Introduction to Longitudinal Data Structure and Visualisation

Nicolás Libuy, nicolas.libuy@ucl.ac.uk

Darío Moreno-Agostino, d.moreno@ucl.ac.uk

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

Thanks to our funders and host institution

Economic and Social Research Council

Funded by

www.esrc.ukri.org

Hosted by

www.ucl.ac.uk

Outline for today

CENTRE FOR LONGITUDINAL STUDIES

Time	Duration	What
12:00-12:05	5m	Welcome & introduction
12:05-12:15	10m	What is longitudinal data?
12:15-12:55	40m	Data structure with demonstration in Stata (incl. Q&A)
12:55-13:00	5m	Break
13:00-14:00	60m	Data visualization with demonstration in Stata (incl. Q&A)
14:00-14:30	30m	Extended Q&A session

Housekeeping

CENTRE FOR LONGITUDINAL STUDIES

- Please keep your cameras off and mics muted at all times.
- If you have a question, please use the chat function, and please note your question will be visible to all attendees.
- Technical issues – please email us: ioe.clsevents@ucl.ac.uk
- We would be grateful for your feedback. Please follow the link in the chat and complete the short survey at the end.
- We are recording today's event.
- We will remove from the video any participant names to protect privacy.

Thank you for joining us today.

What is longitudinal data?

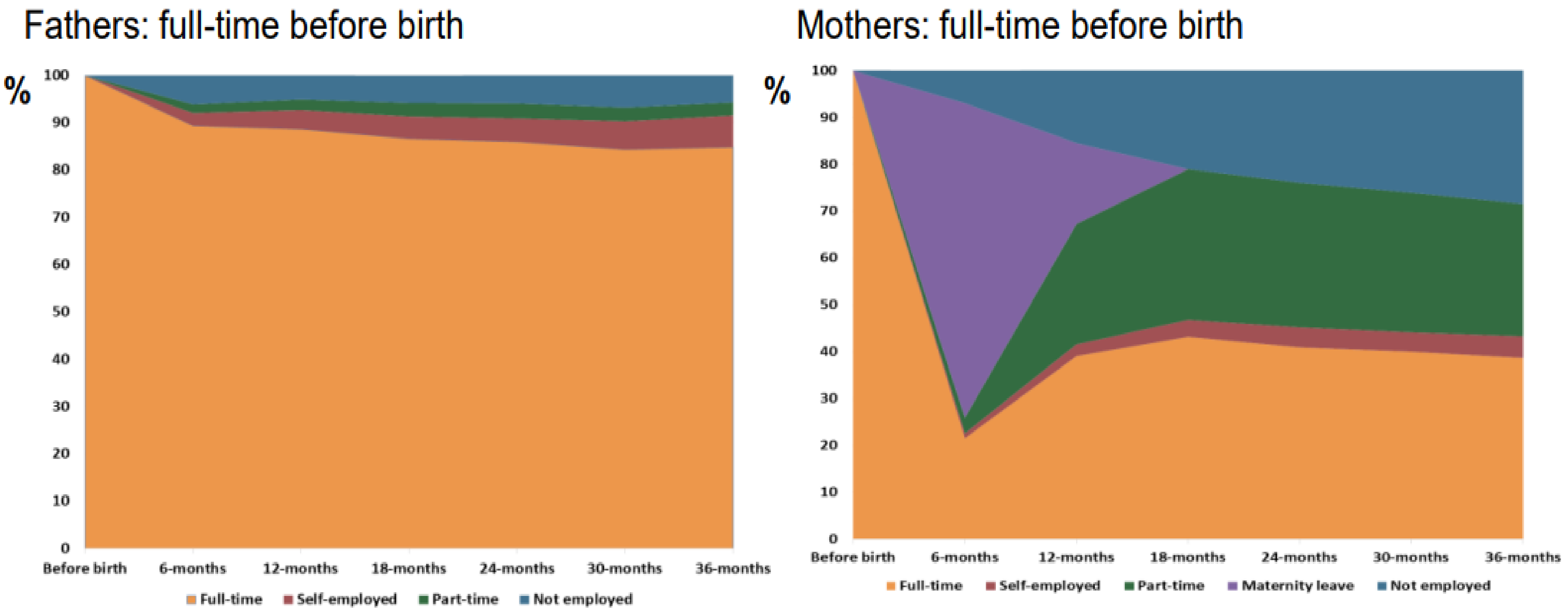
CENTRE FOR
LONGITUDINAL
STUDIES

Economic
and Social
Research Council

Longitudinal data

- Data collected over time **within the same units** (individuals, households…)
  - Describe and visualise trajectories
  - Compare developmental growth / change over time

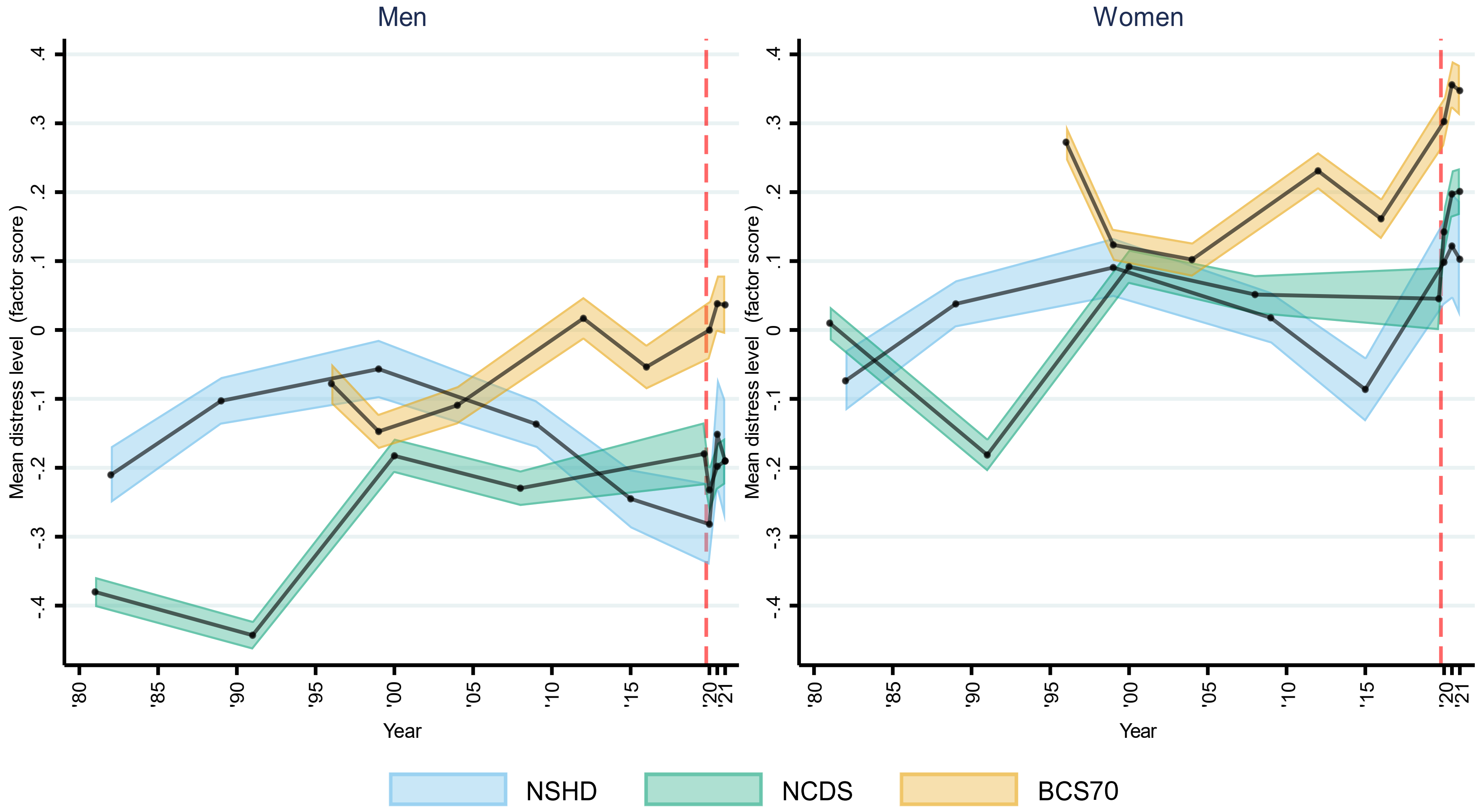
Figure 1: Mothers and fathers' employment patterns (%) in the 3 years after birth



Source: Harkness et al 2019 Employment pathways and occupational change after childbirth https://dera.ioe.ac.uk//34421/

CENTRE FOR LONGITUDINAL STUDIES

Figure 2. Inequalities across women and men in disruption of long-term trajectories of psychological distress with COVID-19 pandemic



Source: Moreno-Agostino et al 2023, Long-term psychological distress trajectories and the COVID-19 pandemic in three British birth cohorts: A multi-cohort study. https://doi.org/10.1371/journal.pmed.1004145

Longitudinal data

- Data collected over time **within the same units** (individuals, households…)
  - Describe and visualise trajectories
  - Compare developmental growth / change over time

As opposed to what?

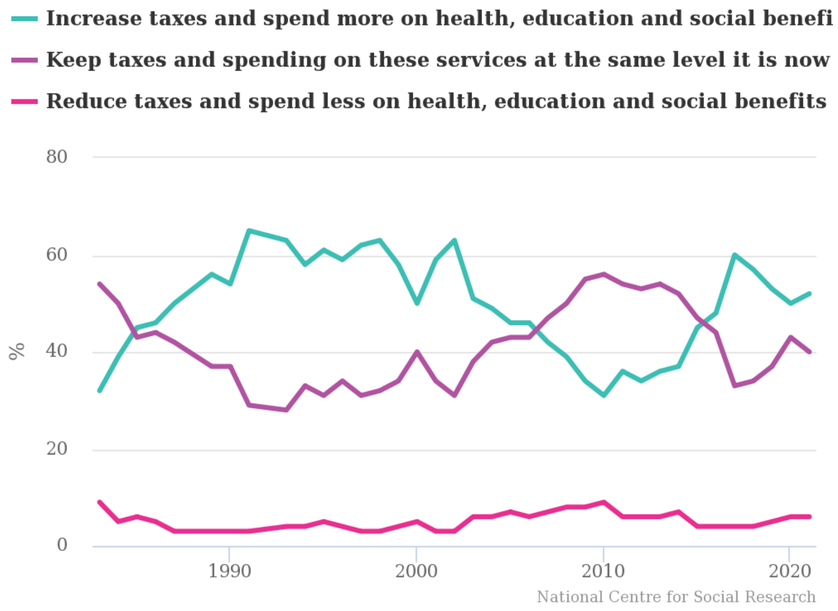
Repeated cross-sectional data

CENTRE FOR
LONGITUDINAL
STUDIES

- Data collected over time *across different units*
- Macro-level trends (e.g., in prevalence)…

Attitudes towards taxation and spending on health, education and social benefits, 1983-2021

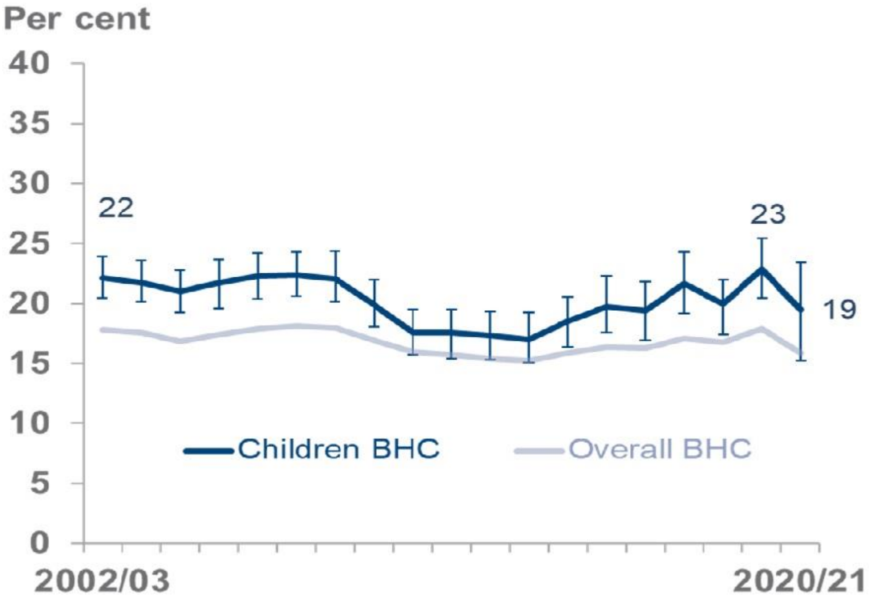
Source: British Social Attitudes



https://www.bsa.natcen.ac.uk/latest-report/british-social-attitudes-39/taxation-welfare-and-inequality.aspx

Children in Low Income Households

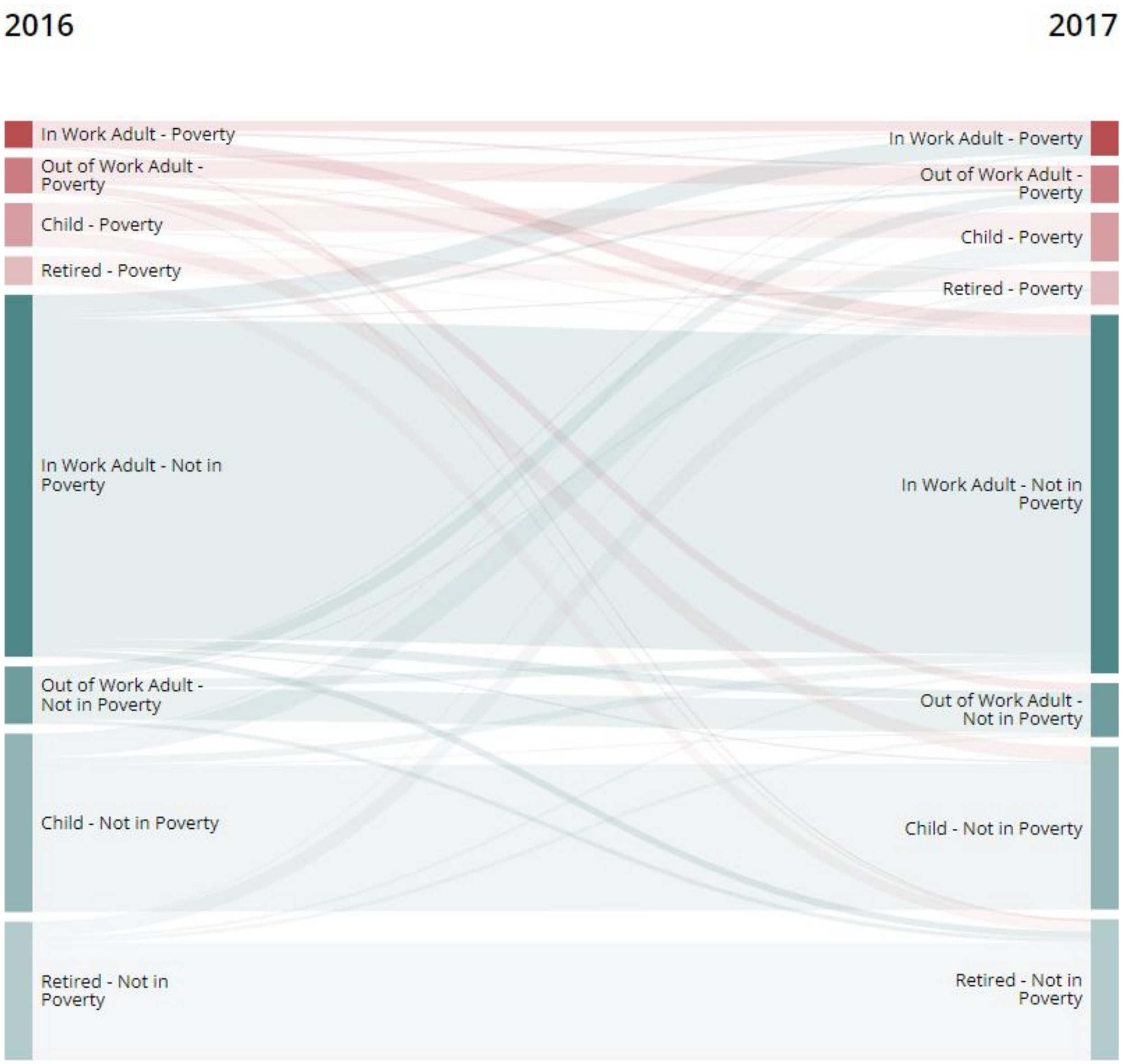
Source: Households Below Average Income series, Family Resources Survey



https://www.gov.uk/government/statistics/households-below-average-income-for-financial-years-ending-1995-to-2021

CENTRE FOR
ITUDINAL
ES

- … but can't capture micro-level (individual) change & transitions

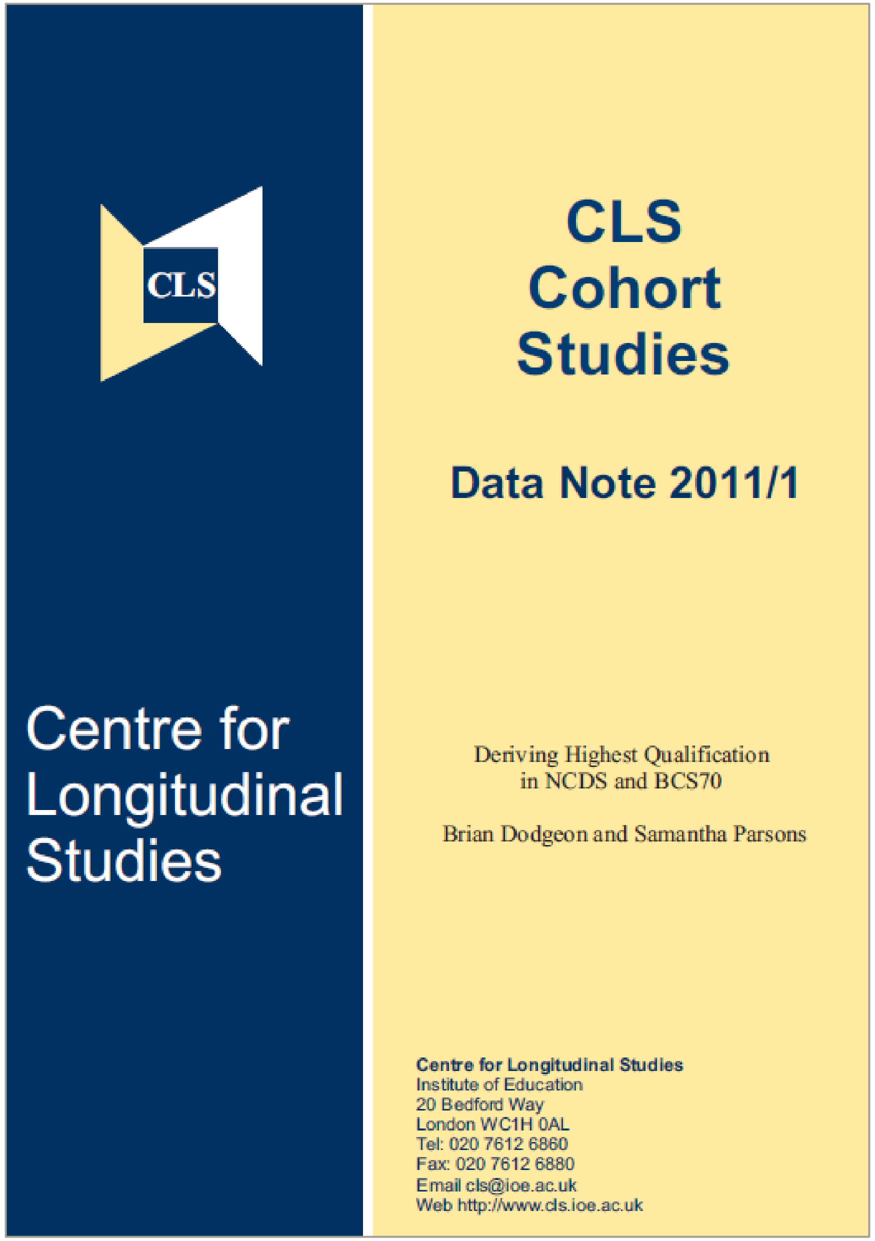


Source: Office for National Statistics https://www.ons.gov.uk/visualisations/dvc617/sankey/

Longitudinal data

- Data collected over time **within the same units** (individuals, households…)
  - Describe and visualise trajectories
  - Compare developmental growth / change over time
  - Sometimes, *same* measures over time
    (but sometimes need harmonisation/linking approaches)

CENTRE FOR
LONGITUDINAL
STUDIES



CLS

CLS Cohort Studies

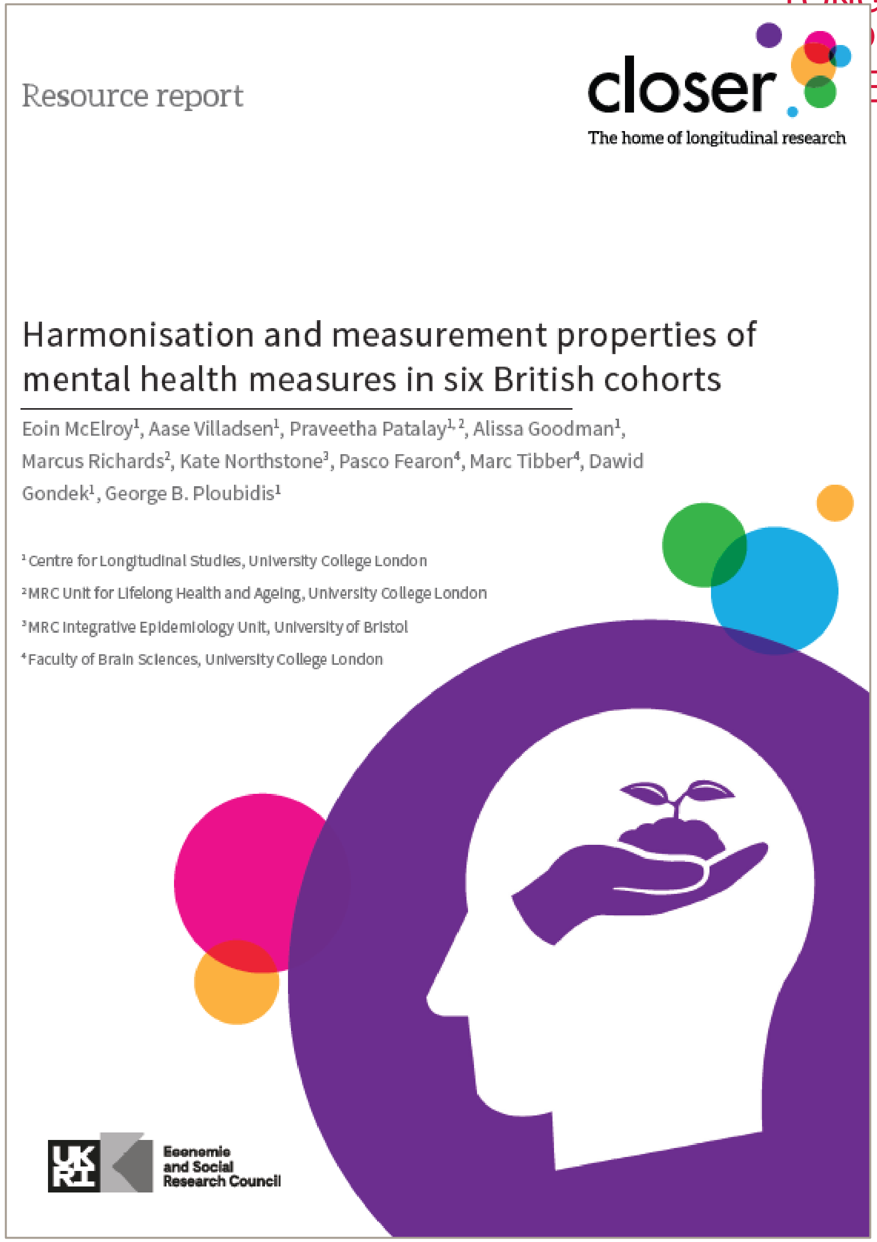
Data Note 2011/1

Centre for Longitudinal Studies

Deriving Highest Qualification in NCDS and BCS70

Brian Dodgeon and Samantha Parsons

Centre for Longitudinal Studies
Institute of Education
20 Bedford Way
London WC1H 0AL
Tel: 020 7612 6860
Fax: 020 7612 6880
Email cls@ioe.ac.uk
Web http://www.cls.ioe.ac.uk



Resource report

closer
The home of longitudinal research

Harmonisation and measurement properties of mental health measures in six British cohorts

Eoin McElroy[1], Aase Villadsen[1], Praveetha Patalay[1,2], Alissa Goodman[1], Marcus Richards[2], Kate Northstone[3], Pasco Fearon[4], Marc Tibber[4], Dawid Gondek[1], George B. Ploubidis[1]

[1] Centre for Longitudinal Studies, University College London
[2] MRC Unit for Lifelong Health and Ageing, University College London
[3] MRC Integrative Epidemiology Unit, University of Bristol
[4] Faculty of Brain Sciences, University College London

UKRI Economic and Social Research Council

Longitudinal data

CENTRE FOR LONGITUDINAL STUDIES

- Data collected over time **within the same units** (individuals, households…)
  - Describe and visualise trajectories
  - Compare developmental growth / change over time
  - Sometimes, *same* measures over time (but sometimes need harmonisation/linking approaches)
  - Designs can vary on multiple aspects (duration, frequency of repeated assessments, number of cohorts…)

CENTRE FOR
LONGITUDINAL
STUDIES

Different types of data

- Cross-sectional: multiple variables in a single time-point
- Time-series: single variable in multiple time-points
- Panel: multiple variables in multiple-time points

Longitudinal data

CLS cohort studies

CENTRE FOR
LONGITUDINAL
STUDIES

1958 National Child Development Study

Following the lives of 17,000 people born in a single week in 1958 in Great Britain.

