APPLIED ECONOMETRICS

An Introduction

This text is intended to introduce Applied Statistics and Econometrics using simple reasoning and many examples. It is ideal as a course template for Accounting and Finance students. Questions and exams are also available on request.

Duarte Trigueiros
ISTAR-University Institute of Lisbon,
2022

Table of Contents

Chapter 1 Data organization

Data[1] are numbers, texts, images, sounds, textures, and smells, which are recognized in the same way by most observers (data are "objective").

1.1 Attributes, objects, tables, entities

Nothing that is difficult or impossible to describe by different persons in the same way is data.

"Subjective" views (emotions, feelings…) are not data.

Non-transferable content are not data.

New observations that nobody witnessed before are not data.

"Observation" is the datum on a fact of interest. The fact of interest is the "object", "subject", "case" or "record". In the analysis of a specific firm, a high liability ratio or a low profitability ratio are not just data: they are observations.

In statistics, we consider not just data but also the relationships between data. The relationship between the Euro and the Yuan or between age and probability of default. Data and relationships are the two basic building blocks of any statistical analysis.

"Information" is data capable of removing uncertainty. The financial analyst is often faced with missing yet required data. For example, results of an auditing trial are not yet available. Such missing, required data are information since, once obtained, uncertainty is removed.

Data, which we do not know neither need to know, or data which we already know, is not information. The only missing, required data is information.

Information once received is no longer information because it is made known.

To inform is to send data; to be informed is to receive required data.

Upon reception of required data, uncertainty is removed, if not totally, at least it decreases.

With the demise of uncertainty, data ceases to be informative.

Data continues to be informative once completed its role in the removal of uncertainty - but in a different way. Data allows discovering statistical relationships, and the testing of hypotheses via statistical inference. In fact,

Data can reveal relationships between their component attributes or between objects. Even the mechanisms underlying the genesis of data can be unveiled.

Data can also test whether certain hypotheses are likely or not, thus allowing scientific inference.

The discovery of relationships (also known as "modelling") and inference are the two major tasks of statisticians and econometrists.

- / -

The two elements of an observation are the "attribute" and the "object". Mr. Thomas is 64 years old. This age is an observation, where the

attribute is 64 years of age, the

[1] Data is the plural and "datum" is the singular form.

object is Mr. Thomas

On the same object, Thomas, we observe other attributes, such as annual income, personal assets, liabilities, and so on. In more than one objects, Thomas, Smith, Brown, or in the same object at different periods in time (a time-sequence of 3 consecutive years), we observe the same attribute. These observations form a table.

A "table" is the collection containing attributes observed in objects. The table is the most basic and general way of establishing a relationship between different data.

A table's "entity" is what identifies its objects according to a common, relevant characteristic. There are tables containing personal data, city data, drugs' data, returns data, prices data... Each table has an entity.

In the example of table given below, the entity is "bank customers".

Name	Age	Annual Income	Mortgage	Married	...	Smoke	Loan Default
Thomas, S.	67	150	90	Y	...	Yes	Yes
Smith, T.	58	120	80	N	...	No	No

In a table,

each attribute occupies a column, and

each object occupies a line.

A table with 120 rows and 8 columns is a collection of seven attributes observed for all the 120 objects. A row is also known as a "record" or "case"; and each column except the one that identifies the object, is an attribute of the object, that is, is an observation.

Many call "variables" to attributes; but, in fact, not all attributes are variables. The concept of variable is narrower than that of attribute. The column that identifies each object, the leftmost in the above table, is not a variable and is usually not called "attribute" but "reference", "key" or "index", and has a role in organizing data.

Each observation found in an attribute is a "state" or "occurrence" of that attribute.

Attribute "Sex" has two states, "masculine" and "feminine".

Attribute "Income" has an endless number of different states.

- / -

A comprehensive description of entities involved and the relationship between them is the initial step of any statistical analysis. Such description can then go for the testing of hypotheses and for the building of models.

Many of the relationships provided by economists and used by financial analysts appear in the form of a table. Relationships that link together different entities can be tables as well: the link between customers and their cities or between customers and the employee who knows them.

A table is a relationship -- but there are relationships that are not in the form of a table. They are, for instance, in the form of an equation.

1.2 Data types

There are three major types of data:

1. Names (nominal data)
2. Orders (ordinal data) and

3. Scales (scalar data).

Names are the simplest of all forms of observation. Labels distinguish one name from the other. Thus, we can distinguish the sexes by their names (male and female) and two types of bonds (convertible and non-convertible) by their names. We distinguish a collection of seven brokers by the name of each. Death distinguishes itself from life by naming each condition; and a name separates the response to stimuli from no response. A, B and C are three names that have been given to three types of bankruptcy, as well as to three strains of fungi.

An important type of nominal observation is the "binary" observation. An attribute is binary when its possible states are only two. This is the case of sex (two names, male and female) or response to recovery attempts (response, no-response). Binary attributes allow the use of simple and reliable modeling tools and inference tests, quite unlikely those tools and tests that admit more than two names.

A binary attribute where the two states are logically complete is a logical attribute. In a logical attribute, a state implies the negation of the other. It is worth distinguishing

the impossibility called "logic"

from the "physical" impossibility.

A physical impossibility is that which is found in real life but could not be necessarily verified. A logical impossibility is one that, neither nor in the intellect can be conceived because each condition is the negation of the other, and nothing can be, and not-be, at the same time.

Sex is binary but it is not logical because one can think of three or even fourteen sexes; and one may even think of sexes that are not mutually exclusive. But the response to treatment is a logical attribute since a bankrupt firm cannot respond and not-respond at the same time. The states of life and death are logical because death implies the logical negation of life.

- / -

Ordinal attributes are those where names imply order, for example, a sequence from the smallest to the largest. When the response of a firm to a recovery plan is described as strong, median, or weak, such observation is not just a label. It indicates an order. The names now used, strong, median, weak, are not just names. They indicate a quantitative differentiation between states, no longer a qualitative differentiation.

In certain circles, it is customary to call categories (in plural) to the nominal and ordinal data, with the latter sometimes called "ordered categories" but it is important to separate nominal attributes from those with an underlying order.

A purely nominal observation requires a simple type of measurement called "classification" or "recognition" (separating what is different and gathering what is similar) while.

an ordinal observation, besides requiring recognition, also requires "sorting" all possible states by magnitude.

When we take an order as though it were a simple name, important information is lost. This happens, for example, when using the Chi-Square test to infer about the difference between the number of weak, good, and very good response to a recovery plan. The response attribute is ordinal in this case. Since Chi-Square treats ordered categories as if they were names, it ignores useful information.

The existence of truly ordinal attributes, that is, attributes where the distance between states are unknown and may be large or small, requires the use of specific tools for modeling and inference.

Where an order has just two or three states as in the Chi-Square example above, it is customary (but not good practice) to ignore its ordinal condition.

- / -

Scales are ordered observations where the distance between observations is known. There are two types of scales: “interval” and “rational” scales.

A scale is interval when zero does not exist. The Centigrade temperature is interval because zero is a convention, not real absence of temperature. Zero is there where the water freezes but it could be any other temperature; it does not indicate absence. Many economic indicators are interval because in them the zero value does not mean absence.

In a rational scale, there is a real zero. Zero money in my pocket, a weight of zero, zero customers or zero returns means the total absence of money, weight, customers.

The distinction between rational and interval scales is important. It does not make sense to perform certain arithmetic operations on interval scales. The calculation of percentages, for example, should never apply to interval data since the very notion of percentage is meaningless (and indeed misleading) in this case. For instance, to say that, with the administration of a drug, the average temperature of patient decreases 14% is nonsense. However, to say that it may decrease two degrees may make sense. In the case of interval attributes, instead of percentages, we should use differences.

Rational scales allow using percentages. Indeed, in rational scales we can perform all arithmetic operations. An increase of 10% in weight makes sense. Therefore, one should not use an interval scale to measure observations where it would not make sense a total absence of what we measure; and one must use rational scales where zero has its true sense in reality: a total absence is indeed possible. The difference between rational and interval scales is often ignored but this may lead to misleading results.

One should also be on guard for the existence of attributes that seem ordinal but are in fact scalar. A bad, median, and good profitability ratio, in fact is not an ordinal attribute because we know the distances between states. Profitability is a scale, a rational scale, even after being discretized into three categories.

Treating orders as though they are scales is an error as well. Suppose that an analyst performs a few observations purely ordinal. In the end, the analyst says that X is the best customer and Y is the worst; then, each one of three states high, median, and low, is assigned to customers based on the comparison of the two extreme cases X and Y. In the end, the analyst calculates an index ranging from -10 to +10, depending on the classification obtained. Does this index turn out to be a scale? No.

Most widely used indexes are not scales. They are manipulated to pretend that scales exist where they do not.

- / -

Besides names, orders and scales, there are other ways to classify observations. One with practical importance divides attributes into

discrete and

continuous.

The latter has endless states. Money, returns, prices, most ratios, Betas, are examples of continuous attributes. A discrete attribute is one with only a limited number of states.

Names and orders are always discrete but, as mentioned, the fact that an attribute has a discrete number of states may not indicate that it is nominal or ordinal. A continuous attribute is generally a scale; a discrete attribute can also be a scale. The age of children between 4 and 16 years form a discrete set with only 13 states; but it is continuous.

1.3 Measurement, informational productivity of data, accuracy

To "measure" is to compare an observation with another, taken as the standard. A balancing scale measures mass by comparison with calibrated weights placed in the opposite plate. Money, ratio values, are measured by comparison with values taken as standard. One million, or a ROE of 13%, are meaningless without a term of comparison.

Without a standard, or, at least, a term of comparison, it is not possible to observe. An illness with symptoms never seen before, and which does not have any manifestation known as pathological, is invisible to the doctor. The financial analyst can detect only what he or she already knows.

The different ways of measuring follow the data types of what is measured: nominal, ordinal, interval scale, or rational scale.

In the case of names and orders, to measure is the same as recognize (classify): choose, from a finite collection of standard states, the one that best approximates that observed.

Besides recognition, the order measurement requires another operation, namely "sort" observations according to magnitude relative to the other observations.

In the case of scales, certain measurements require calculation of a difference, which can change measurement in an unexpected way. Logarithms, for example, are widely used in the measurement of survival, returns, and the growth of firms (and indeed of children). When comparing two observations, which were made using logarithmic scales, such measurement is not a difference but a relative change (percentage change).

It was mentioned that measuring implies comparing something observed (a ratio, a return, response of a firm, sex of the client) with something that was previously agreed to be a standard (ratio standard, scale associated with each response characteristics, features associated with each sex). Without this comparison to an accepted standard, we cannot measure. Being accepted is not the same as being arbitrary.

An accepted standard is appropriate to a given measure. For a measurement to make sense and become useful it is essential that the characteristics of the observation are reflected in the agreed standard. Neither the sex of a client is recorded in dollars, nor annual income as Male / Female. From this need for consistency between observations and the corresponding standards, follow the different data types and measurement operations. If we are dealing with observations that vary continuously, it is convenient to use a range of real numbers to measure it. But if an attribute is only a finite collection of states, then measurement consists only in recognition, sorting, or both.

In the case of scales, both types, there is the added problem of having to harmonize the unit of measurement, and that of having to deal with a variety of standards. In some places in the world dominates the pound and ounce or foot and yard; in other places, standards are decimal. In the economy, such differences in standards and terms of comparison may assume some subtle facets.

- / -

An important type of measurement is the “counting” (reckoning). It uses specific terms and therefore requires some attention.

To count is to observe the number of elements in a set. In a bank, 6 of the totals of 50 loans are not performing and the remaining 44 loans are performing. This is counting. The attribute “to perform” (a binary and logical attribute) has two possible states.

Classes are names employed in counting. In most cases, classes are the same as states of the attribute. Regarding sex, suppose that there are 18 men and 17 women in an office. Thus, we use two classes. By contrast, in the age of children attribute used above, there are 13 different age classes, and we count the number of children in each of them. Even where classes are arbitrary, we still call them classes. We divide firm response to a recovery plan into three classes, bad, median, and good, and we count the number of cases in each class.

The division into classes may be arbitrary, also in the case of nominal and ordinal attributes. In the first, it may happen that two or more names be grouped in one class: 14 different counties can be, for purposes of counting, grouped into five classes or zones. In the second case, the class division is entirely arbitrary: the observed income of 450 customers can be divided into 6 or into 16 classes according to convenience.

In statistics, the result of a count made by class is a “frequency”. The term “frequency” is associated with rhythms and oscillations, not with counts. Therefore, its use in statistics is a delusion, able to cause confusion. We should get used to say, using the example above, that the frequency of responses is 40 and that of non-response is 10. In a city with 500,000 residents, the frequency of accidents by age was 2 for the class of young adults, 8 for the class of middle-aged and 37 for the elderly class.

Relative frequency is a frequency expressed as a percentage of the total. In 50 loans, 40 are performing and 10 are not performing; the relative frequency of performing loans is 80% and that of not performing loans is 20%.

- / -

Relative frequencies are an example of proportion. The proportion is a percentage of the whole: 5% of the loans are not performing; the firm has lost 5% of its customers or recovered 10% of accumulated losses.

But not all percentages are proportions. In a proportion, we cannot observe values above 100%. We should carefully distinguish proportions from other percentages, which can take values above 100% since they are not bound by a whole.

Relative change is any difference expressed as a percentage of one of the two states under comparison: a price grows by 10% (price is now 110% of previous price) when the observed change, when expressed as a percentage of the price previously observed, is 10%. Relative changes are also called relative differences.

When an observation can grow but never regress, a relative change is called relative growth. This is the case, for example, of accumulations, and children’s height. It is worthwhile noting whether observations are the type that can only grow or can grow as well as shrink. This influences the statistical significance of expected differences.

- / -

The amount of information that an observation brings with it may be large or small; it varies according to type. Each type of observation, nominal, ordinal, interval or rational, carries with it a greater or lesser amount of information.

Names, orders, intervals, or rational scales have different informational richness. A simple number, 3 for example, can be poor or rich in information depending on whether 3 represents a simple name, an order, or a scale. In the first case it is just a label, as in the Roman name "Tertio Severus"; in the second case it indicates an order, the third place among N magnitudes; in the third case 3 shows a well-defined position on a scale.

It is important to know the informational abundance of data, as it affects the use of statistical tools.

Purely nominal data occupies the lowest level in informational richness: a name can classify, it can putt each observation in the rightful place and meet it with their like; but a name tells us nothing about order, distance or any other relationship that may exist. A name uniquely identifies an observation as member of a set, nothing else.

The number of digital bits needed to convey nominal information depends on the number of possible states. Binary attributes require 1 bit. A nominal attribute with 8 possible states requires 3 bits. To let it be known that a customer lives on one amongst 16 possible places, 4 bits are needed.

Ordinal data is richer in information because it assumes that states can be sorted according to criteria forming the basis of observation. But the fact that it is known that a category is bigger or smaller than another does not tell by how much. No information about the distance between categories is provided in the case of ordinal data. When someone says that three competitors reached the winning post in 1st, 2nd and 3rd place, that's all we can say about what really happened. It may be that the competitor who came second has crossed the post glued to the first and the third arrived half an hour later. 1, 2 and 3 reveal nothing about the big gap between competitors or 2 and 3 or on the tiny distance between competitor 1 and 2. Thus, there is little information in orders.

Classification and ordination are the only operations applicable to ordinal observations. Arithmetic operations are not allowed. It would be misleading to add or subtract ranked categories represented by numbers, and even worse to multiply them or divide them as if they were scales. However, there are studies that have made this mistake: use ordered categories as explanatory variables in regressions and other tools where attributes are assumed to be rational, and therefore all kinds of arithmetic operations take place using available data, no other. Trying to view information where there is none, leads to conclusions far removed from reality. As mentioned, there are widely used economic and financial indices based on information that do not exist.

The interval scale adds to ordination and recognition an objective measure of the interval or distance between categories. Still, intervals are not as rich in information as rational scales. With interval scales, it is possible to study, not just magnitude but distance. Thus, addition and subtraction are valid arithmetic operations in intervals. As was said, however, it does not make sense to use percentages to compare proportions and other observations nor, indeed, any data manipulation involving multiplication and division.

Finally, rational scales possess a zero value in an objective sense. In addition to having the properties of classification, ordering and distance, rational measurements satisfy all properties of the integer or rational number system: they can add, subtract, multiply, and divide.

From this point on, it is important to note the different informational richness of the data. We should always assign to data its informational richness, neither more nor less. This requires meeting the following two steps:

1. identify the type of data, nominal or ordinal, interval or rational.
2. use the analysis tool designed to extract from such data all the information that they possess, no more, no less.

- / -

In the case of scales, informational richness shows itself through "precision". Precision or "accuracy" is the number of digits used to measure, register, or report scale observations. Observation 2.34 has 3 digits and is more accurate than observation 2.3 which has only two; or than observation 2, which has only one digit. Precision measures the amount of information.

Information neither should emerge from nothing nor disappear into nothing. If 3 digits are used to measure solvency and 4 digits to report it, then information has been unduly invented. If 3 digits are used to measure solvency and then only 2 digits are used to report it, information has been unduly suppressed.

Zeros that may exist on the left side of a measurement add nothing to precision: a measurement of 2.34 is as accurate as 0.0234, as both have only three significant digits. Such measurement is also more accurate than 2.3 or 0.0023, which have only two significant digits. Zeros on the right side do not add to precision. The numbers 0.0023 and 23.000 indicate generally the same two-digit precision. However, when trailing zeros are truly measured, that is, when they are not just rounding, they add to precision.

Precision with which data are recorded and presented should always reflect precision with which attributes are measured and the meaning analysts attach to it. It is absurd to record solvency with more than two or three digits because solvency, and many other indexes, are just an estimation involving the prediction of future, uncertain events.

It is also important to use the precision, or the type of measurement, which conveys true economic or financial distinction. To register the age of a client as 69 years 10 months instead of just a particular age group may have interest for retirement purposes but not in Banking.

Age is indeed a good example of how measurement should conform to reality: in most cases, medical doctors prefer to use an order (age group) instead of a scale (lifetime), knowing that the use of continuous variable not only adds nothing to the diagnosis as it may introduce potential sources of error. Thus, the statistician or econometrist should use common sense and experience to choose precision of measurements and records.

And never think that too much precision is innocuous or looks good. Besides inducing analysts to commit error, it shows ignorance. It can even lead to incorrect classification of data and the inappropriate use of analytical tools.

- / -

Certain transformations reduce the richness of data scales. The logarithm of a number is less informative than the number itself. Notice how, in fact, a logarithm is an interval scale and therefore, zero ceases to be objective. Therefore, when applying logarithms, information is lost.

- / -

The "rounding" of scales is often required after using algorithms, to bring back numbers to their original precision. Suppose that measurements are made with a 3 digit precision. Then, after

rounding, a number reported as 3.78769 should be rounded to 3.7877 and then to 3.788 and then to 3.79. Number 3.78442 would be rounded to 3.7844 and then to 3.78. Typically, digits which are above 5 are rounded up and digits which are below 5 are rounded down, and the digit 5 can be rounded either up or down.

Chapter 2 Probability and distributions

2.1 Probability, operations with probabilities

If, in 20 bankrupt firms, 15 of them (75%) respond to recovery plans while 5 (25%) do not respond, we are left with the view that response is more likely than non-response. The concept of "probability" (likelihood, plausibility) relies on the principle that what we observe should relate to what we expect to observe in similar cases or in the future.

For instance, the fact that it rains makes us suppose that rain is plausible; but we would never have believed in the plausibility of a rainbow, or even a peacock, if its existence were only a legend and we had never seen one. Implausible, out of the way facts do happen although less frequently than rain. Financial analysts, economists, managers and medical doctors know this quite well.

Plausibility is quantified using "probabilities": an expected relative frequency, that is, an expected count, expressed as a percentage of the total.

Frequencies, like everything else, can be

observed, or

expected.

When expected, relative frequencies are called probabilities.

Since a probability is not an observed, but expected frequency, we know probabilities in two possible ways:

1. Through experience. It is known that distribution of sexes is approximately 50% and 50%. This is so, even if in a family seven girls were born in a row. Note that a judgment is made that what is observed in the future will be like that observed in the past; this judgment, in economics, is overconfident in the least.
2. Through knowledge of the underlying mechanism: it is known, for example that, if there are 80 white balls and 20 black balls in an urn, every time that a ball is withdrawn and put back, the expected frequency will be 80 white balls in 100. This is so, even in the unlikely event of 20 black balls come out in succession.

Probability associated with one single state has little interest. What is informative is the collection of probabilities associated with all the states of an attribute. This collection of probabilities is called "distribution". The sum of these probabilities is equal to 1 or 100% or the total number of expected cases.

If, in the distribution, one of the probabilities is 1 then all the others must be zero.

1. A probability of 1 or 100% or the total number of cases indicates a "sure" event, that is, the state in question occurs with certainty.
2. A probability of zero indicates an "impossible" event: the state in question is impossible or never seen.

The probability distribution associated with an attribute measures the uncertainty that analysts face, or the risk that managers face, when dealing with such an attribute:

1. When a distribution is flat (all probabilities are similar) the uncertainty regarding which of the states will come out is maximal. Information is minimal. This is what happens with games of chance (dice, coin tossing) where all outcomes are equally likely. It is what happens to the expected sex of a baby.

2. When, by contrast, a given probability is greater than others are, the uncertainty faced by analysts and bank managers regarding the future outcome is smaller.

Thus, the probability distribution reveals both uncertainty and its opposite, information. One shows the others. Probability distributions and uncertainty are strictly equivalent.

- / -

Two types of mechanisms or a mixture of both dictates the plausibility of events:

- Additive: when expected frequencies stem from the addition of elementary likelihoods. In Nature, additive mechanisms are the majority. Adult weight, height, and other observations obey additive mechanisms.
- Multiplicative: when expected frequency stems from multiplying elementary likelihoods. All phenomena associated with growth (firm size, weight, height of children for instance), sequences of events (survival) or accumulations (income, capital, spread of news) obey this multiplicative mechanism.

Statisticians and econometrists should keep such distinction in mind as it has important implications. An analysis based on risk factors, for instance, assumes that the underlying mechanism is additive, not multiplicative. As an example, the likelihood of a firm going bankrupt may be as high as 90%, stemming from the addition of the following elementary probabilities:

1. Low profitability adds 40% to the probability of bankruptcy.
2. Liquidity problems add another 30% to that probability.
3. The High Liability to Assets ratio adds 20% to that probability.

Elementary probabilities add together when their random mechanisms act independently from each other. This is generally the case with risk factors.

A typical example of multiplicative mechanism is survival. Suppose elementary chances of not surviving a given year (hazard rate), are as follows:

1. elementary mortality rate associated with the first year of the disease is 27%; Thus, survival rate associated with the first year of the disease is 73% = 1 - 27%.
2. Elementary mortality rate associated with the second year is 32%; Thus, survival rate associated with the end of the second year is 68% = 1-32%.
3. Elementary mortality rate associated with the third year is 40%; Thus, the survival rate associated with the end of the third year is 60% = 1 - 40%.
4. Elementary mortality rate associated with the fourth year is 20%; Thus, the survival rate associated with the end fourth year of the disease is 80% = 1 - 20%.

What is the probability of the patient being alive?

... At the end of the first year? It is 73%

... At the end of the second year? It is 73% times 68% = 50%.

... At the end of third year? It is 73% times 68% times 60% = 30%.

The multiplication of each passing year's survival rate reflects the sequential nature of the process: a patient is alive at the end of year N only if he/she survived, in sequence, to all previous years. Such cumulative probability shows the likelihood of a patient surviving the first, then the second, then the third ... up to N-year disease. Each mortality rate is linked to the earlier year by a sequential, causal chain.

Survival mechanisms are important in Insurance. In the Banking industry, the same mechanism applies to bankrupt firms subject to a recovery process, where survival is, in that case, the fact that the firm did not relapse into insolvency.

2.2 Randomness, variability

An attribute is "random" when its states / occurrences cannot be predicted exactly.

A bank analyst can predict the personal income of a customer from age, sex, education, race, and from genetic and environmental factors such as parental education and country; still, there remains much unexplained income and therefore income is a random attribute.

The sex of a yet unborn baby cannot be predicted at all, although it may be observed. Sex is also random but is more random than income because it is not possible to predict sex from other attributes. Moreover, as seen, uncertainty associated with sex is maximal because probabilities associated with the two possible events are equal.

Randomness brings uncertainty to observations. However, uncertainty is not the lack of knowledge or pure chance. On the contrary, randomness typically brings with it one or even several certainties. Most random phenomena exhibit great regularity. Consider the two extreme cases of randomness:

- Unlimited randomness: imagine a living being in the shape of an amoeba with sizes from microscopic to several cubic kilometers.
- Absence of randomness: imagine living beings so regular that it is possible to predict exactly their height from sex and age.

The amount of randomness is contained within the two extreme cases above. It is not zero, but it is not as big as that. Height, for example, is amazingly stable: no bank clients run under the tables as though they were mice or walk through buildings.

If we look carefully, we notice that two causes limit the randomness of attributes:

Their span or range (the distance between lower and upper value) is non-infinite.

The tendency for states to be more frequent, more plausible, the closer they are to a "central value". Height, for example, or firm profitability within an industry, or the Betas of quoted companies, have central values in their distributions. Regarding this,

- the closer to a central value, the more likely a state is,
- The further away from the central value, the less likely a state is.

A large departure from the bounding limits or from the central value of a distribution determines the impossibility of certain states of an attribute. An adult 3 yards tall or 900 pounds weight is in practice impossible to find; a gestation period of 49 weeks is not seen every day, a profitability of 50% is not sustainable, a Beta of 100 is never seen. Randomness, which seems to indicate vagueness, leads in fact to strict conclusions about what is expectable and what is not.

- / -

The "variety" of an attribute informally denotes its number of states (or "occurrences"). The opposite is constancy, and constant attributes are useless. One sex only would be constant and not interesting for analytic purposes. If people would come with two or one head, then the number of heads of individuals would be an important attribute. As it is, it is not.

Variety is not the same as randomness. Attributes may vary but without being random. What characterizes randomness is unpredictability, not variety. We can predict age from the number of

turns that the Earth revolved around the Sun since the moment of birth of the object. Age is not random. This number of turns is stable. Yet age, alas, varies.

Sex may be random or not according to its role in the analysis. When sex is given, then it is not random; but when trying to predict the sex of embryos or customers, then sex is random.

Where there is variability, random or otherwise, there is place for using its distribution.

A distribution counts the number of objects by state. It thus highlights:

1. The number of states, that is, whether states are just a few (two sexes, two responses) or many (several cities of a county).
2. How frequent states are, and what is the most and least frequent state.

Distribution is a collection of the frequencies either observed or expected for each state.

If frequencies are expected, the distribution is called a “probability distribution”;

If frequencies are observed, the distribution is called a “frequency distribution”.

In a distribution,

frequencies may appear in the form of counts,

or in the form of a percentage of the total (“relative frequencies”),

or as either accumulated frequencies or percentages.

- / -

Distributions are the foundation of any statistical analysis.

The correct building and interpretation of distributions depends on whether the attribute is nominal, ordinal or scale.

For nominal attributes, each state usually is a class for the purpose of counting; the position of classes is not important nor is the cumulative frequency:

Number of clients by profession		Relative frequency	Cumulative frequency
18	Liberal	19,4%	19,4%
42	Technical	45,2%	64,6%
27	Manual	29,0%	93,5%
6	Artistic	6,5%	100%
93	Total clients	100%	Total

In ordinal attributes, the position of classes in the table becomes important and the cumulative frequency becomes meaningful. Here is the example of response to stimuli.

Response	Frequency	Relative frequency	Cumulative frequency %
no response	64	44,8%	44,8
Traces of a response	26	18,2%	62,9
Small response	27	18,9%	81,8
Median response	12	8,4%	90,2
Good response	9	6,3%	96,5
Strong response	1	0,7%	97,2
very strong response	4	2,8%	100
Total	143	100%	

In the case of scale attributes, class division is arbitrary since possible states are endless. In the example below, we arbitrarily divide the height of 58 children in four classes of 20 cm and then we find the frequency of children whose height falls into each class:

	Height of children 8-14 years old	Relative frequency	Cumulative frequency %
8	[1,00 m até 1,20 m[	13.8%	13.8%
36	[1,20 m até 1,40 m[	62.1%	75.9%
11	[1,40 m até 1,60 m[	19.0%	94.8%
3	[1,60 m até 1,80 m[	5.2%	100%
58	total	100%	Total

In order to build a distribution, it becomes necessary to choose classes with a width that best suits the ends in sight. This width will depend on the number of objects:

many objects allow using many narrow classes, but

if the number of objects is small, few, classes are used to avoid blurred counts.

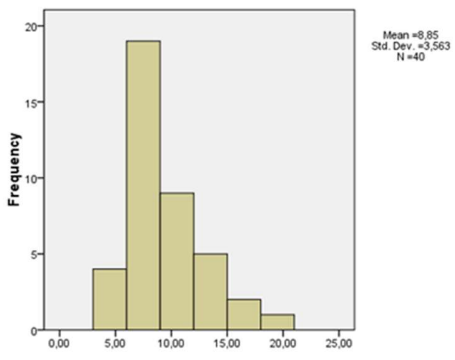
When attributes are scalar, how do we know how many classes to use? Ultimately, what we want to obtain with a distribution is a picture of variety. Total number of objects determines the resolution of such picture and number of classes determines detail of the picture. If we ask for detail but have little resolution, we get a blurred image. Therefore, the number of classes should wisely reflect the total number of objects.

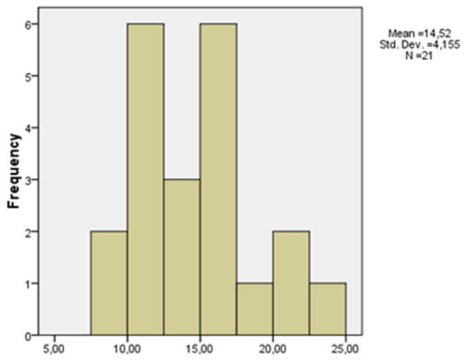
Less than 100 objects do not allow more than 5 classes. Some 500 objects allow using 9 or 10 classes - at least 20 objects per class always.

- / -

The graphical representation of distributions can either help or hinder interpretation. For discrete attributes with few classes, graphics are generally useful. The distributions below compare the variety of a wealth index among customers for two places, one rural and the other urban. Lower variety and lower mean wealth is observed in rural areas (left) when compared with urban areas (right).

Graphical representations of distributions as below, are known as “histograms”.





In an histogram, frequencies are intuitively described as the length of bars: the most common classes are those under long bars. Histograms may show absolute frequencies, as above, or relative frequencies (percentages).

Use this distribution of children's ages 5-18 years old in a town to understand how cumulative frequencies are interpreted:

Age	Frequency	Relative frequency	Cumulative frequency %
5	16	5,9%	5,9%
6	18	6,6%	12,5%
7	22	8,1%	20,7%
8	23	8,5%	29,2%
9	29	10,7%	39,9%
10	23	8,5%	48,3%
11	19	7%	55,4%
12	24	8,9%	64,2%
13	24	8,9%	73,1%
14	17	6,3%	79,3%
15	15	5,5%	84,9%
16	17	6,3%	91,1%
17	17	6,3%	97,4%
18	7	2,6%	100%
Total	271	100%	

There are 16 children 5 years old representing 5.9% of the total. 48.3% of children are under 11 years, almost half. The median age should therefore be around 10 years. "Modal" class, the most common class, is 9 years. Mode and Median are useful when studying ordinal distributions; in the case of scales, mean values are generally preferred.

- / -

We call "variability" to the variety in relation to some fixed standard. The appropriate method to measure variability depends on the type of attribute, that is, ultimately on economic or financial reasons. Nominal attributes are variable according to the number of states they admit. Sex, having only two categories, is less variable than the cities of a county where there are 14 categories. For the inhabitants of a planet where there are 5 different sexes, life on Earth would certainly seem monotonous, lacking in variety.

In order, variability is measured using sorts and counts as these are the only possible operations. This is "non-parametric" comparison. Percentiles and median values can compare the variability of ordinal attributes.

When a variable is continuous, there are two practical ways to measure variability.

In the first, the term of comparison is the span, that is, the range defined by two boundaries, the lower and higher, beyond which it is unlikely to find objects.

In the second, such term is the central value, for example, the mean.

The second way is more convenient but there are variables where comparisons involve ranges, never central values, because their distributions do not show central values.

Height illustrates both the existence of ranges for to what is likely and the existence of a central value. Indeed,

It is unlikely to find adults less than one meter tall or taller than two and a half meters.

It is also evident that all height observations seem to cluster round the mean, which is the central value in this case.

Variability around a central value is measured as the "SSQ", the sum of all squared differences observed with respect to the mean:

$$SSQ = \sum_{j=1}^{n} \left(x_j - mean\right)^2$$

Consider the following example: heights of 3 objects are 145cm, 153cm and 170cm with a central value *a priori* assumed to be 160. In this case,

Object	Height	Difference from 160	Squared difference
1	145	-15	225
2	153	-7	49
3	170	10	100
	total = SSQ		374

The sum of the 3 squared differences is SSQ =225+49+100 = 374.

The symbol $\sum(\dots)$ is the Greek letter Sigma in caps, an abbreviation to the sum of several parts. In the above case it is indicated that the sum must be performed starting from j = 1 and ending at j = n. The total number of additions is N.

Note that the value taken as central in the above example, 160, is the mean but it is not calculated from the three observations given (average 156) although it is close. We may have obtained the central value from a survey table or perhaps from groups with larger number of objects. The use of SSQ is linked to modelling and inference OLS methods.

The above variability measure, SSQ, produces more variability for many objects and less variability for few objects. It cannot compare variability directly.

2.3 Degree of freedom

Since we need to quantify and compare variability, the notion of "degree of freedom", central to statistical end econometric reasoning, is now introduced. The degrees of freedom are the effective number of sources of variability available and are measured as the number of objects less the number of restrictions to the variability of those objects.

Any group of 20 adults of the same sex from whom we observe the height contains 20 sources of variability, one for each object. However, the attribute height of adults of the same sex has a central value, i.e., the distribution of this attribute is restricted to plausible values around the mean height. Therefore, we should subtract one degree of freedom from 20. Thus, what is left is 19 degrees of freedom.

The notion of degree of freedom applies to a type of statistical modeling and hypothesis testing called "parametric" because it describes frequency distributions using a small number of parameters. When distributions of 500 objects can be described using two parameters only, then we note that there remains 500-2 = 498 degrees of freedom. All that is regular and that is modeled using one parameter takes out a degree of freedom.

"Variance" or MSQ (mean squares) is SSQ divided by degrees of freedom. It is an average SSQ. Using the notation where degrees of freedom is written "d.f." and the sum of squares is written SSQ, then:

$$variance = \frac{\sum_{j=1}^{n}\left(x_j - mean\right)^2}{n-1} = \frac{SSQ}{d.f.}$$

where n is the number of objects and *x* is the attribute with a given mean. In the above example variance is 374 / (3-1 = 2) = 187. Using variance, it is possible to compare variability of groups with different number of objects.

The use of SSQ and variance comes from its role in the Normal function, where the latter is a component parameter. Given the preponderance of distributions with shapes like the Normal function, this type of measurement of variability is handy.

“Standard deviation” is the square root of variance. The advantage of standard deviation is that it expresses the same measurement units as the attribute in question: one standard deviation of height is centimeters, not squared centimeters.

- / -

When it is required to describe variability in detail, not in the form of a parameter, the whole distribution is used as the variability measure. In such cases, classes are the states observed in the attribute. This is what we must do in the case of attributes, which do not fit into any parametric distribution.

In the case of orders, the most practical way of assessing variability is using ranges and their limits. It was noted that variability could be assessed by comparing upper and lower limits of what is considered likely. This method is based on the reading of “quantiles” such as deciles and the distance thereof in relation to the median. Cumulative frequency distribution offers an approximation of what these values are.

2.4 Quantiles

A “quantile” is each of any set of values, which divide a frequency distribution into groups containing the same fraction of the total number of objects. From the cumulative frequencies above, it is easy to have an idea about the “median”, the “quartile”, the “quintile”, the “decile”, and other “quantile” values.

The first decile is the value encompassing the initial 10% of cases.

The first quintile is the value encompassing the initial 20% of cases.

The first quartile is the value encompassing the initial 25% of cases.

The median is the value encompassing 50% of cases.

... and so on.

The different “percentiles” arise from dividing sorted observations in 100 groups, each with 1/100 (1%) of cases. Percentiles have a role in estimating extreme values at risk.

Note that

Deciles divide objects in 10 groups, each having 1/10 objects.

Quintiles divide objects in 5 groups, each having 1/5 (20%) objects.

Quartiles divide objects in 4 groups, each having 1/4 (25%) of objects.

The median divides objects in 2 groups, each having 1/2 (50%) of objects.

Importantly, quantiles clearly tell which values of an attribute are close to the limits of likelihood. To obtain them accurately it is necessary to order all states from the smallest to the largest and then divide states into equal groups: 100 groups to get percentiles, 10 groups to get deciles, 5 groups to get quintiles, 4 to get quartiles, 2 to get the median.

Take the case of the growth in family income observed in a group of 140 newly arrived immigrants during a period of 5 years. After ordering from the smallest to the highest income growth, we divide cases into 10 equal groups of 14 families each. Of the 14 resettled families (10%) which experienced the lowest income growth, none is higher than 88 money units, and these are the ones we should observe.

10% of cases are below	88 money units	is 1st decile
20% of cases are below	92 money units	is 2nd decile = 1st quintile
30% of cases are below	96 money units	is 3rd decile
40% of cases are below	99 money units	is 4th decile = 2nd quintile
50% of cases are below	100 money units	is 5th decile = median
60% of cases are below	104 money units	is 6th decile = 3rd quintile
70% of cases are below	107 money unis	is 7th decile
80% of cases are below	112 money units	is 8th decile = 4th quintile
90% of cases are below	117 money units	is 9th decile

It may be difficult for the economist to understand what constitutes a low growth in the income of a resettled family over a 5-year period. However, after looking at a table like the one above, the economist knows that.

10% of the families got an income increase of 88 or less. If a specific family is, say, 80 money units better, then the economist knows that only less than 10% of families are stuck in such a low performance.

Also, since only 10% of families experienced an increase above 117 money units, if a family is, say 120 money units better, it means that less than 10% of families got such an income rise.

Note that the 5th decile (or the 50th percentile) is the median, showing the value of the attribute for which there are as many cases larger as smaller.

We use the median instead of the mean when the attribute's informational content is weak, as in multiplicative attributes (accumulations, growth, and sequences) or orders. In general, the median and specific quantiles are a good substitute for the mean and the variance (central value and variability) when the frequency distribution of an attribute is far from Normal. These are non-parametric or distribution-free measurements.

A median value is as likely to occur as not to occur. Indeed, the number of cases which are smaller than the median is the same as the number of cases which are larger than the median. Therefore, the median is the value about which *a priori* information is nil.

Another useful measurement, which relates directly to the observed distribution, is the "mode" or modal value. The mode is the most likely value in a distribution. It is the state or observation with the highest number of observations.

Except as a term of comparison, the mode is seldom used in statistics. However, there is a term which is derived from the mode and is often used to describe distributions: when a distribution has two clearly observed peaks, it is known as a "bimodal" distribution.

Chapter 3 Random processes

3.1 The Distribution Function

Mathematicians develop theoretical mechanisms to replicate random events found in real life. These theoretical mechanisms are known as "random processes".

An urn with 80 white balls and 20 black balls from which a ball is drawn at random with replacement is one such mechanism. From those theoretical mechanisms, it is possible to deduce the analytical form of distributions. That is, it is possible to deduce a function, known as "probability distribution function", capable of finding, from a small number of parameters, the probabilities of each state of an attribute.

It is important to describe mechanisms underlying randomness and their distribution functions. In this way, we can estimate probabilities analytically. This chapter shows the most basic mechanisms, and, in later chapters, we introduce time-driven mechanisms, both discrete and continuous. The three most basic random mechanisms are:

1. the Bernoulli process, which explains and describes the distribution known as "Binomial", which shows the number of outcomes in series of events.
2. the Poisson process, which explains distributions with the same name associated to counts per unit time, and
3. the process known as "Central Limit Theorem", which explains and describes randomness in many scale attributes with central values.

3.2 Bernoulli and Poisson processes

The Bernoulli process and the corresponding Binomial distribution applies to binary attributes where the probability of response is *p* (non-response is *1-p*). This process allows calculating the probability of obtaining *k* responses in a series of *N* draws or cases. For example:

- What is the probability of, in 7 newborns, *k* of them being male? In this case *N = 7* and *k = 0, 1, 2, … 7*. As for elementary probabilities *p = 1 - p = 50%*.
- What is the probability of *k* responses in 10 subsidized firms? In this case *N = 10*, *k = 0, … 10*. Let's suppose that *p = 80%* and thus *1-p = 20%*.

Binomial distribution answers the above questions for all possible values of *k*. In the case of newborns, *k* can be 0, 1, 2, 3, 4, 5, 6 and 7. Thus, there are 8 classes (N + 1). Classes with the highest frequency are classes 3, 4 and 5. Classes associated with 0, 1, 6 and 7 newborns being of the same sex in 7 cases, have much lower frequencies.

When *p*, the underlying elementary probability, is clearly different from *1-p*, then the Binomial distribution predicts higher frequencies associated with initial or final classes. In the above example of number of responses in 10 patients treated where *p = 80%* high frequencies would most likely be found in the k = 8 and nearby classes. In general, the closer *p* is to 50%, the more symmetrical the Binomial distribution is. After a series of *N* experiments, *k* responses are observed. This number of responses is a random variable *X* with values ranging from 0 to N. *X* has the following probability distribution:

$$P(X=k) = \binom{N}{k} p^k (1-p)^{N-k}$$

where $\binom{N}{k}$ indicates the combination of *N, k* to *k*:

$$\binom{N}{k} = \frac{N!}{k!\,(N-k)!}$$

for *k = 0, 1, 2, ..., N. P (X = k)* is the probability of *k* responses in a series of *N* experiments.

To "parameterize" or to "model" is to replace a large collection of observations by a small number of parameters of the analytical expression describing regularities found in such collection of observations. Models show, in a simple yet approximate way, the most important features of data. The Binomial distribution has a mean value of *N p* and a variance of *N p (1-p)*. These are its parameters. To model a collection of real-world observations we replace such observations by values derived from the math equations above using two parameters.

k	p=50%	p=20%	p=80%
0	0,10%	10,74%	0,00%
1	0,98%	26,84%	0,00%
2	4,39%	30,20%	0,01%
3	11,72%	20,13%	0,08%
4	20,51%	8,81%	0,55%
5	24,61%	2,64%	2,64%
6	20,51%	0,55%	8,81%
7	11,72%	0,08%	20,13%
8	4,39%	0,01%	30,20%
9	0,98%	0,00%	26,84%
10	0,10%	0,00%	10,74%

This table is the result of applying the Binomial formula for different *p* and *k*. It shows the *P (X = k)* probabilities of observing *k = 0, 1, 2, ... 10* responses in the series of *N = 10* experiments when the elementary probabilities are *p = 50%, p = 20%* and *p = 80%* . Observing the column *p = 50%* it is concluded that the probability of, for instance, a couple having 10 boys in a row is *0.1%*: there is one chance in a thousand of this happening. Although rare, it is not impossible.

- / -

Poisson processes are observed in the count of events per unit time. Knowledge of the Poisson distribution allows answering questions like these:

- What is the probability of up to 8 loan defaults per month?
- What is the likelihood of having to meet 12 or more telephone calls per hour in the bank's hotline?
- What is the probability of observing up to 5 bankruptcies per year?

The probability *P (X = k)* of observing *k = 0, 1, 2, 3, ...* events within a given time interval is expressed mathematically as:

$$P(X=k) = \mathrm{M}^k \, exp\,(-\mathrm{M}) \,/\, k!$$

where M is the average number of events in the same time interval. This is the Poisson distribution, which has a mean of M, the expected frequency per time interval, and a variance equal to the mean. Thus, the Poisson distribution has just one parameter.

k	M=2	M=3	M=4	M=5
0	13,53%	4,98%	1,83%	0,67%
1	27,07%	14,94%	7,33%	3,37%
2	27,07%	22,40%	14,65%	8,42%
3	18,04%	22,40%	19,54%	14,04%
4	9,02%	16,80%	19,54%	17,55%
5	3,61%	10,08%	15,63%	17,55%
6	1,20%	5,04%	10,42%	14,62%

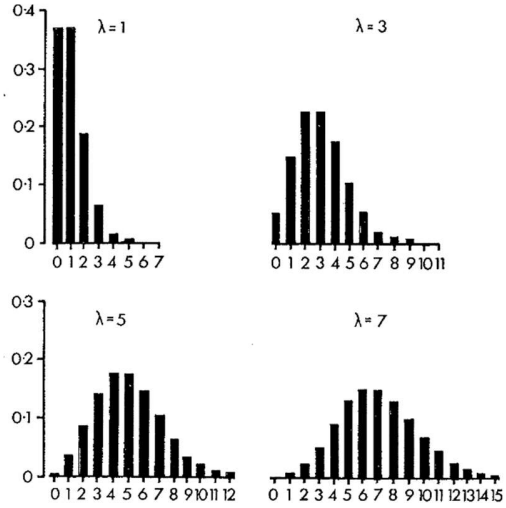
7	0,34%	2,16%	5,95%	10,44%
8	0,09%	0,81%	2,98%	6,53%
9	0,02%	0,27%	1,32%	3,63%
10	0,00%	0,08%	0,53%	1,81%

The table shows results from applying the above equation to $k = 0...10$ possible states for M frequencies ranging from 2 to 5 states per time interval.

Any time interval can be used: minute, hour, day, month, or year. For instance, what is the probability $P(X = k)$ of observing $k = 0, 1, 2, ...$ states during a given time interval when the expected frequency of such states, for the same range, is M = *2, 3, 4* or *5*? As shown in the table there are 8 chances in one thousand that up to 10 defaults may occur in each week, even when the expected frequency of such defaults is only 3 per week.

The sum of all the $P(X = k)$ for a given M should equal to 1. In the table above this sum is less than 1 since classes $k = 11, 12, ...$ are absent. These classes are associated with probabilities that, although small, are not negligible especially for M larger than 3.

To answer questions such as: what the probability is of experiencing 8 or more defaults, it is worth noting that this probability is equal to 1 minus the probability of experiencing seven or less defaults. The latter is obtained by adding the probabilities associated to classes $k = 1, 2, ... 7$. The figure below shows he Poisson distribution for values of M = 1, 3, 5 and 7 The symbol λ (Lambda), a Greek letter in small case, is used here for M.



When, in the e Binomial distribution, the average number of responses k is much smaller than N, then this process approaches what is known as the Poisson process. This may happen when p, the elementary probability of response, is much smaller than 1, and N is large. Therefore, the Poisson process can be approximated, in some cases, as a Binomial process.

3.3: Central Limit, Normal distribution function

Anyone who performs data analysis implicitly accepts that there are regularities in the data being analyzed and it is possible to discover such regularities. Regularities, once discovered, allow "modeling", that is,

> using one formula with parameters instead of the whole collection of cases to explain part of observed variability, but not all of it.

The most basic, important modeling task consists of finding the probability distribution associated with each attribute to use, and then replace it with an appropriate formula that requires not many parameters. For example,

> Instead of using the distribution of market returns observed in 120 months, only one parameter is used, the central value or mean. Thus, after modeling, each return, *j* is explained as a mean value plus a deviation from that mean:
>
> $$return\ of\ month\ j\ =\ mean\ return\ +\ deviation\ of\ j\ in\ relation\ to\ mean$$

Instead of a distribution, only one parameter is now used to approximate return. After modeling, part of the variability is explained, and part is unexplained. In the above example, return of month *j*, which may be 18 percent, has an explained portion (the mean return of of 17.5 percent), and another portion that remains unexplained (the 0.5 percent that month *j* has in excess of the mean return).

- / -

Thus far, we have mentioned probabilities associated with discrete attributes with a small number of states. In such cases, states have associated probability. How can we describe the likelihood associated with events that can take on endless values, as in the case of continuous attributes?

If the number of states of a given attribute tends to infinity, then probabilities associated with each of them must tend to zero: what is the probability of observing an object with a height of exactly one meter and 70 centimeters? It is zero. Even if we could find someone with exactly 1.70 meters, he or she could not tell anything in terms of relative frequency. There are billions of other cases with heights that are not exactly this value.

For continuous attributes, therefore, it only makes sense to speak of probabilities of intervals, not individual states. What is the probability of finding a height contained in the range 1.60 to 1.70 m? This makes sense and such probability is not zero.

In the case of interval scales and specifically for any continuous attributes, the notion of distribution can be naturally adapted to fit the fact that now there is an endless number of possible states. Imagine a distribution where

1. the number of objects increases to infinity and, as consequence,
2. it is also possible to increase the number of intervals classes to infinity.

Instead of a collection of probabilities, the distribution thus obtained is a continuous function where each state (which now is a continuous value) is associated with a probability "density" function. The area defined by an interval in that function will be probability.

These probability density functions are known as PDF and the most common of them is the "Normal" (or Gauss) function, often associated with continuous attributes such as heights, returns or IQs.

The Normal density function stems from a random process known as the "Central Limit Theorem". In a Bernoulli process, N is the number of experiments carried out in each series. When a coin is tossed in series of seven times each, *N* is 7. After repeating these series, we observe that mean number of times of a given outcome is $N\ p$ with variance $p\ (1\text{-}p)$ where p is the underlying elementary probability. Of course, the Binomial distribution only cares for small values of *N*: never more than 20. If we consider series of experiments with *N* larger than 20, and if we calculate the average number of responses, it turns out that the distribution of mean responses approaches the Normal, whatever the distribution of *p* may be. Even where original distributions are skewed, the distribution of their averages is Normal. This is called the Central Limit Theorem.

The random process stemming from collecting many mean values and observing their distribution is also called the same way.

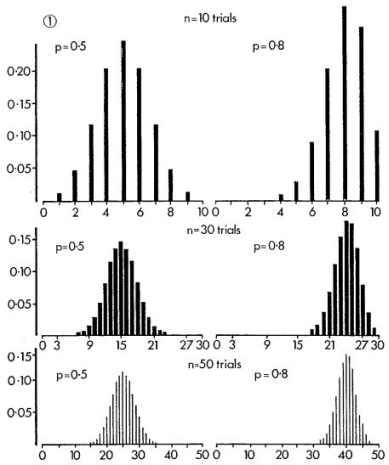
The figure in the coming page shows how increasing *N* makes the Binomial distribution more and more like the Normal distribution, however skewed *p* may be. Both for $p = 0.5$ (50%) and for $p = 0.8$ (80%) the effect of increasing *N* from 10 to 50 is the same.

The importance of the Central Limit Theorem stems from the fact that it shows that the distribution of attributes resulting from various additive influences tends to be Normal no matter what distribution, which these influences may have. This explains why so many phenomena closely follow the Normal distribution.

Mathematically, the density probability function $p\ (X=x)$ associated to *x* is, for a Normal random attribute *X*:

$$p\ (X=x) = \frac{1}{\sqrt{2\pi s^2}}\ exp\left(-\frac{(x-M)^2}{2s^2}\right)$$

where *M* is the mean and s^2 is the variance of *X*. The meaning of *p* is, as mentioned, that of a density or thickness of probability. *p* indeed indicates the probability of finding cases in the neighborhood of *x* but its meaning is not exactly that of probability *P*. It may be interpreted in a manner like *P* since the higher *p* is, the more likely events in the neighborhood of *x* will be.



