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

# 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.

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

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

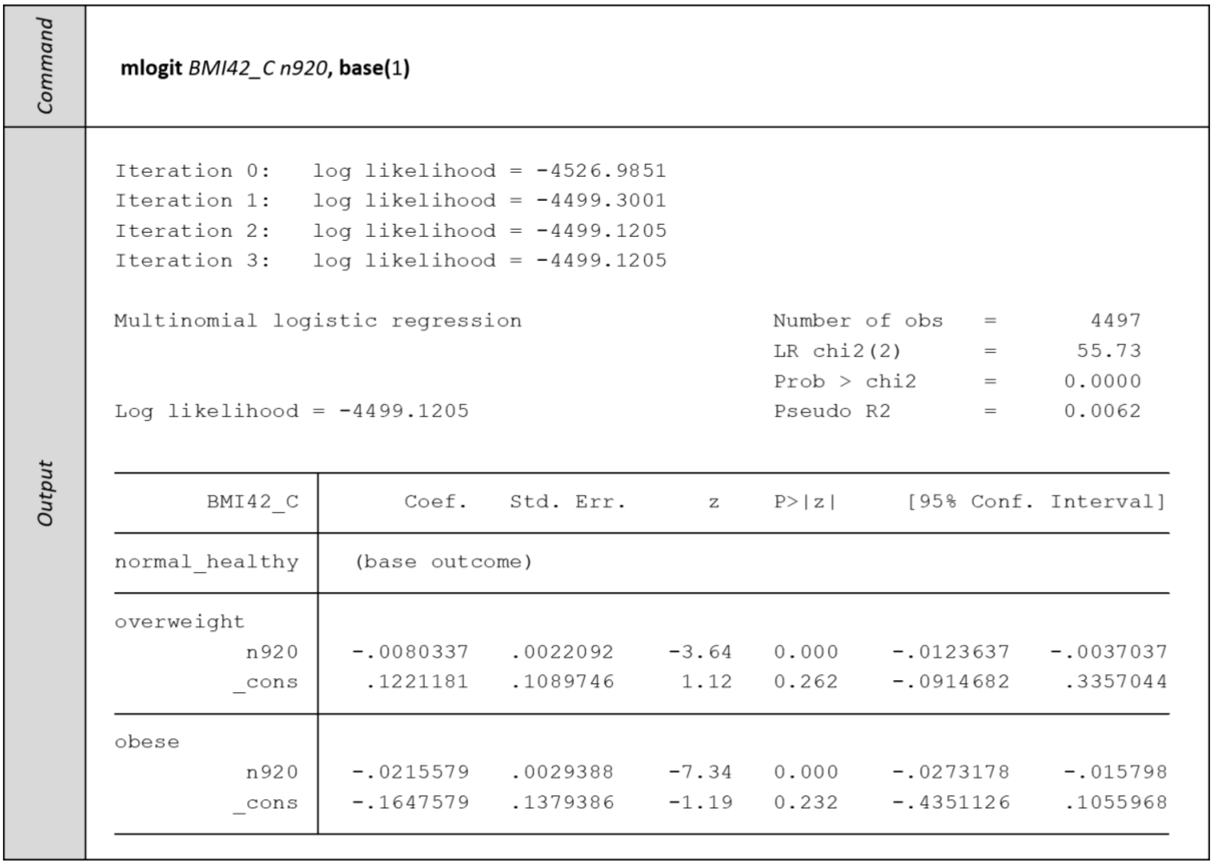
**Suggested citation:** Moulton, V., O’Neill, D., Park, A. & Ploubidis, G.B. (2020). *Regression analysis of longitudinal data*. CLOSER Learning Hub, London, UK: CLOSER

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’.

