Data management for multilevel analysis

Importing and aggregating
Centering and standardizing

Dimitri Mortelmans
Centre for Longitudinal and Life Course Studies (CLLS)
University of Antwerp

OVERVIEW

- **At cluster level (country level)**
  - Importing external data
  - Aggregating internally (or externally)
- **At the individual level**
  - Dichotomizing

0. Converting the ESS-variable cntry

Turning **cntry** into a numeric variable

```
* 0.1 Sorting and creating a new numeric country-variable.
* ------------------------------------------------------------------------------------------.

SORT CASES BY cntry(A).

AUTORECODE VARIABLES=cntry
  /INTO cntry2
  /PRINT.

* Checking the result.
FREQUENCIES VARIABLES=cntry cntry2.
```

Overview of the ML data management

What sorts of questions might arise when preparing your data for a multilevel-analysis ?

Overview of ML data management

- Type 1 = creating raw variables
- Type 2 = centering and standardizing

Overview of ML data management

- **Type 1 = Creating raw variables**

  - Creation at level 1

    Usually unnecessary because they are already in your dataset.

  - Creation at level 2

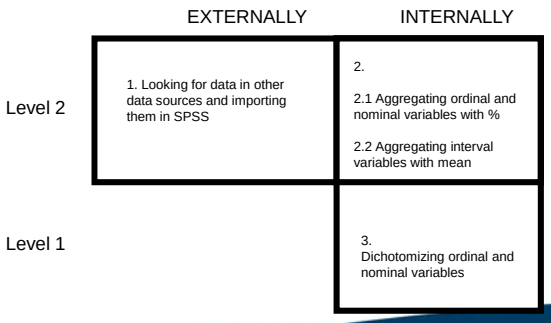
    EITHER *Searching external meta-data and import them*
    (E.g. BNP_PPS of BNP)

    OR *Aggregate internally*
    (bringing variables in your dataset from level 1 to level 2)
    (E.g. % Catholics or Mean Religiosity)

Overview of ML data management

- **Type 1 = Creating raw variables**

  - Aanmaken op level 2



Overview of ML data management

- **Type 2 = Centering and standardizing raw variables**

  - All variables in your multilevel model need to be **centered**.
    (exception = dummies)

  - All variables in your multilevel model need to be **standardized** (see step 11).
    (no exceptions)

Data management at the country level

TYPE 1: Creating raw variables at level 2

1. Importing external data

Where to get external data?

- New Cronos – Eurostat
  - http://europa.eu.int/newcronos/
- Statistics Division - United Nations
  - http://unstats.un.org/unsd/databases.htm
- Data and Statistics - IMF
  - http://www.imf.org/external/data.htm
- Data Centre - UNESCO
  - http://stats.uis.unesco.org/unesco/TableViewer/document.aspx?ReportId=143&IF_Language=eng
- Max Planck Institute for Demographic Research
  - http://www.demogr.mpg.de/databases/cdb/

Data in the example: New Cronos

GDP per capita in PPS

GDP per capita in Purchasing Power Standards (PPS) (EU-27 = 100)

Please be aware that this indicator has been rescaled, i.e. data is expressed in relation to EU-2

Gross domestic product (GDP) is a measure for the economic activity. It is defined as the Standards (PPS) is expressed in relation to the European Union (EU-27) average set to eq PPS, i.e. a common currency that eliminates the differences in price levels between countri = 100, is intended for cross-country comparisons rather than for temporal comparisons.

For more information...

	1997	1998	1999	2000	2001	2002
EU (27 countries)	100.0	100.0	100.0	100.0	100.0	100.0
EU (25 countries)	104.9	105.0	105.0	105.0	104.8	104.6
EU (15 countries)	115.5	115.5	115.4	115.2	114.8	114.2
Euro area	116.2	116.3	116.1	115.6	113.8	112.9
Euro area (13 countries)	114.5	114.6	114.4	113.9	113.6	112.7
Euro area (12 countries)	114.8	114.8	114.7	114.2	113.8	112.9
Belgium	126.2	123.4	123.5	126.4	124.0	125.6
Bulgaria	26.6 (e)	27.1 (e)	27.1	27.9	29.4	31.1
Czech Republic	73.3 (e)	70.8 (e)	69.9	68.7	70.5	70.8
Denmark	133.8	132.6	131.5	132.2	128.4	129.0