The Normal distribution is symmetric: the average *M*, the mode and the median have the same value. *M* and s^2 completely determine the Gaussian distribution. Normal distributions thus are described by two parameters. The quotient

$$Z = \frac{x - M}{s}$$

is known as the "standardized" *x* or "Z-score" of *x*. *Z* is a replica of *x* where the average is made to be zero and the variance is made to be 1. By standardizing an attribute, it becomes comparable with others, which have also been standardized and can be directly used to obtain probabilities associated with intervals.

- / -

In the case of continuous variables, the probability distribution is no longer useful to calculate probabilities. Instead, there are now two functions.

1. probability density function PDF already mentioned, and
2. "Cumulative density function" known as CDF, which is the integration of the density function.

It should be understood that

1. Density function is a generalization of the frequency distribution concept. We interpret it intuitively in a similar manner but with the caveat that frequencies are no longer probabilities. They are just density, intensity, strength, or thickness of the probability in the neighborhood of each observation.
2. Cumulative distribution function is a generalization of the cumulative frequency distribution, and we interpret it in the same way.
3. When attributes are continuous, it no longer makes sense to talk about collections of counts and thus distributions. Density functions are true functions.

- / -

How to estimate probabilities associated with intervals? Given state *x*, we standardize it by subtracting the mean and dividing the result by the standard deviation. This yields a *Z*. Then a Normal table such as that presented below, will give the probabilities.

Normal tables show density and distribution function (P) values for different Z.

1. P is the probability of finding an occurrence smaller than *Z* (standardized *x*).
2. the probability of occurrence of any value greater than *x* is 1 - P.
3. the probability associated with any interval x_1 to x_2 is given by the difference between the probabilities P (Z_2) - P (Z_1).

Z score	PDF (density) %	CDF (cumulative) %
-4	0,01%	0,00%
-3,5	0,09%	0,02%
-3	0,44%	0,13%
-2,5	1,75%	0,62%
-2	5,40%	2,28%
-1,5	12,95%	6,68%
-1	24,20%	15,87%
-0,5	35,21%	30,85%
0	39,89%	50,00%
0,5	35,21%	69,15%
1	24,20%	84,13%
1,5	12,95%	93,32%
2	5,40%	97,72%
2,5	1,75%	99,38%
3	0,44%	99,87%
3,5	0,09%	99,98%
4	0,01%	100,00%

Suppose we wish to calculate the probability associated with the interval 45-50 kg in an attribute whose average is 55 kg and standard deviation is 5 Kg. Z-scores corresponding to 45 and 50 Kg are -2 and -1 respectively. According to the table above, probabilities associated with Z of -2 and Z of -1 are 2.28% and 15.87% respectively. The difference between these probabilities, 13.59% is the probability of finding cases in the interval.

In the case of an interval defined by $Z_1 = -1$ and $Z_2 = +1$:

$Z_1 = -1$ P (Z_1) = 15,86%

$Z_2 = +1$ P (Z_2) = 84,13%.

as the table above shows. Thus, the probability of finding an occurrence in such interval is 84,13 – 15,86 = 68,27%. There are approximately 68 chances in 100 of finding cases with an attribute,

which is smaller than 1 standard deviation above the mean and bigger than 1 standard deviation below the mean.

The same calculations can be made for intervals of -2 to +2 and -3 to +3 standard deviations. Probabilities associated to such intervals are relevant: the probability of observed values in the range.

$$M - 2s \; and \; M + 2s$$

is around 95%. That is, the probability of observing an x bigger, in absolute terms, than two standard deviations above or below the mean is 1 - 95% = 5%, 2.5% for each side.

The probability of observed values within the range

$$M - 3s \; and \; M + 3s$$

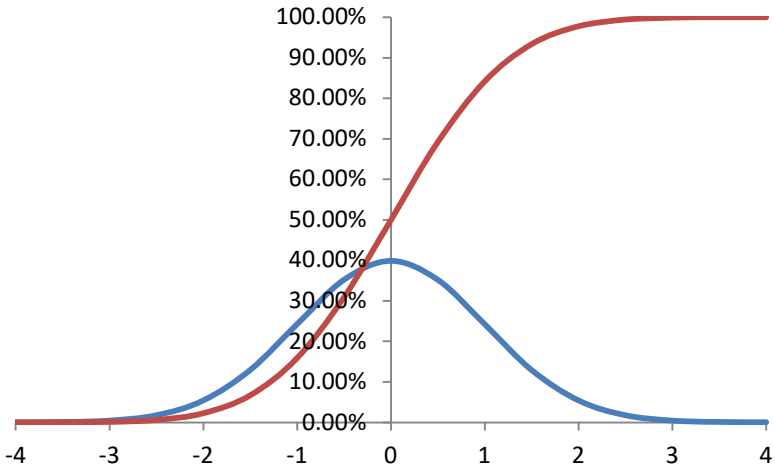
is nearly 99%. Thus, the probability of observing an x bigger, in absolute terms, than three standard deviations above or below the mean is 1 - 99% = 1%, 0.5% for each side.

When, for instance, mean height is 170 cm and standard deviation is 15 cm, there are only 5 chances in 100 of finding objects taller than 200 cm = 170 + 2 * 15 cm or shorter than 140 cm = 170 - 2 * 15 cm.

Graphically, Normal density and distribution functions look as below.

Burgundy-colored function is the distribution function. Such function is obtained by accumulating areas covered by the blue function, the Normal Density Function.

Thus, the area under the Normal probability density function (PDF in blue below) from minus infinity to zero (average value) is 50%. It is at P = 50% where the distribution function crosses the y-axis where z = 0. The area under the whole of the Normal function is 100%



Normal functions show probability densities with no immediate use. What is useful in practice is the integration or accumulation of these densities for a specific interval, and this accumulation is not analytically immediate. You cannot write it as a formula. But any table or computer program will give the value of the desired CDF probability.

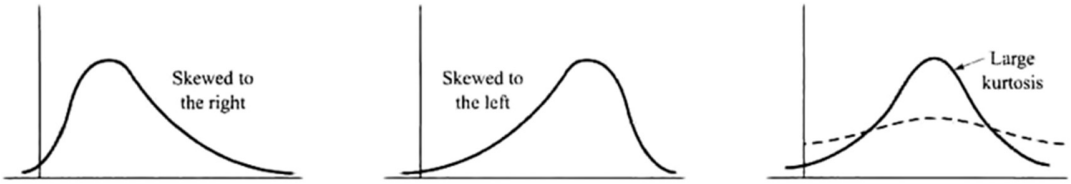
Values obtained from such tables or computer programs are the probability of finding an occurrence lower than the input value. Once obtained, the distribution function shows, for direct reading, how likely are events smaller or equal to any input value.

- / -

The Normal function is a benchmark against which other distributions are compared. So, in statistics, it is important to describe any deviation that is observed in the distribution of an attribute in relation to the Normal function.

Deviations which can be observed in relation to the normal form are of two kinds:

1. Asymmetry (skewedness)
2. Kurtosis.



Each of these two forms of divergence allows, in turn, two types:

- asymmetry might be to the right (positive) or to the left (negative).
- kurtosis can be a narrowing of intermediate or extreme values.

Distributions asymmetric to the right are common and indicate growth, accumulation, a sequence of dependent events or instability. The Lognormal function is of this type. In an attribute asymmetrically distributed to the right, the likelihood of large occurrences is greater than in the Normal case. Asymmetry to the left is typical of attributes whose randomness is truncated. This is the case of proportions, which are limited to 100%.

Kurtosis, by narrowing the intermediate zone, causes peaks and tails more pronounced than those in the Normal function. This is observed in ratios of Lognormal attributes. The name used to designate it is “leptokurtosis”. This type of distribution is also associated with rare events. It is perhaps the most difficult of all distributions to model exactly. Although there have been many attempts to parameterize this type of functions, success is rare.

Kurtosis by narrowing extreme values occurs when the attributes are limited to a given range. It's the case of probability itself, which can only have values between zero and 1. These distributions have peaks and tails less pronounced than in Normal function. They look like bowls. The name used to designate these forms is “mesokurtosis”.

Both asymmetry and kurtosis are quantified by comparison with the Normal function. The interpretation of these two indicators that measure distortion (one for skewedness and kurtosis to another) varies depending on the statistical tool used. In the SPSS tool, for instance, the skewedness and kurtosis should be considered excessive when such indicators have absolute values above 1. Moreover, for the same asymmetry measure provided by SPSS, skewedness will have the value of zero for the normal function (that is, perfect symmetry of the distribution) and will be

1. The more negative the skewedness value is (the greater the tail to the left),
2. The more positive the skewedness value is (the greater the tail to the right).

In the case of kurtosis, the indicator will be set to zero for Normal function and will be

1. Positive for leptokurtosis (exaggerated peaks and tails),
2. Negative for mesokurtosis ("bowl" with no peaks or tails).

Most random phenomena obey one of the three noted above distributions - Binomial, Poisson and Normal. There are however many other distributions, which are used in specific circumstances, namely in inference, that is, the testing of hypotheses. For example, the

- "t" distribution
- Fisher's F distribution
- Chi-Square distribution

are widely used.

The *t*, also known as “Student’s *t*” distribution applies the Central Limit Theorem to the case of small samples (less than 30-40 observations) and its importance comes, not just from its use in determining levels of significance.

The Fisher’s F distribution is widely used in the testing of hypotheses in regressions, analyses of variance and many other cases.

The Chi-Square distribution approaches the Normal when comparing proportions.

Exponential, Uniform, Weibull, Gamma, Lorenz, Hypergeometric, etc., approximate specific processes and relationships. Multinomial, Multinormal, etc. are generalizations of the above, for sets of related attributes.

- / -

One should never forget the cost and usefulness of the event whose likelihood one is trying to assess. The simple knowledge of probabilities is rarely a conclusion of any analysis and leads to no practical results. Probabilities are one step among others to get to determine costs, risks, utilities, the consideration of which lead to the best decision.

Chapter 4 Inference, mean comparison

4.1 Inference, sampling

We have learned to recognize the different types of data and the relationships that can exist between them. We know what a random attribute is (unpredictable states), and we can describe the basic random mechanisms. We are well acquainted with distributions and their parameters. We have also noted the difference between what is observed, and what is expected to occur.

Now we are going to open a new line of thought: we shall start thinking in a way that has not been explored before. The idea with which we open this new line of thought is “inference”. To infer is to generalize or extrapolate.

Through inference we accept that regularities observed in a limited set of objects are also present in all the other objects with the same characteristics.

An explorer arrives on an island and spots groups of natives with feathers adorning their heads. He goes back a few more times and always observes natives wearing feathers. Thus, the explorer decides to make an inference: he accepts as valid the claim that those natives wear feathers, despite having observed only a small sample.

We are continually inferring. Without inference it is impossible to make decisions. Any classification of attributes in classes requires an inference. Of course, inference has its risks and can lead to wrong conclusions. Those natives, for example, may use feathers only when they are in the beach; or they use them to scare away the evil spirits, which boats of explorers inevitably bring. In their day-to-day life, they may never use feathers. Thus, feathers can even be caused by the explorer himself, an undesirable but common situation in our research.

Despite its dangers, inference is the basis of the scientific method and in many cases, it is the only available way of advancing knowledge. It is reliable if used with caution.

- / -

Statistics has contributed enormously to increase the reliability of inference. It was the Statistical science, in fact, which brought mathematical respectability to the inferences practiced in many seemingly less exact sciences such as Sociology or Economics. How did Statistics bring respectability to inference? Statistics can assign a confidence level to inferred values. But to understand what constitutes that level of confidence or trust, it is first necessary to distinguish between “sample” and “population”.

Population is the set of all objects that possess the characteristics under study. Sample is any subset of *N* objects belonging to the same population and obtained randomly. The population of bank customers is the set of all objects borrowing from the bank. A sample of 20 borrowers is a subset of the same population randomly obtained.

In most cases, populations are too widespread to be assessed or parameterized. Attributes such as sex, age, income, and others that may characterize a given population are thus impossible to observe directly.

The goal of the scientific method is none other than infer, as accurately as possible, characteristics of the population without having to study all cases. Statistical inference allows the extraction of conclusions about populations from the observation of small samples. It is impossible to observe all

default cases, but it is not impossible to observe a sample of 50 defaults where the condition of randomness is met.

-/-

To distinguish inferred or expected from observed values, we use the term “estimate” to designate the value or parameter that has been the result of an inference.

An estimated mean is inferred from a sample rather than from the whole population.

An inference brings with it

a statistical “confidence level” in the form of a probability.

Such confidence level, in turn, leads to an interval around the estimated value, which is the “confidence interval”.

For example, the average return of 23 percent enjoyed by banks in the Cayman Islands is estimated with a confidence level of 95%. This level is associated with a confidence interval for returns of 19 to 27. Therefore, there are 95 chances in 100 that the average bank return (that of the population), be contained within the interval 19 to 27 percent.

Confidence intervals become single values where the confidence level refers, not to an estimated parameter, but to an estimated boundary. A given estimation may lead to the following confidence level and boundary: there are 95 chances in 100 that the ROA of banks will not exceed 25%.

- / -

The first and most basic condition for being able to infer from a sample is that the sample be obtained “at random”.

A sample or subset of N cases belonging to a population is obtained at random when the probability of any member of the population be included in this subset is independent from the attribute under consideration and others that are related to it. For a sample to be random, therefore, attributes and relationships relevant to the study should be strictly independent from the probability of a given object being included in the sample.

When a random trial aims at comparing high return banks with normal return ones, the former cannot be more likely than other banks to get into the sample; if the matter being studied is the differences created among customers by the two sexes, one gender should not be more likely than the other to get into the sample.

To randomize is to obtain from a population a subset of N object such that such choice meets the requirement of randomness. There are two ways to randomize:

- random sampling or simply sampling
- random assignment

Random sampling is typical of sciences where experimentation is difficult, as is the case of Economics, Psychology and Sociology. It consists of picking objects at random from the whole population. To avoid correlation between the likelihood of an object being chosen and the attributes being observed, sampling must avoid hidden limitations to entry or other biases.

Random assignment is typical of experimental sciences, but it is being increasingly used in Finance. It consists of randomly assigning objects to two groups, and then compare the groups based on a group quality. The typical random assignment would consist of randomly determine which patients receive a drug and which receive placebo, but this can seldom work. Instead, “semi-experimental”

methods are used in Finance whereby the two groups are pre-existing qualities, and the assignment is not random.

4.2 Distribution of the mean, mean confidence interval

Think of an attribute, for example the Total Capital ratio (CAR). Mean CAR observed in a random sample of *N* banks is an estimator of the true, unknown, population mean.

Let us now replicate the random mechanism known as the Central Limit Theorem. First imagine several samples, all of them with N objects, randomly drawn from the same population. Each of these samples would have a different mean and such mean values would indeed represent different possible estimates of the true population mean.

The collection of mean values obtained above has a distribution. This is the distribution of sample means. According to the Central Limit Theorem, such distribution is Normal, no matter what the individual CAR distributions may be.

Let's assume that the various samples obtained from the same population have the same number of objects, *N*, and a similar variance, *V*. It can be shown that the distribution of their means has itself a mean value that is also a good estimate of the population mean. Moreover, the observed variance is an estimate of the population variance, *V*, divided by the number of cases *N* in each sample.

The standard deviation of the distribution of means is the “standard error of the mean”.

$$St.error = \sqrt{\frac{V}{N}}$$

where *V* is the observed variance of a sample, and *N* is the number of objects.

Standard errors measure the inaccuracy associated with an estimate, in this case, the mean. One should not mix up standard error with standard deviation, the latter measures variability around the mean, which is just one of the factors adding to standard errors.

The accuracy of an estimated mean can be measured by the length of the “one standard error confidence interval”:

$$Mean \pm St.error$$

Therefore, the error with which a mean is estimated depends on N and V: larger errors are associated with small N and large V. In statistics, the smaller the sample is, the larger the standard error will be, and the greater the variance, the greater the error.

If an attribute varies a lot, the sample needs to be bigger; if an attribute is not much varied, then the sample can be smaller. That is why 20 objects are enough to estimate the mean height with a reasonable level of confidence but, in the case of CAR or ROE, we need 40 objects to reach the same confidence level.

- / -

What is the confidence level? And how does it relate to the confidence interval? Mean obtained from the distribution of means is an estimate of the population mean. But an estimate can be more or less accurate. The standard error associated with the mean lead to the probability that the true mean of the population be within a given interval.

The distribution of means is Normal. Due to the properties of the Normal distribution, there are 95 chances in 100 of finding the population mean within the interval

$$Mean \pm 2 \times St.errors$$

Also, there are 99 chances in 100 of finding the population mean within the interval

$$Mean \pm 3 \times St.errors$$

These 95 in 100 (95%) or 99 in 100 (99%) are examples of different "confidence level" associated to an estimated mean.

In short, by sampling N cases at random and then observing the mean and the variance of the sample, it is easy to obtain, not only an estimate of the true population mean, but also the probability (confidence level) that such an estimate be within a given interval. Easy to compute intervals are thus:

- a 95% confidence interval is the range defined by the observed mean, adding and subtracting two standard errors (mean ± 2 standard errors).
- a 99% confidence interval is the range defined by the observed mean, adding and subtracting three standard errors (mean ± 3 standard errors).

Let's assume that the average income observed in a sample of N = 25 female adults is 67,000 money units and the observed variance of V = 343 squared money units. Then:

$$St.error = \sqrt{\frac{343}{25}} = 3.7\ money\ units$$

and the 95% confidence interval of the mean is defined by the following two values:

$$67{,}000 + 2 \times 3.7 = 67{,}007.4\ money\ units$$
$$67{,}000 - 2 \times 3.7 = 66{,}992.6\ money\ units$$

Therefore, the probability that the actual mean - the population's mean - is within the interval [67,007.41 and 66,992.6] is 95%. Yes, there are 95 chances in 100 that the true mean, which cannot be observed, is contained within the above range. Or, if you prefer, for 95 in every 100 times that we accept as good, trustworthy, the estimated mean, we are not being cheated. For 5 in 100 times that we accept that mean, we are cheated.

- / -

The higher the confidence level, the bigger the interval. To have a confidence level of 99% (99 chances in 100 of the true mean population being within the interval) it is now necessary to multiply by three, not by two, the standard error and add / subtract it to the estimated mean. This increases the interval. A certainty or 100% confidence level would require an interval of infinite length, which is a tautology.

It is necessary to find a balance between the length of the confidence interval and the confidence level associated with it. This balance was fixed by tradition in the 95% confidence level. 95% eventually become the confidence level above which it is usual to accept estimated values. In practice:

1. When there are 95 or more chances in 100 of something happening, it is a tradition to accept that such something can indeed happen.
2. When there are less than 5 in 100 chance of something happening, it is a tradition to accept that such something may not happen.

With this tradition, Statistics has built a concrete threshold for distinguishing what is credible and what is not. For instance,

1. an estimation is accepted when the associate confidence level is 95% or higher; note, however, that the fact that an estimate is accepted does not mean that it is the true value, but just that it is a credible estimate.

2. an estimate is rejected as improbable when the associate confidence level is less than 95%; note, however, that improbable does not mean impossible, it means that there is not enough evidence to accept it.

Note also that 5 chances in 100 signify something that, far from never happening, does happen five times in 100 on average. Other sciences or circumstances may dictate the need to work with higher levels of confidence. In Banking, namely, specific confidence levels may be much higher.

4.3 Difference between two means, the null hypothesis, the t-test.

We have 60 clients and assign 30 to group A randomly; the remaining 30 go to group B. Then we make a promotional offer to A and nothing to B. After a time, we measure the attribute we want to track and calculate means and variances for each of the two groups. It turns out that attribute being assessed, e. g. the number of complaints, has decreased in average for A clients. Can the difference between the two mean values be due to chance, poor randomization, or other factors? What is the confidence one can have in that the observed difference between the two means is caused by the promotion?

The mean confidence interval allows answering this question. Consider,

not the distribution of means,

but the distribution of the difference between two means.

From this point on, we are already dealing with "hypothesis testing".

Given two classes A and B (promotion and non-promotion clients), we want to know whether observed difference between the two means M_A and M_B may or may not be plausibly caused by promotion. Samples have N_A and N_B cases.

If there were in the population no difference between A and B, the distribution of the difference between the two means would have a mean of zero, a variance equal to the sum of the variances of A and B, and a standard error of such mean difference of

$$St.error\ of\ dif\ |M_A - M_B| = \sqrt{\frac{V_A}{N_A} + \frac{V_B}{N_B}}$$

Remember, even when variables subtract, not add, their variances always add. The 95% confidence interval would be the two standard errors above and below the zero value:

$$95\%\ conf.interval\ for\ dif\ |M_A - M_B| = \pm 2\ Standard\ error$$

thus

1. When, in the population, the difference between means M_A and M_B is zero, there are 95 chances in 100 of observing a difference, in absolute value, smaller than two standard errors of the distribution of this difference, above and below zero.
2. When, in the population, the difference between the two means M_A and M_B is zero, there are only five chances in 100 of observing a difference between mean values greater in absolute value than two standard errors of the distribution of this difference, above and below zero.

The practical conclusion to be drawn is,

when the observed difference between two means is smaller, in absolute value, than two standard errors of the distribution of the differences, then we are not allowed to reject the "null hypothesis" that this difference is zero in the population.

Only when the observed difference is greater in absolute value than two standard errors, are we be allowed to reject the "null hypothesis" that such a difference is zero in the population.

Consider family income growth. We observe two samples: the first, with 12 families, is from a rural hinterland; the second, with 29 families, comes from a fishing village in the coast. Will it be possible to infer that there are differences between the income growth (IGR) of the families on the coast and in the hinterland? IGR have mean and variance:

Zone	N	Mean	Variance
Hinterland	12	8.17	5.79
Coastal	29	14.59	18.39

From the above it is possible to test the hypothesis that the difference between IGR in the two zones does not really exist in the population (null hypothesis). The difference between means is 6.42 and, using the above we see that the standard error associated with this difference is 1.06. Twice 1.06 is 2.12. Since the difference between the means is outside the range [+2.12; -2.12] we can discard the hypothesis: there are less than 5 chances in 100 of observing a difference as large as 6.42 when, in the population, such a difference does not exist. And, in so rejecting the null hypothesis, we are authorized to accept the differences as being significant.

Intermediate calculations:

Zone	N	Mean	Variance	Standard error of the mean	95% Confidence interval of the mean		Difference between means	Standard error of the difference between means	95% Confidence interval for the difference between means		
Hinterland	12	8,17	5,79	0,69	9,56	6,78					
							6,42	1,06	2,11	-2,11	It is outside
Coastal	29	14,59	18,39	0,80	16,18	12,99					

- / -

The above assessment of the significance of differences in IGR is an example of testing the null hypothesis. Its application extends beyond the test of likelihood of differences between two means. In Statistics, the null hypothesis is the tool used for inference, that is, to infer about the population from samples.

The methodology used to test the plausibility of null hypotheses follows these steps:

1. Differences among two classes of an attribute are observed. They may refer to mean values or other parameters. Degrees of freedom at play are also assessed.
2. The probability of observing as large a difference as that observed in the sample when, in the population, such difference is zero, is then calculated using the appropriate distribution of differences.
    a. If this probability is less than 5%, then the null hypothesis of differences being zero in the population is discarded. This leads to inference.
    b. If this probability is more than 5%, the null hypothesis cannot be rejected, which precludes inference.

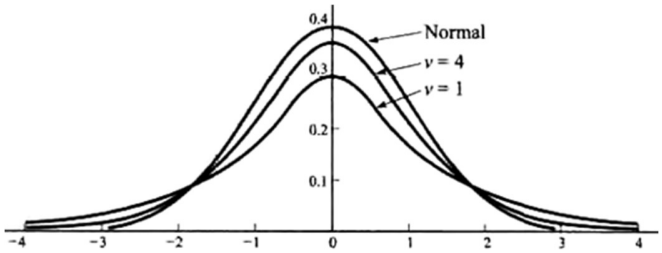
The example of inference just explained, involving mean values, is known as “t-test”.

4.4 Model design, error types, sample power and one tail tests

This section contains complementary, important information.

When a sample is small (N<30) it is no longer possible to estimate variances from samples validly. Hence the approximation that assumes a normal distribution for the difference between means is not tenable.

Instead of the Normal, it is then used a distribution known as "t" or "Student's t" which tends to Normal with sample sizes greater than 30. For values of N lower than 30 the "t" distribution allows estimating variances from samples. The variance of an attribute which obeys the distribution "t" is N / (N-2) with N degrees of freedom involved. Below, t distributions with different degrees of freedom (v) are compared to the Normal.



number of deviates from mean:	-5	-4	-3	-2	-1
Normal distribution	0,99999	0,99997	0,99865	0,97725	0,84134
t distribution, d.f. (v) = 6	0,99880	0,99644	0,98800	0,95379	0,82204
t distribution, d.f. (v) = 4	0,99625	0,99193	0,98003	0,94194	0,81305

For instance, in the Normal case, the probability of a case being 4 standard deviations or more below the mean is 1 – 0.99997 = 0.00003: such values may be observed, in average, 3 times in 100,000 cases. For the t distribution with v = 6 the same probability is 1 – 0.99644 = 0.00356: such values may now be observed 356 times in 100,000, a lot more. In the assessment of bank risk, this is a crucial subject to consider.

- / -

We now introduce the two basic types of design: "between" and "within" groups.

Of 60 clients, 31 of them are randomly assigned to group A while the remaining 29 go to group B. Then a promotion is made to A but not to B. After a time, we measure the attribute in question and calculate means and variances for each of the groups. It is found that, in A clients the mean attribute, for instance, the number of complaints, is smaller than the mean B. By applying the t-test we can then reject or not reject the null hypothesis that mean values in the population are equal. This experimental design is called "between groups" test, but it is also known variously as "separate groups" test, "independent groups" test or "one-variable" test.

The same trial might have been carried out with a different design: we first count previous complaints on all the 60 clients and calculate mean and variance. Then, all clients receive the promotion, and, after a time, new complaints are counted, and the new mean and variance is calculated. The two means and two variances (before and after promotion) can then be compared. This second experimental design is called "within groups" test or "groups of mutually dependent" test or "two-variables" test.

The test of significance of the difference between two means equally applies to the two designs. It is valid either when the sample is divided into two groups or when, using the same group, an attribute is observed for two different moments in time. In the latter, the number of objects is the same; the test is performed on pairs of variables. In the former, the test deals with groups, categories, or classes for the same variable.

- / -

When, on objective grounds, it is not possible for an attribute to have their respective role reversed, then, when assessing confidence levels, we should count half of the distribution only. This is the case of studies that measure the increase in the height of children. The second measurement made on the same group can only show higher averages than the first measurement, never smaller, because growing children cannot regress. There are other instances in which this may occur. When using hypothesis tests, only one tail of the distribution should be considered, not both.

When it cannot be assumed that samples have the same variance, the test of differences between means (t-test) must be adjusted by separately calculating these variances.

The statistical tools used to carry out t-tests have options allowing the user to define one tail testing and different variances testing.

- / -

When inference is made, two types of errors may be made, known as:

1. Type I error: to accept that there are differences when there are not. Claiming for example that there are differences between two means when, in the population, means are equal.
2. Type II error: to accept that there is no difference when there is. For example, say that means cannot be viewed as different when, in the population, they are.

Type I error is what we are most tempted to commit. It consists of seeing differences where there is none or seeing relationships where there are none. Type II error, by contrast, consists of calling equal to what is different or failing to see a relationship.

Significance tests, namely tests of the null hypothesis, tell us what exactly the probability is of seeing differences (or relationships) when they do not exist in the population, only in our sample. A confidence level of 95% would indicate a probability of error of 5%. This is therefore the probability that we are committing a Type I error when we accept a difference, that is, when we reject the null hypothesis.

- / -

The size of a sample, that is, its number of objects, greatly determines the possibility of making inferences and the type of error made:

1. A large sample tends to find significant differences and relationships that, in fact, exist; but which are not of practical importance and should be ignored.
2. A small sample may not show as significant, differences of practical importance, not because these differences are small but only due to sample size.

We should never be confined to significance tests. We should also build a model and compare it with reality. Even when two means are significantly different, we should examine the standard error of the difference in the light of real-world experience.

- / -

Costs associated with Type 1 and Type 2 error may be very different. First, it must be understood that a probability says nothing about the cost associated with the respective uncertainty. If we place 100 cups on a table, 99 of these have wine and one has poison, the random choice of one glass amongst the 100 means that there is one chance in 100 of immediate death. It is possible to quantify the expected benefit from playing this game, given the prize money they offer me to play it. But if instead of poison it is water or if, instead of a prize of 1 million, the reward is small, then everything changes.

Cost associated with type I error may be small when it leads to taking actions which are useless. Cost associated to denying useful drugs or tests due to a type II error can be large. However, this logic is

inverted, e. g. when it comes to taking irreversible actions without any benefit (radical mastectomy or other mutilations).

- / -

The “power” of a sample is its ability to reveal relationships that may exist. In practice, the power depends on the number of objects: the more objects, the more powerful a sample is. A randomized trial should have sufficient power to find what it is expected to find. It should also be able to uncover meaningful relationships in the study if they exist. If a trial has only 40 clients but a minimum of 60 clients are needed to make inferences about parameters or relationships under study, then this study is a waste of time.

Chapter 5 Comparison of proportions, the Chi-Square

We know how to compare mean values and how to test whether the mean difference is significant or not (whether such difference can be trusted as existing in the population).

Can we perform the same test when comparing, not mean values, but frequencies, that is, counts? In Finance, counts and proportions are often used. For instance, we compare the number of responses of clients who received a promotion offer to the frequency of responses of clients who did not receive any offer. We compare the frequency of one of the sexes with respect to the total, or to the other sex. We perform these comparisons using proportions (relative frequencies): 78% response or 34% male.

5.1 Confidence interval for proportions

If we want to compare counts, first we need to estimate the confidence interval for counts or proportions (absolute or relative frequencies): four responses in 10, 6 male clients in 8. The Binomial distribution can determine the likelihood of counts above and below that observed and allow us to test null hypotheses. But to work with the Binomial distribution is computationally costly, especially for large samples. Therefore, it is customary to use the Normal distribution as an acceptable approximation in this case.

It is usual to accept a Normal approximation of the Binomial distribution when

$$N\,(1-p)\;>\;5$$

For example, if *N* = 20 and the probability of response is 80%, then 20 (1 - 0.8) = 4 < 5, which leads us to conclude that the Normal approximation is not valid in this case – we should use the Binomial distribution and the so-called exact tests. But if we could get 25 clients instead of 20, then we would be able to use the Normal approximation.

Note however that in cases where the Normal approximation is valid, the frequency of responses is a discrete random variable while Normal occurrences are continuous. Therefore, frequencies may lose their meaning as in “seven clients and a half”.

- / -

The Normal distribution Binomial approximation has the following parameters:

1. The mean is M = *N p*.
2. The variance is V = *N p (1-p)*.

Given 40 clients and a probability of response of 80%, instead of using the Binomial distribution, we can use a Normal approximation where the mean is 40 times 0.8 (=32) and the variance is 32 times 0.2 (=6.4).

Under this Normal approximation the standard error of the mean is

- when working with counts (absolute frequencies) $St.error = N\sqrt{\frac{p(1-p)}{N}}$
- when working with proportions $St.error = \sqrt{\frac{p(1-p)}{N}}$

From this point on, the process of determining the confidence intervals is the usual:

The 95% confidence interval is M ± 2 standard errors,
and the 99% confidence interval is M ± 3 standard errors.

A conservative way to determine the confidence interval for proportions consists of assessing the standard error just as $\sqrt{\frac{1}{N}}$

In a trial involving 43 clients the expected proportion of responses is p = 80%, thus the average expected responses are M = 43 × 0.8 = 34.4. Since 33 responses were obtained, will this value be within or without the 95% confidence interval?

The standard error is, in this case, $43\sqrt{\frac{0{,}8(1-0{,}8)}{43}} = 2{,}62$ and twice this, is 5.25

Two standard errors above and below M = 34.4 will give the confidence interval

$$34{,}4 - 5{,}25 = 29{,}15 \quad \text{and} \quad 34{,}4 + 5{,}25 = 39{,}65$$

The 33 responses are therefore well within the confidence interval of 95%.

5.2 Difference between proportions, the Chi-Square test

Once a confidence interval for proportions is known, it becomes possible to compare two proportions and to assess the significance of their difference.

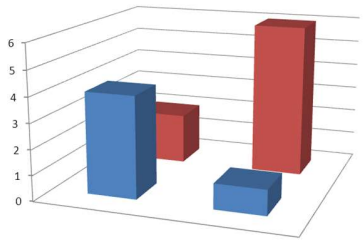
The commonly used method is to arrange frequencies in "crosstabs" and then measure the difference between the two proportions there, based on a Normal approximation.

We want to test two promotions to ascertain which elicits more responses from clients. N = 13 clients are divided in two groups. One gets promotion "a" and the other gets "b". The "a" group has 6 clients, and the "b" group has 7 clients. At the end of the test, responses (positive, "p") and non-responses (negative "n") are:

	Observed frequency of positive (p)	Observed frequency of negative (n)	Marginal total (row)
Frequency of "a"	ap = 4	an = 2	a = 6
Frequency of "b"	bp = 1	bn = 6	b = 7
Marginal total (column)	p = 5	n = 8	N= 13

Of the 6 "a" clients, there were 4 (66.6%) responses. Of the seven "b", there was only 1 client (14.2%) who responded. Is it possible to find a difference between proportions as large as this (from 66.6% to 14.2%) when, in the population, the difference between "a" and "b" is nil (null hypothesis)?

What matters here is whether a promotion can change responses or not. An interaction between promotion and response should exist, otherwise the promotion is not effective in changing responses. Cell frequencies above are depicted graphically as



Is the difference between the proportion of clients who respond and those who do not respond inside or outside the 95% confidence interval for a difference of zero?

If it is inside, we cannot reject the null hypothesis that such difference, observed in the sample, does not exist in the population.

If it is outside, we may reject the null hypothesis. We are allowed to believe that such difference may have come from true differences in the population.

- / -

To test the null hypothesis, we use the Chi-Square test where observed frequencies in each cell of a crosstab are compared with the frequencies that would be expected if, in the population, there were no differences.

The mere consideration of the number of states (“marginal”) lacks practical interest. What is sought with crosstabs is whether, yes or no, the existence of certain categories significantly modifies the proportion of states in others.

Before crossing, each attribute had marginal frequencies:

1. Promotion: 2 classes, a = 6 and b = 7
2. Response: 2 categories, p = 5 and n = 8

The same marginal frequencies of 7 and 8 are obtained as a result of other combinations of counts in the cells. In the table below, for example, counts observed in cells are not the same as before yet marginal frequencies are the same.

	Observed frequency of positive (p)	Observed frequency of negative (n)	Marginal total (row)
Frequency of “a”	ap = 2	an = 4	a = 6
Frequency of “b”	bp = 3	bn = 4	b = 7
Marginal total (column)	p = 5	n = 8	N= 13

Therefore, knowledge of marginal frequencies is not enough to determine frequencies in cells. There is one source of variability and degree of freedom that allows for example a promotion “a” to reverse its response when compared with “b”.

Starting from the null hypothesis that the actual probability (population) of response “a” and “b” are the same - thus there is no difference between the proportion of responses for both promotions - what is the probability of finding in the sample a difference in proportions as large as the observed? Recall original counts:

	Observed frequency of positive (p)	Observed frequency of negative (n)	Marginal total (row)
Frequency of “a”	ap = 4	an = 2	a = 6
Frequency of “b”	bp = 1	bn = 6	b = 7
Marginal total (column)	p = 5	n = 8	N= 13

The observed proportion of responses (“p”) is 5/13. 5 out of N = 13 clients respond:

$$Pp = p / N = 5/13$$

Of these 13 clients, 6 are “a” and 7 are "b." Therefore, assuming the null hypothesis, the expected number of responses in “a” clients, “ape”, is:

$$ape = Pp\ a = 5/13 \times 6 = 2.3$$

and the expected number of “b” clients, “apb”, is:

$$apb = Pp\ b = 5/13 \times 7 = 2.7.$$

Note that 2.3 + 2.7 = 5 which is the frequency of clients who responded to promotion.

The observed proportion of non-responses, the “n”, is 8 in 13:

$$Pn = n / N = 8/13.$$

But 6 are “a” and 7 are “b”. Therefore, assuming the null hypothesis, the expected number of non-responses in “a” clients, “ane”, is

$$ane = Pn\ a = 8/13 \times 6 = 3.7$$

while the expected number of non-responses in “b” clients, “bne” is

$$bne = Pn\ b = 8/13 \times 7 = 4.3$$

Note that 3.7 + 4.3 = 8, which is the frequency of clients who did not respond.

Frequencies ape, bpe, ane, bne are called “expected frequencies” (Fe)

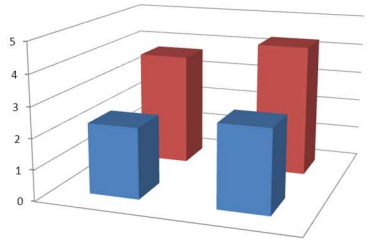
Frequencies ap, bp, in, bn (the original) are called “observed frequencies” (Fo).

Frequencies ape, bpe, ane, bne, were obtained using marginal frequencies only: of 13 total clients, 6 are “a” and 7 are “b”, 5 responded and 8 did not respond.

Frequencies ape, bpe, ane, bne are those that would be observed if promotions “a” and “b” had similar effects in the population, that is, if the null hypothesis were verified. Should this happen, then the crosstab above would be:

	Expected frequency of positive (p)	Expected frequency of negative (n)	Marginal total (row)
Frequency of (a)	ap = 2.3	an = 3.7	a = 6
Frequency of (b)	bp = 2.7	bn = 4.3	b = 7
Marginal total (column)	p = 5	n = 8	N= 13

with a graphical aspect



Remember how to calculate expected frequencies: multiply, for each cell, the respective marginal frequencies and then divide by N.

-/-

According to the Normal approximation to the Binomial distribution, the squared difference between the observed and the expected frequency in a cell,

$$(Fo - Fe)^2$$

shows how such cell deviates from the null hypothesis. $(Fo - Fe)^2$ is a deviation, not an error. It is not something that remains to be explained.

The “Chi-Square” of a crosstab is the sum, for all cells in the table, of the above squared deviation, expressed as a proportion of its respective expected frequency:

$$Chi\ Square = \sum \frac{(Fo - Fe)^2}{Fe}$$

In our example, adding the 4 square deviations obtained for each cell we get:

Observed frequencies (Fo)	Expected frequencies (Fe)	Squared Fo - Fe	Squared deviations divided by Fe
ap = 4	ape = 2,3	$(+1,7)^2 = 2,89$	2,89 / 2,3 = 1,256

bp = 1	bpe = 2,7	$(-1,7)^2 = 2,89$	2,89 / 2,7 = 1,070
an = 2	ane = 3,7	$(-1,7)^2 = 2,89$	2,89 / 3,7 = 0,781
bn = 6	bne = 4,3	$(+1,7)^2 = 2,89$	2,89 / 4,3 = 0,672
	Chi Square =		3,78

Value 3,78 is the Chi-square. The higher a Chi-Square is, the greater the deviation is from the null hypothesis. If Fe = Fo in all cells, then the Chi-square is zero.

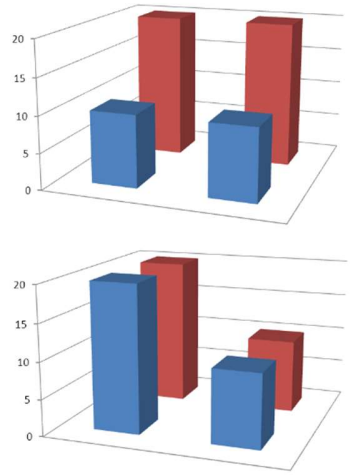
The test of the hypothesis that a Chi-Square as high as 3.78 is observed in a sample when the probability of a case belonging to a cell is given by the expected frequency, can be obtained using the probability distribution of the Chi-square for the degrees of freedom involved (in this case, 1 d. f.). This distribution is obtained squaring the standardized Normal distribution. The probability of observing such a Chi-square is the area to the right of this value. If it is low, the null hypothesis cannot be rejected.

There are tables available with Chi-square distributions for various degrees of freedom. The smallest Chi-square value that provides a probability of 5% in distributions with 1 degree of freedom is 3.84. If the Chi-square is higher than 3.84, then the null hypothesis is rejected.

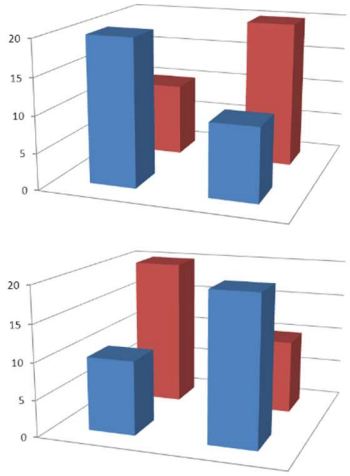
In our case, the value obtained is close to 3.84 but does not allow us to reject the null hypothesis. In situations like this, the person in charge of the trial would introduce new clients in the trial as it is likely that the promotion is having an effect on responses but, being just a few objects, significance cannot be reached.

The Chi-Square increases with the effect of the interaction between attributes in rows and columns in a cross-tabulation. Intuitively, interactions have the following meaning

No interaction



Yes, interaction



The use of this hypothesis testing is valid when the Normal approximation to the Binomial distribution stands. The "Chi" in the Chi-Square indicates a variable Normally distributed with mean zero and variance equal to 1. In practice, the use of the Chi-Square is unacceptable when one or more expected frequencies are less than 5.

- / -

An alternative way, computationally heavy, to determine the probability associated with the null hypothesis, would be to add up all the possible probabilities which, in a Binomial distribution, are against the equality of proportions in question. Values thusobtained are accurate and such test is

known as Fisher's exact test. When expected frequencies are low (less than 5 clients) this is the only way out.

- / -

When one or the two attributes that we crosstab have more than two classes, the degrees of freedom are no longer 1. They are the number of different sources of variation. A crosstab with P rows by Q columns, instead of P x Q degrees of freedom, will have, for the purpose of testing the null hypothesis, just

$$(P-1)(Q-1) = d.f.$$

Notice how this value coincides with the number of cells whose frequency you can change arbitrarily without changing marginal frequencies: in a table as the example above with 2 rows and 2 columns, there will be one degree of freedom. If we crosstab sex (two sexes) with location (3 cities) there are two degrees of freedom.

When we crosstab more than two attributes, the population will be divided into a multi-dimensional grid of cells, each with its expected frequency.

When, as above, the Chi-square tests the null hypothesis that observed interactions are explained by accidents of sampling, marginal frequencies are given, that is, they are not at issue (they are not being modelled) and cannot be taken as sources of variability. Thus, frequencies in each cell, when added by rows or by columns, should always give the same result, that is, the observed marginal frequencies are the same. Marginal frequencies are not considered random: they are fixed. There are other statistical tools, however, where marginal frequencies are considered as random.

Two nominal attributes are "independent" if the probability of an object belonging to a given cell is given by the product of the marginal probabilities of belonging to each class (row and column) that generate frequencies in the cell.

The expected frequency is the application of this concept of independence to the sample in question: if a sample contains N objects, the probability of belonging to row i is Pi and the probability of belonging to column j is Pj. Then, the expected frequency Fe in cell ij, the intersection of row i with column j, will be:

$$Fe\ of\ cell\ ij\ =\ Pi\ Pj\ N.$$

That is, the Chi-Square test compares observed and expected frequencies in the event of independence. Therefore, the hypothesis of independence between nominal attributes is identical to the non-significance of the difference between proportions.

Note that independence admits, in general, graduations: two attributes can be almost synonymous; they can be highly dependent but not synonyms; or they may have a weak dependence. Do not use the Chi-Square as a measure of the degree of dependence because there are better tools to carry out this task.

Chapter 6 Survival

Survival is the length of time between two related events, the initial and the terminal. The typical example of survival is the length of time observed from the appearance of a disease in a patient (initial event) and death from the same desease (terminal event). For the Insurance Industry, however, being alive is seen as an illness and survival is the length of time between today and death. Besides its widespread use in Insurance, the actuarial method is also used to study bankruptcy, loan default and other processes.

6.1 Actuarial method

As a random attribute, survival presents two major difficulties, both theoretical and practical, which makes this object matter an entirely separate chapter of statistical analysis. The difficulties are:

1. Frequency distribution and the probability density of this variable is positively asymmetric and hardly definable: each process has a different distribution.
2. Samples involving survival are biased because the number of cases for which the initial but not the terminal event is known, is usually large. The term used to designate such cases is “censored” and its existence greatly affects results.

The practice traditionally employed to deal with censored cases is called the “actuarial method”: an object that disappeared (we know the initial event date but not the terminal event date) counts as if its survival contributes half-an-object to the period immediately following that when, for the last time, was heard of. Thus, a censored object survives in practice half a period after the last period it was last seen.

There are good reasons to consider the actuarial method as the less bad: it is proven that the elimination of censored cases leads to a biased estimate of survival with larger deviations than when employing the actuarial method.

Survival curves are graphical representations of the probability of survival versus time. There are several types and the most used are:

1. “Cumulative survival rate”: shows, at the end of each time period, the proportion of objects that survived till that date.
2. “Probability density of survival”: shows, for each time period, the probability of an object surviving that period.
3. “Hazard function”, also known as “Fatality Rate” curve: shows the probability per unit time that an object who has survived up to the beginning of a period, dies during such period.

The cumulative survival rate is also known as “Survival Curve”, “Actuarial”, or “Experimental Survival Curve”. It is intuitive to read, and its applicability is wide.

The actuarial method assumes, as it turns out, the division of the survival time in periods; the statistical modelling of survival therefore requires data grouped by classes where each class is a given year, month, or week. There is a variant of the actuarial method for non-grouped data, called “Kaplan-Meier (K-M) method”. In this method, the above curves are used as the starting points of each period but, while the actuarial curve shows steps, the K-M is continuous. In most cases K-M is just an interpolation.

The drawing of an observed actuarial curve requires the following steps:

1. counting, for each time period *i*, the

   number of survivors at the beginning of the period *ni*

   number of deaths that occurred during the period, *di*, and

   number of censored *ci* during the period.

2. based on such data, calculate, for each period, the

   number of objects at risk in the period $Nri = ni - ci/2$

   survival rate in the period $P_i = 1 - di/Nri$

   cumulative survival rate from initial event till the period $S_i = S_{i-1}P_i$

According to the actuarial method, the number of objects at risk during a period is calculated by subtracting from survivors at the beginning of each period half of those who were lost track of (censored) during the period.

The table below shows data obtained over several years to draw the survival curve of premenopausal metastatic breast cancer in Scottish patients treated with combination chemotherapy.

	interval		alive at	dying in	censored
Year	start	end	start	interval	alive at start
1978	0	1	10	4	
	1	2	6	1	
	2	3	5	2	
	3	4	3	1	
	4	5	2	1	
	5	6	1		
	6	7	1		1
1979	0	1	25	5	
	1	2	20	5	
	2	3	15	5	
	3	4	10	3	
	4	5	7	3	
	5	6	4	1	3
1980	0	1	26	8	
	1	2	18	8	
	2	3	10	5	
	3	4	5		
	4	5	5	2	3
1981	0	1	11	2	
	1	2	9	3	
	2	3	6	1	
	3	4	5		5
1982	0	1	19	7	
	1	2	12	3	
	2	3	9	4	5
1983	0	1	4		
	1	2	4	2	2

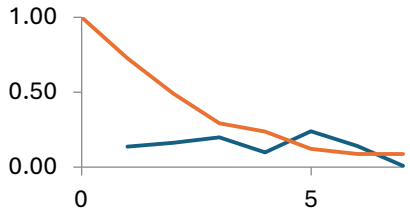
In every year between 1978 and 1983 a new group of patients enters the trial. The first year had 10 patients, then 25 new patients entered and so on. After data collection was finished, the estimation of the survival curve was then made according to calculations explained above, giving the following results:

alive at start	dying in interval (d)	censored alive at start	patient at risk (n)	1-d/n	accumulated survival rate
95	26	0	95	0.73	0.73
69	22	2	68	0.68	0.49
45	17	5	42.5	0.60	0.29
23	4	5	20.5	0.80	0.24
14	6	3	12.5	0.52	0.12
5	1	3	3.5	0.71	0.09
1	0	1	0.5	1.00	0.09

To understand the process leading to the cumulative survival rate, note that each period in the table above contains patients who entered the trial in different years. Counts are made using the same

relative periods but different years. For example, the 95 patients who were alive at the beginning of the trial stem from adding 10 patients who entered in 1978 plus 25 who entered in 1979 and so on.

The two rightmost columns of the table above can be presented in this graphic.



Survival rate P_i (blue) shows the proportion of survivors in each period, not the total proportion of survivors in the period. Such total proportion of survivors is given by the cumulative survival rate S_i whose collection for all periods is the survival curve.

Note that each S_i is obtained by multiplying all P_i taken from the initial period until period i. This was explained earlier when referring to multiplicative probabilities: chains of mutually dependent events, each with a given probability of occurring, lead to an overall probability which is the product of such individual probabilities.

S_i and P_i frequencies are observed or experimental, but the cumulative survival curve can be a good estimator of the underlying cumulative survival function.

Another much used curve is the Hazard rate curve, estimated from successive fatality rates. Each can be calculated by dividing the number of deaths (per unit time) in each period by the mean number of survivors at the midpoint of the same period.

Particularly useful is the fact that, when a survival curve is presented in a logarithmic scale, its slope is proportional to mortality rate. The hazard function is akin to mortality and in some situations one can be used instead of the other.

- / -

The validity of the actuarial method is based on hardly verifiable assumptions. This type of methods, much more than others, rests on shaky ground and requires, in the part of the analyst, a certain degree of skepticism.

There are situations, unfortunately frequent, in which the actuarial method should not be used because it leads to erroneous conclusions. Namely,

1. If the mechanism that leads objects to disappear before the outcome has been recorded relates to survival time, it is no longer possible to apply the actuarial method because, obviously, mortality relates to the rate of censored cases. Often patients go home to die, or they disappear for the opposite reason, i.e., because they are cured, thus becoming censored precisely at the end of the process and because of such end.
2. It is also necessary to ascertain whether the reason that leads to the terminal event is the outcome of the process under study, not any other. Otherwise, the data is misleading. It is often difficult to distinguish the causes of death and is fair to assume that there were several. Hence this problem is difficult to solve.
3. Finally, it is important to consider the cumulative degradation that censored cases introduce in actuarial curves. Such degradation shows up in large confidence periods at the final periods of the study, which testify to the fact that the trial is losing significance. Indeed, final

periods are those most interesting, those with higher relevance. If their significance is low, then the trial is useless.

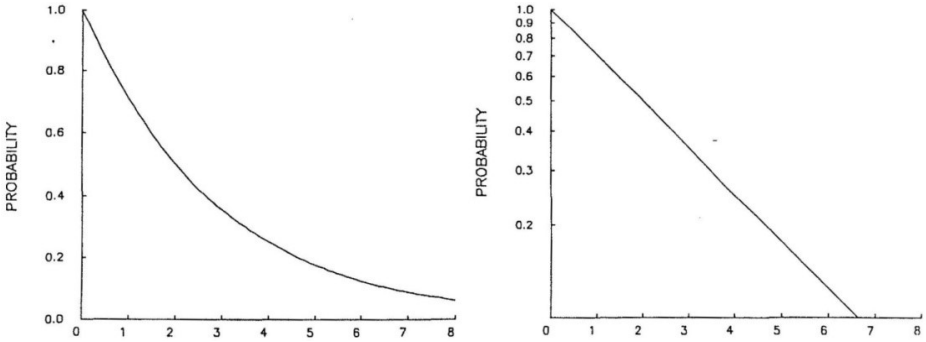
- / -

It is desirable to parameterize the experimental survival curve for each disease, i.e., to describe the typical survival through a function. Here, the diversity of processes is the biggest obstacle as each of them requires different parameterization: the shape of the actuarial curve greatly varies from one process to the other.

