# LIS Self Teaching Package 2024

# R 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[1] 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.

[1] 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 coding
   - 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

**Additional guideline on saving your temporary data files at the LIS directory**

# 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 hourly wage in the first job (**hwage1**).

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 order to make your code easier to read, it may be helpful to store the list of variables you will be using in separate vectors, which you can refer to when you open the data file. You will need two vectors, one for household-level variables and one for person level variables. You can also store the list of datasets in a vector, and refer to it when constructing your loop.

- When loading data sets, you can use the parameter "labels" of the

**read.LIS** function to tell R whether to load the data with the LIS-supplied labels, or with numeric codes. For this and subsequent exercises, we will use **labels = FALSE**.

- The simplest way to merge datasets in R is to use the **merge** command. First, load the person and household files into separate data frames, and then combine them. If you have stored the person and household files in data frames called **dp** and **dh**, you can use:

  **df <- merge(dh, dp, by = c('hid'))**

- 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.

- Like 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 home ownership 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

```
setups <- function(ccyy) {
  # READ DATASETS
  varh   <- c('hid', 'dname', 'own')
  varp   <- c('hid', 'dname', 'ppopwgt', 'age', 'sex', 'relation', 'emp', 'ptime1', 'ageyoch',
'partner', 'status1', 'hwage1', 'educ', 'immigr')
  subset <- 'age >= 25 & age <= 54 & relation <= 2200'
  dh     <- read.LIS(paste(ccyy, 'h', sep = ''), labels  = FALSE, vars = varh)
  dp     <- read.LIS(paste(ccyy, 'p', sep = ''), labels = FALSE, vars = varp, subset =
subset)
  df     <- merge(dh, dp, by = c('hid'))
  # MAP NEW VARIABLES
  df$home <- ifelse(df$own %in% 100:199, 1, ifelse(df$own %in% 200:299, 0, NA))
  return(df)
}
#-------------------------------------- RUN SCRIPTS ---------------------------------------
datasets <- c('us04', 'be04', 'gr04')
for (ccyy in datasets) {
  df  <- setups(paste(ccyy, sep = ''))
  res <- round(with(df[!is.na(df$home),], sum(home*ppopwgt) /
sum(ppopwgt[!is.na(home)])) *100, digits=2)
  print(c(ccyy, res))
}
```

## 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 may offer you substantial advantages to make use of the data.

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 **lfs**). 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 **lfs** 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 line, place a hash (**#**) at the beginning of the line.

- The **read.LIS** command can automatically stack datasets if you call it with a vector of identifiers instead of a single string:

  **df <- read.LIS(c('us04h','be04h','gr04h'))**

  This will return a single data frame containing data from the United States, Belgium, and Greece. Note, however, that you cannot mix person and household file identifiers in the same call, or you will get an error. Thus you may want to first create two stacked datasets (one for person and one for household data) and then merge them. If you do this, however, note that you must merge based on two variables, **dname** and **hid**, since the household identifier by itself will no longer uniquely identify cases. This is easily done by passing a vector of merging variables as the third argument of the merge() command.

- In base R, weighted proportions do not exist as such. One option is to use the function **tapply(X, INDEX, FUN, …)** – that applies a function or operation on subset of the vector broken down by a given factor variable (or a discrete numeric variable).

  For example, apply the following code to calculate the employment rate among women:

  **with(df[df$sex==2 & !is.na(df$emp),], tapply(emp*ppopwgt, list(dname), sum) / tapply(ppopwgt, list(dname), sum)) * 100**

  This code generates the weighted proportion of the **emp** variable subdivided by dataset name for women. In addition, the results are multiplied by 100 to get the percentage.

## Program

```
get_stack <- function(datasets, varp, varh, subset) {
  # READ DATASETS
   pp <- read.LIS(paste(datasets, 'p', sep = ''), labels = FALSE, vars = varp, subset = subset)
   hh <- read.LIS(paste(datasets, 'h', sep = ''), labels = FALSE, vars = varh)
   df <- merge(pp, hh,  by = c("dname", "hid"))
   return(df)
}
#------------------------------ RUN SCRIPTS -----------------------------
varh  <- c('hid', 'dname', 'own')
varp  <- c('hid', 'dname', 'ppopwgt', 'age', 'sex', 'relation', 'emp', 'ptime1')
subset   <- 'age >= 25 & age <= 54 & relation <= 2200'
datasets <- c('us04', 'be04', 'gr04')
df       <- get_stack(datasets, varp, varh, subset)


'Female Employment Rate'
round(with(df[df$sex==2 & !is.na(df$emp),], tapply(emp*ppopwgt, list(dname), sum) / tapply(ppopwgt, list(dname), sum)) *100, digits=2)


'Partime Employment rate among employed women'
round(with(df[df$sex==2 & df$emp==1 & !is.na(df$ptime1),], tapply(ptime1*ppopwgt, list(dname), sum) / tapply(ppopwgt, list(dname), sum)) *100, digits=2)
```

