An Integrated Database to Measure Living Standards

Federico Perali, Martina Menon and Elena Dalla Chiara

University of Verona, Italy

34th IARIW General Conference
Session 3: Well-being I

Discussed by Roberto Zelli - Sapienza University of Rome

Main goal of the paper

- Generate an **integrated micro database to measure living standards** in Italy that combines information available from different data sources, using **propensity score matching**: IILS (Italian Integrated Living Standard) dataset
- Use the integrated database to measure the multiple dimensions of well-being and use it in case-control studies (epidemiological studies)

Why an integrated living standard database? I

- The GDP and beyond Commission and the Stiglitz-Sen-Fitoussi Commission raised awareness about the need to review and update the current system of statistics in order to address new challenges and to support policy-making.
- In particular, the measure of well-being is the result of a multidimensional evaluation process no longer associated with a single indicator.
- The social statistical infrastructure (in Italy and, generally, in OECD countries) is organised around specific surveys covering many relevant aspects of the users' demand: income, consumption, health, education, labour market, social participation (the ability to develop personal relationships, to enjoy a clean environment and to invest in activities creating social capital).
- However, no single survey can cover all the requested aspects of well-being.

Why an integrated living standard database? II

- **Statistical matching** (a.k.a. data fusion, data merging or synthetic matching) is a model-based approach for providing joint statistical information based on variables and indicators collected through two or more sources.
- In this paper **a micro-approach** was followed. The micro approach refers to the creation of a complete micro-data file where data on all the variables is available for every unit.
- Similar experience of data fusion in the United States for the Levy Institute Measure of Economic Well-Being (LIMEW).

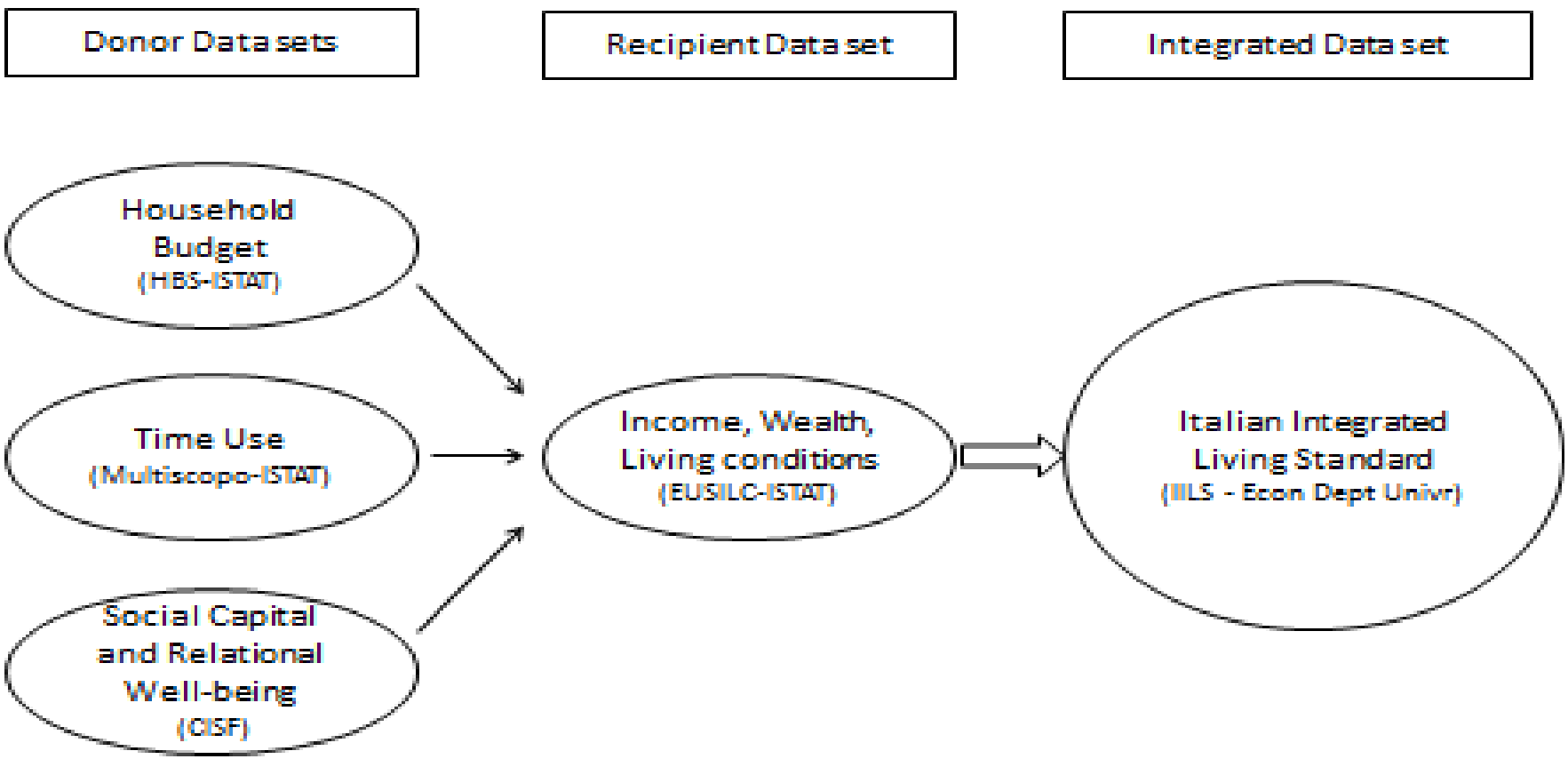
Matched data contains information collected in four surveys - reference year 2009: I

