Unit 4: Study Guide
An Introduction to combining macro and micro data

MIMAS
The University of Manchester

4.1. Introduction

This unit will discuss the kinds of research questions that require linked micro and macro data, methods of assessing micro data for variation between macro units prior to linking the data, how to link micro and macro data, weighting linked micro and macro data for particular kinds of comparative analysis, checking the quality of micro data with macro data and adjusting weights to account for discrepancies, and basic approaches to the analysis of linked micro and macro data.

Most of the activities in this unit are based on the analysis of a subset of the European Social Survey which can be found at http://www.mimas.ac.uk/limmd/. The data are in Stata format and the activities use Stata code.

Those who have access to Stata but have never used the programme before should find that for most of the activities it is simple to issue the commands in the Command window and view the results. For those without access to either Stata or the data (or simply those without the time to do the activities), even if you are not familiar with Stata, you should find that the Stata command language is sufficiently straightforward that it is possible to understand the commands just by inspection and comparison with the example output. Similarly, those who use other statistical analysis software should find it relatively easy to identify how to change the Stata code into some other language that they are familiar with. For instance, SPSS users should notice that the Stata command 'tabulate' is similar to the `crosstab' command in SPSS. The main purpose of the activities in this unit is not to provide instruction on the use of Stata, but to add clarity and aid learning of the statistical analysis and data manipulation procedures with practical examples.

5.2. Learning objectives

By the end of this unit you will be able to:

- understand the potential for combining macro and micro data to solve specific research questions.
- test for variation between macro units within an micro-level dataset
- link a macro level data set to a micro level data set using statistical software
- understand the different roles of certain weighting schemes for linked micro-macro data
- produce a weight for linked micro and macro data set, so that each macro unit contributes equally to the analysis
- produce a weight for a linked micro and macro data set, so that the analysis reflects the population as a whole
- understand that there are a variety of reasons why the frequency distribution of an individual-level survey data may not match macro-level data for the same macro unit
- adjust a weight variable to account for differences between the distribution of a key variable in the survey from the known population distribution
- adjust a weight variable to account for differences, within each macro unit, between the distribution of a key variable in the survey from the known population distribution
- test a hypothesis that outcomes on a particular micro-level variable depend on both micro and macro variables
- test a hypothesis that the effect of one micro-level variable on another depends on the value of a macro-level variable

4.3 Identifying whether it will help to link macro data to micro data

Two things need to be true for you to want to link macro and micro data: you need to have a research question that requires linked data and you need to have variation at the micro level between macro units.

This section will discuss some of the basic kinds of research question that require linked macro and micro data. Since different kinds of question regard different kinds of variation between macro units, the next subsection will discuss different analyses that can be used identify these different kinds of between macro unit variation, before the macro data has even been linked to the micro data. It is worthwhile spending time considering the analyses in this section, since it is a good idea to understand some of the main characteristics of the variation between macro units in the micro data before you try to use macro level variables to explain that variation. This may also help alert you to the presence of any unusual macro units or any problems with your data.

4.3.1 Research questions that required linked macro and micro data

If your research question requires you to assess whether the variation in some individual-level factor depends on some macro-level variable, then you will need to linked macro and micro data to answer that question. The following are examples of such questions.

Does the level of turnout in an election depend on the electoral system in use?

If controls for other relevant factors are to be included in the analysis, this question requires individual level information on turnout, and macro level data on the electoral system in use in different countries.

Does the level of trust in politicians depend on the frequency of coalition governments?

This question requires individual level information on trust in politicians, and macro (country) level information on the frequency of coalition government.

Does the level of turnout in British general elections depend on the marginality of the constituency?
Again, if controls for other relevant factors are to be included in the analysis, this question requires both individual level information on turnout, and macro level information on the marginality of the constituency.