## 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

- You will again need to recode LIS variables, as in exercise 2.1. See that exercise for more information. In this exercises, you will be creating variables **achildcat** and **partner**.

- As in the last exercise, you can use the **tapply** function to calculate weighted proportions over subsets of the data. You can extend this command by adding additional variables to the list of categorical variables, which will allow you to analyze subsets within subsets. For example:

  ```
  with(df[df$sex == 1 & !is.na(df$emp), ], tapply(emp * ppopwgt, list(dname, achildcat, partner), sum) / tapply(ppopwgt , list(dname, achildcat, partner), sum)) * 100
  ```

  This code will produce a tabulation of employment rates among women, which is separated by country, by age of the youngest child in the family, and by partnership status.

## Program

```
get_stack <- function(datasets, varp, varh, subset) {
  # READ DATASETS
  pp <- read.LIS(paste(datasets, 'p', sep = ''), labels = FALSE, vars = varp, subset = subset)
  hh <- read.LIS(paste(datasets, 'h', sep = ''), labels = FALSE, vars = varh)
  df <- merge(pp, hh, by = c("dname", "hid"))
  # MAP NEW VARIABLES


df$achildcat <- ifelse(df$ageyoch < 6, 1, ifelse( df$ageyoch > 5 & df$ageyoch < 18, 2, 0))
df$achildcat [is.na(df$achildcat)] <- 0


  return(df)
}
#---------------------------------------- RUN SCRIPTS ----------------------------------------
datasets <- c('us04', 'be04', 'gr04')
varh     <- c('hid', 'dname', 'own')
varp     <- c('hid', 'dname', 'ppopwgt', 'age', 'sex', 'relation', 'emp', 'ptime1', 'ageyoch', 'partner', 'status1')
subset   <- 'age >= 25 & age <= 54 & relation <= 2200'
df       <- get_stack(datasets, varp, varh, subset)


round(with(df[df$sex == 2 & !is.na(df$emp), ], tapply(emp * ppopwgt, list(dname, achildcat, partner), sum) / tapply(ppopwgt , list(dname, achildcat, partner), sum)) * 100, digits = 2)
```

## 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% |
| GR04 | 57.6% | 68.6% | 64.3% | 81.4% | 49.9% | 57.7% | 60.5% |
| US04 | 72.0% | 80.1% | 66.9% | 78.8% | 76.2% | 58.4% | 71.7% |

*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 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 technique 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**

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

Next, use the LIS hourly wage variable **hwage1** 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: EXP [log Q1 -3*(logQ3 –logQ1)] for the lower boundary and EXP [log Q3 + 3*(logQ3 – logQ1)] 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

- Create a new variable for dependent employment using the recoding techniques from the earlier exercises.

- You can use the **wNtile(var, wgt, split)** function used in the part I of the self-teaching package to calculate the weighted median for the purpose of top-coding.

```
topline <- with(df[!is.na(df$hrwg) & df$dname==datasets[i],], 10 *
wNtile(hrwg, ppopwgt, 0.5))
```

- As in the last exercise, you can use the **tapply** function to calculate the employment rates.

- Use the **wNtile(var, wgt, split)** function to calculate median wages for women and men by country with a for loop technique.

```
for (i in (1:length(datasets))) {
 for (j in 1:length(unique(df$sex))) {
   mat[j, i] <- …
  }
 }
```

- Then allocate the results to a (2LX3C) matrix and calculate the ratio of the values per line:

```
round(mat[2, ] / mat[1, ], digits = 2)
```

