

Section 2 - Getting started and exploring the data

Longitudinal data analysis can be used to explore how characteristics and experiences from early life can influence later outcomes, while taking account of other childhood factors. In this module, we will use an extract of data from the NCDS CLOSER Training Dataset (CTD) to examine the relationship between intelligence test scores at the age of 11 years and BMI at age 42 years. This section will provide you with guidance on accessing relevant data, undertaking exploratory data analysis and preparing the data for the more advanced statistical modelling covered in subsequent sections.

Background

Individuals who gain lower scores on tests of intelligence in childhood or adolescence are more likely to report poorer health outcomes in middle to later life. Studies have shown, for example, that lower intelligence is related to obesity, high blood pressure, coronary heart disease, symptoms of psychological distress, and diagnosis of depression. Hypotheses put forward to explain these associations include the possibility that childhood measures of intelligence are (i) predictive of advantageous social circumstances in later life, (ii) associated with general bodily 'system integrity' (i.e. scoring well on cognitive ability tests might be a marker for a more general tendency for complex systems in the body to be efficient), or (iii) a proxy for stress management skills and the acquisition of behaviours conducive to health (i.e. not smoking, physical activity and prudent diet). The latter has been suggested as an explanation of the association between BMI and intelligence, where higher IQ scoring individuals interpret and respond to health advice in more positive ways. Using data from the CTD and by applying general linear, logistic and multinomial regression, we will test the relationship between childhood intelligence and adult body-mass index (BMI).

Main variables of interest

Our outcome variable is body mass index (BMI) in middle-age. The CTD includes a BMI variable based on self-reported measures of height and weight at age 42. BMI is calculated in metric units, and is based on weight (kg) divided by the square of height (m^2).

Our childhood explanatory variable is 'general ability'. At age 11, the cohort were given a general ability test, which required the child to recognise patterns in either words or pictures, and correctly identify the next word/picture in a sequence. For each child, the general ability test gave a total score that ranged between 0 and 80.

Accessing and preparing the dataset

To access the CTD, we must download it from the UK Data Service (UKDS; <https://discover.ukdataservice.ac.uk/catalogue/?sn=8205&type=Data%20catalogue>). We will need to register/login to access the data and then choose the Stata format download.

Screenshot of download options for the CTD

File Format	File Size (mb)	Download	
Dataset: National Child Development Study: CLOSER Training Dataset, 1958-2013			
SPSS	0.66	Download	<input type="checkbox"/>
STATA	0.64	Download	<input type="checkbox"/>
TAB	0.79	Download	<input type="checkbox"/>
			Download selected

The download is in the format of a zipped (compressed) folder. After unzipping the folder, we can open the 'CLOSER_training_dataset_complete_cases.dta' file in Stata.

The Stata syntax file that accompanies this guide includes preliminary code to prepare the data for the analyses we want to perform. We first drop the variables we are not interested in currently. This variable selection is done using Stata's '**keep**' command as shown below (note that in the code snippets below and throughout this module, Stata commands are in **bold** font and the variable names are in *italics*)."

Command	keep <i>ncdsid bmi42 n920 n622 n016nmed n716dade n1171 bmi11</i>
---------	---

For these analyses, we are adopting a complete case analysis approach. That means that in preparing the dataset, we are excluding any cases where there are missing data on any of the variables of interest. (Missing data can be handled in alternative ways, such as through the use of data imputation techniques). To remove the incomplete cases, we first want to ensure that all of the variables use the same missing value code (".") as illustrated in the Stata code snippet below.

Command	<pre> foreach x of varlist <i>n622-bmi42</i>{ replace `x'=. if inrange(`x',-9,-1) } replace <i>n1171</i>=. if <i>n1171</i>==8 </pre>
---------	---

We then need to run the following set of commands in Stata to create a temporary variable denoting cases with incomplete data (*miss1*). We can then remove cases with any incomplete data using the ‘drop if’ command.