1. EU Statistics on Income and Living Conditions (EU-SILC) by Istat - recipient survey
2. Household Budget Survey (HBS) by Istat - donor survey
3. Time Use Survey (TUS) by Istat - donor survey
4. Household Conditions and Social Capital Survey by International Center of Family Studies (CISF) - donor survey

Recipient and donors I

- EU-SILC chosen as recipient (so the integrated dataset has the same size of EU-SILC sample)
- HBS adds household expenditure (detailed in 9 categories, 5 food and 4 non food categories) to EU-SILC
- TUS adds time employed in daily activities for each hh member (which variables exactly?)
- CISF adds information on social capital and relational well-being (which variables exactly?)

Data sets used to create an integrated database (Figure A1)



Statistical matching: an overview I

- An essential feature of statistical matching (SM) is that, although the units in the concerned data sets should come from the same population, they are usually not overlapping. This is the basic difference compared with record linkage (record linkage deals with identical units).
- Statistical matching, or synthetic linkage, deals with **similar units**, i.e. SM identifies and links records from different sources that correspond to similar units.
- **Similarity between units is based on a set of common variables X**.
- In practice, matching procedures can be regarded as an imputation problem of the target variables from a donor to a recipient survey (literature stems from causal inference literature).

Statistical matching: an overview II

- Y, Z are collected through two different samples drawn from the same population; X variables are collected in both samples and they are correlated with both Y and Z.
- In this case, the relation between these common variables with the specific variables observed only in one of the data sets - the **donor data set** - will be explored and used to impute to the units of the other data set - the **recipient data set** - the variables not directly observed. Thus a synthetic dataset is generated with complete information on X, Y and Z

Statistical matching exemplification

Sample A (Donor)	Sample B (Recipient)	Synthetic dataset
X, Y		
	X, Z	$X, \widehat{Y}, Z$

Assumptions I

Basic conditions:

- Both samples refer to the same target population
- The common variables X are defined in the same way

Assumptions:

- *Conditional independence assumption (CIA)*: measures of association between Y and Z conditional on X cannot be estimated and they are usually assumed to be 0.
- The CIA implies that the variable Y to be imputed is independent of the treatment (being in one data set or in the other) conditional on the selected set of covariates X.

  When this condition holds, matching algorithms will produce accurate estimates that reflect the true joint distribution of variables that were collected in multiple sources. It will give the same results as a perfect linkage procedure.

Assumptions II

- *Common support*: observations with the same covariate values should have the same probability of being both treated (that is to belong to the recipient data set) and untreated (that is, to belong to the donor data set).

  This assumption ensures that there is sufficient overlap between the characteristics of the two data set.

Propensity score I

- The key concern is that of similarity.
- How can we find individuals who are similar on all observable characteristics in order to match treated (recipient) and non-treated (donor) households?
- With a single measure, we can readily compute a measure of distance between a treated unit and each candidate match. With multiple measures defining similarity, how are we to balance similarity along each of those dimensions?
- To deal with this dimensionality problem, Rosenbaum and Rubin (1983) suggest to use so-called balancing scores. They show that if potential outcomes are independent of treatment conditional on covariates X, they are also independent of treatment conditional on a balancing score b(X).
- The method of propensity score matching (PSM) allows this matching problem to be reduced to a single dimension: that of the propensity score.

Propensity score II

- That score is defined as the probability that a unit in the full sample receives the treatment, given a set of observed variables.
- The propensity score is defined as the estimated conditional probability of a unit to belong to one of the two data set, given X.
- in practice, both data sets are extended with an additional variable taking value 1 for all the records in file A and value 0 for all the records in data set B. Putting both files together a **logit or probit model is estimated**, taking as dependent variable the added one, and as independent variables the common variables X (and including the regression constant).
- Thus, rather than matching on all values of the variables, individual units can be compared on the basis of their propensity scores alone.

Steps adopted to implement the propensity score based statistical matching: I

1. **Choose a set of common target variables X having a significant relationship with variables of interest Y.**
2. **Compare the distribution of X in the recipient and control (donor) group (the two sample surveys should represent the same population)**
3. **Estimate the propensity score using the selected explanatory variables.**
4. **Validate the propensity score procedure by:**
   - 4.1 computing balancing tests
   - 4.2 checking the overlap and region of common support between the two groups
5. **Choose the matching algorithm and match each observation of the recipient group to the observation in the control group using the propensity score value.**
6. **Assess the statistical matching quality by:**

Steps adopted to implement the propensity score based statistical matching: II

6.1 inspecting distributions of the added information in the two databases

6.2 comparing the ratio of mean in the two groups by the set of X covariates .