There are some questions that require both micro and macro data, but the macro data can be generated from the micro data, and so it isn't clear that you are required to link two different sources of micro and macro data, when you might instead be able to generate the macro data using the micro data that you have already. For example, suppose you are interested in whether the level of trust in politicians varies between countries according to the level of interpersonal trust. It could be that you have two separate data sources, one at the individual level showing the level of trust in politicians, and one at the macro level providing the level of interpersonal trust in each country. But since the level of interpersonal trust is simply the aggregation of an individual level survey question to the country level, it maybe possible to generate the country level of interpersonal trust from the individual level survey data if the relevant question is included in the same survey as the 'trust in politicians' question.

Sometimes cross-national aggregate statistics from micro data are already available, e.g. for the **Eurobarometer**, **Latinobarometer** and some World Bank datasets. You might find it easier to link the pre-prepared aggregates than to write code to generate them yourself.

In cases such as these, you need to consider whether there is better quality macro data available from sources separate from the micro data source to make it worthwhile linking the macro data to the micro data, rather than generating macro data from the micro data. For instance, suppose you are interested in whether there are neighbourhood (or contextual) effects in operation so that people are more likely to vote Labour in more working class constituencies. Here there is a choice between generating a measure of the class composition of the constituency using micro level information on the class membership of the respondents to the survey (e.g. the **British Election Study**) and aggregating it to the macro (constituency) level, or linking the census data on the class composition of constituencies to the micro

(survey) data. In practice it is clear that the census data is of much higher quality than macro data generated from any survey that could be used to analyse voting behaviour, since in any suitable survey there are only a small number of respondents in each constituency.

In addition to assessing whether the level of some particular micro level outcome vary between macro units according to characteristics of the macro units, it may be that patterns of association between two or more micro level variables might differ between macro units because of some macro level factors. Such effects are sometimes described as cross-level interactions, because the effect of one micro variable on another depends on the level of a macro variable. Research questions which posit these kinds of effects also require linkage between micro and macro data. The following are examples.

Does the level of inequality increase the political polarization of social classes?
The political polarization for a given macro unit is measured by the association between two individual-level variables (social class and vote choice). This question asks whether that association is stronger in macro units where inequality (a macro-level variable, often measured with the Gini coefficient) is higher.

Is the effect of political knowledge on turnout stronger in countries with first-past-the-post electoral systems?
This requires both micro survey data on political knowledge and turnout to be linked to macro (country) level data on the electoral system.

4.3.2 Identifying whether there is variation in the micro data between macro units
Given that the research questions and hypotheses which require linked macro and micro data are ones that postulate variation between the macro units either in the distributions of some of the macro variables or in the patterns of association between micro variables, it is frequently sensible to ask whether or not there is variation of this kind, before proceeding to link micro and macro data. This section discusses how this can be done. The main aim of these analyses is to identify the micro level variation that may be explicable with macro level factors. However, you should not conclude that if there is little between macro unit variation, there is no need to

proceed to linking the macro data. It is possible in multivariate analyses that different processes cancel each other out, but the processes may be visible with careful modelling of the linked macro and micro data.

4.3.2.1 Identifying whether there is variation between macro units in the distribution of a micro variable

There are three methods for testing for variation in an individual level variable between macro units that increase in complexity. The first is to compare the distribution of the micro level variable across macro units. With ratio or interval level measures, this typically takes the form of comparing the means for each macro unit.

Activity 1 : Comparing the level of trust in politicians between different countries using Stata

You will need to go online to complete this activity. This exercise will use a subset of the European Social Survey(ESS) data which can be opened within Stata with the command:
http://www.mimas.ac.uk/limmd/materials/LIMMD-unit4/data/ESSsubset.dta

You will need access to the statistical software programme Stata to complete this task.

Suppose that you wished to compare the **level of trust in politicians** (`trstplt`) between different countries. In Stata the relevant command would be.

```
bysort country: summarize trstplt
```

This will list the means and standard deviations of the `trstplt` for each value of the variable '`country`'. The first few lines of the output should look like the following.

```
_____________________________________________________________________________
-----------------------------------------------------------------------------
-> country = AT

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     trstplt |       295    3.386441    2.351686          0         10

-----------------------------------------------------------------------------
-> country = BE

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     trstplt |       297    4.124579     2.15951          0          8

-----------------------------------------------------------------------------
-> country = CH

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     trstplt |       292    4.732877    1.992395          0          9
```

If you are interested in a categorical variable then it would usually be more appropriate to inspect the frequency distribution of the micro variable separately for each macro unit.

