Quantitative analysis in social sciences: An brief introduction for non-economists

Niño-Zarazúa, Miguel

World Institute for Development Economics Research

May 2012

Quantitative analysis in social sciences:
An brief introduction for non-economists

Miguel Niño-Zarazúa[1]
World Institute for Development Economics Research
United Nations University

Abstract

In this paper, I present an introduction to quantitative research methods in social sciences. The paper is intended for non-Economics undergraduate students, development researchers and practitioners who although unfamiliar with statistical techniques, are interested in quantitative methods to study social phenomena. The paper discusses conventional methods to assess the direction, strength and statistical significance of the correlation between two or more variables, and examines regression techniques and experimental and quasi-experimental research designs to establish causality in the analysis of public interventions.

Introduction

In social sciences, quantitative analysis is used to determine, in a given population, the relationship between an independent variable –a type of quantities that capture observed values of things that can be manipulated, and a dependent variable, the observed results of the manipulation of the independent variable. For instance, you could be interested in determining the relationship between years of schooling (in this case the independent variable) and wages paid to workers employed in the garment industry (the dependent variable). So you would need to collect information about the number of years of schooling among workers and their wages to determine that relationship. You might also be interested in determining the relationship between a balanced diet amongst school children (the independent variable) and their cognitive achievements (the dependent variable).[2] In this case, you would need to collect information about food intake, for example, the number of meals per day, and carry out cognitive achievement tests in reading, mathematics and written language skills.

Formally, the relationship between the independent and dependent variables is expressed as a function, $y = f(x)$, where the letter "x" is used to represent the value of the independent variable, the letter "y" is used to represent the dependent variable, and the letter "f" is used

[1] Contact details at: United Nations University World Institute for Development Economics Research (UNU-WIDER), Katajanokanlaituri 6 B, FI-00160 Helsinki, Finland. Email: miguel@wider.unu.edu

[2] The independent variable is also referred to as 'regressor', 'explanatory variable', 'predictor', and/or 'exogenous variable', whereas the dependent variable is also referred to as 'regressand', 'explained variable', 'target variable' and/or 'outcome variable'.

to denote “a function of”. So the term $y = f(x)$ can be read as *y*, the dependent variable, is a function of *x*, the independent variable. Following the previous example, researchers would be interested in hypothesising that “the levels of earnings (the dependent variable) is a function of years of schooling (the independent variable), that is, earnings = *f* (years of schooling), and then investigate the direction, strength and significance of that relationship. In statistics that relationship is known as *correlation*.

1. Pearson’s product-moment correlation coefficient

A conventional method to assess the direction and strength of a correlation between the two variables is the *Pearson’s product-moment correlation coefficient*, also known as *Pearson’s rho*, o simply *Pearson’s r,* which is the product of the linear relationship between a dependent and independent variable.[3] Formally, the Pearson’s correlation coefficient is defined as the *covariance* of two variables divided by the product of their *standard deviation.* The concept of covariance can be understood as a measure of how much the two variables change together whereas the standard deviation is a measure of how much variation exists between the observed values of the two variables and the mean. In order to determine the significance of such correlations, you will need to use hypothesis testing and confidence levels methods. I return to this issue below in Section 3.

The use of the Pearson’s r requires that the independent and dependent variables are measured continuously in interval or ratio scales.[4] An interval variable allows us to identify equally spaced values in intervals although they have no meaningful zero value. A commonly referred example of an interval variable is temperature measured on the Fahrenheit scale, where a zero degree does not mean there is no temperature, but rather a very cold temperature, about -17.78 degrees Celsius. However, we rarely use interval variables in Social Sciences, so I focus instead on ratio scales.

Ratio scale variables have the same properties of interval variables but they are distinct from interval variables in the sense that they have an identifiable zero point. Income, expenditure, revenues, costs, profits, and other continuous variables are examples of ratio scale variables. Important properties of ratio scales is that they can be measured, compared, and ranked in terms of size differences, so if, for instance, in a dataset you have two individuals, A and B, and A reports a monthly income of US $2000 while B, has an income of US $1000,

[3] It is called “Pearson correlation coefficient” in honour of Karl Pearson, a British mathematician who developed the method.

[4] I discuss the case of nominal and ordinal variables in Section 4

we can say with certainty that individual A enjoys a greater income than individual B. Time and spatial measures can also be classified as ratio scale variables. For example, ‘age’ can be measured in months or years, whereas ‘distance’ can be measured in meters or kilometres. Following the previous example, if individual A is aged 40 and individual B, 20, we could hypothesise that there is a correlation between income and age.

The direction of the correlation coefficient can take values between -1 and +1, going through zero. A positive correlation coefficient means that increases in the independent variable, x, is associated with an increase in the dependent variable, y, whereas a negative correlation coefficient means that an increase in the independent variable is associated with a decrease in the dependent variable. If your data shows that workers with more years of schooling receive higher wages, you could determine that there is a positive correlation between schooling and wages. This hypothetical case is illustrated in Figure 1, where the plotted points move from the lower left-hand corner upwards to the upper right-hand corner of the scattergram. Similarly, if your data shows that higher consumption of alcohol among workers is correlated with premature death, you may determine that there is a *negative* correlation between alcohol consumption and life expectancy. This is illustrated in Figure 2, where the plotted points move from the upper left-hand corner downwards to the lower right-hand corner of the scattergram.

