Documento de Trabajo - 2017/15

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

Julio López-Laborda
(Universidad de Zaragoza and FEDEA)

Carmen Marín-González
(FEDEA)

Jorge Onrubia
(Universidad Complutense de Madrid (ICEI), FEDEA and GEN)

fedea

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

Julio López-Laborda[a], Carmen Marín-González[b] and Jorge Onrubia[c]

Abstract: There are several ways to match SILC-HBS surveys, with the most common technique involving the estimation of Engel curves using Ordinary Least Squares in logs with HBS data to impute household expenditure in the income dataset (SILC). The estimation in logs has certain advantages, as 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. The presence of intrinsic heteroskedasticity in household expenditure therefore calls for another technique, as the smearing estimate produces a bias. Generalized Linear Models (GLMs) are presented as the best option.

JEL codes: C15, C51, C52.

Keywords: Statistical matching surveys, Engel curve, household expenditure, heteroskedasticity, Generalized Linear Models (GLMs).

Acknowledgments: The authors acknowledge the support from Spain's Ministry of the Economy – projects ECO2014-5916P (Carmen Marín-González) and ECO2016-76506-C4-3R (Julio López-Laborda and Jorge Onrubia)–, and the Regional Government of Aragón and the European Regional Development Fund (Public Economics Research Group, Julio López-Laborda). The authors would also like to thank Michael Savage. Any remaining errors are entirely our responsibility.

[a] Universidad de Zaragoza and FEDEA.

[b] FEDEA. Author for correspondence: cmarin@fedea.es.

[c] Universidad Complutense de Madrid (ICEI), FEDEA and GEN.

1. Introduction

Most countries collect data on household expenditure and household income in separate surveys, which are, respectively, the Household Budget Survey (HBS) and the Statistics on Income and Living Conditions (SILC). HBS provides information about household spending, while SILC reports on household income, the main direct taxes, and social contributions. A single database with microdata on income and expenditure is therefore essential for studying the household tax burden including direct and indirect taxes.

There are several ways to merge SILC and HBS. Decoster (2007) divides these techniques into two different groups: explicit and implicit methods. Explicit methods use estimations of Engel curves to impute household expenditure into the income dataset. The model is usually estimated by Ordinary Least Squares (OLS) (O'Donoghue et al., 2004; Decoster et al., 2013, 2014; Savage and Callan, 2015,). The implicit methods or *hot deck* method is a non-parametric approach. The procedure finds records in the donor file and matches them with records in the recipient file, based on a distance function (D'Orazio et al., 2006; Donatiello et al., 2014).

The matching technique most widely used in the literature is the estimation of Engel curves in HBS to impute household expenditure into SILC data using regression coefficients. The dependent variable is usually the logarithm of household expenditure, and the independent variables are the logarithm of income and a set of specific household dummy variables. This model is estimated by OLS. The estimation in logs has certain advantages, as it can deal with skewness in data and reduce heteroskedasticity. However, we are interested in the estimation in levels (euros), and not in logarithms. This problem is flagged in the literature as the retransformation problem,[1] and it is solved using a smearing estimate (Duan et al., 1983). Manning (1998) has shown that the correction of the smearing estimate only works for homoscedastic or heteroskedastic errors due to categorical variables. In the presence of heteroskedasticity, the smearing estimate produces a bias. In our case, the HBS expenditure estimation process records heteroskedasticity of an intrinsic nature.

Generalized Linear Models (GLMs) have been reported in recent literature for estimating health expenditure (Blough et al., 1999; Manning and Mullahy, 2001; Manning et al., 2005; Jones, 2010). They have been proposed as an alternative to OLS regression in logs. However, Başer and Yuce (2010) and Manning and Mullahy (2001) have noted that GLMs are less accurate when kurtosis increases.

The aim of this paper is to select the most suitable model for estimating HBS expenditure in order to impute these results in SILC in a statistical-matching procedure. The selection process is not so obvious. We develop the estimation process in HBS, as this database has the original expenditure variable. The expenditure is then imputed for all SILC households.

This article presents seven different estimation models involving an OLS regression of the expenditure in levels, the regression of the expenditure in logs, and different GLM alternatives. We propose two estimation techniques for each model: a simple regression and a simple regression adding a Chi-squared error term. This latter method increases the variance, skewness and kurtosis of the prediction. However, it reduces the estimation's accuracy.

The paper is structured as follows. Section 2 explains the different estimation techniques and reports the estimation results. Section 3 presents the in-sample and out-sample predictions. The analysis in sections 2 and 3 leads us to choose the GLM 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. Section 4 presents the HBS

[1] This result is a consequence of Jensen´s inequality: "the convex transformation of a mean is less than or equal to the mean applied after the convex transformation" ($\rho[E(x)] \leq E[\rho(x)]$).

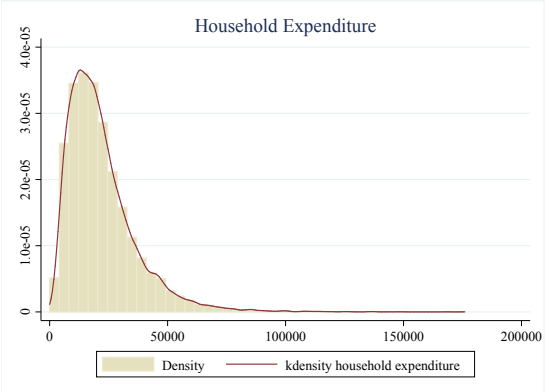
estimation and SILC imputation results per expenditure centiles. Finally, Section 5 contains the main conclusions.

2. Modelling the HBS household expenditure estimation

This section presents different estimation models and two estimation procedures for assessing each method's pros and cons. The models considered involve the OLS regression of the expenditure in levels, the regression of the expenditure in logs, and different GLM alternatives. The two estimation procedures involve a simple OLS regression and a simple regression adding a Chi-squared error term with zero mean and a standard deviation, whereby the new variable has the same standard deviation as the original one. The statistical figures computed are the mean, standard deviation, skewness, kurtosis, the Root Mean Squared Error (RMSE) and the R^2. The estimation results, which include a 1,000 times bootstrap estimation, are presented in tables 2.1 to 2.4. The database used is the Spanish HBS for 2013.