Activity 2: Comparing electoral turnout across different countries using Stata

You will need to go online to complete this activity. This exercise will use a subset of the European Social Survey (ESS) data. You will need access to the statistical software programme Stata to complete this task, and you may need your Athens username and password to access the ESS. Suppose that you wished to compare electoral turnout (where 1 = voted, 0 = did not vote) across different countries, in Stata the relevant command would be.

```
bysort country: tabulate turnout
```

This will list the percentages who voted and did not vote for each value of the variable '`country`'. The first few lines of the output should be as follows.

```
-------------------------------------------------------------------------------
-> country = AT

    turnout |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         50       18.38       18.38
          1 |        222       81.62      100.00
------------+-----------------------------------
      Total |        272      100.00

-------------------------------------------------------------------------------
-> country = BE

    turnout |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         30       10.91       10.91
          1 |        245       89.09      100.00
------------+-----------------------------------
      Total |        275      100.00

-------------------------------------------------------------------------------
-> country = CH

    turnout |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         77       30.31       30.31
          1 |        177       69.69      100.00
------------+-----------------------------------
      Total |        254      100.00
```

While direct inspection of the distribution of a micro variable separately for each macro unit may be appropriate with a relatively small number of macro units, it might be tedious and unhelpful with a large number.

A second method helps in situations where there are large number of macro units, and it is known as analysis of variance (or ANOVA). This technique can be used for ratio or interval level variables, and it can be used to calculate how much of the variance in a micro variable is due to variation between macro units and how much is variance within macro units.

Activity 3: Analysis of trust in politicians between different countries

You will need to go online to complete this activity. This exercise will use a subset of the European Social Survey (ESS) data. You will need access to the statistical software programme Stata to complete this task, and you may need your Athens username and password to access the ESS. Returning to the analysis of trust in politicians (`trstplt`) between different countries, in Stata the relevant command would be.

```
anova trstplt country
```

```
                           Number of obs =      6165     R-squared     =  0.1327
                           Root MSE      =  2.15594     Adj R-squared =  0.1299

      Source |  Partial SS    df       MS           F     Prob > F
-------------+----------------------------------------------------
       Model |  4369.97651    20  218.498826      47.01     0.0000
             |
     country |  4369.97651    20  218.498826      47.01     0.0000
             |
    Residual |   28557.872  6144  4.64809114
-------------+----------------------------------------------------
       Total |  32927.8485  6164  5.34196115
```

The final figure in the row labelled 'country' shows the p-value for the test for significant variation between countries in the level of trust in politicians. Since this is very small you can easily reject the hypothesis that there is no between country variation.

Not only can ANOVA give the researcher a good idea of how much variation there is at the macro level (both absolutely and relative to that at the micro level), but it will calculate whether the macro level variation is statistically significant, i.e. whether it is sufficiently large that you can be confident it didn't simply arise by chance assuming there are no differences in the means of the micro variable for each macro unit in the real world.
ANOVA cannot be used with categorical dependent variables.

A third method for assessing the between macro unit variation for a micro variables is to fit a multilevel model with no explanatory variables but random effects for the macro units. This technique can be extended for use with categorical as well as interval and ratio level micro variables. Multilevel modelling is covered in Unit 6 of this series.

4.3.2.2 Identifying whether there is variation between macro units in the pattern of association between micro variables

If our research question postulates that the correlation between two micro variables might be different in different kinds of macro unit, before linking the micro to the macro data, you can see whether there is any variation between macro units in the correlation between to variables.

Activity 4: Is the correlation between trust in politicians and generalized trust in other citizens is stronger in some countries than others

You will need to go online to complete this activity. This exercise will use a subset of the European Social Survey (ESS) data. You will need access to the statistical software programme Stata to complete this task, and you may need your Athens username and password to access the ESS.

