APPLICATION NOTE

Estimating Engel curves: A new way to improve the SILC-HBS matching process using GLM methods

Julio López-Laborda[a] Carmen Marin[b] and Jorge Onrubia[c]

[a]Department of Public Economics, Universidad de Zaragoza, Zaragoza, Spain and FEDEA; [b]FEDEA and Universidad Complutense de Madrid, Madrid, Spain;[c]Complutense Institute of International Studies (Universidad Complutense de Madrid), FEDEA and GEN

ARTICLE HISTORY
Compiled July 3, 2020

ABSTRACT

Microdata are required to evaluate the distributive impact of the taxation system as a whole (direct and indirect taxes) on individuals or households. However, in European Union countries this information is usually distributed into two separate surveys: the Household Budget Surveys (HBS), including total household expenditure and its composition, and EU Statistics on Income and Living Conditions (EU-SILC), including detailed information about households' income and direct (but not indirect) taxes paid. We present a parametric statistical matching procedure to merge both surveys. For the first stage of matching, we propose estimating total household expenditure in HBS (Engel curves) using a GLM estimator, instead of the traditionally used OLS method. It is a better alternative, insofar as it can deal with the heteroskedasticity problem of the OLS estimates, while making it unnecessary to retransform the regressors estimated in logarithms. To evaluate these advantages of the GLM estimator, we conducted a computational Monte Carlo simulation. In addition, when an error term is added to the deterministic imputation of expenditure in the EU-SILC, we propose replacing the usual Normal distribution of the error with a Chi-square type, which allows a better approximation to the original expenditures variance in the HBS. An empirical analysis is provided using Spanish surveys for years 2012-2016. In addition, we extend the empirical analysis to the rest of the European Union countries, using the surveys provided by Eurostat (EU-SILC, 2011; HBS, 2010).

KEYWORDS
Statistical matching surveys; Engel curve; household expenditure; heteroskedasticity; Generalized Linear Models (GLMs); Monte Carlo.

CONTACT Carmen Marín Email: cmarin@fedea.es.

ACKNOWLEDGMENTS The authors are very grateful to the referees and the Editor of the Journal for very useful suggestions that have helped them to improve the paper. Earlier versions of this paper were presented at workshop organised by the Joint Research Center, European Commission, Seville, Spain, on February 12, 2020; at the 24th Public Economics Meeting, Toledo, Spain, January 26–27, 2017; and at the 6th World Congress of the International Microsimulation Association, Torino, Italy, June 21-23, 2017.The authors gratefully acknowledge Daniel Santín (UCM) for his valuable suggestions in designing the Monte Carlo experiment, and the participants in the three conferences for their comments. All the remaining errors are entirely our responsibility.

FUNDING The authors acknowledge the financial support of Spanish Ministry of Science, Innovation and Universities (formerly, Ministry of Economy and Competitiveness), Project ECO2016-76506-C4-3-R (Julio López-Laborda and Jorge Onrubia) and Project ECO2017-87862-P (Carmen Marín). Julio López-Laborda also acknowledges the financial support of the Government of Aragon and the European Regional Development Fund (Public Economics Research Group).

JEL CLASSIFICATION
C15; C51; C52

1. Introduction

Most European Union countries collect data on household expenditure and household income in separate surveys, which are, respectively, the Household Budget Survey (HBS) and the European Union Statistics on Income and Living Conditions (EU-SILC). HBS provides information about household spending, while EU-SILC reports on household income, the main direct taxes and social contributions (in addition to certain public benefits, as well as other variables related to living conditions). In the case of the HBS, its design and content is established by the national statistical office of each country, while the income surveys are part of the EU-SILC project, designed and coordinated by the European Statistical Office (Eurostat)[1].

A single database with microdata on income and expenditure is therefore essential for studying the impact distribution of household tax burdens including direct and indirect taxes. In the case of indirect taxation, estimating the VAT and excise tax paid by households requires a microsimulation exercise based on the information on total expenditure and its composition contained in the HBS. However, as we have said, this lack of information is a real problem, common to practically all the countries of the European Union, insofar as this represents an important shortcoming in carrying out redistributive analyses of tax-benefit policies.

Although there are several ways to match household expenditure and income surveys ([9], [16] and [17], lately the matching problem has been solved using parametric matching methods, or in other words, regression imputation techniques. In this approach, the first step is to estimate in HBS the total expenditure of households (the so-called Engel curves), and then the household expenditure is imputed in EU-SILC using the regression coefficients (deterministic imputation) ([1], [22] and [23]).

Normally, in the parametric specification of Engel curves, the dependent variable is the logarithm of household expenditure, and the explanatory variables are the logarithm of income (linear, squared and cubed) and a set of specific household categorical dummy variables. The use of a log transformation to estimate the household expenditure is a common practice for dealing with skewness and excess kurtosis, besides reducing heteroskedasticity and diminishing the influence of outliers ([2] and [15]). However, in order to impute the expenditure in the EU-SILC, the researcher is interested in household expenditure in euro and not in logarithms. This problem is flagged in the literature as the retransformation problem and it is usually solved using a smearing estimate ([10]). But we must realize that in the presence of heteroskedasticity the smearing estimate does not work and produces a biased estimation ([18], [20] and [21]). Engel curves are traditionally estimated using OLS with robust standard errors. However, both the continuous (expenditure and income) and categorical variables that have a bearing on these expenditure functions usually produce intrinsic heteroskedasticity in linear estimates.

The aim of this paper is to select the most suitable method for estimating HBS expenditure in order to impute these results in the EU-SILC, taking into account both the problem of heteroskedasticity noted above and the need to retransform the re-

[1]The EU-SILC project entered into force in 2004 and currently covers all EU countries, Iceland, Norway, and Switzerland. For more information, see https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions.

gressors estimated in logarithms. This article presents six alternative estimate models involving an OLS regression of the expenditure in logarithms, and five different Generalized Linear Models (GLMs) alternatives, and concludes that GLM models become a better alternative than the traditionally used OLS method, since they do not suffer from the retransformation problem (the predictions are made on the raw cost scale, instead of the log-scale), and they allow to treat heteroskedasticity through the choice of distributional family, as explained by [15] and [20]. In addition, since an error term must be added to the deterministic imputation of the expenditure in the EU-SILC, we propose to replace the usual Normal distribution of the error used in the literature with a Chi-square type, which allows a better approximation to the HBS' original expenditures variance. This method improves the variance, skewness and kurtosis of the prediction, although it also increases the Root Mean Squared Error (RMSE). To evaluate the performance of GLM estimators compared to the traditional OLS method, we conducted a computational Monte Carlo simulation using a randomized data generation process based on the theoretical characteristics of Engel's relationship between disposable income and household consumption.

The proposed methodology is applied to the Spanish HBS (base 2006) data from 2012 to 2016 (INE [13]) and EU-SILC (base 2013) from 2013 to 2017 and ([14]), as the EU-SILC annual variables are referring to the previous year. The period considered coincides with the latest methodology update of EU-SILC from Eurostat, which improves the quality of the income variables using administrative tax registers. The analysis leads us to choose the GLMs log gamma under the Chi-squared procedure as the preferred model for estimating household expenditure in order to incorporate the results in the matching process. In addition, to test the robustness of the proposed methodology, we extend the empirical analysis to the rest of the European Union countries, using the micro data from the surveys provided by Eurostat (EU-SILC, 2011; HBS, 2010). Main results are provided on the Appendix of the paper [2].