Importing external data

```
DATA LIST
 / cntry2 1-2  BNP2002 3-12 (F,1)  BNP2002PPS 15-20 (F,1).
BEGIN DATA
1   220840,9  127,9
2   267652,4  125,6
3   296018,0
4   80003,6   70,8
5  2143180,0  115,7
6   184743,6  129,0
7   729206,0  100,9
8   143974,0  115,7
9  1548559,0  116,5
10 1678980,0  118,9
11  157586,0   91,1
12   70713,7   61,7
13  130214,0  138,5
14
15 1295225,7  112,4
16   23992,3  241,3
17  465214,0  134,0
18  204073,6  155,4
19  209617,4   48,5
20  135433,6   77,4
21  258877,9  119,2
22   24134,2   81,3
END DATA.
```

Define your variable names here

These are the actual data (from three variables).

Importing external data

```
VARIABLE LABELS BNP2002 "GDP at market current prices (mio euro) 2002 - New Chronos EU"
 / BNP2002PPS "GDP in PPS EU27 - 2002 - New Chronos EU".
```

Create some VARIABELE LABELS

```
SAVE OUTFILE='C:\ESS\landvars1.sav'  /COMPRESSED.
```

Save the file as "landvars1.sav"

```
* Checking the import routine.
SUMMARIZE
  /TABLES=cntry2 BNP2002 BNP2002PPS
  /FORMAT=LIST NOCASENUM .
```

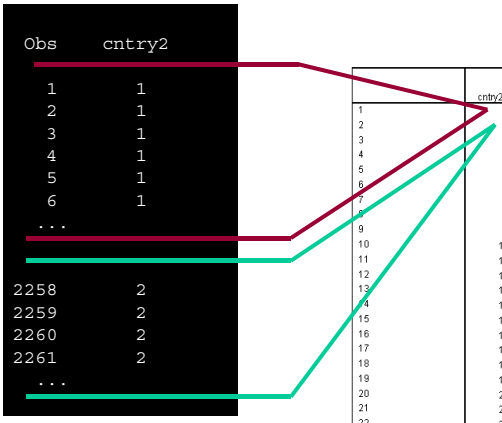
Always use a check afterwards.

Importing external data

Report

		cntry2	GDP at market current prices (mio euro) 2002 - New Chronos EU	GDP in PPS EU27 - 2002 - New Chronos EU
1		1	220840,9	127,9
2		2	267652,4	125,6
3		3	296018,0	.
4		4	80003,6	70,8
5		5	2143180,0	115,7
6		6	184743,6	129,0
7		7	729206,0	100,9
8		8	143974,0	115,7
9		9	1548559,0	116,5
10		10	1678980,0	118,9
11		11	157586,0	91,1
12		12	70713,7	61,7
13		13	130214,0	138,5
14		14	.	.
15		15	1295225,7	112,4
16		16	23992,3	241,3
17		17	465214,0	134,0
18		18	204073,6	155,4
19		19	209617,4	48,5
20		20	135433,6	77,4
21		21	258877,9	119,2
22		22	24134,2	81,3
Total	Mean	11,50	488963,805	114,090
	N	22	21	20
	Std. Deviation	6,494	620620,4262	40,7395

Adding the meta data to your dataset



Obs	cntry2
1	1
2	1
3	1
4	1
5	1
6	1
...	
2258	2
2259	2
2260	2
2261	2
...	

Report

		cntry2	GDP at market current prices (mio euro) 2002 - New Chronos EU	GDP in PPS EU27 - 2002 - New Chronos EU
1		1	220840,9	127,9
2		2	267652,4	125,6
3		3	296018,0	.
4		4	80003,6	70,8
5		5	2143180,0	115,7
6		6	184743,6	129,0
7		7	729206,0	100,9
8		8	143974,0	115,7
9		9	1548559,0	116,5
10		10	1678980,0	118,9
11		11	157586,0	91,1
12		12	70713,7	61,7
13		13	130214,0	138,5
14		14	.	.
15		15	1295225,7	112,4
16		16	23992,3	241,3
17		17	465214,0	134,0
18		18	204073,6	155,4
19		19	209617,4	48,5
20		20	135433,6	77,4
21		21	258877,9	119,2
22		22	24134,2	81,3
Total	Mean	11,50	488963,805	114,090
	N	22	21	20
	Std. Deviation	6,494	620620,4262	40,7395