Suppose that you are interested in whether the correlation between trust in politicians (`trstplt`) and generalized trust in other citizens (`ppltrst`) is stronger in some countries than others, in Stata the relevant command would be.

```
bysort country: pwcorr trstplt ppltrst, star(5)

-------------------------------------------------------------------------------
-> country = AT

             |  trstplt  ppltrst
-------------+------------------
     trstplt |   1.0000
     ppltrst |   0.1572*  1.0000

-------------------------------------------------------------------------------
-> country = BE

             |  trstplt  ppltrst
-------------+------------------
     trstplt |   1.0000
     ppltrst |   0.3635*  1.0000

-------------------------------------------------------------------------------
-> country = CH

             |  trstplt  ppltrst
-------------+------------------
     trstplt |   1.0000
     ppltrst |   0.2297*  1.0000
```

This will list the correlation coefficient between `trstplt` and `ppltrst` for each value of the variable '`country`', and place an asterix by those that are statistically significant at the 5% level.

When one or both of the variables of interest is categorical there are a variety of different strategies for assessing the pattern of association. If both are categorical, the most basic is to consider a cross-tabulation of the two variables. This could then

be produced separately for each macro unit. However, comparing a series of tables is often awkward and so researchers frequently use summary statistics that measure the strength of certain aspects of the pattern of association. If both of the variables are binary, then a logistic regression analysis can be used to calculate the odds ratio for the table. If separate bivariate logistic regression analyses are run for each macro unit it is possible to compare the odds ratios and their confidence intervals. This approach can also be used if there is an interval or ratio level explanatory variable, and a binary dependent variable.

As with inspection of the macro unit means or frequency distributions of micro variables, assessing large numbers of correlation coefficients or odds ratios becomes increasingly difficult as the number of macro units grows. Anova does not provide a solution in this case, however, there are multilevel models, known as random-coefficient models that allow the researcher to assess the extent to which a linear or logistic regression coefficient varies between macro units. These models are discussed in depth in Unit 6 of this series.

4.4. Linking macro and micro data

4.4.1. Preparing the data sets

In order to link a data set with macro data to one with micro data, both data sets must have a variable which is a macro unit identifier. This identifier must be coded in exactly the same way in both data sets and have the same variable name, say `macrounit`. Some statistical programmes, such as Stata, require that both data sets are sorted by the macro-unit identifier before linkage is attempted.

5.4.1.1 Creating a macro unit identifier to match on

It is quite common to find that you have macro unit identifiers in both your macro and micro data sets but they aren't coded in quite the same way. There are various strategies for dealing with this situation. If they are both numeric, you will probably have to recode one so that it is consistent with another, and it needs to be coded in the same way in both data sets, but it doesn't matter how this is achieved.. If they are both alphanumeric or 'string' variables, you will have to edit one so that it is consistent with the style of the other like the other. It will normally be easier for you to edit the macro data set since has only one row per macro unit. If one dataset has an alphanumeric macro unit identifier, and the other has a numeric identifier with value labels, it is usually best to edit the value labels so that they are consistent with the alphanumeric coding and then convert the numerical variable into an alphanumeric one using the former value labels. In Stata this is done using the decode command.

4.4.2. Linking data in Stata

Having prepared the data, linking is usually straightforward. All you need to be clear about is which files you are linking and which variable is the macro identifier. In Stata linking two datasets is done with the merge command. Suppose you wish to merge two files `microdata.dta` and `macrodata.dta`, both including the `macrounit` identifier but with no other variable in common, then you can use the following steps.

1. Save both files in the same folder.
2. Open the macro data and sort it by `macrounit` and save it with the commands

```
sort macrounit
save macrodata, replace
```

3. Open the micro data and sort it by `macrounit`

```
use microdata
sort macrounit
```

4. Use the following command to link/merge the two files.

```
merge macrounit using macrodata
```

In fact these commands work just as well if you swap the order in which you work on the files, i.e. replace 'macrodata' with 'microdata' and vice versa in the commands.