The distribution of the household expenditure (the dependent variable in the estimation process) presents heavily right-skewed data (Figure 2.1.). The household expenditure median is less than the mean (€17,754.8 vs. €20,979.4). The skewness statistic is 2.05 (compared to 0 for symmetric data), and the kurtosis is 10.95 (compared to 3 for normal data).

Figure 2.1. Household expenditure distribution



Source: author's own work based on HBS data.

We start with the simple OLS regression of household monetary expenditure[2] (E_i) over disposable income (y_i) and the household-specific dummy variables (x_i), which are as follows: population density, household members, household type, householder labour status, and household tenure. These variables in the matching process need to meet certain criteria: they must exist in both the HBS and SILC surveys; they must have the same definition in both surveys; they must contribute significantly to explain the total expenditure, and they must have similar distributions in both surveys. The variable choice process is explained in Appendix 1.

The OLS model is as follows:

$$E_i = \alpha + \gamma_1 y_i + \gamma_2 y_i^2 + \gamma_3 y_i^3 + x_i'\beta + \varepsilon_i \quad (1)$$

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

The OLS regression records a R^2 of 45.59, and the errors are highly heteroskedastic (the Breusch-Pagan test rejects the hypothesis of homoscedasticity with a Chi-squared statistic of 1148). One possible way to reduce the heteroskedasticity and treat the high skewness is to take the expenditure in logarithms. We then run an OLS regression of the logarithm of expenditure over the logarithm of income and the dummy specific variables:

$$\ln(E_i) = \alpha + \gamma_1 \ln(y_i) + \gamma_2 \ln(y_i)^2 + \gamma_3 \ln(y_i)^3 + x_i'\beta + \varepsilon_i \quad (2)$$

Using logs, heteroskedasticity is presented in a smaller dimension (Chi-squared statistic is 128). However, to include the results in the matching process the estimation results must be presented in the scale of interest (euros). As there is still heteroskedasticity, the smearing estimate presents a bias in the retransformation process (See Table 2.1). This heteroskedasticity can be caused by misspecification due to the exclusion of certain important variables. As we have already mentioned, the explanatory variables must meet certain conditions if they are to be used in the matching process. Another source of heteroskedasticity could arise from survey measurement error. These causes of heteroskedasticity cannot be corrected because they are a limitation inherent to the matching process. Nonetheless, the smearing estimate cannot be implemented in the matching process, as SILC has insufficient information for this, which is computed using the regression residuals. The SILC expenditure imputation consists of a deterministic equation using the coefficients from the HBS expenditure regression.

GLMs can be used to avoid the bias in the estimation caused by the retransformation problem. Cameron and Trivedi (2009) have defined the GLMs estimators as a subset of maximum likelihood estimators. They are generalizations of Non-Linear-Squares that are ideally suited to a nonlinear regression model with homoscedastic errors or with some kind of heteroskedasticity. GLMs provide a number of estimation alternatives depending on the link function and the distributional family specified. The link function refers to the relation between the dependent variable and the explanatory variables. The conditional mean function $\mathrm{E}[E_i/y_i, x_i]$ in equation 3 is a function of μ_i independent variables:

$$\mathrm{E}[E_i/y_i, x_i] = f(\mu_i) \quad (3)$$

If a log link is specified, then the previous equation is transformed into:

$$\mathrm{E}[E_i/y_i, x_i] = \exp(\mu_i) \quad (4)$$

Distributional family refers to the relationship between the conditional variance and the conditional mean:

$$Var[E_i/y_i, x_i] \propto (\mathrm{E}[E_i/y_i, x_i])^v \quad (5)$$

The distributional family is determined by the value of v. The most common options are as follows: Gaussian (constant variance; $v = 0$); Poisson (the variance is linearly related to the mean; $v = 1$); Gamma (the variance is explained by the square of the mean; $v = 2$); Inverse Gaussian (the variance is explained by the cube of the mean; $v = 3$). The alternatives studied include log and square root link functions, and Gaussian, Poisson and Gamma distributional families.

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. Basu and Rathouz (2005) have extended the standard GLM using Box-Cox transformation for the link function and two parameters for the variance to choose the correct distributional family:

$$\mathrm{E}[E_i/y_i, x_i] = \frac{\mu_i^{\lambda}-1}{\lambda} \;\; if\ \lambda \neq 0 \tag{6}$$

$$\mathrm{E}[E_i/y_i, x_i] = \log(\mu_i) \;\; if\ \lambda = 0 \tag{7}$$

$$Var[E_i/y_i, x_i] = \theta_1 \mathrm{E}[E_i/y_i, x_i]^{\theta_2} \tag{8}$$

This new model is called Extended Estimating Equations (EEE), which avoids the problems of misspecification due to the wrong choice of a family distribution or link function. The above equations reveal that the GLM options previously defined are special cases of EEE. In our case, the estimation of the equation using EEE presents a value of $\lambda = 0.59$ and a value of $\theta_2 = 1.49$. This means that the link function should be a square root, and the distributional family is between Poisson and Gamma.

Tables 2.1 to 2.4 show the comparison of all the models, including the two estimation procedures: the simple regression and a regression adding a Chi-squared error term. We call the first one the simple procedure, and the second one the Chi-squared procedure.

The Chi-squared procedure presents a similar bias to the simple regression. This procedure's main advantage is that its moments are closer to the real data. By definition, the standard deviation of the Chi-squared estimation is similar to the original expenditure one. The skewness and kurtosis of the Chi-squared procedure are greater than the simple procedure ones, so they are nearer to the real expenditure data. This holds for all the estimation models except for the GLM log gamma, which presents similar skewness and lower kurtosis after adding the Chi-squared error term. The drawback of the Chi-squared procedure is the loss of precision. The RMSE of the Chi-squared procedure is nearly 40% higher (from around 10,600 to around 14,800). 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. Section 4 explains that the Chi-squared procedure is, on average, more accurate for households with lower and higher expenditures.

