Cross-country comparison using the ECHP
Descriptive statistics and Simple Models

Cheti Nicoletti

Institute for Social and Economic Research

Comparing income variables across countries

- Income variables are measured in different currencies.
- Until 2001 there was not a common currency.
- How to measure income variables in a common currency for EU countries in the ECHP?
- Using the Purchasing-Power-Parity (PPP)

Purchasing-Power-Parity (PPP)

Information download from
http://encyclopedia.thefreedictionary.com/Purchasing%20power%20parity

- **PPP exchange rates are useful for comparing living standards between countries.**
- **Actual exchange rates can give a very misleading picture of living standards.**
- **For example, if the value of the Italian lira (Euro now) falls by half compared to the GB pound, the average household income observed in the ECHP for Italy measured in pounds will also halve.**
- **However, this does not necessarily mean that Italians are any poorer - if incomes and prices measured in lira (Euro) stay the same - they will be no worse off assuming that imported goods are not essential to the quality of life of individuals.**
- **Measuring income in different countries using PPP exchange rates helps to avoid this problem.**

Basic idea of the PPP

- $EXa=\sum pa_i\, x_i$ expenditure to buy a bundle of goods $x=(x_1, x_2, \ldots, x_n)$ at prices for country A, say $pa=(pa_1, pa_2, \ldots, pa_n)$
- $EXb=\sum pb_i\, x_i$ expenditure to buy a bundle of goods $x=(x_1, x_2, \ldots, x_n)$ at prices for country B, say $pb=(pb_1, pb_2, \ldots, pb_n)$
- PPP exchange rate consists in computing the rate between expenditure in country A and in country B.
- PPP= EXa/ EXb
- At this exchange rate we can assure that a person can buy the same amount of bundle of goods when measured at domestic prices and at foreign prices.

Criticisms of PPP

Information download from
http://encyclopedia.thefreedictionary.com/Purchasing%20power%20parity

- Critics say it is wrong to assume that the prices of goods should be equal in all countries. People in different countries usually put different values on the same goods.
- The exchange rate says how much you can buy in another country with one unit of your own currency. But the PPP does not.
- Most sources do not state the goods used to measure the PPP.

PPP from the country file

Country	93	94	95	96	97	98	99	00	01
Germany, mark	2.22	2.16	2.15	2.13	2.09	2.09	2.04	1.95	1.99
Ireland, punt	0.73	0.71	0.70	0.74	0.73	0.78	0.81	0.85	0.89
UK, pound	0.70	0.70	0.73	0.71	0.72	0.73	0.74	0.72	0.72

Comparing personal income

- Pi100: TOTAL NET PERSONAL INCOME (DETAILED, NC, TOTAL YEAR PRIOR TO THE SURVEY) pi100

- To compare income variables across countries we have to divide them by the purchasing power parity rate for the reference year.

- pi100 collected in first wave, 1994, refers to the year 1993.

pi100/ppp93

Comparing household income taking account of different household sizes

- Hi100 = TOTAL NET HOUSEHOLD INCOME (TOTAL YEAR PRIOR TO THE SURVEY, Amount in National Currency) hi100
- hd003 = Number of household members age <=14
- hd001 = Household size
- hd005 = EQUIVALISED SIZE, MODIFIED-OECD SCALE

  hd005=[1+0.5*(hd003-1)+0.3*(hd001-hd003)]
- Equivalized household income=hi100/hd005

Nominal and real income

- Nominal income = Income measured at current prices.
- Real income = Income measured at constant prices (as if the same prices applied each year)
- If the inflation >0 and a person has the same nominal personal income in two consecutive waves, then the person is becoming poorer.
- In the second period the person is not able to buy the same amount of goods and services because of the prices increase.

Comparing income variables across waves

- The income variables in the ECHP are nominal.
- To compare income variables across waves we need to use the consumer price indexes.
- Pi100: total net personal income in national currency (total year prior o the survey
- ICP index of consumer prices
- Let pi100 the nominal personal income for 1993 and collected in 1994
- Real personal income = pi100*100/ICP93

Harmonised ICP

Eurostat:

- “The harmonised indices of consumer prices (HICP’s) provide the best statistical basis for comparisons of consumer price inflation within the EU. The methodology ensures comparability between Member States.”
- For comparability the HICP for each country has a common base year, 1996=100.