If there are other variables in common between the two files other than `macrounit`, you can still merge the files but you will need to specify with the options to the merge command (see the Stata Help system or manuals) whether you want values of the variables to be updated (in the case of missing data in one file) or replaced (in the case of having one more authoritative file) or left unchanged.

Activity 5: Linking exercise using Stata

The following is an example of linking macro and micro data in Stata. You will need to down load the following .dta file on to your computer:

natturnout.dta

The following commands are an example of linking macro and micro data in Stata. The macro file, nattturnout.dta, includes ESS survey-estimates of national-level turnout but they aren't accurate or official (this is just an exercise!).

1. From within Stata open the data file **natturnout.dta**, use the command:
   http://www.mimas.ac.uk/limmd/materials/LIMMD-unit5/data/natturnout.dta
2. Have a look at the data:
   list
3. Now, sort the data by the macro-unit identifier '**country**':
   sort country
4. To save the macro data in the route directory, use the following command.
   save natturnout
   Note that if you don't have write access to your current route directory, you can either change that directory using standard MS DOS commands, or specify the complete path in the save command.
5. Open the micro data **ESSsubset.dta**, use the command:
   http://www.mimas.ac.uk/limmd/materials/LIMMD-unit5/data/ESSsubset.dta
   and sort it by the macro-unit identifier:
   sort country
6. Now, you can merge in the macro data. Merge country using natturnout If you saved the natturnout file somewhere other than the route directory, you will have to specify the full path.
7. You should see a new variable **_merge** appear in your variable list
8. Check that this always takes the value 3 to see that the merge has worked properly. tabulate _m
9. If the merge has worked properly you probably want to drop the **_merge** variable with the following command.
   drop _merge
10. You can also browse and explore your data in other ways to see what's happened.

4.4.3 Linking data in SPSS

Having prepared the data, linking is usually straightforward. All you need to be clear about is which files you are linking and which variable is the macro identifier. The process of linking data in SPSS is fairly similar. First sort (using the command "SORT

CASES BY macrounit.") and save each file. You can either use the drop down menus or use the following commands in a syntax file.

```
MATCH FILES FILE=microdata.sps
/TABLE=macrodata.sps
/BY macrounit
```

There are also various options you can use if you want the resulting file to include only particular variables.

4.4.4. Checking the linking process has worked

After linking the micro and macro data you should perform some basic checks to your new combined file. These include checking that, you have the right number of micro unit cases; there are is no macro unit data unlinked to micro data, all the variables from each of the macro and micro data sets are present.

Stata creates a variable called _m in the process of linking/merging. This variable takes the value 3 if a row from one data set was successfully merged with a row from another, or 1 if there was no macro data for a given macro unit, or 2 if there was no micro data for a given macro unit. You will probably want to delete cases where there is no micro data with the command:

```
drop if _m ==2
```

The _m variable can take other values if you are using the merge command to update or over-ride variables in the microdata.dta file with those in the macrodata.dta file (see Stata Help system or manuals). It is well worthwhile inspecting this variable to check the merging has worked in the way you expected.

4.5. Weighting linked macro-micro data

This section first explains why your choice of weights will depend on the kind of research question you're asking, and then explains how to calculate each of the two main types of weights used in comparative research.

4.5.1. Different weighting strategies in comparative analysis

With micro data from different countries it is usually the case that the survey in each different country is a different size, and the different sample sizes may have nothing to do with the number of people in each country. When analysing this kind of cross-national data you usually interested in either saying something about the group of countries as a whole, or in treating each country as having equal weight in comparison with others. Both situations require you to weight the data for analysis, but in different ways.

Another complicating factor is that there are sometimes survey sampling weights for the micro data that account for either the structure of the sampling scheme or the pattern of unit non-response, or both. For what follows you can assume that these weights are in a variable called `surveyweight`. If there are no such weights, you can set this variable equal to 1 and the following commands will still work.

4.5.2 Producing weights to equalize the size of each macro unit

The following steps produce a weight variable that you can use for analysis where you want each macro unit to contribute equally.

