

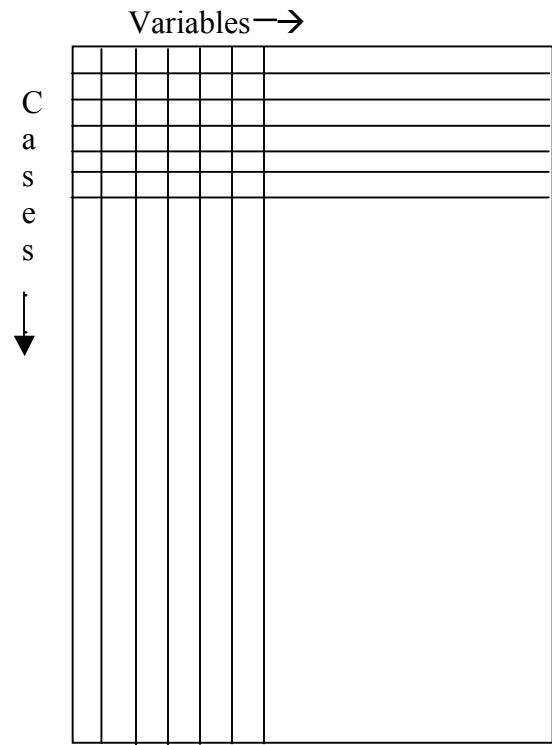
Vocabulary of data analysis

Data – data matrix – dataset

Cases – variables – codebook – system file

I should not have to talk much about these since these concepts are covered in the chapter you should have read before the class.

In a data matrix each variable occupies a column, with a different value for each case. Each case occupies a row, with values for each variable specific to that case. If the cases are respondents to a sample survey, then the columns will contain values for the ages of each person, the gender, educational attainment, etc.



The nature of the information in each column, what variable, how measured, etc. is contained in a codebook – a sort of dictionary or crib sheet that is needed in order to make sense of the data matrix. The data matrix and codebook taken together are known as a ‘data set’ (often written as one word). When we use a statistical package such as SPSS or STATA the codebook is an integral part of a so-called ‘system file’ (a file with the suffix ‘.sps’ for spss system files and ‘.dta’ for stata system files). In such system files the format of the data is proprietary and we are dependent on the manufacturers of the statistical packages to continue to support the data formats concerned. Statistical packages can often read the system files belonging to other statistical packages, however, and there is a utility program called ‘Stat Transfer’ that can translate data from one system format to another. In the last analysis, data formatted as a spreadsheet constitutes a common denominator at least for the data matrix in a system file, even if not for information about variables.

How does this look in STATA?

Double click on the file you downloaded from my website. If you have STATA installed on your computer, it will open, showing a screen that looks something like...

The menus are pretty standard Windows-style menus. You can find out what the various buttons do by hovering your cursor over each of them in turn. I have shown only five buttons...

Log begin/close/suspend/resume

Clicking this button is the first thing you should do at the start of a session. In the standard file browsing window that opens, change file type to .log

“Eye in the sky” – this is the “help” button. It brings up a window in which you can browse STATA’s extensive documentation. In the command window type ‘help summarize’

Spreadsheet icon – this brings up the data editor. In this window you can scroll around your whole data matrix and make changes to individual cells.

Same but with magnifying glass superimposed – data browser. Here you can scroll around the data but do not risk accidentally changing anything.

X in a circle – stop. Interrupts whatever STATA is doing (very useful if you inadvertently asked for more output than you bargained for or more computation than you bargained for).

Menus, buttons and how commands are constructed

Strategies of STATA usage: ALWAYS LOG YOUR SESSION – why...

The basic syntax of a STATA command

[prefix:] command [varlist] [=expression] [if...] [in...] [weight] [using...], options
[using filename]

The prefix allows you to do things like having the command repeated for each of the values of a variable - eg ‘by country:’ to have a command repeated for each country in the file.

The command itself is not in brackets because it is the one thing that is required.

