Regression analysis for longitudinal data

Authors: Vanessa Moulton, Dara O'Neill, Alison Park and George B. Ploubidis

Methods of analysis of data from longitudinal studies allow us to make use of their rich data and to explore the temporal relationships between measures collected across different life stages. Regression analysis is an important and widely-used technique for exploring the relationship between an outcome (e.g. later-life health) and possible explanatory variables (e.g. early-life circumstances). We can gain important insights in social science, biomedical and health research by studying a range of factors throughout the life course, including physical and mental health, and socioeconomic and behavioural factors

In this module you will learn about:

- The advantages of longitudinal data over cross-sectional data analysis
- How to explore a longitudinal dataset and prepare it for analysis
- How to apply general linear, logistic and multinomial regression techniques

Challenge level: advanced

Key concepts:

- Answering research questions with a longitudinal dimension
- Preparing data for longitudinal data analysis
- Examining associations between outcomes and potential explanatory variables
- Adapting analyses for different types of outcome variable
- Updating and comparing statistical models

Suggested citation: Moulton, V., O'Neill, D., Park, A. & Ploubidis, G.B. (2020). *Regression analysis for longitudinal data.* CLOSER Learning Hub, London, UK: CLOSER

Authors: Vanessa Moulton, Dara O'Neill, Alison Park and George B. Ploubidis

1 Introduction and overview

This section introduces some of the important fundamentals of analysing data from longitudinal studies and describes how regression techniques can be used to explore variables relating to different points in an individual's life course.

1.1 Analysing data from longitudinal studies

The utility of longitudinal studies and the differences between longitudinal and cross-sectional designs are described more fully in the Learning Hub's Introduction to Longitudinal Studies. There are data analysis methods that allow us to make use of the rich data collected by longitudinal studies and to explore the temporal relationships between measures collected across different life stages. Each of these is suited to the analysis of different types and combinations of variables. Some variables are continuous (e.g. age) and others are categorical (e.g. a list of occupations). We call categorical variables with two levels 'dichotomous' (e.g. deceased or living) and, where they are coded as 0 or 1, we can also call them 'binary'. This guide will teach you about different analytic approaches to exploring how certain types of outcomes are associated with potential explanatory factors.

Dissimilar outcomes can occur even among people who share the same characteristics. The term 'heterogeneity' is often used to refer to differences like these. Longitudinal data can help control for such differences by including a wide range of explanatory variables across the life course in statistical models. The problem of 'omitted variable bias' is also improved by using longitudinal data, but always remains, as there are connections between the outcome and explanatory variables that have not or could not be included as they are unmeasurable.

Suggested citation: Moulton, V., O'Neill, D., Park, A. & Ploubidis, G.B. (2020). *Regression analysis for longitudinal data*. CLOSER Learning Hub, London, UK: CLOSER

Authors: Vanessa Moulton, Dara O'Neill, Alison Park and George B. Ploubidis

We will use an extract from the National Child Development Study (NCDS) CLOSER Training Dataset to illustrate some of the different methods that can be used in analysing longitudinal data. The NCDS is a cohort study of people born in England, Scotland and Wales during a single week of 1958. In the NCDS, detailed information has been collected on participants from childhood, through adolescence into early adulthood and later life, allowing us to look at different outcomes and potential explanatory variables.

Measurements that have been collected over time include assessments of physical health (e.g. Body Mass Index (BMI) measured at ages 7, 11, 16, 23, 33, 42 and 50), as well as a series of mental health (e.g. Malaise inventory), socio-economic position, and behavioural factors (e.g. smoking), measured at ages 23, 33, 42, and 50. These measures are examples of the variety of data available in the NCDS and other longitudinal studies.

1.2 Overview of this guide

In the following sections, we will present a variety of longitudinal data techniques you can apply to longitudinal data and repeated measures. First, we will explore and prepare the dataset before demonstrating how to apply general linear, logistic and multinomial regression approaches which are commonly used in the analysis of longitudinal study data. In future updates to this module, we will also illustrate how to transfer data to a format suitable for repeated measures analysis. We will also be adding guidance on techniques for analysing such repeated measures data, including multilevel regression, fixed effects, and latent growth models.

We will guide you through these methods as performed in the STATA statistical software package, and we will provide documented syntax to explain the steps involved. Guidance for other statistical software packages is forthcoming.

Suggested citation: Moulton, V., O'Neill, D., Park, A. & Ploubidis, G.B. (2020). *Regression analysis for longitudinal data*. CLOSER Learning Hub, London, UK: CLOSER

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.

2.1 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 the efficient functioning of other complex systems in the body) 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 body-mass index (BMI) and intelligence, where higher IQ scoring individuals interpret and respond to health advice in more positive ways.

Consequently, in this modules, we will use data from the CTD and apply a series of different analytic techniques to explore the relationship between childhood intelligence and adult BMI.

2.2 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 (m2).

2.2.1 Potential explanatory variable

Our childhood explanatory variable, i.e. our predictor of interest, is 'general ability'. At age 11, the NCDS cohort were given a general ability test, which required the children participating in the study to recognise patterns in either words or pictures and correctly identify the next word/picture in a sequence. Their total score on this task represents their 'general ability' at that age, and this total score can range between 0 and 80. This variable is also available in the CTD.

2.3 Accessing and preparing the dataset

2.3.1 Accessing the CTD 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 formatted data from the download options. This can be completed on the UKDS website.

File Format	File Size (mb)	Download	
Dataset: National Child Development Study: CLOSER Training Dataset, 1958-2013			
SPSS	0.66	Download	☐
STATA	0.64	Download	☐
TAB	0.79	Download	☐
			Download selected

Figure 1: Screenshot of download options for the CTD

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.

2.3.2 Accessing the complete Stata syntax

We have prepared a Stata syntax file (a .do file) to accompany this module. It includes all of the commands discussed in the following sections and we recommend you open it up in Stata alongside the CTD data.

Download the syntax file

2.3.3 Preparing the data for our analyses

Now we have the data, our first step will be to simplify the dataset by dropping the variables not currently relevant to us. 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	
Command	**keep** *ncdsid bmi42 n920 n622 n016nmed n716dade n1171 bmi11*

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

```
foreach x of varlist n622-bmi42{
replace `x'=. if inrange(`x',-9,-1)
}
replace n1171=. if n1171==8
```

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	
	gen *miss1*=.
	replace *miss1*=0 **if missing(***bmi42*, *n920*, *n622*, *n016nmed*, *n716dade*, *n1171*, *bmi11***)**
	replace *miss1*=1 **if!missing(***bmi42*, *n920*, *n622*, *n016nmed*, *n716dade*, *n1171*, *bmi11***)**
	drop if *miss1*==0
	drop *miss1*

The data are now ready for some initial exploration of the variables of interest.

2.4 What does the dataset contain?

Now that the dataset is loaded and initial preparation is complete, we can begin exploring the data.

2.4.1 Looking at the contents of the dataset

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
-------------------------------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------------------------------------
ncdsid          str21   %21s                  ncdsid serial number
n622            double  %12.0g     n622       Sex of NCDS cohort member
n016nmed        double  %12.0g     n016nmed   Mother left education at min age or not [derived from age 0 and 16]
n716dade        double  %12.0g     n716dade   Father left education at min age or not [derived from age 7 and 16]
n1171           double  %12.0g     n1171      2P 1970-style Social Class of father or male head at CM age 11 (1969)
n920            double  %12.0g     n920       2T Total score on general ability test, CM age 11
bmi11           double  %12.0g                CM's body-mass index at age 11 (kg/m2)
bmi42           double  %12.0g                CM's body-mass index at age 42 (kg/m2)
-------------------------------------------------------------------------------------------------------------
Sorted by:  ncdsid
```

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.

2.4.2 Looking at the contents of the variables

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

Command	**summarize** *bmi42 n920 bmi11 n622 n016nmed n716dade n1171*
Output	see below

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       bmi42 |      4497    25.86068    4.431863   14.74405   51.71761
        n920 |      4497    46.64421    14.93775          0         79
       bmi11 |      4497    17.46035    2.573711   11.66545   37.74945
        n622 |      4497    1.523905    .4994838          1          2
    n016nmed |      4497    .2781855    .4481551          0          1
-------------+---------------------------------------------------------
    n716dade |      4497    .2739604    .4460385          0          1
       n1171 |      4497     3.75517    1.562278          1          7
```

As you can see from the output table above, 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*).

2.5 Examining the predictor and outcome variables

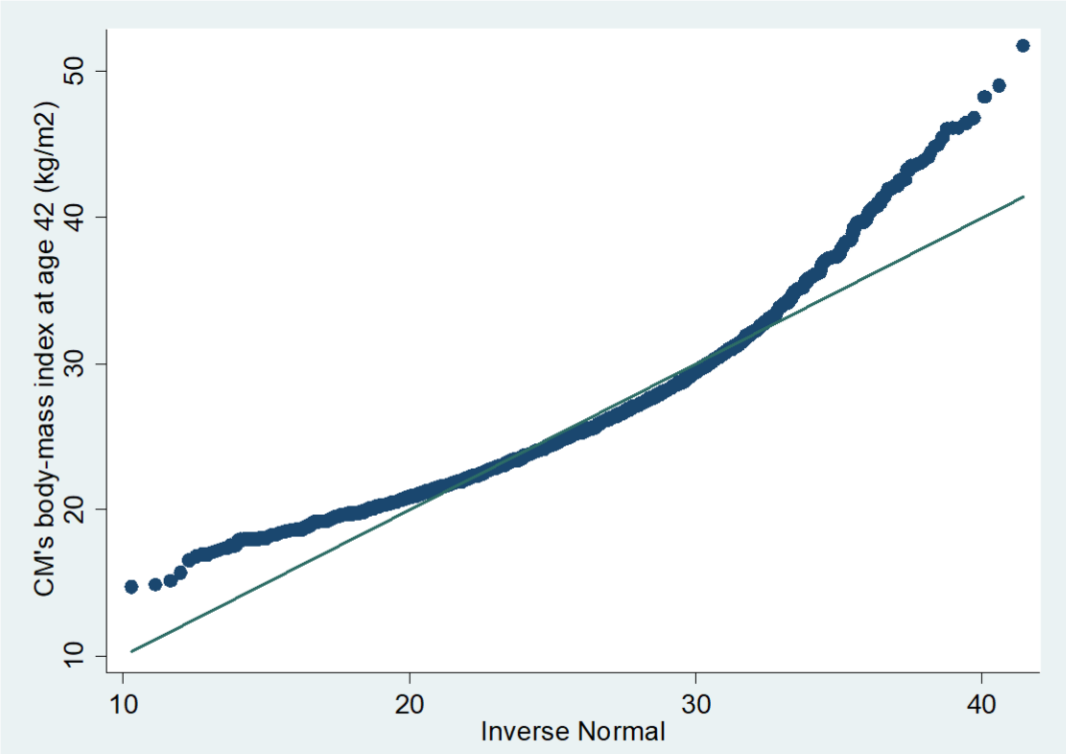
We can also use the **'summarize'** command to get even more detailed information on our two main variables of interest – our outcome, BMI at age 42 (*bmi42*), and our predictor variable, 'general ability' at age 11 (*n920*). You should note that **'summarize'**, as well as other Stata commands, can 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**