1. Generate a variable that, for every micro unit within a given macro unit, takes the value of the sum of `surveyweight` within that macro unit. In Stata this can be done using the following command.

```
bysort macrounit: gen muess = sum(surveyweight)
```

You can use whatever variable name you like, but `muess` stands for macro unit effective sample size.

2. Generate a variable that gives the average number of micro units per macro unit. You can do this by hand. Find out the number of micro units and the number of macro units in the data (without weighting) and divide the former by the latter. Note that this variable will be a constant for all units in your data set, so having calculated the number, x, you can generate the variable, in Stata, with the command.

   generate avepermu = x

3. Generate the analysis weight using the following formula.

   generate analysisweight = (avepermu/muess)*surveyweight

4.5.3 Producing weights to respect to population as a whole

The following steps are similar to those in the previous subsection but produce a weight variable that you can use for analysis where you want to make inferences about the population as a whole, so micro units within a given macro unit are weighted according to the true population for that macro unit. For this you will need a variable that gives the population figure for each macro unit (`mupop`).

1. Generate a variable that, for every micro unit within a given macro unit, takes the value of the sum of `surveyweight` within that macro unit. In Stata this can be done using the following command.

   bysort macrounit: gen muess = sum(surveyweight)

You can use whatever variable name you like, but `muess` stands for macro unit effective sample size.

2. Generate the analysis weight using the following formula.

   analysisweight = (avepermu/mupop)*surveyweight

In Stata you can create this variable by simply using the generate command and then typing the above formula.

4.6. Assessing the quality of survey data

This section outlines how to check survey data with macro data and then looks at weighting strategies to adjust for discrepancies.

Comparing the micro survey distribution with macro data

One of the virtues of some macro data sources is that they allow you to check the quality of micro survey data. For example, you can use UK census data to check the quality of UK survey data by comparing the marginal distributions from the census and the survey for variables that both have in common, such as age or gender. Another example is that you can check reported electoral turnout in the **European Social Survey** with official turnout data. This is often done by running a simple frequency distribution and comparing the results with some published source of macro data.

4.6.1 Interpreting differences

The distribution of any particular variable from a survey nearly never matches that from the best macro source. There are various reasons why there might be a discrepancy between your survey and the more official or higher quality macro data.

First, there is random sampling error. Since even the best surveys are random samples from the population, any particular sample is likely to differ from the population average just by chance. The nature of sampling error is well known and you can easily test whether the distribution of a variable in your survey is statistically significantly different from that specified by your high quality macro data.

Second, there may be differences in the sampling frame for the macro and micro data. Survey organizations are frequently unable to sample randomly from the entire population, and the choice of sampling frame can make a difference. For electoral turnout, surveys frequently sample all adults in the country, but official turnout figures often refer to the registered electorate or the voting eligible population.

Third, the nature of the measurement may not be quite the same for the different sources, e.g. the question wording may differ. Fourthly, there maybe effects of the

way people are interviewed (e.g. by phone or face-to-face) and whom they are interviewed by.

Finally, there may be particular unit or item non-response bias in either the survey or the macro source.

4.6.2. Strategies for dealing with survey error

Strategies for dealing with survey error depend on the source of the error. So you should try to investigate the extent to which each different type of error contributes to any discrepancies between the survey and official data you observe in order to formulate the best strategy to adjust for the discrepancies.

Techniques for tackling survey error are many, varied and often complex (see, for instance, Groves 2004). The remainder of this section considers one particularly common form of adjustment: weighting. This is a particularly useful strategy for dealing with unit non-response and relatively little or no information on the nature of the non-respondents. If you know that the distribution of your sample on a key variable (maybe but not necessarily the dependent variable) is different from official (accurate) sources, even if you are unsure of the source of the discrepancy or whether it is in fact an error in the survey, it is usually reasonable to weight your data to match the official marginal distribution for that variable. This could be simply for presentational purposes, but if the use of weights affects your substantive conclusions it is important to report how and why.