Figure 1 Perfect positive correlation
Pearson's r= 1

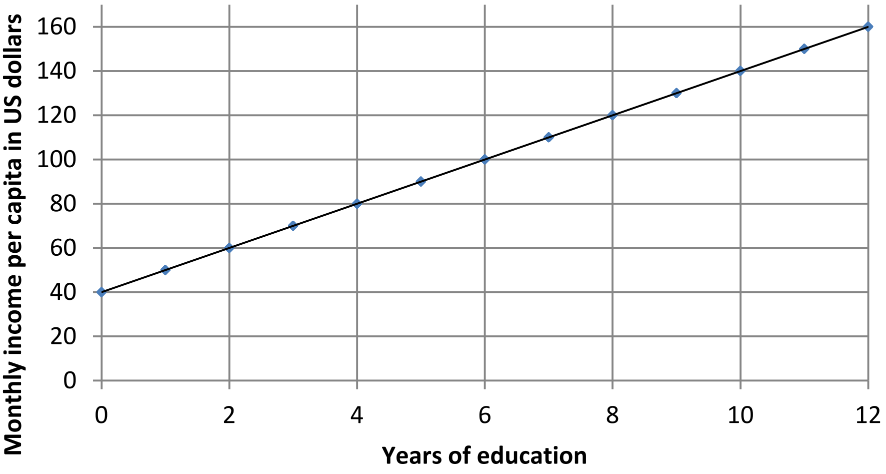
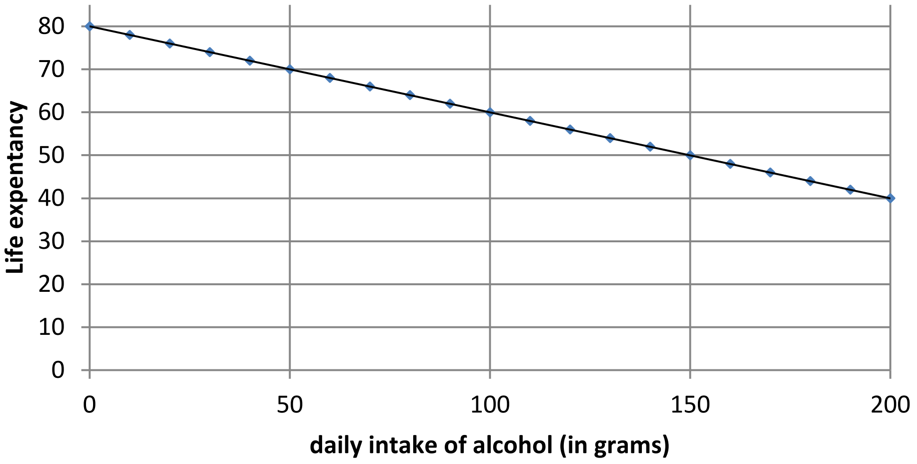
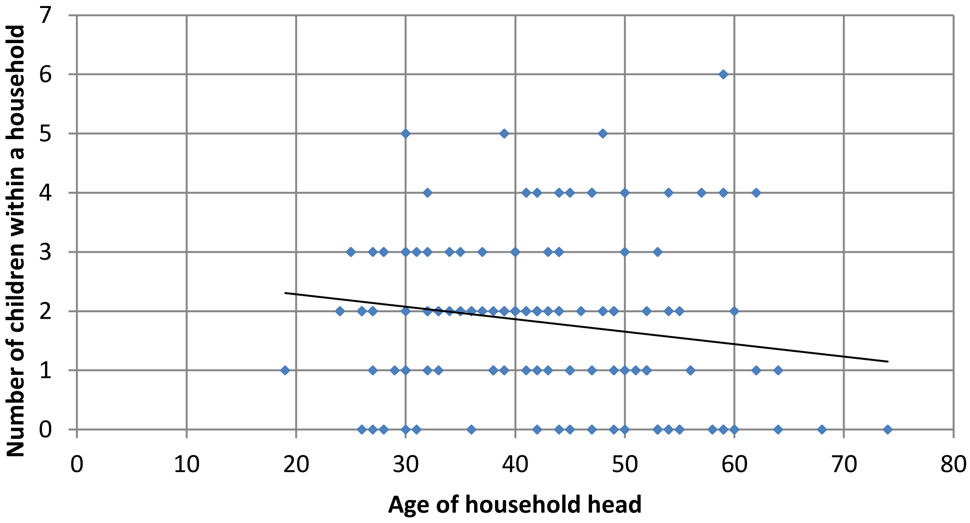


Figure 2 Perfect negative correlation
Pearson's r = -1



You must be aware of the fact that in many cases, the level of correlation between two variables may not be as clear as you might have thought, in either a positive or negative direction. For example, Figure 2 shows the correlation between the numbers of children living in a household, and the age of the household head. Clearly, the relationship between the two variables is not a straight line, which stems from the fact that the Pearson's correlation coefficient shows a weak and negative association (r= -0.17).

Figure 3 Weak correlation
Pearson's r (146) = -0.17



Researchers often make use of scattergrams to visually explore the relationship between dependent and independent variables. Scattergrams help you to inspect patterns and trends in the data, and detect potential problems that can invalidate the use of correlations as a statistical technique. In particular, there are two problems that require special attention:

First, the presence of extremely low and high values in the scattergram, what researchers commonly refer to as outliers. The presence of few outliers in a dataset, especially when the sample size of your data is not large, can lead to misleading results. For example, if you were examining the relationship between the numbers of children living in a household, and the age of the household head and you spotted two extreme values on the Y axis (Number of children within the household): one very high, 999, and the other, negative, -3, what could you conclude? The first value is far away from the rest, which would most probably be the result of a coding error, failing to declare in the statistical package, the missing value code of '999' for households that did not answer that particular question. Can you think of a household having 999 children! The second value, -3, would most probably indicate a data entry error, that is, the person entering the information from the questionnaires to a spreadsheet, typing by mistake -3 instead of 3. It would not make sense having a household with -3 children! Both coding errors and entry errors cause errors of measurement that can affect the interpretation of our results. Unfortunately, the identification of these errors is subjected to nature of the variable itself, and must be made on an individual basis. In that context, scattergrams become useful tools before they are fixed or removed.

It is important emphasize the effect of outliers on the sample mean. If, for example, you have a sample of 10 families reporting the number of children, and two extreme outliers as described above are not treated appropriately, you could end up with a misleading interpretation of families having an average of 101.5 children (see Table 1). After correcting for outliers, the mean value goes down to the reasonable level of 2.4 children per family. Note that the median captures better the number of children in the sampled families than the mean, reflecting the fact that the median is particularly useful when we have small samples and skewed datasets. I discuss these and other measures of central tendency in Box 1 below.

Table 1 The effect of outliers on sample the mean

	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	Mean	Median	SD
Before correcting for outliers	2	3	4	3	2	1	3	1	-3	999	101.5	2.5	315.3
After correcting for outliers	2	3	4	3	2	1	3	1	3		2.4	3.0	

In addition, you could also look at the standard deviation (SD) of your sample (315.35 in our example) and exclude those observations with values falling outside an range around the mean. For instance, if you decide to establish a range of ±2 standard deviations around the mean (101.5), you will notice that H9 and H10 fall outside that range. If you exclude those observations from the sample, the mean will drastically fall from 101.5 to 2.4 children per family, a similar figure to the one obtained after correcting for outliers.[5]

A second potential data issue may emerge from the presence of curvilinear or nonlinear associations. The statistics of correlation are built on the assumption of linear relationships between dependent and independent variables; however, in some cases, we may find a strong relationship between two variables even when your data do not show a straight line, and scattergrams can help you to detect this type of curvilinear trends. For example, researchers often find U-shaped curves when analysing the relationship between health care utilisation and age. This reflects the fact that small children and the elderly make use of health care services more intensively than young adults. On the contrary, researchers often find an inverted U-shaped curve when analysing the relationship between the probability of pregnancy and age of women. The likelihood of pregnancy is zero with small girls but began to rise steeply after they become teenagers to reach the peak in the mid-twenties and then began to gradually fall after reaching the thirties and again towards zero when they get into the late forties and fifties. The value of the Pearson's r would in such cases reflect a low *linear* relationship, when in fact it shows a strong *curvilinear* association.

Box 1. Measures of spread and central tendency

It is often useful to find the midpoint or average of a set of values. The bad news is that there are three types of average in mathematics, known together as 'measures of central tendency:

- The *mean* is what most people are talking about when they refer to 'the average.' You calculate it by adding up all the values for the respondents in your dataset and then dividing them by the number of respondents (excluding those for whom there is a missing answer). How well the mean can represent a set of values therefore depends on how spread out or dispersed the values are around the mean. In addition, it is affected strongly by any extremely large or small numbers, and where this happens the distribution is said to be skewed.
- The median (or middle point) is the middle value of a set, when the values are arranged in order of size. If there is an even number of cases, such as 12, the median falls between the two middle numbers, and the convention is to take the mean of the two. The median can be useful when your data is skewed, and also works better than the mean with small numbers of values.
- The third type of average, is the mode, and is the value that occurs most frequently in a data set. You find it by counting the number of times each value occurs. There can be two or more modes and be used with any kind of data. And like the median, is not affected by extreme values.