Missing data for the HICP

- Eurostat releases the harmonised annual average consumer price indices for all countries belonging to the European Union.
- BUT the time series are available only from 1995.
- Solution suggested: Impute the 1993 and 1994 missing data by using the ICP previously released by Eurostat and correct them to take account of a different base year.

Comparing income across waves and countries

- To have comparable measures of income variables across countries and waves we have to:

1. Measure the income at constant prices of the base year (1996)
2. Use the purchasing power parity exchange rate in 1996 (the base year of the HICP) to convert national incomes in a common purchasing power

Descriptive statistics by countries and wave for continuous variables

- Using variables whose definition in harmonized is possible to compare descriptive statistics computed by countries and waves.
- For continuous variables like the personal income, say pincome

Stata command:

```
table wave country,
    c(median pincome
    mean pincomec sd
    pincome)
```

wave	ireland	germany	uk
1994	6187.361	11558.65	9964.104
	9257.311	12822.71	11636.19
	10643.02	10952.85	9394.72
1995	6817.87	11138.07	10341.71
	9797.414	12197.67	12604.08
	12605.01	10264.89	16998.33
1996	7132.72	11440.62	10611.98
	10103.7	12485.42	12538.02
	13978.68	10046.53	9979.383
1997	7676.388	11654.95	11164.35
	10807.33	12858.99	13262.75
	15598.05	10162.55	10720.12
1998	8386.793	11758	11520.25
	11638.18	12942.27	13698.26
	15842.15	10005.06	11797.57
1999	9061.613	12014.92	11702.44
	12440.88	13381.97	14001.36
	18825.76	10532.38	13634.21
2000	9558.211	12385.05	12096.94
	12728.57	13933.37	14641.47
	13161.46	10995.37	14318.81
2001	10222.72	12547.1	12849.17
	13593.19	14134.62	15297.28
	14014.3	11098.95	13884.94

Descriptive statistics by countries and age for continuous variables

- Even if ECHP does not have complete personal life histories is possible to have an idea of the profile by age of some variables.
- For continuous variables like the personal income, say pincome

Stata command:

table wave ageg, c(median pincome mean pincomec sd pincome)

country	16	26	36	46	56	66	76
ireland	5968	11080	11764	8781	6135	5739	5624
	6860	12146	14070	13020	10427	8578	7523
	5381	11538	18396	17425	14626	9599	8201
germany	5529	12578	13958	13740	10482	10270	11393
	6708	12759	15198	15187	12245	11636	13050
	5896	8744	11113	12182	11388	8643	8760
uk	7750	12829	14152	12994	10380	8554	8341
	7928	14003	16448	15284	12587	10592	9948
	6560	10713	17425	13637	11830	8831	7415

Descriptive statistics by countries and age for discrete variable variables

- Even if ECHP does not have complete personal life histories is possible to have an idea of the profile by age of some discrete variables.
- Let us consider the main activity status self-defined pe002, and age (ageg) then we can use the following Stata command

  bys country: tab ageg ls, row

- Table reports only the UK case

ageg	normally	unemploye	inactive	Total
16	2,718	251	888	3,857
	70.47	6.51	23.02	100
26	7,388	296	1,984	9,668
	76.42	3.06	20.52	100
36	8,055	243	2,014	10,312
	78.11	2.36	19.53	100
46	7,008	276	2,319	9,603
	72.98	2.87	24.15	100
56	2,585	152	3,777	6,514
	39.68	2.33	57.98	100
66	308	4	5,187	5,499
	5.6	0.07	94.33	100
76	12	2	3,050	3,064
	0.39	0.07	99.54	100
Total	28,074	1,224	19,219	48,517
	57.86	2.52	39.61	100

Advantages of panel data

It is possible:

1. to analyse labour, income and other dynamics in the life course,
2. to estimate the duration of some events such as unemployment,
3. to identify people moving to and out from a status (ex. unemployment), so that both gross and net changes are identified,
4. to control for unobserved heterogeneity due to personal unobservable characteristics which do not change across time (by considering random and fixed effects)

Issues for panel data analysis

- Missing data: besides the item and unit nonresponse in a single wave we have also to deal with the problem of people non responding in some of the waves (attrition in particular)
- The assumption of constant parameters across individuals may be inadequate, then random (or fixed) coefficient models must be considered.