Adding the meta data to your dataset

```
* ALWAYS sort your data before you merge !.

DATASET ACTIVATE ess1e05_1.
SORT CASES BY cntry2(A).

GET FILE='C:\ESS\landvars1.sav'.
DATASET NAME landvars1 WINDOW=FRONT.
SORT CASES BY cntry2(A).

* THEN you start to merge.

DATASET ACTIVATE  ess1e05_1.
MATCH FILES /FILE=*
  /IN=Control_1
  /TABLE='C:\ESS\landvars1.sav'
  /IN=Control_2
  /BY cntry2.
EXECUTE.

VARIABLE LABELS Control_1 " Record comes from the original ESS-file "
                / Control_2 " Record comes from the external landvars-file".
```

Adding the meta data to your dataset

```
* ALWAYS sort your data before you merge !.

DATASET ACTIVATE ess1e05_1.
SORT CASES BY cntry2(A).

GET FILE='C:\ESS\landvars1.sav'.
DATASET NAME landvars1 WINDOW=FRONT.
SORT CASES BY cntry2(A).

* THEN you start to merge.

DATASET ACTIVATE  ess1e05_1.
MATCH FILES /FILE=*
  /IN=Control_1
  /TABLE='C:\ESS\landvars1.sav'
  /IN=Control_2
  /BY cntry2.
EXECUTE.

VARIABLE LABELS Control_1 " Record comes from the original ESS-file "
                / Control_2 " Record comes from the external landvars-file".
```

Always sort first !
otherwise, the BY does not work

Make a new variable that checks from which dataset each record comes from.

The cntry2-variable is used a an ID-variable between the two datasets.

Adding the meta data to your dataset

```
*  Checking the merge-routine;
CROSSTABS  /TABLES=Control_1 BY Control_2.
```

Possible form of the table “Control_1* Control_2”.

Control_1 \ Control_2	0	1
0	The record comes from neither of the two datasets	The record comes from landvars1 and not from ESS
1	The record comes from ESS and not from landvars1	The record comes from landvars1 and ESS.

Adding the meta data to your dataset

```
* Checking the merge-routine;
CROSSTABS /TABLES=Control_1 BY Control_2.
```

Possible form of the table “Control_1* Control_2”.

Control_1 \ Control_2	1
1	N

This means that each country in LANDVARS was correctly merged whit the countries in ESS.

Adding the meta data to your dataset

Record comes from the original ESS-file * Record comes from the external landvars-file Crosstabulation

Count

		Record comes from the external landvars-file	Total
		1	
Record comes from the original ESS-file	1	42359	42359
Total		42359	42359

2. Aggregating internally

2.1 Aggregating categorical variables
2.2 Aggregating interval variables

TYPE 1: Creating raw variables at level 2

2.1 Aggregating categorical variables

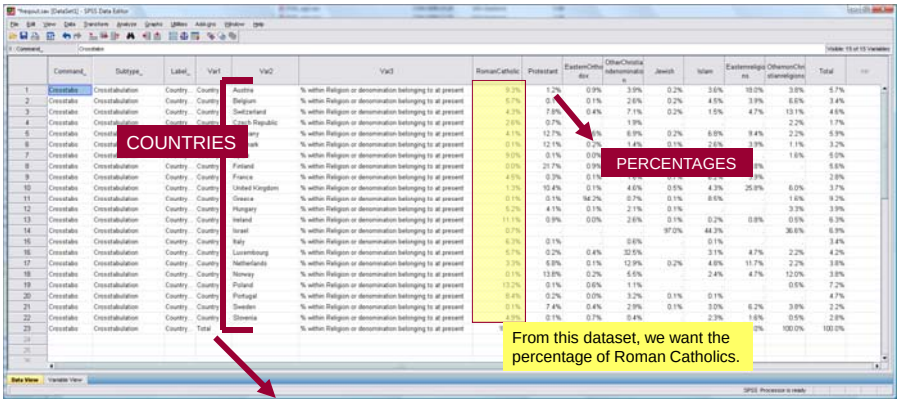
```
* A. Percentage of Roman Catholics in a country.
* (Example of aggregating a categorical variable in a data-file).

* Method: Save a table of the variable to the file "freqout".

DATASET ACTIVATE ess1e05_1.
SORT CASES BY cntry2(A).

OMS SELECT TABLES
 /DESTINATION format = sav OUTFILE = "C:\ESS\freqout.sav"
 /IF commands = ['Crosstabs'] subtypes = ['Crosstabulation']

CROSSTABS
  /TABLES=cntry BY RLGDNM
  /CELLS=COLUMN.

OMSEND.
```