'Measures of spread' are logically linked to averages. They describe the way the data is arranged around the midpoint.

[5] There are other statistical techniques available to deal with outliers. A full discussion of these techniques is beyond the scope of this paper. However I suggest to consult the following texts: Barnett and Lewis (1985) Outliers in Statistical Data. New York: John Wiley & Sons, 2d Ed, and Rousseeuw and Leroy (1987) Robust Regression and Outlier Detection. New York: John Wiley & Sons; and Iglewicz and Hoaglin (1993) How to Detect and Handle Outliers, American Society for Quality Control, Milwaukee, WI/

- The simplest measure is very easy to work out. It is called the *range*, and is found by subtracting the lowest value from the highest one. However, as with the mean, the range is affected strongly by extreme values. For example, if only one family has nine children and everyone else has one or two, the range is still nine.
- There are a number of other ways of describing spread, including by the shape of the distribution in relation to the *standard deviation*. The standard deviation shows how much variation exists from the mean. A low standard deviation indicates that your data points tend to be close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values.

Unfortunately there is no clear solution to this problem as there are not equivalent methods to the Pearson's r when analysing nonlinear relationships. For some type of economic data that follows decreasing or decreasing slope curves, (e.g. income, expenditure, revenues, costs, or profits), you could adopt logarithmic functions to 'linearise' those variables that observe large range of values. However, you will need to identify the function that best describes your data, whether it follows a linear or nonlinear trend and then test its goodness of fit by applying hypothesis testing. I return to this issue in Section 3.

2 Establishing the strength of a correlation

A question that often emerged, and which is linked to the second objective of the Pearson's *r*, relates to the issue of how to establish the *strength* of the correlation. In order words, how can you identify whether the Pearson's correlation coefficient shows a strong or weak relationship? There is no strict guidance with regard to the interpretation of the Pearson's *r*, as it may vary depending on the context and characteristics of the phenomenon that you investigate. As I discussed earlier, the Pearson's r can range from high negative, -1, to high positive, +1, correlations. A correlation of r= 0.7 would obviously be stronger than a correlation of r= 0.5. In physics, researchers often expect very high correlations when testing natural phenomena, so they would be surprised to find, for instance, a Pearson's r of 0.90 when testing the linear relationship between volume of a gas and temperature. However, in social and behavioural sciences, it is much harder to established clear parameters for the size effect of correlations. It is often more of a matter of judgement by experienced researchers to establish parameters for weak or strong correlations; however, there are some hints that you may want to consider. Cohen, for example, has provided ambiguous guidelines, suggesting that Pearson's correlation coefficients lying between -0.09 to 0.0, and 0.0 to 0.09 would have zero effect; between -0.23 and -0.10, and 0.10 and 0.23 would have a small effect; between -0.36 and -0.24, and 0.24 and 0.36 would have a medium effect, and between -1.0 and -0.37, and 0.37 and 1.0 would have a large effect.[6]

[6] Cohen, J (1988) Statistical Power Analysis for the Behavioural Sciences. (2nd Ed), Hillsdale, NJ: Erlbaum

One indicator that can help you to determine the strength of a correlation is the square of the correlation coefficient, or coefficient of determination, r^2, which provides an indication of the variance of the dependent variable that may be associated with the observed values of the independent variable. So if you look at the correlation coefficient of -0.17 between the numbers of children living in a household, and the age of the household head, depicted in Figure 3 above, you will find that r^2= $(-0.17)^2$ = 0.029, meaning that only 2.9 percent of variance of children living in a household is associated with the age of the household head. In that context, it would be difficult to argue that this correlation is of any considerable importance. If you go back to the cut-offs provided by Cohen, and look at Table 2, you will see that a correlation coefficient of r = 0.31would be needed to explain a 10 percent variation between two variables.

Table 2 Size effect of correlations

	r	r^2	Explained variance
No effect	0.00	0.00	0%
	0.05	0.00	0%
	0.07	0.00	0%
	0.09	0.01	1%
Small effect	0.10	0.01	1%
	0.15	0.02	2%
	0.20	0.04	4%
	0.23	0.05	5%
Medium effect	0.24	0.06	6%
	0.25	0.06	6%
	0.31	0.10	10%
	0.36	0.13	13%
Large effect	0.37	0.14	14%
	0.50	0.25	25%
	0.75	0.56	56%
	1.00	1.00	100%

Source: Cohen (1988)

3. Establishing the statistical significance of a correlation

In order to establish the statistical significance of a correlation, researchers use *statistical hypothesis testing* to find out whether the value of the test statistic is sufficiently small or sufficiently large to reject the null hypothesis (denoted by H_0) that the values of the dependent and independent variables result purely from chance, and hence are *unrelated*. To do so, you can adopt a *one-tailed test* if you have an expectation about the sign of a correlation, either positive or negative. For example, you may expect *a priori*, a positive correlation between schooling and wages, and reject the null if the test statistic satisfies that expectation. In order words, you can accept the alternative hypothesis (denoted by H1) that the values of the dependent and independent variables are influenced by non-random forces. However, if you have no theory underpinning your expectations about the sign of the correlation coefficient, then you should adopt a two-tailed test.

A full discussion on the properties of these test statistics lies beyond the scope of this book; although we can briefly say that both the one-tailed and two-tailed tests depend on the size of the Pearson's coefficients and the sample size.7 The larger the sample size, the higher the reliability of the correlation coefficients even if the value of the Pearson's correlation coefficients, r, is small in size. This is illustrated using the critical values of the Person's r in Table 3. The first column from left to right shows the degrees of freedom, df, resulting from subtracting the number of variables involved in the correlation (i.e. two, the dependent and independent variables) from the number of observations, N, (your sample size). The degrees of freedom thus, show the number of values that are 'free' to vary in the determination of the statistical significance, and are expressed as df = N-2.

Following the example presented in Figure 3 that shows a Pearson's r (146)= -0.178, you could determine its level of statistical significance by following a few steps: first, decide whether you need a one-tailed or two-tailed test. Second, calculate the degrees of freedom and locate them in Table 3. As 148 households were interviewed for that study, df = 146 (148-2). Third, look at the critical values corresponding to the Pearson's r. To do so, read across the row that corresponds to the df, from left to right until you find a critical value greater than the estimated r. The p-values for that coefficient (i.e. the probability that the null hypothesis is true) will be found on the top of the column that corresponds to the critical value. So, if for example, *r* (146) = -0.17, you will see that its critical value (0.195) has a p<0.05 for a two-tailed test, and p<0.025, for a one-tailed test. In order words, the p-values are telling us that the probability of errors arising from declaring *r* to be determined by non-random factors occurs 5 times in 100 by chance, meaning that we can be 95% confident (under the two-tailed test) that a relationship between the independent and dependent variable do exist.

[7] I remit the interested reader to statistics texts. For a general discussion, see Markin, M (2005) Statistics for the Social Sciences. London: Sage Publications

[8] Correlation coefficients are read in absolute values, irrespective of their sign, which only indicates the direction of the relationship.