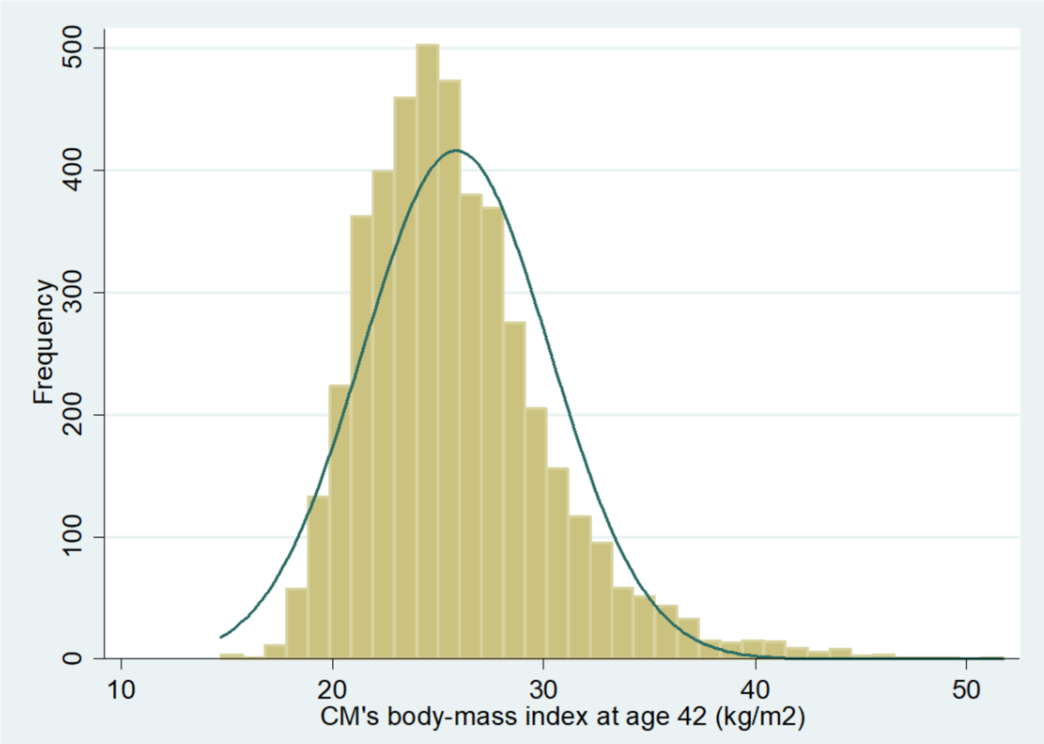
Output

```
              CM's body-mass index at age 42 (kg/m2)
-------------------------------------------------------------
      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
-------------------------------------------------------------
      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 50th 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	**qnorm** *bmi42*
Output	

Suggested citation: Moulton, V., O'Neill, D., Park, A. & Ploubidis, G.B. (2020). *Regression analysis of longitudinal data*. CLOSER Learning Hub, London, UK: CLOSER

Command	**histogram** *bmi42*, **frequency normal**
Output	

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

2.6 Preparing the data for modelling

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

Command	**codebook** *n622*

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	**gen** *sex* **=** . **replace** *sex* **=** 1 **if** *n622*==2 **replace** *sex* **=** 0 **if** *n622*==1 **label define** *sexL* 0 "male" 1 "female" **label values** *sex sexL*

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

Command

codebook *n1171*

Output

```
              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
```

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

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*

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.

Authors: Vanessa Moulton, Dara O'Neill, Alison Park and George B. Ploubidis

3 General linear regression

This section introduces a method, known as general linear regression, that can be used to examine how an outcome that has been measured on a continuous scale is associated with potentially explanatory variables. We offer a step-by-step illustration of how we can use this important statistical analysis approach to explore such associations in longitudinal data.

3.1 What is general linear regression?

General linear regression enables us to evaluate the association between a continuous outcome variable and one or more continuous or categorical predictor variables. The model we fit is linear, which means we summarise the data with a straight line that best describes the data by minimising the distance between the actual data and the predictions of the regression line. Multiple regression allows us to determine the overall fit of the model and the relative contribution of each of the predictors to the variance explained. With our longitudinal data, we can try and explain a later life outcome for a particular person by whatever model we fit to the data using information about that person from earlier in their life.

3.2 Example research question: Is childhood intelligence related to body-mass index (BMI) in middle age?

In this regression, the outcome variable *bmi42* is a continuous variable that includes all values of BMI at age 42. In the first model we will analyse, there is only one predictor variable 'general ability' at age 11 (*n920*), which is also a continuous variable.

It is always important to explore the data before running statistical models. If you have not yet done so, please first look at exploring the data to learn how you can examine the data. You will also need to have first derived a few of the explanatory variables, see main variables of interest, before proceeding with the regression modelling. In this work, we will adopt a significance threshold of *p*=.05, meaning that we will infer statistical significance for p-values that fall below this cutoff.

3.3 Running the regression

In Stata, linear regressions can be run with the **‘regress’** command. This can be abbreviated to **‘reg’** in our code to keep our commands concise. To run the **‘reg’** command appropriately, we must specify the outcome variable immediately after the **‘reg’** command in our syntax, followed by the predictor variable(s). This is the order used in the code snippet below:

Command

reg *bmi42 n920*

Output

```
      Source |       SS           df       MS      Number of obs   =      4497
-------------+----------------------------------   F(  1,   4495)  =     61.29
       Model |  1187.90472         1  1187.90472   Prob > F        =    0.0000
    Residual |  87119.8689      4495  19.3815059   R-squared       =    0.0135
-------------+----------------------------------   Adj R-squared   =    0.0132
       Total |  88307.7736      4496  19.6414087   Root MSE        =    4.4024

