UNIVERSITY OF NAIROBI

STA 602: MODELLING AND ANALYSIS OF SOCIAL DATA

MSC- SOCIAL STATISTICS

CHIROMO CAMPUS

BY

DR. OEBA VINCENT ONGUSO

vongusoeba@gmail.com

MOBILE: +254-720-475053

OR

+254-733-244911

COURSE OUTLINE

1. Sources, nature, standardization and uses of social statistics

- Educational statistics,
- Housing statistics
- Crime statistics
- Social security statistics
- Health statistics
- Labour statistics
- Environmental statistics
- Food and Agricultural statistics

2. **Techniques for modeling social data:**

- Models for discrete data and
- Multivariate techniques of social data-social statistics

3. Social Indicators

4. Design and analysis of comparative studies

5. Man power surveys and Man power projection techniques

Aim of the course

This course aims to equip students/learners with knowledge and experience of the principles, theory and practical skills of statistics applicable to social and behavioural sciences.

Course outcomes

By the end of this course you should be able to:

i) Distinguish different sources of social statistics and their application in social and behavioural sciences
ii) Apply appropriate statistical techniques in analysing data from education, health, agriculture, labour, environment and crime statistics

iii) Design social indicators on various sources of social statistics
iv) Design surveys of social statistics with appropriate projects techniques on social and behavioural sciences
v) Support research in the social and behavioural sciences with appropriate statistical techniques for effective decision making and policy development

Mode of assessment

Assignments =10%

CAT =20%

Examination =70%

References

- The Methods and Materials of Demography by Henry S . Shryock, Jacob S. Siegel
- Computer Aided Multivariate Analysis 4th edition by Abdelmonen Afifi, Virginia A. Clark and Susanne May, 2004
- The analysis of Time series 5th edition, Chris Chatfield

Chapter 1: Sources, nature, standardization and uses of social statistics

Learning outcomes

By the end of this chapter you should be able to:

i) Distinguish types of data sources of social statistics

ii) Explain use of social statistics and the principles of standardization

iii) Employ appropriate statistical methods in analysis of data from social science research

1.0 Introduction

We begin this chapter by understanding what is data? What is social data? What is the nature of social statistics? What are the standardization needs in social statistics? What are the uses of the social statistics? Then much of the content will be devoted to various topics on sources of statistics in social sciences such as education, crime, health, environment, food and agriculture, housing, social security and labour. Each of these topics will be dealt in detail to cover the principles and methods used in analyzing data from such sources.

1.1 What is data?

Data are values (measurements or observation) that variables can assume. A collection of data values forms a data set. Each value of the data set is called **data value** or **datum**. In this case a variable is a characteristic or attribute that can assume different values. The variables whose values are determined by chance are called **random variables.**

1.2 Variables, Measurements and Data Types

Introduction

We hardly talk without making reference to numbers, rates or confidence. For instance:

- The unemployment rate increases at 12 percent per year;
- The chances of raining tomorrow are very high;
- I am 99 % sure that our gross yield for this year will exceed government target;
- The media always tend to quote percentages or give figures; etc.
- The failure rate in Kenya Certificate of Secondary Education is very high
- The crime rates in urban centres of Kenya are very high as compared to rural areas

- The agricultural land productivity is declining in Kenya. farmers are only able to harvest about 0.8 t/ha
- The rate of rural urban migration is high in Kenya
- The housing prices in Kenya is increasing by 5% every year

All the above statements are meant to garner support for some assertions. One can conclude that we are living in a world of numbers. Many of these statements are subjective.

Statistics therefore is the art of

- ❖ collecting,
- ❖ summarising,
- ❖ presenting and
- ❖ interpreting data about a population of interest.

The part of statistics that deals with the collection and summarisation of the data is referred to as ***descriptive statistics*** and that which relates to inference about a population of interest based on the information contained in the sample is called ***statistical inference.***

Statistics acts as a magnifying glass that helps us to see the inside of data for the purpose of understanding the system. The prediction of the future is based on these statistics. Hence, we predict future with more confidence. Most of the statistical tools in use are based on mathematics, especially probability theory. Statistics can therefore be defined as a theory of information gathered through experimentation by sampling techniques.

Almost everyone is confronted with statistics in their day to day living, but few people have much of a notion about the discipline of statistics. Three examples are given in the box below:

Box 1: Examples of everyday statistical practice

(i) Suppose that a manufacturer of light bulbs produces roughly a half million bulbs per day. Suppose the firm wishes to determine the fraction of the bulbs produced on a given day that are defective. The firm can undertake two approaches: Could insert the entire half million bulbs into sockets and test them. The cost would be very high for such undertaking. Could take say 1000 bulbs from the half million produced and test each one and record the defectives. The fraction of the defective in the 1000 bulbs could estimate the fraction of defective in the entire day's production. Statistics also deals with how the 1000 bulbs will be selected.

(ii) Suppose a researcher wishes to investigate the effect of a new drug on the simulation of a patient's heart. Interest is in the effect of the drug on all future patients treated with the same drug. Suppose fifty heart patients are selected and

each treated with the new drug. Based on the results from the fifty patients the researcher may infer the effect of the new drug on future heart patients.

(iii) Suppose a veterinarian wishes to investigate the extent of spread of taro leaf blight disease in the island of Savaii, Samoa. Fifty farms were visited and the incidence of the disease recorded. The proportion of incidence can be used to estimate the incidence of the disease in the entire Savaii Island.

Statistics is essentially concerned with the application of logic and objectivity in the understanding of events. Necessarily, the events have to be identified in terms of relevant characteristics or occurrences that could be measured in numeric or non-numeric expressions.

The characteristic or occurrence is technically referred to as a variable because it may assume different values or forms within a given range of values or forms known as the domain of the variable. The act of attaching values or forms to variables is known as measurement.

1.2.1 Variables and types of data

A variable may be qualitative or quantitative; and a quantitative variable may be continuous or discontinuous, i.e. discrete. The convention is to identify a random variable by the upper case letter e.g., X and an observation by a corresponding lower case e.g., x. Qualitative variables are variables than can be placed into distinct categories to some characteristics or attribute.

1.2.1.1 Qualitative variable

A qualitative variable X may have forms x that could only be described but could not be measured numerically. It is a non-numeric entity. Nutrient when identified only as nitrogen, phosphorus, potassium or zinc is a qualitative variable or factor. Variety of maize is qualitative; sex of a farmer is qualitative. Yield scores when recorded as high, moderate or low is qualitative. A qualitative variable has states or forms, which can be described, categorized, classified, or qualified.

1.2.1.2 Quantitative variable

A quantitative variable Y is numeric; it may assume values y that can be quantified. Age, weight, height, volume are all quantitative variables. Production, prices, costs, sales, sizes, etc., are also quantitative.

A qualitative X may be recorded as if a quantitative variable; for instance in a farm survey, a male farmer may be coded as 0 and a female as 1. In an entomological experiment, response may be classified as 0 if the plant survives a chemical treatment and 1 if otherwise. Resistance of plants after exposure to a plant disease may be scored as 0, 1, 2, 3, 4, and 5 in ascending order of infection; and in a taste experiment, the tasters may be asked to grade dishes as 1, 2, 3, and 4 respectively. And on the other hand, a quantitative variable Y may be coded qualitatively; for example in an agronomic trial, maize plot response may be coded 'A' for yields exceeding 7500 kg per ha. 'B' for yields in the bracket 6001 – 7500 kg, 'C' for yields between 5001 and 6000 and 'D' for yields 5000 and below. Forms A, B, C, and D are of the descriptive type.

Continuous and discrete variables

A quantitative variable may be continuous or discontinuous i.e. discrete. The family size of the farming household, X, in a crop survey may take between values 1,2,3,4,5,6,7, and so on. In the interval (0, 100) the number of values which X can take is countable (finite), it is 101; but the size of farms Y cultivated by the household may take an uncountable (infinite) number of values within even the shortest of intervals when measurement is not limited by degree of accuracy nor by the precision of the measuring equipment. Family size variable X is a typical discrete variable while farm size Y is a typical continuous one.

Let y_1 and y_2 be two distinct values of a continuous Y, the variable may still take an uncountable number of values between y_1 and y_2, however close the two values might have been to each other. On the other hand, a discrete variable X may not take more than a known finite number of values between two distinct values x_1 and x_2, however distant the two values might have been from each other. In a continuous situation, there is always another value between two values while in the discrete environment, there may not be any other value between two values.

1.3 Types of data

The type of statistical analysis that is appropriate for a particular variable depends on the data type (scale of measurement) used for the random variable. There are four data types, namely nominal, ordinal, interval and ratio. These levels of measurement can be distinguished on the basis of the following criteria:

- Magnitude or size
- Direction
- Distance or interval
- Origin

- Equality of points
- Ratios of intervals
- Ratio of points

1.3.1 Nominal-scaled data

Examples are the labels sex (male, female), marital status (single, married, widowed, divorced), religious affiliation (Anglican, Methodist, Catholic, Protestant, etc.), employment status (employed, unemployed), farm size (small, medium, large), education-level (primary, secondary, tertiary). These labels are only used to identify an attribute of the random variable. Under nominal measurements, 'numbers' are simply to identify, classify, categorize or distinguish. For instance, scores 0 for male and 1 for female are simply used to identify, distinguish, categorize or classify the subject by sex. The score has size or magnitude, it has equality because two subjects are similar (equal) if they score same number But the scores has only these two properties and none else. Score 1 is not necessarily greater than score 0, nor score 0 less than score 1; Concept of distance or interval, would not make any sense and the four arithmetic operations (addition, subtraction, multiplication and division) are not feasible. Nominal level of measurement is very poor; it is the weakest level.

1.3.2 Ordinal-scaled data

Ordinal data are associated with qualitative random variables and are generated from ranked responses (also generated from a counting process). The data have the properties of nominal data and the order or rank of the data is meaningful (e.g., where customer service is interested in getting views related to their services from the customers). Suppose the customers are asked to provide rating for six different variables namely, food, drinks, service, waiter, manager and hostess. The response categories are excellent, good, and poor for each variable. The observation for each variable possesses characteristics of nominal data in that each response rating is a label for excellent, good or poor quality. In addition, the data can be ranked with respect to quality. An ordinal variate may also be represented by scores such as 1 = poor, 2 = fair, 3 = good, etc.

The data can be numeric or nonnumeric. The scores have magnitude, equality, direction, and none else. The ordinal level is the second weakest level of measurement, and the four arithmetic operations do not also apply for ordinal data. There is however a wider range of valid statistical methods available for the analysis of ordinal-scaled data than there is for nominal-scaled data.

1.3.3 Interval-scaled and Ratio-scaled data

These are really numeric. They each have magnitude or size, direction, distance or interval, and origin. Any two intervals of values on one scale of measurement maintain the same ratio on any other scale of measurement. The distinguishing properties between the interval and ratio scale are:

- Interval scale has no absolute origin that is independent of system of measurement, while the ratio scale has an absolute origin that is not system dependent.
- At the interval level of measurement, ratio of any two points is system dependent while at the ratio scale, such ratio is system independent.

Freezing, body and boiling points are 0, 37 and 100°C and 32, 98.6 and 212°F. Intervals between freezing and body temperatures are 37°C or 66.6°F, and freezing and boiling points are 100°C and 180°F. The ratio of the two intervals is 37:100 = 0.37 under Celsius and 66.6:180 = 0.37 under Fahrenheit. The ratios are therefore system independent.

On the other hand, 0°C is not the same temperature as 0°F neither is 0°F equal to 0°C, i.e., there is no absolute origin. The ratio body and boiling temperature points is 37:100 = 0.37 under Celsius but 98.6: 212 = 0.465 under Fahrenheit. It is shown therefore that temperature variable is measurable only at the interval scale.

Now consider this: Let the weights of dry matter from 3 plots A, B, and C be given as 50, 100 and 200 kg, respectively; these are equivalently given as 114.8 and 229.6, 459.2 lbs, respectively. Ratio of differences (weight of B – weight of A): (weight of C – weight of B) is 2 on the metric system and same number 2 on the imperial system. Also, ratio weight of C: weight of A gives 4 on both systems. The ratios are system independent. This is a property of the ratio scale of measurement. *Height, area, age, money, etc. are variables that could be recorded at the ratio scale of measurement*.

Let x be the value of a quantitative characteristic on one system of measurement e.g., metric system and y the equivalent on an alternative system; y and x would be related differently depending on the scale of measurement.

- Interval scale: $y = a + bx$, where constants a and b are each non-zero
- Ratio scale: $y = bx$, where b is non-zero

For instance, F = 32 + 1.8C, 1lb = 0.4356kg, 1ft = 0.3048m, 1 gal = 4.4561L. These relationships confirm that while temperature is of the interval scale, weight, length, and volume are each of the ratio scale.

1.4 Cross-sectional and Time series Data: This terminology identifies the timeframe of the data. A survey or experimental data could be cross-sectional or time series.

1.4.1 Cross-sectional Data

A cross-sectional or latitudinal data has a single time scope or frame. The data consists of observations made on sampling or experimental units over a single point in time which may be a growing season, a month, a quarter, a year.

- A single experiment gives a cross-sectional data.

- A census data is cross-sectional.
- A farm survey in which the figures are observations made at a point in time is cross-sectional.
- A series of experiments or surveys in which observations are not repeated on same experimental units constitute a cross-sectional data.

1.4.2 Time series

The time series or longitudinal data are repeated measures on same experimental or survey units. Sets of observations made on each of the experimental units of an experiment at p different times constitute a time series or repeated measures.

- A long-term experiment in which yields are taken at successive periods on the same experimental units gives a kind of time series data.
- Retrospective or prospective surveys in which observations in respect of specified medical conditions of a patient are made over time are time series data.
- Economic data that are observations on same economic units, like households, political entities, etc. constitute an economic series that are time series.

Box 2: Survey versus design of experimental study

The difference between a survey study and design of experiment study is mainly in the study objectives. The researcher should understand the difference before he or she undertakes the study. Failure to make the distinction between the two forms of studies leads to complicated data analyses whose results may fail to tie with the study objectives. The primary objective in survey study is to observe the characteristics of the population of interest. For instances, is the disease common across the different communities? Is the level of education equally distributed among race? Is the distribution of land even across different communities? What is the opinion across the residence regarding new rules in rubbish disposal? In such situations we would be concerned about the level of distribution rather than the actual difference. Where is the variability high? would also be a question of great interest.

In designed experimental study the primary interest is to investigate on the relative performance of certain factors. The key questions to be answered are generally expressed as a statement of hypothesis that has to be verified or disproved through experimentation. The interest would be in answering questions such as: *Are the three methods for treating the disease different? If so, by how much? Is the new teaching method significantly different from the old method?*

In a survey study the researcher has no control over the responses. He/she acts as an observer. The outcomes are mainly considered as random. Survey study can be classified **into two types** namely exploratory (or informal) survey and formal survey. The exploratory survey is mainly used in obtaining information about population of interest,

e.g., farmer circumstances. The approach places interviewer in direct contact with the subject and allows the interviewers to observe the characteristics of the population. An exploratory survey allows for quick gathering of information through informal interviews with many people. The information from exploratory survey is used to design a well-focused formal survey by:

- identifying important topics bearing on research planning that should be the focus of the formal survey;
- ensuring that written questions in the formal survey are asked in a way that can be understood; and
- designing and testing a sampling scheme.

1.5 Data structure

A data may be structured as a:

(i) Simple set of univariate data,
(ii) Simple set of multivariate data,
(iii) Designed experimental data,
(iv) Simple regression data,
(v) Multiple regression data,
(vi) Nonlinear regression data,
(vii) Classical functional data,
(viii) Classical structural data,
(ix) Contingency data

1.6 What is social data?

The understanding of social data is based on research undertaken in social and behavioral sciences that can take any type of data described in the previous section. It can also follow the various types of data structure identified in section 1.5. Therefore the types of data sources in social statistics will consists of measurements from education, health, social securities, food and agriculture, labour, housing, crime among others.

1.7 Types of data sources for social statistics

The increased emphasis on overall human development rather than only economic well-being has increased the demands of data from the National Statistical Systems.
Not only is there a demand for more frequent and more disaggregated data but also for a much wider set of indicators such as energy, environment, and dissemination of technology. The National Statistical Offices have a number of sources through which they try to meet these demands. The sources of data are particularly diverse for social

statistics as compared to economic and financial statistics. Each source however has a few drawbacks.

The common types of data sources for social statistics include the following;

i) Census
ii) Surveys
iii) Civil society registration system
iv) Administrative records
v) International data sources
vi) Data produced by NGOs and private sector

Each of these types of data sources are described below

a) Census

- The Census is considered one of the most reliable and comprehensive sources of socio-economic status. However, in many developing countries it is not carried out on a regular basis due to their social or political conditions

- Census data may be of limited use for monitoring the millennium development goals (MDGs) of achievement vision of developing and developed nations such as that of Kenya's Vision 2030 as they are generally conducted once in ten years, whereas the MDGs need to be monitored annually.

- Moreover, there is considerable delay in processing of census data once it is collected. For example, in Kenya the 2009 census was realized in 2011? due to various reasons including unavailability/inconsistency of the complete data when the report was prepared.

b) Surveys

- Sample surveys are the most frequently used sources of information, specifically for human development data, since as compared to the Census, they are relatively more cost effective.
- However, the results of the survey depend heavily on the sampling techniques used, size of the sample and the extent of bias on the part of those conducting the survey and the responses of those surveyed.
- In specific cases such as HIV/AIDS, the stigma and discrimination associated with HIV may lead to poor reporting.

c) Civil society registration system

- Complete, timely and accurate registration of births and deaths is considered crucial for the understanding of population dynamics at the local level and planning of effective health and development programmes.
- Civil Registration Systems that exist in most countries provide vital information necessary for the estimation of mortality rates and life expectancy.

- If functioning efficiently, civil registration systems can be of immense help in generating human development data at the disaggregated level.
- It is a pity that civil registration systems in most of the developing countries suffer from inadequate coverage and under reporting.

d) Administrative records

- Each of the administrative departments in the national governments collects and records data for its own monitoring and reporting purposes.
- Very often it is this data that is put to use when reporting on human development becomes essential.
- Programmatic data collected by government functionaries has an in-built bias towards highlighting achievements and is considered unreliable by experts.
- What is even more disconcerting is that even as data on a variety of indicators is collected by the programme implementing agencies, they are not accessible to the very people who are the main stakeholders in development planning.
- In view of the movement on Right to Information in several countries, it is important that the data is disseminated in a user-friendly format to policy makers as well as stakeholders.

e) International data sources

- This points out that it is ironical that often international sources of data are considered more authentic than the concerned national data sources.
- It is well known that international data is either derived from a national source or estimated on the basis of projection/ extrapolation using data from countries with a similar profile.
- Another issue that is of concern in this regard is that of ownership of data.
- International data sources are easily accessible, but whether they depict the true conditions and priorities of the people concerned is an aspect that needs to be considered while using them.

f) Data produced by NGOs and private sector

- Often data is collected by NGOs and the private sector such as private hospitals and health care systems for their own administrative and monitoring purposes.
- This data often can be very detailed and dis-aggregated as per the requirement of the concerned organization.
- Despite its use in specific contexts, its use for general planning and monitoring is limited, as it cannot be used for wider and more diverse sets of population.
- However, a systematic effort at developing appropriate data formats and insisting on reporting of a minimum set of indicators can contribute to the generation of micro level data on a variety of indicators of interest to human development practitioners.

1.8 Standardization of Social Statistics

- The need for deeper harmonization of social statistics on international level is the consequence of globalization processes and progressing institutionalization of economic and social policy in more open, market driven economy.
- Effective approach to harmonization of social statistics on international and superanational level -is the development and implementation of integrated complex of standards of metadata used for representation of relevant social surveys and data.
- The operational tool of harmonization is the Standard Social Data Interchange System.
- The concept of political, economic and social *transition designates the process of accelerated and comprehensive political, social and economic changes, transforming* the societies and economies from non-democratic political systems and centrally planned economies, to more democratic political systems and to market - driven, more opened economies, integrated with global economic system.
- The *transition is initiated,* coordinated and controlled by governments of particular countries. First stimulus for starting the processes of *transition is of purely political nature. The governments of* formerly *centrally - planned economies are made - under social and political pressure* strengthened by the inefficiency (and in some countries - the bankruptcy) of centrally - planned economy - to introduce *institutional changes of political system.*
- *The* consequence of political changes is the transformations of social and economic systems. Those institutional political changes generate specific social and economic processes, which are commonly known as transition processes.
- The standards in social statistics should be developed on the basis of "best practices" of official statistical agencies, but should not copy them.
- "Best practice" may be "not good enough" for harmonization and international interchange of social data.
- There are needed the standards that create a "***common metainformational denominator"*** harmonizing formats and documentation of semantics of social statistical data.