Table 3 Critical values for Pearson's correlation coefficient *r*

df =N-2	Level of significance for a one-tailed test			
	.05	.025	.01	.005
	Level of significance for a two-tailed test			
	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
20	.360	.423	.492	.537
30	.296	.349	.409	.449
40	.257	.304	.358	.393
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.302
80	.183	.217	.256	.284
90	.173	.205	.242	.267
100	.164	.195	.230	.254
∞	.073	.087	.103	.114

4. Cross-tabulations

I have discussed in Section 3 some methods to assess the direction, strength and statistical significance of a linear correlation between two interval or ratio scale variables. However, more often than expected, datasets are integrated by discrete (or categorical) variables. These variables are called discrete because they take on discrete (or finite) values in the form of categories. As some of these categories are different in nature, it is useful to classify them into nominal or ordinal variables.

Nominal variables are measured in terms of arbitrary distinctive categorical values that are not quantifiable, and which do not follow a given rank ordering. Examples of nominal variables are: gender (1=Male, 2=Female); ethnicity (e.g. 1=White, 2=Black, 3=Asian, 4=Latino); marital status (1=Married, 2=Unmarried, 3=Divorced, 4=Widow). An important point to make here is that we cannot perform arithmetic operations (e.g. Male ÷ Female = ?) or apply logic functions (e.g. Married > Divorced) on nominal variables as they don't follow equal intervals or ranks.

Researchers often transform ratio scales into a special case of nominal variables, dichotomous variables commonly known as *dummy* variables, in which responses take the values 1 to indicate the presence of a response, or 0, its absence. For example, if in your data you have a continuous ratio scale variable measuring the years of education of garment workers, running from zero to 12 years of schooling, you could separate those with completed secondary education from the rest to find out whether there is any relationship between 'completed levels' (instead of years) of schooling and income. In that case, you would need to generate a dummy with value 1 if worker *i* completed secondary education (12 years of schooling) and zero, otherwise. Researchers can also include in survey questionnaires nominal variables of this type, with a 'yes' or 'no' answer to capture behavioural responses where a choice or decision is involved. For example, when asking interviewees 'are you employed?' 'Is your child attending school?' or 'Did you vote in the last general election?', you would expect a dummy-type nominal response variable with values 1 for a 'yes' response, and 0 for a 'no' response.[9]

Ordinal variables are discrete variables too, but unlike nominal variables, they have non-arbitrary distinctive values with meaningful ranks or orders. For example, the World Bank uses continuous data on countries' gross national income per capita to classify countries by income groups: low income for $995 or less; lower middle income, $996–3,945; upper middle income, $3,946–12,195; and high income, $12,196 or more. Based on that classification you could generate an ordinal variable to rank countries in terms of per capital income: 1=low income, 2= lower middle income, 3= upper middle income and 4=high income.[10] Similarly, demographic and health surveys usually collect subjective information about people's health status. They ask: how would you describe your health?, followed by a set of possible answers: 1=very bad; 2=bad, 3=average, 4= good, and 5= very good. Although we cannot perform arithmetic operations on ordinal variables, because the interval between their scale points do not show equal distances (e.g. low income country + upper middle income country = ?), we can still perform logical operations (e.g. very good health is *preferred* to very bad health).

In addition, categorical variables also allow us to combine two or more variables, for example, a dependent and independent variables, in a way that their frequencies can be *cross-tabulated* in a table. Cross-tabulations show the number of cases that occur jointly in each combination of categories, allowing us to identify associations (or correlations) between

[9] Survey questionnaires often use the number 2, instead of 0, to capture a 'no' response

[10] World Bank's country classifications, available at: http://data.worldbank.org/about/country-classifications

the cross-tabulated categories. In other words, they help us to answers the following questions: Is there any relationship between two discrete variables? And if there is, how strong is it? For example, you may want to know whether there is a statistically significant correlation between households headed by women and the area of residence. The results from the cross-tabulation in Table 4 show a higher percentage of households headed by women in rural areas than in urban areas. You will need to back to your theories or think about the factors that may explain that difference before you investigate further into the data. Remember that a correlation that might be expected but is *not* found may be important to report as well. For example, most people would expect gender to affect school examination performance in different ways in relation to each subject – or that girls (or indeed boys) would simply do worse at everything. If girls and boys do equally well, this could be important.

When we set up cross-tabulations, we usually have an idea about what the relationship between the data will mean. But there is always a possibility that any relationship is simply the result of chance, that is, it would not recur if you did the research again. This is why *tests of significance* are used: to help us see how reliable our findings are. But unlike the use of one-tailed or two-tailed tests, as done in Section 3, you will need now to use the Chi-square test to establish the statistically significant association between two categorical variables. The Chi-square statistic works in a similar fashion to the one-tailed and two-tailed tests, by allowing you to determine whether you should reject (or accept) the null hypothesis, H_0, that the values of two categorical variables occur entirely by chance. The Chi-square statistic in the cross-tabulation presented in Table 4 shows a value of 7.28, which is statistically significant at 0.05 levels, meaning that we can be at least 95% certain that there is an association between households headed by women and the area of residence.

A problem with the Chi-square statistic is that it becomes an unreliable indicator for measuring the strength of a correlation, especially if you have small samples.[11] That problem stems from the dependency of the Chi-square statistic upon the sample size and the number of cells in the tabulation table. For that reason, researchers often use the Cramér's V statistic to measure the strength of a correlation. The Cramér's V statistic ranges from 0 to 1, with 0 representing no correlation and 1, a perfect correlation. In Table 4, the Cramér's V statistic shows a value of -0.058, indicating a negative relationship between households headed by women and urban areas. Overall, the results show that although the association between

[11] In a 2 by 2 tabulation table, the Chi-square will test the underlying probabilities in each cell; and when the frequencies fall below 5 observations, those probabilities cannot be estimated with sufficient precision. That is why the Chi-square statistic should not be estimated with very small samples.

these two categorical variables is statistically significant, the strength of the association (measured by the Cramér's V coefficient) is very weak.

Table 4 Correlations between households headed by women and the area of residence

Sex of household head?	Area of residence			
	Rural = 0		Urban = 1	
	Frequencies	%	Frequencies	%
Men =0	631	(53.84)	578	(59.65)
Women =1	541	(46.16)	391	(40.35)
Total	1172	(100.00)	969	(100.00)

Chi-square statistic = 7.28, $p < 0.05$
Cramér's V = -0.0583

So unless you are working with a large, reasonably representative sample, it is unlikely that any correlation you identify will be statistically significant. It is bad practice to publish correlational data unless it reaches a reasonable level of significance. If you do so, you can seem to be misleading readers. There are statistical software packages that run cross-tabulations easily (see Box 2 below). Using a computer is a much easier way of working out correlations for large datasets.

Box 2. Popular Data analysis software

A number of options are available in terms of computer software for quantitative analysis.