Common analytical forms used to approximate survival curves are:

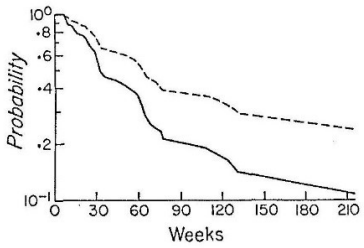
- Negative Exponential: when, in logarithmic scale, the cumulative survival rate is linear, the survival curve can be approximated using a negative exponential, which means a constant mortality rate over time.
- Weibull: this family of curves allows modeling a monotonically increasing risk (survival decreasing over time).
- "Proportional Hazard" or "Cox" method: this modelling approach consisting of supposing that the outcome is the result of known risk factors. The effect of each factor would be to multiply risk by a constant value. This is the method often used in Finance.

Figure below shows a negative exponential survival, in both the usual and logarithmic scale. The steeper the line on the right side, the higher the constant mortality rate is.



In the Proportional Hazard method, survival curves have little interest. What matters is the set of risk factors (in bankruptcy prediction, a rescue plan and other beneficial factors are negative factors) such that this method, rather than parameterize a curve, proceeds to explain the process itself, using a regression. Well known risk factors are required.

In this case, the shape of the survival curve is not changed when any of the risk factors decreases or increases, as below. The Cox method is applicable only in such cases.



Frequently, none of the above applies. When the mortality rate shows an increase to a peak (maximum mortality at a given time) and then it decreases, a most common case, the family of Weibull curves cannot be used. But when, in many curves of this type, peaks are located at different periods, and mortality at the end of the trial is identical for all curves (suggesting large errors) then

the hypothesis of a single common mechanism commanding the evolution of the process, as well as the hypothesis of risk factors with constant action, cease to be sustainable.

6.2 Survival comparison

Is it possible to estimate confidence periods for survival curves and thus compare two such curves in the same way mean values or proportions are compared? Yes, in a sense.

Each period in a survival curve is a proportion. Based on this, methods have been developed which estimate standard errors and confidence periods for experimental actuarial curves. Standard errors show how close each period's survival is from the true survival probability (of the population). As usual, confidence periods have an associated confidence level, typically 95%.

When the sample has two or more classes, for example two drugs used, each class will have an experimental curve. This is the case of comparative trials studying the evolution of a given type of process in objects treated in three different ways to find out what would be the most advantageous.

How to test the significance of the difference between survival curves? A multitude of tests have been developed, none of them convincing enough to be unconditionally accepted. The common starting point of all such tests is the repeated use of 2 by 2 crosstab, one for each time period, comparing the mortality rates. Thus, a series of Chi-Square values is obtained, each with its level of significance. From this series, an index value and significance level are then computed.

The way to weigh the collection of Chi-Squares determines the significance of the resulting index. Some tests are more sensitive to differences in the late periods while other tests are more sensitive to differences observed early in the process. No consensus exists on the quality of any of the existing tests. Trial results, therefore, are often reported mentioning at least two tests.

There are two main types of tests:

The Log-Rank test with variations:

Proportional Hazard,

Mantel-Haenszel,

Generalized Savage,

Exponential Scores and others.

The Wilcoxon Generalized test with variations:

Modified Gehan

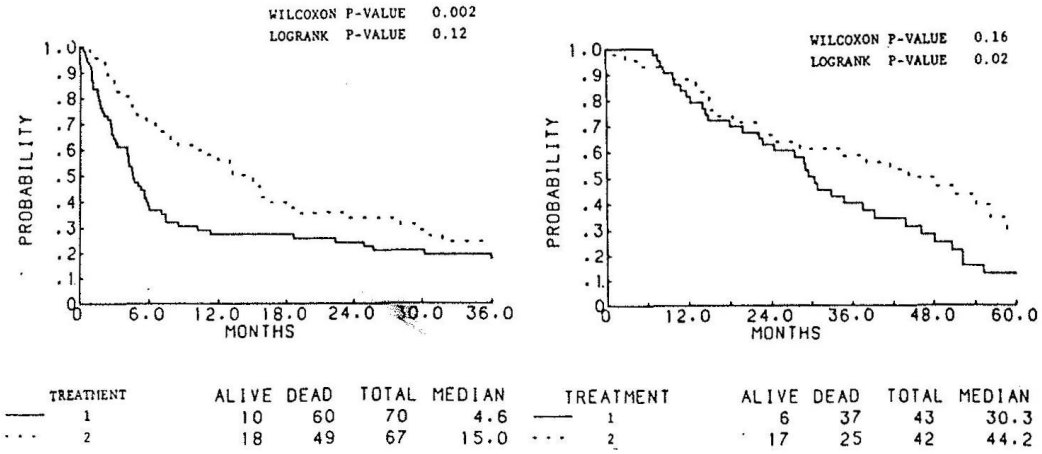
Modified Gilbert

Breslow and others.

Log-Rank test is sensitive to differences in late periods while

Wilcoxon test is sensitive to differences in early periods.

This is because Log-Rank gives the same weight to all Chi-squares no matter the period even though, in latter periods, the number of cases is much smaller. Thus, the weight of the final part of the curve is inflated in this case.



TREATMENT	ALIVE	DEAD	TOTAL	MEDIAN
1	10	60	70	4.6
2	18	49	67	15.0

TREATMENT	ALIVE	DEAD	TOTAL	MEDIAN
1	6	37	43	30.3
2	17	25	42	44.2

The figure above compares Log-Rank and Wilcoxon tests.

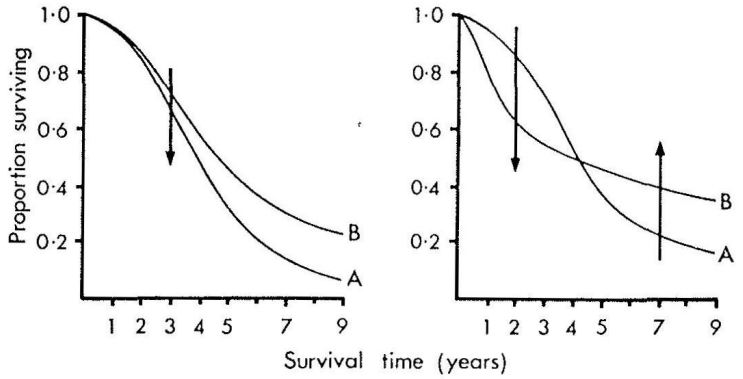
- On the left, one of the treatments delays death but does not reduce incidence. In this case, the Wilcoxon test would likely reject the null hypothesis that, in the population, both treatments have the same effect. The Log-Rank test would not find any significant differences between treatments.
- On the right, one treatment does not reduce mortality during the initial periods of the disease but ultimately it reduces mortality. In this case, the Wilcoxon test would not indicate significant differences while the Log-Rank rejects the null hypothesis.

These tests have incorrect significance levels when the survival time and censored are statistically dependent.

- / -

The comparison of actuarial curves is a problem that has little to do with Statistics.

For example, in the example below, the left shows an undeniable improvement but the situation depicted on the right is ambiguous: which of the two curves is the best?



Treatment B saves the lives of patients but does so at the expense of a higher mortality at the beginning of the disease. Patients who overcome the initial phase of the disease are likely to be cured. Treatment A has lower mortality during initial periods but then overall mortality is higher.

Chapter 7 Linear Modelling

Suppose that random variable x_i observed in object *i* is explained by the model:

$$x_i = M + e_i$$

where

- *M* is the mean population. This is what the model "explains" or "predicts".
- e_i is what remains unexplained on object *i* of x_i, It is random with zero mean. It is called the "error" or "residual".

The above is the simplest "linear model" possible, where observations are explained by their average or mean. What the model does not explain, e_i, is unexplained variability. The term "linear" here, refers to the additivity of "effects": indeed, e_i adds to *M*.

7.1 Analysis of variance

Another, slightly less simple model, would add another effect to the effect of the mean. Suppose, for instance, that three different promotion types are compared, to find out which of them leads to higher increases on customer satisfaction.

In general, whenever it can be assumed that, in a population, observations group around mean values forming strata or cells, analysis of variance (ANOVA) is the appropriate tool for modelling and hypothesis testing. ANOVA can be applied provided that:

1. The sample is divisible in strata (classes, categories, cells)
2. The population is Normal in each stratum
3. The variance and the number of objects is similar in each stratum.

ANOVA posits that random variable x_i observed in object *i* can be explained using a linear (additive) type of model:

$$x_i = M + d_k + e_i$$

where

- *M* is the mean the whole population without considering strata. *M* influences in the same way all x_i objects.
- d_k is the deviation from *M* which the x_i suffer due to the fact that, in stratification K, object *i* belongs to category *k* in such stratification.
- e_i is what in x_i (on the object *i*) remains unexplained, and it is supposed to be random with zero mean and variance identical, no matter the stratum.

Therefore, the analysis of variance divides the variability of *x* in two parts:

1. the variability of observations within their stratum, which is measured relative to m_k, the mean of the stratum. This is SSW ("sum of squares within groups") or unexplained variability. $SSW = \sum_K (x - m_k)^2$.
2. the variability of the mean values m_k of strata in relation to the overall mean *M* is SSB, multiplied by the number of objects in strata, is the sum of squares between groups or explained. $SSB = N/K \sum_K (m_k - M)^2$. SSB shows how averages of each stratum differ among themselves.

For *N* objects and *K* strata, the degrees of freedom involved in modelling are

SSW has *N-K* degrees of freedom.

SSB has *K-1* degrees of freedom.

After divided by their respective degrees of freedom, both SSW and SSB express variances. These variances are called "mean sum of squares", MSQ:

MSW = SSW/N-K

MSB = SSB/K-1

The quotient between MSW and MSB is called "Fisher's F" or just F coefficient.

How to test hypotheses with the analysis of variance? The model, that is, what will replace the sample, consists of either

a collection of m_k mean values, as many as the number of strata considered.

or, what is the same, the overall mean *M* plus the *K-1* deviations d_k from this medium-base (one stem from the others and *M*).

The significance of the model is assessed by testing the null hypothesis that strata do not introduce any extra variability in the model. Now, this is equivalent to saying that all strata have identical mean in the population and therefore the observed differences should not be real differences.

If SSW, after corrected for the degrees of freedom as explained above, is greater than the similarly corrected SSB, then there is more variability within strata than between strata. It is possible to determine the probability of obtaining a quotient F as high as that observed when the null hypothesis of identical averages in the population is accepted. When such probability is less than 5% we reject the null hypothesis of the mean values being identical in the population. The distribution of F allows assessing the probability of obtaining a value of F as big as that observed, when, in the population, this value is 1, i.e., MSW = MSB, thus, no difference in variability.

- / -

A client satisfaction score with values ranging between 0 and 30 is measured in 12 clients. Of these, 6 are city dwellers and 6 are from the countryside. We use the analysis of variance to test the null hypothesis that there is no difference between client satisfaction in cities and in the countryside.

The 6+6 satisfaction scores observed have the values as follows.

City	13.24	14.07
	12.76	13.93
	13.00	14.00
Country	12.15	13.30
	11.85	12.70
	12.00	13.00

The overall mean is 13 and mean values by strata are.

City	13.50
Country	12.50

We calculate SSW by subtracting 13.5 to all scores from cities, subtracting 12.5 to all scores from the countryside, squaring these differences and adding them all up.

We calculate SSB by subtracting 13.5, to the overall mean and squaring, then subtracting 12.5 from the overall mean and squaring. Adding these two squared differences and then multiplying the total by the number of cases per stratum N/K, 6. From here, and bearing in mind the degrees of freedom, the value of F is calculated:

	Sum of Squares	df	Mean Square	F
Between strata	3.00	1	3.00	8.96

Within strata	3.35	10	0.34
Total	6.35	11	

F is, in this case, significant as the probability associated with the null hypothesis is 0.014 or 1.4%, thus less than 0.05. The null hypothesis that scores show no differences from the city to the country is rejected.

-/-

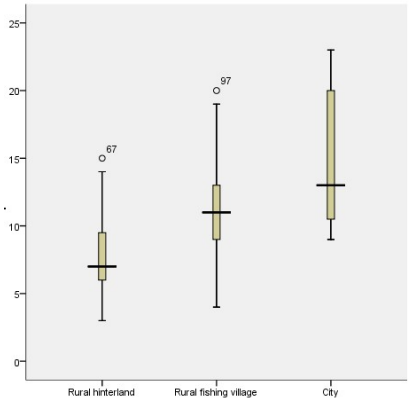
In the following example, we use three strata, not two: rural hinterland, rural coastal area, and urban area. Moreover, strata are “unbalanced”, that is, the number of objects per strata is not the same, which make calculations less straightforward (we omit them). Observed frequencies are:

Rural hinterland	Fishing village	City	Total
43	89	19	151

Means and standard deviations are as follows:

Place		Satisfaction
Rural hinterland	Mean	7.88
	Std. Deviation	2.994
Rural fishing village	Mean	10.9
	Std. Deviation	3.378
City	Mean	14.5
	Std. Deviation	5.113

The “box plot” shows the quartiles and the median of a sample. With a box plot, it is possible to compare central values and dispersions in different strata. Inside the "box" there are 50% of cases and this, coupled with the size of the “whiskers”, shows the variability of the sample. In a symmetrical distribution, median and mean are the same.



It is not surprising that the differences between zones is significant. The table shows, step-by-step, the analysis of variance.

		Sum of Squares (SSQ)	Degrees of freedom	Mean squares	F quotient	Significance
Satisfaction	Between	596	2	298	24,016	0,000 high
	Within	1.825	149	12.2		
	Total	2.421	150			

- / -

In most cases analysts are not interested in parameter values but in ascertaining whether there are grounds to reject the null hypothesis. This is the case of the examples testing the effect of a new promotion on different groups of clients. One of the groups does not receive a promotion; the second

receives a well-known promotion, and the third group receives the new promotion, the one we wish to test. The conclusion that observed difference in mean values is significant is, by itself, the most important result.

When it is desirable to know if there are significant differences, not just overall but from stratum to stratum, the "multiple comparisons" test is used: successive "t" tests of differences between mean values should not be used because the probability of spurious significance increases very rapidly with the number of tests: 5 strata would require 10 comparisons; however, even if the mean values were strictly equal in the population, the probability of obtaining significance at least in one of these ten tests is 30%. Using multiple comparisons, significance levels account for the number of tests.

7.2 Two-way analysis of variance, interaction effect

In the two-way analysis of variance, we consider two stratifications that cross each other feely: region by sex of client, for example. The analysis is the same as for the one-way case: variability is divided in two parts – between and within strata; but, instead of just strata, now there are "cells" formed by the crossing of the two stratifications; and "within" or "between" now refer to such cells.

The parameterization of the two-way model is as follows:

$$x_i = M + d_k + d_q + I_{kq} + e_i$$

where it is assumed that each observation x_i on object i stems from

- the effect of an overall mean M
- the effect d_k from x_i belonging to the stratum k in K
- the effect d_q from x_i belonging to the stratum q in Q
- the effect of the "interaction" I_{kq} between these two stratifications
- a random effect e_i showing what is unexplained in object i. It is expected to have zero mean and similar variance regardless of which cell is considered.

The two-way ANOVA is often used in the experimental sciences, in Sociology and Psychology, but is not a concern of Econometrics. The reason why it is studied here has to do with the fact that interactions, an important Econometric concern, become better understood after studying the two-way ANOVA.

Although the principle that presides over the analysis of variance is always the same, that is, variability is divided in two parts, one within and another between cells, when there is more than one stratification, it is useful to examine in a different way the sub-partitions of variability. In particular, the sum of squares explained, SSB, should be partitioned to show the variability due to each stratification.

In a two-way ANOVA with stratifications K and Q, then SSB is

= effect of stratification K

+ effect of stratification Q

+ effect of the "interaction" of K with Q

The effects above, which are parts of SSB, are called

- "main effects", zero-order effects or additive effects, or
- higher-order effects, cross-effects, or interactions.

Sub-partitions of the SSB enable testing the null hypothesis individually for each of the effects. Quotients F are computed as usual. All effects are all compared with SSW.

For computing the MSB's the following degrees of freedom should be considered:

- The degrees of freedom associated with a main effect are K-1, where K is the number of strata of that effect. For the main effect Q it is Q-1.
- The degrees of freedom associated with an interaction are (K-1)(Q-1) which is the product of the degrees of freedom engaged by each effect.
- The residual variability, SSW, has N-KQ degrees of freedom, and the total variability has N-1 degrees of freedom.

How to calculate the sub-partitions of variability?

- Main effects are obtained considering stratifications one at a time and ignoring other stratifications. The effect of "sex", for example, is obtained by calculating the variability due to the existence of both sexes and ignoring the variability due to the existence of four types of treatment.
- Interactions are computed by exclusion: sums of squares that cannot be ignored and are not explained as main effects are interactions. Considering each cell, the difference between the overall mean and the observed mean is first explained by the main effects:

  since the cell belongs to K, deviation d_k is expected

  since the cell belongs to Q, deviation d_q is expected.

The overall average M plus the two deviations d_k, + d_q are calculated beforehand. Thus, whatever in the observed cell mean is different from such addition, is the interaction.

When an interaction is not significant the two stratifications do not interact with each other. It is as if the analysis is made using two separate models, one for each of the stratifications. When, on the contrary, there are significant interactions, cell means are less than or greater than that expected if main effects were purely additive.

- / -

A two-way version of the balanced example observes satisfaction in 12 clients, 6 from the city and 6 from the countryside. The effect of sex is also noted. Thus, there are two stratifications (place and sex), each having two strata and there are 3 clients in each cell:

	Male	Female
City	13.24	14.07
	12.76	13.93
	13.00	14.00
Country	12.15	13.30
	11.85	12.70
	12.00	13.00
	12.50	13.50

The average values per cell, per stratum and the overall mean, are:

	Male	Female	
City	13.00	14.00	13.50
Country	12.00	13.00	12.50
	12.50	13.50	13.00

The two-way ANOVA is

		S. Square	df	M. Square	F	Sig.
Main Effects	Both	6.00	2	3.00	68.57	0.00
	Place	3.00	1	3.00	68.57	0.00
	Sex	3.00	1	3.00	68.57	0.00
Interaction	Place * Sex	0.00	1	0.00	0.00	1.00
Model		6.00	3	2.00	45.71	0.00
Residual		0.35	8	0.04		
Total		6.35	11	0.58		

Interaction is non-significant: the two main effects do not influence each other.

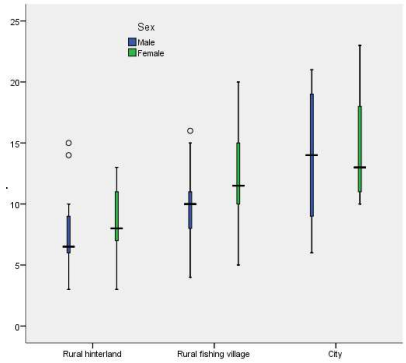
In the example that follows, the calculation of effects is less straightforward because the number of objects per cell is unbalanced. The number of objects by sex and area is:

		Rural	Fish vill.	City	Total
Sex	Male	22	43	6	71
	Female	21	46	13	80
Total		43	89	19	151

After crossing the two effects, mean values per cell are compared.

Sex	Place		Satisfaction
Male	Rural hint.	Mean	7.36
		Std. Dev.	3.02
	Rural fishing	Mean	9.40
		Std. Dev.	2.70
	City	Mean	13.8
		Std. Dev.	6.38
	All 3 places	Mean	9.07
		Std. Dev.	3.49
Female	Rural hint.	Mean	8.43
		Std. Dev.	2.94
	Rural fishing	Mean	12.3
		Std. Dev.	3.37
	City	Mean	14.8
		Std. Dev.	4.82
	All 3 places	Mean	11.7
		Std. Dev.	4.10
Both sexes	Rural hint	Mean	7.88
		Std. Dev.	2.99
	Rural fishing	Mean	10.9
		Std. Dev.	3.38
	City	Mean	14.5
		Std. Dev.	5.11
	All 3 places	Mean	10.5
		Std. Dev.	4.03

The above can be graphically displayed as box plots where interactions are visible.



Results of the analysis of variance are as follows:

			S. Squares	df	M. Squares	F	Sig.
Satisfaction	Main Effects:	Both	753	3	251	22.92	0.000
		Sex	10	1	10	0.93	0.336
		Place	543	2	272	24.80	0.000 * * *
	Interaction	Sex * Place	56	2	28	2.55	0.082
	Model		837	5	167	15.29	0.000
	Residual		1490	136	11		
	Total		2327	141	17		

The partition of variability is very detailed. First appears total main effect and then, one by one, the interaction effect, the sum of squares that the whole model explains (main effects and interaction), the sum of squares that the model doesn't explain (residual) and the total, that is, both the variability that the model explains and what it doesn't explain.

Note that, in this case, the sums of squares do not add up. Only for balanced calls is it possible to use calculation where sums of squares add up. But the total variability equals that of the residual (not explained) plus that of the model (explained).

The differences between objects due to sex, do not reach the significance level of 5% required to reject the null hypothesis. Place explains most variability.

The Sex / Place interaction does not reach the significance level of 5 percent, although it is near. In cases like this, where the significance is smaller than 10 percent but bigger than 5 percent, it is worthwhile increasing the number of objects in the sample so that a higher power may bring about conclusive results. A significant interaction would have shown that the effect of sex on satisfaction would not be independent of the effect of the place so that the two effects would not add up: one influences the other, making mean values smaller or larger than would be expected.

Chapter 8 Correlation, linear regression, OLS assumptions

8.1 Covariance, correlation

We shall now study the modeling and hypothesis testing involving continuous attributes whose variability is explained by other continuous attributes: income may be explained by education, for instance, or systolic blood pressure may be explained by weight.

Attributes that have common variability are called "covariates". Given two continuous attributes *x* and *y* observed in object *i* on a sample with *N* objects, their "covariance" is:

$$COV_{xy} = \frac{\sum_{i=1}^{N}(x_i - M_x)(y_i - M_y)}{N-1}$$

where M_x and M_y are mean values. COV is higher the more pronounced is the linear association between the variability of *x* and *y*.

In the limit case of these attributes having the same variability, for example when one of them is a linear combination the other as in $y = A + B\,x$ where A and B are constant, COV becomes the variance of *x* or *y*.

When, in the opposite end, *x* and *y* do not share variability, being independent, then COV = 0.

The "correlation coefficient" r_{xy} between *x* and *y* is the covariance expressed as a proportion of total variance:

$$r_{xy} = \frac{COV_{xy}}{\sigma_x \sigma_y}$$

where $\sigma_x \sigma_y$ are the standard deviations of *x* and *y*. A given correlation coefficient is higher, in absolute terms, the closer the association between *x* and *y* is. Correlation can be positive or negative varying between -1 and +1 indicating the correlation type:

- for negative r_{xy} correlation is "inverse": when *x* increases, then *y* decreases;
- for positive r_{xy} correlation is "direct": when *x* increases, then *y* increase.

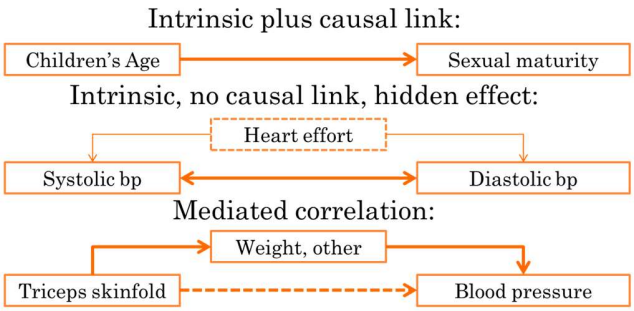
Non-linear association between two attributes may entirely elude measurement based on covariance and correlation, which assume a linear relationship. Non-linearity is indeed a characteristic often found in Finance.

Correlation is just a relationship; strong correlation does not mean causation. It is useful to distinguish between the following links between variables, all leading to correlation:

"Intrinsic" correlation: a "mechanic" or "automatic" link, which is independent of other variables such as place or sex.

"Causal" correlation: one of the variables causes, commands, drives the other.

"Mediated" correlation: the observed correlation is caused by a third, hidden variable, which influences the two variables being observed. Purely mediated correlation should be seen as spurious: no true correlation exists.



Correlation can be used to grasp causation. Suppose that correlations of epidemiologic data from 151 children are as follow:

		Weight	Height	Skinfold	Systolic	Diastolic
Height	Correlation	0,76				
	Significance	0,00				
Skinfold	Correlation	0,41	0,23			
	Significance	0,00	0,01			
Systolic	Correlation	0,12	0,06	0,18		
	Significance	0,15	0,43	0,03		
Diastolic	Correlation	0,09	0,10	0,24	0,69	
	Significance	0,27	0,21	0,00	0,00	
Sexual mt.	Correlation	0,43	0,41	0,08	-0,03	-0,03
	Significance	0,00	0,00	0,34	0,69	0,75

What the table shows is that, beyond trivial cases such as correlation of weight with height, skinfold is positively correlated with blood pressure.

Here, "significance" level is the same as a confidence level calculated exactly. It shows the probability associated with the null hypothesis that correlation between the two attributes does not exist in the population. So, for example,

there are 3 chances in 100 of a correlation of 0.18 be observed between systolic blood pressure and skinfold when in the population correlation is zero.

And, since 0.03 is below 0.05, the null hypothesis of no correlation is rejected.

"Partial correlation" is correlation between two attributes when the effect of a third attribute on them is removed. The relationship between skinfold and blood pressure in objects with different weights can be estimated using partial correlation.

In the example below, correlations are now shown when the effect of sexual maturation is removed from all attributes. Blood pressure is still correlated with skinfold.

		Weight	Height	Skinfold	Systolic
Weight	Correlation				
	Significance				
Height	Correlation	0,73			
	Significance	0,00			
Skinfold	Correlation	0,46	0,27		
	Significance	0,00	0,00		
Systolic	Correlation	0,11	0,05	0,17	
	Significance	0,19	0,60	0,04	
Diastolic	Correlation	0,08	0,11	0,22	0,67
	Significance	0,33	0,21	0,01	0,00

When, not just sexual maturation but also weight is removed, correlation between skinfold and systolic blood pressure ceases to be significant.

		Skinfold	Systolic
Systolic	Correlation	0,14	
	Significance	0,11	
Diastolic	Correlation	0,21	0,67
	Significance	0,01	0,00

This shows that the influence of skinfold on blood pressure is illusory, and stems from the mediated influence of weight and skinfold.

- / -

We have just seen an example involving several, variously correlated random variables $x_1, x_2, \dots, x_k$. The "variance-covariance" matrix *S* of these variables is

$$S = \begin{bmatrix} VAR_1 & COV_{1,2} & \cdots & COV_{1,k} \\ COV_{2,1} & VAR_2 & \cdots & COV_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ COV_{k,1} & COV_{k,2} & \cdots & VAR_k \end{bmatrix}$$

S is a square, symmetrical matrix containing variances in the diagonal and covariances in the other positions. Note that diagonal numbers, the variances, are the covariance of variables with themselves. *S* can easily be calculated using matrix manipulation:

$$S = \frac{1}{N} R^T R - E E^T$$

where *N* is the number of observations, R is the matrix containing observations (each line is an observation, each column is a variable so that *R* has *N* rows and *k* columns). R^T is the transposed *R*, *E* is the column vector containing expected $x_1, x_2, \dots, x_k$ and E^T is the transposed *E*. *E* is line vector, thus EE^T is scalar.

For example, if the $x_1, x_2, \dots, x_k$ are *N* monthly returns of *k* securities forming an *R* matrix in which each line is a month and each column is a security, then it is easy to calculate the variance of a portfolio in which the $x_1, x_2, \dots, x_k$ are present in proportions $\theta_1, \theta_2, \dots, \theta_k$. This variance is just $\theta^T S\, \theta$. In this case, θ is a column vector so that θ^T is a line vector and $\theta^T S\, \theta$ is a scalar. The expected returns of the same portfolio are $\theta^T E$.

Moreover, of all possible θ proportions, those Z proportions which obey

$$Z = S^{-1}(R - C)$$

will form, after being corrected to add to exactly 100%, an efficient portfolio. *C* is an arbitrary constant value: different *C* will lead to different efficient portfolios. S^{-1} is the inverse variance-covariance matrix of returns.

8.2 Linear regression, OLS

The "linear regression" is the tool employed to model a continuous, random attribute, when explained by continuous, non-random attributes. For example, a linear regression can explain blood pressure of children using age as an explanatory attribute, or the market returns of traded securities in terms of their systematic risk.

The "simple" linear regression model has the following form:

$$y_i = A + B\, x_i + e_i$$

- *y* is the attribute that is being explained
- *x* is the attribute that explains *y*

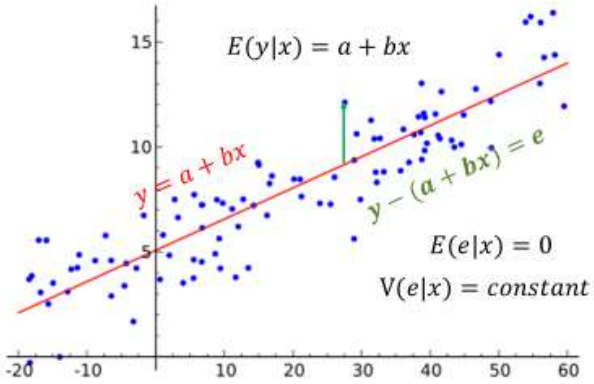
- e_i is the part of *y* that the model does not explain in the specific case of object *i*; it is also the difference between the observed *y* in *i* and *y* predicted by the model for the same object.

A and *B* are the two coefficients or parameters of the model. *A* is called "intercept" or "constant term"; *B* is called "slope".

e_i is the "error term", the unexplained variability, and is also called "residual".

y is also called "dependent variable"; it is the one whose evolution is modelled.

x is the "predictor", "explanatory" or "independent" variable.



The previous regression has only one predictor and that is why is "simple". When more than one predictor is used, we have the "multiple" linear regression

$$y_i = A + B_1 x_{1i} + B_2 x_{2i} + B_3 x_{3i} + \ldots + B_p x_{pi} + e_i$$

- *y* is the attribute that is being explained
- x_j are the *p* attributes that explain *y*
- e_i is the part of observation y_i left unexplained; it is also the difference in object *i*, between the observed *y* and the *y* predicted by the model.
- *A*, B_1 B_2 B_3 ... B_p are coefficients or parameters of the model.

  A is called intercept or constant,

  the B_1 B_2 B_3 ... B_p are called slopes of their respective attributes.

In a multiple regression with *p* predictors, the degrees of freedom engaged are K=*p+1*.

-/-

How to estimate *A* and the *B*? When the following assumptions are met:

1. Mean of *y* for successive intervals of *x* is linear
$$E(y|x) = a + bx$$
2. The sample is obtained at random.
3. Variance of *e* is constant for any interval of *x*
$$V(e|x) = constant$$
4. The y_i are independent of each other
5. The mean residuals *e* for any interval of *x* is zero (implies e independent of *x*)
$$E(e|x) = 0$$
6. *y* is normally distributed for any interval of *x*

it is possible to model *y* as a function of *x* using the "ordinary least squares" (OLS) method. OLS uses simple formulae to find the regression coefficients *A* and *B* above that minimize the variability unexplained by the model (the SSQ of all e_i). For instance, in the case of the simple linear regression, *A* and *B* are estimated as:

$$B = \frac{COV_{xy}}{VAR_x}$$

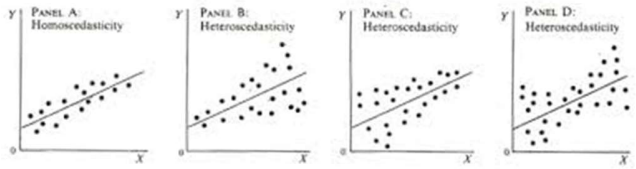
$$A = mean\ y - B\ mean\ x$$

Later on, we shall present the same assumptions in a more systematic way.

A simple way to check the validity of the OLS assumptions is to observe the residuals e_i

Must be normally distributed with zero mean and variance equal to *y*.

Variance cannot change for different *y* - residuals cannot be "heteroscedastic":



Must be independent of each other - cannot have a trend, for instance.

-/-

When, in a simple regression, *y* and *x* are standardized, the model

$$y_i = A + B\,x_i + e_i$$

becomes

$$y_i = \beta\,x_i + e_i$$

since, in that case, *A = 0*. The slope of this regression becomes independent of scale.

A measure of the risk of a security *e* relative to the market is Beta (β):

$$\beta = \frac{COV(r_e, r_m)}{VAR(r_m)}$$

r_m is the expected return on the market, and

r_e is the expected return on security *e*.

Therefore, the Betas used in the capital asset pricing and other models, are slopes of OLS regressions where market returns explain individual securities' returns.

8.3 Model hypotheses testing

Hypothesis testing for regressions is of two types:

1. Null hypothesis that A or B coefficients have irrelevant values: A=0, B = 0 in the population, regardless of observed values.
2. Null hypothesis that the model as a whole does not exist in the population.

With no linear relationship between *y* and *x*, the distribution of quotients

$$t_B = \frac{B}{\sigma_B} \qquad \text{and} \qquad t_A = \frac{A}{\sigma_A}$$

obeys a *t* distribution with *N-K* degrees of freedom. *K* is the degrees of freedom that the model engages; for simple regression, *K*=2. t_B and t_A are standard deviations associated with the estimation of *B* and of *A*, allowing calculating standard errors and confidence intervals for *A* and *B*

with a given level of confidence. In this way, the null hypothesis that regression coefficients are zero in the population can be tested.

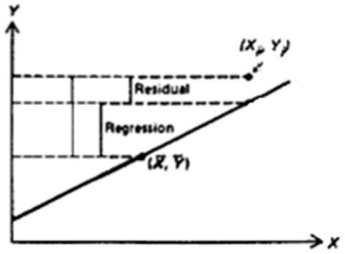
Using an ANOVA where the observed variability of *y* is divided in two parts:

1. the regression SSQ containing variability explained by the regression,
2. the residual SSQ containing variability unexplained by the regression,

we test the hypothesis that, in the population, there is no linear relation between *y* and *x*. To do this test, the distance of each y_i to the mean of *y* is divided in the same two parts:

the error or residual part, from y_i to the value of *y* predicted by the model,

and the regression or predicted part, from this value to the mean of *y*.



Each of these parts adds to SSQ due to regression and SSQ due to residuals respectively and allows making the ANOVA's F test, which will look like this:

ANOVA	SSQ	df	MSQ	F	Significance
Regression	456.25	1	456.25	4.70	0.03
Residual	14356.12	148	97.00		
Total	14812.37	149			

Then, using the *F* distribution for *1* and *N-2* degrees of freedom where *N* is the number of objects, it is possible to obtain a significance level associated with this *F*. Such level will then decide whether we accept or reject the null hypothesis that, in the population, the model explains no variability at all. In the above example the hypothesis would be rejected because the significance level is smaller than 0.05.

Note that, in the case of the simple regression, $F = t^2$. This *t* is the t_B referred to above. To evaluate the efficiency of a linear regression in explaining y_i we use the "coefficient of determination", also called "R Squared" (R^2), the square of the correlation coefficient between *y* and *x*. In simple linear regressions, the R Squared is also the square of the correlation coefficient between the observed *y* and the *y* predicted by the model.

The R Squared is obtained from the ANOVA table above:

$$R^2 = 1 - \frac{SSQ\ residual}{SSQ\ total}$$

The R Squared measures the explanatory power of the model as a percentage of the total variability. An R Squared of 0.63 means that, in the attribute to be explained, 63% of the variability is indeed explained.

When the sample is small, instead of the R Squared calculated according to the formula above, we use the "adjusted R Squared":

$$adjusted\ R^2 = 1 - \frac{SSQ\ residual\ (N-1)}{SSQ\ total\ (N-K)}$$

Instead of SSQs, now we use MSQs, that is, the SSQ divided by the respective degrees of freedom. The residual SSQ is divided by N-K and the total SSQ is divided by N-1 where *K* is the degrees of freedom engaged in explaining *y*. In simple linear regressions *K = 2* but in multivariate regressions *K* is the number of predictors plus 1, the constant term. The R Squared without adjustment would give an optimistic view of the explained variability as, in small samples, model fits the population worse than it fits the sample.

Continuing with the epidemiologic example, blood pressure is predicted by skinfold as:

Systolic BP = 100.87 + 0.434 Skinfold + residual

This formulation is used when we want to correct blood pressure for different skinfolds. The *t* and its significance, standard errors, and the standardized slope are:

OLS	Coefficient	Std. Error	Standardized	t	Significance.
Intercept	100.87	2.24		44.95	0.00
Slope	0.43	0.20	0.18	2.17	0.03

The standard error is the error of each coefficient and allows calculating confidence intervals. The R^2 is 3.1%, and the adjusted R^2 is 2.4%. The ANOVA test is:

ANOVA	SSQ	df	MSQ	F	Significance
Regression	456.25	1	456.25	4.70	0.03
Residual	14356.12	148	97.00		
Total	14812.37	149			

The degrees of freedom engaged by simple regressions are 2 because the constant term also engages one degree of freedom. Multiple regressions engage K degrees of freedom, being K-1 predictors plus the constant term. In that case,

- The 149 degrees of freedom associated with the total SSQ are *N-1*, with N being the number of effective objects in the sample, 150 in this case.
- The degrees of freedom associated with the residual SSQ are *N-K*: 148=150-2
- The degrees of freedom associated with the regression SSQ has *K-1*, that is, 2-1 in this case. In general, K are the number of predictors plus the constant term.

8.4 OLS assumptions testing

The testing of the OLS assumptions requires, in the least, the following:

A: Normality of residuals is verified with a "Normal probability plot" where the distribution of residuals is compared with the Normal cumulative function. When these two distributions are identical the result is a straight, diagonal line.

B: diagrams of dispersion of *y* with *x* (no non-linearity should be visible) or of values predicted by the model with residuals (no trends should be visible).

C: Homogeneity of variance of *y*: it can be detected in the same way. A triangular scatter would indicate that the variance decreases or increases with *x*.

D: Independence of the *y*: sort the *y*

a. by order of collection,
b. by date of collection,
c. by order of magnitude of other factors that may influence *y*,

see "case-wise plot", where no trend should be visible. The "Durbin-Watson" test is used to detect violations of independence of the *y*. It should be near 2.

In addition to this simple verification of the OLS assumptions, it is important to

- check for "influential cases" capable of distorting regression coefficients: the "Cook distance" is used, measuring the effect of each object on *A* and *B* through its omission and the successive computation of regressions. When *A* or *B* vary greatly due to this omission, then the omitted object is influential.
- and to make sure that there are no "endogenous" variables. Endogeneity may arise due to misspecification of predictors. If residuals are correlated with the predicted variable, a basic OLS assumption is violated. Omitted variables should be included in the model in the form of proxy variables or else other estimation methods should be used.

When the OLS assumptions are violated or when there are influential cases, it is worth trying to transform the attributes, both explained and explanatory, to see if the difficulty disappears. The major transformations to apply to the *x*, to *y*, or to both are:

1. Logarithm of the attribute instead of the attribute to reduce the asymmetry and positive "leptokurtosis" and influential cases. It is effective when attributes are multiplicative, as is often the case in Finance.
2. Use the *x* squared to linearize relationships that are curved (concave or convex), to decrease negative asymmetry or heteroskedastic residuals.

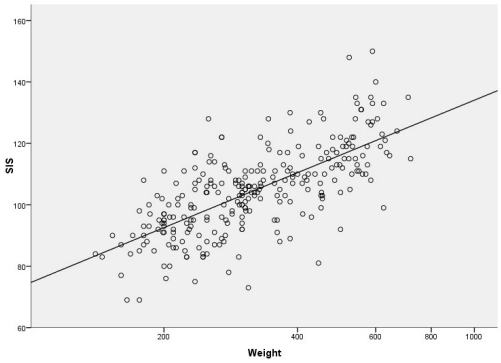
For instance, when we want to correct blood pressure for different weights of children, we use the logarithm of weight instead of just weight. The regression model is:

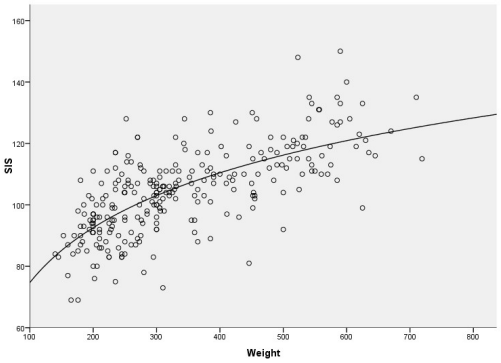
Systolic BP = -18.6 + 48.89 Log Weight + residual Adj. R Sq = 23%

Blood pressure "bends" downwards with increasing weight. Thus

- given heteroscedasticity (non-homogeneous variance), OLS coefficients are no longer optimal, but this is not the more damaging feature of heteroscedasticity.
- residuals (errors) associated with higher blood pressures are bigger than others,
- larger weight values become influential they may distort the model severely.

The OLS method is thus called into question but, when using the logarithm of weight, the relationship becomes linear and homogeneous, R^2 is bigger, the standard error of coefficients decreases, and no influential cases are observed.





If a model is made on purpose to study each independent variable x_j so that x_j appears as dependent and the other predictors as independent variables, then, the value of R^2 found will be the proportion of variability of x_j explained by the other predictors. An R^2 high, above 75%, shows that x_j is synonymous with other predicting attributes. Is a linear combination of the remaining x_j.

$1\text{-}R^2$, the proportion of variability of x_j not explained by other independent variables, is called "tolerance of the model to the entrance of x_j". Tolerance is circumstantial: in a model that contains

several independent variables, the input of one more x_j will lead to an indicator that expresses the tolerance of this model, as it stands, to the entry of x_j.

A low tolerance (less than 0.1) shows that:

1. x_j upon entering, will not add anything that the model does not have already.
2. The standard error of the coefficient B_j will be big
3. Computing difficulties and instability will emerge in the least-squares algorithm because, when there are independent variables which are linear combinations of others, the correlation matrix approaches singularity and cannot be reversed.

This is called "multicollinearity". To prevent it, tolerance of the attributes candidates to enter the model; or already entered, must be calculated. If multicollinearity associated with one of these attributes drops to less than 0.1, then the attribute or its synonym must be removed from the model. Therefore, the effect of each predictor x_j on y should be analyzed, and the appropriate tool to do this is the "partial plot", which calculates:

1. y residual when predicted by all attributes except x_j
2. x_j residuals when predicted by all other predictors.

The partial plot shows these two values for all objects in a scatterplot where y and x_j appear and the effects of all other independent variables have been removed.

The partial plot has some interesting properties:

1. The slope of the regression line is equal to the coefficient B_j that will appear in the final regression, i.e., after all variables have entered.
2. The correlation coefficient is the partial correlation coefficient between y and x_j.
3. The residuals of the regression are the same as those of the final regression.

Thus, the partial plot is good at detecting violations of OLS assumptions.

How to choose the attributes that should be included in a multiple regression? Faced with a collection of explanatory variables, the question is, which should be included in the regression, and which should be rejected.

A good model is one where a maximum of variability is explained by a minimum of predictors. In principle, any phenomenon can be approximated with a desired detail; just input more and more predictors. However, the degrees of freedom engaged should be much smaller than the number of objects in that sample, otherwise.

the gain in generalization ability that the model brings is lessened, and

the reliability of inference is low.

Criteria commonly used to select attributes as predictors in a multiple regression are:

1. The magnitude of the coefficients B_j is not the most appropriate criterion: the B_j are difficult to interpret as measures of the magnitude of the relationship between y and x_j because the difference in dimension of various x_j hides the true slope. When standardized attributes are used this difficulty vanishes.
2. The C_j (standardized slopes) resulting from the use of standardized attributes, can be used as a direct measurement of the interest of variable x_j to explain y but depend on the attributes already entered the model and are strongly affected by the possible correlation between the x_j. Therefore, they do not reflect the relative importance of each independent variable.
3. Variation of R^2: when x_j enters the model, the change then observed in the R^2 is a good indicator of its importance. It is a useful indicator that lends to testing the null hypothesis that

x_j (after removing the effects of other independent variables) is uncorrelated with y. R_j^2 being R^2 after x_j entered the model, and R_{j-1}^2 its value before x_j enters, then the j change in R^2 is R_{j-1}^2 -R_j^2. Thus, the use of the

R_j^2 partial = $\sqrt{\frac{j^{th}\ change\ in\ R^2}{1-R_j^2}}$ is equivalent to using partial plots but with the advantage of percentages being objective.

4. The *t* of each variable has the same inconveniences pointed to the C_j but, being a score of the significance of the variable, are more readily interpretable.

Algorithms that perform multiple linear regressions using the OLS method are equipped with procedures to carry out the automatic selection of predicting attributes namely,

1. Selecting "forward": attributes are chosen to enter one by one, according to their explanatory potential. Before entering, a new attribute is compared with the model, so that it may be rejected for lack of significance or low tolerance.
2. Elimination "Backward": at the beginning the forced entry of all candidate attributes. Then such attributes are removed from the model, one by one, the less significant attributes first.
3. Selection "Stepwise", similar to forward but where, after each attribute having entered, the model is evaluated in the perspective of eliminating attributes that may have lost significance.

In the example we've been using, blood pressure is explained by several other attributes using multiple linear regression. Only skinfold is significant.

The table below shows the coefficients of the model, the *t* values, the 95% confidence intervals (below and above the estimated coefficient) and a diagnosis of possible multicollinearity, the tolerance of the model to the entrance of each attribute. Weight and height have lower tolerances but none approaching 0.1.

OLS	Coefficient	Std. Error	Standardized Coefficient	t	Significance	95% lower in.	95% upper in.	Tolerance
Intercept	41.96	19.60		2.14	0.03	3.20	80.72	
Skinfold	0.61	0.23	0.25	2.65	0.01	0.15	1.06	0.77
Weight	-0.02	0.02	-0.14	-0.93	0.35	-0.07	0.02	0.33
Height	0.02	0.02	0.14	1.06	0.29	-0.02	0.05	0.40
Sexual mt.	-0.29	0.61	-0.04	-0.48	0.63	-1.49	0.91	0.79

-/-

It may happen that the predictors employed in a regression to explain the predicted variable are not the most appropriate ones. For instance,

1. One or several predictors are not significant. They do not apportion significant variability explained, being, in fact, useless.
2. Significant predictors are not being used in the regression. The regression is unable to explain variability that should be there but is missing.
3. Predictors are not appropriately transformed, leading to non-linear and / or heteroskedastic residuals.
4. The predicted variable is divided in strata but this is not being considered.
5. And so on.

The above cases are known as "model misspecification". Case 1 above is considered as non-important and case 2, known as "endogeneity" is viewed as extremely damaging. When studying more advanced topics, we shall address model misspecification.

-/-

How to interpret the coefficients of a linear regression?

In the first place, a positive, significant coefficient indicates that the influence of the predictor on the predicted variable is sizeable and direct. Increases in the predictor lead to increases in the predicted variable.

On the contrary, a negative, significant coefficient indicates that the influence of the predictor on the predicted variable is sizeable and inverse. Increases in the predictor lead to decreases in the predicted variable.

Finally, non-significant coefficients should not be the object of interpretation.

As for the magnitude of effects, coefficients directly show their influence in the overall model, but the *t* statistic is better suited to show statistical importance. The *t* statistic is directly linked to the probability of finding such a large coefficient value when, in the population, the coefficient is not different from zero.

The "ceteris paribus" (other things being kept unchanged) assumption allows coefficient interpretation. It consists of assuming that effects are independent of each other and, therefore, can be observed independently of each other. This is often not the case.

8.5 General Linear modelling

The analysis of variance may include, in addition to strata, some "covariate" attributes, which are continuous variables able to explain y_i. Or, conversely, a regression may also include, as explanatory variables, "dummy variables" to perform the role of strata.

Typically, in Sociology and Biology, the modelling of data is carried out using linear models with "factors", that is, ANOVA's strata plus covariates, that is, regressions. By contrast, in econometric studies the use of dummy variables is preferred to ANOVA.

-/-

For example, in the case of teenage (13-16 years old), when carrying out the analysis of variance of blood pressure by sex and by area, it is appropriate to also use the child's age as covariate attribute to correct blood pressures for the effect of different ages.

The analysis of variance and covariance is called "general linear model". If there are two stratifications (factors) and one covariate, the model has the following form:

$$y_i = M + d_k + d_q + I_{kq} + B\,x_i + e_i$$

where, for object *i*,

y is the explained or dependent attribute

M is the overall average or constant

d are strata K and Q;

I is the interaction between K and Q;

B is the coefficient denoting covariance of *x* with *y*;

x is the covariant attribute

e_i is what, in object *i* is left unexplained; it is the residual or error of the model.

There are several different assumptions and the corresponding algorithms available to estimate parameters of general linear models. Given this, general linear modeling may not be just an ANOVA followed by a regression. Results may vary according to the assumptions / algorithm used. Therefore,

in most cases, general linear models do not yield exactly the same results as regressions cum dummies.

In the usual example, blood pressure in children is explained by place (3 strata), gender (2 strata), weight, skinfold, sexual maturation.

Results are as follow:

			SSQ	df	MSQ	F	Significance
Systolic BP	Covariates	Both	516.99	2	258.50	2.64	0.07
		Skinfold	296.07	1	296.07	3.03	0.08
		Weight	51.84	1	51.84	0.53	0.47
	Main Effects	Both	320.80	3	106.93	1.09	0.35
		Sex	20.19	1	20.19	0.21	0.65
		Place	320.79	2	160.40	1.64	0.20
	Interaction	Sex * Place	4.46	2	2.23	0.02	0.98
	Model		933.28	7	133.33	1.36	0.23
	Residual		13879.09	142	97.74		
	Total		14812.37	149	99.41		

Despite previous results, it is now clear that skinfold is the sole significant effects to consider in explaining blood pressure for this age group.

-/-

When dummy variables are used in regressions to introduce stratification, each stratum originates one variable that may take on values of 1 and zero according to its being true or false. Thus, one stratification requires two or more dummy variables to be modelled in regressions. For example, the dummies required to model the attribute “sex” are:

Sex_is_M, which takes on the value 1 if sex is male and zero otherwise.

Sex_is_F, which takes on the value 1 if sex is female and zero otherwise.

The attribute “zone” would require three dummy variables, and so on.

In practice, since regressions already have a constant term, it is possible to use, for each stratification, one less dummy variable. Sex would require one dummy variable only, and zone would require two. But if the goal is to build interpretable models, not just to test hypotheses, this reduction in the number of dummies should be used with care to avoid meaningful effects to be mingled together in the constant term.

In the case of 11- to 12-year-old children, the effect of sex can be introduced in the regression explaining blood pressure. Dummy variable called “boy” is created to signal observations where the subject is boy. The R-Squared is low and the only significant effect is the dummy signalling the fact that the child’s mother is fat (uppermost quintile:

	coefficient	St. Error	Beta	t	P (t)
(Constant)	96.068	35.260		2.725	0.007
Log of Weight	5.697	20.192	0.043	0.282	0.778
Height	-0.007	0.018	-0.049	-0.364	0.716
Triceps Skinfold	0.388	0.257	0.161	1.509	0.134
boy	-1.053	1.861	-0.052	-0.566	0.573
Fat mother	7.816	2.629	0.254	2.973	0.004

It is frequent to use interactions as predictors. An interaction is created by multiplying

two dummies, or

one dummy and a continuous attribute, or

two continuous attributes.