------------------------------------------------------------------------------
       bmi42 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        n920 |  -.0344106   .0043954    -7.83   0.000    -.0430277   -.0257935
       _cons |   27.46574   .2152731   127.59   0.000      27.0437    27.88778
------------------------------------------------------------------------------
```

Looking at the output table above, we can see that the p-value of the F-test (=61.29, *p*<.001 is below our adopted significance threshold of which means we can say that the model statistically significant. r-squared value approximately variance bmi at age accounted for by model. as there only one predictor this aria-describedby="tt" class="glossaryLink" data-cmtooltip="General ability is a term used to describe cognitive ability, and is sometimes used as a proxy for intelligent quotient (IQ) scores.">general ability’ at age 11 explains only 1.4% of the variance of BMI at age 42. The coefficient for *n920* is -.0344106 or approximately -.03, meaning that for 1 unit increase in general ability, we would expect a .03 decrease in BMI at age 42. Put more simply, a study participant with a general ability score of 60 at age 11 would have a 1 unit lower BMI score at age 42 than a study participant with a general ability score of 30 at age 11. The intercept (or constant) is 27.47 and this is the predicted value of BMI at age 42 when 'general ability' equals zero.

In the next section, we will look at how we can plot our results.

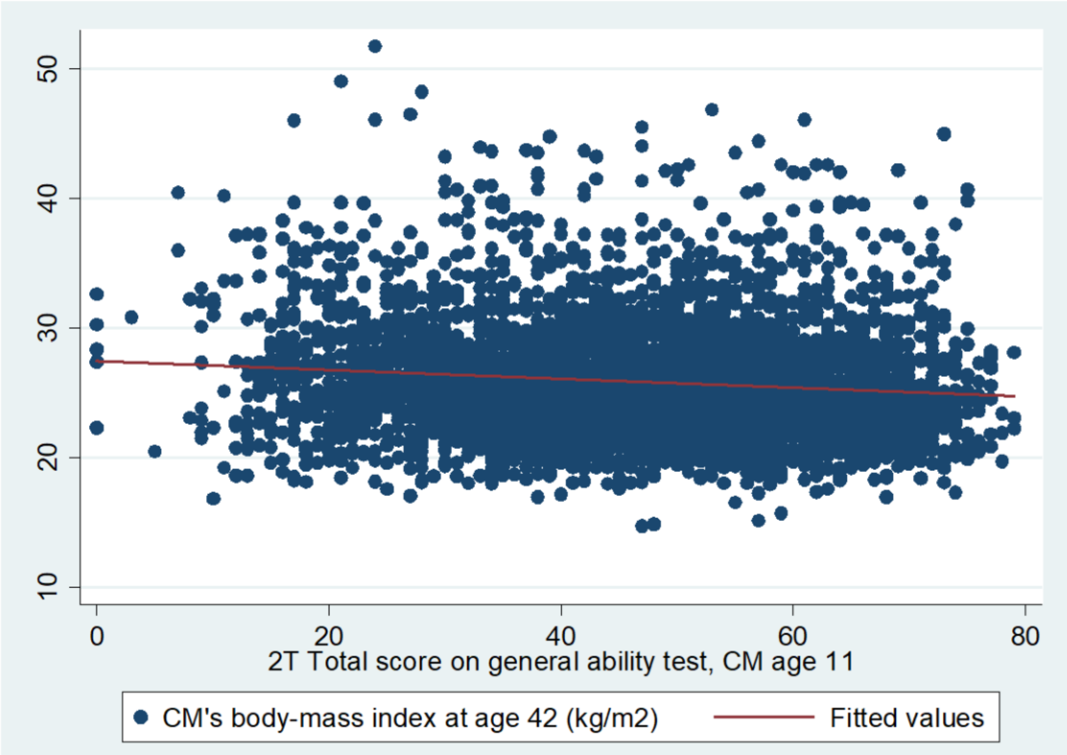
3.4 Plotting the results

To help visualise our results, we can create a scatterplot of the outcome and the predictor variables with the regression line plotted on top. This involves two steps:

1. After running the regression, we create a variable containing the predicted values (which we have named *bmi_iq1*) using the **'predict'** command.

Command	**predict** *bmi_iq1*

2. Then to create the plot, we use the Stata **'twoway (scatter …)'** graph command, in combination with the **'(lfit …)'** command to overlay the regression line.

Command	**twoway (scatter** *bmi42 n920***) (lfit** *bmi42 n920***)**
Output	

Running the above commands with our data, the plot we generate has 'BMI at age 42' on the Y axis and 'general ability at age 11' on the X axis. The fitted regression line slopes from the left of the plot (where the intercept for 'BMI at age 42' is 27.5) to the right (where a 'general ability' score of 80 equals a 'BMI at age 42' of 24.7). However, the slope is fairly flat, which is to be expected given the small regression coefficient (-.03) we obtained in the previous step when we ran the **'reg'** command.

What we have run here is often called a simple regression, as it contains only one predictor variable. We may get a more informative insight if we extended our model to consider other variables that may influence the association between our predictor and outcome variables, and that is exactly what we will do in the next section.

3.5 Updating the regression model

3.5.1 Including potential confounding variables

We are now going to extend our model to consider variables that may influence or confound the association between our predictor and outcome variables. These new variables being considered are: sex, parents' education and family social class.

The *sex* variable has already been recoded to be binary (see the Preparing the data for modelling section) and in this regression analysis we are using the category 'male' as the reference group.

In addition, we are going to include a few family background factors in the model. These include two parental education measures that denote whether the participant's mother (*n016nmed*) and father (*n716dade*) left school at the minimum age or not; these are also binary variables. For both of these variables, we are using the 'left school at the minimum age' as the reference group.

The final potential confounder we are including is the social class of the study participant's father (*n1171_2*). This is a categorical variable with 5 values. In Stata you can automatically create dummy variable(s) for each of the values in a multi-category variable by appending the prefix of **'i.'** to the variable name, e.g. **i.***n1171_2***.** In this instance, it means that the model will compare each of 'III Skilled non-manual', 'III Skilled manual', 'IV Partly skilled' and 'V unskilled' against the 'I/II Prof & Managerial' category. Stata will use 'I/II Prof & Managerial' as the reference category simply because it is the first category in the variable.

Command

reg *bmi42 n920* **i.***sex n016nmed n716dade* **i.***n1171_2*

Output

Source	SS	df	MS			
Model	3544.33638	8	443.042048	Number of obs	=	4,497
Residual	84763.4372	4,488	18.8866839	F(8, 4488)	=	23.46
				Prob > F	=	0.0000
				R-squared	=	0.0401
				Adj R-squared	=	0.0384
Total	88307.7736	4,496	19.6414087	Root MSE	=	4.3459

bmi42	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
n920	-.0202903	.0046853	-4.33	0.000	-.0294757	-.0111049
sex						
female	-1.143607	.130116	-8.79	0.000	-1.398698	-.8885155
n016nmed	-.4688374	.1600576	-2.93	0.003	-.782629	-.1550457
n716dade	-.2127517	.169213	-1.26	0.209	-.5444926	.1189892
n1171_2						
III Skilled non-manual	-.0447093	.2370468	-0.19	0.850	-.5094379	.4200192
III Skilled manual	.6271419	.1837793	3.41	0.001	.266844	.9874398
IV Partly skilled	.5702515	.2271172	2.51	0.012	.1249899	1.015513
V unskilled	1.0015	.3332483	3.01	0.003	.348169	1.654831
_cons	27.19543	.2863722	94.97	0.000	26.63399	27.75686

From the output table above, we can see that including the study participant's sex and family background factors have not markedly changed the model. A small proportion, 4%, of the variance of BMI at age 42 is accounted for by family background, general ability at age 11 and the sex of the study participant. The participant's general ability is still significant; for a 1 unit increase in general ability, we can expect a .03 decrease in BMI at age 42. The average BMI for females at age 42 is 1.14 lower than males, taking account of general ability at age 11. If the participant's mother did not leave school at the minimum age, on average the participant's BMI at age 42 was .47 lower than a participant whose mother left school. The father staying on at school was not significant, as this was explained by the father's social class which was also included in the model. Social class and education are highly correlated; an individual's educational attainment will in part reflect later occupational status which determines social class (You can explore this yourself as the syntax for the model above with social class excluded has been provided in the Stata .do file that accompanies this module). Compared to a participant whose father was in the highest social classes (I and II), having a father in the skilled and partly skilled manual social classes increased a participant's BMI by .63 and .57 respectively (if all other factors remained equal). If the participant's father was instead in the unskilled class, the increase in BMI was on average higher by 1.

3.5.2 Including a childhood measure of BMI

In our final model we add *bmi11*, the BMI of the study participants when they were aged 11. By adding BMI at age 11 we adjust for earlier measures of BMI, thereby focusing on the change in BMI from age 11 to age 42. This allows us to measure BMI and general ability over a comparable duration from the age of 11 to 42 years.

Command	**reg** *bmi42 n920* **i.***sex n016nmed n716dade* **i.***n1171_2 bmi11*
Output	(see below)

Source	SS	df	MS
Model	22791.6156	9	2532.40174
Residual	65516.158	4487	14.6013278
Total	88307.7736	4496	19.6414087

Number of obs	= 4497
F(9, 4487)	= 173.44
Prob > F	= 0.0000
R-squared	= 0.2581
Adj R-squared	= 0.2566
Root MSE	= 3.8212

bmi42	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
n920	-.020955	.0041196	-5.09	0.000	-.0290314	-.0128785
sex						
female	-1.423287	.1146651	-12.41	0.000	-1.648087	-1.198487
n016nmed	-.2417224	.1408715	-1.72	0.086	-.5178999	.0344552
n716dade	-.0690629	.1488352	-0.46	0.643	-.3608533	.2227274
n1171_2						
III Skilled non-manual	.1682037	.2085088	0.81	0.420	-.2405762	.5769836
III Skilled manual	.7120927	.1616071	4.41	0.000	.3952632	1.028922
IV Partly skilled	.6390438	.1997045	3.20	0.001	.2475246	1.030563
V unskilled	1.069299	.2930186	3.65	0.000	.4948385	1.64376
bmi11	.8075547	.0222425	36.31	0.000	.7639485	.851161
_cons	13.09728	.4627985	28.30	0.000	12.18997	14.0046

