log(e_i) = \beta_0 + \beta_1 \log(I) + \sum_j \beta_j \log(p_j) + \sum_k \beta_k X_k + \varepsilon_i \tag{25}$$

Thus, the method consists in estimating a linear regression, and then predicting expenditures based on a set of socio-demographic characteristics. The estimation can be done by population groups, as rural/urban or by decile groups. Note that, **MCEMA** enables the user to select among different functional forms. They can be for instance:

- $\log(e_i) = \beta_0 + \beta_1 \log(I) + \sum_j \beta_j \log(p_j) + \sum_k \beta_k X_k + \varepsilon_i$
- $e_i = \beta_0 + \beta_1 I + \sum_j \beta_j \log(p_j) + \sum_k \beta_k X_k + \varepsilon_i$
- $e_i = \beta_0 + \beta_1 \log(I) + \sum_j \beta_j \log(p_j) + \sum_k \beta_k X_k + \varepsilon_i$

When the price of the services is not available, the Working-Leser model is reduced to: $\log(e_i) = \beta_0 + \beta_1 \log(I) + \sum_k \beta_k X_k + \varepsilon_i$. Typically, expenditures, prices and incomes follow log-normal distributions, and this justifies their log-transformation. However, and depending on the study case and nature of data, the user can select the appropriate functional form. An advantage of using regression models is to consider real needs according to household characteristics. However, when we only use the predicted part, we likely underestimate the variability of the estimated expenditures, due to omitting the residual components. Also, this method will increase the associations between variables because it imputes values that are perfectly correlated with one another (see also Craig Enders book "Applied Missing Data Analysis" (2010)).[8]

2.3.4 Imputation with expenditures model and random imputation of residuals (Stochastic Regression Imputation)

As indicated in subsection 2.3.3, among the disadvantages of using the deterministic components from linear regressions to implement imputations is the omission of the residual component and, subsequently, the reduction of the estimated variability. In contrast with the method discussed in subsection 2.3.1, imputing missing values using regression estimates will overestimate the correlations between imputed and predicted values.

To address this, and following the literature of missing imputation, we suggest to randomly impute the residual component using the hotdeck approach (consult the subsection 2.3.2 for more information on the hot deck method). The hotdeck command can also be used only to perform the random imputation of the residual and can be implemented by population (e.g., percentile) subgroups.[9] This relies on the assumption of normal distribution of the residuals, with zero mean and the predicted standard deviation. Using predicted residuals is a better approach when many observations are available, and the distribution of residuals is unknown.

However, the imputation approach also has shortcomings. For instance, it may lead to poor results if the predictive power of the model is poor or if the data is heteroscedastic. To overcome these drawbacks, other complementary computation techniques are included as part of the **MCEMA** module. First, by performing the stochastic regression imputation by population groups (deciles for instance), we aim to reduce the heteroscedasticity problem. Also, to guarantee (at least) positive predicted values of expenditures, the random

[8] Usually, we do not have information on prices of the other goods or even that of the good of interest. Obviously, this will imply strict restrictions on the demand model, and even, on the demand system (see for instance Young (1982)).
[9] In several ocassions, for missing data imputation, instead of using the predicted residuals, we randomly generate them with computational algorithms.

imputation of the residual is performed more than once. In each iteration, the random imputation is done for only the remaining non-plausible values (negative values).[10]

2.3.5 Imputation based on expenditures models and quantile regressions

Theoretical and empirical literature on expenditure models confirm the importance of income and prices as relevant determinants in expenditure models. However, the response of consumers after a change in income or in price (beta parameters) may vary according to the level of well-being (income). In this case, the quantile regression model, and more precisely, the percentile weight regression (PWR) model (see Araar 2016)) can be helpful to refine the estimates. Indeed, the Araar's (2016) approach enables to estimate for a given percentile/quantile of interest, and this even if the percentile-order is based on another variable rather the dependent variable of the model (for instance, income instead of expenditures on the good of interest). Compared to the OLS regression, the PWR is relevant when we focus on the impact by level of well-being/income.

In some words, the idea behind the PWR approach is to attribute large weights for those with levels of incomes close to that of the percentile of interest and low weights for those with farther levels of incomes. For this end, we start by constructing a normal Gaussian distribution around the percentile of interest using the kernel method. The bandwidth size help control for the importance attributed to observations around of the percentile of interest. The steps of the PWR model are:

- Estimate the Gaussian density around the percentile of interest. These densities are what we call the percentile weights. Let's assume that the corresponding income of the percentile of interest p_i is m_i (e.g., the first quartile). If we denote the attributed weight to observation j (with income m_j) by ω_j, we find that:

$$\omega_j(m_j|m_i) = \frac{\exp(-0.5 * v(m_j))}{Nh\sqrt{2\pi}} \text{ and } v(m_j) = \left(\frac{m_j - m_i}{h}\right)^2 \quad (26)$$

  where *h* is the bandwidth that acts as a "*smoothing parameter*".

- Running a weighed OLS regression with the generated percentile weights. Formally, in a matrix form, the weighted OLS estimate can be written as:

[10] In recent years, another closely related method to the stochastic regression imputation has become popular and widely adopted: Predictive Mean Matching (PMM) (see Rubin (1986) and Litlle (1988)). Without going into details, its breakthrough point resides in assigning one random value among the nearest non-missing values (among 3 to 10 of the closet values) to the predicted value of a given missing (including the variability of the residual) to that missing. Note that the imputation with this method can be performed with Stata. For more information, see the Stata help for the syntax : mi impute pmm depvar indepvars, replace knn(number of closest non-missing values).

$$\beta = (X'\Omega X)^{-1}X'\Omega Y \quad (27)$$

where Ω is a diagonal matrix, containning in our case the percentile weights. The advantage of the quantile (PWR) model compared to the OLS is that in the latter the coefficients are estimated to average the returns of the individual of reference. The quantile model will give more precise coefficients for a given percentile-group. This could also help in reducing the imputation of large residuals to observations with small expenditures (poorer group). As discussed in subsection 2.3.4, this problem can arise when there is heteroskedasticity.

2.3.6 Imputation with quantile regression model and random imputation of residuals

As we discuss below, the Stata **MCEMA** module enables the selection of one among the six different methods, which were presented in subsections 2.3.1 to 2.3.6.

2.4 Changes in welfare

In this section we discuss the estimation of the impact of price changes on well-being of entrants or new consumers of a given good or service. Notice that the ***MCEMA*** Stata module allows the user to estimate the change in the proportion of consumers after a change in price or after an equivalent change in income, as well as the change in the well-being of current users and entrants.

2.4.1 Changes in welfare among current users using Taylor approximation

Following a first order Taylor approximation, the impact of price change on well-being can be approximated by:[11,12]

$$dw = -e_i * dp_i, \quad (28)$$

where dw denotes the change in well-being, e_i represents the expenditures on the good i, and dp_i the observed proportional change in price. For instance, if the price increases by 1 percent, the money-metric utility decreases by $0.01 * e_i$.

[11] Note that we have two hot-deck approaches. The first is the distance function approach, and the second the pattern matching approach. The first imputes the missing value on the basis of the smallest squared distance statistic to the case with the missing value. The matching pattern method that we use assumes that the the sample can be stratified in separate -homogenous- groups and imputation is made randomly within each group (Fox-Wasylyshyn & El-Masri, 2005).

[12] See among others: Ahmad and Stern (1984), Ahmad and Stern (1991), as well as the works of Newbery (1995), Araar (1998), Creedy (1998), Yitzhaki and Lewis (1996) and Araar and Verme (2016).

Box 1. A simple implementation of the Taylor approximation method

Notice that information on the substitution between goods after the price change is not needed.[13] To clarify the first order Taylor approximation approach, assume that the consumer has a disposable income of $200, and has two goods (such as apples (good *i)* and oranges (good *j*)). Prices are initially normalized to one. Also, assume that the optimal consumption bundle is 100 units of good *i* and 100 units of good *j*.

Now, let the price of good *i* increase by 1%. With the old $100 devoted to good *i*, the consumer can buy approximately 99 units. Thus, the consumer loses one money-metric unit of utility. The total utility becomes 199 instead of 200. The loss is also approximately equal to $-\exp_i * dp_i = -100 * 0.01 = -1$, this being the case without any substitution or changes in consumption patterns.

Assume that the consumer decides to substitute two additional units of good i by approximately two units of good j, therefore loosing 3 units of good *i*. In sum, the money-metric utility becomes approximately equal to:

$$w' = 97*1.0 + (200-97*1.01)*1.0 = 199.03$$

The red colors represent the quantity and the blue values the initial utility drawn from the consumption of the last unit of good j. However, it can be seen that the substitution effect will not have a large impact on improving the well-being in this case (-0.97 instead of -1.00), which is explained by the fact that each of the last consumed units—of any good—will approximately generate the same utility (the blue values). This simple example tries to summarize the first order Taylor approximation approach, where the substitution effect (mainly captured by the second term of the Taylor approximation of the money-metric utility function) is neglected.

2.4.2 Change in welfare among new users

2.4.2.1 Taylor approximation and change in basket of goods

The Taylor approximation allows to assess the impact of price changes on well-being under the assumption of a predefined basket of goods. Assuming the continuity of the demand function in income and prices, theoretically, the consumer will not have zero expenditures on some goods. Empirically, we observe that consumers select a subset among the total possible set of goods, based on the availability of goods, on their income, the price vector, etc. To model the set of preferences with some potential nil expenditures using a Selective Demand Models (SDM). Jackson (1984) introduces the foundations of the hierarchic demand system where the change in a bundle of goods can be modelled. We introduce the use of the Taylor approximation with SDM in the following **Box 2**.

[13] As a demonstration, assume that all prices are normalized to one. Thus, the utility optimization condition requires that the marginal utility from the last consumed unit is equal for every good. It follows that the substitution between the last unit will not change the level of well-being significantly. This explains why we only focus on the good of interest to estimate the impact on well-being.

Box 2. New entrants and changes in welfare: an illustrative example

Assume that the consumer preferences are modelled with a discontinuous Cobb-Douglas function and that the selection of goods depends on real income ($\tilde{\mathrm{m}}$) or nominal income, and prices. The consumer purchases good i with an expenditure share α_i if real income falls under a specific range. The observed expenditure share is thus:

$$\alpha_i^* = \frac{\alpha_i \mathrm{I}[\tilde{\mathrm{m}} \in [\mathrm{l_i}, \mathrm{u_i}]]}{\sum_{j=1}^{J} \alpha_j \mathrm{I}[\tilde{\mathrm{m}} \in [\mathrm{l_j}, \mathrm{u_j}]]}$$

Where $I[condition] = 1$, if the income condition is met and $I[condition] = 0$, otherwise. J is the number of available goods in the market, and $\sum_{j=1}^{J} \alpha_j = \sum_{j=1}^{J} \alpha_j^* = 1$. In other words, the consumer starts to purchase good i if real income is higher than a given lower limit l_i (for instance, the consumer starts to use public transportation instead of a bicycle) and stops the consumption of that good if the consumer becomes rich and real income is higher than u_i (for instance, buying a car). Real income is denoted by $\tilde{\mathrm{m}} = \frac{m}{\Gamma(p)}$, where m denotes the nominal income and $\Gamma(p) = \prod_{i=1}^{K} p_i^{\alpha_i}$ is the price index. Assume that we have two goods in the market. Let $\alpha_1 = 0.2$ and $\alpha_2 = 0.8$. Also, let $(\mathrm{l}_1 = 100, \mathrm{u}_1 = \infty)$ and $(\mathrm{l}_2 = 0, \mathrm{u}_2 = 200)$. The utility function is:

$$U(x_1, x_2) = \begin{cases} {x_2}^{\theta \alpha_2^*} & if\ x_1 = 0; \\ {x_1}^{\alpha_1^*} {x_2}^{\alpha_2^*} & \text{otherwise} \end{cases}$$

The scalable parameter θ enables to smooth the marginal change in utility when discontinuity is almost met. In other words, the utility level must be the same when real income converges to 96 from both sides. In our example: $((96)^{\theta} = (19.2^{0.2} 76.8^{0.8}))$, and then $\theta = .89036717$. The other natural condition is that the marginal utility of income just after starting to consume the new good will be higher than that if we decide to continue to consume only one good. In our example, this condition is implicitly satisfied since $\theta < 1$.

The Marshallian demand functions are: $x_1 = \frac{\alpha_1^* m}{p_1}$, $x_2 = \frac{\alpha_2^* m}{p_2}$. Initially, prices are normalized to one thus, $p_1 = p_2 = 1$. With an income of \$96, $\alpha_1^* = 0$ and the consumed quantity of the first good is zero (x_1=0; $x_2 = 96$).[14]

Now, let's assume that the price of the first good decreases to 0.8. Thus, the consumer starts to consume the first good (x_1=24, $x_2 = 76.8$). The equivalent and compensated gains can be explained by the following:

$$EV = \left(\frac{1}{\prod_{i=1}^{K} p_i^{\alpha_i^*}} - 1 \right) * m \text{ and } CV = \left(1 - \prod_{i=1}^{K} p_i^{\alpha_i^*} \right) * m$$

$$EV = \left(\frac{1}{0.8^{0.2} 1^{0.8}} - 1 \right) * 96 = \mathbf{4.38} \text{ and } CV = (1 - 0.8^{0.2} 1^{0.8}) * 96 = \mathbf{4.19}$$

[14] The suggested function in which the consumer is not able to consume a small quantity of a given good can be explained by different things. For instance, one cannot buy a small quantity, or a fraction of a given good (half of a mobile phone for instance). Furthermore, this can also be explained by the arrival of a new product (before the arrival, the consumer is constrained to not consume that good). In other cases, the consumer can leave totally the consumption of a given good and replace it by another. This explains the use of upper bound for inferior goods.

The Taylor approximation of the impact on well-being is then equal to 0.2*24 = **4.80**, which is –moderately close to the EV and CV measurements. Note that the previous example is simply used to illustrate the Selective Demand Models, but it is used explicitly in the ***MCEMA*** module.