In this way, the significance of effects involving more than one attribute can be tested. For example, when multiplying “boy” by “fat mother” dummies, we obtain a new dummy, “boy with fat mother”, which will test the hypothesis that boys with fat mothers have significantly higher blood pressure:

	coefficient	St. Error	Beta	t	P (t)
(Constant)	101.233	35.094		2.885	0.005
Log of Weight	2.167	20.171	0.016	0.107	0.915
Height	-0.004	0.018	-0.032	-0.235	0.815
Triceps Skinfold	0.445	0.238	0.185	1.871	0.064
Boy with fat mother	7.968	3.931	0.177	2.027	0.045

Similarly, when multiplying “fat mother” by the continuous attribute “Height”, we obtain a variable able to test the hypothesis that the influence of fat mothers on their children’s height has a significant effect on children’s blood pressure, and so on.

Chapter 9 Time-series overview

A time-series of, say, the daily price of a security or the yearly inflation in an economy, is indeed random. However, randomness differs in two characteristics from randomness that we find in cross-section. First,

- a time series is a "sequence": an ordered set of observations made in subsequent time periods. Second,
- a time series is "unique": contrary to cross-sectional samples, it is impossible to obtain more samples from the same population.

These two characteristics compromise the meaning of estimated parameters in models.

9.1 The variety of time series models

It is all important to understand from the outset that there are many different types of modelling and estimation problems, all of them covered by the same "time-series" label. Broadly speaking, the modelling of time-series seeks two distinct goals:

1. Forecasting, that is, the prediction of future observations, and
2. Model parameter estimation for the sake of finding significant effects.

Moreover, the approaches used by different areas of knowledge when studying the same models and estimation problems can be disconcertingly distinct and even opposed in their goals and methods. Major approaches are

- The Econometric approach, linked to Economics, seeks estimation.
- The Classical, forecasting approach, which basically consists of the Box-Jenkins method, seeks both estimation and forecasting.
- The "ad-hoc" forecasting approaches, linked to operations research.

Regarding the variety of modelling and estimation approaches, we have

on the one end, the pure "time-series" type of model

$$Y_t = M\ (t, e)$$

which aims at finding a model *M* with a set of parameters and functional form, capable of explaining a series Y_t in a parsimonious way, and so that residuals *e* are random, independent from each other, and time-independent.

The final goal of these models is either to forecast future vales of Y_t, or to discover the underlying random mechanism and parameters that governs Y_t.

Examples are series of prices, rainfall, temperature, all types of individual series.

On the other end, we have the purely static model

$$Y_t = M\ (x_1, x_2, \dots x_k, e)$$

which aims at finding a model *M* with a set of parameters and functional form, capable of explaining a series Y_t from other, equally time-dependent variables $x_1, x_2, \dots x_k$ in a parsimonious way, and so that residuals *e* are random, independent from each other, and time-independent.

The goal is to explain Y_t in terms of $x_1, x_2, \dots x_k$, the same as in regressions.

Time, here, underlies the predicted and some of the predicting variables, being an endogenous variable, not necessarily damaging to the model explaining Y_t . Most econometric models are of this second type.

Between these two extremes, we have all types of mixtures of the time-series and the static models

$$Y_t = M\,(t,u) + M\,(x_1, x_2, \dots x_k, e)$$

which aim at explaining a series Y_t from $x_1, x_2, \dots x_k$ and from time, in a parsimonious way, so that residuals are random, independent from each other, and time-independent.

$M\,(t)$ and $M\,(x_1, x_2, \dots x_k)$ may mingle together and may include "lagged" values of Y_t, for example, Y_{t-1}, or of other variables as predictors.

Time may be endogenous, or it may be explicitly accounted for in the model.

Some econometric models are of this type.

Basic time-series types are:

The "white noise" $Y_t = e$

The simple regression in time process or linear trend: $Y_t = a + b\,t + e$

The "random walk" process: $Y_t = Y_{t-1} + \ e$

The "autoregressive" of first order, AR (1) process:

$$Y_t = a + b\,Y_{t-1} + \ e \text{ with } b < 1.$$

The "static" regression model:

$$Y_t = a + b_1 x_1 + b_2 x_2 + \ \dots + e$$

… as opposed to the "dynamic" model:

$$Y_t = a + b_0 Y_{t-1} + b_1 x_1 + b_2 x_2 + \ \dots + e$$

The "moving average" of second order MA (2) process:

$$Y_t = x_t + a_1 x_{t-1} + a_2 x_{t-2}$$

…which is just the weighted average of a series with its two lags.

9.2 Time-series modelling

If we wish to understand time-series modelling, we need to become familiar with

Specific characteristics of time-series,

The tools used in modelling when modelling involves time, and

operations performed as part of time-series modelling, namely differencing and integrating, and moving averaging of residuals.

The specific characteristics are

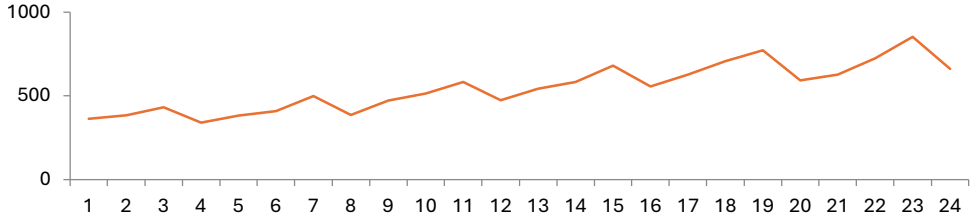
- "Trend": expected values of the series increase, decrease, or are steady with time
- "Seasonality": the time-series shows a pattern which repeats itself over fixed periods, namely yearly.
- Stationarity, weak dependence, its opposite, persistence, to be explained later on.

Time-series are also defined by not having any trend, not having seasonality, not being stationary, not being weakly dependent.

The tools are:

- Lagged predictors (also known as lagged dependent variables) lagged predicted variables, lagged error terms. For example, Y_{t-1}, Y_{t-2}, x_{t-2}, or e_{t-2}. A "first lag" is the value at the *t-1* period, the "second lag" is the value at the *t-2* period.
- "Autocorrelation" is the correlation between a variable and its lag. For example, the "second order" autocorrelation between Y_t and Y_{t-2} An "Autocorrelation function" (Correlogram, ACF) is the set of all the autocorrelations for a given variable, from zero-order to some higher-order.
- "Partial autocorrelation", "partial autocorrelation function", PACF, are similar concepts if autocorrelation is adjusted to account for the other autocorrelations.

For example, in a time-series of sales volume evolving over time, sales may show a growing "trend" plus a "seasonal" effect, that is, quarterly patterns which repeat themselves each year, plus some unexplained variability, as shown here:



The initial 8 periods are shown below, together with their lags:

period	Quarter 1	Quarter 2	Quarter 3	Quarter 4	Quarter 5	Quarter 6	Quarter 7	Quarter 8
sales	362	385	432	341	382	409	498	387
first lag of sales		362	385	432	341	382	409	498
second lag of sales			362	385	432	341	382	409
third lag of sales				362	385	432	341	382

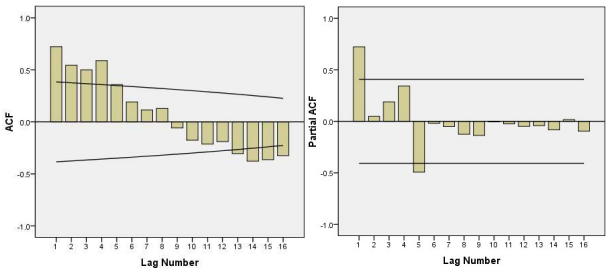
Note that lagged sales are a displacement to the right-hand side of the original sales values. Indeed, for period 2, the t-1 sales is 362, and so on.

The correlation between sales and the first lag of sales is the correlation between the first and second lines of the above table. It is the "first-order autocorrelation" of sales.

The "correlogram" or "autocorrelation function" (ACF) is the collection of all these autocorrelations, from lag zero- to some higher-order lag. An example is in the left.

The "partial autocorrelation function" (PACF) is the collection of all these correlations after adjusting each of them for the influence of all the other autocorrelations (right).

The graphics also show the 95% upper and lower confidence intervals for the estimation of the correlations.



Individual values from which the above graphics were obtained are:

lag	auto correla-tion	std. error	partial auto correla-tion	std. error
0	1.00	0.00	1.00	0.00

1	0.72	0.19	0.72	0.20
2	0.55	0.19	0.05	0.20
3	0.50	0.18	0.19	0.20
4	0.59	0.18	0.34	0.20
5	0.36	0.17	-0.49	0.20
6	0.19	0.17	-0.02	0.20
7	0.11	0.17	-0.05	0.20
8	0.13	0.16	-0.13	0.20
9	-0.06	0.16	-0.14	0.20
10	-0.18	0.15	-0.01	0.20
11	-0.21	0.14	-0.02	0.20
12	-0.19	0.14	-0.05	0.20
13	-0.31	0.13	-0.04	0.20
14	-0.38	0.13	-0.08	0.20
15	-0.36	0.12	0.02	0.20
16	-0.32	0.11	-0.09	0.20

9.3 Stationarity and weak dependence

Time-series are “stationary” or “on-stationary” depending on whether characteristics of its distribution change over time or not. The above example is non-stationary because the mean, variance, and time-dependence increase over time. Stationarity requires that the series be

> “identically distributed” and, for
>
> any time-periods, covariance does not change with time.

“Covariance-stationary” series are stationary up to the second moment only (covariance between contiguous periods). Strictly stationary time-series should also keep higher moments steady, not just the mean and variance between contiguous periods.

Moreover, time-series may or may not be

> “weakly dependent” if the lags of a series tend to be independent from each other with increasing lags between them. In practice, weak dependence means that autocorrelation functions tend to zero with increasing lags.

Weak dependence is important in time-series analysis because it is a requirement for the validity of OLS estimation. To say that residuals of a given model are weakly dependent is almost equivalent to posit that OLS estimation methods can be used. This is why time-series makes such intensive use of auto- and partial correlation functions.

Chapter 10 Estimation and inference with time-series

Here we revise and detail the set of assumptions under which the OLS method can be validly used in regressions. Then we introduce the corresponding set of assumptions applicable to time-series, so that comparisons may be made between the two cases.

10.1 Detailed description of OLS assumptions

First, let us introduce two definitions:

- "Bias" means "distortion". A biased regression coefficient is a model parameter that do not obey any desirable quality.
- "Consistency" means the steadiness of a desirable quality. A consistent regression coefficient is a model parameter that remains unbiased when estimation conditions change.

For example, a consistent estimation remains unbiased when the number of objects in the sample is small. An inconsistent estimation may become biased when the number of objects in the sample increases or decreases.

Estimators must be unbiased and consistent in the least.

The 4 assumptions under which OLS estimators are unbiased and consistent for the population are:

1. "MLR.1": Regression model is linear in parameters.
$$E(predicted \mid predictors) = a + bx$$
This assumption does not preclude the use of transformations.
2. "MLR.2": Random sampling.
3. "MLR.3": No perfect collinearity: no variable is constant, and no exact linear relationship exists among predictors. Collinear variables should be dropped.
4. "MLR.4": Zero conditional mean of residuals for any value of predictors:
$$E(residuals \mid predictors) = 0$$
Expected residuals are zero for any value of predictors. Mis-specified regression models include regressions where a variable is missing, some transformation is omitted, residuals are serially correlated, an interaction has not been accounted for or other omitted specification of the regression. All of this may lead to MLR.4 assumption being violated.

Based on whether, for a given regression predictor, MLR.4 is tenable or not, a central distinction is now introduced:

- When MLR.4 holds, we have an "exogenous" predictor.
- A predictor correlated with residuals is "endogenous".

The inclusion of irrelevant variables in regressions is known as "over-specification" of the model and does not affect the unbiasedness of the OLS estimators.

The exclusion of a relevant variable is known as "under-specification" of the model and causes OLS estimation to be biased via MLR.4. This is the "omitted variable bias".

After ensuring the unbiasedness and consistency of OLS estimation, there are two more assumptions worth considering.

5. "MLR.5": Homoscedasticity:

$$V(residuals \mid predictors) = C$$

That is, residuals have the same variance for any values of predictors.

When MLR.5 is violated, estimation, though possible, becomes less precise. That is, if standard errors associated with parameters can be estimated, these will be bigger.

Assumptions MLR.1 to 5 are the "Gauss-Markov" assumptions for small samples.

The Gauss-Markov theorem states that, under the same 5 assumptions, OLS is the best estimator that can be found.

6. "MLR.6": normality assumption. Residuals are independent of predictors and are Normal (we have accepted that the mean of residuals is zero and variance is stable). By accepting this assumption about the distribution of residuals, we can compute the standard errors of parameters, and therefore confidence intervals.

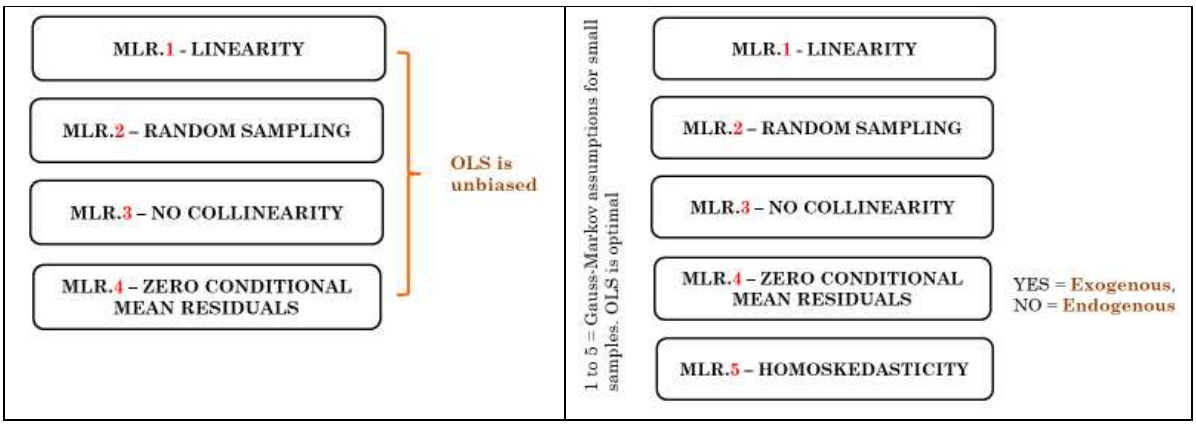
Assumptions MLR.1 to 6 are the "Classical Linear Modelling" assumptions and lead to the smallest possible variance of errors among all the possible unbiased estimators.

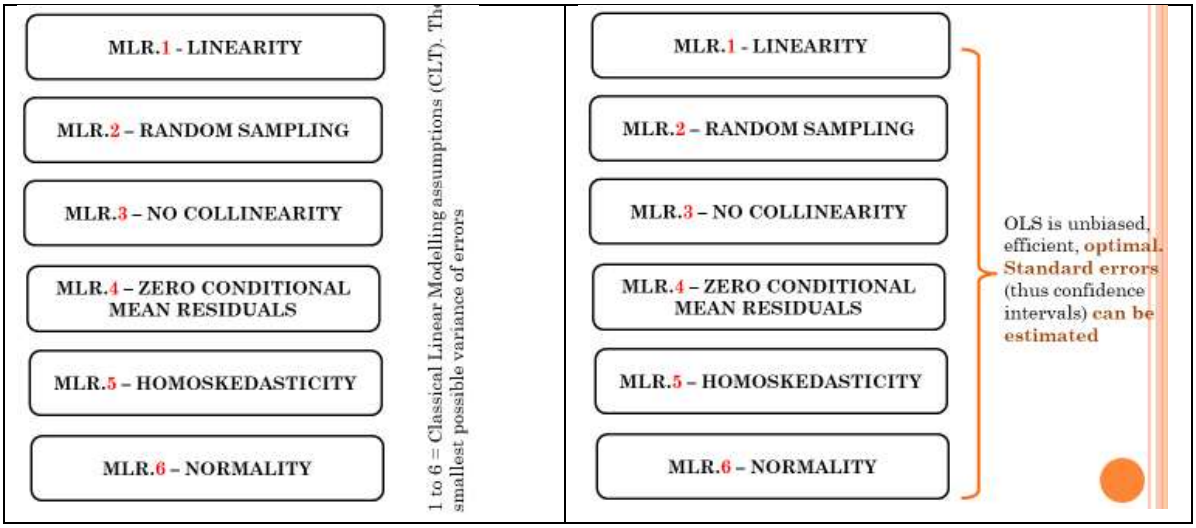
Assumptions MLR.1 to 6 are "finite sample" assumptions, meaning that they hold for small sample sizes, and not just when samples obey the law or large numbers. For large samples, MLR.6 may be relaxed if errors go to zero with increasing sample sizes. This is because of the central limit theorem, according to which, large samples would tend to be Normal if predictors act on the predicted variable in separate, additive ways, without interactions. However, correlated predictors are viewed as acceptable.

The fact that, in a proposed estimation method or solution, normality can be overlooked when the size of the sample increases, is known as the "consistency" of that solution.

Consistent methods, therefore, allow overlooking normality for large samples.

A graphical description of MLR.1 to MLR.6 follow:





There is yet another useful consideration to make. The "reduced" MLR.4 assumption requires that mean residuals, and all covariance of predictors with residuals, be zero:

$$E(e) = 0$$
$$COV(x|e) = 0$$

The reduced MLR.4 is contained in MLR.4, that is, where MLR.4 is verified, the reduced MLR.4 is also verified. But the reduced MLR.4 overlooks the possible mis-specification of the regression model plus serial correlation of residuals.

In some cases, it is sufficient for MLR.4 to hold in its reduced, not full form. However,

If mean residuals are not zero, MLR.4 fails and OLS estimates are biased,

when covariance is not zero, MLR.4 fails and OLS estimates are inconsistent, therefore, bias persists for large samples.

Models can be unbiased and consistent yet inference further require that the distribution of residuals be Normal, otherwise *p*-values from *t* and *F* statistics, and confidence intervals, cannot be estimated.

However, the Central Limit Theorem leads to "asymptotic normality of OLS": as the sample size grows, the sampling distribution of model parameters tend to Normal. In other words, so long as the estimation is consistent, Normality can be asymptotically presupposed, and estimation is possible.

10.2 OLS assumptions for time-series

Now we extrapolate the previous assumptions to include time-related variability. Time, a non-random count, may bring about correlation between successive occurrences of time-dependent variables, thus compromising the random sampling assumption.

When time-related variables are included as predictors in a regression, the following are the OLS assumptions:

"TS.1": The time-series process follows a linear model.

"TS.2": No perfect collinearity.

Note that, since MLR.2 (random sampling) is not verified, residuals are not necessarily independent of each other.

"TS.3": Zero mean residuals for any value of predictors, for all time periods.

$$E(u_t|x) = 0,\ t = 1,2,\dots$$

This is the crucial assumption:

Not just contemporary exogeneity but strict exogeneity.

Not just omitted variables, also measurement errors.

Moreover, residuals must not influence future values (we can control for the influence on past values).

Besides assumptions required for the unbiasedness and consistency of estimators, we further have:

"TS.4": Homoscedasticity.

"TS.5": Residuals are uncorrelated, also for different time periods.

$$CORR(e_t, e_s) = 0, for\ all\ t \neq s$$

This is required here because of the absence of random sampling.

TS.1 to 5 leads to unbiased estimators.

Gauss-Markov theorem: under TS.1 to 5, OLS linear unbiased estimators are optimal.

"TS.6": Residuals are independently and identically normally distributed.

TS.1 to 6 leads to the Classical Linear Modelling assumptions: the OLS estimators are normally distributed. *F* and *t* statistics have *F* and *t* distributions that can be computed.

Therefore, wherever TS.1 to TS.6 hold, all estimation and inference issues in cross-section also apply to time-series.

-/-

We have mentioned that time-series are stationary if distributions do not change over time, and weakly dependent if lags tend to be independent from each other with increasing lags between them. Otherwise, they are "strongly dependent".

Weak dependence means that autocorrelation functions tend to zero with increasing lags. The opposite situation is called "persistence" and it denotes strong dependence.

Stationarity and weak dependence are required for TS.1 to TS.3 to hold. If TS.4 and TS.5 also hold, then OLS estimators are asymptotically Normal.

If residuals are stationary, weak-dependent, then OLS estimation applies. This is why time-series models make such intensive use of autocorrelation functions: to make sure that weak dependence is verified.

Chapter 11 Types of time-series

This chapter offers examples of time-series, comparing them to the white noise type and its opposite, the random walk type, and showing their characteristics.

White noise is a time-series of independent, normally distributed observations. ACF and PACF are very small. The white noise series

$$Y_t = a + \sigma e$$

(a, σ constant) has mean *a* and standard deviation σ, being stationary: the mean and standard deviation can be predicted from past values. It is also weakly-dependent. Market returns r_t, that is, the relative differences between successive market prices,

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

approach white noise and expected values can be predicted from past values. Expected returns reflect the expected profitability of each security.

11.1 Trends

The white noise with a trend is

$$Y_t = a + b\,t + e$$

where residuals are independent from each other, random with zero mean. The ACF is not small because it denotes the trend, but the PACF is small.

An important point worth making now is that trends are unrelated to persistence, in spite of the fact that trends have the same effect on ACF as persistence, namely increasing and expanding higher-order autocorrelations. Therefore, autocorrelations that will not go to zero quickly can stem from two different causes, persistence or trends.

The white noise with linear trend is an example of a regression-in-time.

The slope of the regression, *b*, is the series' growth.

The process is non-stationary with an increasing mean, *a* and *b* can be predicted.

How to deal with linear trends?

These processes can be modelled simply by running an OLS regression with time as the predictor. In this way, the process is adequately described by its two parameters *a* and *b*. We can also use differencing $Y_t - Y_{t-1}$ to the same effect.

So long as residuals are independent and random, the model is unbiased.

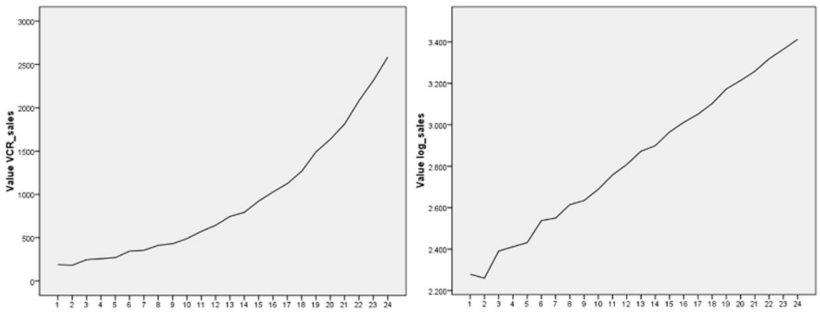
-/-

The white noise with an exponential trend is a non-linear regression-in-time:

$$Y_t = \exp(a + b\,t + e)$$

where *e* are random, zero-mean and independent from one another residuals. This process is a time-regression in logarithmic space:

$$\log(Y_t) = a + b\,t + e$$



The white noise with exponential trend is very common, as it reflects the growth in the sales of new products and other processes.

It can be dealt with simply by taking logarithms of the series and then fitting an OLS regression to the transformed series.

The slope of the transformed regression, *b*, is the growth rate of the series. It is constant.

The transformed process often is stationary, with *a* and *b* being predictable.

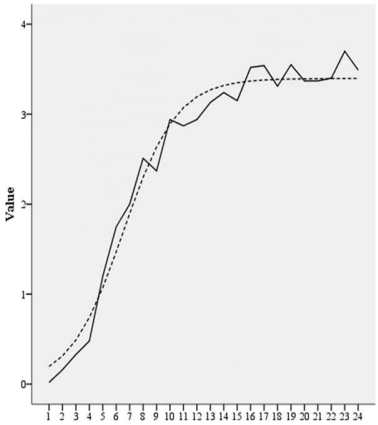
So long as residuals are independent and random, any OLS model is unbiased. This can be ascertained by examining ACF and PACF, which should be small.

-/-

The white noise with logistic trend is a logistic curve with white noise superimposed:

$$Y_t = \frac{a_1}{a_2 + \exp(kt + e)}$$

a_1, a_2, and *k* are constant parameters that can be fit using non-linear optimization algorithms. The *e* are random, independent zero-mean residuals.



The white noise with a logistic trend is very common, as it reflects the growth in the sales of new products until they attain market maturity.

This time-series can be dealt with by fitting sales to a logistic curve by means of a non-linear regression algorithm. The fitting process is iterative and akin to “maximum likelihood” estimation, whereby a set of parameters is chosen because of being optimal in some sense. Here, optimality consists of the most likely fit.

The parameters of the non-linear regression have interesting meanings and residuals often are stationary.

So long as residuals are independent and random, the model is unbiased. This can be ascertained by examining the ACF and PACF of residuals, which should be small.

11.2 Persistence, Integration, Differencing

Having exemplified some common processes with trends, we now turn to another type of series where persistence, not the trend, is the important characteristic. Here, we also explain the two operations performed in time-series, differencing and integration.

The "random walk"

$$Y_t = Y_{t-1} + e$$

may have a drift when $Y_t = a + Y_{t-1} + e$ but what matters most is its persistence.

Random walks are not just non-stationary. They have no trend. Except for the $t + 1$ case, mean values cannot be forecast from the past.

The variance of random walks increases with *t*.

Random walks are "strongly" rather than weakly dependent: they are persistent.

Persistence is not the same as a trend, in spite of the fact that trends have almost the same effect on ACF as persistence, namely increasing and expanding higher-order autocorrelations.

Therefore, autocorrelations that will not go to zero quickly can stem from two different causes, persistence or trends.

Due to persistence, random walks as well as AR (1) processes with coefficients near the unit, may generate illusory trends. When we introduced the econometric approach, we implicitly accepted that the time-series under analysis had a trend, and not an illusory trend caused by a unit root (extreme persistence).

Regressions performed on nonstationary time-series give spurious results if the trend is illusory, being the result of persistence.

Many time-series have true trends. If a process returns to its trend after a shock, then it is not persistent and the trend is not illusory. True trends are called "trend-stationary". Trend-stationary time-series can be made stationary by removing the trend.

Random walks in continuous time are known as "Brownian motion", to be studied later.

Market prices obey a type of random walk in logarithmic space of this form:

$$Y_t = Y_{t-1} \exp(a + \sigma e)$$

with *a* (the drift) and σ (the spread) being constant.

"Differencing" is an operation that can be performed on series. It consists of building a new time-series from another, subtracting lagged values from current values:

$$new\ Y_t = Y_t - Y_{t-1}$$

Differencing a time-series which approaches random walk characteristics can bring the series to stationarity. For instance, market prices are random walks in logarithmic space:

$$Y_t = Y_{t-1} \exp(a + \sigma e)$$

with *a* (drift) and σ (spread) being constant for small periods of time. Therefore,

$$\frac{Y_t}{Y_{t-1}} = \exp(a + \sigma e)$$

$$\log \frac{Y_t}{Y_{t-1}} = a + \sigma e$$

$$logY_t - logY_{t-1} = a + \sigma e$$

To estimate the distribution moments of returns we difference logs of market prices.

"Integrating" is reversed differencing. It consists of building a new time-series from another, successively adding lagged values to current values:

$$new\ Y_t = Y_t + Y_{t-1}$$

Integration followed by differencing, or the reverse, do not change a time-series. This table shows an example of differencing and integrating a time-series of closing prices:

Date	Jan 31	Feb 28	Mar 31	Apr 30	May 31	Jun 30
Closing price	54	24	35	12	8	40
Integration	54	78	113	125	133	173
Difference		-30	+11	-23	-4	+32

It is usual to identify time-series by the number of times they have to be differenced to become similar to white noise, that is, to obey classical linear modelling assumptions:

Time-series are said to be "integrated of order zero", I (0), if they do not need any differencing to obey the classical linear modelling assumptions (in practice, if they are stationary and weakly dependent). This is the case of white noise.

Time-series are said to be "integrated of order one", I (1), if they require differencing to obey the classical linear modelling assumptions. This is the case of random walk and market prices, which when differenced, become stationary, weak dependent.

Time-series are said to be "integrated of order 2", I (2), if they need double differencing to obey classical linear modelling assumptions, and so on.

Conversely, the label "integrated of order N", or I (N), also shows how many times the white noise series must be integrated to become similar to the actual series. Market prices, for instance, are said to be integrated of order 1 because white noise must be integrated once to become similar to prices.

Time series that are integrated of order one exhibit illusory trends. These series are termed "difference-stationary", or are referred to as having a "unit root". In short, series are termed as

Trend-stationary if they have true trends, or

Difference-stationary is their trend is illusory.

The unit root test is a statistical test aimed at ascertaining whether a series is trend- or difference-stationary.

11.3 The meaning of AR, MA, I, ARMA, ARIMA

To understand the meaning of widely used AR, MA, I, ARMA, and ARIMA, remember that these labels do not mean operations that a given model performs. Rather, the labels are the list of operations that a given time-series requires until it becomes stationary and weakly dependent, that is, similar to white noise. Therefore,

- AR means to account for auto-regression: an AR (1) time series must be striped of its autoregression by means of the inclusion of first lag in the model before becoming similar to white noise.
- I means to account for integration: A I (1) time-series must be differenced before becoming similar to white noise.
- MA means to account for a moving average: the residuals of an MA (1) time-series must be smoothed by averaging residuals and the lag of residuals before becoming similar to white noise.

The first-order autoregressive process or AR (1), is

$$Y_t = a + bY_{t-1} + e$$

with $b < 1$. Values of the slope equal and above the unit are no longer autoregressive.

The goal of using a lagged predicted variable as a predictor in a regression is to subdue autocorrelation and thus obtain identically independent residuals.

When residuals are not weakly-dependent after the introduction of a first lag, higher order autoregressive terms are included. AR (2), for instance, would be:

$$Y_t = a + b_1Y_{t-1} + b_2Y_{t-2} + e$$

and so on.

Autoregressive processes are used extensively, namely in dynamic models and in the Box-Jenkins toolbox as a way to obtain well-behaved residuals.

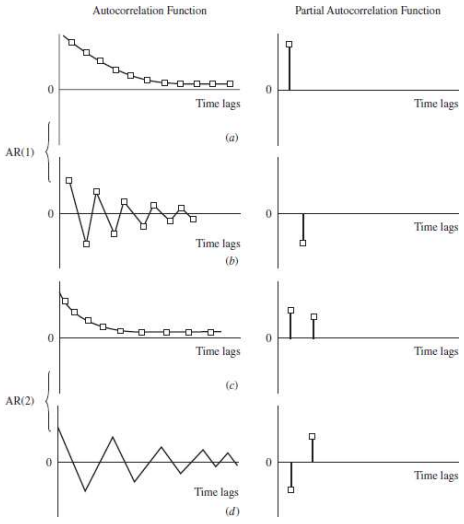
The random walk is a special case of AR (1) where $b = 1$.

AR (1) processes with b near 1 are no longer weakly dependent. In this case,

no matter how far in the future we look, our best prediction always is the previous value of a series and the future depends on the past.

Autocorrelation will not vanish with time: this series is persistent, instead of asymptotically uncorrelated. Persistent series may also exhibit pseudo-trends, that is, apparent trends that, in fact, are not caused by a trend.

Autoregressive AR (1) have typical ACF: large, decreasing autocorrelation functions denoting persistence, and PACF where only the first autocorrelation is big.



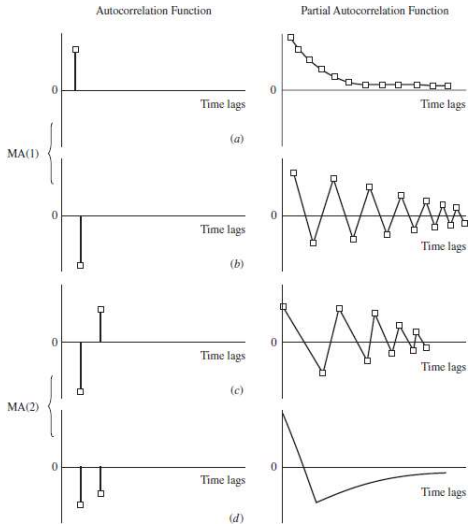
Weakly dependent processes are integrated of order zero, I (0), because they do not need any transformation to obey the classical linear modelling assumptions. Persistent processes, by contrast are I (1) if they need first differencing, or I (2) if they require first and second differencing to become weakly dependent and obey classical assumptions.

Moving average is an operation performed to smooth away cycles in residuals. Given identically, independently distributed e_t, each with a Normal distribution having zero mean and the same variance, the 1st order moving average model MA (1) is:

$$x_t = \mu + e_t + \theta_1 e_{\{t-1\}}$$

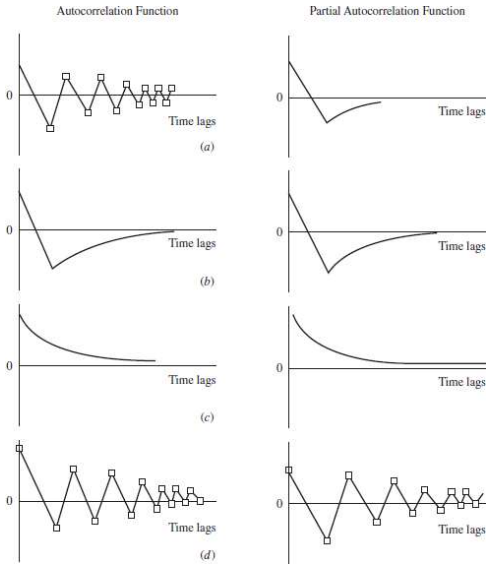
MA (2), MA (3) and so on, are extensions of MA (1). Basically, a moving average is a weighted average of a series and their lags, where the θ_k are the weights.

Moving averages MA (1) and (2) have typical ACF, the first autocorrelation being large and the other order autocorrelations are small with persistent PACF as shown here:



Specific time-series where a weighted average of cases and their lags can account for much variability, are also known as “moving average processes”.

Time-series where AR (1) and MA (1) features exist are called ARMA (1,1) processes. Their typical ACF and PACF may look like this:



11.4 Econometric models

Until now, we have studied time-series in isolation, and we endeavored to characterize their most relevant types. However, it is worth analyzing, not just isolated time-series but also the relationships that may exist among time-series, and the role that time-series may play in other models. Indeed, time-series can be correlated with other time-series and they can also be part of a more complex analysis. Although multivariate problems involving time-series are not going to be covered here, it is worth mentioning a few relevant cases.

In the static, as opposed to dynamic regression model, time is present only implicitly and is not modelled because one or more of the predictors is a proxy for time. In

$$Y_t = a + b_1 x_{1t} + b_2 x_{2t} + \ \ldots + e$$

both the predictors x_i and the predicted variable Y_t are influenced by time. So long as OLS assumptions are not violated, namely the *e* are independent random, then it is appropriate to use them for prediction and inference.

In case OLS assumptions are violated, the regression model with time

$$Y_t = a + b_1 x_1 + b_2 x_2 + \ \ldots + b_k t + e$$

may, or may not, solve the problem. If not,

it is worth trying autoregressive and other modelling techniques.

An epidemiological example follows: 821 children aged 5- to 18-year-old. We are interested in identifying fat children, so we introduce height, skinfold and sexual maturation as control variables and then examine the upper decile of weight.

coefficients		Std. Error	Beta	t	sig.
constant	-373.903	18.161		-20.588	0.000
height	0.431	0.018	0.564	23.412	0.000
triceps skinfold	3.585	0.308	0.137	11.642	0.000
sexual maturatic	18.351	0.934	0.408	19.658	0.000
dummy for boy	6.167	2.759	0.024	2.235	0.026
dummy for buti	2.327	3.942	0.009	0.590	0.555
dummy for caxi	13.731	3.833	0.053	3.582	0.000
age	0.383	0.988	0.010	0.388	0.698

It is observed that the addition of time (age) did not add to significance, variability explained, or the randomness / independence of residuals. This is because height and other variables in the model are a "proxy" for age.

In static models, the use of proxy variables which, not being time themselves, are able to account for the passing of time, avoids omitted variables biases.

The dynamic model adds an AR (1) term to the static model:

$$Y_t = a + b_0 Y_{t-1} + b_1 x_1 + b_2 x_2 + \ \ldots + e$$

where b_0 accounts for first-order autocorrelation in the predicted variable, measuring the persistence of the *Y*. The following example models the determinant factors of CR, the Tier 1 ratio of East-Asian banks. CR_LAG is the lagged CR.

coefficient values:		Std. Error	Beta	t	Sig.
constant	0.279	0.167		1.664	0.10
CR_LAG	0.696	0.040	0.750	17.327	0.00
ROA	0.033	0.014	0.157	2.342	0.02
ROE	-0.004	0.001	-0.207	-3.315	0.00
CI	0.000	0.000	0.022	0.547	0.58
LLR	0.000	0.002	0.005	0.146	0.88
NIM	0.007	0.004	0.077	1.660	0.10
LA	-0.00	0.000	-0.144	-3.781	0.00
GDP	0.007	0.011	0.025	0.640	0.52
DIM	0.004	0.007	0.026	0.588	0.56

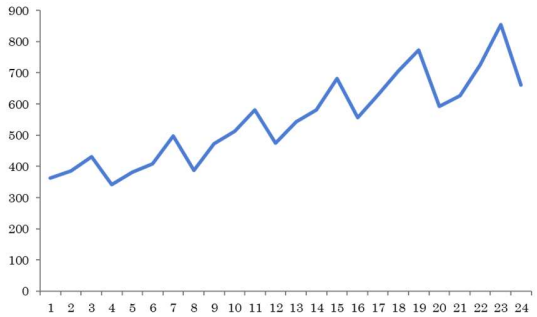
The coefficient of CR_LAG measures the persistence of CR, being positive and highly significant in this case. LA, loans to assets ratio, and the return on assets and equity ratios are also significant in explaining the capital ratio of banks. However, ROA and ROE show opposite effects on CR.

Chapter 12 Time-series methods compared

Consider the already mentioned 24-month time-series in the graphic:

Sales shows a trend, not exactly linear

There is a clear "seasonal" effect: quarterly patterns repeat themselves yearly.



This example is going to be used here to illustrate three different approaches available to deal with series exhibiting trend / seasonality, namely:

- The Econometric method based on AR terms and regressing in time.
- Ad-hoc forecasting methods, where there is no model, no parameters, but the fit can be good and residuals weakly dependent.
- Classical forecasting (Box-Jenkins) based on AR, MA, and differencing.

-/-

In the Econometric approach we define a model

$$SALES_t = a + b_0 SALES_{t-1} + b_T t + b_1 mar_t + b_2 jun_t + b_3 sep_t + b_4 dec_t + e$$

This is an AR (1) plus time, plus 4 seasonality dummies indicating the quarter to which an observation belongs. The R-Squared is very high, around 0.9, but the autoregressive component is non-significant in this case. The other model parameters are

coefficients		Std. Error	Beta	t	Sig.
Const.	303.70	14.536		20.894	0.000
period	18.088	0.829	0.911	21.827	0.000
jun	32.746	16.033	0.103	2.042	0.055
sep	97.825	16.097	0.308	6.077	0.000
dec	-54.929	16.204	-0.173	-3.390	0.003

The residual's ACF and PACF are small (weak dependent) and the fit is also good. The model formulation is improved by including the term $b_0 SALES_{t-4}$, not just $b_0 SALES_{t-1}$ to account for autoregressive seasonality.

-/-

In the Ad-hoc approach, the trend is found using smoothing, then, from the detrended series we calculate the seasonal effect. We follow this sequence:

1. the series is smoothed using one-year (4 quarters) "moving" averages calculated from quarter 3 onwards,
2. then the resulting series is centered using 2-month "moving" average of quarters 3 and 4. The resulting series is the trend effect.
3. The seasonal factor is found by dividing the trend effect by the original series.

4. Seasonal factors from the same quarter are brought together and averaged to become the seasonal effect.
5. The predicted sales = trend effect / seasonal effect.
6. "Residual" sales = predicted sales – sales.

In this case, residuals seem to be heteroscedastic.

Trends and seasonal effects can be calculated in different ways. It is common to use "additive", not "multiplicative" seasonality, which adds, not multiplies, the trend.

-/-

The Box-Jenkins approach combines three tools to fit the data:

Integration, which is the use of differencing. I (1), that is $Y_t - Y_{t-1}$. This is the first operation to apply.

Auto-regression, which is the use of lagged values to predict current values. For example, use AR (1) and AR (2). This is the second tool to be used

Moving average, the use of current and lag residuals to predict present values. MA (1) and MA (2) smooth away the cycles in residuals. This is the final tool.

An ARIMA (1,1,1) is a model where the series is first differenced, then AR (1) explains auto-correlation and then MA (1) smooth residuals of this model.

To decide which ARIMA orders to include, it is customary to examine the ACF and PACF. Ultimately, it is trial and error. The goal is IID residuals, no persist ACF PACF.

How to interpret ACF and PACF plots of a series?

- If ACF declines gradually and PACF drops instantly, use AR terms. The more gradually ACF declines, the closer the AR series is to the random walk (a random walk has not declining ACF and small PACF except for lag 1).
- If ACF drops instantly and PACF declines gradually, use MA.
- If ACF and PACF decline gradually, combine AR and MA models (ARMA).
- If both ACF and PACF drop instantly, series are white noise.

The Box-Jenkins model applied here is an AR (1), I (1), MA (1) model, or ARIMA (1,1,1), that is, the series is differenced (which cuts 1 quarter plus 1 month), applied an AR (1), plus an MA (1) to reduce cycle. Results are as follow:

Sales series, ARIMA (1,1,1)			Estimate	SE	t	Sig.
sales	Constant		17.86	10933	0.002	0.999
	AR	Lag 1	-0.612	0.501	-1.223	0.245
	Diff.		1			
	MA	Lag 1	-0.997	13.357	-0.075	0.942
	AR Season	Lag 1	-0.992	0.275	-3.610	0.004
	Season Diff.		1			
	MA Season	Lag 1	-0.887	1.917	-0.463	0.652
year not periodic	numerator	Lag 0	-0.001	5.419	0.000	1.000
quarter period 4	numerator	Lag 0	-6.014	3.766	-1.597	0.136

Note that AR and MA terms are made to agree with the seasonal plus the year effects. This is why two parameters exist for each AR, MA, and I, the latter being 1. The residual's ACF and PACF are small (weak dependent) and the fit is also good.

Chapter 13 Model misspecification, instrumental variables

Here, and in the following chapters, we discuss the remedies which should be used to overcome difficulties that faulty or very specific data may cause to inference.

When modelling relationships, we often face

Missing, intractable or unobserved variability of attributes,

missing objects or intractable distributions.

Both frailties may cause biases and inconsistencies in OLS estimation.

First, we discuss how to overcome biases caused by missing variability. When studying probit models, we shall tackle the problem of missing data and intractable distributions.

13.1 Proxies and lagged dependent variables

Model misspecification may occur in all types of predictive models.

If key variability is not accounted for, correlation between predictors and residuals may occur. This leads to bias and inconsistency in all the model.

When missing variability is an algebraic function of a predictor (for example, it is a logarithm), this is known as **"**functional form misspecification".

Failing to include dummies or interactions in a model may lead to functional form misspecification as well.

To test whether a given predictor or transformation should be included, we can compare ANOVA's F tests before and after inclusion. For instance, we want to predict wages of workers using years of education and experience plus a dummy indicating female worker, as predictors (file WAGE1"). Transformations (logarithms, powers) and interactions (one dummy multiplied by one predictor) are tested for inclusion as in:

$$\log(wage) = a + b_1\, educ + b_2\, exper + b_3\, exper^2 + b_4\, female + b_5\, female * educ$$

We compare F statistics before and after inclusion. Specifically, we test the significance of F changes before and after inclusion of each predictor. Results are:

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.431[a]	.186	.184	.48008	.186	119.582	1	524	.000	
2	.548[b]	.300	.298	.44550	.114	85.510	1	523	.000	
3	.594[c]	.353	.349	.42893	.052	42.192	1	522	.000	
4	.632[d]	.400	.395	.41345	.047	40.817	1	521	.000	1.776

a. Predictors: (Constant), educ

b. Predictors: (Constant), educ, female

c. Predictors: (Constant), educ, female, exper

d. Predictors: (Constant), educ, female, exper, square of exper

e. Dependent Variable: log of wages

All F changes are significant. The marginal contribution of the entered variables to the model's overall significance (F value) is shown here to be significant. One non-significant variable, the female–education interaction, did not enter the model:

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	exper	.264[a]	6.653	.000	.279	.910
	square of exper	.187[a]	4.549	.000	.195	.890
	female	-.339[a]	-9.247	.000	-.375	.993
	female_educ	-.332[a]	-9.008	.000	-.366	.993
2	exper	.240[b]	6.496	.000	.273	.906
	square of exper	.166[b]	4.341	.000	.187	.887
	female_educ	.000[b]	-.004	.996	.000	.044
3	square of exper	-.795[c]	-6.389	.000	-.270	.074
	female_educ	-.046[c]	-.273	.785	-.012	.044
4	female_educ	-.007[d]	-.046	.963	-.002	.044

a. Predictors in the Model: (Constant), educ

b. Predictors in the Model: (Constant), educ, female

If models are misspecified due to endogenous (missing) variability we can use "proxy" variables. A proxy variable is an available variable that is used in a model in lieu of an unavailable variable. We simply plug-in the proxy as a predictor.

Suppose that we want to explain the same wages level and we wish to introduce ability as a model predictor (file "WAGE2"):

$$\log(wage) = a + b_1\, educ + b_2\, exper + \cdots + b_k\, ability$$

Now, we don't have data on worker's ability, so we use IQ as a proxy. We trust that IQ contains variability on ability. For proxy variables to be validly used, it is essential that

they contain variability in common to the proxied feature.

they are not correlated to the other predictors in the model.

The results of predicting the logarithm of wages without using IQ as a proxy are:

```
Coefficients:  Estimate Std. Error t value Pr(>|t|)
  Intercept    5.395497   0.113225  47.653  < 2e-16 ***
  educ         0.065431   0.006250  10.468  < 2e-16 ***
  exper        0.014043   0.003185   4.409 1.16e-05 ***
  tenure       0.011747   0.002453   4.789 1.95e-06 ***
  married      0.199417   0.039050   5.107 3.98e-07 ***
  south       -0.090904   0.026249  -3.463 0.000558 ***
  urban        0.183912   0.026958   6.822 1.62e-11 ***
  black       -0.188350   0.037667  -5.000 6.84e-07 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Residual standard error: 0.3655 on 927 degrees of freedom
Multiple R-squared:  0.2526,       Adjusted R-squared:  0.2469
F-statistic: 44.75 on 7 and 927 DF,  p-value: < 2.2e-16
```

Estimated return to education is 6.5%. If we think that the omitted ability is positively correlated with education then we assume that this estimate is too high. The results of predicting the logarithm of wages using IQ as a proxy are:

```
Coefficients:  Estimate  Std. Error  t value Pr(>|t|)
  Intercept)   5.1764391  0.1280006  40.441  < 2e-16 ***
  educ         0.0544106  0.0069285   7.853 1.12e-14 ***
  exper        0.0141459  0.0031651   4.469 8.82e-06 ***
  tenure       0.0113951  0.0024394   4.671 3.44e-06 ***
  married      0.1997644  0.0388025   5.148 3.21e-07 ***
  south       -0.0801695  0.0262529  -3.054 0.002325 **
  urban        0.1819463  0.0267929   6.791 1.99e-11 ***
  black       -0.1431253  0.0394925  -3.624 0.000306 ***
  IQ           0.0035591  0.0009918   3.589 0.000350 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Residual standard error: 0.3632 on 926 degrees of freedom
Multiple R-squared:  0.2628,       Adjusted R-squared:  0.2564
F-statistic: 41.27 on 8 and 926 DF,  p-value: < 2.2e-16
```

Therefore, when IQ is added to the equation, the return to education falls to 5.4%, which corresponds with our prior beliefs about omitted ability bias.

When we further introduce, besides IQ, the interaction between education and IQ as a predictor, results are:

```
Coefficients:    Estimate   Std. Error t value Pr(>|t|)
  Intercept     5.6482478  0.5462963  10.339  < 2e-16 ***
  educ          0.0184560  0.0410608   0.449 0.653192
  exper         0.0139072  0.0031768   4.378 1.34e-05 ***
  tenure        0.0113929  0.0024397   4.670 3.46e-06 ***
  married       0.2008658  0.0388267   5.173 2.82e-07 ***
  south        -0.0802354  0.0262560  -3.056 0.002308 **
  urban         0.1835758  0.0268586   6.835 1.49e-11 ***
  black        -0.1466989  0.0397013  -3.695 0.000233 ***
  IQ           -0.0009418  0.0051625  -0.182 0.855290
  I(educ*IQ)    0.0003399  0.0003826   0.888 0.374564
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3632 on 925 degrees of freedom
Multiple R-squared:  0.2634,   Adjusted R-squared:  0.2563
F-statistic: 36.76 on 9 and 925 DF,  p-value: < 2.2e-16
```

IQ has a statistically significant, positive effect on earnings, after controlling for several other factors. Everything else being equal, an increase of 10 IQ points is predicted to raise monthly earnings by 3.6%. The standard deviation of IQ in the U.S. population is 15, so a one standard deviation increase in IQ is associated with higher wages of 5.4%. This is identical to the predicted increase in wage due to another year of education. Education has an important role in increasing earnings, even though the effect is not as large as originally estimated.

Adding IQ to the equation only increases the R-squared from .253 to .263. Most of the variation in log (wage) is not explained by predictors. Also, adding IQ to the equation does not eliminate the estimated earnings difference between black and white men: a black man with the same IQ, education, experience, and so on, as a white man is predicted to earn about 14.3% less, and the difference is very statistically significant.

The interaction term I (educ*IQ) allows for the possibility that *educ* and IQ interact in determining log (wage). We might think that the return to education is higher for people with more ability, but this turns out not to be the case: the interaction term is not significant, and its addition makes *educ* and *IQ* individually insignificant while complicating the model. (Wooldridge p. 300).

-/-

A "lagged dependent variable" is a particular type of proxy variable where past value of the variable being explained is used as predictor. Historical factors may explain current variability in the predicted variable. If we want to control for this variability, the use of lagged dependent variables is fitting.

Suppose that we want to explain the log of the crime rate in 46 US cities in 1987 as a function of unemployment rate and expenditure in law enforcement as in "CRIME2":

$$\log(crmrte) = a + b_1\, unem + b_2 \log(lawexpc) + b_3 \log(crmrte_{1983})$$

To account for past crime rate and other specific variability, we include the log of the crime rate in 1983. This is the lagged dependent variable. Results show how important the lagged variable is. Before including, the R square is 0.13 and coefficients are

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3.343	1.251		2.673	.011
	unemployment rate	-.029	.032	-.135	-.897	.375
	log of lawexpc	.203	.173	.177	1.178	.245

a. Dependent Variable: log of crime rate

After inclusion, the R square becomes 0.62 and the coefficients are

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.076	.821		.093	.926
	unemployment rate	.009	.020	.040	.442	.661
	log of lawexpc	-.140	.109	-.121	-1.285	.206
	log of 1983 crime rate	1.194	.132	.871	9.038	.000

Without the lagged crime rate in the equation, the effects of the unemployment rate and expenditures on law enforcement are counterintuitive; neither is statistically significant, although the *t* statistic on *log (lawexpc)* is 1.17. One possibility is that increased law enforcement expenditures improve reporting conventions, and so more crimes are reported. But it is also likely that cities with high recent crime rates spend more on law enforcement.

Adding the log of the crime rate from five years earlier has an effect on the expenditures coefficient. The elasticity of the crime rate with respect to expenditures becomes -0.14, with t = -1.28. This is not strongly significant, but it suggests that a more sophisticated model with more cities in the sample could produce significant results. Not surprisingly, the current crime rate is strongly related to the past crime rate. The estimate indicates that if the crime rate in 1982 was 1% higher, then the crime rate in 1987 is predicted to be about 1.19% higher. We cannot reject the hypothesis that the elasticity of current crime with respect to past crime is unity t = (1.194-1)/.132 is approximately 1.47. Adding the past crime rate also increases the explanatory power of the regression but this is no surprise. The primary reason for including the lagged crime rate is to obtain a better estimate of the *ceteris paribus* effect of *log (lawexpc)* on *log (crmrte)*. (Wooldridge p. 300).