4.6.2.1. Adjusting weights to account for survey error on one variable

Suppose that you have a survey where reported turnout is x% but you know from official sources that turnout is actually y% and the discrepancy is not due to differences in the sampling frame and it is statistically significant. You may decide that you wish to weight your data so that in your analysis the reported turnout matches the official turnout. To calculate the relevant weight variable, you can use the following formulae, where the variable turnout is a dummy variable to indicate whether someone did (1) or did not (0) vote.

```
Weight = y/x if turnout==1
Weight = (100-y)/(100-x) if turnout==0
```

If the main variable of interest is a categorical variable with more than two categories, the same principles apply to creating the weight. Respondents in each category need to be assigned a weight equal to the ratio of their frequency in reality to their frequency in the sample.

4.6.2.2. Adjusting weights to account for survey error on one variable separately for each macro unit

Suppose you have a survey where turnout is recorded in several different countries, such as the **European Social Survey**, and you want to make sure that in your analysis reported turnout matches the official turnout in each country separately. Let's say that the variable with official turnout is called `offturnout` and is measured as a proportion rather than as a percentage. You can use the following steps to create a suitable weight.

1. Create a variable which provides the reported level of turnout for each country (`muturnout`), perhaps with the following command in Stata.

```
bysort country: generate muturnout = mean(turnout)
```

2. Create the weight using the following Stata commands or similar.

```
gen weight = .
replace weight = offturnout/muturnout if turnout==1
replace weight = (1-offturnout)/(1-muturnout) if turnout==0
```

4.7. Analysing linked macro and micro data

This section discusses techniques for analysing your linked micro and macro data, starting with exploratory data analysis and moving on to regression analysis with both macro and micro variables, and cross-level interactions.

4.7.1. Starting simple: separate analyses for each macro unit

As discussed in section 4.3, it is possible to run certain analyses of the micro data separately for each macro unit. Having linked the micro and macro data it is possible to do this in a way that is much more relevant to your research question. For instance, suppose you are interested in whether trust in politicians is higher where there is proportional representation (PR), you could estimate the mean level of trust in politicians where there is PR and compare it with the level where there isn't e.g. the `ESSsubset.dta` has a dummy variable `pr` where 1= PR and 0 = not PR, so you could use the following Stata commands.

```
bysort pr: summarize trstplt
```

```
-------------------------------------------------------------------------------
-> pr = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     trstplt |       883    3.509626    2.264452          0         10

-------------------------------------------------------------------------------
-> pr = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     trstplt |      5282    3.883567    2.314891          0         10
```

or alternatively to get a sense of the variation between PR countries and then between non-PR countries, you could try the following command

```
bysort pr country: summarize trstplt
```

```
_____________________________________________________________________________
-----------------------------------------------------------------------------
-> pr = 0, country = FR

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     trstplt |       298    3.708054    2.148996          0          9

-----------------------------------------------------------------------------
-> pr = 0, country = GB

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     trstplt |       298    3.348993    2.123862          0          9

-----------------------------------------------------------------------------
-> pr = 0, country = HU

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     trstplt |       287    3.470383    2.502794          0         10

-----------------------------------------------------------------------------
-> pr = 1, country = AT

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     trstplt |       295    3.386441    2.351686          0         10

-----------------------------------------------------------------------------
-> pr = 1, country = BE

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     trstplt |       297    4.124579     2.15951          0          8
```

This style of analysis can be extended in various different ways. For instance, if the macro variable of interest is interval or ratio level (e.g. gdp per capita) the level of trust in politicians in each country could be graphed against gdp per capita with the country names as markers for the points. You could then subject the data for such an essentially macro level graph to a regression analysis; this is known as the two-step approach (e.g. Jusko and Shively 2005).