## Program

```
get_stack <- function(datasets, varp, varh, subset) {

# READ DATASETS
pp <- read.LIS(paste(datasets,'p',sep=''),labels=FALSE, vars=varp, subset = subset)
hh  <- read.LIS(paste(datasets,'h', sep=''), labels=FALSE, vars=varh)
df  <- merge(pp, hh,  by = c("dname", "hid"))

# MAP NEW VARIABLES
df$sex  <- ifelse(df$sex == 1, 0, 1)
df$dept <- ifelse(df$status1 %in% c(100:120),1,ifelse(is.na(df$status1),NA,0))
df$hrwg <- df$hwage1
df$hrwg <- ifelse(df$hrwg <= 0, NA, df$hrwg)
for (i in 1:length(datasets))  {
topline <- with(df[!is.na(df$hrwg) & df$dname==datasets[i],], 10 * wNtile(hrwg,
ppopwgt, 0.5))
df$hrwg <- with(df[!is.na(df$hrwg) & df$dname==datasets[i],], ifelse(df$hrwg >
topline, topline, df$hrwg))
}
 return(df)
}
wNtile <- function(var, wgt, split) {
 x   <- var[order(var)]
 y   <- wgt[order(var)]
 z   <- cumsum(y) / sum(y)
 cop <- rep(NA,length(split))
 for (i in 1:length(cop))  {
   cop[i] <- x[Find(function(h)  z[h] > split[i], seq_along(z))]
 }
 return(cop)
}
#------------------------------ RUN SCRIPTS ---------------------------------
datasets <- c('us04', 'be04', 'gr04')
varh     <- c('hid', 'did','dname', 'own')
varp     <- c('hid', 'dname', 'ppopwgt', 'age', 'sex', 'relation', 'emp', 'ptime1', 'ageyoch',
'partner', 'status1', 'hwage1')
subset   <- 'age >= 25 & age <= 54 & relation <= 2200'
```

```
df       <- get_stack(datasets, varp, varh, subset)


# EMPLOYMENT RATES
print('Employment Rate')
round(with(df[df$emp == 1 & !is.na(df$dept), ], tapply(dept * ppopwgt, list(dname,
sex), sum) / tapply(ppopwgt, list(dname,sex), sum)) * 100, digits=2)


# GENDER WAGE GAP
mat <- matrix(NA, 2 ,3)
colnames(mat) <- datasets
for (i in (1:length(datasets))) {
  for (j in 1:length(unique(df$sex))) {
    mat[j,i] <- with(df[df$dname==datasets[i] & !is.na(df$hrwg) & df$sex== j-1,] ,
wNtile(hrwg, ppopwgt, 0.5))
  }
}
'Gender Wage Gap'
round(mat[2, ] / mat[1, ], digits = 2)
```

## Results

**Employment Rates**

| | Men | | Women | |
|---|---|---|---|---|
| Dataset | 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 | 0.95 |
| GR04 | 0.85 |
| US04 | 0.76 |

*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

- Because the program so far has been designed to load variables without their labels, you will have to consult METIS for variable information and codebook in order to determine what the country-specific education codes refer to. If you prefer, you could write your program to load the education variables with labels before tabulating, but this may be somewhat redundant and burdensome because of the way the program has been designed so far.

- As in the previous exercises, you can produce tabulations of the wage ratio for each country by using **tapply()** function and loop technics.