The paper is structured as follows. Section 2 explains the usual methodology employed to estimate total household expenditure in HBS, identifies its weaknesses, and presents the GLM alternative. Section 3 describes the Monte Carlo simulation computed to validate the performance of GLM estimators. Section 4 contains an empirical application for the Spanish case, comparing the different estimate alternatives using the in-sample and out-sample predictions. Finally, Section 5 contains the main conclusions.

2. Methodology

In this section, we explain the usual methodology employed to estimate total household expenditure in HBS, identify its weaknesses, and present the GLM alternative.

In the OLS model proposed in [7] and [23]. The dependent variable is household monetary expenditure in logs $(ln(E_i))$[3] and the independent variables are the linear, square and cube logarithm of disposable income ($ln(y_i)$, $ln(y_i)^2$ and $ln(y_i)^3$) and the following household-specific dummy variables (vector x_i): population density, house-

[2]The extended study does not include all European Countries for several reasons. First, the HBS and EU-SILC surveys of Austria and Netherlands are not provided by Eurostat. Second, the Italian HBS does not include the variable disposable income. And third, United Kingdom is not considered, as this country elaborates a survey with jointly information of household income and expenditure called *Living Cost and Food Survey.*

[3]Monetary expenditure does not include the rental imputed or expenditure from self-supply, self-consumption and wages in kind.

hold members, household type, householder labour status, and household tenure. This model is presented in Equation 1 where t is referring to time (2012,... 2016) and the variables with the superscript B are collected from the HBS:

$$ln(E^B)_i^t = \alpha^t + \gamma_1^t ln(y_i^{B,t}) + \gamma_2^t ln(y_i^{B,t})^2 + \gamma_3^t ln(y_i^{B,t})^3 + x_i'^{B,t}\beta^t + \epsilon_i^t \quad (1)$$

As explained in the literature (see [6], [8] and [16]), the independent variables used in the matching process need to meet certain criteria: they must exist in both the HBS and EU-SILC surveys; they must have the same definition in both surveys; they must contribute significantly to explaining total expenditure, and they must have similar distributions in both surveys. In the empirical application contained in section 4, we have developed a harmonisation process for the independent variables in both surveys using the *Hellinger Distance* to choose the dummy variables. Since we have found that the HBS disposable income is underestimating the real value of disposable income as reflected by the EU-SILC disposable income (data collected from the administrative tax records), as in [7], the EU-SILC disposable income is rescaled in order to present similar mean and variance to the HBS disposable income.

Expenditure is imputed in the EU-SILC using the regression coefficients from the previous equation ($\hat{\alpha}^t$, $\hat{\gamma}_1^t$, $\hat{\gamma}_2^t$, $\hat{\gamma}_3^t$ and $\hat{\beta}^t$ and the independent variables from the EU-SILC (variables with a superscript I) as in Equation 2:

$$ln(\tilde{E}^I)_i^t = \hat{\alpha}^t + \hat{\gamma}_1^t ln(y_i^{I,t}) + \hat{\gamma}_2^t ln(y_i^{I,t})^2 + \hat{\gamma}_3^t ln(y_i^{I,t})^3 + x_i'^{I,t}\hat{\beta}^t \quad (2)$$

This model has two weaknesses. First, we are interested in household expenditure in levels and not in logarithms. As shown in [1] and [23], in order to impute the expenditure in the EU-SILC, total expenditure estimates must be corrected for retransformation bias using smearing estimates. However, if, as we have said, the estimates suffer from heteroskedasticity, the smearing estimates do not work so well and produce a bias in the retransformation process ([18]).

To illustrate these shortcomings, in Table 1 we display the Spanish HBS household expenditure (the dependent variable in the estimate process) in the period 2012-2016, which presents heavily right-skewed data and is leptokurtic. The skewness is around 2 and the kurtosis is around 10. These values are similar in the remaining European Union countries, as shown in Table A1 of the Appendix (data from Eurostat, HBS year 2010). As can be observed in the Figure 1, the HBS expenditure estimate for Spain using a log OLS regression presents a bias higher than 200 euro per household (even 300 euro and 400 euro in 2012 and 2014, respectively). As can be seen in Table 2, both the Breusch-Pagan/Cook-Weisberg test and the White test yield p-values equal to zero in all years, which allows us to reject the null hypothesis of homoskedasticity (constant variance). As Table A1 shows, the same result is obtained for the remaining 22 European Union Member States with HBS available at Eurostat.

Table 1. Spanish HBS Household Expenditure (2012-2016)

Year	Sample size	Population size	Mean (€)	Median (€)	Standard Deviation	Skewness	Kurtosis
2012	21,808	18,091,838	21,881	18,467	14,850	1.89	9.66
2013	22,057	18,212,214	20,979	17,755	14,490	2.05	10.95
2014	22,146	18,303,177	21,032	17,627	14,590	1.95	10.65
2015	22,130	18,374,351	21,439	17,930	14,973	2.06	11.55
2016	22,011	18,444,023	22,330	18,746	15,320	2.08	13.21

Source: Spanish HBS microdata provided by Spain's National Office of Statistics (INE) and own elaboration.

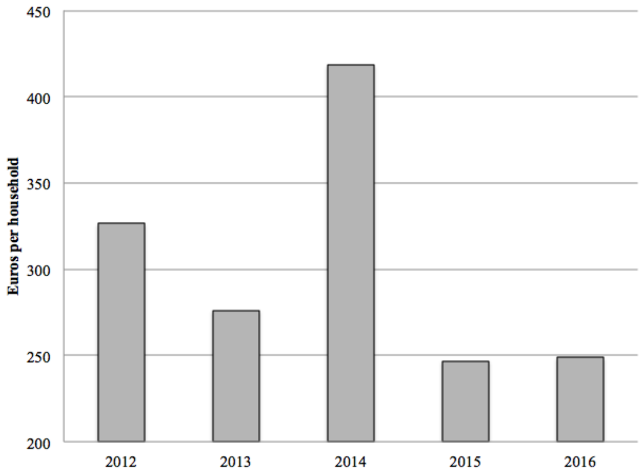

Figure 1. Bias HBS expenditure estimation via OLS in logarithms

Table 2. Breusch-Pagan/Cook-Weisberg Test and White Test for heteroskedasticity.Log OLS. Ho: Constant Variance

Year	Breusch Pagan/ Cook-Weisberg Test		White Test	
	Chi2(1)	p-value	Chi2(df)	p-value
2012	874.12	0.00	2622.69 (299)	0.00
2013	890.31	0.00	3173.04 (299)	0.00
2014	3322.21	0.00	5345.52 (301)	0.00
2015	1057.19	0.00	1490.02 (295)	0.00
2016	1247.21	0.00	4075.93 (302)	0.00

Note: Chi2(1) is a Chi-squared with one degree of freedom and Chi2(df) with df degrees of freedom.
Source: Spanish HBS microdata provided by Spain's National Office of Statistics (INE) and own elaboration.

This led us to reject the OLS model. The Generalized Least Squares (GLS) estimate method is the usual recommended way to solve the heteroskedasticity problem. However, we must realise that we are immersed in a matching procedure comprising an estimate step (HBS) and a deterministic imputation step (EU-SILC), this latter without standard errors. This fact led us to reject the GLS model, since its application is not feasible.