The OLS regression of the logarithm of expenditure has an important bias in both estimation procedures (Table 2.1.). As we have already explained, this is due to heteroskedasticity in the data. The skewness and kurtosis for the OLS regression and the GLM log normal of the simple procedure record very low values (closer to the normal distribution). These results improve in the case of the Chi-squared procedure. In the simple procedure, the GLM log normal and GLM log Poisson have the lowest RMSE. In the Chi-squared procedure, however, the OLS over the logarithm of expenditure and the GLM log gamma have the lowest RMSE. This latter model is the preferred one

Table 2.1. Comparison of model specifications (Mean)

Model	Mean		Bias	
	Regression	*Bootstrap 1000*	*Regression*	*Bootstrap 1000*
OLS. Simple Procedure	20,979	20,978	0	-2
Log OLS. Simple Procedure	21,245	21,242	265	262
GLM sqrt Gamma. Simple Procedure	21,002	21,002	23	22
GLM log Gamma. Simple Procedure	21,025	21,024	46	44
GLM log Poisson. Simple Procedure	20,979	20,978	0	-2
GLM log Normal. Simple Procedure	20,946	20,944	-34	-35
EEE. Simple Procedure	20,981	20,979	1	0
OLS. Chi^2 Procedure	20,933	20,982	-47	2
Log OLS. Chi^2 Procedure	21,240	21,246	261	266
GLM sqrt Gamma. Chi^2 Procedure	20,973	21,006	-6	26
GLM log Gamma. Chi^2 Procedure	20,997	21,028	18	49
GLM log Poisson. Chi^2 Procedure	20,950	20,982	-29	2
GLM log Normal. Chi^2 Procedure	20,917	20,948	-63	-31
EEE. Chi^2 Procedure	20,952	20,983	-28	4

Source: author's own work based on INE data (2014).

Table 2.2. Comparison of model specifications (Standard Deviation)

Model	Standard Deviation		R^2	
	Regression	*Bootstrap 1000*	*Regression*	*Bootstrap 1000*
OLS. Simple Procedure	9,815	9,822	0.459	0.459
Log OLS. Simple Procedure	10,498	10,499	0.525	0.525
GLM sqrt Gamma. Simple Procedure	10,008	10,011	0.477	0.477
GLM log Gamma. Simple Procedure	10,230	10,233	0.498	0.499
GLM log Poisson. Simple Procedure	9,873	9,881	0.464	0.465
GLM log Normal. Simple Procedure	9,931	9,943	0.470	0.471
EEE. Simple Procedure	9,883	9,880	0.465	0.465
OLS. Chi^2 Procedure	14,491	14,487	1.000	1.000
Log OLS. Chi^2 Procedure	14,301	14,487	0.974	1.000
GLM sqrt Gamma. Chi^2 Procedure	14,436	14,487	0.992	1.000
GLM log Gamma. Chi^2 Procedure	14,444	14,487	0.994	1.000
GLM log Poisson. Chi^2 Procedure	14,439	14,487	0.993	1.000
GLM log Normal. Chi^2 Procedure	14,439	14,487	0.993	1.000
EEE. Chi^2 Procedure	14,433	14,487	0.992	1.000

Source: author's own work based on INE data (2014).

Table 2.3. Comparison of model specifications (Skewness, Kurtosis)

Model	Skewness		Kurtosis	
	Regression	*Bootstrap 1000*	*Regression*	*Bootstrap 1000*
OLS. Simple Procedure	0.80	0.80	3.85	3.89
Log OLS. Simple Procedure	1.14	1.14	5.64	5.64
GLM sqrt Gamma. Simple Procedure	1.23	1.22	6.73	6.65
GLM log Gamma. Simple Procedure	1.45	1.44	8.53	8.45
GLM log Poisson. Simple Procedure	1.12	1.12	5.48	5.45
GLM log Normal. Simple Procedure	0.95	0.97	4.39	4.50
EEE. Simple Procedure	1.14	1.13	6.04	5.94
OLS. Chi^2 Procedure	1.34	1.37	6.24	6.65
Log OLS. Chi^2 Procedure	1.30	1.36	6.04	6.40
GLM sqrt Gamma. Chi^2 Procedure	1.44	1.47	6.82	7.07
GLM log Gamma. Chi^2 Procedure	1.49	1.51	7.12	7.35
GLM log Poisson. Chi^2 Procedure	1.43	1.46	6.70	6.92
GLM log Normal. Chi^2 Procedure	1.37	1.40	6.42	6.65
EEE. Chi^2 Procedure	1.43	1.46	6.77	7.03

Source: author's own work based on INE data (2014).

Table 2.4. Comparison of model specifications (RMSE)

Model	RMSE		Model Ranking	
	Regression	*Bootstrap 1000*	*Regression*	*Bootstrap 1000*
OLS. Simple Procedure	10,673	10,656	3	3
Log OLS. Simple Procedure	10,703	10,688	5	5
GLM sqrt Gamma. Simple Procedure	10,711	10,697	6	6
GLM log Gamma. Simple Procedure	10,719	10,701	7	7
GLM log Poisson. Simple Procedure	10,649	10,626	2	2
GLM log Normal. Simple Procedure	10,632	10,608	1	1
EEE. Simple Procedure	10,695	10,677	4	4
OLS. Chi^2 Procedure	15,118	15,077	7	7
Log OLS. Chi^2 Procedure	14,581	14,635	1	1
GLM sqrt Gamma. Chi^2 Procedure	14,834	14,981	4	4
GLM log Gamma. Chi^2 Procedure	14,698	14,833	2	2
GLM log Poisson. Chi^2 Procedure	14,882	15,018	5	5
GLM log Normal. Chi^2 Procedure	14,832	14,963	3	3
EEE. Chi^2 Procedure	14,904	15,054	6	6

Source: author's own work based on INE data (2014).

3. The in-sample and out-sample estimation

The aim of this analysis is to determine the most accurate model for the matching between HBS and SILC. This section describes a tenfold cross-validation process used to test the accuracy of out-sample forecasts. Firstly, the sample is randomly split using a uniform distribution into 10 subsamples, of which a single subsample is retained as validation data for testing the model, and the remaining nine subsamples are used as training data. We estimate the model using the training data (90% of the sample), and predict the

household expenditure for the whole sample. The estimated expenditure using the training data is called the in-sample prediction. The estimated expenditure using the validation data is called the out-sample prediction. Another subsample is then retained as the validation data for testing the model; the training data are the remaining subsamples. We repeated this dynamic with a different subsample in each case until completing the whole data sample (10 times). As a result, the out-sample estimation is the same size as the whole sample, and the in-sample estimation is repeated nine times for each household, so we take the average value.

Tables 3.1 to 3.4 show the results of the in-sample and out-sample 1,000 bootstrap regressions. The statistical figures computed are as follows: the mean, standard deviation, skewness, kurtosis, the RMSE and the R^2.

