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     4,495  19.3815059   R-squared       =    0.0135
-------------+----------------------------------   Adj R-squared   =    0.0132
       Total |  88307.7736     4,496  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.

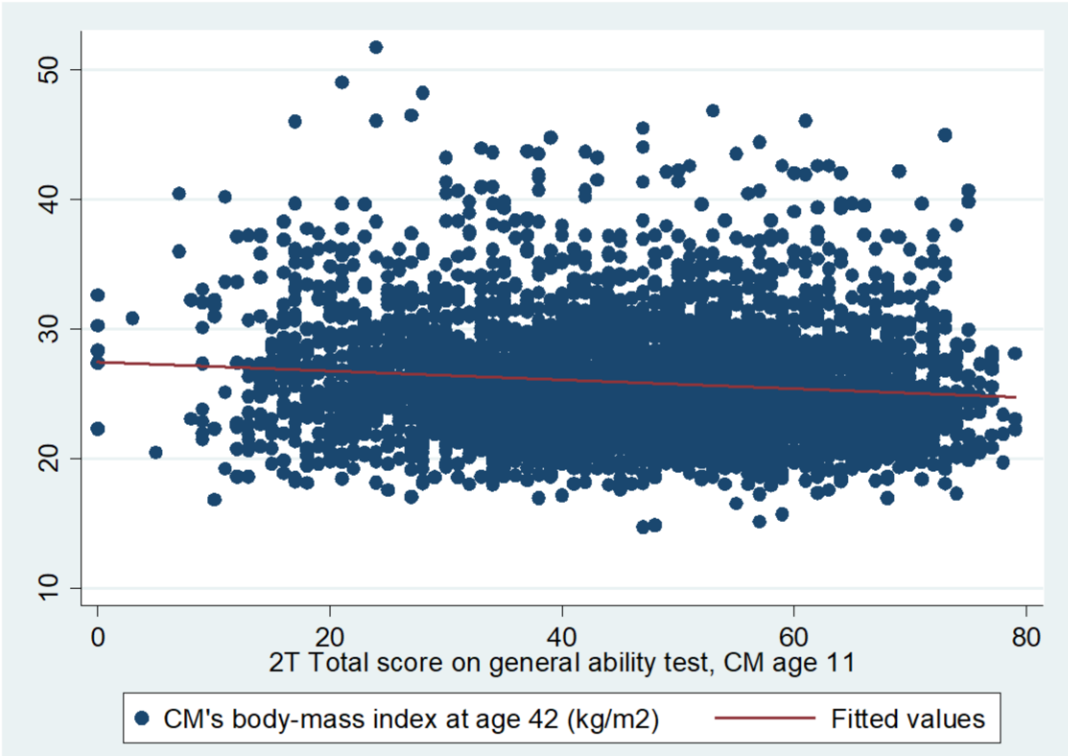
Command	**predict** *bmi_iq1*

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      Number of obs   =     4,497
-------------+----------------------------------   F(8, 4488)      =     23.46
       Model |  3544.33638         8  443.042048   Prob > F        =    0.0000
    Residual |  84763.4372     4,488  18.8866839   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

```
      Source |       SS           df       MS      Number of obs   =      4497
-------------+----------------------------------   F(  9,   4487)  =    173.44
       Model |  22791.6156         9  2532.40174   Prob > F        =    0.0000