1.9 What to consider when defining standards for social statistics

The following principles need to be followed when defining standards to follow for social statistics.

i) New social and economic phenomena, which appear in many countries, and are explicitly visible in transition countries, require new socio - economic concepts, statistical methods and indicators. Some "old" social indicators may lead to misinterpretation, to erroneous use or misuse of statistical data.

ii) Control of comparability and integrity gaps of social statistical data in transition countries needs special methods of identification and elimination of those gaps.

iii) Problems with accessibility and reliability of sources of social data are the result of enlargement of the spheres of shadow economy, non-registered and ill

registered social and economic activities, lower quality of administrative registers**.**

iv) Problems of proper interpretability of "old" social indicators in dynamically changing social and economic environment. Social statistical indicators in transition period cannot be interpreted correctly without proper understanding of their economic, social and cultural context of concrete phases of transition of different spheres of social and economic life and good understanding of national, regional and local specificity.

The description below provides an overview of what need to considered in standardization of various sources of social statistics and corresponding implications

1.9.1 Demographic statistics

High dynamic of demographic processes is observed in particular in the beginning of transition period. This dynamics is caused both by political and economic reasons. In some regions dramatic changes of demographic situation is observed. For instance we can be confronted with the following specific phenomena

- Significant dynamic changes of basic vital indicators in first years of transition: decrease of rate of birth, *overmortality* of men, rapid decrease of rate of marriages, increase of rate of divorces, changes life expectancy, changes of fertility rate etc.)
- Types of internal and international migrations of different character:
- *Ethnic re-emigration*: after the creation of new independent states ethnic migrations of people to their ethnic states,
- *Political migrations*: migrations caused by political reasons (e.g. migrations stimulated by introducing new laws on citizenship, on official national language etc.)
- *Economic migrations*: economic disturbances in first years of transition have caused economic international migrations: permanent, long term, short term and periodical,
- *Refugees*: permanent, long term, short term and periodical migrations caused by military actions,
- *Migrations caused by disasters*: natural, ecological.
- Causes of internal and international migrations:
- Re-emigration of formerly displaced population,
- Ethnic conflicts and discrimination,
- Social (including religious) conflicts,
- Economic reasons (poverty, unemployment),
- Security (i.e. wars, military operations, criminality),
- Ecological and natural disasters (i.e. emigration of people from heavily polluted areas)
- Refugees as special category of migrants

1.9.1.1 Implications of demographic phenomena

Population censuses conducted every ten years. Provides exceptional opportunity to get the information on population and housing after several years of dynamic changes. Therefore it should be recommended to compile demographic information in basic social cross sections (age, sex, civil status, ethnicity, religion, education, profession, and economic activity - employment, housing conditions).

- ***Current population survey*** producing population estimates through registries of vital statistics (births, deaths, marriages, divorces, migrations) by sex, age, ethnicity, regions.
- ***Migration statistics*** (***international and internal***). It is recommended to pay special attention to ***migration statistics*** by: ***reasons of migration***
  - ❖ age,
  - ❖ sex,
  - ❖ ethnicity,
  - ❖ education,
  - ❖ profession,
  - ❖ duration of migration (permanent vs. temporal migrations: periodicity, migration cycles)

1.9.2 Housing statistics

The specific phenomena here includes the following

- Changes in housing conditions of population caused by market economy and cuts of government subsidies (significant increase of the share of rents for flats in family budgets).
- Substance of housing: in some regions political disturbances (including military actions) cause losses in the quantity and quality substance of housing.
- Regionally concentrated migrations caused by housing conditions.
- Consequences of demographic process (rate of births, marriages, divorces, migrations) on housing situation.
- Homelessness.
- Commercialization and privatization of housing and its social and economic consequences for households (change of income for disposal, negative income for disposal etc.)

The implication of this phenomena include the following

- Continuation of ***population and housing censuses*** (every 10 years).
- Special ***ad-hoc surveys of housing in special regions***, in which specific factors influencing exceptional demographic processes and changes in housing conditions (refugees, displaced population, extensive migrations), changes of housing resources in regions unaffected by:
  - ▪ wars, military actions,
  - ▪ ecological disasters,
  - ▪ natural disasters,
  - ▪ concentrated migrations (e.g. camps for refugees),

- Careful ***monitoring of migrations*** (international and internal) caused by the changes of labor markets (data from labor force surveys), ***and its influence on housing*** conditions.
- Current population survey based on censuses, registration of population (administrative records) and the results of special ad-hoc surveys concatenated with data on housing.
- Estimates of ***homeless population*** by regions, with special reference to urban areas and big cities, social structure of homeless people.
- Level and dynamics of rents for flats and their share in incomes of households.
- Development of dwelling market (prices of flats, rents for flats ratio of the price to the average income and salary). The data are necessary to evaluate the mobility of labor force in the country.

1.9.3 Labour statistics

This is based on the following phenomena:
- Shadow and ill-registered employment and self-employment.
- Incidental, unstable, short time and part time employment.
- Unemployment generated by the processes of the restructuring of industry.
- Very high rate of discouraged workers.
- High non-registered unemployment.
- Many, relatively small, relatively isolated, autonomous *local labor markets,* i.a. because of underdevelopment of modern commuting infrastructure, high costs of flats on free market and the structure of urbanization of the country ("onefactory dependent" towns).
- Law mobility of labor force because of economic reasons.
- Underdevelopment of administrative infrastructure organizing labor markets and unemployment
-Information gaps in labor market and unemployment of rural areas.
- Changes of administrative criteria of registration of unemployed persons, time series data on registered unemployment and not comparable.
- Ethnic changes of employment and unemployment caused by administrative decisions (e.g. introduction of laws on official national language eliminates national minorities from then jobs requiring fluent knowledge of official language).
- Different quality of data on employment and unemployment collected by statistics from different groups of businesses.
- Unsatisfactory reliability and stability of administrative registers of unemployment and changes of quality, i.a. caused by changes of laws and regulations of labor market and of social security system.

1.9.3.1 Implication of the labour phenomena on social statistics

-Basic relatively reliable and uniformed source of information on labor market is the labor force survey based on data collected from households.
Despite of high costs and organizational problems of this type of survey in transition countries, it is recommended to use the labor force survey as basic source of information on labor market.

- The sample of households should allow to produce data (employment, unemployment) by all important cross sections specified above i.e.:
    - Age,
    - Sex,
    - Ethnicity, knowledge of official language (and religion - if relevant),
    - Education (with special reference to observation of school leavers),
    - Region, urban/rural areas, local labor market,
    - Duration of employment and/or unemployment, with special reference to short and "incidental" employment, part time employment,
- Migration aspect (place of permanent residence, place of work),
- Estimation of discouraged workers,
- Shadow employment and unemployment.
- Quarterly periodicity of labor force surveys is recommended.
- Statistical map of *local labor markets* to help the governments to evaluate the situation on regional and local labor markets and to chose most effective measures to preventing and fighting against unemployment and its social consequences.
- Because of frequent changes of laws and regulations of labor market and unemployment, the contents and interpretation of data driven from administrative registers of unemployment is unstable. Systematically updated methodological metadata are necessary for proper interpretation of statistical indicators on jobs, employment and unemployment (e.g. unemployment rate and its changes based on administrative unemployment registers) produced on the basis of administrative registers.
- Validation and quality control of data on employment and jobs collected directly form businesses.
- Needs for good statistics of costs of labor:
- Costs of labor for employers.

1.9.4 Health statistics

The specific phenomena under which health statistics is supposed to be standardized is based on the following:

-Fundamental changes (implemented or under preparation) of public health care systems, from public services provided by government to market - driven services.
-Changes of the system of financing of health care: from government budget to health insurance separated from government budget and direct financing of health services by households.
- Significant reductions of government subsidies for public health care system, for medicines and health services.
- Development of commercialized (formal or informal) and privatized market of health services.
- Changes and polarization of accessibility and availability of health services. Some social groups and households are loosing the access to medical care because of formal or economic reasons (e.g. homeless people, non-registered unemployed people, workers employed in shadow economy etc.)
- Increase of costs of health services paid directly by households.

- Changes in health status of different classes of population, with special reference to (a) social groups around and below poverty line, (b) the unemployed, (c) people who do not benefit from health insurance (secondary effect of economic and social polarization of population).
- Significant share of foreign aid in providing health services, especially in case of disasters (medical equipment, medicines etc.).
- Specific need for medical care in regions of disasters, conflicts and for specialsocial groups (displaced population, homeless people).

Implication on standardization of health on social statistics

- Official statistics should cover all units providing medical care both public and private commercial health services and other organizational forms of medical care. Adjustment of the definition of the health care units and their classification is necessary.
- Official statistics should provide data for estimation of real costs of health services. It is needed for proper budgetary policy of governments, reform of health insurance systems and for proper transition of national health systems.
- Statistics of direct expenditures for health services paid by households, on the basis of household surveys; if necessary, the classification of expenditures by kind should be extended.
- Health insurance administrative data and social insurance administrative data may be very useful to compile statistics on health services.
- Health status surveys conducted every 3-5 years are recommended.
- In household surveys special attention should be paid to the data on health care, expenditures for health services and government subsidies addressed to the population with law income, to social groups below poverty line, regions of disasters, as the part of surveys on poverty and health.

1.9.5 Education statistics

The phenomena under which education statistics needs to be standardized is based on the following:

- Institutional changes of national systems of education as the entire component of transition process, adjustment of education system to market - driven economy and to national tradition, especially in new independent states (e.g. religious and ethnic schools).
- Organizational changes of education system to new organization of regional and local self-government.
- Changes in financing the education system: relative decrease of financing from central government budget, increase of financing by local self - governments, social organizations, including NGO`s, churches and other religious organizations.
- Privatization and commercialization of some segments of education system.
-Commercialization of vocational training, of high and university level education.

- Gaps between profiles of education in school system developed under centrally planned economy (particularly vocational education) and the needs of the market - driven economy in transition. The gaps appear on local labor markets and on national level.
- Gaps between the profiles of education and skills of population and the needs of market driven economy in transition. The re-training of large groups of workers in restructuring branches of the economy.
- Changes in accessibility to education, e.g. low accessibility to university level education for population from non-academic towns (rural areas, small towns) and for household with lower incomes.
- New forms of education and re-training, e.g. training and re-training of school leavers and unemployed persons.
- Rapid changes of accessibility of pre-school education.

The implications of standardization of education on social statistics

- Changes of education systems require the adjustment definitions of terms, classifications of schools and classifications of vocational profiles of training to new institutional, organizational and economic situation.
- The adoption of international methodological standards (definitions, classifications) is recommended, but national specificity should be also represented. Gateways (correspondence tables) between national and international classifications are necessary.
- ***Priority: official statistics should identify the "education profile gaps" between the education system and the labor market in short, middle and long terms of transition process and the projection for post - transition period:*** (comparative analysis of vocational profile structure of the education system and of the needs of the market, for both local labor markets and for the national economy).
- Official statistics should provide data on education profiles and the profiles of jobs to support the elaboration of programs of re-training of workers, to adopt their skills to the situation on local and national labor market.
- Statistics of accessibility to education and attendance rate:
    - by age,
    - by sex,
    - by socio - economic groups (with special reference to low-income households),
    - by ethnicity (with special reference to minorities),
    - by regions and places of inhabitance
    - by levels and profile of education.

- Special statistics of school leavers and their entrance into the labor market.
- Statistics of conditions of education (number of children per classroom, teachers, laboratories, computer and internet laboratories, other facilities in schools)
    - Pre-school education
    - Costs of education (schools as economic units)
    - Level and structure of financing the education (sources of financing)

- Costs of education paid by households.

1.9.6 Environment statistics

This is based on the following phenomena

- Environmental "heritage" of centrally planned economy - devastation of environment:
- Concentration of polluting industries in selected regions
- Ecologically destructive industrial technologies
- Low level of environment protection facilities
- Large areas durably or permanently devastated (some regions excluded from economic and social use forever)
- Relatively low level of "environmental culture"
- High density of population of most polluted areas
- Economic and social degradation of polluted areas: in the process of transition the most polluting branches should be restructures and their production should be

Implications on social statistics

- Basic statistical indicators characterizing social aspects of environment
    - Positives
    - Negatives
- Status of ecologically wasted areas by:
- Level of ecological degradation
    - Kinds of pollution or degradation
    - Branch structure of the economy
    - Structure of land
- Population living on ecologically wasted areas (number, percent density of population) by:
    - Level of ecological degradation
    - Kinds of pollution or degradation
    - Structure of population living in regions (age, sex, economic activity status)
    - Special indicators characterizing standard of life and health conditions of population in polluted areas.

1.9.7 Social security, crime and justice

- Processes of transition in some regions are accompanied by dramatic political and social events (military actions, ethnic disturbances, connected with violence of human and civil rights).
- Official statistics should deliver basic social indicators these processes, on population living in the areas of military, social or ethnic disturbances, displacement of population and social consequences of these events and processes.
- Attention should be paid to information on human rights violence connected with those events.
- Official statistics helps the governments to evaluate and forecast social consequences of those events for population living in the areas, to estimate the number and structure of population and to organize respective measures.

Implication on social statistics

- Priority should be given to statistics of (a) population living in the areas of conflicts, (b) number and kind of human rights violence and crimes, (c) population suffering because of security problems.
- Demographic (data on population and housing) geographic database systems (identification of territorial units) are recommended the basis for operational informing of central and regional governments on social consequences of those disturbances.
- Quality of statistics is major problem of data. Careful editing and validation of data on security, crimes and human rights violence received from administrative sources is necessary.
- Official statistics should be involved in the evaluating and disseminating reliable data on human rights violence collected by governments, NGO` and other organizations.
- Statistically measurable social consequences of the situation in security, crime, conflicts and human rights violence should be monitored: e.g. migrations incl. refugees, expenditures of households for safety.

1.9.8 Nutrition [Food and agriculture statistics]

This is based on the following phenomena:

- Polarization of the quality of nutrition of households by incomes and by incomes for disposal.
- Changes in nutrition caused by significant changes of structure of prices (cuts of government subsidies to basic food products have changed the level and structure of prices and consumption preferences).
- Regional and local diversification of food and nutrition (regions of social or economic disturbances etc.).
- Nutrition of households with low income for disposal and bellow national poverty line.
- Aggregated data on expenditures for nutrition (e.g. consumption compiled on the basis on retail sales *of per capita)* may not represent real nutrition standards of households.

Implications for social statistics

- Classification of consumption goods and services used in household surveys should represent products characterizing standard of nutrition.
- In household surveys special attention should be paid to the analysis of nutrition of population with law income for disposal, social groups below poverty line, regions of disasters, as the part of surveys on poverty and health.

ASSIGNMENT

Uses of social statistics

With relevant examples state the uses socials statistics (Hand in by 17th October 2013).

CHAPTER TWO: TECHNIQUES FOR MODELING SOCIAL DATA (MULTIVARIATE AND DISCRETE MODELS)

The focus of this chapter is to apply various techniques relevant in modeling and analysis of social data from different types & sources of social statistics. The use of models for discrete and multivariate data will be emphasized. In each of the types of social statistics (education, crime, health, food and nutrition, environment etc) various models with examples will be demonstrated.

Modelling is concept popularly used in many disciplines. It requires to follow some processes or stages in order to develop or formulate valid model based on the data and underlying assumptions. The diagram (Figure 2.1) below illustrates some of the common stages/steps followed when developing any mathematic model. This approach will employed in various sources social data.

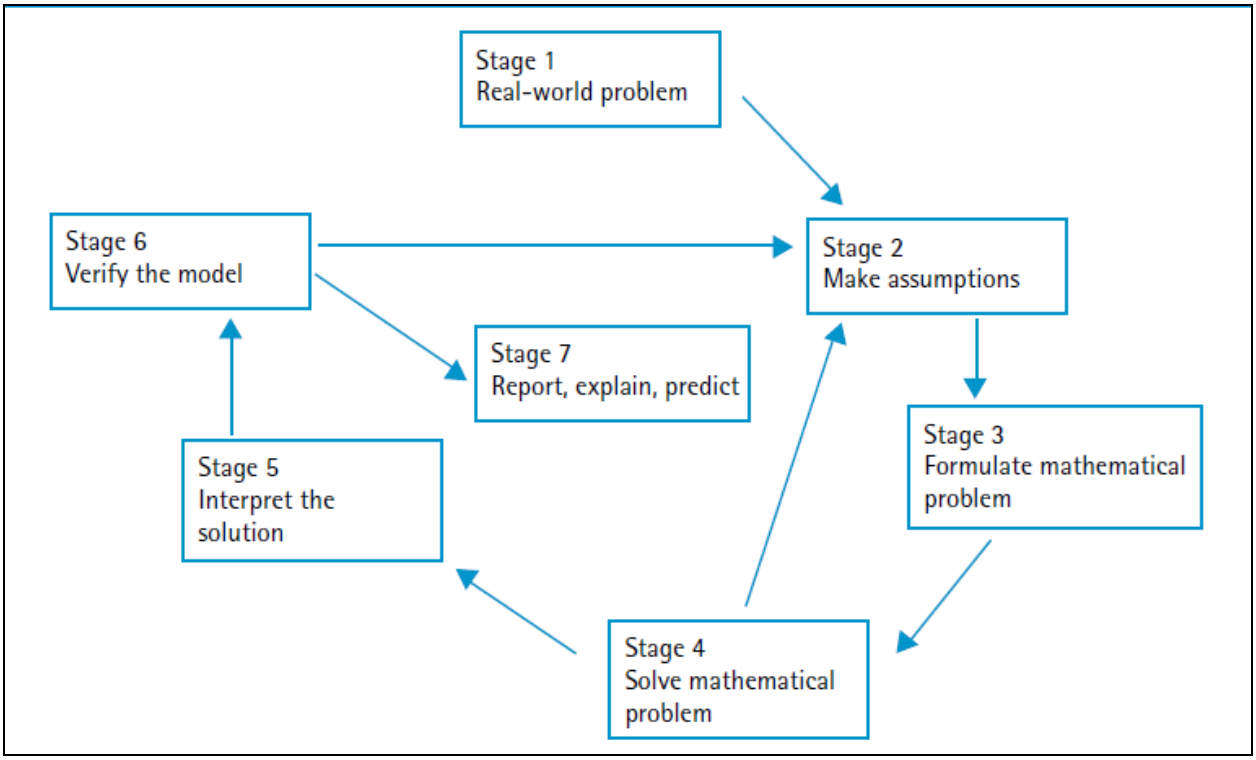


Figure 2.1. Stages involved in modeling process

Stage 1. Real world problem

The problem statement should be very general and free of as much data as possible, as later stages of the modelling process will consider and gather what is needed.

Stage 2: Making assumptions

It consists of listing all the variables involved and then trying to simplify or modify the list. In this process, it becomes obvious that there is a need to obtain certain information that will constitute the initial conditions of the problem.

Stage 3: Formulating mathematical model

This involves algebraically construction of the model with associated parameters that will yield meaningful interpretation during data analysis

Stage 4: Solving mathematical problem

This stage describes the process used when applying a procedure to given data. Using the modelling process may mean a return to the initial assumptions in order to modify the problem being considered.

Stage 5: Interpret the solution

After obtaining the solutions get back to the problem and check to ensure that model has enabled you to answer the problem within the assumptions made. Interpretations made should make explicit the assumptions and initial conditions. This is an important step in helping to realise that solutions to problems are constrained by the context and are not easily transferable to other situations.

Stage 6: Verify the model

In this stage the strengths and weaknesses of the model are discussed. This involves reflecting upon the mathematics that has been used. The statement that "all models are wrong, but some are useful" is an important reminder of the dangers of oversimplification and of ignoring the underlying assumptions. Models should be evaluated in terms of the variables used and, more importantly, those omitted.

Stage 7: Report, explain, predict

This is a valuable part of the process because it builds experience in using language to express mathematical ideas. It is here that we reflect upon the quality of thinking. It should include documentation of the progress through the stages of the cycle as well as final predictions and answers. The structure of the modelling process provides a good organizing device for the report.

2.1 Models in education statistics

Educational activity is becoming more complex and planning its development more difficult. In order to understand the function of an educational system, it is useful to

study the long term implications of the present educational structure. Mathematical modeling has gained prominence as a means of improving educational planning. At present there are many types of models, some deal with the whole system, some of the particular sectors of the system and others with specific institutions.

The functional forms of the models are similarly varied. These are;

a) Regression models

b) Stochastic models

c) Linear programming models among others

2.1.1 Regression models

Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another—the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence. The investigator also typically assesses the "statistical significance" of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship. Regression techniques have long been central to the field of economic statistics ("econometrics"). Increasingly, they have become important to lawyers and legal policy makers as well.

The following are some of the regression models in use;

a) Simple regression
b) Multiple regression
c) Log-linear regression models
d) Logistic regression model

2.1.1.1 Simple regression

This involves modeling two quantitative variables. For instance, we can regress education and earnings in various organizations. However, in reality, any effort to quantify the effects of education upon earnings without careful attention to the other factors that affect earnings could create serious statistical difficulties (termed "omitted variables bias"). But for now let us assume away this problem. We also assume, again quite unrealistically, that "education" can be measured by a single attribute—years of schooling.

We thus suppress the fact that a given number of years in school may represent widely varying academic programs.

