Section 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.

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.

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 regression" 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 the "Getting started and exploring the data" section. You will also need to construct a few of the explanatory variables before creating your regression model (as explained in that introductory section).

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.

Running the regression

In the first 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 displays the untransformed beta coefficients, which are 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 '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 1 unit decrease in general ability, the log odds of being obese (compared to not being obese) increases by 0.018.

Updating the regression model

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. 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.

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.

Exploring predictors' influence and predicted probabilities on the outcome

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<0.05).

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.

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**

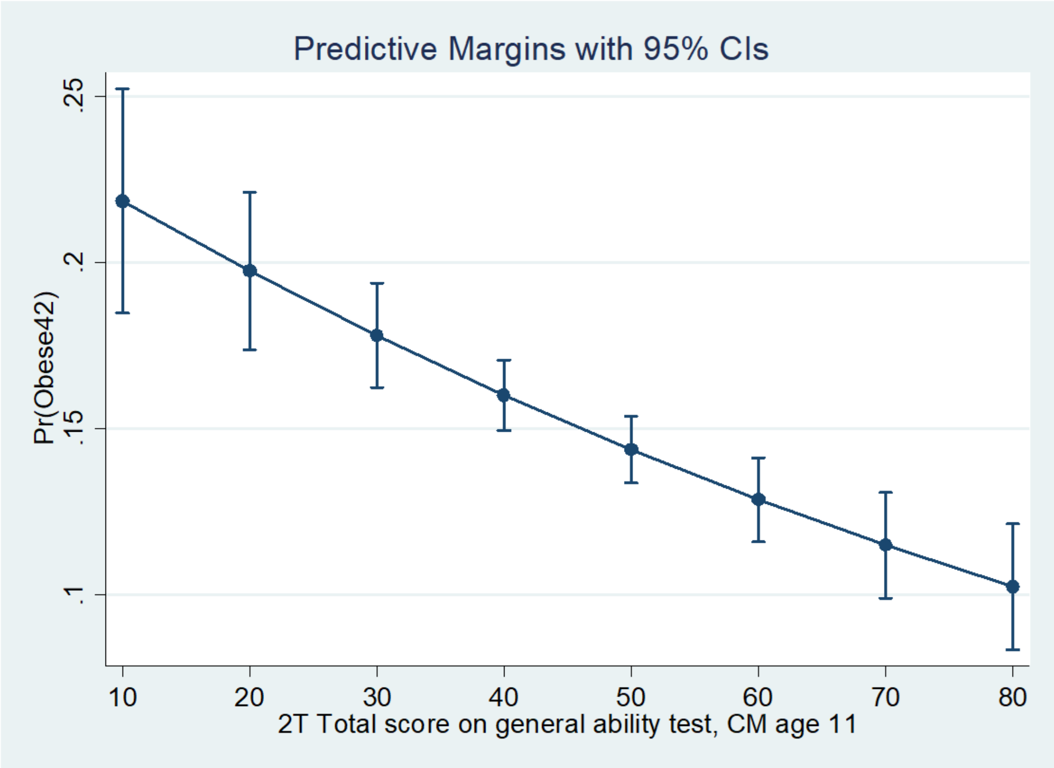
Output

```
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%.

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.

<table>
<tr><td>Command</td><td>marginsplot</td></tr>
<tr><td>Output</td><td>
</td></tr>
</table>

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.

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 BMI at age 11 in *M3* makes a notable further improvement to the model fit (chi-square = 526.14, *p*=<.001). The final test *M0* v *M3* compares the original model with no explanatory variables and our final model; and 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).

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.