Descriptive dynamic analysis

- Since the same individuals are followed across waves it is possible to compute changes rates or first differences for personal variables.
- Dynamics analysis obviously requires that an individual be respondent in both waves for which we want to compute differences or change rates.
- There are different way to compute differences between waves.

1st method to compute differences

Let pfile1 and pfile2 two personal files with the following variables:
country hid pid pincome wave (real personal income already in PPP)

```
use pfile1, clear
append using pfile2
sort country pid wave
by country pid: gen pincome_1=pincome[_n-1]
gen dpincome=pincome-pincome_1
gen lpincome=log(pincome)
gen lpincome_1=log(pincome_1)
gen dlpincome=lpincome-lpincome_1
gen chrpincome=(pincome-pincome_1)/pincome_1
```

2nd method to compute differences

Let pfile1 and pfile2 two personal files with the following variables:
country hid pid pincome wave (real personal income already in PPP)

use pfile1, clear
append using pfile2
reshape wide pincome, i (country pid) j(wave)
gen dpincome=pincome2-pincome1
gen lpincome2=log(pincome2)
gen lpincome1=log(pincome1)
gen dlpincome=lpincome2-lpincome1
gen chrpincome=(pincome2-pincome1)/pincome1

Balanced and unbalanced panel data

- A balanced panel is given by the subsample of people that are responding in all waves.
- An unbalanced panel is given by the sample of all people responding in at least one wave.
- The size of the balanced panel is smaller and the potential bias due to selection might be bigger.

Response patterns

- A response pattern can be described by the 8-dimensional vector D = (D_1,D_2,D_3,D_4,D_5,D_6,D_7,D_8).
- 255 (2^8 - 1) participation patterns are possible
- continued participation: D=(1,1,1,1,1,1,1,1);
- **monotone attrition**: D=(1,0,0,0,0,0,0,0), D=(1,1,0,0,0,0,0,0), …
- **new entry**: D=(0,1,1,1,1,1,1,1), D=(0,0,1,1,1,1,1,1), …
- occasional nonresponse: D=(1,0,1,1,1,1,1,1), D=(1,0,1,1,1,1,1,1), …
- occasional response: D=(0,1,0,0,0,0,0,0), D=(0,0,1,0,0,0,0,0),
- very irregular response: all other participation patterns.

Response patterns in first 5 waves

	Continued Particpation	Monotone Attrition	New Entry	Occasional Nonresponse	Occasional Response	Very irregular response
Denmark	46.8	31.9	8.1	5.1	4.9	3.2
France	58.1	26.6	8.1	2.6	3	1.7
Greece	55.5	27.6	10.6	1.7	2.8	1.8
Ireland	44.7	40	9.1	1.1	3.8	1.3
Italy	62.4	19.5	11	3.3	2.2	1.7
Portugal	62.4	16	14.6	3	2.6	1.5
Spain	50.4	29.6	10.9	3.9	2.9	2.3

Causes of nonparticipation

1. Ineligibility reasons

- Natural demographic events: death or 16th birthday.
- Movement from in to out of scope of the survey, or vice versa: it includes institutionalization, migration to a foreign country, movement of a nonsample person to a household without sample individuals, etc.

2. Nonresponse reasons

- Absence of the person at the address.
- Other types of contact failure: it includes the case of incomplete number of callbacks or interview not attempted for some reason, person omitted by error, inability to contact the person because address non residential or non existent, inability to locate the address, or other reasons.
- Lack of cooperation (refusal to respond): it includes definite or temporary refusal to participate, individuals unable to respond because of physical or language problems, and failure to return a self-completed questionnaire.