1970 British Cohort Study

Following the lives of 17,000 people born in a single week in 1970 in Great Britain.

Next Steps

Following the lives of 16,000 people in England born in 1989-90.

Millennium Cohort Study

The most recent of Britain's cohort studies, following 19,000 young people born in the UK at the start of the new century.

CENTRE FOR LONGITUDINAL STUDIES

CLS cohort studies: Typical content

Infant	Child	Adult
Household composition Parental social class Birth history Pregnancy & labour Birth outcomes	Household composition Parental social class Parental employment Financial circumstances Housing Health Cognitive tests Emotions & behaviour School experience & attainment Attitudes & expectations	Household composition Social Class Employment Income Housing Relationships & children Health & Mental health Training & qualifications Cognitive tests Attitudes & expectations

Getting started: An introduction to four British cohort studies

15 Nov 2023

WEBINAR

This 90-minute session gives first-time users an overview of the 1958, 1970, Next Steps and millennium cohort studies – unique data resources available for researchers across the biomedical and social sciences.

Different types of data

- Cross-sectional: multiple variables in a single time-point
- Time-series: single variable in multiple time-points
- **Panel: multiple variables in multiple-time points**

Data structure with demonstration in Stata

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

Applied example for this session: MCS

CENTRE FOR
LONGITUDINAL
STUDIES

1958 National Child Development Study

Following the lives of 17,000 people born in a single week in 1958 in Great Britain.

1970 British Cohort Study

Following the lives of 17,000 people born in a single week in 1970 in Great Britain.

Next Steps

Following the lives of 16,000 people in England born in 1989-90.

Millennium Cohort Study

The most recent of Britain's cohort studies, following 19,000 young people born in the UK at the start of the new century.

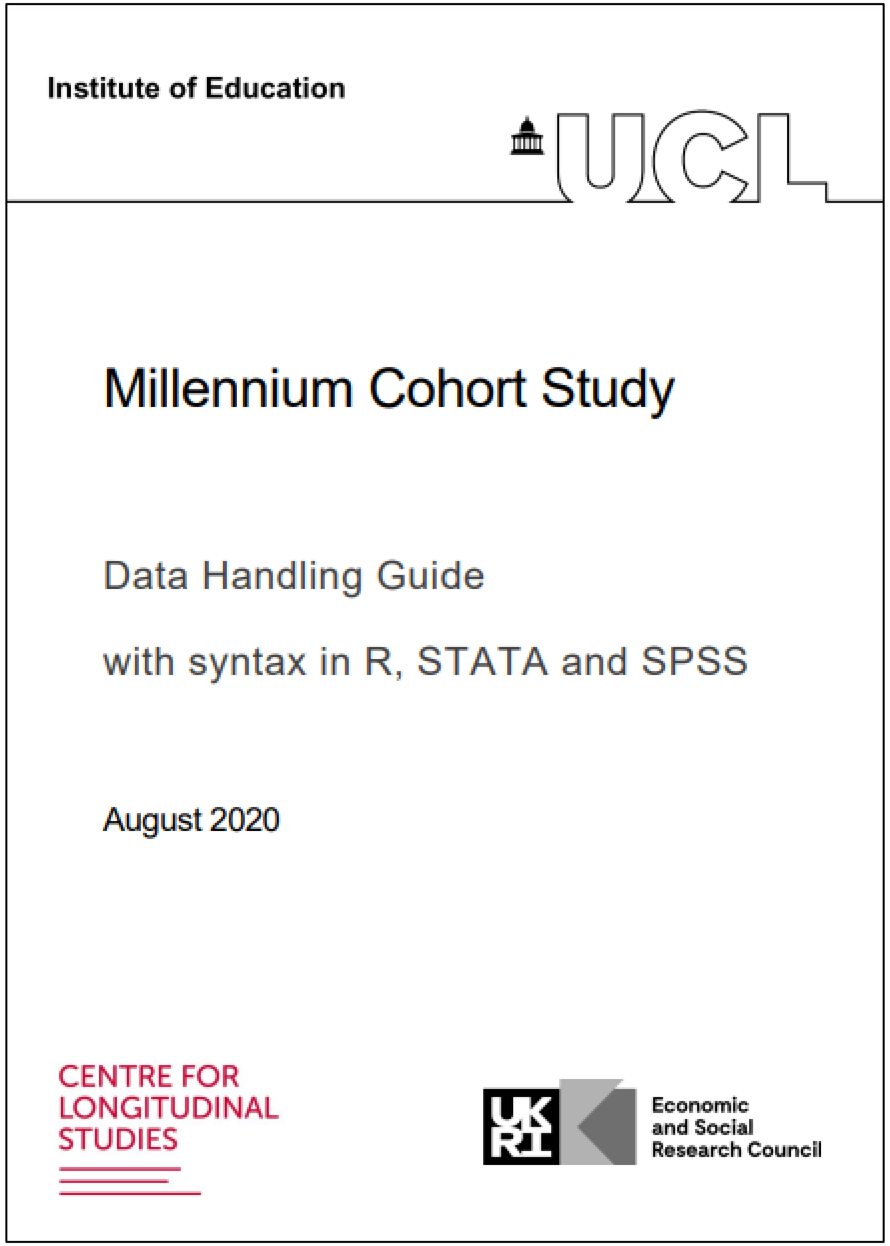
Year	2001	2004	2006	2008	2012	2015	2018	2023
Age	9 months	3	5	7	11	14	17	23

CENTRE FOR
LONGITUDINAL

Important notes!

- Simplified examples:
  - Survey features (clustering, survey weights) have been ignored
  - Only first cohort member of the family or singletons between sweeps 1-7 (ages 9 months to 17 years)
- Although all data shown is anonymous, we have also changed the IDs in the files
- Very helpful documentation →

  https://doc.ukdataservice.ac.uk/doc/4683/mrdoc/pdf/mcs_data_handling_guide_ed1_2020-08-10.pdf


Institute of Education

UCL

Millennium Cohort Study

Data Handling Guide

with syntax in R, STATA and SPSS

August 2020

CENTRE FOR
LONGITUDINAL
STUDIES

UKRI Economic and Social Research Council

CENTRE FOR LONGITUDINAL STUDIES

UK Data Service

Millennium Cohort Study

Abstract FAQ's Resources Access data

Access data

GN 33359
Millennium Cohort Study – Survey and Biomeasures Data

SN	Study description	Explore online	Select
8756	Millennium Cohort Study, Sweeps 1-7, 2001-2019: Exact Participation Dates: Secure Access		☐
8755	Millennium Cohort Study, Sweeps 1-7, 2001-2019: Demographics, Language and Religion: Secure Access		☐
8754	Millennium Cohort Study, Sweeps 1-7, 2001-2019: Self-Reported Health, Behaviour and Fertility: Secure Access		☐
8753	Millennium Cohort Study, Sweeps 1-7, 2001-2019: Socio-Economic, Accommodation and Occupational Data: Secure Access		☐
8682	Millennium Cohort Study: Age 17, Sweep 7, 2018		☐
8172	Millennium Cohort Study: Sweeps 1-7, 2001-2018: Longitudinal Family File		☐
8156	Millennium Cohort Study: Age 14, Sweep 6, 2015		☐
7464	Millennium Cohort Study: Age 11, Sweep 5, 2012		☐
7261	Millennium Cohort Study: Age 9 months, Sweep 1, 2001-2003: Health Visitor Survey		☐
7238	Millennium Cohort Study: Age 7, Sweep 4, 2008: Physical Activity		☐
6411	Millennium Cohort Study: Age 7, Sweep 4, 2008		☐
5795	Millennium Cohort Study: Age 5, Sweep 3, 2006		☐
5559	Millennium Cohort Study: Age 9 months, Sweep 1, 2003: Survey of Mothers who Received Assisted Fertility Treatment		☐
5350	Millennium Cohort Study: Age 3, Sweep 2, 2004		☐
4683	Millennium Cohort Study: Age 9 months, Sweep 1, 2001		☐

https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000031#!/access-data

CENTRE FOR LONGITUDINAL STUDIES

Different types of data

- Cross-sectional: multiple variables in a single time-point
- Time-series: single variable in multiple time-points
- Panel: multiple variables in multiple-time points

Deposited data may be “cross-sectional” (by *sweep*)

CENTRE FOR LONGITUDINAL STUDIES

UK Data Service

Millennium Cohort Study

Abstract | FAQ's | Resources | Access data

Access data

GN 33359
Millennium Cohort Study – Survey and Biomeasures Data

SN	Study description	Explore online	Select
8756	Millennium Cohort Study, Sweeps 1-7, 2001-2019: Exact Participation Dates: Secure Access		☐
8755	Millennium Cohort Study, Sweeps 1-7, 2001-2019: Demographics, Language and Religion: Secure Access		☐
8754	Millennium Cohort Study, Sweeps 1-7, 2001-2019: Self-Reported Health, Behaviour and Fertility: Secure Access		☐
8753	Millennium Cohort Study, Sweeps 1-7, 2001-2019: Socio-Economic, Accommodation and Occupational Data: Secure Access		☐
8682	Millennium Cohort Study: Age 17, Sweep 7, 2018		☐
8172	Millennium Cohort Study: Sweeps 1-7, 2001-2018: Longitudinal Family File		☐
8156	Millennium Cohort Study: Age 14, Sweep 6, 2015		☐
7464	Millennium Cohort Study: Age 11, Sweep 5, 2012		☐
7261	Millennium Cohort Study: Age 9 months, Sweep 1, 2001-2003: Health Visitor Survey		☐
7238	Millennium Cohort Study: Age 7, Sweep 4, 2008: Physical Activity		☐
6411	Millennium Cohort Study: Age 7, Sweep 4, 2008		☐
5795	Millennium Cohort Study: Age 5, Sweep 3, 2006		☐
5559	Millennium Cohort Study: Age 9 months, Sweep 1, 2003: Survey of Mothers who Received Assisted Fertility Treatment		☐
5350	Millennium Cohort Study: Age 3, Sweep 2, 2004		☐
4683	Millennium Cohort Study: Age 9 months, Sweep 1, 2001		☐

https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000031#!/access-data

Merging datasets

CENTRE FOR
LONGITUDINAL
STUDIES

Merging datasets

CENTRE FOR LONGITUDINAL STUDIES

- Need to find the variable (or set of variables) that uniquely identify the individual (unit)
- Unique identifier: e.g., NCDSID, BCSID, NSID, MCSID *
    - MCSID is the identifier of the family!
    - In the examples, due to focus on first cohort member of the family or singleton births, MCSID is unique to the individual

CENTRE FOR
LONGITUDINAL
STUDIES

Merging datasets

- mcs1_cm_derived.dta
- mcs2_cm_derived.dta
- mcs3_cm_derived.dta
- mcs4_cm_derived.dta
- mcs5_cm_derived.dta
- mcs6_cm_derived.dta
- mcs7_cm_derived.dta

	MCSID	ACNUM00	ACNOBA00
1	11985	1st Cohort Member of the family	One baby
2	11986	1st Cohort Member of the family	One baby
3	11991	1st Cohort Member of the family	One baby
4	11992	1st Cohort Member of the family	Twins
5	11992	2nd Cohort Member of the family	Twins
6	11995	1st Cohort Member of the family	One baby
7	11998	1st Cohort Member of the family	One baby
8	11999	1st Cohort Member of the family	One baby

```
. isid MCSID
variable MCSID does not uniquely identify the observations
r(459);

. isid MCSID ACNUM00
```

Merging datasets

CENTRE FOR LONGITUDINAL STUDIES

- mcs1_cm_derived.dta
- mcs2_cm_derived.dta
- mcs3_cm_derived.dta
- mcs4_cm_derived.dta
- mcs5_cm_derived.dta
- mcs6_cm_derived.dta
- mcs7_cm_derived.dta

	MCSID	ACNUM00	ACNOBA00
1	11985	1st Cohort Member of the family	One baby
2	11986	1st Cohort Member of the family	One baby
3	11991	1st Cohort Member of the family	One baby
4	11992	1st Cohort Member of the family	Twins
5	11992	2nd Cohort Member of the family	Twins
6	11995	1st Cohort Member of the family	One baby
7	11998	1st Cohort Member of the family	One baby
8	11999	1st Cohort Member of the family	One baby

```
. keep if ACNUM00 == 1
(246 observations deleted)

. isid MCSID
```

CENTRE FOR
LONGITUDINAL
STUDIES

Merging datasets

Only first cohort members

mcs1_cm_derived_first.dta →

Variables

Filter variables here

Name	Label
MCSID	MCS Research pseudo-ID - Anonymised Family/House...
ACNUM00	Cohort Member number within an MCS family
ACNOBA00	Number of CMs in household
ADCEEA00_R30	DV Cohort Member Ethnic Group merged (England) [...
ADCEWA00_R30	DV Cohort Member Ethnic Group (merged) (Wales) [...
ADCESA00_R30	DV Cohort Member Ethnic Group (merged) (Scotland)...
ADCENA00_R30	DV Cohort Member Ethnic Group (merged) (N Ireland)...
ADC06E00	DV Cohort Member Ethnic Group - 6 category Census ...
ADC11E00	DV Cohort Member Ethnic Group - 11 category Census...
ADC08E00	DV Cohort Member Ethnic Group - 8 category classific...
ADBWGT00	DV Cohort Member birth weight in kilos
ADERLT00	DV Birth of Cohort Member: Number of days early or l...
ADGEST00	DV Cohort Member Gestation time in days
ADLSTW00	DV Cohort Member most recent weight in kilos
ADAGLW00	DV Cohort Member Age post-term in days when last ...

CENTRE FOR
LONGITUDINAL
STUDIES

Merging datasets

Only first cohort members

mcs2_cm_derived_first.dta →

Variables

Filter variables here

Name	Label
MCSID	MCS Research pseudo-ID - Anonymised Family/House...
BCNUM00	Cohort Member number within an MCS family
BDNOBA00	Number of CMs in household
BDCEEA00_R30	DV Cohort Member Ethnic Group (England) - new fam...
BDC06E00	DV Cohort Member Ethnic Group - 6 category Census ...
BDC11E00	DV Cohort Member Ethnic Group - 11 category Census...
BDC08E00	DV Cohort Member Ethnic Group - 8 category classific...
BDCEWA00_R30	DV Cohort Member Ethnic Group (merged) (Wales) [...
BDCESA00_R30	DV Cohort Member Ethnic Group (merged) (Scotland...
BDCSBI00	Child Social Behaviour Questionnaire (Independence-...
BDCSBE00	Child Social Behaviour Questionnaire (Emotional-Dysr...
BDMPIA00	DV CM Child-Parent Relationship Scale (CPRS) MAIN
BDMVLD00	DV CM CPRS Number of valid responses (max 15) MAI...
BDMMPT00	DV CM CPRS Number of imputed responses (max 3) ...
BDPPIA00	DV CM Child-Parent Relationship Scale (CPRS) PARTN...

Merging datasets

CENTRE FOR LONGITUDINAL STUDIES

Only first cohort members

mcs1_cm_derived_first.dta → Open (this is now the *master* –or active- dataset)

mcs2_cm_derived_first.dta → This is the *using* dataset (a.k.a.: the dataset we want to merge to the *master* or active one)

Merging datasets

Only first cohort members

mcs1_cm_derived_first.dta → Open (this is now the *master* –or active- dataset)

mcs2_cm_derived_first.dta → This is the *using* dataset (a.k.a.: the dataset we want to merge to the *master* or active one)

The main command

```
merge 1:1 MCSID using mcs2_cm_derived_first
```

Merging datasets

Only first cohort members

mcs1_cm_derived_first.dta → Open (this is now the *master* –or active- dataset)

mcs2_cm_derived_first.dta → This is the *using* dataset (a.k.a.: the dataset we want to merge to the *master* or active one)

The main command

```
merge 1:1 MCSID using mcs2_cm_derived_first
```

This means that the variable(s) used as identifiers identify only 1 observation in both the *master* and *using* dataset

Merging datasets

Only first cohort members

mcs1_cm_derived_first.dta → Open (this is now the *master* –or active- dataset)

mcs2_cm_derived_first.dta → This is the *using* dataset (a.k.a.: the dataset we want to merge to the *master* or active one)

The main command

```
merge 1:1 MCSID using mcs2_cm_derived_first
```

This means that the variable(s) used as identifiers identify only 1 observation in both the *master* and *using* dataset

This is the variable used as identifier

Merging datasets

Only first cohort members

mcs1_cm_derived_first.dta → Open (this is now the *master* –or active- dataset)

mcs2_cm_derived_first.dta → This is the *using* dataset (a.k.a.: the dataset we want to merge to the *master* or active one)

The main command

```
merge 1:1 MCSID using mcs2_cm_derived_first
```

This means that the variable(s) used as identifiers identify only 1 observation in both the *master* and *using* dataset

This is the variable used as identifier

This precedes the name of the dataset to be merged to the *master* –or active- one

CENTRE FOR LONGITUDINAL STUDIES

Merging datasets

Only first cohort members

mcs1_cm_derived_first.dta → Open (this is now the *master* –or active- dataset)

mcs2_cm_derived_first.dta → This is the *using* dataset (a.k.a.: the dataset we want to merge to the *master* or active one)

The main command

```
merge 1:1 MCSID using mcs2_cm_derived_first
```

This is the name of the *using* dataset

This means that the variable(s) used as identifiers identify only 1 observation in both the *master* and *using* dataset

This is the variable used as identifier

This precedes the name of the dataset to be merged to the *master* –or active- one

Merging datasets

CENTRE FOR LONGITUDINAL STUDIES

Only first cohort members

mcs1_cm_derived_first.dta → Open (this is now the *master* –or active- dataset)

mcs2_cm_derived_first.dta → This is the *using* dataset (a.k.a.: the dataset we want to merge to the *master* or active one)

```
merge 1:1 MCSID using mcs2_cm_derived_first

    Result                           Number of obs
    -----------------------------------------
    Not matched                          4,343
        from master                      3,652  (_merge==1)
        from using                         691  (_merge==2)

    Matched                             14,888  (_merge==3)
    -----------------------------------------
```

Observations that have not been matched (present only in the *master* or *using* datasets)

Observations that have been matched (present both in the *master* and *using* datasets)

Unless instructed not to do so (with the option *nogenerate*), Stata will create a variable flagging the result of the merge

CENTRE FOR LONGITUDINAL STUDIES

Merging datasets

Merged dataset contains variables from both datasets, stored under the same MCSID

If needed, this dataset can be saved

Variables

Filter variables here

Name	Label
MCSID	MCS Research pseudo-ID - Anonymised Family/House...
ACNUM00	Cohort Member number within an MCS family
ACNOBA00	Number of CMs in household
ADCEEA00_R30	DV Cohort Member Ethnic Group merged (England) [...
ADCEWA00_R30	DV Cohort Member Ethnic Group (merged) (Wales) [...
ADCESA00_R30	DV Cohort Member Ethnic Group (merged) (Scotland)...
ADCENA00_R30	DV Cohort Member Ethnic Group (merged) (N Ireland)...
ADC06E00	DV Cohort Member Ethnic Group - 6 category Census ...
ADC11E00	DV Cohort Member Ethnic Group - 11 category Census...
ADC08E00	DV Cohort Member Ethnic Group - 8 category classific...
ADBWGT00	DV Cohort Member birth weight in kilos
ADERLT00	DV Birth of Cohort Member: Number of days early or l...
ADGEST00	DV Cohort Member Gestation time in days
ADLSTW00	DV Cohort Member most recent weight in kilos
ADAGLW00	DV Cohort Member Age post-term in days when last ...
BCNUM00	Cohort Member number within an MCS family
BDNOBA00	Number of CMs in household
BDCEEA00_R30	DV Cohort Member Ethnic Group (England) - new fam...

Merging datasets

mcs1_cm_derived_first.dta
mcs2_cm_derived_first.dta
mcs3_cm_derived_first.dta
mcs4_cm_derived_first.dta
mcs5_cm_derived_first.dta
mcs6_cm_derived_first.dta
mcs7_cm_derived_first.dta

Note:

- The dataset has one row per individual
- Variables from different waves are denoted by the starting letter (this may be different in other scenarios)

'wide' layout

- May need to `rename` variables before merging if they have the same name in both *master* and *using* datasets

```
save mcs_cm_derived_first_wide, replace
```

Longitudinal data layout: wide vs long

Longitudinal data layout: wide vs long

CENTRE FOR LONGITUDINAL STUDIES

- Wide (or unstacked):
  - Single row per unit
  - Observations from different time points denoted with different names (e.g., bmi5, bmi7, bmi11…)

ID	var1_t1	var2_t1	…	var1_t2	var2_t2	…	var1_t3	var2_t3	…
1									
2									
3									
…									

Longitudinal data layout: wide vs long

CENTRE FOR LONGITUDINAL STUDIES

- Wide (or unstacked):
  - Single row per unit
  - Observations from different time points denoted with different names (e.g., bmi5, bmi7, bmi11…)

ID	var1_t1	var2_t1	…	var1_t2	var2_t2	…	var1_t3	var2_t3	…
1									
2									
3									
…									

- Long (or stacked or narrow):
  - Multiple rows per unit
  - Time is a variable (e.g., age = 5, 7, 11…)

ID	time	var1	var2	…
1	1			
1	2			
1	…			
2	1			
2	2			
2	…			
3	1			
3	2			
3	…			

Longitudinal data layout: <u>wide</u>

mcs_cm_derived_first_wide.dta

```
browse MCSID BEMOTION CEMOTION
       DDEMOTION FEMOTION GEMOTION
```

	MCSID	BEMOTION	CEMOTION	DDEMOTION	FEMOTION	GEMOTION
1	11985	4	3	4	.	.
2	11986	2	2	1	4	2
3	11991	0	1	0	2	1
4	11992	1	.	.	.	.
5	11995	2	1	1	.	.
6	11998	0	.	.	.	.
7	11999	1	3	3	2	3

- One row per unit (in this case, individual)
- Repeated assessments of the same variables (e.g., parent-reported emotional subscale of the SDQ questionnaire, range: 0 [minimum] – 10 [maximum]) denoted by different names
  - Often variable names in panel studies contain a letter or number denoting the wave; if not, may want to include it before merging
- 'Wide', 'unstacked'

Longitudinal data layout: long

CENTRE FOR LONGITUDINAL STUDIES

Harmonised Height, Weight and BMI in Five Longitudinal Cohort Studies: Millennium Cohort Study

Details | Documentation | Resources | Access data

Details

Title:	Harmonised Height, Weight and BMI in Five Longitudinal Cohort Studies: Millennium Cohort Study
Alternative title:	CLOSER Work Package 1; MCS
Study number (SN):	8550
Access:	These data are safeguarded
Persistent identifier (DOI):	10.5255/UKDA-SN-8550-1
Series:	CLOSER
Data creator(s):	Cohort and Longitudinal Studies Enhancement Resources

- Harmonised data from multiple sweeps

https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8550

CENTRE FOR LONGITUDINAL STUDIES

Longitudinal data layout: long

mcs_closer_wp1.dta

```
browse mcsid visitage sex bmi
```

	mcsid	visitage	sex	bmi
1	11985	0	Female	.
2	11985	1	Female	.
3	11985	3	Female	16.9
4	11985	5	Female	16.2
5	11985	7	Female	15.61
6	11985	11	Female	.
7	11986	0	Male	.
8	11986	1	Male	.
9	11986	3	Male	17.72
10	11986	5	Male	15.72
11	11986	7	Male	17.32
12	11986	11	Male	20.3
13	11991	0	Male	.
14	11991	1	Male	.
15	11991	3	Male	18.55
16	11991	5	Male	18.17
17	11991	7	Male	15.92
18	11991	11	Male	17.01

- Units (individuals) have multiple rows
- ‘Long’, ‘stacked’, ‘narrow’

CENTRE FOR LONGITUDINAL STUDIES

Longitudinal data layout: long

	mcsid	visitage	sex	bmi
1	11985	0	Female	.
2	11985	1	Female	.
3	11985	3	Female	16.9
4	11985	5	Female	16.2
5	11985	7	Female	15.61
6	11985	11	Female	.
7	11986	0	Male	.
8	11986	1	Male	.
9	11986	3	Male	17.72
10	11986	5	Male	15.72
11	11986	7	Male	17.32
12	11986	11	Male	20.3
13	11991	0	Male	.
14	11991	1	Male	.
15	11991	3	Male	18.55
16	11991	5	Male	18.17
17	11991	7	Male	15.92
18	11991	11	Male	17.01