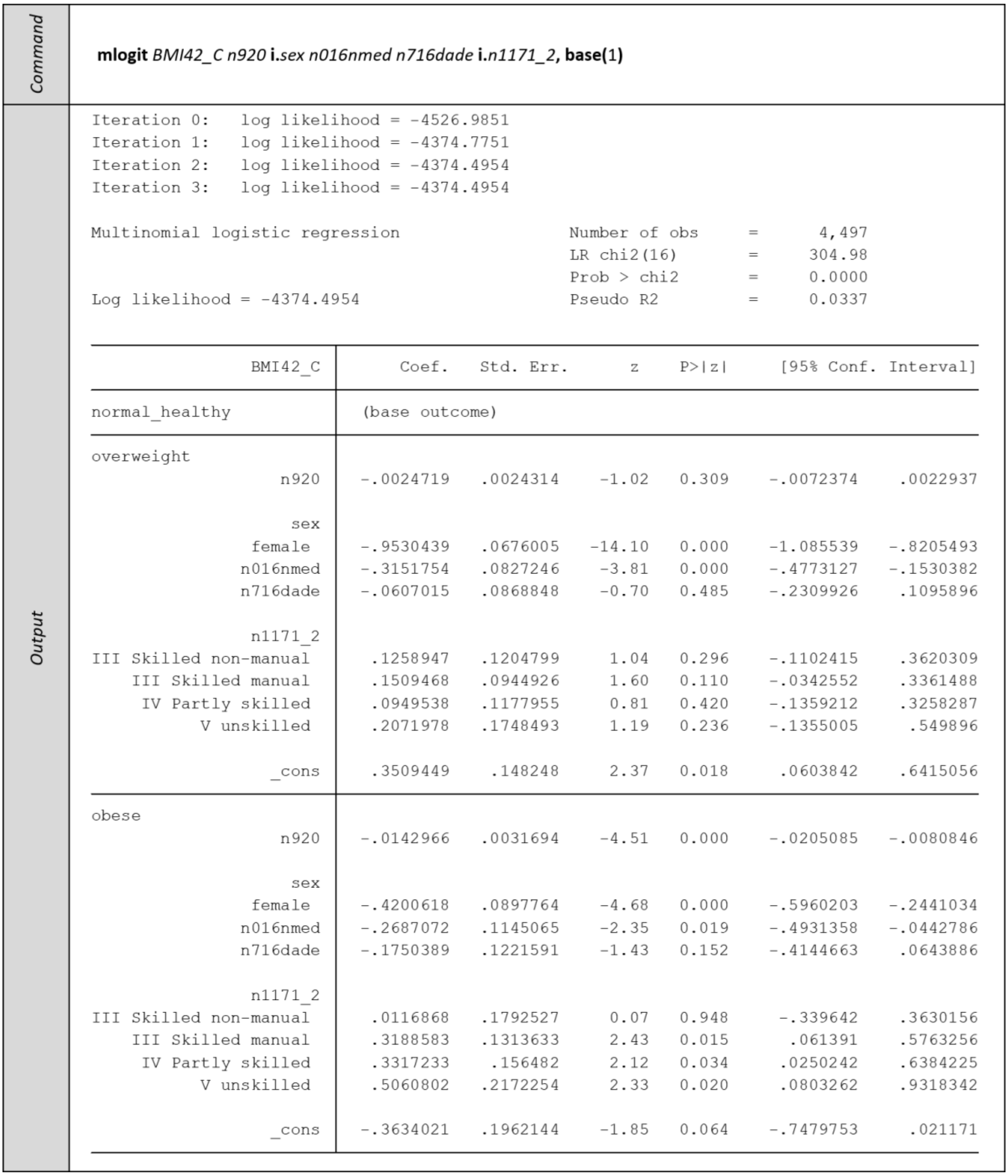
In the output above, we also get a tabulation of the coefficient, standard error, the z statistic, associated p-values and the 95% confidence intervals of the coefficients. This table is in two parts, labelled with the categories of the outcome variable *BMI42_C***.** In both outputs, 'general ability 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.



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

```
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
------------------------------------------------------------------------------------------
```