13.2 Repeated measures, fixed effects, differencing

CRIME2 dataset has two cross-sections: the same objects are observed in the year 1982 and then again in 1987. Data are about crime and unemployment rates for 46 US cities.

What happens if we use the 1987 cross-section and run a simple regression of *crmrte* on *unem*? We obtain $\widehat{crmrte} = 28.38 - 4.161\, unem$ with N = 46 objects and R Sq. = 0.033. If we interpret the estimated equation causally, it implies that an increase in the unemployment rate lowers the crime rate. This is certainly not what we expect. The coefficient on *unem* is not statistically significant at standard significance levels: at best, we have found no link between crime and unemployment rates.

This simple regression equation likely suffers from omitted variable problems. One possible solution is to try to control for more factors, such as age distribution, gender distribution, education, law enforcement efforts, and so on, in a multiple regression analysis. But many factors might be hard to control for.

We now explore another remedy, known as “repeated measures”, which is part of a more general “fixed effects” modelling.

We showed before how including the *crmrte* from a previous year—in this case, 1982— help control for the fact that different cities have historically different crime rates. This is one way to use repeated observations to estimate a causal effect. An alternative way is to view the unobserved factors affecting the predicted variable as being of two types:

those that are constant over time but change from object to object (*OBJECT*),

those that vary over time but are the same for all objects (*TIME*).

If *i* denotes the cross-sectional object and *t* the time period, we can write a linear model

$$y_{it} = a + OBJECT_i + TIME_t + b_1 x_{it} + e_{it}$$

where predicted variable y_{it} is explained by a two-ways ANOVA with factors $OBJECT_i$ and $TIME_t$ plus covariates x_{it}. In this model, *i* denotes a person, firm, city, bank, and so on, and *t* denotes the time period. Remember,

The factor $OBJECT_i$ is constant for the different time periods, which is why it has no *t* subscript,

the factor $TIME_t$ is constant for the different objects, which is why it has no *i* subscript.

In our example, overall trends will cause crime rates in all US cities to change, perhaps markedly, over a five-year period. This effect will be captured by the intercept term *a*. Moreover, in some US cities the crime rate will increase more or less than the US average over the same period. The factor $OBJECT_i$ will capture all the city-specific unobserved factors that affect y_{it} and did not change over time. Indeed, the fact that $OBJECT_i$ has no t subscript tells us that it does not change over time. $OBJECT_i$ is called an "unobserved" heterogeneity or "fixed effect".

The factor $TIME_t$, being constant for the different cities in the US, will change *a*, the intercept term, to reflect the time-period t in y_{it}. The error e_{it} is often called the "idiosyncratic" error or time-varying error, as it captures unobserved factors that change over time and affect y_{it}.

A simple fixed effects model for city crime rates for 1982 and 1987 is

$$crmrte_{it} = a + OBJECT_i + TIME_t + b_1 unem_{it} + e_{it}$$

Results of this ANOVA plus covariate modelling are as follows:

Source	Type III SSQ	d. f.	MSQ	F
Corrected Model	72201	47	1536	7.64
Intercept	27140	1	27140	135.02
city	71211	45	1582	7.87
year	2157	1	2157	10.73
unem	1283	1	1283	6.38
Error	8845	44	201	
Total	1015653	92		
Corrected Total	81046	91		

The R Square is 0.89, the adjusted R Square is 0.77. The estimated coefficients are:

Parameter	B	Std.	t	Sig.	95%	95%
Intercept	63.614	12.045	5.281	0.000	39.338	87.890
[city=1]	3.277	14.517	0.226	0.822	-25.979	32.533
[city=2]	20.569	14.382	1.430	0.160	-8.416	49.554
[city=3]	7.966	14.292	0.557	0.580	-20.838	36.769
. . .	. . .	. . .	. . .	. . .	. . .	. . .
[city=45]	-14.476	14.526	-0.997	0.324	-43.751	14.799
[city=46]	0.000 a	.	.	.	.	.
[year=82]	-15.402	4.702	-3.276	0.002	-24.879	-5.926
[year=87]	0.000 a	.	.	.	.	.
unem	2.218	0.878	2.527	**0.015**	0.449	3.987

a. This parameter is set to zero because it is redundant.

Now that unobserved, city-specific variability has been properly accounted for, it turns out that unemployment rate is significant and positively related to the crime rate. The use of fixed effects is indeed a powerful method to face omitted variables bias when more than one period of data is available.

Note the degrees of freedom engaged by the fixed effects of cities: 45, which means that 44 degrees of freedom only are available to play the part assigned to randomness. This is a drawback of the method.

If we had estimate $crmrte_{it}$ by pooling the two periods together, there would be 46 cities and two years for each city or 92 total observations. We might then use a dummy to denote the 1987 cases. The result would be $crmrte = 93.4 + 7.94\ d_{87} + 0.43\ unem$ with a very small R Squared of 0.012. The coefficient on *unem*, though positive, is far from significant. Using pooled OLS on the two years has

not substantially changed anything from using a single cross section. This is not surprising since using pooled OLS does not solve the omitted variables problem.

-/-

In most applications, the main reason for collecting a panel containing more than one period of data is to allow for the unobserved effect, $OBJECT_i$ to be correlated with the predictors. For example, in the crime equation, we want to allow the unmeasured city factors in $OBJECT_i$ that affect the crime rate, also to be correlated with unemployment rate. It turns out that this is simple to allow: because $OBJECT_i$ is constant over time, we can difference the data across the two years, getting rid of $OBJECT_i$, ending up with a new intercept term which is the difference between the two $TIME_t$ effects:

$$\Delta\, crmrte_{it} = \Delta\, TIME_t + B_1 \Delta\, unem_{it} + u_{it}$$

This is the "FD" equation, which is estimated by OLS and may account for unobserved city effects as well. Results look like this:

```
             Estimate  Std. Error t value Pr(>|t|)
Intercept    15.4022     4.7021   3.276  0.00206 **
cunem         2.2180     0.8779   2.527  0.01519 *
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Residual standard error: 20.05 on 44 degrees of freedom
46 missing observations deleted.
Multiple R-squared: 0.1267, Adjusted R-squared: 0.1069
F-statistic: 6.384 on 1 and 44 DF, p-value: 0.01519
```

Again, the changes in unemployment rate significantly explain the crime rate. Actually, the coefficient is the same here and in the fixed effects models, and the intercept term is the difference between the two time-effects in the fixed effects model.

13.3 Instrumental variables

When faced with the prospect of omitted variables bias or unobserved heterogeneity in regressions, we have so far discussed three possible solutions:

(1) we can ignore the problem and suffer the consequences of having biased and inconsistent estimators,

(2) we can try to use a suitable "proxy" variable for the unobserved variable, or

(3) we can assume that the omitted variable does not change over time and use fixed effects or differencing to account for the unobserved variability.

The first solution can be satisfactory if the estimates follow the direction of the biases for the key parameters. For example, if we can say that the estimator of a positive parameter, say, the effect of job training on subsequent wages, is biased toward zero and we have found a statistically significant positive estimate, we have still learned that job training has a positive effect on wages. Unfortunately, the opposite case, where our estimates may be too large in magnitude, often occurs, which makes it very difficult for us to draw any useful conclusions. Proxy variables can also produce satisfying results, but it is not always possible to find a good proxy. Differencing and fixed effects are effective but are expensive in terms of degrees of freedom engaged.

Another approach is based on the idea of leaving the unobserved variable in the error term, but rather than estimating the model by OLS directly, use an estimation method that recognizes the presence of the omitted variable. This is what the method of "instrumental variables" (IV) does.

Given a regression with a predicted variable *y* and some predictors, of which *x* is endogenous, an "instrument" of *x* is any variable *z* that satisfies two requirements:

(1) *z* is uncorrelated with residuals of the regression, that is, *z* is uncorrelated with the unexplained portion of *y*

(2) *z* is correlated with *x*, the endogenous, missing predictor in the regression

The requirement that the instrument *z* satisfies (1) is summarized by saying "*z* is exogenous in the regression" and so we often refer to (1) as "instrument exogeneity" requirement.

In the context of omitted variables, instrument exogeneity means that *z* should have no partial effect on the predicted variable *y* (after predictors and omitted variables have been controlled for), and *z* should be correlated with the omitted variables.

Requirement (2) means that *z* must be related, either positively or negatively, to *x*, the regression predictor that we cannot observe directly. This condition is sometimes referred to as "instrument relevance".

There is an important difference between the above requirements for an instrumental variable. Because (1) involves the covariance between *z* and the unobserved error, we cannot generally hope to test this assumption: in the vast majority of cases, we must maintain it by appealing to economic behavior or introspection. Of course, if we had a good proxy for an important element of residuals, we might just add the proxy as an explanatory variable and estimate the expanded equation by ordinary least squares.

By contrast, the condition that the instrument *z* be correlated with the endogenous predictor *x* can be tested, given a random sample from the population. The easiest way to do this is to estimate regressions between the predictor and its instrument.

When relevant and exogenous instruments are available, then we can use algorithms to estimate regression parameters using other variables as instruments. In this way, instruments are able to "mirror" the endogenous variability and "send" such mirrored variability into the regression where it is missing. However, it is worth noting that

- The use of instruments can lead to high standard errors in coefficients, especially when instruments are "weak", that is, when the correlation between instruments and the predictors that they follow is weak.
- The use of instruments amplifies the sensitivity of models to multicollinearity.

As an example, consider the usual regression explaining wages:

$$Log(wage) = a + b_1 educ + b_2 abil + e$$

Here, *educ* is "endogenous": it leaves much variability of *wage* unexplained. Therefore, we call "*abil*" to this unexplained variability and include it in the regression but, as said, we cannot observe *abil* directly. Previously, we used *QI* as proxy for *abil*, but we know now that we can explore another possibility: find a variable which is, at once, correlated to *educ*, and uncorrelated to the unexplained variability in *y*. Such variable would be an instrument for *educ*.

Instruments use two stages to solve the problem of unobserved variability:

1. First, instrument *z* explains the endogenous predictor *educ*:
$$educ = A + B_1 z + u$$
2. Then, the b_1 in the *log (wage)* regression is written as a function of the *A* and B_1 above. IV algorithms find the new b_1 as a function of *A* and B_1. In other words, the $educ$ fitted in (1) is used as a proxy for $educ$.

We are going to change notation to incorporate these two stages more easily. We start with the usual OLS where we want to predict y_1 using z_1, y_2 as predictors:

$$y_1 = a + b_1 y_2 + b_2 z_1 + e$$

y_1 and y_2 are endogenous (y_1 is the former *y*) and z_2 is exogenous.

Now we include an instrument for y_2, the exogenous z_2, and, as a first stage, we explain the endogenous y_2 using the exogenous variables available as predictors:

$$y_2 = A + B_1 z_1 + B_2 z_2 + u$$

This regression on y_2 is called the "reduced form equation".

> The reduced form is estimated via OLS because all predictors are exogenous.
>
> The sole requirement is that B_2 be different from zero.
>
> And, if so, it is possible to estimate the b_i from the B_i.

This same reasoning can be extended to more than one predictor, either endogenous or exogenous so that a regression containing endogenous variables can be estimated using two OLS regressions in two stages:

1. In the first stage, we run the multi-instrument equivalent to the reduced form as above, using one regression per endogenous variable to obtain the fitted values of the endogenous variables where correlation with the residuals is not present.
2. Then we let these fitted values explain the predictor of interest, y_1 via a second stage regression where the $\widehat{y_2}$ (the y_2 predicted by the first stage) are used instead of y_2, that is, $y_1 = a + b_1 \widehat{y_2} + b_2 z_1 + e$

This is the 2SLS ("two stages least squares") method.

When applying 2SLS, the rules to apply concerning instruments are:

- One instrumental variable is required per endogenous predictor.
- If 2 predictors are endogenous and we use 12 more exogenous variables to control for other effects, then we only need to find 2 instruments.
- The 12 control variables are used both as predictors and as instruments. Indeed, when instruments coincide with predictors, 2SLS and OLS give the same results.

In fact, in an 2SLS algorithm, a predictor can always be used as an instrument of itself because, when instruments are the predictors themselves, the 2SLS algorithm calculates the OLS solution. It does not matter if control variables also used as instruments are irrelevant as instruments or not.

Using the CARD data file on wages *vs* education, we explain the logarithm of wages via education plus other 14 control variables (Wooldridge example 15.4, p. 526). As an instrument of *educ*, we use "*nearc4*" a dummy denoting proximity to a college.

Note the two stages:

- Stage 1, the model explains education:
  $$\boldsymbol{educ} = a + b_1 nearc4 + b_2 exper + b_3 exper^2 + b_4 black + \cdots$$
  and the fitted education, $\widehat{\boldsymbol{educ}}$, is kept.
- Stage 2, the model uses explained education to predict wage:
  $$\log(wage) = a + B_1 \widehat{\boldsymbol{educ}} + B_2 exper + B_3 exper^2 + B_4 black + \cdots$$

Obviously, in the second stage OLS, $nearc4$ (the instrument) is not included.

Besides $nearc4$, the other exogenous variables are also used as instruments. These variables are present in the model in order to control for other effects.

Instead of performing the two stages separately, we can use the algorithms available to perform 2SLS and thus avoiding having to explicitly write the reduced form equation. It is sufficient to indicate which variables are endogenous. This is the description of the command to use, first in an OLS and then in the 2SLS algorithm:

OLS regression of log (*wage*) on ***educ**, exper, expersq, black, smsa, south, smsa66, reg662, reg663, reg664, reg665, reg666, reg667, reg668, reg669*

...where *educ* is endogenous, all the others are exogenous.

IV regression of log (wage) on ***educ**, exper, expersq, black, smsa, south, smsa66, reg662, reg663, reg664, reg665, reg666, reg667, reg668, reg669* **| instruments *nearc4*** *exper, expersq, black, smsa, south, smsa66, reg662, reg663, reg664, reg665, reg666, reg667, reg668, reg669*

The sign "|" separates the list of predictors from the list of instruments. Remember, nothing changes if a variable is the instrument of itself. Results are:

```
OLS:            Estimate  Std. Error t value  Pr(>|t|)
Intercept     4.6208067  0.0742327  62.248  < 2e-16 ***
educ          0.0746933  0.0034983  21.351  < 2e-16 ***
exper         0.0848320  0.0066242  12.806  < 2e-16 ***
I(exper^2)   -0.0022870  0.0003166  -7.223 6.41e-13 ***
black        -0.1990123  0.0182483 -10.906  < 2e-16 ***
smsa          0.1363845  0.0201005   6.785 1.39e-11 ***
south        -0.1479550  0.0259799  -5.695 1.35e-08 ***
smsa66        0.0262417  0.0194477   1.349  0.17733
reg662        0.0963672  0.0358979   2.684  0.00730 **
reg663        0.1445400  0.0351244   4.115 3.97e-05 ***
reg664        0.0550756  0.0416573   1.322  0.18623
reg665        0.1280248  0.0418395   3.060  0.00223 **
reg666        0.1405174  0.0452469   3.106  0.00192 **
reg667        0.1179810  0.0448025   2.633  0.00850 **
reg668       -0.0564361  0.0512579  -1.101  0.27098
reg669        0.1185698  0.0388301   3.054  0.00228 **
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Residual standard error: 0.3723 on 2994 degrees of freedom
Multiple R-squared:  0.2998,     Adjusted R-squared:  0.2963
F-statistic: 85.48 on 15 and 2994 DF, p-value: < 2.2e-16

2SLS:           Estimate  Std. Error t value Pr(>|t|)
Intercept     3.6661511  0.9248295   3.964 7.54e-05 ***
educ          0.1315038  0.0549637   2.393  0.01679 *
exper         0.1082711  0.0236586   4.576 4.92e-06 ***
I(exper^2)   -0.0023349  0.0003335  -7.001 3.12e-12 ***
black        -0.1467758  0.0538999  -2.723  0.00650 **
smsa          0.1118083  0.0316620   3.531  0.00042 ***
south        -0.1446715  0.0272846  -5.302 1.23e-07 ***
smsa66        0.0185311  0.0216086   0.858  0.39119
reg662        0.1007678  0.0376857   2.674  0.00754 **
reg663        0.1482588  0.0368141   4.027 5.78e-05 ***
reg664        0.0498971  0.0437398   1.141  0.25406
reg665        0.1462719  0.0470639   3.108  0.00190 **
reg666        0.1629029  0.0519096   3.138  0.00172 **
```

```
  reg667       0.1345722  0.0494023   2.724  0.00649 **
  reg668      -0.0830770  0.0593314  -1.400  0.16155
  reg669       0.1078142  0.0418137   2.578  0.00997 **
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Residual standard error: 0.3883 on 2994 degrees of freedom
Multiple R-Squared: 0.2382,      Adjusted R-squared: 0.2343
Wald test: 51.01 on 15 and 2994 DF, p-value: < 2.2e-16
```

The IV estimate of the return to education is almost twice as large as the OLS estimate, but the standard error of the IV estimate is over 18 times larger than the OLS standard error. The 95% confidence interval for the IV estimate is between .024 and .239, which is a very wide range.

The presence of big confidence intervals is the price we pay to get a consistent estimator of the return to education when we think that education is endogenous .

13.4 Errors in variables, the over identification of restrictions

OLS presupposes that the predictors are non-random, exact measurements. However, it often happens that this is not the case: predictors are measurements subject to errors or they may be random. In the presence of randomness in predictors, OLS becomes biased and inconsistent. This is the problem of "errors in variables".

It can be shown that IV methods allow the use of OLS estimation in the case of errors in variables. We can use either repeated observations or other exogenous variables as instruments.

As an example, if we think that *educ* in the MROZ dataset is not accurately measured, we can use *motheduc* and *fatheduc* as instruments for *educ*. As a result, so long as *motheduc* and *fatheduc* are not correlated with the measurement error affecting *educ*, then the IV estimator for *educ* will not suffer from measurement error.

IV methods can also be adopted when using "test scores" to control for unobserved characteristics. We showed that, under certain assumptions, proxy variables can be used to solve the omitted variables problem. IQ, say, was used as proxy for unobserved ability in a regression explaining wages. This simply entails adding IQ to the model and performing an OLS regression.

There is an alternative that works when IQ does not fully satisfy the proxy variable assumptions. If we have two test scores that are indicators of ability, since it is ability that affects the variable to be predicted, wages, we can assume that the two test scores are uncorrelated with the error term of the regression predicting wages. If we write *abil* in terms of the first test score and plug the result into the equation explaining the second test score in terms of ability, it can be shown that one of the tests is an IV for the other.

We can use the WAGE2 dataset to exemplify the preceding procedure, where IQ plays the role of the first test score and KWW (knowledge of the world of work) is the second test score. The explanatory variables are *educ, exper, tenure, married, south, urban*, and *black*. Rather than adding IQ and doing OLS, we add IQ and use KWW as instrument of KWW. When explaining log (wage), the coefficient on *educ* now is 0.025.

-/-

Some types of models introduce more instruments than the endogenous variables that they are supposed to mirror. The use of multiple instruments raises the problem of "over identifying restrictions".

Here, the term “restrictions” refers to the parameters of the second-stage model. Indeed, a model parameter is a restriction in the sense that it engages one degree of freedom, thus restricting the free or unexplained variability.

Over identifying restrictions can be thought of as over fitting parameters, and leads to the same consequences. We have more instruments than required and, therefore, there are more reduced form coefficients than those required to model the second-stage parameters.

Models with over identified restrictions have too many degrees of freedom. They “over fit” the data, yielding significance where (in the population) there is none, while being bad at generalizing results to other datasets.

There are tests available (for example, the Sargan test) to check whether a regression over identifies restrictions or not.

Finally, it is worth emphasizing the close connection between 2SLS and systems of two equations. For almost all purposes, these two modelling situations are the same. When introducing systems of equations, we show a few more examples of 2SLS.

Chapter 14 Inference using panels

A "panel" is a time-series of cross-sectional data. We mentioned the case of repeated measures, which is a panel where the time-series is small.

Suppose that we observe the capital ratio of a bank plus other bank ratios and economic factors in s specific time-period, say, year 2007.

> This is a cross-section. A regression can explain a capital ratio of that period.
>
> Now, if we have this same cross-section data, not for just one period but for several subsequent periods, say, years 2007, 2008, ..., 2021, we have a panel.

For example,

> 5 ratios from 120 different bank reports in 2009 – is a cross-section.
>
> 5 ratios from 120 different bank reports in 2007, 2008, ..., 2021 – is a panel.

A cross-section is defined by one unique index, which identifies each object. A panel is defined by two indexes: the object, and the time period.

14.1 Types of panel design

Models can be built from panel data and there are many designs available. We consider two extreme cases where we ignore the panel nature of our data:

- First, we may decide to build separate cross-section models for each period.
  - If we predict the capital ratio of banks from other ratios, we may run a separate cross-section regression for each year, thus ending up with as many regressions as the number of years available in the panel.
  - Similarly, we might use time-series to build ARIMA or other models in time. We will end up with as many time-models as banks in the panel.
- A second way to ignore the panel nature of the dataset consists of pooling all the years together to build one sole regression explaining the capital ratio of banks from other bank ratios. For each bank ratio, there will be 11 observations in the sample, one for each year. Trends and autocorrelation would lead to biases.

In between these two extremes, we have designs that take advantage of the panel nature of the dataset.

Useful panels result from applying restrictions to an over specified case where there are as many variables as observations (observations act as variables), that is, for $i = 1, \dots N$ objects and $t = 1, \dots M$ time-periods,

$$y_{ij} = a_{it} + \Sigma b_{it} x_{it} + \varepsilon_{it}$$

where y_{it} and x_{it} are respectively the predicted and the predicting observations, a_{it} and b_{it} are model parameters and ε_{it} are zero-mean errors. For 100 objects observed during 5 time-periods, the model above would have 500 slope coefficients b_{it} plus 500-1 a_{it} intercept terms. This is the meaning of the model being over specified.

As mentioned, the strongest restriction that can be applied to the above case consists of pooling all time-periods and objects together. Given $k = 1, \dots K$ predictors, this "pooled" regression model is

$$y_{ij} = a + \Sigma b_k x_{it}^k + \varepsilon_{it}$$

where slopes b_k are the same across all objects and time-periods. Pooled models ignore the panel (repeated measures) character of observations and can be biased. "Poolability" tests and "Lagrange multiplier" (LM) tests are used to ascertain whether or not model parameters can be validly pooled together. If not, other, designs must be put in place.

When time-periods can be pooled together but the objects being modelled cannot, a fixed effects model

$$y_{ij} = a_t + \Sigma b_k x_{it}^k + \mu_i + \varepsilon_{it}$$

accounts for non-homogeneity of objects. The fixed effects μ_i are dummy variables that are constant across objects. Age, sex, race and other variables do not change or change at a constant rate over time. They are "fixed effects": any change that they cause to an object is the same over time. Therefore, the μ_i in the fixed effects model are object-dependent but time-independent deviation from the mean of the predicted variable, being known as the "within" transformation.

If we want to predict, say, a bank's capital ratio from other ratios, we may assume that some of the endogenous variability that we cannot observe can be ignored because it is created by bank features that do not change with time. In a cross-section regression, we would have to account for the unobserved effect of individual banks by using proxies or instrumental variables. Using a fixed effects design, so long as unobserved bank effects are assumed to be the same for different years, they will be accounted for.

If we have N banks in our dataset, the within transformation is numerically identical to including N – 1 dummies in the regression: fixed effects account for unobserved bank heterogeneity by means of constant terms μ_i, as many as banks minus 1.

Slopes b_k, one for each predictor, are the same across objects. The effect of μ_i and b_k is that of producing a set of parallel regression lines, one for each object's time-histories. All time-invariant variability is accounted for by the μ_i so that only overall changes in *y* are captured by the b_k.

Since fixed-effects models account for any variability that is constant over time, model parameters are not biased by omitted variables that are constant over time. However, this design is expensive in terms of degrees of freedom engaged. Therefore, unless the dataset has many objects and time-units, the balance between observations and degrees of freedom engaged by the model may be insufficient to draw useful inferences.

The fixed effects design cannot include time-invariant predictors in the model because these would be removed together with other unobserved variability. Besides, model consistency requires strict exogeneity of predictors, which is not satisfied when the estimated model includes, say, the lagged predicted variable (a dynamic model).

14.2 What is the appropriate panel?

As a first example, the dataset PRISON provides data on prison population and crime rates in the form of a panel of annual, aggregate data for each of the 51 US states, for each year from 1980 to 1993. In PRISON, therefore, 51 objects (states) are observed for 14 periods (years) in a total of 714 cases (Woodridge p. 573, example 16.8).

Our goal is to estimate the causal effect of prison population growth on crime rates at the US state level and we can control for several other variables available, namely

1. the log of *pris*, the prison population per 100,000 residents

2. the log of *incpc*, the per capita income for that state and year, nominal
3. the log of *polpc*, the number of policemen per 100,000 residents
4. the proportion of unemployed residents, *unem*
5. the proportion of residents that are *black*
6. the proportion of residents that live in metropolitan areas (cities)
7. 4 proportions showing the age distribution of the population of that state / year.

The formulation

$$\log(criv) = a + b_1 \log(pris) + b_2 \log(incpc) + b_3 \log(polpc) + b_4 unem + b_5 black + b_6 metro + b_7 ag0 + b_8 ag15 + b_9 ag18 + b_{10}\, ag25$$

explains the log of *criv*, the rate of violent crime in an US state during the year, in terms of the log of *pris*, the rate of prison population per resident, plus control variables.

This formulation can be used as a starting point of our discussion.

First, we should dismiss the use of one unique OLS regression because residuals would be correlated. Indeed, the time-series would bring into estimation the 51 groups of 14 trendy observations. In such a pooled OLS, the coefficient of the log of *pris*, when explaining the log of *criv*, is positive and highly significant.

Also, the simultaneity that exists between growth in crime and prison population would make OLS estimation generally inconsistent. Indeed, given that criminals, when in prison, raise their crime prospects, it is not possible to say whether it is crime that drives prison population or the other way round.

Finally, the above formulation is about absolute values, not changes.

When the fixed effects design is used, results are:

```
             Estimate Std. Error t-value  Pr(>|t|)
log(pris)  -0.165880   0.039508 -4.1987 3.065e-05 ***
log(incpc) -0.260020   0.163929 -1.5862  0.113193
log(polpc)  0.216361   0.069076  3.1322  0.001814 **
unem       -1.988183   0.459440 -4.3274 1.750e-05 ***
black      -1.818856   1.219916 -1.4910  0.136462
metro       1.610980   0.366746  4.3926 1.310e-05 ***
ag0_14      2.483919   1.401205  1.7727  0.076754 .
ag15_17     5.309049   3.119787  1.7017  0.089291 .
ag18_24     3.660076   1.665141  2.1981  0.028302 *
ag25_34     9.201941   1.181240  7.7901 2.712e-14 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
Total Sum of Squares: 9.7391, Residual Sum of Squares: 7.8092
R-Squared: 0.19816, Adj. R-Squared: 0.1067
F-statistic: 15.8166 on 10 and 640 DF, p-value: < 2.22e-16
```

where the coefficient of the log of *pris* now is negative and significant, which suggests a negative causal relationship between imprisonment and crime rate.

-/-

Another useful panel is the “random effects” design, known as “variance-component” model. In statistics, a random effect is an effect that is random, not exactly predictable. If we can observe all the 7 different bacteria that cause a given illness, then the effect of each one of those bacteria can be measured exactly. However, if we are limited to sampling at random to find 7 among so many different bacteria that may cause the same illness, then the effect of each of such bacteria is random.

When it is assumed that the set of x_{it} are independent from the μ_i dummies mentioned above, the panel is random effects. In random-effects models, some of the parameters (effects) that define systematic components of the model exhibit a random variation while the estimated μ_i are no longer deterministic (fixed). Rather, they are supposed to be drawn at random from a probability distribution.

While fixed-effects account for (remove) any time-invariant object-specific variability of *y,* random effects will evidence (show) such variability. Therefore, random effects designs require all variables with explanatory power to be explicitly specified. For this reason, random effects cannot be considered in any study where endogenous variability may be present. Fixed and random effects may be compared via the Hausman test, the null hypothesis being that the random effects μ_i are uncorrelated with the predictors. When the null hypothesis is rejected, random-effects estimation is not tenable.

The random effects design is more restrictive than the fixed effects. With fixed effects models we do not estimate the effects of variables whose values do not change across time. Instead, we get rid of them (control for them). In random effects designs we do not control for time-invariant variability. Therefore, we need to estimate it, and the model may be biased because we are no longer controlling all the time-invariant effects.

Contrary to fixed effects, under a random effects design there may be bias in the coefficient estimates if predictors are correlated with the unobserved objects’ effects. In short, the random effects assumption is that the unobserved variability of objects is uncorrelated with predictors. The fixed effect assumption is that the same unobserved variability of objects is correlated with predictors.

If the random effects assumption holds, estimation is more efficient than in fixed effects, less degrees of freedom are engaged in modelling. However, if this assumption does not hold, the random effects estimator is not consistent.

When the random effects design is used in the PRISON panel, results are:

```
Effects:          var std.dev share

idiosyncratic 0.01520 0.12328 0.181

individual    0.06875 0.26220 0.819

theta: 0.8753

              Estimate Std. Error z-value  Pr(>|z|)

(Intercept) -10.561950   1.173333 -9.0017 < 2.2e-16 ***

log(pris)    -0.020923   0.039983 -0.5233 0.6007625

log(incpc)    0.643825   0.092623  6.9511 3.625e-12 ***

log(polpc)    0.488877   0.071695  6.8189 9.176e-12 ***

unem         -1.013079   0.377332 -2.6849 0.0072562 **

black         2.199123   0.354121  6.2101 5.296e-10 ***
```

```
metro         1.362132   0.171334  7.9501 1.863e-15 ***
ag0_14        6.899626   1.079641  6.3907 1.652e-10 ***
ag15_17       7.283966   2.670257  2.7278 0.0063755 **
ag18_24       3.998789   1.115419  3.5850 0.0003371 ***
ag25_34      -0.933864   0.801790 -1.1647 0.2441311
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Total Sum of Squares: 21.644, Residual Sum of Squares: 11.754
R-Squared: 0.45693, Adj. R-Squared: 0.44921
Chisq: 591.499 on 10 DF, p-value: < 2.22e-16
```

The coefficient of the log of *pris* is negative but non-significant. The idiosyncratic share of variability, that is, the state effect, is around 20%. The Hausman test is significant, which shows that random effects are not consistent here.

-/-

When the restriction of well-behaved errors is not tenable and ε_{it} is allowed to be heteroskedastic and serially correlated over time, OLS is inconsistent. For instance, if μ_i are correlated with predictors, then fixed-effects estimates would be valid but the use of feasible generalized least squares (FGLS) is recommended instead of OLS.

-/-

Another way of allowing for non-homogeneous variability in banks consists of first differencing y_{ij}, so that fixed-effects μ_i will cancel-out, and then using OLS estimation. This is the "first difference" (FD) estimation, useful when errors are persistent over time, but changes are serially uncorrelated. First differencing is especially adequate for the PRISON panel, since it will model changes directly. Results are:

```
Observations used in estimation: 663
             Estimate Std. Error t-value  Pr(>|t|)
(Intercept) -0.002215   0.015065 -0.1470 0.8831587
log(pris)   -0.171494   0.050870 -3.3712 0.0007925 ***
log(incpc)   0.776708   0.168617  4.6064 4.929e-06 ***
log(polpc)   0.083917   0.061323  1.3684 0.1716432
unem         0.133041   0.305682  0.4352 0.6635413
black       -1.170415   3.732594 -0.3136 0.7539510
metro       -0.166856   1.061656 -0.1572 0.8751626
ag0_14       3.748051   2.111715  1.7749 0.0763834 .
ag15_17     10.530331   3.041789  3.4619 0.0005714 ***
ag18_24      0.734674   2.217401  0.3313 0.7405077
ag25_34     -3.247766   1.727748 -1.8798 0.0605853 .
Total Sum of Squares: 5.0884, Residual Sum of Squares: 4.6678
R-Squared: 0.082658, Adj. R-Squared: 0.068588
F-statistic: 5.87489 on 10 and 652 DF, p-value: 1.5892e-08
```

The coefficient associated with changes in the log of *pris* is negative and significant.

-/-

A dynamic panel design introduces the lagged dependent variable in the fixed-effects:

$$y_{ij} = a_i + b_0 y_{i(t-1)} + \Sigma b_k x_{it}^k + \mu_i + \varepsilon_{it}$$

Dynamic designs account for persistence in *y* and measure it via b_0. For example, when banks adjust their capital ratios to comply with a regulatory target, b_0 will estimate the speed of such adjustment. But when, as said, the time-period M is small, the presence of $y_{i(t-1)}$ among the set of predictors raises estimation problems.

The Generalized Method of Moments (GMM) is often used in dynamic models where the number of periods is small. GMM at once addresses three relevant issues:

- the presence of unobserved effects, which are removed by differencing;
- the need to introduce lagged *y* among explanatory variable to account for the dynamic nature of *y*;
- the endogeneity of predictors, which is accounted for by using lagged variables as instruments.

Our second example is CAR2, a dataset of bank ratios and economic data for 253 retail banks from East-Asia jurisdictions for each of the 11 years from 2004 to 2014. We want to find out how persistent the Tier 1 ratio is but here, too, we face simultaneity issues because this ratio is subject to tight regulation and bank managers try to keep the Tier 1 ratio in line with regulatory targets while, at the same time, the risk-taking policies of the same managers may make the ratio diverge from targets. Moreover, we have only 11 years of data, which is not much. The formulation

$$Log\ (T1R)_t = a + b_0 \log(T1R)_{t-1} + b_1 NIM + b_2 LA + b_3 \log(TA)$$

explains the log of the Tier 1 capital ratio *T1R* of banks in terms of the lagged log of *T1R* and we control for *NIM*, the net interest margin of banks, *LA*, the loans to assets ratio, and a proxy for bank size, the log of assets,. Banks are affected by unobserved effects, so we rely on panel designs that account for these. GMM results are:

```
Two ways effects One-step model Difference GMM
Balanced Panel: n = 253, T = 11, N = 2783, number of Observations Used: 2277
                 Estimate  Std. Error z-value  Pr(>|z|)
lag(log(T1R))  0.52462006  0.04952041 10.5940 < 2.2e-16 ***
NIM            0.00084166  0.00056192  1.4978 0.1341798
LA            -0.00952035  0.00202386 -4.7041 2.550e-06 ***
log(TA)       -0.24189032  0.04554342 -5.3112 1.089e-07 ***
2006           0.05082674  0.02012903  2.5250 0.0115683 *
. . .
2013           0.26878629  0.05210654  5.1584 2.491e-07 ***
2014           0.27657782  0.05033467  5.4948 3.912e-08 ***
Sargan test: chisq(16) = 26.4 (p-value = 0.05)
Autocorrelation test (1): normal = -6.55 (p-value < 0.000)
Autocorrelation test (2): normal = 0.412 (p-value = 0.68)
Wald test for coefficients: chisq (4) = 525.91 (p-value < 0.000)
Wald test for time dummies: chisq (9) = 91.67 (p-value < 0.000)
```

In East Asia, therefore, persistence of T1R is around 0.52. In the coming section we show results of tests leading to the decision of using GMM. Suffice to say that, when other designs are used, persistence varies from 0.87 (pooled, random effects) to 0.53 (differencing and fixed effects). GMM results are akin to differencing and fixed effects.

We have mentioned the following types of panel designs:

- Pooled,
- Fixed Effects,
- Random Effects,

- First Difference, and
- Dynamic

Which of the above would be the appropriate choice in a specific case?

14.3 Panel designs

Prior to the modelling, it is necessary to ascertain which of the available panel designs and estimation methods is adequate to the problem at hand.

Basically, there are two kinds of information in panel data:

The cross-sectional information reflected in the differences "between" objects.

The time-series information reflected in the changes "within" objects.

In a panel, both the variable to be predicted and the predictors, can potentially vary over both time and by object. Therefore, we label the variability in a panel as being

"Within", if the variability is caused by the same objects over time, or

"Between" if the variability is caused by different objects at the same time. A between effects use only the cross-sectional information and asks: "What is the expected difference in a capital ratio between two banks that differ by 1 in x_{it}^{k}?"

This labelling is similar to the one we use when describing the possible t-test designs, and in the one-way ANOVA. The fixed effects design in panel models is the same as the "repeated measures" or "within" effects design in t-tests and ANOVA. In fixed effects, the same objects are observed several times, in different periods in time, and we are interested in finding out how changes from one period to the other explain or affect the predicted variable. Cross-sectional regressions, by contrast, model "between" effects.

The 3 basic types of panel models are the consequence of the way the two main effects, object and time, are treated:

- Pooled panel modelling, when the effects of objects and time periods are not accounted for prior to estimation. The effects of time and object are ignored.
- Fixed effects, where the effects of objects are viewed as non-random and are accounted for prior to estimation. Fixed effects use the time-series information and asks, "What is the expected change in a capital ratio if x_{it}^{k} increases by 1?"
- Random effects, where the effects of objects, time-periods, and interactions are viewed as random and included in estimation. These are similar to "variance component" ANOVA modelling.

Statistically, a fixed effects model is a reasonable thing to do with panel data (it always gives consistent results), but it may not be the most efficient description. A random effects model will give lower standard errors as it is a more efficient estimator.

Moreover, in general panel designs can also be

- static or pulled, when time is not explicitly modelled, or
- not static. If not static, panels may include
    - a "first difference" (FD), where first or higher-order differencing is applied to the predicted variable prior to estimation, or
    - "dynamic" or AR (1), where the lag of the predicted variable is included as a predictor
    - or both FD and dynamic.

The estimation method used is one of the usual methods: OLS, WLS (weighted least squares), GLS (generalized least squares) FGLS (feasible generalized least squares), or GMM (generalized method of moments). In practice, OLS, FGLS and GMM are used.

Here, as in everything else, before selecting the panel we should have it clearly in our mind what is the question we want to answer by using the panel.

> If what we want is, say, to describe the basic factors underlying bank's capital ratios, and if we have enough data and time periods, then we should account for as many randomness as possible. GMM models with time and bank effects accounted for would be the choice.
>
> If, instead, what we want to find out is whether a given feature has changed over time, then a "within" panel (fixed effects) would be better.
>
> Finally, if we want to examine bank features or if our data is limited in the number of observations but not in the variety of predictors available, we may select random effects.

In most cases, however, the selection of the most appropriate panel is carried out for reasons that have little to do with the above. The model is chosen simply in order to be unbiased and consistent, given the specific characteristics and limitations of the data. Econometric reasoning prevails over other considerations.

Typically, therefore, tests are conducted, beginning with the most basic model, pooling, and proceeding till the most sophisticated, GMM. More sophisticated models should not be used unless previous designs are discarded.

For example, to find out the adequate panel to use in the CAR2 panel, the following tests are performed:

1. First, Wooldridge test of unobserved effects is run, rejecting the null hypothesis that there are no unobserved effects ($z = 3.13$, p value < 0.01).
2. Lagrange multiplier (LM) tests (Honda, 1985) are run next, the null hypothesis being that year- and bank-specific effects are non-significant. Both hypotheses are rejected (z=10.67, p value < 0.000 for year-specific effects and z=14.02, p value < 0.000 for bank-specific effects).
3. Two-ways effects are also tested via another LM test (Gourieroux, Holly & Monfort, 1982). The null hypothesis is rejected (Chi-Square=310, p-value < 0.000). It is concluded that bank- and year-specific effects cannot be ignored.
4. The fixed-effects design with year- and unit-specific effects is tested next. The null hypothesis that errors are uncorrelated to explanatory variables is rejected by the Hausman, test (Chi-Square=1250, 32 d. f., p value < 0.000).
5. Since the test for serial correlation in fixed-effects errors is significant (Chi-Square=18, p-value < 0.000), the FD design (differencing) is tested next.

Due to limitations of the FGLS algorithm, the FD design is estimated via OLS.

6. The Wooldridge's first-difference test compares fixed-effects (serial correlation) to FD (no correlation). The Chi-Square is 7.09, p value = 0.008, thus showing serial correlation in differenced errors.

Given that FD and previous designs are non-satisfactory, GMM is adopted.

-/-

Estimation using panel designs require specialized software, for example, R or STATA or SAS or EVIEWS. However, SPSS can also handle most panel designs. See for a detailed example.

https://stats.idre.ucla.edu/spss/library/spss-librarypanel-data-analysis-using-gee/

Chapter 16 Limited dependent variables

Regression and panel estimation are inadequate to predict variables whose randomness is restricted in range. For example, the fact that a loan may default or not, will require a attribute with two nominal states only, default and not default, to be modelled. This, and similar types of predicted attributes are known as “limited dependent variables" (LDV).

Other limited dependent variables include proportions such as the percentage of pension plan members (which must assume values between zero and 100), college grading (which is between zero and 4 at most colleges) or the number of children in a family.

Most economic variables that we would like to explain are limited in some way, often since they must be positive or discrete or both. For example, market prices and nominal interest rates must be greater than zero. Not all such variables need special treatment but in specific cases the direct use of linear models to treat limited dependent variables has clear drawbacks.

Here, we introduce the “Logit” (logistic regression) and “Probit” modelling algorithms which surmount the shortcomings of dealing with binary predicted variables directly. In the following chapters we apply these models to specific limited range problems.

16.1 Binary response models

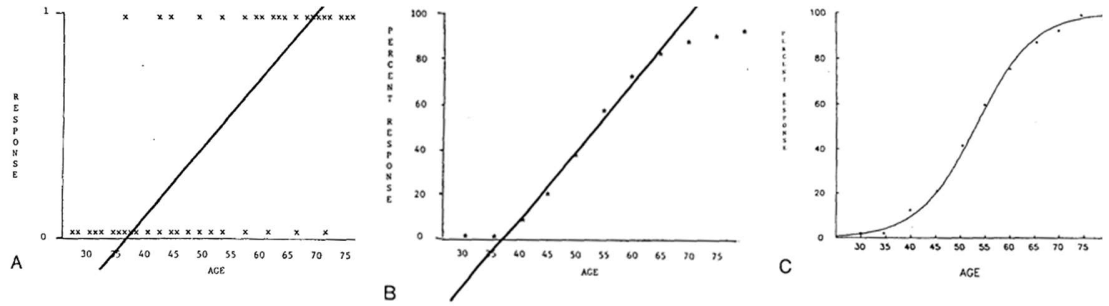
Logit modelling, that is, logistic regressions, overcome the awkwardness of having to predict binary attributes using a linear regression.

The two most important disadvantages of using a linear regression in this context are that the fitted probabilities can be less than zero or greater than one, while the partial effect of any explanatory variable (appearing in level form) is constant.

These limitations are surmounted by using “binary response” models able to deliver a response probability as illustrated in the figure “C” below. The two most popular types of response probabilities are the logistic function and the Normal probability function. The first leads to the Logistic regression algorithm and the second to the Probit one.

Figure “A” shows how a regression would look like when trying to fit a binary response of 0 or 1, for example 0 is “not affected” and 1 is “affected” by an illness, using age as predictor. It is clear that the elder are more affected than the younger.

Figure “B” shows the same regression trying to fit, not the binary response variable directly, but the proportions of affected cases over non-affected cases, formed from responses counted in fixed intervals. Youths show an almost zero proportion of the total, the older show much higher proportions.



Finally, figure “C” shows the result of fitting an appropriate probability function to the proportions calculated in “B”. Now, all awkwardness of the fit has vanished.

Since the model in “C” requires a non-linear fit, OLS cannot be used here and iterative estimation methods that maximize the likelihood of the fit are used instead. Recall that, besides OLS, other, often used estimation methods are WLS, GLS, FGLS, and GMM.

16.2 Maximum likelihood estimation

“Maximum likelihood” estimation (MLE) is a method to estimate the parameters of a model. It is an alternative to OLS, WLS and so on, when the relationship is non-linear or is intractable analytically. In maximum likelihood estimation, the parameters are chosen so to maximize the likelihood that the assumed model fits the observed data.

When studying trends, we have offered an example of data fit to a logistic function by means of a non-linear optimization algorithm. That fit basically shows how MLE works.

To implement maximum likelihood estimation, we follow these three steps:

> First, we choose an analytical model, also known as a “data generating process”, and we assume that such model is the generative mechanism of our data.
>
> Next, we derive the “likelihood function” for our data, given the assumed model.
>
> Once the likelihood function is derived, the finding of the maximum likelihood estimates is a simple optimization problem.

In the case of the Logistic regression, for instance,

- we first assume that the logistic function is the data generating process that governs the frequency (the likelihood) of zeros and ones observed in the data.
- Then we derive the likelihood function corresponding to the logistic curve and we apply our observations to that function, thus ending up with an equation with a few unknown parameters. These are the parameters we wish to discover.
- Finally, we find out, using optimization, what are the values of those parameters that maximize the likelihood function, that is, the likelihood of our fit.

Parameters that maximize the likelihood function are maximum likelihood estimates. The method has become a dominant means of statistical inference when OLS is not tenable or when probabilities themselves are the object of estimation, as in the case of Logit and Probit modelling.

In practice, it is often convenient to work with the natural logarithm of the likelihood function, called the “log-likelihood”. Since the logarithm is a monotonic function, the maximum of the log-likelihood occurs at the same value of the likelihood function.

Finding the set of parameters that maximize the likelihood function is asymptotically equivalent to finding the set of parameters that defines a probability distribution that has a minimal distance to the

probability distribution from which our data was generated. Even if the model that we use is misspecified, still the MLE will give us the "closest" distribution to the real distribution. Moreover, the MLE also asymptotically minimizes the “cross entropy”, that is, the informational uncertainty of the estimate.

In the case of Normal, identically independent errors, the ML method coincides with the OLS method. Importantly, the error in the logarithm of likelihood values for estimates from multiple independent observations is asymptotically Chi Square-distributed, which enables determination of confidence intervals for any estimate of the parameters.

Maximum likelihood estimation requires presuming a distribution for data. Likelihoods are similar to the probability density function. However,

> A probability density function is the probability of observing our data given the underlying distribution parameters. It assumes that the parameters are known.
>
> The likelihood function is the likelihood of parameter values occurring given the observed data. It assumes that the parameters are unknown.

In the case of a linear regression, we assume that the model residuals are identical and independently normally distributed $\epsilon = y - \hat{\beta}x \sim N(0, \sigma^2)$. Based on this assumption, the log-likelihood function for the unknown parameter vector, $\theta = \{\beta, \sigma^2\}$ conditional on the observed *n* cases is

$$\ln L\,(\theta|y, x) = -\frac{1}{2}\sum_{i=1}^{n}\left[\ln \sigma^2 + \ln(2\pi) + \frac{y - \hat{\beta}x}{\sigma^2}\right]$$

and maximum likelihood estimates are those that maximize this log-likelihood function.

MLE cannot be calculated analytically, and appropriate iterative methods are employed to discover the optimal set of parameters. Widely used among such methods are.

- Gradient descent
- Newton Raphson
- Fisher’s information score

When dealing with non-linear modelling we shall return to these iterative methods.

16.3 Logistic regression

The logistic function made to fit the curve in figure “C” is of the form

$$y = \frac{1}{1 + e^{-(b_1x_1 + b_2x_2 + \cdots + e)}}$$

so that the range of *y* is between 0 and 1. Observed *y* are random collections of zeros and ones, each of them associated with predictors $x_1, x_2, \ldots$ After successful modelling, the predicted $\hat{y}$ mimics the behavior of the probability associated with its specific set of predictors’ values $x_1, x_2, \ldots$ and can be seen as an expected probability of *y* assuming the value of 1 given $x_1, x_2, \ldots$.

Probit and Logit models give similar results as the two underlying probability curves are not much different from one another. However, two issues require more attention in the context of Probit models, namely non-normality and heteroscedasticity of errors.

Issues concerning endogenous predictors in linear models also arise here. Probit models can be made to handle unobserved effects. It is possible, for instance, to test and correct for endogenous explanatory variables using methods related to 2SLS as explained later.

-/-

An example ensues, where the probability of companies going bankrupt during the year following the publication of their accounting reports is estimated from a collection of financial ratios, the $x_1, x_2, \ldots$ above, taken from 55 bankrupt ($y = 1$) and 91 non-bankrupt ($y = 0$) same-industry same-year companies. The file is "TAFFLER".

To explain *y* the 4 financial ratios used as predictors in the logistic regression are:

cash flow from operating activities to total assets (CF_TTA),

revenues to total liabilities (S_TL),

short term liabilities to net capital employed (STL_NCE), and

working capital to total assets (WC_TTA).

The model output $\hat{y}$ is the probability of bankruptcy one year ahead.

The following coefficients are estimated:

	B	S.E.	Wald Chi-Sq.	df	Sig.	Exp(B)
CF_TTA	-35.656	11.479	9.649	1	.002	.000
S_TL	-2.735	1.135	5.803	1	.016	.065
STL_NCE	2.408	.805	8.938	1	.003	11.111
WC_TTA	-9.780	2.827	11.969	1	.001	.000
Constant	4.738	1.986	5.689	1	.017	114.180

This means that the fitted model is of the form

$$\hat{y} = \frac{1}{1 + e^{-Z}}$$

with

$$Z = 4.738 - 35.656\, CF\_TTA - 2.735\, S\_TL + 2.408\, STL\_NCE - 9.780\, WC\_TTA$$

That is, model parameters estimated by the logistic regression allow calculating Z. In turn, Z is the "log-odds" of a company going bankrupt in the year after presenting the report from which ratios are drawn. The log-odds Z is a different way of expressing probabilities. In Z, likelihoods are presented as a score, not as a proportion:

$$\log\left(\frac{\hat{y}}{1 - \hat{y}}\right) = Z$$

Expected probabilities $\hat{y}$ and log-odds Z have a monotonic, increasing relationship. Therefore, any increment in Z leads to an increment in $\hat{y}$ and coefficients of the logistic regression can be readily interpreted as in any linear model.

Moreover, Z spans the interval –∞ to +∞, being better behaved than probabilities, which span the interval 0 to 1. Z will range from large negative values for financially sound companies to large positive values in companies highly likely to go bankrupt.

The probability linked to bankruptcy, $\hat{y}$, as predicted by the model, is computed from Z. In general, this probability is made available by algorithms as an output of fitted model.

From the probability linked to bankruptcy, it is then possible to build a binary variable which tells whether a company is expected to go bankrupt or not. For instance, we may assign a value of zero or one if $\hat{y} < 0.5$ or $\hat{y} \geq 0.5$ respectively.

We end up with classification results that show the accuracy of the model in predicting the state of the attribute, bankrupt or non-bankrupt. In the present case,

		Predicted		
		Status		
	Observed	SOUND	FAIL	Percentage Correct
Status	SOUND	89	2	97.8
	FAIL	4	51	92.7
Overall Percentage				95.9

This classification table is known as the “confusion matrix”, being basically similar to the 2 by 2 tables that we studied in relation to the comparison of proportions and the Chi-Square. It shows that the model has classified 95.9 percent of the objects correctly. More specifically, of the 146 companies in the sample,

- 89 sound companies (97.8%) were correctly classified as sound. The number of sound companies with negative diagnostic are the “true negative” cases (TN). The true negative rate (TNR) is TN divided by the number of sound cases.
- 51 failed companies (92.7%) were correctly classified as failed. The number of failed companies with positive diagnostic are the “true positive” cases (TP). The true positive rate (TPR) is TP divided by the number of failed cases.
- 2 sound companies were classified as failed by the model. The number of sound companies with positive diagnostic are the “false positive” cases (FP). The false positive rate is FP divided by the number of sound cases. The analysis of what may cause FP cases is important in classification.
- 4 failed companies were classified as sound by the model. The number of failed companies with negative diagnostic are the “false negative” cases (FN) and the analysis of what may cause FN cases is important in classification. FN divided by the number of failed companies is the false negative rate (FNR).