Table 3.1. Comparison of model specifications (Mean)

Model	Mean		Bias	
	In sample	*Out sample*	*In sample*	*Out sample*
OLS. Simple Procedure	20,978	20,974	-2	-5
Log OLS. Simple Procedure	21,243	21,242	264	263
GLM sqrt Gamma. Simple Procedure	21,002	21,001	22	22
GLM log Gamma. Simple Procedure	21,024	21,023	45	44
GLM log Poisson. Simple Procedure	20,978	20,977	-2	-3
GLM log Normal. Simple Procedure	20,944	20,943	-35	-37
EEE. Simple Procedure	20,979	20,978	0	-1
OLS. Chi^2 Procedure	20,977	20,973	-2	-6
Log OLS. Chi^2 Procedure	21,242	21,241	263	262
GLM sqrt Gamma. Chi^2 Procedure	21,001	21,000	22	21
GLM log Gamma. Chi^2 Procedure	21,023	21,022	44	43
GLM log Poisson. Chi^2 Procedure	20,977	20,976	-2	-4
GLM log Normal. Chi^2 Procedure	20,943	20,942	-36	-38
EEE. Chi^2 Procedure	20,978	20,977	-1	-2

Source: author's own work based on INE data (2014).

Table 3.2. Comparison of model specifications (Standard Deviation)

Model	Standard Deviation		R^2	
	In sample	*Out sample*	*In sample*	*Out sample*
OLS. Simple Procedure	9,822	9,850	0.459	0.462
Log OLS. Simple Procedure	10,499	10,506	0.525	0.526
GLM sqrt Gamma. Simple Procedure	10,011	10,013	0.477	0.477
GLM log Gamma. Simple Procedure	10,232	10,241	0.499	0.499
GLM log Poisson. Simple Procedure	9,881	9,890	0.465	0.466
GLM log Normal. Simple Procedure	9,944	9,950	0.471	0.472
EEE. Simple Procedure	9,879	9,883	0.465	0.465
OLS. Chi^2 Procedure	14,478	14,477	0.998	0.998
Log OLS. Chi^2 Procedure	14,479	14,478	0.998	0.998
GLM sqrt Gamma. Chi^2 Procedure	14,478	14,478	0.998	0.998
GLM log Gamma. Chi^2 Procedure	14,479	14,478	0.998	0.998
GLM log Poisson. Chi^2 Procedure	14,478	14,478	0.998	0.998
GLM log Normal. Chi^2 Procedure	14,478	14,478	0.998	0.998
EEE. Chi^2 Procedure	14,478	14,478	0.998	0.998

Source: author's own work based on INE data (2014).

Table 3.3. Comparison of model specification (Skewness, Kurtosis)

Model	Skewness		Kurtosis	
	In sample	*Out sample*	*In sample*	*Out sample*
OLS. Simple Procedure	0.80	0.65	3.89	9.33
Log OLS. Simple Procedure	1.14	1.15	5.63	5.77
GLM sqrt Gamma. Simple Procedure	1.22	1.22	6.64	6.69
GLM log Gamma. Simple Procedure	1.44	1.46	8.42	8.80
GLM log Poisson. Simple Procedure	1.12	1.14	5.44	5.74
GLM log Normal. Simple Procedure	0.98	0.99	4.51	4.65
EEE. Simple Procedure	1.13	1.13	5.93	6.04
OLS. Chi^2 Procedure	1.37	1.31	6.67	8.26
Log OLS. Chi^2 Procedure	1.36	1.36	6.42	6.48
GLM sqrt Gamma. Chi^2 Procedure	1.47	1.47	7.10	7.14
GLM log Gamma. Chi^2 Procedure	1.51	1.52	7.37	7.49
GLM log Poisson. Chi^2 Procedure	1.46	1.47	6.95	7.03
GLM log Normal. Chi^2 Procedure	1.40	1.41	6.68	6.72
EEE. Chi^2 Procedure	1.46	1.47	7.05	7.10

Source: author's own work based on INE data (2014).

Table 3.4. Comparison of model specifications (RMSE)

Model	RMSE		Model ranking	
	In sample	*Out sample*	*In sample*	*Out sample*
OLS. Simple Procedure	10,653	10,713	3	5
Log OLS. Simple Procedure	10,685	10,712	5	4
GLM sqrt Gamma. Simple Procedure	10,695	10,715	6	6
GLM log Gamma. Simple Procedure	10,698	10,730	7	7
GLM log Poisson. Simple Procedure	10,623	10,663	2	2
GLM log Normal. Simple Procedure	10,604	10,651	1	1
EEE. Simple Procedure	10,674	10,700	4	3
OLS. Chi^2 Procedure	15,062	15,091	7	7
Log OLS. Chi^2 Procedure	14,620	14,642	1	1
GLM sqrt Gamma. Chi^2 Procedure	14,966	14,986	4	4
GLM log Gamma. Chi^2 Procedure	14,818	14,842	2	2
GLM log Poisson. Chi^2 Procedure	15,002	15,031	5	5
GLM log Normal. Chi^2 Procedure	14,947	14,982	3	3
EEE. Chi^2 Procedure	15,039	15,062	6	6

Source: author's own work based on INE data (2014).

The out-sample estimation performs well with all the models, as it has a similar bias, skewness and kurtosis to the in-sample. The exception is the OLS model, which records greater kurtosis in the out-sample prediction. The standard deviation and the RMSE are slightly greater in the out-sample prediction, and the model ranking remains unchanged.

The out-sample estimation results are similar to the ones reported in the previous section. In the simple procedure, the GLM log gamma has the highest skewness and kurtosis, in line with the real expenditure data. The remaining models have a high correction of these moments with the Chi-squared procedure. The models with the lowest RMSE in the simple

procedure are the GLM log Normal and the GLM log Poisson. In the Chi-squared procedure, however, the models with the lowest RMSE are the OLS over the logarithm of expenditure and the GLM log gamma.

As previously explained, we preferred the Chi-squared procedure. The model with the lowest RMSE under the Chi-squared procedure is the OLS over the logarithm of expenditure. However, this model is rejected because of the high bias. This led us to choose the GLM log gamma as the best model for estimating household expenditure and including the results in the matching process.