Time is a variable

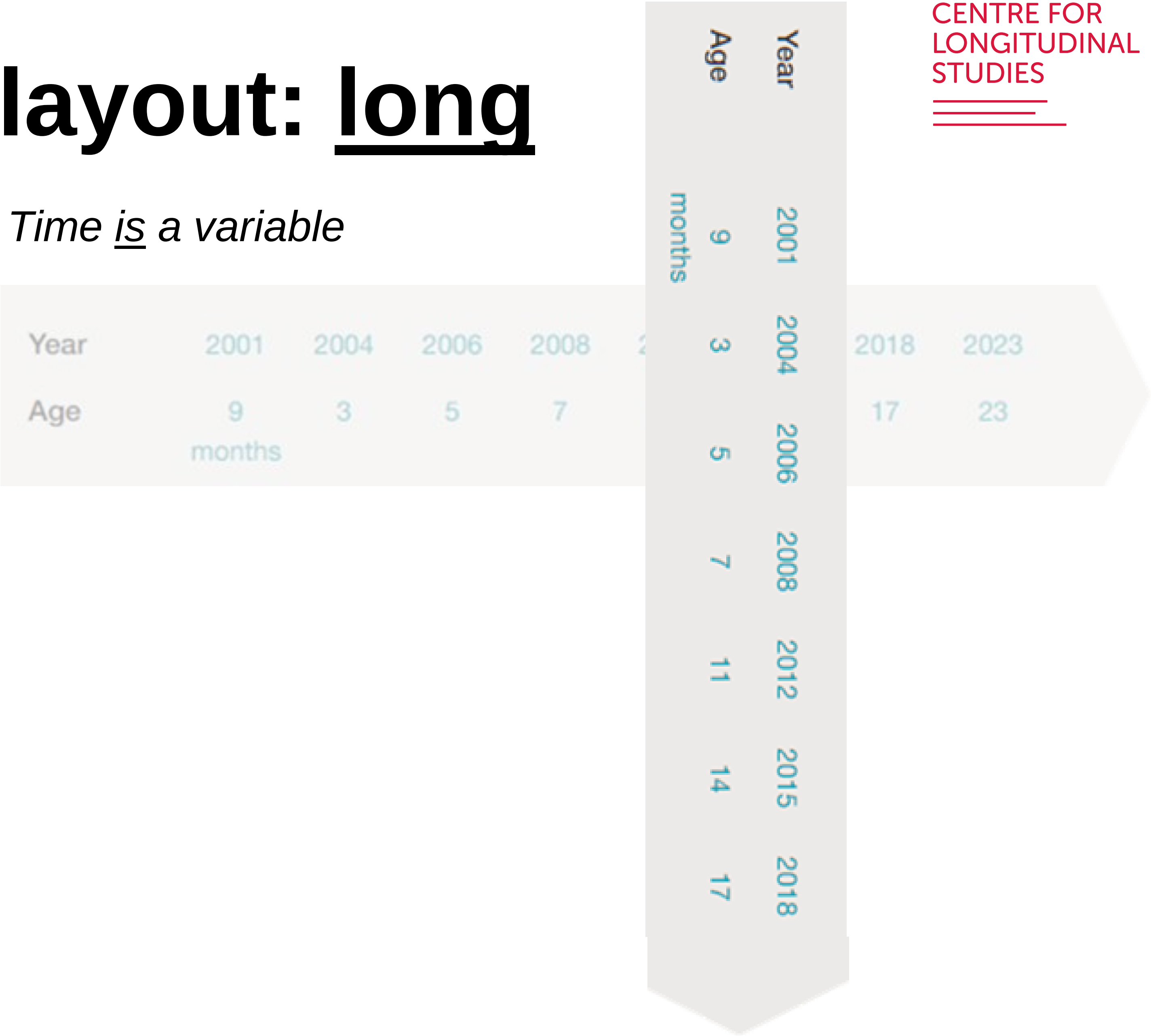
Year	2001	2004	2006	2008	2012	2015	2018	2023
Age	9 months	3	5	7	11	14	17	23

CENTRE FOR LONGITUDINAL STUDIES

Longitudinal data layout: <u>long</u>

	mcsid	visitage	sex	bmi
1	11985	0	Female	.
2	11985	1	Female	.
3	11985	3	Female	16.9
4	11985	5	Female	16.2
5	11985	7	Female	15.61
6	11985	11	Female	.
7	11986	0	Male	.
8	11986	1	Male	.
9	11986	3	Male	17.72
10	11986	5	Male	15.72
11	11986	7	Male	17.32
12	11986	11	Male	20.3
13	11991	0	Male	.
14	11991	1	Male	.
15	11991	3	Male	18.55
16	11991	5	Male	18.17
17	11991	7	Male	15.92
18	11991	11	Male	17.01

Time <u>is</u> a variable



Longitudinal data layout: long

	mcsid	visitage	sex	bmi
1	11985	0	Female	.
2	11985	1	Female	.
3	11985	3	Female	16.9
4	11985	5	Female	16.2
5	11985	7	Female	15.61
6	11985	11	Female	.
7	11986	0	Male	.
8	11986	1	Male	.
9	11986	3	Male	17.72
10	11986	5	Male	15.72
11	11986	7	Male	17.32
12	11986	11	Male	20.3
13	11991	0	Male	.
14	11991	1	Male	.
15	11991	3	Male	18.55
16	11991	5	Male	18.17
17	11991	7	Male	15.92
18	11991	11	Male	17.01

Time-invariant variables

Sex assigned at birth
(as binary variable)

CENTRE FOR
LONGITUDINAL
STUDIES

CENTRE FOR LONGITUDINAL STUDIES

Longitudinal data layout: long

	mcsid	visitage	sex	bmi
1	11985	0	Female	.
2	11985	1	Female	.
3	11985	3	Female	16.9
4	11985	5	Female	16.2
5	11985	7	Female	15.61
6	11985	11	Female	.
7	11986	0	Male	.
8	11986	1	Male	.
9	11986	3	Male	17.72
10	11986	5	Male	15.72
11	11986	7	Male	17.32
12	11986	11	Male	20.3
13	11991	0	Male	.
14	11991	1	Male	.
15	11991	3	Male	18.55
16	11991	5	Male	18.17
17	11991	7	Male	15.92
18	11991	11	Male	17.01

Time-varying variables

3 repeated observations

4 repeated observations

4 repeated observations

Examples using wide layout

mcs_cm_derived_first_wide.dta

CENTRE FOR
LONGITUDINAL
STUDIES

Examples using <u>wide</u> layout

CENTRE FOR LONGITUDINAL STUDIES

- `summarize` variables from multiple time-points

```
. sum BEMOTION CEMOTION DDEMOTION FEMOTION GEMOTION
```

Variable	Obs	Mean	Std. dev.	Min	Max
BEMOTION	15,579	1.254188	1.573782	-1	10
CEMOTION	15,236	1.315371	1.636546	-1	10
DDEMOTION	13,847	1.463277	1.800602	-1	10
FEMOTION	11,717	1.947768	2.17288	-1	10
GEMOTION	10,831	1.606961	2.335044	-1	10

Examples using wide layout

- `recode` multiple variables that require the same recoding

```
. recode BEMOTION CEMOTION DDEMOTION FEMOTION GEMOTION (-1 = .)
(846 changes made to BEMOTION)
(519 changes made to CEMOTION)
(414 changes made to DDEMOTION)
(388 changes made to FEMOTION)
(1,560 changes made to GEMOTION)

. sum BEMOTION CEMOTION DDEMOTION FEMOTION GEMOTION
```

Variable	Obs	Mean	Std. dev.	Min	Max
BEMOTION	14,733	1.383629	1.520021	0	10
CEMOTION	14,717	1.397024	1.605305	0	10
DDEMOTION	13,433	1.539195	1.774631	0	10
FEMOTION	11,329	2.048725	2.138997	0	10
GEMOTION	9,271	2.045626	2.243622	0	10

Examples using wide layout

- `rename` multiple variables

```
. rename (BEMOTION CEMOTION DDEMOTION FEMOTION GEMOTION) (EMOTION3 EMOTION5 EMOTION7
>  EMOTION14 EMOTION17)

. sum EMOTION*

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
    EMOTION3 |     14,733    1.383629    1.520021          0         10
    EMOTION5 |     14,717    1.397024    1.605305          0         10
    EMOTION7 |     13,433    1.539195    1.774631          0         10
   EMOTION14 |     11,329    2.048725    2.138997          0         10
   EMOTION17 |      9,271    2.045626    2.243622          0         10
```

CENTRE FOR LONGITUDINAL STUDIES

Examples using wide layout

- `generate` variables that are conditional on variables from multiple time-points

```
. gen flag = 1 if EMOTION3 != . & EMOTION5 != .
(6,386 missing values generated)

. tab flag, mis

       flag |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |     12,845       66.79       66.79
          . |      6,386       33.21      100.00
------------+-----------------------------------
      Total |     19,231      100.00
```

	MCSID	EMOTION3	EMOTION5	flag
1	11985	4	3	1
2	11986	2	2	1
3	11991	0	1	1
4	11992	1	.	.
5	11995	2	1	1
6	11998	0	.	.

Examples using <u>wide</u> layout

- Obtain correlations across variables over time

```
. pwcorr EMOTION*
```

	EMOTION3	EMOTION5	EMOTION7	EMOTI~14	EMOTI~17
EMOTION3	1.0000				
EMOTION5	0.4212	1.0000			
EMOTION7	0.3503	0.5135	1.0000		
EMOTION14	0.2343	0.3322	0.4081	1.0000	
EMOTION17	0.1890	0.2677	0.3231	0.5677	1.0000

Examples using long layout

mcs_closer_wp1.dta

CENTRE FOR LONGITUDINAL STUDIES

Examples using long layout

- `summarize` all the repeated observations of the same variable (e.g., grand mean)

```
. sum bmi
```

Variable	Obs	Mean	Std. dev.	Min	Max
bmi	44,468	17.22816	2.616493	7.3	61.72

Examples using long layout

- `summarize` the repeated observations by time-point

```
. tabstat bmi, by(visitage)

Summary for variables: bmi
Group variable: visitage (Age at visit/interview)

visitage |      Mean
---------+----------
       0 |         .
       1 |         .
       3 |  16.86595
       5 |  16.38981
       7 |  16.66427
      11 |  19.21847
---------+----------
   Total |  17.22816
--------------------
```

Examples using long layout

- `recode` the repeated observations of the same variable

```
recode bmi ///
    (min/24.9 = 0 "BMI below 25") ///
    (25/29.9 = 1 "BMI between 25-29.9") ///
    (30/max = 2 "BMI 30+"), ///
    gen(bmi_cat)
```

```
. tab bmi_cat
```

RECODE of bmi (Body mass index (kg/m2))	Freq.	Percent	Cum.
BMI below 25	43,553	97.94	97.94
BMI between 25-29.9	788	1.77	99.71
BMI 30+	127	0.29	100.00
Total	44,468	100.00	

Examples using long layout

CENTRE FOR
LONGITUDINAL
STUDIES

- `tabulate` across time-points

```
. tab visitage bmi_cat
```

Age at visit/interview	RECODE of bmi (Body mass index (kg/m2))			
	BMI below	BMI betwe	BMI 30+	Total
3	11,319	26	6	11,351
5	11,926	30	4	11,960
7	10,863	80	8	10,951
11	9,445	652	109	10,206
Total	43,553	788	127	44,468

Examples using long layout

- `generate` new time variables

```
. tab visitage if bmi != .
```

Age at visit/inter view	Freq.	Percent	Cum.
3	11,351	25.53	25.53
5	11,960	26.90	52.42
7	10,951	24.63	77.05
11	10,206	22.95	100.00
Total	44,468	100.00	

Examples using <u>long</u> layout

- `generate` new time variables

```
. gen visityear = visitage + 2001

. tab visityear if bmi != .
```

visityear	Freq.	Percent	Cum.
2004	11,351	25.53	25.53
2006	11,960	26.90	52.42
2008	10,951	24.63	77.05
2012	10,206	22.95	100.00
Total	44,468	100.00	

Examples using long layout

CENTRE FOR LONGITUDINAL STUDIES

- Check number of repeated observations available per individual

```
. keep if bmi != .
(36,394 observations deleted)

. sort mcsid visitage

. by mcsid: gen nobs = _n

. by mcsid: gen totobs = _N

. browse mcsid visitage bmi nobs totobs
```

	mcsid	visitage	bmi	nobs	totobs
1	11985	3	16.9	1	3
2	11985	5	16.2	2	3
3	11985	7	15.61	3	3
4	11986	3	17.72	1	4
5	11986	5	15.72	2	4
6	11986	7	17.32	3	4
7	11986	11	20.3	4	4
8	11991	3	18.55	1	4
9	11991	5	18.17	2	4
10	11991	7	15.92	3	4
11	11991	11	17.01	4	4

Moving across long and wide layouts

CENTRE FOR
LONGITUDINAL
STUDIES

Moving across long and wide layouts

CENTRE FOR LONGITUDINAL STUDIES

- `reshape` is a useful command to move across long and wide layouts
- Requires some dataset preparation for reshape
- Keep the variables of interest

reshape: from long to wide

Full dataset

mcs_closer_wp1.dta

Name	Label
mcsid	MCS Research pseudo-ID
stid	Study identifier
visitage	Age at visit/interview
sex	Sex of study member
xage	Actual age
wt	Harmonised weight
wtself	Weight, measured or self-report
wtimp	Weight, imperial or metric
wtpre	Weight (precision)
ht	Harmonised height
htself	Height, measured or self-report
htimp	Height, imperial or metric
htpre	Height (precision)
bmi	Body mass index (kg/m2)

CENTRE FOR
LONGITUDINAL
STUDIES

reshape: from long to wide

Full dataset

mcs_closer_wp1.dta

	mcsid	stid	visitage	sex	xage	wt	wtself	wtimp	wtpre	ht	htself	htimp	htpre	bmi
1	11985	MCS	0											
2	11985	MCS	1											
3	11985	MCS	3											
4	11985	MCS	5											
5	11985	MCS	7											
6	11985	MCS	11											
7	11986	MCS	0											
8	11986	MCS	1											
9	11986	MCS	3											
10	11986	MCS	5											
11	11986	MCS	7											
12	11986	MCS	11											
13	11991	MCS	0											
14	11991	MCS	1											
15	11991	MCS	3											
16	11991	MCS	5											
17	11991	MCS	7											
18	11991	MCS	11											

reshape: from long to wide

Full dataset

mcs_closer_wp1.dta

```
. tabstat bmi, by(visitage)

Summary for variables: bmi
Group variable: visitage (Age at visit/interview)
```

visitage	Mean
0	.
1	.
3	16.86595
5	16.38981
7	16.66427
11	19.21847
Total	17.22816

reshape: from long to wide

Reduced dataset

mcs_closer_wp1.dta

```
. keep mcsid visitage sex bmi

. keep if inrange(visitage,3,11)
(26,954 observations deleted)

. tabstat bmi, by(visitage)

Summary for variables: bmi
Group variable: visitage (Age at visit/interview)

visitage |      Mean
---------+----------
       3 |  16.86595
       5 |  16.38981
       7 |  16.66427
      11 |  19.21847
---------+----------
   Total |  17.22816
--------------------
```

```
. recode bmi ///
>          (min/24.99 = 0 "BMI below 25") ///
>          (25/29.99 = 1 "BMI between 25-29.9") ///
>          (30/max = 2 "BMI 30+"), ///
>          gen(bmi_cat)
(44,468 differences between bmi and bmi_cat)
```

CENTRE FOR LONGITUDINAL STUDIES

reshape: from long to wide

Reduced dataset

mcs_closer_wp1.dta

	mcsid	visitage	sex	bmi	bmi_cat
1	11985	3	Female	16.9	BMI below 25
2	11985	5	Female	16.2	BMI below 25
3	11985	7	Female	15.61	BMI below 25
4	11985	11	Female	.	.
5	11986	3	Male	17.72	BMI below 25
6	11986	5	Male	15.72	BMI below 25
7	11986	7	Male	17.32	BMI below 25
8	11986	11	Male	20.3	BMI below 25
9	11991	3	Male	18.55	BMI below 25
10	11991	5	Male	18.17	BMI below 25
11	11991	7	Male	15.92	BMI below 25
12	11991	11	Male	17.01	BMI below 25

CENTRE FOR LONGITUDINAL STUDIES

reshape: from long to wide

Reduced dataset

mcs_closer_wp1.dta

	mcsid	visitage	sex	bmi	bmi_cat
1	11985	3	Female	16.9	BMI below 25
2	11985	5	Female	16.2	BMI below 25
3	11985	7	Female	15.61	BMI below 25
4	11985	11	Female	.	.
5	11986	3	Male	17.72	BMI below 25
6	11986	5	Male	15.72	BMI below 25
7	11986	7	Male	17.32	BMI below 25
8	11986	11	Male	20.3	BMI below 25
9	11991	3	Male	18.55	BMI below 25
10	11991	5	Male	18.17	BMI below 25
11	11991	7	Male	15.92	BMI below 25
12	11991	11	Male	17.01	BMI below 25

Main command

```
reshape wide bmi bmi_cat, i(mcsid) j(visitage)
```

reshape: from long to wide

Reduced dataset

mcs_closer_wp1.dta

	mcsid	visitage	sex	bmi	bmi_cat
1	11985	3	Female	16.9	BMI below 25
2	11985	5	Female	16.2	BMI below 25
3	11985	7	Female	15.61	BMI below 25
4	11985	11	Female	.	.
5	11986	3	Male	17.72	BMI below 25
6	11986	5	Male	15.72	BMI below 25
7	11986	7	Male	17.32	BMI below 25
8	11986	11	Male	20.3	BMI below 25
9	11991	3	Male	18.55	BMI below 25
10	11991	5	Male	18.17	BMI below 25
11	11991	7	Male	15.92	BMI below 25
12	11991	11	Male	17.01	BMI below 25

To reshape into a 'wide' layout

Main command

```
reshape wide bmi bmi_cat, i(mcsid) j(visitage)
```

reshape: from long to wide

CENTRE FOR LONGITUDINAL STUDIES

Reduced dataset

mcs_closer_wp1.dta

	mcsid	visitage	sex	bmi	bmi_cat
1	11985	3	Female	16.9	BMI below 25
2	11985	5	Female	16.2	BMI below 25
3	11985	7	Female	15.61	BMI below 25
4	11985	11	Female	.	.
5	11986	3	Male	17.72	BMI below 25
6	11986	5	Male	15.72	BMI below 25
7	11986	7	Male	17.32	BMI below 25
8	11986	11	Male	20.3	BMI below 25
9	11991	3	Male	18.55	BMI below 25
10	11991	5	Male	18.17	BMI below 25
11	11991	7	Male	15.92	BMI below 25
12	11991	11	Male	17.01	BMI below 25

Exhaustive list of time-varying variables

To reshape into a 'wide' layout

Main command

```
reshape wide bmi bmi_cat, i(mcsid) j(visitage)
```

CENTRE FOR
LONGITUDINAL
STUDIES

reshape: from long to wide

Reduced dataset

mcs_closer_wp1.dta

Unique identifier

	mcsid	visitage	sex	bmi	bmi_cat
1	11985	3	Female	16.9	BMI below 25
2	11985	5	Female	16.2	BMI below 25
3	11985	7	Female	15.61	BMI below 25
4	11985	11	Female	.	.
5	11986	3	Male	17.72	BMI below 25
6	11986	5	Male	15.72	BMI below 25
7	11986	7	Male	17.32	[illegible]
8	11986	11	Male	20.3	[illegible]
9	11991	3	Male	18.55	[illegible]
10	11991	5	Male	[illegible]	[illegible] below 25
11	11991	7	Male	15.92	BMI below 25
12	11991	11	[illegible]	17.01	BMI below 25

Exhaustive list of time-varying variables

To reshape into a 'wide' layout

Main command

Variable(s) that uniquely identify each unit

```
reshape wide bmi bmi_cat, i(mcsid) j(visitage)
```

CENTRE FOR
LONGITUDINAL
STUDIES

reshape: from long to wide

Reduced dataset

mcs_closer_wp1.dta

Time variable

Unique identifier

	mcsid	visitage	sex	bmi	bmi_cat
1	11985	3	Female	16.9	BMI below 25
2	11985	5	Female	16.2	BMI below 25
3	11985	7	Female	15.61	BMI below 25
4	11985	11	Female	.	.
5	11986	3	Male	17.72	BMI below 25
6	11986	5	Male	15.72	BMI below 25
7	11986	7	Male	17.32	[illegible]
8	11986	11	Male	20.3	[illegible]
9	11991	3	Male	18.55	[illegible]
10	11991	5	Male	[illegible]	[illegible] below 25
11	11991	7	Male	15.92	[illegible]
12	11991	11	[illegible]	17.01	[illegible]

Exhaustive list of time-varying variables

To reshape into a 'wide' layout

Main command

Variable(s) that uniquely identify each unit

Existing variable capturing time

```
reshape wide bmi bmi_cat, i(mcsid) j(visitage)
```

reshape: from long to wide

Reduced dataset

mcs_closer_wp1.dta

```
. reshape wide bmi bmi_cat, i(mcsid) j(visitage)
(j = 3 5 7 11)

Data                               Long   ->   Wide
-----------------------------------------------------------------------------
Number of observations           53,908   ->   13,477
Number of variables                   5   ->   10
j variable (4 values)          visitage   ->   (dropped)
xij variables:
                                    bmi   ->   bmi3 bmi5 ... bmi11
                                bmi_cat   ->   bmi_cat3 bmi_cat5 ... bmi_cat11
-----------------------------------------------------------------------------
```

reshape: from long to wide

CENTRE FOR LONGITUDINAL STUDIES

Reduced dataset

mcs_closer_wp1.dta

	mcsid	visitage	sex	bmi	bmi_cat
1	11985	3	Female	16.9	BMI below 25
2	11985	5	Female	16.2	BMI below 25
3	11985	7	Female	15.61	BMI below 25
4	11985	11	Female	.	.
5	11986	3	Male	17.72	BMI below 25
6	11986	5	Male	15.72	BMI below 25
7	11986	7	Male	17.32	BMI below 25
8	11986	11	Male	20.3	BMI below 25
9	11991	3	Male	18.55	BMI below 25
10	11991	5	Male	18.17	BMI below 25
11	11991	7	Male	15.92	BMI below 25
12	11991	11	Male	17.01	BMI below 25

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

reshape: from long to wide

CENTRE FOR
LONGITUDINAL
STUDIES

Reduced dataset

mcs_closer_wp1.dta

	mcsid	visitage	sex	bmi	bmi_cat
1	11985	3	Female	16.9	BMI below 25
2	11985	5	Female	16.2	BMI below 25
3	11985	7	Female	15.61	BMI below 25
4	11985	11	Female	.	.
5	11986	3	Male	17.72	BMI below 25
6	11986	5	Male	15.72	BMI below 25
7	11986	7	Male	17.32	BMI below 25
8	11986	11	Male	20.3	BMI below 25
9	11991	3	Male	18.55	BMI below 25
10	11991	5	Male	18.17	BMI below 25
11	11991	7	Male	15.92	BMI below 25
12	11991	11	Male	17.01	BMI below 25

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

CENTRE FOR LONGITUDINAL STUDIES

reshape: from long to wide

Reduced dataset

mcs_closer_wp1.dta

	mcsid	visitage	sex	bmi	bmi_cat
1	11985	3	Female	16.9	BMI below 25
2	11985	5	Female	16.2	BMI below 25
3	11985	7	Female	15.61	BMI below 25
4	11985	11	Female	.	.
5	11986	3	Male	17.72	BMI below 25
6	11986	5	Male	15.72	BMI below 25
7	11986	7	Male	17.32	BMI below 25
8	11986	11	Male	20.3	BMI below 25
9	11991	3	Male	18.55	BMI below 25
10	11991	5	Male	18.17	BMI below 25
11	11991	7	Male	15.92	BMI below 25
12	11991	11	Male	17.01	BMI below 25

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

reshape: from long to wide

CENTRE FOR LONGITUDINAL STUDIES

Reduced dataset

mcs_closer_wp1.dta

```
. pwcorr bmi3 bmi5 bmi7 bmi11

             |     bmi3     bmi5     bmi7    bmi11
-------------+------------------------------------
        bmi3 |   1.0000
        bmi5 |   0.6726   1.0000
        bmi7 |   0.5820   0.7960   1.0000
       bmi11 |   0.4340   0.6385   0.7929   1.0000
```

reshape: from long to wide

Reduced dataset

mcs_closer_wp1.dta

- The dataset in wide layout can now be saved (e.g., mcs_harmonised_bmi_3to11_wide.dta)

```
. save mcs_harmonised_bmi_3to11_wide
```

- Can go back to long layout by typing `reshape long` (only works if the dataset has been reshaped in the same session)

reshape: from wide to long

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

Name	Label
mcsid	MCS Research pseudo-ID
bmi3	3 bmi
bmi_cat3	3 bmi_cat
bmi5	5 bmi
bmi_cat5	5 bmi_cat
bmi7	7 bmi
bmi_cat7	7 bmi_cat
bmi11	11 bmi
bmi_cat11	11 bmi_cat
sex	Sex of study member

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

reshape: from wide to long

CENTRE FOR LONGITUDINAL STUDIES

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

Name	Label
mcsid	MCS Research pseudo-ID
bmi3	3 bmi
bmi_cat3	3 bmi_cat
bmi5	5 bmi
bmi_cat5	5 bmi_cat
bmi7	7 bmi
bmi_cat7	7 bmi_cat
bmi11	11 bmi
bmi_cat11	11 bmi_cat
sex	Sex of study member

In *wide* layout there is no variable capturing time at each observation

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

reshape: from wide to long

CENTRE FOR LONGITUDINAL STUDIES

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

Name	Label
mcsid	MCS Research pseudo-ID
bmi3	3 bmi
bmi_cat3	3 bmi_cat
bmi5	5 bmi
bmi_cat5	5 bmi_cat
bmi7	7 bmi
bmi_cat7	7 bmi_cat
bmi11	11 bmi
bmi_cat11	11 bmi_cat
sex	Sex of study member

In *wide* layout there is no variable capturing time at each observation

Rather, it is contained in the variables' names (in this case, suffixes)