These rates are probabilities of observing, when using the model, positive or negative responses when the true response is failed or sound. They are not probabilities of failure or soundness when a positive or negative response is given by the model.

The Wald Chi Square measures how far each parameter of the model is from the value of zero, allowing the testing of the null hypothesis that the true parameter value, in the population, is indeed zero. The significance associated with this Wald Chi Square tells how likely such hypothesis is.

The significance of the model is measured using one of the pseudo-R-squared available:

-2 Log likelihood	Cox & Snell	Nagelkerke
41.373[a]	.647	.881

16.4 Probit regression

The Probit regression model assumes that there exists an underlying "latent" (hidden) variable y^* driving the discrete, observed outcomes of zeros and ones.

This latent variable follows a Normal distribution such that:

$$y^* = x\theta + \epsilon$$
$$\epsilon \sim N(0,1)$$

and the observed outcomes are

$$y_i = \begin{matrix} 0\ if\ y_i^* \leq 0 \\ 1\ if\ y_i^* > 0 \end{matrix}$$

In this case, the log-likelihood function is

$$\ln L\,(\theta) = \sum_{i=1}^{n} \left[y_i \ln \Phi\,(x_i\theta) + (1 - y_i) \ln\left(1 - (x_i\theta)\right) \right]$$

for *n* cases, where $\ln \Phi\,(x_i\theta)$ is the Normal cumulative distribution function. The θ are the parameters of the model that maximize $\ln L\,(\theta)$.

Due to the above normality assumptions, Probit modelling is extensively employed in the correction of data-generated biases in regressions.

A modelling example follows where Probit regression modelling is used instead of Logit modelling in the same TAFFLER dataset. Results are quite similar (notice the reversion of signs, which does not change predictive results):

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-2.746	1.0060	-4.717	-.774	7.448	1	.006
CF_TTA	19.884	5.8048	8.507	31.261	11.734	1	.001
S_TL	1.527	.5773	.396	2.659	7.001	1	.008
STL_NCE	-1.237	.3867	-1.995	-.479	10.238	1	.001
WC_TTA	5.552	1.4884	2.635	8.469	13.914	1	.000

Besides bankruptcy prediction, other areas of Finance that use logistic regressions to build predictive models are the detection of falsified accounts of companies, the forecasting of earnings increases or decreases one year ahead, the prediction of take-overs, and more. These models are widely used by financial analysts, investment banks and practitioners, in spite of the fact that

econometrists refuse to accept their validity on grounds that the samples employed to build them are not random.

16.5 Probabilities as Logit and Probit output

We mentioned that the *y* predicted by a logistic or Probit regressions are interpreted as probabilities of response, conditional on the $x_1, x_2, \ldots$ values, but what is the relationship between this $\hat{y}$ output and the real-world probability of response given $x_1, x_2, \ldots$?

To answer the question, three features must be taken into consideration:

1. the distribution of response and no response in the sample from which the model is built. In a sample where responses and no responses are represented with ω and 1- ω frequencies, the meaning of $\hat{y}$ will depend on ω.
2. the type of inference made by the algorithm, which dictates how the sample distribution ω and 1- ω will be interpreted. In this respect, logistic and Probit modelling whare maximum likelihood estimation is used, are different from multiple discriminant analysis (MDA) and other algorithms.
3. The prior probabilities associated with response and no response, that is, the "prevalence" of these states in the real world. Prevalence is the probability of a given state when nothing else is known. For example, the bankruptcy prevalence of mature companies of medium size is around 6 percent.

It may be difficult to answer to the above question. When the sample frequencies ω and 1- ω are made to be similar to prevalence, for example, when the sample is obtained at random from the population so that the proportion of bankrupt companies is 6 percent, the same as prevalence, logistic and Probit algorithms give as output an $\hat{y}$ which is the probability in the real world as desired. In this specific case, models can be used in practice to predict important financial attributes directly. However, we should be aware of the following difficulties.

Except in rare cases, the sample cannot be obtained at random from population because the sample must contain a good number of companies that belong to the minority class, say, bankruptcy, and in the real world these are just a few. If we sample at random from the population, the number of bankrupt companies in the sample will not be sufficient to allow inference. A sample with, say, 6 cases of bankrupt companies and 94 cases of non-bankrupt companies will be unable to identify bankrupt features accurately for lack of a sufficient number of patterns. Therefore, the common practice consists of including as many bankrupt cases as possible in the analysis and then choose non-bankrupt cases at random to match bankrupt cases. This type of sampling is known as "response-based". How is this going to affect results?

When response-based sampling is used, the predicted probability of bankruptcy will be much higher than that in the real world. Indeed, in the real world, the *a-priori* probability of a mature company going bankrupt in the short-term (if nothing else is known) is, as said, around 6 percent whereas in the sample used to build the model, bankrupt companies are much more frequent. In the TAFFLER example, bankrupt companies are some one-third of cases.

Third, the sample employed to build the model is the same as used to measure its performance. This type of "in-sample" performance assessment will inevitably offer an overoptimistic picture of performance. It is possible to obtain a trustworthy performance

assessment by using one sample to build the model and then testing its performance using a different sample, in what is known as “out of sample” assessment.

The random sampling issue is difficult to circumvent because, as said, samples obtained at random would be unworkable. Modelling algorithms require a minimum number of patterns to recognize an outcome. If, at best, 6 percent of cases are bankruptcies, this is insufficient for algorithms to work properly.

There is a number of solutions available to solve this problem, namely resampling and the correction of the logistic regression intercept term to reflect prevalence but in academic circles, econometrists disagree about their use. By contrast, practitioners use these solutions every day with good predictive results.

Chapter 17 Probabilistic reasoning, the Bayes rule

When employed in classification tasks, Logit and Probit provide probabilities of binary events being verified or not. For example, the probability of a firm going bankrupt conditional on the values of a few accounting ratios, can be obtained from appropriate Logit modelling.

The question is, after obtaining those probabilities of bankruptcy, what should we do to take the appropriate decisions? How do we reason using probabilities? Probabilistic reasoning underpins decision making in the presence of uncertainty.

17.1 Prevalence

To decide in the presence of uncertainty requires foremost that a given state is identified from among others as being the issue or problem at hand. A bank analyst, for instance, wants to know what are the causes that may lead clients to default in their obligations. This analyst is required to identify a specific "illness" among other possible illnesses.

Financial distress is one possible state, (the opposite of financial health) capable of leading to loan default and other undesirable consequences. There are other undesirable states equally possible, and the analyst wishes to find out which, if any, of those states may afflict a given client. In this, analysts seek to "diagnose", to make a diagnostic of each client, much in the same way practitioners do with presumptive illnesses.

Probabilistic reasoning requires knowledge of the expected frequency of undesirable states. These probabilities are known as "prevalence". For example, if the expected frequency of loan defaults is 10 in 1,000 loans, prevalence of loan default is 1 percent.

17.2 Conditional probability

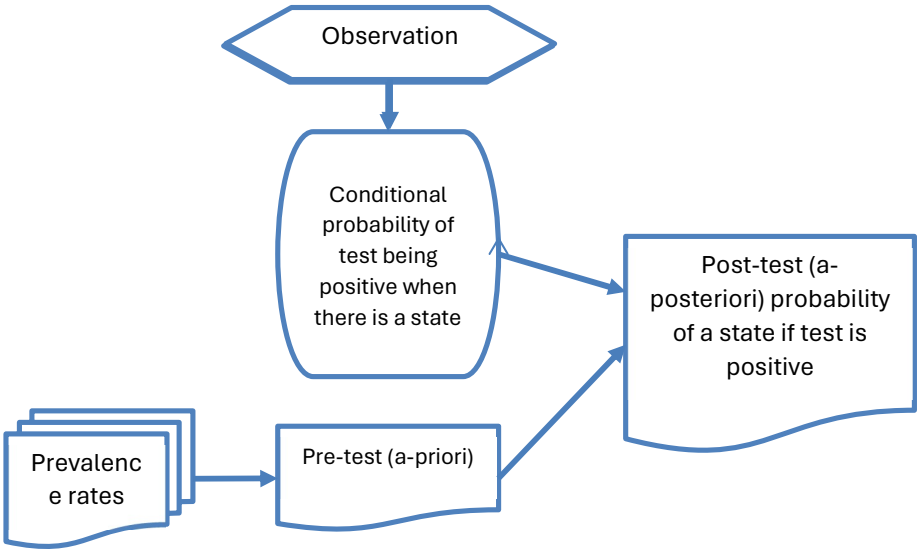
Prevalence is an a-priori probability, indicating the probability of a given state when nothing else is known. Possessing knowledge about different likelihoods (prevalence) associated with possible unfavorable states is the first step in the process of discovering the state which actually afflicts a particular firm. This is process is to "diagnose".

To identify a state, for example financial distress in a firm or in a family, bank analysts perform observations (measurements). Starting from prevalence, each new observation (solvability, profitability, liquidity ...) increases or decreases the odds of a state being verified. The reasoning that leads to the final diagnostic is probabilistic:

1. It begins with the knowledge of a collection of prevalence, which predispose the analyst to *a-priori* accept the most probable states and reject the less probable states. Prevalence rates are also known as "pre-test" in experimental sciences. Hereafter,
2. each new observation (or test, in the case of experimental sciences) brings a new likelihood, called "conditional probability" of that observation being positive in objects with the state in question. For example, a logistic regression may bring in new probabilities of default.
3. Analysts then combine the a-priori knowledge (prevalence) and the conditional probability, thus obtaining the probability, called "a-posteriori" of the object having the state when the observation is positive (or otherwise).

4. The probability a-posteriori is then used as a new prevalence rate, a new starting point, and the process repeats itself for other observations or tests on the subject.

Finally, there comes a time when the resulting probability a-posteriori is high, thus the decision can be made with some degree of assurance.



This is the process followed by analysts, practitioners, and other decision-makers to diagnose in the presence of uncertainty.

Practitioners can make tests (for example, blood tests), which is something that analysts cannot do with companies. Analysts can run logistic regressions or other diagnosing tools. Otherwise, the probabilistic reasoning used by bank analysts to come to the most likely state is the same as that used by other professionals that make decisions under uncertainty. It involves a diagnosis.

Suppose that a practitioner will see a patient.

1. A-priori, he doesn't know anything about this patient. Just knows the prevalence rates of diseases and this predisposes for accepting diseases that are more likely.
2. Before allowing the patient in, the doctor realizes that the name is a male name. Therefore, multiple conditional probabilities are introduced and, once combined with prevalence rates, increase and decrease a-posteriori (post-test) probabilities of various diseases:
    a. A-posteriori probabilities of women's disease when sex is male become zero: conditional probability of sex being male when diseases are women's, is zero.
    b. A-posteriori probability of the object having cardiovascular disease or others when sex is male, increase: conditional probability of being male when having those diseases is larger than average.
3. In possession of a-posteriori probabilities the doctor makes further observations. The patient is found to be middle-aged and over-weighted.
    a. A-posteriori probability of cardiovascular disease increases: conditional probability of middle-aged obese person having such disease is high.
4. When estimating a-posteriori probability the doctor uses previous a-posteriori probabilities for male patients (male prevalence), not all, general, prevalence.
5. Once a a-posteriori probability stands out among all others, the doctor measures blood pressure and test other risk factors of the suspected disease.

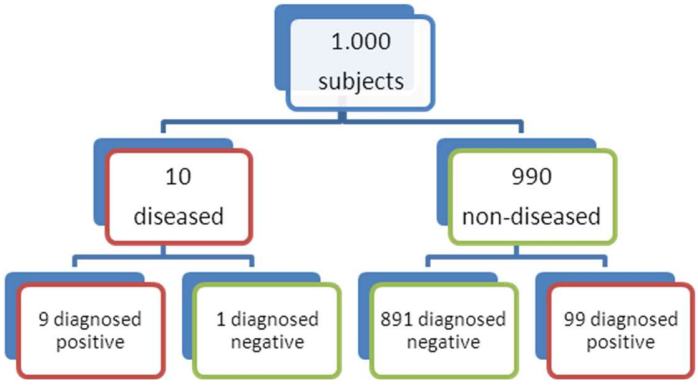
6. Such observations and tests, if positive, add likelihood to the reasoning: high blood pressure when disease is cardiovascular, cholesterol high when disease is cardiovascular and others. These conditional probabilities are joined to previous a-posteriori probability and give rise to new a-posteriori probabilities.

Finally, the doctor decides that a given likelihood is high enough to diagnose, and he makes a diagnostic pointing to the disease whose a-posteriori probability is highest. In practice, this process experiences relapses and restarts. Each step involves a diagnosis that may or may not be reinforced by subsequent observations. In all steps, however, doctors conditionally accept the disease with higher a-posterior probability while performing new observations and tests to try to reinforce or dismiss such hypothesis.

The advantage of explicitly knowing probabilistic reasoning underlying diagnosis is to allow quantification of costly decisions (costly in cash or in discomfort for the patient) as is the case of decision to perform expensive tests. There are cases in which such tests cannot increase the likelihood a-posteriori and, being useless, they should be avoided.

17.3 Post-test probabilities

The question analysts need to answer is, what is the probability of this object having X state when diagnosis is positive? In the example below, 1,000 objects are first divided in diseased and non-diseased; then, both such groups are classified into truly and falsely diagnosed. At the top of the hierarchy, the only information available is that there are 1,000 objects in total. No references to disease yet. The count, 1,000 objects, is used to calculate frequencies. At the second level of the hierarchy, it appears the disease and its prevalence: the probability of disease is, from the outset, 10 in 1,000 or 1%.



Previous level provides the denominator. On the third level of the hierarchy appears the diagnosis. Frequencies, as before, are calculated by comparison with previous level:

- "True positive" rate, TPR: 9 in 10 diseased, 90% are diagnosed as diseased (positive) when they are diseased. TPR is the probability of an observation or test being positive when there is disease.

$$TPR = \frac{Number\ of\ diseased\ patients\ with\ positive\ diagnostic}{Number\ of\ diseased\ patients}$$

- "False negative" rate FNR: 1 in 10 diseased, 10%, are diagnosed as non-diseased (negative) when, in fact, they are diseased.

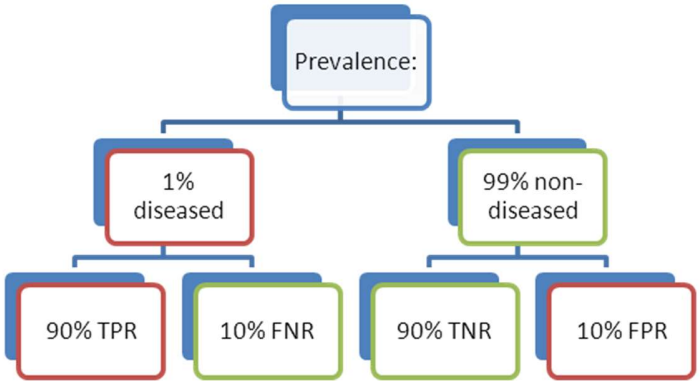
- "True negative" rate TNR: 891 in 990 objects, 90%, are diagnosed as non-diseased (negative) when they are non-diseased. TNR is the probability of a test being negative when there is disease.

$$TNR = \frac{Number\ of\ nondiseased\ patients\ with\ negative\ diagnostic}{Number\ of\ nondiseased\ patients}$$

- "False positive" rate FPR: 99 in 990 objects, 10%, are diagnosed as diseased (positive) when, in fact they, are non-diseased.

These rates are basically probabilities of observing positive or negative responses when, in reality, the response is diseased or not diseased. They are not probabilities of disease or not disease when a positive or negative response was observed.

The above hierarchy can be written with reference to TPR, FNR, TNR and FPR thus:



The same can be presented in the crosstab form as a confusion matrix:

diagnostic/ patients	Diseased	non-diseased	Total
positive	9 TP	99 FP	108=P
negative	1 FN	891 TN	892=N
total	10 = D	990 = ND	1.000

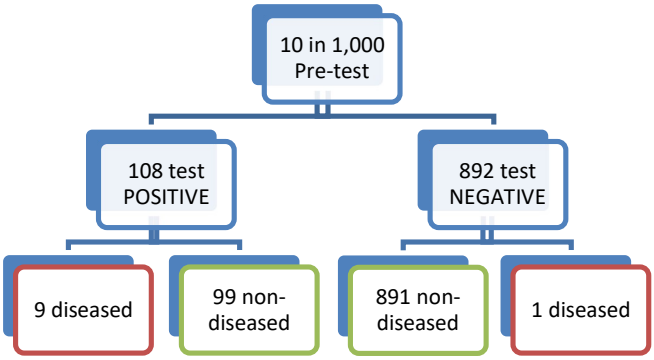
Rates TPR FNR TNR FPR are calculated dividing frequency of each cell by respective totals (total of diseased or total of non-diseased on the line below the table). Only two of such rates are independent since TPR + FNR = 1 and TNR + FPR = 1.

Practitioners do not need to know the probability that a test be positive in truly diseased objects (TPR) or negative in truly non-diseased objects (TNR), but the probability of disease when the test is positive or non-disease when the test is negative.

Such probabilities can be obtained from the above table using as denominator the line of total positive and total negative instead of the line of total diseased and non-diseased:

1. Total positives are 108 = 9 + 99; of these, 9 are diseased and the others are non-diseased. The likelihood of being diseased when tested positive is 9 in 108 or 8.3% and it's called "post-test probability of disease for positive result" or "post-test prevalence". The likelihood of disease rises from 1% (the prevalence, also known as "probability a-priori" or "pre-test probability of disease") to 8.3% due to the test being positive.
2. What is the probability of a object being diseased when tested negative? The total number of negatives is 892 = 891 + 1; of these, 1 is false non-diseased and 891 are true non-diseased. The likelihood of someone being diseased when tested negative is 1 in 892 or 0.112% and is called "post-test probability of disease for negative result". The likelihood of disease goes down from 1% (prevalence) to 0.1% due to test being negative.

3. Other two probabilities that may be considered are: probability of non-disease for positive result, 99 in 108 or 91.7%; and probability of non-disease for negative result, 891 in 892 or 99.88%.



Note how the above numbers link with the probabilistic reasoning presented above:

1. Prevalence is the probability a-priori or "pre-test" of disease. It is the starting point of any probabilistic reasoning.
2. TPR and TNR are conditional probabilities already referred to. They show the plausibility of an observation or test being
    a. positive when the object is diseased,
    b. negative when the object is not diseased.

    Note the terms used (in epidemiology, not Finance) to designate TPR and TNR:
    a. TPR is known as “sensitivity”.
    b. TNR is known as “specificity”.
3. FNR and FPR are the reverse of TPR and TNR respectively:
    c. FNR = 1-TPR
    d. FPR = 1-TNR.
4. The arrival point of probabilistic reasoning, what decision-makers want to know
    e. is the post-test probability of disease for positive result, the a-posteriori probability of disease when the diagnosis is positive. When the doctor accepts the observation or test as good and is willing to continue down this path, such post-test probability becomes the new starting point for probabilistic reasoning, being known as “post-test prevalence”.
    f. is the post-test probability of disease for negative result, the a-posteriori probability of disease when the diagnosis is negative.
5. The two a-posteriori probabilities or “post-test” of non-disease when the test
    g. is positive
    h. is negative.
    i. are complement of previous case (hypothesis of independence accepted).

The above is the quantification of probabilistic reasoning leading to diagnoses.

17.4 Bayes’ rule

So far, frequencies have been used to illustrate probabilistic reasoning. How to proceed when only probabilities are available? The formula to calculate a-posteriori probabilities from a-priori probabilities plus conditional probabilities is called the “Bayes Rule”. Suppose two binary, independent attributes:

1. real situation (with states, D and D') and
2. observation (with states T and T').

In such case Bayes' rule is:

$$P\ (D|T) = \frac{P(D)P(T|D)}{P(D)P(T|D) + [1 - p(D)]P(T|D')}$$

where

1. P(D) is a-priori probability of situation D. P(D) may be, e.g., the probability of disease, that is, the prevalence, or the overall probability of bankruptcy.
2. P(T|D) is conditional probability of observation T given situation D. P(T|D) is the probability of a test being positive when the object is diseased, as in TPR. Or the TPR obtained from the output of an algorithm that predicts bankruptcy.
3. P(T|D') is conditional probability of observation T given situation D'. P(T|D') is FPR, that is, probability of test being positive when object is non-diseased (D'); False positive rate is 1-TNR as predicted by an algorithm.
4. P(D|T) is what we wish to know. The "a-posteriori" probability of situation D given observation T. P(D|T) is, e.g., post-test probability of disease for positive result; the probability of bankruptcy when the algorithm says "bankrupt".

Using epidemiological names:

post-test probability of disease for positive result $= \frac{Prev \times TPR}{Prev \times TPR + (1 - Pre\ ) \times FPR}$

post-test probability of disease for negative result $= \frac{Prev \times FNR}{Prev \times FNR + (1 - Pre\ ) \times TNR}$

Two post-test probabilities of non-disease for positives / negatives complement this.

For example:

Prob. a-priori is prevalence or pre-test probability of disease = 0.02

Prob. conditional of D is TPR or sensitivity = 0.9

Prob. conditional of D' is FPR = 490/4,900 = 0.1

Prob. a-posteriori = (0.02 * 0.9) / (0.02 * 0.9 + 0.98 * 0.1) = 0.018 / (0.018 + 0.091) = 0.018 / 0.116 = 0.155 or 15.5%, the same as calculated using crosstab.

-/-

Another example with higher prevalence:

diagnostic/patients	diseased	non-diseased	total
positive	160 TP	80 FP	240 P
negative	40 FN	720 TN	760 N
total	200 D	800 ND	1.000

In this case:

total diseased diagnosed as positive: 160 + 80 = 240 of which 160 are TP.

post-test probability of disease for positive result: 160 in 240 = 66.6%.

This probability, 66.6%, is lower than 80% of sensitivity but the likelihood of correctly diagnosed disease rises from 20% (prevalence) to 67%, a nice gain of 47%.

When using the Bayes ' rule:

Prevalence, 200/1,000 = 0.2 or 20%.

TPR, sensitivity, 160/200 = 0.8 or 80%

FPR, 80/800 = 0.1 or 10%.

TNR, specificity, 720/800 = 0.9 or 90%. TNR + FPR = 1.

FNR, 40/200 = 0.2 or 20%. TPR + FNR = 1.

post-test probability of disease for positive result =

(0.2 * 0.8) / (0.2 * 0.8 + 0.8 * 0.1)

= 0.16 / (0.16 + 0.08) = 0.16 /0.24 = 0.666 or 66.6% as above.

post-test probability of disease for negative result =

(0.2 * 0.2) / (0.2 * 0.2 + 0.8 * 0.9)

= 0.04 / (0.04 + 0.72) = 0.04 /0.76 = 0.052 or 5.2%.

Uncertainty has decreased in relation to the previous situation where only prevalence is known.

-/-

The above examples show that prevalence influences the amount of uncertainty that can be removed with a diagnostic test / logistic model of bankruptcy.

1. When prevalence is low, only tests with high sensitivity can remove uncertainty. Means of diagnose whose sensitivity is not at that level almost ideal, should not be used because it will not remove uncertainty. An example is the PSA, whose sensitivity in detection of prostate cancer is not high and therefore the post-test probability of disease when the PSA is positive, is low.
2. When the prevalence is high, the diagnostic tests are often unnecessary because the uncertainty that the doctor faces is too low. Should not therefore use.
3. Finally, when the prevalence is low but not too much, then the diagnostic tests can remove a substantial amount of uncertainty even when their sensitivity, being high, is not absolute.

-/-

The last example concerns the PSA test for prostatic cancer. Data (approximate) are as follows:

1 in 100 diseased: prevalence = 1%.

30 in 100 positives are diseased: TPR = 30%, sensitivity. Since TPR + FNR = 1

70 in 100 negatives are diseased: FNR = 70%,

15 in 100 positives are non-diseased: FPR = 15%. Since TNR + FPR = 1

85 in 100 negatives are non-diseased TNR = 85%, specificity.

Thus,

Post-test prob. of disease for pos. result =

(0.01 * 0.3) / (0.01 * 0.3 + 0.99 * 0.15)

= 0.03/(0.03 + 0.148) = 0.03/0.1785 = 0.168, half of sensitivity.

Post-test prob. of disease is neg. result =

(0.01 * 0.7) / (0.01 * 0.7 + 0.99 * 0.85)

= 0.07/(0.07 + 0.8415) = 0.07 / 0.911= 91%, higher than specificity.

False positives are in a worrying proportion then.

To estimate expected costs stemming from false-positives or false negatives, the above probability of non-disease for positive result should be multiplied by the respective unit cost and the probability of disease for negative result should also be multiplied by the respective unit cost.

Chapter 18 Inverse Mill's ratio, Tobit models

One of the attractive features of the Probit regression is that it can be used to account for biases caused by limitations in samples, namely missing observations.

We have studied censored cases in survival analysis, which are objects that disappear from the otherwise random sample and may lead to a severe increase in standard errors in estimates. Another example is the birth weight dataset (file "BWGHT"), where 197 of the 1388 cases have missing father education. Now, the probability that education is missing may be higher in low education fathers, and this invalidates random sampling. Also, the WAGE1 dataset was made from objects where IQ scores were available. Unavailable IQ cases were excluded, but low IQ objects make it more difficult to collect IQ data. As a consequence, WAGE1 may be unusable.

Sample selection bias is a challenge to face, but it is one among other cases where datasets and their limitations will compromise inference.

18.1 Selection bias

Missing data often occurs in samples. Any object that has missing attributes cannot be modelled, but that is not an econometric concern *per se*. The econometric problem with missing data is that the sample may not be random. There may arise a "selection bias".

When the missing observations are in predictors only, we call it the "exogenous sample selection" and no *a-priori* bias exists: the sample is selected according to predictors, e. g., all man below 35 years old are excluded. However,

- In the "BWGHT" dataset, the probability that education is missing is higher in low education parents than in educated parents. Missingness, that is, the fact that an observation is missing or not missing, influences education and therefore there is a breach of the random sampling assumption MLR.2.
- In the WAGE2 dataset, unavailable IQ cases were excluded but these excluded cases are likely to be low IQ. The sample is not random for the same reason.

MLR.2 is tenable only if expectations regarding the predicted variable do not change with the missingness of observations, namely if $E(predicted \mid predictors)$ is constant in any subset of the population: expected predicted variable, conditional on predictors' values, is constant. Clearly, this is not the case in the above examples.

When missing observations are present in the predicted variable, we say that we face "endogenous sample selection". In most cases, this condition leads to biased inference.

This chapter shows how to circumvent sample selection problems and exemplifies cases of limited dependent (predicted or *y)* variables and how to correct biases that may upset OLS estimation due to this limited *y*. Specifically, the following four cases are worth noting, of which we show two relevant examples:

- Data on *y* is "censored" if, for part of the range of *y*, we observe only that *y* is in that range, rather than observing the exact value of *y*. e. g. income is top coded at $75,000 per year. Censoring can arise for distributions other than the normal. For time duration data, e. g. the length of a spell of unemployment or a survival time, a separate treatment of censoring is

needed, due to its peculiar censoring random mechanism, as mentioned in relation to the comparison of proportions

- Data on *y* is "truncated" if for part of the range of *y* we do not observe *y* at all. e. g. people with income above $75,000 per year are excluded from the sample.
- Data on *y* is a "corner solution" response, which is zero for a nontrivial fraction of the population but is continuously distributed over the other positive values. This is also known as "left-censored" data, and there may be cases of "right-censored" and both left- and right-censored data.
- Data on *y* is "incidentally truncated" if we do not observe *y* at all for a nontrivial fraction of the population, but the fact that *y* are missing does not depend on the outcome of another variable. Incidental truncation is a clear case of non-random sample selection leading, in most cases, to a sample selection bias.

In all of these cases, we face missing *y* except for a restricted subsample. Meaningful analysis requires extrapolation from the restricted subsample into the population as a whole, but running regressions on censored or truncated data, without controlling for censoring or truncation, leads to inconsistent parameter estimates.

Using the dataset MROZ on the labour of women in the US, we offer two examples of how to tackle the bias which limited dependent variables bring into OLS estimation. These examples concern

incidentally truncated *y* and

corner solution *y*.

18.2 The Heckit method for incidentally truncated predicted variables

Heckman proposed a two-stage estimation method to correct for selection bias caused by incidentally truncated *y*. The procedure is known as the "Heckit" model. In the first stage, a Probit model is used to discriminate between two groups of objects, namely

those where the dependent variable is observed, and

those where the dependent variable is missing.

Then, the "inverse Mill's ratio" is computed from the estimated Probit model's output.

For a given Probit output score *z*, the inverse Mill's ratio is the quotient of two standard normal functions, the density ϕ (PDF) and the cumulative Φ (CDF):

$$IMR(z) = \frac{\phi(z)}{\Phi(z)} = \frac{PDF(z, 0,1)}{CDF(z, 0,1)}$$

The inverse Mill's ratio transforms a zero-cantered *z* into a positive variable that is zero for positive *z* and increasingly positive for increasingly negative *z*. As a result, when the inverse Mill's ratio is used as predictor in models explaining incidentally truncated *y*, it will be zero for observed *y* and positive for missing *y*. The inverse Mill's ratio is widely used in statistics.

The inverse Mill's ratio is then included as explanatory variable in the OLS estimation, causing it to be unbiased regarding missing observations in the predicted variable.

In the MROZ dataset, 325 out of 753 respondents are not in the labour market and therefore earn no salary and work during zero hours per day.

Since the salary observation is missing for 325 women, if we want to predict the salary based on other variables, this will be a case of incidental truncation to be tackled by the Heckit model.

Also, in case we want to predict the number of hours of work, since, for 325 women the number of working hours is zero, this will be a case of left-censored data (corner solution), to be tackled using the Tobit model.

We now show how to apply the Heckit model to the prediction of salaries (wages). The dataset MROZ contains, among other attributes, `Inlf, hours, kidslt6, age, educ, wage.`

where *inlf* is the dummy indicating employed (1) or unemployed (0) cases, with frequencies of 428 and 325 respectively. We wish to predict the logarithm of wages but only employed cases have observed wages. For unemployed cases, wages is missing.

In the first stage of the Heckit method we run a Probit model explaining absence from work in terms of available predictors:

$$Inlf = Probit(a + b_1 educ + b_2 exper + b_3 expersq + b_4 kidslt6)$$

and we obtain the predicted probabilities associated with *inlf*, conditional on *educ*, *exper*, *expersq* (squared *exper*) and *kidslt6* values.

The following are the respective Probit results:

```
Probit binary choice model/Maximum Likelihood estimation
Newton-Raphson maximisation, 4 iterations
Return code 1: gradient close to zero (gradtol)
Log-Likelihood: -434.1902
Model: Y == '1' in contrary to '0'
753 observations (325 'negative' and 428 'positive') and 5 free parameters (df = 748)
Estimates:
               Estimate  Std. error t value   Pr(> t)
(Intercept) -1.97888924  0.29349938 -6.7424 1.558e-11 ***
educ         0.11522425  0.02286361  5.0396 4.664e-07 ***
exper        0.12321757  0.01800496  6.8435 7.726e-12 ***
expersq     -0.00243875  0.00058447 -4.1726 3.011e-05 ***
kidslt6     -0.49547330  0.10027364 -4.9412 7.764e-07 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
Significance test:
chi2(4) = 161.3659 (p=7.446758e-34)
```

From the obtained probabilities, we now compute the inverse Mill's ratio *imr* and we run the desired OLS regression,

$$\log(wage) = a + b_1 educ + b_2 exper + b_3 expersq + b_4 imr$$

where *imratio*, from the first stage, is included as predictor. This is the second stage of the Heckit model. Results are as follows,

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1761710  0.5506569  -0.320    0.749
educ         0.0977484  0.0202379   4.830 1.91e-06 ***
exper        0.0274730  0.0247323   1.111    0.267
expersq     -0.0005259  0.0005781  -0.910    0.363
imratio     -0.1836831  0.2727257  -0.674    0.501
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
Residual standard error: 0.6669 on 423 degrees of freedom
(325 observations deleted due to missingness)
Multiple R-squared:  0.1577,     Adjusted R-squared:  0.1498
F-statistic:  19.8 on 4 and 423 DF, p-value: 5.884e-15
```

where the bias caused by missing data has been removed. The R command "hackman" can implement the above two stages automatically and can account for the fact that a variable generated

from the first step, the inverse Mills ratio, is used in the second stage as predictor, by adjusting the standard error and *t* value.

18.3 The Tobit method for corner solutions

Using the same MROZ dataset, we now present an example of how the Tobit model solves the problem of biases caused by a "corner solution" dependent variable.

A corner solution response is zero for a fraction of the population but is continuously distributed over the other positive values. Examples are models explaining the amounts of dividends paid (many don't pay), the amount which an individual spends on alcohol in a given month (many don't drink) or the number of working hours of woman (many work at home). This problem is known as the Tobit model or left-censored.

A linear fit might be a good approximation, but we would obtain negative predictions for *y*. It is important to have a model that implies nonnegative predicted values for *y*, and which has sensible partial effects over a wide range of the explanatory variables. Plus, we sometimes want to estimate features of the distribution of *y* given $x_1, x_2, \ldots$ predictors. The Tobit model is convenient for these purposes.

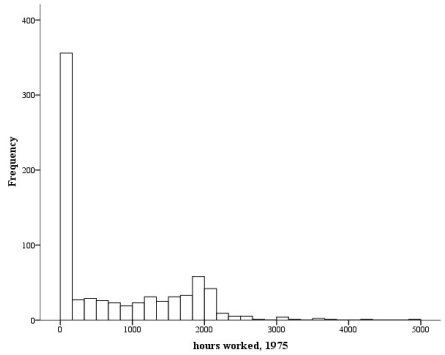
The Tobit model expresses the observed response of *y* in terms of a "latent" variable y* that satisfies the classical linear model assumptions; in particular, it has a normal, homoscedastic distribution with a linear conditional mean.

The observed variable *y* equals y* when y* ≥ 0, but y = 0 when y* < 0.

Because y* is normally distributed, *y* has a continuous distribution for positive values. The density of *y* given *x* is the same as the density of *y** given *x* for positive values.

The procedure used in the case of a Tobit model is similar to the two-stages Heckit method, but the maximum-likelihood estimation (MLE) is applied differently.

In the example now provided, we wish to explain hours of work in the MROZ dataset using several predictors. The problem, of course, is that hours of work is zero for many cases which are unemployed, as illustrated in the histogram below.



We wish to predict *hours* thus:

$$hours = a + b_1 nwifeinc + b_2 educ + b_3 exper + b_4 exper^2 + b_5 age + b_6 kidslt6 + b_7 kidsge6$$

Results are as follows:

```
Observations:
          Total     Left-censored     Uncensored     Right-censored
            753               325            428                  0
```

```
Coefficients:
              Estimate Std. error t value  Pr(> t)
(Intercept)  965.30528  446.43603   2.162 0.030599 *
nwifeinc      -8.81424    4.45910  -1.977 0.048077 *
educ          80.64561   21.58324   3.736 0.000187 ***
exper        131.56430   17.27939   7.614 2.66e-14 ***
expersq       -1.86416    0.53766  -3.467 0.000526 ***
age          -54.40501    7.41850  -7.334 2.24e-13 ***
kidslt6     -894.02174  111.87803  -7.991 1.34e-15 ***
kidsge6      -16.21800   38.64139  -0.420 0.674701
logSigma       7.02289    0.03706 189.514  < 2e-16 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Newton-Raphson maximisation, 7 iterations
Return code 1: gradient close to zero (gradtol)
Log-likelihood: -3819.095 on 9 Df
```

The variable *logSigma*, significant in this case, corrects for the bias introduced by the undesired number of zero cases in *hours*, having been obtained following a two-stage procedure similar to the Hackit method. Using the reading provided, it is possible to replicate in the R package each of the two stages separately.

Chapter 19 The modelling of volatility

Option pricing and related modelling techniques require that the volatility of returns on securities underlying these options be assessed. It is possible to estimate the volatility of prices / returns implied by a given option price. However, it is often desirable to forecast option prices, and, in that case, the behavior of volatility should be forecasted first.

To understand the need to forecast volatility, the important facts to retain are:

In efficient markets, current returns are indeed independent of past returns (no information about the past is relevant) but the same cannot be said of the current volatility of these returns, which may be predicted from past volatility.

However, returns are heteroscedastic, with volatility that changes over time, and it is important to find the time-series models that fit the volatility of returns.

The pricing of options has uncovered the fact that the volatility of the underlying returns is, not just required but also a major driver of option prices.

19.1 Estimating and predicting volatility

Let us first use the NYSE dataset to illustrate the predictability of returns' volatility. First, we predict returns from an expected or mean return *a*, plus lagged returns:

$$r_t = a + b\, r_{(t-1)} + e_t$$

As seen, *b* is non-significant. Next, we square the residuals e_t and we fit the model

$$e_t^2 = a + b\, r_{(t-1)} + e_t + \varepsilon_t$$

where squared residuals are explained by previous period returns and current residuals:

```
             Estimate Std. Error t value Pr(>|t|)
Intercept     4.6565     0.4157  11.201  < 2e-16 ***
return_1     -1.1041     0.1958  -5.640 2.49e-08 ***
residuals    -1.2597     0.1964  -6.413 2.66e-10 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Residual standard error: 10.87 on 686 degrees of freedom
Multiple R-squared:  0.0961,     Adjusted R-squared: 0.09346
F-statistic: 36.47 on 2 and 686 DF, p-value: 8.9e-16
```

The coefficient of $r_{(t-1)}$ now is highly significant, indicating heteroscedasticity, and is negative, showing that the volatility of stock returns is lower when the previous return was high and vice-versa. Therefore, expected returns may not depend on past returns, but the variance of returns does.

Autoregressive conditional heteroscedasticity (ARCH) is a model of the volatility of a time-series, not of the time series itself. ARCH predicts the squared residuals of time series models using ARMA or similar procedures to account for heteroscedasticity in these squared residuals. For example, a typical but simple ARCH model would be

$$e_t^2 = a + b\, e_{t-1}^2 + e_t + \varepsilon_t$$

where current volatility is predicted from past volatility and from non-squared residuals of the return-predicting AR (1) model. In the same example, results are:

```
             Estimate Std. Error t value Pr(>|t|)
Intercept     3.0999     0.4345   7.134 2.49e-12 ***
res1sq_lag    0.3028     0.0361   8.389 2.80e-16 ***
```

```
res1         -0.9345     0.1953  -4.784 2.11e-06 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Residual standard error: 10.59 on 685 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1423,     Adjusted R-squared:  0.1398
F-statistic: 56.81 on 2 and 685 DF, p-value: < 2.2e-16
```

Big t-statistics in lagged volatility indicates strong ARCH.

The term "conditional" in ARCH means that the current volatility is conditional on past information. High volatility persists for a while, the same as low volatility, Shocks in market prices create more or less extended periods of high volatility. Thence the use of AR (1) of squared volatility to explain current volatility.

19.2 ARCH and GARCH modelling

ARCH is the name given to ARMA models that predict the volatility of time-series, not the time-series themselves. Formally, the ARCH model assumes that the conditional mean of the error term in a series model is zero (therefore, constant) but the conditional variance is not. Such model can be described as follows: given y_t, for example a series of daily returns, then the residuals e_t of an appropriate modelling of y_t, say ϕ,

$$y_t = \phi + e_t$$

are Normal with variance h_t, conditional on information I_{t-1} about past returns, that is,

$$e_t | I_{t-1} \sim N(0, h_t)$$

The variance h_t is what we want to estimate, and we use ARMA models to that effect. In the simple case of AR (1), h_t would be estimated by two parameters α_0 and α_1 in

$$h_t = \alpha_0 + \alpha_1 e_{t-1}^2, \quad \alpha_0 > 0, \ \ 0 \leq \alpha_1 < 1$$

and the predicted variance would have constant mean. This variance-estimating model is the same we used above to verify the presence of heteroscedasticity in residuals.

The generalized autoregressive conditional heteroscedasticity (GARCH) is an extension of ARCH, incorporating moving average components together with the autoregressive component. GARCH, therefore, includes lagged variance terms plus lagged residual errors from a smoothing-of-the-mean variance process, that is, in general,

$$h_t = \omega + \sum_{i}^{q} \alpha_i e_{t-i}^2 + \sum_{1}^{p} \beta_i h_{t-i}$$

where

p is the number of lags of the variance to include in the model, and

q is the number of lags of residuals to include in the model.

The introduction of moving average components allow modelling conditional change in variance over time, not just changes in the time-dependent variance. These components lead themselves to adaptive learning. Adaptive, real-time learning nowadays is part of the variance-forecasting algorithms available.

The notation widely accepted for GARCH is GARCH (*p*, *q*). GARCH (1, 1) is a first order ARCH model with first order moving average. For *p* = *q* = 0 we model white noise. In ARCH (0, *q*), the conditional

variance is specified as a linear function of past sample variances only, whereas the GARCH (*p*, *q*) process allows lagged conditional variances to enter as well.

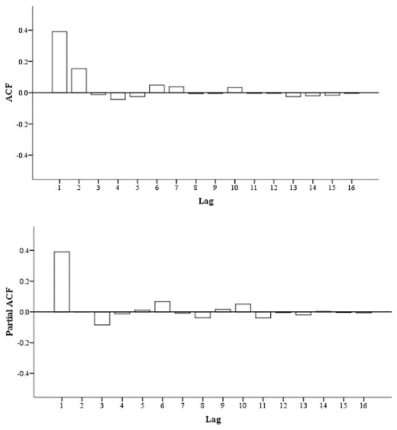
19.3 Applying GARCH

Similar to ARMA, the first step of GARCH modelling consists of finding the proper configuration. The configuration for a GARCH model can be found as that of ARMA models by examining ACF and PACF plots of the variance of the time series.

We just follow these two steps:

1. subtract the mean from observations in the series and square the result or square the residuals from a simple regression model of the series with their lag.
2. interpret ACF and PACF plots to estimate values for *p* and *q*.

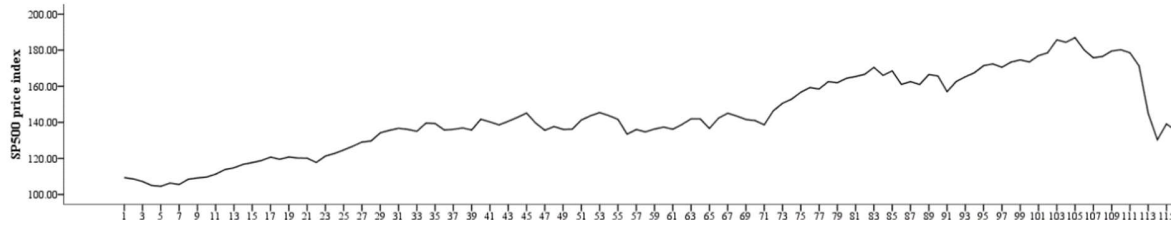
The NYSE dataset has the following ACF and PACF of squared residuals of an AR (1) model of returns. The plots suggest that both AR (1) and MA (1) effects are present.

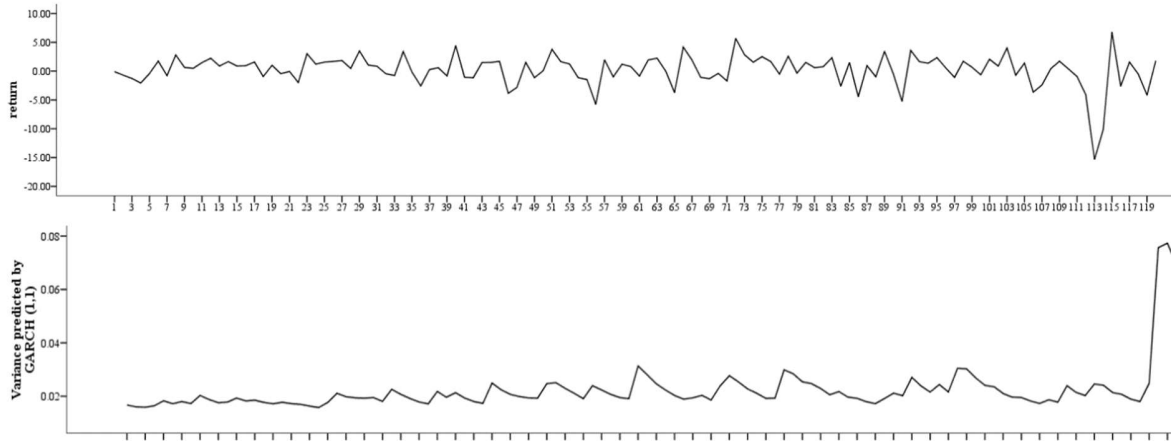


The next step, therefore, consists of fitting GARCH (1,1). Results are as follows:

```
     Estimate  Std. Error  t value Pr(>|t|)
a0 7.465e-05   2.611e-05    2.860  0.00424 **
a1 1.897e-01   2.293e-02    8.273 2.22e-16 ***
b1 6.527e-01   6.982e-02    9.348  < 2e-16 ***
   Jarque Bera Test, data: residuals
X-squared = 285.13, df = 2, p-value < 2.2e-16
   Box-Ljung test, data: squared residuals
X-squared = 2.0464, df = 1, p-value = 0.1526
```

where “a0” is the constant, “a1” is the AR (1) term and “b1” is the MA (1) term. These coefficients are positive and significant. “Jarque-Bera” test verifies that residuals are not normal. “Box-Ljung” Q test, the hypothesis that residuals are not auto correlated cannot be rejected. These three plots compare prices, returns and the fitted variance.





It is worth noting how predicted variance persists after every shock, decaying slowly.

The ARCH model replaces assumption of constant volatility by conditional volatility, recognizing that past volatility influences future volatility—that is the definition of auto-regression. The conditional heteroscedasticity portion of ARCH simply refers to the observable fact that volatility in financial markets is not constant—all financial data, whether stock market values, oil prices, exchange rates, or GDP, go through periods of high volatility. Economists have always known that volatility varies but they often kept it constant because they lacked a better option when modeling markets.

ARCH models can forecast beyond the volatility clusters that are seen in the market during periods of financial crisis or other events. For example, volatility for the S&P 500 was unusually low for an extended period during the bull market from 2003 to 2007, before spiking to record levels during the 2008 crisis. ARCH models are able to correct for the statistical problems that arise from this type of pattern in the data.

ARCH models work best with high-frequency data (hourly, daily), being thus ideal for market data. Financial institutions use GARCH to estimate the volatility of stocks, bonds, and market indices to determine pricing, judge which assets will potentially provide higher returns, and forecast the returns of current investments to help in asset allocation, hedging, risk management, and portfolio optimization decisions.

ARCH was developed by R. Engle in the 1980s in response to Friedman's conjecture that it was the uncertainty about what the rate of inflation would be rather than the actual rate of inflation that had a negative impact on the economy. Where there is heteroscedasticity in error terms, periods of high volatility are followed by higher volatility and periods of low volatility are followed by lower volatility. Volatility tends to cluster in specific time-periods, which is useful to investors when considering the risk of holding an asset over different time periods.

-/-

GARCH has spawned many related models that are also widely used in research and in finance, including EGARCH, STARCH, and others. These variant models introduce changes in terms of weighting and conditionality in order to achieve more accurate forecasting ranges. For example, EGARCH, or exponential GARCH, gives a greater weighting to negative returns in a data series as these have been shown to create more volatility. Put another way, volatility in a price chart increases more after a large drop than after a large rise. Most ARCH model variants analyze past data to adjust weights (parameters of the model) using a maximum likelihood approach. This technique results in

an adjustable, dynamic model that can forecast near-term and future volatility with increasing accuracy.

This area is at the forefront of financial research. In order to proceed from here, it is necessary to add two ingredients to our understanding of how prices behave in markets. The first is Continuous Time finance, a branch of Stochastic Calculus, and the second is Monte Carlo methods, specifically how to sample at random to obtain the desired distribution. In the coming chapters we will highlight these topics.

Chapter 20 Continuous time finance

In Chapter 3 we introduced random processes and explained why it is worthwhile to develop analytical mechanisms capable of replicating random events found in real life. It is indeed important to describe mechanisms underlying randomness and their distribution functions because, in this way, we can estimate probabilities analytically.

The first example then offered, an urn with 80 white balls and 20 black balls from which a ball is drawn at random with replacement, is the Bernoulli process, a "discrete" mechanism in the sense that events are separated from each other and do not form a continuum in time.

We now introduce continuous-time random processes widely used in Finance, namely in valuation. From these processes it is possible to deduce the analytical forms of distributions, however, their application lies, as said, in pricing, or in simulation.

Continuous time processes are first described as elementary, unobservable changes and only after aggregating elementary changes it is possible to obtain distribution functions.

20.1 Brownian motion and the Itô lemma

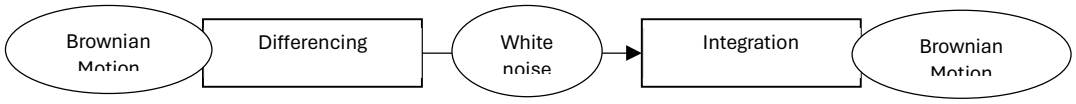
"Brownian" motion is the most important building block for continuous-time asset pricing models. It has a long history in the modelling of random events in science. Around 1830 R. Brown, a botanist, discovered that molecules of water in a suspension perform an erratic movement under the buffeting of other water molecules. While Brown's research had no relation to mathematics this observation gave Brownian motion its name. In 1900 Bachelier introduced Brownian motion as model for stock prices and in 1905 Einstein proposed Brownian motion as a model to describe the movement of particles in a suspension. The first rigorous theory of Brownian motion is due to Norbert Wiener (1923) and is often referred to as the "Wiener" process.

Random time series are called "stochastic" processes. A stochastic process $X_t, t \geq 0$ is a standard one-dimensional Brownian motion if the following conditions are met:

1. $X_0 = 0$
2. X has independent "increments": for all $t, u \geq 0$, the increment $X_{t+u} - X_t$ is independent of X_s for all $s \leq t$.
3. X has stationary, normally distributed increments: $X_{t+u} - X_t \sim N(0, u)$.
4. X has continuous sample paths.

Without condition (4), Brownian motion is just the usual random walk.

This definition is based on the fact that increments (differences) of Brownian motion inevitably is white noise and conversely: when white noise is integrated, that is, when observations are accumulated sequentially, the resulting time series is Brownian motion. Integration is indeed the reverse of differencing and this is why the presented definition of Brownian motion is, in fact, a statement about differenced white noise.



Both the random walk and Brownian motion are examples of “Markov” processes where, to predict the future, only the present observation is of interest. Past information is incorporated into current information, as is the case of prices in efficient markets. Note, however, the following difference:

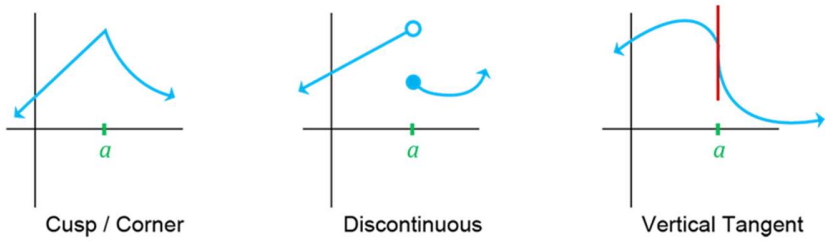
In the case of the random walk and other discrete variable, individual numbers emerge at random.

In the case of Brownian motion, it is the whole “trajectory” (continuous curve) that emerges at random.

What makes Brownian motion remarkable and indeed extremely useful is the fact that it is a continuum in time, that is, it is uninterrupted, without any breaches of continuity. However, due to its stochastic nature, Brownian motion trajectories are full of peaks being nowhere differentiable, at least in the classical Calculus sense. In Calculus,

a function is “continuous” if it can be drawn without picking up the pencil, that is, if there is no break in the graph of the function at any point; and

any “differentiable” function must be continuous, but the converse does not hold: a continuous function may not be differentiable, for example, if the function has cusps, peaks (left), or a vertical tangent (right) as shown below.



It can be proved that Brownian motion can exist: Brownian motion is not a preposition, continuous-time processes like this are not impossible.

The fact that the Brownian motion is continuous in time is important because, in this way, a specific type of Calculus can be used to allow its differentiation and integration. This specific Calculus is called Stochastic Calculus and is based on the Itô formulas.

No discrete process can be treated as a continuous function, and thus discrete processes can neither be differentiated nor integrated. It is this possibility of differentiating and integrating the Brownian motion and some other continuous-time processes that make it possible to obtain pricing formulas for many assets. In other words, continuous time has opened up a new avenue for the pricing of assets, thence its importance.

-/-

This chapter describes Brownian motion and briefly introduces “Stochastic Calculus”, which is the specific type of Calculus applicable to these processes, as a way to obtain “closed forms” (usable formulas) from elementary mechanisms.

Consider variable X, not yet observed because it is being formed over time from many, elementary increments dx. Observed prices in markets, for example, are formed from elementary bid and ask orders, the current, observable wealth of an individual is formed from many elementary receipts and payments, and the numbers that are published in accounting reports are formed from many, elementary transactions that occur during a given period.

Also consider very small, contiguous periods of time dt, each of them associated with an elementary change in *x*, which we call dx, the increment of *x*. Now write

$$dx = \mu\, dt + \sigma\, dz$$

where μ (the "drift") and σ (the "diffusion" coefficient) follow some trajectory in time, and dz is Normal with zero mean and variance equal to dt.

When μ and σ are constant over the interval t from 0 to t, then it is easy to add (accumulate, integrate) all the elementary *dx* along the elementary dt into observed *X*:

$$X_t = X_0 + \mu t + \sigma Z_t$$

X_0 is the initial ($t = 0$) value of *X*. Moreover,

$$X_t \sim N[X_0 + \mu t, \sigma^2 t]$$

Z_t, a Wiener process with zero mean and variance t. X_t is a Brownian motion with drift.

-/-

When μ and σ, rather than being constant, evolve over time, then it must be possible to deduce the observed X_t from elementary mechanisms as above. This is what the "Itô" lemma does. The Itô formula performs the same role as the chain rule for finding the derivative of composite functions in Calculus. Indeed, the Itô lemma allows calculating partial derivatives of stochastic processes.

We begin, considering as before the elementary process of a Brownian motion with drift

$$dx = \mu\, dt + \sigma\, dz$$

Let the function $f(t, x)$ be given and define the stochastic process Z_t by $Z_t = f(t, X_t)$.

What does $df(t, X_t)$ look like? The answer is given by the Itô formula, which stems from making a Taylor expansion of $f(t, X_t)$, including the second order term:

$$df(t, X_t) = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dX_t + \frac{1}{2}\left(\frac{\partial^2 f}{\partial x^2}\right) dX_t^2$$

Now we plug into dX_t the elementary process which we are interested in. For example, we may be interested in using the simple Brownian motion. In that case,

$$df(t, X_t) = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} (\mu\, dt + \sigma\, dz) + \frac{1}{2}\left(\frac{\partial^2 f}{\partial x^2}\right) (\mu\, dt + \sigma\, dz)^2$$

The three remarkable facts that make Stochastic Calculus possible, stem from the fact that, when dt tends to zero, elementary changes dt^2, $dZ_t dt$, and $(dZ_t)^2$ above exhibit different speeds of convergence to zero so that, for all purposes, we can say that

$$dt^2 = 0$$
$$dZ_t dt = 0$$
$$(dZ_t)^2 = dt$$

This fact greatly simplifies the above expression. We get rid of terms that are zero, and then integrate. For the above Brownian motion, this leads to its closed form for X_t.

20.2 Geometric Brownian motion

The Brownian motion is not a good approximation of financial processes.

As with economic phenomena such as income, wealth, the size of firms, and market prices, financial numbers cannot be described as resulting from the kind of additive stochastic process that underlies

normal variables because they stem from money accumulation, not from the added effect of many, independent random influences.

Therefore, the amount that is reported in a financial statement item, or a market price, is generated by a multiplicative rather than an additive law of probabilities. While each transaction adding to the amount reported as, say, Total Sales can be modelled as a random event, the transaction contributes to the reported aggregate not in a manner that could lead to either an increase or decrease in Sales, but by accumulation (increase) only. Accumulations are multiplicative as opposed to additive because their likelihood is conditional on the occurrence of a chain of prior events and stems, therefore, from the multiplication rather than from the addition of the individual probabilities.

Since early, economists have noticed that those processes where the multiplications of probabilities play a major role, lead to peculiar type of randomness where observations, rather than gathering around an expected value plus or minus two standard deviations, may take on values from the extremely small to the extremely large. Differences in, say the income of individuals, can vary from a few thousand dollars to billions annually.

The "Gibrat" law, known as the law of "proportionate effect", states that, in processes governed by the multiplicative law of probabilities, it is the rate of change of a variable, not the variable itself, which is Normal. Given a random multiplicative variable *x* and its change *dx*, the Gibrat law says that

$$\frac{dx}{x} \sim N(m, s)$$

It is easy to see that the distribution of *x* is "lognormal": the logarithm of *x* is Normal.

Multiplicative variables tend to be lognormal, rather than normal. They cannot be treated as distortions of normality. No distorting mechanism would be able to create, from additive events, the wide range of values found in multiplicative variables. For instance, the larger values observed in a lognormal sample are likely to be many hundreds of times bigger than the smaller ones. Such extreme proportions have no counterpart in additive variables where the likelihood of observations two or three standard deviations above or below the mean is very small.

When the multiplicative character of financial data is ignored, features of the data that would otherwise be considered as commonplace (such as positive skewness and extreme values) are likely to be seen as extraordinary. In fact, the presence of so-called outliers in empirical frequencies is most likely to be a consequence of multiplicative skewness in financial aggregates.

We now describe two Brownian motions known as "multiplicative" or "geometric" Brownian motions, where relative changes, not changes themselves are Normal. First, we introduce the elementary process used by Fisher Black and Myron Scholes in 1973 to value options, that is,

$$\frac{dx}{x} = \mu\, dt + \sigma\, dz$$

where relative changes dx/x stem from a deterministic term $\mu\, dt$, which increases with t, plus a random term $\sigma\, dz$ with standard deviation σ. The summation of all *dt*, t, reflects the length of the period during which the generation of an observed X takes place. By writing

$$dx = x\, \mu\, dt + x\, \sigma\, dz$$

and then applying the Itô formula to the logarithm of dx, this elementary random process becomes

$$d \log x = \left(\mu - \frac{1}{2}\sigma^2\right) \mathrm{dt} + \sigma\, dz$$

which, after integrating over the t period, leads to the closed form

$$X = X_0 \exp\left(\mu t - \frac{1}{2}\sigma^2 t + \sigma Z\right)$$

where X_0 is an arbitrary constant level, as above, and Z is the Wiener processes, *i.e.* the time-transformed Brownian motion with independent, normally-distributed increments with zero mean, present in Markov chains.

If the X are market prices, then we can write the market "log returns" as

$$\log\frac{X}{X_0} = \mu t - \frac{1}{2}\sigma^2 t + \sigma Z$$

From here, by taking expectations, we can estimate the mean of log returns of specific securities. In general,

$$E\left(log\frac{X}{X_0}\right) = \left(\mu - \frac{1}{2}\sigma^2\right)t$$

Therefore, mean returns will increase with the time elapsed and with the variance. When using this formulation in practice, attention should be paid to the way changes in time are modelled: the closed form above contemplates time, not changes in time.

20.3 Continuous compounding Brownian motion

The above process is based on the assumption that changes in X are essentially discrete, so that, when expressed in relative (proportionate) terms dx/x, they obey a random walk process. If, instead of relative changes dx/x we consider that dt is proportional to continuously compounded increments $dlog\ x$, we have the elementary mechanism

$$dlog\ x = \mu\ dt + \sigma\ dz$$

where continuously compounding increments $dlog\ x$ stem from a deterministic term $\mu\ dt$, which is constant, plus a random term $\sigma\ dz$ with standard deviation σ. The summation of all dt, t, reflects the length of the period during which the generation of an observed X takes place. If we now use the Itô lemma to calculate the exponential of $dlog\ x$, this elementary process can be written,

$$\frac{dx}{x} = \left(\mu + \frac{\sigma^2}{2}\right)dt + \sigma\ dz$$

Which, after integration, yields the closed form

$$X = X_0 \exp(\mu t + \sigma Z)$$

where X_0 is an arbitrary constant level, as above, and Z is the Wiener processes, *i.e.* the time-transformed Brownian motion with independent, normally-distributed increments that is present in Markov chains.

The most remarkable difference in relation to the process used by Black and Scholes is that, in the current process, the drift is not influenced by the term

$$\frac{1}{2}\sigma^2 t$$

and, therefore, is not influenced by the variance. Indeed, the drift is constant over time.

This process, which is another Geometric Brownian motion, has the vital quality that ratios of variables generated thus, remove any component of μ that is present in both numerator and denominator. In the case of the Y/X ratio, for instance,

$$\frac{Y}{X} = \frac{Y_0}{X_0} e^z$$

Therefore, the ratio will not drift. The term z is also a Wiener process with variance

$$(\sigma_y^2 + \sigma_x^2 - 2\rho_{yx}\sigma_y\sigma_x)t$$

ρ being the correlation coefficient between z_y and z_x.

In light of the development just presented, it is now possible to discuss the validity of financial ratio usage and offer guidelines to obtain normally distributed ratios.

As mentioned, when data are additive, distributions are preserved if variables are added or subtracted. In multiplicative data, distributions are preserved when variables are multiplied or divided. The simplest additive formulation is $x = \mu + z$, where x is equal to an expected value, μ, plus a random deviation, z. The multiplicative equivalent would be $x = x_0 f$, where a realisation of x is explained by the pre-specified level, x_0, multiplied by a random factor f specific to each case.

When an additive variable x is explained not only by an expected value, μ, but also by d_j, an extra component of the variance of x, then $x_j = \mu + d_j + z$, where d_j is the expected deviation from μ introduced by the j^{th} level of d. If the same d_j is present in two variables, y and x, it is possible to remove it by subtracting variables. For instance, when a medical trial is carried out in the same group of patients both before and after treatment, the difference between y and x measure the effect of the treatment and is free from spurious influences such as those of the sex of the patient because such factors, being present in both observations, cancel out when subtracted.

The financial ratio is the multiplicative equivalent of this example where firm size is the 'spurious' influence to be removed from the ratio components y and x. Wherever this influence of size is not totally removed by a ratio, econometric models using that ratio as predictors are biased because an endogenous effect is present.

Based on the reasoning just presented and bearing in mind that accounting identities such as Assets – Liabilities = Owner's Equity may constrain some ratios, that is, for example, the numerator of the ratio may be an upper limit to the denominator, we now present the list of transformations that should be used in statistical models that include financial ratios as predictors or predicted variables.

Original ratio form	Apply transformation to	Original boundaries	Examples	
$\frac{Y}{X}$	r	0, ∞	Current Ratio	Current Assets / Current Liabilities
$\frac{X-Y}{X}$	$1-r$	-∞, 1	Sales Margin	Sales - Costs / Sales
$\frac{Y-X}{X}$	$1+r$	-1, ∞	Chg. in Capital Employed	Closing Capital - Opening Capital / Opening Capital
$\frac{Y+X}{X}$	$r-1$	1, ∞	Interest Cover	Earnings + Interest Paid / Earnings
$\frac{X}{Y+X}$	$\frac{1}{r}-1$	0, 1	Liabilities Ratio	Liabilities / Equity + Liabilities
$\frac{X}{Y-X}$	$1+\frac{1}{r}$	0, ∞	Leverage Ratio	Liabilities / Total Capital - Liabilities

After transformed, ratios become well behaved and no longer generate influential cases.

20.4 Stochastic volatility

The most basic attempt to model volatility consists of viewing σ^2 as constant. In case it is not tenable that volatility σ^2 is constant, the following step consists of assuming that volatility is a deterministic function of the stock price and time. This is called the "local volatility" model and it encompasses GARCH. In fact, GARCH is an attempt to model returns' volatility where σ^2 is viewed as a deterministic time-varying process similar to those covered by ARMA procedures.

If volatility is random and is described by a σ^2 equation driven by a specific type of Brownian motion, then the model is called a "stochastic volatility" model. A whole area of Finance is devoted to the study of stochastic volatility in relation to option pricing and other key topics.

Suppose that asset prices are driven by the Black-Scholes geometric Brownian motion

$$\frac{dx}{x} = \mu\, dt + \sigma\, dz$$

so that the closed form will be

$$X = X_0 \exp\left(\mu t - \frac{1}{2}\sigma^2 t + \sigma Z\right)$$

as explained above. The maximum likelihood estimator of a constant volatility for given stock prices at different periods in time has an expected value of

$$E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$$

n being the number of cases. This constant volatility model of σ is the starting point for non-stochastic volatility models such Black–Scholes and Cox–Ross–Rubinstein models. Stochastic volatility models replace the constant σ by function ν_t of the variance of X_t.

This function ν_t is also Brownian motion, and the form of ν_t depends on the particular assumptions made. In the Heston model, for example, variance is a stochastic process with tendency to revert towards a long-term mean at a given rate, being proportional to the square root of its level, and whose source of randomness is correlated with the randomness of the underlying price.

The Heston model is different from GARCH in that it uses the square root of variance.

The whole of stochastic volatility modelling rests on the extensive use of Monte Carlo simulation and a type of Bayesian reasoning not covered by this course.

Chapter 21 Financial simulation

"Monte Carlo" methods aim at finding the output of intractable processes using random simulation instead of analytical prowess. Monte Carlo methods use extensive computing power to randomly replicate the inputs to a process and then observing its output.

For example, a spreadsheet contains an articulate fishing project in Africa with the capital expenditure duration of 5 years, including financial statements, economic functions, and other models, all of them conditional on a set of inputs such as the prime interest rate, the demand, the inflation of labor, supplies, and fuel, the expected sales price, market share and so on. The desired output is the net present value of the project (NPV), plus some auxiliary liquidity and solvency ratios, and cash-flows along the period.

It would be extremely difficult to find out, analytically, the project's most likely NPV and the other outputs, plus the variance (risks) of those outputs in terms of the available inputs. But it is quite feasible to posit the distribution of the various inputs and their correlation. Therefore, Monte Carlo methods can "simulate" many individual scenarios obeying the input's distributions and correlations, thus being able to obtain a distribution of the future NPV of the project, together with distributions of ratios and cash-flows during the period. These distributions allow determining the likely values and their confidence intervals.

In the pricing of derivatives and in other applied models, Monte Carlo methods are also used to sample at random from time series, to obtain a desired distribution.

-/-

In order to operate the Monte Carlo method, we need to know how to

- generate, at random, the distributions that will be needed as inputs,
- introduce correlations and other relationships among such distributions,
- build the project in a spreadsheet.

Spreadsheets are equipped with generators of random numbers distributed uniformly over the zero to 1 interval. From these, we can obtain

- Random numbers obeying distributions where the inverse of accumulated distribution functions is available. For example, normally distributed random numbers can be obtained by calculating the inverse of the accumulated Normal function of uniformly distributed numbers between zero and one (in Excel this would be **`=norm.inv(rand())`**.
- Discrete random numbers assuming, for instance, the value zero with probability 0.9 and 1 with probability 0.1 are generated using truncating functions and if-else functions, for example, **`=if((truncate(10*rand())),>8,1,0)`**.
- It is also easy, using the same procedures, to generate from uniformly distributed numbers, other numbers that assume integer values between, say, 3 and 9.
- Correlated random numbers are made available through the Cholesky matrix as explained below.

The spreadsheet file "BPLAN" exemplifies a few among these procedures.

-/-

Now that we know how to generate the most widely used distributions, we show how to introduce correlations among the generated numbers.

Given a matrix *X* of independent random numbers with *N* rows (objects) by *k* columns (attributes), and given *L*, the Cholesky decomposition of *A*, a desired k by *k* variance-covariance matrix, then *X L* is the *N* by *k* matrix containing correlated, random numbers obeying *A*, the desired *k* by *k* variance-covariance matrix. In other words, it is possible to generate *k* correlated variables from *k* uncorrelated, random variables by multiplying the latter by the Cholesky decomposition of the respective variance-covariance matrix.

When $k = 3$, the Cholesky decomposed of the variance-covariance matrix *A* can be computed from *A* and from previously computed *L*:

$$\begin{pmatrix} \sqrt{A_{(11)}} & 0 & 0 \\ A_{(21)}/L_{(11)} & \sqrt{A_{(22)} - L_{(21)}^2} & 0 \\ A_{(31)}/L_{(11)} & \left(A_{(32)} - L_{(31)}L_{(21)}\right)/L_{(22)} & \sqrt{A_{(33)} - L_{(31)}^2 - L_{(32)}^2} \end{pmatrix}$$

For $k > 3$, the following formulas can be used to obtain the entries of *L*, provided that calculations are performed from top to bottom and left to right:

$$L_{kk} = \sqrt{A_{kk} - \sum_{j=1}^{k-1} L_{kj}^2}$$

$$L_{ik} = \frac{1}{L_{kk}}\left(A_{ik} - \sum_{j=1}^{k-1} L_{ij}\, L_{kj}\right), i > k$$

The same BPLAN example shows a 2 by 2 case.

Chapter 22 Simultaneous equations

This chapter introduces two popular types of simultaneous equations models, namely simple "structural" equations and the more general case of "system" of equations. These two instances are a necessarily limited view of simultaneity in Econometrics, but they will hopefully open the door to further exploring in the part of students.

22.1 Structural equations

Until this point, we have assumed that our models have only one predicted variable, *Y*, and any number of predictors, *X*:

$$Y = a + b_1X_1 + b_2X_2 + \ldots$$

We have further assumed that $X_1, X_2 \ldots$ cause *Y*. In actual fact, causation is difficult to demonstrate empirically but theory often makes the direction of causation fairly clear. For instance, in the sentence "your income at age 30 is partly determined by your gender", the causation is unlikely to run the other way because, if your income can indeed change, the same is not true of your gender, which is unlikely to change. Or if "your income is partly determined by your age", the causation is unlikely to run the other way because, again, your income may shrink but your age will not shrink.

There are cases, however, where there is room to assume a reversed causation, namely when one of the predictors actually is affected by the explained (dependent) variable. In other cases, two or more variables are "simultaneously" determined by each other and by the other variables in the model.

In previous units, we showed how the method of instrumental variables can solve two kinds of endogeneity problems: omitted variables and measurement error. In both cases, we could estimate the parameters of interest by OLS if we could collect better data.

Here, we study "simultaneity", which is another form of endogeneity of predictors. It arises if one or more of the predictors is jointly determined with the dependent variable.

In a "system of structural equations", for example,

$$Y_1 = b_0 + b_1X_1 + b_2X_2 + b_3Y_2$$
$$Y_2 = c_0 + c_1X_1 + c_2X_3 + c_3Y_1$$

Y_1, Y_2 are dependent "endogenous" (i. e. determined within the model) variables.

X_1, X_2, X_3 are independent "exogenous" variables (i. e. from outside the model).

This model is a "multiple equation system". It is easy to imagine cases where structural equations are required because the mutual influence between some variables is evident. Take, for example, the system of two equations where loyalty is explained by tenure, age and sex, while tenure is explained by loyalty, age and income:

$$loyalty = b_0 + b_1\, age + b_2\, sex + b_3\, tenure$$
$$tenure = c_0 + c_1\, age + c_2\, income + c_3\, loyalty$$

Here, loyalty, age and tenure are present in both equations, but sex and income are present in one equation only.

Let us suppose that we are interested in assessing the effect of sex on tenure or loyalty. If we try to run a regression on the first model without taking any account of the second, the coefficients we get from that regression would be a mixture of other coefficients in both models.

This can be verified by solving the second equation and then plugging the tenure results into the tenure variable in the first equation and solving, thus obtaining a "reduced form" for loyalty, containing exogenous predictors only.

It would be impossible to correctly identify the results of that regression, and this is why the sole use of such reduced form raises what is known as the "identification problem": results do not allow estimating the parameter we are seeking to estimate, for example, the impact of income on employee loyalty.

The requirements for the parameters of a system of two equations to be "identified" are:

- The "rank condition": for the first equation to be identified, the second equation contains at least one exogenous variable that is not present in the first equation.
- The "order condition": at least one exogenous variable is excluded from the first equation.

Ideally, we want our equation to be exactly identified. Often, we are interested in one of the two equations and it does not matter if the other equation is not exactly identified.

If G is the number of endogenous variables considered in all equations and K is the number of variables (endogenous and exogenous) that are missing from an equation,

If $K = G - 1$, the equation is exactly identified,

If $K > G - 1$, the equation is over-identified,

If $K < G - 1$, the equation is under-identified.

In the above example,

$$Y_1 = b_0 + b_1 X_1 + b_2 X_2 + b_3 Y_2$$
$$Y_2 = c_0 + c_1 X_1 + c_2 X_3 + c_3 Y_1$$

X_3 is missing from the first equation and X_2 is missing from the second equation. This is what allows the system to be identified:

For the first equation, $K = 1 = G - 1$ and the equation is exactly identified.

For the second equation $K = 1 = G - 1$ and the equation is exactly identified

Since equations are exactly identified, it is possible to estimate any of the coefficients simply by using OLS in two stages (2SLS). The two stages are:

Stage 1: Estimate the reduced form equations by OLS and obtain the predicted values for the endogenous variables.

Stage 2: Replace the right-hand-side endogenous variables with these predicted values and estimate the equation via OLS.

In the given example we would run an OLS regression explaining *tenure* in term of all the exogenous variables available. Note that *loyalty* is not present in this reduced form.

$$tenure = d_0 + d_1\, age + d_2\, sex + d_3\, income$$

We would save the predicted tenure $\hat{T}$ from this equation and then we would predict *loyalty* using the first equation where, instead of tenure, we would use predicted tenure.

$$loyalty = b_0 + b_1\, age + b_2\, sex + b_3 \hat{T}$$

The coefficients from this regression should be reliable estimations of the respective predictors' effect on the dependent variable.

22.2 Example of structural equation

We wish to predict hours of work from wages, from education and from other variables,

$$hours = a + b_1 \log(wage) + b_2 educ + b_3 age + b_4 kids + \ldots$$

but we are aware that wages can be predicted from hours of work and other variables,

$$\log(wage) = A + B_1 hours + B_2 educ + B_3 exper + \ldots$$

because hours of work influences wages. We are satisfied that these two equations are plausible but are the parameters of this system of two equations identified? The answer is yes, because each equation has variables which are not present in the other equation.

We use TROZ, a truncated version of the MROZ where the cases with zero hours have been removed. This modelling problem is similar to applying 2SLS to the prediction of salaries with education as endogenous variable (Wooldridge p. 563).

When all variables are included, the system of two equations is:

$$hours = a + b_1 \log(wage) + b_2 educ + b_3 age + b_4 kidslt6 + b_5 nwifeinc + e$$

$$\log(wage) = A + B_1 hours + B_2 educ + B_3 exper + B_4 expersq + e$$

The first equation above, by itself, gives the following OLS model:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1523.7748   305.5755   4.987 8.99e-07 ***
log(wage)     -2.0468    54.8801  -0.037  0.97027
educ          -6.6219    18.1163  -0.366  0.71491
age            0.5623     5.1400   0.109  0.91295
kidslt6     -328.8584   101.4573  -3.241  0.00128 **
nwifeinc      -5.9185     3.6833  -1.607  0.10884
Residual standard error: 766.6 on 422 degrees of freedom
Multiple R-squared:  0.03611,    Adjusted R-squared:  0.02469
F-statistic: 3.162 on 5 and 422 DF,  p-value: 0.008173
```

Now we perform the 2SLS. The reduced form, where *hours* is not present, is

$$\log(wage) = A + B_1 educ + B_2 age + B_3 kidslt6 + B_4 nwifeinc + B_5 exper + B_6 expersq + e$$

The first stage consists of modelling this reduced form using OLS. Results are:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4471607  0.2852028  -1.568  0.11766
educ         0.1011113  0.0149618   6.758 4.68e-11 ***
age         -0.0025561  0.0051920  -0.492  0.62275
kidslt6     -0.0532185  0.0884411  -0.602  0.54767
nwifeinc     0.0055600  0.0033104   1.680  0.09378 .
exper        0.0418643  0.0132377   3.162  0.00168 **
expersq     -0.0007625  0.0004008  -1.903  0.05779 .
Residual standard error: 0.6662 on 421 degrees of freedom
Multiple R-squared:  0.1633,     Adjusted R-squared:  0.1514
F-statistic: 13.69 on 6 and 421 DF, p-value: 3.158e-14
```

We now use $log(\widehat{wage})$, which are the fitted values obtained from the first stage, and we plug this $log(\widehat{wage})$ into the first equation above instead of *log (wage)*:

$$hours = a + b_1 log(\widehat{wage}) + b_2 educ + b_3 age + b_4 kidslt6 + b_5 nwifeinc + e$$

We now perform the second stage to find the OLS estimate of *hours* in this regression:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2225.662    310.328   7.172 3.34e-12 ***
wage.hat    1639.556    254.163   6.451 3.05e-10 ***
educ        -183.751     31.920  -5.757 1.65e-08 ***
```

```
age           -7.806      5.065  -1.541  0.12403
kidslt6     -198.154     98.802  -2.006  0.04554 *
nwifeinc     -10.170      3.573  -2.846  0.00464 **
Residual standard error: 731.4 on 422 degrees of freedom
Multiple R-squared:  0.1226,     Adjusted R-squared:  0.1122
F-statistic:  11.8 on 5 and 422 DF p-value: 1.081e-10
```

It is worthwhile comparing these results to those obtained when the first equation was estimated by itself. It turns out that, with 2SLS, *log (wages)* is highly significant, which shows how endogenous wage is.

Instead of performing the two stages separately, it is possible to use algorithms that make 2SLS, such as the "ivreg" command in r (package AER). In that case, however, a clear separation between these types of variable is required beforehand:

- Purely endogenous variables, in this case it is *log (wage)*
- Purely exogenous variables,
- Both exogenous and instrumental variables,
- Instrumental variables only

In the R, such separation is made using the "|" delimiter. In the present instance, the command would be (note that *hours* is the predicted variable):

```
Library (AER)

form <- (hours ~ log (wage) + educ + age + kidslt6 + nwifeinc |
           educ + age + kidslt6 + nwifeinc + exper + expersq)

step <- ivreg (form, data = mroz)

summary (step, vcov = sandwich, diagnostics = TRUE)
```

