

# **LIS Self Teaching Package 2022**

**Stata version**

**Part II**

**Gender, employment, and wages**





## **Part II: Gender, employment, and wages**

### **Overall Plan and Structure of the Exercise**

The exercises in Part I demonstrated the use of household income data along with useful programming techniques for working with the LIS data. Part II emphasises the use of person-level data, including wages, demographics, and labour market information. Whereas Part I consisted entirely of calculating descriptive statistics, Part II introduces regression modelling in the final exercises.

The program that was written in the first set of exercises is now completed and can be set aside. Starting with the next exercise, you will begin the process of building up an entirely new program for Part II. Many of the techniques shown in the previous part will be useful again. In addition, users will learn how to combine LIS datasets by merging household and person files, and by concatenating multiple country-year datasets into a single file.

The general approach of the exercises is the same as in Part I. After beginning a new program in the initial exercise, each subsequent exercise will add new code to the existing program. Within each exercise, results<sup>1</sup> will be produced to illuminate the central research themes of this section.

### **Research Questions**

The analysis of poverty and inequality using household income, which was covered in Part I, has always been central to research using LIS data. Over the years, however, there has been an increasing volume of work that examines individual outcomes in the labour market. The richness of the labour market data available in LIS has increased over time, and today it is possible to address many types of questions about wages and employment.

Labour market outcomes for women are one especially popular area of research. Women's rate and intensity of work shows much wider cross-country variation than men's. At the same time, on average women consistently earn lower wages.

---

<sup>1</sup> Please note that results calculated by you might differ from the ones presented here given that the LIS data is subject to updates from time to time.

For the exercises, we will examine three countries, using data from LIS Wave VI: The United States, Belgium, and Greece. As we will see, labour market outcomes for women show distinctly different patterns in each of these countries. Looking at persons of prime working age (which we will define as ages 25-54), our central questions will be:

- How does the percentage of prime-aged women employed in paid work vary across these three countries?
- Among those who are employed, how does the rate of *part-time* employment among women vary across the countries?
- How does employment vary by partnership and family status?
- How do wage differentials between men and women vary between across countries, across levels of educational attainment, and between immigrants and non-immigrants?

In the exercises, we will begin by producing tabulations of employment and wages for various population subgroups. In the concluding exercises, we will use linear regression to study multiple determinants of wages simultaneously, in order to better understand how family structure, education, and immigrant status are related to wages for men and for women.

## **Before you begin**

Before you begin the exercises, take a look at the [2019 Template LIS User Guide](#), which can be accessed through *LIS Website* → *Our Data* → *LIS Database*. The User Guide provides an overview of the structure of LIS data and some data management practices, such as missing values policy and aggregation rules, which will be useful for working with LISSY.

In addition to this, an overview of the datasets and variables is provided through the METadata Information System ([METIS](#)) without having to login to LISSY. You can access METIS via the *LIS Website* → *METIS* → *Enter METIS* → *LIS*. After selecting the datasets and variables, consult the *Results tab* for information on variables and definitions, dataset-specific information and variable availability across datasets.

## **Contents**

### **1. Merging person and household data, selecting a sample**

- Selecting a sample of prime-age workers
- Homeownership rates by country

### **2. Stacking data, employment rates by gender**

- Employment rates by country and gender
- Part-time employment rates by country and gender

### **3. Family structure and employment**

- Family/partnership status and employment rates

### **4. Dependent employment and hourly wages**

- Non-dependent employment as a proportion of employment, by country
- Gender wage gaps among dependent employees, by country

### **5. Hourly wages, education, and country-specific variables**

- Harmonised vs. country-specific codings
- Gender wage gaps by country and educational attainment

### **6. Immigration and wages, understanding harmonisation**

- Understanding harmonization choices
- Gender wage gaps by country and immigrant status

### **7. Wage regressions**

- Regressing log wages on demographics, separately by gender and country

### **8. Pooled regressions and normalised weights**

- PPP adjustments for income
- Pooled regression with multiple countries, using normalised weights

# 1. Merging person and household data, selecting a sample

## Goal

While the exercises in Part I only used data at the household level, Part II uses data from both the household and person level files. In this exercise, we will combine the person and household files in order to create a single dataset for each country in which household data is appended to each person record.

This exercise also selects the universe of persons that we will be studying in the subsequent exercises. Since we are interested in labour market outcomes, we will restrict our attention to people of *prime age*: those who are likely to be old enough to have completed schooling and young enough to not yet be retired. In these exercises, we will define prime age persons as those between 25 and 54 years old. This is a commonly used range in statistics from the United States government and other sources, but other definitions are also possible.

Some of the variables we will be using are not always available for household members other than the head and spouse. For that reason, we will further restrict our universe to heads and partners only.

All of the variables that will be needed in the subsequent exercises are introduced here. However, for now we will only analyse one: we will summarize home ownership, which is a household-level variable, in order to measure the rate of homeownership in each of the three countries under analysis. Homeownership will be included in our later multivariate analysis, because it serves as a rough proxy for wealth, which we otherwise have no information about.

## Activity

Go to *LIS website* → [\*Login LISSY\*](#) tab with your LISSY account. Write a program to loop through three datasets: United States 2004 (**us04**), Belgium 2004 (**be04**) and Greece 2004 (**gr04**). In each country, merge the person file to the household file and keep the following variables:

- Household: Unique household identifier (**hid**) and owned/rented housing indicator (**own**)
- Person: Unique household identifier (**hid**), dataset name (**dname**), normalized person weight (**pwgt**), relationship to the household head (**relation**), partnership status (**partner**), age of youngest own child living in the household (**ageyoch**), age (**age**), sex (**sex**), immigrant

indicator (**immigr**), 3-category recoded educational attainment (**educ**), country-specific educational attainment (**educ\_c**), indicator for employment (**emp**), status in employment (**status1**), indicator for part-time employment (**ptime1**), and gross hourly wage in the first job (**gross1**).

Keep only those cases that are in the prime age range (between 25 and 54), and which are defined as either household heads or spouses in the variable **relation**.

Create an indicator variable equal to 1 if a person owns their house (with or without a mortgage), and 0 otherwise. Summarize this new variable to find the homeownership rate among the prime-aged persons for each country, and complete the following table:

Dataset	Homeownership %
BE04	
GR04	
US04	

### Questions

- 1.1. In which country do the largest percentage of persons of prime working age own their houses, and in which country are homeownership rates lowest?
- 1.2. Can we consider all non-homeowners as tenants? If not, what other housing tenure status are possible?

### Guidelines

- In order to make your code easier to read, it may be helpful to store the list of variables you will be using in global macros. You will need two global macros, one for household level variables and one for person level variables. You can also store the list of datasets in a macro, and refer to it when constructing your loop.
- In the current version of Stata, merging two files can be done with one line of code. First, open one of the datasets, and then use the **merge** command. The code below performs a *many-to-one* merge. This means that each case in the second file can have its information copied to more than one record in the first file:

```
use <variable list> using <file>, clear
```

```
merge m:1 <merge variables> using <file>, keepusing(<variable list>)
```

In this case, the first file you open should be the person file. You will then attach household-level information to the person-level records. The *merge variable* (or variables) is a variable that exists in both files, indicating which household level cases go with which person level records. In LIS datasets, the variable **hid** uniquely identifies households, and can be found in both the person and household files. The **keepusing()** clause can include a list of household level variables that you wish to keep in the merged file.