The variable list can be created by clicking names of variables in the variable window there are various facilities for abbreviating the names of variables. Use the eye in the sky to tell you about variable lists.

The easiest way to see how commands are constructed is to use the menus in order to create a few simple ones

Choose Statistics → summaries → summary and descriptive → summary statistics

Then click on the names of the variables you want summarized (I suggest ‘elect’ ‘turnout’ and ‘timeleft’ – these are all numeric variables for which a summary makes some sense)

Click ‘OK’ to have the command performed. You will see the text of the command printed in the Results window before the output from the command. You can retrieve the command into your command line by pressing the ‘page up’ button on your keyboard. This brings the command into your command line where you can edit it (maybe changing the list of variables) which would be much easier than creating the command again from scratch using the menus.

Change the variable list, eg by double clicking on ‘yrsleft’ and replacing it with ‘natturn’

The results window contains the mean and standard deviation of each variable, as well as the minimum, maximum, and the number of observations (cases) over which these statistics were calculated.

What is a mean? In statistical notation the mean is indicated by putting a bar across the top of the letter standing for the variable whose mean it is. So the mean of X is

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N}$$

Here, i stands for “individual” and the big Greek sigma stands for “sum”.

The i=1 and N at bottom and top of the sigma tells us we are summing across all individuals from 1 to N (the number of individuals). The N beneath the long bar tells us that the whole equation is divided by the same N umber of individuals as we summed X over in the first place.

So the mean is just the average, but notation is important.

The letter sigma and how we use it represents one of about five pieces of mathematical notation that you will need to master.

Standard deviation. Consider the following table of numbers. Each column has the same mean, but the extent to which that mean accurately portrays the set of numbers in the column is different...

Individual	X	Y	Z
1	5	4	2
2	5	5	5
2	5	6	8
Mean =	5	5	5

With large Ns we need a measure that tells us how much the values are dispersed – the range is a measure of dispersion, but does not take account of how values are dispersed between the extremes.

The standard deviation is represented by the roman letter S, for Standard deviation, and is generally subscripted with the letter naming the variable whose standard deviation we are talking about. So

$$S_x = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N}}$$

is the standard deviation of X.

