# Module 3: Multiple Regression

# Stata Practical

*George Leckie* [1]
Centre for Multilevel Modelling

**Pre-requisites box**

- Modules 1-2

## 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 ................................................... 9

**P3.2 Comparing Groups: Regression with a Single Categorical Explanatory Variable 18**

P3.2.1 Comparing attainment for girls and boys .......................................... 18
P3.2.2 Attainment by parental social class .................................................. 19
P3.2.3 Fitting a non-linear relationship to attainment and cohort ...................... 22

**P3.3 Regression with More than One Explanatory Variable (Multiple Regression) ... 25**

**P3.4 Interaction Effects ........................................................................... 29**

P3.4.1 Model with fixed cohort effect for boys and girls ................................. 29
P3.4.2 Fitting separate models for boys and girls .......................................... 33
P3.4.3 Allowing for sex-specific trends in a pooled analysis: interaction effects ..... 34
P3.4.4 Allowing the trend in attainment to depend on social class ..................... 36

**P3.5 Checking Model Assumptions in Multiple Regression ................................ 41**

P3.5.1 Checking the normality assumption ................................................. 42
P3.5.2 Checking the homoskedasticity assumption ....................................... 43

[1] This Stata practical is adapted from the corresponding MLwiN practical: Steele, F. (2008) Module 3: Multiple Regression MLwiN Practical. LEMMA VLE, Centre for Multilevel Modelling. Accessed at http://www.cmm.bris.ac.uk/lemma/course/view.php?id=13.

**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 Q 1 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.

# 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 33,988 students in the dataset.

[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

Load "3.1.dta" into memory and open the do-file for this lesson:

From within the LEMMA Learning Environment

- Go to **Module 3: Multiple Regression**, and scroll down to ***Stata* Datasets and Do-files**
- Click " 3.1.dta" to open the dataset

and use the `describe` command to produce a basic description of the dataset which includes some general information on the number of variables and observations, along with a description of every variable in the dataset:

```
. describe

Contains data from 3.1.dta
  obs:        33,988
 vars:             5                          4 Sep 2009 10:40
 size:       543,808 (99.9% of memory free)
--------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
caseid          float  %9.0g                  Case ID
score           byte   %9.0g                  Score
cohort90        byte   %9.0g                  Cohort
female          byte   %9.0g                  Female
sclass          byte   %9.0g                  Social class
--------------------------------------------------------------------------------
Sorted by:
```

The dataset contains 33,988 observations on 5 variables and each variable has been given a variable label.

You can view individual values in the data using the `list` command. Here we use the `in` range qualifier to restrict the scope of the command to the first 20 observations in the data (to view, for example, the last 20 observations change the range to `-20/-1`). Note, we have not typed any variable names after the command and so the values of all the variables are displayed:

```
. list in 1/20

     +---------------------------------------------+
     | caseid   score   cohort90   female   sclass |
     |---------------------------------------------|
  1. |    339      49         -6        0        2 |
  2. |    340      18         -6        0        3 |
  3. |    345      46         -6        0        4 |
  4. |    346      43         -6        0        3 |
  5. |    352      17         -6        0        3 |
     |---------------------------------------------|
  6. |    353      29         -6        0        2 |
  7. |    354      15         -6        0        3 |
  8. |    361      19         -6        0        2 |
  9. |    362      45         -6        0        3 |
 10. |    363      12         -6        0        1 |
     |---------------------------------------------|
 11. |   6824       0         -4        0        1 |
 12. |   6826       0         -4        0        3 |
 13. |   6827      20         -4        0        2 |
 14. |   6828      32         -4        0        1 |
 15. |   6829       0         -4        0        2 |
     |---------------------------------------------|
 16. |   6834      24         -4        0        3 |
 17. |   6836      23         -4        0        2 |
 18. |  13206       7         -2        0        3 |
 19. |  13209      38         -2        0        3 |
 20. |  13215      46         -2        0        1 |
     +---------------------------------------------+
```

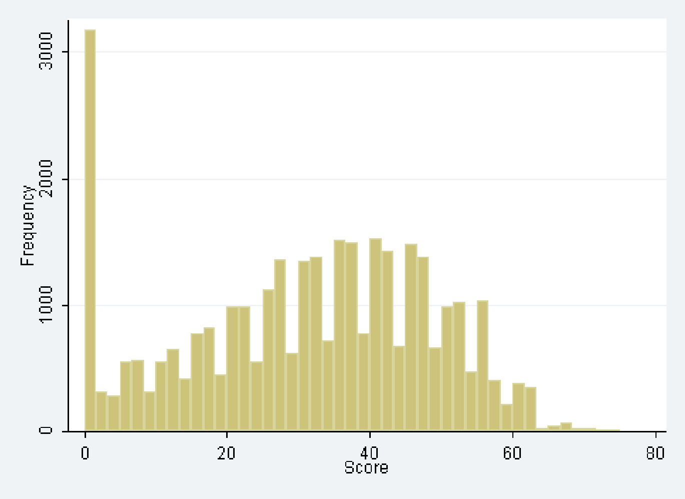
For example, the 10th student in the data belongs to the 1984 cohort and scored 12 out of 75. This student is a boy from a managerial social class background.