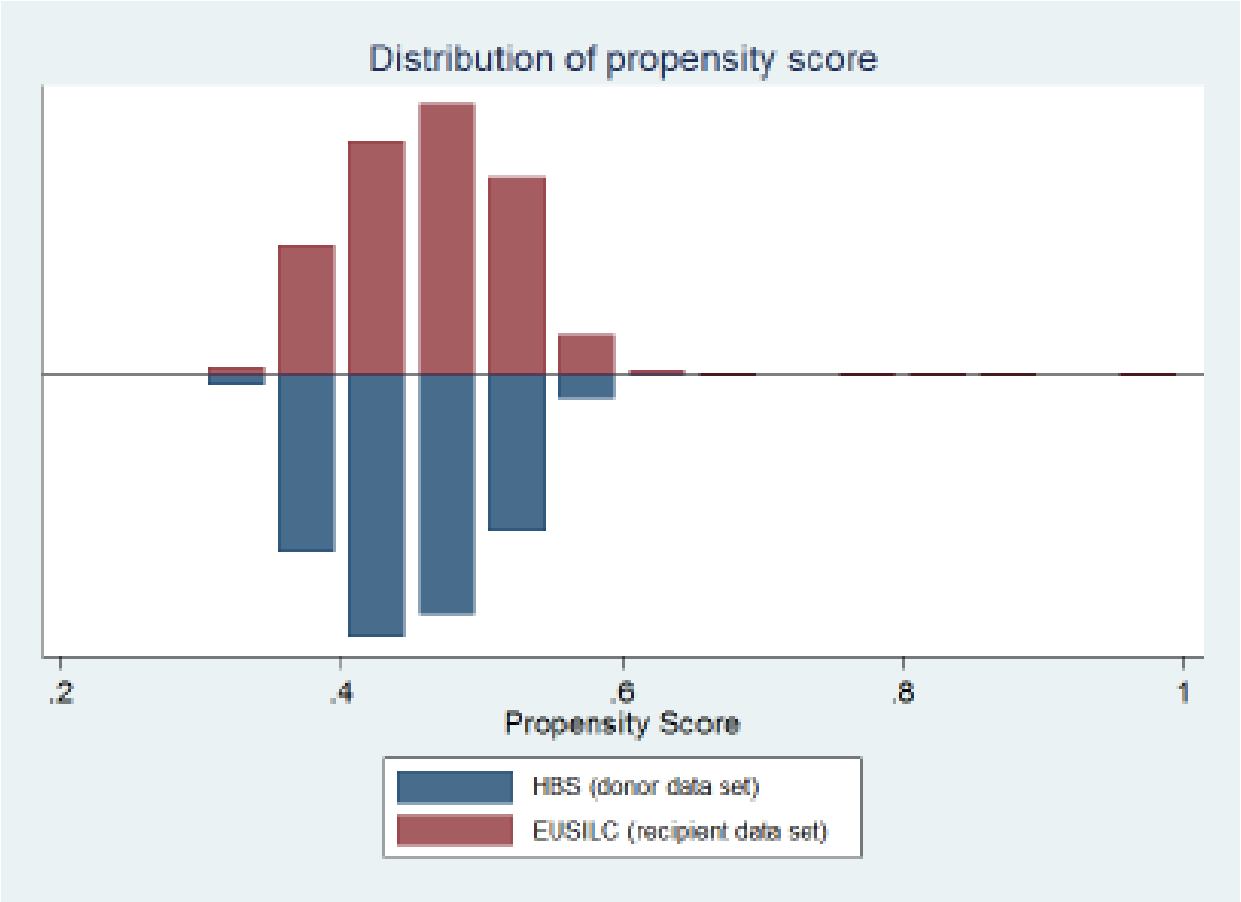
7 **Assess the economic matching quality using Engel curve analysis.**

Comparison of the Distributions of Common Variables (Tab.A1)