- To create a new variable of homeownership based on the categories of an existing variable, you can use the Stata command **recode**:

**recode own (100/199=1) (200/299=0), gen(homeowner)**

- Remember that when recoding variables, you can find a listing of the possible values of the original variable in [METIS](#). In this case, go to the LIS Database information (*Enter METIS → LIS*). Select BE04, GR04 and US04. Select the variable **own**. Go to *Results → Crossed-compare* and click on the variable name to see the statistics and labels of the variables.
- As the household file, the person file contains weight variables. These variables can be used to weight by person, as an alternative to the method of multiplying household weights by number of household members that was used in the Part I. Although homeownership is a household level variable, you will want to use the person weight to determine the proportion of *persons* who live in owner-occupied dwellings. For now, use the variable **ppopwgt**, which inflates to the total population size.

## Program

```
global varshh "hid own"  
global varspp "hid dname pwgt ppopwgt relation partner ageyoch age sex  
immigr educ educ_c emp status1 ptime1 gross1"  
global datasets "us04 be04 gr04"  
  
foreach ccyy in $datasets {  
    use $varspp using $`ccyy'p, clear  
    merge m:1 hid using $`ccyy'h, keepusing($varshh)  
    keep if inrange(age,25,54) & relation<=2200  
    recode own (100/199=1) (200/299=0), gen(homeowner)  
    sum homeowner [aw=ppopwgt]  
}
```

## **Results**

Dataset	Homeownership %
BE04	69.5%
GR04	67.6%
US04	71.5%

### *Solutions*

- 1.1. In which country do the largest percentage of persons of prime working age own their houses, and in which country are homeownership rates lowest?
  - Homeownership rates are highest in the United States at 71.5%. Belgium and Greece have similar rates (69.5% and 67.6%, respectively).
- 1.2. Can all non-homeowners be considered as tenants? If not, what other housing tenure status are possible?
  - No, in these three datasets there is also a category labelled as "free housing". According to the LIS variable definitions available in [METIS](#) this category can include "housing provided by employer, government or others, or illegal occupation".

### *Comments*

- You will notice that in this exercise the merge worked perfectly, i.e. all observations of the merging file were uniquely linked to one observation in the using file. This is always the case with LIS household and individual level files from the same dataset because all individuals belong to at least one and no more than one household.

## 2. Stacking data, employment rates by gender

### Goal

So far, we have performed all the analysis separately for each dataset, working with only one country at a time. For this and all subsequent analyses, however, we will create a “stacked” dataset that contains information for all three countries in a single file. This means your dataset will have as many value observations as your countries altogether have stored within one file. This is necessary if you wish to analyse several countries simultaneously, which we will be doing in the regressions in [Exercise 8](#) (Part II). In case of independent cross country comparisons, the use of a stacked dataset is an alternative technique to the looping through datasets.

After creating a combined dataset, we will examine rates of employment and part-time employment of women, and see how they differ among these three countries. As in the previous exercise, we will be looking only at prime-aged persons who are defined as household heads or partners of household heads.

We will be using the LIS variable **emp**, an indicator that reports whether or not a person is currently employed. This variable will contain the current main employment status (as derived from LIS variable **Ifs**). If the current main employment status is not available in the original dataset, employment status during income reference time or employment status according to the ILO criteria in the current period will be used in LIS variable **Ifs** and hence also **emp** (you can check dataset-specific notes in [METIS](#) for information about each dataset). By this definition, a person may be considered as employed as soon as he/she has carried out any work.

Rates of employment and full-time employment among prime-age men tend to be similar and consistently high across countries. Due to this, we will be examining differences in employment outcomes among women.

### Activity

Modify your program so that it first creates a combined data file for the United States, Belgium, and Greece, and then performs any necessary recoding and produces descriptive statistics.

Create a set of cross-tabulations that shows the rates of prime-age employment of women within each country. Create another set of cross-tabulations showing the rates of part-time work of women within each country, among those who are employed. Use your results to complete the following table below.

You should write your code so that your overall program is broken down into three subroutines. The first subroutine should contain only the code needed to create the merged, stacked dataset. The second subroutine should contain all of the data-preparation and recoding. The third subroutine should contain the code that produces the summary statistics. Your overall program can then simply call these two subroutines to make the dataset and output the results. Breaking up your code in this way will be important for making the program compact and efficient in later exercises.

Dataset	Female employment rate	Part-time employment rate among employed women
BE04		
GR04		
US04		

### *Question*

- 2.1 Contrast these countries in terms of their rates of female employment (high or low) and their rates of part-time employment among employed women (high or low).

### **Guidelines**

- You do not need to remove the code that you used to produce the descriptive in the last exercise (on homeownership), but you can comment it out to make your job run slightly faster. To comment out a single line, place an asterisk (\*) at the beginning of the line. To comment out multiple lines of code, place the characters /\* at the point where you want to begin the comment, and \*/ where you want the comment to end.
- When stacking datasets in Stata, you will need to save some temporary files. When saving temporary files, you can place them in a directory at LIS. In order to ensure that your filename differs from others saving files at the same time, your filename should have a unique name. For example, you could include your LIS username in the filename:

**save \${mydata}exercisepart2\_<myusername>, replace**

- Remember that to put all of your country files together, you will need to append the datasets. Here is an example of the code that will perform the person-to-household merge shown in the previous exercise, and

then stack the three merged datasets, keeping only the cases and variables needed:

```
foreach ccyy in $datasets {  
    use $varspp using $`ccyy'p, clear  
    merge m:1 hid using $`ccyy'h, keepusing($varshh)  
    keep if inrange(age,25,54) & relation<2200  
    if "`ccyy'" != "us04" {  
        append using ${mydata}exercisepart2_<myusername>  
    }  
    save ${mydata}exercisepart2_<myusername>, replace  
}
```

This example assumes that you have stored your list of datasets in the global macro **\$datasets**, and that **us04** is the first dataset listed. When the code above processes United States 2004 data, it will overwrite the stacked file if such a file did already exist. For subsequent countries, it will append the stacked file to the current dataset before saving the new file.

Once the stacked file is created, it will remain on the server, so you will be able to comment out the file-creation code in the next exercises.

- Once you have produced the stacked data file, you can produce descriptive statistics for multiple countries at once by using the **dname** variable and the Stata command **bysort**, which performs the same command repeatedly for subsets of the data. For example:

```
bysort dname: tabulate emp [aw=ppopwgt] if sex==2
```

This code will produce the summary statistics on homeownership by country that we produced in the previous exercise.

- You can get the row percentages in a cross-tabulation (such as part-time employment among all employed women by country) by using the option **row nofreq** on the command:

```
tabulate dname ptime1 [aw=ppopwgt] if emp==1 & sex==2, row nofreq
```

## Program

```
global varshh "hid own"
global varspp "hid dname pwgt ppopwgt relation partner ageyoch age sex immigr educ
educ_c emp status1 ptime1 gross1"
global datasets "us04 be04 gr04"

program define make_data
foreach ccyy in $datasets {
    use $varspp using $`ccyy'p, clear
    merge m:1 hid using $`ccyy'h, keepusing($varshh)
    keep if inrange(age,25,54) & relation<=2200
    if "`ccyy'" != "us04" {
        append using ${mydata}exercise2_LIS
    }
    save ${mydata}exercise2_LIS, replace
}
end

program define get_descriptives
bysort dname: tabulate emp [aw=ppopwgt] if sex==2
tabulate dname ptime1 [aw=ppopwgt] if emp==1 & sex==2, row noref
end

quietly make_data
use ${mydata}exercise2_LIS, clear
get_descriptives
```

## **Results**

Dataset	Female employment rate	Part-time employment rate among employed women
BE04	69.2%	37.9%
GR04	57.6%	21.3%
US04	72.0%	17.7%

### *Solution*

- 2.1. Contrast these countries in terms of their rates of female employment (high or low) and their rates of part-time employment among employed women (high or low).
- Employment rates among prime age women are relatively high in the United States and in Belgium, and lower in Greece. In the United States, most employed women work full time, while more than one-third of employed Belgian women work part time. Greece combines low employment rates with high rates of full time employment among those women who are employed.