## Program

```
get_stack <- function(datasets, varp, varh, subset) {
  # READ DATASETS
  pp      <- read.LIS(paste(datasets, 'p', sep = ''), labels = FALSE, vars = varp, subset = subset)
  hh      <- read.LIS(paste(datasets, 'h', sep = ''), labels = FALSE, vars = varh)
  df      <- merge(pp, hh,  by = c("dname", "hid"))
  # MAP NEW VARIABLES
  df$sex  <- ifelse(df$sex == 1              , 0, 1)
  df$hrwg <- df$hwage1
  df$hrwg <- ifelse(df$hrwg <= 0, NA, df$hrwg)
  for (i in 1:length(datasets)) {
    topline <- with(df[!is.na(df$hrwg) & df$dname == datasets[i], ] , 10 * wNtile(hrwg, ppopwgt, 0.5))
    df$hrwg <- with(df[!is.na(df$hrwg) & df$dname == datasets[i], ] , ifelse(df$hrwg > topline, topline, df$hrwg))
  }
  return(df)
}
wNtile <- function(var, wgt, split) {
  x  <- var[order(var)]
  y  <- wgt[order(var)]
  z  <- cumsum(y) / sum(y)
  cop  <- rep(NA,length(split))
  for (i in 1:length(cop)) {
    cop[i] <- x[Find(function(h) z[h] > split[i], seq_along(z))]
  }
  return(cop)
}
#--------------------------------- RUN SCRIPTS --------------------------------------
datasets <- c('us04', 'be04', 'gr04')
varh     <- c('hid', 'dname', 'own')
varp     <- c('hid', 'dname', 'ppopwgt', 'age', 'sex', 'relation', 'emp', 'ptime1', 'ageyoch', 'partner', 'status1', 'hwage1', 'educ')
subset   <- 'age >= 25 & age <= 54 & relation <= 2200'
df       <- get_stack(datasets, varp, varh, subset)
```

```
# GENDER WAGE GAP
ctry_list <- list()
for (k in (1:length(datasets))) {
mat <-
matrix(NA,length(unique(df$sex[!is.na(df$sex)])),length(unique(df$educ[!is.na(df$educ)
])))
  colnames(mat) <- c('Low', 'Medium', 'high')
  for (j in (1:length(unique(df$educ[!is.na(df$educ)])))) {
    for (i in (1:length(unique(df$sex[!is.na(df$sex)])))) {
mat[i,j] <- with(df[df$dname == datasets[k] & !is.na(df$hrwg) & !is.na(df$sex) &
!is.na(df$educ) & df$sex == i - 1 & df$educ == j, ], wNtile(hrwg, ppopwgt, 0.5))
    }
  }
  ctry_list[[datasets[k]]] <- round(mat[2, ] / mat[1, ], digits = 2)
}

'Gender Wage Gap'
ctry_list
```

## Results

**Gender wage gaps by educational attainment**

| | Low education | Medium education | High education |
|---|---|---|---|
| BE04 | 0.84 | 0.89 | 0.89 |
| GR04 | 0.66 | 0.78 | 0.96 |
| US04 | 0.74 | 0.73 | 0.79 |

*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 the one used in the previous exercise.

## Program

```
get_stack <- function(datasets, varp, varh, subset) {
# READ DATASETS
pp <- read.LIS(paste(datasets, 'p', sep = ''), labels=FALSE, vars=varp, subset=subset)
hh <- read.LIS(paste(datasets, 'h', sep = ''), labels = FALSE, vars = varh)
df <- merge(pp, hh, by = c("dname", "hid"))

# MAP NEW VARIABLES
df$sex <- ifelse(df$sex == 1, 0, 1)
df$dept <- ifelse(df$status1 %in% c(100:120), 1, ifelse(is.na(df$status1) , NA, 0))
df$hrwg <- df$hwage1
df$hrwg <- ifelse(df$hrwg <= 0, NA, df$hrwg)
for (i in 1:length(datasets)) {
topline <- with(df[!is.na(df$hrwg) & df$dname==datasets[i],], 10*wNtile(hrwg,
ppopwgt,0.5))
df$hrwg <- with(df[!is.na(df$hrwg) & df$dname==datasets[i],], ifelse(df$hrwg >
topline, topline, df$hrwg))
}
 return(df)
}
wNtile <- function(var, wgt, split) {
  x <- var[order(var)]
  y <- wgt[order(var)]
  z <- cumsum(y) / sum(y)
  cop <- rep(NA,length(split))
  for (i in 1:length(cop)) {
    cop[i] <- x[Find(function(h) z[h] > split[i], seq_along(z))]
  }
  return(cop)
}
```

```
#------------------------------------ RUN SCRIPTS ---------------------------------
datasets <- c('us04', 'be04', 'gr04')
varh     <- c('hid', 'dname', 'own')
varp     <- c('hid', 'dname', 'ppopwgt', 'age', 'sex', 'relation', 'emp', 'ptime1', 'ageyoch',
'partner', 'status1', 'hwage1', 'educ', 'immigr')
subset   <- 'age >= 25 & age <= 54 & relation <= 2200'
df       <- get_stack(datasets, varp, varh, subset)

# GENDER WAGE GAP
ctry_list <- list()
 for (k in (1:length(datasets))) {
   mat <- matrix(NA, length(unique(df$sex[!is.na(df$sex)])),
   length(unique(df$immigr[!is.na(df$immigr)])))
   colnames(mat) <- c('Non Immigrant', 'Immigrant')
   for (j in (1:length(unique(df$immigr[!is.na(df$immigr)])))) {
     for (i in (1:length(unique(df$sex[!is.na(df$sex)])))) {
       mat[i,j] <- with(df[df$dname == datasets[k] & !is.na(df$hrwg) & !is.na(df$sex)
       & !is.na(df$immigr) & df$sex== i-1 & df$immigr== j-1,],
       wNtile(hrwg,ppopwgt,0.5))
     }
   }
   ctry_list[[datasets[k]]] <- round(mat[2, ] / mat[1, ], digits = 2)
 }

 'Gender Wage Gap'
 ctry_list
```