The R-squared value in the output table above tells us that a quarter (25.8%) of the variance of BMI at age 42 is accounted for when we include BMI at age 11, as well as family background, general ability at age 11 and the sex of the participant, in the model. We can infer from the fact that mother's education is no longer a significant predictor in this updated model that childhood BMI explains its significance in the earlier model. However, all other factors that were significant in the earlier less-adjusted model remain significant in this updated model, including our 'general ability' predictor variable. It may be that the influence of mother's education on the participant's midlife BMI, for example, reflects the family's early eating habits, physical activity and health behaviours, which would be more influential in a child's early life and therefore be reflected in their childhood BMI. For a 1 unit increase in general ability, we would expect a .02 decrease in BMI at age 42. In other words, a participant with a general ability score of 60 at age 11 would have a .63 lower BMI score at age 42 than a study participant with a general ability score of 30 at age 11, after controlling for BMI at age 11 and other factors.

However, we have still only explained a quarter (25.8%) of the variance in BMI at age 42. There are other factors, not included in this analysis which may play a role in that unexplained variance as they are known to be associated with BMI, such as physical activity, diet, sleep duration, socio-economic factors in later life, parent's BMI and genetic factors.

3.6 Regression diagnostics

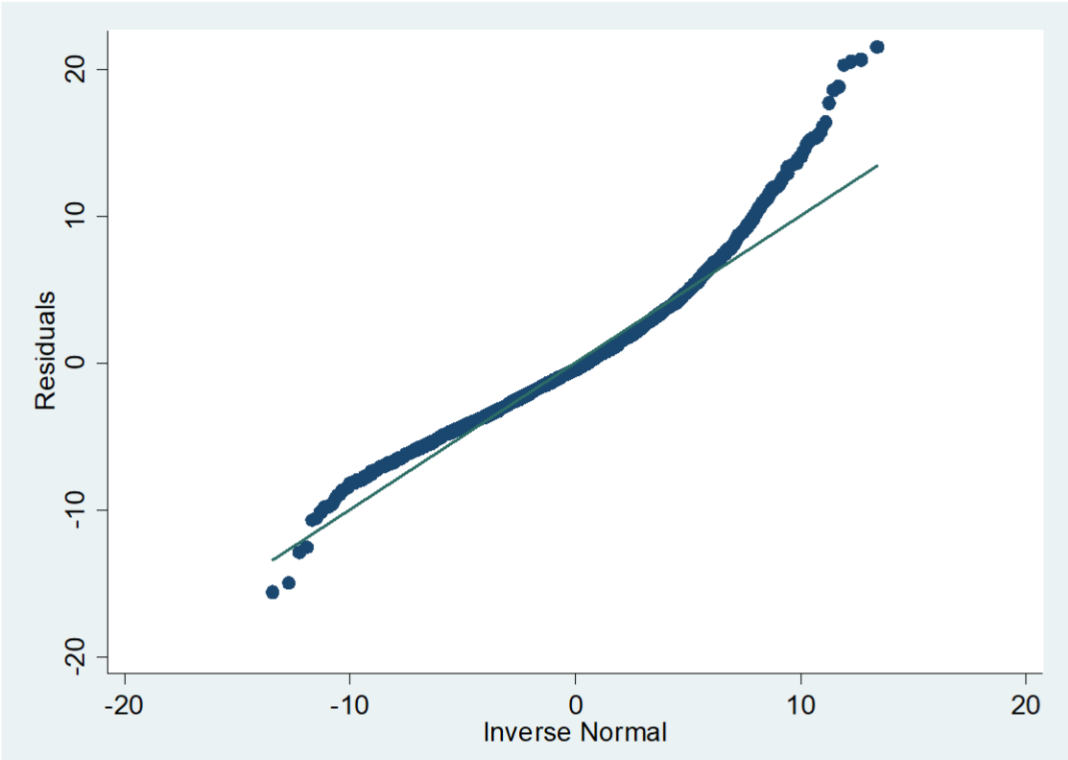
We have conducted this analysis without checking whether the data we have been using have met the assumptions underlying an ordinary least squares (OLS) linear regression. Three main assumptions we will however now briefly explore are normality, homogeneity of variance (homoscedasticity) and independence. Normality of residuals is only required for valid hypothesis testing, where we need to ensure the p-values are valid; it is not required to obtain unbiased estimates of the regression coefficients. OLS requires that the residuals are identically and independently distributed, i.e. the observed error (the residual) is random.

3.6.1 Normality

First, we will formally test the normality of residuals to identify if we can use our analysis for valid hypothesis testing. After running our final regression analysis, we can use the **'predict'** command with the **'resid'** option to calculate the residuals. We can store these residual values as a variable, which in this case we will call bmi_iq2, and we can then use this variable to then check the residuals' normality.

Command	**predict** ***bmi_iq2*****, resid**

We can plot the residuals against a normal distribution, using either the **'pnorm'** (which is sensitive to non-normality in the middle range of data) or **'qnorm'** (which is sensitive to non-normality near the tails) commands. We are going to look at the **'qnorm'** method, as we suspect that BMI is non-normal at the tails of the distribution. Previous research indicates that BMI is not symmetrical but is always skewed to the right, toward a higher ratio of weight (body mass) to height.

Command	**qnorm** *bmi_iq2*
Output	

In the above output, the **'qnorm'** command has plotted quintiles of the residuals of BMI at age 42 (the thicker dotted line) against the quintiles of a normal distribution (the thin diagonal line). If the two lines were exactly the same, the residuals of BMI at age 42 would be normally distributed. The plot shows that the residuals of BMI at age 42 deviate from the norm, particularly at the upper tail and are therefore not normally distributed.

To numerically test for normality, we can use the **'swilk'** test. This performs the Shapiro-Wilk test which tests whether the distribution is normal.

Command	**swilk** *bmi_iq2*

Output

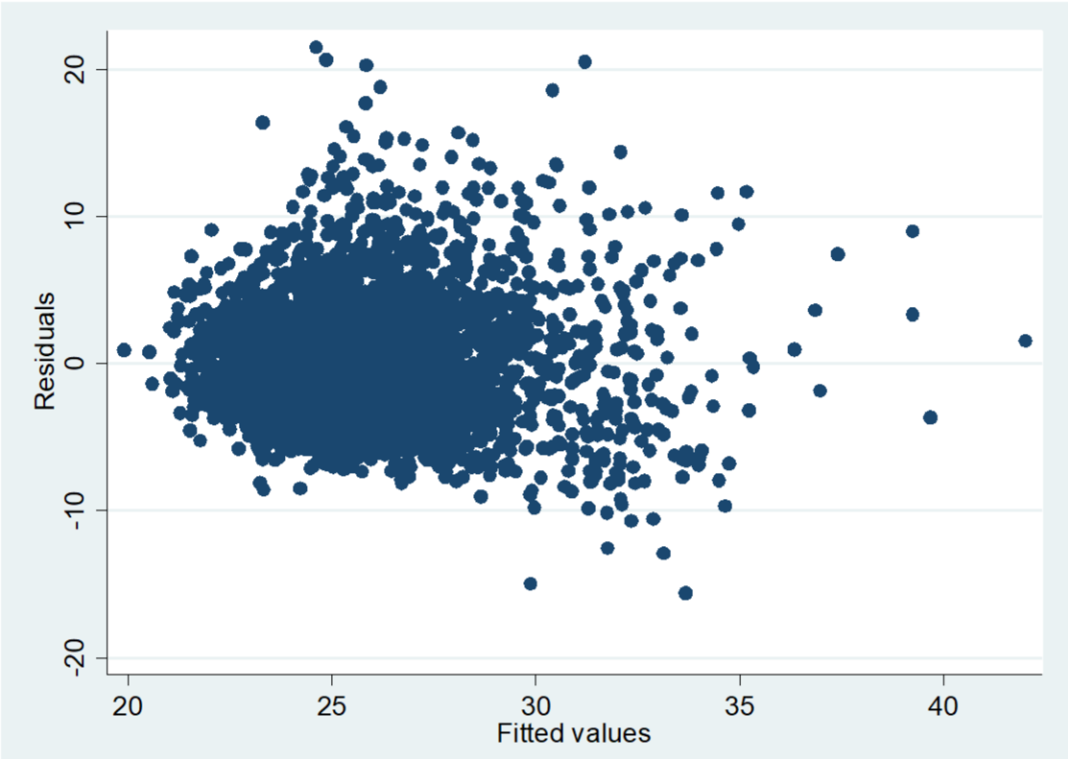
Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
bmi_iq2	4497	0.96110	95.878	11.933	0.00000

In the **‘swilk’** output, we can see that the test’s p-value is <.001 and therefore we can reject the null hypothesis that residuals in model are normally distributed. our general linear regression is not appropriate for valid testing. models categorising aria-describedby="tt" class="glossaryLink" data-cmtooltip="In analysis, the dependent variable is the variable you expect to change in response to different values of your independent (or predictor) variables. For example, a students’ test results may be (partially) explained by the number of hours spent on revision. In this case, the dependent variable is students’ test score, which you expect to be different according to the amount of time spent revising.">outcome variable BMI at age 42, into the top and or bottom tails may better reflect the distribution of the data. For example, the top of the distribution tail represents higher BMI, so transforming our continuous variable into a dichotomous variable (such as ‘obese’ versus ‘not obese’) would capture this feature of the distribution. Likewise, if we were interested in lower BMI, by transforming the bottom tail of the distribution into an ‘underweight’ versus ‘not underweight’ dichotomous variable, we would capture the opposite end of the distribution.

3.6.1 Homogeneity of variance (homoscedasticity of residuals)

A commonly used graphical method for evaluating the model fit is to plot the residuals against the predicted values. If the model is well-fitted, there should be no pattern evident in the plot. We can create such a plot by using the **‘rvfplot’** command.

Command	**rvfplot**
Output	

We can see the pattern of the data points is getting wider towards the right end which is an indication that the model is not well fitted. This implies that our linear regression model would be unable to accurately predict BMI at age 42 consistently across both low and high values of BMI.

3.6.1 Independence

The assumption of independence states that the errors associated with one observation are not correlated with the errors of any other observation. This assumption is often violated if measures of the same variable such as the BMI of an individual are collected over time. Measurements nearer in time are especially likely to be more highly correlated. However, in this example we note BMI of an individual may be very different at age 11 than at age 42, some 31 years later.

Authors: Vanessa Moulton, Dara O'Neill, Alison Park and George B. Ploubidis

4 Logistic regression

This section discusses a method that can be used to analyse the association between a dichotomous (two-category) outcome measure and potentially explanatory variables. This method is a widely used approach and the following guide provides a detailed illustration of how we can use this logistic regression method to answer research questions with longitudinal data.

4.1 What is logistic regression?

Logistic regression is an analysis method that allows us to test the association between an outcome variable that is dichotomous (categorical with two levels) and predictor variables that are either continuous or categorical. We can use logistic regression to predict which of two categories a person is likely to belong to given certain other information. With our longitudinal data, we can use logistic regression to test the probability of an event occurring in later life or not, based on events in early life.

4.2 Example research question: Is lower intelligence in childhood related to obesity in middle age?

In this regression, the outcome variable will be a dichotomous variable, 'not obese' or 'obese' at age 42, as explained below.

All the predictor variables are the same as those used in the general linear and multinomial logistic regression sections. It is always important to explore the data before running statistical models, so if you have not yet done so, please first look at exploring the data. You will also need to construct a few of the explanatory variables before creating your regression model, see main variables of interest.

Suggested citation: Moulton, V., O'Neill, D., Park, A. & Ploubidis, G.B. (2020). *Regression analysis of longitudinal data*. CLOSER Learning Hub, London, UK: CLOSER

4.3 Preparing the outcome variable: Obese or not at age 42

For this regression, we are going to derive an outcome variable, *obese42*, that is dichotomous (comprised of two groups): 'not obese' and 'obese'. We do this derivation using the variable *bmi42*, a continuous variable that we also use in the [general linear regression]() section. The definition of obesity that we are using as the basis of our categorisation is from the World Health Organisation (WHO) standards (http://apps.who.int/bmi/index.jsp?introPage=intro_3.htm). A BMI of 30 and over was defined as obese; a BMI below 30 as not obese. Creating the *obese42* variable requires a series of commands as illustrated below.

Command