At the outset of any regression study, one formulates some hypothesis about the relationship between the variables of interest, here, education and earnings. Common experience suggests that better educated people tend to make more money. It further suggests that the causal relation likely runs from education to earnings rather than the other way around. Thus, the tentative hypothesis is that higher levels of education cause higher levels of earnings, other things being equal.

To investigate this hypothesis, imagine that we gather data on education and earnings for various individuals. Let *E* denote education in years of schooling for each individual, and let *I* denote that individual's earnings in dollars per year. We can plot this information for all of the individuals in the sample using a two-dimensional diagram, conventionally termed a "scatter" diagram. Each point in the diagram represents an individual in the sample.

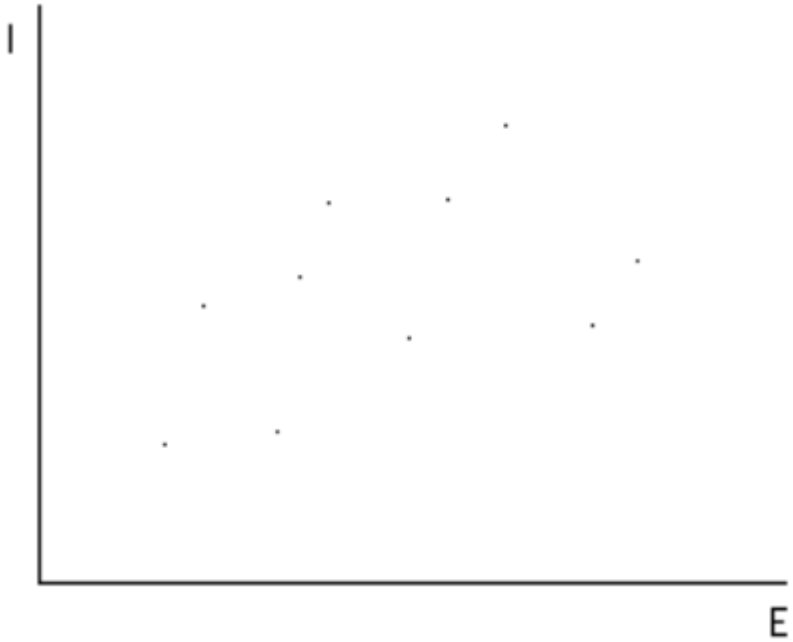


Figure 2.2 Relationship between education and earnings

The diagram indeed suggests that higher values of *E* tend to yield higher values of *I*, but the relationship is not perfect—it seems that knowledge of *E* does not suffice for an entirely accurate prediction about *I*. We can then deduce either that the effect of education upon earnings differs across individuals, or that factors other than education influence earnings. Regression analysis ordinarily embraces the latter explanation. Thus, pending discussion below of omitted variables bias, we now hypothesize that

earnings for each individual are determined by education and by an aggregation of omitted factors that we term "noise."

To refine the hypothesis further, it is natural to suppose that people in the labor force with no education nevertheless make some positive amount of money, and that education increases earnings above this baseline. We might also suppose that education affects income in a "linear" fashion—that is, each additional year of schooling adds the same amount to income. This linearity assumption is common in regression studies but is by no means essential to the application of the technique, and can be relaxed where the investigator has reason to suppose a priori that the relationship in question is nonlinear. Then, the hypothesized relationship between education and earnings may be written

$I = a + bE + e$

where
a = a constant amount (what one earns with zero education);
b = the effect in dollars of an additional year of schooling on income, hypothesized to be positive; and
e = the "noise" term reflecting other factors that influence earnings.
The variable *I* is termed the "dependent" or "endogenous" variable; *E* is termed the "independent," "explanatory," or "exogenous" variable; a is the "constant term" and b the "coefficient" of the variable *E*.

Remember what is observable and what is not. The data set contains observations for *I* and *E*. The noise component e is comprised of factors that are unobservable, or at least unobserved. The parameters a and b are also unobservable. The task of regression analysis is to produce an *estimate* of these two parameters, based upon the information contained in the data set and, as shall be seen, upon some assumptions about the characteristics of e.

To understand how the parameter estimates are generated, note that if we *ignore* the noise term e, the equation above for the relationship between *I* and *E* is the equation for a line—a line with an "intercept" of a on the vertical axis and a "slope" of b.
Returning to the scatter diagram, the hypothesized relationship thus implies that somewhere on the diagram may be found a line with the equation $I = a + bE$.
The task of estimating a and b is equivalent to the task of estimating where this line is located.
What is the best estimate regarding the location of this line? The answer depends in part upon what we think about the nature of the noise term e. If we believed that e was usually a large negative number, for example, we would want to pick a line lying above

most or all of our data points—the logic is that if e is negative, the true value of *I* (which we observe), given by $I = a + bE + e$, will be less than the value of *I* on the line $I = a + bE$. Likewise, if we believed that e was systematically positive, a line lying below the majority of data points would be appropriate. Regression analysis assumes, however, that the noise term has no such systematic property, but is on average equal to zero—I will make the assumptions about the noise term more precise in a moment. The assumption that the noise term is usually zero suggests an estimate of the line that lies roughly in the midst of the data, some observations below and some observations above.

But there are many such lines, and it remains to pick one line in particular. Regression analysis does so by embracing a criterion that relates to the *estimated* noise term or "error" for each observation. To be precise, define the "estimated error" for each observation as the vertical distance between the value of *I* along the estimated line $I = a + bE$ (generated by plugging the actual value of *E* into this equation) and the true value of *I* for the same observation. Superimposing a candidate line on the scatter diagram, the estimated errors for each observation may be seen as follows:

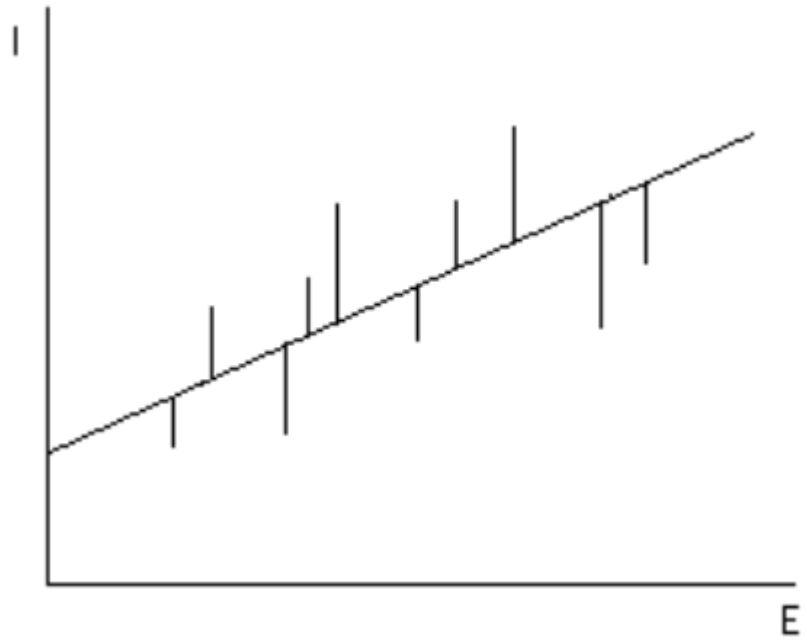


Figure 2.3 Illustration on the effect of estimate in each of the observed value

With each possible line that might be superimposed upon the data, a different set of estimated errors will result. Regression analysis then chooses among all possible lines by selecting the one for which the sum of the squares of the estimated errors is at a minimum. This is termed the minimum sum of squared errors (minimum SSE) criterion

The intercept of the line chosen by this criterion provides the estimate of a, and its slope provides the estimate of b. It is hardly obvious why we should choose our line using the minimum SSE criterion. We can readily imagine other criteria that might be utilized (minimizing the sum of errors in absolute value, for example). One virtue of the SSE criterion is that it is very easy to employ computationally. When one expresses the sum of squared errors mathematically and employs calculus techniques to ascertain the values of a and b that minimize it, one obtains expressions for a and b that are easy to evaluate with a computer using only the observed values of *E* and *I* in the data sample.

But computational convenience is not the only virtue of the minimum SSE criterion—it also has some attractive statistical properties under plausible assumptions about the noise term. These properties will be discussed in a moment, after we introduce the concept of multiple regression.

Problem: Use the data set on parent-child academic ability to formulate appropriate simple regression model and interpret the results.

2.1.1.2 Multiple regression model

"Multiple regression" is a technique that allows additional factors to enter the analysis separately so that the effect of each can be estimated. It is valuable for quantifying the impact of various simultaneous influences upon a single dependent variable. Further, because of omitted variables bias with simple regression, multiple regression is often essential even when the investigator is only interested in the effects of one of the independent variables.

For purposes of illustration, consider the introduction into the earnings analysis of a second independent variable called "experience." Holding constant the level of education, we would expect someone who has been working for a longer time to earn more.

Let *X* denote years of experience in the labor force and, as in the case of education, we will assume that it has a linear effect upon earnings that is stable across individuals. The modified model may be written:

$I = a + bE + gX + e$

The task of estimating the parameters a, b, and g is conceptually identical to the earlier task of estimating only a and b. The difference is that we can no longer think of regression as choosing a line in a two-dimensional diagram—with two explanatory variables we need three dimensions, and instead of estimating a line we are estimating a plane. Multiple regression analysis will select a plane so that the sum of squared errors—

the error here being the vertical distance between the actual value of *I* and the estimated plane—is at a minimum. The intercept of that plane with the I-axis (where *E* and *X* are zero) implies the constant term a, its slope in the education dimension implies the coefficient b, and its slope in the experience dimension implies the coefficient g.
Multiple regression analysis is in fact capable of dealing with an arbitrarily large number of explanatory variables. Though people lack the capacity to visualize in more than three dimensions, mathematics does not. With *n* explanatory variables, multiple regression analysis will estimate the equation of a "hyperplane" in n-space such that the sum of squared errors has been minimized. Its intercept implies the constant term, and its slope in each dimension implies one of the regression coefficients. As in the case of simple regression, the SSE criterion is quite convenient computationally.

Formulae for the parameters a, b, g . . . can be derived readily and evaluated easily on a computer, again using only the observed values of the dependent and independent variables.
The interpretation of the coefficient estimates in a multiple regression warrants brief comment.

In the model $I = a + bE + gX + e$,

a -captures what an individual earns with no education or experience,
b- captures the effect on income of a year of education, and
g -captures the effect on income of a year of experience.

To put it slightly differently, b is an estimate of the effect of a year of education on income, holding experience constant. Likewise, g is the estimated effect of a year of experience on income, holding education constant.

Example: Wage, Education and Experience
In this case one could think of wages as a function of education and work experience and formulate as follows:

Wage = *f* (*Education*, *Experience*).
This is because education is not the only factor that affects pay as even for workers with the *same* education, there is remarkable variation in wages. Some of this variation is due to work experience, unionization, industry, occupation, region, and demographics, such as gender, race, marital status, etc. These easily can be accounted for using multiple regression.

The description below provides an illustration and interpretation of multiple regression model

The longer one spends on a job, the better one gets. If people are paid for their productivity, then workers with more work experience should be more productive, and therefore, paid more. That is,

$$\frac{\Delta Wage}{\Delta Experience} > 0,$$

other things equal.

The complete relationship between wages, education, and experience can be written as

$$\ln(Wage_i) = \beta_1 + \beta_2 Education_i + \beta_3 Experience_i + u_i, \qquad (1)$$

where wages are measured in natural logs. This is a multiple regression model of wages. Because there is more than one explanatory variable, each parameter is interpreted as a partial derivative, or the change in the dependent variable for a change in the explanatory variable, holding all other variables constant. For example,

$$\beta_3 = \frac{\partial \ln(Wage)}{\partial Experience} \approx \frac{\Delta \ln(Wage)}{\Delta Experience}\bigg|_{Education} \qquad (2)$$

is the effect of experience on the log wage, *holding education constant*. Other ways of saying "holding experience constant" are "controlling for experience" or "accounting for the effect of experience." Because pay is measured in natural logs, β_3 also can be interpreted as

$$\beta_3 = \frac{\%\Delta Wage}{\Delta Experience}\bigg|_{Education}, \qquad (3)$$

or the "return to experience" in the labor market.

If we group all workers according to their education level (less than high school, high school, some college, college graduates, and more than college), we can compare wages and work experience *within* education categories. This is really what multiple regression does. By looking *within* categories, you are *holding education constant*. From the univariate analysis in we know that wages increase with educational level. Table 5.1 shows that within any education category (i.e reading across rows), hourly wages rise with greater work experience. This suggests β_3 is positive, so that wages increase with work experience controlling for education, but also work experience explains some of the residual variation in wages within education levels.

Table 5.1

Means, Standard Deviations and Frequencies of Hourly Wages

Years of Education	Years of Work Experience					
	exp<=5	5<x<=10	10<x<=20	20<x<=30	exp>30	Total
Educ<12	6.610577	8.3096154	8.506556	8.6632116	11.499039	9.310918
	2.2335515	4.5823646	3.2871646	3.8071621	9.4367496	6.0386959
	4	10	22	25	25	86
Educ=12	8.5617234	10.222842	11.647422	15.137898	13.069812	12.522641
	3.8917755	6.1940197	7.0772693	7.2318973	6.6605462	6.9705726
	26	45	102	93	96	362
Educ=13	6.0346955	11.595442	12.601342	17.252274	16.064233	13.730448
	2.2878627	5.0236677	6.807096	8.68351	9.4877654	8.0749169
	18	27	67	52	38	202
13<Educ<=16	12.018377	11.547343	19.680886	18.701486	18.666967	16.868053
	5.1686776	4.4477838	10.56787	8.6071216	12.906913	9.5829901
	40	38	78	66	31	253
Educ>16	17.574786	22.328942	28.116649	22.697912	26.153953	24.389087
	6.5705128	11.500545	13.368682	10.918406	10.881327	11.893388
	9	15	35	32	9	100
Total	10.274019	12.073586	15.587707	16.724456	14.907941	14.769702
	5.5195622	7.2312446	10.452645	8.8239055	9.4999288	9.257249
	97	135	304	268	199	1003

Likewise,

$$\beta_2 = \frac{\partial \ln(Wage)}{\partial Education} \approx \frac{\Delta \ln(Wage)}{\Delta Education}\bigg|_{Experience} \tag{4}$$

is the effect of education on the log wage, holding experience constant. β_2 also can be expressed as

$$\beta_2 = \frac{\%\Delta Wage}{\Delta Education}\bigg|_{Experience}, \tag{5}$$

or the return to education in the labor market.

If we group workers according to years of work experience (0-5, 5-10, 11-20, 21-30, >30), we can compare wages and education *within* work experience categories. Again, this is what multiple regression does. By looking *within* experience categories, we are *holding experience constant*. In Table 5.1, within any given experience category (reading down columns), the hourly wage rises with education. This suggests β_2 is positive, so that wages increase with education even when controlling for work experience.

Importantly, multiple regression recognizes possible *interdependence* among explanatory variables. For example, for any individual, education and work experience are determined *in part* by the underlying decision to allocate time. Individuals can go to school or work. Those with more education will have less work experience, and vice versa, holding other factors such as age constant. Thus, education and experience are interdependent. In fact, they are inversely correlated since the sample correlation coefficient, r= –0.186.

This interdependence implies that some of the population variation in education and experience is common. The Venn diagram in Figure 5.1 illustrates this. The two circles represent the variation in education and experience, respectively. Area B is the intersection and represents the variation shared by the variables. This is the co-variation between education and experience. Area A is the remaining variation in education and is due to influences other than experience, and hence, is *independent* of experience. Similarly, area C is the remaining variation in experience, *independent* of education.

When estimating parameters, least squares uses only the *independent* variation in each explanatory variable to estimate that variable's parameter. To estimate β_2, only the independent part of education is used. The formula for the least squares estimator of β_2 is

$$\hat{\beta}_2 = \frac{\widehat{Cov}(\ln(Wage), \text{Independent Part of } Education)}{\widehat{Var}(\text{Independent Part of } Education)}. \qquad (6)$$

and for β_3,

$$\hat{\beta}_3 = \frac{\widehat{Cov}(\ln(Wage), \text{Independent Part of } Experience)}{\widehat{Var}(\text{Independent Part of } Experience)}. \qquad (7)$$

Table 5.2 shows parameter estimates, standard errors and 95% confidence intervals for simple and multiple regression models of the log wage.

Table 5.2 Regression of Log Wages against Education and Experience

Explanatory Variable	(1)	(2)	(3)
Education	0.0933		0.1035
	(0.0067)	----	(0.0066)
	(0.0801, 0.1064)		

Experience	----	0.0084	0.0129
		(0.0017)	(0.0015)
		(0.0051, 0.0117)	(0.0098, 0.0159)
Constant	1.2597	2.3456	0.8629
	(0.0919)	(0.0389)	(0.1008)
	(1.0793, 1.4399)	(2.2692, 2.4220)	(0.6651, 1.0607)
R^2	0.162	0.024	0.217

For the simple regression model 1,

$$\ln(Wage_i) = \beta_1 + \beta_2 Education_i + u_i,$$

an additional year of education is estimated to raise log wages by 0.0933 or in terms of relative change in wages, by a factor of exp(0.0933)=1.098 with a 95% confidence interval of (exp(0.0801), exp(0.1064)) = (1.08, 1.11). Economists often say that the increase in percent wages is 9.3%, an approximation. This is a moderately good return to a one-year investment!

Alternatively, for model 2,

$$\ln(Wage_i) = \beta_1 + \beta_2 Experience_i + u_i,$$

an additional year of experience is estimated to raise the log wage by 0.0084 or to raise the wage by a factor of exp(0.0084)=1.0084 or 0.84%. To put this finding in a more meaningful context, an additional 10 years of experience raises the log wages by 0.084, or raises wages by a factor of exp(0.084)=1.088 or 8.8%. R^2 is 0.024, which means that variation in experience alone explains just 2.4% of the sample variation in log wages.

The estimates for the multiple regression model 3,

$$\ln(Wage_i) = \beta_1 + \beta_2 Education_i + \beta_3 Experience_i + u_i,$$

show that together, education and experience explain 21.7% of the variation in log wages. This is much more than both explain individually (16.2% and 2.4%). So, the whole is greater than the sum of its parts!

Accounting for the effect of experience on wages, an additional year of education is estimated to raise the log wage by 0.1035 or the actual wage by a factor of exp(0.1035)=1.109 or 10.9% (Economists 10.4%). In addition, accounting for the effect of education on wages, an additional year of experience is estimated to raise the log wage by 0.0129 or wages by a factor of 1.013 or 1.3%. Surprisingly, the return to an additional year of experience is significantly less than the return to an additional year of education. In fact, based on these estimates, it would require an additional 8.4 years of work experience to raise wages by the same percent as an additional year of education (10.9/1.3=8.4). Education seems like a good deal!

Interestingly, the estimated effects of education and experience on wages change substantially from simple to multiple regression. An additional year of education is estimated to raise the wages by 9.8% in model (1) but by 10.9% in model (3). That is, the estimated return rises by more than a percentage point once differences in work experience are taken into account.

Because this is a big difference in the return on an investment---you would much prefer a 10.4% to a 9.3% return---it is natural to ask: "Why did this happen?"

The answer is at the heart of multiple regression. There are many less-educated (but more-experienced) workers that earn as much as more-educated (but less-experienced) workers. Without accounting for differences in experience, the better educated appear to get a lower return to education. Is this "low return" *really* because of education? No, it is because of experience.

Simple linear regression does not account for experience; however, multiple regression does. Once differences in experience across workers are taken into account, an additional year of education has a much bigger payoff and the estimated return to education rises. Because education and experience are correlated (or *interdependent*), simple regression confuses or "confounds" the effect of education on wages with the effect of experience on wages. By acknowledging potential correlation between the explanatory variables, multiple regression neatly sorts out each variable's independent effect. Section 6 will discuss "confounding effects" in more detail.

5.2. Gender and Wages

A question of great public interest is whether there is gender inequality in earnings, and, if so, what accounts for it. The basic comparison of *average* wages for men and women in section 3 showed that women earn $4.90 per hour less than men. Because men's average earnings were $17.05, this implies that, on average, women earn about 28.7% less than men (4.90/17.05=0.287).

If pay is based solely on productivity, then this differential could be economically rational only if there were some innate underlying difference in productivity between the sexes. In this case, men would have to be more productive than women to justify their higher wages. If one believes the sexes are equal, then gender difference in wages must be caused by something else. One view is that there is labor market discrimination against women. Another is that there are other, *confounding* factors that affect wages but happen to be correlated with gender. Multiple regression can account for these additional factors. If gender-based wage differentials exist even after controlling for many possible confounding influences, then more credence might be given to the discrimination explanation.

The effect of gender on wages can be modeled simply as

$$\ln(Wage_i) = \beta_1 + \beta_2 Female_i + u_i, \qquad (8)$$

where *Female* is an indicator variable that is 1 if the worker is female and 0 otherwise (which, of course, means male). A 0-1 indicator variable such as this is frequently referred to as a categorical or dummy variable. β_2 is the relative change in the wage from going from the 0 category to the 1 category. The way to think about β_2 in this context is to observe the wage of worker first as a male, and then "transform" the worker into a female and observe the wage. If the wage rises, β_2 is positive and women earn more than men. That is, there is a labor marker "premium" to being a woman. If the wage does not change, β_2 is zero and gender has no effect on wages. Finally, if the wage falls, β_2 is negative and men earn more than women. That is, there is a labor market discount to being a woman.