## Results

**Gender earnings ratios by immigration status**

| | Non-immigrants | Immigrants |
|---|---|---|
| BE04 | 0.94 | 1.02 |
| GR04 | 0.89 | 0.82 |
| US04 | 0.75 | 0.83 |

*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

If you have followed the instructions up to this point, you should not need to create any additional variables to run regression models predicting **logwage**.

You should run the regressions separately in each country. In addition, within each country you should run a separate model for men and women. Please use the following model:

**logwage = f (age agesq meduc heduc immigr partn ychild ochild ptime1 homeowner)**

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

*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 R is done with the **glm()** function. Since this

command accepts weights, you can use it instead of the **svyglm()** function from the **survey** package, which is necessary for more complex sample designs.

- R will automatically detect categorical variables if they are coded as factors, meaning that you do not need to manually create dummy variables. You can also include mathematical operations directly in the definition of your regression formula if you enclose them in the **I()** function, so you do not need to manually create log wages or age squared. For example, the following estimates the log wage regression for men in the United States:

```
glm(I(log(hrwg))~age+I(age^2)+educ+immigr+partner+achildcat+ptime1
+homeowner, data=df, weights=df$ppopwgt, subset=sex=="Male" &
dname=="us04")
```

- When performing several regression models in a single program, one strategy is to again write a loop that estimates each regression in turn, and then prints a summary of the model

```
model <- formula(I(log(hrwg))~age +
I(age^2)+meduc+heduc+immigr+partn+ychild+ochild+ptime1+homeowne
r)
for(s in …) for(d in datasets) {
  res <- glm(model, data = df, weight = df$ppopweight, subset = sex &
dname = d)
  print(summary(res))
}
```

Here the definition of the model has been done before calling the loop, which makes the call to **glm()** itself somewhat easier to read. The summary function will output coefficients, standard errors and R-squared statistics.

## Program

```
get_stack <- function(datasets, varp, varh, subset) {
# READ DATASETS
  pp <- read.LIS(paste(datasets, 'p', sep=''), labels=FALSE, vars=varp, subset=subset)
  hh <- read.LIS(paste(datasets, 'h', sep = ''), labels = FALSE, vars = varh)
  df <- merge(pp, hh, by = c("dname", "hid"))


# MAP NEW VARIABLES
df$homeowner <- ifelse(df$own %in% 100:199, 1, ifelse(df$own %in% 200:299, 0,
NA))
df$achildcat <- ifelse(df$ageyoch < 6, 1, ifelse( df$ageyoch > 5 & df$ageyoch < 18, 2,
0))
df$achildcat [is.na(df$achildcat)] <- 0
df$ychild <- ifelse(df$achildcat==1, 1, ifelse(is.na(df$achildcat), NA, 0))
df$ochild <- ifelse(df$achildcat==2, 1, ifelse(is.na(df$achildcat), NA, 0))
df$meduc <- ifelse(df$educ==2 , 1, ifelse(is.na(df$educ), NA, 0))
df$heduc <- ifelse(df$educ==3 , 1, ifelse(is.na(df$educ), NA, 0))
df$hrwg <- df$hwage1
df$hrwg <- ifelse(df$hrwg <= 0, NA, df$hrwg)
for (i in 1:length(datasets)) {
 topline <- with(df[!is.na(df$hrwg) & df$dname==datasets[i],],
 10*wNtile(hrwg,ppopwgt,0.5))
 df$hrwg <- with(df[!is.na(df$hrwg) & df$dname==datasets[i],],
 ifelse(df$hrwg>topline, topline, df$hrwg))
}
  return(df)
}
wNtile <- function(var, wgt, split) {
  x <- var[order(var)]
  y <- wgt[order(var)]
  z <- cumsum(y) / sum(y)
  cop <- rep(NA,length(split))
  for (i in 1:length(cop)) {
    cop[i] <- x[Find(function(h) z[h] > split[i], seq_along(z))]
  }
  return(cop)
```

```
}
#---------------------------------------- RUN SCRIPTS -----------------------------
options(scipen  = 2)
datasets <- c('us04', 'be04', 'gr04')
varh     <- c('hid', 'dname', 'own')
varp     <- c('hid', 'dname', 'ppopwgt', 'age', 'sex', 'relation', 'emp', 'ptime1',
             'ageyoch', 'partner', 'status1', 'hwage1', 'educ', 'immigr')
subset   <- 'age >= 25 & age <= 54 & relation  <= 2200'
df       <- get_stack(datasets,  varp, varh, subset)

# REGRESSION  MODEL
gender <- c('Male', 'Female')
model <- formula(I(log(hrwg))~age  + I(age^2) +
meduc+heduc+immigr+partner+ychild+ochild+ptime1+homeowner)
for(s in  (1:length(unique(df$sex[!is.na(df$sex)])))  for(d in  datasets) {
  res <- glm(model,  data = df, weights  = df$ppopwgt, subset = sex == s & dname ==
d)
  print('------------------------------------------------------------------------')
  print(paste(gender[s],  d, sep = " : "))
  print(summary(res),digits=1)
 }
```