2.4.3 Total changes in wellbeing (current consumers and new users)

In addition to measuring the impact of price changes on the welfare of current consumers, the change in well-being of entrants also needs to be considered to measure the welfare impacts adequately. At group level, the average expected expenditures of entrants (e.g., decile 2) on good i is denoted by $e_{g,i}^{entrants}$ and those of current consumers of the same group by $e_{g,i}^{old}$. Note that $e_{g,i}^{old}$ is the average expenditures within group g at the initial period. The component $e_{g,i}^{entrants}$ precisely denotes the change in average expenditures within group g with price changes. The component $e_{g,i}^{non-users}$ denotes the average expected expenditures within group g when the expenditures of current users is set to zero. If the estimated proportion of entrants within group g is denoted by $\pi_{g,i}$ (estimated with the probabilistic model), then, we have that $e_{g,i}^{entrants} = \pi_{g,i} e_{g,i}^{non-users}$. In case of providing the exogenous number total entrants by region, we have that: $e_{g,i}^{entrants} = \frac{\sum_{h=1}^{N_g} \kappa_{g,h}\, e_{g,h,i}}{\sum_{h=1}^{N_g} sw_{g,h} hs_{g,h}}$, and where $sw_{g,h}$ and $hs_{g,h}$ denote the sampling weight and the household size of the household h in region g, respectively (remember that $\kappa_{g,h} = 0$ for current users, and it is equal to the normalized probability of use times the exogenous proportion of entrant for the non-users).

In sum, the initial average expenditures within group g is $e_{g,i}^{old}$ and this average after prices changes and entrants is $e_{g,i}^{old} + e_{g,i}^{entrants}$. The total change in well-being of group g is then:

$$dw_g = -\left(e_{g,i}^{old} + e_{g,i}^{entrants}\right) dp_i \qquad (29)$$

2.5 Prices information availability and estimated proportion of entrants

Usually, when working with data from household income and expenditure surveys, we typically do not observe the vector of prices. Thus, the probabilistic model described in the previous section can only be estimated with information on household income (welfare). The aim of this section is to find a simple way to estimate the proportion of entrants in the market given a price change, linking the probability of use and income.

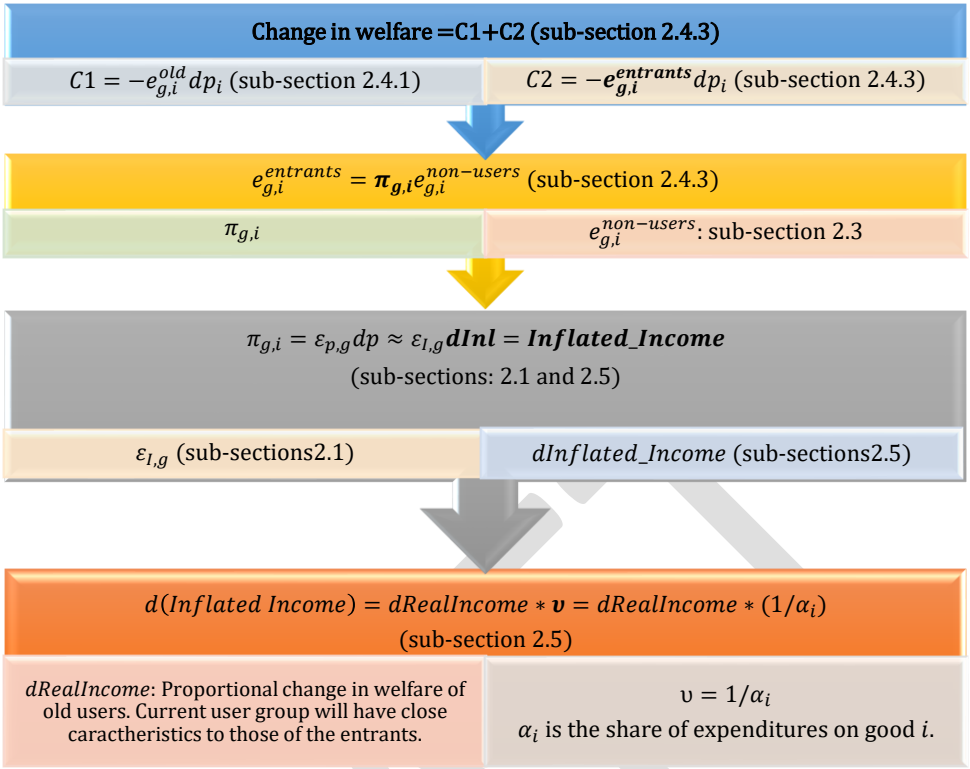
To do this, we need to assume that we can estimate the proportional change in real income (welfare) implied by a given price change. The basic condition that must be satisfied is that the change in quantity implied by the price change must be somehow equivalent to a change in income. However, a 1% increase in price will not have the same impact on quantity than a 1% increase in income. Therefore, we need to adjust the change in income by a scalar, that we denote by υ. Assuming that the utility function is monotonic, and that: $\varepsilon_i = -1$, and $\eta_i = \alpha_i$, while a decrease in price by $x\%$ will increase the quantity by $x\%$, an increase in income by $x\%$ will subsequently increase quantity by $\alpha_i x\%$. Thus, the suggested parameter of adjustment is defined as: $\upsilon = 1/\alpha_i$.

Box 3.

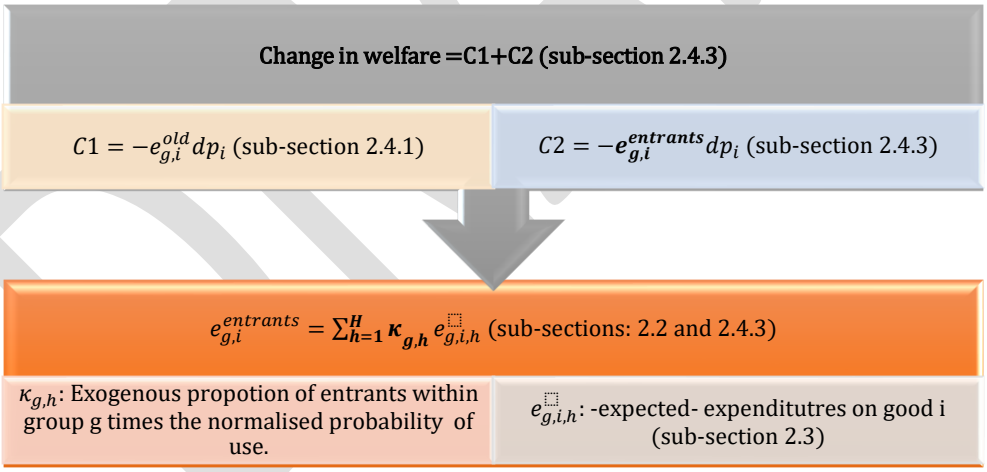
For example, let $U(x_1, x_2) = x_1^{0.1} x_2^{0.9}$ and $p_1 = p_2 = 1$. $m = 100$. Initially, $x_1^0 = \frac{\alpha_1 m}{p_1} = 10$. After the change in price of $\frac{dp_1}{p_1} = -0.01$, $x_1^1 = \frac{0.1*100}{0.99} \approx \mathbf{10.1}$. With a moderate change in price the money metric welfare change results in: $EV \approx CV \approx CS = -expenditures * dp_1 = -10 * -0.01 = 0.1$ Thus, the adjusted income change is (0.1/0.1)% = 1%. We can check that $x_1^1 = \frac{0.1*101}{1} = \mathbf{10.1}$.

2.6 Change on welfare and computation steps

Next, we present a scheme with the steps for computing the welfare change with entrants and for the case of using the probabilistic model to estimate the proportion of entrants, as well as the case where the prices' microdata are not available.



Now, we present a scheme when using the exogenous information on the number of entrants.



2.7 Value added and areas of improvement

As with any microsimulation tool, **MCEMA** has limitations which must be considered by the users. First, the tool focuses on current consumer expenditure of the item of interest to estimate spending patterns of entrants. Thus, it does not consider personal preference of consumers who would decide to not consume despite a decrease in prices. Second, sub-group expenditure imputations can be based on small number of

observations, yielding biased expenditure estimates of entrants. Thus the users must consider with caution which expenditure estimation model to use, considering any potential data limitations the underlying data could have. Third, predictions of new entrants are sensitive to consumption model, potentially resulting in noisy predictions if consumption patterns are not properly explained. Thus, there is the potentially of over allocating new entrants to a market that normally would not consume the good or services being analyzed due to individual and household characteristics.

3 The **MCEMA** Stata module

The Stata **MCEMA** module of the package WELCOM is designed to estimate a wide array of statistics of interest in the case of price changes implied by market competition, resulting in new entrants in a market. As it is explained in the theoretical section, the main statistics of interest are, the proportion of entrants, their expected expenditures and the change in well-being of moving (entrant) and staying (current) consumers. It follows that the main steps of the ***MCEMA*** WELCOM module are:

a. Estimate a *probit* model to measure the likelihood of consuming the good or service of interest. The user can select the significance level for the stepwise selection process of explanatory variables;
b. Predict the probability of consuming the good of interest with initial prices/incomes (if prices are available, the models will include the vector of prices by default);
c. Computing the proportion of entrants based on the expected variation in predicted probabilities;
d. Predict expenditures of entrants using one of six models presented;
e. Estimating the impacts on well-being of current and new consumers. The estimation of the impact can be done in a given population group partition.
f. Estimate the aggregate impact of current and new consumers on poverty headcount, poverty gap, and Gini index.
g. If indicated, export the results into an excel file.

In the rest of this document, **MCEMA** is introduced and applied to the case of Mexico's cell phone market using data from the *Encuesta Nacional de Ingresos y Gasto de Hogares* (ENIGH) for 2014. The premise of this example is that the removal of the market power will reduce prices, increase real incomes, and increase the use cell-phones (paying monthly plans or buying phone cards).[15] The first step is to use WELCOM's ***mcwel***

[15] The proportion of cell-phone users was about 63.7 percent in 2014.

module to estimate the change in price—in average—and the change in well-being of moving towards a more competitive market (see Rodriguez Castelan et., al 2019 for details).[16]

3.1. Installation

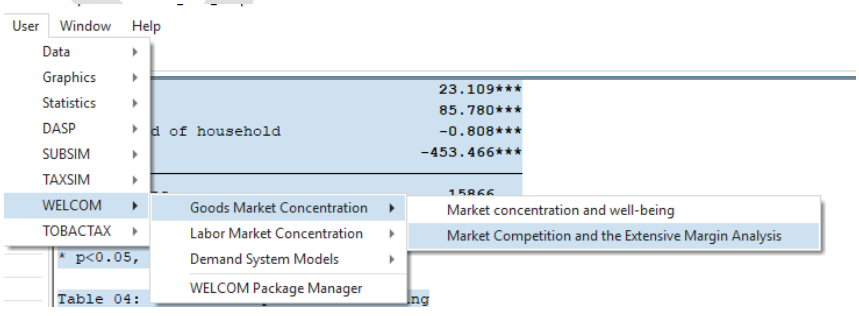
To install the **MCEMA** module for the WELCOM tool, execute the following commands in the Stata command line. Note that it is possible to either copy and paste these lines directly in the command window or in the dofile editor preferred by the User:

Commands 01

```
clear all
set more off
net from http://dasp.ecn.ulaval.ca/welcom/Installer34
net install welcom_p1, force
net install welcom_p2, force
net install welcom_p3, force
net install welcom_p4, force
cap additemenu profile.do _welcom_menu
_welcom_menu
```

After executing these steps, the User should close all Stata sessions and restart the program. After opening a new window, the User can go to the menu bar in Stata, click on the User option, choose the WELCOM package, select the "Goods Market concentration" option and launch the WELCOM tool by clicking on "Market concentration and Extensive Margin Analysis" (**MCEMA**).

*Figure 03: The **MCEMA** –tab Main - dialog box*



[16] For this example, we assume that the telecommunication market in Mexico behaves as an oligopoly composed of four main firms and has a price elasticity of demand of -0.32. The following examples are for illustrative purposes only and should be interpreted with caution. Note that the mcwel module can generate automatically a variable of the impact on real income.

3.2. Implementation

3.2.1. Step 1: estimating the change in well-being with the WELCOM mcwel module

Commands 02

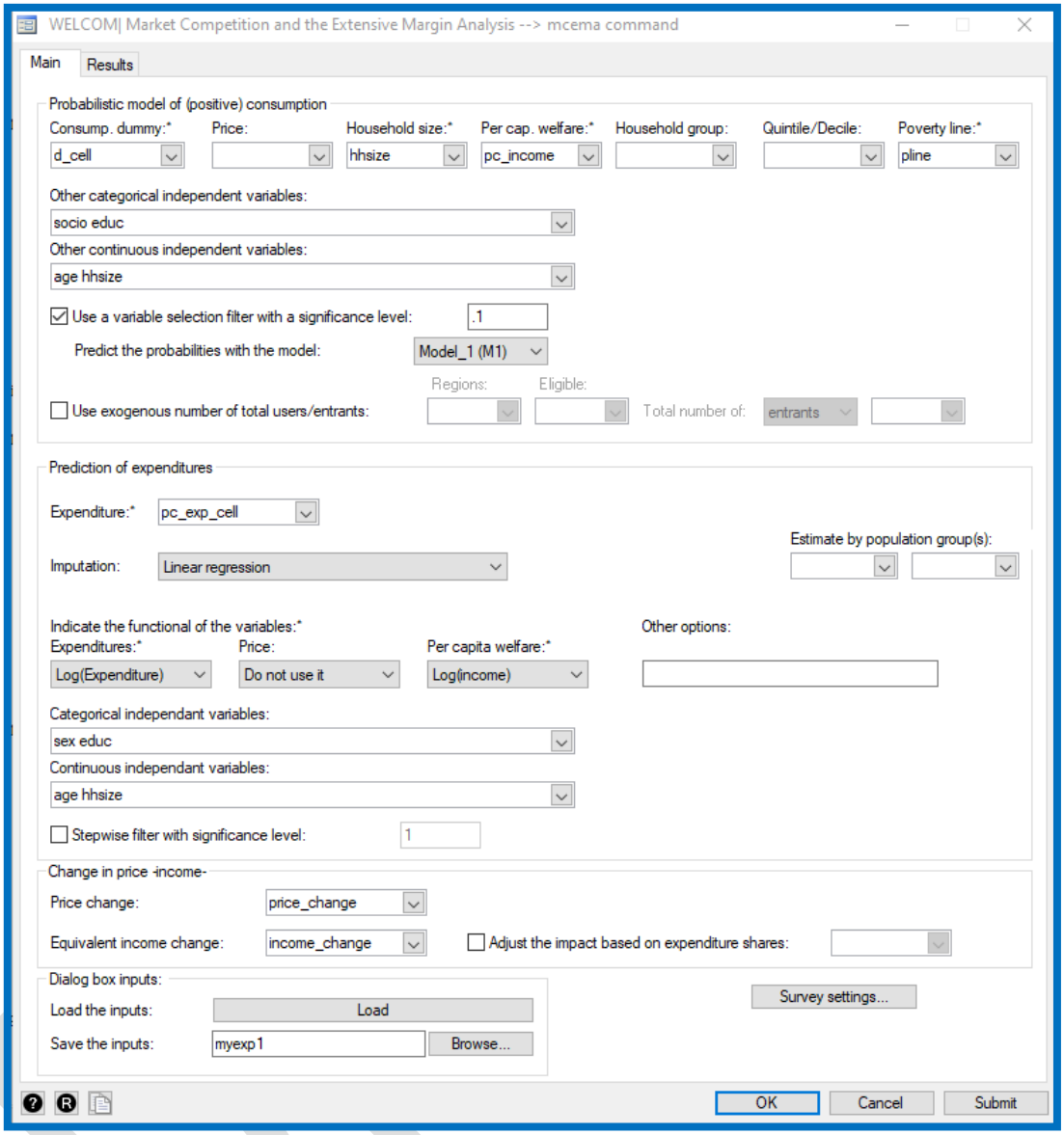
```
clear
use http://dasp.ecn.ulaval.ca/welcom/examples/mc/Mexico_2014_Cellphones.dta , replace
mcwel pc_income, hsize(hhsize) pline(pline) gvimp(1)  nitems(1) mpart(0) gscen(0)        ///
   it1( sn(Cell_phone)  vn(pc_exp_cell) el(-0.32) st(2) nf(4) ) move(-1)          ///
   epsilon(.5) tjobs(11) gjobs(off) gvpc(1)

*Rename price and income change variables
rename __impwell_pc_exp_cell_step_1 income_change
ren __pricech_pc_exp_cell_step_1 price_change

*Convert price change to negative as we simulate a price reduction
replace price_change = price_change*-1
```