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

CENTRE FOR
LONGITUDINAL
STUDIES

reshape: from wide to long

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

help reshape

```
[D] reshape — Convert data from wide to long form and vice versa
              (View complete PDF manual entry)


Syntax

    Overview

          long
                                              wide
        i  j  stub
                                             i   stub1  stub2
        1  1   4.1          reshape
        1  2   4.5        <--------->        1    4.1    4.5
        2  1   3.3                           2    3.3    3.0
        2  2   3.0


        To go from long to wide:

                                           j existing variable
                                          /
                reshape wide stub, i(i) j(j)

        To go from wide to long:

                reshape long stub, i(i) j(j)
                                          \
                                           j new variable
```

reshape: from wide to long

CENTRE FOR LONGITUDINAL STUDIES

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

Main command

```
reshape long bmi bmi_cat, i(mcsid) j(age)
```

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

CENTRE FOR LONGITUDINAL STUDIES

reshape: from wide to long

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

To reshape into a 'long' layout

Main command

```
reshape long bmi bmi_cat, i(mcsid) j(age)
```

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

reshape: from wide to long

CENTRE FOR LONGITUDINAL STUDIES

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

To reshape into a 'long' layout

Exhaustive list of **stubs** of time-varying variables

Main command

```
reshape long bmi bmi_cat, i(mcsid) j(age)
```

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

reshape: from wide to long

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

To reshape into a 'long' layout

Exhaustive list of **stubs** of time-varying variables

Variable(s) that uniquely identify each unit

Main command

```
reshape long bmi bmi_cat, i(mcsid) j(age)
```

Unique identifier

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

CENTRE FOR
LONGITUDINAL
STUDIES

reshape: from wide to long

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

To reshape into a 'long' layout

Exhaustive list of **stubs** of time-varying variables

Variable(s) that uniquely identify each unit

New variable to be created capturing time

Main command

```
reshape long bmi bmi_cat, i(mcsid) j(age)
```

Unique identifier

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

reshape: from wide to long

CENTRE FOR LONGITUDINAL STUDIES

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

To reshape into a 'long' layout

Exhaustive list of **stubs** of time-varying variables

Variable(s) that uniquely identify each unit

New variable to be created capturing time

Main command

```
reshape long bmi bmi_cat, i(mcsid) j(age)
```

Unique identifier

Which will take on these values

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

reshape: from wide to long

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

```
. reshape long bmi bmi_cat, i(mcsid) j(age)
(j = 3 5 7 11)

Data                               Wide   ->   Long
-----------------------------------------------------
Number of observations           13,477   ->   53,908
Number of variables                  10   ->   5
j variable (4 values)                     ->   age
xij variables:
                      bmi3 bmi5 ... bmi11  ->   bmi
          bmi_cat3 bmi_cat5 ... bmi_cat11  ->   bmi_cat
-----------------------------------------------------
```

	mcsid	age	bmi	bmi_cat	sex
1	11985	3	16.9	BMI below 25	Female
2	11985	5	16.2	BMI below 25	Female
3	11985	7	15.61	BMI below 25	Female
4	11985	11	.	.	Female
5	11986	3	17.72	BMI below 25	Male
6	11986	5	15.72	BMI below 25	Male
7	11986	7	17.32	BMI below 25	Male
8	11986	11	20.3	BMI below 25	Male
9	11991	3	18.55	BMI below 25	Male
10	11991	5	18.17	BMI below 25	Male
11	11991	7	15.92	BMI below 25	Male
12	11991	11	17.01	BMI below 25	Male

reshape: from wide to long

CENTRE FOR
LONGITUDINAL
STUDIES

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

- The dataset in long layout can now be saved (e.g., mcs_harmonised_bmi_3to11_long.dta)

- Can go back to wide layout by typing `reshape wide` (only works if the dataset has been reshaped in the same session)

CENTRE FOR LONGITUDINAL STUDIES

reshape: from wide to long

Reduced dataset

mcs_harmonised_bmi_3to11_wide.dta

Name	Label
mcsid	MCS Research pseudo-ID
bmi3	3 bmi
bmi_cat3	3 bmi_cat
bmi5	5 bmi
bmi_cat5	5 bmi_cat
bmi7	7 bmi
bmi_cat7	7 bmi_cat
bmi11	11 bmi
bmi_cat11	11 bmi_cat
sex	Sex of study member

In *wide* layout there is no variable capturing time at each observation

Rather, it is contained in the variables' names (in this case, suffixes)

If not, need to make sure it is

mcsid	bmi3	bmi_cat3	bmi5	bmi_cat5	bmi7	bmi_cat7	bmi11	bmi_cat11	sex
11985	16.9	BMI below 25	16.2	BMI below 25	15.61	BMI below 25	.	.	Female
11986	17.72	BMI below 25	15.72	BMI below 25	17.32	BMI below 25	20.3	BMI below 25	Male
11991	18.55	BMI below 25	18.17	BMI below 25	15.92	BMI below 25	17.01	BMI below 25	Male

Remember this?

- `rename` multiple variables

```
. rename (BEMOTION CEMOTION DDEMOTION FEMOTION GEMOTION) (EMOTION3 EMOTIONS5 EMOTION7
>  EMOTION14 EMOTION17)

. sum EMOTION*

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
    EMOTION3 |     14,733    1.383629    1.520021          0         10
    EMOTION5 |     14,717    1.397024    1.605305          0         10
    EMOTION7 |     13,433    1.539195    1.774631          0         10
   EMOTION14 |     11,329    2.048725    2.138997          0         10
   EMOTION17 |      9,271    2.045626    2.243622          0         10
```

Remember this?

- `rename` multiple variables

```
. rename (BEMOTION CEMOTION DDEMOTION FEMOTION GEMOTION) (EMOTION3 EMOTION5 EMOTION7
>  EMOTION14 EMOTION17)

. sum EMOTION*

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
    EMOTION3 |     14,733    1.383629    1.520021          0         10
    EMOTION5 |     14,717    1.397024    1.605305          0         10
    EMOTION7 |     13,433    1.539195    1.774631          0         10
   EMOTION14 |     11,329    2.048725    2.138997          0         10
   EMOTION17 |      9,271    2.045626    2.243622          0         10
```

- Set up the suffixes in a way that is helpful and informative: e.g., this reflects the age at the specific sweep (they don't necessarily need to be at the end of the variable name, see `help reshape`)

CENTRE FOR
LONGITUDINAL
STUDIES

Remember this?

- `rename` multiple variables

```
. rename (BEMOTION CEMOTION DDEMOTION FEMOTION GEMOTION) (EMOTION3 EMOTION5 EMOTION7
>  EMOTION14 EMOTION17)

. sum EMOTION*
```

Variable	Obs	Mean	Std. dev.	Min	Max
EMOTION3	14,733	1.383629	1.520021	0	10
EMOTION5	14,717	1.397024	1.605305	0	10
EMOTION7	13,433	1.539195	1.774631	0	10
EMOTION14	11,329	2.048725	2.138997	0	10
EMOTION17	9,271	2.045626	2.243622	0	10

- Make sure the *stubs* are consistent within variable

Longitudinal data structure: summary

CENTRE FOR LONGITUDINAL STUDIES

- Importance of unique identifiers
- Merging datasets (1:1, additional options)
- Wide layout: one row per unit, no variable capturing time
- Long layout: multiple rows per unit, variable(s) capturing time
- Different layouts → different data management and analysis opportunities
    - Some analyses require specific layouts (e.g., long data for multilevel modelling, wide data for structural equation modelling)

CENTRE FOR
LONGITUDINAL
STUDIES

Questions?

Longitudinal Data Visualisation with demonstration in Stata

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

CENTRE FOR
LONGITUDINAL
STUDIES

Motivation: Describing patterns of change over time

- Qualitative and quantitative change
- Change within a unit (intraindividual change) versus differences between units (interindividual differences)
- Intergenerational change
- Short-term versus long-term change
- Focus on description, prediction, or explanation of change

Outline

- Introduce simple and useful tools to descriptive and visualise longitudinal data
  - (Artificial) distinction between categorical and continuous variables
  - Longitudinal description, rather than cross-sectional (e.g. Table 1 of academic article)
- Categorical Variables:
  - Descriptive statistics, transition probability matrix
  - Visualization tools
    - Lasagne plots
- Continuous Variables:
  - Descriptive statistics, correlation and linear regression
  - Visualization tools
    - Box/Violin/Spaghetti/Lasagne plots
    - Diagnostic plots (histograms, kernels, symmetry/quintile/Qnorm plots)

Before we start

- Objective: Demonstrate the use of tools that are useful to understand your data before moving into more complex modelling techniques.

- Set up:
  - Data: The Millennium Cohort Study
  - Balanced sample (or no attrition) n=6427; ages: 3, 5, 7, 14 ,17
  - No missing data
  - Variables represent the truth and are measured without errors

Working example:

- Understand longitudinal patterns of children's Body Mass Index.

Descriptive and Visualisation tools

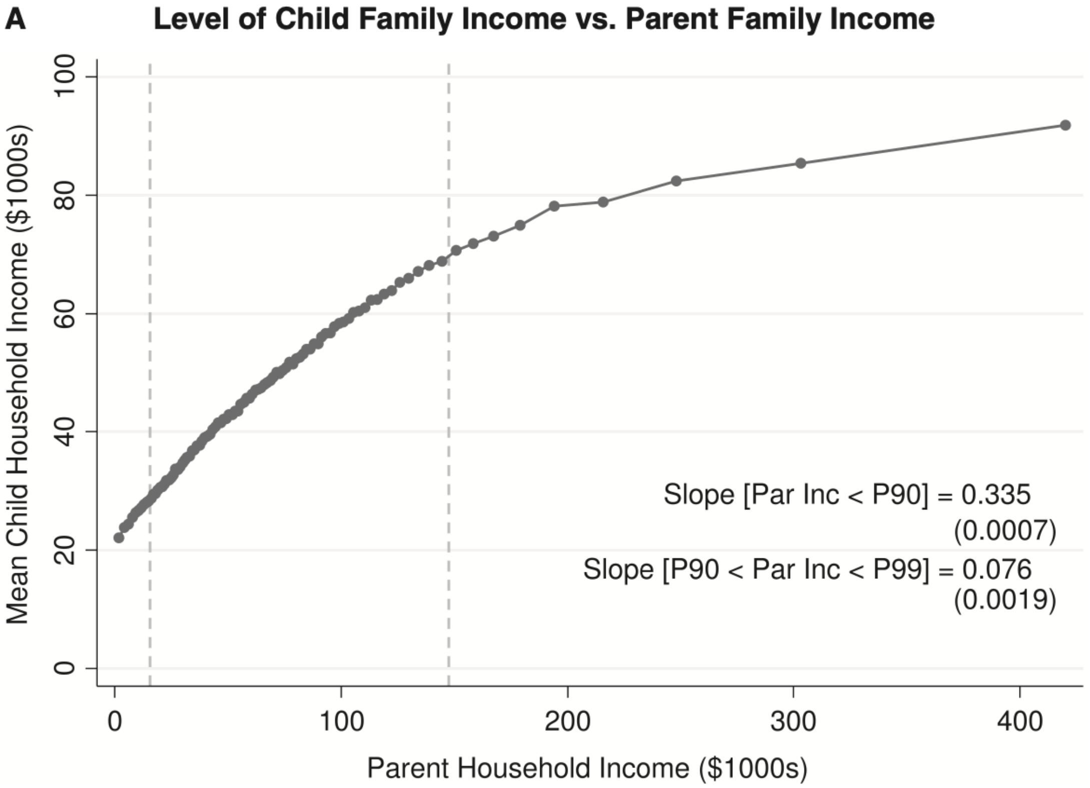
CENTRE FOR LONGITUDINAL STUDIES

Categorical variables

Plot and marginal distribution table of weight level over survey wave

Jones et al (2014), Visualising and modelling changes in categorical variables in longitudinal studies, BMC Medical Research Methodology.

Continuous variables



Association between Children’s and Parents’ Incomes.

Chetty et al (2014), Where is the land of opportunity? The geography of intergenerational mobility in the United States, Quarterly Journal of Economics

Categorical variables

Include a limited number of possible values that represent a qualitative property of units (Socioeconomic Status, Ethnicity, Income quantiles, among others.)

- Descriptive statistics, transition probability matrix
- Lasagne plot

CENTRE FOR LONGITUDINAL STUDIES

Descriptive statistics

- Frequency tables
- Cross tables

MCS

mcs_bmi_wide_clswebinar.dta

	mcsid	sex	bmi3	bmi5	bmi7	bmi11	bmi14	bmi17
1	1	male	17.36	15.72	17.32	20.3	19.64	19.6
2	2	male	18.26	18.17	15.92	17.01	18.95	21.89

	mcsid	sex	bmic3	bmic5	bmic7	bmic11	bmic14	bmic17
1	1	male	Normal	Normal	Normal	Normal	Normal	Normal
2	2	male	Overweight	Overweight	Normal	Normal	Normal	Normal

mcs_bmi_long_clswebinar.dta

	mcsid	age	sex	bmi	bmic
1	1	3	male	17.36	Normal
2	1	5	male	15.72	Normal
3	1	7	male	17.32	Normal
4	1	11	male	20.3	Normal
5	1	14	male	19.64	Normal
6	1	17	male	19.6	Normal
7	2	3	male	18.26	Overweight
8	2	5	male	18.17	Overweight
9	2	7	male	15.92	Normal
10	2	11	male	17.01	Normal
11	2	14	male	18.95	Normal
12	2	17	male	21.89	Normal

```
table age, stat(fvpercent bmic ) nformat(%5.2f)
```

	BMI categories		
	Normal	Overweight	Obese
Age at MCS sweep			
3	83.55	12.56	3.89
5	80.89	14.98	4.12
7	82.37	13.15	4.48
11	75.59	19.37	5.04
14	75.56	18.39	6.05
17	72.27	18.36	9.37
Total	78.37	16.14	5.49

MCS

CENTRE FOR
LONGITUDINAL
STUDIES

mcs_bmi_wide_clswebinar.dta

	mcsid	sex	bmi3	bmi5	bmi7	bmi11	bmi14	bmi17
1	1	male	17.36	15.72	17.32	20.3	19.64	19.6
2	2	male	18.26	18.17	15.92	17.01	18.95	21.89

	mcsid	sex	bmic3	bmic5	bmic7	bmic11	bmic14	bmic17
1	1	male	Normal	Normal	Normal	Normal	Normal	Normal
2	2	male	Overweight	Overweight	Normal	Normal	Normal	Normal

mcs_bmi_long_clswebinar.dta

	mcsid	age	sex	bmi	bmic
1	1	3	male	17.36	Normal
2	1	5	male	15.72	Normal
3	1	7	male	17.32	Normal
4	1	11	male	20.3	Normal
5	1	14	male	19.64	Normal
6	1	17	male	19.6	Normal
7	2	3	male	18.26	Overweight
8	2	5	male	18.17	Overweight
9	2	7	male	15.92	Normal
10	2	11	male	17.01	Normal
11	2	14	male	18.95	Normal
12	2	17	male	21.89	Normal

```
. tab bmic5 bmic7, cell nofreq
```

BMI categories (age 5)	BMI categories (age 7) Normal	Overweigh	Obese	Total
Normal	75.73	4.79	0.37	80.89
Overweight	6.13	7.34	1.51	14.98
Obese	0.51	1.01	2.60	4.12
Total	82.37	13.15	4.48	100.00

CENTRE FOR LONGITUDINAL STUDIES

Transition probability matrix

Sweep 2

Sweep 1

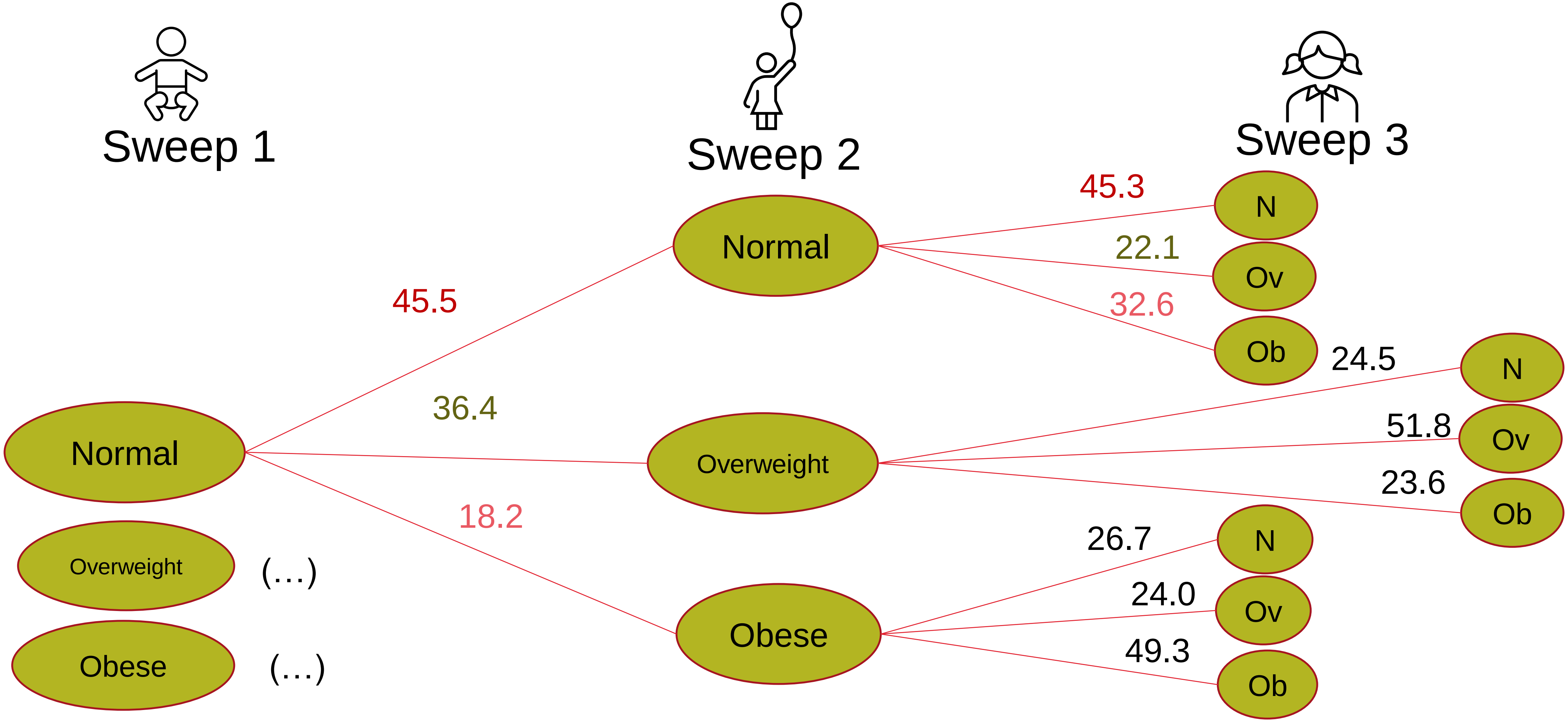
n	Normal	Overweight	Obese	Total
Normal	50	40	20	110
Overweight	40	50	25	115
Obese	5	20	30	55

	Normal	Overweight	Obese	Total
Normal	45.5	36.4	18.2	100.0
Overweight	34.8	43.5	21.7	100.0
Obese	9.1	36.4	54.5	100.0

Pr(High class | Low class) = Probability of being obese at sweep 3 given that I had normal weight in sweep 1
= (20/110)x100 = 18.2

Transition probability matrix

CENTRE FOR
LONGITUDINAL
STUDIES



Transition probability matrix

- Useful to study stability of states over time
- Continuous variables can be categorised (e.g. deciles). Although we lose information, it gives us results that are easier to explain.
- Capture non-linearities as opposed to correlations
- It is possible to compute transition probabilities for panels with t>2

Transition probability matrix - Stata

CENTRE FOR
LONGITUDINAL
STUDIES

```
. xtset mcsid age

Panel variable: mcsid (strongly balanced)
 Time variable: age, 3 to 17, but with gaps
         Delta: 1 unit


. xttab bmic
```

bmic	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
Normal	30222	78.37	6112	95.10	82.41
Overweig	6222	16.14	2780	43.26	37.30
Obese	2118	5.49	992	15.43	35.58
Total	38562	100.00	9884	153.79	65.02
			(n = 6427)		

The total within of 65% is the normalized between weighted average of the within percents, that is, (6112 * 95% + 2780*43 + 992 * 15%)/ 9884.
It is a measure of the overall stability of the BMI category variable.

Transition probability matrix - Stata

CENTRE FOR LONGITUDINAL STUDIES

```
. xttrans bmic, freq
```

BMI categories	BMI categories 1	2	3	Total
1	23,102 90.32	2,306 9.02	169 0.66	25,577 100.00
2	1,602 31.77	2,713 53.81	727 14.42	5,042 100.00
3	148 9.76	396 26.12	972 64.12	1,516 100.00
Total	24,852 77.34	5,415 16.85	1,868 5.81	32,135 100.00

```
. bys sex: xttrans bmic,
```

-> sex = female

BMI categories	BMI categories 1	2	3	Total
1	89.91	9.49	0.60	100.00
2	30.53	55.40	14.07	100.00
3	7.85	27.40	64.76	100.00
Total	75.84	18.13	6.02	100.00

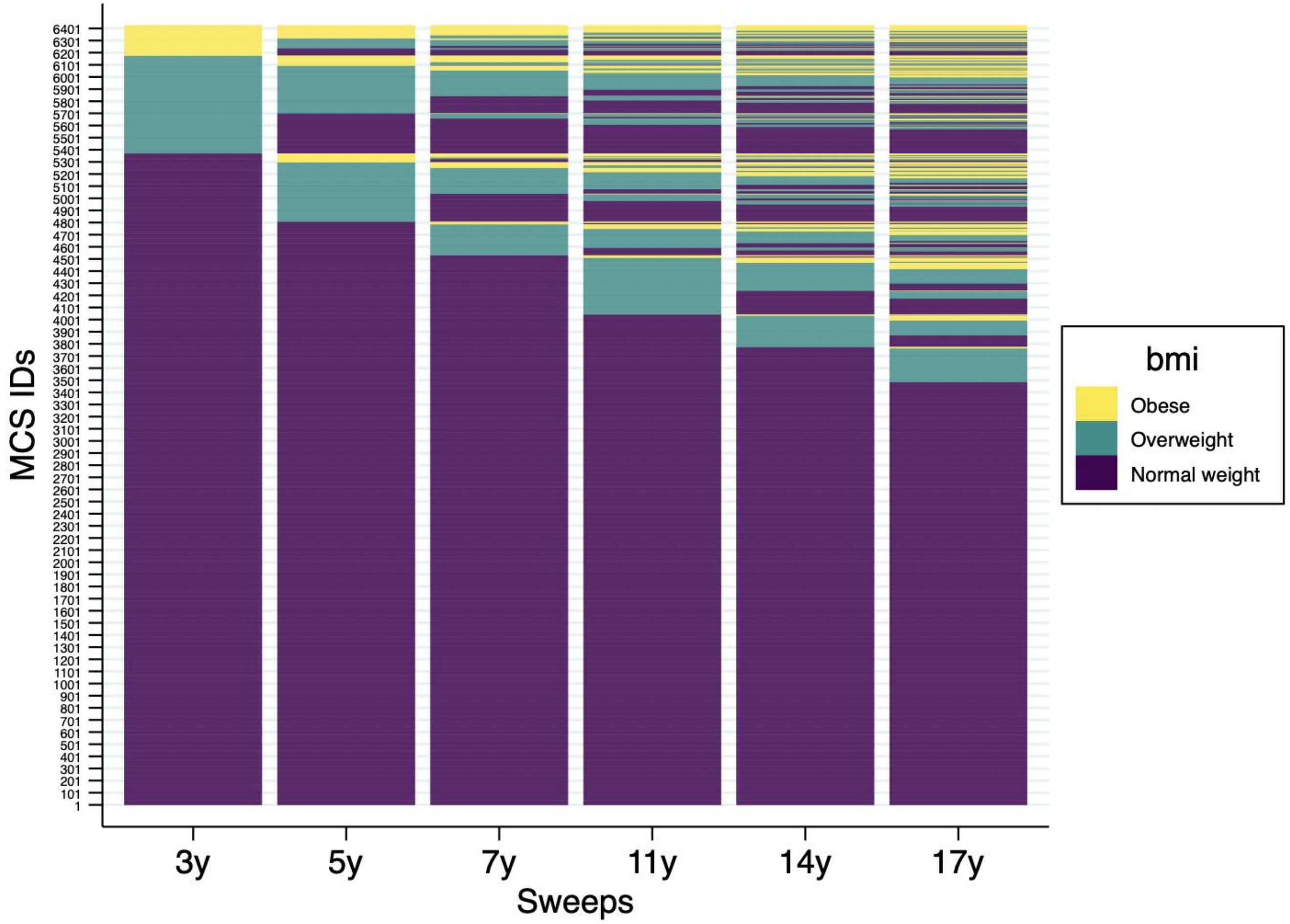
-> sex = male

BMI categories	BMI categories 1	2	3	Total
1	90.74	8.54	0.72	100.00
2	33.29	51.87	14.84	100.00
3	11.92	24.68	63.39	100.00
Total	78.89	15.52	5.59	100.00

Lasagne plot for categorical variables

CENTRE FOR
LONGITUDINAL
STUDIES

- A type of heat map
- Each row is a unit (e.g., girls)
- Variable of interest is sorted across time
- Summarise all possible transitions
- Useful when small number of categories, not much with larger number and long periods
- It gives a first look of the data



Lasagne plot for categorical variables

CENTRE FOR LONGITUDINAL STUDIES

mcs_bmi_wide_clswebinar.dta

	mcsid	sex	bmic3	bmic5	bmic7	bmic11	bmic14	bmic17
1	1	male	Normal	Normal	Normal	Normal	Normal	Normal
2	2	male	Overweight	Overweight	Normal	Normal	Normal	Normal

```
sort bmic3 bmic5 bmic7 bmic11 bmic14 bmic17
gen id_sbmic=_n
```