Command	<pre> gen <i>miss1</i>=. replace <i>miss1</i>=0 if missing(<i>bmi42</i>, <i>n920</i>, <i>n622</i>, <i>n016nmed</i>, <i>n716dade</i>, <i>n1171</i>, <i>bmi11</i>) replace <i>miss1</i>=1 if !missing(<i>bmi42</i>, <i>n920</i>, <i>n622</i>, <i>n016nmed</i>, <i>n716dade</i>, <i>n1171</i>, <i>bmi11</i>) drop if <i>miss1</i>==0 drop <i>miss1</i> </pre>
---------	--

What does the dataset contain?

Now that the dataset is loaded and initial preparation is complete, we can begin exploring the data. By running the Stata command ‘describe’, we will get a summary of the dataset, including the number of observations and a table of the variable names and labels.

Command	describe
Output	Contains data from D:\CLOSER\Method 1\Feb 2018\CTD_1.dta obs: 4,497 vars: 8 2 Mar 2018 13:04 size: 346,269

There are 4,497 observations and 8 variables. The *ncdsid* variable comprises unique identifier codes for each study participant. Other variables in the dataset include the study participant’s family background, whether their mother and father left education at the minimum age or not (*n016nmed*, *n716dade*) and their father’s social class (*n1171*). *n622* is the sex of the study participant, while early

life factors include their ‘general ability’ (*n920*) and body-mass index at age 11 (*bmi11*) and our outcome variable body-mass index at age 42 (*bmi42*). Note that ‘CM’ in some of the variable labels stands for ‘cohort member’, i.e. the participants in the study.

We can use the ‘**summarize**’ command to learn more about the variables we will employ in our analyses.

Command	summarize <i>bmi42 n920 bmi11 n622 n016nmed n716dade n1171</i>					
Output	Variable	Obs	Mean	Std. Dev.	Min	Max
	<i>bmi42</i>	4497	25.86068	4.431863	14.74405	51.71761
	<i>n920</i>	4497	46.64421	14.93775	0	79
	<i>bmi11</i>	4497	17.46035	2.573711	11.66545	37.74945
	<i>n622</i>	4497	1.523905	.4994838	1	2
	<i>n016nmed</i>	4497	.2781855	.4481551	0	1
	<i>n716dade</i>	4497	.2739604	.4460385	0	1
	<i>n1171</i>	4497	3.75517	1.562278	1	7