Specialist statistical software

These are specifically designed for this task and will therefore give you the most help with your project. They are set up to expect survey data and will easily carry out all the calculations you need (and many more):

- SPSS (Statistical Package for the Social Sciences) allows you to analyse large quantities of data. In a spreadsheet format, each column represents a variable and each row represents a case. You can put in labels that keep information about the meaning of the data close to the numbers involved. SPSS produces high quality tables and graphs and the latest versions are quite user-friendly. It is worth buying if you expect to do more than one substantial survey (perhaps 60 or more respondents) with a quantitative element.[12]

- STATA is a popular alternative to SPSS. It has equivalent functionality and is generally as widely used in academic institutions around the world. It features a relatively straightforward, user-friendly interface.[13]

- R is a free and powerful software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. R provides a wide variety of statistical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc. [14]

- PSPP is an open source substitute for commercial statistics programs – particularly SPSS, whose interface intentionally mimics. It offers a strong set of basic capabilities, including frequencies, cross-tabs comparison of means, linear regression and others.[15] In recent years, PSPP has proven popular in the South.[16]

[12] For guidance on how to use SPSS, see, for example, Argyrous (2005). SPSS is available from SPSS UK Ltd., St. Andrews House, West St., Woking, Surrey, GU21 1EB. Tel: +44 (0)1483 71920, IBM worldwide and through sales offices located in a number of countries, listed at www.spss.com/worldwide/.

[13] The University of California, Los Angeles maintains a useful website for guidance and links on using STATA at www.ats.ucla.edu/stat/stata/. STATA is available through a number of global distributors, listed at www.stata.com/worldwide/.

[14] R can be download for free at http://www.r-project.org/

[15] PSPP can be downloaded for free at www.gnu.org/software/pspp/.

A very important note of caution: You must remember that you cannot be sure, even if you find a correlation that is statistically significant, which way causation goes (what is the cause and what is the effect). Just because variable A is associated with variable B that does not mean that changes in the former has caused changes in the latter. For example, if more charges are brought in areas where there are more police officers, is a higher level of crime the explanation or are more officers simply able to prosecute more cases? Other factors may be acting on the related variables and affect their behaviours. So, while correlations can establish associations between two (or more) variables, they cannot tell us anything about *causality.* In order to establish causality, you will need to employ regression analysis.

5. Regression analysis

As with correlation analysis, regression analysis is concerned with the study of the relationship of two (or more variables), but in addition, it helps you to test the proposition that that one (or more) independent variables are the cause, at least partly, of the way the dependent variable behaves.[17] For example, you may be interested in analysing the extent to which the levels of formal education explain workers' income. You may also be interested in exploring how much the price of a staple, say rice, affects its quantity demanded, or how much the demand for microcredits is affected by the rate of interest. Regression analysis will help you to test empirically your propositions, say, that the interest rate on loans has a linear (or direct) effect on the demand for microcredits.

Formally, a regression equation is expressed as $Y_i = \alpha + \beta X_i + u_i$, where Y represents the dependent variable, and X, the independent variable. The subscript i refers to the ith population subgroup under analysis, for instance, individuals, households, villages, firms, countries, etc. If you have a multiple variable equation, with more than one independent variable, you can simply use subscripts to indicate each explanatory variable: X_1, $X_2, \ldots, X_n$.

The Greek letters α and β are known as the *intercept* and the *slope* coefficient, respectively. The intercept captures the *conditional* mean value of Y when X has zero value. For example, if you were analysing the effect of rice prices on household's demand for that staple, the

[16] The South African Statistical Association, for instance, presented a 2006 conference which included an analysis of how PSPP can be used as a free replacement to SPSS.

[17] In this section I discuss the two-variable regression model, and refer the reader to more specialised texts for the discussion of multiple linear regression analysis. For a basic introduction, see Gujarati, D (2003) Basic Econometrics. New York: McGraw Hill. Fourth Edition

intercept would tell you how much rice would be demanded by an average household if its price were zero. The slope coefficient provides the most important information in regression analysis, as it measures the *rate of change* in the mean value of Y relative to one unit change in the value of X. In order works, it gives us information about the way the dependent variable is expected to behave conditional upon given values of the independent variable(s).

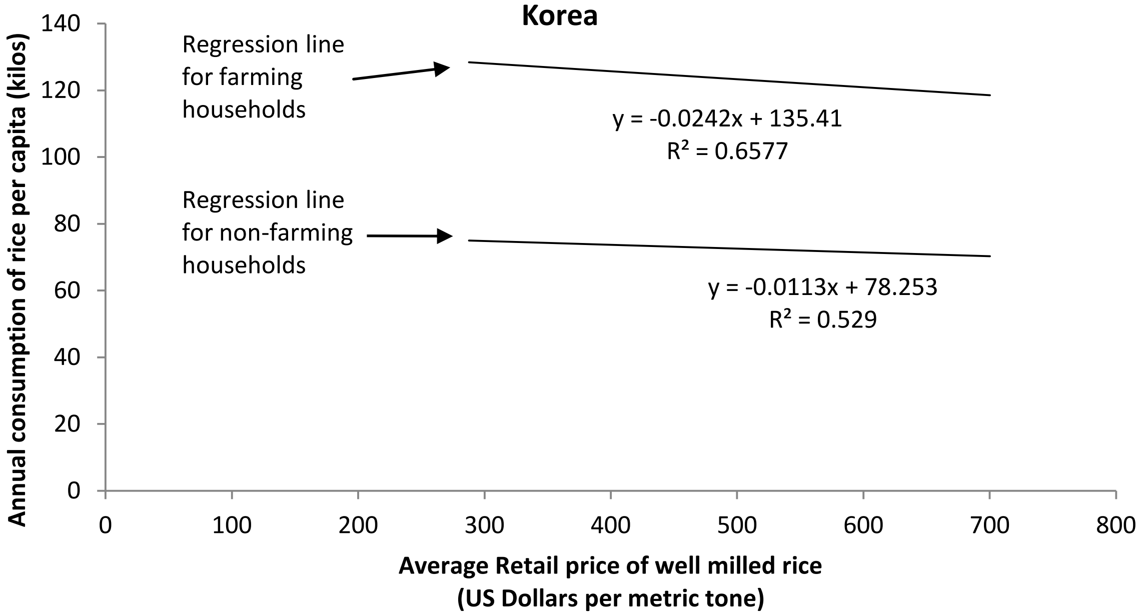
The letter u captures the *residual* or *error term,* a random variable determined by unknown factors that affect the value of the dependent variable. For example, in South Korea, data suggests that at the average retail price of US$ 287.8 per metric ton, the *mean* annual consumption of well milled rice per capita is 77 kilos.[18] If there are individuals willing to consume 80 kilos of rice at the same price, so their quantity demanded exceeds the mean value of consumption (77 kilos) we can say the corresponding error term for these individuals is equal to 3 kilos or rice, implying that there are factors other than price that are driving up their demand for rice. The error term also captures measurement errors arising from the presence of extreme outliers in datasets.

Now, before moving on to the interpretation of a regression equation, it is important to point out that regression analysis can only establish causality when there is an underlying theory explaining the relationship. In Economics, the *law of demand* provides the analytical foundations to predict a *negative* effect of a rice price increase on the demand of that staple, *ceteris paribus*, i.e. holding constant all other factors that may affect demand, e.g. household income; the price of substitutes, consumers' taste, etc. In this sense, regression analysis does not imply causality *per se*; but rather, help you to test the underlying theory empirically, a process that methodologists called deductive analysis.

