Module 3: Multiple Regression Concepts

Fiona Steele[1]
Centre for Multilevel Modelling

Contents

[1] With additional material from Kelvyn Jones. Comments from Sacha Brostoff, Jon Rasbash and Rebecca Pillinger on an earlier draft are gratefully acknowledged.

All 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

All of the sections within this module have practicals so you can learn how to perform this kind of analysis in MLwiN or other software packages. To find the practicals:

EXAMPLE

From within the LEMMA learning environment
- Go down to the section for Module 3: Multiple Regression, then

Either
- Click "3.1 Regression with a Single Continuous Explanatory Variable" to open Lesson 3.1
- Click MLwiN practical ()

Or
- Click Print all Module 3 MLwiN Practicals

Pre-requisites

- Understanding of types of variables (continuous vs. categorical variables, dependent and explanatory); covered in Module 1.
- Correlation between variables
- Confidence intervals around estimates
- 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

Introduction

What is Multiple Regression?

Multiple regression is a technique used to study the relationship between an outcome variable and a set of explanatory or predictor variables.

Motivation

To illustrate the ideas of multiple regression, we will consider a research problem of assessing the evidence for gender discrimination in legal firms. Statistical modelling can provide the following:

- A quantitative assessment of the size of the effect; e.g. the difference in salary between women and men is £5000 per annum;
- A quantitative assessment after *taking account of* other variables; e.g. a female worker earns £6500 less after taking account of years of experience. This *conditioning* on other variables distinguishes multiple regression modelling from simple 'testing for differences' analyses.
- A measure of uncertainty for the size of the effect; e.g. we can be 95% confident that the female-male difference in salary in the population from which our sample was drawn is likely to lie between £4500 and £5500.

We can use regression modelling in different modes: 1) as description (what is the average salary for men and women?), 2) as part of causal inference (does being female result in a lower salary?), and 3) for prediction ('what happens if' questions).

Conditioning

The key feature that distinguishes multiple regression from simple regression is that more than one predictor variable is involved. Even if we are interested in the effect of just one variable (gender) on another (salary) we need to take account of other variables as they may compromise the results. We can recognise three distinct cases where it is important to *control or adjust for* the effects of other variables:

i) *Inflation* of a relationship when not taking into account extraneous variables. For example, a substantial gender effect could be reduced after taking account of type of employment. This is because jobs that are characterized by poor pay (e.g. in the service sector) have a predominantly female labour force.

ii) *Suppression* of a relationship. An apparent small gender gap could increase when account is taken of years of employment; women having longer service and poorer pay.

iii) *No confounding*. The original relationship remains substantially unaltered when account is taken of other variables. Note, however, that there may be unmeasured confounders.

Data for multiple regression analysis

Statistical analysis requires a quantifiable outcome measure (dependent variable) to assess the effects of discrimination. Possibilities include the following, differentiated by the nature of the measurement: a continuous measure of salary, a binary indicator of whether an employee was promoted or not, a three-category indicator of promotion (promoted, not promoted, not even considered), a count of the number of times rejected for promotion, the length of time that it has taken to gain promotion. All of these outcomes can be analysed using regression analysis, but different techniques are required for different scales of measurement.

The term 'multiple regression' is usually applied when the dependent variable is measured on a continuous scale. A dichotomous dependent variable can be analysed using logistic regression and multinomial logistic and ordinal regression can be applied to nominal and ordinal dependent variables respectively. There are also methods for handling counts (Poisson regression) and time-to-event data (event history analysis or survival analysis). These techniques will be described in later Modules.

The explanatory variables may also have different scales of measurement. For example, gender is a binary categorical variable; ethnicity is categorical with more than two categories; education might be measured on an ordinal scale (e.g. <11, 11-13, 14-16 and >16 years of education); years of employment could be measured on a continuous scale. Multiple regression can handle all of these types of explanatory variable, and we will consider examples of both continuous and categorical variables in this Module.

Introduction to Dataset

The ideas of multiple regression will be introduced using data from the 2002 European Social Surveys (ESS). Measures of ten human values have been constructed for 20 countries in the European Union. According to value theory, values are defined as desirable, trans-situational goals that serve as guiding principles in people's lives. Further details on value theory and how it is operationalised in the ESS can be found on the ESS education net (http://essedunet.nsd.uib.no/cms/topics/1/).

We will study one of the ten values, *hedonism*, defined as the 'pleasure and sensuous gratification for oneself'. The measure we use is based on responses to the question "How much like you is this person?":

- *He (sic) seeks every chance he can to have fun. It is important to him to do things that give him pleasure.*
- *Having a good time is important to him. He likes to "spoil" himself.*

A respondent's own values are inferred from their self-reported similarity to a person with the above descriptions. Each of the two items is rated on a 6-point scale (from "very much like me" to "not like me at all"). The mean of these ratings is calculated for each individual. The mean of the two hedonism items is then adjusted for individual differences in scale use[2] by subtracting the mean of all value items (a total of 21 are used to measure the 10 values). These *centred* scores recognise that the 10 values function as a system rather than independently. The centred hedonism score is interpreted as a measure of the *relative* importance of hedonism to an individual in their whole value system.