With OMS, you can divert tables towards a dataset.

We will export this table

Cross tabulation from religion and country.

2.1 Aggregating categorical variables



2.1 Aggregating categorical variables

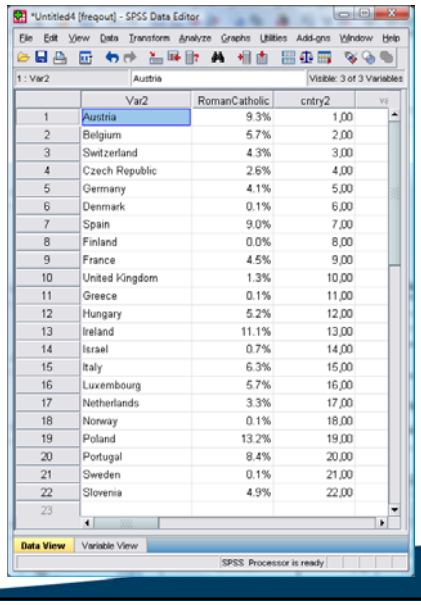
```
GET FILE='C:\ESS\freqout.sav'.
DATASET NAME freqout WINDOW=FRONT.

FILTER OFF.
USE ALL.
SELECT IF (Var1 = "Country").
EXECUTE.

SAVE OUTFILE='C:\ESS\freqout.sav'
  /DROP=Command_ Subtype_ Label_ Var1 Var3 Protestant EasternOrthodox
OtherChristiandenomination Jewish Islam Easternreligions OthernonChristianreligions Total
  /COMPRESSED.

GET FILE='C:\ESS\freqout.sav'.
DATASET NAME freqout WINDOW=FRONT.
COMPUTE cntry2=$CASENUM.
SAVE OUTFILE='C:\ESS\freqout.sav'.

* Checking.
SUMMARIZE
  /TABLES=cntry2 RomanCatholic
  /FORMAT=LIST NOCASENUM .
```

Only keep the cases with the percentages. (= remove the Total-regel underneath the dataset.)

Only keep the variables Var2 (=country) and RomanCatholic

Make a new variable cntry2 based on the case number (= $CASENUM).

2.1 Aggregating categorical variables



2.1 Aggregating categorical variables

Report

		cntry2	Religion or denomination belonging to at present Roman Catholic
1		1,00	9,3%
2		2,00	5,7%
3		3,00	4,3%
4		4,00	2,6%
5		5,00	4,1%
6		6,00	,1%
7		7,00	9,0%
8		8,00	,0%
9		9,00	4,5%
10		10,00	1,3%
11		11,00	,1%
12		12,00	5,2%
13		13,00	11,1%
14		14,00	,7%
15		15,00	6,3%
16		16,00	5,7%
17		17,00	3,3%
18		18,00	,1%
19		19,00	13,2%
20		20,00	8,4%
21		21,00	,1%
22		22,00	4,9%
Total	Mean	11,5000	4,545%
	N	22	22
	Std. Deviation	6,49359	3,8616%