3.2.2. Step 2: Using the WELCOM **MCEMA** module

The main dialogue box of the **MCEMA** module is composed of different panels:



The minimum required variables are:

- *The dummy variable of household consumption;
- *The variable of household/individual expenditures on the item of interest;
- *The household' disposable income (or total expenditures);
- *The household size;
- *The national poverty line defined by the survey;
- The household's price or unit value.

The other optional variables for the probabilistic model are:

- The welfare aggregate partition variable (as, the decile, the quintile, etc.);
- The group variable (example: the geographical area);
- The other categorical explanatory variables (example: sex, living_area, etc.);

- The other continues explanatory variables (example: age, etc.);

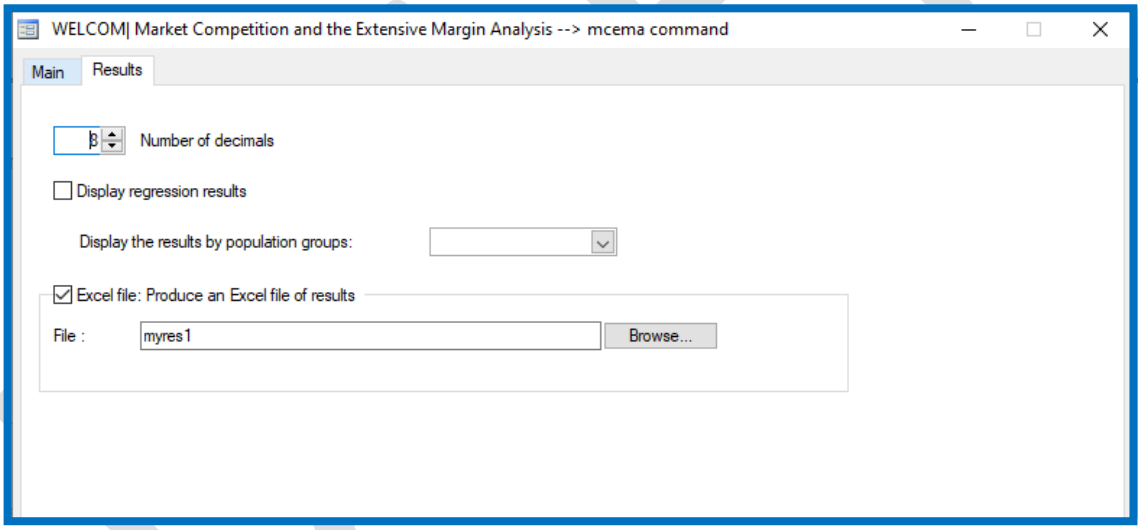
Optional variables for the linear expenditure models are:

- Additional categorical explanatory variables (example: sex, living_area, etc.);
- Other continuous explanatory variables (example: age, etc.);

In the Results TAB, the user can specify

- Number of decimals of the numerical results;
- Display the full regression results;
- Display the results by a desired population groups (by default is by quintiles). Users are highly encouraged to include their desired population group, as they may be different from the default manner in which quintiles are being calculated or from the ones used when running MCWEL.
- Save the estimated results in an Excel file (*.xml format).

*Figure 04: The **MCEMA** –tab: Results - dialog box*



3.3. Applications

3.1.1. Example 1

```
#delimit ;
use http://dasp.ecn.ulaval.ca/welcom/examples/mc/Mexico_2014_Cellphones.dta, replace;
mcwel pc_income, hsize(hhsize) pline(pline) gvimp(1)  nitems(1) mpart(0) gscen(0)
   it1( sn(Cell_phone)  vn(pc_exp_cell) el(-0.32) st(2) nf(4) ) move(-1)
   epsilon(.5) tjobs(11) gjobs(off) gvpc(1) ;

/*Rename variables generated by MCWEL module */
cap drop income_change;
cap drop price_change;
rename __impwell_pc_exp_cell_step_1 income_change ;
ren __pricech_pc_exp_cell_step_1 price_change ;
replace price_change = price_change*-1 ;

/*Run MCEMA*/
mcema d_cell pc_exp_cell, expmod(3) welfare(pc_income) hsize(hhsize) indcat(socio educ)
indcon(age hhsize)
pswp(.1) pchange(price_change) ichange(income_change) ewgr(quintile) um(1) dec(4) fex(2)
fpr(3) fin(2) cindcat(sex educ)
cindcon(age hhsize) inisave(myexp1) xfil(myres1) pline(pline) ;
```

- **Part 1 of results**

Table 01: Estimates of probability of consumption mode

	Model 1
0/1 dummy: use of cellphone	
LnInc	**0.4159*** **
Socio_2	**0.1188*** **
Number of household members	**0.0931*** **
Educ_10	**0.1860** **
Age of head of household	**-0.0063*** **
Educ_9	**0.2402***
Educ_4	**0.0859**
Educ_5	**0.2100***
Educ_6	**0.1459*** **
Educ_8	**0.1272***
Constant	**-3.1491*** **
Observations	**15866**
Pseudo R-squared	**0.050**

*** p<0.05, ** p<0.01, *** p<0.001**

Table 02: Estimated impact on the proportions of consummers (Predictions with the model 1)

	[1] Proportion of change in welfare (adjusted by share) (in %)	[2] welfare Elasticity	[3] Observed proportion of consumers/users	[4] Predicted probability with initial welfare(s)	[5] Predicted probability with final welfare(s)	[6] Absolute change in probability ([5] - [4])	[7] Relative change in probability (in %) ([6]/[4])*100
Quintile 1	4.9439	0.2379	0.4785	0.5052	0.5112	0.0059	1.1749
Quintile 2	4.6270	0.2379	0.5868	0.5975	0.6041	0.0066	1.1038
Quintile 3	5.0600	0.2379	0.6599	0.6437	0.6506	0.0070	1.0839
Quintile 4	5.3035	0.2379	0.6978	0.6847	0.6917	0.0070	1.0243
Quintile 5	4.9658	0.2379	0.7638	0.7688	0.7750	0.0062	0.8104
Population	5.0127	0.2379	0.6373	0.6565	0.6631	0.0066	1.0060

Notes:
[1]: Estimated based on averages of the population/groups.
[2]: Estimated based on the reference individual (at means of X covariates).
[3]: Estimated based on population/groups.
[4]: Based on predicted individual -housecurrent- probabilities with initial welfares.
[5]: Based on predicted individual -housecurrent- probabilities with final welfares.
[6]: Based on [4] and [5] Statistics.
[7]: Based on [4] and [6] Statistics.

Note that the relative change in probability is estimated using the predicted probabilities, which is about 1.00 (column 7). This is equivalent to an absolute increase in users of 0.7 percentage points, from a baseline of 63.7 percent current users. If we use the reference individual, this can be approximated to the proportional change in income times the elasticity of income, which it is equal to .238. Note that, the table 02 shows that the welfare

elasticity is the same across the quintiles. This is because of we use model 1, and where the quintile dummies are not used.

- **Part 2 of results: Aggregate monetary impacts**

This part shows the estimated impacts on well-being, where on average welfare impact from current users (derived from the MCWEL module) is Mex$ 175.4 and from new users is Mex$ 1.27 (Column 7 and column 8). The total welfare impact from both current and new users would thus be Mex$ 176.67 (Columns 9).

Table 03: Estimated impact on well-being

Group	[1] Expenditures (current users)	[2] Predicted expenditures (new users)	[3] Proportion of current users in (%)	[4] Proportion of new users in (%)	[5] Impact on well-being (current users)	[6] Impact on well-being (new users)	[7] Impact on well-being with current users	[8] Impact on well-being with new users	[9] Impact on well-being with all users
Quintile_1	22.7125	13.8026	47.8488	0.5936	81.1160	49.2949	38.8130	0.2926	39.1056
Quintile_2	32.6740	25.0193	58.6844	0.6596	116.6929	89.3546	68.4805	0.5893	69.0699
Quintile_3	47.0292	35.8786	65.9914	0.6977	167.9615	128.1378	110.8402	0.8940	111.7342
Quintile_4	71.3539	53.6040	69.7770	0.7013	254.8353	191.4429	177.8164	1.3426	179.1590
Quintile_5	176.4317	146.2424	76.3753	0.6230	630.1132	522.2945	481.2510	3.2540	484.5051
Population	77.0594	50.5967	63.7316	0.6604	275.2121	180.7024	175.3972	1.2742	176.6714

Notes:
[1]: Estimated based on average expenditures of current consumers by population/groups.
[2]: Estimated based on average usage (expenditures>0) by population/groups.
[3]: Estimated based on average predicted expenditures of the new consumers by population/groups.
[4]: Estimated based on the predicted expected change in proportion of consumers by population/groups.
[5]: Estimated average impact on well-being of the current consumers by population/groups.
[6]: Estimated average impact on well-being of the new consumers by population/groups.
[7]: ([3]/100)*[5].
[8]: ([4]/100)*[6] // For the population, the statistic is based on the case where we assign the expected group's benefit to each entrant.
[9]: [7] + [8].

- **Part 3 of results: Aggregate impacts of current and new users on poverty and Gini**

Based on the estimates from part 2, MCEMA transforms these results into welfare indicators such as the poverty headcount ratio, poverty gap and Gini coefficient, to better gauge the impact of reforms. Table 04 reports the breakdown of the welfare effects of both competition and adopter on poverty, where poverty rates are expected to decline by 2.60 percentage points from a baseline of 53.35 percent, driven by a reduction in poverty among current users of 2.57 percentage points and a reduction in poverty equivalent to 0.03 percentage points among new users, albeit not significant for this specific scenario likely driven by the low uptake of new users. Table 05 reports similar results for the poverty gap, which is expected to decrease by 1.4 percentage points. Finally, Table 06 reports the estimated impact on the Gini coefficient, however, for this specific scenario they are not significant as it can be see from the p-values.

Table 04: Poverty headcount & market power

Step	Poverty level	The change in poverty	Standard error	P-Value
Initial/Concentrated Market	53.3519	.	.	.
With benefits of current users	50.7814	-2.5705	0.1535	0.0000
With benefits of new users	53.3154	-0.0365	0.0365	0.3172
Final	50.7448	-2.6071	0.1576	0.0000

Table 05: Poverty gap & market power

Step	Poverty level	The change in poverty	Standard error	P-Value
Initial/Concentrated Market	21.0865	.	.	.
With benefits of current users	19.6996	-1.3869	0.0314	0.0000
With benefits of new users	21.0740	-0.0125	0.0041	0.0024
Final	19.6871	-1.3994	0.0314	0.0000

Table 06: Gini index & market power

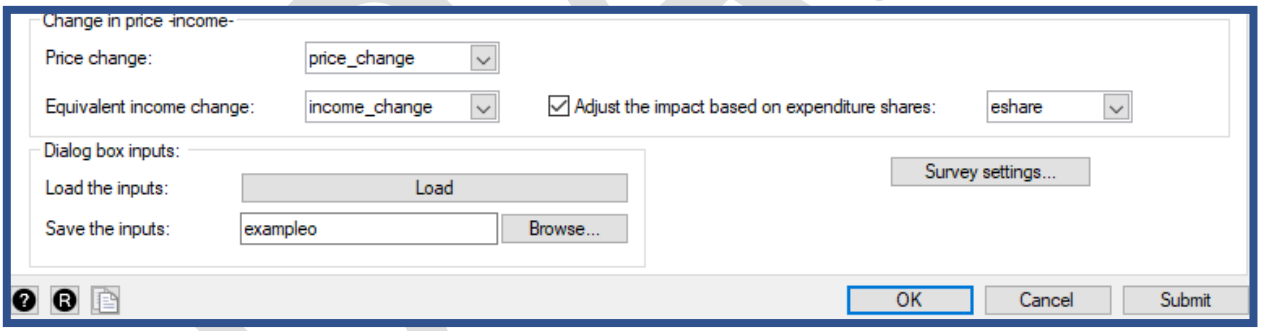
Step	The Gini index	Variation in the Gini index	Standard error	P_Value
Initial/Concentrated Market	49.6999	.	.	.
With benefits of current users	49.7626	0.0627	0.0413	0.1292
With benefits of new users	49.7020	0.0022	0.0042	0.6078
Final	49.7639	0.0641	0.0410	0.1186

3.3.2. Example 2

In this example, we consider the nature of change in the determinant of expenditures (price) and, thus, we adjust using expenditure shares, as discussed in sub-section 2.5. Furthermore, we analyze the estimated impact on the proportion of consumers using a household group variable (tam_loc: categories of the number of inhabitants). In **MCEMA**, to consider this adjustment, just indicate the expenditure share, shown in the figure below. Please refer to Annex II for the full list of commands to recreate this exercise.

```
#delimit ;

mcema d_cell pc_exp_cell, expmod(3) welfare(pc_income) hsize(hhsize) hgroup(tam_loc)
indcat(socio educ) indcon(hhsize) pswp(.05) pchange(price_change)
ichange(income_change) expshare(eshare) ewgr(quintile) um(3) dec(3) fex(2) fpr(3)
fin(2) cindcat(socio educ) cindcon(age) inisave(example2) xfil(myres2) pline(pline)
disgr(tam_loc) ;
```



- **Results**

We now show the results of the third model available where elasticities are estimated according to the concentration of population. We can observe that the income elasticity is lower in big cities, since most households already own and use the cell-phones. This is the opposite in less populated cities. It must be pointed out to users that it is not recommended to interact income with other covariates that are highly correlated to it, such as deciles. This will decrease the predictive power of income, and thus, underestimate its impact.