Based on the comparison of mean wages above, one expects β_2 to be negative. The least squares estimate is shown in column (1) of Table 5.3.

Table 5.3 Regression of Log Wages against Female, Education and Experience

Explanatory Variable	(1)	(2)	(3)
Female	0.0933 (0.0067) (0.0801, 0.1064)	----	0.1035 (0.0066)
Education	0.0933 (0.0067) (0.0801, 0.1064)	----	0.1035 (0.0066)
Experience	----	0.0084 (0.0017) (0.0051, 0.0117)	0.0129 (0.0015) (0.0098, 0.0159)
Constant	1.2597 (0.0919) (1.0793, 1.4399)	2.3456 (0.0389) (2.2692, 2.4220)	0.8629 (0.1008) (0.6651, 1.0607)
R^2	0.162	0.024	0.217

$\hat{\beta}_2 = -0.3193$, which says that females earn 31.9% less than males. The 95% confidence interval is (-0.3918, -0.2468) and does not include zero. Therefore, this discount to being female is statistically significantly different from zero.

In theory, one explanation for this wage differential could be differences in education between men and women. If men had more education than women on average, then that could explain the gender differential. Unfortunately, this is not the case. From section 3 above, the sample mean education is 13.48 years for men, but 13.42 years for women. That is, men and women have *almost identical* education levels! In fact, the sample correlation between *Education* and *Female* is -0.01. Effectively, the variables are uncorrelated. Therefore, differences in education will not explain the gender difference in wages.

This can be seen in two ways. First, Table 5.4 compares mean hourly wages for males and females *within* education categories. Remember, this is akin to what least squares does to estimate the effect of *Female* on wages holding *Education* constant in a multiple regression.

Table 5.4

Means, Standard Deviations and Frequencies of Hourly Wages

Years of Education	Gender		
	Male	Female	Total

Educ<12	10.609443	7.225408	9.310918
	6.9104135	3.46186	6.0386959
	53	33	86
Educ=12	14.280749	10.601934	12.522641
	7.5401576	5.7210515	6.9705726
	189	173	362
Educ=13	16.397839	10.955284	13.730448
	8.9813645	5.8753599	8.0749169
	103	99	202
13<Educ<=16	19.477352	14.110256	16.868053
	10.390313	7.7854763	9.5829901
	130	123	253
Educ>16	26.977737	20.165499	24.389087
	13.00519	8.371838	11.893388
	62	38	100
Total	17.048445	12.14377	14.769702
	10.240742	7.132306	9.257249
	537	466	1003

It is easy to see that *within* all categories, males are paid more females. Therefore, education cannot explain the gender wage differential.

Second, column (2) of Table 5.3 gives parameter estimates for the multiple regression model

$$\ln(Wage_i) = \beta_1 + \beta_2 Female_i + \beta_3 Education + u_i . \quad (9)$$

β_2 measures the effect of being female on wages (in relative terms), *controlling for education*. $\hat{\beta}_2 = -0.3137$, which says that females earn 31.4% less than males, even after accounting for any differences in education between the sexes! This estimate is virtually unchanged from its value in column (1). The 95% confidence interval is (-0.3797, -0.2477) and does not include zero. Therefore, this discount to being female is statistically significantly different from zero.

The estimated relationship between wages, education, and gender also can be illustrated graphically. Figure 5.1 demonstrates that regressing log wages on education and gender has the effect of fitting separate parallel lines to the relationship between log hourly wages and education for males and females. Parallel lines mean that the increase in log wages for an additional year of education (or the return to education) is the same for males and females, and averages about 0.093 (or 9.3%). The distance between the lines for males and females represents the effect of gender: the line for males is 0.3137 log dollars (or 31.37%) higher than the line for females. Figure 5.2 shows the same relationship, but expressed in terms of the wage rather than the log wage.

Another potential explanation for the gender wage differential could be differences in work experience between men and women. If men had more experience than women on average, then that could explain the gender differential. Again, this is not the case. From section 3, the sample mean experience is 19.72 years for men, but even more, 20.63 years, for women. That is, women have slightly more experience than men. The sample correlation between *Experience* and *Female* is 0.041. Effectively, the variables are uncorrelated. Therefore, differences in experience will not explain the gender difference in wages.

This is illustrated nicely in column (2) of Table 5.2, which gives parameter estimates for the multiple regression model

$$\ln(Wage_{ie}) = \beta_1 + \beta_2 Female_i + \beta_3 Education + \beta_4 Experience + u_i . \quad (9)$$

β_2 measures the effect of being female on wages (in relative terms), *controlling for education and experience*. $\hat{\beta}_2 = -0.3252$, which says that females earn 32.5% less than males, even after accounting for any differences in education and experience between the sexes. The 95% confidence interval is (-0.3887, -0.2617) and does not include zero. Therefore, this discount to being female is statistically significantly different from zero. The R^2 is 0.289, which means that variation in education, experience, and gender explains 28.9% of the sample variation in log wages.

2.1.2 Stochastic models

A stochastic process is one in which evolves with time or space according to probabilistic laws. This means that we cannot predict its future behavior with certainty, the most that we can do is to attach probabilities to the various possible future outcomes.

One of the early Markovian models for educational planning was proposed by Gani (1963), who used it to forecast enrolment and degrees awarded in Australian universities. Since then many similar models have been discussed. Among the more substantial contributions was that of Thonstad (1969), who in his book on educational planning makes extensive use of stochastic models. Other contributions include that of Uche (1980) who applied the Markovian model to the Nigerian educational system.

Education can be considered as a hierarchical organization. Students usually remain in a given school grade for one academic year and then move to the next grade, repeat the same grade or leave the system as graduates or dropouts. This is the basic idea of the Markovian model. When students graduate and leave the system, or when they drop out due to illness, death or poor academic performance, then situation is akin to transition into an **absorbing state**. Transition between grades is similar to that between **non-absorbing states.** Consequently we shall use an absorbing Markov Chain. Then grades and final educations form the states of the process.

Example: Transition probabilities in old primary education system in Kenya

In this case we shall concern ourselves with the estimation of the school staying ratios, the dropout and completion ratios, the expected length of schooling, the school survival time and the cost of educating an individual upto completion. We shall accomplish this by studying a grade cohort, which means a group of pupils, regardless of age, entering a certain grade at school in a given year.

THE MODEL

Education is one of the systems which exist only in one of a finite number of states and change only at discrete point of time. The states of the education system will be partitioned into two, non-absorbing (corresponding to the various final educations, mortality and others).

Let the states of the education system be denoted by integers 1, 2, 3 ,...N, where N is the number of possible states of the education system. We shall assume that the education system is time homogenous so that there is a fixed probability P_{ij} that a student in state i at time $(t-1)$ will transfer to state j at time t. this will give rise to the transition matrix

$$P = (P_{ij}),\ i, j, = 1, 2, \ldots, N \qquad (1)$$

Suppose we now define $n_{ij}(t)$ to be the number of students in state i at time $(t - 1)$ who move to state j at time t and $n_i(t)$ to be the number of students in state i at time t.

Then $\sum_{J=1}^{N} n_{ij}(t) = n_{i}\ (t-1)$ and each $n_{ij}(t)$ is associated with the probability P_{ij} of moving from state i at time $(t - 1)$ to state j at time t, so that $\sum_{j=1}^{N} P_{ij} = 1.$ Assuming the multinational probability distribution, the transition probabilities may be estimated from

$$P_{ij} = \frac{n_{ij}(t)}{n_i\ (t-1)}, i, j = 1, 2, \ldots\ldots, N \qquad (2)$$

Which is the proportion of students in state I at time (t – 1) who move to state j at time t.

If we further assume that the system has r absorbing and s non-absorbing states, such that $r+s$ = N, then the transition matrix will have the canonical form,

$$P = \begin{bmatrix} I & O \\ R & Q \end{bmatrix} \qquad (3)$$

Where,

I is an $r \times r$ identity matrix giving transition probabilities between absorbing states,

O is an $r \times s$ matrix of zeros giving transition probabilities from absorbing to non-absorbing states,

R = (r_{ik}), is an $s \times r$ matrix r_{ik} being the probability that a student in grade the probability that a student in grade I at time $(t - 1)$ will graduate with final education k at time $t, i = 1, 2, \ldots.s$ and $k = 1, 2, \ldots.r$, *and*

Q = (q_{ij}) , is an $s \times s$ matrix with q_{ij} as the probability that a student in grade i *at* time $(t - 1)$ will be in grade j at time $t, i, j = 1,2,\ldots.s$.

The i-th diagonal element of Q, q_{ij}, is the probability of a student is in one of the r final educations we shall say that he or she is absorbed so that he or she will not leave that state.

The fundamental matrix

An important property of an absorbing Markov chain is that the probability of the system being absorbed tends to one as the number of trails gets large. That is, the elements of the matrix R tends to one as the number of trails increase. Thus $\lim_{n \to \infty} Q^{\square} n = \mathbf{0}$ and the matrix series $I + Q + Q^2 + \ldots$ is convergent with sum $(I - Q)^{-1}$, which we shall denote by L. the matrix

$$L = (I - Q)^{-1} \qquad (4)$$

is known as the fundamental matrix of the absorbing Markov chain.

APPLICATIONS OF THE MODEL

In this section we apply the model described in the preceding section, to obtain expressions for the school staying ratio, the drop out and completion ratios, the expected length of schooling, the school survival time and the cost of educating an individual up to completion.

The expressions so obtained are then used to calculate numerical results, based on the primary school system in Kenya, which consists of seven school grades (up to 1984) called standards as the non-absorbing states of the education system. The numerical results presented are purely for illustrative purposes. The details of computation may be found in Owino (1982).

The school staying ratio

The probability that a student now in any school grade i will be in another school grade j, n years later, $n = 0,1,2,\ldots$.will be

$$Q_{ij}^{(n)}, i,j = 1,2,\ldots,s,$$

Which is the (i,j)-th entry of Q^n. This is interpreted as the fraction of students now in school grade I who will be in school grade j after n years. In particular

$$Q_{ij}^{(o)} = 1, i = j \qquad (5)$$

$$= 0, \text{otherwise}$$

Therefore the probability that a student now in school grade i will still be in any of the s school grades n years later is the sum.

$$Q_i^{(n)} = \sum_{j=1}^{s} q_{ij}(n)^{\square}, i, 2, \ldots, s \qquad (6)$$

Which is the *i*-th entry of the column vector $Q^{(n)}$ 1 where 1 is an ($s \times 1$) column vector of ones. This is interpreted as the fraction of students now in school grade some *n* years later and is called the school staying ratio.

Table 1 shows the school staying ratios for boys (B) and Girls (G).

The school drop-out and completion ratios

The probability for a student now at the beginning of school grade *i* to graduate *n* years later with final education k is given by

$$B_{ik}^{(n)} = \sum_{j=1}^{s} q \; [\![_{ij} n - 1^{\square} \qquad i = 1,2,\ldots..,s \qquad (7)$$

$$k = 1,2,\ldots.,r$$

where,

$q_{ij}^{(n-1)}$ is the probability for a student in school grade *i* to be in school grade *j* after (n-1) years

and r_{jk} is the probability for a student now in school grade j to graduate with final education *k* in the next year.

It is clear that $b_{ik}^{(n)}$ is the (*i,k)*-th entry of the product $Q^{n-1} R$, n= 1,2,.... It is called the drop out ratio. Summing the drop out ratios for *n* equal to one to *x* gives us the probability for a student now in grade *I* to graduate with final education *k* within *x* years, which we denote by $b_{ij}^{(x)}$i.e

$$B_{ik}^{(x)} = \sum_{n=1}^{x} b_{ik}(n)^{\square} \quad i = 1, 2, \ldots.,s \text{ and}$$

$$k = 1,\ldots,r \qquad (8)$$

Table 1. Fraction of pupils now in standard 1 who n years later will be in standard j, j = 1,2, ...,7; n – 1,2,....,10; and the school-staying ratio.

Year	Grade	1	2	3	4	5	6	7	Staying ratio
1	B	0.1346	0.6588						0.7934
	G	0.128	0.6594						0.7874
2	B	0.0181	0.1784	0.5458					0.7423
	G	0.0164	0.1723	0.5507					0.7394
3	B	0.0024	0.0362	0.2187	0.4616				0.7189
	G	0.0021	0.0338	0.2154	0.4715				0.7228
4	B	0.0003	0.0065	0.0584	0.2448	0.3789			0.6889
	G	0.0003	0.0059	0.0562	0.2479	0.3891			0.6994
5	B		0.0011	0.013	0.0812	0.249	0.3246		0.6689
	G		0.001	0.0122	0.0815.	0.2572	0.3226		0.593
6	B		0.0002	0.0026	0.0215	0.0982	0.2622	0.2538	0.6385
	G		0.0002	0.0024	0.0214	0.102	0.267	0.2196	0.6126
7	B			0.0005	0.005	0.0301	0.1236	0.2468	0.406
	G			0.0004	0.0049	0.0315	0.1291	0.2147	0.3806
8	B			0.0001	0.0011	0.0079	0.0444	0.1373	0.1908
	G			0.0001	0.001	0.0083	0.0476	0.12	0.177
9	B				0.0002	0.0019	0.0135	0.0573	0.0729
	G				0.0002	0.002	0.0148	0.0504	0.0674
10	B					0.0004	0.0036	0.02	0.024
	G					0.0004	0.0041	0.0177	0.0222

~

In fact $b_{ik}^{(x)}$ is the (i,j) –th entry of the matrix

$$(I+ Q+ Q^2+ ....Q^{X-1}) R \qquad (9)$$

It is called the school completion ratio and is an important parameter in manpower planning.

The probability that a student now in school grade *i* will sooner or later graduate with final education *k* is therefore the infinite sum given by

~

$$b_{ik} = \sum_{n=1}^{\infty} b_{ik}(n) \quad i = 1,2,\ldots,s \qquad (10)$$

$$k = 1,2,\ldots,r$$

and is the (i,k)- th entry of the matrix L.R where $L = (I-Q)^{-1}$ is the fundamental matrix of the absorbing Markov chain, represented by the transition matrix P. We shall call b_{ik} the school absorbing ratio.

Table 2 shows the school dropout ratios for boys and girls, for example, the proportion of boys now in grade 4 who will drop out of school in 5years time is 0.2681, which is the first entry in the row marked B of the (5, 4)th square. The corresponding proportion for girls is 0.2457.

The completion and absorbing ratios can be obtained from table 2 by comulation. For instance, the proportion of boys who will leave the school after completing grade 3, within four years, is 0.1393. This value is obtained as the sum of the first four entries in the column marked 3, corresponding to boys (B)

Table 2. Drop out ratios for boys (B) and girls (G) in year n, n = 1,2,.....,10

Year (n) \ Grade		1	2	3	6	5	6	7
1	B	0.2066	0.0353	0.0245	0.0498	0.0165	0.0674	0.8353
	G	0.2126	0.0325	0.0127	0.0404	0.0357	0.1526	0.8502
2	B	0.0511	0.0251	0.0453	0.02	0.0598	0.6633	0.1376
	G	0.0486	0.0149	0.0363	0.0349	0.1313	0.6042	0.1274
3	B	0.0234	0.0409	0.0228	0.0517	0.5759	0.2075	0.0227
	G	0.0161	0.0323	0.0346	0.1131	0.5186	0.1874	0.0191
4	B	0.0301	0.0245	0.0467	0.4794	0.2508	0.049	0.0037
	G	0.0233	0.0332	0.1014	0.4432	0.2255	0.0442	0.0029
5	B	0.0202	0.042	0.4115	0.2681	0.0737	0.3246	0.0006
	G	0.0249	0.089	0.3931	0.2457	0.0672	0.3226	0.0004
6	B	0.0304	0.3466	0.2801	0.0953	0.0182	0.2622	0.0001
	G	0.0619	0.3398	0.2617	0.0884	0.0168	0.267	0.0001
7	B	0.2325	0.2793	0.117	0.0273	0.004	0.0004	
	G	0.232	0.2636	0.1098	0.0258	0.0038	0.0004	
8	B	0.2153	0.135	0.0383	0.0069	0.0008	0.0001	
	G	0.2035	0.1268	0.0364	0.0066	0.0008	0.0001	

	B	0.1179	0.05	0.0108	0.0016	0.0002		
9	**G**	0.1097	0.0473	0.0104	0.0016	0.0002.		
	B	0.0489	0.0157	0.0027	0.0003			
10	**G**	0.0452	0.015	0.0027	0.0003			

The school survival time

Let $q_{ij}^{(k)}$ be the probability of going from school grade i to school grade j in exactly $k$yaers. We note that $q_{ij}^{(k)}$ is the (i,j) –th element of the matrix Q^k. for k = 0 we define $Q^0 = 1$, since after zero years the system remains where it is. If T_n is the number of years a student spends in grade j during the first n years of schooling, if he is initially in grade i, then $E(T_n) = \sum_{k=0}^{n} q(k)_{ij}$

Which is the (i,j)-th element of the series $I+ Q+ Q^2+ ....Q^n$ Now, let $\ell_{ij} = \lim_{n\to\infty} E(T_n)$

Then $\ell_{ij}$is the (i,j) –th entry of the fundamental matrix L is given in question (4). Thus $\ell_{ij}$is the expected length of time , in years, spent in school grade j by those students initially in school grade i.

The expected number of school yaers left for those students now entering grade i, before graduating with any of the i-th row of the fundamental matrix given in equation (4).

Table shows the survival times for boys and girls.

The expected length of schooling (ELS)

The probability distribution of students in the various grades at time t is given by

$P(t) = (p_1(t), p_2(t),.....p_n(t))'$ (11)

Where $p_i(t)$ is the probability of a student being in grade i at time t, i = 1,2,....,N.

Table 3. The expected length of time, in years, spent in grade j by a student now in grade I and the school survival times

Grade i	Grade j	1	2	3	4	5	6	7	Survival time
1	B	1.1555	0.8813	0.8391	0.8151	0.7661	0.7728	0.7234	5.9533
	G	1.1468	0.873	0.8368	0.8285	0.7907	0.7866	0.6298	5.8922
2	B		1.1577	1.1022	1.0707	1.0063	1.0152	0.9503	6.3024
	G		1.1538	1.1066	1.0956	1.0457	1.0403	0.8329	6.2749
3	B			1.1491	1.1163	1.0492	1.0585	0.9908	5.3639
	G			1.1497	1.1383	1.0864	1.0808	0.8653	5.3205
4	B				1.1486	1.0796	1.0891	1.0195	4.3368
	G				1.1551	1.1565	1.1505	0.9211	4.3832
5	B					1.1451	1.1552	1.0813	3.3816
	G					1.1565	1.1505	0.9211	3.2281
6	B						1.1774	1.1022	2.2796
	G						1.2001	0.9608	2.1609
7	B							1.1972	1.1972
	G							1.176	1.176

B = boys
G = girls

Assuming the multinomial distribution on n_1 (t), n_2 (t),....,n_N(t), we can estimate p_i (t) from

$$\hat{p}_i(t) = \frac{n_i(t)}{\sum_{i=1}^{N} n_i(t)} \quad ,i = 1,2,.....,N \qquad (12)$$

When t = 0, then p (0) = (p_1(0), p_2(0)....., p_N(0)′, is the initial probability vector. The initial probability vector may be partitioned, writing the absorbing states first, as follows.

$$\underline{p}'(0) \ (u'(0), q'(0)) \qquad (13)$$

Where $\underline{u}(0) = (u_1(0), u_2(0),....., u_r(0))'$, so that u_k (0) is the probability of a student being in the final education category K′ at some initial time t = 0, for K= 1,2,...,r and q(0)= (q_1(0), q_2(0),...., q_s(0))′, such that q_1(0), is the probability of a student being in school grade I at some initial time t = 0, for I = 1,2,...,s.

The length of stay in any school grade j, may be considered as a random variable which takes values $\ell_{1j}, \ell_{2j,\ldots..}, \ell_{sj,}$ for j=1,2,….,s, with respective probabilities $q_1(0), q_2(0),\ldots., q_s(0)$, where $\ell_{ij}$is the (i,j)-th entry of the fundamental matrix (4). The expected length of stay in school grade j before graduation, by student who is in any of the school grades at some initial time, t = 0 is then given by $q'(0)\ell_{j,}$ where $\underline{\ell}_j = (\ell_{1j}, \ell_{2j,\ldots..}, \ell_{sj})'$. The expected length of stay in school before completion by any student who is in any of the school grades at some initial time is therefore given by $q'(0)L\ \underline{1},$ where $\underline{1}$ denotes an $s_x\ell$ column vector of ones and L is the fundamental matrix given in equation (4). This is the expected length of schooling (ELS) of a student picked at random from the education system. It is a good indicator of the sustenance property of the education system.
Table 4 shows, for illustrative purpose only, the expected length of stay in school grade j and the expected length of schooling (ELS). For instance, the expected length will spend in grade 5 is 0.770 of a year for a boy and 0.8167 of a year for a girl. These values are the entries in the column marked 5.