The GLMs have been reported in recent literature for estimating health expenditure, since estimates of health expenditure functions usually suffer from heteroskedasticity ([5], [15], [19], [20]). GLMs are generalizations of Non-Linear-Squares that are ideally suited to a nonlinear regression model with homoskedastic errors or with some kind of heteroskedasticity. They have been proposed as an alternative to OLS regression in logs. However, Baser [2] and Manning and Mullahy [20] have noted that GLMs are less accurate when kurtosis increases.

GLMs provide a number of estimate alternatives depending on the link function and the distributional family specified. GLMs do not suffer from the retransformation problem, and they allow dealing with heteroskedasticity through distributional families. The main disadvantage of these models is that the appropriate link function and distributional family need to be used for more accurate results. Extended Estimating Equations model (EEE) is a generalization of the GLMs proposed by Basu and Rathouz [4] to avoid the problems of misspecification due to the wrong choice of a family distribution or link function[4].

The second weakness concerning the methodology summarised in Equations 1 and 2 is that it results in a deterministic imputation of household expenditure. The main drawback is that the imputed expenditure has a lower standard deviation than the HBS expenditure. In our case, the R^2 of the regression is slightly higher than 0.5 in the whole period. To solve this problem, we have added an error term to the estimated and imputed expenditure with zero mean and a standard deviation such that the new variable generated has the same standard deviation as the original one. This method is called in [9] and [17] as *Stochastic Regression Imputation* and it is used in [23]. As the error terms of the regression are not normal, we propose to add an error term with a Chi-squared distribution with one degree of freedom[5], instead of a Normal distribution as in [23]. We will refer to this adjustment as Chi-squared procedure.

Figure 2 for the year 2013 shows the kernel density of the HBS expenditure and its estimate using GLM with a logarithm link and a family Gamma after adding an error term with a Normal and a Chi-squared distribution. As we see, on the left hand side of Figure 2, the shape of the estimated expenditure adjusted using a normally distributed error term is not similar to the shape of the HBS expenditure. However, the HBS expenditure and its estimate using the Chi-square procedure have similar kernel density distributions, as shown on the right hand side of Figure 2. We have obtained similar graphs for the whole period considered (from 2012 to 2016).

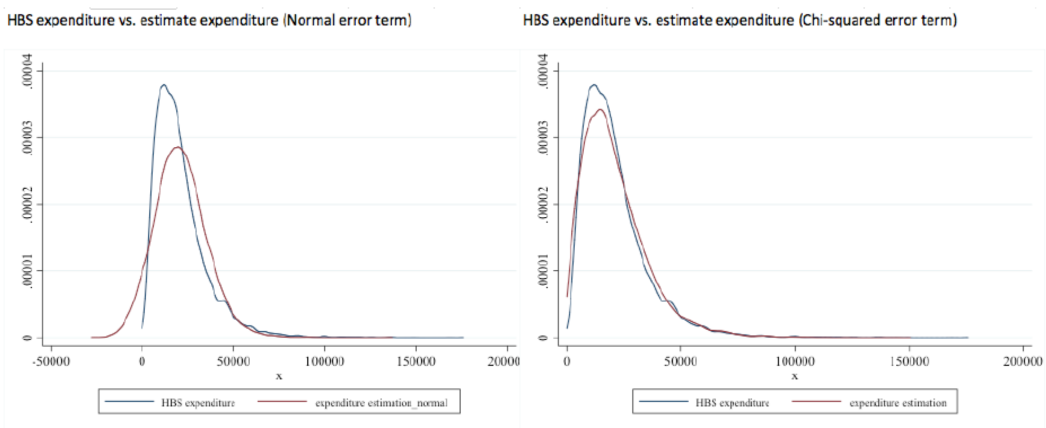

Figure 2. HBS expenditure and estimated expenditure Kernel density distribution. Year 2013

[4]However, as warned by [3] and [15], there is a high probability that the EEA estimation algorithm will not converge.

[5]A Chi-Squared distribution with one degree of freedom has a skewness of 2.82 and a kurtosis of 12. The HBS expenditure has a skewness of around 2 and a kurtosis of around 12.

This procedure's main advantage is that its moments are closer to the real data. By definition, the Chi-squared procedure presents a similar bias to the simple regression and the standard deviation of the Chi-squared estimate is similar to that of the original expenditure. The skewness and kurtosis of the Chi-squared procedure are higher than in the simple procedure, so they are nearer to the HBS expenditure data. Nevertheless, the drawback of the Chi-squared procedure is the loss of precision. The Root Mean Squared Error (RMSE) of the Chi-squared procedure is nearly 40% higher (from around 11,000 to around 15,000). In spite of its greater RMSE, we consider the Chi-squared procedure to be superior to the simple one, as it produces similar moments in the prediction to the original expenditure data and it presents, on average (by centiles), more accurate expenditure for households with lower and higher expenditures (Figure 3 shows these results for the year 2013. We have obtained similar results for the rest of the years of the period covered).

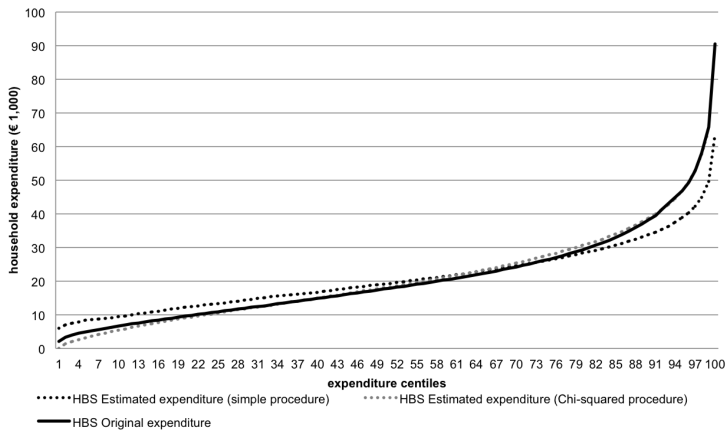

Figure 3. Average household expenditure centiles. Year 2013

3. Monte Carlo Simulation

In order to evaluate the performance of GLM estimators compared to the traditional OLS method, in this section we conducted a computational Monte Carlo simulation using a randomized data generation process based on the theoretical characteristics of Engel's relationship between disposable income and household consumption[6]. As it is well known, the relationship between both variables usually displays intrinsic heteroskedasticity in the data, with the variance of OLS residuals increasing with income ([12]). This is due to the greater variability shown by the average propensity to consume as household disposable income increases. This computational experiment consists of the following phases. Firstly, we use a data generation process (DGP) to create 2,000 random replications (R) of a database consisting of 20,000 observations (households) with two variables, disposable income (y_i^R) and consumption (E_i^R).