```
gen obese42 = .
replace obese42 = 0 if inrange(bmi42,14,29.99999)
replace obese42 = 1 if inrange(bmi42,30,52)
label define obese42L 0 "not obese" 1 "obese", modify
label values obese42 obese42L
```

We can then use the **'tabulate'** command (abbreviated to **'tab'**) to get the frequency of the new variable.

Command

```
tab obese42
```

Output

obese42	Freq.	Percent	Cum.
not obese	3,815	84.83	84.83
obese	682	15.17	100.00
Total	4,497	100.00	

The output shows that, at age 42, approximately 1 in 6 (15.2%) of the sample are obese.

4.4 Running the regression

In the first logistic regression we are going to run, there will only be one predictor variable, 'general ability' at age 11 (*n920)*, which is a continuous variable. We are going to use the **'logit'** command which will provide us with the untransformed beta coefficients (in log-odd units) and their confidence intervals. These are often difficult to interpret, so are sometimes converted into odds ratios. If we wanted to get the odds ratios we could use the command **'logistic'** instead of **'logit'** or add the **'or'** option (**', or'**) to the **'logit'** example below. The odds ratio is the odds of success for one group divided by the odds of success for the other group, where in this example 'success' is the odds of being obese or not obese. When running a logistic regression in Stata, the dependent variable should be specified immediately after the **'logit'** command, followed by the predictor variable(s).

Command

logit *obese42 n920*

Output

```
Logistic regression                             Number of obs     =      4497
                                                LR chi2(1)        =     42.48
                                                Prob > chi2       =    0.0000
Log likelihood = -1892.5587                     Pseudo R2         =    0.0111

------------------------------------------------------------------------------
     obese42 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        n920 |  -.0179825   .0027596    -6.52   0.000    -.0233912   -.0125738
       _cons |  -.9080353   .1280344    -7.09   0.000    -1.158978   -.6570925
------------------------------------------------------------------------------
```

The output above shows that the log likelihood of the fitted model is -1892.56. The number itself does not have much meaning, but when used in comparisons with other models, it can help to identify if the reduced model fits significantly better than the full model (which we will come back to later when we include other predictors in the model). The overall model is statistically significant (chi-square = 42.48, *p*=<.001 which means the model including aria-describedby="tt" class="glossaryLink" data-cmtooltip="General ability is a term used to describe cognitive ability, and is sometimes used as a proxy for intelligent quotient (IQ) scores.">general ability at age 11' fits the data statistically significantly better than the model without it, i.e. a model with no predictors. The 'pseudo R-squared' gives a very general idea of the proportion of variance accounted for by the model; however it is not a reliable statistic, hence its name ‘pseudo’.

In the table, we can see the coefficient, the standard error, the z statistic, associated *p*-values and the 95% confidence intervals of the coefficients. ‘General ability at age 11’ is statistically significant ($Z=-6.52$, $p<.001$ for every unit decrease in aria-describedby="tt" class="glossaryLink" data-cmtooltip="General ability is a term used to describe cognitive ability, and is sometimes used as a proxy for intelligent quotient (IQ) scores.">general ability, the log odds of being obese (compared to not being obese) increases by 0.018.

4.5 Updating the regression model

4.5.1 Including potential confounding variables

In the next model (*M2*), we will add a number of possible confounding variables to the regression: sex, parents’ education and family social class. First we will add sex, where 0=Male and 1=Female. As mentioned previously, this type of binary variable is also known as a dummy variable. In our regression analysis, the reference group is ‘male’. We are also going to include a few family background factors in the model: whether the cohort’s mother (*n016nmed*) and father (*n716dade*) left school at the minimum age or not, and the social class of the study participant’s father (*n1171_2*). Social class *n1171_2* has 5 categories: ‘I/II Prof & Managerial’, ‘III Skilled non-manual’, ‘III Skilled manual’, ‘IV Partly skilled’ and ‘V unskilled’. In Stata we can use the prefix of ‘**i.**’ in the variable name **i.***n1171_2* which will automatically create dummy variable(s). The first category ‘I/II Prof & Managerial’ will be treated as the reference category for that variable.

Command	**logit** *obese42 n920* **i.***sex n016nmed n716dade* **i.***n1171_2*

Output

```
Iteration 0:   log likelihood = -1913.7973
Iteration 1:   log likelihood = -1882.6997
Iteration 2:   log likelihood = -1882.1624
Iteration 3:   log likelihood = -1882.1622
Iteration 4:   log likelihood = -1882.1622

Logistic regression                             Number of obs     =      4,497
                                                LR chi2(8)        =      63.27
                                                Prob > chi2       =     0.0000
Log likelihood = -1882.1622                     Pseudo R2         =     0.0165

------------------------------------------------------------------------------------------
                 obese42 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------+----------------------------------------------------------------
                    n920 |  -.0132103   .0029688    -4.45   0.000     -.019029   -.0073916
                         |
                     sex |
                  female |   .0089026   .0840545     0.11   0.916    -.1558413    .1736464
                n016nmed |  -.1364699   .1094626    -1.25   0.212    -.3510127    .0780729
                n716dade |  -.1500477    .116528    -1.29   0.198    -.3784384     .078343
                         |
                 n1171_2 |
III Skilled non-manual   |  -.0390942   .1715503    -0.23   0.820    -.3753266    .2971382
    III Skilled manual   |   .2543346   .1250572     2.03   0.042     .0092269    .4994422
     IV Partly skilled   |   .2924959   .1480166     1.98   0.048     .0023887    .5826032
          V unskilled    |   .4145009   .2009449     2.06   0.039     .0206562    .8083456
                         |
                   _cons |   -1.24043     .18368    -6.75   0.000    -1.600436   -.8804239
------------------------------------------------------------------------------------------
```

'General ability' is still significant after controlling for the other predictor variables. For every 1 unit decrease in general ability, the log odds of being obese (compared to not being obese) increases by 0.013. In addition, if the participant's father was in the manual or unskilled social classes, by age 42 the participant was more likely to be obese, compared to participants whose fathers were professional or managerial. In this model, the coefficients for sex and mother's and father's education were not significant, that is to say, we have not found that the log odds of being obese or not obese at age 42 differ between men and women, or according to parental educational level.

4.5.2 Including a childhood measure of BMI

For our final model (*M3*), we will also add *bmi11*, the BMI of the participants when they were aged 11. Doing so means that we will be adjusting for participant's baseline BMI, and that will allow us to focus on the subsequent change in BMI from age 11 to age 42, and therefore to measure both BMI and general ability over a comparable period, from childhood to middle age.

Command	**logit** *obese42 n920* **i.***sex n016nmed n716dade* **i.***n1171_2 bmi11*

Output

```
Logistic regression                               Number of obs   =      4497
                                                  LR chi2(9)      =    589.41
                                                  Prob > chi2     =    0.0000
Log likelihood =  -1619.092                       Pseudo R2       =    0.1540

------------------------------------------------------------------------------------------
                  obese42 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------------+---------------------------------------------------------------
                     n920 |  -.0151402    .003208    -4.72   0.000    -.0214277   -.0088526
                          |
                      sex |
                   female |  -.1851705   .0919714    -2.01   0.044    -.3654311     -.00491
                 n016nmed |  -.0254165   .1182051    -0.22   0.830    -.2570942    .2062611
                 n716dade |  -.0761896    .125169    -0.61   0.543    -.3215162    .1691371
                          |
                  n1171_2 |
  III Skilled non-manual  |   .0961269    .183298     0.52   0.600    -.2631305    .4553843
      III Skilled manual  |   .3367054   .1351669     2.49   0.013      .071783    .6016277
       IV Partly skilled  |    .392927   .1602766     2.45   0.014     .0787906    .7070635
             V unskilled  |   .5620454   .2179275     2.58   0.010     .1349153    .9891755
                          |
                    bmi11 |   .3529736   .0168074    21.00   0.000     .3200318    .3859155
                    _cons |  -7.578431   .3636064   -20.84   0.000    -8.291087   -6.865776
------------------------------------------------------------------------------------------
```

The results above show that for a 1 unit increase in BMI at age 11, the log odds of being obese at age 42 increases by 0.353. After controlling for BMI at age 11 and all the other predictors, being female compared to male decreases the log odds of obesity by 0.185. In addition, having a father in the lower social classes compared to one with a professional/managerial occupation increases the odds of obesity at age 42.

4.6 Exploring predictors' influence and predicted probabilities on the outcome

4.6.1 Testing the influence of a specific categorical variable

We can examine the overall effect of social class using the '**test**' command. To specify which levels of the categorical *n1171_2* social class variable we wish to compare to the reference category ('I/II Prof & Managerial'), we include a prefix denoting the numeric code for each other category (e.g. 'III Skilled non-manual' is the second category so this is denoted as **2.***n1171_2*).

Command

test 2.*n1171_2* 3.*n1171_2* 4.*n1171_2* 5.*n1171_2*

Output

```
( 1)  [obese42]2.n1171_2 = 0
( 2)  [obese42]3.n1171_2 = 0
( 3)  [obese42]4.n1171_2 = 0
( 4)  [obese42]5.n1171_2 = 0

         chi2(  4) =   10.32
       Prob > chi2 =    0.0354
```

From the output of the ‘**test**’ command above, we can see that the overall effect of social class is statistically significant (p<.05>

We can also examine the differences in the coefficients for each of the different social classes compared to the reference category. For instance, we could again use the ‘**test**’ command, as shown in the example below, to evaluate whether the coefficient for social class ‘III Skilled non-manual’ is equivalent to the coefficient for social class ‘III Skilled manual’.

Command

test 2.*n1171_2* 3.*n1171_2*

Output