### 3. Family structure and employment

#### Goal

In the previous exercise, we examined cross-national differences in women's employment. In this exercise, we will examine the variation in employment rates among women, based on their partnership and family status. We will contrast partnered and single women. Within each of those two categories, we will contrast women without children in the household, women with young children, and women with older children. The variables created in this exercise will be useful later, when we combine family structure with other personal characteristics in a multivariate analysis of wages.

#### Activity

Since you already created the merged and stacked dataset in the previous exercise, you do not need to create it again. Modify your code so that the subroutine that merges and stacks the data is commented out, and add a line that simply loads the merged and stacked file at the beginning of the program.

Create a variable **achildcat**, to indicate the age of the youngest own child living in the household. This variable should be equal to 0 if there are no children under 18, equal to 1 if the youngest child is under 6 years old, and equal to 2 if the youngest child is between 6 and 17. You can create this variable based on the information in the variable **ageyoch**.

Produce summary statistics to complete the table below using **achildcat**, an indicator of whether a person is currently living with a partner (=1) or not (=0) (**partner**) and the employment indicator you used in the previous exercise (**emp**).

#### Employment Rates

Dataset	All women	Single			Partnered		
		No children under 18	Child under 6	Child 6-17	No children under 18	Child under 6	Child 6-17
BE04							
GR04							
US04							

*Question*

- 3.1 Within each country, which subpopulation of prime age women has the lowest employment rates?

**Guidelines**

- To produce a three-way table with employment rates among women, you can use the **table** command:

```
table dname achildcat partner [aw=ppopwgt] if sex==2, contents(mean emp) format(%9.3f)
```

The **format** option ensures that only the desired number of digits is displayed.

## Program

```
global varshh "hid own"
global varspp "hid dname pwgt ppopwgt relation partner ageyoch age sex immigr educ
educ_c emp status1 ptime1 gross1"
global datasets "us04 be04 gr04"

program define make_data
foreach ccyy in $datasets {
    use $varspp using $`ccyy'p, clear
    merge m:1 hid using $`ccyy'h, keepusing($varshh)
    keep if inrange(age,25,54) & relation<=2200
    if "`ccyy'" != "us04" {
        append using ${mydata}exercise2_LIS
    }
    save ${mydata}exercise2_LIS, replace
}
end

program define recode_data
recode ageyoch (. 18/max= 0 "no children <18") (0/5 = 1 "<6 years") (6/17 = 2 "6-17 years"), gen(achildcat)
label var achildcat "Lowest age of own children"
end

program define get_descriptives
table dname achildcat partner [aw=ppopwgt] if sex==2, contents(mean emp)
format(%9.3f)
end

*quietly make_data
use ${mydata}exercise2_LIS, clear
quietly recode_data
get_descriptives
```

## **Results**

### **Employment Rates**

Dataset	All women	Single			Partnered		
		No children under 18	Child under 6	Child 6-17	No children under 18	Child under 6	Child 6-17
BE04	69.2%	67.4%	40.2%	58.2%	67.8%	71.3%	75.6%
US04	72.0%	80.1%	66.9%	78.7%	76.2%	58.4%	71.7%
GR04	57.6%	68.6%	64.2%	81.4%	49.7%	57.7%	60.5%

### *Solutions*

**Question:** Within each country, which subpopulation of prime age women has the lowest employment rates?

- In the United States, partnered women with young children have the lowest employment rates. In Greece, partnered women without children have the lowest employment rates. In Belgium, however, single mothers with children have lower employment rates. This may be because of more generous child policy in Belgium that makes it easier for mothers of young children to support themselves without paid employment.

### *Comments*

- There is no clear-cut definition of a single-mother household. In this exercise, we allow other adult members to be present (as long as they are not defined as her partner). An alternative approach would be to limit the sample to households composed of a single female adult and her children. Another possibility is to limit single mother households to those with children under a specified age limit.
- When subdividing subsets of the population as has been done here, pay attention to sample sizes. In small datasets, estimates for narrowly defined groups may become very small, making estimates less reliable. The estimate for single Greek women with young children in this exercise, for example, is based on only 17 cases!

## 4. Dependent employment and hourly wages

### Goal

In the next several exercises, we will shift from considering employment to analysing the earnings of those who are employed. We will focus our analysis on the hourly wages and thus restrict our sample to those in dependent employment only — that is, those who are employees. The self-employed, along with several other small categories of workers, are excluded.

In this exercise we will first determine how many workers are excluded from the analysis when the sample is restricted to those in dependent employment. We will then measure the gap in hourly wages for men and women, in each of the three countries in our study.

We will be using a measure of gross hourly wages, which is available in the three datasets we are using. In other datasets, however, it could be that only annual wages were available. In such cases, researchers must account for variations in employment over the year, perhaps by restricting the sample to full-year, full-time workers.

In part I, we have introduced bottom- and top-coding as a technique to deal with extreme values. This is especially important when calculating measures that are not defined for non-positive values (such as logarithmic measures). In later exercises we will convert the hourly wages into logs, and thus we need to make sure that the sample that we analyse at this stage is the same that we will keep for our final analysis.

### Activity

Recode the variable **status1** to create a new variable **depemp** that indicates whether a person is in dependent employment. Using this variable, produce summary statistics reporting the proportion of dependent employment among prime-age male workers and among prime-age female workers, and complete the following table.

## Employment Rates

Dataset	Men		Women	
	Non-dependent Employment (%)	Dependent employment (%)	Non-dependent Employment (%)	Dependent employment (%)
BE04				
GR04				
US04				

Next, use the LIS hourly wage variable **gross1** to construct a new hourly wage variable **hourwage**, where the bottom and the top of the distribution are corrected as follows:

- we will carry out the same bottom- and top-code as used in Part I, interquartile range (IQR): first, hourly wages is log transformed and used to calculate the log values for the interquartile range; second, the exponential of the log values in the original hourly wages before the log transformation:  $\text{EXP} [\log Q1 - 3*(\log Q3 - \log Q1)]$  for the lower boundary and  $\text{EXP} [\log Q3 + 3*(\log Q3 - \log Q1)]$  for the upper boundary.

Using this new hourly wage variable, calculate the *gender wage gap* for dependent employees in each country. The gender wage gap is defined here as the ratio of the median wages of women to the median wages of men. Use your results to complete the table below.

	Gender wage gap for dependent employees
BE04	
GR04	
US04	

## Questions

- 4.1 Does the percentage of workers not in dependent employment differ substantially across countries? Does it differ between men and women?
- 4.2 Which country has the most wage inequality between men and women, among dependent employees?