Variable of interest is sorted across time

In *long* layout there is no variable capturing time at each observation

mcs_bmi_long_clswebinar.dta

	mcsid	age	bmic	id_sbmic	t
1	1	3	Normal	1	1
2	1	5	Normal	1	2
3	1	7	Normal	1	3
4	1	11	Normal	1	4
5	1	14	Normal	1	5
6	1	17	Normal	1	6
7	5165	3	Normal	2	1
8	5165	5	Normal	2	2
9	5165	7	Normal	2	3
10	5165	11	Normal	2	4
11	5165	14	Normal	2	5
12	5165	17	Normal	2	6
13	4723	3	Normal	3	1
14	4723	5	Normal	3	2
15	4723	7	Normal	3	3
16	4723	11	Normal	3	4
17	4723	14	Normal	3	5
18	4723	17	Normal	3	6

heatplot [z] y x

```
heatplot bmic id_sbmic t, statistic(asis) ///
  discrete(0.9) ylabel(0(100)6400,labsize(tiny))  ///
  xlabel(1 2 3 4 5 6) cut(1 1.01 2.001  3.1)  ///
  keylabels( , range(1)  ) ///
  xlabel( 1 "3y" 2 "5y" 3 "7y" 4 "11y" 5 "14y" 6 "17y") ///
  xtitle(Sweeps ,size(large)) ytitle(MCS IDs, size(large)) ///
  ylabel(,angle(horizontal )   ) ///
  plotregion(fcolor(white)) graphregion(fcolor(white)) ///
  legend( region(lcolor(black)) subtitle("BMI categories") ///
  ring(0) position(4) col(1) size(large))
```

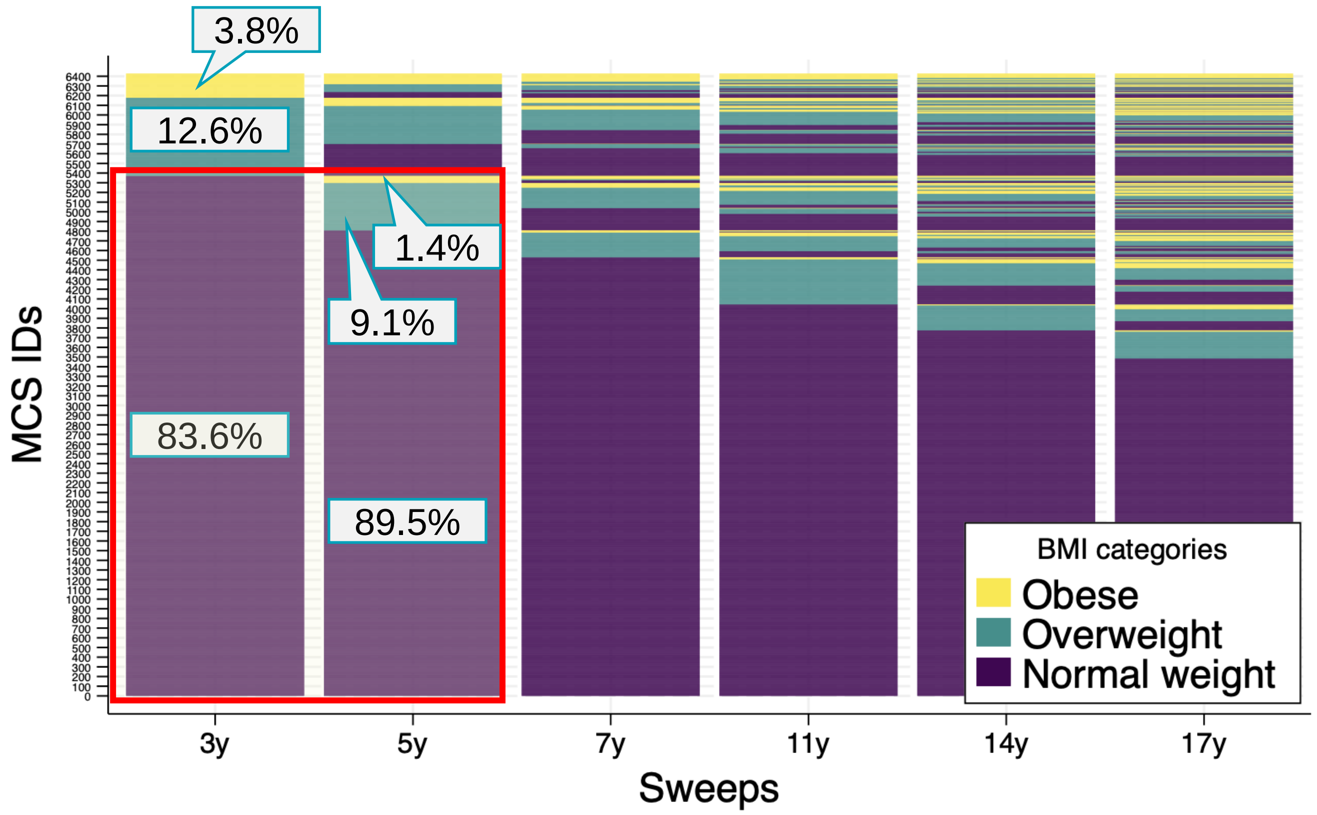
Lasagne plot for categorical variables

CENTRE FOR LONGITUDINAL STUDIES

mcs_bmi_wide_clswebinar.dta

```
heatplot bmic id_sbmic t, statistic(asis) ///
  discrete(0.9) ylabel(0(100)6400,labsize(tiny))  ///
  xlabel(1 2 3 4 5 6) cut(1 1.01 2.001  3.1)  ///
  keylabels( , range(1)  ) ///
  xlabel( 1 "3y" 2 "5y" 3 "7y" 4 "11y" 5 "14y" 6 "17y") ///
  xtitle(Sweeps ,size(large)) ytitle(MCS IDs, size(large)) ///
  ylabel(,angle(horizontal )  ) ///
  plotregion(fcolor(white)) graphregion(fcolor(white)) ///
  legend( region(lcolor(black)) subtitle("BMI categories") ///
  ring(0) position(4) col(1) size(large))
```

```
gr_edit legend.plotregion1.label[1].text = {}
gr_edit legend.plotregion1.label[1].text.Arrpush Obese
gr_edit legend.plotregion1.label[2].text = {}
gr_edit legend.plotregion1.label[2].text.Arrpush Overweight
gr_edit legend.plotregion1.label[3].text = {}
gr_edit legend.plotregion1.label[3].text.Arrpush Normal weight
```



Continuous variables

Can take on any value within a range (Income, body weight)

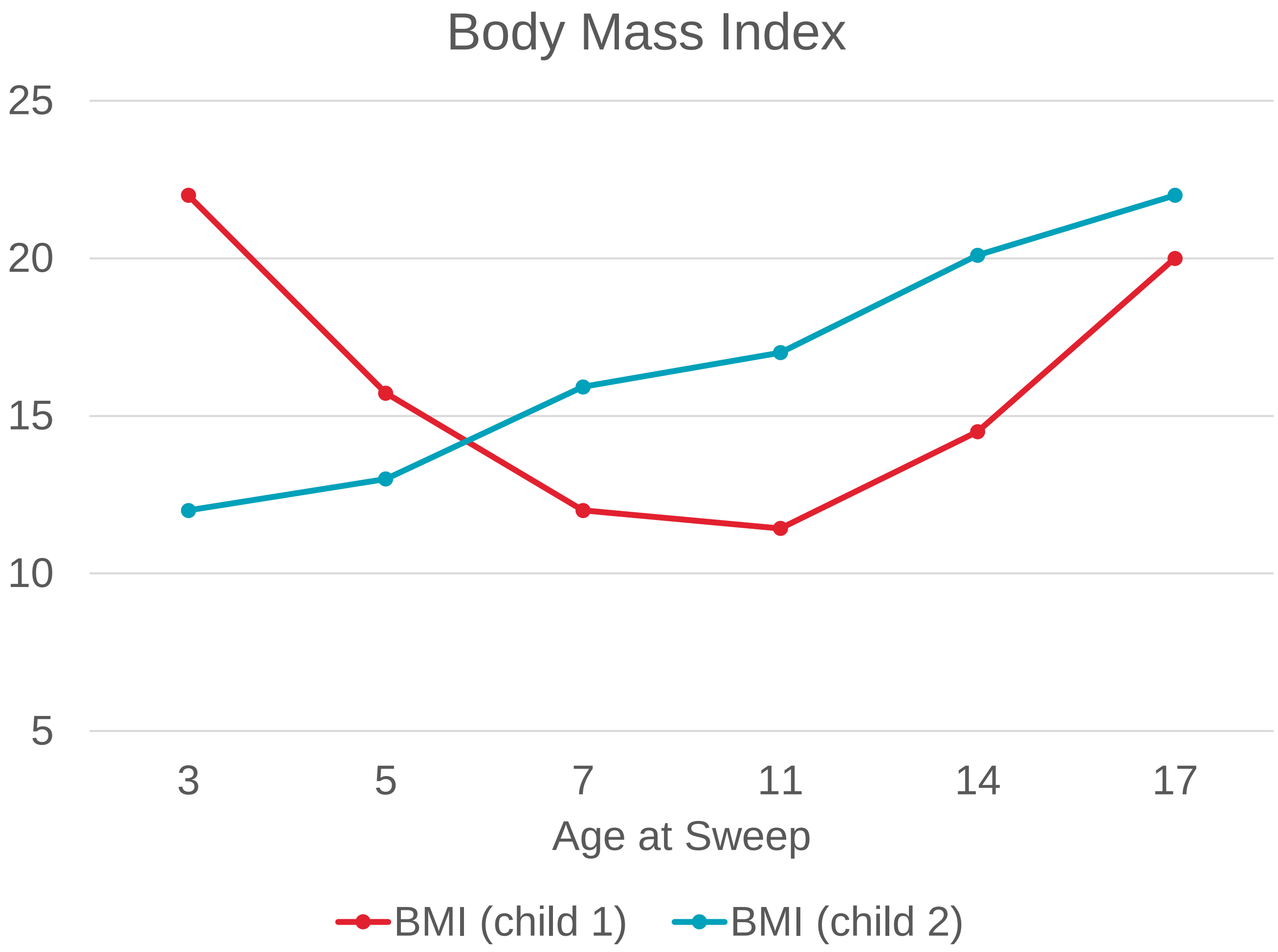
- Descriptive statistics, correlation and linear regression
- Graphical tools
  - Box plots
  - Violin plots
  - Spaghetti and Lasagne plots
  - Diagnostic plots (histograms, kernel plot, symmetry plot, quintile plot, Qnorm plot)

CENTRE FOR LONGITUDINAL STUDIES

Descriptive statistics

1) Cross-sectional Mean/Average

Age at Sweep	BMI (child 1)	BMI (child 2)	?
3	22	12	
5	16	13	
7	12	16	
11	11	17	
14	15	20	
17	20	22	

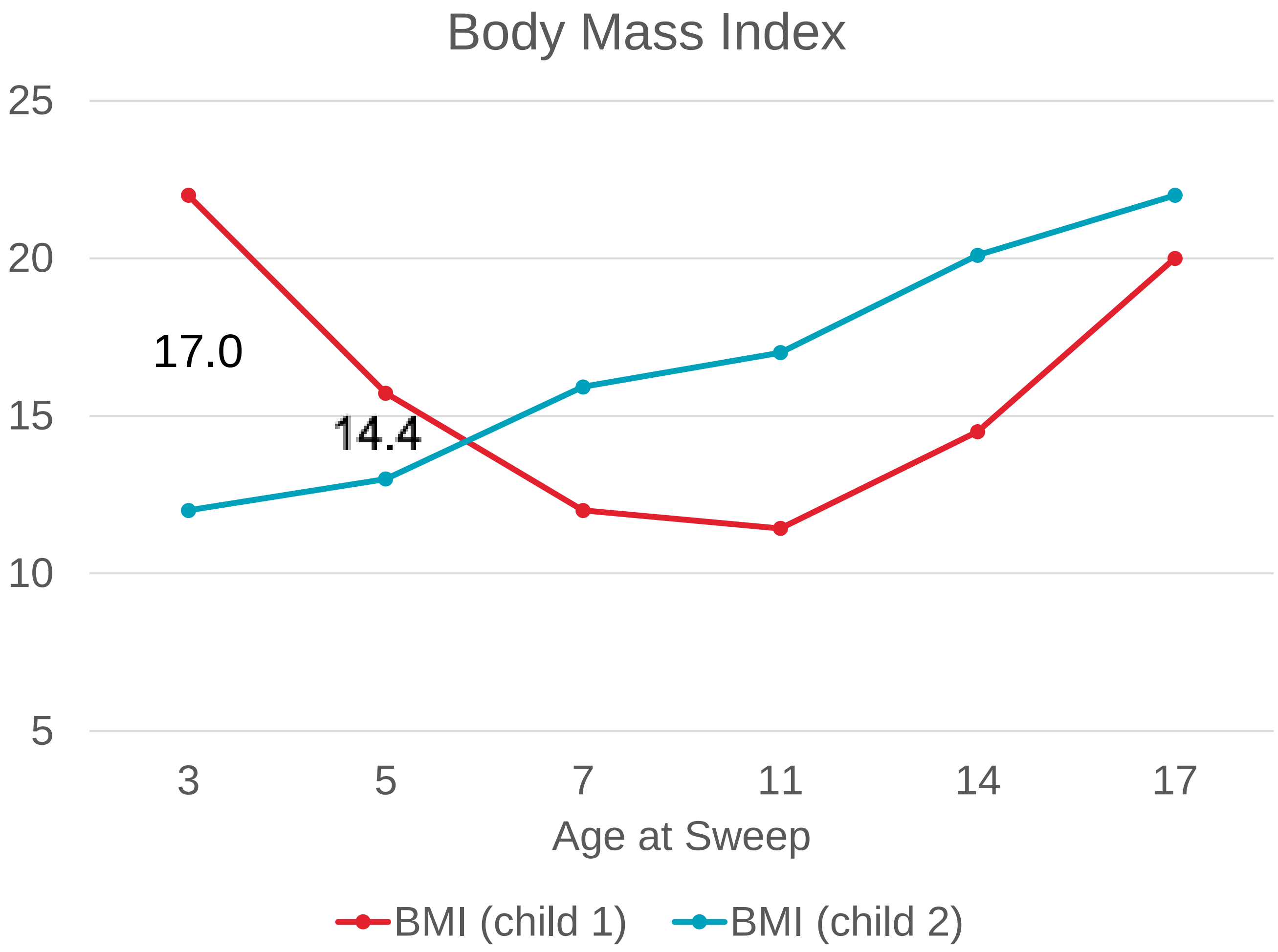


Descriptive statistics

CENTRE FOR
LONGITUDINAL
STUDIES

1) Cross-sectional Mean/Average

Age at Sweep	BMI (child 1)	BMI (child 2)	Mean
3	22	12	17.0
5	16	13	14.4
7	12	16	14.0
11	11	17	14.2
14	15	20	17.3
17	20	22	21.0



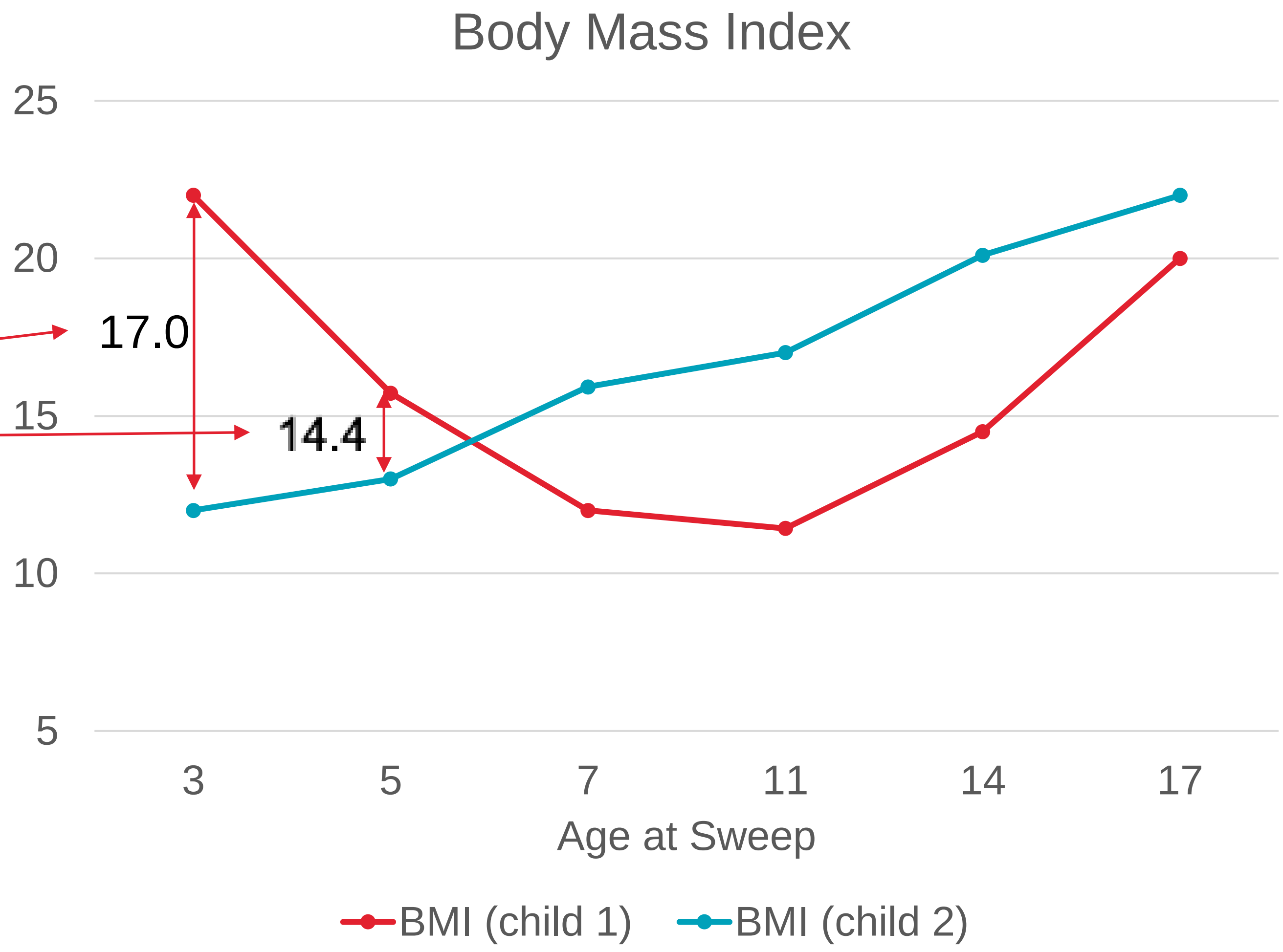
Dividing the sum of the values in each sweep by the total number of observations

Descriptive statistics

CENTRE FOR
LONGITUDINAL
STUDIES

2) Cross-sectional Standard Deviation

Age at Sweep	BMI (child 1)	BMI (child 2)	SD
3	22	12	7.1
5	16	13	1.9
7	12	16	2.8
11	11	17	3.9
14	15	20	4.0
17	20	22	1.4



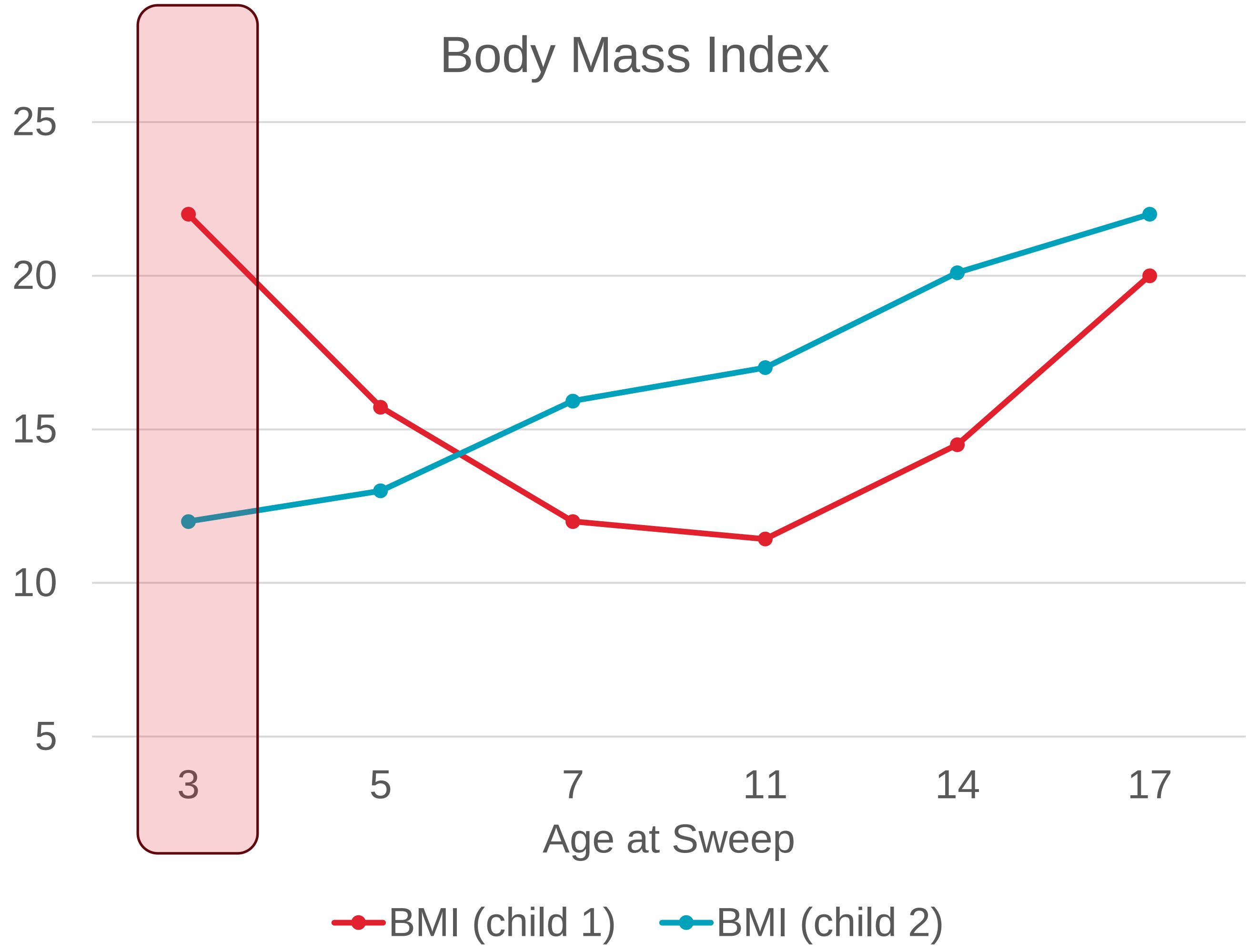
Dispersion of a dataset relative to its mean

Descriptive statistics

CENTRE FOR LONGITUDINAL STUDIES

3) Cross-sectional: Others

- Coefficient of Variation or (Normalized Root-Mean-Square Deviation)
- Median
- Min
- Max
- Quantiles Q1 Q2 Q3 Q4 Q5
- Kurtosis
- Skewness



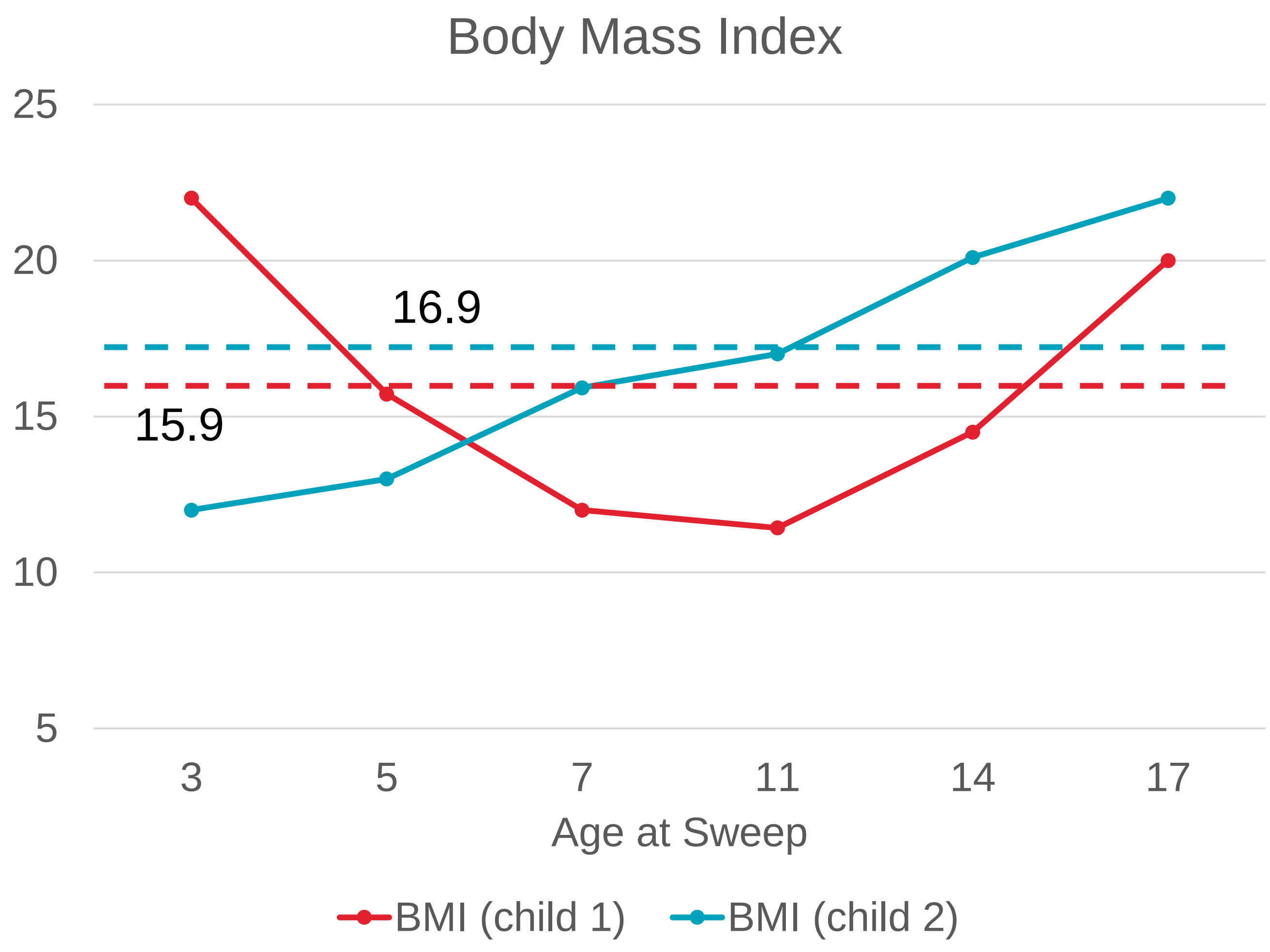
CENTRE FOR LONGITUDINAL STUDIES

Descriptive statistics

1) Longitudinal Mean/Average

Age at Sweep	BMI (child 1)	BMI (child 2)
3	22	12
5	16	13
7	12	16
11	11	17
14	15	20
17	20	22