## Results

```
 [1] "------------------------------------------------------------------------"
[1] "Male : us04"

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.15350    0.08477      14   <2e-16 ***
age          0.04864    0.00438      11   <2e-16 ***
I(age^2)    -0.00049    0.00005      -9   <2e-16 ***
meduc        0.27095    0.01244      22   <2e-16 ***
heduc        0.64845    0.01255      52   <2e-16 ***
immigr      -0.06106    0.00927      -7    5e-11 ***
partner      0.04137    0.00980       4    2e-05 ***
ychild       0.05711    0.00985       6    7e-09 ***
ochild       0.06468    0.00885       7    3e-13 ***
ptime1      -0.31086    0.01811     -17   <2e-16 ***
homeowner    0.23740    0.00856      28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Null deviance: 17420924  on 30019  degrees of freedom

[1] "------------------------------------------------------------------------"
[1] "Male : be04"

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.71803    0.21033     8.2    6e-16 ***
age          0.01757    0.01078     1.6     0.10
I(age^2)    -0.00005    0.00013    -0.4     0.70
meduc        0.15079    0.02040     7.4    2e-13 ***
heduc        0.43824    0.02059    21.3   <2e-16 ***
immigr      -0.05663    0.02376    -2.4     0.02 *
partner      0.02774    0.02092     1.3     0.19
ychild       0.02794    0.02274     1.2     0.22
ochild       0.01135    0.02039     0.6     0.58
ptime1       0.07681    0.03216     2.4     0.02 *
```

```
homeowner    0.09860    0.01894     5.2    2e-07 ***---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Null deviance: 173862  on 1590  degrees of freedom



[1] "------------------------------------------------------------------------"
[1] "Male : gr04"

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)   0.7955     0.3069     2.6     0.01 **

age           0.0338     0.0158     2.1     0.03 *

I(age^2)     -0.0002     0.0002    -1.3     0.21

meduc         0.1665     0.0249     6.7    3e-11 ***

heduc         0.4295     0.0277    15.5   <2e-16 ***

immigr       -0.2473     0.0321    -7.7    3e-14 ***

partner      -0.0126     0.0366    -0.3     0.73

ychild        0.1172     0.0322     3.6    3e-04 ***

ochild        0.1125     0.0292     3.9    1e-04 ***

ptime1        0.3249     0.0502     6.5    1e-10 ***

homeowner     0.0355     0.0230     1.5     0.12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Null deviance: 179062  on 1175  degrees of freedom



[1] "------------------------------------------------------------------------"
[1] "Female : us04"

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  1.15011    0.08814      13   <2e-16 ***

age          0.04106    0.00457       9   <2e-16 ***

I(age^2)    -0.00043    0.00006      -8    5e-14 ***

meduc        0.31098    0.01496      21   <2e-16 ***

heduc        0.73557    0.01502      49   <2e-16 ***

immigr      -0.02597    0.01038      -2     0.01 *
```

```
partner      -0.00831     0.00836       -1      0.32
ychild        0.03534     0.01072        3     1e-03 ***
ochild       -0.04822     0.00842       -6     1e-08 ***
ptime1       -0.18379     0.00918      -20     <2e-16 ***
homeowner     0.17773     0.00901       20     <2e-16 ***---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null deviance: 16262501  on 28653  degrees of freedom


[1] "------------------------------------------------------------------------"
[1] "Female : be04"

Coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.1673     0.2281     5.1     3e-07 ***
age            0.0459     0.0119     3.9     1e-04 ***
I(age^2)      -0.0005     0.0001    -3.0     0.002 **
meduc          0.1806     0.0260     6.9     6e-12 ***
heduc          0.4865     0.0253    19.2     <2e-16 ***
immigr        -0.0394     0.0278    -1.4     0.157
partner        0.0033     0.0220     0.1     0.882
ychild        -0.0062     0.0254    -0.2     0.806
ochild        -0.0506     0.0217    -2.3     0.020 *
ptime1         0.0117     0.0175     0.7     0.504
homeowner      0.0825     0.0214     3.9     1e-04 ***---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null deviance: 172900  on 1491  degrees of freedom

[1] "------------------------------------------------------------------------"
[1] "Female : gr04"

Coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2666     0.3403    -0.8     0.433
age            0.0710     0.0180     4.0     8e-05 ***
I(age^2)      -0.0007     0.0002    -2.9     0.004 **
```

```
meduc         0.3483     0.0336    10.4   <2e-16 ***
heduc         0.7634     0.0350    21.8   <2e-16 ***
immigr       -0.2595     0.0408    -6.4    3e-10 ***
partner      -0.0649     0.0350    -1.9    0.064 .
ychild        0.0549     0.0379     1.4    0.148
ochild        0.0225     0.0330     0.7    0.496
ptime1        0.2259     0.0310     7.3    6e-13 ***
homeowner    -0.0080     0.0274    -0.3    0.769    ---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null deviance: 205080  on 966  degrees of freedom
```

