Instructions for obtaining the data

European Community Household Panel (ECHP)

The European Community Household Panel (ECHP) is a confidential panel survey which contains information about individual statistical units. Access to anonymised microdata is available at Eurostat only for scientific purposes. ECHP includes a wide range of topics concerning living conditions, including detailed income information, financial situation in a wider sense, working life, housing situation, social relations, health and biographical information of the interviewed. The total duration of the ECHP was 8 years, running from 1994-2001 (8 waves).

Contact information: European Commission, Eurostat – Unit D4, BECH Building – Office A3/102

5, rue Alphonse Weicker, 2721 Luxembourg; estat-microdataaccess@ec.europa.eu

Fees: The price is € 250 for the full data set (1994-2001; 8 waves).

Procedures: Due to the confidential character of the ECHP microdata, direct access to the anonymised data is only provided by means of research contracts. Access is in principle restricted to universities, research institutes, national statistical institutes, central banks inside the EU, as well as to the European Central Bank. Please note that individuals cannot be granted direct access. For other kind of organisations inside the EU and any requests from any organisations outside the EU, approval for access needs to be requested first from the Committee of Statistical Confidentiality by written procedure, which takes about 6 months. More detailed information is available at Eurostat's website in the sub domain statistics (http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/echp).

Other information: The timespan between the moment of reception of an access request to the ECHP UDB (User Data Base) at Eurostat and the final delivery to the contractor is on average 10 weeks. This time span includes amongst others the consultation with the Member States, which has a duration of 6 weeks as specified in the Commission Regulation and some weeks for applying the necessary contractual administrative procedures at Eurostat's and the contractor's sides.

Medical Expenditure Survey (MEPS)

The Medical Expenditure Panel Survey (MEPS) is a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States, consisting of two major components: the Household Component and the Insurance Component. The MEPS began in 1996.

Contact information: AHRQ Publications Clearinghouse, P.O. Box 8547, Silver Spring, Maryland 20907, United States. Tel: 1-800-358-9295

Fees: The MEPS can be obtained free of charge.

Procedures: All public use data files can be downloaded from the MEPS website: http://www.meps.ahrq.gov/mepsweb/index.jsp. Alternatively, one can also obtain the data on a CD Rom by contacting AHRQ Publications Clearinghouse (see contact information).

Description of the code files and program files to replicate the tables and figure in the paper.

All analyses have been done using STATA version 10.1; on a windows operating system.

Monte Carlo simulations

To replicate the Monte Carlo simulations and the response surface estimates, one should

1. Run "monte carlo FOC and SObia.do". Each run of this file corresponds to one set of Monte Carlo simulations (i.e. 20000 replications) for a given distribution function (beta, lognormal or uniform) with a specific value for the parameters of the distribution function. The global variables "distribution", "alpha", "beta", "variance" and "distri" can be used to change the distribution function and the parameter values. All five global variables are defined in the first section "SETUP".
2. Each run of "monte carlo FOC and SObia.do" gives two output files. The first is a stata dataset called "_mc_NAME OF DISTRIBUTION FUNCTION WITH PARAMETERS_N.dta" and the second is a graph "_mc_NAME OF DISTRIBUTION FUNCTION WITH PARAMETERS_N.wmf". The latter gives the empirical density function of the distribution function, while the former gives all bootstrap samples (N=10.000 and S=20.000 in the paper) for a given distribution function. The summary statistics of these bootstrap files for the uniform and lognormal distributions have been reported in table 1 in the paper, columns Mean and St.Dev (load the file with all bootstrap files into STATA and type sum). The other columns (First-order and Cov) have been calculated in an excel-sheet ("final calculations.xls", sheet monte carlo, rows 1-22, columns A-Y).
3. In order to create the response surface estimates for the beta distribution in table 2, one should run the stata do-file "response surface.do". The latter program calls the bootstrap samples described above and reports (in a log file "responsesurface.log") the ouput of the OLS regressions. These outputs have been used to construct table 2.
4. In order to create table 3, we have used the excel sheet "final calculations.xls", sheet monte carlo, rows 24-39

Analysis based on ECHP

For the analyses using the ECHP, one should take the following steps.

1. Obtain a CD Rom with the raw data from EUROSTAT (see also "instructions for obtaining the data"). We have used the ECHP version of December 2003.
2. Run the file "ECHP_merging.do" on the raw data available on the CD Rom obtained from EUROSTAT. The *.dta files created in this do-file can be used in step 3.
3. Run the file "ECHP_workingfile.do" for each country separately. At the top of the file one should specify the file to be used, i.e. masterfile1_9.dta (for ECHP countries 1 to 9), masterfile10_57.dta (for countries 10 to 18), and masterfile.dta. At the bottom of the file, one should only save the relevant datasets (depending on the choice of the masterfile AND depending on the country code).