According to the previous premises, the DGP used is defined by the following expression, which characterizes a third grade primary Engel's function,

[6]On the microeconomic theoretical foundations of the Engel's curve, see [11]

$$E_i^R = \alpha + \beta_1 y_i^R + \beta_2 y_i^{R^2} + \beta_3 y_i^{R^3} + u_i^R \tag{3}$$

where u_i^R follows a random generation process based on a Normal distribution with zero mean and increasing variance with disposable income. The variance values are obtained from the residues of the OLS estimate of the above equation, segmented by income deciles (D):

$$u_i^R \sim N(0, \sigma^2_{\hat{u}_i/y^R_{iD}}) \tag{4}$$

In order to facilitate the comparison of our empirical findings with Monte Carlo evidence about the performance of the regression models analysed, the values of the parameters used in the DGP come from the OLS estimate of equation (3) using the Spanish HBS 2016 data. For the random generation of data corresponding to the disposable income (independent variable) a log-normal functional form is assumed, imposing mean and variance values equal to those offered for this variable in the HBS 2016 ($\bar{y}^{B,2016}$=23,010; $\sigma_y^{2B,2016}$=15,561). The values assigned to the parameters are: α=4125.328, β_1=0.940542, $\beta_2=-5.05e^{-6}$, β_3=1.27e^{-11}.

Secondly, for each of the 2,000 pairs of y_i^R and E_i^R generated, the different estimate models compared are run: GLM Log Gamma, GLM Square root Gamma, GLM Log Poisson, GLM Log Normal (with disposable income in logarithms) and the Log OLS model (with expense and disposable income in logarithms) which includes the estimate of the expense after undoing the logarithm using the smearing estimator technique. As expected, the estimate of the EEE model has presented convergence problems.

Finally, for the 2,000 estimates of each model, the values of bias, skewness, kurtosis and RMSE are obtained. The results, expressed in mean, lower and upper limits of the 95% confidence interval, are shown in Table 3.

Table 3. Estimated household expenditure (Equation 1). Log-transformed OLS regression versus GLM models comparison (statistical moments) using a simulation excercise of household disposable income and expenditure (20,000 observations and 2,000 replicates). 95% Cofindence intervals.

Model		Log OLS	GLM Sqrt Gamma	GLM Log Gamma	GLM Log Poisson	GLM Log Normal
Bias (€)	Lower limit	-778.23	-5.33	-3.05	0.00	-1.30
	Mean	-7.09	-1.41	-0.85	0.00	-0.33
	Upper limit	-640.00	0.55	0.49	0.00	1.00
Skewness	Lower limit	1.21	1.23	1.16	1.18	1.19
	Mean	1.40	1.39	1.30	1.32	1.33
	Upper limit	1.61	1.57	1.45	1.46	1.48
Kurtosis	Lower limit	4.83	5.24	4.61	4.74	4.81
	Mean	5.98	6.28	5.42	5.58	5.76
	Upper limit	7.57	7.58	6.40	6.53	6.86
RMSE	Lower limit	11314	11274	11273	11273	11273
	Mean	11444	11405	11405	11404	11404
	Upper limit	11574	11536	11537	11536	11535

Note: The EEE Model is not shown because of lack of convergence
Source: Own elaboration.

As can be seen from the results of the Monte Carlo experiment carried out, on the one hand, the traditional use of a Log OLS model to estimate the Engel curve specified in equation (1) produces biased estimates of household expenditure. This is due to the non-correction of the intrinsic heteroskedasticity caused by the structure of the income and consumption data, as well as the necessary use of the smearing estimator for the retransformation of the logarithms. On the other hand, GLM estimators offer in all cases a better performance than Log OLS estimate, since they satisfactorily mitigate the bias introduced by heteroskedasticity, and avoid the retransformation of logarithms. Poisson GLM model provides the best result, with an unbiased estimate.

4. Empirical Application

In this section, an empirical analysis is carried out to determine the most accurate model for the matching between HBS and EU-SILC, using Spanish SILC and HBS for years 2012-2016. Firstly, we present the HBS estimated expenditure (Equation 1) for six different models: OLS regression in logarithms and five GLM alternatives: GLM Square root Gamma, GLM log Gamma, GLM log Poisson, GLM log Normal and the EEE model. Then, the EU-SILC imputed total expenditure (Equation 2) statistics are shown only for the chosen alternative. All models are run using a Chi-square procedure.

Table 4 shows the HBS estimated expenditure moments for each model using a tenfold cross-validation process to test the accuracy of out-sample forecasts. With respect to the bias, the high bias of the OLS estimate in logarithms can be observed. We consider that the other models have an assumable bias. By definition of the Chi-square procedure, the standard deviation of the estimated expenditure is the same as the HBS expenditure (the dependent variable). Skewness and kurtosis present similar values for all the models. The lowest RMSE value is the criterion used to choose our preferred estimation model. We can observe that the model with the lowest RMSE for the period considered is the OLS regression in logarithms; however, we reject this model because of the high bias, as we have already anticipated. Thus, the model which presents the second lowest RMSE is the GLM with a log link function and a Gamma distribution family. This is our chosen alternative to estimate expenditure in the HBS and impute the results in the EU-SILC.

Table 4. Spain's HBS estimated household expenditure (Equation 1) (2012-2016). Log-transformed OLS regression versus GLM models comparison (statistical moments). (Bootstrap: 100 replicates).

Model	Bias (€)		Skewness		Kurtosis		RMSE	
	In sample	Out sample	In sample	Out sample	In sample	Out sample	In sample	Out sample
2012								
Log OLS	326.55	318.27	1.30	1.30	6.16	6.21	15,029	15,045
GLM sqrt Gamma	38.49	13.39	1.40	1.40	6.77	6.69	15,411	15,391
GLM log Gamma	55.82	31.19	1.44	1.43	6.89	6.82	15,266	15,248
GLM log Poisson	-21.91	-13.95	1.44	1.45	6.84	6.99	15,405	15,437
GLM log Normal	13.61	-11.81	1.47	1.46	7.02	6.92	15,487	15,476
EEE	-10.07	-7.68	1.40	1.40	6.85	6.82	15,560	15,581
2013								
Log OLS	276.15	273.91	1.36	1.36	6.40	6.44	14,643	14,645
GLM sqrt Gamma	56.38	47.65	1.48	1.47	7.18	7.13	14,994	14,993
GLM log Gamma	77.14	68.39	1.52	1.52	7.46	7.49	14,850	14,853
GLM log Poisson	17.67	10.00	1.47	1.47	7.00	7.03	15,045	15,056
GLM log Normal	-4.00	-13.33	1.40	1.39	6.71	6.67	14,982	14,995
EEE	-5.03	-6.07	1.46	1.47	7.03	7.13	15,039	15,061
2014								
Log OLS	418.52	402.87	1.34	1.33	6.31	6.22	14,751	14,759
GLM sqrt Gamma	36.24	36.42	1.43	1.44	6.79	6.91	15,119	15,133
GLM log Gamma	49.74	50.11	1.47	1.48	6.92	7.04	15,004	15,020
GLM log Poisson	6.33	0.32	1.45	1.46	6.93	6.98	15,256	15,279
GLM log Normal	-20.21	-20.30	1.41	1.43	6.76	6.89	15,218	15,244
EEE	-5.09	-43.42	1.43	1.43	6.82	6.88	15,245	15,254
2015								
Log OLS	246.86	242.75	1.33	1.33	6.43	6.36	15,326	15,350
GLM sqrt Gamma	-8.83	-10.54	1.44	1.44	6.94	6.93	15,710	15,712
GLM log Gamma	16.83	15.26	1.45	1.45	6.87	6.85	15,559	15,566
GLM log Poisson	-9.30	-6.72	1.42	1.44	6.71	6.85	15,587	15,627
GLM log Normal	-42.30	-43.78	1.41	1.41	6.77	6.75	15,588	15,606
EEE	20.93	25.01	1.44	1.44	6.88	6.93	15,630	15,657
2016								
Log OLS	248.76	257.48	1.29	1.30	6.04	6.14	15,411	15,439
GLM sqrt Gamma	11.14	4.47	1.41	1.41	6.80	6.77	15,911	15,930
GLM log Gamma	53.45	47.01	1.42	1.41	6.65	6.62	15,696	15,719
GLM log Poisson	-3.14	-23.31	1.43	1.43	6.78	6.79	15,771	15,762
GLM log Normal	14.50	7.82	1.43	1.42	6.77	6.73	15,759	15,792
EEE	-20.88	-21.23	1.40	1.41	6.69	6.74	15,737	15,771

Source: Spanish HBS microdata provided by Spain's National Office of Statistics (INE), and own elaboration.