Causes of nonparticipation in first 5 waves

	Demographic event	Out of scope	Collection problems	Absence	Lack of cooperation
	Causes of non participation before entry				
New entry	42.6	45.5	5.1	2.3	4.5
Occasional response	22.2	58.9	7	4.2	7.8
	Causes of drop out				
Attrition	9.7	4.5	50.9	4.6	30.3
Occasional nonresponse	0	7.7	41.5	18.1	32.6
Occasional response	3.7	8.5	59.3	5.8	22.7
Very irregular response	0.5	8.6	35.5	15	40.5

Description of participation patterns

sort country pid wave
bys country: xtdes , i(pid) t(wave)
Ireland

Freq.	Percent	Cum.	Pattern
2948	24.65	24.65	11111111
1850	15.47	40.12	1.......
1233	10.31	50.43	11......
929	7.77	58.20	111111..
813	6.80	65.00	111.....
741	6.20	71.20	11111...
640	5.35	76.55	1111....
443	3.70	80.26	1111111.
164	1.37	81.63	.1......
2197	18.37	100.00	(other patterns)
11958	100.00		XXXXXXXX

Controlling for unobserved heterogeneity

$y_{i,t}=\alpha+ x_{i,t}\beta+z_i\gamma+u_{i,t}$ i=1,…N, t=1,…,T

- Let be z unobservable variables, then if z and x are correlated the OLS will be inconsistent.
- Panel data allows to control for ***unobserved heterogeneity*** by considering the ***first differences*** or the ***deviations from the mean***, i.e. by considering ***fixed effects models***.
- When the regression is not linear (ex. probit model) is not in general possible to consider fixed effects, ***random effects estimation*** is the only solution in those cases (few exceptions exist as for example in the case of the logit model).
- Random effects estimators are consistent if and only if random effects are uncorrelated with the explanatory variables.

Stata command for fixed and random effects models.

Fixed, between and random-effects, and population-averaged linear models

GLS Random-effects model

xtreg depvar [varlist] [if exp] [, re i(varname) sa theta level(#)]

xttest0 (testing if the variance of the random effects is 0)

Between-effects model

xtreg depvar [varlist] [if exp] , be [i(varname) wls level(#)]

Fixed-effects model

xtreg depvar [varlist] [if exp] , fe [i(varname) level(#)]

ML Random-effects model

xtreg depvar [varlist] [weight] [if exp] , mle [i(varname) noconstant level(#)]

Population-averaged model

xtreg depvar [varlist] [weight] [if exp] , pa [i(varname) noconstant level(#) offset(varname) xtgee_options]

How to choose between random effects and fixed effects linear models

- Hausman test:

  H_0 random effects uncorrelated with explanatory vars

  Under H_0 both random and fixed effects estimators are consistent and the random effects model is more efficient,

  Under H_1 only random effects estimator is consistent

- Stata commands

  xtreg y x1 x2 x3, fe

  est store fixed

  xtreg y x1 x2 x3, re

  hausman fixed

- Under H_0 the test is distributed as a chi2(1)

Dynamic models

$y_{i,t} = \rho\, y_{i,t-1} + x_{i,t}\,\beta + \mu_i + u_{i,t}$ *level*

$dy_{i,t} = \rho\, dy_{i,t-1} + dx_{i,t}\,\beta + du_{i,t}$ *first differences*

- Fixed effects estimator applied to dynamic models is inconsistent because $dy_{i,t-1}$ *and* $du_{i,t}$ *are correlated*
- The solution is to use lagged *y* as IV and apply a GMM estimator.
- Arrelano-Bond linear, dynamic panel-data estimation
- Stata command: xtabond

Logit model for panel data

Stata commands

Fixed-effects, random-effects, and population-averaged logit models

Random-effects model

```
xtlogit depvar [varlist] [weight] [if exp] [in range] [, re i(varname) ]
```