- **Part 1 of results**

	Model 1	Model 3
0/1 dummy: use of cellphone		
LnInc	0.349***	
Socio_2	0.130***	0.090**
Number of household members	0.086***	0.087***
Educ_10	0.466***	0.468***
Educ_11	0.414***	0.419***
Educ_3	0.116*	0.120*
Educ_4	0.253***	0.258***
Educ_5	0.435***	0.440***
Educ_6	0.383***	0.390***
Educ_7	0.389***	0.395***
Educ_8	0.385***	0.388***
Educ_9	0.534***	0.538***
LnInc_tam_loc_1		0.350***
LnInc_tam_loc_2		0.373***
LnInc_tam_loc_3		0.365***
LnInc_tam_loc_4		0.349***
Constant	-3.098***	-3.130***
Observations	19477	19477
Pseudo R-squared	0.056	0.058

* p<0.05, ** p<0.01, *** p<0.001

Table 02: Estimated impact on the proportions of consummers (Predictions with the model 3)

	[1] Proportion of change in welfare (adjusted by share) (in %)	[2] welfare Elasticity	[3] Observed proportion of consumers/users	[4] Predicted probability with initial welfare(s)	[5] Predicted probability with final welfare(s)	[6] Absolute change in probability ([5] - [4])	[7] Relative change in probability (in %) ([6]/[4])*100
100000 or more	170.890	0.200	0.679	0.677	0.766	0.089	13.216
15000 to 99999	193.404	0.209	0.691	0.689	0.786	0.097	14.099
2500 to 14999	173.973	0.220	0.643	0.642	0.738	0.095	14.857
2500 or less	158.363	0.264	0.513	0.518	0.602	0.084	16.251
Population	172.528	0.219	0.637	0.637	0.727	0.090	14.167

Notes:
[1]: Estimated based on averages of the population/groups.
[2]: Estimated based on the reference individual (at means of X covariates).
[3]: Estimated based on population/groups.
[4]: Based on predicted individual -household- probabilities with initial welfares.
[5]: Based on predicted individual -household- probabilities with final welfares.
[6]: Based on [4] and [5] Statistics.
[7]: Based on [4] and [6] Statistics.

- **Part 2 of results: Aggregate monetary impacts**

This part shows the estimated impacts on well-being. As we can observe from the results above, on average, the estimated increase in proportion of users is about 9.0 percentage points (instead of the 0.70 percentage points without correction). This also changes the estimated impact on well-being (15.99 instead of 1.27). From Table 03, it is estimated that the average welfare impact from current users (derived from the MCWEL module) is Mex$ 175.4 and from new users is Mex$ 15.99 (Column 7 and column 8). The total welfare impact from both current and new users would thus be Mex$ 191.38 (Columns 9).

Table 03: Estimated impact on well-being

Group	[1] Expenditures (current users)	[2] Predicted expenditures (new users)	[3] Proportion of current users in (%)	[4] Proportion of new users in (%)	[5] Impact on well-being (current users)	[6] Impact on well-being (new users)	[7] Impact on well-being with current users	[8] Impact on well-being with new users	[9] Impact on well-being with all users
100000_or_more	99.675	68.016	67.949	8.945	355.981	242.914	241.884	21.728	263.612
15000_to_99999	65.604	43.121	69.132	9.718	234.301	154.003	161.978	14.967	176.944
2500_to_14999	52.003	34.716	64.305	9.540	185.724	123.985	119.430	11.828	131.258
2500_or_less	44.746	24.630	51.279	8.420	159.806	87.964	81.947	7.406	89.353
Population	77.059	49.213	63.732	9.023	275.212	175.761	175.397	15.986	191.383

Notes:
[1]: Estimated based on average expenditures of current consumers by population/groups.
[2]: Estimated based on average usage (expenditures>0) by population/groups.
[3]: Estimated based on average predicted expenditures of the new consumers by population/groups.
[4]: Estimated based on the predicted expected change in proportion of consumers by population/groups.
[5]: Estimated average impact on well-being of the current consumers by population/groups.
[6]: Estimated average impact on well-being of the new consumers by population/groups.
[7]: ([3]/100)*[5].
[8]: ([4]/100)*[6] // For the population, the statistic is based on the case where we assign the expected group's benefit to each entrant.
[9]: [7] + [8].

- **Part 3 of results: Aggregate impacts of current and new users on poverty and Gini**

In terms of poverty, as it can be seen from Table 04, poverty headcount is expected to decline in 2.86 percentage points, driven by a reduction in poverty among current users of 2.57 percentage points and equivalent to 0.29 for new users. A similar decline is found for Mexico's poverty gap, which is expected to

decrease by 1.59 percentage points. Inequality is also expected to decrease in 0.048 Gini points, driven by new users. No significant effect was found for current users.

Table 04: Poverty headcount & market power

Step	Poverty level	The change in poverty	Standard error	P-Value
Initial/Concentrated Market	53.3519	.	.	.
With benefits of current users	50.7814	-2.5705	0.1535	0.0000
With benefits of new users	53.0607	-0.2912	0.0847	0.0006
Final	50.4901	-2.8618	0.1743	0.0000

Table 05: Poverty gap & market power

Step	Poverty level	The change in poverty	Standard error	P-Value
Initial/Concentrated Market	21.0865	.	.	.
With benefits of current users	19.6996	-1.3869	0.0314	0.0000
With benefits of new users	20.8825	-0.2040	0.0131	0.0000
Final	19.4956	-1.5909	0.0322	0.0000

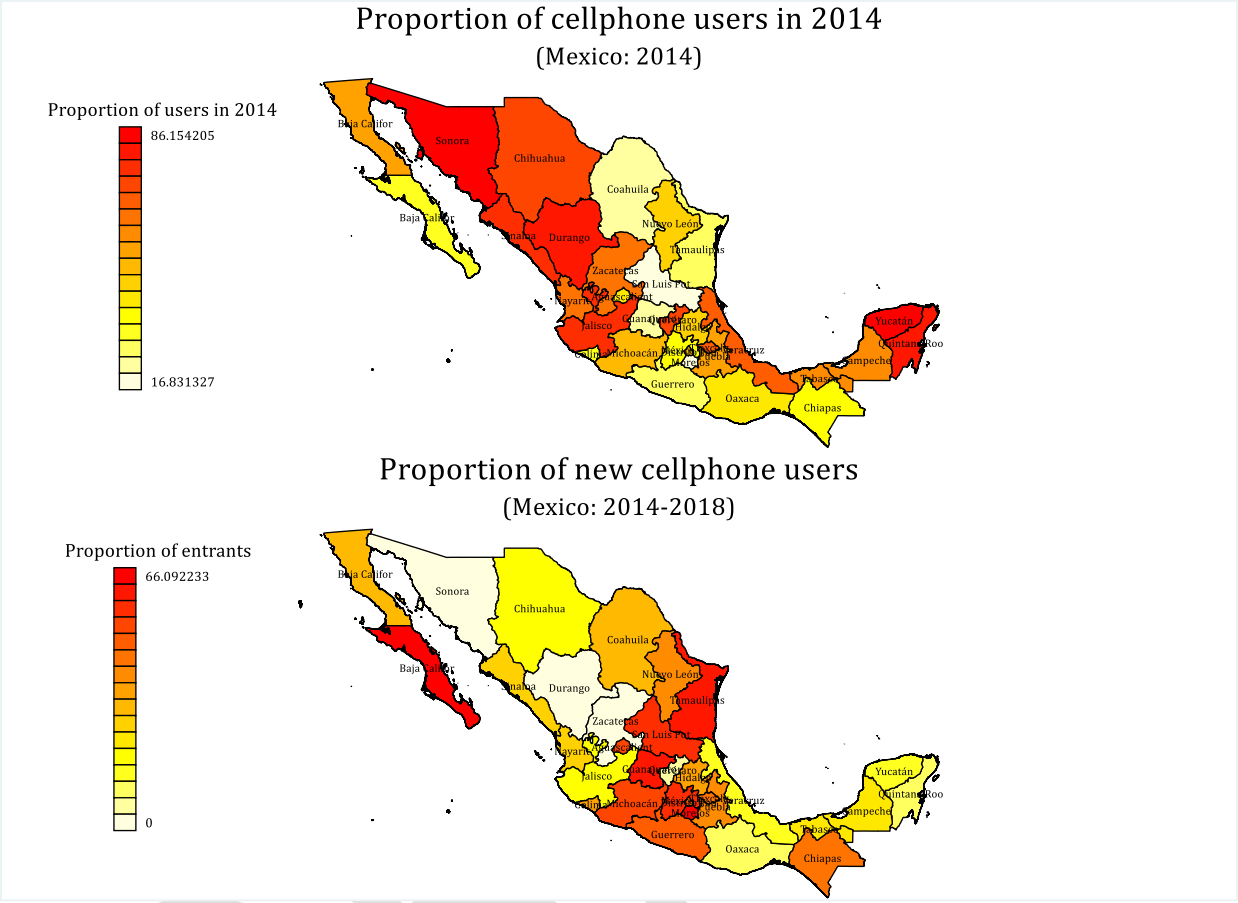
Table 06: Gini index & market power

Step	The Gini index	Variation in the Gini index	Standard error	P_Value
Initial/Concentrated Market	49.6999	.	.	.
With benefits of current users	49.7626	0.0627	0.0413	0.1292
With benefits of new users	49.6516	-0.0482	0.0170	0.0045
Final	49.7049	0.0051	0.0383	0.8947

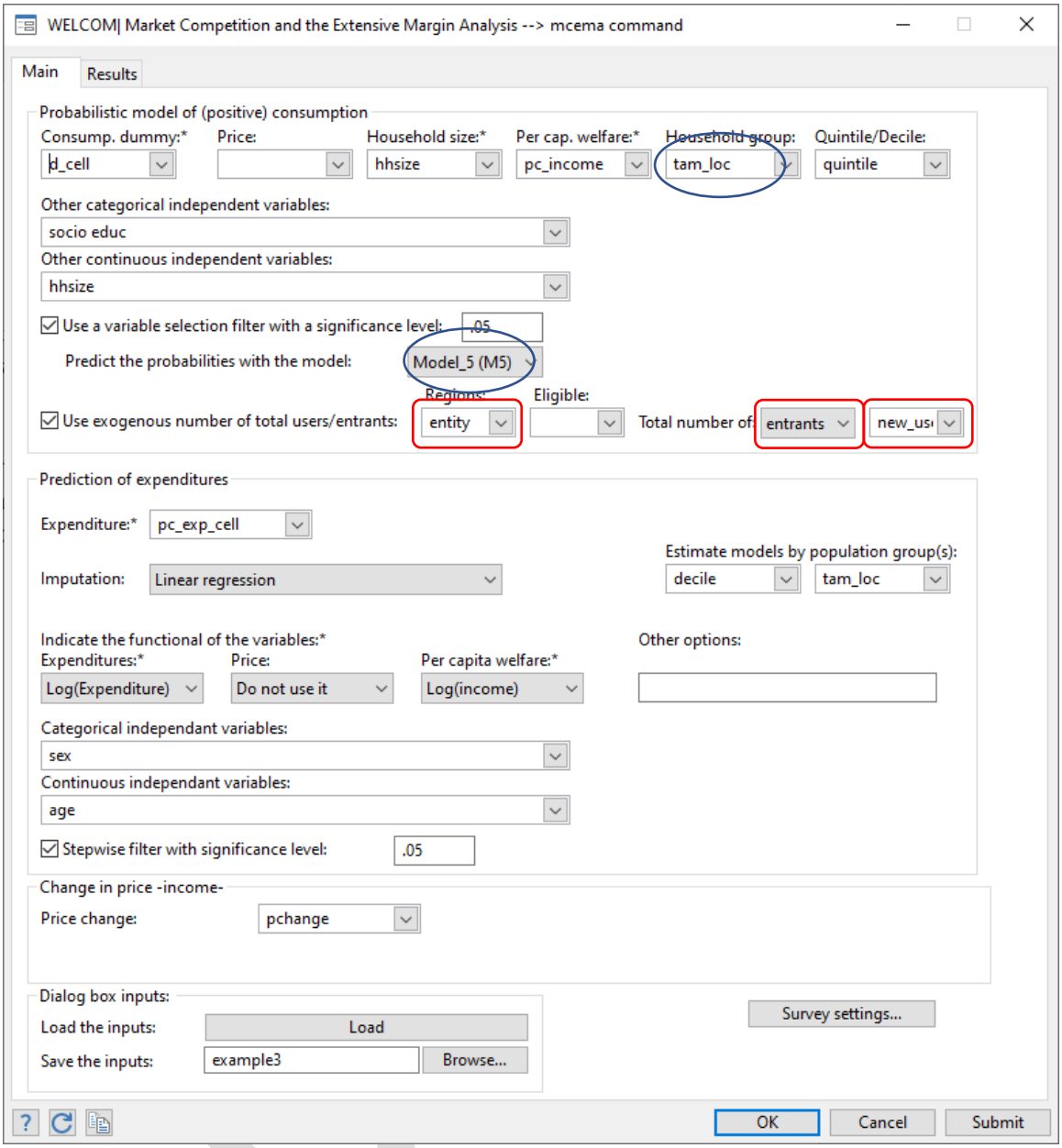
3.3.3. Example 3

In this third example, we assume that the user disposes exogenous information on the regional number of entrants. For example, if there is a targeted policy with the goal of increasing the number of cell phone users by administrative region by a specified amount. The data implemented for this scenario- the number of new cellphone users between 2014 and 2018- was estimated using household survey data.