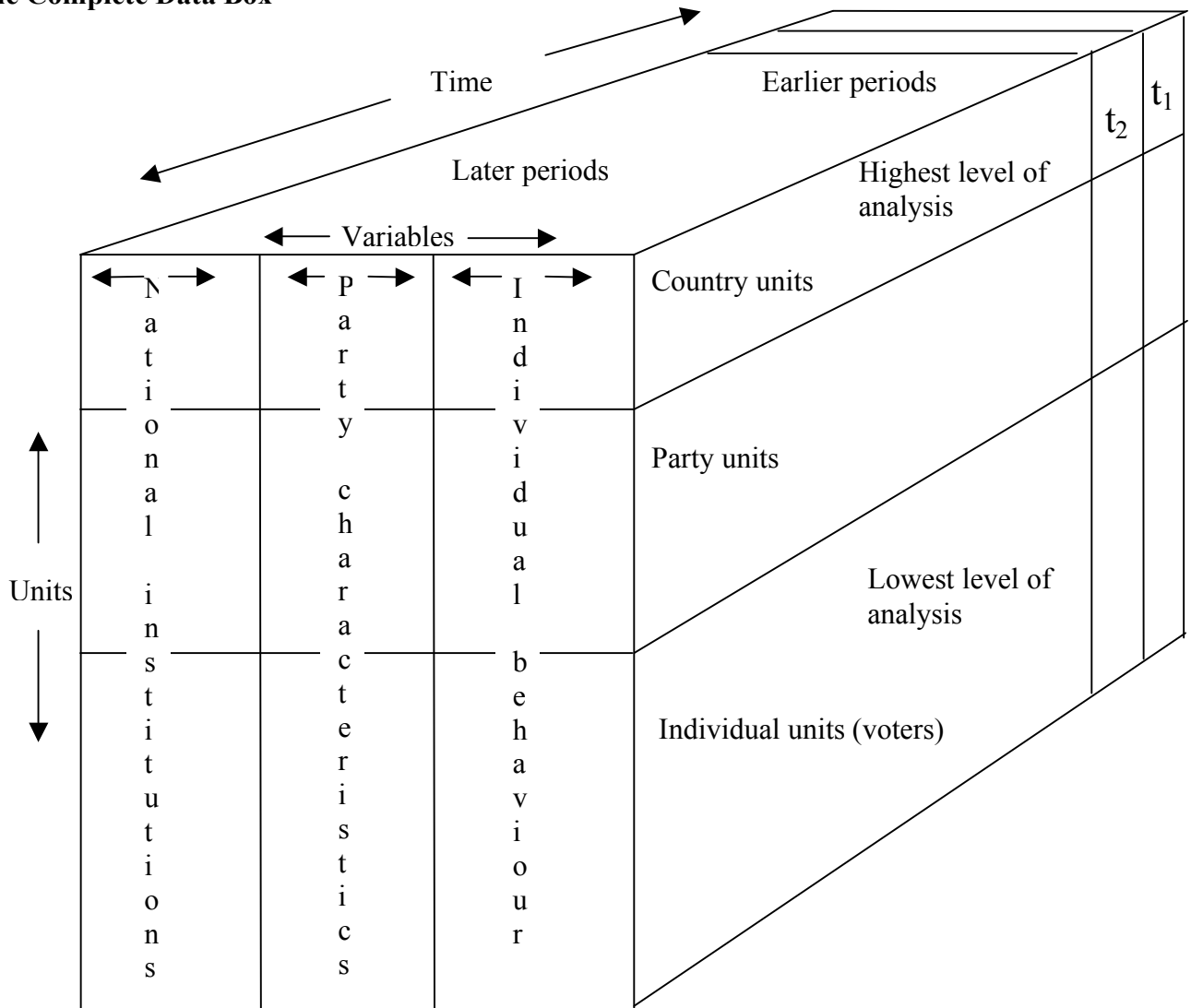
Take it in chunks...

First, ignore the square root symbol that encloses the whole of the expression. We consider that last. Next consider the summation symbol. We saw that before, when we calculated the mean. This time it has no $i=1$ or N below and above, but these are assumed when there is no ambiguity about which variable is being summed. Here we only have an X_i so it is usually taken for granted that this is the variable involved in the summation. The \bar{X} we have already met, in the previous equation. It is the mean of X . This time it has a dot after it. This indicates that it stays the same for every different individual as we sum over those individuals (but the dot is frequently not shown when the statistic, like a mean, is calculated over all cases).

Digression on Levels of analysis

A data matrix generally contains data at a particular 'level of analysis', often the individual level if we are studying respondents to a sample survey, but often at other levels such as the party level. If the information in the data matrix related to different countries, country would be the level of analysis. A dataset can contain information from different levels of analysis, but then it no longer contains a data matrix but something more complicated. Below is pictured a so-called 'data box' with information about different levels of analysis and different points in time.

The Complete Data Box



Dealing with data in this format is virtually impossible. In practice we take slices or chunks out of the data box (e.g. time-series, party data, survey data – see over) and investigate just those slices. But then it is important to verify that different ways of slicing and dicing the data box yield the same story. Sometimes additional aspects of the full data structure can be accommodated in a single analysis, in which case it is called a ‘multi-level analysis’. Multi-level analysis is currently at the cutting edge of methodological techniques in political science and is yielding very important findings.

Multi-level modeling is a response by behaviorists to the objections of the new institutionalists who complained about the focus of behavioral studies on individual-level data. Their criticisms were well-taken and behaviorists and institutionalists now work together on data collected at multiple levels of analysis.