4. Run the file “group_data_probleml_ECHP_gini.do”. This file returns a log-file for each ECHP country containing all information needed to construct the columns “Gini” and “FOC” in table 4 (in the log file, these columns are named “gini” and “our”), and the descriptive statistics in table A.1 (in the log files reported on top of the table).
5. We have also applied the POVCAL software tool upon our data. To this means, one should
    a. Run “POVcal_ECHP.do” to obtain the files that will serve as input for the POVCAL software. The names of these files are echpx_y.dat where x refers to the country and y to the number of income groupings.
    b. Download and install the POVCAL software tool. This tool can be obtained from the Poverty Research website of the World Bank http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTPROGRAMS/EXTPOVRES/0,,menuPK:477905~pagePK:64168176~piPK:64168140~theSitePK:477894,00.html or from the PovcalNet website http://iresearch.worldbank.org/PovcalNet/jsp/index.jsp.
    c. POVCAL is now a very old program written in Fortran, but it can still be run through a DOS type interface in Windows XP or VISTA. However, it is important that the POVCAL executable file be stored in the same directory as the program. When you execute POVCAL you will be asked a series of questions:
        i. INPUT THE NAME OF YOUR DATA FILE: Type in any of the names of “.dat”, for example “echp1_5.dat” (i.e. first ECHP country with 5 income groups), and then hit Enter.
        ii. INPUT THE NUMBER OF SUBGROUPS: Type 1 and return
        iii. INPUT THE NUMBER OF RECORDS: For this example “5”
        iv. CHOOSE YOUR TYPE OF DATA SET: Indicate 5 as the current data sets are of the format percentage of each class interval and then the class means so echp1_5.dat contains the following data:

            % in class Mean of group
            20.021299 12017.405
            20 21185.881
            19.98935 27201.384
            20.021299 35218.907
            19.968051 61506.654

        v. Next POVCAL asks if the data have percentage or proportion format: Indicate 1 for proportions
        vi. INPUT NAME OF YOUR RESULTS FILE: Any input is possible
        vii. Then a series of screens are displayed which end with asking for you to INPUT POVERTY LINE Z WHICH IS WITH A SPECIFIED RANGE. POVCAL supplies a range so choose any number within this range. If you use so echp1_5.dat, then 5000 is fine.
        viii. It then produces an estimated Gini 31.2 (general quadratic) and 31.2 (Beta) which is reported in Table 4. This procedure is repeated for every data set.
    d. Estimates of the group bias corrected are summarized in the sheet “POVCAL_Comparison” in final calculations.xlsx. Generally the General quadratic Lorenz curve was the preferred specification indicated from POVCAL diagnostics tests.

Analysis based on MEPS

For the analyses using MEPS, one should take the following steps.

1. Obtain the Household Component Full-Year file for 2000 from http://www.meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-050 (see also "instructions for obtaining the data"), and convert to a STATA data file (extension *.dta). The file should be renamed into h50.dta
2. Run the file "group_data_probleml_MEPS_gini.do". This file is similar to the file "group_data_probleml_ECHP_gini.do" (see also step 4 under ECHP), but runs on MEPS rather than ECHP. It allows constructing the results for the US, i.e. the columns "Gini" and "FOC" in table 4, and the descriptive statistics in table A.1.
3. We have also applied the POVCAL software tool upon the MEPS data. To this means, one should
    a. Run "POVcal_MEPS.do" to obtain the files that will serve as input for the POVCAL software. The names of these files are "mepsx.dat" where x refers to the number of income groupings.
    b. Download and install the POVCAL software tool (see step 5b under ECHP)
    c. The same procedure was followed for ECHP analysis (see step 5c under ECHP).
    d. Generally the General quadratic Lorenz curve was the preferred specification indicated from POVCAL diagnostics tests.

Procedure to remove the second-order bias using an empirical estimate of the covariances (see also equation 16 in the paper)

One should take the following steps.

1. One has to create a STATA data file (name: "gini.dta") based on the log files for all countries obtained under step 4 of the ECHP analyses and step 2 of the MEPS analysis. This STATA data file has the variables in columns and the observations (49 per country) in rows. In other words, there are 784 observations (i.e. 16 countries times 49 observations). The columns are
    a. groups: the number of income groups (50 to 2)
    b. biasrel: the LHS of equation 16, i.e. G^k_n / G_n with k the number of groups
    c. gini: the Gini based on the full sample, i.e. G_n.
    d. country: the name of the country, i.e. Austria, Belgium, Denmark, Finland, France, Germany, Greece, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, UK and US.
2. Run the do file "SOC_reg.do". This do-file calls the file "gini.dta" and reports the estimation results of equation (16). These coefficient estimates are reported in the paper in table A.2.
3. The same coefficient estimates are used to construct the columns SOC_r in table 4. This is done in the sheet "SOC" in the excel-file "final calculations.xlsx". We have copy-pasted the covariance estimates from step 2 in the first column and the columns FOC from Step 4 in the ECHP analysis and step 2 in the MEPS analysis. The resulting second order correction is reported under SOC_reg, and is presented in the column SOC_r in table 4.

Procedure to obtain table 5 and figure 1

Table 5 and figure 1 are *solely* based on the information presented in table 4.

1. The exact construction of table 5 is illustrated in the sheet “Table 5” in the file “final calculations.xlsx”
2. The construction of figure 1 is illustrated in sheet “Figure “ in the file “final calculations.xlsx”