We have conducted in the Appendix an extended study to show that the proposed approach can work on similarly on different datasets from another European Countries (See Table A2).

To conclude this section, we compare the statistics of the HBS expenditure with the EU-SILC imputed expenditure using a GLM with log link and Gamma distribution family. As can be observed in Table 5, EU-SILC imputed expenditure presents a similar mean and standard deviation; however, the skewness and kurtosis values are smaller than in HBS expenditure. Similar results have been obtained in the extension of the estimates for the remaining European Union countries (see Table A3 in the Appendix). For each country, the family of GLM that offers the best results in terms of bias reduction has been used.

Table 5. Spain's HBS expenditure (dependent variable) vs. GLM log gamma Spain's SILC imputed expenditure (2012-2016) (Equation 2)

Year	Mean (€)		Standard Deviation		Skewness		Kurtosis	
	HBS expenditure	SILC imputation	HBS expenditure	SILC imputation	HBS expenditure	SILC imputation	HBS expenditure	SILC imputation
2012	21,881	22,075	14,850	15,056	1.89	1.59	9.66	7.30
2013	20,979	20,960	14,490	14,458	2.05	1.41	10.95	6.38
2014	21,032	21,173	14,590	14,909	1.95	1.56	10.65	6.83
2015	21,439	21,627	14,973	15,246	2.06	1.61	11.55	7.94
2016	22,330	22,358	15,320	15,451	2.08	1.50	13.21	6.91

Source: Spanish HBS and SILC microdata provided by Spain's National Office of Statistics (INE), and own elaboration.

5. Conclusion

The distributive analysis of household tax burden, including direct and indirect taxes, is essential for choosing appropriate tax policies, including the choice of the tax-mix. However, in the European Union, the vast majority of National Statistical Institutes do not usually create surveys combining information about household income and expenditures. In fact, in those countries this information is presented in two separate surveys. Given this limitation, statistical matching techniques are the only option for creating a survey that presents household income and expenditure together.

Against the backdrop of contributing to the literature with a matching procedure for Spanish data from 2012 to 2016, in this article we present a suitable method for estimating HBS expenditure in order to impute these results in the EU-SILC. Lately, the most common technique involves estimating Engel curves using Ordinary Least Squares in logs with HBS data to impute household expenditure in the income data set (EU-SILC). Estimation in logs has certain advantages, since it can deal with skewness in data and reduce heteroskedasticity. However, the model needs to be corrected with a smearing estimate to retransform the results into levels (euros). The presence of intrinsic heteroskedasticity in household expenditure requires another estimate technique, as the smearing estimate produces a bias.

As shown in the paper, our proposal to estimate Engel curves using GLM estimators is a superior alternative to the traditional OLS method, since it is an option that corrects the usual bias problems caused by the intrinsic heteroskedasticity of the data used, while making it unnecessary to retransform the logarithms of the regressors. This methodological proposal has been validated by conducting a Monte Carlo experiment, in which 2,000 replicates of disposable income and consumption data have been generated for 20,000 households. In the empirical application for the Spanish case, the GLM log gamma under the Chi-squared procedure is selected as the best option. In the exercise carried out, our model presents an accurate level of expenditure for low and high-income households. As we have empirically tested, the best performance of the GLM estimators also happens in the estimates of the Engel curves for the statistical fusion of the SILC and HBS of the rest of the European Union countries.

Appendix A. Tables

Table A.1. HBS Household Expenditure in European countries (2010) and Breusch Pagan/Cook-Weisberg and White Test for heteroskedasticity

Country	Sample size	Mean (€)	Standard Deviation	Skewness	Kurtosis	Breusch Pagan/ Cook-Weisberg Test		White Test	
						Chi2(1)	p-value	Chi2(df)	p-value
1.Belgium	7,168	34,302	21,601	2.50	15.60	32.36	0.00	566.62 (121)	0.00
2.Bulgaria	2,982	4,657	2,615	1.38	6.05	31.23	0.00	464.48 (116)	0.00
3.Cyprus	2,702	39,427	25,667	1.50	7.11	63.86	0.00	1005.18 (113)	0.00
4.Czech Republic	2,932	9,791	5,065	1.45	7.46	0.09	0.76	254.16 (100)	0.00
5.Germany	53,996	29,199	20,118	2.78	18.65	87.71	0.00	6853.48 (102)	0.00
6.Denmark	2,484	39,793	22,739	1.58	8.35	7.15	0.01	209.13 (106)	0.00
7.Estonia	3,632	7,776	6,146	2.66	18.25	34.52	0.00	49.34 (101)	0.99
8.Greece	3,512	28,143	19,386	1.97	9.23	36.44	0.00	233.58 (121)	0.00
9.Finland	3,551	32,608	21,920	1.82	9.15	14.59	0.00	198.46 (112)	0.00
10.France	15,797	30,330	19,161	1.85	9.33	146.10	0.00	3479.17 (126)	0.00
11.Croatia	3,459	12,941	7,053	1.12	4.79	95.93	0.00	518.11 (120)	0.00
12.Hungary	9,937	8,485	4,454	1.99	11.68	54.02	0.00	1076.29 (124)	0.00
13.Ireland	5,891	38,908	22,280	1.29	5.67	387.01	0.00	822.73 (126)	0.00
14.Lithuania	6,103	9,343	5,861	2.05	11.62	41.71	0.00	339.89 (116)	0.00
15.Latvia	3,798	8,020	6,270	3.56	28.32	0.01	0.91	306.14 (111)	0.00
16.Malta	3,732	20,518	15,362	2.91	20.30	64.70	0.00	135.36 (44)	0.00
17.Poland	37,412	9,202	6,116	3.91	38.34	113.78	0.00	2848.19 (126)	0.00
18.Portugal	9,484	20,391	14,963	1.96	8.60	129.37	0.00	374.07 (124)	0.00
19.Romania	31,336	5,513	3,200	2.85	31.36	587.57	0.00	3935.70 (94)	0.00
20.Sweden	2,047	28,299	16,751	2.36	18.84	17.63	0.00	270.57 (48)	0.00
21.Slovenia	3,924	21,922	12,708	1.87	10.04	27.45	0.00	388.41 (124)	0.00
22.Slovakia	6,143	10,550	6,365	5.82	92.65	50.43	0.00	530.59 (118)	0.00

Note: Chi2(1) is a Chi-squared with one degree of freedom and Chi2(df) with df degrees of freedom.
Source: HBS microdata provided by Eurostat (HBS, 2010) and own elaboration.