Levels of measurement

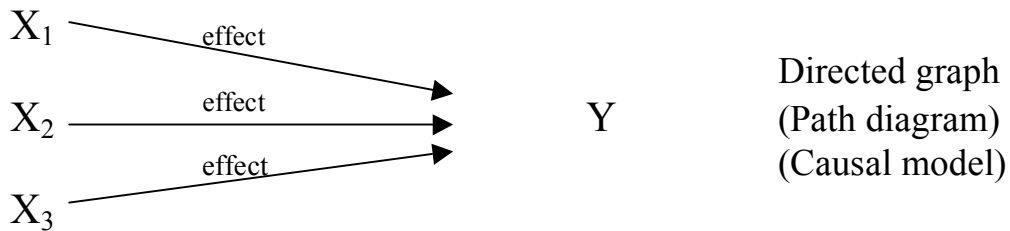
Variables measure things about cases, but different things are measured in different ways. The most obvious type of variable is a so-called ‘interval-level’ variable whose values are measured on an interval scale – a variable like age (measured in years) or income (measured in Euros). In political science and sociology we have very few proper interval-level variables at the individual level, though at higher levels of analysis so called ‘aggregate data’ almost intrinsically contains mainly interval-level variables derived from the process of aggregation to the higher level of analysis. As explained in my chapter in the departmental textbook, when you aggregate individual-level data the result is often a mean or a percentage, both interval-level types of measurement.

At the individual-level, however, we often have to make do with variables that are not quite interval-level. Either they are ordinal (where successive numeric codes indicate more of something, but not in terms of any measured scale) or they are nominal, in which case they have to be converted to dummy variables which measure the presence or absence of an attribute.

As I tried to make clear in my chapter in the department reader, dummy variables are the last resort in our search for variables that behave as though they were interval. By a sort of sleight of hand, dummy variables obey the principle rules for interval-level variables: a score of 1 is greater than 0, and the interval is the same as any other interval on the scale – there being no other intervals on the scale!

Dependent and independent variables

Independent → Dependent



X_2 and X_3 might be viewed as ‘control variables’

Each arrow can be accompanied by a coefficient indicating the strength of the effect carried by that path.

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e$$

Prediction equation
(regression equation)

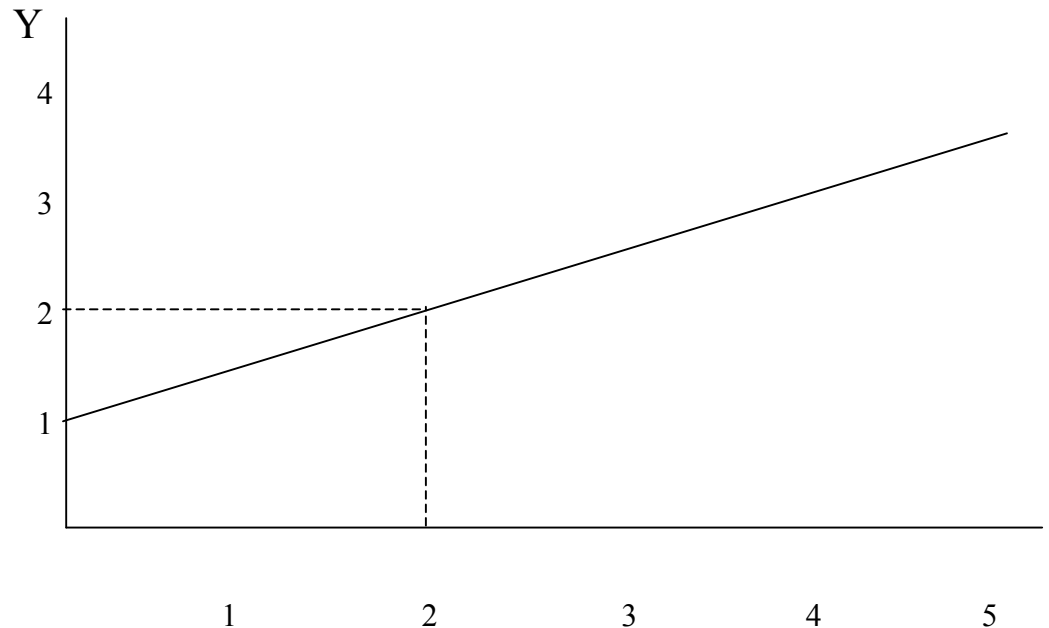
A prediction equation contains the same coefficients as would be placed on the arrows of the corresponding path diagram. Whether on a path diagram or a regression equation, the dependent variable is always the variable affected by the independent variables. In a path diagram the dependent variable generally appears to the far right, at the end of a set of rightward pointing arrows leading from the independent variables; in a regression equation it appears to the left of an equals sign and is followed by the set of independent variables.