Table4. The expected length of stay in school grade j and the expected length of schooling, for a student in the education system at some initial time, t = 0

Grade / Sex	1	2	3	4	5	6	7	ELS
Boys	0.3015	0.4108	0.548	0.6809	0.777	0.9188	0.9595	4.5965
Girls	0.308	0.4153	0.5572	0.7048	0.8167	0.9445	0.8376	4.5841

The cost of educating an individual upto completion
The cost educating is an important factor for planning purposes. To estimate the cost from the model described above we proceed as follows. Suppose that the yearly cost is represented in vector form as

$c = (c_1, c_{2,} \ldots.,c_s)'$ (14)

Where $c_j$is the cost of educating an individual in school grade j per academic year, j = 1,2,….,s. we know from the above that the average time spent in grade j by an individual now in grade i is ℓ_{ij} which is the (i,j)-th entry of the fundamental matrix $L = (I-Q)^{-1}$. Therefore the cost of educating an individual who is now in grade I within grade j is $c_j\ell_{ij}$ and so the expected cost of educating an individual now in grade I upto graduation will be the sum

$$C^*_i = \sum_{j=1}^{s} c_j\ell_{ij} \quad i=1,2,\ldots.,s \qquad (15)$$

Next, let $\underline{n}(t) = (n_1(t), n_2(t),....,n_s(t))'$ be the enrolment vector at time t. then the total number of student years spent in grade j by those in the school system at time t is the sum
$\Sigma^s_{i=1}\ell_{ij}n(t)$.

Therefore the total cost of educating the lot of students in the school system at time t is given by

$$C(t) = \sum_{j=1}^{s}\sum_{i=1}^{s} c_j\, n_i(t)\ell ij \qquad (16)$$

Let $\hat{\underline{n}}(t+h) = (\hat{n}_1(t+h), \hat{n}_2(t+h),...., \hat{n}_s(t+h))'$ denote the expected enrolment vector at time (t+h) given the enrolment at time t as n(t) i.e.

$\hat{n}_k(t+h) = E[n_k(t+h)/\underline{n}(t)]$, k=1,2,.....,s.

Then the expected total cost op educating all the students who will be enrolled in the school system at time (t+h) upto graduation is given by

$$C(t+h) = \sum_{j=1}^{s}\sum_{i=1}^{s} c_j\, n_i(t+h)\ell ij \quad , h=1,2,.... \qquad (17)$$

This follows if we assume homogeneity and stability in the transition process of the educating system.
Table 5 shows the cost of educating an individual upto completion based on the primary school system in Kenya. The cost is currently approximately Ksh. 20 per child per year including the teacher salaries, or approximately Ksh. 9 excluding teachers' salaries. This cost excludes the expenditure on physical facilities, currently being undertaken by parents associations.

Table 5. The cost of educating an individual upto completion (in Ksh)

Sex \ Grade	1	2	3	4	5	6	7
Boys	119.07	126.05	107.28	86.74	67.63	45.59	23.94
Girls	117.84	124.68	106.41	87.83	64.56	43.22	23.52

2.2 MODELS FOR HOUSING STATISTICS

2.2.1 Introduction

Housing statistics can be simply defined as collections of numerical facts or data on the state of housing. The collection and dissemination of housing statistics is the responsibility of governments due to the following reasons:

.

i) The government itself generates a lot of statistics in its normal operations in various sectors including housing. In this way, data collection costs and mechanisms are simplified or reduced.
ii) The data collection is imposed as a legal obligation both for the government as collector and private sources as givers. It is easier for the government agency to get statistical information from private enterprises than for an enterprise to divulge such information to a rival enterprise or an ordinary member of the public.
iii) The collection of data by one central agency would prevent wastage of resources through duplication hence data would be provided en mass and at a cheaper cost to secondary users. In most cases, one would easily get the data required by purchasing a government statistical publication compared to the cost of mobilizing resources to collect that data individually.
iv) It is presumable that only the government agencies can claim the distinction for collection and provision of housing statistics since it is beyond the powers of private investigators to carry out housing investigations on an adequate national scale.

The housing market commands a significant position in today's market economy. The value of the stock of dwellings and real estate is considerable and economic cycles have caused major changes in it over time. The housing market provides a significant number of jobs in the building, agency and maintenance sectors, among others. In addition, housing concerns in practice every individual, either through owner-occupancy or renting. In spite of this, from the international perspective, the compilation of statistics on this topic has been relatively deficient. The common type of information obtained from this sector includes but not limited to the following:

i) Current construction reports
ii) Housing starts and housing completions
iii) New one-family houses sold and for Sale
iv) Price indexes of new one-family houses Sold
v) Housing Units Authorized by Building Permits
vi) Expenditures for Residential Upkeep and Improvements
vii) Value of New Construction put in Place

viii) Current Housing Reports
ix) Sale of existing houses
x) Occupancy rates
xi) Rental and capital values
xii) Demolitions/obsolescence/condemnation
xiii) Destruction, e.g. by natural disasters.

Of notable interest is the price index which varies because of;

i) Housing is not a homogeneous good, but it depends on **where it is located**, **how big it is**, **how its structure**, **quality of the materials**, etc
ii) The characteristics of housing change in time.

This forms the basis of housing statistics and modeling. For instance, researcher might be interested to assess the price per unit square metre of commercial houses in urban areas or rental rates in a given country. This could be extended to assess the price trends over time in order to project the house prices in order to inform the government for developing appropriate policies on housing to improve the country's economy. Depending on the objective of the study and type of data collected, the available statistics to use will include descriptive and inferential analysis to test hypothesis of interest. The latter will require the use of statistical models to support decision making on the overall findings of the study.
The common type models used in housing statistics include the following;

a) Regression models
b) Structural/Markovian models
c) Time series models-ARIMA
d) Dynamic programming models
e) Static models

2.2.2 Discrete and multivariate models for housing statistics

The classes of discrete and multivariate models for housing are as structured above. The following sub sections provides highlights on the data generated from housing will utilize the existing statistical models with required underlying assumptions.

2.2.2.1 Regression Models

Consider housing data on unit price per square foot/metre in urban area. Both the cost or price and unit area continuous variable. In this sense, we can always attempt to check whether there exists a relationship between the cost and size/unit area. The use of

scatter plot will be the first check to whether such relationship exists as illustrated below.

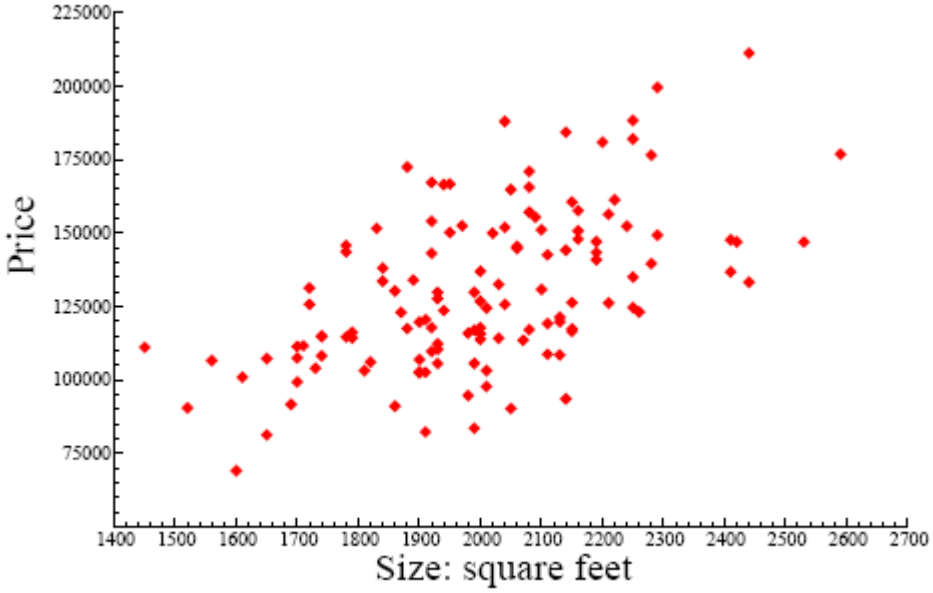


Figure 2.1 Relationship between size in square feet and price.

Now,

We can define a random variable Y_i to be the price of the i-th house!

It is a random variable because before the house is sold we do not know how much it will sell for.

Then, our data $(y_1, y_2, \ldots, y_n)$ is the outcome of a sequence of random variables $(Y_1, Y_2, \ldots, Y_n)$.

We can model this data with an *i.i.d.* Normal model.

$$p(y_1, y_2, \ldots, y_n) = p(y_1) * p(y_2) * \cdots * p(y_n)$$

In other words, each p.d.f. $p(y_i)$ is a normal density and each observation is a normal random variable $Y_i \sim N(\mu, \sigma^2)$.

Suppose you are going to sell your house in this city.

You are interested in the average price μ of a house.

We can perform

statistical inference for the parameters of interest, for example the mean $\mathrm{E}[Y] = \mu$.

For example, we could use $\overline{x}$ as an estimator of μ.

We can also use $\overline{x}$ to predict what the next house Y_i will sell for.

BUT! We are ignoring information about house sizes!!

We are currently only looking at the marginal distribution $P(Y = y)$.

To incorporate information on house sizes, we need to define a second random variable X_i, which is the size of the i-th house.

In reality, we not only observe outcomes of random variables $(Y_1, Y_2, \ldots, Y_n)$ but we observe pairs of outcomes on a pair of random variables (X_i, Y_i).

We believe that a house's size is clearly related to its price.

Which distribution do you think we are interested in?

1. $P(Y = y|X = x)$
2. $P(Y = y, X = x)$

Since we believe that a house's size is clearly related to its price, we think that $P(Y = y|X = x) \neq P(Y = y)$.

KEY POINT: Regression provides a simple way to model the conditional distribution of Y given $X = x$.

This will allow us to incorporate our information on house sizes. This may help us improve our prediction of house prices.

In this case,

Our data looks like: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

In our housing example, this is a fancy way of saying we observe a sample of n houses and:

$$x_i = \text{square footage of the } i\text{-th house}$$
$$y_i = \text{price of the } i\text{-th house}$$

We imagine that (x_i, y_i) are the outcomes of a pair of random variables (X_i, Y_i).

Regression looks at the *conditional* distribution of Y given X.

Instead of coming up with a story for the joint distribution $p(x, y)$:

What do I think the next (x, y) pair will be?

Regression just talks about the conditional distribution $p(y|x)$:

Given a value for x, what will the next y be?

To model *two* numeric variables, our first option is to come up with a story for the joint distribution $p(x_i, y_i)$.

If we think of our data as *i.i.d.* draws from $p(x_i, y_i)$, this option means that we model the data as

$$p(x_1, y_1, x_2, y_2 \ldots, x_n, y_n) = p(x_1, y_1) * p(x_2, y_2) * \cdots * p(x_n, y_n)$$

This is one way of solving the problem. Another way is regression.

When doing regression, we care about the conditional distribution $P(Y = y|X = x) = p(y|x)$.

We use the following terminology:

- We call Y the dependent variable.
- We call X the independent variable, the explanatory variable, or sometimes just the regressor.

For the housing data, what kind of model should we use?

Consider modeling the house prices y_i as an "approximate" linear function of their size x_i.

$$y_i \quad = \quad \text{linear function of } x_i + \text{"error"}$$

This is the starting point for the linear regression model. We need the "errors" because this linear relationship is not exact. y depends on other things besides x that we don't observe in our sample.

Why are we approaching the problem in this way?

Here are three reasons.
1. Sometimes you know x and just need to predict y as in the housing price problem
2. The conditional distribution is an excellent way to think about the relationship between two variables.
3. Linear relationships are easy to work with and are a good approximation in lots of real world problems.

The *simple linear regression model* is

$$Y_i = \alpha + \beta X_i + \varepsilon_i.$$

$$\varepsilon_i \sim N(0, \sigma^2) \text{ } i.i.d.$$

ε_i is independent of X_i.

- The intercept is α.
- The slope is β.
- We use the normal distribution to describe the "errors".

- The parameters of our model are α, β, and σ.
- The slope β measures the change in y when x increases by 1 unit.
- The intercept α is the value y takes when $x = 0$.
- The linear relationship holds for each pair (X_i, Y_i). Consequently, it is common to drop the subscripts and write $Y = \alpha + \beta X + \varepsilon$ instead of $Y_i = \alpha + \beta X_i + \varepsilon_i$.
- The assumption that X is independent of ε is important. It implies that they are uncorrelated.

- The parameters of our model are α, β, and σ.
- These are parameters just like p from the $i.i.d.$ Bernoulli(p) model or μ from the $i.i.d.$ Normal(μ, σ^2) model.
- Just like p and μ, the "true" parameters are unknown, when using real data.

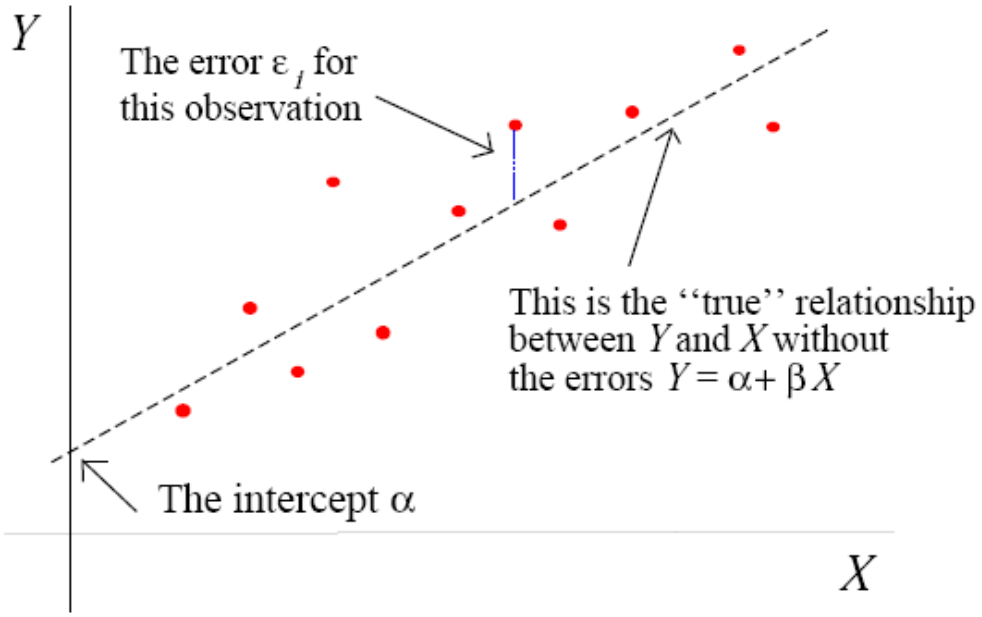
Given a specific value $X = x$, how do we interpret α, β, and σ?

β tells us: if the value we saw for X was one unit bigger, how much would our prediction for Y change?

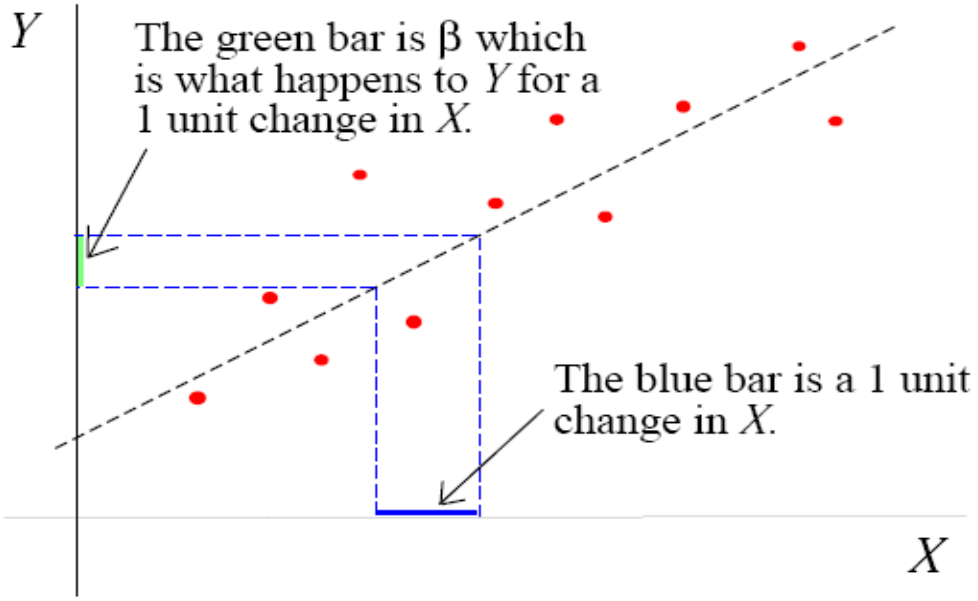
α tells us: what would we predict for Y if $x = 0$?

σ tells us: if $\alpha + \beta x$ is our prediction for Y given x, how big is the error associated with this prediction?

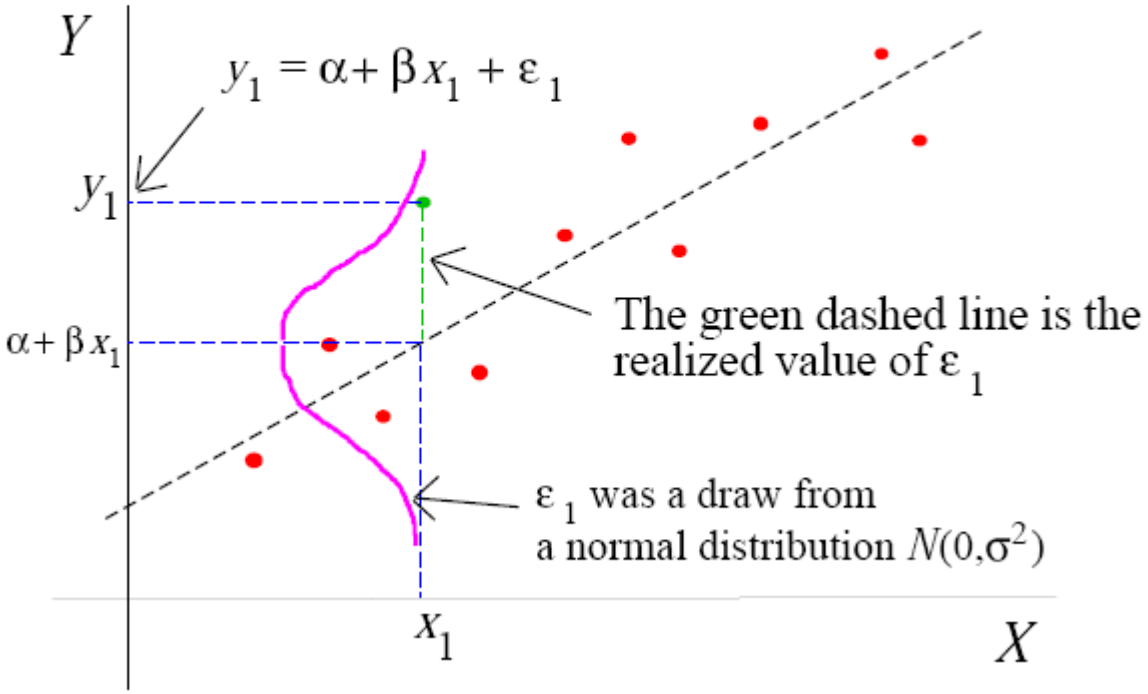
Here is a picture of our model. We are simply drawing a line through the data. α is the intercept.



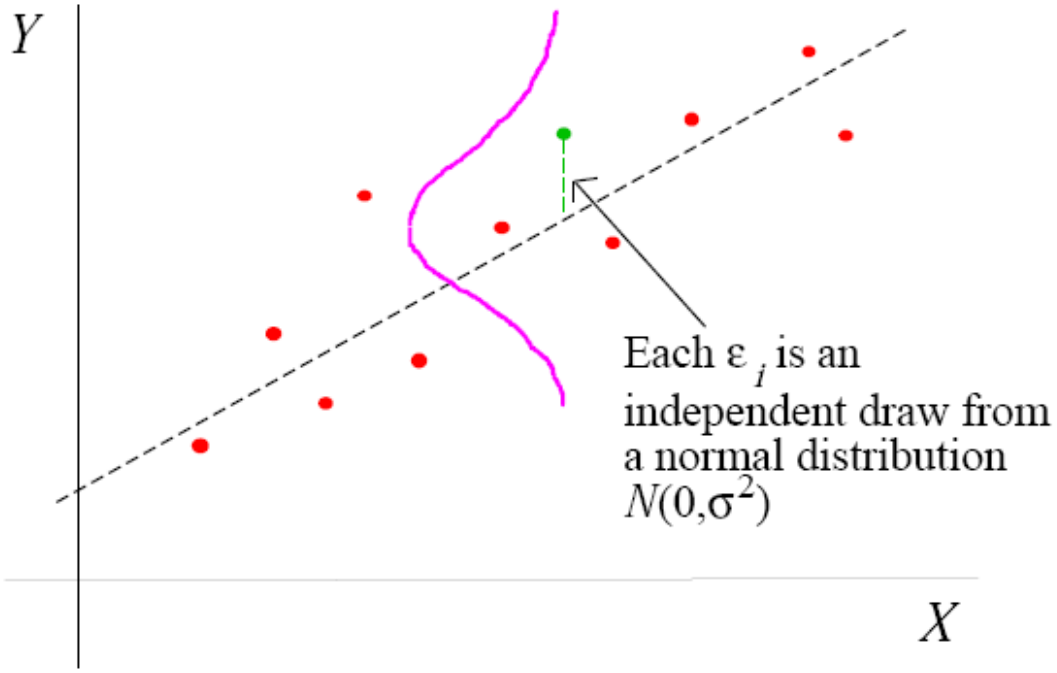
β measures the slope of the line.