Table A.2. European Union countries' HBS estimated household expenditure (Equation 1) (2010). Log-transformed OLS regression versus GLM models comparison (statistical moments). (Bootstrap: 100 replicates).

Model	Bias (€)		Skewness		Kurtosis		RMSE	
	In sample	Out sample	In sample	Out sample	In sample	Out sample	In sample	Out sample
1. Belgium								
Log OLS	198.35	172.26	1.64	1.66	9.36	9.74	22,946	22,977
GLM sqrt Gamma	215.07	188.95	1.58	1.58	8.25	8.16	22,988	22,988
GLM log Gamma	95.44	92.40	2.15	2.23	14.92	17.10	22,639	22,684
GLM log Poisson	39.58	0.38	1.55	1.56	7.76	7.97	23,277	23,260
GLM log Normal	3.08	37.76	1.43	1.43	6.93	6.99	23,354	23,459
EEE	44.60	-38.78	1.52	1.55	7.16	7.59	23,640	23,603
2.Bulgaria								
Log OLS	32.86	34.93	1.34	1.86	6.11	34.18	1,819	1,946
GLM sqrt Gamma	14.66	12.84	1.13	1.27	5.00	5.22	1,941	1,948
GLM log Gamma	21.09	14.75	1.38	1.38	6.38	6.35	1,833	1,827
GLM log Poisson	-15.11	-6.96	1.23	1.26	5.36	5.64	1,951	1,967
GLM log Normal	9.74	14.95	1.25	1.25	5.53	5.52	1,996	2,010
EEE	NA	NA	NA	NA	NA	NA	NA	NA
3.Cyprus								
Log OLS	413.56	362.32	1.02	1.03	5.17	5.29	23,467	23,454
GLM sqrt Gamma	196.20	192.26	1.17	1.20	6.07	6.27	23,547	23,678
GLM log Gamma	172.97	211.61	1.37	1.42	8.23	8.73	23,564	23,658
GLM log Poisson	32.96	17.95	1.06	1.06	5.38	5.46	24,437	24,452
GLM log Normal	-162.36	-84.09	0.94	0.92	4.96	4.85	23,844	24,004
EEE	NA	NA	NA	NA	NA	NA	NA	NA
4.Czech Republic								
Log OLS	-14.14	-13.85	0.86	0.86	4.44	4.47	4,226	4,236
GLM sqrt Gamma	-3.18	6.43	0.86	0.91	4.43	4.83	4,213	4,255
GLM log Gamma	1.68	-3.93	0.86	0.86	4.40	4.38	4,191	4,200
GLM log Poisson	10.48	-5.85	0.86	0.86	4.30	4.28	4,146	4,146
GLM log Normal	-1.69	-15.96	0.87	0.88	4.18	4.19	4,101	4,106
EEE	NA	NA	NA	NA	NA	NA	NA	NA
5.Germany								
Log OLS	14.66	25.51	1.42	1.43	6.29	6.36	18,404	18,418
GLM sqrt Gamma	153.66	143.29	1.43	1.43	6.22	6.25	17,875	17,875
GLM log Gamma	128.91	134.12	1.64	1.64	7.59	7.57	17,651	17,662
GLM log Poisson	2.39	6.78	1.36	1.36	6.07	6.06	18,534	18,533
GLM log Normal	-54.19	-79.02	1.25	1.24	5.60	5.57	18,455	18,454
EEE	12.86	17.38	1.31	1.31	5.79	5.79	18,531	18,536
6.Denmark								
Log OLS	180.59	211.47	1.19	1.16	5.81	5.60	21,132	21,193
GLM sqrt Gamma	-17.02	-39.53	1.28	1.29	6.42	6.52	20,760	20,953
GLM log Gamma	37.39	26.16	1.20	1.21	5.82	6.00	20,937	21,166
GLM log Poisson	78.64	14.74	1.08	1.10	5.23	5.35	21,102	21,353
GLM log Normal	-151.66	-152.83	1.00	1.00	4.82	4.87	20,567	20,903
EEE	NA	NA	NA	NA	NA	NA	NA	NA

Source: Microdata provided by Eurostat (HBS, 2010), and own elaboration.

Table A.2. (continued). European Union countries' HBS estimated household expenditure (Equation 1) (2010). Log-transformed OLS regression versus GLM models comparison (statistical moments). (Bootstrap: 100 replicates).

Model	Bias (€)		Skewness		Kurtosis		RMSE	
	In sample	Out sample	In sample	Out sample	In sample	Out sample	In sample	Out sample
7.Estonia								
Log OLS	38.74	29.00	1.49	1.52	6.58	6.82	6,225	6,260
GLM sqrt Gamma	16.37	16.07	1.47	1.50	6.55	6.82	6,232	6,257
GLM log Gamma	18.74	15.41	1.48	1.47	6.65	6.55	6,236	6,267
GLM log Poisson	3.37	12.27	1.51	1.51	6.69	6.73	6,159	6,243
GLM log Normal	-15.83	-22.35	1.56	1.52	6.98	6.74	6,120	6,169
EEE	NA	NA	NA	NA	NA	NA	NA	NA
8.Greece								
Log OLS	154.86	145.78	1.76	1.79	8.33	8.58	14,828	14,860
GLM sqrt Gamma	4.97	20.13	1.62	1.62	7.23	7.21	15,607	15,677
GLM log Gamma	98.23	92.46	1.72	1.77	8.03	8.37	15,353	15,428
GLM log Poisson	-31.80	-25.71	1.58	1.61	6.97	7.27	15,829	15,947
GLM log Normal	-28.16	-37.07	1.53	1.59	6.65	7.50	15,809	15,875
EEE	NA	NA	NA	NA	NA	NA	NA	NA
9.Finland								
Log OLS	257.89	274.00	1.45	1.57	7.62	9.04	19,845	19,979
GLM sqrt Gamma	94.89	93.99	1.43	1.48	7.28	8.15	19,792	19,854
GLM log Gamma	222.39	152.96	1.65	1.75	9.68	11.06	19,612	19,677
GLM log Poisson	7.04	-39.64	1.31	1.35	6.15	6.66	20,544	20,574
GLM log Normal	-37.83	-34.59	1.20	1.21	5.57	5.66	20,402	20,506
EEE	NA	NA	NA	NA	NA	NA	NA	NA
10.France								
Log OLS	455.26	447.59	1.65	1.68	7.67	8.12	18,278	18,318
GLM sqrt Gamma	129.88	102.95	2.08	2.11	16.91	17.34	18,814	18,814
GLM log Gamma	233.64	257.62	1.93	2.00	12.66	10.16	18,360	18,436
GLM log Poisson	28.37	55.28	1.65	1.81	8.87	12.31	19,532	19,543
GLM log Normal	-127.50	-102.91	1.32	1.34	6.48	6.61	19,170	19,240
EEE	53.06	64.09	1.56	1.57	8.15	8.33	19,326	19,342
11.Croatia								
Log OLS	147.63	138.31	0.96	0.95	4.57	4.49	6,054	6,054
GLM sqrt Gamma	54.12	70.48	0.94	0.95	4.54	4.62	6,173	6,198
GLM log Gamma	34.93	45.54	1.02	1.02	4.78	4.76	6,140	6,177
GLM log Poisson	-3.71	-27.39	1.05	1.06	5.01	5.14	6,372	6,376
GLM log Normal	-4.41	-6.86	1.36	1.35	6.33	6.26	4,143	4,155
EEE	NA	NA	NA	NA	NA	NA	NA	NA
12.Hungary								
Log OLS	27.57	29.25	1.49	1.51	7.30	7.47	4,058	4,065
GLM sqrt Gamma	35.71	30.29	1.47	1.47	7.04	7.02	4,071	4,079
GLM log Gamma	21.20	23.29	1.70	1.71	9.08	9.23	4,010	4,018
GLM log Poisson	7.87	5.17	1.49	1.51	7.14	7.31	4,157	4,169
GLM log Normal	-4.41	-6.86	1.36	1.35	6.33	6.26	4,143	4,155
EEE	NA	NA	NA	NA	NA	NA	NA	NA

Source: Microdata provided by Eurostat (HBS, 2010), and own elaboration.