    Residual |   65516.158     4487  14.6013278   R-squared       =    0.2581
-------------+----------------------------------   Adj R-squared   =    0.2566
       Total |  88307.7736     4496  19.6414087   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

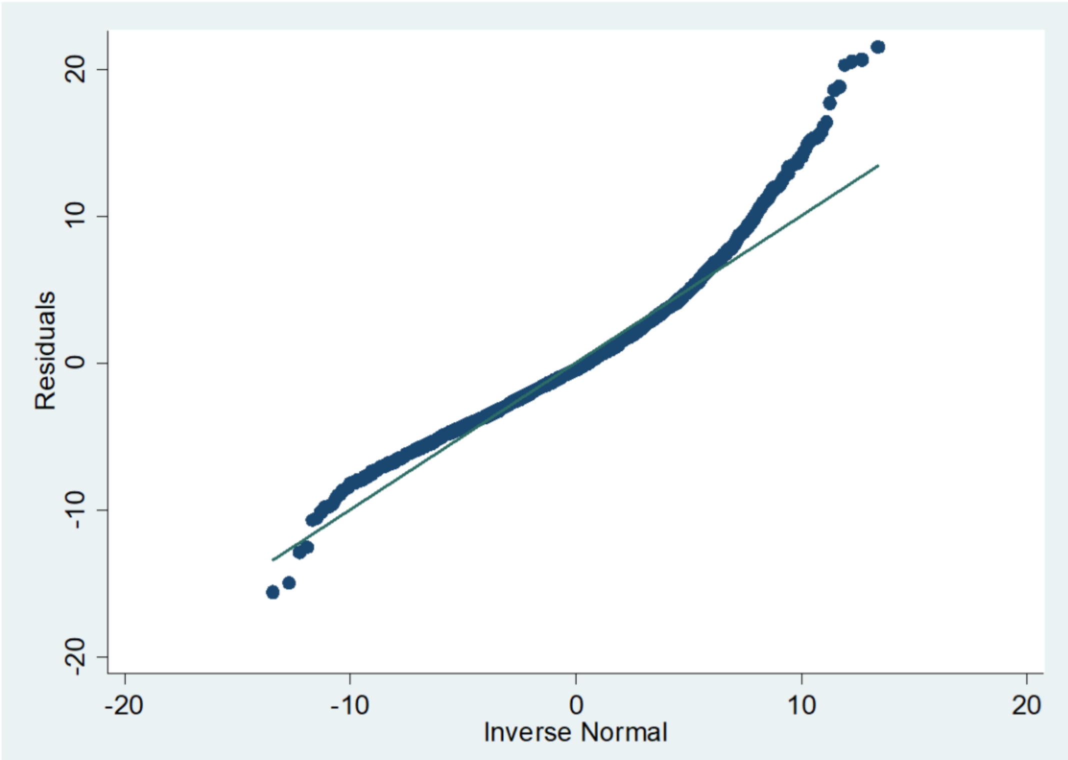
We have conducted this analysis without checking whether the data we have been using have met the assumptions underlying an ordinary least squares (OLS) linear regression. Three main assumptions we will however now briefly explore are normality, homogeneity of variance (homoscedasticity) and independence. 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.

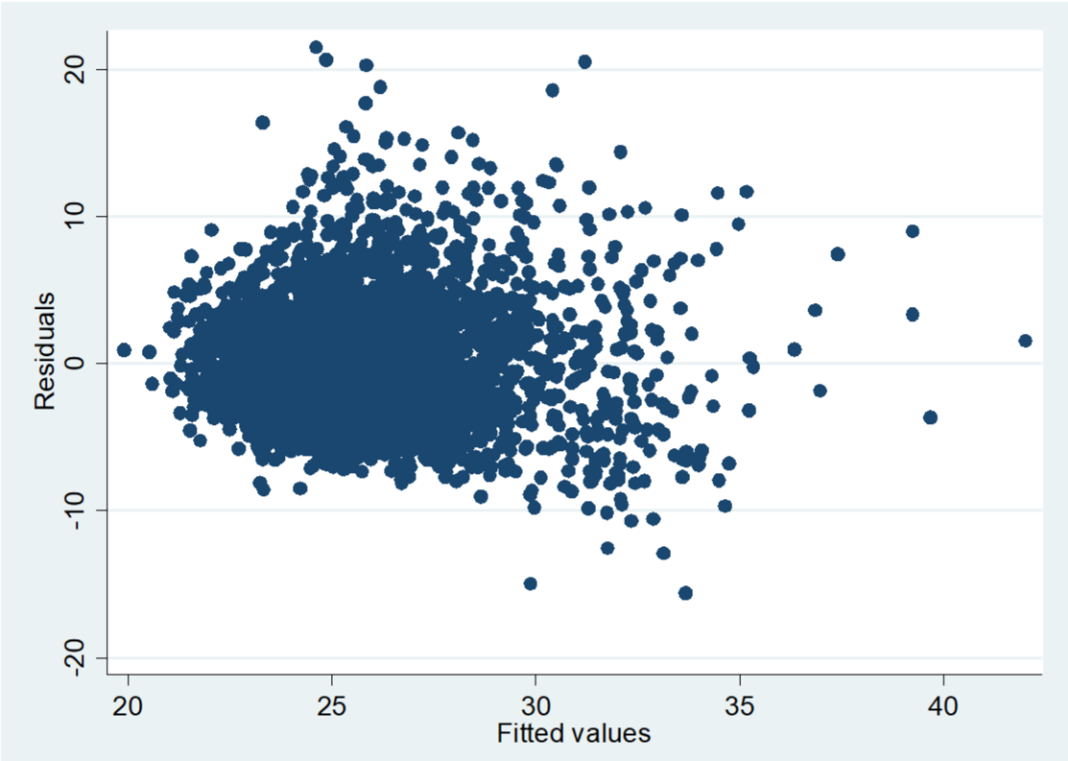
Command	**swilk** *bmi_iq2*
Output	Shapiro-Wilk W test for normal data

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.

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

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.