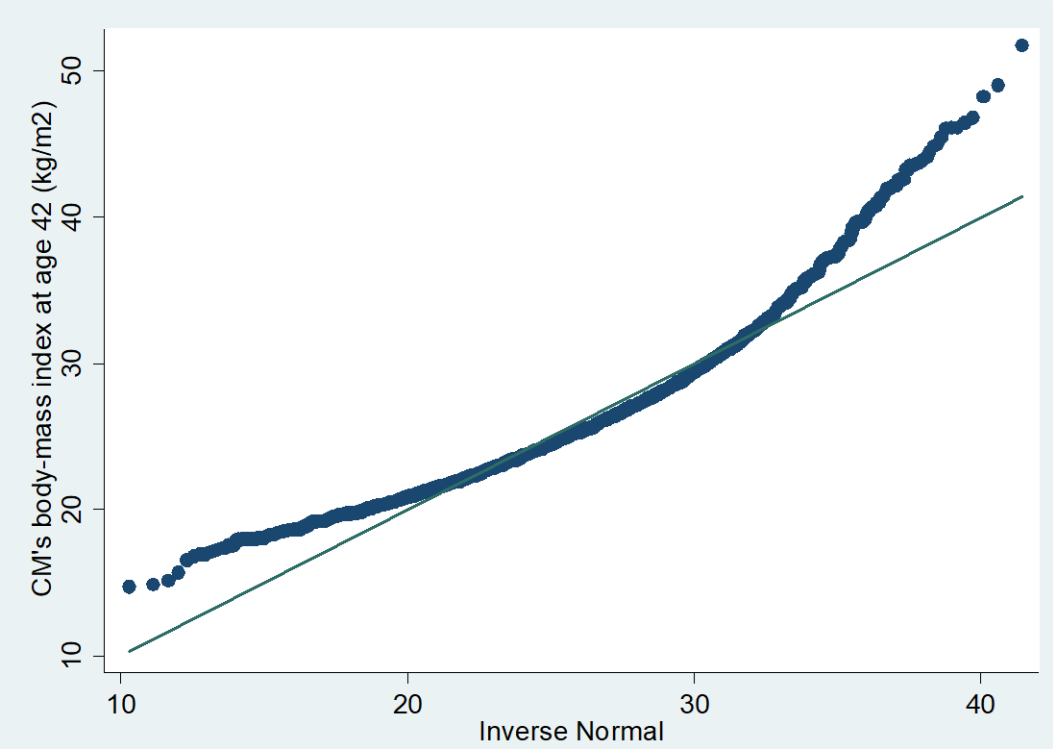
As you can see there are no missing data; each variable has 4,497 observations. Although survey datasets will usually have at least some missing data, we have already removed any study participants with missing data for the purposes of our analyses. As indicated by the minimum and maximum values in the output table, the dataset has 3 continuous variables (*bmi42*, *n920* and *bmi11*), 3 dichotomous variables (*n622*, *n016nmed*, and *n716dade*), and 1 categorical variable (*n1171*).

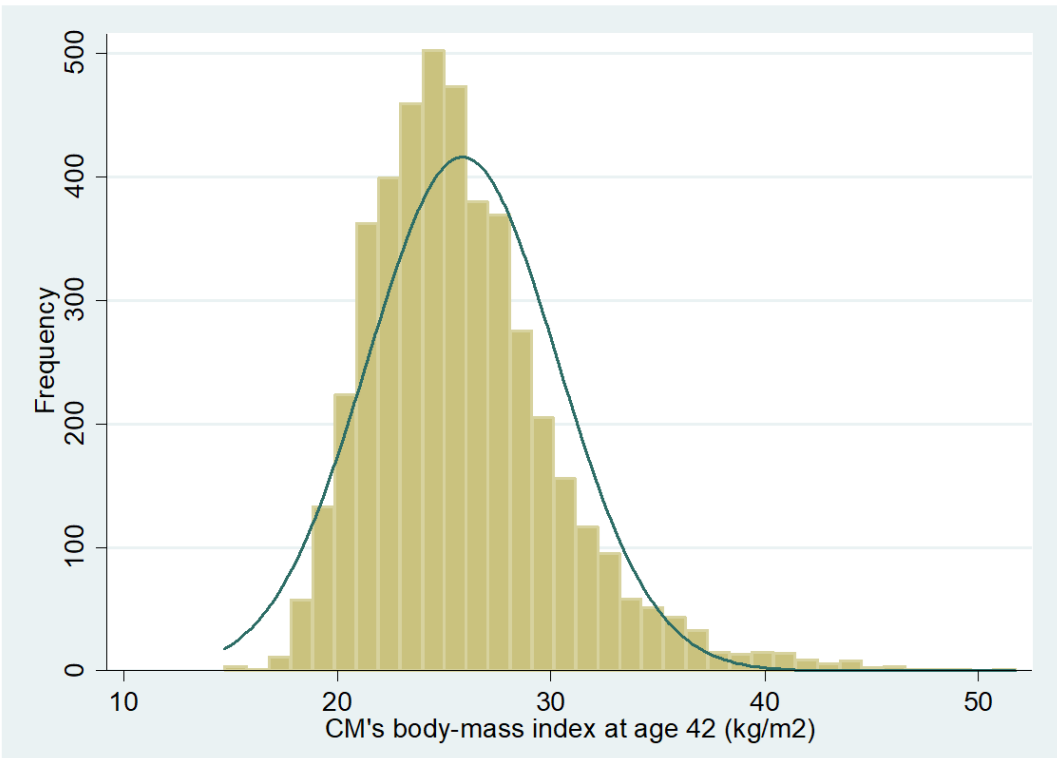
Examining the predictor and outcome variables

