Module 3: Multiple Regression
MLwiN Practicals

Fiona Steele[1]
Centre for Multilevel Modelling

Contents

Module 3 (Practice): Multiple Regression
Some of the sections within this module have online quizzes for you to test your understanding. To find the quizzes:

EXAMPLE

From within the LEMMA learning environment
- Go down to the section for **Module 3: Multilevel Modelling**
- Click "3.1 Regression with a Single Continuous Explanatory Variable" to open Lesson 3.1
- Click Q1 to open the first question

Pre-requisites

- Understanding of types of variables (continuous vs. categorical variables, dependent and explanatory); covered in Module 1.
- Correlation between variables
- Confidence intervals
- Hypothesis testing, p-values
- Independent samples t-test for comparing the means of two groups

Online resources:

http://www.sportsci.org/resource/stats/
http://www.socialresearchmethods.net/
http://www.animatedsoftware.com/statglos/statglos.htm
http://davidmlane.com/hyperstat/index.html

The aim of these exercises is to gain practical experience of the application and interpretation of multiple regression. The MLwiN software will be used throughout.

Introduction to the Scottish Youth Cohort Trends Dataset

You will be analysing data from the Scottish School Leavers Survey (SSLS), a nationally representative survey of young people. We use data from seven cohorts of young people collected in the first sweep of the study, carried out at the end of the final year of compulsory schooling (aged 16-17) when most sample members had taken Standard grades[2]. These are subject-based examinations, typically taken in up to eight subjects. Each subject is graded on a scale from 1 (highest) to 7 (lowest). The dependent variable is a total attainment score calculated by assigning 7 points for a '1', 6 for a '2' and so on.

The analysis dataset contains the following five variables:

Variable name	Description and codes
CASEID	Anonymised student identifier.
SCORE	Point score calculated from awards in Standard grades. Scores range from 0 to 75, with a higher score indicating a higher attainment.
COHORT90	The sample includes the following cohorts: 1984, 1986, 1988, 1990, 1996 and 1998. The COHORT90 variable is calculated by subtracting 1990 from each value. Thus values range from -6 (corresponding to 1984) to 8 (1998), with 1990 coded as zero.
FEMALE	Sex of student (1=female, 0=male).
SCLASS	Social class, defined as the higher class of the mother or father (1=managerial and professional, 2=intermediate, 3=working, 4=unclassified).

There are 33988 students in the data file.

---

[2] We are grateful to Linda Croxford (Centre for Educational Sociology, University of Edinburgh) for providing us with these data. The dataset was constructed as part of an ESRC-funded project on Education and Youth Transitions in England, Wales and Scotland 1984-2002. Further analyses of the data can be found in Croxford, L. and Raffe, D. (2006) "Education Markets and Social Class Inequality: A Comparison of Trends in England, Scotland and Wales". In R. Teese (Ed.) *Inequality Revisited*. Berlin: Springer.

P3.1 Regression with a Single Continuous Explanatory Variable

We will begin by looking at the relationship between attainment (SCORE) and cohort (COHORT90). Has attainment changed over time and, if so, is the trend linear?
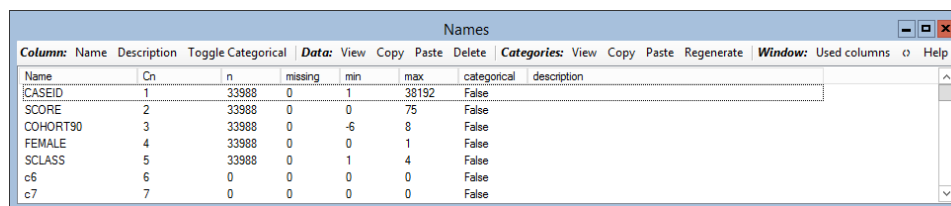
P3.1.1 Examining the data

To access the data files associated with this tutorial, you must have an account with LEMMA. To open the first data file,

> From within the LEMMA Learning Environment
> - Go to **Module 3: Multiple regression**, and scroll down to **MLwiN Datafiles**
> - If you do not already have MLwiN to open the datafile with, click **(get *MLwiN*)**.
> - Click " 3.1.wsz"

When the worksheet is opened, the filename will appear in the title bar of the main window. The **Names** window will also appear, giving a summary of the data in the worksheet:

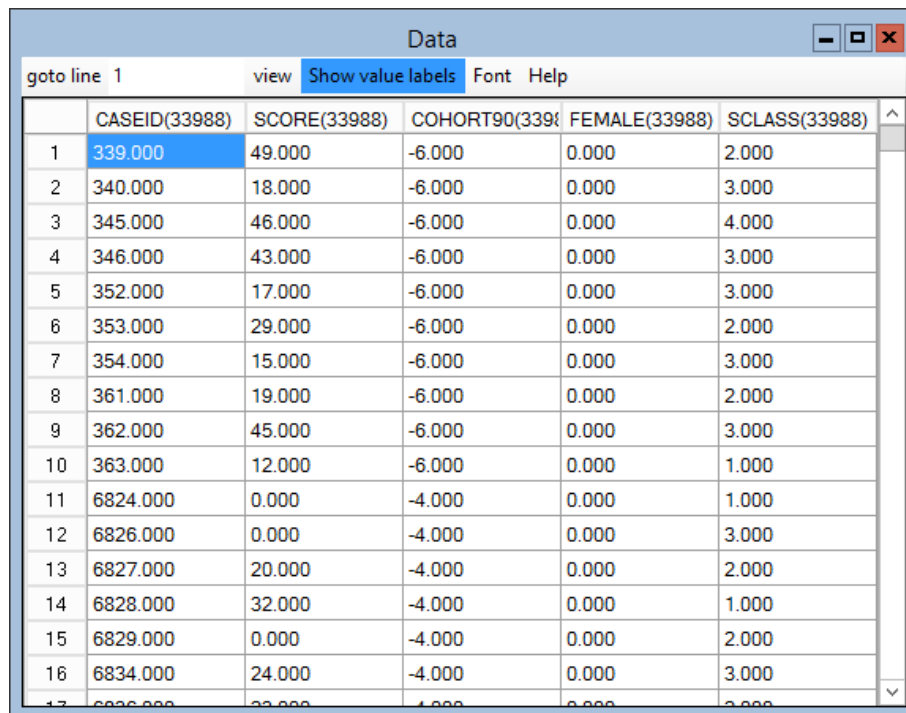| Name	Cn	n	missing	min	max	categorical	description
CASEID	1	33988	0	1	38192	False	
SCORE	2	33988	0	0	75	False	
COHORT90	3	33988	0	-6	8	False	
FEMALE	4	33988	0	0	1	False	
SCLASS	5	33988	0	1	4	False	
c6	6	0	0	0	0	False	
c7	7	0	0	0	0	False	

The MLwiN worksheet holds the data and other information in a series of columns, as on a spreadsheet. There are initially named c1, c2, etc. but we recommend that they be given meaningful names to show what their content relates to. This has already been done in the worksheet that you have loaded.

Each line in the body of the **Names** window summarises a column of data. In the present case only the first five of the 400 columns of the worksheet contain data. Each column contains 33988 values, one for each student represented in the data set. There are no missing values, and the minimum and maximum value in each column are shown. It is possible to define a variable as categorical (we shall do this later) and to add variable descriptions.

You can view individual values in the data using the **Data** window as follows:

> - From the **Data Manipulation** menu, select **View or edit data**

Module 3 (Practice): Multiple Regression
The following window appears:

Because there are only five variables in the data file, all columns can be seen. When there are more variables, you can view any selection of columns, spreadsheet fashion, as follows:

- Click the **View** button
- Select columns to view
- Click **OK**

You can select a block of adjacent columns either by pointing and dragging or by selecting the column at one end of the block and holding down 'Shift' while you select the column at the other end. You can add to an existing selection by holding down 'Ctrl' while you select new columns or blocks. Use the scroll bars of the **Data** window to move horizontally and vertically through the data, and move or resize the window if you wish. You can go straight to line 1000, for example, by typing 1000 in the **goto line** box, and you can highlight a particular cell by pointing and clicking. This provides a means to edit data.

Having viewed the data we will examine SCORE and COHORT90, the variables to be considered in our first regression analysis.
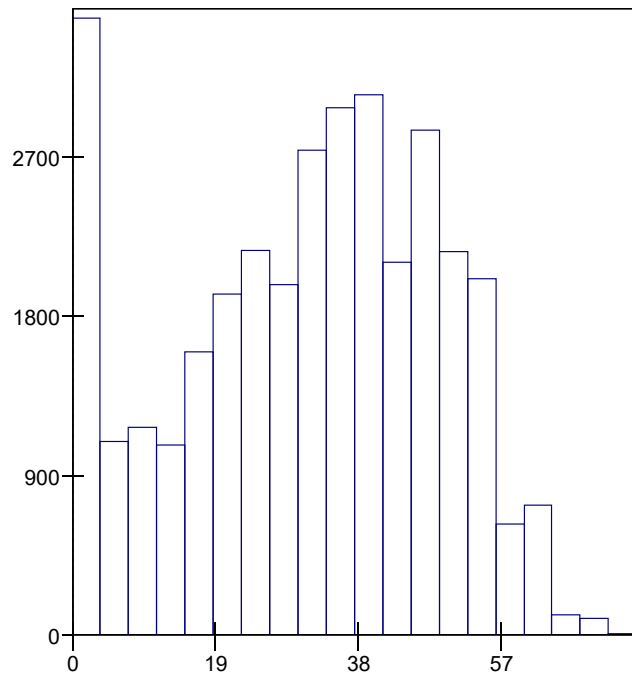
Distribution of SCORE

We will begin by obtaining a histogram and descriptive statistics for the dependent variable, SCORE.

To obtain a histogram:

Module 3 (Practice): Multiple Regression

The histogram should look like the above figure. Apart from a peak at around zero, the distribution looks approximately normal. Remember that in a linear regression model it is the residuals that are assumed to be normal; we will check this assumption at the end of the exercise.

To obtain descriptive statistics for SCORE:

An **Output** window opens, showing the number of cases, number of missing values, mean and standard deviation of SCORE. The mean is 31.095 and the standard deviation is 17.314.

Distribution of COHORT90

Because COHORT90 contains only six distinct values, we will look at its distribution in a frequency table rather than graphically.

This document is only the first few pages of the full version.
To see the complete document please go to learning materials and register:
http://www.cmm.bris.ac.uk/lemma
The course is completely free. We ask for a few details about yourself for our research purposes only. We will not give any details to any other organisation unless it is with your express permission.