	Child 1	Child 2
Mean	15.9	16.7
SD	4.3	3.9
Coefficient of Variation	0.27	0.23



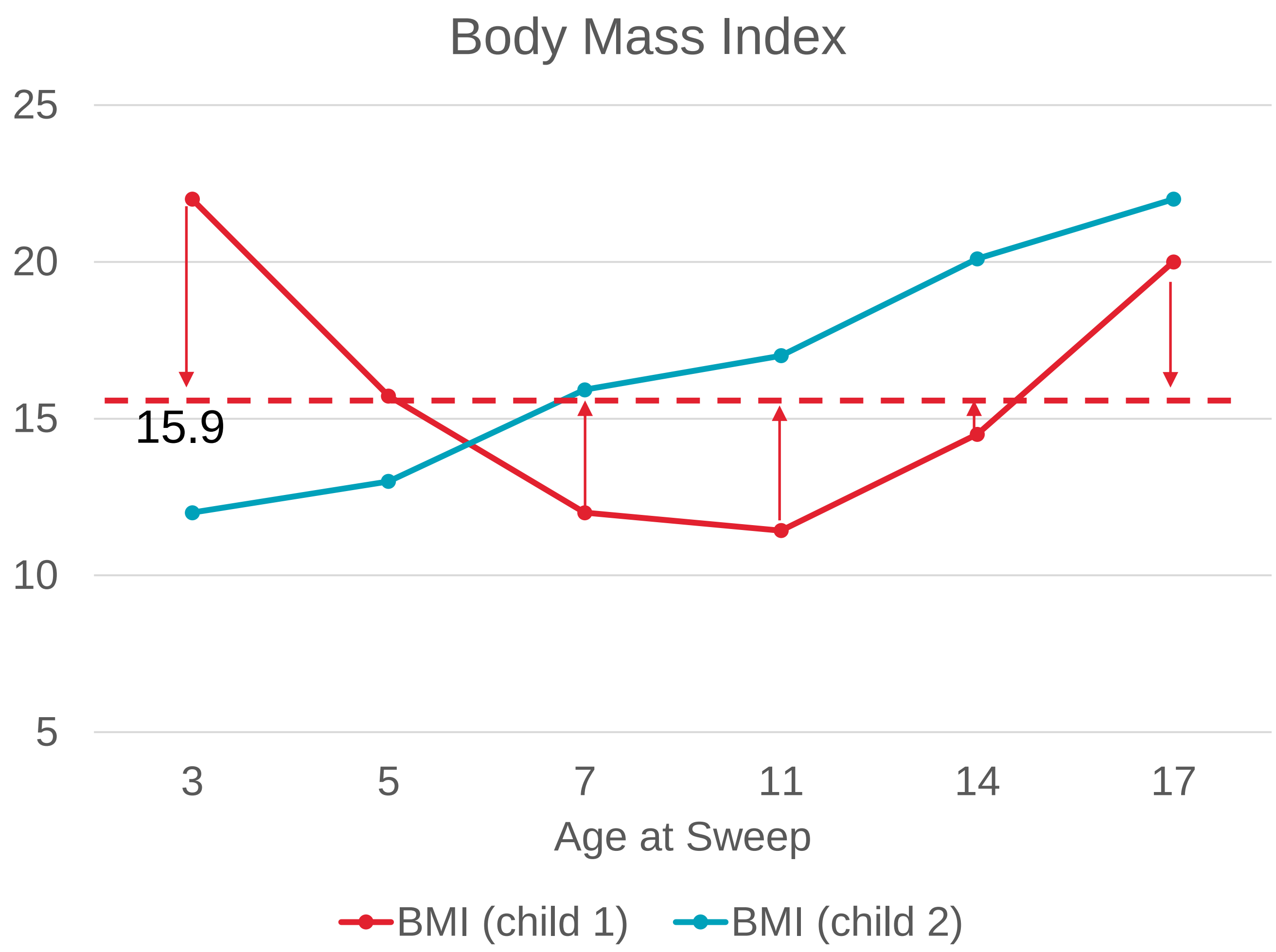
Descriptive statistics

CENTRE FOR LONGITUDINAL STUDIES

2) Longitudinal Standard Deviation

Age at Sweep	BMI (child 1)	BMI (child 2)
3	22	12
5	16	13
7	12	16
11	11	17
14	15	20
17	20	22

	Child 1	Child 2
Mean	15.9	16.7
SD	4.3	3.9
Coefficient of Variation	0.27	0.23



Descriptive statistics

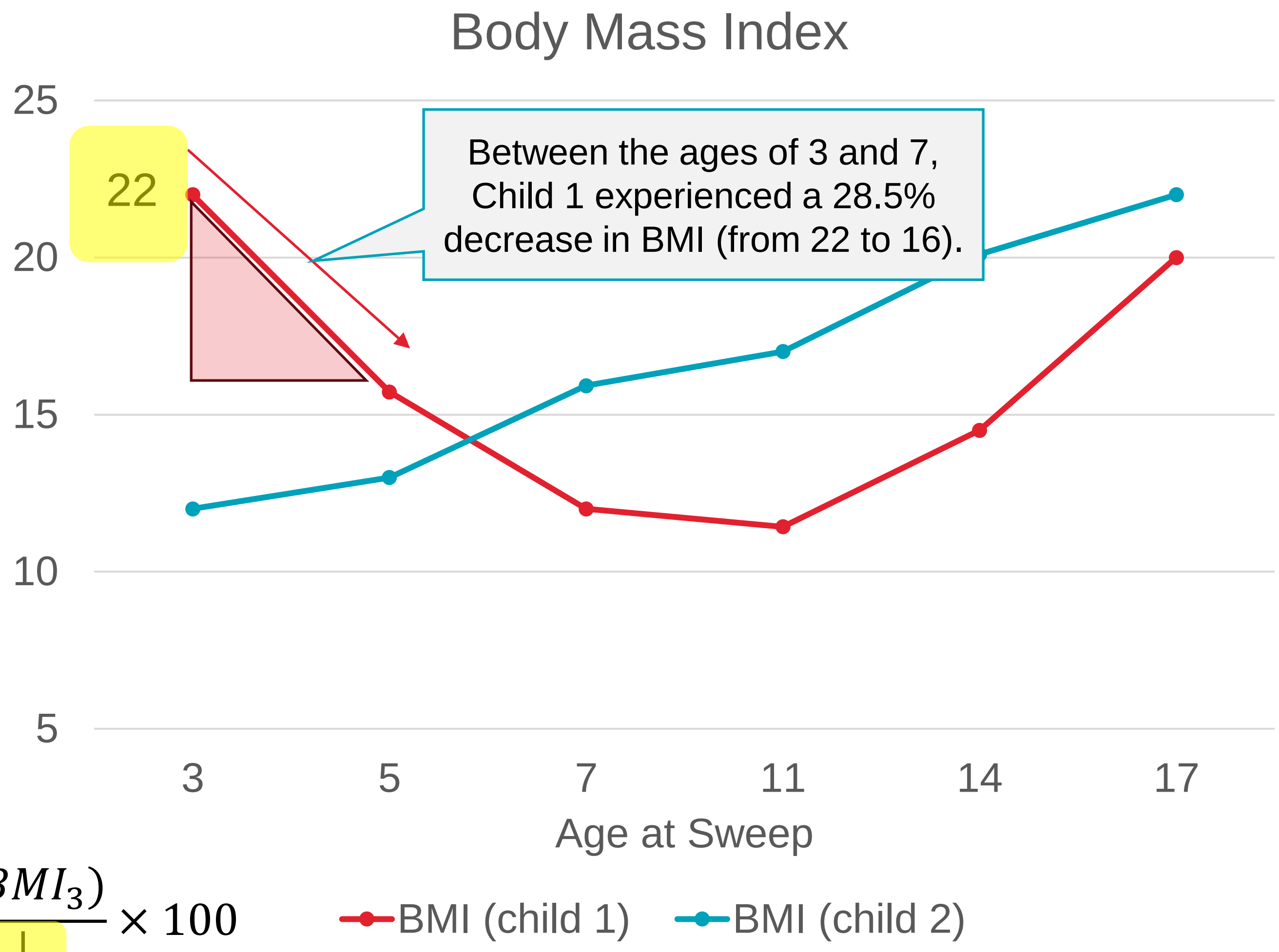
CENTRE FOR LONGITUDINAL STUDIES

3) Individual Percentage Change

Age	BMI (child 1)	BMI (child 2)	PCh (child 1)	PCh (child 2)
3	22	12		
5	16	13	-28.5	8.3
7	12	16	-23.7	22.5
11	11	17	-4.8	6.8
14	15	20	26.9	18.2
17	20	22	37.9	9.5

Change in BMI divided by the absolute value of the initial BMI, multiplied by 100.

$$\frac{(BMI_5 - BMI_3)}{|BMI_3|} \times 100$$



Descriptive statistics

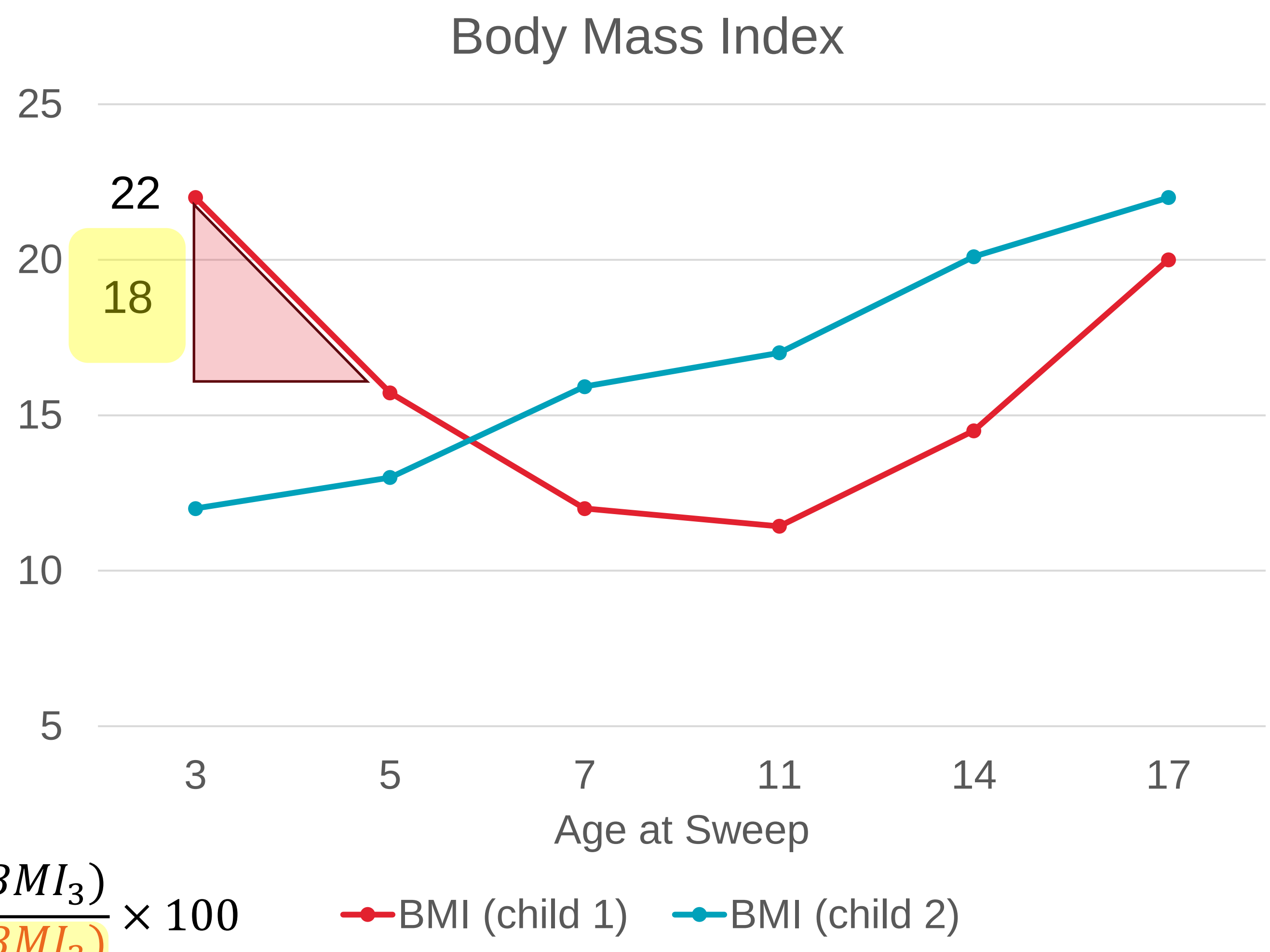
CENTRE FOR LONGITUDINAL STUDIES

4) Individual Arc Percentage Change

Age	BMI (child 1)	BMI (child 2)	APCh (child 1)	APCh (child 2)
3	22	12		
5	16	13	-33	8
7	12	16	-27	20
11	11	17	-5	7
14	15	20	24	17
17	20	22	32	9

Change in BMI divided by the midpoint, multiplied by 100.

$$\frac{(BMI_5 - BMI_3)}{\frac{(BMI_5 + BMI_3)}{2}} \times 100$$

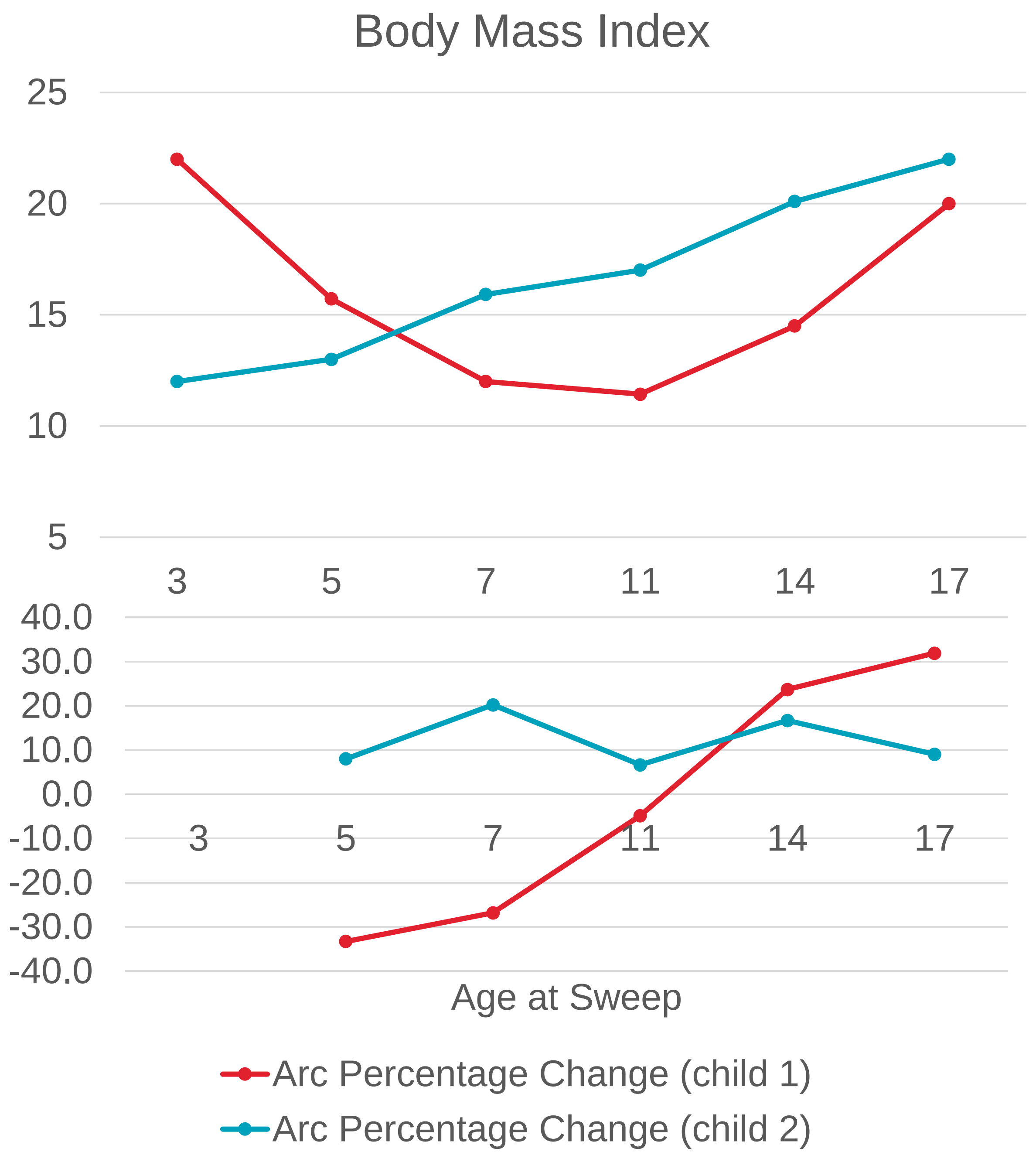


Descriptive statistics

5) Individual SD of Arc Percentage Change

Age	BMI (child 1)	BMI (child 2)	APCh (child 1)	APCh (child 2)
3	22	12		
5	16	13	-33	8
7	12	16	-27	20
11	11	17	-5	7
14	15	20	24	17
17	20	22	32	9

	Child 1	Child 2
Mean	15.9	16.7
SD	4.3	3.9
CV	0.27	0.23
SD Arc PCh	29.2	6.0



Descriptive statistics - Stata

CENTRE FOR
LONGITUDINAL
STUDIES

```
xtset mcsid t
* Lag of BMI
gen bmi_l=l1.bmi
* Change in BMI
gen ch_bmi=(bmi-bmi_l)
* Percentage Change
gen pch_bmi=((bmi-bmi_l)/bmi_l)*100
* Arc Percentage Change in BMI
gen apch_bmi=((bmi-bmi_l)/((bmi+bmi_l)/2))*100
* SD of Arc Percentage Change in BMI
bys mcsid: egen sd_apch_bmi=sd(apch_bmi)
```

mcs_bmi_long_clswebinar.dta

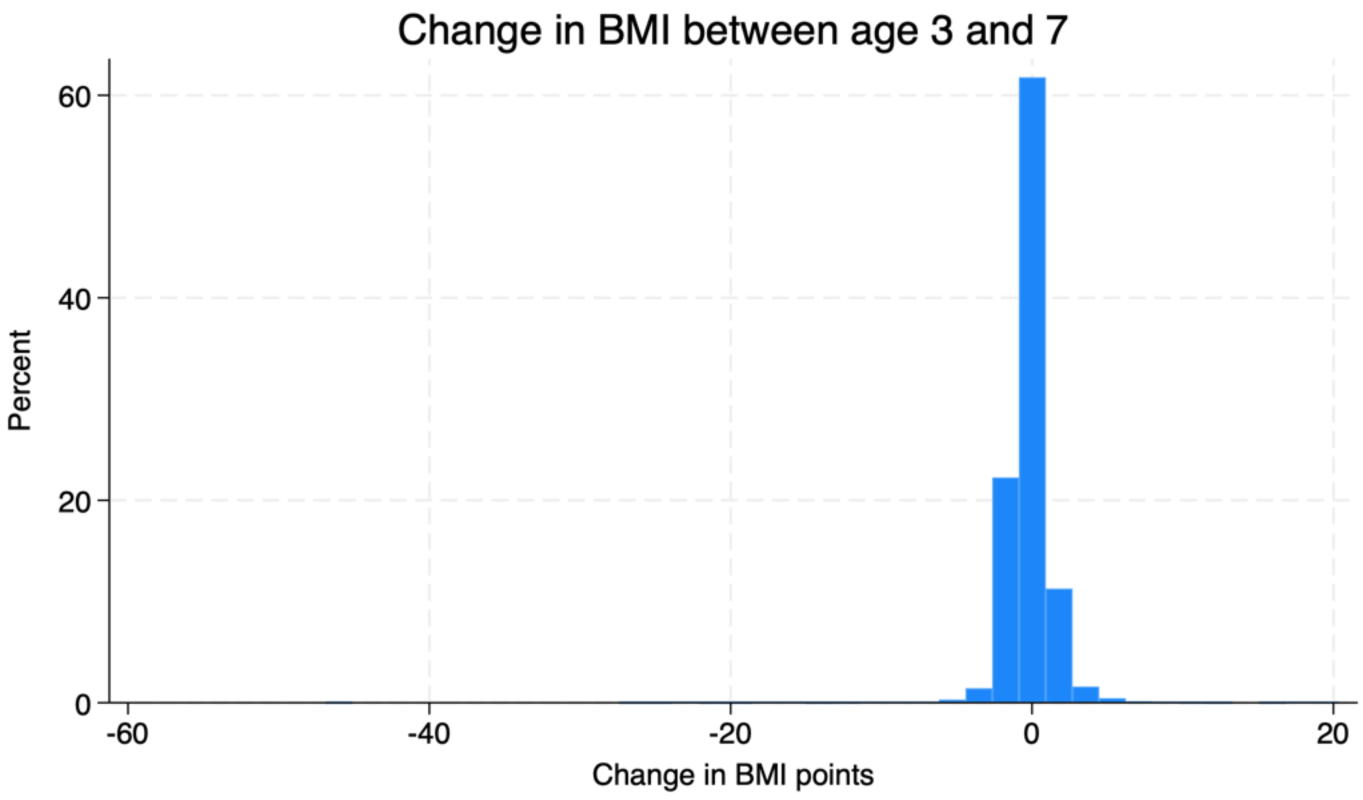
mcsid	bmi	t	bmi_l	ch_bmi	pch_bmi	apch_bmi	sd_apch_bmi
1	17.36	1	.	.	.	.	10.29573
1	15.72	2	17.36111	-1.636619	-9.426923	-9.893236	10.29573
1	17.32	3	15.72449	1.595506	10.14663	9.656713	10.29573
1	20.3	4	17.32	2.98	17.20555	15.84264	10.29573
1	19.64	5	20.3	-.66362	-3.269064	-3.323386	10.29573
1	19.6	6	19.63638	-.0394115	-.2007068	-.2009084	10.29573
2	18.26	1	.	.	.	.	10.91704
2	18.17	2	18.26221	-.0921402	-.5045403	-.5058163	10.91704
2	15.92	3	18.17007	-2.250067	-12.38337	-13.20072	10.91704
2	17.01	4	15.92	1.09	6.846733	6.620103	10.91704
2	18.95	5	17.01	1.936836	11.38646	10.77312	10.91704
2	21.89	6	18.94684	2.944019	15.53832	14.41815	10.91704

Descriptive statistics - Stata

CENTRE FOR LONGITUDINAL STUDIES

3) Individual Change in BMI

```
hist ch_bmi if age==5 , ///
title(Change in BMI between age 3 and 7) ///
plotregion(fcolor(white)) graphregion(fcolor(white)) ///
percent xtitle(Change in BMI points)
```