After having that information at hand, we can now run a linear regression on the annual consumption of rice per capita in South Korea for a given range of prices ranging from US $287.81 per metric ton to US$ 520.55 in the 2005-2010 periods. This is illustrated in Figure 4. The scattergram shows that at the price of US$ 287.8, the *average* annual consumption of rice per capita among farming households is 130.8 kilos vis-à-vis 77 kilos among non-farming households, and as prices increase, the quantity demanded goes down.

[18] This information is based on South Korea's Consumer Price Surveys and the Household Income and Expenditure Surveys, which are available on the Korea's National Statistics Office(Kostat) website: http://kostat.go.kr/portal/english/surveyOutlines/4/1/index.static

Figure 4. Regression line of the price and consumption of rice in South Korea



Source: Author' calculations with data from Korea's National Statistics Office

The regression line depicted in the scattergram reveals an inverse linear relationship between price and quantity demanded, meaning that the average annual consumption of rice per capita decreases with a price increase. But how large is that effect? In the case of farming households, the estimated regression equation is $Y= 135.41 - 0.024X_i$, which means that, other things held constant, for every dollar increase in the price of well milled rice, the *average* annual consumption per capital goes down by only 0.024 kilos (or 2.4 grams). In the case of non-farming households, the slope coefficient shows an even smaller effect (0.0113 kilos). The results reveal a very weak responsiveness of both farming and non-farming households to rice price increases, what Economists call inelastic *price elastic of demand*.

Note that the coefficient of determination, r^2, indicates that about 66% of the total variation in rice consumption among farming households, and 53% among non-farming households can be explained by the regression model. The intercepts α also show interesting information: Farming households would consume more rice every year (135.4 kilos per capita) relative to non-farming households (78.25 kilos per capita) if the prices were put down to zero. Given such difference in behaviour, you could hypothesise about the possible reasons that explain the results based on theory and the existing literature on the subject before considering introducing other explanatory variables to test your propositions empirically.

A large number of regression techniques have been developed to determine causality. These techniques are used according to the research questions of the investigation, the type of research design, and type of available data.[19] In particular, you will need to be aware of a major issue that arises from the analysis of the cause-effect relationship between two or more variables: the problem of *confounding*.

6. The problem of confounding

Confounding factors are variables that are causally correlated with the dependent and/or independent variables, and therefore can influence the observed relationship. For example, if you were interested in establishing the effect of a microcredit programme on women's income, you would need to account for those observed and unobserved factors that can determine income. For example, *observable* factors can be associated with the level of women's education, the number of children they have, or the physical infrastructure in the localities where they live. *Unobservable* factors can be attributed to individual entrepreneurial abilities and attitudes towards risks that can potentially affect women's capacity to generate income.

If women's participation in the microcredit programme is based on a non-random selection process, some the unobserved factors could be determining selection. For example, if entrepreneurial and risk-loving women actively *self-select* to participate in the microcredit programme, they are more likely to achieve better outcomes, in terms of income, than women that do not actively seek to participate in the programme because of their risk-averse behaviour. Self-selection would mean thus that more able and entrepreneurial women that are also more likely to generate higher income would be more likely to join the microcredit programme. Under these circumstances, the comparison of the *self-selected* women with *eligible* women that did not seek participation would lead to spurious results.

In addition, it is often the case that programme managers choose certain communities or groups under the expectation that these groups would benefit more from the intervention. Many microcredit programmes, for instance, embrace the goal of poverty reduction, and as result, their managers often operate in deprived locations where the poor live. But the non-random process of programme placement can generate cofounding if there are omitted

[19] A full discussion of these techniques is beyond the scope of this book; however, I strongly recommend to discuss your research needs with quantitative specialists or consult the following texts: Wooldridge. J (2002) Econometric Analysis of Cross Section and Panel Data Cambridge Massachusetts: The MIT Press, and Greene, W. (2011) Econometric Analysis, New Jersey: Prentice Hall, 7th Ed.

factors that are correlated with the dependent (or outcome) variables. These factors can be associated with local characteristics (e.g. public infrastructure, proximity to local markets, public security, etc.) that somehow can determine women's income. So, if you were thinking to carry out a quantitative study to establish causation (in the development jargon this is called *impact assessment* or *impact evaluation*), you would need to keep in mind the problems of self-selection and non-random programme placement. It is in that context that quantitative research design becomes critical.

7. Quantitative research designs

There are different types of research design used to establish causality, mainly through the identification of the *counterfactual*. The concept of counterfactual refers to question of 'what would have happened to programme beneficiaries in the absence of the intervention?' The identification of the counterfactual can be done through experimental or quasi-experimental research designs that allow researchers to attribute (or not attribute) observed changes in outcomes to a programme or policy intervention. In the remaining of this paper, I briefly discuss these design protocols.

7.1 Experimental research designs

Experimental research designs make use of randomisation to select both the treatment group, i.e. the group to benefit from a policy intervention, and the control group, the comparison group, or placebo, which although probabilistically equivalent, does not benefit from policy. The most reliable research design in terms of internal validity is the *pretest-posttest randomized experimental design*. With sufficiently large samples, the pretest-posttest randomized experimental design allows you to assess, at the *posttest* stage, the differences in mean outcomes between the treatment and control groups, and if such differences are statistically significant, they can be attributed confidently to policy.

The *pretest-posttest randomized experimental design* is illustrated in Figure 5. The letter R denotes a random assignment to treatment; the letter X represents the explanatory or impact variable that captures the effect of policy whereas the letter O represents the dependent or outcome variables measured before the intervention (O_1) i.e. during the pretest stage, and after the intervention (O_2), i.e. during the posttest stage. The pretest measure is of particular interest, as it allows to control for the effect of covariates, i.e. variables that are likely to affect the outcome of interest. As I discuss below, the pretest measure is obtained through a baseline survey that collects information on covariates, particularly the pre-treatment value of the outcome. This will enable you to control for the variance observed in outcomes at the initial stage and thus improve the accuracy of your results.

Figure 5. The pretest-posttest randomized experiment

		Pretest measure	Treatment	Posttest measure
Treatment group	R	O_1	X	O_2
Control group	R	O_1		O_2

To illustrate this let us assume, following the previous example, that you are interested in operationalising a pretest-posttest randomized experimental design to assess the impact of a microcredit programme on women's income. What would you need to do? Here are some guidelines: First, use recent census data to list your 'subjects' in the sample population (eligible women and their families). Second, use randomization to draw a statistically representative sample of the population while dividing the sample randomly into two groups: women selected to receive a microcredit (your treatment group) and eligible women that will not receive the microcredit, at least for the time being (your control group). You could divide the sample into more than two groups if you were interested in testing the effect of different programme features. Third, collect a *baseline survey* on the treatment and control groups *before* microcredits are given to women.