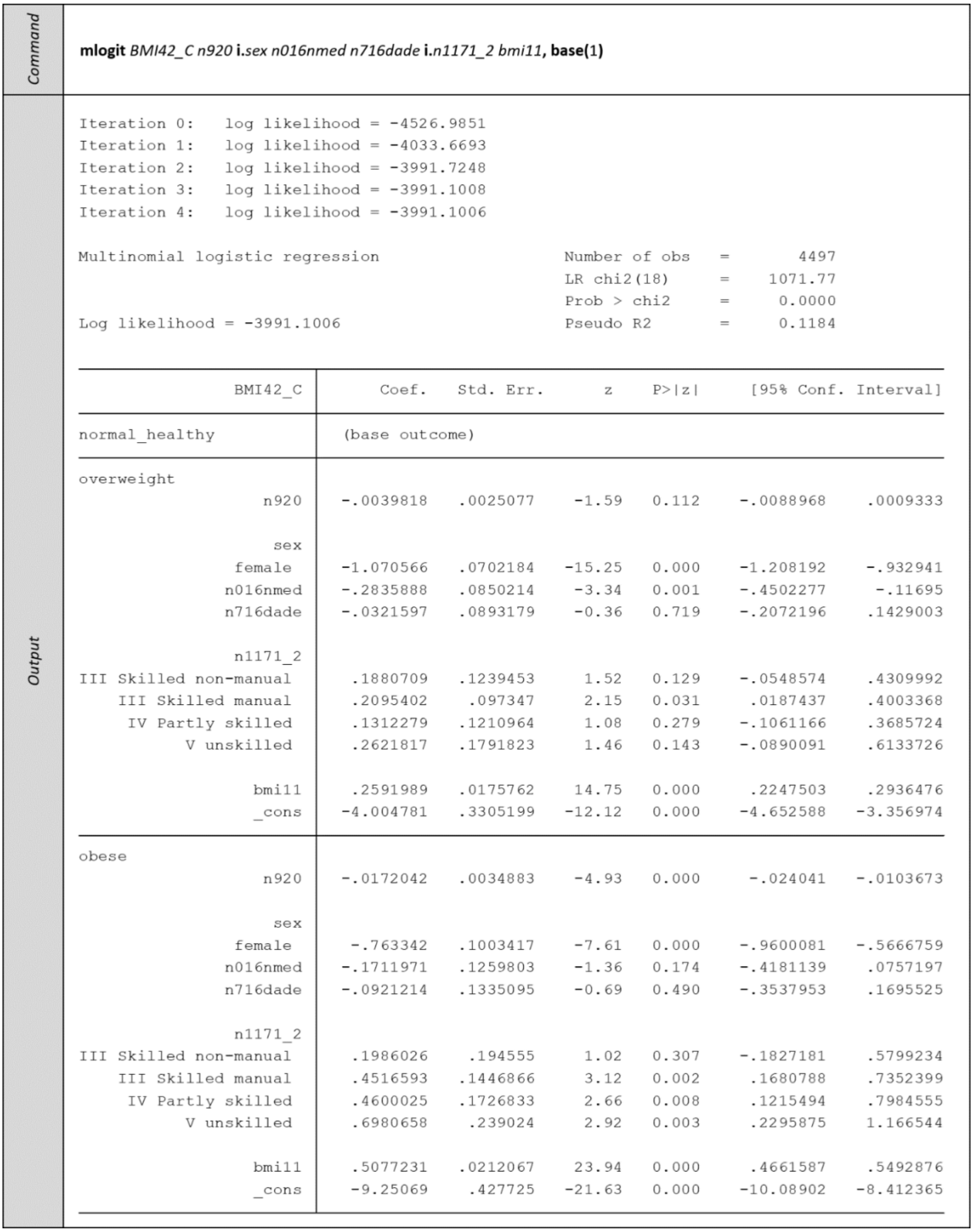
Interestingly in the output we can see that ‘general ability’ is significant in the ‘obese’ versus ‘normal/healthy’ BMI comparison, but not in the ‘overweight’ versus ‘normal/healthy’ BMI comparison after controlling for all the other predictors. A 1 unit decrease in ‘general ability’ test score is associated 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

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

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.



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

```
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
```

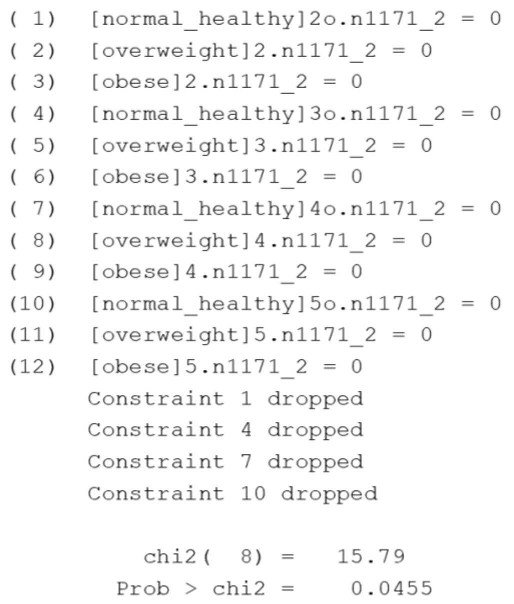
We can see that there is no significant difference between the association of when the participant’s mother left education and the participant’s own BMI in later life.