Table A.2. (continued). European Union countries' HBS estimated household expenditure (Equation 1) (2010). Log-transformed OLS regression versus GLM models comparison (statistical moments). (Bootstrap: 100 replicates).

Model	Bias (€)		Skewness		Kurtosis		RMSE	
	In sample	Out sample	In sample	Out sample	In sample	Out sample	In sample	Out sample
13.Ireland								
Log OLS	555.04	534.61	1.68	1.74	11.40	12.32	18,827	18,820
GLM sqrt Gamma	38.41	105.32	1.31	1.32	6.51	6.68	20,119	20,243
GLM log Gamma	179.01	145.49	1.84	1.89	12.79	13.50	19,523	19,519
GLM log Poisson	-19.62	-7.89	1.43	1.46	7.39	7.75	20,375	20,413
GLM log Normal	-27.64	5.04	1.20	1.23	5.82	6.01	20,502	20,560
EEE	NA	NA	NA	NA	NA	NA	NA	NA
14.Lithuania								
Log OLS	74.56	72.70	1.37	1.34	6.47	6.23	6,037	6,035
GLM sqrt Gamma	25.14	19.25	1.38	1.36	6.73	6.48	6,137	6,148
GLM log Gamma	19.72	17.80	1.44	1.41	6.93	6.62	6,062	6,069
GLM log Poisson	9.33	10.18	1.48	1.47	7.05	6.97	6,184	6,216
GLM log Normal	10.22	8.90	1.49	1.52	7.22	7.49	6,205	6,270
EEE	NA	NA	NA	NA	NA	NA	NA	NA
15.Latvia								
Log OLS	14.99	29.32	2.37	2.40	15.03	15.70	6,000	6,066
GLM sqrt Gamma	-5.37	1.03	1.83	1.85	8.98	9.20	6,392	6,421
GLM log Gamma	34.18	28.70	2.53	2.59	17.46	18.49	5,966	5,987
GLM log Poisson	-11.66	-5.77	2.11	2.15	11.76	12.52	6,161	6,215
GLM log Normal	-12.70	-8.06	2.25	2.29	13.31	14.25	6,123	6,254
EEE	NA	NA	NA	NA	NA	NA	NA	NA
16.Malta								
Log OLS	254.24	213.27	1.63	1.70	8.15	9.69	17,863	17,816
GLM sqrt Gamma	23.16	24.11	1.84	1.82	9.74	9.59	18,124	18,137
GLM log Gamma	51.09	34.13	1.77	1.78	8.82	8.93	18,297	18,270
GLM log Poisson	-8.95	45.06	1.81	1.91	9.00	10.70	18,188	18,313
GLM log Normal	16.40	5.94	1.89	1.95	9.44	10.76	18,266	18,422
EEE	NA	NA	NA	NA	NA	NA	NA	NA
17.Poland								
Log OLS	-8.54	-7.01	1.96	2.01	13.07	14.17	6,419	6,429
GLM sqrt Gamma	30.11	30.11	2.06	2.06	14.74	14.73	6,294	6,296
GLM log Gamma	46.53	48.97	5.03	5.38	139.55	160.70	6,258	6,254
GLM log Poisson	0.64	-0.11	1.67	1.70	8.51	8.86	6,465	6,469
GLM log Normal	-37.39	-38.04	1.46	1.46	7.03	7.03	6,367	6,371
EEE	6.22	4.15	1.69	1.69	8.75	8.78	6,460	6,463
18.Portugal								
Log OLS	225.19	229.61	1.52	1.53	6.65	6.70	14,233	14,266
GLM sqrt Gamma	41.88	49.71	1.63	1.66	7.26	7.52	14,424	14,512
GLM log Gamma	11.46	18.03	1.61	1.60	7.16	7.03	14,515	14,578
GLM log Poisson	31.89	33.39	1.54	1.54	6.75	6.72	14,674	14,698
GLM log Normal	-95.78	-95.81	1.51	1.53	6.46	6.73	14,513	14,594
EEE	-24.59	-12.66	1.58	1.58	7.10	7.12	14,685	14,737

Source: Microdata provided by Eurostat (HBS, 2010), and own elaboration.

Table A.2. (conclusion). European Union countries' HBS estimated household expenditure (Equation 1) (2010). Log-transformed OLS regression versus GLM models comparison (statistical moments). (Bootstrap: 100 replicates).

Model	Bias (€)		Skewness		Kurtosis		RMSE	
	In sample	Out sample	In sample	Out sample	In sample	Out sample	In sample	Out sample
19.Romania								
Log OLS	32.98	32.21	1.58	1.58	10.15	10.10	2,719	2,718
GLM sqrt Gamma	22.28	22.25	1.57	1.57	9.48	9.47	2,645	2,644
GLM log Gamma	27.13	27.04	2.70	2.75	32.13	33.55	2,624	2,625
GLM log Poisson	-0.94	-1.55	1.37	1.36	7.03	6.91	2,831	2,830
GLM log Normal	-11.37	-10.66	1.16	1.16	5.63	5.58	2,802	2,812
EEE	NA	NA	NA	NA	NA	NA	NA	NA
20.Sweden								
Log OLS	142.83	136.89	1.87	2.37	15.04	27.47	16,431	16,401
GLM sqrt Gamma	192.91	211.47	1.76	1.79	12.15	12.69	16,217	16,240
GLM log Gamma	205.20	199.08	3.19	3.74	46.72	61.77	16,285	16,387
GLM log Poisson	-77.92	-44.69	1.28	1.34	6.14	6.81	16,949	17,005
GLM log Normal	-112.52	-67.44	1.14	1.16	5.71	5.84	16,633	16,787
EEE	NA	NA	NA	NA	NA	NA	NA	NA
21.Slovenia								
Log OLS	184.09	177.81	1.45	1.47	7.47	7.72	12,360	12,399
GLM sqrt Gamma	82.87	108.03	1.27	1.28	6.09	6.18	12,735	12,759
GLM log Gamma	101.09	113.33	1.52	1.52	8.00	8.00	12,356	12,395
GLM log Poisson	-2.96	-17.06	1.42	1.42	6.91	6.92	12,712	12,746
GLM log Normal	-17.44	-29.30	1.32	1.31	6.50	6.39	12,896	12,915
EEE	NA	NA	NA	NA	NA	NA	NA	NA
22.Slovakia								
Log OLS	-30.49	-39.79	1.43	1.44	7.22	7.37	6,573	6,579
GLM sqrt Gamma	-7.61	-16.45	1.42	1.42	7.07	7.06	6,531	6,535
GLM log Gamma	4.17	-3.56	1.45	1.45	7.35	7.45	6,472	6,488
GLM log Poisson	-18.15	-9.01	1.47	1.50	7.64	8.09	6,322	6,371
GLM log Normal	-3.05	-6.12	1.50	1.59	7.65	9.07	6,426	6,498
EEE	NA	NA	NA	NA	NA	NA	NA	NA

Source: Microdata provided by Eurostat (HBS, 2010), and own elaboration.

