1.1 Database structure

Databases can be saved in different formats:

- **Long format** the database has repeated measurements in separate records.
- **Wide format** the database has repeated measurements in separate columns of the same record.
- **Sheet format** the database has repeated measurements in separate sheets of the same record.

According to its format, the database should have its correspondent structure. See below the correspondent characteristics.

1.1.1 Long format

1. The file containing the database should have the extension ".xlsx"
2. The first column of the table should have the subjects ids, while the second column should have the time in numbers (i.e. $1, 2, 3, 4, \dots$). From the third column on the dataset should store first the features that remain unchanged through time (e.g. age of the subjects, classification factors, fixed numeric factors), then the features that change through time.
3. All times must be present for each subject. This means that if a subject misses the values at time n, it should be added an empty row indicating the subject id, the time n, all the values time independent and empty values for the time dependent features.
4. The decimal numbers should be written with the point indicating the decimal, not the comma.
5. The names of the variables should be as simple and as intuitive as possible. It is better to avoid capital letters, too long names and punctuation characters such as ! " # $ % & ' ( ) * + , − / : ; < = > ? @ [ ] ^ _ { | }
6. It is advisable to avoid columns with comments or information about the data, once imported in R, those information become useless.

See Table 1 as an example.

1.1.2 Wide format

1. The file containing the database should have the extension ".xlsx"
2. The first column of the table should have the subjects ids. From the second column on the dataset should store first the features that remain unchanged through time (e.g. age of the subjects, classification factors, fixed numeric factors), then the features that change through time. The same feature must be repeated according to the time stamps of the longitudinal study. The name of each feature through time should be the same, except for the time stamp indicator, that should be expressed as "*.number*". For example *var_name.1, var_name.2, var_name.3*, for the variable scores from time 1 to time 3.

subj	time	fact1	fact2	fact3	fact4
s1	1	0	2	23.4	12
s1	2	0	2	34	56
s1	3	0	2	21	32.5
s2	1	1	6	12.2	11
s2	2	1	6	32	23
s2	3	1	6	14	22
s3	1	0	3	7.5	6.3
s3	2	0	3	32	24
s3	3	0	3		
s4	1	1	8	12	8
s4	2	1	8		
s4	3	1	8	42	21

Table 1: Long format dataset.

subj	fact1	fact2	fact3.1	fact3.2	fact3.3	fact4.1	fact4.2	fact4.3
s1	0	2	23.4	34	21	12	56	32.5
s2	1	6	12.2	32	14	11	23	22
s3	0	3	7.5	32		6.3	24	
s4	1	8	12		42	8		21

Table 2: Wide format dataset.

3. The decimal numbers should be written with the point indicating the decimal, not the comma.

4. The names of the variables should be as simple and as intuitive as possible. It is better to avoid capital letters, too long names and punctuation characters such as ! " # $ % & ' ( ) * + , − / : ; < = > ? @ [ ] ^ _ { | }

5. It is advisable to avoid columns with comments or information about the data, once imported in R, those information become useless.

See Table 2 as an example.

1.1.3 Sheet format

1. The file containing the database should have the extension ".xlsx"

2. The first column of the table should have the subjects ids, while the second column the time, in numbers, displayed in the according sheet (e.g. 1 for the first sheet, 2 for the second sheet and so on). From the third column on the dataset should store first the features that remain unchanged through time (e.g. age of the subjects, classification factors, fixed numeric factors), then the features that change through time. Each sheet of the ".xlsx" file represents a time. It is important to maintain fixed the variable names throughout the sheets.

3. The decimal numbers should be written with the point indicating the decimal, not the comma.

subj	time	fact1	fact2	fact3	fact4
s1	1	0	2	23.4	12
s2	1	1	6	12.2	11
s3	1	0	3	7.5	6.3
s4	1	1	8	12	8

Sheet format dataset, sheet 1.

subj	time	fact1	fact2	fact3	fact4
s1	2	0	2	34	56
s2	2	1	6	32	23
s3	2	0	3	32	24
s4	2	1	8		

Sheet format dataset, sheet 2.

subj	time	fact1	fact2	fact3	fact4
s1	3	0	2	21	32.5
s2	3	1	6	14	22
s3	3	0	3		
s4	3	1	8	42	21

Sheet format dataset, sheet 3.

Table 3

4. The names of the variables should be as simple and as intuitive as possible. It is better to avoid capital letters, too long names and punctuation characters such as ! " # $ % & ' ( ) * + , − / : ; < = > ? @ [ ] ^ _ { | }
5. It is advisable to avoid columns with comments or information about the data, once imported in R, those information become useless.

See Table 3 as an example.