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

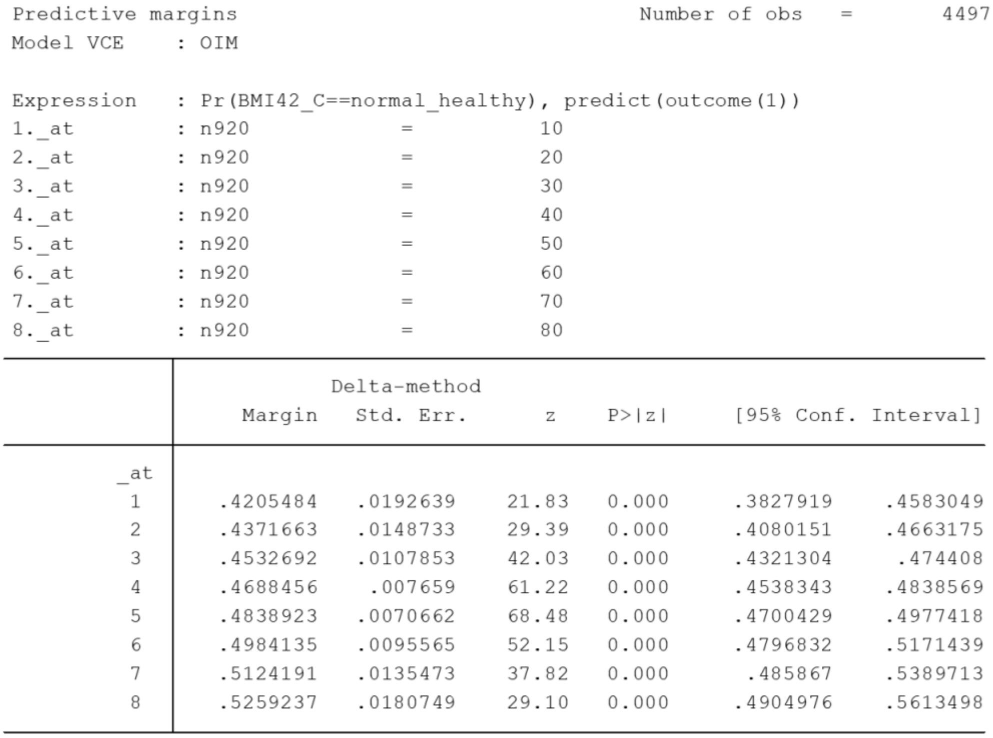
Focusing on our predictor of interest 'general ability', we can use predicted probabilities to help understand the relationship between 'general ability' and obesity, overweight and normal/healthy BMI in the model. In this example we want to calculate the predicted probability of the three BMI categories 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.

**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

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

Output



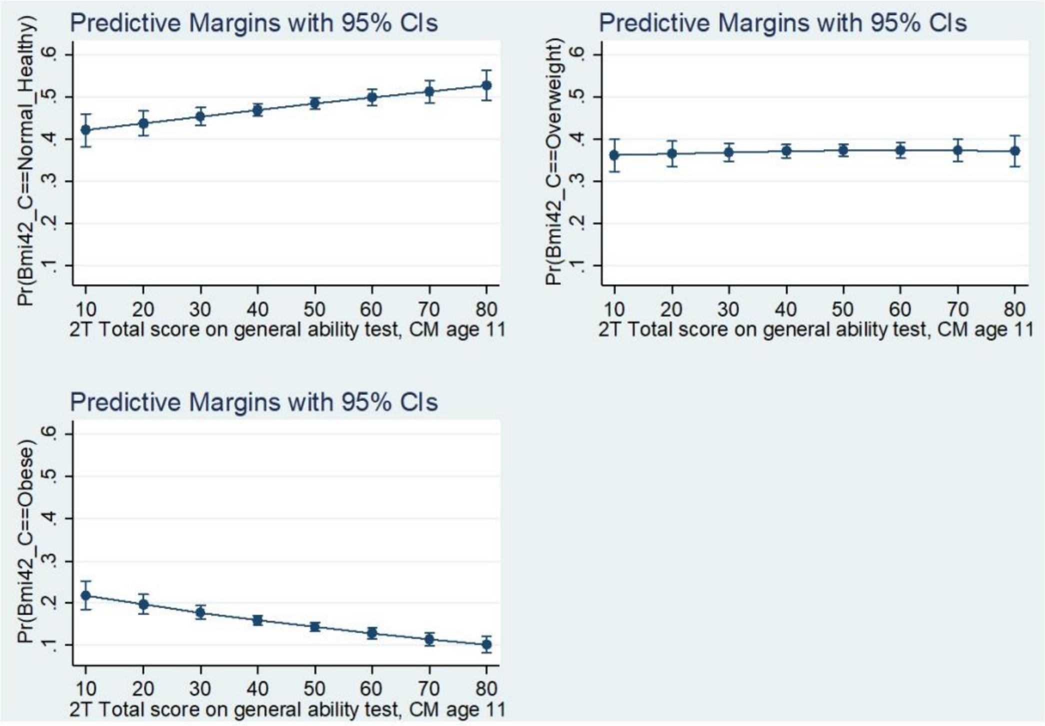
```
Predictive margins                              Number of obs     =      4497
Model VCE    : OIM

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

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   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.

**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* | **margins, at(***n920***=(**10(10)80**)) predict(outcome(**1**)) vsquish**<br>**marginsplot, name(**healthy**)**<br>**margins, at(***n920***=(**10(10)80**)) predict(outcome(**2**)) vsquish**<br>**marginsplot, name(**overweight**)**<br>**margins, at(***n920***=(**10(10)80**)) predict(outcome(**3**)) vsquish**<br>**marginsplot, name(**obese**)**<br>**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.

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

## 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