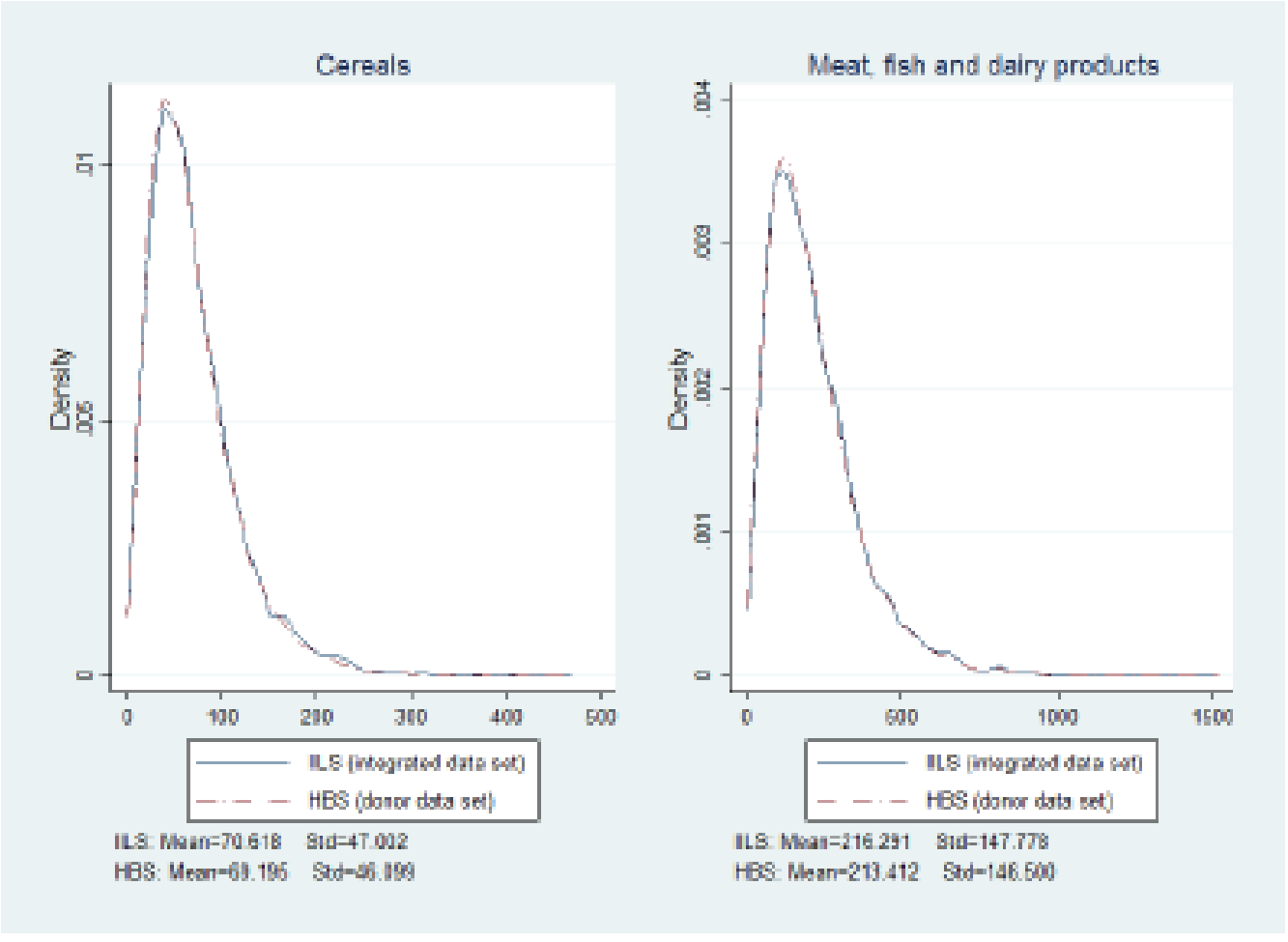
	EUSILC	HBS	Absolute difference	Cramer's V *
Single-parent				0.025
No	91.41	92.78	1.37	
Yes	8.59	7.22	1.37	
Owner occupancy				0.009
No	25.50	24.72	0.78	
Yes	74.50	75.28	0.78	
Average family education				0.036
Primary	26.95	26.83	0.12	
Middle	24.28	27.24	2.96	
Middle-High	19.16	18.15	1.01	
High	23.16	21.48	1.68	
University	6.44	6.29	0.15	
Family income				0.025
1st quintile	19.51	20.41	0.90	
2nd quintile	19.48	20.44	0.96	
3rd quintile	20.17	19.85	0.32	
4th quintile	19.85	20.13	0.28	
5th quintile	20.99	19.17	1.82	

*Acceptance threshold of weak relationship is 0.15

Distribution of propensity score across recipient and donor data sets (Fig. A2)



Matching quality outcome: Expenditure by category in IILS and HBS (Fig.A3)



Matching quality outcome: Cereals - ratio of means of covariates (Tab. A6a)