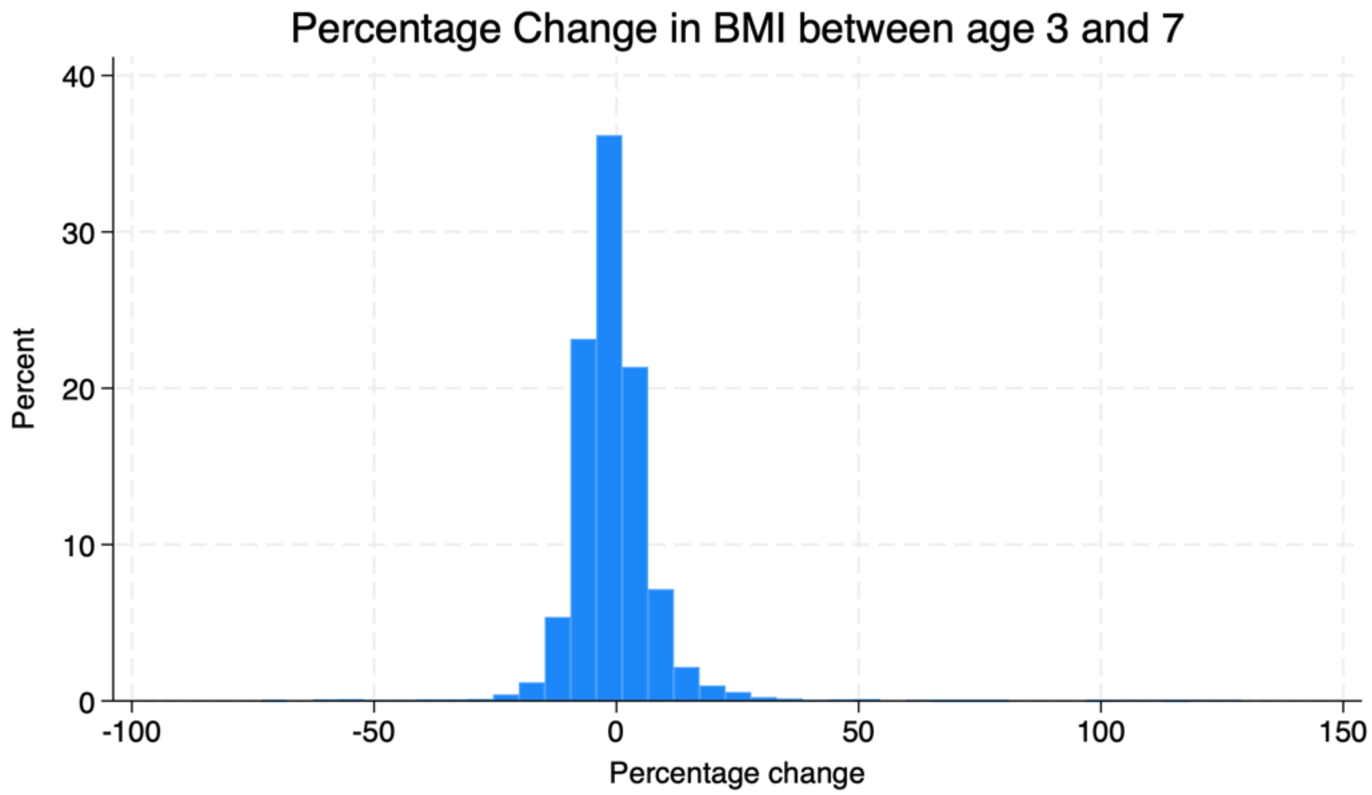


Descriptive statistics - Stata

CENTRE FOR LONGITUDINAL STUDIES

3) Individual Percentage Change

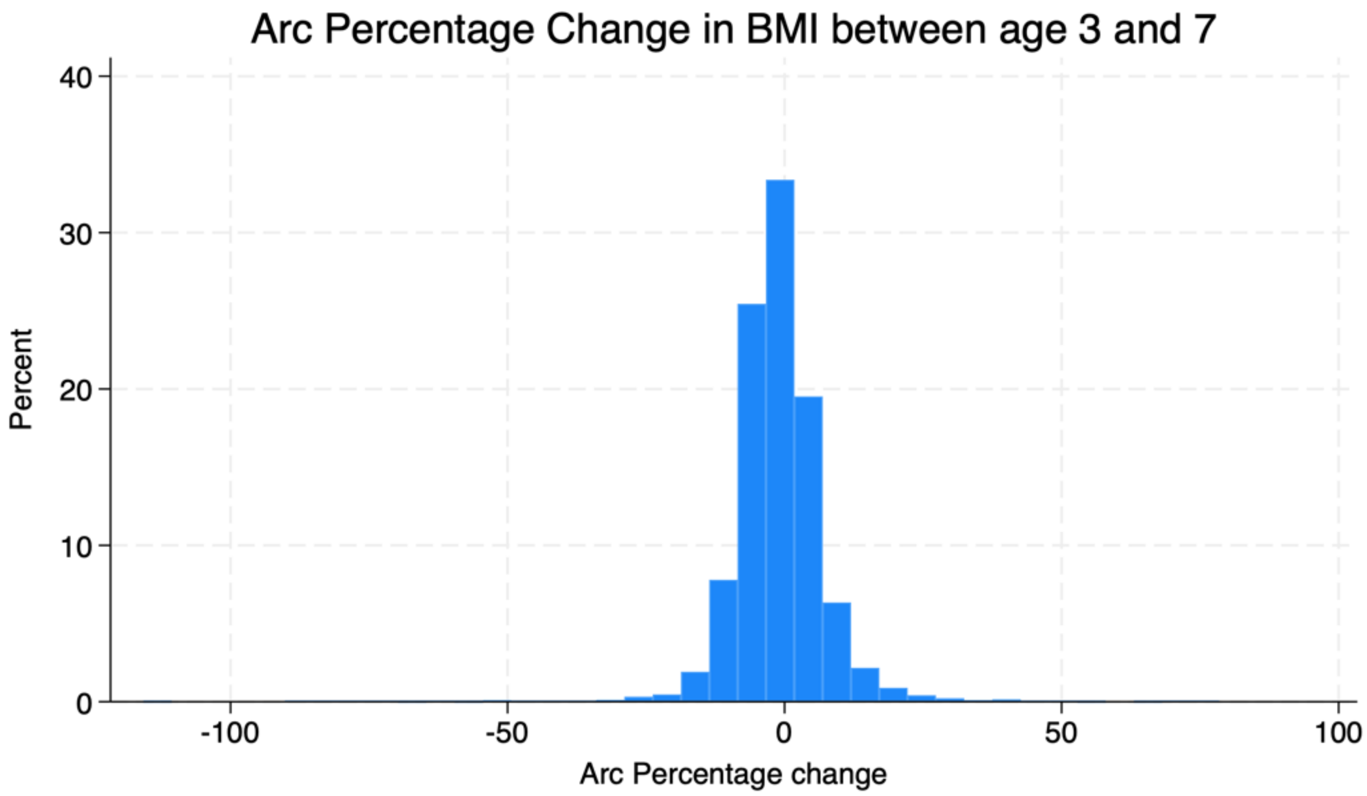
```
hist pch_bmi if age==5 , ///
title(Percentage Change in BMI between age 3 and 7) ///
plotregion(fcolor(white)) graphregion(fcolor(white)) ///
percent xtitle(Percentage change)
```



Descriptive statistics - Stata

4) Individual Arc Percentage Change

```
hist apch_bmi if age==5 , ///
title(Arc Percentage Change in BMI between age 3 and 7) ///
plotregion(fcolor(white)) graphregion(fcolor(white)) ///
percent xtitle(Arc Percentage change)
```

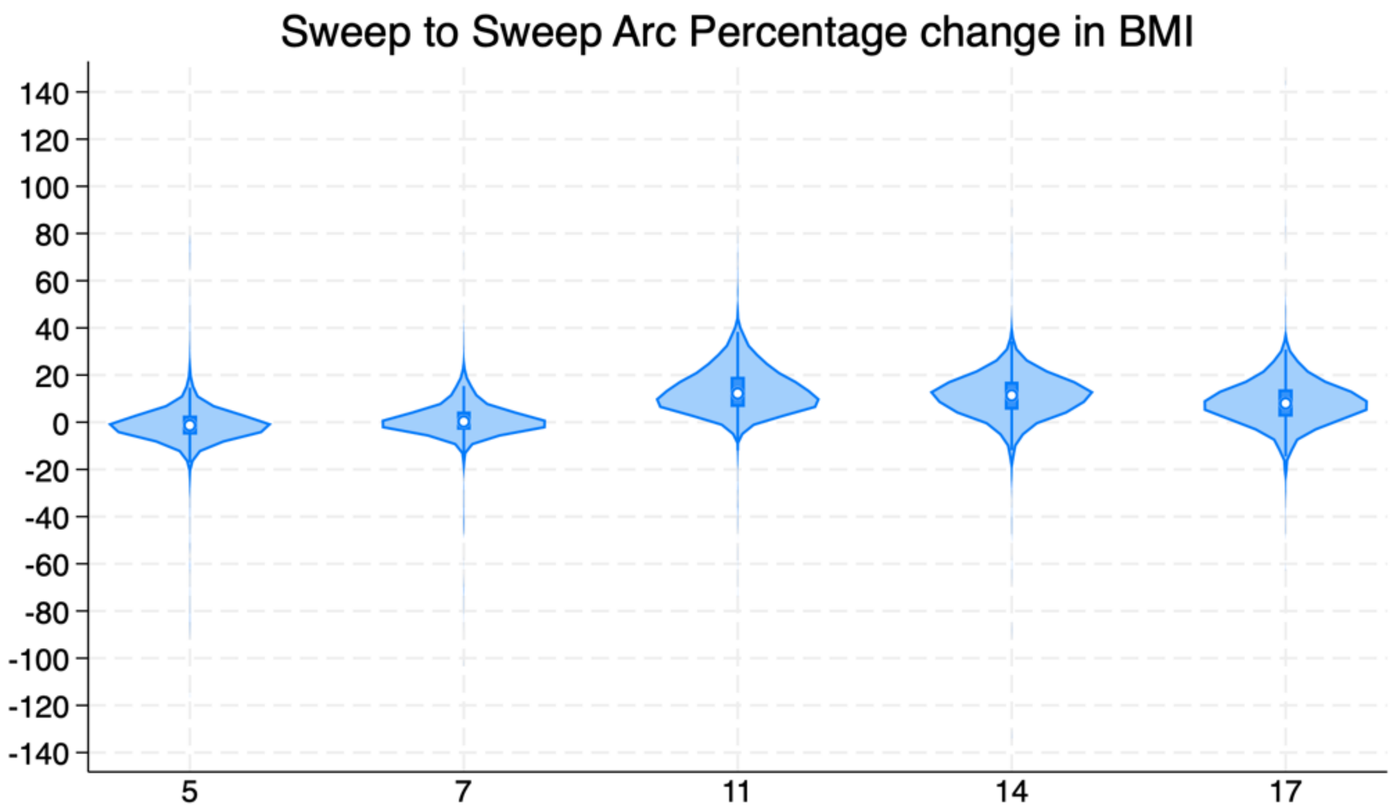


Descriptive statistics - Stata

CENTRE FOR
LONGITUDINAL
STUDIES

4) Individual Arc Percentage Change

```
vioplot apch_bmi if age>3 , over(age) ///
title(Sweep to Sweep Arc Percentage change in BMI) ///
 ylabel(-140(20)140, angle(horizontal ) )  ///
 yscale(range(-140 140))
```

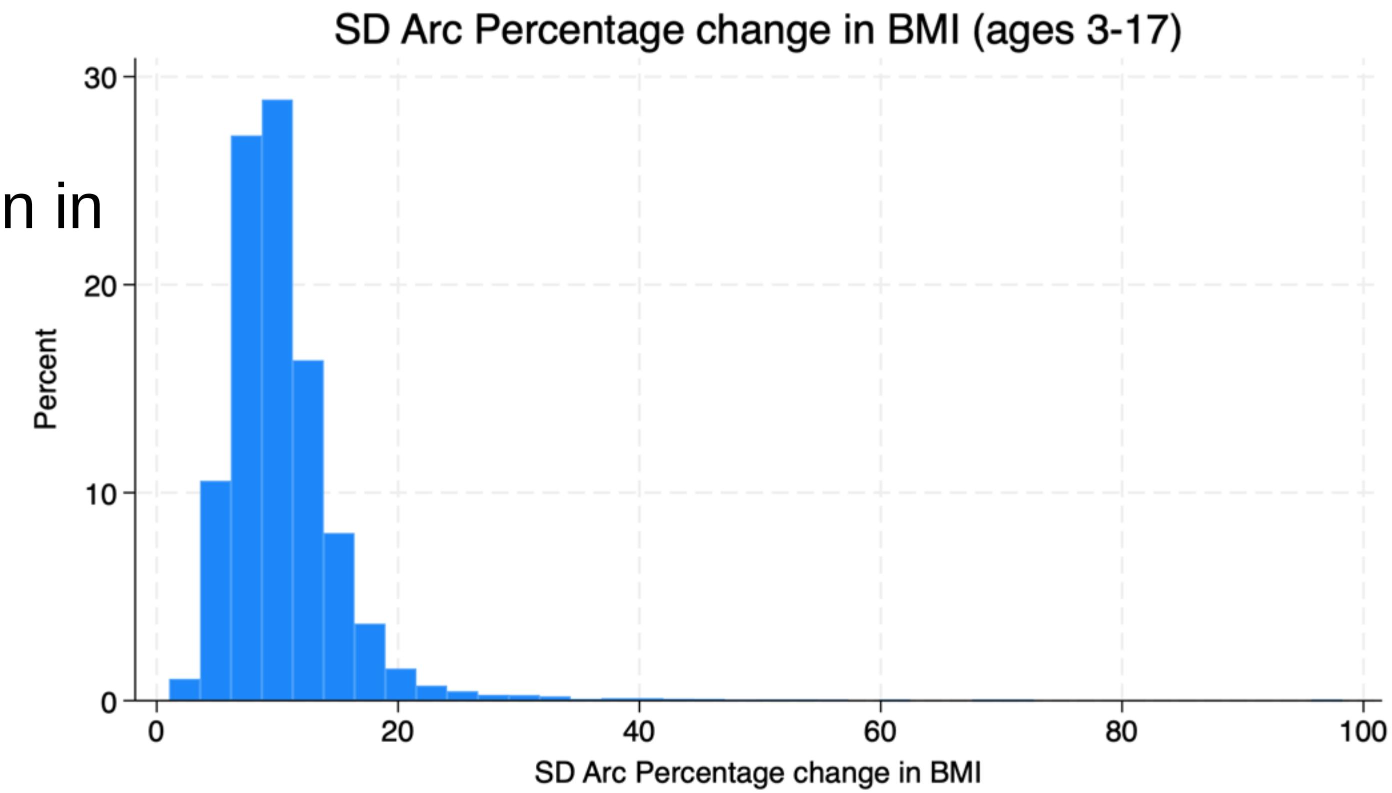


Descriptive statistics - Stata

5) Individual SD Arc Percentage Change

It characterises BMI trajectories based on within individual variation in BMI over the lifecycle (ages 3-7).

```
hist sd_apch_bmi if age==5 , ///
title(SD Arc Percentage change in BMI (ages 3-17)) ///
plotregion(fcolor(white)) graphregion(fcolor(white)) ///
percent xtitle(SD Arc Percentage change in BMI)
```

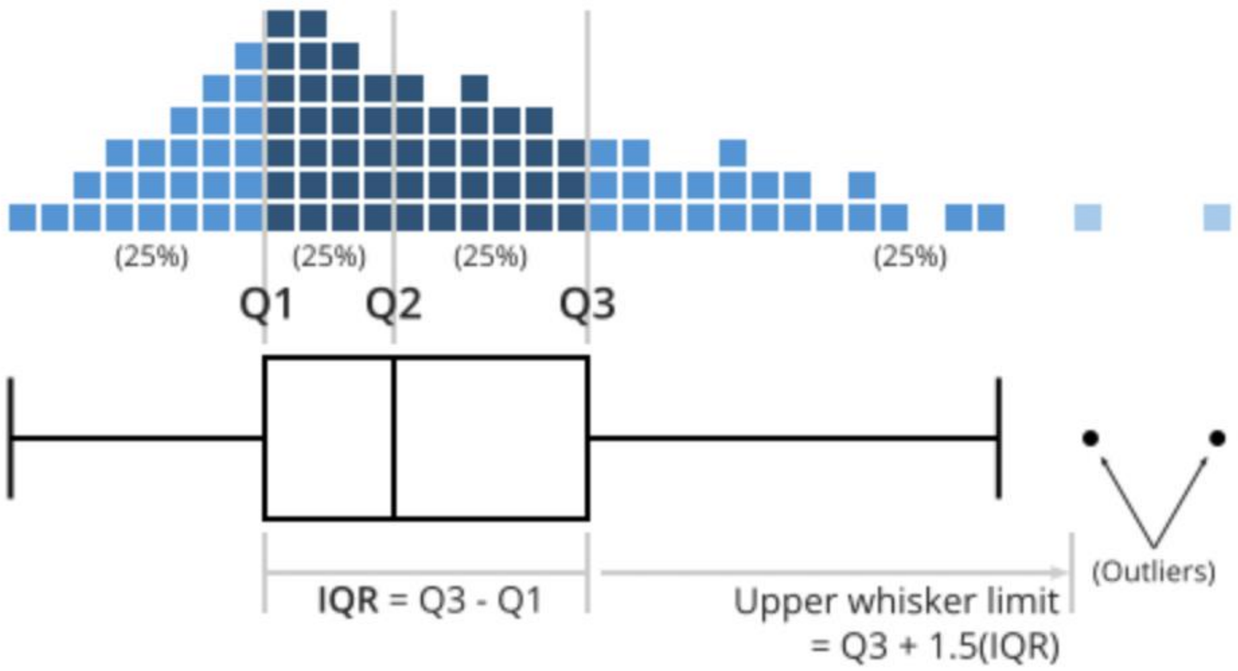


Graphical tools: Box and Violin plot

CENTRE FOR
LONGITUDINAL
STUDIES

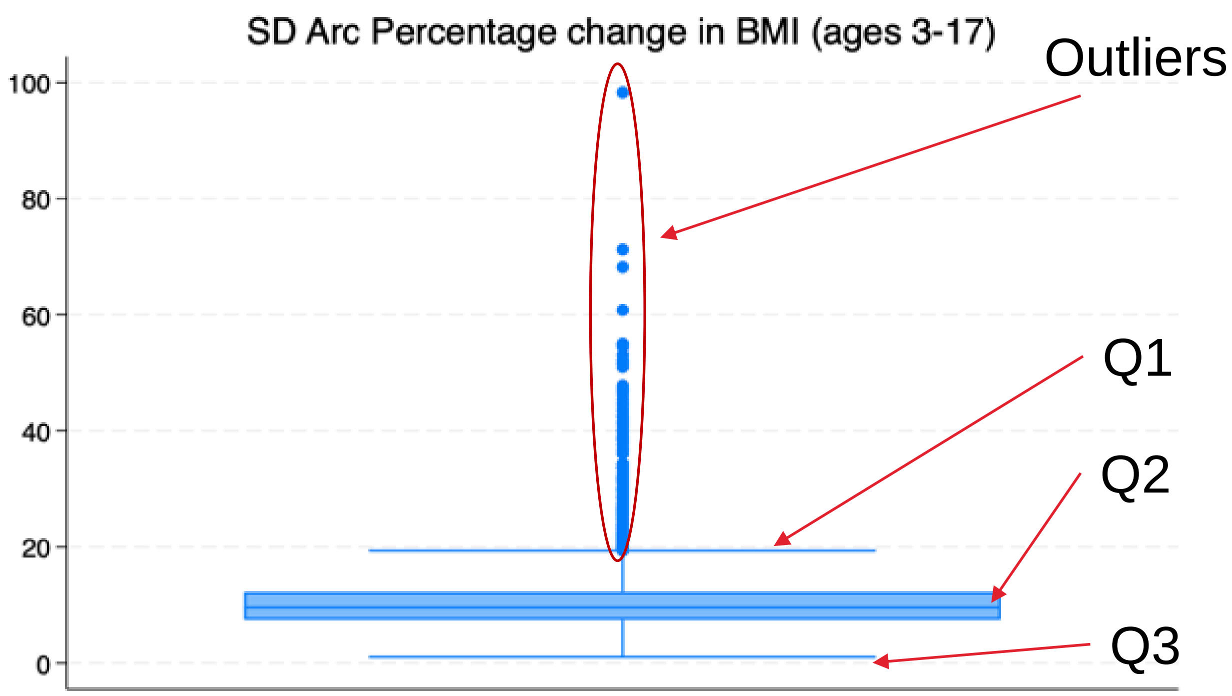
It shows:

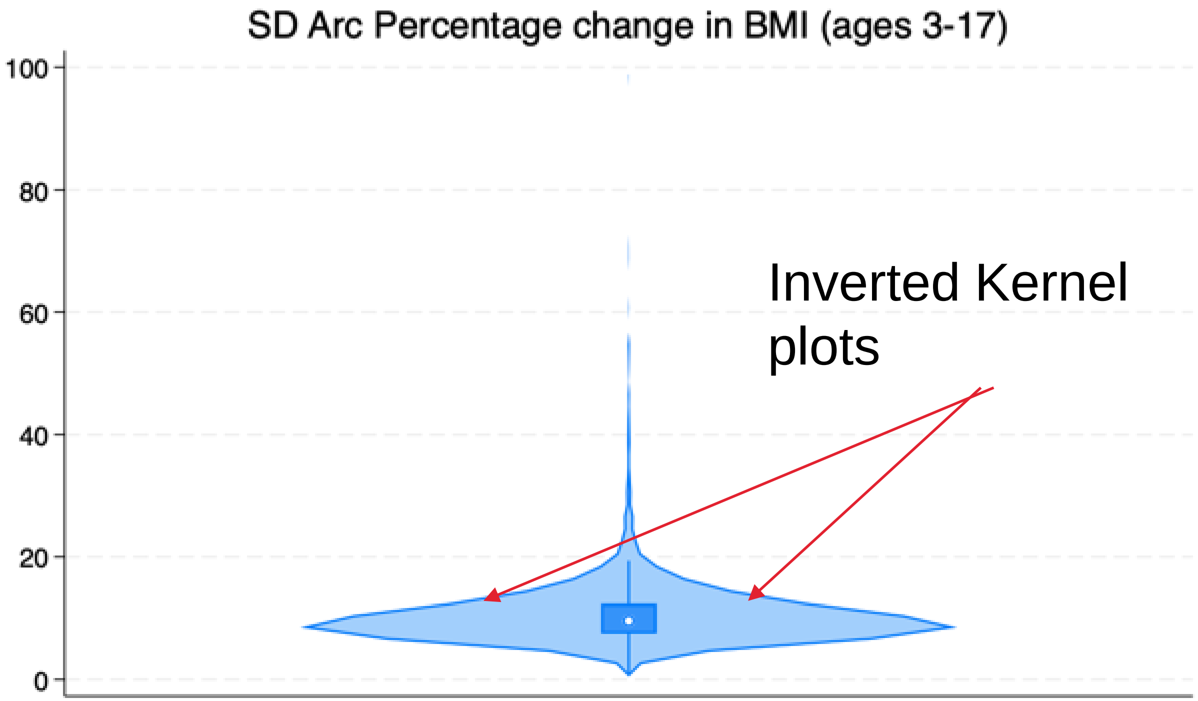
- Median
- Quartiles divide a sample on 4 intervals, with cut points dividing the range of sample in intervals with equal probabilities.
- Interquartile range (IQR): the distance between the upper and lower quartiles
- The bottom/top wiskers show the Q1-1.5*IQR and Q3+1.5*IQR
- Outliers



Graphical tools: Box and Violin plot – Stata

CENTRE FOR LONGITUDINAL STUDIES





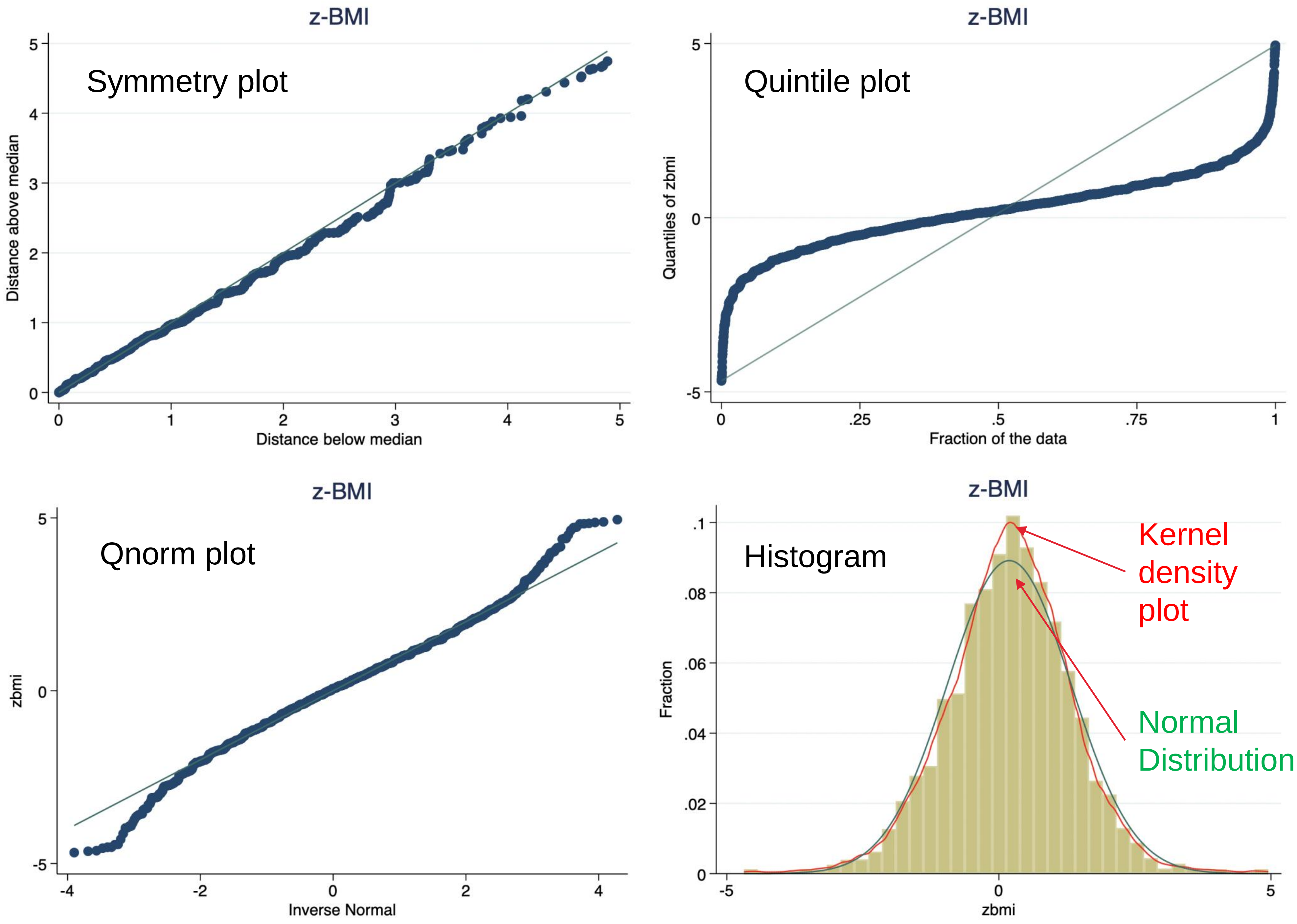

```
graph  box sd_apch_bmi if age==5 , ///
title(SD Arc Percentage change in BMI (ages 3-17)) ///
plotregion(fcolor(white)) graphregion(fcolor(white)) ///
 ytitle("")
```

```
vioplot sd_apch_bmi if age==5 , ///
title(SD Arc Percentage change in BMI (ages 3-17)) ///
plotregion(fcolor(white)) graphregion(fcolor(white)) ///
 xlabel("")
```