2.2 Aggregating continuous variables

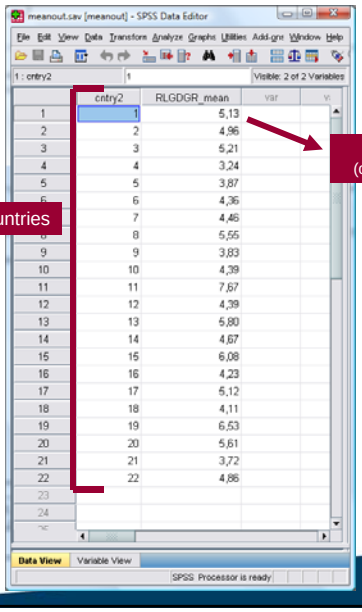
```
DATASET ACTIVATE ess1e05_1.

SORT CASES BY cntry.
AGGREGATE
  /OUTFILE='C:\ESS\meanout.sav'
  /PRESORTED
  /BREAK=cntry2
  /RLGDGR_mean=MEAN(RLGDGR).

GET FILE='C:\ESS\meanout.sav'.
DATASET NAME meanout WINDOW=FRONT.


* Checking.
SUMMARIZE
  /TABLES=cntry2 RLGDGR_mean
  /FORMAT=LIST NOCASENUM .
```

Always sort first !
otherwise, the BREAK does not work

Save the mean (for each country) in a new dataset, called *meanout*.

The name of the variable with the mean is called *RLGDGR_mean*.

2.2 Aggregating continuous variables



RLGDGR_mean
(one mean per country)

All 22 ESS-countries

2.3 Merging both variables into the ESS-dataset

```
DATASET ACTIVATE ess1e05_1.
SORT CASES BY cntry2(A).

DATASET ACTIVATE freqout.
SORT CASES BY cntry2(A).

DATASET ACTIVATE meanout.
SORT CASES BY cntry2(A).

DATASET ACTIVATE  ess1e05_1.
MATCH FILES /FILE=*
    /IN=Control_4
  /TABLE='freqout'
    /IN=Control_5
  /TABLE='meanout'
    /IN=Control_6
 /BY cntry2.
EXECUTE.

VARIABLE LABELS Control_4 "Record comes from the original ESS-file" / Control_5 "Record comes from the freqout-file" / Control_6 "Record comes from the meanout-file".

* Checking the merge-routine;
CROSSTABS  /TABLES=Control_4 BY Control_5.
CROSSTABS  /TABLES=Control_4 BY Control_6.
```

First, we sort all three datasets (again)

The file receiving the indicators comes first (= ess1e05_1)

Freqout and meanout can be merged to the ESS-dataset in one step. For all datasets, we create a control-variable.

Centering and standardizing

TYPE 2: Centering and standardizing raw variables

Why centering ?

- Interpretation intercept = What happens when the independent variables are equal to 0 ?

  BUT: The value 0 needs a proper meaning !

  Centering gives a meaning to the value of 0 (being: the mean value of that variable).

- Interpretation interaction-effect = What happens with the slope of one variable when the other is equal to 0 ?

Why centering ?

Centering gives a meaning to the value of 0

So ALL variables in a multilevel analysis (dependent, independent X and independent Z) have to be centered and included as such in the models.

EXCEPTION: Dummies: here the value 0 has already a meaningful interpretation.

Why centering and standardizing ?

- Centering
  - Better interpretation of the intercept.
  - Possible interpretation of the variance components
  - Possible interpretation of interaction-effects
  - Speeds up the estimation of the models
- Standardization
  - Makes parameters comparable

What is centering and standardizing ?

- Centering
  - Subtract the grand mean of a variable from each value.
  - Result: mean = 0
  - Formula: $X_i - \bar{X}$
- Standardizing
  - Subtract the grand mean of a variable from each value and divide by the standard deviation.
  - Result: Mean = 0, Standard deviation =1
  - Formula:

$$\frac{X_i - \bar{X}}{\sigma}$$

Moeilijkheid in SPSS

- SPSS cannot center automatically
- Function MEAN does not work
  - because MEAN() calculates a mean *over* variables.
  - We need a mean *within* one variable
- SOLUTION: centering by hand

  Calculate the mean and standard deviation with the AGGREGATE-procedure and save them immediately in your ESS dataset.