We can also use the ‘**summarize**’ command to get more detailed information on our two main variables of interest – our outcome, BMI at age 42 (*bmi42*), and our predictor variable, ‘general ability’ at age 11 (*n920*). You should note that ‘**summarize**’, as well as other Stata commands, can often be abbreviated to keep your command syntax concise. So instead of typing out the full ‘**summarize**’ command, we can instead use ‘**sum**’, which Stata will interpret in the exact same way. Stata commands also often allow us to specify additional options to customise the output we get when we run the command. If we use the ‘**detail**’ option with the ‘**sum**’ command for example, the Stata output will also include percentiles, measures of central tendency and variance.

Command	sum bmi42 n920, detail				
Output	CM's body-mass index at age 42 (kg/m2)				
	<hr/>				
		Percentiles	Smallest		
	1%	18.44472	14.74405		
	5%	20.04742	14.87977		
	10%	20.96727	15.14303	Obs	4497
	25%	22.79416	15.73226	Sum of Wgt.	4497
	50%	25.21589		Mean	25.86068
			Largest	Std. Dev.	4.431863
	75%	28.08403	46.83073		
	90%	31.44282	48.2391	Variance	19.64141
	95%	34.17019	49.01731	Skewness	1.132382
	99%	40.67343	51.71761	Kurtosis	5.243019
	2T Total score on general ability test, CM age 11				
	<hr/>				
		Percentiles	Smallest		
	1%	13	0		
	5%	20	0		
	10%	26	0	Obs	4497
	25%	36	0	Sum of Wgt.	4497
	50%	48		Mean	46.64421
			Largest	Std. Dev.	14.93775
	75%	58	78		
	90%	66	79	Variance	223.1363
	95%	69	79	Skewness	-.2827034
	99%	74	79	Kurtosis	2.40289

From the output, we can see that BMI at age 42 ranges from 14.74 to 51.72, with a mean of 25.86 and a median of 25.22 (the 50% percentile). General ability at age 11 ranges from 0 to 79, with a mean of 46.64 and a median of 48. The distribution of BMI at age 42 is not symmetrical (skewness = 1.13) and is heavy on the tails of the distribution (kurtosis = 5.24) which we can examine graphically using the **'qnorm'** and **'histogram'** commands, as shown in the plots below.

Command	<code>qnorm bmi42</code>
Output	 <p>The figure is a Q-Q plot comparing the empirical distribution of BMI at age 42 against a theoretical normal distribution. The x-axis, labeled 'Inverse Normal', has major ticks at 10, 20, 30, and 40. The y-axis, labeled 'CM's body-mass index at age 42 (kg/m2)', has major ticks at 10, 20, 30, 40, and 50. A solid green line represents the theoretical normal distribution. The data points, represented by dark blue dots, follow this line closely for values between approximately 10 and 30 kg/m². For values greater than 30 kg/m², the data points curve upwards above the green line, suggesting that the actual distribution has a heavier right tail than a normal distribution.</p>

Command	<code>histogram bmi42, frequency normal</code>
Output	 <p>The figure is a histogram showing the frequency distribution of BMI at age 42. The x-axis is labeled 'CM's body-mass index at age 42 (kg/m2)' and ranges from 10 to 50. The y-axis is labeled 'Frequency' and ranges from 0 to 500. The histogram bars are olive green. A dark teal normal distribution curve is overlaid on the histogram, peaking at a frequency of approximately 420 around a BMI of 25. The distribution is slightly right-skewed, with a long tail extending towards higher BMI values.</p>

We will examine these in more detail when we investigate the regression diagnostic at the end of the general linear regression example.