4.7.2 Regression analysis with both micro and macro variables

In order to test whether there is a statistically significant effect of a particular macro level variable on a particular micro level outcome, you will probably wish to use some from of regression analysis. For example, suppose you are interested in whether PR increases trust in politicians and you want to control for the effect of generalized trust in people, you could then run the following regression.

```
regress trstplt ppltrst pr
```

```
      Source |       SS       df       MS              Number of obs =    6148
-------------+------------------------------           F(  2,  6145) =  402.38
       Model |  3801.87099     2   1900.9355           Prob > F      =  0.0000
    Residual |  29030.2259  6145  4.72420276           R-squared     =  0.1158
-------------+------------------------------           Adj R-squared =  0.1155
       Total |  32832.0969  6147  5.34115779           Root MSE      =  2.1735

------------------------------------------------------------------------------
     trstplt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ppltrst |   .3159584   .0112979    27.97   0.000     .2938106    .3381062
          pr |    .196908   .0793005     2.48   0.013     .0414513    .3523648
       _cons |   2.082625   .0891842    23.35   0.000     1.907793    2.257457
------------------------------------------------------------------------------
```

The results suggest that there is a significant positive effect of PR on trust in politicians even after controlling for trust in people generally. Note that macro and micro variables are introduced into the regression in entirely the same way.

> You should be aware that linear regression takes no account of the clustering of micro units within macro units, and so there is good reason to believe that the standard errors of the coefficients might be incorrect and t-values inflated as a consequence. This problem can be dealt with by using a multilevel model as described in Unit 5 in this series.

4.7.3 Regression analysis to test cross-level interactions

A cross-level interaction is an interaction between a variable measured at the micro level and one measured at the macro level. Otherwise it is no different from a standard interaction effect in that we test for the interaction by considering the coefficient of a variable that is the product of two other variables. For example, suppose you are interested in whether trust in people in general has a stronger effect on trust in politicians in PR countries than elsewhere, then you could first create a variable which is the product of `ppltrst` and `pr`.

```
gen ppltrstXpr = ppltrst*pr
```

and then run the following regression with the new variable as the interaction term.

```
regress trstplt ppltrst pr ppltrstXpr
```

```
      Source |       SS       df       MS              Number of obs =    6148
-------------+------------------------------           F(  3,  6144) =  268.37
       Model |  3803.81214     3  1267.93738            Prob > F      =  0.0000
    Residual |  29028.2848  6144  4.72465573            R-squared     =  0.1159
-------------+------------------------------           Adj R-squared =  0.1154
       Total |  32832.0969  6147  5.34115779           Root MSE      =  2.1736

------------------------------------------------------------------------------
     trstplt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ppltrst |   .2969988   .0316635     9.38   0.000     .2349273    .3590703
          pr |   .0972149   .1745838     0.56   0.578    -.2450304    .4394603
  ppltrstXpr |   .0217259   .0338948     0.64   0.522    -.0447198    .0881715
       _cons |   2.168255   .1606279    13.50   0.000     1.853368    2.483141
------------------------------------------------------------------------------
```

The coefficient of `ppltrstXpr` is positive and statistically significant, indicating that the correlation between trust in people in general and trust in politicians is stronger in PR countries. Again this analysis doesn't take the clustering of people into different countries into account, and strategies to deal with this are discussed in Unit 5.

4.8. Summary

This unit has covered various analytical techniques and methods of data management associated with hierarchically structured data, i.e. data at both the micro and macro levels. These have started with the kinds of research questions that require linked micro and macro data, assessing micro data for variation between macro units, linking micro and macro data, weighting linked micro and macro data, checking the quality of micro data with macro data and adjusting weights to account for discrepancies, and basic approaches to the analysis of linked micro and macro data. Unit 6 moves on from this unit by introducing multilevel modelling which is explicitly designed for hierarchical data of which linked micro and macro data is an example.

4.9 References/Further Reading

Agresti, A and B Finlay (1997) Statistical Methods for the Social Sciences. Prentice Hall.

Groves, R. M. (2004) Survey Methodology. Wiley, Hoboken, NJ.

Karen Long Jusko and W. Phillips Shively, Applying a Two-Step Strategy to the Analysis of Cross-National Public Opinion Data. Political Analysis 2005 13: 327-344; doi:10.1093/pan/mpi030