REGRESSION EQUATION:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e$$

(Prediction equation)

The regression coefficients (bs) are the effects in a path diagram. But a regression equation gives us the raw materials for more than a **directed graph**, it also gives us the raw material for a **Cartesian graph**



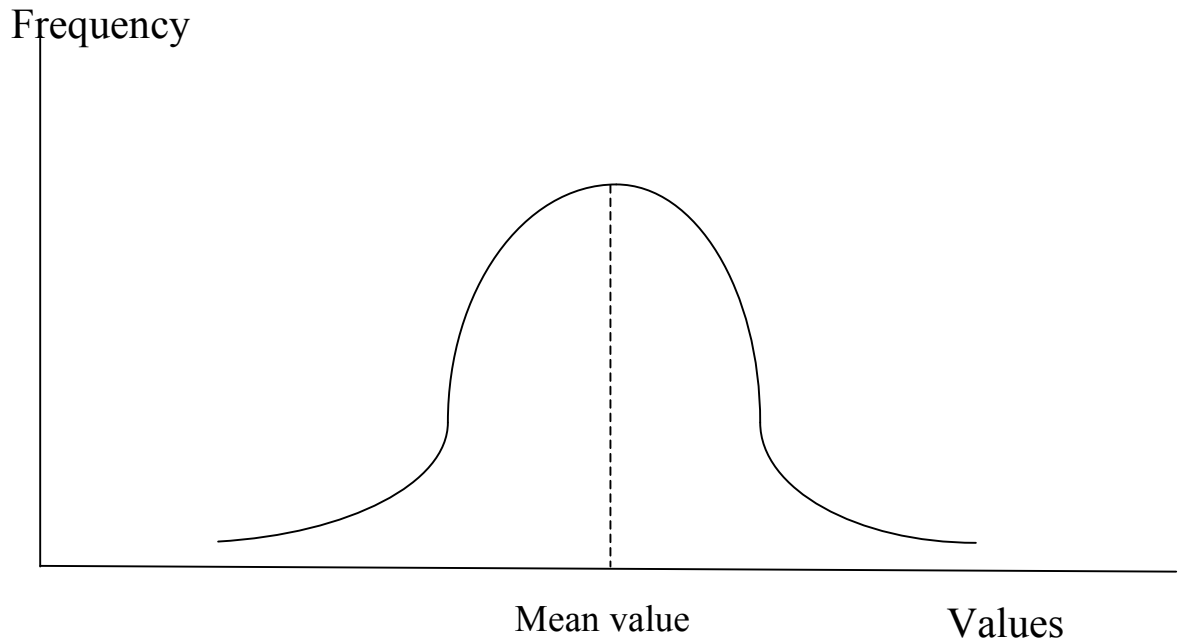
This is the Cartesian graph for $\hat{Y} = 1 + 0.5 X$

When a Cartesian graph is drawn on the basis of a regression analysis, the sloping line is called the “best-fitting line”. Notice the hat on top of the Y - plays the same role as the bar on top of a symbol, indicating an estimated value, but this time estimated in a different way. How?

It is called this because it is the line that “best fits” the cloud of points that will be found in real world data, where definitive relationships are rare or impossible to find. It is chosen so that the variance of the points about the line is minimized – in other words, so that unexplained variance is minimized.

