Module 3: Multiple Regression
MLwiN Practicals

Fiona Steele
Centre for Multilevel Modelling

Contents

Introduction to the Scottish Youth Cohort Trends Dataset .................................... 3
P3.1 Regression with a Single Continuous Explanatory Variable ........................... 4
  P3.1.1 Examining the data ....................................................................... 4
  P3.1.2 A simple linear regression analysis .............................................. 11
P3.2 Comparing Groups: Regression with a Single Categorical Explanatory Variable 23
  P3.2.1 Comparing attainment for girls and boys ........................................... 23
  P3.2.2 Attainment by parental social class ................................................. 24
  P3.2.3 Fitting a non-linear relationship to attainment and cohort ............... 28
P3.3 Regression with More than One Explanatory Variable (Multiple Regression).... 31
P3.4 Interaction Effects ................................................................. 35
  P3.4.1 Model with fixed cohort effect for boys and girls ................................ 35
  P3.4.2 Fitting separate models for boys and girls .................................... 40
  P3.4.3 Allowing for sex-specific trends in a pooled analysis: interaction effects..... 42
  P3.4.4 Allowing the trend in attainment to depend on social class .......... 46
P3.5 Checking Model Assumptions in Multiple Regression ................................. 54
  P3.5.1 Checking the normality assumption ............................................. 55
  P3.5.2 Checking the homoskedasticity assumption .................................... 56

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. 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:

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description and codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASEID</td>
<td>Anonymised student identifier.</td>
</tr>
<tr>
<td>SCORE</td>
<td>Point score calculated from awards in Standard grades. Scores range from 0 to 75, with a higher score indicating a higher attainment.</td>
</tr>
<tr>
<td>COHORT90</td>
<td>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.</td>
</tr>
<tr>
<td>FEMALE</td>
<td>Sex of student (1=female, 0=male).</td>
</tr>
<tr>
<td>SCLASS</td>
<td>Social class, defined as the higher class of the mother or father (1=managerial and professional, 2=intermediate, 3=working, 4=unclassified).</td>
</tr>
</tbody>
</table>

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.
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:

- From the Graphs menu, select Customised Graph(s)
- Next to y, select SCORE from the drop-down list
- Next to plot type, select histogram
- Click Apply

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:

- From the Basic Statistics menu, select Averages and Correlations
- Under Operation, retain the default of Averages
- Highlight SCORE in the variable list
- Click Calculate
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.

- From the Basic Statistics menu, select Tabulate
- Under Output Mode, retain the default of Counts
- Under Display, check Percentages of row totals
- From the drop-down list next to Columns, select COHORT90
- Click Tabulate

You should see the following table in the Output window.

<table>
<thead>
<tr>
<th>COHORT90</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6</td>
<td>6478</td>
<td>19.1</td>
</tr>
<tr>
<td>-5</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>-4</td>
<td>6325</td>
<td>18.6</td>
</tr>
<tr>
<td>-3</td>
<td>5245</td>
<td>15.4</td>
</tr>
<tr>
<td>-2</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>0</td>
<td>1231</td>
<td>3.7</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>6</td>
<td>4244</td>
<td>12.6</td>
</tr>
<tr>
<td>7</td>
<td>7325</td>
<td>21.6</td>
</tr>
<tr>
<td>8</td>
<td>33988</td>
<td>100.0</td>
</tr>
</tbody>
</table>

The number of observations in each category from -6 (year 1984) to 8 (year 1998) are shown. Some are empty because the cohorts in the sample are not from consecutive years. Shown below the number of students (N) is the percentage in each category. The largest proportion of students are from the 1998 cohort, with somewhat smaller proportions from 1990 and 1996.

**Relationship between SCORE and COHORT90**

Before fitting a linear regression model with attainment and cohort, we will examine the nature of their relationship using a scatter plot.

- From the Graphs menu, select Customized Graph(s)
- From the drop-down list labelled plot type, select point
- From the drop-down list labelled y, select SCORE
- From the drop-down list labelled x, select COHORT90
- Click Apply

Clicking anywhere in the Graph display window will bring up the Graph options window from which you can, for example, add titles and change the axes scales.

Although there is some suggestion of a positive linear trend, it is difficult to see the relationship because of the small number of distinct values of COHORT90. We will therefore supplement the scatterplot with a table of the mean attainment score for each value of COHORT90.

Before requesting this table, we will compute a recoded version of COHORT90 with consecutive values 1, 2, ..., 6. This will avoid empty cells in the table where there are gaps between cohorts.
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.