4. SILC imputation results

This section presents the HBS estimation and the subsequent SILC imputation of the chosen model using the GLM log gamma. Table 4.1 shows the HBS original expenditure data moments compared to the imputed SILC expenditure. The mean and standard deviation of the SILC imputed expenditure with the Chi-squared procedure are very similar to the original expenditure ones. However, skewness and kurtosis are lower in the imputed expenditure.

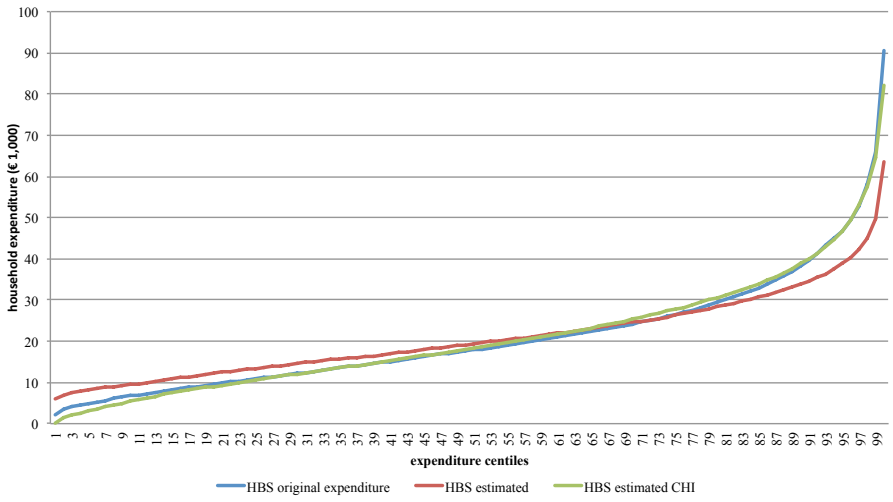
Table 4.1. HBS real expenditure vs. GLM log gamma SILC imputation

	Mean	Standard Deviation	Skewness	Kurtosis
HBS Real Expenditure	20,979	14,490	2.05	10.95
SILC Imputation. Simple Procedure	20,964	10,302	1.35	6.75
SILC Imputation. Chi^2 Procedure	20,921	14,572	1.55	7.36

Source: author's own work based on INE data (2014, 2015).

Figure 4.1 shows the average centiles of the original expenditure and the GLM log gamma estimation via the simple procedure (the variable is called "HBS estimated") and the Chi-squared procedure (the variable is called "HBS estimated CHI"). The simple procedure overestimates the original expenditure for lower household spending and underestimates it for higher spending. The introduction of the Chi-squared error improves the fit. This explains why we have preferred to use the Chi-squared procedure despite the greater RMSE observed in the previous section.

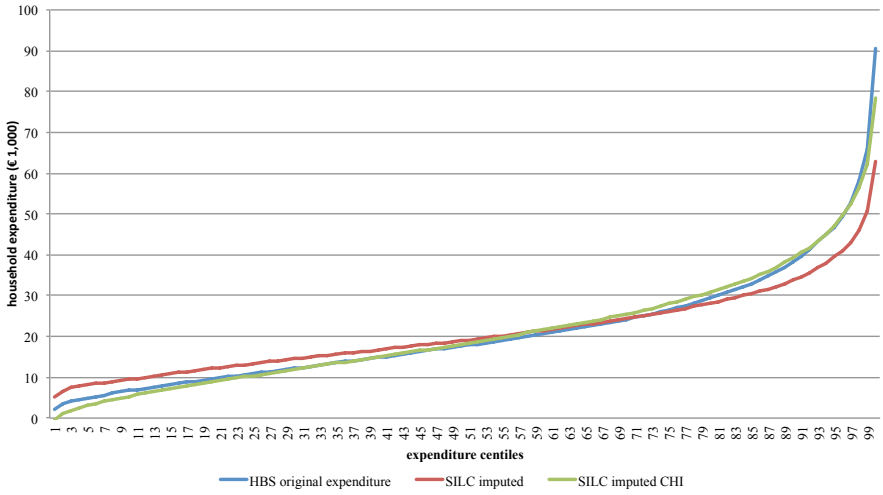
Figure 4.1. Average Household Expenditure Centiles. HBS Estimation



Source: author's own work based on INE data (2014).

Figure 4.2 shows the average centiles of the original expenditure and the GLM log gamma SILC imputation using the simple procedure (called "SILC imputed") and the SILC imputation using the Chi-squared procedure (called "SILC imputed CHI"). As in the HBS estimation, the SILC imputed overestimates the original expenditure for lower household spending and underestimates it for higher household spending. The "SILC imputed CHI" has a similar shape to the original HBS expenditure.

Figure 4.2. Average Household Expenditure Centiles. SILC Imputation



Source: author's own work based on INE data (2015).

5. Concluding Remarks

There are several ways to match the SILC-HBS surveys. 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 (SILC). The estimation in logs has certain advantages, as can it deal with the skewness in data and reduce heteroskedasticity. However, the model needs to be corrected with a smearing estimate to retransform the results into levels. The presence of intrinsic heteroskedasticity in household expenditure requires another estimation technique, as the smearing estimate produces a bias. The Generalized Linear Model (GLM) log gamma under the Chi-squared procedure is selected as the best option. The paper shows that the estimated HBS and imputed SILC results by expenditure centiles are, on average, very close to the original HBS expenditure when the GLM log gamma technique is applied.

Appendix 1. The choice of explanatory variables

Two types of explanatory variables were used in the estimation process explained in section 2, HBS and SILC disposable income and specific household characteristics. Three criteria must be met to obtain a consistent and statistically valid merge, being as follows: 1) The variables chosen must have the same definition in both surveys; 2) They must significantly help to explain the dependent variable - household expenditure; 3) All these variables must have similar distributions in both surveys. Even if these conditions are not met, disposable income must remain in the process after some adjustments.

Figure A.1 shows HBS and SILC disposable income grouped by percentiles. HBS underestimates the real value of disposable income as reflected by SILC disposable income (data collected from the administrative files). SILC average disposable income was €26,154 in 2013, while for HBS it was €21,800 (see Table A.1). The differences in the income

(Figure A.2.) can be explained by a different definition of disposable income or by a different method of data collection. Firstly, we adjusted SILC disposable income to present a similar definition to the one for HBS disposable income. This new variable, called *adjusted SILC disposable income,* has a similar distribution to the original SILC disposable income (Figure A.3.), so the differences are due to different data collection methods. SILC disposable income was then rescaled to present a similar mean and variance to HBS disposable income (Figure A.4.)[3]. This new variable is called *rescaled SILC disposable income,* and it was used solely for statistical matching.