```
( 1)  [obese42]2.n1171_2 = 0
( 2)  [obese42]3.n1171_2 = 0

         chi2(  2) =    6.72
       Prob > chi2 =    0.0347
```

The output above shows that the p-value is under <.05 (our threshold for inferring statistical significance) and we can consequently say the coefficients for these two categories are different.

4.6.2 Testing predicted probabilities of our explanatory variable of interest on our outcome variable

Focusing on our predictor of interest ‘general ability’, we can use predicted probabilities to help understand the relationship between general ability and obesity in the model. In this example we want to calculate the predicted probability of obesity for a given score on the general ability test. Predicted probabilities can be calculated using the **'margins'** command. We can use this command to create the predicted probabilities for values of the general ability test (*n920* which ranges from 0 to 79) from 10 to 80 in increments of 10. The **'margins'** command uses the sample values of other predictor variables to calculate the average predicted probabilities on our predictor of interest. We can also use the **'vsquish'** option in the command to help tidy up the output as this removes blank lines in output tables.

Command

margins, at(*n920*=(10(10)80)) vsquish

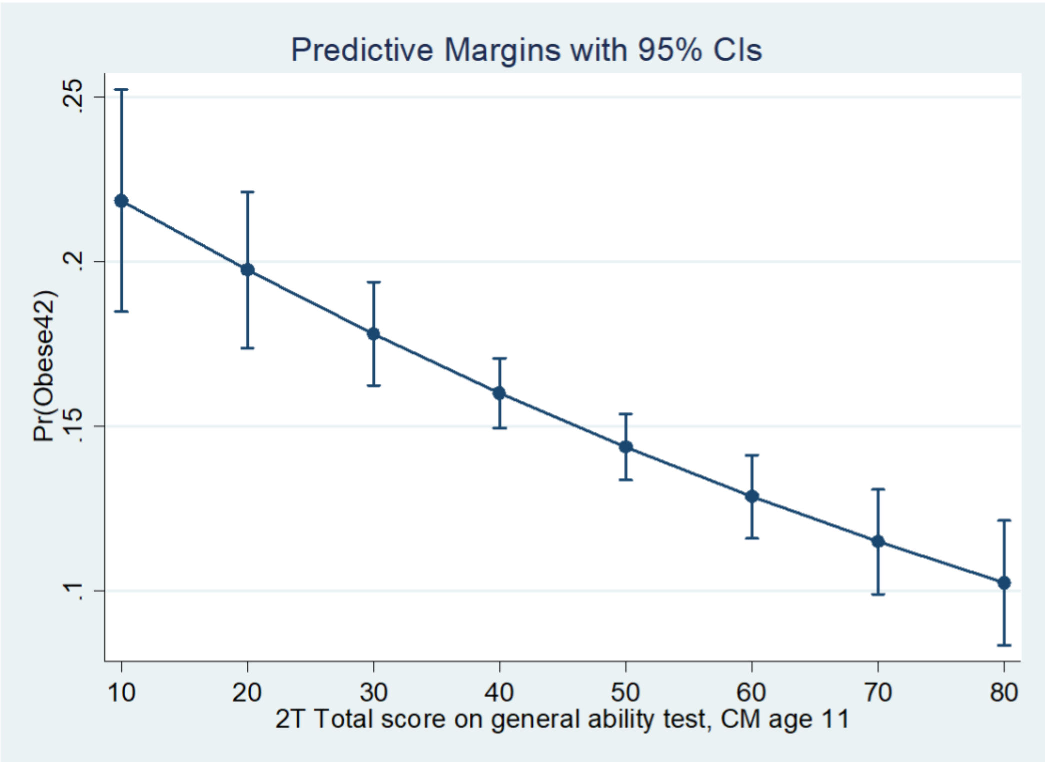
Output

```
Predictive margins                              Number of obs   =       4497
Model VCE    : OIM

Expression   : Pr(obese42), predict()
1._at        : n920            =          10
2._at        : n920            =          20
3._at        : n920            =          30
4._at        : n920            =          40
5._at        : n920            =          50
6._at        : n920            =          60
7._at        : n920            =          70
8._at        : n920            =          80

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .2184704   .0171803    12.72   0.000     .1847976    .2521432
          2  |    .197483   .0120891    16.34   0.000     .1737889    .2211771
          3  |   .1780491   .0079934    22.27   0.000     .1623823    .1937159
          4  |   .1601343    .005402    29.64   0.000     .1495466    .1707221
          5  |   .1436885   .0050912    28.22   0.000     .1337099     .153667
          6  |   .1286489   .0064315    20.00   0.000     .1160433    .1412545
          7  |   .1149441   .0081215    14.15   0.000     .0990263     .130862
          8  |   .1024967   .0096418    10.63   0.000     .0835991    .1213942
------------------------------------------------------------------------------
```

The first part of the output above tells us which row is associated with which general ability test score. Row 1 corresponds to a test score of 10, while row 8 is equal to a test score of 80. We can interpret from the table that as the test score at age 11 increases, the probability of obesity at age 42 is decreasing from 21.8% to 10.2%.

4.6.3 Plotting the predicted probabilities

We can present the results as a graph by using the **'marginsplot'** command, which plots both the predicted probabilities and their confidence intervals.

Command	**marginsplot**
Output	

In the output plot above, the 'predicted probability of obesity at age 42' is on the Y axis and the 'general ability test score at age 11' is on the X axis. The fitted line decreases from left to right, indicating that as general ability scores increase, the probability of obesity decreases. The predicted probability of obesity at age 42 would be 17.8% with a test score of 30 at age 11, compared to 12.9% with a test score of 60.

4.7 Comparing model fit of the logistic regression models

As we mentioned earlier, the log likelihood of the fitted model is used to compare to other models, to identify if the reduced model fits significantly better than the full model. In order to compare models, in Stata we can use the **'estimates store'** and **'lrtest'** commands. We will re-run the same models we have just completed in the previous logistic regression examples. Each model is estimated and stored using the command **'est store'** under an arbitrary name; in this example we are labelling them *M0* to *M3*. You can use the **'quietly'** command in front of the **'logistic'** command to run the models in the background (i.e. Stata stores the output rather than writing it out at the time the command is run). It is possible to include code comments or annotations (text that explains the code you are running) in the Stata command window by starting the comment line with an asterisk (**'*'**).

Command

```
*Model 0: Intercept only
quietly logit obese42
est store M0
*Model 1: 'general ability' added
quietly logit obese42 n920
est store M1
*Model 2: general ability', sex and family background
quietly logit obese42 n920 i.sex n016nmed n716dade i.n1171_2
est store M2
*Model 3: general ability', sex, family background and BMI at age 11
quietly logit obese42 n920 i.sex n016nmed n716dade i.n1171_2 bmi11
est store M3
```

We will then use the **'lrtest'** command to test whether the log likelihoods for each model are significantly different to each other.

Command

```
*Model 1 versus Model 0
lrtest M1 M0
*Model 2 versus Model 1
lrtest M2 M1
*Model 3 versus Model 2
lrtest M3 M2
*Model 3 versus Model 0
lrtest M3 M0
```

Output

```
. *Model 1 versus Model 0
. lrtest M1 M0

Likelihood-ratio test                                 LR chi2(1)  =     42.48
(Assumption: M0 nested in M1)                         Prob > chi2 =    0.0000

. *Model 2 versus Model 1
. lrtest M2 M1

Likelihood-ratio test                                 LR chi2(7)  =     20.79
(Assumption: M1 nested in M2)                         Prob > chi2 =    0.0041

. *Model 3 versus Model 2
. lrtest M3 M2

Likelihood-ratio test                                 LR chi2(1)  =    526.14
(Assumption: M2 nested in M3)                         Prob > chi2 =    0.0000

. *Model 3 versus Model 0
. lrtest M3 M0

Likelihood-ratio test                                 LR chi2(9)  =    589.41
(Assumption: M0 nested in M3)                         Prob > chi2 =    0.0000
```

In the output above, the log-likelihood test for *M1* v *M0* is the same result as the first model we ran in this set of **‘logit’** examples. This is because we are comparing the empty model (*M0*) with *M1* which has only one predictor variable: general ability (chi-square = 42.48, p=<.001 in the second comparison above>M2 v *M1*), we can see that the addition of sex and family background variables to the model marginally improves the fit (chi-square = 20.79, p=<.01 while adding a single predictor at age in m3 makes notable further improvement to the model fit p="<.001)." final test>M0 v *M3* compares the original model with no explanatory variables and our final model; unsurprisingly given the other results, this again shows that adding all the predictors improves the fit over the empty model (chi-square = 589.41, p=<.001>

4.8 Regression diagnostics

When modelling a binary outcome variable, unlike in linear regression there are no typically agreed statistical tests that can be used in the diagnostic process. However, you can find out more from the following sources:

- Menard, S. (2010). Logistic regression: From introductory to advanced concepts and applications. Thousand Oaks, CA: SAGE.
- Hilbe, J.M. (2009). Logistic regression models. Boca Raton, FL: Chapman & Hall/CRC.
- Hosmer, D.W. & Lemeshow, S. (2000). Applied logistic regression (2nd edition). New York, NY: Wiley.

5 Multinomial logistic regression

This section provides guidance on a method that can be used to explore the association between a multiple-category outcome measure and other potentially explanatory variables. Multinomial logistic regression can offer us useful insights when we are working with longitudinal data and this section breaks down and discusses each of the key steps involved.

5.1 What is multinomial logistic regression?

Multinomial regression is an extension of logistic regression that is used when a categorical outcome variable has more than two values and predictor variables are continuous or categorical. We can use multinomial regression to predict which of two or more categories a person is likely to belong to, compared to a baseline (or reference) category and given certain other information. With our longitudinal data we can use multinomial logistic regression to test the probability of an event occurring (A) in later life compared to other potential outcomes (B, C), applying information gathered in earlier life. In order to make comparisons, we can use any of the events (A, B or C) as the baseline category.

5.2 Example research question: Is childhood intelligence related to normal/healthy body-mass index (BMI) compared to being overweight or obese in middle age?

In this regression, we will again explore the links between childhood intelligence and body mass index (BMI) at age 42, but this time we will categorise participants' BMI score into three groups: 'normal/healthy', 'overweight' and 'obese'. We are going to treat this variable as being nominal and so we will use a method called multinomial logistic regression that is appropriate for use with outcome variables with multiple categories.

In the next section, we will show you how to create the variable for use in the analysis.

5.3 Preparing the outcome variable: BMI categories

We will group the categories together based on the World Health Organisation (WHO) standards (http://apps.who.int/bmi/index.jsp?introPage=intro_3.htm). Few of the sample were underweight (n=54, <1%) so in this example they will be included in the normal or healthy category.

Command

```
gen BMI42_C = .
replace BMI42_C = 1 if inrange(bmi42,14,24.99999)
replace BMI42_C = 2 if inrange(bmi42,25,29.99999)
replace BMI42_C = 3 if inrange(bmi42,30,52)
label define BMI42_CL 1 "normal/healthy" 2 "overweight" 3 "obese", modify
label values BMI42_C BMI42_CL
```

Once we have created the variable, we can use the **'tab'** command to look at the number of participants that fall into each BMI category .

Command