Conditional fixed-effects model

```
xtlogit depvar [varlist] [weight] [if exp] [in range] , fe [i(varname) ]
```

Population-averaged model

```
xtlogit depvar [varlist] [weight] [if exp] [in range] , pa [i(varname)]
```

Probit model for panel data

Random-effects and population-averaged probit models

Random-effects model

xtprobit depvar [varlist] [weight] [, re i(varname)]

Population-averaged model

xtprobit depvar [varlist] [weight] , pa [i(varname) robust]

Harmonized index of consumer prices

country	cpi93	cpi94	cpi95	cpi96	cpi97	cpi98	cpi99	cpi00	cpi01
1	94.453	97.022	98.8	100	101.5	102.1	102.8	104.2	106.2
2	94.08	96.04	98	100	101.9	103.3	105.4	108.3	110.7
3	94.064	96.727	98.6	100	101.9	103.7	105.8	108.2	113.8
4	94.663	96.924	98.3	100	101.5	102.4	103.6	106.4	109
5	94.848	96.923	98.8	100	101.4	102.4	103.4	107.3	109.9
6	94.668	96.334	98	100	101.3	102	102.5	104.4	106.3
7	92.134	94.379	97.6	100	101.8	103.4	104.8	105.6	106.9
8	93.299	95.453	97.9	100	101.2	103.4	106	111.5	116
9	87.927	91.486	96.2	100	101.9	103.9	105.7	108.4	110.9
10	76.478	84.821	92.7	100	105.4	110.25	112.6	115.8	120.1
11	90.47	94.736	99.2	100	101.9	102.9	103.4	104.8	107.6
12	88.646	93.312	97.2	100	101.9	104.2	106.4	109.4	114.2
13	93.385	96.137	98.3	100	101.2	102	102.5	104.5	106.9
14	96.922	97.911	98.9	100	101.2	102.6	103.9	107	109.8
15	94.637	96.72	99.2	100	101.9	102.9	103.4	104.8	107.6

Reading the HICP data file and merge it with the country file

```
insheet using hcpi.csv, clear
sort country
save hcpi.dta, replace
use train_ctyvar.dta, clear
sort country
merge country using hcpi.dta
drop _m
keep if country==1 |country==7
    |country==8
recode country 7=57 1=51
```

```
rename cpi00 cpi2000
rename cpi01 cpi2001
rename ppp00 ppp2000
rename ppp01 ppp2001
local i=93
while `i'<=99{
local s=1900+`i'
rename cpi`i' cpi`s'
rename ppp`i' ppp`s'
local i=`i'+1
}
keep country cpi* ppp*
sort country
save country.dta, replace
```

Reshaping the country file and computing lagged ppp and cpi

```
keep country ppp* cpi*
gen pppbase=ppp1996
reshape long ppp cpi, i(country) j(wave)
replace wave =wave-1993
sort country wave
by country: gen ppp_1=ppp[_n-1]
by country: gen cpi_1=cpi[_n-1]
sort country wave
save countryl.dta, replace
```

Comparing personal income (pi100) across countries for a same wave

```
use country hid pid wave pi100 using trn_w1p.dta, clear
keep if country==8 | country==51|country==57
rename pi100 pincome
sort country
merge country using country
tab _m
keep if _m==3
replace pincome=pincome/ppp1993
```

Comparing household income controlling for household size

```
use country hid hd005 wave hi100 using trn_w1h.dta, clear
keep if country==8 | country==51|country==57
*dividing household income by the EQUIVALISED SIZE,
gen ehincome=hi100/hd005
rename hi100 hincome
rename hd005 hsize
*Computing mean and median by country
collapse (mean) hincome ehincome hsize , by(country)
sort country
merge country using country
tab _m
keep if _m==3
replace hincome=hincome/ppp1993
replace ehincome=ehincome/ppp1993
```