Table A.1. HBS and SILC Disposable Income (€)

Variable	Obs.	Mean	Standard Deviation	Min.	Max.
HBS disposable income	22,057	21,800	15,277	0	340,032
SILC disposable income	11,965	26,154	19,928	-27,082	309,796
Adjusted SILC disposable income	11,965	26,183	20,008	0	360,426
Rescaled SILC disposable income	11,965	21,800	15,277	1,808	277,010

Source: INE (2014, 2015) and own elaboration.

The variables representing specific household characteristics must meet the three conditions outlined above. We have chosen a number of variables with similar definitions in both surveys, which significantly help to explain household expenditure. These variables are as follows: population density, household members, household type, householder labour status, household tenure, and householder education level. These variables as a whole explain the expenditure with a R^2 of 0.47.

The *Hellinger Distance* (HD)[4] was computed to determine whether the specific characteristic variables have a similar distribution in both surveys:

$$HD(V,V') = \sqrt{\frac{1}{2}\sum_{i=1}^{K}\left(\sqrt{\frac{n_{Oi}}{N_O}} - \sqrt{\frac{n_{Pi}}{N_P}}\right)^2} \quad \text{(A1)}$$

The variable V is the donor dataset (in our case, HBS), V' is the recipient (in our case, SILC), k is the number of categories of the variable, $\frac{n_{Oi}}{N_O}$ is the relative frequency in the donor dataset, and $\frac{n_{Pi}}{N_P}$ is the relative frequency in the recipient data. It is generally accepted that an HD of over 5% should raise concerns about the similarities in the distributions.

The HD for each variable is presented in Table A.2. The HD is above 5% for three variables: householder education level, household type, and householder labour status. The householder education level has an extremely high HD (10.5%), and was dropped from the regression estimation. However, the remaining variables were kept because the HD is near the limit considered (5%).

[3] As in Decoster (2014) for the Belgium case.

[4] See Eurostat (2013) and Donatiello et al. (2014).

Table A.2. Sample Descriptive variables: HBS and SILC

		HBS Frequency	SILC Frequency	Hellinger Distance
Population Density	1. High	51.71%	52.13%	1.66%
	2. Medium	23.70%	21.91%	
	3. Low	24.59%	25.97%	
Household members	1. One	24.22%	24.63%	0.48%
	2. Two	30.45%	30.61%	
	3. Three	21.25%	21.07%	
	4. Four	17.99%	17.81%	
	5. Five	4.63%	4.49%	
	6. Six or more	1.46%	1.39%	
Household Type	1. One person aged >65	10.29%	10.39%	6.40%
	2. One person aged 30-65	12.46%	13.01%	
	3. One person aged <30	1.47%	1.23%	
	4. Couple with no dependent children, one aged >65	10.21%	13.87%	
	5. Couple with no dependent children, both aged <65	12.72%	14.60%	
	6. Couple with one dependent child	10.97%	11.26%	
	7. Couple with two dependent children	11.53%	11.00%	
	8. Couple with three dependent children	2.34%	2.24%	
	9. Single-parent with at least one dependent child	2.78%	3.28%	
	10. Other households or without information	25.23%	19.12%	
Household Tenure	1. Owner without a mortgage	46.74%	49.66%	4.50%
	2. Owner with a mortgage	30.54%	28.39%	
	3. Renter at market price	15.23%	12.44%	
	4. Renter at a reduced price	1.46%	2.49%	
	5. Free or nearly free assignment	6.03%	7.02%	
Householder Labour Status	1. Employed	52.81%	54.12%	5.60%
	2. Unemployed	10.82%	11.04%	
	3. Pensioner	27.85%	24.37%	
	4. Student	0.18%	0.45%	
	5. Homemaker	4.65%	5.59%	
	6. Permanent job disability	1.35%	3.24%	
	7. Other	2.33%	1.19%	
Householder Education Level	1. No education or Primary	17.76%	28.00%	10.53%
	2. Below Upper Secondary	33.43%	22.78%	
	3. Upper Secondary	18.69%	17.55%	
	4. Tertiary	30.12%	31.67%	

Source: INE (2014, 2015) and author's own work.

Figure A.1. HBS vs. original SILC

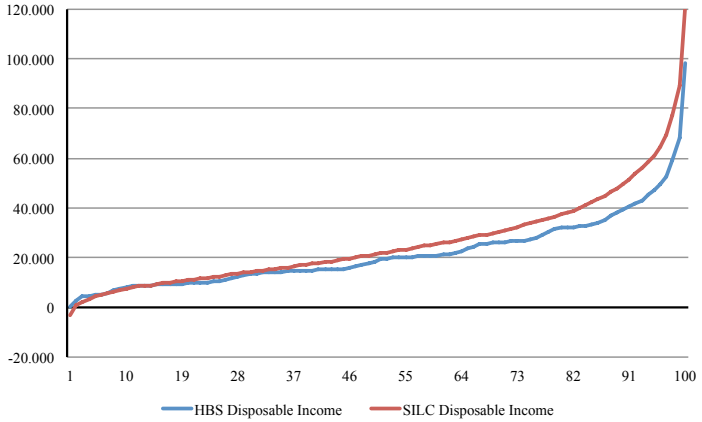
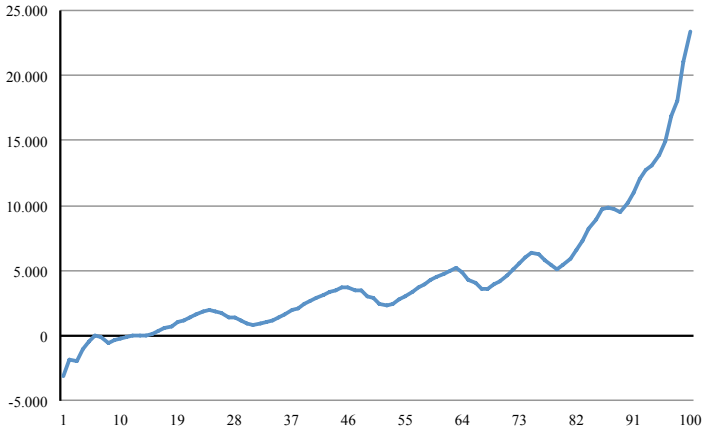


Figure A.2. Differences between HBS and original SILC



Source: INE (2014, 2015) author's own work.