Centering and standardizing the country variables

Centering and standardizing the country variables

STEP 1: Calculating the means and standard deviations

```
DATASET ACTIVATE ess1e05_1.

AGGREGATE
  /PRESORTED
  /BREAK=name
  /m1=MEAN(BNP2002)
  /m2=MEAN(BNP2002PPS)
  /m3=MEAN(RomanCatholic)
  /m4=MEAN(RLGDGR_mean)
  /std1=SD(BNP2002)
  /std2=SD(BNP2002PPS)
  /std3=SD(RomanCatholic)
  /std4=SD(RLGDGR_mean).

* Checking.
FREQUENCIES / VARIABLES= m1 m2 m3 m4 std1 std2 std3 std4.
```

MEAN asks for the mean of this variable.

SD asks for the stan dev of this variable.

REMARK: We use cntry2 not as the BREAK-variable because you want the mean over *all* individuals and countries (= GRAND MEAN centering). That is why we use the constant "name".

Centering and standardizing the country variables

STEP 2: Centering and standardizing

```
* 1.2 Computing the centered and standardised variables.
* ------------------------------------------------------------------------------.

COMPUTE bnp_centr       = BNP2002 - m1.
COMPUTE bnppps_centr    = BNP2002PPS - m2.
COMPUTE romkath_centr   = percent - m3.
COMPUTE relig_centr     = meanrelig - m4.
COMPUTE bnp_std         = bnp_centr / std1.
COMPUTE bnppps_std      = bnppps_centr / std2.
COMPUTE romkath_std     = romkath_centr / std3.
COMPUTE relig_std       = relig_centr / std4.

VARIABLE LABELS bnp_centr "BNP 2002 gecentreerd"
                / bnppps_centr  "BNP-PPS 2002 centered"
                / romkath_centr "Perc. Rooms-Kath. centered"
                / relig_centr "Mean religiosity centered"
                / bnp_std "BNP 2002 standardised"
                / bnppps_std "BNP-PPS 2002 standardised"
                / romkath_std "Perc. Roman Catholics standardised"
                / relig_std "Mean religiosity standardised".

FREQUENCIES
/ VARIABLES= bnp_centr bnppps_centr romkath_centr relig_centr bnp_std bnppps_std romkath_std relig_std.
```

Here, we center the variables.

Here, we standardize the variables

To finish, we give each new variable a variabele label.

Centering and standardizing the individual variables

Centering and standardizing the individual variables

STEP 1: Calculating the means and standard deviations

```
DATASET ACTIVATE ess1e05_1.

AGGREGATE
  /PRESORTED
  /BREAK=name
  /m5=MEAN(hinctnt)
  /m6=MEAN(geslacht)
  /m7=MEAN(edulvl)
  /std5=SD(hinctnt)
  /std6=SD(geslacht)
  /std7=SD(edulvl).

* Checking.
FREQUENCIES / VARIABLES= m5 m6 m7 std5 std6 std7.
```

MEAN asks for the mean of this variable.

SD asks for the stan dev of this variable.

REMARK: We use cntry2 not as the BREAK-variable because you want the mean over *all* individuals and countries (= GRAND MEAN centering). That is why we use the constant "name".

Centering and standardizing the individual variables

STEP 2: Centering and standardizing

```
* 2.1 Computing the means and standard deviations of income, gender and education with AGGREGATE
* * ----------.

COMPUTE ink_centr = hinctnt - m5.
COMPUTE gesl_centr= geslacht - m6.
COMPUTE opl_centr = edulvl - m7.
COMPUTE ink_std = ink_centr / std5.
COMPUTE gesl_std = gesl_centr / std6.
COMPUTE opl_std = opl_centr / std7.

VARIABLE LABELS ink_centr  "Income centered"
                /  gesl_centr "Gender centered"
                /  opl_centr  "Education centered"
                /  ink_std    "Income standardised"
                /  gesl_std   "Gender standardised"
                /  opl_std    "Education standardised".

* Checking.
FREQUENCIES /VARIABLES= hinctnt edulvl ink_centr gesl_centr opl_centr ink_std gesl_std
opl_std.
```

Here, we center the variables.

Here, we standardize the variables

To finish, we give each new variable a variabele label.

Data management for multilevel analysis

Importing and aggregating
Centering and standardizing

Dimitri Mortelmans
Centre for Longitudinal and Life Course Studies (CLLS)
University of Antwerp