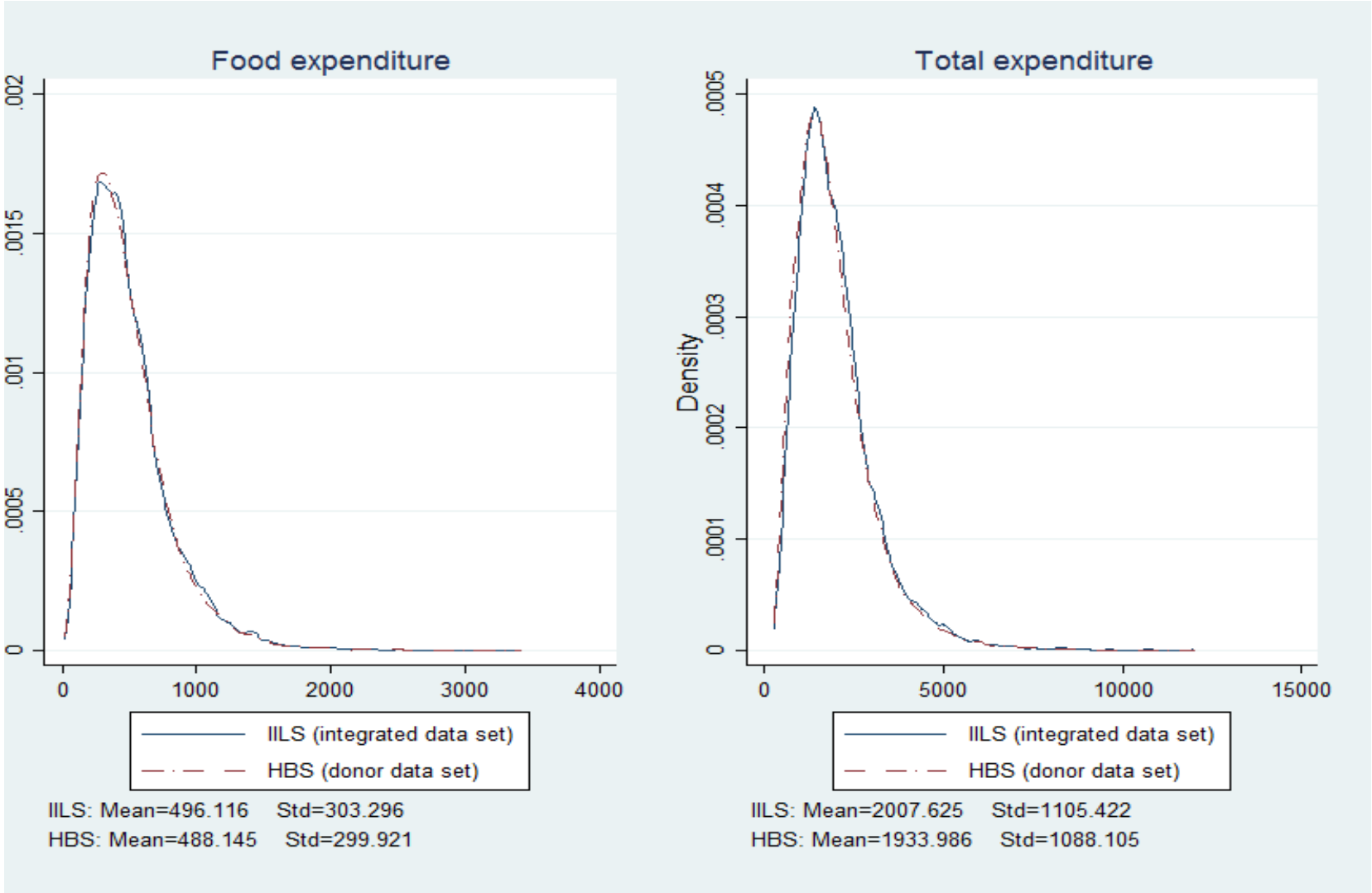
	HBS	IILS	Ratio *
Single-parent			
No	150.49	153.89	102.26
Yes	140.69	161.90	115.07
Owner occupancy			
No	143.90	156.06	108.45
Yes	151.71	154.07	101.55
Average family education			
Primary	101.78	148.86	146.25
Middle	153.10	153.05	99.96
Middle-High	182.39	158.95	87.15
High	173.84	159.28	91.63
University	163.87	154.30	94.16
Family income			
1st quintile	86.39	144.97	167.80
2nd quintile	124.63	149.16	119.68
3rd quintile	149.82	158.72	105.94
4th quintile	184.34	157.12	85.23
5th quintile	207.77	162.14	78.04

* The closer is the ratio to 100, the higher is the similarity of the extra information in the two samples