*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

Adjust wages for Purchasing Power Parities by dividing **hrwage** by the following deflators before taking the natural logarithm (see https://www.lisdatacenter.org/resources/ppp-deflators/).

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

Estimate the following model, for men and women separately:

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

The model is the same as in the last exercise, except that it includes an indicator for country. This time, however, 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, so you may not want to use a loop.

## Program

```
get_stack <- function(datasets, varp, varh, subset) {
  # READ DATASETS
  pp      <- read.LIS(paste(datasets, 'p', sep = ''), labels = FALSE, vars = varp, subset = subset)
  hh      <- read.LIS(paste(datasets, 'h', sep = ''), labels = FALSE, vars = varh)
  df      <- merge(pp, hh,  by = c("dname", "hid"))
  # MAP NEW VARIABLES
  df$homeowner <- ifelse(df$own %in% 100:199, 1, ifelse(df$own %in% 200:299, 0, NA))
  df$achildcat <- ifelse(df$ageyoch < 6, 1, ifelse( df$ageyoch > 5 & df$ageyoch < 18, 2, 0))
  df$achildcat [is.na(df$achildcat)] <- 0
  df$ychild    <- ifelse(df$achildcat == 1, 1, ifelse(is.na(df$achildcat), NA, 0))
  df$ochild    <- ifelse(df$achildcat == 2, 1, ifelse(is.na(df$achildcat), NA, 0))
  df$meduc     <- ifelse(df$educ == 2, 1, ifelse(is.na(df$educ), NA, 0))
  df$heduc     <- ifelse(df$educ == 3, 1, ifelse(is.na(df$educ), NA, 0))
  df$ppp       <- ifelse(df$dname == 'be04', 0.79, ifelse(df$dname == 'gr04', 0.65, 1))
  df$hrwg      <- df$hwage1
  df$hrwg      <- ifelse(df$hrwg <= 0, NA, df$hrwg)
  df$Belgium   <- ifelse (df$dname == 'be04', 1, 0)
  df$greece    <- ifelse (df$dname == 'gr04', 1, 0)
  for (i in 1:length(datasets)) {
    df$hrwg <- df$hrwg / df$ppp
    topline <- with(df[!is.na(df$hrwg) & df$dname==datasets[i],], 10 * wNtile(hrwg,ppopwgt,0.5))
    df$hrwg <- with(df[!is.na(df$hrwg) & df$dname==datasets[i],],
              ifelse(df$hrwg>topline,topline,df$hrwg))
  }
  return(df)
}
wNtile <- function(var, wgt, split) {
  x  <- var[order(var)]
  y  <- wgt[order(var)]
  z  <- cumsum(y) / sum(y)
  cop  <- rep(NA,length(split))
  for (i in 1:length(cop)) {
```

```
    cop[i] <- x[Find(function(h) z[h] > split[i], seq_along(z))]
  }
  return(cop)
}
#---------------------------------------- RUN SCRIPTS ----------------------------------
options(scipen = 2)
datasets <- c('us04', 'be04', 'gr04')
varh <- c('hid', 'dname', 'own')
varp <- c('hid','dname','pwgt','ppopwgt','age','sex','relation','emp','ptime1','ageyoch',
          'partner','status1','hwage1','educ','immigr')
subset   <- 'age >= 25 & age <= 54 & relation <= 2200'
df       <- get_stack(datasets, varp, varh, subset)

# REGRESSION MODEL
gender <- c('Male', 'Female')
model <- formula(I(log(hrwg))~age + I(age^2) + meduc + heduc + immigr + partner
+ ychild + ochild + ptime1 + homeowner + Belgium + greece)
for(s in (1:length(unique(df$sex[!is.na(df$sex)])))) {
  res <- glm(model, data = df, weights = df$pwgt, subset = sex == s)
  print('--------------------------------------------------------------------------')
  print(gender[s])
  print(summary(res),digits=1)
}
```