## Guidelines

- For the purpose of bottom- and top-coding, you can loop over the datasets (after they have been stacked) in order to apply a separate upper and lower bound for each country:

```
foreach ccyy in $datasets {  
    sum hourwage_log [aw=ppopwgt] if dname==``ccyy'', de  
    if ``ccyy'' == "us04" {  
        gen iqr=r(p75)-r(p25) if dname==``ccyy''  
        * detect upper bound for extreme values  
        gen upper_bound=r(p75) + (iqr * 3) if dname==``ccyy''  
        gen lower_bound=r(p25) - (iqr * 3) if dname==``ccyy''  
    }  
    if ``ccyy'' != "us04" {  
        replace iqr=r(p75)-r(p25) if dname==``ccyy''  
        * detect upper bound for extreme values  
        replace upper_bound=r(p75) + (iqr * 3) if dname==``ccyy''  
        replace lower_bound=r(p25) - (iqr * 3) if dname==``ccyy''  
    }  
    * top code income at upper bound for extreme values  
    replace hourwage=exp(upper_bound) if hourwage>exp(upper_bound) &  
    !mi(hourwage) & dname==``ccyy''  
    * bottom code income at lower bound for extreme values  
    replace hourwage=exp(lower_bound) if hourwage<exp(lower_bound) &  
    !mi(hourwage) & dname==``ccyy''  
}
```

For further discussion of bottom- and top-coding, see Part I, exercise 3.

- As before, you can use the **bysort** command to produce summary statistics subdivided by country and gender.
- You may find it helpful to attach labels to your new variables using the **label define** and **label values** commands, as in the following example:

```
label define depempl 0 "not in dependent employment" 1 "in dependent  
employment"  
label values depemp depempl
```

## Program

```
global varshh "hid own"
global varspp "hid dname pwgt ppopwgt relation partner ageyoch age sex immigr educ
educ_c emp status1 ptime1 gross1"
global datasets "us04 be04 gr04"

program define make_data
foreach ccyy in $datasets {
    use $varspp using $`ccyy'p, clear
    merge m:1 hid using $`ccyy'h, keepusing($varshh)
    keep if inrange(age,25,54) & relation<=2200
    if "`ccyy'" != "us04" {
        append using ${mydata}exercise2_LIS
    }
    save ${mydata}exercise2_LIS, replace
}
end

program define recode_data
recode status1 (100/120=1) (200/240=0) (else=.), gen(depemp)
label define depempl 0 "not in dependent employment" 1 "in dependent employment"
label values depemp depempl
gen hourwage = gross1
replace hourwage=0 if gross1<0
gen hourwage_log=log(hourwage)
* keep negatives and 0 in the overall distribution of non-missing dhi
replace hourwage_log=0 if hourwage_log==. & hourwage!=.
foreach ccyy in $datasets {
    sum hourwage_log [aw=ppopwgt] if dname=="`ccyy'", de
    if "`ccyy'" == "us04" {
        gen iqr=r(p75)-r(p25) if dname=="`ccyy'"
        * detect upper bound for extreme values
        gen upper_bound=r(p75) + (iqr * 3) if dname=="`ccyy'"
        gen lower_bound=r(p25) - (iqr * 3) if dname=="`ccyy'"
    }
}
```

```

if "`ccyy'" != "us04" {
    replace iqr=r(p75)-r(p25) if dname=="`ccyy'"
    * detect upper bound for extreme values
    replace upper_bound=r(p75) + (iqr * 3) if dname=="`ccyy'"
    replace lower_bound=r(p25) - (iqr * 3) if dname=="`ccyy'"

}
* top code income at upper bound for extreme values
replace hourwage=exp(upper_bound) if hourwage>exp(upper_bound) &
!mi(hourwage) & dname=="`ccyy'"
* bottom code income at lower bound for extreme values
replace hourwage=exp(lower_bound) if hourwage<exp(lower_bound) &
!mi(hourwage) & dname=="`ccyy'"

}
end

program define get_descriptives
bysort dname: tab sex depemp [aw=ppopwgt] if emp==1, row norefq
bysort dname sex: sum hourwage [aw=ppopwgt], de
end

quietly make_data
use ${mydata}exercise2_LIS, clear
quietly recode_data
get_descriptives

```

## **Results**

### **Employment Rates**

Dataset	Men		Women	
	Non-dependent Employment (%)	Dependent employment (%)	Non-dependent Employment (%)	Dependent employment (%)
BE04	15%	85%	10%	90%
GR04	34%	66%	29%	71%
US04	14%	86%	8%	92%

### **Gender wage gaps**

	Gender wage gap for dependent employees
BE04	0.96
GR04	0.85
US04	0.75

### *Solutions*

- 4.1 Does the percentage of workers not in dependent employment differ substantially across countries? Does it differ between men and women?
- Greece has a much higher rate of non-dependent employment (which is primarily self-employment). In all the countries, women have higher rates of dependent employment than men do. Keep in mind, therefore, that the results in the subsequent exercises may be unrepresentative, particularly for Greece, because they exclude a substantial number of workers.
- 4.2 Which country has the most wage inequality between men and women, among dependent employees?
- The United States shows the largest gender wage gap. Among prime age workers, the median hourly wage of women is only 75% that of men.

### *Comments*

- The wage gap calculated here is based on the median, but some researchers calculate an alternative version based on the mean, which will give slightly different results.

## 5. Hourly wages, education, and country-specific variables

### Goal

This exercise continues the analysis of gender wage gaps in hourly wages among those in dependent employment, which we started to program in the previous exercise. In this exercise, we will see how gender disparities in wages differ by educational attainment.

This exercise also demonstrates the use of two different LIS variables measuring educational attainment. One is fully standardised for cross-national compatibility, but contains few categories. The other may contain more information, but has country-specific codes, and thus requires researchers to perform their own standardisation.

The standardised variable is called **educ**, which is based on the International Standard Classification of Education (ISCED). The non-standardised version (from which **educ** is constructed) is **educ\_c**. This is one of many attributes for which LIS provides both a standardised and country-specific variable. Any variable ending in **\_c** is non-standardised, meaning that it can have different contents in different datasets. It is important to carefully examine the dataset-specific documentation before using such variables.

### Activity

Add code to your program to create a table cross-tabulating the variables **educ** and **educ\_c** for each country. This will show how the standardised variable was constructed in each case. Be sure to:

- include missing values in your table, so that you can see whether any of the cases in the original education variable could not be allocated to a category in the standardised version;
- remove the value labels from the tabulation of the **educ\_c** variable (since the value labels of the **\_c** variables are by definition dataset-specific, in a stacked dataset with observations from several LIS datasets, the value labels of those variables will be incorrect, as they can only refer to one specific LIS dataset – usually the last one that was used to construct the stacked data, see further details in the comments section of this exercise).

Using the hourly wage variable **hourwage** that you created in the last exercise, calculate the gender wage gap by education for each country, and complete the table below. The gender wage gap is defined as it was in the

previous exercise. To obtain the earnings ratio by education, simply calculate the ratio separately for individuals in each of the three categories of the standardised education variable.

### **Gender wage gaps by educational attainment**

	Low education	Medium education	High education
BE04			
GR04			
US04			

### *Questions*

- 5.1 For each of the three countries, what are the categories in the original dataset that are recoded as “high education” in the standardised education variable?
- 5.2 Are there any categories in the original **educ\_c** variable that could not be translated into the standardised form?
- 5.3 In general, which educational attainment category shows the greatest earnings inequality between genders? How do the patterns differ by country?