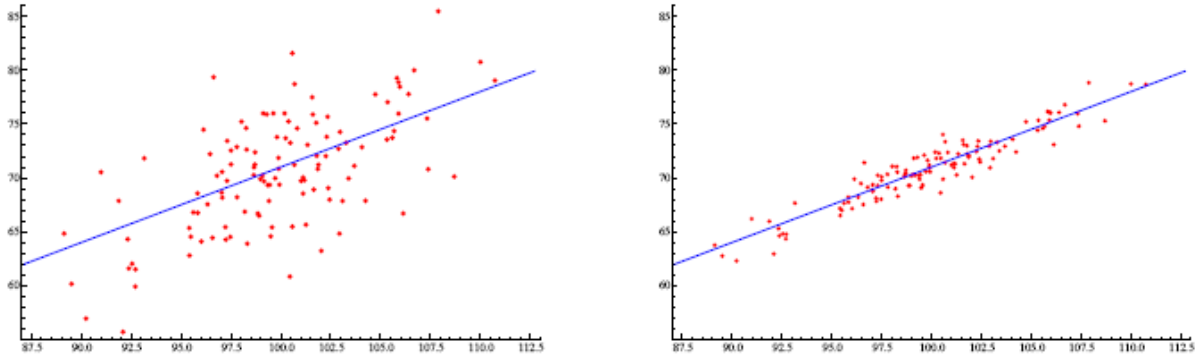
How do we get y_1 given a specific value for $X_1 = x_1$?



Each ε_i is *i.i.d.* $N(0, \sigma^2)$. The variance σ^2 measures the spread of the normal distribution, i.e. the size of our errors.

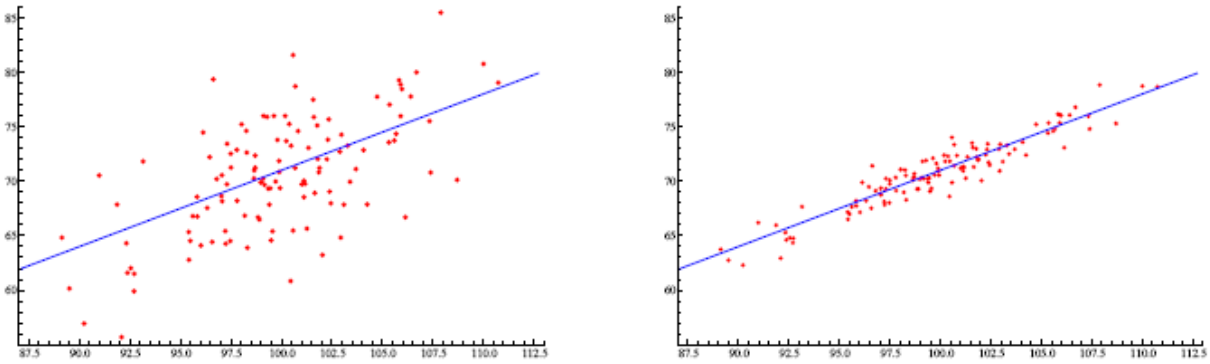


What role does the variance σ^2 play?



The variance of the error term σ^2 describes how big the errors are on average. Notice that when σ^2 is smaller (right) the data are closer to the "true" regression line.

What role does the variance σ^2 play?



The variance will determine how "wide" (or narrow) our predictive intervals are.

Regression as a model of P(Y = yjX = x)

Our model is:

$$Y = \alpha + \beta X + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

where ε is independent of X.

Earlier, we said that regression was a model for the conditional distribution $P(Y = y|X = x)$.

How can we determine the mean and variance of the conditional distribution?

- $\mathrm{E}[Y|X = x]$
- $\mathrm{V}[Y|X = x]$

Since our model is linear

$$Y = \alpha + \beta X + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

we can use our formulas for linear functions!

First, we can compute the conditional mean

$$\begin{aligned} \mathrm{E}[Y|X = x] &= \mathrm{E}[\alpha + \beta X + \varepsilon | X = x] \\ &= \alpha + \beta x + \mathrm{E}[\varepsilon | X = x] \\ &= \alpha + \beta x \end{aligned}$$

Since our model is linear

$$Y = \alpha + \beta X + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

we can use our formulas for linear functions!

And, we can compute the conditional variance

$$\begin{aligned} \mathrm{V}[Y|X = x] &= \mathrm{V}[\alpha + \beta X + \varepsilon | X = x] \\ &= \mathrm{V}[\varepsilon | X = x] \\ &= \sigma^2 \end{aligned}$$

Another way of thinking about our model is:

$$P(Y|X = x) \quad = \quad N(\alpha + \beta x, \sigma^2)$$

In other words,

$$Y|X = x \quad \sim \quad N(\alpha + \beta x, \sigma^2)$$

The conditional distribution of Y is normal with

$$\text{mean: } \mathrm{E}[Y|X = x] = \alpha + \beta x$$

$$\text{variance: } \mathrm{V}[Y|X = x] = \sigma^2$$

Suppose for the moment, we know α, β, and σ.

Given a specific value for $X = x$ and our model,

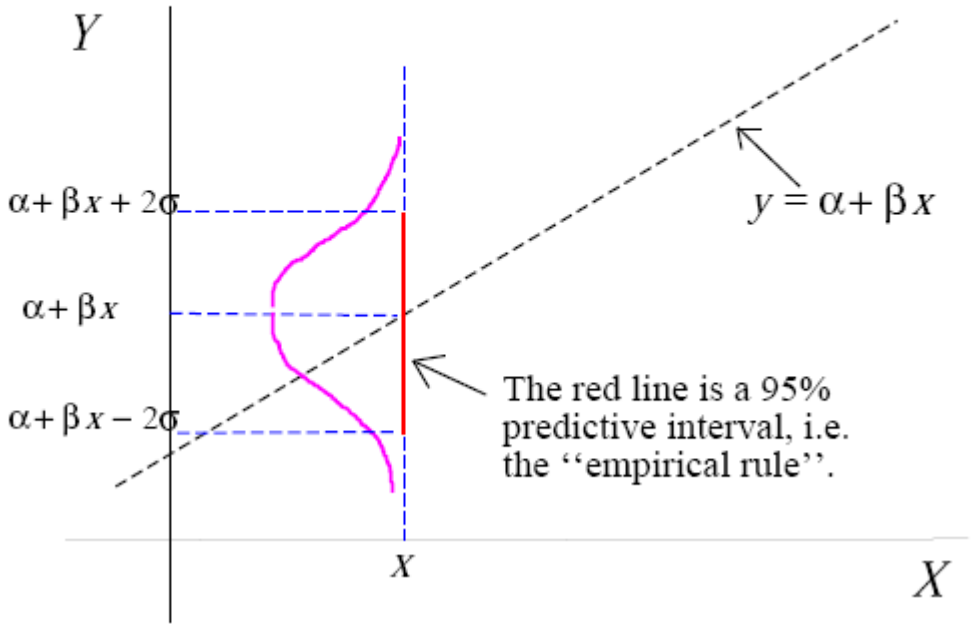
$$Y|X = x \quad \sim \quad N(\alpha + \beta x, \sigma^2)$$

what is our **prediction** of Y?

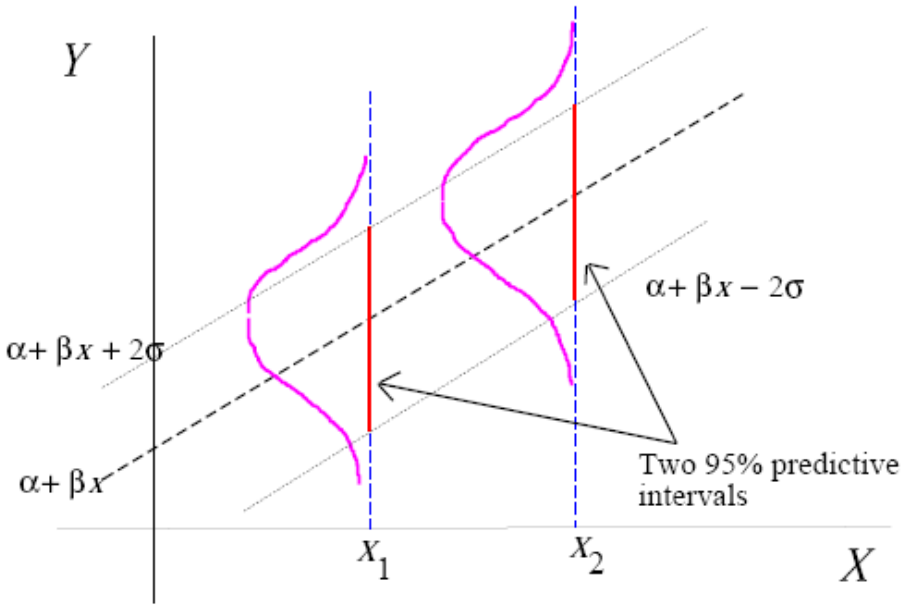
Our prediction is the mean: $\alpha + \beta x$

Since Y has a (conditional) normal distribution, we know that there is a 95% probability that the observed y will be within 2σ.

Given a specific value for $X = x$, we can predict.



Consider two different values x_1 and x_2. Note that since σ^2 is the same for both, the size of the intervals is the same.



Given a specific value for x, our prediction is the conditional mean

$$\alpha + \beta x$$

and with 95% probability the observed value y will lie in the interval

$$(\alpha + \beta x - 2\sigma, \alpha + \beta x + 2\sigma).$$

In practice, we do not know the "true" parameters α, β, and σ.

We have to estimate them from the observed data!

Running the analysis

Following the description of the regression model as outlined in each of the steps above, we can fit the model using any suitable statistical software.

The output will be as follows

The results for a regression of house price on house size.

Results of simple regression for price

Summary measures

Multiple R	0.5530
R-Square	0.3058
StErr of Est	22.4755

ANOVA table

Source	df	SS	MS	F	p-value
Explained	1	28036.3631	28036.3631	55.5011	0.0000
Unexplained	126	63648.8512	505.1496		

Regression coefficients

	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	-10.0911	18.9661	-0.5321	0.5956	-47.6245	27.4422
size	70.2263	9.4265	7.4499	0.0000	51.5716	88.8810

The Fitted regression line

Given our estimates of a and b, these determine a *new* regression line

$$y = a + bx$$

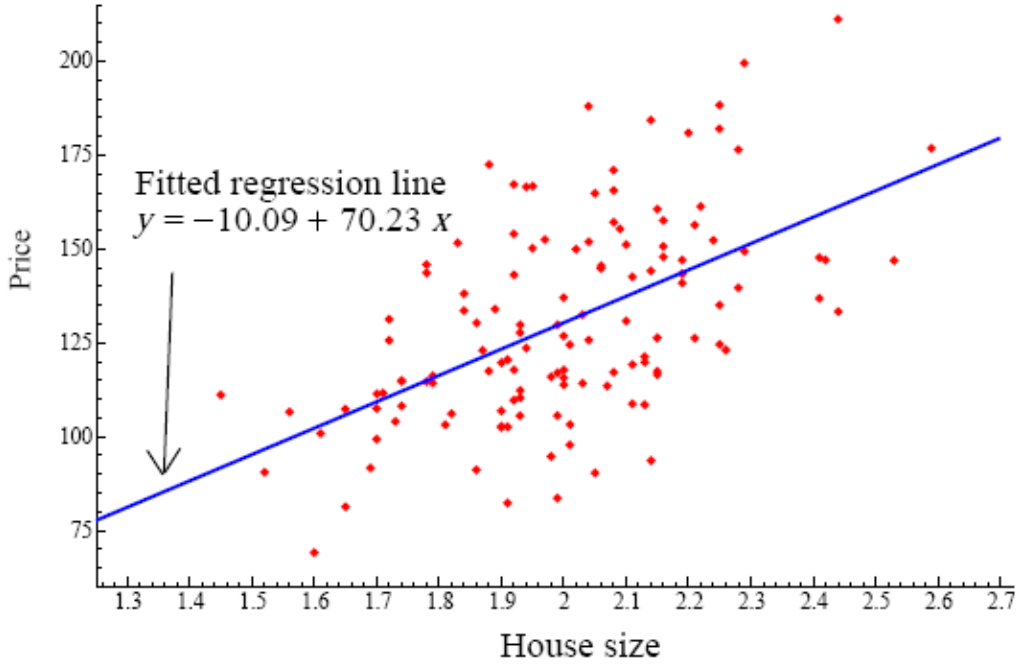
which is called the fitted regression line.

Remember that due to sampling error, our estimate a is not going to be exactly equal to α and our estimate b is not going to be exactly equal to β.

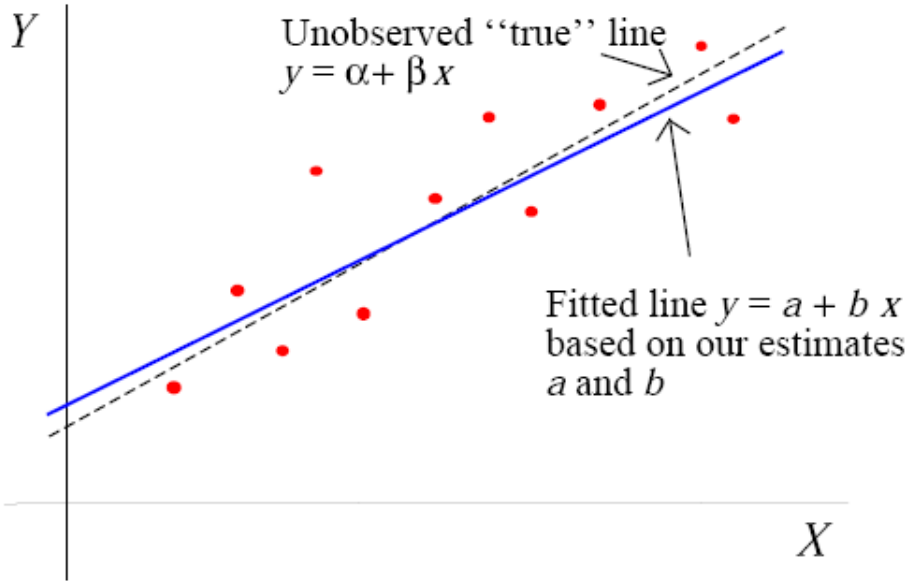
Consequently, the fitted regression line is not going to be exactly equal to the "true" regression line:

$$y = \alpha + \beta x$$

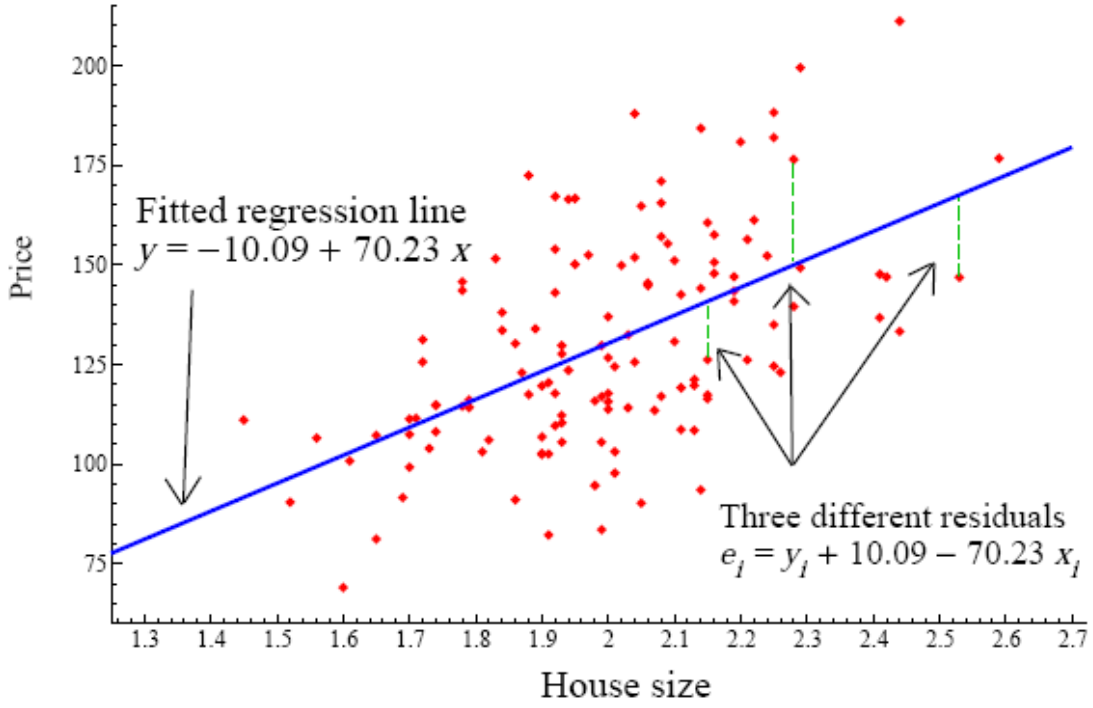
What does the fitted regression line look like? On real data, we can't see the true line.



On simulated data, we can see that the fitted regression line is not the same as the "true" line.

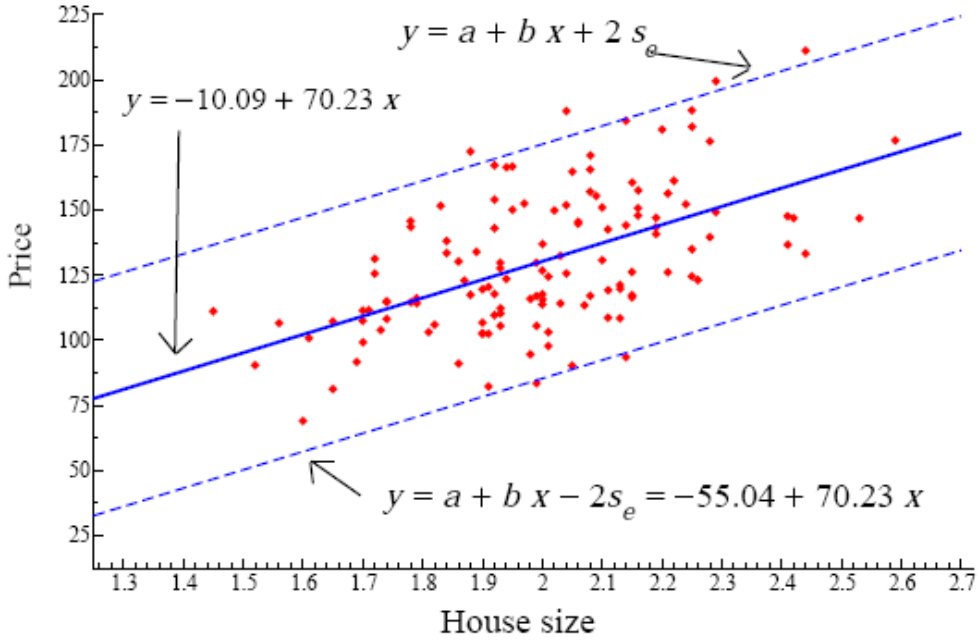


What do the residuals look like?