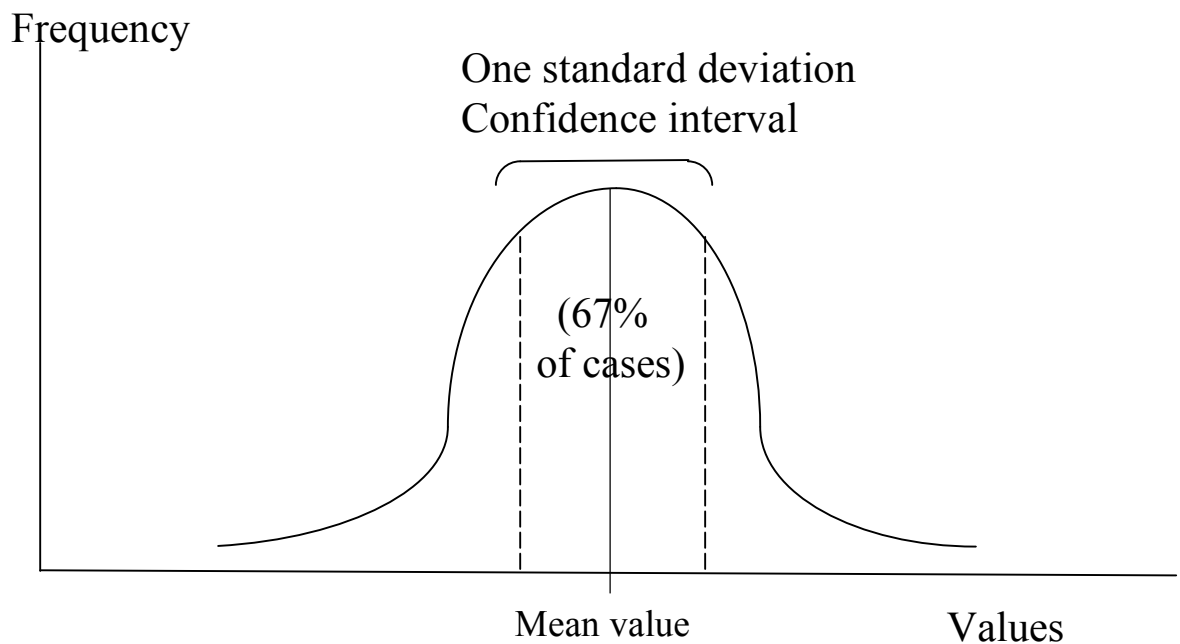
Not only a regression line but also differences of means can minimize the unexplained variance. If we sort people according to any criterion (categorical or numeric) this may reduce the unexplained variance.

Frequency distribution and normal curve (not drawn accurately)...



How **fat** the normal curve is depends on the standard deviation and indicates accuracy of the mean.

The "normal" curve is designed so that 66.66666% of cases fall within one standard deviation above or below the mean, 95% of cases fall within two standard deviations of the mean. The two-standard deviation



confidence interval is the interval that corresponds with the industry standard of 95% confidence (or a 5% chance of being wrong). This is the confidence interval that corresponds to the “**margin of error**” talked about in newspaper reports that describe opinion poll findings. However, the margin of error is based not on the specific sample but on an estimate of what you would get from an infinite number of samples. This is called the "standard error of estimate" (s.e.e., or s.e. for short).

In much the same way, we also talk about the margin of error about the best fitting line – a margin of error that tells us how accurately the best-fitting line actually fits the data. Two standard errors of estimate yields the 95% confidence interval around the best-fitting line.

From this it is a small step to variance explained (R^2) – the proportion of the variance in the dependent variable explained by the independent variable(s). The variance about the best-fitting line.

We know the total variance (the variance originally present in the dependent variable), we also know the unexplained variance (the variance about the best-fitting line). By subtraction this yields the explained variance (the variance of the best-fitting line):

Explained variance = Total variance – Unexplained variance. From this we get

$$R^2 = \frac{\text{Explained variance}}{\text{Total variance}}$$

Significance level: The probability of being wrong when you conclude that there is a relationship between two variables (or an effect of one variable on another). We say that a coefficient is significant at the .05 level when the 95% confidence interval does not include the 0 point (when the coefficient is significantly different from 0).

This is all you really need to know about statistics, the rest is embellishment and detail. But the devil is in the details!