Figure A.3. HBS vs. adjusted SILC

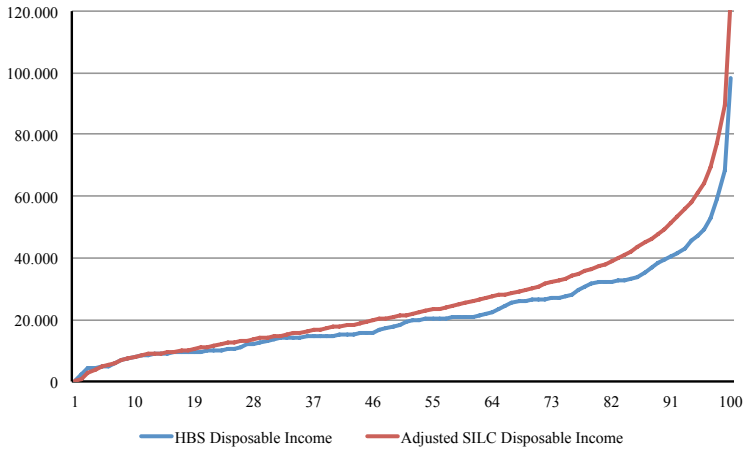
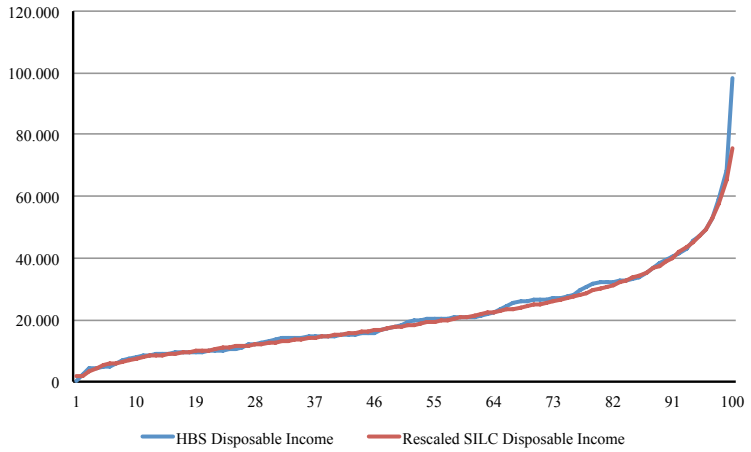


Figure A.4. HBS vs. rescaled SILC



Source: INE (2014, 2015) author's own work.

References

Başer, O., and H. Yuce (2010). "PCV156 Modeling transformed health care costs with unknown heteroskedasticity". *Value in Health*, *13*(3), A179.

Basu, A., and P. J. Rathouz, (2005). "Estimating marginal and incremental effects on health outcomes using flexible link and variance function models", *Biostatistics*, 6: 93–109.

Blough, D. K., C. W., Madden, and M. C. Hornbrook (1999). "Modelling risk using generalized linear models", *Journal of Health Economics*, 18: 153-71.

Cameron A. C., and P. K. Trivedi (2009). *Microeconometrics using Stata*. College Station, Texas: Stata Press.

Decoster A., B. D., Rock, K. D., Swerdt, J., Loughrey, C. O'Donoghue, and D. Verwerft (2007). "Techniques to impute expenditures into an income data set EUROMOD AIM-AP deliverable 3.4", Institute for Social and Economic Research.

Decoster, A., R. Ochmann, and K. Spiritus (2013). "Integrating VAT into EUROMOD. Documentation and results for Germany", *EUROMOD Working Paper Series*, EM20/13. https://www.iser.essex.ac.uk/research/publications/working-papers/euromod/em20-13.pdf

Decoster A., R. Ochmann, and K. Spiritus (2014). "Integrating VAT into EUROMOD. Documentation and results for Belgium", *EUROMOD Working Paper Series*, EM12/14. https://www.iser.essex.ac.uk/research/publications/working-papers/euromod/em12-14.pdf

D'Orazio M., M. Di Zio, and M. Scanu (2006). *Statistical Matching, Theory and Practice*. New York: Wiley.

Donatiello G., D. Frattarola, A. Rizzi, and M. Spaziani (2014), "Statistical Matching of IT-SILC and HBS: Some critical issues", *International workshop and conference on comparative EU statistics on income and living conditions*, Lisbon, 15-17 October 2014.

Duan, N, W. G. Manning, C. N. Morris, and J. P. Newhouse, (1983). "A Comparison of Alternative Models for the Demand for Medical Care", *Journal of Business & Economic Statistics*, 1(2):115-126.

Eurostat (2013). "Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation", *Eurostat Methodologies and Working Papers.* Luxembourg: Publications Office of the European Union. http://ec.europa.eu/eurostat/documents/3888793/5857145/KS-RA-13-007-EN.PDF/37d4ffcc-e9fc-42bc-8d4f-fc89c65ff6b1

INE (Instituto Nacional de Estadística – Spain's National Office of Statistics) (2014). "Household Budgetary Survey. Base 2006 -Year 2013".

INE (Instituto Nacional de Estadística) (2015). "Survey of Income and Living Conditions. Base 2013-Year 2014".

Jones A. (2010). "Models for Health care", *Health Econometrics and Data Group Working Paper*, 10/01. University of York. https://www.york.ac.uk/media/economics/documents/herc/wp/10_01.pdf

Jones A., N. Rice, T. Bago d'Uva, and S. Balia (2013). *Applied Health* Economics, 2nd ed. Abingdon, Oxon: Routledge.

Manning W. G. (1998). "The logged dependent variable, heteroscedasticity, and the retransformation problem", *Journal of Health Economics*, 17: 283-295.

Manning, W. G., and J. Mullahy (2001). "Estimating log models: to transform of not to transform?", *Journal of Heath Economics*, 20: 461-94.

Manning W. G., A. Basu, and J. Mullahy (2005). "Generalized modelling approaches to risk adjustment of skewed outcomes data", *Journal of Health Economics,* 24: 465-488.