## Results

```
[1] "------------------------------------------------------------------------"
[1] "Male"
Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept)  1.46997    0.06953    21.1   <2e-16 ***
age          0.03857    0.00359    10.7   <2e-16 ***
I(age^2)    -0.00032    0.00004    -7.2    8e-13 ***
meduc        0.16900    0.00730    23.1   <2e-16 ***
heduc        0.51452    0.00751    68.5   <2e-16 ***
immigr      -0.11006    0.00759   -14.5   <2e-16 ***
partner      0.03503    0.00766     4.6    5e-06 ***
ychild       0.06267    0.00772     8.1    5e-16 ***
ochild       0.05178    0.00696     7.4    1e-13 ***
ptime1       0.01129    0.01225     0.9      0.4
homeowner    0.14108    0.00629    22.4   <2e-16 ***
belgium      0.47829    0.00603    79.4   <2e-16 ***
greece       0.45709    0.00683    66.9   <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null deviance: 1160.88  on 32786  degrees of freedom

[1] "------------------------------------------------------------------------"
[1] "Female"

Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept)  0.90550    0.07093    12.8   <2e-16 ***
age          0.05512    0.00371    14.8   <2e-16 ***
I(age^2)    -0.00056    0.00005   -11.9   <2e-16 ***
meduc        0.25107    0.00882    28.5   <2e-16 ***
heduc        0.64611    0.00878    73.6   <2e-16 ***
immigr      -0.07830    0.00846    -9.3   <2e-16 ***
partner     -0.00335    0.00686    -0.5     0.63
ychild       0.01500    0.00819     1.8     0.07 .
ochild      -0.04345    0.00681    -6.4    2e-10 ***
ptime1       0.00046    0.00633     0.1     0.94
homeowner    0.09740    0.00666    14.6   <2e-16 ***
belgium      0.65724    0.00635   103.6   <2e-16 ***
greece       0.61045    0.00729    83.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null deviance: 1227.88  on 31112  degrees of freedom
```

*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.

# Additional guideline on saving your temporary data files at the LIS directory

When working on your datasets in R, you may need to save some data files that you created from LIS/LWS files for a latter (re-)analysis. You can save these data files via LISSY in the LIS directory (see the code below). In order to ensure that your data file(s) are not overwritten by other users, you should save your file(s) with a unique name (e.g. you could include your LIS username in the filename). You can also ask the LIS User Support to create a specific folder in the LIS directory where you will be able to save your data files (created from LIS/LWS data) via LISSY.

## Program

```
require(foreign)
require(readstata13)
#read LIS data
data <- read.LIS('lu10h')
#save the above data in your folder that LIS created for you*
save.dta13(data,paste(USR_DIR,"/name_of_your_folder/name_of_your_data_file_that_you_want
_to_save.dta", sep = ""), version = 117,convert.factors=FALSE)
#open the above data that you've just saved
data1 <-read.dta13(paste(USR_DIR,
"/name_of_your_folder/name_of_your_data_file_that_you_saved_in_your_folder.dta",
sep=""),convert.factors=FALSE)
```

Note: in order to have your specific folder, please write to the LIS user support at usersupport@lisdatacenter.org .