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:

```
. histogram score, frequency
```



The histogram should look like the above figure. The `frequency` option causes the histogram to be scaled in frequencies (as opposed to the default of scaling the histogram in density units). 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.

The `summarize` command can be used to calculate and display a variety of univariate summary statistics for the variables in the dataset. To obtain summary statistics only for **score**:

```
. summarize score

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       score |     33988    31.09462    17.31437          0         75
```

We see that there are 33,988 observations and that **score** has a mean of 31.095, a standard deviation of 17.314 and can range between a minimum and maximum value of 0 and 75.

### *Distribution of cohort90*

Because **cohort90** contains only six distinct values, we will look at its distribution in a frequency table rather than graphically. The `tabulate` command produces one-way (and two-way) tables of frequency counts.

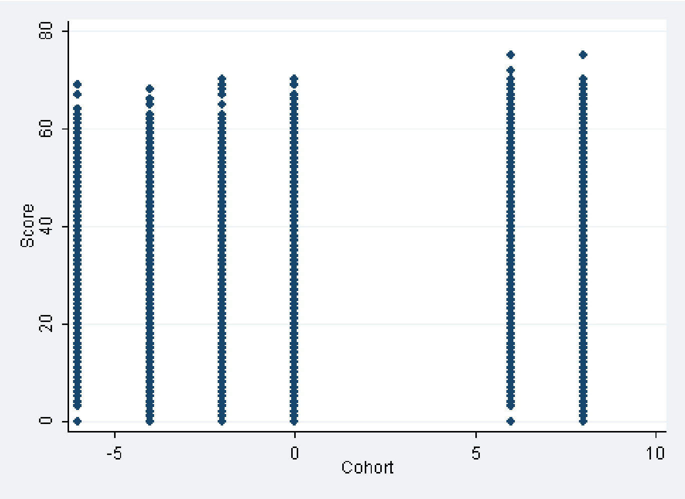
```
. tabulate cohort90

     Cohort |      Freq.     Percent        Cum.
------------+-----------------------------------
         -6 |      6,478       19.06       19.06
         -4 |      6,325       18.61       37.67
         -2 |      5,245       15.43       53.10
          0 |      4,371       12.86       65.96
          6 |      4,244       12.49       78.45
          8 |      7,325       21.55      100.00
------------+-----------------------------------
      Total |     33,988      100.00
```

The number of observations in each category from -6 (year 1984) to 8 (year 1998) are shown. The column labelled `Percent` shows the percentage of students in the dataset that belong to the indicated cohort. For example, 12.86% of students in our dataset belong to the 1990 cohort (coded zero). The largest proportion of students is from the 1998 cohort, with somewhat smaller proportions from 1990 and 1996. The final column `Cum.` shows the cumulative percentage. For example, we see that 65.96% of students in our dataset belong to either the 1990 cohort or one of the three earlier cohorts (1984, 1986 and 1988).

### *Relationship between score and cohort90*

Before fitting a linear regression model with attainment and cohort, we use the `scatter` command to examine the nature of their relationship using a scatterplot.

```
. scatter score cohort90
```



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**.

To tabulate the mean of **score** for each value of **cohort90**, we use the `table` command. The `contents()` option specifies the contents of the table's cells. For example, `contents(mean score)` would produce a table of the means of the variable **score** for each value of **cohort90**. Below, we specify `contents(freq mean score sd score)` to additionally produce the frequencies and standard deviation of the variable **score** for each value of **cohort90**. Finally, the `row` option specifies that a row be added to the table reflecting the total across the rows. This can be seen in the last line of the output below:

```
. table cohort90, contents(freq mean score sd score) row

--------------------------------------------------
   Cohort |       Freq.  mean(score)    sd(score)
----------+---------------------------------------
       -6 |       6,478     23.65545     18.07995
       -4 |       6,325     24.77265     17.37533
       -2 |       5,245      28.5245     15.93629
        0 |       4,371     29.10044     15.76355
        6 |       4,244     39.43473     13.55147
```