### **Guidelines**

- Stata reminder: in order to remove the values labels from a variable, use the command **label values** and assign a missing value label; in this way, any existing value label is detached from that variable:  
**label values <varname> .**
- If you are uncertain about the meaning of a code in a categorical variable, consult [METIS](#) for variable information and codebook.
- As in the previous exercises, you can produce tabulations for each country by placing **bysort dname**: before your commands, and then calculate ratios by hand from your resulting output. However, an alternative method that will produce more compact output is to loop through countries and education levels and store results in macros, outputting only the final ratio of interest using the **display** command, as in the following example code:

```
foreach ccyy in $datasets {
    foreach e in 1 2 3 {
```

```
quietly sum hourwage [aw=ppopwgt] if dname=="`ccyy'" & educ==`e' &
sex==2, de
local fwage : di %9.2f r(p50)
quietly sum hourwage [aw=ppopwgt] if dname=="`ccyy'" & educ==`e' &
sex==1, de
local wageratio : di %9.2f `fwage'/r(p50)
display "`ccyy' educ=`e': `wageratio'"
}
}
```

## Program

```
global varshh "hid own"  
global varspp "hid dname pwgt ppopwgt relation partner ageyoch age sex immigr educ  
educ_c emp status1 ptime1 gross1"  
global datasets "us04 be04 gr04"  
  
program define make_data  
foreach ccyy in $datasets {  
    use $varspp using $`ccyy'p, clear  
    merge m:1 hid using $`ccyy'h, keepusing($varshh)  
    keep if inrange(age,25,54) & relation<=2200  
    if "`ccyy'" != "us04" {  
        append using ${mydata}exercise2_LIS  
    }  
    save ${mydata}exercise2_LIS, replace  
}  
end  
  
program define recode_data  
gen hourwage = gross1  
replace hourwage=0 if gross1<0  
gen hourwage_log=log(hourwage)  
* keep negatives and 0 in the overall distribution of non-missing hourwage  
replace hourwage_log=0 if hourwage_log==. & hourwage!=.  
foreach ccyy in $datasets {  
    sum hourwage_log [aw=ppopwgt] if dname=="`ccyy", de  
    if "`ccyy'" == "us04" {  
        gen iqr=r(p75)-r(p25) if dname=="`ccyy"  
        * detect upper bound for extreme values  
        gen upper_bound=r(p75) + (iqr * 3) if dname=="`ccyy"  
        gen lower_bound=r(p25) - (iqr * 3) if dname=="`ccyy"  
    }  
    if "`ccyy'" != "us04" {
```

```

replace iqr=r(p75)-r(p25) if dname=="`ccyy'"
* detect upper bound for extreme values
replace upper_bound=r(p75) + (iqr * 3) if dname=="`ccyy'"
replace lower_bound=r(p25) - (iqr * 3) if dname=="`ccyy'"

}

* top code income at upper bound for extreme values
replace hourwage=exp(upper_bound) if hourwage>exp(upper_bound) &
!mi(hourwage) & dname=="`ccyy'"

* bottom code income at lower bound for extreme values
replace hourwage=exp(lower_bound) if hourwage<exp(lower_bound) &
!mi(hourwage) & dname=="`ccyy'"

}

label values educ_c .

end

program define get_descriptives
bysort dname: tab educ_c educ, mi
foreach ccyy in $datasets {
    foreach e in 1 2 3 {
        quietly sum hourwage [aw=ppopwgt] if dname=="`ccyy'" &
        educ==`e' & sex==2, de
        local fwage : di %9.2f r(p50)
        quietly sum hourwage [aw=ppopwgt] if dname=="`ccyy'" &
        educ==`e' & sex==1, de
        local wageratio : di %9.2f `fwage'/r(p50)
        di "`ccyy' educ=`e': `wageratio'"
    }
}
end

quietly make_data
use ${mydata}exercise2_LIS, clear
quietly recode_data
get_descriptives

```

## **Results**

### **Gender wage gaps by educational attainment**

	Low education	Medium education	High education
BE04	0.85	0.90	0.90
GR04	0.67	0.78	0.96
US04	0.71	0.71	0.78

### *Solutions*

- 5.1. For each of the three countries, what are the categories in the original dataset that are recoded as “high education” in the standardised education variable?
- In the United States, high education combines those with associate degrees, bachelor's degrees, and advanced degrees (masters, professional school, or doctorate).
  - In Belgium, high education combines those classified as having “higher vocational” (of any kind) or “higher education”.
  - In Greece, high education includes those with tertiary graduate level education, postgraduate level education or Ph.D.
- 5.2. Are there any categories in the original **educ\_c** variable that could not be translated into the standardised form?
- In the United States, all values of **educ\_c** receive a value in **educ**. In Belgium, a small number of persons categorised as “inadequately, other diploma” or “still in education” are set to missing. In Greece, a small number of persons listed as “still in education” are set to missing.
- 5.3. In general, which educational attainment category shows the greatest earnings inequality between genders? How do the patterns differ by country?
- In all three countries, there is a smaller gender wage gap among highly educated workers. This is particularly notable in Greece. In that country, wage inequality is greater among the low-educated than in the United States and Belgium, but there is near equality among the highly educated.

### *Comments*

- You may have some doubts and questions why we did advise you in the exercise to not show the labels of **educ\_c** for the cross-tabulation of **educ** and **educ\_c**. As you are aware we did append values for several countries for each variable to get the stacked file. In your stacked file, for standardised variables these values have all the same meaning, as the values and labels are completely standardised. However, it is more complicated for non-standardized values and labels of \_c variables, as each dataset has its own values and own meaning, as indicated by the labels attached to the data.
- Be aware that while appending the data, your programming software will very likely overwrite the label automatically. Thus we do in general advise you to drop the labels from the variables **\_c** within your code. You always have the full information on the labels in the codebooks. However, if you prefer to keep the full labels somewhere in the data there are several solutions.
- A simple solution is to tabulate each country separately (see for example exercise 2.1 of part II) before you generate the stacked version. Alternatively, you can also easily rename the variables **\_c** to the specific **\_`ccyy'** of each dataset - this way you will append a separate variable **\_`ccyy'** for each of your datasets, which does have only observations for the specific 'ccyy' with the country specific labels attached. Be aware that you then need to tabulate for each 'ccyy' separately to get the right percentage of missing values!

## 6. Immigration and wages, understanding harmonisation

### Goal

Each of the countries we are examining has a significant immigrant population, and their labour market outcomes are often very different from those of the non-immigrant population. In this exercise, we will compare the wages of immigrants and non-immigrant men and women, just as we compared individuals of different educational levels in the last exercise.

LIS provides a variable indicating whether someone is an immigrant, which we will be using in this exercise. However, the choices that go into constructing this variable are complex, because the information available to construct it varies widely from country to country. It is important to understand the assumptions behind variables such as this one, because in some cases researchers may prefer to develop their own standardisation procedures based on their particular needs.

### Activity

Using the bottom- and top-coded hourly wage variable, calculate the gender earnings ratio by immigration status for each country, and complete the table below. The gender earnings ratio is computed just as in the previous exercise, except that you will now subdivide the population into immigrants and non-immigrants, rather than by educational attainment categories.

#### **Gender earnings ratios by immigration status**

	Non-immigrants	Immigrants
BE04		
GR04		
US04		

### *Questions*

- 6.1 What information is used to construct the **immigr** variable? If you wanted to determine how the indicator was constructed in a particular dataset, what other variables would you need to look at?
- 6.2 Is gender earnings inequality larger among immigrants or non-immigrants? Does this differ by country?