O'Donoghue, C., M. Baldini and D. Mantovani (2004). "Modelling the redistributive impact of indirect taxes in Europe: an application of EUROMOD"*, EUROMOD Working Paper Series*, EM7/01.

Savage M. and T. Callan (2015), "Modelling the impact of direct and indirect taxes using complementary datasets", *Discussion Paper* 8897, Bonn: Institute for the Study of Labor (IZA).
http://ftp.iza.org/dp8897.pdf

ÚLTIMOS DOCUMENTOS DE TRABAJO

2017-15: "Estimating Engel curves: A new way to improve the SILC-HBS matching process", **Julio López-Laborda, Carmen Marín-González y Jorge Onrubia.**

2017-14: "New Approaches to the Study of Long Term Non-Employment Duration in Italy, Germany and Spain", **B. Contini, J. Ignacio Garcia Perez, T. Pusch y R. Quaranta**.

2017-13: "Structural Scenario Analysis and Stress Testing with Vector Autoregressions", **Juan Antolín-Díıaz y Juan F. Rubio-Ramírez**.

2017-12: "The effect of changing the number of elective hospital admissions on the levels of emergency provision", **Sergi Jimenez-Martin, Catia Nicodemo y Stuart Redding**.

2017-11: "Relevance of clinical judgement and risk stratification in the success of integrated care for multimorbid patients", **Myriam Soto-Gordoa, Esteban de Manuel, Ane Fullaondo, Marisa Merino, Arantzazu Arrospide, Juan Ignacio Igartua y Javier Mar**.

2017-10: "Moral Hazard versus Liquidity and the Optimal Timing of Unemployment Benefits", **Rodolfo G. Campos, J. Ignacio García-Pérez y Iliana Reggio**.

2017-09: "Un análisis de modelos para financiar la educación terciaria: descripción y evaluación de impacto", **Brindusa Anghel, Antonio Cabrales, Maia Gu̇ell y Analía Viola**.

2017-08: "Great Recession and Disability Insurance in Spain", **Sergi Jiménez-Martín, Arnau Juanmarti Mestres y Judit Vall Castelló**.

2017-07: "Narrative Sign Restrictions for SVARs", **Juan Antolín-Díaz y Juan F. Rubio-Ramírez**.

2017-06: "Faster estimation of discrete time duration models with unobserved heterogeneity using hshaz2", **David Troncoso Ponce**.

2017-05: "Heterogeneous Household Finances and the Effect of Fiscal Policy", **Javier Andrés,José E.Boscá, Javier Ferri y Cristina Fuentes-Albero**.

2017-04: "Statistical Discrimination and the Efficiency of Quotas", **J. Ignacio Conde-Ruiz, Juan-José Ganuza y Paola Profeta**.

2017-03: "Cargos por Azar", **Emilio Albi**.

2017-02: "Should pensions be redistributive? The impact of Spanish reforms on the system's sustainability and adequacy", **Concepció Patxot, Meritxell Solé y Guadalupe Souto**.

2017-01: "El Modelo de Perfilado Estadístico: una herramienta eficiente para caracterizar a los demandantes de empleo", **Yolanda F. Rebollo-Sanz**.

2016-10: "Family Job Search and Wealth: The Added Worker Effect Revisited", **J. Ignacio García-Pérez y Sílvio Rendon**.

2016-09: "Evolución del Gasto Público por Funciones durante la crisis (2007-2014): España vs UE", **José Ignacio Conde-Ruiz, Manuel Díaz , Carmen Marín y Juan F. Rubio-Ramírez**.

2016-08: "Thinking of Incentivizing Care? The Effect of Demand Subsidies on Informal Caregiving and Intergenerational Transfers", **Joan Costa-Font, Sergi Jiménez-Martín y Cristina Vilaplana-Prieto**.

2016-07: "The Pruned State-Space System for Non-Linear DSGE Models: Theory and Empirical Applications", **Martin M. Andreasen, Jesús Fernández-Villaverde y Juan F. Rubio-Ramírez**.

2016-06: "The effects of non-adherence on health care utilisation: panel data evidence on uncontrolled diabetes", **Joan Gil, Antonio Sicras-Mainar y Eugenio Zucchelli**.

2016-05: "Does Long-Term Care Subsidisation Reduce Unnecessary Hospitalisations?", **Joan Costa-Font, Sergi Jiménez-Martín y Cristina Vilaplana-Prieto**

2016-04: ""Cultural Persistence" of Health Capital: Evidence from European Migrants", **Joan Costa-Font y Azusa Sato**.

2016-03: "Like Mother, Like Father? Gender Assortative Transmission Of Child Overweight", **Joan Costa-Font y Mireia Jofre-Bonet**.

2016-02: "Health Capacity to Work at Older Ages: Evidence from Spain", **Pilar García-Gómez, Sergi Jimenez Martin y Judit Vall Castello**.

2016-01: "Monte Carlo evidence on the estimation of AR(1) panel data sample selection models", **Sergi Jiménez-Martín y José María Labeaga**.

2015-13: "On the Treatment of Foreigners and Foreign-Owned Firms in Cost–Benefit Analysis", **Per-Olov Johansson y Ginés de Rus**.

2015-12: "Evaluating Options for Shifting Tax Burden to Top Income Earners", **Jorge Onrubia, Fidel Picos y María del Carmen Rodado**.

2015-11: "Differences in Job De-Routinization in OECD countries: Evidence from PIAAC", **Sara De La Rica y Lucas Gortazar**.

2015-10: "Bad times, slimmer children?", **Cristina Belles-Obrero, Sergi Jimenez-Martín y Judit Vall-Castello**.

2015-09: "The Unintended Effects of Increasing the Legal Working Age on Family Behaviour", **Cristina Belles-Obrero, Sergi Jimenez-Martín y Judit Vall-Castello**.

2015-08: "Capital Humano y Productividad", **Ángel de la Fuente**.

2015-07: "The effect of changes in the statutory minimum working age on educational, labor and health outcomes", **Sergi Jiménez-Martín, Judit Vall y Elena del Rey**.

2015-06: "The Effects of Employment Uncertainty, Unemployment Insurance, and Wealth Shocks on the Retirement Behavior of Older Americans", **Hugo Benítez-Silva, J. Ignacio García-Pérez y Sergi Jiménez-Martín**.