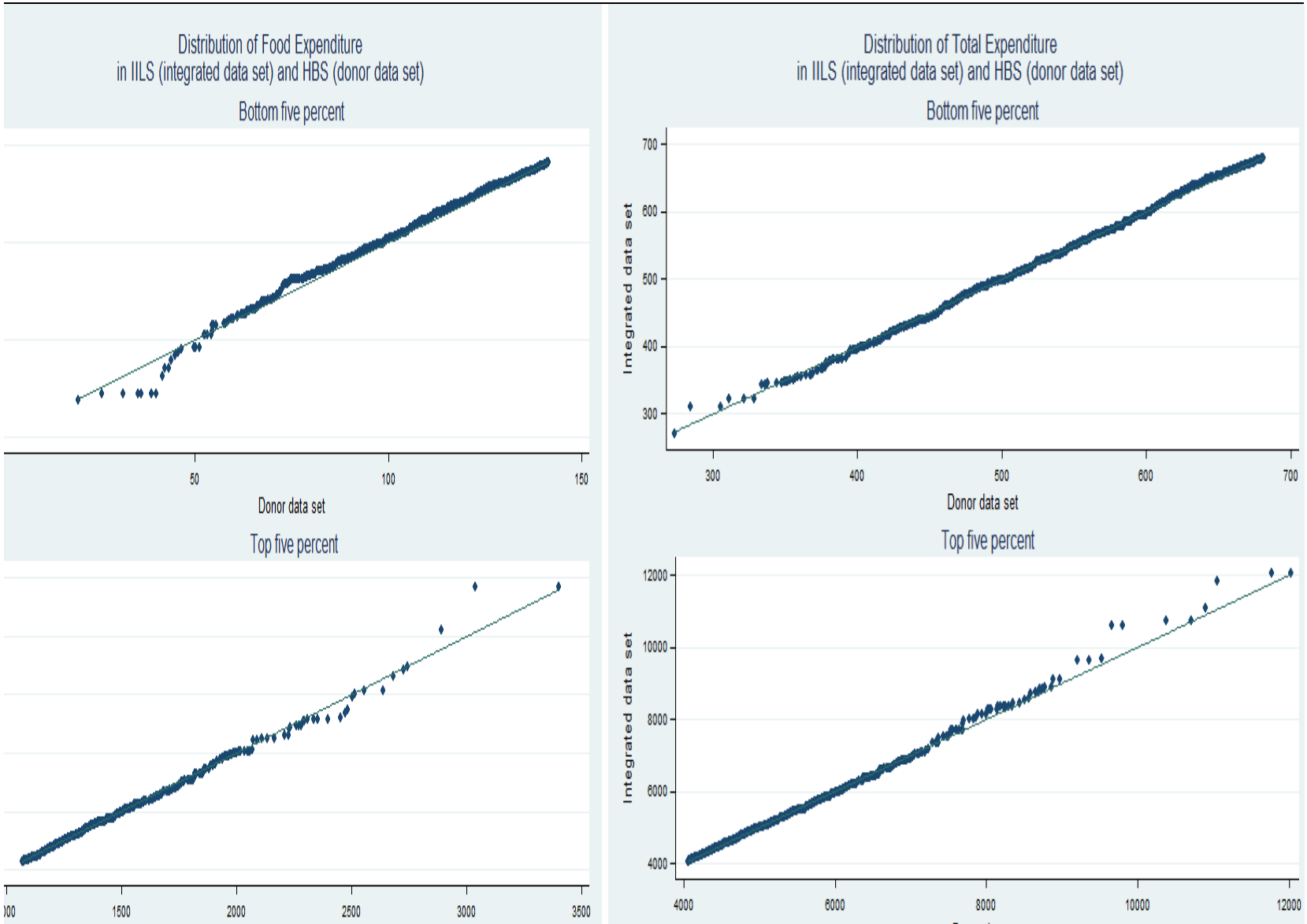
Matching quality I

- Matching quality outcome: relevant contribution of the paper is that it evaluates both the statistical AND **economic** robustness of the matching.
- To this end, robust economic tests based on the fundamental Engel relationship
  - Compare the distributions in the two databases
  - Zoom on bottom and top five percent of the distributions
  - Test statistical differences using: dispersion indexes; inequality and poverty indexes
  - Estimate the Engel relationship linking the food share and the logarithm of total expenditure
  - Evaluate the influence of extreme values comparing estimated coefficients with OLS and quantile regression

Distribution of Food expenditure and Total expenditure in integrated and donor data sets (Fig.B1)



Q-Q plot: focus on the tails (Fig.B2-B3)



Dispersion, inequality, poverty: statistical differences (Tab.B3)

Dispersion indexes for Food Expenditure

	p90/p10	Gini coefficient
IILS (integrated data set)	4.7902	0.3189
HBS (donor data set)	4.7309	0.3206
DIFFERENCE	-0.0592	0.0017
std. err.	*0.0339*	*0.0022*
*p-value**	*0.0802*	*0.4474*