## **Guidelines**

- The coding required for this exercise is essentially the same as in the previous one. You can produce your results either with **bysort** or with loops. See the guidelines for the previous exercise for details.
- Remember that if you want to comment out the loop from the previous exercise, you can do so by surrounding it with **/\*** and **\*/**.

## **Program**

```
global varshh "hid own"  
global varspp "hid dname pwgt ppropwgt relation partner ageyoch age sex immigr educ  
educ_c emp status1 ptime1 gross1"  
global datasets "us04 be04 gr04"  
  
program define make_data  
foreach ccyy in $datasets {  
    use $varspp using `$ccyy'p, clear  
    merge m:1 hid using `$ccyy'h, keepusing($varshh)  
    keep if inrange(age,25,54) & relation<=2200  
    if "`ccyy'" != "us04" {  
        append using ${mydata}exercise2_LIS  
    }  
    save ${mydata}exercise2_LIS, replace  
}  
end  
  
program drop recode_data  
program define recode_data  
gen hourwage = gross1  
replace hourwage=0 if gross1<0  
gen hourwage_log=log(hourwage)  
* keep negatives and 0 in the overall distribution of non-missing dhi  
replace hourwage_log=0 if hourwage_log==. & hourwage!=.
```

```

foreach ccyy in $datasets {
    sum hourwage_log [aw=ppopwgt] if dname=="`ccyy'", de
    if "`ccyy'" == "us04" {
        gen iqr=r(p75)-r(p25) if dname=="`ccyy'"
        * detect upper bound for extreme values
        gen upper_bound=r(p75) + (iqr * 3) if dname=="`ccyy'"
        gen lower_bound=r(p25) - (iqr * 3) if dname=="`ccyy'"
    }
    if "`ccyy'" != "us04" {
        replace iqr=r(p75)-r(p25) if dname=="`ccyy'"
        * detect upper bound for extreme values
        replace upper_bound=r(p75) + (iqr * 3) if dname=="`ccyy'"
        replace lower_bound=r(p25) - (iqr * 3) if dname=="`ccyy'"
    }
    * top code income at upper bound for extreme values
    replace hourwage=exp(upper_bound) if hourwage>exp(upper_bound) &
    !mi(hourwage) & dname=="`ccyy'"
    * bottom code income at lower bound for extreme values
    replace hourwage=exp(lower_bound) if hourwage<exp(lower_bound) &
    !mi(hourwage) & dname=="`ccyy'"
}
label values educ_c .
end

```

```

program define get_descriptives
foreach ccyy in $datasets {
    foreach i in 0 1 {
        quietly sum hourwage [aw=ppopwgt] if dname=="`ccyy'" &
        immigr==`i' & sex==2, de
        local fwage : di %9.2f r(p50)
        quietly sum hourwage [aw=ppopwgt] if dname=="`ccyy'" &
        immigr==`i' & sex==1, de
        local wageratio : di %9.2f `fwage'/r(p50)
        di "`ccyy' immigr=`i': `wageratio'"
}

```

```
    }  
}  
end  
  
quietly make_data  
use ${mydata}exercise2_LIS, clear  
quietly recode_data  
get_descriptives
```

## **Results**

### **Gender earnings ratios by immigration status**

	Non-immigrants	Immigrants
BE04	0.95	1.02
GR04	0.89	0.82
US04	0.76	0.82

### *Solutions*

- 6.1 Question: What information is used to construct the **immigr** variable? If you wanted to determine how the indicator was constructed in a particular dataset, what other variables would you need to look at?
- As indicated in the Variable Definition of **immigr** using [METIS](#) documentation tool, Immigrants are defined by LIS as all persons who have that country as country of usual residence and (in order of priority):
    - whom the data provider defined as immigrants;
    - who self-define them-selves as immigrants;
    - who are the citizen/national of another country;
    - who were born in another country.
  - The definition of immigrant used in **immigr** may differ substantially from dataset to dataset. The variables that may be used in its construction include **citizen**, **ctrybrth**, **yrsresid**, **ethnic\_c** and **immigr\_c** (you can check dataset-specific notes in [METIS](#) for information about each dataset).
- 6.2 Question: Is gender earnings inequality larger among immigrants or non-immigrants? Does this differ by country?
- In the United States and Belgium, the gender wage gap is greater among non-immigrants, but in Greece, it is greater among immigrants.

## 7. Wage regressions

### Goal

We have seen how employment varies by gender and family structure, and how gender wage gaps vary by education and immigration status. In this exercise, we will investigate the impact of all these variables on wages, using a multivariate regression.

Wages are generally not normally distributed. We will therefore apply a logarithmic transformation in order to create an outcome variable that is approximately normal, which is more suitable for regression modelling.

In addition to the variables we have already seen, we will also control for age, which has a strong relationship with earnings. Since the relationship between age and income is not necessarily linear, we will also add a term for age-squared.

### Activity

Create a new variable, **logwage**, which is the natural logarithm of the **hourwage** variable you created earlier.

Create a variable **agesq**, which is equal to **age** squared.

Create dummy variables based on the **achildcat** and **educ** variables you have used previously. You will need the following variables:

- **youngchild**, equal to 1 if the person has a child under 6 in the household, 0 otherwise.
- **oldchild**, equal to 1 if the person has a child age 6-17 in the household, 0 otherwise.
- **hieduc**, equal to 1 if the person has high educational attainment, 0 otherwise.
- **mededuc**, equal to 1 if the person has medium educational attainment, 0 otherwise.

When including categorical variables in a regression, one category is always left out and used as the reference category. The reference categories in this case are having no non-adult children, and having low educational attainment.

After creating all of the variables, run regression models predicting **logwage**. You should run the regressions separately in each country, and within each country you should run a separate model for men and women.

Use the following model:

```
logwage = f (age agesq mededuc hieduc immigr partner youngchild  
oldchild ptime1 homeowner)
```

Produce a table of the six resulting models, with coefficients, standard errors, sample sizes, and r-squared values.

It is recommended that you create a new sub-routine to run the regressions, which can be called from your main program.

### Questions

- 7.1 Who receives a higher wage premium from being highly educated, men or women?
- 7.2 When controlling for other individual characteristics, what is the relationship between immigrant status and wages?
- 7.3 When controlling for other individual characteristics, do women with young children make more or less than women without children?

### **Guidelines**

- Linear regression in Stata is done with the **regress** command:

```
regress <outcome> <predictors> [pw=<weight>], vce(robust)
```

Use of the **robust** option is recommended to correct standard errors for possible heteroskedasticity.

- When performing several regression models in a single program, Stata allows results to be stored as they are calculated and then printed out later, using the **eststo** command:

```
foreach g in 1 2 {  
    foreach ccyy in $datasets {  
        quietly eststo: regress <outcome> <predictors> [pw=<weight>] if  
        sex==`g' & dname=="`ccyy'", vce(robust)  
    }  
    esttab, b(a2) se(a2) r2(a3) mtitles($datasets)  
    eststo clear  
}
```

The code above estimates regressions for men in each country and then for women in each country, and prints out two tables of estimates using the **esttab** command, along with standard errors and R-squared

statistics. It is also possible to produce table output in other formats, such as CSV (for spreadsheets) and TeX. See the documentation at <http://repec.org/bocode/e/estout/esttab.html> for more information.