Results of using these command lines in TROZ would be:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2225.662    574.564   3.874 0.000124 ***
log(wage)   1639.556    470.576   3.484 0.000545 ***
educ        -183.751     59.100  -3.109 0.002003 **
age           -7.806      9.378  -0.832 0.405664
kidslt6     -198.154    182.929  -1.083 0.279325
nwifeinc     -10.170      6.615  -1.537 0.124942
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Residual standard error: 1354 on 422 degrees of freedom
Multiple R-Squared: -2.008,      Adjusted R-squared: -2.043
Wald test: 3.441 on 5 and 422 DF,  p-value: 0.004648
```

which are equal to those obtained previously.

The leading method for estimating simultaneous equations models is the method of instrumental variables and, as we have just illustrated, the solution to the simultaneity problem is basically the same as the solution to the omitted variables and measurement error problems.

22.3 System of equations

As seen, econometric models may include more than one equation. They become a "system" of simultaneous equations when they express a unique reality, for example, 4 equations predicting 4 variables from 4 other variables, all of them correlated.

$$y_1 = a_1 + b_{11}x_1 + b_{21}\,x_2 + b_{31}x_3 + b_{41}x_4 + e_1$$
$$y_2 = a_2 + b_{12}\,x_1 + b_{22}\,x_2 + b_{32}\,x_3 + b_{42}\,x_4 + e_2$$
$$y_3 = a_3 + b_{13}\,x_1 + b_{23}\,x_2 + b_{33}\,x_3 + b_{43}\,x_4 + e_3$$
$$y_4 = a_4 + b_{14}\,x_1 + b_{24}\,x_2 + b_{34}\,x_3 + b_{44}\,x_4 + e_4$$

Some predicted variables may be predictors in other equations in the system. The error terms e_i are contemporaneously correlated because unconsidered factors that influence the error term in one equation will influence the error in other equations too. Ignoring this contemporaneous correlation and estimating these equations separately leads to inefficient estimates of the coefficients.

When OLS is applied to an equation in a simultaneous equations system, the estimate is generally biased and inconsistent but when estimating all equations simultaneously with a "generalized least squares" (GLS) estimator, which is capable of taking the covariance structure of the residuals into account, resulting estimates are efficient. This estimation procedure is called "seemingly unrelated regression" (SUR).

Another reason to estimate a system of equations simultaneously is the existence of cross-equation restrictions on coefficients. The estimation of coefficients under cross-equation restrictions and the testing of restrictions requires simultaneous estimation.

SUR models can contain variables that appear on the left-hand side in one equation and on the right-hand side of another equation, but if we ignore the endogeneity of these variables, estimates can become inconsistent. This simultaneity bias can be corrected for by applying 2SLS estimation to each equation. Combining 2SLS estimation with the SUR method results in a simultaneous estimation of the system of equations by what is known as the "three-stage least squares" (3SLS) method.

As an example, we use the IPO file containing data on initial public offerings (IPO) in the Hong Kong exchange during the period between 2003 and 2007. Our goal is to test for the existence of significant relationships, namely between the initial return of an IPO and other variables.

The methodology employed consists of building a system of 5 simultaneous equations which are seemingly unrelated, and where % variables,

initial returns (IR),

subscription rate (SR),

number of underwriters (WR_N),

reputation of underwriters (WR_RP) and

reputation of auditors (AU_RP)

are simultaneously explained by the same 5 variables plus other 3 predictors,

logarithm of the issue's size (SIZE)

offering price (OP)

H share dummy (H_DUM)

that is, 5 of the predictors in one equation are also used as predicted variables in the other 4 equations. This is the system of equations:

$$initial.return = a_1 + b_{11}\,subscript.rate + b_{21}\,sponsor.no + b_{31}\,rep.1.sponsor + b_{41}\,auditor.rp + b_{51}ln.fund + b_{61}\,offer.price + b_{71}H.share + e_1$$
$$subscript.rate = a_2 + b_{12}\,initial.return$$

$$+ b_{22}\ sponsor.no\ +\ b_{32}\ rep.1.sponsor\ +\ b_{42}\ auditor.rp$$
$$+ b_{52}\ ln.fund\ +\ b_{62}\ offer.price\ +\ b_{72}\ H.share + e_2$$
$$sponsor.no\ =\ a_3 + b_{13}\ initial.return\ +\ b_{23}\ subscript.rate$$
$$+ b_{33}\ rep.1.sponsor\ +\ b_{43}\ auditor.rp$$
$$+ b_{53}\ ln.fund\ +\ b_{63}\ offer.price\ +\ b_{73}\ H.share + e_3$$
$$rep.1.sponsor\ =\ a_4 + b_{14}\ initial.return\ +\ b_{24}\ subscript.rate$$
$$+ b_{34}\ sponsor.no\ +\ b_{44}\ auditor.rp$$
$$+ b_{54}\ ln.fund\ +\ b_{64}\ offer.price\ +\ b_{74}\ H.share + e_4$$
$$auditor.rp\ =\ a_5 + b_{15}\ initial.return\ + b_{25}\ subscript.rate$$
$$+ b_{35}\ sponsor.no\ +\ b_{45}\ rep.1.sponsor$$
$$+ b_{55}\ ln.fund\ +\ b_{65}\ offer.price\ +\ b_{75}\ H.share + e_5$$

Simultaneity is required firstly because of the reverse causality that may exist, namely between IPO initial returns, the rate of subscription, the offered price, underwriters' number, reputation, or auditors' reputation. Indeed, it may be argued that expectations on future IPO returns are likely to influence the choice of underwriters and their number, the choice of auditors and the offered price. The same reasoning applies to expectations on the rate of subscription.

Simultaneity is also required because, importantly, the influence of variables such as OP, WR_N, WR_RP and AU_RP on initial returns may not be directly observable. Indeed, such influence may be entirely absorbed into the rate of subscription, SR, which is known by investors in advance of the offer day. If this were so, then a pooled regression explaining initial returns in terms of SR and the other explanatory variables would be unable to unveil any direct link between such initial returns and information which investors, in fact, use to place their orders. This would be so, notwithstanding the fact that the influence of explanatory variables on initial returns may be significant.

One possible remedy might be to build a structural model contemplating two predicted variables SR and IR but this would prevent the comparison with previously employed methodologies while not addressing all causality issues. The choice of a system of simultaneous equations is thus justified.

Besides the system of 5 simultaneous equations, other 5 pooled OLS regressions are run, similarly explaining IR, SR, WR_N, WR_RP and AU_RP in terms of all the other variables. In this way, the two methodologies' results can be compared.

Below, we compare coefficients and overall model significance of pooled regressions and the system of seemingly unrelated simultaneous equations.

Results confirm that, while pooled OLS regressions explain initial returns (IR) solely in terms of subscription rate (SR), all other independent variables being non-significant, the system of 5 simultaneous equations shows that, besides subscription rate, the number of underwrites (WR_N) and their reputation (WR_RP), significantly explain initial returns. Offer price (OP) is near significance at 0.06.

The signs of coefficients show that underwriter reputation tends to increase initial returns while both the number of sponsors and offer price reduces it in average. For the remaining 4 equations explaining SR, WR_N, WR_RP and AU_RP, results are similar in pooled and in simultaneous equations.

WR_N emerges as a separate underwriter's feature with specific characteristics. By contrast, WR_RP and AU_RP share most characteristics.

The signs of coefficients are consistent across the different roles assigned to variables.

Depend. Var:	IR		SR		WR_N		WR_RP		AU_RP	
Model:	Pooled	Simultaneous	Pooled	Simultaneous.	Pooled	Simultaneous	Pooled	Simultaneous	Pooled	Simultaneous

IR			2.880***	4.986***	-0.094	-0.175*	0.099	0.191*	-0.046	-0.089
SR	0.050 ***	0.086***			-0.005	-0.007	0.003	0.010	-0.014	-0.022*
SIZE	0.027	0.012	0.463***	0.410**	0.096***	0.140***	-0.089**	-0.222***	0.240***	0.225***
OP	-0.036	-0.042.	0.147	0.213	-0.120***	-0.102**	0.024	0.043	0.046	-0.002
WR_N	-0.049	-0.091*	-0.142	-0.206			0.538***	0.853***	-0.266***	-0.581***
WR_RP	0.037	0.072*	0.069	0.215	0.388***	0.614***			0.679***	0.868***
AU_RP	-0.020	-0.039	-0.353	-0.558*	-0.224***	-0.488***	0.793***	1.013***		
H_DUM	0.012	0.023	-0.135	-0.116	0.109 .	0.113*	-0.044	-0.110	0.070	0.087
Constant	-0.442	-0.140	-5.996*	-4.709	-0.220	-0.344	-1.760**	0.774	-2.457***	-1.365**
Adj. R-Sq.	0.190	0.109	0.272	0.204	0.344	0.274	0.798	0.755	0.870	0.853
F stat.	11.570***		18.090***		24.920***		182.000***		306.600***	

Initial Returns (IS), Subscription Rate (SR), Number of Underwriters (WR_N), Underwriter Reputation (WR_RP) and Auditor Reputation (AU_RP), plus IPO size (SIZE), IPO Offer Price (OP) and Type of Share on Offer (H_DUM). Initial returns are the difference between the closing price on the first trading day and offering price divided by the offering price.

"***" depicts significance at the 0.1% level, "**" at the 1% level, "*" at the 5% level and "." depicts near-significance below the 10% level.

Results show that the influence of reputation and number of sponsors plus offer price on initial returns is hidden from direct scrutiny because the explanatory power of these three variables is absorbed into the previously known rate of subscription. Thus, a structural link may exist, where the rate of subscription is ex-ante significantly predicted by reputation, offer price and other variables, but then, after subscriptions are entered and the allotment results are known, the rate of subscription becomes the sole ex-ante predictor of initial returns, containing two types of variability: a large amount of variability of its own, strongly influencing IPO returns, plus variability imported from previous sources, some of which influence IPO returns. In summary, of all possible predictors, the two features of underwriters (WR_N and WR_RP) plus offer price (OP) significantly influence initial returns through the corresponding increases and decreases in SR, the rate of subscription, which is also an important source of variability of its own, able to explain initial returns.

Chapter 23 Value at risk

Banks expect that loans and other assets may yield, in due course, a positive return. In the case of loans, a positive return is achieved when amounts lent are paid back, and interest is equally paid.

However, banks face situations where assets' value decrease instead of increasing. This stems from the fact that outflows associated with assets, albeit expected to be favorable, are uncertain. Loans, for instance, may default.

It is possible to estimate the likelihood of default of a loan. As a general rule, the higher the expected return of a loan, the higher its likelihood of default. Therefore, banks which, in order to increase profitability, take higher risks in their lending, may indeed face higher losses in their investments.

This chapter is about regulatory and other measures that banks are required to put in place, in order to face likely losses without collapsing.

23.1 The distribution of losses

Risk is defined as the distribution of losses. Risk, therefore, is the range of likely losses, some more likely, others less likely. Each loss has an associated likelihood (probability).

Consider a portfolio of 1,000 similar loans of 2 million each. The likelihood of default of each individual loan is known. As a result, there is a range of likely losses in the portfolio and, to each of these loss amounts, there is an associated likelihood. Risk is thus the table showing the relation between each loss amount and the corresponding likelihood. For instance:

Loss	Likelihood over a period of 1 year
No material losses	80 chances in 100 (80%)
2 million	15 chances in 100 (15%)
20 million	3 chances in 100 (2%)
200 million	15 chances in 1,000 (1.5%)
2,000 million	5 chances in 1,000 (0.5%)

The risk of the portfolio losing 2,000 million is small: there are only 5 chances in 1,000 of such loss. Unfavorable chances faced by the bank in this case are the same as those faced by a person who chooses at random one ticket from 1,000 tickets where 5 of them lead to a penalty while the other 995 offer a reward.

Not all likelihoods have the same meaning. Likelihood may refer to

unique events (occurring only once), for instance winning a lawsuit or any event which covers the whole life of a portfolio; or to

events occurring regularly during time periods, for instance, annual events.

These two types of likelihood stem from distinct mechanisms and may lead to different loss amounts. In the example given above, likelihoods refer to the time period of 1 year.

Likelihoods associated with unique events lose interest once the event has taken place. By contrast, likelihoods associated with time periods remain in effect as they indicate the likelihood that the event may occur during such period or during coming periods. Thus the above table may also be interpreted in the following way:

- 80 out of every 100 years (4 of each 5 years) are expected to register no losses in average;
- 15 out of 100 years (3 years in 20) are expected to register a loss of 2 million;

and so on. Above, instead of using percent values to express likelihoods, the counting of years was used but the relationship didn't change. Note the use of the term expectation in relation to losses. Over several time periods, losses are said to be expected.

The above table is an example of a distribution: the relationship between values such as losses and their likelihood. Risk, therefore, is a distribution. The exact amount of future losses of a portfolio cannot be known in advance - but distribution can.

Of particular interest is the mean loss, which estimates an expectation. In the table above, the mean is obtained by multiplying losses by their corresponding likelihoods:

$$0 \times 0.8 + 2 \times 0.15 + 20 \times 0.02 + 200 \times 0.015 + 2{,}000 \times 0.005 = 13.7\ Million.$$

Thus, *13.7 Million* is the 1-year expected loss of the portfolio. When considering a large number of such portfolios (or many time periods) the observed yearly loss is near 13.7 million in average.

The tails of a distributions, that is, the smallest and largest values, are the less likely. They have particular interest for risk analysis. A distribution of gains and losses with fat (long) left-hand tail, indicates that extremely large losses are likely to occur.

Losses are often shown as positive values in spite of being negative numbers.

> In the table, and where losses are observed alone, separated from gains, they are depicted as positive values and the higher losses are those in the right-hand tail of the distribution.
>
> When losses and gains are shown together, that is, where the distribution of gains and losses, not just that of losses is shown, then losses are negative and the higher losses are in the left-hand tail.

The skewedness and kurtosis of a distribution are measured against the Normal, which is symmetrical. Zero kurtosis is that of the Normal distribution.

When distributions are Normal, 68% of cases are expected to lie within the interval defined by the mean plus and minus one standard deviation. 95% of cases are expected to lie in the interval defined by the mean plus and minus two standard deviations. 99% of cases are contained within the mean plus and minus three standard deviations.

Random events strictly obeying the Normal distribution are mostly contained within three standard deviations above and below the mean. Events above or below the third standard deviation are rare; those above or below four standard deviations are almost impossible to occur.

23.2 The components and types of risk

Risk, the distribution of losses, is often viewed as the result of a given "exposure" to one or several "risk factors":

$$Risk = Exposure \cdot Risk\ Factor$$

> Exposure is the economic value at risk. It is measured in money units and depends on will: A risk manager takes as much exposure as he/she desires.
>
> Risk factors do not depend on manager's will. They are exogenous (coming from outside) and uncertain. Risk factors, in fact, are distributions, that is, relationships between returns, yields, rates, percentages or other dimensionless multipliers and the corresponding likelihoods.

Risk stems from risk factors, not from exposures. Exposures only quantify risk.

In the case of the most common assets (stocks, bonds, commodities, bank loans and others) the analysis of risk based on the separation of the two components is appropriate and illuminating. However, there are other, not so frequent cases where any attempt to explain risk as stemming from two such components may be misleading.

The task of measuring risk consists of discovering the distribution of losses for a given asset. In order to measure risk it is thus required,

not just to know exposures and risk factors that may affect them,

but also the relationships between such exposures and such risk factors.

Relationships can be expressed either as tables, as in the example above, or as functions. Functions are also known as "analytical models" or "analytical solutions". To "model" a relationship is to find a function, that is, an analytical solution as opposed to a simple table, capable of expressing that relationship. The Black-Scholes formula or the CAPM are examples of functions, that is, analytical solutions expressing relationships.

Often, risk managers have to be content with tables because no analytical solutions are available.

We now show simple examples of the way the two components of risk, exposure and risk factors, are first isolated and then their relationship is used to determine the distribution of losses, that is, risk.

In the case of bonds, the risk of a loss is an unfavorable change in price *dP*. Now, it is possible to write

$$dP = -(Duration \,.\, Price) \cdot (change\ in\ Yield)$$

Duration times Price, the dollar duration, depends on will; it is the exposure;

change in Yield, uncertain, exogenous, dimensionless variable obeying a given distribution, is the risk factor.

In the case of stocks, a loss is also associated with *dP*, an unfavorable change in price; in this case:

$$dP = (Beta \,.\, Price) \cdot (Market\ return)$$

Beta times Price is the exposure in dollars;

Market return, a random, dimensionless distribution, is the risk factor.

In the case of options, the above separation between constant, money units of exposure and dimensionless distributions is not so clear. Changes in price *dP* can be written as:

dP = (Delta) · (change in price of underlying asset)

Delta is the first partial derivative of the option price *P*.

Change in price is exogenous, depending on Market factors; it is also a distribution.

-/-

Unfavorable changes in the value of assets, that is, losses, may be of two types:

linear when the relationship between losses and risk factors is a straight line;

nonlinear when the relationship between losses and risk factors is nonlinear.

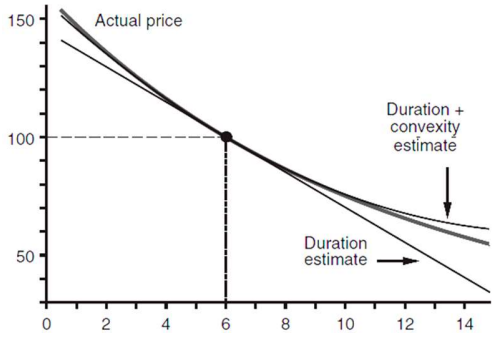
Equities are linear because the relationship between price and risk is a straight line; other assets namely bonds may be treated as linear or nonlinear depending on accuracy desired; options, swaps and other derivatives must always be treated as nonlinear.

It is easy to model risks in portfolios of linear assets such as equities. But in order to model nonlinear risks (as in derivatives) Taylor series expansion or similar devices must be used in order to deal with nonlinearity. In the case of nonlinear relationships Taylor series expansions may include:

the 1st derivative only. Thus, higher order nonlinearity is ignored. Methods which assess risk using only the 1st derivative are known as "delta" methods;

both the 1st, 2nd and maybe higher order derivatives (depending on whether non-linearity is weak or strong). These are known as the "delta-gamma" methods.

The graphic below depicts the relationship between price (Y-axis) and yield (X-axis) of a 10-year, 6 percent coupon bond. The graphic shows the true relationship and also two estimates of such relationship obtained from the 1st derivative of price (duration) and the 1st and 2nd derivatives (duration + convexity).



Traditional risk analysis separately considers three types of risk:

the risk of loan default or other material breach of credit obligations including the risk of a default during settlement. This risk is known as "credit risk" and it is by far the largest source of risk for banks.

losses from accidents, malfunction, fraud, lawsuits, relating to bank operations. It is known as "operational risk" and it is the second largest source of risk in banks.

losses caused by volatility or sudden shocks on market prices. This is known as "market risk" and it pertains to equities, FX trade, fixed income assets, commodities and other assets traded in open markets.

Market, credit and operational losses add up to the total risk a bank faces. These three types of risk are supposed not to interact, but this is not so: operational losses caused by fraud may depend on previous market losses. Market risk influences credit losses in the medium term. Presently, bank risk management techniques and regulation ignore such interactions.

Financial institutions have their own distinct patterns of risk. In the case of investment banks, for instance, market risk is larger than for retail and commercial banks. But all face an overall risk of loss which must be balanced by corresponding capital reserves.

23.3 Economic and regulatory capital

In order to cope with unexpected losses, banks and other financial institutions "reserve" (keep or put aside) capital in proportion to risks they incur. Unexpected losses cannot be predicted but the underlying risks may be assessed. Capital that a bank should keep in order to face unexpected losses is known as "economic capital".

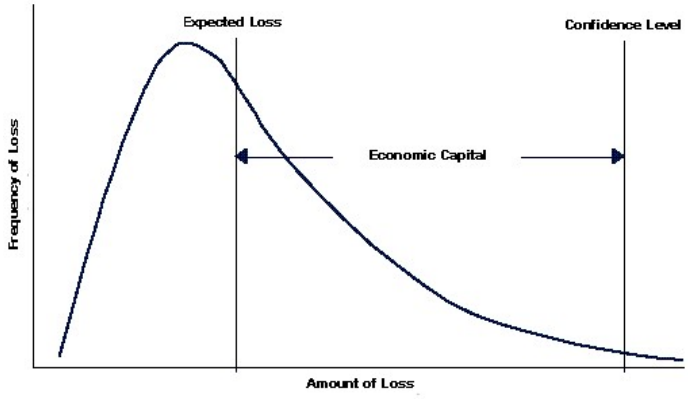
For each risky asset, banks estimate losses and the corresponding likelihoods. Based on such distributions, banks determine the probability that the loss may exceed a given magnitude.

Given a predefined, small probability, say 0.02%, banks can also determine the "maximum loss" they are exposed to with a confidence level of 1 – 0.02% = 99.98%. Such maximum loss should be balanced by a proportion of reserved capital. In this way, banks ensure solvency with the confidence level used.

Economic capital is computed using the difference between the maximum loss for a given confidence level and the expected loss. The bank's expected loss is the anticipated average loss over the period. It is one of the costs of doing business; thus such loss is part of operating expenses (it is a provision in most cases) and should not be balanced by capital. Notice that economic capital is a misnomer, being a measure of risk, not of capital held and distinct from accounting and regulatory capital measures.

The confidence level used to estimate economic capital is called "financial strength" it is the probability of the bank remaining solvent over a given time period.

Typically banks use a confidence level of between 99.96 percent and 99.98 percent which is 1 – 0.04% or 1 – 0.02% (2 in 10,000 chances) respectively. This is the same as the insolvency rate expected for AA/Aa credit ratings over one year.



The relationship between the likelihoods of losses and economic capital is depicted in the above distribution where losses are taken as positive values.

Economic capital is costly, reducing banks' profitability. Therefore, it is all important for banks to hold as much economic capital as required but not more. Regulators, on the other hand, require banks to hold capital of an amount equal (at least) to economic capital.

-/-

"Regulatory capital" is the amount of capital that bank regulators enforced by central banks according to the Basel Accords, require to cover the risks that a bank is running or collecting as a going concern, such as market risk, credit risk, and operational risk. It includes

the amount of money needed to secure survival in a worst case scenario, and

after-bankruptcy provisions known as "gone concern" capital to pay obligations once the bank is insolvent.

Thus regulatory capital is placed at a higher level than economic capital. Economic capital may differ from regulatory capital for other reasons such as diverging treatment of diversification benefits among different businesses and risks and the inclusion of economic capital estimates for risks that have no regulatory measurement, for example, business or strategic risks.

An advantage of the Basel accords enforced by central banks is that most of the largest banks are allowed to determine the amount of regulatory capital that they should keep. This is so because regulatory capital is determined using banks' own computation of economic capital. Therefore, good computations leading to accurate economic capital estimations are of strategic importance for banks.

Economic capital is extensively used to measure the risk-adjusted performance of banks and to make decisions on risk taking. Thus, the importance of its reliable estimation.

When measuring how much capital the bank should hold to face unexpected losses over a time period, there are different concepts of bank capital that can be adopted, namely:

book value of capital, the difference between the book value of assets and liabilities.

market value of capital, the difference between the value of assets and the value of liabilities when both are valued at mark-to-market prices.

market capitalization, the value of the bank on the stock market at current prices.

Since there are different concepts of capital, the bank may measure capital according to each of the different views. Yet book capital and market capitalization are usually the two key measures. Market capitalization has the best economic underpinning as it should reflect the complex evaluation the market may make about the market value of the bank's assets and liabilities, the bank's business mix and the perspective of each business the bank is in, its competitive position and other factors.

23.4 Value at risk

It was mentioned that banks are required to put aside capital in proportion to risks they take, to ensure that banks stay solvent over a certain time period with a pre-specified confidence level. Economic capital is based on "value at risk" estimates.

Value at risk was conceived as an attempt to summarize relevant information which is present in the tail (the most severe, least likely values) of a loss distribution. Risk is indeed a distribution; but distributions, either in the form of functions or tables, are not a practical way to convey information. A summary measure of risk is desirable. In order to summarize the distribution of gains and losses, classical Finance uses the variance of returns, rates or yields; but variance fails to highlight the particular characteristics of tails, thus being inadequate to assess unexpected losses. Variance is basically expected risk, leading to expected losses' estimation, not unexpected losses.

Banks use an *ad hoc* method to get a summarized risk, known as "value at risk" (VaR) and its use is widespread. When adequately estimated, VaR may help answering the banking industry concerns, which are not so much about expected but rather about unexpected losses (distribution tails).

VaR is not intended to tell the magnitude of the highest possible loss but rather to allow the management of economic and regulatory capital.

VaR is an estimation of the maximum loss for a given period of time and confidence level. For example, a VaR of 2 million within N=10 days and P=95% confidence level is the highest loss

expected over the coming 10 days with a 95% confidence level that such loss will not be larger than the specified 2 million. In other words, there are 95 chances in 100 that the largest loss to occur within the coming 10 days will not exceed 2 million.

Consider a position of US$ 4 billion short the yen, long the dollar (a bet that the yen will fall in the short term against the dollar). A confidence level of 95 chances in 100 that the loss will not exceed a maximum value is considered as sufficient here. How to estimate VaR, the maximum likely loss over a period of 1 day?

First, gather historical data to build the observed distribution of the risk factor: in the present case, several years of daily yen/dollar rates are needed.

Next, calculate daily returns from such rates and multiply by the exposure, 4 billion, to get the expected distribution of gains and losses for the coming day.

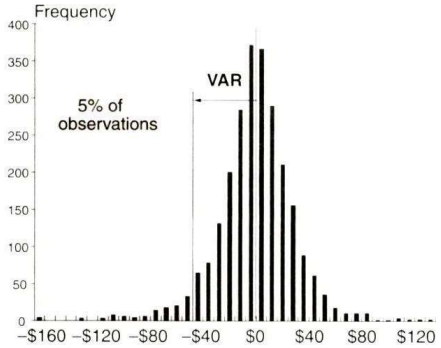
Let us suppose that expected losses are zero. But whatever be such expected loss, a 95% VaR is the portfolio value larger than 95% of losses below expected. VaR is the value which divides the distribution of losses in two so that 95% of cases exhibit smaller losses (but larger than expected) and 5% exhibit larger losses.

Such value, the 95th percentile of higher-than-expected losses is, say, 40 million. This is the VaR within N=1 day, for a confidence level P=95%.

VaR increases in proportion to the square root of time and to the confidence level:

- A 99% VaR is proportionally higher than a 95% VaR;
- a 4 days' VaR is twice that of a 1-day VaR;
- a 16 days' Var is four times that of a 1-day VaR;
- to obtain a 10 days' VaR, a 1-day VaR is multiplied by the square root of 10.

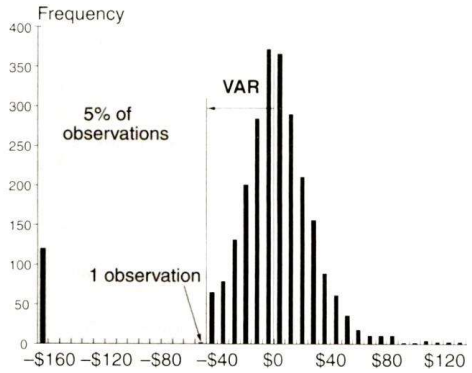
VaR is subject to a standard error which depends on the number of observations used to estimate it: Few observations mean large error, large confidence interval. e.g. a VaR which may vary between 3 million and 30 million. This is why it is so important to use as many past cases as possible to draw the distribution of gains and losses. The figure below depicts the distribution of gains and losses in the example.



Value at risk is subject to several limitations, the most relevant being:

VaR is blind to losses beyond the confidence level used. In the distribution depicted below, VaR would be misleading as there are highly likely losses to the left of such 5th percentile.

VaR is not "sub-additive", that is, the VaR of a portfolio is not smaller or equal to the added VaR of its components - and it may be higher. Therefore, individual assets' VaR cannot be used to estimate VaR of portfolios.



In order to address the two limitations above, an estimate known as CVaR, “conditional VaR” or “expected shortfall” was devised. CVaR is the average loss in the tail, to the left of VaR. CVaR is higher than VaR and better accounts for extreme events that are beyond the traditional VaR. Moreover, CVaR is sub-additive; it can be used to estimate the VaR of portfolios from components’ VaR.

The recently updated Basel accords adopt CVaR instead of VaR.

CVaR estimation requires, not just determining the value corresponding to the desired confidence level, but also the estimation of the mean loss beyond such value.

In the following, we apply the VaR definitions to the three basic types of bank risk, the market, credit and operational risk.

23.5 Market risk

Market risk is not the major source of risk for retail or commercial banks, but it is the easiest to estimate, and this is why it will be studied in the first place.

There are 4 types of market risk corresponding to the 4 asset types traded in markets: currencies, bonds, equities, and commodities. For each type, the mechanism causing risk is:

Change in the relative value of currencies or a sudden devaluation of a currency (“Forex FX / currency risk").

Change in the term structure of interest rates, in inflation, in a credit spread, in the number of liquidations anticipated and others ("fixed income risk").

Volatility or shocks in prices of equities traded in markets or in the reported profit of companies ("equity risk").

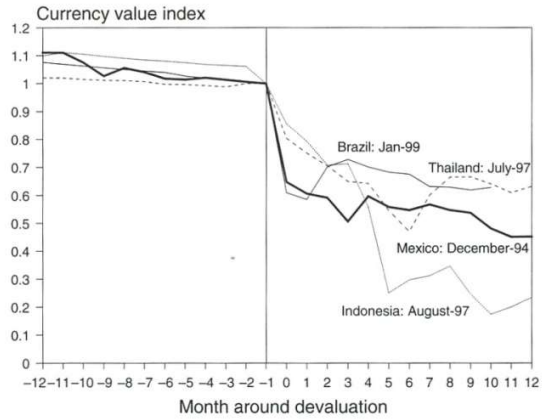
Volatility or shocks in prices of goods traded in global markets ("commodity risk").

Thus, market risk comes from two sources:

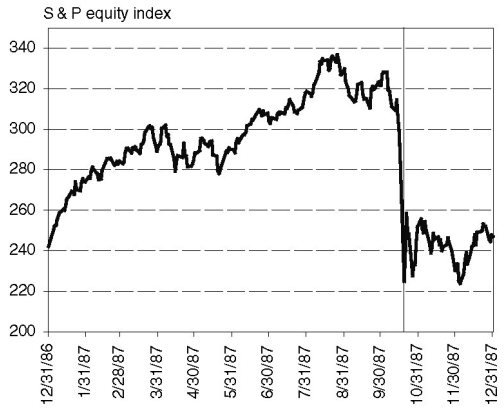
Volatility: the result of random, common price fluctuations. It is the traditional risk.

Shocks: sporadic, strong movements in prices. It is the banking industry risk.

Finance theory deals satisfactorily with exposure to volatility, the common type of risk but not so well with exposure to shocks. A risk measure able to account for extreme losses is thus needed in the banking industry. As mentioned, this need has prompted the use of VaR. Figures below depict two well-known shocks.



FX shocks suffered by four currencies when detached from the USD



An equity shock (in the price of NYSE stocks) due to a sudden panic amongst investors.

It should be stressed that VaR methods are only capable of accounting for shocks where there is some past experience regarding those shocks, so that the distribution of gains and losses incorporates such experience. Now, since shocks are rare, past experience is scarce thus rendering VaR methods fragile.

Market volatility can be partially diversified thus reducing exposure. It is also possible to use marketable securities to build effective hedged positions guarding against most types of market risk. For this reason, in banks, losses stemming from volatility are expected, thus being, in general, treated as provisions. There are, however, cases where even a common type of risk exposure may lead to severe losses. When markets become shallow, that is, with little liquidity, apparently common exposures such as hedges cannot be “undone” in time leading to uncommon losses. During a crisis, markets inevitably are shallow

Risk analysts should have an intuitive knowledge of the magnitude of volatilities. Let us thus compare some typical standard deviations. Standard deviation, the square root of variance, is preferable for comparing cases because it is the same dimension as the underlying value.

Small volatilities are typical of most currency rates. FX risk is just 6-12% per year, with large variation from one currency to the other and strong correlations amongst currencies. FX risk is also easy to offset.

Fixed-Income volatility greatly depends on maturity.

Short term debt has low volatility, lower indeed than the typical FX rate.

For maturities up to 10 years, volatility is, in average, similar to the FX volatility.

For 20 years and more, volatility can go up to 30% per year, the same as equity risk. High correlations exist amongst all fixed income securities.

Equity risk is subject to high volatility, 30% or more per year, and to the occasional shock. Equities are correlated thus most of their risk diversifies but the remaining risk, that of the market as a whole, is still significant.

Commodity Risk: the highest volatility of all. It can reach 60% per year (fuels). Correlation amongst commodities is small.

Aggregate market VaR is estimated using either “local” or “full” valuation methods.

Rather than using the whole distribution of losses, local methods use parameters of the distribution of gains and losses such as the variance. Local methods are thus available where the distribution of gains and losses can be analytically described as is the case of the Normal distribution.

Full valuation methods use the whole distribution of losses to estimate VaR, being thus independent from type of distribution. They are used where the distribution of gains and losses cannot be easily modelled.

Note that the use of local VaR methods is somehow contradictory. Indeed, if the banking industry has reverted to non-traditional risk measures it is because traditional measures, namely the variance of the distribution of gains and losses in equity portfolios, are almost blind to shocks (extreme events). Thus local VaR methods, which use variances instead of the whole distribution, are as blind to extreme events as the variance is.

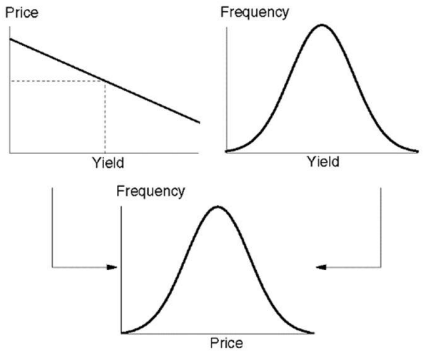
Market VaR estimation methods may also be linear or nonlinear.

Local linear valuation methods are analytically straightforward. VaR is estimated from variances and co-variances of risk factors (not from the whole distribution of gains and losses) and from a linear exposure to risk. Typically the distribution of gains and losses is Normal because risks are Normal and exposures are linear. This happens in the case of equity portfolios but also applies to FX and commodity risk.

Local nonlinear methods are used where risk factors may be Normal but gains and losses are not Normal because exposures are non-linear. This is the case of fixed income securities (bonds). As mentioned, local nonlinear methods may use either: the first derivative (Delta) or first and second derivatives (Delta Gamma) of an expansion in Taylor series of nonlinear exposures.

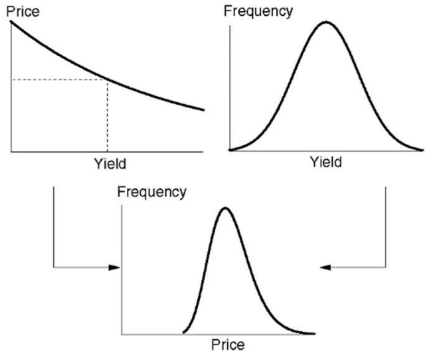
The figure below compares the way nonlinearity is introduced in the distribution of gains and losses of bonds' portfolios with the linear case, that of equities (in the left hand side).

Local linear VaR estimation method – in the case of market returns, prices are linearly related to returns and these are Normal thus the distribution of gains and losses is also Normal



Stocks' portfolio

Nonlinear local VaR method – in the case of bonds, prices are nonlinearly related to yields so that, in spite of the distribution of yields being Normal, the distribution of gains and losses is non-Normal.



Bonds' portfolio

Finally, full valuation VaR methods are used to assess the risk of derivative portfolios and also in the case of credit and operational risk as explained in due course. Risk factors are, in such cases, non-Normal while the relationship between such risk factors and exposures is also non-linear. Full valuation methods are based on simulation as well as in historical data. They are not necessarily

complex analytically and, importantly, information conveyed by tail distributions is fully used in VaR estimation.

In the case of equity, FX and commodity portfolios, market VaR is often estimated using local, linear methods: a distribution of daily returns multiplied by the square root of number of days and by the exposure (value or function) leads to the distribution of gains and losses which, in turn is supposed to be Normal. Then a simple formula using standard deviation of such distribution will lead to the appropriate VaR. An example is "RiskMetrics" from Morgan Stanley, also known as the Delta Normal method. RiskMetrics uses, as the only risk factor, the distribution of daily market returns in their logarithmic form. Such logarithmic returns are nearly Normal. Since the relationship between changes in exposure and the risk factor is linear, the resulting distribution of gains and losses is also Normal.

The distribution of gains and losses being Normal, the intrinsic characteristics of a Normal distribution apply. Thus:

VaR at 95% = 1.645 standard deviations.

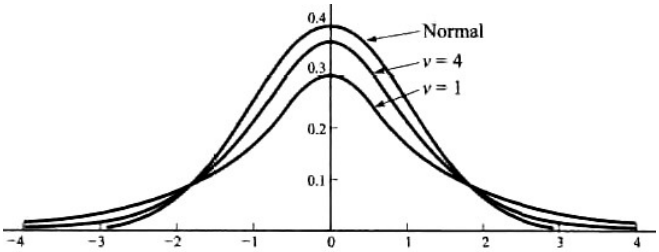
VaR at 99% = 2.326 standard deviations.

In the case of the Normal distribution, the exact relationship between the number of standard deviations below the mean (threshold) and the likelihood of an observed value turning out to be smaller than such threshold is:

Likelihood:	99.99%	99.9%	99%	97.72%	97.5%	95%	90%	84.13%	50%
Threshold:	-3.715	-3.09	-2.326	-2	-1.96	-1.645	-1.282	-1	0

Under strict Normality it is thus easy to estimate VaR from returns in the portfolio. As mentioned, this method adds nothing to traditional risk assessment.

Normality, however, is not verified exactly in returns: "tails" of distributions are fat, more in line with Student's "t" distribution than with the Normal distribution. The t distribution is similar to the Normal except that, when the number of sampled cases v is small (v < 20) it exhibits leptokurtosis becoming peaky with fatter tails than that expected in the Normal case. The smaller v is the more leptokurtic the t distribution becomes. For v > 20 there is no difference between t and the Normal; but for v = 4 or smaller the difference is noticeable:



For t distributions with small v, losses below 3 standard deviations become likely. This fact, not any theoretical underpinning, is what prompts risk analysts to use t instead of the Normal. Used v range from 4 to 8. The table below compares the Normal and Student's t tails in terms of likelihood of finding cases 1 to 5 standard deviations (or more) below the mean. This time likelihood is shown as a probability, not as a percentage

number of deviates from mean:	-5	-4	-3	-2	-1
Normal distribution	0,99999	0,99997	0,99865	0,97725	0,84134
t distribution, v = 6	0,99880	0,99644	0,98800	0,95379	0,82204

t distribution, v = 4	0,99625	0,99193	0,98003	0,94194	0,81305

For instance, in the Normal case, the probability of a case being 4 standard deviations or more below the mean is 1 – 0.99997 = 0.00003: such values may be observed, in average, 3 times in 100,000 cases. For the t distribution with v = 6 the same probability is 1 – 0.99644 = 0.00356 which means that values may be observed 356 times in 100,000.

Since VaR estimation greatly depends on the lower tail, assumptions that may change perceptions about tail probabilities will affect VaR significantly. Thus, VaR estimation using purely analytical methods (as depicted above) may not be in a sure footing.

23.6 Credit risk

As mentioned, in nonlinear cases such as derivatives, the distribution of gains and losses is estimated from exposures, distributions of risk factors and their relationships. This is the full valuation method where VaR is estimated from the whole distribution of losses. Historical data, simulated data or both, is used to draw such distribution.

Full valuation VaR methods consist of the following steps:

For each asset in the portfolio, find a function (or at least a table) able to relate gains and losses to a reduced number of risk factors. This is called "mapping".

Add all mappings so as to get a model of gains and losses for the entire portfolio.

Vary risk factors in time and magnitude (this may be carried out using simulation or historical values) observe the response of the model. Use the distribution thus obtained to get VaR.

As mentioned, in order to simplify VaR estimate, a reduced number of risk factors, not all of them, is used. For example, in the case of a bond portfolio with 25 different maturities, only 9 amongst such maturities are considered.

-/-

Credit risk is the largest source of risk for banks, much larger than other types of risk. Exposure, in this case, is the replacement value of cash-flows uncollected due to a default.

Attempts to model credit risk are recent and *ad-hoc* or arbitrary practice prevail. Moreover, data available is weak as loan ratings and financial reports of companies to which the bank lends money may be misleading. Analytical complexity is high, and this is why analysts often revert to simulation.

Credit risk includes not just the usual

loan default risk but also

settlement default such as the "contract counterparty default" occurring where, in an over the counter (OTC) derivatives contract, a bank's counterparty fails to settle.

Thus, there are two types of default:

default before settlement (this is the common type).

default at settlement amongst banks. In turn, this type of default can be:

contract counterparty default generally associated with an OTC derivative contract settlement

"Herstatt" default: during money settlement between two banks where only one of them pays, the other seeks bankruptcy protection after receiving.

Settlement risk is very short term; some may be reduced with the use of netting systems involving intermediaries, but it is of a large magnitude and may lead to dangerous situations for the whole economy, namely by raising fears amongst depositors ("systemic" risk).

Default refers to

a single vehicle or contract without affecting others; this is not common

a company or organization, in which case all the obligations of the company default

all debt obligations of a country, in which case all the obligations of that country, private as well as public, default

Traditionally, the risk of default is viewed as a function of three components:

PD, the "probability of default", of entities to which the bank has lent money, or individual vehicles. PD is estimated internally by the bank or by rating agencies

EaD, the "exposure at default" also known as CE, credit exposure. It is the economic value at risk, e.g., the notional; and it is estimated by the bank.

LgD, "loss given default" is the total estimated loss in case of default. LgD is expressed in the form of a percentage of EaD, e.g., 70%

Losses in a portfolio of bank loans depend on the quality but also on the number of loans: one unique loan of 100 million may cause a total loss of 100 million or more; while 100 loans of 10 million have a much smaller expected and unexpected loss because the likelihood that all of them default is also much smaller.

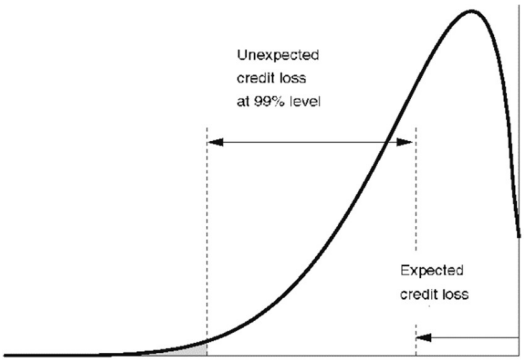
Expected losses in a portfolio (EL) can be calculated as the summation, for all loans in the portfolio, of the product of individual PD by EaD by LgD. In mathematical notation,

$$EL = \sum PD \cdot EaD \cdot LgD$$

The distribution of losses of a portfolio may be estimated from individual loans' PD, EaD and LGD using combination or simulation techniques as illustrated below. Unexpected losses (UL) may then be estimated from such distribution for a given confidence level. Credit value at risk, CVaR, will be the difference between expected and unexpected losses:

$$CVaR = EL - UL$$

The graphical description below shows a typical distribution of credit losses. In this case, losses are depicted as negative values. CVaR is the unexpected credit loss at the 99% level



Credit loss distributions are not symmetrical but rather skewed to the left (if losses are depicted as negative values as above) with a typically fat tail.

An example using all combinations of loan defaults will now illustrate how to draw a distribution of losses in a small portfolio: consider a 100 million portfolio with 3 loans, A, B and C. Exposures and probabilities of default are:

loan	exposure	probability of default
A	25m	5%
B	30m	10%
C	45m	20%

During a given period, 0, 1, 2 or 3 defaults may occur in a total of 8 possible combinations:

no defaults with probability P = (1-0.05) (1-0.1) (1-0.2) = 0.684

only A defaults with probability P = 0.05 (1-0.1) (1-0.2) = 0.036

only B defaults with probability P = 0.1 (1-0.05) (1-0.2) = 0.076

only C defaults with probability P = (1-0.05) (1-0.1) 0.2 = 0.171

both A and B default with probability P = 0.05 · 0.1 · (1-0.2) = 0.004

both A and C default with probability P = 0.05 · (1-0.1) 0.2 = 0.009

both B and C default with probability P = (1-0.05) 0.1 · 0.2 = 0.019

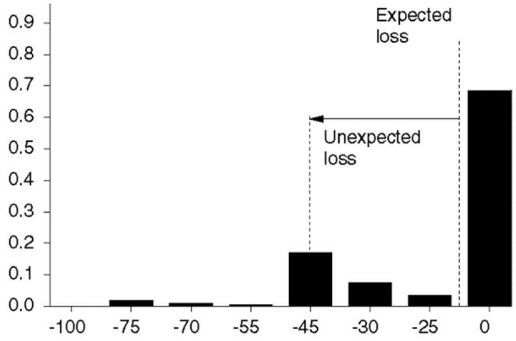
the three loans A, B and C default with probability P = 0.05 · 0.1 · 0.2 = 0.001

The 8 probabilities above add to 1 as independence between them is presumed.

The whole distribution of losses is given by the following table where each accumulated probability stems from multiplying all previous and present probabilities.

loan defaults	loss million	probability	accumulated probability	expected loss	variance
no	$0	0.6840	0.6840	0.000	120.08
A	$25	0.0360	0.7200	0.900	4.97
B	$30	0.0760	0.7960	2.280	21.32
C	$45	0.1710	0.9670	7.695	172.38
A,B	$55	0.0040	0.9710	0.220	6.97
A,C	$70	0.0090	0.9800	0.630	28.99
B,C	$75	0.0190	0.9990	1.425	72.45
A,B,C	$100	0.0010	1.0000	0.100	7.53
			total loss expected:	13.25	434.7

From this distribution CVaR would be determined as the difference between EL = 13.25 million and the nearly 5th percentile of the distribution, 45 million (accumulated probability of 96.7%) or the nearly 1st percentile, 75 million (accumulated probability of 99.9%). A graphic description of such distribution follows (losses depicted as negative values).



For larger portfolios, distributions are obtained from individual loans' PD, EaD and LgD by simulating many cases obeying random processes governed by such PDs:

First, each loan in the portfolio is assigned a random number generator which will assume the value of 1 with probability PD and zero with probability 1 – PD.

Then a good number of simulations, say 10 million simulations, calculate total portfolio loss based on individual EaD, LgD and on this random value of 0 or 1. Only where 1 turns out, will that particular loan contribute a loss to the portfolio.

The 10 million total portfolio losses are a distribution of losses from which CVaR can be computed as the n^{th} percentile.

The method is apparently straightforward, but it requires that all individual PD be estimated. Inside a given portfolio, PDs are correlated, which increases CVaR. PDs are also correlated to EaD and to LgD. Correlations make simulation and stress-testing more complex.

23.7 Probability of Default

Capability to estimate PD using a bank's own "internal" methods is what distinguishes banks possessing true risk management capabilities. Bank regulators also divide approaches made available to banks to estimate capital charges, according to whether such capabilities are or not in place. The estimation of PD being a difficult step in determining expected and unexpected losses of a loan portfolio, every means is used to make it trustworthy.

PD is estimated using two possible methods:

actuarial or

market-based methods.

Actuarial methods, in turn, are of 2 types:

External to the bank, e.g., purchased from rating agencies.

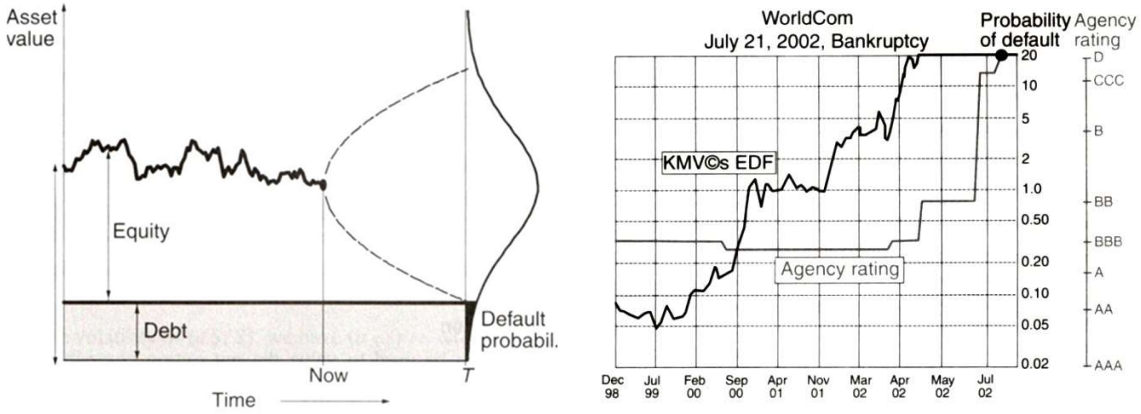
Internal to the bank: the bank estimates PD using, for instance, Z-score models.

There is not much to say about external methods except that they are expensive, often inaccurate and apply only to traded loans or large companies. Ratings, moreover, are not static. They evolve: an AAA may become BBB in just a few days. This is called "migration", an unusual name and Markov chains are used to model such changes. Paths to default and respective probabilities are also modeled from current ratings.

Internal methods are based on loans' attributes available for analysis, such as ratios from financial statements of companies. The Altman Z-score model is a well-known instance. Banks often develop

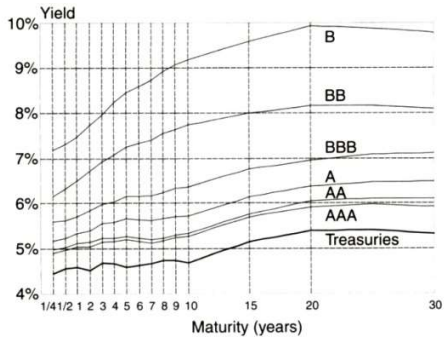
their own models based on past experience (loans' attributes and final outcome, default or not default), using tools such as "Logistic" regression where PD is the predicted variable and financial ratios are the attributes used as predictors. The advantage of Z-scores, Logistic and other methods is that they may easily estimate PDs of smaller, unlisted companies and even of individuals.

Market methods use the debt-to-equity ratio of a company (at market prices) and the Black-Scholes option pricing formula. The relationship between the market value of long-term debt and that of equity determines the probability of default because owners of companies have the option to let the company default when equity is lower than debt (the Merton model). Thus, option pricing can be used to estimate PD as depicted in the left-hand figure.



"Estimated default frequencies" (EDF) sold by KMV are based on Merton's model. They seem to be more accurate than rating agencies in predicting default as depicted in the above graphic, right-hand side.

Another instance of how PD is related to values observed in the market.is the rating of bonds: it closely follows the position of the term structure of interest rates (yield curve) relative to treasuries, as observed in the graphic below.



After PD is estimated, it is required that such values be "calibrated", that is, adjusted and corrected. Two common adjustments are

corrections for a-priori default probabilities (observed or taken from tables);

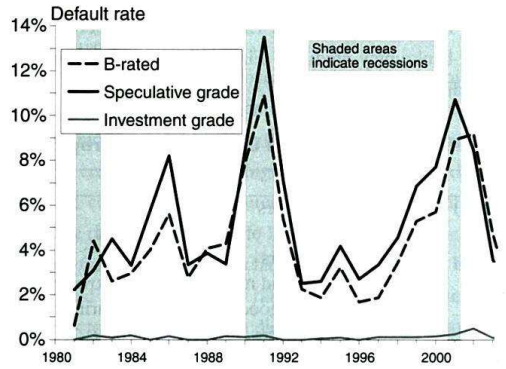
corrections for model biases.

Adjustments for prior probabilities aim at correcting PD for real-life default frequency. Indeed, most methods of estimating PD produce "conditional" probabilities which reflect frequencies present in sampled cases where default is much more frequent than in real-life.

An example of real life, prior default probabilities by rating and by the number of years considered follows, where prior probabilities are expressed as percentages.

rating	1	2	3	4	5	6	7	8	9	10
Aaa	0.00	0.00	0.02	0.09	0.19	0.29	0.41	0.59	0.78	1.02
Aa	0.07	0.22	0.36	0.54	0.85	1.21	1.60	2.01	2.37	2.78
A	0.08	0.27	0.57	0.92	1.28	1.67	2.09	2.48	2.93	3.42
Baa	0.34	0.99	1.79	2.69	3.59	4.51	5.39	6.25	7.16	7.99
Ba	1.42	3.43	5.60	7.89	10.16	12.28	14.14	15.99	17.63	19.42
B	4.79	10.31	15.59	20.14	23.99	27.12	30.00	32.36	34.37	36.10
C	14.74	23.95	30.57	35.32	38.83	41.94	44.23	46.44	48.42	50.19
overall	1.50	3.09	4.62	6.02	7.28	8.41	9.43	10.38	11.27	12.14

Overall, the probability of credit default in the coming 4 years is 6%. This figure is used in most calculations as a general prior probability. PDs are also dependent on the economic cycle. During economic downturns the probability of default increases. This should be taken into account when calibrating PDs for the present period.



The graphic above depicts default rates for different types of bonds over a period of more than 20 years. It is verified that, during recessive periods such as years 1990-1992, bond default increased significantly. In short, a loans' PD depends, not just on intrinsic quality as denoted by ratings (which is available only for large companies or large bond issues) or Z-scores, but also on economic risk. Company size, company age and business risk also play a role in determining PD. Small firms or recently established banks are much likelier to go bankrupt than larger and older firms. Some lines of business are more risky than others. These are indeed priors worth considering.

The "Bayes rule" can calibrate PDs for priors. According to Bayes rule, probabilities of default when specific states are observed in attributes stem from considering priors P (default) as well as conditional probabilities P (states given default) as follows:

$$\text{P (default given an observed state)} = \frac{P(state\ when\ default)\ P(default)}{P(state\ when\ default)\ P(default)+P(state\ when\ nodef)\ P(nodef)}$$

where "nodef" means no default. If, for instance, a 5% prior default probability is deemed as appropriate, then a conditional probability of 95% of default (obtained from loan states by means of Z-scores or other) will translate into a posterior probability of default of just 50% or less. Notice that the two conditional probabilities considered here, that of observing states (e.g., Z score values) in defaulted cases and observing the same states in no default cases, don't have to add to 1: they may indeed add to less than 1.

PDs are also corrected for systematic errors present in models used to estimate them. This is carried out using simulated cases and also well-known past experience not used to build the model, e.g., a collection of rated loans, to measure the model's bias and then correcting individual PDs for such bias.

23.8 Loss given Default

For banks, an increase in the accuracy of loss given default (LgD) estimation represents an important opportunity to improve credit and capital management. The key challenge here is that most banks do not have the internal data to generate credible statistical analyses of LgD. Where the bank has some loss data it is often difficult to link any record of recovery cash flows to specific collateral types and track and allocate recovery costs. If enough data exists in a sector to make an analysis, results may not be enough to differentiate properly between the LgDs of other sectors.

Many banks will also want to make sure their approach complies with strict Basel rules for advanced approaches, including the need for one source of data to cover a complete economic cycle—no shorter than seven years—and for LgD estimates to reflect downturn conditions.

The latest trend in LgD analysis is toward "hybrid approaches" that can address these various challenges. Hybrid approaches combine statistical analysis—based on internal and external data—and expert judgment in a rigorous and transparent way. The "decision tree" is one such approach where distinct paths for the LgD analysis are chosen depending on the fundamental risk drivers in each structurally distinct portfolio:

The "collateral path" is designed to assess LgD where a loan is associated with a particular collateral type, and to take into account collateral value and allocation;

the "seniority path" is used for loans to large corporations secured by a general charge on obligor assets, where the key risk factors are often the degree to which the loan is secured against all assets, the loan's seniority and the amount of debt above and below the loan;

the "specialized path" is used for exposures to structurally distinct sectors such as project finance, trade finance and so on.

The decision tree allows the bank to use a combination of internal and external data to explore the effect of each relevant risk factor. Results of statistical analyses can then be benchmarked against industry-wide data and improved using panels of experts from the relevant divisions of the bank. these experts may also suggest specific adjustments to help make the analysis more specific (e.g., to take account of the relative LgD of construction real estate versus permanent real estate lending, or the effect of different types of collateral valuation in the case of loans secured by equipment).

Another useful hybrid approach where LgD data is sparse is the "asset-based" approach. The attraction of this approach is that it builds up a probability distribution of the value of all assets that represent a possible source of repayment—not only those that represent the primary collateral to the lender. The approach therefore does not depend on historical loss data and can be especially useful for low-default portfolios and for specialized lending. The approach also accommodates adjustments to the LgD estimates based on the bank's own experience and policies and takes into account additional risk factors such as jurisdiction.

Rather than developing their own approach, banks may also purchase LgD scorecards based on rating companies' data. Scorecards can help where there is insufficient data to build statistical

models; but they must be transparent, with detailed regulatory-related documentation otherwise they will not be accepted by Basel's Pillar 2. A range of industry and asset class specific off-the-shelf scorecards are available using statistical modelling where data exists or expert judgment where it does not.

23.9 Securitization and other concerns

A security is a financial instrument with specific legal characteristics. Publicly traded stocks or bonds are securities but there are many other securities, every one of them being some sort of equity or debt interest. "Securitization" is the act of transforming assets which originally are not securities into one or several securities. Basically, it is a financial transaction in which assets are pooled together and securities representing interests in the pool are issued. An example would be a financing company that has issued a large number of loans for final consumers to purchase cars. Rather than waiting for these loans to run their natural course, this company wants to raise cash so that it can issue more loans. One solution would be to sell off its existing loans, but there isn't a liquid secondary market for individual car loans. Instead, the firm pools a large number of its loans and sells interests in the pool to investors. For the financing company, this raises capital and gets the loans off its balance sheet, so it can issue new loans. For investors, it creates a liquid investment in a diversified pool of car loans, which may be an attractive alternative to a corporate bond or other fixed income investment. The ultimate debtors—the car owners—need not be aware of the transaction. They continue making payments on their loans, but now those payments flow to the new investors as opposed to the financing company. All sorts of assets are securitized: car loans, student loans, mortgages, credit card receivables, lease payments, accounts receivable, corporate and sovereign debt, etc.

Securitized assets are often -and misleadingly- called "collateral". They are more than collateral. In a typical arrangement, the owner—or "originator"—of assets, sells those assets to a "special purpose vehicle" (SPV). This may be a corporation, US-style trust or some form of partnership. It is established specifically to facilitate the securitization. It may hold the assets—collateral—on its balance sheet or place them in a separate trust. In either case, it sells bonds to investors. It uses the proceeds from those bond sales to pay the originator for the assets.

Most collateral requires the performance of ongoing servicing activities. With credit card receivables, monthly bills must be sent out to credit card holders; payments must be deposited, and account balances must be updated. Similar servicing must be performed with car loans, mortgages, accounts receivable, etc. Usually, the originator is already performing servicing at the time of a securitization, and it continues to do so after the assets have been securitized. It receives a small, ongoing servicing fee for doing so. Because of that fee income, servicing rights are valuable. The originator may sell servicing rights to a third party. Whoever actually performs servicing is called the "servicing agent".

Cash flows from the assets—minus the servicing fees—flow through the SPV to bond holders. In some cases, there are different classes of bonds, which participate differently in the asset cash flows. This is a structured securitization and, in this case the different bonds are called "tranches". If the securitization is unstructured, it is known as a "pass-through": there is only one tranche and all investors participate proportionately in the net cash flows from the assets.

When assets are transferred from the originator to the SPV, it is critical that this be done as a legal sale. If the originator retained some claim on those assets, there would be a risk that creditors of the originator might try to seize the assets in a bankruptcy proceeding. If a securitization is correctly implemented, investors face no credit risk from the originator. They also face no credit risk from the SPV, which serves merely as a conduit for cash flows. Whatever cash flows the SPV receives from the collateral are passed along to investors and whatever party is providing servicing.

Collateral will typically pose credit risk. For example, people may fail to make their credit card payments, so credit card receivables entail credit risk. This can be addressed with some sort of credit enhancement such as "over-collateralization" or a third party guarantee. Tranches (structuring) are also widely used to allocate credit risk among investors.

Credit ratings are often obtained for securitizations that entail credit risk. If a securitization is structured into different tranches, each may receive a different credit rating.

Securitization of assets should be regarded as an extremely dangerous transaction whereby true risks are hidden.

> First, with a securitization, the party underwriting credit risks is not the party taking that credit risk. This opens the door to various abuses.
>
> Moreover, the rating of structured instruments has often been misleading, referring to the best tranches and ignoring highly risky tranches.
>
> Finally and most importantly, the probability of default of securitized assets may increase in unison and exponentially during recession periods or as a consequence of panics or shocks. By pooling together assets, such as mortgages or car loans, securitization multiplies rather than diversifies risks.

Such risks, especially with regard to securitizations of subprime residential mortgages, were the primary cause of the 2008 financial crisis.

The most usual securitizations are

> mortgage-backed securities (MBS), which are backed by mortgages;
>
> asset-backed securities (ABS), which are mostly backed by consumer debt;
>
> collateralized debt obligations (CDO), which are mostly backed by corporate bonds or other corporate debt.

When securitization goes together with some type of synthetic instrument it is called "synthetic securitization". In general, the term "synthetic" is used to indicate the mimicking of assets using other assets. For instance, it is possible to mimic the return behavior of equity using treasuries and options. Synthetic securitization thus includes some type of derivative embedded in the contract.

Procedures used to assess risk of securitizations are similar to those used for other portfolios but, given the lack of sufficient information on final consumers such as their PD and LGD, securitizations are often badly rated.

Moreover, there is no agreement between two blocks of countries on how to recognize such assets:

> some countries regard securitized assets as though they were investment funds;
>
> others treat them as independent businesses or entities.

Recognition in turn affects risk assessment. This topic should be further developed in connection with the Basel regulatory framework.

Other, important topics relating to Credit Risk such as CVA and WWR are dealt with when discussing Basel regulation. This is because, although these concepts are quite general and basic, their designation became inextricably linked to regulation.

Chapter 24 The Basel accords – Pillar 1

The Basel accords have three pillars, of which a course in econometrics should cover the first pillar only, as the other regulatory recommendations do not relate to data analysis.

24.1 Bank capital requirements

According to Basel accords' pillar 1, banks are required to hold capital in readiness to cover unexpected losses or to pay the bank's obligations after bankruptcy. Pillar 1 contemplates two types of requirements regarding capital:

risk charges, that is, capital in proportion to risk and

extra buffers in proportion of risk charges,

Every risk identified by banks should be translated into capital "charges" that is, in an increase in the amount of capital held aside. Namely, banks should hold capital to cover

- at least 8% of the charges that banks may incur from loan default and other credit risk. This is CRC, "credit risk charges",
- plus total charges that banks may incur from market losses: this is MRC, "market risk charges",
- plus total charges that banks may incur from operating losses: this is ORC, "operation risk charges".

Thus,

$$minimum\ capital = 8\%\ CRC + MRC + ORC$$

In practice, Basel accords use RWA, the "risk-weighted assets" as a weighted addition

$$RWA = CRC + 12.5\ MRC + 12.5\ ORC$$

so that CAR, the "capital adequacy ratio" of banks,

$$\frac{minimum\ capital}{risk - weighted\ assets} > 8\%$$

is required by regulators to be 8% or above. MRC and ORC formulas or tables leading to the RWA components already include the 12.5 factors.

The Basel accords provide detailed rules governing the way risk charges (CRC, MRC and ORC) should be computed. In general, risk charges relate to risky assets' attributes in a way similar to VaR in traditional risk analysis. Basel accords allow banks to adopt one amongst a predefined set of methodologies to compute risk charges.

Minimum capital should be understood in the sense of a ready source of liquid funds, a reserve. Not all types of capital can be used to cover Basel's minimum capital requirements. Basel accords define two major types of capital with acceptable liquidity; they are known as tier 1 and tier 2 capital. In the Basel accords, a description of what is eligible to be tier 1 or 2 capital and what these tiers may cover is extremely detailed.

Tier 1 capital is also known as "going concern" capital. It is intended to prevent banks from failing. It consists of capital not tied to any obligation:

common equity (CET1), the best quality capital, to which distribution constraints are attached by the regulator

retained earnings,

restrictions on dividends, share handouts and other.

Tier 2 capital is also known as “gone concern” capital. It is intended to ensure that depositors and senior creditors are paid in case of bank failure. Tier 2 capital consists of long-term liabilities with maturity > 5 years.

Once the above specificities concerning minimum capital quality are taken into account, the capital adequacy ratio becomes

$$CAR = \frac{Tier\ 1 + Tier\ 2}{Risk - Weighted\ Assets}$$

Besides risk charges, the third version of the Basel accords introduced extra capital requirements in the form of buffers. For most banks, such buffers are the “capital conservation buffer” and the “countercyclical buffer”. Both are intended to be capital readily available to cover losses.

The capital conservation buffer (CC) is 2.5% of RWA. It is simply an extra charge aimed at making banks more resilient to unexpected losses. This buffer had its origin in the 2008 financial crisis.

The countercyclical buffer may go from 0% to a further 2.5% of RWA depending on whether economic conditions are good or feeble. It aims to be an automatic response to increases in the likelihood of losses due to economic downturns.

The two mentioned buffers are required to be tier 1 and CET1 (free common equity).

The larger banks, which are viewed as “systemically important financial institutions” (SIFI) face a further CET1 capital surcharge of 1% to 2.5% of their RWA.

Note that under Basel III, capital requirements still amount to 8% of RWA but now

only 2% of RWA is allowed to be covered by tier 2 capital; the remaining 6% must be tier 1.

CET1, the best quality tier 1 capital, must cover 4.5% of RWA; the 1.5% needed to get to 6% can be lower quality tier 1 capital.

CET1 must also cover the capital conservation buffer, an extra 2.5%

CET1 therefore now amounts to 6% + 2.5% = 8.5% of RWA.

Total capital requirements are thus 8% + 2.5% = 10.5% of which only 2% (10.5 – 8.5) are not CET1. If countercyclical buffer and SIFI requirements are added, then the capital requirements become huge.

Basel accords allow banks to choose among different approaches to measure MRC, CRC and ORC. In order to understand such variety of approaches it is first necessary to consider the difference between long and short-term risks. These are indeed very distinct risks, so Basel accords built such distinction into risk charge methods. The accords divide bank assets in two books: the “banking book” and the “trading book”.

Assets in the banking book (or long-term assets)

are held to maturity;

are valued at historical cost;

their VaR is estimated at 99.9% confidence level with 1 year time horizon

Assets in the trading book (or short term assets)

- are traded regularly;
- are to be marked to market daily (mark to market: assets are priced by comparison with actual market values of similar assets);
- VaR is estimated at 99% confidence level with 10 days horizon.

Given assets' attributes,

- most of credit and loans are kept in the banking book as these assets are to be held till maturity; but there are important exceptions to this rule
- most of assets traded in markets such as equities and bonds are kept in the trading book; but some FX and commodities positions are held to maturity thus they pertain to the banking book
- derivatives and operational risk may pertain to the banking as well as to the trading book, depending on whether assets are held to maturity or not.

Since banks incur in lower risk charges when assets are held in their trading books, Basel accords require an incremental risk charge for credit-sensitive assets that are held in the trading book. Indeed, such extra charge makes them similar, in terms of risk charges, to those held in the banking book. This is now a basic feature of the accords, having been implemented to avoid loopholes available in previous Basel versions.

24.2 Estimating Basel's risk charges

The approaches available to estimate risk charges are of two types: "standardized" and "advanced".

- standardized approaches are easy to implement but lead to higher capital charges; they are useful to small retail banks with no sophisticated risk analysis.
- advanced approaches lead to a lesser need of capital, but they require the use of sophisticated risk analysis and IT systems for risk assessment.

Some approaches are known by abbreviations. The complete list is as follows:

risk category	allowed approach
credit (CRC)	standardized (from 1988 accord)
	foundation internal ratings-based (IRB)
	advanced internal ratings-based (AIRB)
market (MRC)	standardized
	internal models (IMA)
operational (ORC)	basic indicator
	standardized

Banks have incentives to implement advanced risk management and liquidity control tools to reduce minimum capital.

The first risk charges to be described here are the simplest, the market risk charges. Regarding market risk, Basel allows banks to choose between two approaches, standard and internal models, known as IMA, or their mixture.

The standard approach simply uses "add-ons" instead of risk factors:

$$MRC = \sum position \cdot addon$$

where each type of risk (FX, Equity…) has specific add-on taken from tables. Assets are separated into portfolios according to type and subtype. Then, portfolio positions or exposure are multiplied by the designated add-on and added. MRC all from portfolios are finally added. The standard approach is crude, not recognizing differences in volatility nor diversification effects. Thus, it is expensive in terms of minimum capital.

Internal model IMA: MRC is basically an average of previous 60 days' VaR at a 10 days horizon and 99% confidence level, multiplied by a factor k. In case previous day's VaR is higher than such average, then previous day's VaR is taken as MRC.

$$MRC = SRC + k \cdot \max\left[\left(\frac{1}{60}\sum daily\ VaR\right), (previous\ day\ VaR)\right]$$

k is initially set at 3 but if the model is unable to signal losses beyond predicted, k is increased according to tables. SRC is an additional, specific risk charges, set by banks. This approach is a great bonus to banks, relying entirely on their analytical capabilities. It can spare significant amounts in MRC capital. But it requires an explicit permission and the whole process must be regularly examined by supervisors as part of Tier 2 framework.

Basel accords allow banks to choose one of 3 approaches to estimate CRC:

Standardized approach (from Basel I);

Foundation internal ratings-based approach (IRB)

Advanced internal ratings-based approach (AIRB)

The standardized approach is based on ratings attributed to each loan by a rating agency and on loan type. Both type and rating then determine a weight or add-on; such weights are multiplied by each loan's exposure and all added to obtain the portfolio risk charge.

$$CRC = \sum loan\ notional \cdot weight$$

The table shows how the standardized approach determines the CRC weights based on rating and type.

rating: / type:	AAA to AA-	A+ to A-	BBB+ to BBB-	BB+ to BB-	below B-	unrated
sovereign loans	0%	20%	50%	100%	150%	100%
bank loan option 1	20%	50%	100%	100%	150%	100%
bank loan option 2	20%	50%	50%	100%	150%	50%
short term loan	20%	20%	20%	50%	150%	20%
corporate loan	20%	50%	100%	150%	150%	100%

Percentage values indicate the weight that should be applied. Option 1 refers to rating based on the sovereign country in which it is incorporated; option 2 refers to rating based on an external assessment. Short-term means an original maturity less than three months.

In the foundation internal ratings-based approach (IRB) the bank internally estimates the probabilities of default, PD, for each of its loans. From these, Basel standard tables set weights and these lead to CRC as in previous approach:

$$CRC = \sum loan\ notional \cdot weight$$

Weights are taken from tables such as this:

PD	Corporate	Mortgage	Retail
0.03%	14.44%	4.15%	4.45%
0.10%	29.65%	10.69%	11.16%
0.25%	49.47%	21.30%	21.15%
0.50%	69.61%	35.08%	32.36%
0.75%	82.78%	46.46%	40.10%
1.00%	92.32%	56.40%	45.77%
2.00%	114.86%	87.94%	57.99%

...and so on. Basel III increases the risk weights on exposures to other banks relative to the non-financial corporate sector in the IRB approach: a multiplier of 1.25 is introduced for exposures to other banks.

In the advanced internal rating-based approach (AIRB), besides PD, banks may also estimate internally exposure at default EaD and the percentage lost in case of default LgD for each loan. A combination of PDs with LgD's for each class of exposure is then mapped into a table of weights. CRC then stems from multiplying each loan's EaD by the weight, adding all loans' results as usual. Tables are calibrated so that losses are made to equate capital required to support them during one year with a confidence level of 99.9%.

The AIRB approach is not to be applied in small and retail loans: only to corporation credit, other banks' credit and sovereign credit.

Banks are highly vulnerable to the default of a contract's counterparty. For instance, during settlement of OTC interest rate swaps or other derivative contracts, where huge sums are at stake, counterparties may default. Basel III counts counterparty credit risk in two ways:

- as an expected exposure adding to credit risk, to be reckoned in the banking book;
- as adjustments to market risk charges reflecting creditworthiness of counterparties; such adjustment is to be reckoned in the trading book, as market risk.

First, CCR is estimated basically like any other credit risk. For instance, where banks adopt an internal ratings-based approach, the corresponding weighting formula for estimating risk charges is applied, i.e.,

$$CCR = \sum EaD \cdot weights\ mapped\ from\ table$$

where PD, LGD and other attributes are introduced as explained for credit risk charges.

But whatever the approach used, specific rules now need to be applied to the estimation of EaD because exposure, in this case, is not a loan notional but a contracted settlement amount. Basel accords stipulate that EaD may be computed in two possible ways:

- In the standardized approach EaD = MtM + add-on where MtM is marked to market (market value of the settlement). This adds to a tabled add-on to get EaD.

In the internal approaches EaD is estimated using an “effective expected positive exposure” (EEPE or effective EPE): EaD is the higher of such EEPE at current market values and stressed values.

EEPE is the exposure to CCR expected over 1 year weighted by the proportion that an individual expected exposure represents of the entire time interval. EEPE estimation is made at portfolio level.

In addition to EEFE, a charge is introduced in Basel III to reflect, not the risk of default, but the deterioration in the creditworthiness of a contract’s counterparty in OTC derivatives and other contracts. The charge, known as credit valuation adjustment CVA, covers the risk of mark to market losses on expected counterparty risk. Thus it is market, not credit risk.

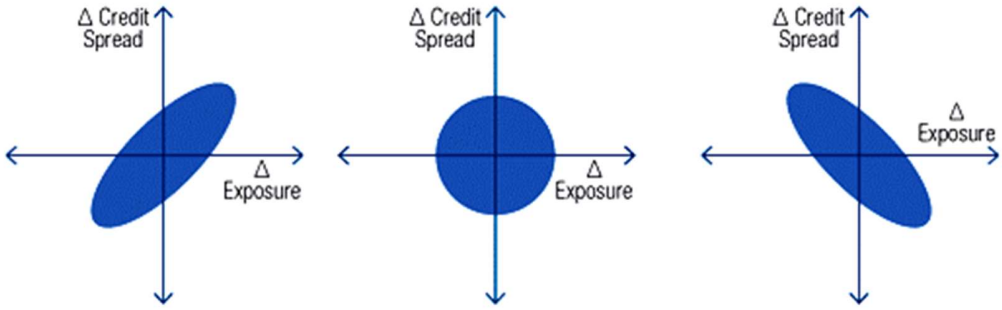
CVA is the difference between the risk-free portfolio value and the true portfolio value that takes into account the possibility of counterparty’s default. CVA is thus the market value of counterparty’s credit risk.

CVA estimation approaches are the same as any market risk exposure: standardized and internal model (IMA).

Besides CVA capital charges, CCR requires examining whether there exists wrong way risk or not. A risk is called “wrong-way” (WWR) when exposure tends to increase when the counterparty credit quality gets worse. Wrong-way risk is a risk that occurs when exposure to counterparty is correlated with the credit quality of that counterparty. The terms “wrong-way risk” and “wrong-way exposure” are often used interchangeably.

An example would be a forward contract with a gold producer in which the bank pays the spot price of gold and receives a fixed price. Suppose the price of gold were to decrease. That would worsen the credit quality of the gold producer, since their revenues would decrease, making their business less profitable and viable. It would also increase the value of the forward contract to the bank since the bank is paying the spot price; therefore, the bank's exposure would increase. If these two effects tend to happen together, then that co-dependence will increase the CVA on the forward contract and it will make the CVA larger than if the effects were independent.

Basel III introduces a capital charge for wrong way risk. Banks must have procedures in place to identify, monitor and control cases of wrong way risk. To calculate this capital charge, the instruments for which there exists a legal connection between the counterparty and the underlying issuer, and for which specific wrong way risk has been identified, are not considered to be in the same netting set as other transactions with the counterparty. Furthermore, for single-name credit default swaps where a legal connection exists between the counterparty and the underlying issuer, and where specific wrong way risk has been identified, EAD counterparty exposure equals the full expected loss in the remaining fair value of the underlying instruments assuming the underlying issuer is in liquidation.



The representation above illustrates the difference between WWR (left), independent (center) and right-way risk (right). Data needed to quantify WWR correlations is difficult to obtain and is ephemeral. But it is crucial that the bank's CVA framework should be able to recognize and handle WWR cases.

We hope that we have managed to offer a detailed view of market and credit risk in banks. Among the subjects not covered by these notes are operational risk, both under the viewpoint of risk management and Basel Pillar 1 approaches, then Basel Pillars 2 and 3 plus all the vast concerns of Stress testing.