Proportion of current and new cellphone users in Mexico by State 2014.



Source: ENIGH 2014
Notes: Top panel shows current users and lower panel shows entrants.



Based on the selected options in the dialog box, we note that:

1- We indicate that the exogenous information on the total number of entrants is available at the level of the entity group variable, which contains the code of 32 Mexican regions.
2- We indicate that we would like to predict the probabilities with the M5 model (we estimate the probability models by tam_loc group variable – according to the density of population-);
3- We indicate the variable that contains the total entrants by regions (new_users). Note that the user can indicate the number of total final users instead of entrants by region.
4- We do not indicate the variable eligible (by default all observations are eligible (potentially can use the service).

- **Probability results:**

```
. welcom_examples ex_mcema_db_03

. mcema d_cell pc_exp_cell, expmod(3) grmod1(decile) grmod2(tam_loc) welfare(pc_income) hsize(hhsize) hgroup(tam_loc) incpar(quintile) indcat(socio educ) ind
> con(hhsize) pswp(.05) pswe(.05) pchange(pchange) um(5) dec(3) fex(2) fpr(3) fin(2) exnum(1) grmac(entity) totentr(new_users) cindcat(sex) cindcon(age) inis
> ave(example3) xfil(myres3)
```

Table 01: Estimates of probability of consumption model(s)

	Model 1	Model 2	Model 3	Model 4
0/1 dummy: use of cellphone				
LnInc	0.351***			
Socio_2	0.128***	0.120***	0.088**	0.061*
Number of household members	0.086***	0.086***	0.087***	0.088***
Educ_10	0.465***	0.480***	0.466***	0.477***
Educ_11	0.413***	0.443***	0.416***	0.449***
Educ_3	0.120*	0.117*	0.124*	0.116*
Educ_4	0.255***	0.252***	0.259***	0.253***
Educ_5	0.435***	0.432***	0.439***	0.431***
Educ_6	0.385***	0.379***	0.390***	0.378***
Educ_7	0.391***	0.383***	0.397***	0.387***
Educ_8	0.388***	0.384***	0.390***	0.379***
Educ_9	0.532***	0.538***	0.536***	0.541***
LnInc_01		0.340***		
LnInc_02		0.341***		
LnInc_03		0.350***		
LnInc_04		0.348***		
LnInc_05		0.340***		
LnInc_tam_loc_1			0.351***	
LnInc_tam_loc_2			0.374***	
LnInc_tam_loc_3			0.365***	
LnInc_tam_loc_4			0.350***	
LnInc_Quintile_01_tam_loc_1				0.317***
LnInc_Quintile_02_tam_loc_1				0.302***
LnInc_Quintile_03_tam_loc_1				0.310***
LnInc_Quintile_04_tam_loc_1				0.314***
LnInc_Quintile_05_tam_loc_1				0.311***
LnInc_Quintile_01_tam_loc_2				0.346***
LnInc_Quintile_02_tam_loc_2				0.328***
LnInc_Quintile_03_tam_loc_2				0.334***
LnInc_Quintile_04_tam_loc_2				0.332***
LnInc_Quintile_05_tam_loc_2				0.328***
LnInc_Quintile_01_tam_loc_3				0.323***
LnInc_Quintile_02_tam_loc_3				0.319***
LnInc_Quintile_03_tam_loc_3				0.322***
LnInc_Quintile_04_tam_loc_3				0.329***
LnInc_Quintile_05_tam_loc_3				0.328***
LnInc_Quintile_01_tam_loc_4				0.282***
LnInc_Quintile_02_tam_loc_4				0.303***
LnInc_Quintile_03_tam_loc_4				0.337***
LnInc_Quintile_04_tam_loc_4				0.329***
LnInc_Quintile_05_tam_loc_4				0.303***
Constant	-3.111***	-3.050***	-3.137***	-2.796***
Observations	19477	19477	19477	19477
Pseudo R-squared	0.056	0.056	0.058	0.060

* p<0.05, ** p<0.01, *** p<0.001

The rest of_Table 01: Estimates of probability of consumption model 5

	100000_or_~e	15000_~99999	2500_t~14999	2500_or_less
0/1 dummy: use of cellphone				
LnInc_tam_loc_1	0.265***			
Socio_2	-0.464**			
Socio_3	-0.513***			
Socio_4	-0.607***			
Educ_2	1.945***			
Number of household members	0.071***	0.102***	0.093***	0.104***
Educ_11	0.383**	0.992***		
Educ_5	0.261*	0.605***	0.504*	
Educ_6	0.230***	0.555***	0.336***	0.220**
Educ_7	0.328**		0.403*	0.474*
Educ_8	0.275***	0.351*	0.559***	
Educ_9	0.422**	0.583*	0.856***	
Educ_10	0.367***	0.641***	0.541*	
LnInc_tam_loc_2		0.275***		
Educ_3		0.289*		
Educ_4		0.298*	0.331***	0.169**
LnInc_tam_loc_3			0.330***	
LnInc_tam_loc_4				0.598***
Constant	-1.694***	-2.456***	-2.825***	-4.849***
Observations	8289	2840	3097	5251
Pseudo R-squared	0.030	0.044	0.052	0.084

* p<0.05, ** p<0.01, *** p<0.001

Table 02: Estimated impact on the proportions of consummers

	[1] Observed proportion of consumers/users	[2] Predicted probability with initial welfare(s)
Quintile 1	0.478	0.478
Quintile 2	0.587	0.606
Quintile 3	0.661	0.654
Quintile 4	0.698	0.690
Quintile 5	0.764	0.760
Population	0.638	0.637

Notes:
[1]- Estimated based on population/groups.
[2]- Based on predicted individual -household- probabilities.

- **Results:**

Table 03: Number of current and new consumers

	[1] Population (Micro)	Current Users (Micro)	New Users (Macro/Micro)	New Users (Macro)	All Users (Macro/Micro)
1_Aguascalientes	1342587.04321	809146.94351	243019.04688	243019.04688	1052165.99039
2_Baja California	3648115.80396	2009107.18188	1205168.75000	1205168.75000	3214275.93188
3_Baja California Sur	839675.01950	564343.68411	95587.33594	95587.33594	659931.02005
4_Campeche	952279.02863	660596.55362	61980.49219	61980.49219	722577.04581
5_Coahuila	3073512.86072	1836191.64905	496526.15625	496526.15625	2332717.80530
6_Colima	763300.01712	573487.23672	46807.74609	46807.74609	620294.98282
7_Chiapas	5463341.92139	2781058.28955	666776.75000	666776.75000	3447835.03955
8_Chihuahua	3826986.84229	2139996.78088	493810.21875	493810.21875	2633806.99963
9_Distrito Federal	8781314.86761	5816391.12567	1537166.75000	1537166.75000	7353557.87567
10_Durango	1820740.01483	1490171.85754	0.00000	0.00000	1490171.85754
11_Guanajuato	5964968.30273	3059870.81628	1833289.12500	1833289.12500	4893159.94128
12_Guerrero	3630052.22180	1929717.89893	634457.31250	634457.31250	2564175.21143
13_Hidalgo	2990302.97449	1912653.01111	395952.06250	395952.06250	2308605.07361
14_Jalisco	8222690.85132	6575177.02417	507054.18750	507054.18750	7082231.21167
15_México	17674774.93774	10108913.69531	4039790.25000	4039790.25000	14148703.94531
16_Michoacán	4695278.19226	3077116.82019	921789.25000	921789.25000	3998906.07019
17_Morelos	1994132.96429	335639.03931	1317967.00000	1317967.00000	1653606.03931
18_Nayarit	1296946.95537	922231.03745	135594.48689	136153.95313	1057825.52433
19_Nuevo León	5321301.11951	3432064.55042	798881.37500	798881.37500	4230945.92542
20_Oaxaca	3602455.02332	2199868.69727	78453.14844	78453.14844	2278321.84570
21_Puebla	6877010.31067	4568215.86816	1025304.43750	1025304.43750	5593520.30566
22_Querétaro	2100098.05505	1558688.96930	1766.10242	1766.10242	1560455.07172
23_Quintana Roo	1722567.02014	1407358.63214	9278.35352	9278.35352	1416636.98566
24_San Luis Potosí	2831855.13245	1343210.34778	800288.68750	800288.68750	2143499.03528
25_Sinaloa	3066439.11371	2340002.13202	287693.96875	287693.96875	2627696.10077
26_Sonora	3061962.89648	2638009.80469	0.00000	0.00000	2638009.80469
27_Tabasco	2460881.99286	1717055.77362	158298.23438	158298.23438	1875354.00800
28_Tamaulipas	3672328.04114	1901516.66406	1072605.37500	1072605.37500	2974122.03906
29_Tlaxcala	1334965.02087	973683.73062	219991.23438	219991.23438	1193674.96500
30_Veracruz	8236730.34741	6040894.18481	272285.31250	272285.31250	6313179.49731
31_Yucatán	2207155.92505	1858011.00629	49512.96094	49512.96094	1907523.96722
32_Zacatecas	1615039.91797	1176092.77338	0.00000	0.00000	1176092.77338
Population	125091790.73589	79756483.77985	19407096.11079	19407655.57703	99163579.89063

Notes:

[1]: Estimation based on micro data (ex. sampling weight and househcurrent size variables).

[2]: Estimation based on the micro data (ex. observed users in the micro data).

[3]: Estimation based on micro/macro data (new users : based on total users (Macro/Micro) and current users (Micro).)

[4]: Estimation based on macro data (new users: based on the new users (Macro)).)

[5]: Estimation based on micro/macro data (total users: based on new users (Macro) and current users (Micro)).)

Table 04: Percentages of current and new consumers

	Current Users	New Users	All users
1_Aguascalientes	0.647	0.194	0.841
2_Baja California	1.606	0.963	2.570
3_Baja California Sur	0.451	0.076	0.528
4_Campeche	0.528	0.050	0.578
5_Coahuila	1.468	0.397	1.865
6_Colima	0.458	0.037	0.496
7_Chiapas	2.223	0.533	2.756
8_Chihuahua	1.711	0.395	2.105
9_Distrito Federal	4.650	1.229	5.879
10_Durango	1.191	0.000	1.191
11_Guanajuato	2.446	1.466	3.912
12_Guerrero	1.543	0.507	2.050
13_Hidalgo	1.529	0.317	1.846
14_Jalisco	5.256	0.405	5.662
15_México	8.081	3.229	11.311
16_Michoacán	2.460	0.737	3.197
17_Morelos	0.268	1.054	1.322
18_Nayarit	0.737	0.108	0.846
19_Nuevo León	2.744	0.639	3.382
20_Oaxaca	1.759	0.063	1.821
21_Puebla	3.652	0.820	4.472
22_Querétaro	1.246	0.001	1.247
23_Quintana Roo	1.125	0.007	1.132
24_San Luis Potosí	1.074	0.640	1.714
25_Sinaloa	1.871	0.230	2.101
26_Sonora	2.109	0.000	2.109
27_Tabasco	1.373	0.127	1.499
28_Tamaulipas	1.520	0.857	2.378
29_Tlaxcala	0.778	0.176	0.954
30_Veracruz	4.829	0.218	5.047
31_Yucatán	1.485	0.040	1.525
32_Zacatecas	0.940	0.000	0.940
Population	63.758	15.514	79.273

Table 05: Estimated impact on well-being

Group	[1] Expenditures (current users)	[2] Predicted expenditures (new users)	[3] Impact on well-being current users	[4] Impact on well-being new users	[5] Impact on well-being all users
1_Quintile 1	22.75	14.73	2.24	0.80	3.04
2_Quintile 2	32.72	24.31	3.95	1.09	5.04
3_Quintile 3	47.08	34.74	6.41	1.29	7.70
4_Quintile 4	71.48	54.14	10.27	1.81	12.08
5_Quintile 5	176.17	141.46	27.69	3.81	31.49
Population	77.04	43.76	10.11	1.76	11.87

Notes:
[1]- Computed based on average expenditures of current consumers by population/groups.
[2]- Estimated based on predicted expenditures of entrants by population/groups.
[3]- Computed based on average expenditures of current consumers and price change.
[4]- Estimated based on estimated expenditures of entrants, price change and the probability of use.
[5]- [3]+[4].

Table 06: Poverty headcount & market power

Step	Poverty level	The change in poverty	Standard error	P-Value
Initial/Concentrated Market	53.3197	.	.	.
With benefits of current users	53.1209	-0.1988	0.0415	0.0000
With benefits of new users	53.3039	-0.0158	0.0097	0.1039
Final	53.1051	-0.2147	0.0426	0.0000

Table 07: Poverty gap & market power

Step	Poverty level	The change in poverty	Standard error	P-Value
Initial/Concentrated Market	21.0411	.	.	.
With benefits of current users	20.9521	-0.0890	0.0021	0.0000
With benefits of new users	21.0166	-0.0245	0.0008	0.0000
Final	20.9276	-0.1135	0.0021	0.0000

Table 08: Gini index & market power

Step	The Gini index	Variation in the Gini index	Standard error	P_Value
Initial/Concentrated Market	49.6124	.	.	.
With benefits of current users	49.6113	-0.0011	0.0025	0.6423
With benefits of new users	49.6048	-0.0076	0.0010	0.0000
Final	49.6036	-0.0088	0.0022	0.0001

References

[1] Ahmad, E. and N. Stern (1984): "The Theory of Reform and Indian Indirect Taxes," Journal of Public Economics, 25, 259–98.

[2] Ahmad, E. and N. Stern (1991): The Theory and Practice of Tax Reform in Developing Countries, Cambridge.

[3] Allison, P. (2015), Imputation by Predictive Mean Matching: Promise & Peril, Statistical Horizons: https://statisticalhorizons.com/predictive-mean-matching

[4] Araar Abdelkrim (2016), Percentile Weighed Regression , PEP-PMMA technical note series 2016-01.