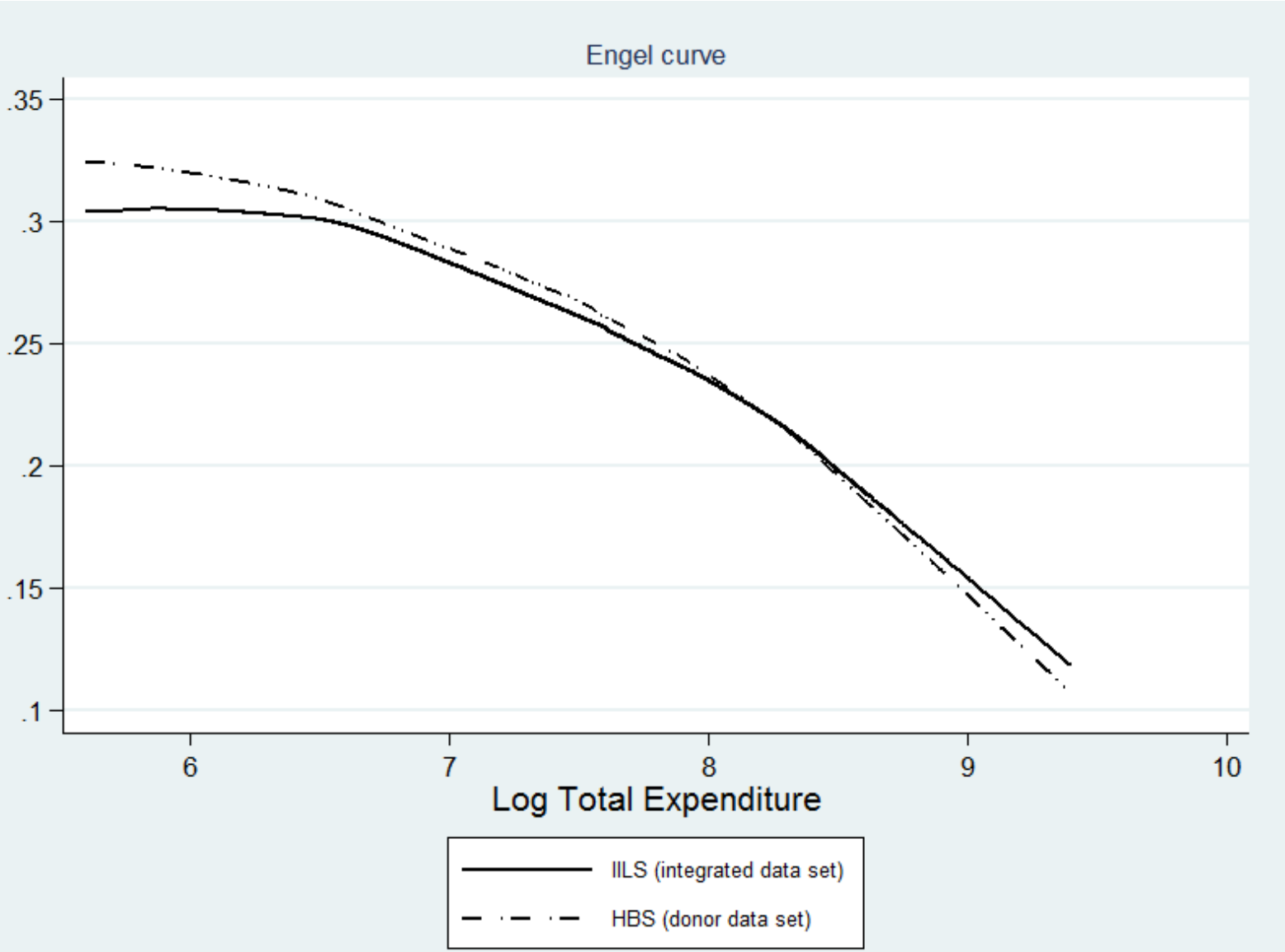
* If the p-value is greater than the chosen significance level (by convention equal to 0.05 or 0.01) the null hypothesis cannot be reject.

Inequality and poverty indexes for Total Expenditure

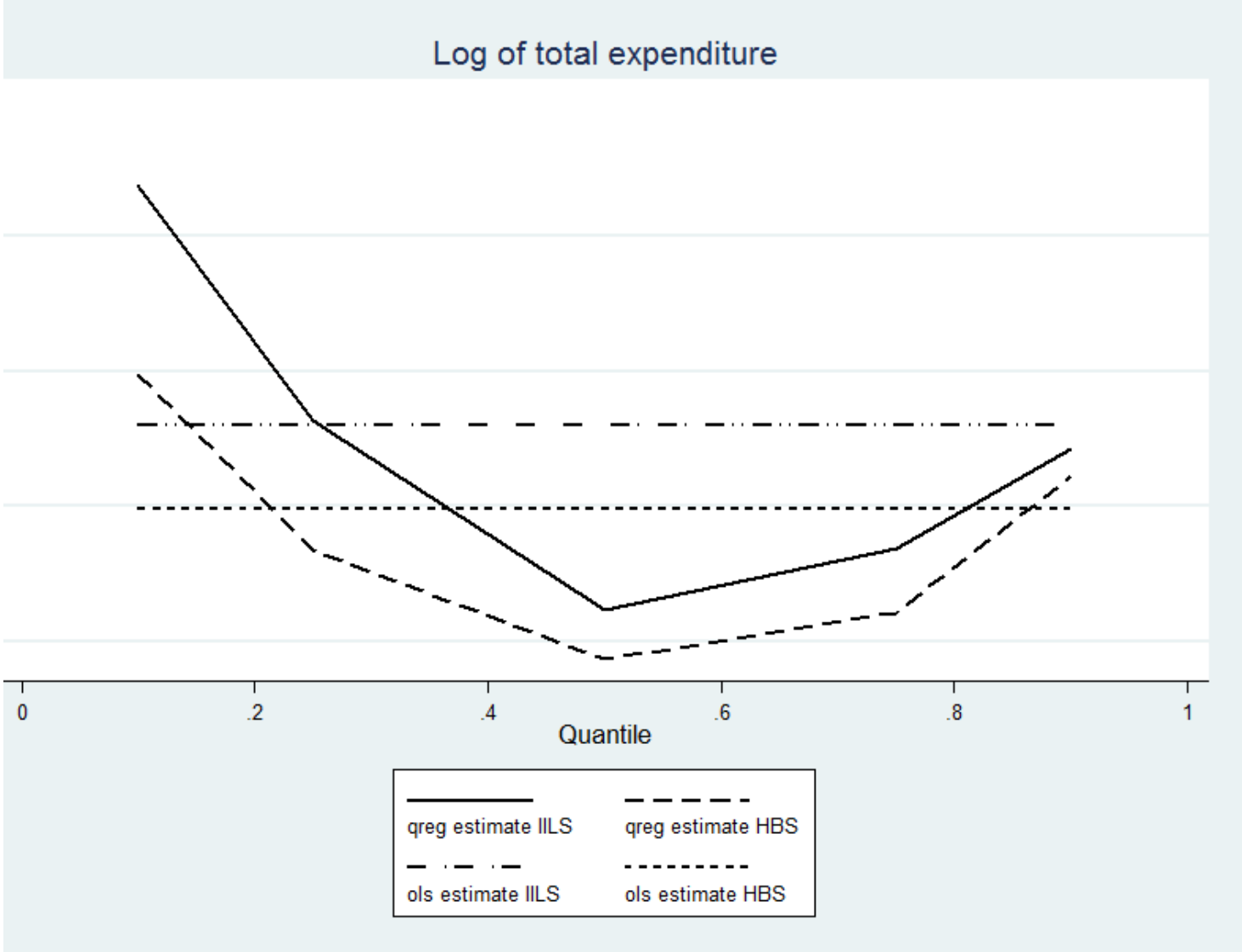
	p90/p10	*Gini coefficient*	*FGTα poverty index***		
			α=0	*α=1*	*α=2*
IILS (integrated data set)	*3.6310*	*0.2766*	*0.1568*	*0.0334*	*0.0101*
HBS (donor data set)	*3.7593*	*0.2816*	*0.1658*	*0.0354*	*0.0106*
DIFFERENCE	*0.1283*	*0.0050*	*0.0090*	*0.0020*	*0.0005*
std. err.	*0.0218*	*0.0021*	*0.0035*	*0.0012*	*0.0006*
*p-value**	*0.0000*	*0.0175*	*0.0103*	*0.1004*	*0.3832*