The scores on the hedonism variable range from -3.76 to 2.90, where higher scores indicate more hedonistic beliefs.

We consider three countries – France, Germany and the UK – with a total sample size of 5845. That is, we use a subsample of the original data.

Hedonism is taken as the outcome variable in our analysis. We consider three explanatory variables:

- Age in years
- Gender (coded 0 for male and 1 for female)
- Country (coded 1 for the UK, 2 for Germany and 3 for France)
- Years of education.

An extract of the data is given below.

Respondent	Hedonism	Age	Gender	Country	Education
1	1.55	25	0	2	10
2	0.76	30	0	2	11
3	-0.26	59	0	2	9
4	-1.00	47	1	3	10
.	.	.	.	.	.
.	.	.	.	.	.
5845	0.74	65	0	1	9

[2] Some individuals will tend to select responses from one side of the scale ("very much like me") for *any* item, while others will select from the other side ("not like me at all"). If we ignore these differences in response tendency we might incorrectly infer that the first type of individual believes that all values are important, while the second believes that all values are unimportant.

C3.1 Regression with a Single Continuous Explanatory Variable

We will begin with a description of simple linear regression for studying the relationship between a **pair** of continuous variables, which we denote by Y and X. Simple regression is also commonly known as *bivariate* regression because only two variables are involved.

Y is the **outcome** variable (also called a response or dependent variable)
X is the **explanatory** variable (also called a predictor or independent variable).

C3.1.1 Examining data graphically

Before carrying out a regression analysis, it is important to look at your data first. There are various assumptions made when we fit a regression model, which we will consider later, but there are two checks that should always be carried out before fitting any models: i) examine the distribution of the variables and check that the values are all valid, and ii) look at the nature of the relationship between X and Y.

Distribution of Y

We can examine the distribution of a continuous variable using a histogram. At this stage, we are checking that the values appear reasonable. Are there any 'outliers', i.e. observations outside the general pattern? Are there any values of -99 in the data that should be declared as missing values? We also look at the shape of the distribution: is it a symmetrical bell-shaped distribution (normal), or is it skewed? Although it is the residuals[3] that are assumed to be normally distributed in a multiple regression model, rather than the dependent variable, a skewed Y will often produce skewed residuals. If the residuals turn out to be non-normal, it may be possible to transform Y to obtain a normally distributed variable. For example, a positively skewed distribution (with a long tail to the right) will often look more symmetrical after taking logarithms.

Figure 3.1 shows the distribution of the hedonism scores. It appears approximately normal with no obvious outliers. The mean of the hedonism score is -0.15 and the standard deviation is 0.97.

Distribution of X

[3] The residual for each observation is the difference between the observed value of Y and the value of Y predicted by the model. See C3.1.2 for further details.

For a regression analysis the distribution of the explanatory variable is unimportant, but it is sensible to look at descriptive statistics for any variables that we analyse to check for unusual values.

We will first consider age as an explanatory variable for hedonism. The age range in our sample is 14 to 98 years with a mean of 46.7 and standard deviation of 18.1.

Relationship between X and Y

In its simplest form, a regression analysis assumes that the relationship between X and Y is linear, i.e. that it can be reasonably approximated by a straight line. If the relationship is nonlinear, it may be possible to transform one of the variables to make the relationship linear or the regression model can be modified (see C3.3.3). The relationship between two variables can be viewed in a scatterplot. A scatterplot can also reveal outliers.

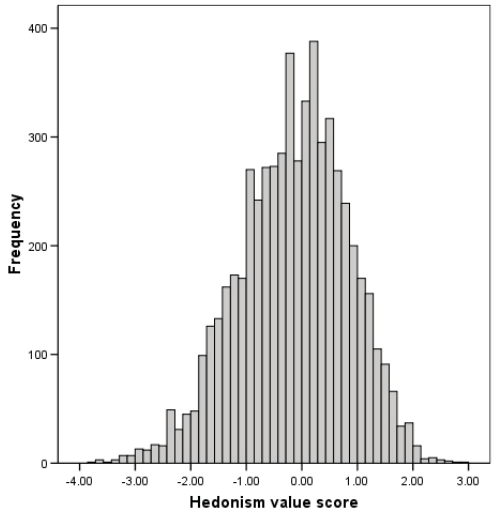

Figure 3.1. Histogram of hedonism

Figure 3.2 shows a scatterplot of hedonism versus age, where the size of the plotting symbol is proportional to the number of respondents represented by a particular data point. Also shown is what is commonly called the *line of best fit*, which we will come back to in a moment. The scatterplot shows a negative relationship: as age increases then hedonism decreases. The Pearson correlation coefficient for the linear relationship is -0.34.

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.