[5] Araar, Abdelkrim; Verme, Paolo. 2016. Prices and welfare. Policy Research working paper; no. WPS 7566. Washington, D.C. : World Bank Group.

[6] Andridge RR, Little RJ. A Review of Hot Deck Imputation for Survey Non-response. Int Stat Rev. 2010;78(1):40–64

[7] Creedy, J. (1998): "Measuring the Welfare Effects of Price Changes: A Convenient Parametric Approach," Australian Economic Papers, 37, 137–51.

[8] Cragg JG. (1971), Some statistical models for limited dependent variables with applications to the demand for durable goods. Econometrica 1971;39:829–44.

[9] Deaton A., and Muellbauer J. 1999. 8conomics and Consumer Behavior. Cambridge University Press, Cambridge.

[10] Dow, W.H., Norton, E.C. Choosing Between and Interpreting the Heckit and Two-Part Models for Corner Solutions. *Health Services & Outcomes Research Methodology* 4, 5–18 (2003).

[11] Eekhout, I., H. C. W. d. Vet, et al. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. Journal of Clinical Epidemiology, 67(3), 335-42.

[12] Enders (2010). Applied Missing Data Analysis. The Guilford Press.

[13] Fox-Wasylyshyn, S. M. and M. M. El-Masri (2005). Handling missing data in self-report measures. Res.Nurs.Health. 28(6): 488-495.

[14] Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. Structural Equation Modeling: A Multidisciplinary Journal, 7, 319–355.

[15] Jackson, Laurence Fraser, 1984. "Hierarchic Demand and the Engel Curve for Variety," The Review of Economics and Statistics, MIT Press, vol. 66(1), pages 8-15, February.

[16] Jakobsen, J.C., Gluud, C., Wetterslev, J. et al. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Med Res Methodol 17, 162 (2017).

[17] Leser, C. 8., 1963. Forms of Engel Function. Econometrica, 31: 6945703. Pangaribowo, 8H, 2010, Food Demand Analysis of Indonesian Households: Do They 8at Better or Smoke When They Get Richer? International Conference on Eurasian Economies 2010.

[18] Little, Roderick J. A. (1988) "Missing-data adjustments in large surveys." Journal of Business & Economic Statistics 6: 287-296.

[19] Mander, Adrian and Clayton, David, (2000), Hotdeck imputation, Stata Technical Bulletin, 9, issue 51.

[20] Nargis, Nigar & Ruthbah, Ummul & Hussain, Akm Ghulam & Fong, Geoffrey & Huq, Iftekharul & Ashiquzzaman, S. (2013). The Price Sensitivity of Cigarette Consumption in Bangladesh: Evidence from the International Tobacco Control (ITC) Bangladesh Wave 1 (2009) and Wave 2 (2010) Surveys. Tobacco control. 23. 10.1136/tobaccocontrol-2012-050835.

[21] Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. Organizational Research Methods 6, 328–362.

[22] Rubin, D. B. 1976. 1987. *Multiple imputation for non-response in surveys*. New York: John Wiley & Sons.

[23] Rubin, D. B. (1986) "Statistical matching using file concatenation with adjusted weights and multiple imputations." Journal of Business & Economic Statistics 4: 87-94

[24] Saha, Atanu, Capps, Oral and Byrne, Patrick, (1997), Calculating marginal effects in models for zero expenditures in household budgets using a Heckman-type correction, *Applied Economics*, 29, issue 10, p. 1311-1316.

[25] Siddique J, Belin TR. Using an Approximate Bayesian Bootstrap to Multiply Impute Nonignorable Missing Data. Comput Stat Data Anal. 2008;53(2):405–415. doi:10.1016/j.csda.2008.07.042

[26] Yitzhaki, S. and J. Lewis (1996): "Guidelines on Searching for a Dalton-Improving Tax Reform: An Illustration with Data from Indonesia," The World Bank Economic Review, 10, 541–562.

[27] Young, T., Single equation demand estimation, European Review of Agricultural Economics, Volume 9, Issue 1, 1982, Pages 103–105.

[28] Working, H.1943. Statistical Laws of Family 87penditure. Journal of the American Statistical Association, 32: 43516.

Annex I

Reproducing main results from Example 1

```
/*
mcema d_cell pc_exp_cell, expmod(3) welfare(pc_income) hsize(hhsize) indcat(socio
educ) indcon(age hhsize)
pswp(.1) pchange(price_change) ichange(income_change) ewgr(quintile) um(1) dec(4)
fex(2) fpr(3) fin(2) cindcat(sex educ)
cindcon(age hhsize) inisave(myexp1) xfil(myres1) pline(pline) */


*Run MCWEL to get price and income change
clear all
*Import Data
use http://dasp.ecn.ulaval.ca/welcom/examples/mc/Mexico_2014_Cellphones.dta , replace
mcwel pc_income, hsize(hhsize) pline(pline) gvimp(1)  nitems(1) mpart(0) ///
gscen(0) it1( sn(Cell_phone)  vn(pc_exp_cell) el(-0.32) st(2) nf(4) ) move(-1)    ///
   epsilon(.5) tjobs(11) gjobs(off) gvpc(1)

/*Rename variables generated by MCWEL module */
cap drop income_change
cap drop price_change
rename __impwell_pc_exp_cell_step_1 income_change
ren __pricech_pc_exp_cell_step_1 price_change
replace price_change = price_change*-1



*****************************************************************
*Table 1: Estimates of the probability of consumption model(s).*
*****************************************************************

*******************
*Generate variables*
*******************

*Ln of welfare variable
gen ln_income=ln(pc_income)

*Categorical variable dummies. This is done differently in the ado. Same results hold
tab educ,gen(educ_)
tab socio,gen(socio_)
```

```
tab tam_loc,gen(tam_)
tab sex,gen(sex_)

*Interaction variables as stepwise wont let us do it automatically
foreach x in tam_1 tam_2 tam_3 tam_4 {
    cap drop ln_inc_`x'
        gen ln_inc_`x' = `x'*ln_income
}

*Income changes
gen vchange = income_change

*Run Probit, remember to make sure to multiply hhweight by hhsize. We need to make sure
this is in module and helpfile

*This gives us the result for  Model 01
stepwise, pr(0.1): probit   d_cell ln_income hhsize    age educ_2 educ_3 educ_4 educ_5
educ_6 educ_7 educ_8 educ_9 ///
educ_10 educ_11  socio_2 socio_3 socio_4  [pw=sweight*hhsize]

************************************************************
*Table 2:  Estimated impact on the proportions of consummers*
************************************************************

*Generate temporary variables

*The probability elasticity with respect to price or welfare
tempvar elap
qui gen `elap'  = 0

*Proportional change in price
tempvar prop_ch
qui gen `prop_ch' = 0
tempvar prop_change

*Proportion of change in welfare (in %)
qui gen `prop_change' = 0
tempvar dif prdif

*The estimated change in the probability of use -or consumption- of the good
qui g `dif' = 0

*The estimated change in the price of use -or consumption- of the good
```

```
qui g `prdif' = 0

*Probability from probits are stored
cap drop `pr1'
cap drop `pr2'
tempvar pr1 pr2

*Predict probabilitys and add change in income from model
qui predict `pr1'
qui replace ln_income=log(exp(ln_income)+vchange)
qui predict `pr2'
qui replace ln_income=log(exp(ln_income)-vchange)

*Find margins vis-a-vis covariates
qui margins [aw=sweight*hhsize] , dydx(ln_income) atmeans

*Store elasticity of income
local mar1 = el(r(table),1,1)
if "`mar_1'" == "" {
local  local mar_`1' = 0
}

*Calculate price elasticity
qui sum d_cell [aw=sweight*hhsize]
local elap1 = `mar1' / r(mean)
qui replace `elap' = `elap1'
qui drogen tot_elap = `elap'

*Store change in income from mcwel in mu1
qui sum vchange  [aw=sweight*hhsize] , meanonly
local mu1 = r(mean)

*Store average pc income in mu2
qui sum pc_income   [aw=sweight*hhsize] ,   meanonly
local mu2 = r(mean)

*Calculate proportional change in welfare (%) and store in macro
qui replace `prop_change' = `mu1'/`mu2'*100

*Calculate average predicted probability and store in mu1
qui sum `pr1'  [aw=sweight*hhsize]  , meanonly
local mu1 = r(mean)
```

```
drogen pr1 =  `pr1'

*Calculate average predicted probability and store in mu2
qui sum `pr2'  [aw=sweight*hhsize]  , meanonly
local mu2 = r(mean)
drogen pr2 =  `pr2'

*Calculate difference in probabilities
qui replace `prdif' = `pr2'-`pr1'
drogen difa = `prdif'
cap drop dif
qui gen dif = `prdif'
qui gen double difa = pr2-pr1

* Calculate the estimated change in the probability of use -or consumption- of the good
qui replace `dif' = `mu2'-`mu1'

*Calculate Proportional change in price
qui replace `prop_ch' = `dif'/`mu1'*100

*Col (1)
forvalues i=1/5 {
qui sum vchange [aw=sweight*hhsize] if (quintile==`i') , meanonly
local mu1 = r(mean)

*Calculate average predicted probability and store in mu2
qui sum pc_income  [aw=sweight*hhsize]  if (quintile==`i') , meanonly
local mu2 = r(mean)
local  nprop_`i' = `mu1' / `mu2'*100
dis  `i' " : " `nprop_`i'' " : "
}

dis `prop_change'

*Col (2)
table quintile[aw=sweight*hhsize],c(mean tot_elap) row

*Col (3)
table quintile[aw=sweight*hhsize],c(mean d_cell) row

*Col (4)
table quintile   [aw=sweight*hhsize] , c (mean pr1) row  format(%4.3f)
```

```
*Col (5)
table quintile   [aw=sweight*hhsize] , c (mean pr2) row format(%4.3f)

*Col (6)
table quintile   [aw=sweight*hhsize] , c (mean difa) row format(%14.6f)

dis `mu2'-`mu1'
local nprop_p = `mu2'-`mu1'

forvalues i=1/5 {
qui sum `pr1'  [aw=sweight*hhsize] if (quintile==`i') , meanonly
local mu1 = r(mean)

*Calculate average predicted probability and store in mu2
qui sum `pr2'  [aw=sweight*hhsize]  if (quintile==`i') , meanonly
local mu2 = r(mean)
local  nprop_`i' = `mu2' - `mu1'
dis  `i' " : " `nprop_`i'' " : "
}

qui sum `pr1'  [aw=sweight*hhsize], meanonly
local mu1 = r(mean)
qui sum `pr2'  [aw=sweight*hhsize], meanonly
local mu2 = r(mean)
dis  `mu2' - `mu1'


*Col (7)
cap drop temp
qui gen temp = 0
forvalues i=1/5 {
qui sum pr1  if quintile==`i' [aw=sweight*hhsize]
local m1 = r(mean)
qui sum pr2  if quintile==`i' [aw=sweight*hhsize]
local m2 = r(mean)
qui replace temp = (`m2'/`m1'-1)*100 if quintile==`i'
}

table quintile   [aw=sweight*hhsize] , c (mean temp)  format(%7.6f) row
```

```
*Population level
qui sum pr1  [aw=sweight*hhsize]
local m1 = r(mean)
qui sum pr2   [aw=sweight*hhsize]
local m2 = r(mean)
dis (`m2'/`m1'-1)*100

*************************************************************************
```

```
*************************************************************
*Table 3:  Estimated impact on Wellbeing*
*************************************************************

*Col (1)
table quintile [aw=hhsize*sweight] if d_cell==1 , c (mean pc_exp_cell) row

*Col (2)
cap drop ln_pc_income
cap drop ln_pc_exp_cell
gen ln_pc_exp_cell = ln(pc_exp_cell)
gen ln_pc_income   = ln(pc_income)
cap drop predicted_exp
gen predicted_exp = .
cap drop ln_pc_exp_cell
qui gen ln_pc_exp_cell = ln(pc_exp_cell)

*Run regression to estimate expenditures
xi:   regress ln_pc_exp_cell ln_pc_income age hhsize  sex_2   educ_2 educ_3 educ_4 educ_5
educ_6 educ_7 educ_8 educ_9 educ_10 educ_11 [pw=hhsize*sweight] if d_cell==1
cap drop preda
predict preda, xb
qui replace predicted_exp = exp(preda)

table quintile [aw=hhsize*sweight] if d_cell==0 , c (mean predicted_exp) row

*Col (3)
drogen d_cell100=d_cell*100
table quintile [aw=hhsize*sweight] , c (mean d_cell100) row format(%10.2f)
```

```
*Col (4)
drogen dif100 = dif*100
table quintile  [aw=hhsize*sweight] , c (mean dif100) row format(%10.2f)

*Col (5)
cap drop old_imp
qui gen  old_imp = -pc_exp_cell*price_change
table quintile [aw=hhsize*sweight] if d_cell==1 , c (mean old_imp ) row format(%10.2f)

*Col (6)
cap drop new_imp
qui gen  new_imp = -predicted_exp*price_change
table quintile [aw=hhsize*sweight] if d_cell==0 , c (mean new_imp ) row format(%10.2f)

*Col (7)
cap drop old_imp
qui gen  old_imp = -pc_exp_cell*price_change
table quintile [aw=hhsize*sweight]             , c (mean old_imp ) row format(%10.2f)

*Col (8)
cap drop tot_imp
qui drogen tot_imp = 0

forvalues i=1/5 {
qui sum dif      [aw=hhsize*sweight]  if (quintile==`i')
local prop = `r(mean)'
qui sum new_imp  [aw=hhsize*sweight]  if (quintile==`i') & d_cell==0
local imp_n = `r(mean)'
dis `i' " : " `prop' " : " `imp_n' " : " `prop'*`imp_n'
qui replace  tot_imp =  `prop'*`imp_n' if (quintile==`i')
}

table quintile  [aw=hhsize*sweight]             , c (mean tot_imp ) row format(%10.2f)

*Col (9)
cap drop tot_imp
qui drogen tot_imp = 0
forvalues i=1/5 {
qui sum dif      [aw=hhsize*sweight]  if (quintile==`i')
local prop = `r(mean)'

qui sum old_imp  [aw=hhsize*sweight]  if (quintile==`i')
```

```
local imp_o = `r(mean)'

qui sum new_imp  [aw=hhsize*sweight]  if (quintile==`i') & d_cell==0
local imp_n = `r(mean)'
qui replace  tot_imp = `imp_o' + `prop'*`imp_n' if (quintile==`i')
}

table quintile  [aw=hhsize*sweight]             , c (mean tot_imp ) row format(%10.2f)

***********
*Table 4-6:*
***********

*Create the 3 types of consumers (Never consume / Already consume / New consumer)
qui   qui drogen      _type_consumer  = 0
qui   replace         _type_consumer  = 1 if d_cell==1

cap lab drop type_consumer
lab define type_consumer 0 "No entrant" 1 "Old user" 2 "New user"
lab val _type_consumer  type_consumer
drogen sw = sweight*hhsize

* Total population
qui sum sw
local ss1 = r(sum)

* Total old users */
 qui sum sw if  _type_consumer ==1
 local ss2 = r(sum)

* proportion of new users at population level : estimated in table 2 */
  local newprop =  `nprop_p'

  /*
  We draw randomelyrandomly from the non-users the porpotionproportion of users,
  but the probability of being drawn depends on the probability
  of use, i.e. the higher is the probability of use, the higher is the
  probability of selection
  */

set seed 6543456
qui drogen double random = uniform()*pr1*sw
```

```
*We do not draw among old users
qui replace random = 0  if  _type_consumer == 1

/* We keep the original order of the data */
qui drogen tkey = _n

/* We order by the actual type_of consumnersconsumers, the groups,
    and then the weighetd random probability of use in decendingdescending orerorder
*/
 gsort _type_consumer  quintile  - random

/* Within each population group, we seek for those:
       1-  with highest probability of selectiionselection;
       2-  prealalbly were nnon users;
       3-   Their3-  their  proortionproportion  is  equal  to  the  new  proportion  of
entransentrants of the group.s
       */
forvalues g=1/5 {
qui sum sw if (quintile==`g')
local   ss1 = r(sum)
local prop = `nprop_`g''
qui  drogen     double     tpsum  =  sum(sw*(_type_consumer!=1)*(quintile==`g')  )   if
(quintile==`g')
   qui       replace   tpsum = `ss1'*1.0001 if  _type_consumer ==1  & (quintile==`g')
   qui           replace   _type_consumer  =   2  if  (tpsum  <=  `prop'*`ss1'*1.0001)  &
(quintile==`g')

}

/*  In  this  first  example,  the  estimated  proportion  of  new  entrants  was  very  low
(practically is equal to zero).
   For that reason, non newnon-new entrants are found */
tab _type_consumer

/* Incomet is the welfare at situation t
       t=0 : intialinitial situtationsituation;
       t=1 : with only benefits of old users;
       t=2 : with only benefits of new users;
       t=3 : final situation.

*/
```

```
drogen      _benefit = 0
replace  _benefit = -pc_exp_cell*price_change   if _type_consumer==1
replace  _benefit = -predicted_exp*price_change if _type_consumer==2
replace _benefit = 0 if _benefit==.
drogen income0 = pc_income
drogen income1 = pc_income +  (_benefit * (_type_consumer==1))
drogen income2 = pc_income +  (_benefit * (_type_consumer==2))
drogen income3 = pc_income +  (_benefit )

local lab0    IntialInitial situtationsituation
local lab1  With only benefits of old users
local lab2  With only benefits of new users
local lab3  Final situation



*****************************
*Table 4:  Impact on poverty*
*****************************
forvalues i=0/3 {
drogen poor`i' = (income`i'<pline)
qui sum poor`i' [aw=sw]
dis "`lab`i'' : " _col(36) %6.4f `r(mean)'*100
}

***********************
*Table 5:  Poverty Gap*
***********************

forvalues i=0/3 {
drogen poor`i' = max(0,(pline-income`i')/pline)
qui sum poor`i' [aw=sw]
dis "`lab`i'' : " _col(36) %6.4f `r(mean)'*100
}

************************
*Table 6:Impact on Gini*
************************

forvalues i=0/3 {
qui fastgini income`i' [pw=sw]
local gini_`i' = r(gini)
dis "`lab`i'' : " _col(36) %6.4f `r(gini)'*100
```

```
}
```