Through the baseline survey, you will be able to collect *pretest* data on women's income (your dependent variable or outcome of interest) and other key independent or explanatory variables. The pretest measures will also allow you to remove the effect of covariates after women receive microcredits, during the posttest measure stage. Fourth, after the microcredits had been distributed wait for a reasonable period to allow for the effects of microcredits (if any) to be observed. Fifth, collect the posttest measures through the endline survey, and finally, apply regression techniques to establish causality, i.e. whether or not microcredit has an effect on women's income.[20]

It is critical to ensure that your sample is representative of the population. You could follow different sampling techniques to strengthen the representativeness of the sample. For instance, adopt a *stratified* and by *cluster* sampling framework. Assuming the microcredit programme is expected to have full coverage of a province, you could randomly select a

[20] For a discussion on regression techniques in experimental studies, see Duflo, Glennerster and Kremer (2007) Using Randomisation in Development Economics Research: A Toolkit. CERP Discussion Paper 6059, London, and Khandker, Koolwal and Samad (2010) Handbook on Impact Evaluation. Quantitative Methods and Practices. Washington DC: The World Bank

sample of districts, wards, and villages within that province while adopting different geographical strata, e.g. urban and rural settings, northern, southern, eastern and western areas, etc. You could also use random stratification to ensure a proportional representation of population subgroups such as ethnic and religious groups, etc. These procedures will allow you to reduce the required sample size and the budget requirements without losing significant statistical power.

A simpler version of a randomised experiment is known as *posttest-only randomised experiment*. Under this design, researchers assume that both the treatment and control groups are equivalent by the virtue of randomisation, so the outcomes of interest are measured and compare only after the treatment group had received the benefit. This is illustrated in Figure 6. The letter R denotes a random assignment to treatment; the letter X represents the explanatory or impact variable that captures the effect of policy, whereas the letter O represents the outcome of interest.

Figure 6. The posttest-only randomised experiment

		Treatment	Posttest measure
Treatment group	R	X	O
Control group	R		O

After the treatment group receives treatment, i.e. in the posttest stage, researchers gather data through survey questionnaires (or other data collection instruments) in order to test for any statistical difference between the mean outcomes of the treatment and control groups.

To illustrate this experimental design, let us assume that you are interested operationalising a randomised experiment to assess the effect of political awareness in an upcoming general election. In order to do so, you have designed a leaflet with information about the importance of exercising the political rights. What would you need to do next? First, you will need to collect a list of households living in your constituency. Second, divide the households into treatment and control by random assignment. Third, distribute the leaflet among the treatment households just before the generation election. Fourth, after the Election Day, collect data on the outcomes of interest, for example, information about whether or not eligible voters did actually vote, and analyse the difference in sample means. If you data show a larger voter turnout in the treatment group, vis-à-vis the control group, you could conclude that political awareness through leaflets is an effective tool to tackle disenchantment and political indifference in election times.

The posttest-only randomized experimental design is an appropriate approach to establishing causality when it is unfeasible to collect pretest measures, i.e. data before the policy intervention. It is also convenient in terms of budgetary and time requirements, relative to the pretest-posttest randomized experimental design; however, it becomes inappropriate in contexts where determining the extent of change in the outcome variables, or measuring the short-medium term effects from policy is the main objective of the study.[21]

Over the last decade, randomised experiments have been designed for causal inference of a growing number of pilot projects, programmes and development policies. These studies range from interventions in the areas of education[22], health care[23], social protection[24], microcredit [25] agriculture[26] and other development fields, highlighting their applicability and importance in quantitative development research.[27] However, even randomised experiments can be contaminated by spillover effects, occurring when members of the treatment and control group receive 'treatment from other policies that affect outcomes. In addition, there are areas where randomized experiments are logistically or ethically unfeasible.

For instance, it would be ethically unacceptable to use randomisation to separate food aid recipients from non-recipients in contexts of humanitarian crisis.[28] Randomised experiments offer limited insights into questions that are concerned with long-term effects of programmes, or generalised impacts at national scale. In those cases, carefully design quasi-experimental

[21] There are other modalities in the way randomized experiments have been designed. A detailed discussion on these designs can be found in: Campbell, D and Stanley, J (1966) Experimental and Quasi-Experimental Designs for Research, Boston: Houghton Mifflin Company; Shadish, W., Cook, T and Campbell, D (2001) Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Belmont, California: Wadsworth Publishing.

[22] Glewwe, P., M. Kremer, S. Moulin, and E. Zitzewitz (2004): "Retrospective Vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya," Journal of Development Economics, 74, 251 268.

[23] Cohen, J. and Dupas, P. (2010) "Free distribution or Cost-sharing? Evidence from a Randomized Malaria Prevention Experiment", Quaterly Journal of Economics, 15(1)

[24] Schultz, T. P. (2004): "School subsidies for the poor: Evaluating the Mexican PROGRESA poverty program," Journal of Development Economics, 74(1), 199–250; and Hoddinott, J. & Skoufias, E. (2004) The impact of PROGRESA on food consumption. Economic Development and Cultural Change, 53, 37-61.

[25] Karlan, D. and Zinman, J (2011) Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation, Science, 332(6035), pp. 1278-1284; Karlan, D. & Zinman, J. (2010) Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts. Review of Financial Studies, 23, 433-464; Banerjee, A., Duflo, E., Glennerster, R. & Kinnan, C. (2009) The Miracle of Microfinance? Evidence from a Randomized Evaluation. Cambridge, MA, MIT Department of Economics and Abdul Latif Jameel Poverty Action Lab.

[26] Duflo, E; Kremer, M. and Robinson, J.(forthcoming): Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya. American Economic Review

[27] For a review see Harrison, G., and J. A. List (2004): "Field Experiments," Journal of Economic Literature, XLII, 1013–1059.

[28] See Bamberger, M, Rao, V., and Woolcock, M. (2010) Using Mixed Methods in Monitoring and Evaluation. Experiences from International Development. Policy Research Working Paper 5245, World Bank, and Bamberger, M. and White, H. (2007) Using Strong Evaluation Designs in Developing Countries: Experience and Challenges, Journal of Multi-Disciplinary Evaluation, Volume 4, Number 8, 58-73

methodologies, along with well grounded theory, can be important alternatives for the identification of the counterfactual to establish the causal inference of a public intervention.[29] In the following section, I discuss some of these quasi-experimental research designs.

7.2 Quasi-experimental research designs

Quasi-experimental studies are used when the identification of the counterfactual cannot be achieved through randomisation. These studies rely on complex econometric methods that aim at tackling cofounding, and can be grouped into two groups: *non-equivalent pretest-posttest groups design and non-equivalent posttest-only groups design.* While the objective of this section is not to review these methods in detail, I briefly describe their main features.

Non-equivalent pretest-posttest groups designs are largely used in development research to resemble the pretest-posttest randomized experiments discussed above, but with the only difference that the counterfactual is not establish through random assignment. This is illustrated in Figure 7, where the letter N denotes non-random assignment to treatment. Once the pretest and posttest measures are collected through baseline and endline surveys, researchers will need to apply econometric methods to deal with cofounding.

Figure 7. The pretest-posttest randomized experiment

		Pretest measure	**Treatment**	**Posttest measure**
Treatment group	N	O_1	X	O_2
Control group	N	O_1		O_2

Difference-in-difference, double difference or simply DID methods use pretest differences in outcomes between the treatment and control group to control for these differences before and after a policy intervention. Thus, the difference-in-difference estimates calculate the observed mean outcomes between the two groups under the assumption that both treatment and control groups would have followed a parallel trend overtime had the policy not been

[29] See Ravallion (2009) Should Randomistas Rule? Economists' Voice. February. Available at www.bepress.com/ev, and Deaton (2009) Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development. NBER Working Paper No. w14690

introduced. Variations of the difference-in-difference method can be found in an increasingly number of studies in development research.[30]

Fixed-effects estimates are a special case, largely used when there are more than one treatment group or more than two observed periods. Note that the robustness of your results will depend on how strong your assumption is with regard to the parallel trend between the two groups. In order words, how credible is that the observed factors that cause cofounding are time-invariant, i.e. fixed overtime.[31]

In contexts where pretest measures are not available, the counterfactual can be generated through complex econometric methods that rely on *observed* characteristics. These techniques belong to n*on-equivalent pottest only comparison-group* studies (see Figure 8.) and include *propensity score matching, regression discontinuity* and *instrumental variables*.