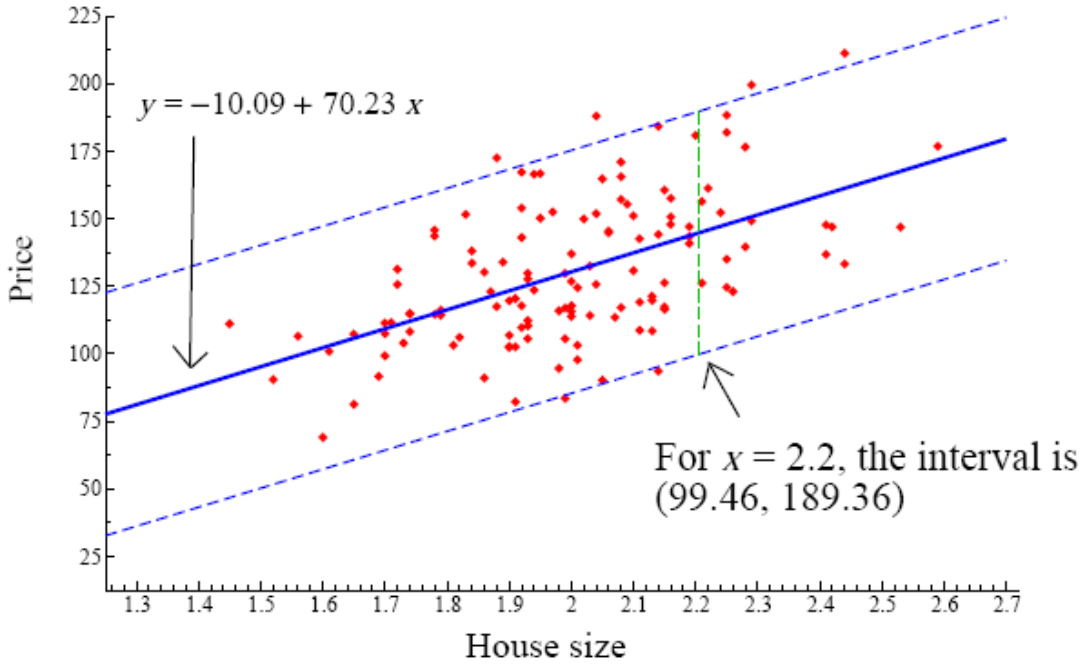


Prediction

Given the estimates of a, b, and s_e, we can get 95% prediction intervals.



Suppose $x = 2.2$. Then, $a + bx = 144.41$ and $2s_e = 44.95$.



Fitted values and residuals

Our model is

$$Y = \alpha + \beta X + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

Conditional on a value x_i, we think of each y_i as a draw from

$$Y_i = \underline{\alpha + \beta x_i} + \underline{\varepsilon_i}$$

the part of y that depends on x

the part of y that has nothing to do with x

We want to ask, "How well does X explain Y"?

We could think about this by "breaking up" Y into two parts:

$$\begin{aligned} \alpha + \beta x_i \quad & \text{(part that's explained by x)} \\ \varepsilon_i \quad & \text{(part that's NOT explained by x)} \end{aligned}$$

But remember, we *don't* know α or β !!

However, we can use our estimates a and b to create estimates of these two parts for each observation in our sample.

So let's suppose we have some data

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

and we've "run a regression".

That is, we've computed the estimates: a, b, and s_e.

For each (x_i, y_i) in the data, we know the following

$$\begin{aligned} \alpha + \beta x_i &\approx a + b x_i \\ \varepsilon_i = y_i - (\alpha + \beta x_i) &\approx y_i - (a + b x_i) = e_i \end{aligned}$$

Define two new variables $\hat{y}_i$ and e_i as follows

$$\begin{aligned} \hat{y}_i &= a + bx_i \\ e_i &= y_i - \hat{y}_i \end{aligned}$$

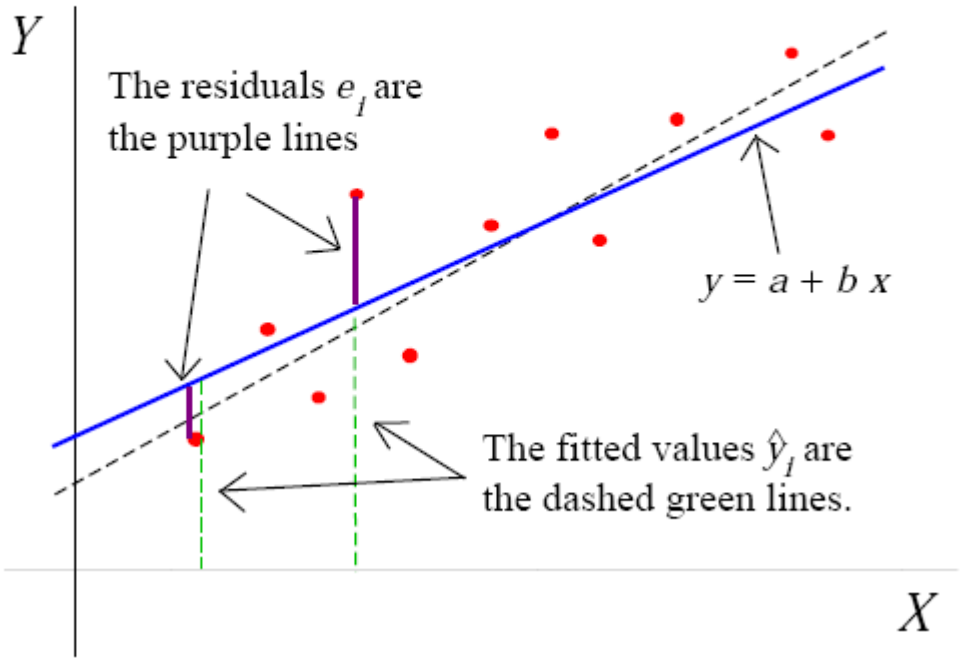
Notice that we have broken up each observation into two parts:

$$y_i = \hat{y}_i + e_i$$

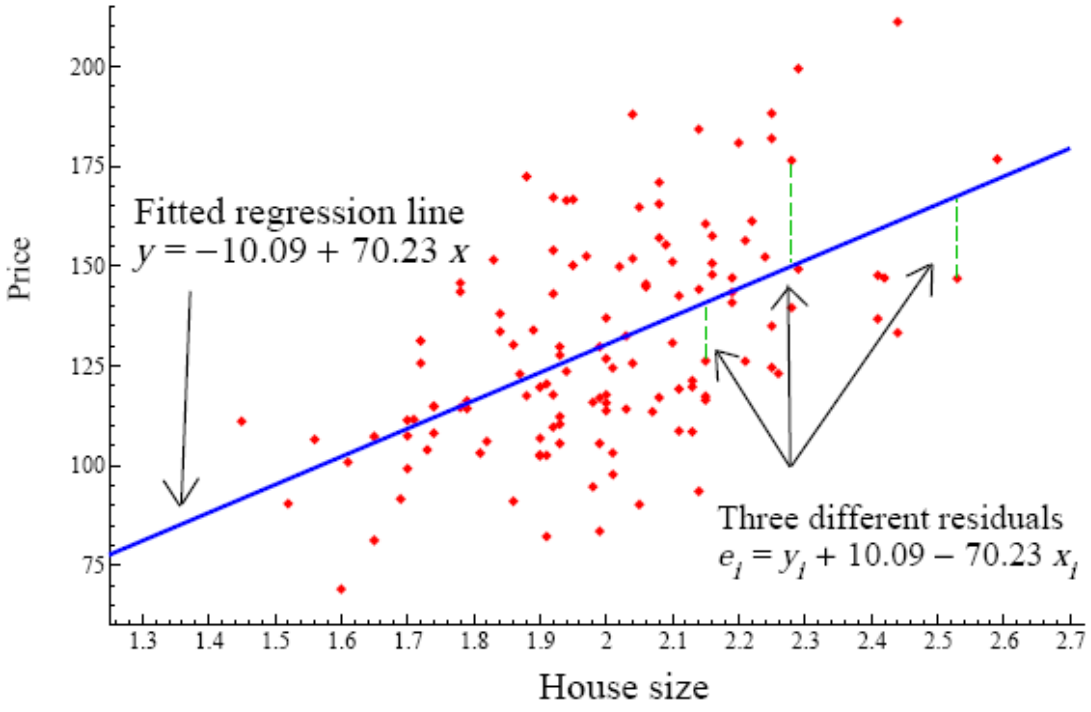
$\hat{y}_i$ is called the *fitted value* for the i-th observation. It is the part of y_i that is "explained" by x_i.

e_i is called the *residual* for the i-th observation. It is the part of y_i that is left unexplained.

What do e_i and $\hat{y}_i$ look like?



Remember the residuals and fitted line for the housing data.



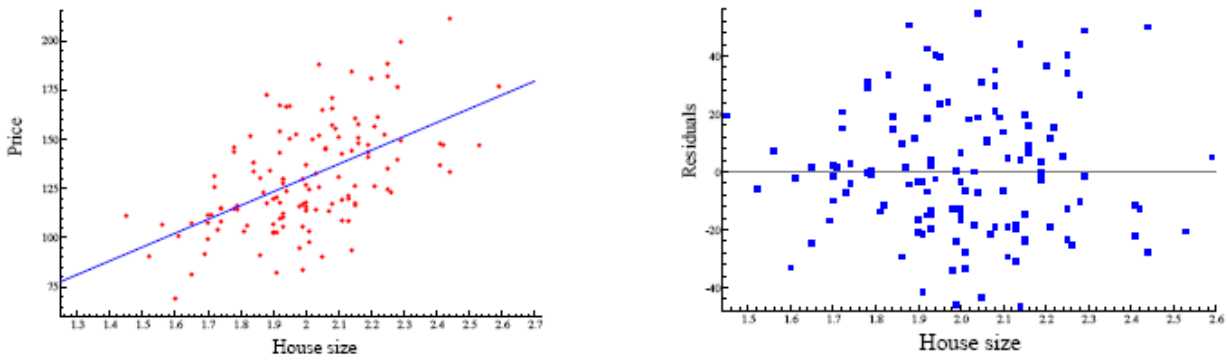
Properties of residuals

Two important properties of the residuals e_i are:

- The sample mean of the residuals equals zero: $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$.
- The sample correlation between the residuals e and the explanatory variable x is zero: $\text{cor}(e, x) = 0$.

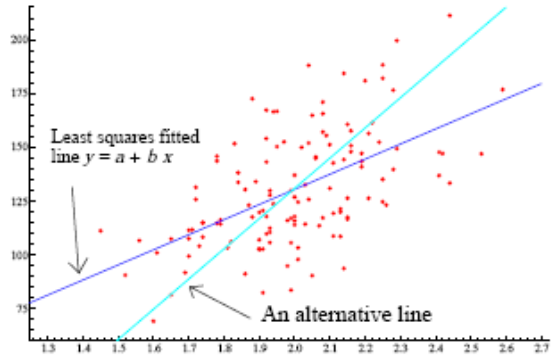
Let's see what this looks like graphically on the housing data.

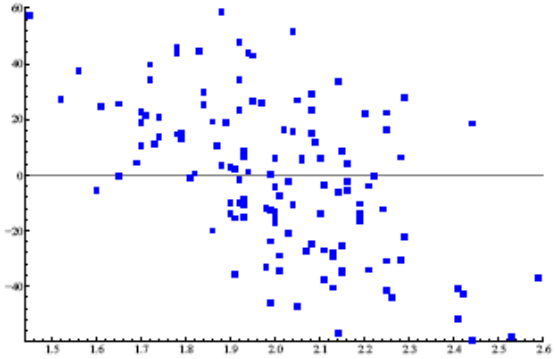
This is the fitted regression line (left) and the residuals (right).



Notice how the residuals demonstrate no obvious pattern and visually look like they have mean zero.

Consider another line that is NOT the least squares line.





Notice how the residuals computed from this alternative line leave a downward right pattern.

We know that $\text{cor}(e, x) = 0$ which means that:

$$\begin{aligned} \text{cor}(e, x) = 0 &\Rightarrow \text{cor}(e, a + bx) = 0 \\ &\Rightarrow \text{cor}(e, \hat{y}) = 0 \end{aligned}$$

In other words, the sample correlation between residuals and fitted values is zero.

Therefore, we now have the three properties:

- $y_i = \hat{y}_i + e_i$
- $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0.$
- $\text{cor}(e, \hat{y}) = 0.$

What does the second property $s_y^2 = s_{\hat{y}}^2 + s_e^2$ mean?

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2 + \frac{1}{n-1}\sum_{i=1}^{n}e_i^2$$

$$\Rightarrow \sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2 + \sum_{i=1}^{n}e_i^2$$

Intuitively, it says that the variance of our dependent variable y can be broken apart into two pieces

- $\sum_{i=1}^{n}(y_i-\overline{y})^2$. This is the total variation in y.
- $\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$. This is the variation in y explained by x.
- $\sum_{i=1}^{n}e_i^2$. This is the unexplained variation in y.

Generalised linear model

The simple regression model can be extended to generalized linear model of the form

$Y=\beta_0+ \beta_1X_1+ \beta_2X_2+.....+\beta_kX_k +e$ .....(1)

depending on the objectives of the study.

For instance in, Y could be the price and X1 =site, X2 = size, etc

2.2.2.3 Time series models-ARIMA

The focus of this model of housing data is assess trend and projections say of housing pricing index. The analysis are based purely on the principles of time series analysis as highlighted below.

A time series is a succession of quantitative observations of a phenomenon in time. Its study is interesting because it allows us to analyze the time evolution of a variable, both in order to construct a descriptive model of the phenomenon's history, and to forecast futures values by means of smoothing methods.

It is important to understand that the assigned order in time to the data is essential in a time series, in this way, each observation should be associate to a certain period. Thus, in fact, a time series is a bi-dimensional frequency distribution (t, yt) where the dependent variable,

yt, is the one we are studying, and t is the independent variable.

But only the variable yt constitutes what is known as univariate model of time series, which itself explain its own past, without the help of an explicative variable, which let us to establish a relation (between) cause-effect as it happens in the regression and correlation. One studies the historical past of de yt (its components) in a descriptive way, assuming that its structure will remain constant, one makes futures predictions. Therefore the sales of a firm in each of the last ten years will be a time series, just like the financial costs, the available incomes of the potential customers, in a descriptive way, assuming that its structure will remain constant, one makes futures predictions. Therefore the sales of a firm in each of the last ten years will be a time series, just like the financial costs, the available incomes of the potential customers,
etc.
Every analysis of time series has to begin with a graphical representation of the series, using the cartesian axes, so that in the abscissas-axis we represent the time and in the other one the observed series y_t, in this way we get a set of points (t, y_t), and by joining them, we see a graphically evolution of its history, from which we can get some conclusions.

Time series components

In the classic study of the time series, it is considered that the studied variable in a certain value and in a certain period is the consequence of the performance of four components or forces, **the trend, the cyclical component, the seasonal component** and **the irregular component**. Now, we are going to define them.

- **Trend (T):** It is the series component that reflects its evolution in the long term. This long term will be according to the series nature, as much numbers of period we have, the analysis will be better. This component, in the set of the whole series, can have a stationary behavior or a constant one (it would be represented by a parallel line to the abscissas-axis), a linear behavior, an exponential behavior, or others.

- **Cyclical component (C):** It is the series component that gathers *the periodic oscillations* of amplitude higher than one year. These oscillations are not regular and they appear in most of economics phenomenons when depression or prosperity stages occur of an alternative form. In a economic series, there are usually more than one cycle of this type, and as consequence, this will be the most difficult component to determine. Since it is natural, the greater the period of a cycle that affects our variable is, the greater the number of observations to consider is.

- **Seasonal component (E):** It is the series component that gathers *the oscillations* in periods of time equal or less than one year. Its name comes from the climatology seasons. If one has the year as period of repetition one can observe the fluctuations of the magnitude through months, quarters, or periods of four months, etc. Just like if one takes the month as a period of repetition, the fluctuations of the magnitude could be observed trough days, weeks, etc. The origin of the seasonal variations can be due to physical-natural factors, like climatologic seasons, or cultural and tradition factors, like Christmas celebrations, holidays, commercial schedules, etc. The climate affects the sale series of certain products, for example, ice creams and soft drinks will fundamentally be sold in summer and warm clothes in winter.

- **Irregular component (A):** It is the series component that gathers the erratic fluctuations due to unforeseeable phenomenon, they affect sporadic and nonpermanent the studied variable (like an extraordinary order to a company, a strike, a catastrophe, etc). They are also known as *residual* or *erratic variable.*

Hypothesis to test

Now it is normal to make a basic question: How do the fourth questions giving as a result of the different values of the series act? In the classic study of time series one deals with two hypothesis of work:

1. Each observation of a certain time series is the result of the **sum** of the four components:

$$y_t = T_t + C_t + E_t + A_t.$$

   This expression is known as additive hypothesis.

2. Each observation of a certain time series is the result of the **multiplication** of the four components:

$$y_t = T_t * C_t * E_t * A_t.$$

   The following variant of the model is used when assuming that the irregular component is independent from the others and it does not follow any regularity periodic, like the others do. This independence implies that the component A appears as an additive element:

$$y_t = T_t * C_t * E_t + A_t.$$

How to decide which hypothesis to follow?
To answer this question, it will be necessary to carry out a previous analysis of the series. An analytical form to determine which model fits best is by means of **the method of the seasonal quotients and differences**. Let us see what seasonal quotients and differences are:

- The *seasonal difference* is calculated as the difference between a certain season of a year and the same season from the year before. It is named $d_{t,i}$.

$$d_{t,i} = y_{t,i} - y_{t-1,i}.$$

- The *seasonal quotient* is calculated as the quotient of a certain season of a year between the same season from the year before. It is named $c_{t,i}$.

$$c_{t,i} = y_{t,i}/y_{t-1,i},$$

  where $y_{t,i}$= is the value of the series in the year t, in the season i.

Once we have gone over the concepts of difference and seasonal quotient we are going to explain the steps that we should follow in order to apply this method.

1. We calculate all seasonal quotients and differences. Obviously, in these calculations we lose the observations corresponding to the first year of the data.

2. We calculate the coefficient of variation (CV) of the seasonal quotients and differences, they are given by the following expressions:

$$CV(d) = \left| \frac{\text{Standar deviation}(d)}{Average(d)} \right|,$$

$$CV(c) = \left| \frac{\text{Standar deviation}(c)}{Average(c)} \right|.$$

3. We apply the following decision rule:
   If $CV(c) > CV(d)$ we choose the additive model.
   If $CV(c) \leq CV(d)$ we choose the multiplicative model.

The series of the seasonal differences is equivalent to the yearly increases one. However, the series of the seasonal quotients has a stronger relation with the global increase series. Therefore what we are implicitly saying is that, if the yearly increase for each season has minor variability that in global terms, this leads to a multiplicative association between trend and stationarity. Otherwise it would be better to choose the additive hypothesis.

Example. Housing data taken in 4 quarters a year

The data below show prices of houses per square meter taken from 1987 to 2003. Compute the seasonal difference and seasonal quotients. Comment on your results. What can of kind time series model will use to fit the data?

Year	Quarter	Price
1987	1	289.89
1987	2	308.64
1987	3	324.99
1987	4	345.55
1988	1	369.13
1988	2	389.79
1988	3	404.39
1988	4	423.12
1989	1	456.58
1989	2	480.17
1989	3	502.72
1989	4	516.43
1990	1	550.4
1990	2	559.73
1990	3	570.77
1990	4	580.6
1991	1	613.42
1991	2	637.9
1991	3	652.8
1991	4	681.23
1992	1	650.49
1992	2	635.7
1992	3	633.78
1992	4	630.72
1993	1	625.44
1993	2	634.83
1993	3	639.69
1993	4	640.61
1994	1	634.72
1994	2	636.8
1994	3	644.36
1994	4	642.63
1995	1	652.92
1995	2	661.06
1995	3	665.46
1995	4	667.47
1996	1	669.98

1996	2	674.79
1996	3	675.18
1996	4	676.45
1997	1	677.74
1997	2	683.06
1997	3	686.64
1997	4	691.78
1998	1	694.34
1998	2	709.66
1998	3	723.95
1998	4	738.58
1999	1	755.21
1999	2	780.25
1999	3	803.89
1999	4	829.81
2000	1	857.25
2000	2	891.76
2000	3	926.36
2000	4	953.42
2001	1	994.5
2001	2	1030.77
2001	3	1065.78
2001	4	1096.57
2002	1	1148.23
2002	2	1193.66
2002	3	1254.09
2002	4	1287.73
2003	1	1349.11
2003	2	1402.57
2003	3	1450.6

Solutions

Year	Price square meter	Seasonal difference	Seasonal quotient
1987 1T	289,89		
2T	308,64		
3T	324,99		
4T	345,55		
1988 1T	369,13	79,24	1,27
2T	389,79	81,15	1,26
3T	404,39	79,40	1,24
4T	423,12	77,57	1,22
1989 1T	456,58	87,45	1,24
2T	480,17	90,38	1,23
3T	502,72	98,33	1,24
4T	516,43	93,31	1,22
1990 1T	550,40	93,82	1,21
2T	559,73	79,56	1,17
3T	570,77	68,05	1,14
4T	580,60	64,17	1,12
1991 1T	613,42	63,02	1,11
2T	637,90	78,17	1,14
3T	652,80	82,03	1,14
4T	681,23	100,63	1,17
1992 1T	650,49	37,07	1,06
2T	635,70	-2,20	1,00
3T	633,78	-19,02	0,97
4T	630,72	-50,51	0,93
1993 1T	625,44	-25,05	0,96
2T	634,83	-0,87	1,00
3T	639,69	5,91	1,01
4T	640,61	9,89	1,02
1994 1T	634,72	9,28	1,01
2T	636,80	1,97	1,00
3T	644,36	4,67	1,01
4T	642,63	2,02	1,00
1995 1T	652,92	18,20	1,03
2T	661,06	24,26	1,04
3T	665,46	21,10	1,03
4T	667,47	24,84	1,04

Year	Price square meter	Seasonal difference	Seasonal quotient
1996 1T	669,98	17,06	1,03
2T	674,79	13,73	1,02
3T	675,18	9,72	1,01
4T	676,45	8,98	1,01
1997 1T	677,74	7,76	1,01
2T	683,06	8,27	1,01
3T	686,64	11,46	1,02
4T	691,78	15,33	1,02
1998 1T	694,34	16,60	1,02
2T	709,66	26,60	1,04
3T	723,95	37,31	1,05
4T	738,58	46,80	1,07
1999 1T	755,21	60,87	1,09
2T	780,25	70,59	1,10
3T	803,89	79,94	1,11
4T	829,81	91,23	1,12
2000 1T	857,25	102,04	1,14
2T	891,76	111,51	1,14
3T	926,36	122,47	1,15
4T	953,42	123,61	1,15
2001 1T	994,50	137,25	1,16
2T	1.030,77	139,01	1,16
3T	1.065,78	139,42	1,15
4T	1.096,57	143,15	1,15
2002 1T	1.148,23	153,73	1,15
2T	1.193,66	162,89	1,16
3T	1.254,09	188,31	1,18
4T	1.287,73	191,16	1,17
2003 1T	1.349,11	200,88	1,17
2T	1.402,57	208,91	1,18
3T	1.450,60	196,51	1,16
	variance	3.865,50	0,01
	stand.deviation	62.17	0,09
	average	128,36	1,10
	stand.deviation/average	0.4843	0,0788

As we can see, the quotient coefficient of variation (coef.variation=|variance/average|) is less than the difference coefficient of variance, therefore it is a model of time series under the **multiplicative hypothesis**.