Annex II

Reproducing main results in Example 2

```
use http://dasp.ecn.ulaval.ca/welcom/examples/mc/Mexico_2014_Cellphones.dta , replace

mcwel  pc_income,  hsize(hhsize)  pline(pline)  gvimp(1)    nitems(1)  mpart(0)  gscen(0)
///
   it1( sn(Cell_phone)  vn(pc_exp_cell) el(-0.32) st(2) nf(4) ) move(-1)          ///
   epsilon(.5) tjobs(11) gjobs(off) gvpc(1)
cap drop income_change
rename __impwell_pc_exp_cell_step_1 income_change
cap drop price_change
gen price_change = -  __pricech_pc_exp_cell_step_1

/* Note that the change in welfare is evaluated with the normalised price reference */
mcema d_cell pc_exp_cell, expmod(3) welfare(pc_income) hsize(hhsize) hgroup(tam_loc) ///
indcat(socio educ) indcon(hhsize) pswp(.05) pchange(price_change)
ichange(income_change) expshare(eshare) ewgr(quintile) um(3) dec(3) fex(2) ///
fpr(3) fin(2) cindcat(socio educ) cindcon(age) inisave(example2)  ///
xfil(myres2) pline(pline) disgr(tam_loc)

*****************************************************************
*Reproducing the results with simple Stata commands.              *
*****************************************************************

clear all
use http://dasp.ecn.ulaval.ca/welcom/examples/mc/Mexico_2014_Cellphones.dta , replace
mcwel  pc_income,  hsize(hhsize)  pline(pline)  gvimp(1)    nitems(1)  mpart(0)  gscen(0)
///
   it1( sn(Cell_phone)  vn(pc_exp_cell) el(-0.32) st(2) nf(4) ) move(-1)          ///
   epsilon(.5) tjobs(11) gjobs(off) gvpc(1)
rename __impwell_pc_exp_cell_step_1 income_change
cap drop
gen price_change =-    __pricech_pc_exp_cell_step_1
gen sw = sweight*hhsize

*************************************************************
*Table 1: Estimates of the probability of consumption model(s).*
*************************************************************
```

```
********************
*Generate variables*
********************

*Ln of welfare variable
qui gen double ln_income=ln(pc_income)

*Categorical variable dummies. This is done differently in the ado. Same results hold.
qui {
tab educ,gen(educ_)
tab socio,gen(socio_)
tab tam_loc,gen(tam_)
tab sex,gen(sex_)
}

*Interaction variables as stepwise wont let us do it automatically
foreach x in tam_1 tam_2 tam_3 tam_4 {
    cap drop ln_inc_`x'
        qui gen  double ln_inc_`x' = `x'*ln_income
}

*Income changes
gen double vchange = income_change

*Interaction variables as stepwise wont let us do it automatically
foreach x in tam_1 tam_2 tam_3 tam_4 {
    cap drop ln_inc_`x'
        qui gen double ln_inc_`x' = `x'*ln_income
}



************************************************************
*Table 1: Estimates of the probability of consumption model(s).*
************************************************************

*Run Probit, remember to make sure to multiply hhweight by hhsize. We need to make sure
this is in the module and the help file
*This gives us the result for  Model 01
qui stepwise, pr(0.05): probit  d_cell ln_income  hhsize  educ_2 educ_3 educ_4 educ_5
educ_6 educ_7 educ_8 educ_9 ///
educ_10 educ_11  socio_2 socio_3 socio_4  [pw=sw]
eststo m1
```

```
*Run probit with interaction, this gives us model 03 adjusting for location size
qui stepwise, pr(0.05): probit  d_cell ln_inc_tam_1 ln_inc_tam_2 ln_inc_tam_3 ln_inc_tam_4
hhsize   educ_2 educ_3 educ_4 educ_5 educ_6 educ_7 educ_8 educ_9 ///
educ_10 educ_11  socio_2 socio_3 socio_4  [pw=sw]
eststo m2
esttab m1 m2, not pr2

************************************************************
*Table 2:  Estimated impact on the proportions of consummers*
************************************************************
/*
We predict with model 3 : um(3)
*/
qui sum d_cell [aw=sw]
local meanprob = r(mean)
qui stepwise, pr(0.05): probit  d_cell ln_inc_tam_1 ln_inc_tam_2 ln_inc_tam_3 ln_inc_tam_4
hhsize   educ_2 educ_3 educ_4 educ_5 educ_6 educ_7 educ_8 educ_9 ///
educ_10 educ_11  socio_2 socio_3 socio_4  [pw=sw]

/* Table 2 : Column 1 */
cap drop adjusted_income_change
qui gen  double adjusted_income_change =  income_change/(eshare)
cap drop tmp
qui gen double double tmp = 0
forvalues i=1/4 {
qui sum  adjusted_income_change [aw=sw]    if  tam_loc == `i'
local mu1 = r(mean)
qui sum  pc_income                 [aw=sw]    if   tam_loc == `i'
local mu2 = r(mean)
qui replace tmp  = (`mu1'/`mu2')*100                      if   tam_loc == `i'
qui replace tmp  = 0 if    `mu2' ==0
}

table tam_loc    [aw=sw] , c (mean tmp )   format(%8.3f)

/* At population level */
qui sum  adjusted_income_change [aw=sw]
local mu1 = r(mean)
qui sum  pc_income                 [aw=sw]
local mu2 = r(mean)
qui replace tmp  = (`mu1'/`mu2')*100
qui replace tmp  = 0 if    `mu2' ==0
```

```
sum tmp [aw=sw]


/* Table 2 : Column 2  : estimated with model 3 (option um(3))*/
foreach x in tam_1 tam_2 tam_3 tam_4 {
    cap drop ln_inc_`x'
        qui gen ln_inc_`x' = `x'*ln_income
}
qui    stepwise,  pr(0.05):  probit    d_cell  ln_inc_tam_1  ln_inc_tam_2  ln_inc_tam_3
ln_inc_tam_4 hhsize   educ_2 educ_3 educ_4 educ_5 educ_6 educ_7 educ_8 educ_9 ///
educ_10 educ_11  socio_2 socio_3 socio_4  [pw=sw]
*Predict the probabilities with and without change in incomes
cap ddrop pr1
qui predict pr1
forvalues i=1/4 {
qui replace ln_inc_tam_`i'=log(exp(ln_inc_tam_`i')+income_change/eshare)
qui replace ln_inc_tam_`i'=ln_inc_tam_`i'*tam_`i'
}
qui predict pr2
forvalues i=1/4 {
qui replace ln_inc_tam_`i'=log(exp(ln_inc_tam_`i')-income_change/eshare)
qui replace ln_inc_tam_`i'=ln_inc_tam_`i'*tam_`i'
}
qui gen double prdif = pr2 - pr1

/* The margin is estimated by tam_loc  */
/* and the dF(XB) by one unit is assumed to be the same */
forvalues i=1/4 {
qui margins      [aw=sw]   , dydx(ln_inc_tam_`i') atmeans
local   margin_`i'   =     el(r(b),1,1)   //   This   is   assumed   to   be   equal   to   :
Norm_Dens(X_bar*Beta)*_Beta_of_Ln_pc_income.
}

cap drop elas
qui  gen double elas=.
forvalues i=1/4 {
qui sum d_cell   [aw=sw]       if  tam_loc == `i'
qui replace elas =  `margin_`i''/r(mean)    if  tam_loc ==`i'
}
/* The elasticities vary by cities*/
table tam_loc    [aw=sw] , c (mean elas ) row format(%4.3f)
```

```
/* Table 2 : Column 3 */
table tam_loc    [aw=sw] , c (mean d_cell ) row format(%4.3f)


/* Table 2 : Column 4 */
table tam_loc   [aw=sw] , c (mean pr1) row  format(%4.3f)

/* Table 2 : Column 5 */
table tam_loc   [aw=sw] , c (mean pr2) row format(%4.3f)

/* Table 2 : Column 6 */
table tam_loc   [aw=sw] , c (mean prdif) row format(%4.3f)

forvalues i=1/4 {
qui sum prdif [aw=sw] if tam_loc == `i'
local nprop_`i' = r(mean)
}

qui sum prdif [aw=sw]
local nprop_p = r(mean)

/* Table 2 : Column 7 */
cap drop temp
qui gen temp = 0
forvalues i=1/4 {
qui sum pr1  if tam_loc==`i' [aw=sw]
local m1 = r(mean)
qui sum pr2  if tam_loc==`i' [aw=sw]
local m2 = r(mean)
qui replace temp = (`m2'/`m1'-1)*100 if tam_loc==`i'
}

table tam_loc   [aw=sw] , c (mean temp) format(%4.3f)

*****************************************************************
*Table 3:  Estimating the impact on Well-being*
*****************************************************************

/* TABLE 03*/
/* Table 3 : Column 1 */
table tam_loc [aw=sw] if d_cell==1 , c (mean pc_exp_cell) row format(%4.3f)
```

```
/* Table 3 : Column 2*/
/* Here we do not use the stepwise because the option pswe(..) is not added */
cap drop ln_pc_income
cap drop ln_pc_exp_cell
gen ln_pc_exp_cell     = ln(pc_exp_cell)
qui gen ln_pc_income   = ln(pc_income)
cap drop predicted_exp
qui gen predicted_exp = .
cap drop ln_pc_exp_cell
qui gen ln_pc_exp_cell = ln(pc_exp_cell)
qui xi:   regress ln_pc_exp_cell ln_pc_income age  i.socio i.educ  [pw=sw] if d_cell==1
cap drop preda
qui predict preda, xb
qui replace predicted_exp = exp(preda)
table tam_loc [aw=sw] if d_cell==0  , c (mean predicted_exp) row format(%4.3f)
cap drop d_cell100
qui gen d_cell100= d_cell*100

/* Table 3 : Column 3*/
table tam_loc  [aw=sw] , c (mean d_cell100) row format(%4.3f)

/* Table 3 : Column 4*/
qui gen dif100 = prdif*100
qui gen difa = prdif
table tam_loc  [aw=sw] , c (mean dif100) row format(%4.3f)

/* Table 3 : Column 5 */
cap drop current_imp
qui gen  current_imp = income_change
table tam_loc  [aw=sw] if d_cell==1 , c (mean current_imp ) row format(%4.3f)

/* Table 3 : Column 6 */
cap drop new_imp
qui gen  new_imp = -predicted_exp*price_change
table  tam_loc  [aw=sw] if d_cell==0 , c (mean new_imp ) row format(%4.3f)

cap drop vec_c2
qui gen vec_c2 = new_imp

/* Table 3 : Column 7 */
cap drop current_imp
```

```
qui gen  current_imp = -pc_exp_cell*price_change
table tam_loc  [aw=sw]             , c (mean current_imp ) row format(%4.3f)

/* Table 3 : Column 8 */
cap drop tot_imp
qui drogen tot_imp = 0
forvalues i=1/4 {
qui sum difa      [aw=sw]  if (tam_loc==`i')
local prop = `r(mean)'
qui sum new_imp  [aw=sw]  if (tam_loc==`i') & d_cell==0
local imp_n = `r(mean)'
qui replace  tot_imp =  `prop'*`imp_n' if (tam_loc==`i')
}
table tam_loc  [aw=sw]             , c (mean tot_imp ) row format(%6.3f)

/* Table 3 : Column 9 */
cap drop tot_imp
qui drogen tot_imp = 0
forvalues i=1/4 {
qui sum difa      [aw=sw]  if (tam_loc==`i')
local prop = `r(mean)'
qui sum current_imp  [aw=sw]  if (tam_loc==`i')
local imp_o = `r(mean)'
qui sum new_imp  [aw=sw]  if (tam_loc==`i') & d_cell==0
local imp_n = `r(mean)'
qui replace  tot_imp = `imp_o' + `prop'*`imp_n' if (tam_loc==`i')
}
table tam_loc  [aw=sw]             , c (mean tot_imp ) row format(%6.3f)


/* TABLE 04  */
/* Create the 3 types of consumers (Never consume / Already consume / New consumer) */
 qui   qui drogen      _type_consumer  = 0
 qui   replace         _type_consumer  = 1 if d_cell==1
 cap lab drop type_consumer
 lab define type_consumer 0 "No entrant" 1 "Old user" 2 "New user"
 lab val _type_consumer  type_consumer


 /* Total population */
 qui sum sw
 local ss1 = r(sum)
```

```
 /* Total current users */
 qui sum sw if  _type_consumer ==1
 local ss2 = r(sum)

 /* Proportions of new users at population level : estimated in table 2 */
  local newprop =  `nprop_p'

  /*
       We randomly draw s from the nonusers the proportion of new users,
       but the probability of being drawn will depend on the probability
       of potential use, i.e. the higher is the probability of use, the higher is the
       probability of selection.
  */

 set seed 6543456
 qui drogen double random = uniform()*pr1*sw
 /* We do not draw among current users */
 qui replace random = 0  if  _type_consumer == 1

 /* We keep the original order of the data */
 qui drogen tkey = _n

 /*  We order by the actual type_of consumers, the groups,
     and then, the weighted random probability of use in descending order
       */
 gsort _type_consumer  tam_loc  - random

/* Within each population group, we seek for those:
       1-  with highest probability of selection;
       2-  initially they were nonusers;
       3-  Their proportion is equal to the new proportion of entrants of groups.
*/
 forvalues g=1/4 {
  qui sum sw if (tam_loc==`g')
  local   ss1 = r(sum)
  local prop = `nprop_`g''
   qui  drogen     double     tpsum  =  sum(sw*(_type_consumer!=1)*(tam_loc==`g')  )   if
(tam_loc==`g')
   qui       replace   tpsum = `ss1'*1.0001 if  _type_consumer ==1  & (tam_loc==`g')
   qui     replace  _type_consumer =  2 if (tpsum <= `prop'*`ss1'*1.0001) & (tam_loc==`g')
}
```

```
/* Incomes_t is the welfare with the situation t.
    t=0 : initial situation;
       t=1 : with only benefits of current users;
       t=2 : with only benefits of new users;
       t=3 : final situation (with current and new user benefits).
*/
drogen   _benefit = 0
replace  _benefit = income_change                      if _type_consumer==1
replace  _benefit = -predicted_exp*price_change        if _type_consumer==2
replace  _benefit = 0 if _benefit==.

/* Callibrating the benefits to meet those estimated with averages in table 2*/
forvalues g=1/4 {
qui sum vec_c2 [aw=sw] if (tam_loc==`g') & _type_consumer==2
local m1 = r(mean)
qui sum _benefit  [aw=sw]   if (tam_loc==`g')  & _type_consumer==2
if `r(mean)'!=0   qui  replace _benefit  = _benefit*`m1'/`r(mean)'  if (tam_loc==`g') &
_type_consumer==2
if `r(mean)'==0   qui  replace _benefit  = 0                        if (tam_loc==`g') &
_type_consumer==2
}



qui gen income0 = pc_income
qui gen income1 = pc_income +  (_benefit * (_type_consumer==1))
qui gen income2 = pc_income +  (_benefit * (_type_consumer==2))
qui gen income3 = pc_income +  (_benefit )

local lab0    Intial situtation
local lab1  With only benefits of current users
local lab2  With only benefits of new users
local lab3  Final situation

/* Table 04: Poverty headcount : 578 households currently being users, but the benefit
is not enough to escape poverty*/
forvalues i=0/3 {
cap drop poor`i'
qui gen  poor`i' = (income`i'<pline)
qui sum  poor`i' [aw=sw]
dis "`lab`i'' : " _col(36) %6.4f `r(mean)'*100
}
```

```
/* Table 05: Poverty gap */
forvalues i=0/3 {
cap drop poor`i'
qui gen poor`i' = max(0,(pline-income`i')/pline)
qui sum poor`i' [aw=sw]
dis "`lab`i'' : " _col(36) %6.4f `r(mean)'*100
}

/* Table 06: Gini */
forvalues i=0/3 {
qui fastgini income`i' [pw=sw]
local gini_`i' = r(gini)
dis "`lab`i'' : " _col(36) %6.4f `r(gini)'*100
}
```