```
tab BMI42_C
```

Output

BMI42_C	Freq.	Percent	Cum.
normal/healthy	2,151	47.83	47.83
overweight	1,664	37.00	84.83
obese	682	15.17	100.00
Total	4,497	100.00	

Just under half (48%) of our sample were normal or healthy weight, over a third (37%) were overweight and 15% were obese.

All of the predictor variables are the same as those used in the general linear and logistic regression sections. It is always important to explore the data before running statistical models. To examine the data, please look at exploring the data. If you have not done so already you will also need to construct a few of the explanatory variables before creating your regression model, see main variables of interest.

5.4 Running the regression

In Stata, we use the **'mlogit'** command to run a multinomial logistic regression. As with the logistic regression method, the command produces untransformed beta coefficients (in log-odd units) along with their confidence intervals. (These are often difficult to interpret, so are sometimes converted into relative risk ratios. If we wanted to get the relative risk ratios we could add the **'rrr'** option (**', rrr'**) to the **'mlogit'** example below). With the **'mlogit'** command, we also include the option **'base'** to specify which category is the reference group. For our analysis, we will use 'normal or healthy' weight as the reference category.

In the first regression we run, there will only be one predictor variable, 'general ability at age 11' (*n920*), which is a continuous variable.

Command

mlogit *BMI42_C n920*, **base**(1)

Output

```
Iteration 0:   log likelihood = -4526.9851
Iteration 1:   log likelihood = -4499.3001
Iteration 2:   log likelihood = -4499.1205
Iteration 3:   log likelihood = -4499.1205

Multinomial logistic regression                 Number of obs     =       4497
                                                LR chi2(2)        =      55.73
                                                Prob > chi2       =     0.0000
Log likelihood = -4499.1205                     Pseudo R2         =     0.0062

------------------------------------------------------------------------------
     BMI42_C |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
normal_healthy|  (base outcome)
-------------+----------------------------------------------------------------
overweight   |
        n920 |  -.0080337   .0022092    -3.64   0.000    -.0123637   -.0037037
       _cons |   .1221181   .1089746     1.12   0.262    -.0914682    .3357044
-------------+----------------------------------------------------------------
obese        |
        n920 |  -.0215579   .0029388    -7.34   0.000    -.0273178    -.015798
       _cons |  -.1647579   .1379386    -1.19   0.232    -.4351126    .1055968
------------------------------------------------------------------------------
```

The iterations 0 through 3 listed in the top left-hand corner of the output above are the log likelihoods at each iteration of the maximum likelihood estimation. Iteration 0 is the log likelihood of the model with no predictors. When the difference between successive iterations is very small, the model has ‘converged’. The final iteration is the log likelihood of the fitted model. The log likelihood of the fitted model is -4499.12. The number itself does not have much meaning, but is used to make comparisons across the models and to identify if the reduced model fits significantly better than the full model. The overall model is statistically significant (chi-square = 55.73, p=<.001 which means the model including aria-describedby="tt" class="glossaryLink" data-cmtooltip="General ability is a term used to describe cognitive ability, and is sometimes used as a proxy for intelligent quotient (IQ) scores.">general ability at age 11’ fits the data statistically significantly better than the model without it, i.e. a model with no predictors. The ‘pseudo R-squared’ value (*Pseudo R2*) gives a very general idea of the proportion of variance accounted for by the model, but it is just an approximation and not very reliable which is why we call it ‘pseudo’.

In the output above, we also get a tabulation of the coefficient, standard error, the z statistic, associated p-values and the 95% confidence intervals of the coefficients. This table is in two

Suggested citation: Moulton, V., O’Neill, D., Park, A. & Ploubidis, G.B. (2020). *Regression analysis of longitudinal data*. CLOSER Learning Hub, London, UK: CLOSER

parts, labelled with the categories of the outcome variable *BMI42_C***.** In both outputs, 'general ability at age 11' (*n920*) is statistically significant. A 1 unit decrease in 'general ability' is associated with a 0.008 decrease in the relative log odds of being overweight compared to a normal/healthy weight, and a 0.022 decrease in the relative log odds of being obese compared to a normal/healthy weight.

In the next step, we will extend the model further to explore the influence of other variables on this association between general ability and the different categories of BMI.

5.5 Updating the regression model

5.5.1 Including potential confounding variables

In the next model, we will add a set of possible confounding variables to the regression: sex, parents' education and family social class. First, we will add *sex* where 0=Male and 1=Female. As explained in previous sections, this type of binary variable is also known as a dummy variable. In our analysis, the reference group will be 'male' (as this group is coded as 0). We are also going to include a few family background factors in the model: whether the cohort's mother (*n016nmed*) and father (*n716dade*) left school at the minimum age or not, and the social class of the study participant's father (*n1171_2*). Social class *n1171_2* has 5 categories: 'I/II Prof & Managerial', 'III Skilled non-manual', 'III Skilled manual', 'IV Partly skilled' and 'V unskilled'. With multi-category variables such as this, you can use the prefix of **'i.'** in the variable name **i.***n1171_2* and Stata will automatically create dummy variable(s) for each category. The first category 'I/II Prof & Managerial' will be treated as the reference category.

Command	**mlogit** *BMI42_C n920* **i.***sex n016nmed n716dade* **i.***n1171_2*, **base**(1)

Output

```
Iteration 0:   log likelihood = -4526.9851
Iteration 1:   log likelihood = -4374.7751
Iteration 2:   log likelihood = -4374.4954
Iteration 3:   log likelihood = -4374.4954

Multinomial logistic regression                 Number of obs     =      4,497
                                                LR chi2(16)       =     304.98
                                                Prob > chi2       =     0.0000
Log likelihood = -4374.4954                     Pseudo R2         =     0.0337
```

BMI42_C	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
normal_healthy	(base outcome)					
overweight						
n920	-.0024719	.0024314	-1.02	0.309	-.0072374	.0022937
sex						
female	-.9530439	.0676005	-14.10	0.000	-1.085539	-.8205493
n016nmed	-.3151754	.0827246	-3.81	0.000	-.4773127	-.1530382
n716dade	-.0607015	.0868848	-0.70	0.485	-.2309926	.1095896
n1171_2						
III Skilled non-manual	.1258947	.1204799	1.04	0.296	-.1102415	.3620309
III Skilled manual	.1509468	.0944926	1.60	0.110	-.0342552	.3361488
IV Partly skilled	.0949538	.1177955	0.81	0.420	-.1359212	.3258287
V unskilled	.2071978	.1748493	1.19	0.236	-.1355005	.549896
_cons	.3509449	.148248	2.37	0.018	.0603842	.6415056
obese						
n920	-.0142966	.0031694	-4.51	0.000	-.0205085	-.0080846
sex						
female	-.4200618	.0897764	-4.68	0.000	-.5960203	-.2441034
n016nmed	-.2687072	.1145065	-2.35	0.019	-.4931358	-.0442786
n716dade	-.1750389	.1221591	-1.43	0.152	-.4144663	.0643886
n1171_2						
III Skilled non-manual	.0116868	.1792527	0.07	0.948	-.339642	.3630156
III Skilled manual	.3188583	.1313633	2.43	0.015	.061391	.5763256
IV Partly skilled	.3317233	.156482	2.12	0.034	.0250242	.6384225
V unskilled	.5060802	.2172254	2.33	0.020	.0803262	.9318342
_cons	-.3634021	.1962144	-1.85	0.064	-.7479753	.021171

Interestingly in the output we can see that 'general ability' is significant in the 'obese' versus 'normal/healthy' BMI comparison, but not in the 'overweight' versus 'normal/healthy' BMI comparison after controlling for all the other predictors. A 1 unit decrease in 'general ability' test score is associated with a .014 increase in the relative log odds of being obese v normal/healthy BMI at age 42. Father's social class also predicts obesity; it is associated with the odds of the study participant being overweight compared to normal/healthy BMI in the study participant. Males (compared to females) and participants whose mothers left education at the minimum age were more likely to be overweight or obese compared to normal/healthy BMI.

5.5.2 Including a childhood measure of BMI

For our final model, we are going to include *bmi11*, the BMI of the participant when they were aged 11. Doing so means that we will be adjusting for participant's baseline BMI, and that will allow us to focus on the subsequent change in BMI from age 11 to age 42, and therefore to measure both BMI and general ability over a comparable period, from childhood to middle age.

Command	**mlogit** *BMI42_C n920* **i.***sex n016nmed n716dade* **i.***n1171_2 bmi11*, **base**(1)

Output

```
Iteration 0:   log likelihood = -4526.9851
Iteration 1:   log likelihood = -4033.6693
Iteration 2:   log likelihood = -3991.7248
Iteration 3:   log likelihood = -3991.1008
Iteration 4:   log likelihood = -3991.1006

Multinomial logistic regression                 Number of obs     =      4497
                                                LR chi2(18)       =   1071.77
                                                Prob > chi2       =    0.0000
Log likelihood = -3991.1006                     Pseudo R2         =    0.1184