Appending personal files

```
local i=1
while `i'<=8{
use country hid pid wave pi100 pd003 pd004 pe002 using
    "D:\home\nicolet\data\echp\epunet\trn_w`i'p.dta", clear
keep if country==8 | country==51|country==57
rename pi100 pincome
sort country
save pfile`i', replace
local i=`i'+1
}
use pfile1, clear
local i=2
while `i'<=8{
append using pfile`i'
local i=`i'+1
}
```

Comparing income across waves and countries

```
sort country wave
merge country wave using countryl
tab _m
keep if _m==3
drop _m
gen
   pincomec=pincome*100/(cpi_1*pppbase)
```

Example: Earnings equation

```
*Hausman test random versus fixed effects
xtreg wage age exp bhealth edu1 edu2
   marst if country==8 & sex==1, fe i(pid)
est store  fixed
xtreg wage age exp bhealth edu1 edu2
   marst if country==8 & sex==1, re i(pid)
hausman fixed
```

Fixed effect model

```
Fixed-effects (within) regression               Number of obs      =      5597
Group variable (i): pid                         Number of groups   =      1965

R-sq:  within  = 0.1838                         Obs per group: min =         1
       between = 0.1024                                        avg =       2.8
       overall = 0.1013                                        max =         8

                                                F(6,3626)          =    136.10
corr(u_i, Xb)  = -0.8145                        Prob > F           =    0.0000

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0914535   .0755286     1.21   0.226    -.0566292    .2395362
         exp |   .0122289   .0755239     0.16   0.871    -.1358446    .1603025
     bhealth |    -.02072   .0935933    -0.22   0.825    -.2042208    .1627807
        edu1 |   .0750615   .0526845     1.42   0.154    -.0282326    .1783556
        edu2 |   .0560552   .0427449     1.31   0.190    -.0277513    .1398618
       marst |   .0892807   .0436175     2.05   0.041     .0037635     .174798
       _cons |   5.909722   1.403445     4.21   0.000     3.158102    8.661342
-------------+----------------------------------------------------------------
     sigma_u |  1.1320419
     sigma_e |  .41145892
         rho |  .88330823   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(1964, 3626) =     5.52                 Prob > F = 0.0000
```

Random effects model

```
Random-effects GLS regression                   Number of obs      =      5597
Group variable (i): pid                         Number of groups   =      1965

R-sq:  within  = 0.1292                         Obs per group: min =         1
       between = 0.1583                                        avg =       2.8
       overall = 0.1554                                        max =         8

Random effects u_i ~ Gaussian                   Wald chi2(6)       =    743.76
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0529253    .006398     8.27   0.000     .0403855     .065465
         exp |  -.0167883   .0061454    -2.73   0.006    -.0288331   -.0047434
     bhealth |  -.0079139   .0884719    -0.09   0.929    -.1813156    .1654878
        edu1 |    .400669   .0391888    10.22   0.000     .3238603    .4774776
        edu2 |   .2503276   .0322392     7.76   0.000     .1871399    .3135152
       marst |  -.0085353    .030604    -0.28   0.780    -.0685181    .0514474
       _cons |   7.339955   .1287644    57.00   0.000     7.087581    7.592328
-------------+----------------------------------------------------------------
     sigma_u |  .65131692
     sigma_e |  .41145892
         rho |  .71475153   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Hausman test

```
hausman fixed
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-
   V_B))
             |     fixed           .          Difference          S.E.
-------------+----------------------------------------------------------
   ---
          age |    .0914535     .0529253        .0385282        .0752571
          exp |    .0122289    -.0167883        .0290172        .0752735
      bhealth |     -.02072    -.0079139       -.0128061        .0305357
         edu1 |    .0750615      .400669       -.3256074        .0352121
         edu2 |    .0560552     .2503276       -.1942723        .0280672
        marst |    .0892807    -.0085353        .0978161        .0310786
------------------------------------------------------------------------
   ---
                            b = consistent under Ho and Ha; obtained from
   xtreg
             B = inconsistent under Ha, efficient under Ho; obtained from
   xtreg
    Test:  Ho:  difference in coefficients not systematic
                  chi2(6) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      452.38
                Prob>chi2 =      0.0000
```