Table A.3. European Union Countries' HBS expenditure (dependent variable) vs. European Union Countries' SILC imputed expenditure using GLM (2010) (Equation 2)

Country	Mean		Standard Deviation		Skewness		Kurtosis	
	HBS expenditure	SILC imputation	HBS expenditure	SILC imputation	HBS expenditure	SILC imputation	HBS expenditure	SILC imputation
1.Belgium	34,302	34,376	21,601	20,939	2.50	1.61	15.60	9.09
2.Bulgaria	4,657	4,721	2,615	2,611	1.38	2.37	6.05	22.09
3.Cyprus	39,427	39,349	25,667	25,432	1.50	1.39	7.11	7.39
4.Czech Republic	9,791	9,869	5,065	5,109	1.45	1.06	7.46	4.78
5.Germany	29,199	29,203	20,118	20,716	2.78	1.68	18.65	8.57
6.Denmark	39,793	39,709	22,739	23,035	1.58	1.20	8.35	6.10
7.Estonia	7,776	7,814	6,146	5,950	2.66	1.32	18.25	5.51
8.Greece	28,143	28,063	19,386	19,580	1.97	2.69	9.23	19.92
9.Finland	32,608	32,684	21,920	21,739	1.82	3.15	9.15	64.21
10.France	30,330	30,282	19,161	19,698	1.85	1.72	9.33	8.41
11.Croatia	12,941	12,926	7,053	7,027	1.12	1.06	4.79	4.87
12.Hungary	8,485	8,440	4,454	4,464	1.99	1.39	11.68	6.33
13.Ireland	38,908	38,398	22,280	23,172	1.29	1.88	5.67	12.60
14.Lithuania	9,343	9,212	5,861	5,958	2.05	1.49	11.62	6.60
15.Latvia	8,020	7,989	6,270	6,343	3.56	1.57	28.32	7.09
16.Malta	20,518	20,362	15,362	16,013	2.91	1.89	20.30	10.22
17.Poland	9,202	9,170	6,116	6,155	3.91	2.51	38.34	24.92
18.Portugal	20,391	20,422	14,963	14,825	1.96	1.47	8.60	6.27
19.Romania	5,513	5,493	3,200	3,209	2.85	1.25	31.36	5.67
20.Sweden	28,299	27,653	16,751	17,134	2.36	2.35	18.84	26.85
21.Slovenia	21,922	21,899	12,708	12,928	1.87	1.37	10.04	6.58
22.Slovakia	10,550	10,592	6,365	6,384	5.82	1.99	92.65	10.97

Source: Microdata provided by Eurostat (HBS, 2010; SILC, 2011), and own elaboration.

References

[1] P.D. Agostini, B. Capéau, A. Decoster, F. Figari, J. Kneeshaw, C. Leventi, K. Manios, A. Paulus, H. Sutherland, and T. Vanheukelom, *Euromod extension to indirect taxation: Final report*, EUROMOD Technical Note Series EMTN-3.0, EUROMOD, 2017.

[2] O. Baser, *Modelling transformed health care cost with unknown heteroskedasticity*, App Econ Res Bull 01 (2007), pp. 1–6.

[3] A. Basu, *Extended generalized linear models: simultaneous estimation of flexible link and variance functions*, Stata Journal 5(4) (2005), pp. 501–516.

[4] A. Basu and P.J. Rathouz, *Estimating marginal and incremental effects on health outcomes using flexible link and variance function models*, Biostat 6(1) (2005), pp. 93–109.

[5] D.K. Blough, C. Madden, and M.C. Hornbrook, *Modelling risk using generalized linear models*, Journal of Health Economics 18(2) (1999), pp. 153–171.

[6] A. Decoster, J. Loughrey, C. O'Donoghue, and D. Verwerft, *Microsimulation of indirect taxes*, International Journal of Microsimulation 4(2) (2011), pp. 41–56.

[7] A. Decoster, R. Ochmann, and K. Spiritus, *Integrating VAT into EUROMOD. Documentation and results for Belgium*, EUROMOD Working Paper Series EM12/14, EUROMOD, 2014.

[8] G. Donatiello, M. D'Orazio, D. Frattarola, A. Rizzi, M. Scanu, and M. Spaziani, *Statistical Matching of Income and Consumption Expenditures*, International Journal of Economic Sciences III (3) (2014), pp. 50–65.

[9] M. D'Orazio, M.D. Zio, and M. Scanu, *Statistical Matching: Theory and Practice*, John Wiley Sons, 2006.

[10] N. Duan, W. Manning, C.N. Morris, and J.P. Newhouse, *A comparison of alternative models for the demand for medical care*, J. Bus. Econom. Statist. 1(2) (1983), pp. 115–126.

[11] W.M. Gorman, *Some Engel curves, in Essays in the Theory and Measurement of Consumer Behaviour: In Honour of Sir Richard Stone, A. Deaton*, Cambridge University Press, New York, 1981.

[12] W.H. Greene, *Econometric Analysis*, 5th Edition, Upper Saddle River, Prentice Hall, NJ, 2003.

[13] INE, *Household Budgetary Survey Base 2006*, Spain's National Office of Statistics, 2012–2016.

[14] INE, *Survey of Income and Living Conditions Base 2013*, Spain's National Office of Statistics, 2013–2017.

[15] A. Jones, *Models For Health Care*, HEDG Working paper 10/01, Health Econometrics and Data Group, 2010.

[16] A. Leulescu and M. Agafitei, *Statistical Matching: a model based approach for data integration*, Eurostat Methodologies and Working papers, 2013.

[17] R.J.A. Little and D.B. Rubin, *Statistical Analysis with missing data*, 2nd edition, Hoboken, NJ: Wiley, 2002.

[18] W.G. Manning, *The logged dependent variable, heteroskedasticity, and the retransformation problem*, Journal of Health Economics 17(3) (1998), pp. 283–295.

[19] W.G. Manning, A. Basu, and J. Mullahy, *Generalized modelling approaches to risk adjustment of skewed outcomes data*, Journal of Health Economics 24(3) (2005), pp. 465–488.

[20] W.G. Manning and J. Mullahy, *Estimating log models: to transform of not to transform?*, Journal of Health Economics 20(4) (2001), pp. 461–494.

[21] J. Mullahy, *Much ado about two: reconsidering retransformation and the two part model in health econometrics*, Journal of Health Economics 17 (1998), pp. 247–281.

[22] C. O'Donoghue, M. Baldini, and D. Mantovani, *Modelling the redistributive impact of indirect taxes in europe: an application of EUROMOD*, EUROMOD Working Paper Series EM7/01, EUROMOD, 2004.

[23] M. Savage, *Integrated modelling of the impact of direct and indirect taxes using complementary datasets*, The Economic and Social Review 48(2) (2017), pp. 171–205.