Annex III

Full derivation of probit models and proportion of entrants when price variable is not available.

MCEMA offers users six alternative probability models to estimate new entrants, giving users a flexible set of tools to best model the decision to of individuals or households to enter a market. Each model is subsequently discussed in this section, with their respective specifications and data requirements.

In the first proposed model (see **M1** below), we exploit the variability of prices or incomes at the household level to estimate a *probability unit* or probit model.[17] Considering the assumption of a normal distribution of the error term in probit models, the probability of use/consumption can be written as:

$$Prob(d_h) = f(\log(I_h), \log(p_h), Z_h) \dots \boldsymbol{M1} \quad (9)$$

Z_h denotes the set of covariates for the household h. In the case where the prices' microdata are not available, the equation (1) is reduced to: $Prob(d_h) = f(\log(I_h), Z_h)$. If X denotes all covariates including the log of income ($\log(I_h) = \log(income_h)$) and potentially prices, we can write:

$$Prob(d = 1|X) = \Phi(X\beta) \quad (10)$$

Alternatively:

$$Prob(use = 1) = \Phi(\beta_0 + \beta_I \ln(income) + \cdots + u) \quad (11)$$

where $\Phi(.)$ denotes the normal cumulative distribution function (CDF). In what follow, we discuss the probabilistic elasticity with respect to income. However, the presented formulas that will follow are valid to price or to any over covariate of the model. Then, the probabilistic elasticity respect to income is given by:

$$\varepsilon_I = \frac{\partial Prob(.)}{\partial Income} * \frac{\overline{Income}}{\overline{Prob}} \quad (12)$$

where $\overline{Income}$ and $\overline{Prob}$ refer to the average income and average probability of having positive consumption at the population level (individual of reference), respectively. Taking derivatives and using the chain rule on the Probit model specification it is possible to rewrite equation (4) as:[18]

[17] In Stata, the stepwise prefix can be used to automatically perform the selection of explanatory variables according to their significance levels

[18] For more information on price elasticity and the Two-Part/Heckman models, see for instance, Saha et. al (1997).

$$\begin{aligned} \varepsilon_I &= \frac{\partial \Phi(XB)}{\partial (XB)} \frac{\partial (XB)}{\partial \ln (Income)} \frac{\partial \ln (Income)}{\partial (Income)} \frac{\overline{Income}}{\overline{Prob}} \\ &= \phi(.)\beta_I \left(\frac{1}{\overline{Income}}\right) \frac{\overline{Income}}{\overline{Prob}} \end{aligned} \tag{13}$$

where $\phi(.)$ denotes the probability density function (pdf) of the normal distribution. Thus:

$$\begin{aligned} \varepsilon_I &= \frac{\partial Prob}{\partial Income} \frac{\overline{Income}}{\overline{Prob}} \\ &= \frac{\partial \Phi(XB)}{\partial (XB)} \frac{\partial (XB)}{\partial \ln (Income)} \frac{\partial \ln (Income)}{\partial (Income)} \frac{\overline{Income}}{\overline{Prob}} \\ &= \phi(.) \frac{\beta_I}{\overline{Prob}} \end{aligned} \tag{14}$$

In order to estimate the probabilistic elasticity with respect to income, it is not necessary to observe incomes after changes (or prices after changes). The absolute change in probability is denoted by:

$$\begin{aligned} A_I &= \frac{\partial Prob}{\partial Income} dincome \\ &= \frac{\partial \Phi(XB)}{\partial (XB)} \frac{\partial (XB)}{\partial \ln(Income)} \frac{\partial \ln(Income)}{\partial (Income)} dincom \\ &= \phi(\bar{X}\beta) * \beta_I * \frac{dIncome}{\overline{Income}} \end{aligned} \tag{15}$$

In addition, we can also allow for interactions between the income and the quintile or decile dummies to estimate the price coefficients by income partitions, thus resulting in our second suggested model (**M2**).

Notice that the distributional effects of competition policies are not necessarily uniform across the income or consumption distribution. As expenditure shares on a specified item might vary across the distribution, it is important to control for the consumer heterogeneity to assess the welfare effects at a granular level. The added value of this model is that it enables the user to calculate income elasticities by welfare levels and, thus, determine the distributional effects of income variations across the income distribution. For the third proposed model (**M3**), we allow the user to interact income with population groups, such as by region. For the model (**M4**), we allow the user to interact income with the quintile/decile partition variable and another categorical variable of interest (population groups) such as rural/urban area. In the equation below, we allow for interactions of the log income variable $\ln(Income)$ with quintiles or deciles or any other group variable (rural/urban, for instance).

Then, the model thus takes the form:

$$Prob(use = 1) = \Phi\left(\beta_0 + \sum_{g=1}^{G} \beta_{I,g} \ln(Income) * I_g \ + \cdots + u\right) \ldots\ldots \boldsymbol{M2, M3, M4} \qquad (16)$$

where I_g is an indicator variable that takes the value of one if the household belongs to the group *g* and, otherwise, zero. More precisely, note that for $\boldsymbol{M2}$ we will have G income partition groups, for instance if we select a quintile income partition, we will have 5 population groups. For $\boldsymbol{M3}$, we will have G depending on the population group variable. For instance, for the *Rural/Urban* variable, $G = 2$. For $\boldsymbol{M4}$, the number of population groups G is equal to the number of group income partitions times the number of population groups (e.g., we have $G = 10$ when using the quintile income partition and the *Rural/Urban* population group variable).

In the rest of this document, we denote $\ln(Income) * I_g$ by $\ln(Income)_g$:

$$\frac{\partial Prob}{\partial \ln(Income)_g} \frac{\partial \ln(Income_g)}{\partial Income_g}\Big|_{Income_g = \overline{Income}_g} = \frac{\phi(.)\beta_{I,g}}{\overline{Income}_g} \qquad (17)$$

We define the probabilistic elasticity with respect to income within the group *g* by:

$$\varepsilon_{I,g} = \frac{\partial \mathrm{Prob}_g(.)}{\partial Income_g} \frac{\overline{Income}_g}{\overline{Prob}_g} \qquad (18)$$

Then, we can write:

$$\begin{aligned} \varepsilon_{I,g} &= \frac{\partial Prob_g}{\partial Prob} \frac{\partial Prob}{\partial \ln(Income_g)} \frac{\partial \ln(Income_g)}{\partial Income_g} \frac{\overline{Income}_g}{\overline{Prob}_g} \\ &= \frac{1}{\varphi_g} \varphi_g \phi(.) \beta_{I,g} \frac{1}{\overline{Prob}_g} \\ &= \phi(.) \frac{\beta_{I,g}}{\overline{Prob}_g} \end{aligned} \qquad (19)$$

where φ_g is the population share of group *g*.

In models 2 ,3 and 4, we assume that we run one population model. However, it is also possible to split the data by groups of interests, modelling each group independently. For example, the fourth model (M5), we assume that the user would like to run the models by population groups (e.g., Rural and then Urban). Thus, can be written as:

$$Prob_g(use = 1) = \Phi_g\left(\beta_{0,g} + \beta_{I,g} \ln(income) * I_g \ + \cdots + u_g\right) \qquad \boldsymbol{M5}$$

The last model **M6** is similar to model **M5,** except that income variable interacts with the variable *income partition groups*.

Based on equation (11) and estimating for each group g, we find that:

$$\varepsilon_{I,g} = \phi_g(.) * \frac{\beta_{I,g}}{Prob_g}$$

After estimating the probit models, we calculate the proportion of entrants in the population, assuming that **the change in the probability of the reference individual**, for whom the determinants have the average values, **represents the proportion of new entrants**. As shown above one can evaluate the change in the probability of the reference individual as:

$$A_I = \phi(\bar{X}\beta) * \beta_I * \frac{dIncome}{\overline{Income}} \quad (20)$$

Another method to estimate the proportion of entrants is to calculate the expected change in probabilities. Because on the nonlinearity, the formula above may be less precise. An alternative approach is to compute the difference between averages of the predicted probabilities with initial and final prices (which could lead to more accurate estimates), resulting in the estimated proportion of entrants. Formally, the change in probability of use of the household h is:

$$A_{I,h} = Prob_h(use = 1|X'_h) - Prob_h(use = 1|X_h) \quad (21)$$

The difference X'_h and X_h is the covariate income increases by the observed change. At population level, the expected change in probability of use is:

$$A_I = E[A_{I,h}] \quad (22)$$

In other words, the expected change in probability is equal to the expected predicted probability after the increase in income (X'), minus the expected probability under the initial values of incomes. The advantage of this second method is that we do not need to evaluate the density for the reference individual, which will give more accurate results as the two measurements will capture the main part of the change in probability. Notice that the computation of the proportion of entrants can also be done by population groups (for instance, rural/urban).