## **Program**

```
global varshh "hid own"
global varspp "hid dname pwgt ppopwgt relation partner ageyoch age sex immigr educ
educ_c emp status1 ptime1 gross1"
global datasets "us04 be04 gr04"

program define make_data
foreach cyy in $datasets {
    use $varspp using $`cyy'p, clear
    merge m:1 hid using $`cyy'h, keepusing($varshh)
    keep if inrange(age,25,54) & relation<=2200
    if "`cyy'" != "us04" {
        append using ${mydata}exercise2_LIS
    }
    save ${mydata}exercise2_LIS, replace
}
end

program define recode_data
recode own (100/199=1) (200/299=0), gen(homeowner)
recode ageyoch (. 18/max = 0 "no children <18") (0/5 = 1 "<6 years") (6/17 = 2 "6-17
years"), gen(achildcat)
label var achildcat "Lowest age of own children"
gen hourwage = gross1
replace hourwage=0 if gross1<0
gen hourwage_log=log(hourwage)
* keep negatives and 0 in the overall distribution of non-missing dhi
replace hourwage_log=0 if hourwage_log==. & hourwage!=.
foreach cyy in $datasets {
    sum hourwage_log [aw=ppopwgt] if dname=="`cyy", de
    if "`cyy'" == "us04" {
        gen iqr=r(p75)-r(p25) if dname=="`cyy'"
```

```

        * detect upper bound for extreme values
        gen upper_bound=r(p75) + (iqr * 3) if dname=="`ccyy'"
        gen lower_bound=r(p25) - (iqr * 3) if dname=="`ccyy'"
    }

    if "`ccyy'" != "us04" {
        replace iqr=r(p75)-r(p25) if dname=="`ccyy'"
        * detect upper bound for extreme values
        replace upper_bound=r(p75) + (iqr * 3) if dname=="`ccyy'"
        replace lower_bound=r(p25) - (iqr * 3) if dname=="`ccyy'"
    }

    * top code income at upper bound for extreme values
    replace hourwage=exp(upper_bound) if hourwage>exp(upper_bound) &
    !mi(hourwage) & dname=="`ccyy'"

    * bottom code income at lower bound for extreme values
    replace hourwage=exp(lower_bound) if hourwage<exp(lower_bound) &
    !mi(hourwage) & dname=="`ccyy'"

}

label values educ_c .

gen logwage = ln(hourwage)
gen agesq=age^2
recode achildcat (1=1) (0 2=0) (else=.), gen(youngchild)
recode achildcat (2=1) (0 1=0) (else=.), gen(oldchild)
recode educ (2=1) (1 3=0) (else=.), gen(mededuc)
recode educ (3=1) (1 2=0) (else=.), gen(hieduc)
end

program define get_estimates
foreach g in 1 2 {
    foreach ccyy in $datasets {
        quietly eststo: regress logwage age agesq mededuc hieduc immigr
partner youngchild oldchild ptime homeowner [pw=ppopwgt] if sex=="`g' &
dname=="`ccyy'", vce(robust)
    }
    esttab, b(a2) se(a2) r2(a3) mtitles($datasets)
    eststo clear
}
end

```

```
quietly make_data  
use ${mydata}exercise2_LIS, clear  
quietly recode_data  
get_estimates
```

## **Results**

### **Men**

	(1) us04	(2) be04	(3) gr04
age	0.050*** (0.0056)	0.018 (0.011)	0.033 (0.017)
agesq	-0.00050*** (0.000070)	-0.000053 (0.00014)	-0.00023 (0.00021)
mededuc	0.28*** (0.015)	0.15*** (0.020)	0.17*** (0.028)
hieduc	0.68*** (0.016)	0.44*** (0.021)	0.43*** (0.037)
immigr	-0.056*** (0.012)	-0.049 (0.029)	-0.25*** (0.044)
partner	0.040** (0.013)	0.028 (0.023)	-0.014 (0.042)
youngchild	0.064*** (0.012)	0.026 (0.023)	0.12*** (0.034)
oldchild	0.071*** (0.011)	0.011 (0.022)	0.11*** (0.032)
ptime1	-0.29*** (0.031)	0.081 (0.046)	0.33*** (0.084)
homeowner	0.24*** (0.012)	0.096*** (0.020)	0.035 (0.026)
_cons	1.10*** (0.11)	1.71*** (0.22)	0.81* (0.32)
N	29996	1615	1175
R-sq	0.220	0.333	0.346

Standard errors in parentheses

\* p<0.05, \*\* p<0.01, \*\*\* p<0.001

## Women

	(1) us04	(2) be04	(3) gr04
age	0.041*** (0.0057)	0.038** (0.013)	0.071*** (0.021)
agesq	-0.00043*** (0.000072)	-0.00036* (0.00016)	-0.00065* (0.00027)
mededuc	0.32*** (0.018)	0.16*** (0.026)	0.35*** (0.040)
hieduc	0.75*** (0.018)	0.47*** (0.025)	0.76*** (0.043)
immigr	-0.019 (0.013)	-0.031 (0.031)	-0.26*** (0.051)
partner	-0.011 (0.010)	-0.0016 (0.023)	-0.067 (0.040)
youngchild	0.041** (0.013)	-0.0060 (0.025)	0.058 (0.042)
oldchild	-0.047*** (0.0100)	-0.047* (0.022)	0.026 (0.036)
ptime1	-0.17*** (0.012)	0.022 (0.019)	0.23*** (0.041)
homeowner	0.18*** (0.011)	0.076*** (0.021)	-0.0083 (0.032)
_cons	1.13*** (0.11)	1.34*** (0.23)	-0.26 (0.39)
N	28605	1507	966
R-sq	0.192	0.265	0.461

Standard errors in parentheses

\* p<0.05, \*\* p<0.01, \*\*\* p<0.001

### *Solutions*

- 7.1 Who receives a higher wage premium from being highly educated, men and women?
- In the US, Belgium and Greece, the coefficient for high education is higher for women, indicating a larger wage premium from having high educational attainment.
- 7.2 When controlling for other individual characteristics, what is the relationship between immigrant status and wages?
- The association between immigrant status and wages is negative in Greece (among men and women) and in the United States (among men)..
- 7.3 When controlling for other individual characteristics, do women with young children make more or less than women without children?
- In general, women with young children have higher wages than women without children in the United States, but there is no association in Belgium and Greece. Higher wages for women with young children could be due to a selection effect, where mothers of young children are more likely to enter the labour market if they have higher earning power.

### *Comments*

- As we have seen, employment rates, particularly among women, vary substantially across countries. The wage regressions shown here do not account for this differential selection into employment. For this reason, many studies of wages apply a technique such as a Heckman correction, which attempts to correct for this selection bias.

## 8. Pooled regressions and normalised weights

### Goal

In the previous exercise, we ran parallel, separate regressions for each country. In this exercise, we see an alternative approach in which all countries are pooled together in a single model. We will continue to use classical OLS regression, but the approach shown here can easily be extended for more complex multilevel estimation approaches.