** If the p-value is greater than the chosen significance level (by convention equal to 0.05 or 0.01) the null hypothesis cannot be reject.*

Engel curve in integrated and donor data sets (Fig.B5)



Estimated coefficient of total expenditure with OLS regression and quantile regression (Fig.B6b)



Remarks, comments, questions I

- Very important and valuable attempt to integrate EU-SILC with other information coming from different sources.
- The potential benefits of this approach lie in the possibility to enhance the complementary use and analysis of existing data sources, without further increasing costs and response burden.
- Origins of statistical matching can be traced back to the mid 1960s and developed in the 1970s. See Ruggles N. and Ruggles R. (1974), A strategy for merging and matching microdatasets, Annals of Economic and Social Measurement, I(3).
- General question: is this the direction for statistical offices? (apparently yes, see Eurostat (2013)). Integrated survey vs surveys specific to each relevant dimension
- Or at least, several aspects could be harmonised ex-ante: the choice of common variables between variables; consider matching jointly with other options for micro-integration (linking and use of administrative data), possibilities of auxiliary information.

Remarks, comments, questions II

- Matching concerns:
  - In case the conditional independence does not hold, and no additional information is available, the model will have identification problems and the artificial datasets produced may lead to incorrect inferences.
  - This assumption cannot be tested from the data sets. However.....
  - Optimal: the common variables X contain ALL the association shared by the target variables and thus fulfill the CIA
  - Therefore a strong explanatory power of X makes the CIA more plausible to hold.
  - Better to have more details on the selection of X (all possible common variables? which selection procedure? strong association?)
  - About the sample weights: can we ignore that the two samples come from two different complex surveys?

Remarks, comments, questions III

- Matching algorithm: Nearest-Neighbour (NN) with replacement is good practice. The number of units used in the comparison set is one (only one donor is used). By using a single closest match, one reduces the bias, but including more matched donors, the variance is reduced whereas bias increases if the addition al observations are inferior matches for the treated (recipient) observations. A partial solution is to use a pre-defined neighbourhood in terms of a radius around the $p(X)$ of the trated (recipient) observation and to exclude matches that lie outside this neighbourhood. In other words, one only uses the better matches. This is called **caliper matching**.
- Guess that the misalignment of the means of Y by family income is due to different definitions and different accuracy in measurement of this covariate in the two surveys.

Remarks, comments, questions IV

- Useful a comparison with other attempts to build integrated data sets in Italy, like:
  - Coli A. *et al.* (2006), "La costruzione di un Archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'Indagine Banca d'Italia sui bilanci delle famiglie italiane", Istat, Technical Report 12/2006.
  - Sisto A. (2006), "Propensity score matching: un'applicazione per la creazione di un database integrato ISTAT-Banca d'Italia", POLIS WP n.63.
  - Conti, P.L., Marella D., Neri A. (2015), "Statistical matching and uncertainty analysis in combining household income and expenditure data", Banca d'Italia, Temi di Discussione n. 1018.
  - Conti, P.L., Marella D., Scanu M. (2016), "Statistical Matching Analysis For Complex Survey Data With Applications", Journal Of The American Statistical Association, forthcoming.
  - Tedeschi S. and Pisano E. (2013), "Data Fusion Between Bank of Italy-SHIW and Istat-HBS", MPRA WP n. 51253.