Preparing the data for modelling

First, we are going to examine gender (*n622*), a dichotomous variable, to look at how this is coded. The **'codebook'** command is particularly useful for looking at categorical variables.

Command	codebook <i>n622</i>
Output	<pre> type: numeric (double) label: n622 range: [1,2] units: 1 unique values: 2 missing .: 0/4497 tabulation: Freq. Numeric Label 2141 1 Male 2356 2 Female </pre>

The *n622* variable is coded 1=Male and 2=Female. There are 2,141 males in our data and 2,356 females.

For our regression analysis, we will recode the data to create a new binary variable (which we will label 'sex' and in which we will recode the values as 0=Male and 1=Female). Such binary variables are often known as dummy variables. Although the coefficients would work out the same if the variable was coded as 1/2 or 0/1, the intercept (labelled as “_cons” in the output) would be less intuitive. In our regression analysis, we will use males as the reference group.

Command	<pre> gen sex = . replace sex = 1 if <i>n622</i>==2 replace sex = 0 if <i>n622</i>==1 label define sexL 0 "male" 1 "female" label values sex sexL </pre>
---------	---

The second variable we are going to look at is father's social class (*n1171*).

Command	codebook n1171
Output	<pre> type: numeric (double) label: n1171 range: [1,7] units: 1 unique values: 7 missing .: 0/4497 tabulation: Freq. Numeric Label 274 1 Social class I 910 2 Social class II 484 3 SC III non-man. 1892 4 SC III manual 75 5 SC IV non-manual 636 6 SC IV manual 226 7 Social class V </pre>

The *n1171* variable has 7 categories ranging from 1='Social class I' to 7='Social class V'. Some of the categories have low numbers of observations. For example, 'SC IV non-manual' has only 75 observations, so we will combine some of the categories to increase the number of observations they capture by creating a new variable with fewer categories using the '**gen**' and '**replace**' commands.

Command	<pre> gen n1171_2 = . replace n1171_2 = 1 if n1171==1 n1171==2 replace n1171_2 = 2 if n1171==3 replace n1171_2 = 3 if n1171==4 replace n1171_2 = 4 if n1171==5 n1171==6 replace n1171_2 = 5 if n1171==7 label define n1171_2L 1 "I/II Prof & Managerial" 2 "III Skilled non-manual" 3 "III Skilled manual" 4 "IV Partly skilled" 5 "V unskilled" , modify label values n1171_2 n1171_2L </pre>
---------	---

We have now created a new variable *n1171_2* which collapses social class I and II from *n1171* into a combined I and II professional and managerial category which we will use as our reference group. These two categories are often combined into a single high social class grouping. The second change we have made is combining the 'SC IV non-manual' category with only 75 observations with the 'SC IV manual' category to create a single IV category with 711 observations. With only 75 observations it may increase the chance that we may find no association with BMI at age 42 in the non-manual unskilled category (compared to the higher social classes) as a consequence of the low sample size, even if there actually is a relationship. We can examine the difference between the original and recoded variable using the '**tab**' command.

Command	tab n1171 n1171_2						
Output	2P 1970-style Social Class of father or male head at CM age 11 (1969)	n1171_2					
		I/II Prof	III Skill	III Skill	IV Partly	V unskill	Total
	Social class I	274	0	0	0	0	274
	Social class II	910	0	0	0	0	910
	SC III non-man.	0	484	0	0	0	484
	SC III manual	0	0	1,892	0	0	1,892
	SC IV non-manual	0	0	0	75	0	75
	SC IV manual	0	0	0	636	0	636
	Social class V	0	0	0	0	226	226
	Total	1,184	484	1,892	711	226	4,497

As you can see from the output table above, social class *n1171_2* now has 5 categories. We can now proceed to the next steps in our analysis, where we will undertake statistical modelling to explore research questions with the data.