Graphical tools: Distributional diagnostic plots

- Compare the distribution of variables against a known distribution (normal, uniform, etc.)
    - Symmetry plot – check if the data is symmetric around the mean
    - Quintile plot - compares with a uniform distribution
    - Qnorm plot - compares normal distribution
    - Histogram – Kernel density plots

- Symmetric
- Not uniform
- Similar to a normal distribution, but right tail is heavier



Graphical tools: Spaghetti and Lasagne plot

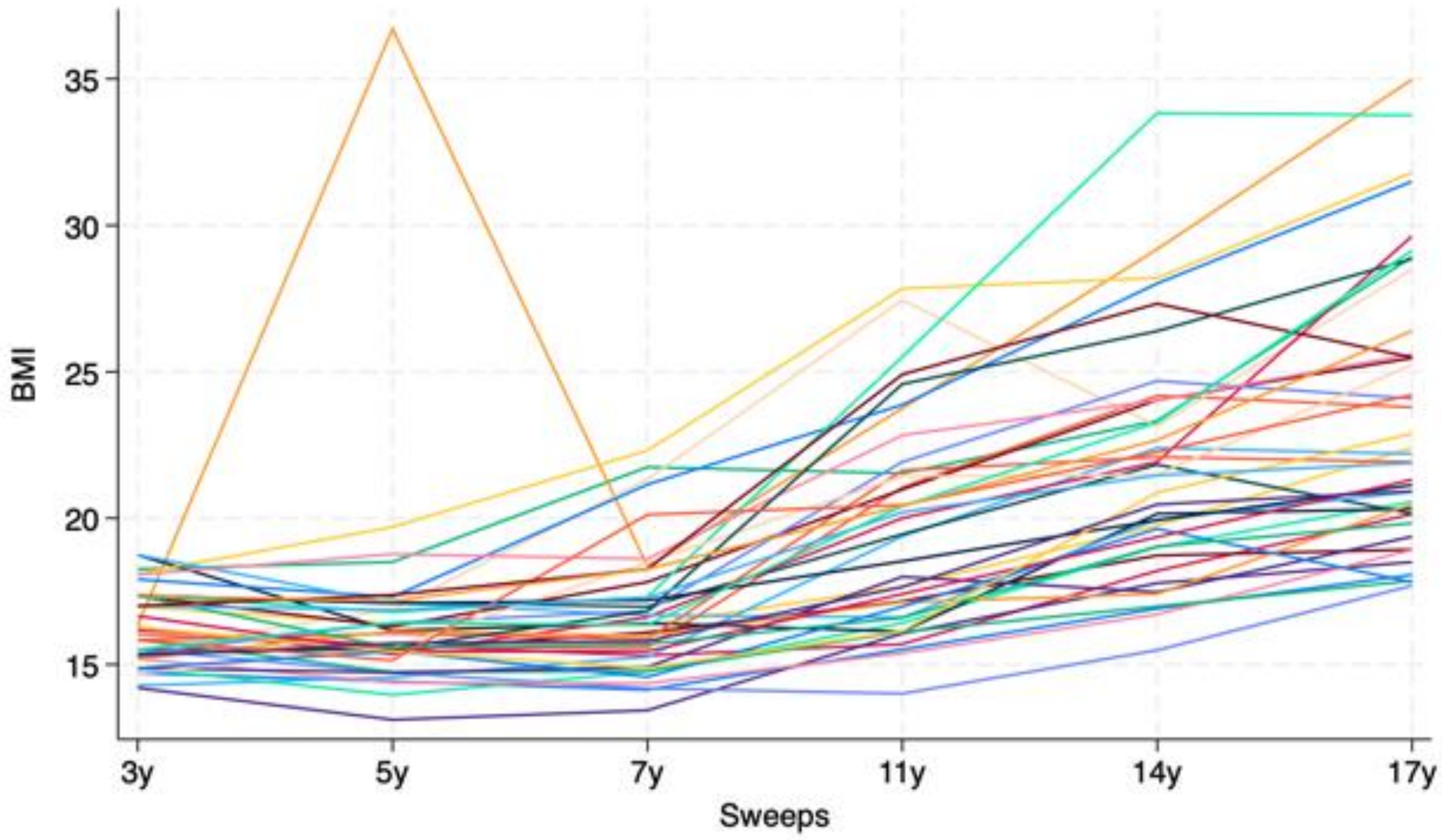
CENTRE FOR
LONGITUDINAL
STUDIES

- Used to represent the observed heterogeneity on unit trends over time.
- Trends are represented using individual lines (spaghettis) that connect observed values over time.
- Lasagne plot are heat plots that use horizontal layers instead of lines to represent individual changes over time.

Graphical tools: Spaghetti plot

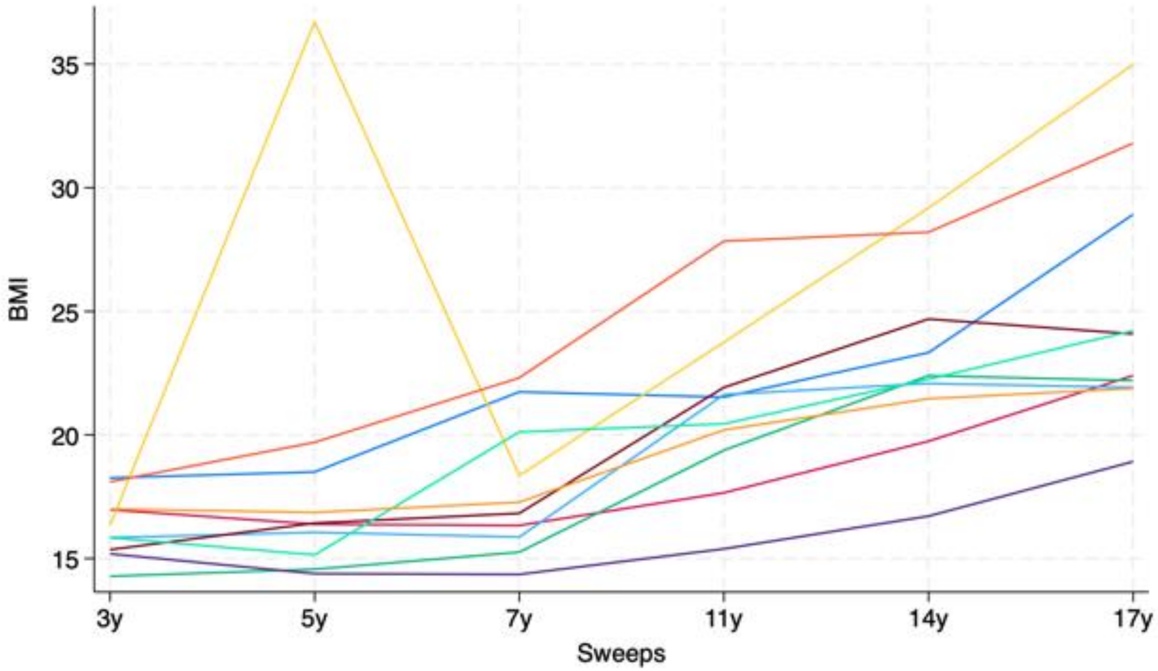
CENTRE FOR LONGITUDINAL STUDIES

- Sort data in wide format by variable of interest (BMI)
- Plot a random sample of units
- Useful to visualise different likely trajectories



Graphical tools: Spaghetti plot - Stata

CENTRE FOR LONGITUDINAL STUDIES




```
set seed 321
generate u1 = runiform()
sort u1
gen sample10=_n<=10
gen sample20=_n<=20
gen sample30=_n<=30
gen sample40=_n<=40
```

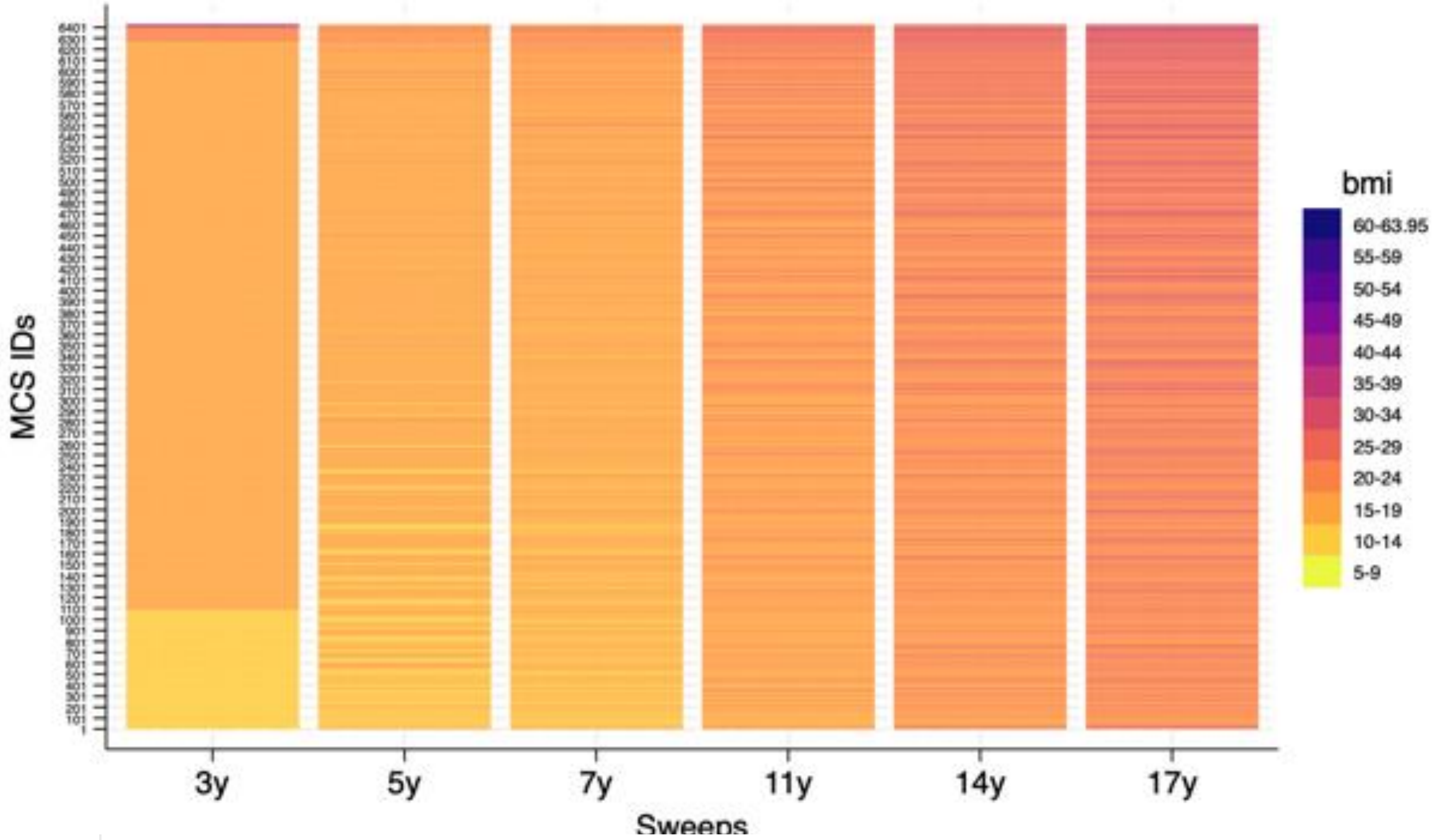
```
xtline bmi if sample40==1, overlay legend(off) ///
xlabel( 1 "3y" 2 "5y" 3 "7y" 4 "11y" 5 "14y" 6 "17y") ///
  xtitle(Sweeps) ytitle(BMI) ylabel(,angle(horizontal ) ) ///
  plotregion(fcolor(white)) graphregion(fcolor(white))
```

Graphical tools: Lasagne plot - Stata

CENTRE FOR LONGITUDINAL STUDIES

- Sort data in wide format by variable of interest (BMI)
  - id_sbmi

- Useful to visualise all possible trajectories

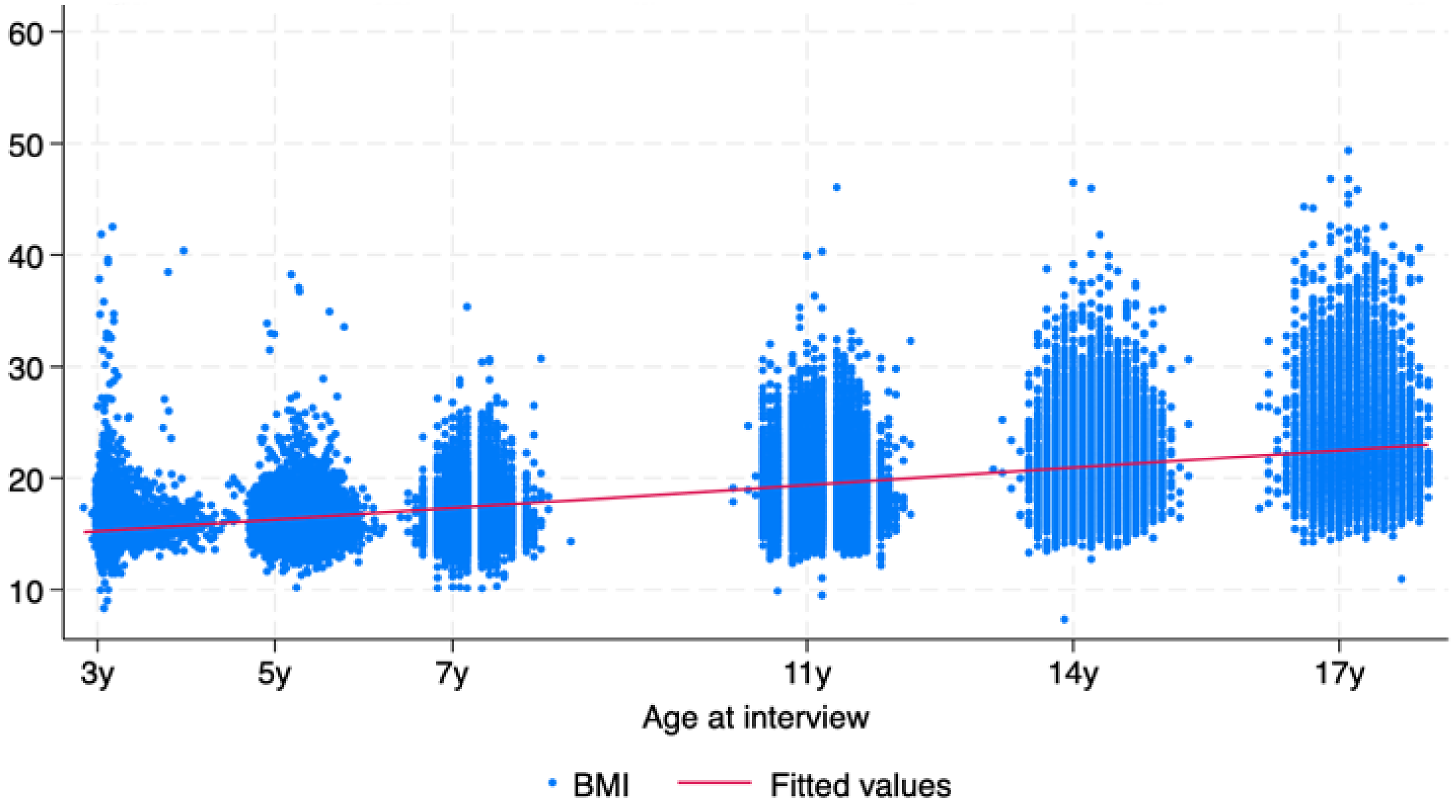



```
sort bmi3 bmi5 bmi7 bmi11 bmi14 bmi17
gen id_sbmi=_n
```

```
heatplot bmi id_sbmi t, statistic(asis) ///
  discrete(0.9) ylabel(1(100)6400,labsize(tiny))  ///
  xlabel(1 2 3 4 5 6)  cut(5(5)@max) keylabels(, range(1))  ///
  xlabel( 1 "3y" 2 "5y" 3 "7y" 4 "11y" 5 "14y" 6 "17y") ///
  xtitle(Sweeps) ytitle(MCS IDs) ylabel(,angle(horizontal ) ) ///
  plotregion(fcolor(white)) graphregion(fcolor(white))  color(plasma, reverse )
```

Descriptive statistics
Correlation and Linear regression

- Is there a linear association between BMI at age 5 and BMI at age 7?
- What is the best fitting line to summarise the trend between age and BMI?



$$BMI_{it} = constant + slope \times Age_{it}$$

Descriptive statistics - Correlation

CENTRE FOR LONGITUDINAL STUDIES

Indicate the degree of linear association between two variables without implying causation.

Wide Structure: Used to analyse association between variables over time

```
corr bmi5 bmi7 bmi11 bmi14 bmi17
```

	BMI 5	BMI 7	BMI 11	BMI 14	BMI 17
BMI 5	1.00				
BMI 7	0.72	1.00			
BMI 11	0.58	0.78	1.00		
BMI 14	0.52	0.70	0.83	1.00	
BMI 17	0.48	0.64	0.75	0.85	1.00

Long Structure: Used to analyse association between outcome variable with time/period

```
corr bmi age
```

	Age
BMI	0.62

Descriptive statistics – Linear regression

mcs_bmi_long_clswebinar.dta

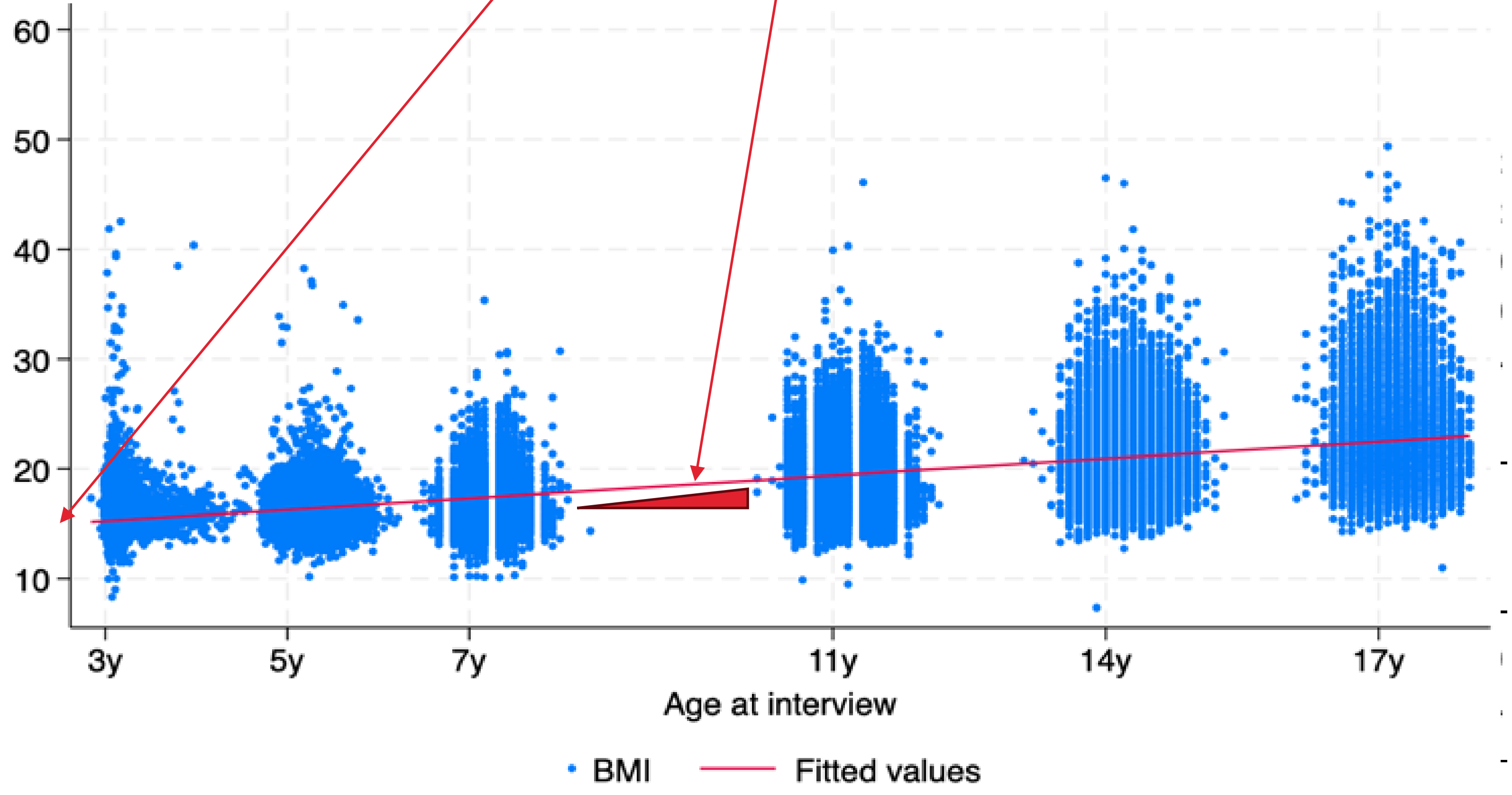
$$BMI_{it} = constant + slope \times Age_{it}$$

$$BMI_{it} = 13.7 + 0.51 \times Age_{it}$$

(0.004) (0.04)

Our best linear prediction summarise the relationship between age and BMI.

- BMI increases with age



Longitudinal Data Visualisation: summary

CENTRE FOR LONGITUDINAL STUDIES

- Introduce simple and useful tools to descriptive and visualise longitudinal data (Categorical/Continuous variables; Box/Violin/Spaghetti/Lasagne)
- Focus: understanding change and trajectories of key variable
    - What about multivariable?

- Descriptive and graphical analyses of longitudinal data help us to better understand our key variables. It’s a useful first step!

CENTRE FOR
LONGITUDINAL
STUDIES

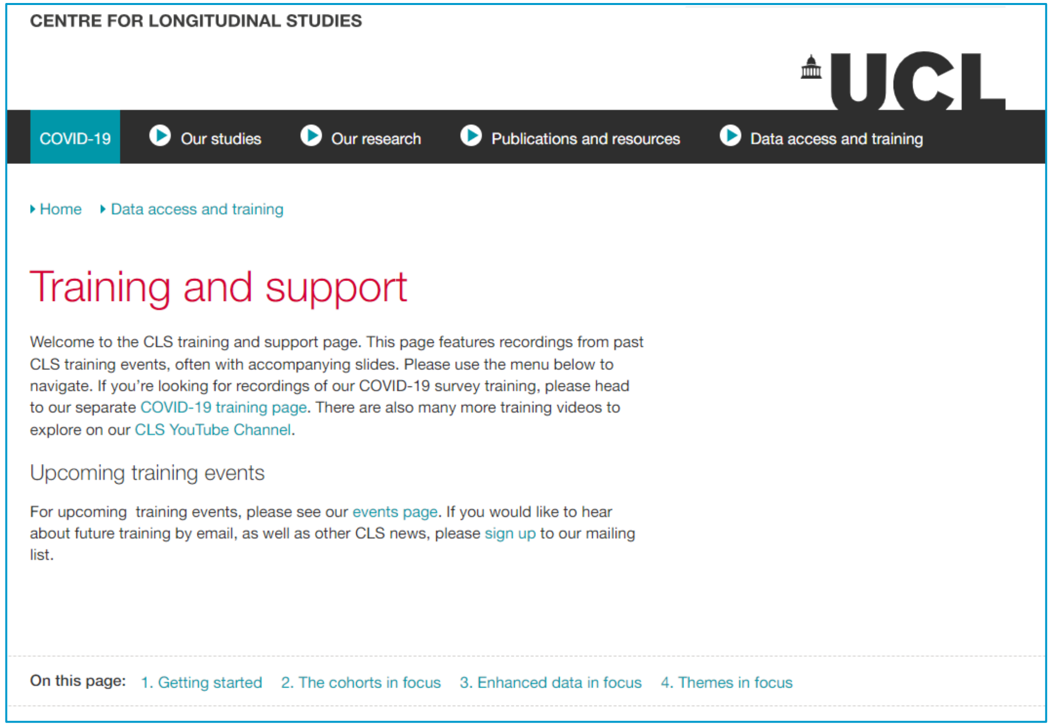
Questions?

Upcoming Training Events in early 2024

CENTRE FOR LONGITUDINAL STUDIES

- Ageing in the British cohort studies: measurement, research and access
- Genetic data: An overview of genetic data in the British cohort studies
- Methods: Cross-cohort analyses

https://cls.ucl.ac.uk/events/



CENTRE FOR LONGITUDINAL STUDIES

UCL

COVID-19 | Our studies | Our research | Publications and resources | Data access and training

Home › Data access and training

Training and support

Welcome to the CLS training and support page. This page features recordings from past CLS training events, often with accompanying slides. Please use the menu below to navigate. If you're looking for recordings of our COVID-19 survey training, please head to our separate COVID-19 training page. There are also many more training videos to explore on our CLS YouTube Channel.

Upcoming training events

For upcoming training events, please see our events page. If you would like to hear about future training by email, as well as other CLS news, please sign up to our mailing list.

On this page: 1. Getting started 2. The cohorts in focus 3. Enhanced data in focus 4. Themes in focus

https://cls.ucl.ac.uk/data-access-training/training-and-support-2/

Introduction to Longitudinal Data Structure and Visualisation

Nicolás Libuy, nicolas.libuy@ucl.ac.uk

Darío Moreno-Agostino, d.moreno@ucl.ac.uk

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council