The income variables in these datasets use different currencies. To compare them, we need to convert them to a common scale. We will apply *Purchasing Power Parity* (PPP) deflators, which are intended to ensure that equal quantities of income correspond to equivalent purchasing power across currencies and national economies. The PPP deflators used in this exercise are retrieved from the World Bank Development Indicators and are constantly updated. However, in order to compare real amounts across countries and over time, LIS provides the LIS PPPs, which combine CPI and PPP deflators retrieved from the World Bank Development Indicators. Since May 2020, with the release of the results from the 2017 cycle by the International Comparison Program ([ICP](#)), two sets of PPP deflators are available by LIS via LISSY: 2017 PPPs and the revised 2011 PPPs. In order to convert LIS monetary values into 2017 USD PPPs, amounts expressed in nominal currency should be divided by the LIS PPP of the corresponding year. For more information on the CPI and PPP deflators, please see [here](#), and check our tutorial videos on [Price deflators in LIS](#), and [Price deflators in LISSY using Stata](#).

Up to this point, we have been using the weight variable **ppopwgt**, which inflates to the total population. If we use this weight in a pooled regression, every household will receive equal weight. However, this would mean that Greece — which has a much smaller population than the United States or Belgium — will not have much influence on the results. In order to give each country equal weight in the model, we will use the alternative normalized weight variable **pwgt**, which always sums to 10,000 within each dataset.

### Activity

Create a new logged hourly wage variable, **logwage\_ppp** by dividing **hourwage** by the following deflators (<https://www.lisdatacenter.org/resources/ppp-deflators/>) and then taking the natural logarithm:

dataset	PPP deflator (2017 USD)
US04	1
BE04	0.86
GR04	0.65

Create dummy variables for **belgium** (=1 if the case is from Belgian dataset, otherwise 0) and **greece** (=1 if the case is from the Greek dataset, otherwise 0). The United States will be the reference category.

Estimate the following model, for men and women separately:

```
logwage = f (age agesq mededuc hieduc immigr partner youngchild
oldchild ptime1 homeowner belgium greece)
```

Make sure to use *normalised*, not inflated weights.

Produce a table of the two resulting models, with coefficients, standard errors, sample sizes, and r-squared values.

### Questions

- 8.1. How can you interpret the meaning of the coefficients for the dummy variables for Belgium and Greece?
- 8.2. In this pooled model, which carries a higher wage penalty: being an immigrant, or working part time?

### **Guidelines**

- Run your regressions as you did in the previous exercise. This time, only two models need to be produced, but it still may be helpful to store your estimates and then produce a table of estimates.

## Program

```
global varshh "hid own"
global varspp "hid dname pwgt ppopwgt relation partner ageyoch age sex immigr educ
educ_c emp status1 ptime1 gross1"
global datasets "us04 be04 gr04"

program define make_data
foreach ccyy in $datasets {
    use $varspp using `$ccyy'p, clear
    merge m:1 hid using `$ccyy'h, keepusing($varshh)
    keep if inrange(age,25,54) & relation<=2200
    if "`ccyy'" != "us04" {
        append using ${mydata}exercise2_LIS
    }
    save ${mydata}exercise2_LIS, replace
}
end

program define recode_data
recode own (100/199=1) (200/299=0), gen(homeowner)
recode ageyoch (. 18/max= 0 "no children <18") (0/5 = 1 "<6 years") (6/17 = 2 "6-17
years"), gen(achildcat)
label var achildcat "Lowest age of own children"
gen hourwage = gross1
replace hourwage=0 if gross1<0
gen hourwage_log=log(hourwage)
* keep negatives and 0 in the overall distribution of non-missing dhi
replace hourwage_log=0 if hourwage_log==. & hourwage!=.
foreach ccyy in $datasets {
    sum hourwage_log [aw=ppopwgt] if dname=="`ccyy'", de
    if "`ccyy'" == "us04" {
        gen iqr=r(p75)-r(p25) if dname=="`ccyy'"
        * detect upper bound for extreme values
        gen upper_bound=r(p75) + (iqr * 3) if dname=="`ccyy'"
        gen lower_bound=r(p25) - (iqr * 3) if dname=="`ccyy'"
    }
    if "`ccyy'" != "us04" {
        replace iqr=r(p75)-r(p25) if dname=="`ccyy'"
```

```

        * detect upper bound for extreme values
        replace upper_bound=r(p75) + (iqr * 3) if dname=="`ccyy"
        replace lower_bound=r(p25) - (iqr * 3) if dname=="`ccyy"
    }

    * top code income at upper bound for extreme values
    replace hourwage=exp(upper_bound) if hourwage>exp(upper_bound) &
    !mi(hourwage) & dname=="`ccyy"

    * bottom code income at lower bound for extreme values
    replace hourwage=exp(lower_bound) if hourwage<exp(lower_bound) &
    !mi(hourwage) & dname=="`ccyy"

}

label values educ_c .

gen logwage = ln(hourwage)
gen agesq=age^2
recode achildcat (1=1) (0 2=0) (else=.), gen(youngchild)
recode achildcat (2=1) (0 1=0) (else=.), gen(oldchild)
recode educ (2=1) (1 3=0) (else=.), gen(mededuc)
recode educ (3=1) (1 2=0) (else=.), gen(hieduc)

gen ppp = .
replace ppp=0.86 if dname=="be04"
replace ppp=0.65 if dname=="gr04"
replace ppp=1 if dname=="us04"
gen hourwage_ppp = hourwage/ppp
gen logwage_ppp = log(hourwage_ppp)
gen belgium=0
replace belgium=1 if dname=="be04"
gen greece=0
replace greece=1 if dname=="gr04"
end

program define get_estimates
quietly eststo: regress logwage_ppp age agesq mededuc hieduc immigr partner
youngchild oldchild ptime1 homeowner belgium greece [pw=ppopwgt] if
sex==1, vce(robust)
quietly eststo: regress logwage_ppp age agesq mededuc hieduc immigr partner
youngchild oldchild ptime1 homeowner belgium greece [pw=ppopwgt] if
sex==2, vce(robust)
esttab, b(a2) se(a2) r2(a3) mtitles(Men Women)
end

```

```
quietly make_data  
use ${mydata}exercise2_LIS, clear  
quietly recode_data  
get_estimates
```

## Results

	(1)	(2)
	Men	Women
age	0.049*** (0.0054)	0.042*** (0.0055)
agesq	-0.00049*** (0.000068)	-0.00044*** (0.000069)
mededuc	0.26*** (0.013)	0.31*** (0.016)
hieduc	0.66*** (0.014)	0.73*** (0.016)
immigr	-0.063*** (0.012)	-0.025* (0.012)
partner	0.040** (0.013)	-0.010 (0.010)
youngchild	0.063*** (0.012)	0.037** (0.012)
oldchild	0.069*** (0.010)	-0.048*** (0.0096)
ptime1	-0.26*** (0.029)	-0.15*** (0.012)
homeowner	0.23*** (0.011)	0.18*** (0.011)
belgium	-0.067*** (0.0099)	0.13*** (0.011)
greece	-0.37*** (0.014)	-0.23*** (0.017)
_cons	1.13*** (0.10)	1.12*** (0.11)
N	32786	31078
R-sq	0.229	0.197

Standard errors in parentheses

\* p<0.05, \*\* p<0.01, \*\*\* p<0.001

## *Solutions*

- 8.1 How can we interpret the meaning of the coefficients for the dummy variables for Belgium and Greece?
- These coefficients represent the overall national level of PPP-adjusted wages, when controlling for the other variables. The negative value for Greece and Belgium reflects the fact that both countries have lower wages than the United States.
- 8.2 In this pooled model, which carries a higher wage penalty: being an immigrant, or working part time?

Wage penalty for working part time is smaller than that for being an immigrant. Male immigrants and part-time workers seem to suffer higher wage penalties than female ones.