------------------------------------------------------------------------------------
             BMI42_C |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------+--------------------------------------------------------------
normal_healthy       |  (base outcome)
---------------------+--------------------------------------------------------------
overweight           |
                n920 |  -.0039818   .0025077    -1.59   0.112    -.0088968    .0009333
                     |
                 sex |
              female |  -1.070566   .0702184   -15.25   0.000    -1.208192    -.932941
            n016nmed |  -.2835888   .0850214    -3.34   0.001    -.4502277    -.11695
            n716dade |  -.0321597   .0893179    -0.36   0.719    -.2072196    .1429003
                     |
             n1171_2 |
III Skilled non-manual |   .1880709   .1239453     1.52   0.129    -.0548574    .4309992
  III Skilled manual |   .2095402    .097347     2.15   0.031     .0187437    .4003368
   IV Partly skilled |   .1312279   .1210964     1.08   0.279    -.1061166    .3685724
        V unskilled  |   .2621817   .1791823     1.46   0.143    -.0890091    .6133726
                     |
               bmi11 |   .2591989   .0175762    14.75   0.000     .2247503    .2936476
               _cons |  -4.004781   .3305199   -12.12   0.000    -4.652588   -3.356974
---------------------+--------------------------------------------------------------
obese                |
                n920 |  -.0172042   .0034883    -4.93   0.000     -.024041   -.0103673
                     |
                 sex |
              female |   -.763342   .1003417    -7.61   0.000    -.9600081   -.5666759
            n016nmed |  -.1711971   .1259803    -1.36   0.174    -.4181139    .0757197
            n716dade |  -.0921214   .1335095    -0.69   0.490    -.3537953    .1695525
                     |
             n1171_2 |
III Skilled non-manual |   .1986026    .194555     1.02   0.307    -.1827181    .5799234
  III Skilled manual |   .4516593   .1446866     3.12   0.002     .1680788    .7352399
   IV Partly skilled |   .4600025   .1726833     2.66   0.008     .1215494    .7984555
        V unskilled  |   .6980658    .239024     2.92   0.003     .2295875    1.166544
                     |
               bmi11 |   .5077231   .0212067    23.94   0.000     .4661587    .5492876
               _cons |   -9.25069    .427725   -21.63   0.000    -10.08902   -8.412365
------------------------------------------------------------------------------------
```

In the output above, we can see that after controlling for BMI at age 11 'general ability' is significant in the comparison of obese versus normal/healthy BMI, but not in the overweight versus normal/healthy BMI comparison. A 1 unit decrease in 'general ability' test score is associated with a .017 increase in the relative log odds of being obese versus normal/healthy BMI at age 42. Lower parental social class, compared to professional and managerial is also important. In addition, as in the previous model, males are more likely than females to be either overweight or obese than to have a normal/healthy BMI.

5.6 Exploring predictors' influence and predicted probabilities on the outcome

5.6.1 Testing the influence of a categorical variable

The above results suggest that there are differences in the association of family background (education and social class) with obesity and being overweight compared to normal/healthy BMI. We can test these formally, by examining the overall effect of mother's education using the **'test'** command.

Command	**test [**overweight**]***n016nmed* **= [**obese**]***n016nmed*
Output	(1) [overweight]n016nmed - [obese]n016nmed = 0 chi2(1) = 0.80 Prob > chi2 = 0.3711

We can see that there is no significant difference between the association of when the participant's mother left education and the participant's own BMI in later life.

We can also test the overall influence of fathers social class using the **'test'** command.

Command	**test** 2.*n1171_2* 3.*n1171_2* 4.*n1171_2* 5.*n1171_2*
Output	(1) [normal_healthy]2o.n1171_2 = 0 (2) [overweight]2.n1171_2 = 0 (3) [obese]2.n1171_2 = 0 (4) [normal_healthy]3o.n1171_2 = 0 (5) [overweight]3.n1171_2 = 0 (6) [obese]3.n1171_2 = 0 (7) [normal_healthy]4o.n1171_2 = 0 (8) [overweight]4.n1171_2 = 0 (9) [obese]4.n1171_2 = 0 (10) [normal_healthy]5o.n1171_2 = 0 (11) [overweight]5.n1171_2 = 0 (12) [obese]5.n1171_2 = 0 Constraint 1 dropped Constraint 4 dropped Constraint 7 dropped Constraint 10 dropped chi2(8) = 15.79 Prob > chi2 = 0.0455

Here we see the overall influence of father's social class on BMI category is statistically significant (chi-square = 15.79, $p<0.05$). (NB the commands 1,4,7 and 10 are constrained as they are the baseline reference category, i.e. normal/healthy weight).

5.6.2 Testing predicted probabilities of our explanatory variable of interest on our outcome variable

Focusing on our predictor of interest 'general ability', we can use predicted probabilities to help understand the relationship between 'general ability' and obesity, overweight and normal/healthy BMI in the model. In this example we want to calculate the predicted probability of the three BMI categories for a given score on the 'general ability' test. Predicted probabilities can be calculated using the **'margins'** command. We create the predicted probabilities for values of the 'general ability' test (*n920* which ranges from 0 to 79) from 10 to 80 in increments of 10. The values in the table are the average predicted probabilities calculated using the sample values of other predictor variables. The example below shows the predicted probability for healthy BMI given the 'general ability' test score.

Command	**margins, at(*n920*=(10(10)80)) predict(outcome(1)) vsquish**

Output

```
Predictive margins                              Number of obs     =      4497
Model VCE    : OIM

Expression   : Pr(BMI42_C==normal_healthy), predict(outcome(1))
1._at        : n920            =          10
2._at        : n920            =          20
3._at        : n920            =          30
4._at        : n920            =          40
5._at        : n920            =          50
6._at        : n920            =          60
7._at        : n920            =          70
8._at        : n920            =          80
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	.4205484	.0192639	21.83	0.000	.3827919	.4583049
2	.4371663	.0148733	29.39	0.000	.4080151	.4663175
3	.4532692	.0107853	42.03	0.000	.4321304	.474408
4	.4688456	.007659	61.22	0.000	.4538343	.4838569
5	.4838923	.0070662	68.48	0.000	.4700429	.4977418
6	.4984135	.0095565	52.15	0.000	.4796832	.5171439
7	.5124191	.0135473	37.82	0.000	.485867	.5389713
8	.5259237	.0180749	29.10	0.000	.4904976	.5613498

The first part of the output tells us which row is associated with which 'general ability' test score. Row 1 (*Expression = 1._at*) relates to a test score of 10, while row 8 equal to a test score of 80. As the test score at age 11 increases, the probability of a healthy BMI at age 42 being a 1 is increasing from a probability of 0.421 to 0.526.

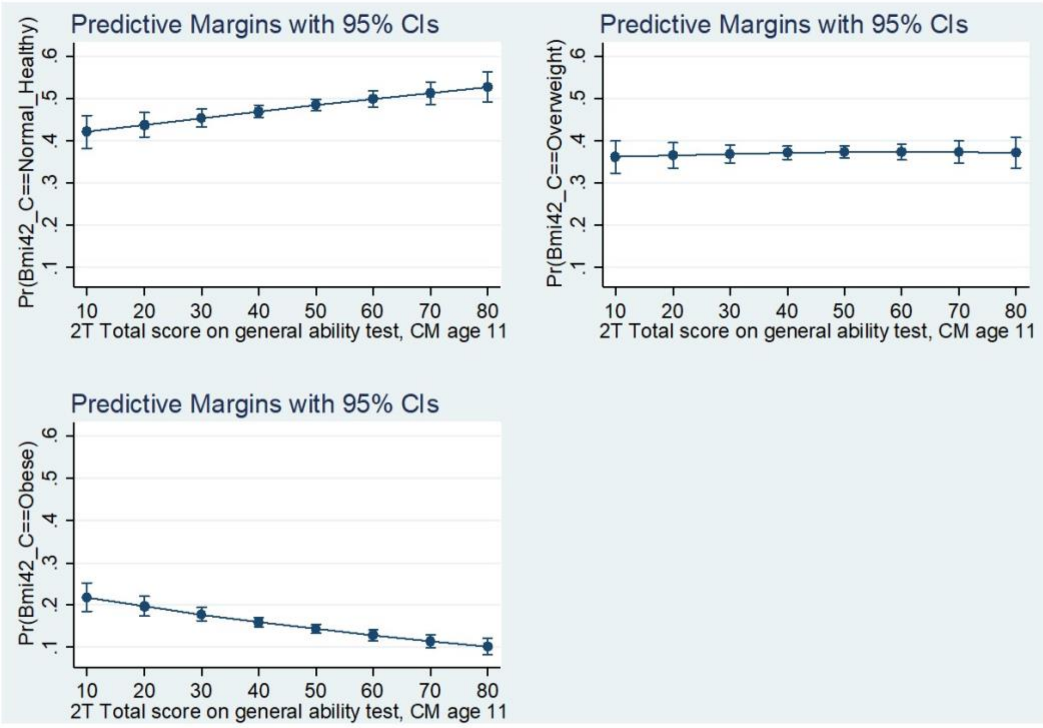
5.6.3 Plotting the predicted probabilities

We can use the **'marginsplot'** command to create a graph of the predicted probabilities and their confidence intervals for each of the BMI categories. We can also combine those graphs using the command **'graph combine'**. This last command has the option **'ycommon'** which we will use to ensure the combined graphs have the same y axis.

Command

```
margins, at(n920=(10(10)80)) predict(outcome(1)) vsquish
marginsplot, name(healthy)
margins, at(n920=(10(10)80)) predict(outcome(2)) vsquish
marginsplot, name(overweight)
margins, at(n920=(10(10)80)) predict(outcome(3)) vsquish
marginsplot, name(obese)
graph combine healthy overweight obese, ycommon
```

Output



The predicted probability of a normal weight (top left graph), overweight (top right graph) or obesity (bottom left graph) at age 42 is on the Y axis and the ‘general ability’ test score at age 11 is on the X axis. The fitted line increases from left to right, is flat and decreases from left to right for normal weight, overweight and obesity respectively as general ability scores increase.

5.7 Regression diagnostics

When modelling a categorical outcome variable, unlike in linear regression there are no typically agreed statistical tests that can be used in the diagnostic process. However, you can find out more from the following sources:

- Menard, S. (2010). *Logistic regression: From introductory to advanced concepts and applications.* Thousand Oaks, CA: SAGE.
- Hilbe, J.M. (2009). *Logistic regression models*. Boca Raton, FL: Chapman & Hall/CRC.
- Hosmer, D.W. & Lemeshow, S. (2000). *Applied logistic regression* (2nd edition). New York, NY: Wiley.

If the purpose of the analysis is to investigate repeated measures over time for example BMI at a number of different time points, the analysis should account for the clustered nature of the data, i.e. allow that measurements within individuals be correlated. Therefore, general linear, logistic and multinomial regression models may not be the most appropriate methods when analysing this type of longitudinal data. We will be adding new sections soon that will illustrate a number of methods that can be applied when analysing repeated measures data.

Suggested citation: Moulton, V., O'Neill, D., Park, A. & Ploubidis, G.B. (2020). *Regression analysis of longitudinal data*. CLOSER Learning Hub, London, UK: CLOSER