Figure 8. The posttest-only randomised experiment

		Treatment	**Posttest measure**
Treatment group	N	X	O
Control group	N		O

Among posttest-only quasi-experimental designs, *propensity score matching* (PSM) are regarded as the best possible alternative. PSM generate the counterfactual by modelling (or estimating) the probability of treatment on the basis of observed characteristics that are not supposed to be affected by the presence of the policy in question. The estimated probability (or propensity score) is then used to match an equivalent control group.[32] The closer the propensity score, the better the match between the two groups. Once the treatment and the control group are identified, you will be able to measure the mean difference in outcomes between the groups even in the absence of randomisation. PSM have been widely used in

[30] Two classic examples are found in Gruber, J. and Poterba, J (2004) “Tax Incentives and the Decision to Purchase Health Insurance: Evidence from the Self-Employed,” Quarterly Journal of Economics, 109: pp:701-34; and Card, D. and Krueger, A (1994) “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania,” American Economic Review, 84, pp:772-93

[31] For a discussion on the limitations of the difference-in-difference method see Bertrand, M., Duflo. E and Mullainathan, S (2004) “How Much Should we Trust Differences-in-Differences Estimates?” Quarterly Journal of Economics, 119. pp: 249-76.

[32] For a discussion on matching techniques, see Hirano, K, Imbens, G. and Ridder, G. (2003) “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score”, Econometrica, 71(4): 1161–89.

development research[33]; however, as matching techniques rely on the assumption that mean differences in outcomes just appear from differences in observables, they require large datasets, usually nationally representative household surveys, to oversample eligible programme beneficiaries. And that condition imposes analytical constraints in contexts where large and reliable datasets are not accessible.

Sometimes the rule for programme eligibility can be exploited to generate a counterfactual. In those cases, the eligibility rule acts as threshold or *cut-off* to separate the treatment from the control group. For example, the retirement age of 65 acts as a cut-off to distinguish those who are eligible to receive a state pension from those who are not. You could exploit that well defined rule and measure the effect of that pension scheme on health outcomes through *regression discontinuity* methods. Regression discontinuity allows you to account for observed and unobserved heterogeneity within and between groups, and compare the mean outcomes of those who are just above the cut-off with those who are below. For example, you could compare the mean health outcomes of people aged 64 (just before retiring) with the mean health outcomes of people aged 66 (just after retiring).[34]

A variation in the regression discontinuity method is the *pipeline approach*, which is used when programme managers decide to postpone the delivery of benefits among a group of *eligible* recipients (the control group) and compare their mean outcomes with those who received the benefit without delay (the treatment group). Pipeline studies thus exploit the variation in the timing of programme implementation and compare the outcomes of the two groups under the assumption that they are similar in terms of observed characteristics. Concerns about the reliability of pipeline studies emerge in contexts where the eligibility rule is not fully or consistently followed. For example, if politicians have incentives to benefit certain communities or groups first, and these groups are better-off than those who have been left to receive benefits later, that would cause selection bias and then spurious results in an impact analysis. To reduce these problems, you could combine regression discontinuity with matching methods to strengthen the internal validity of the study.

[33] See, e.g. Jalan, J., and Ravallion, M. (2003). “Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching.” Journal of Business and Economic Statistics 21 (1): 19–30; Godtland, E., Sadoulet, E., de Janvry, A., Murgai, R., and Ortiz, O. (2004) “The Impact of Farmer-Field-Schools on Knowledge and Productivity: A Study of Potato Farmers in the Peruvian Andes.” Economic Development and Cultural Change 52 (1): 129–58, and Heckman, J., Ichimura, H., and Todd, P. (1997) “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme.” Review of Economic Studies 64 (4): 605–54.

[34] For an illustration, see Duflo, E. (2003) “Grandmothers and Granddaughters: Old Age Pension and Intra-household Allocation in South Africa” World Bank Economic Review 17 (1): 1–26.

Finally, instrumental variables (IV) are econometric methods that allow for endogeneity in the process of causal inference. If you were going to use instrumental variables in your analysis, you would need to follow two-stages. First, to model the probability of participation within your sample population, using one or more exogenous variables (known as instrumental variables or instruments). These instruments must be *correlated* with programme participation, but *uncorrelated* with the outcomes of interest. In the second stage, you will need to estimate the effect of programme participation while including in the impact equation the predicted values from the first stage. Instrumental variables help you to predict the effect of a policy in contexts where the delivery of benefits is made on a non-random basis.

A note of caution: instruments should be selected carefully. Instruments that are found to be correlated with unobserved characteristics that affect outcomes can exacerbate the bias. Fortunately, there are a number of techniques to test for the robustness of your instruments. You should apply these techniques or consult specialists in econometric methods to ensure that your results are reliable.[35] That is why it is critical to understand the factors underlying programme participation. For example, if you were asked to assess the impact of a microcredit programme, and its manager follows the rule of lending to women who live within a given radius around branch, you could use that exogenous rule as instrument. To do so, you would need to collect information on the *distance* between the branch and the residence of active borrowers (your treatment group) and accepted applicants (your control group). Then estimate the correlation between distance (your instrument) and the probability of participation. The predicted values of that correlation would then allow you to control for cofounding when comparing the mean outcomes of the two groups.[36]

Concluding remarks

In this paper, I have briefly covered quantitative research methods that are widely used in development research. Some of these methods are complex and require high level of expertise in statistics, econometrics and research designs. If you are planning to undertake a study but don’t have a background in Economics, Statistics or Behavioural Sciences, it would be advisable to consult an expert in quantitative research methods, but always

[35] There are some textbooks in applied econometrics that can be illustrative, e.g. Wooldridge. J (2002) Econometric Analysis of Cross Section and Panel Data Cambridge Massachusetts: The MIT Press, and Greene, W. (2011) Econometric Analysis, New Jersey: Prentice Hall, 7th Ed.

[36] This approach was actually used in an empirical study conducted in Mexico. For a full detail of the methodology, see: Niño-Zarazúa, M. (2009) Microcredit and Poverty in Mexico: An Impact Assessment in Urban Markets, Munich, Germany, VDM Verlag.

keeping in mind the issues that I have discussed throughout this paper. In particular, remember that you research objectives will guide your methodological approach. Whether or not randomisation is feasible, in terms of context, and/or ethical, budgetary and time considerations, data collection will determine the selection of your methodology. If randomisation is not an option, the process of generating the counterfactual becomes an *art.* You will need to ensure that cofounding, i.e. the problems of self-selection and non-random programme placement, are appropriately addressed through quasi-experimental techniques. A well designed study will provide you with reliable information to make informed decisions about policy.