To calculate these seasonal differences and quotients, we make the difference and the quotient respectively between a certain observation in a certain season and the one of the year before in the same season, that is, we will not be able to calculate these coefficients for the year 1987 (because we do not have the data from 1986). And in the others cases the calculations are the following,

Year	Seasonal difference	Seasonal quotient
1988-1T	369,13-289,89=79,24	369,13/289,89=1,27
1988-2T	389,79-308,79=81,15	389,79/308,79=1,26
1988-3T	404,39-324,99=79,40	404,39/324,99=1,24
1988-4T	423,12-345,55=77,57	423,12/345,55=1,22

We see now the calculation of each component of the series. We have to take into account that there are several methods, but we use classical and simple methods.

Calculation of the trend

1. We calculate the series without stationarity:

   - We calculate a centered moving average, with a yearly period, named MM_t. A moving average is an average of data for a certain number of time periods. It "moves" because for each calculation, we use the latest x number of time periods data. Its calculation is as follows:
     - We have the observed time series y_t.
     - We calculate the average for each y_t with a fixed number of previous and later observations (in our case we take four observations, because of the fact we have quarterly data). If the numbers of used observations is odd, the average $\overline{y}_t$ (it is centered) fits in the period t. If this number is even the y_t does not fit in the period t and it has to be calculated a new average $\overline{\overline{y}}_t$ using the previous and later observations $\overline{y}_t$, in order to obtain a centered moving average.
       As we already said, we take 4 observations, i.e., and the calculation is the following;

$$\overline{y}_{2.5} = \frac{y_1 + y_2 + y_3 + y_4}{4} = \frac{289.89 + 308.64 + 324.99 + 345.55}{4} = 317.27,$$

$$\overline{y}_{3.5} = \frac{y_2 + y_3 + y_4 + y_5}{4} = \frac{308.64 + 324.99 + 345.55 + 369.13}{4} = 337.08,$$
$$\overline{y}_{4.5} = \frac{y_3 + y_4 + y_5 + y_6}{4} = \frac{324.99 + 345.55 + 369.13 + 389.79}{4} = 357.37,$$
$$\overline{y}_{5.5} = \frac{y_4 + y_5 + y_6 + y_7}{4} = \frac{345.55 + 369.13 + 389.79 + 404.39}{4} = 377.22.$$

These averages have fictitious periods, they are $t' = 2.5; 3.5; 4.5; 5.5$, and they are not periods from this series. That means that 317.27 has a fictitious period $t' = 2.5$, and so does 337.08 with $t' = 3.5$, etc. In order to get this **centered** averages with the real periods, we calculate once more the averages of $\overline{y}_{t'}$ with two observations, we obtain $\overline{\overline{y}}_t$ centered in the real periods of the series :

$$\overline{\overline{y}}_3 = \frac{\overline{y}_{2.5} + \overline{y}_{3.5}}{2} = \frac{317.27 + 337.08}{2} = 327.17,$$
$$\overline{\overline{y}}_4 = \frac{\overline{y}_{3.5} + \overline{y}_{4.5}}{2} = \frac{337.08 + 357.37}{2} = 347.22,$$
$$\overline{\overline{y}}_5 = \frac{\overline{y}_{2.5} + \overline{y}_{3.5}}{2} = \frac{357.37 + 377.22}{2} = 367.29.$$

In this way we calculate the Moving (MMA) of the original series. We should take into account that with this method we will lose the two first and the two last observations, (see it third column of 2).

- Since this series is yearly given, we will lose most of the seasonal and irregular component, since adding the observations of a year, the positive and negative values of this last component will be balanced, then we can interpret MM_t as the **multiplication** of the trend and the cyclical component: $T_t * C_t$. Therefore, the quotient between the original series and the moving average, i.e., y_t/MM_t (see it on the fourth column of 2), it is a percentages around 1, with information about the seasonal and irregular component.
- In order to delete the irregular component, we calculate the average of each season of y_t/MM_t (in our case of each quarter, they are 4, M_1, M_2, M_3, M_4). These averages represent the importance of each season. Let us see the calculations;

$$M_1 = \frac{1.005 + 1.0071 + 1.0177 + \cdots + 1.0009}{16} = 0.99955,$$
$$M_2 = \frac{1.074 + 1.006 + 1.0043 + \cdots + 0.9972}{15} = 1.00080,$$

$$M_3 = \frac{0.9933 + 0.9923 + 1.004 + \cdots + 1.0065}{16} = 1.00030,$$
$$M_4 = \frac{0.9952 + 0.9845 + 0.9886 + \cdots + 0.9927}{16} = 0.9974.$$

- We obtain the indexes of the seasonal variation: we calculate the yearly average MA of the seasonal averages, i.e.

$$MA = \frac{M_1 + M_2 + M_3 + M_4}{4} = \frac{0.99955 + 1.00080 + 1.00030 + 0.9974}{4} = 0.99951.$$

Now we are ready to get these indexes in percentages:

$$I_1 = \frac{M_1}{MA} * 100 = \frac{0.99955}{0.99951} \cdot 100 = 100.004,$$
$$I_2 = \frac{M_2}{MA} * 100 = \frac{1.0008}{0.99951} \cdot 100 = 100.128,$$
$$I_3 = \frac{M_3}{MA} * 100 = \frac{1.00030}{0.99951} \cdot 100 = 100.079,$$
$$I_4 = \frac{M_4}{MA} * 100 = \frac{0.9974}{0.99951} \cdot 100 = 99.789.$$

If we obtain an index in one of the season of 80%, that means that in this season the seasonal component is 20% smaller than the average. In our case, these seasonal indexes are almost insignificant, that is the average price of square meter has a similar behavior in each season, that is, the price increases always in time without depending on the season.

- Finally, once we have these indexes, we can take the seasonal component out, dividing each element of the original series by its corresponding season index, expressed in percentages around one (see it on the fifth column of 2).

2. We estimate a linear regression model of this new series, y_{td} and so we will get the **trend**.

We will deal with the **least square method**, which is a statistical approach to estimate an expected value or function with the highest probability from the observations with random errors. The highest probability is replaced by minimizing the sum of the squares of residuals.

Residual is defined as the difference between the observation and an estimated value of a function: $y_t = a + bt$. We form the system of normal equations:

$$\sum_{t=1}^{n} y_t = na + b\sum_{t=1}^{n} t, \tag{1}$$

$$\sum_{t=1}^{n} ty_t = a\sum_{t=1}^{n} t + b\sum_{t=1}^{n} t^2, \tag{2}$$

where n is the total number of observations, that fit in with the numbers of periods of time (we will take in this case the series given year by year).
The system of normal equations (2) is simplified by carrying out the following change of variable $t^{'} = t - O_t$ if we have an odd number of periods (like in our case, due to we have 17 years, so $O_t = 1995$), where O_t is located in the middle of the series. With this change of variable we get $\sum_{t^{'}=1}^{n} t^{'} = 0$. Therefore, the system (2) can be expressed in a simpler way:

$$\sum_{t=1}^{n} y_t = na,$$

$$\sum_{t=1}^{n} y_t t^{'} = b\sum_{t=1}^{n} t^2.$$

When working out the parameter of the straight line, in this case the variables of the system, we obtain:

$$a = \frac{\sum_{t=1}^{n} y_t}{n}, \tag{3}$$

$$b = \frac{\sum_{t=1}^{n} y_t t^{'}}{\sum_{t^{'}=1}^{n} t^{'2}}, \tag{4}$$

so the estimate straight line is:

$$y_t = a + bt^{'},$$

and undoing the change of variable we get the regression straight line:

$$y_t = a + b(t - O_t).$$

When the observations are given in periods of time smaller than a year (as it is our case, quarterly periods) before making the estimation we calculate the yearly averages to eliminate the seasonal component that can distort the result.

Once we have smoothed the regression straight line to the scatter plot, it is important to fix a measure of the goodness of fit of the relationship between the dependent and independent variable. That allows us to decide if the estimation is good or not. This measure in regression analysis is called **coefficient of determination**, defined as

$$R^2 = \frac{(S_{t'y_t})^2}{S^2_{t'} S^2_{y_t}},$$

where $S^2_{t'y_t}$ is the covariance between the variable t' and y_t, i.e.,

$$S_{t'y_t} = \frac{1}{N}\sum_{i=1}^{N} t'_i y_{ti} - \overline{t'}\,\overline{y_t}.$$

Both $S^2_{t'}$ and $S^2_{y_t}$ represent the marginal variances of t' and y_t respectively. They are calculating according to the following formula;

$$S^2_{y_t} = \frac{1}{N}\sum_{i=1}^{N} y^2_{ti} - (\overline{y_t})^2,$$

$$S^2_{t'} = \frac{1}{N}\sum_{i=1}^{N} (t'_i)^2 - (\overline{t'})^2.$$

This value is within $-1 \leq S^2_{y_t} \leq 1$, and the reliability of this model is good if this coefficient approaches to 1, in case of positive correlation, and to -1 in case of a negative correlation.

Carrying out all these steps we obtain the following results;

$$a = 731.27 \qquad b = 50.10.$$

The regression straight line of the series without stationarity is the following:

$$y_t = 731.27 + 50.10(t - 1995) = 50.10t - 99222.45,$$

with a coefficient of determination of

$$R^2 = \frac{1445887.84}{24 \cdot 72629.06} = 0.83.$$

(A detailed calculation of the regression straight line by means of the least squares method is shown in the section 4.1.)

Let us see in table 2 the values of the trend;

Year	Average price	MMA	y_t/MMA	Series without stationarity	Trend
1987 1T	289,89			289,880	325,96
2T	308,64			308,244	338,485
3T	324,99	327,17	0,99333	324,735	351,01
4T	345,55	347,22	0,99519	346,279	363,535
1988 1T	369,13	367,29	1,00501	369,117	376,06
2T	389,79	386,91	1,00744	389,290	388,585
3T	404,39	407,54	0,99227	404,073	401,11
4T	423,12	429,77	0,98453	424,013	413,635
1989 1T	456,58	453,36	1,00711	456,564	426,16
2T	480,17	477,31	1,00599	479,554	438,685
3T	502,72	500,70	1,00403	502,326	451,21
4T	516,43	522,38	0,98862	517,520	463,735
1990 1T	550,40	540,83	1,01770	550,380	476,26
2T	559,73	557,35	1,00426	559,012	488,785
3T	570,77	573,25	0,99567	570,322	501,31
4T	580,60	590,90	0,98257	581,825	513,835
1991 1T	613,42	610,93	1,00408	613,398	526,36
2T	637,90	633,76	1,00653	637,082	538,885
3T	652,80	650,97	1,00281	652,288	551,41
4T	681,23	655,33	1,03952	682,668	563,935
1992 1T	650,49	652,68	0,99665	650,467	576,46
2T	635,70	643,99	0,98713	634,885	588,985
3T	633,78	634,54	0,99880	633,283	601,51
4T	630,72	631,30	0,99908	632,051	614,035
1993 1T	625,44	631,93	0,98973	625,417	626,56
2T	634,83	633,91	1,00146	634,016	639,085
3T	639,69	636,30	1,00532	639,188	651,61
4T	640,61	637,71	1,00455	641,962	664,135
1994 1T	634,72	638,54	0,99402	634,697	676,66
2T	636,80	639,38	0,99597	635,983	689,185
3T	644,36	641,90	1,00383	643,854	701,71
4T	642,63	647,21	0,99292	643,986	714,235
1995 1T	652,92	652,88	1,00006	652,896	726,76
2T	661,06	658,62	1,00370	660,212	739,285
3T	665,46	663,86	1,00241	664,938	751,81
4T	667,47	667,71	0,99964	668,878	764,335

Calculation of cyclical component

- We calculate the quotient between the series without stationarity, y_{td}, and the trend, T_t, in this way we obtain the multiplication of the cyclical and irregular component, $C_t * A_t$. Since both series y_{td} and T_t are measured in the same unit, the multiplication $W_t = C_t * A_t$ is a percentages around one.

- Finally, to separate the cyclical component, we calculate the Moving Average series of W_t giving as result a estimation of the cyclical component.

The table 3 shows us the results of the calculation of the cyclical component in percentages, these calculations are the same as the ones we did in order to obtain the Moving Average of the original series.

Year	series without stationarity/ trend=Ct*At	Moving Average of Ct*At	Second Moving Average of Ct*At*100 = cyclical component, Ct (%)
1987 1T	0,889368928		
2T	0,910638004	0,919410849	
3T	0,925124174	0,942469	93,09399243
4T	0,952512289	0,965257551	95,38632755
1988 1T	0,981601533	0,985817515	97,55375329
2T	1,001792208	1,003956158	99,48868364
3T	1,007364029	1,026409219	101,5182689
4T	1,025066862	1,049245815	103,7827517
1989 1T	1,071413778	1,075720041	106,2482928
2T	1,093138591	1,098442605	108,7081323
3T	1,113260931	1,119515645	110,8979125
4T	1,11595712	1,132143872	112,5829758
1990 1T	1,155705938	1,138238271	113,5191071
2T	1,143651498	1,142322531	114,0280401
3T	1,137638527	1,144754813	114,3538672
4T	1,132294162	1,154390945	114,9572879
1991 1T	1,165435064	1,165711078	116,0051011
2T	1,182196027	1,185266608	117,5488843
3T	1,182919058	1,176021741	118,0644174
4T	1,210516283	1,149949281	116,2985511

Year	Average price	MMA	y_t/MMA	Series without stationarity	Trend
1996 1T	669,98	670,64	0,99902	669,956	776,86
2T	674,79	672,98	1,00269	673,924	789,385
3T	675,18	675,07	1,00016	674,650	801,91
4T	676,45	677,07	0,99908	677,877	814,435
1997 1T	677,74	679,54	0,99735	677,716	826,96
2T	683,06	682,89	1,00025	682,184	839,485
3T	686,64	686,88	0,99965	686,101	852,01
4T	691,78	692,28	0,99928	693,240	864,535
1998 1T	694,34	700,27	0,99153	694,315	877,06
2T	709,66	710,78	0,99842	708,750	889,585
3T	723,95	724,24	0,99960	723,382	902,11
4T	738,58	740,67	0,99717	740,139	914,635
1999 1T	755,21	759,49	0,99436	755,183	927,16
2T	780,25	780,89	0,99919	779,249	939,685
3T	803,89	805,05	0,99857	803,259	952,21
4T	829,81	831,74	0,99768	831,561	964,735
2000 1T	857,25	860,99	0,99566	857,219	977,26
2T	891,76	891,75	1,00002	890,616	989,785
3T	926,36	924,35	1,00217	925,633	1002,31
4T	953,42	958,89	0,99430	955,432	1014,835
2001 1T	994,50	993,69	1,00082	994,464	1027,36
2T	1.030,77	1.029,01	1,00171	1029,448	1039,885
3T	1.065,78	1.066,12	0,99968	1064,944	1052,41
4T	1.096,57	1.105,70	0,99174	1098,884	1064,935
2002 1T	1.148,23	1.149,60	0,99881	1148,189	1077,46
2T	1.193,66	1.197,03	0,99718	1192,129	1089,985
3T	1.254,09	1.246,04	1,00646	1253,106	1102,51
4T	1.287,73	1.297,26	0,99265	1290,447	1115,035
2003 1T	1.349,11	1.347,94	1,00087	1349,061	1127,56
2T	1.402,57			1400,771	1140,085
3T	1.450,60			1449,462	1152,61

Year	series without stationarity/ trend=Ct*At	Moving Average of Ct*At	Second Moving Average of Ct*At*100 = cyclical component, Ct (%)
2000 1T	0,877223783	0,89060753	88,06169706
2T	0,899787817	0,910483906	90,05457179
3T	0,923479484	0,933188996	92,1836451
4T	0,941444539	0,955727368	94,44581819
2001 1T	0,968044144	0,977829308	96,67783377
2T	0,989941304	1,000432224	98,91307656
3T	1,011887244	1,024849742	101,2640983
4T	1,031856203	1,050786203	103,7817973
2002 1T	1,065714219	1,081956565	106,6371384
2T	1,093687146	1,113315036	109,76358
3T	1,136568692	1,146017008	112,9666022
4T	1,157290088		
2003 1T	1,196522108		

Calculation of irregular component

We get an estimation of the irregular component as a quotient between W_t and the estimation of the cyclical component, i.e, $A_t = W_t/C_t$. We can see the results in table 4. Let us see the calculations for the four first iterations;

Year	Irregular component
1987-3T	$0.9251/0.9309 \cdot 100$=99.3752
1987-4T	$0.95251/0.9533 \cdot 100$=99.8583
1988-1T	$0.9815/0.9755 \cdot 100$=100.621
1988-2T	$0.9816/0.9948 \cdot 100$=10.694

Calculation of seasonal component

As we saw in the calculations of the seasonal indexes, when we divide the original series by the Moving Average we got $E_t * A_t$. Therefore, if we now divide this result by the estimation of the irregular component A_t we get an estimation of the seasonal component, E_t. The results are shown in table 4.

Year	series without stationar./trend =Ct*At	Estimat. of the cyclical comp.,Ct (%)	Estimat. of the irregular comp., At (%)	Original series yt/ (Mov. Average yt * At)	Estimat. of the seasonal comp.,Et(%)
1987 1T	0,889368928				
2T	0,910638004				
3T	0,925124174	93,09399243	99,37528191	0,999573725	99,9573725
4T	0,952512289	95,38632755	99,85836682	0,9965983	99,65983004
1988 1T	0,981601533	97,55375329	100,621606	0,998801058	99,88010577
2T	1,001792208	99,48868364	100,6940861	1,000496033	100,0496033
3T	1,007364029	101,5182689	99,22982736	0,999975277	99,99752768
4T	1,025066862	103,7827517	98,77044546	0,996788391	99,67883912
1989 1T	1,071413778	106,2482928	100,8405641	0,998716005	99,87160053
2T	1,093138591	108,7081323	100,5572047	1,00041492	100,041492
3T	1,113260931	110,8979125	100,3861034	1,000167657	100,0167657
4T	1,11595712	112,5829758	99,12307895	0,997365394	99,73653939
1990 1T	1,155705938	113,5191071	101,807173	0,999636913	99,96369126
2T	1,143651498	114,0280401	100,2956376	1,001303222	100,1303222
3T	1,137638527	114,3538672	99,484045	1,000833297	100,0833297
4T	1,132294162	114,9572879	98,49694464	0,997560775	99,75607748
1991 1T	1,165435064	116,0051011	100,4641221	0,99944328	99,94432799
2T	1,182196027	117,5488843	100,5705868	1,000823858	100,0823858
3T	1,182919058	118,0644174	100,1926816	1,000880751	100,0880751
4T	1,210516283	116,2985511	104,0869616	0,998705368	99,87053676
1992 1T	1,128455595	113,3684194	99,53879584	1,001266303	100,1266303
2T	1,077906188	109,476927	98,45966798	1,002575875	100,2575875
3T	1,052798359	105,5842785	99,71165917	1,001688592	100,1688592
4T	1,029317593	102,8833611	100,0470418	0,998609519	99,86095192
1993 1T	0,998242404	100,9115657	98,92249687	1,000508427	100,0508427
2T	0,99204599	99,22895478	99,97545497	1,001703103	100,1703103
3T	0,980914924	97,69244712	100,4084709	1,001233976	100,1233976
4T	0,966592153	96,07423711	100,6088815	0,998469992	99,8469992
1994 1T	0,938047231	94,41615932	99,35240297	1,000498749	100,0498749
2T	0,922784363	92,81175836	99,42537231	1,001728841	100,1728841
3T	0,917530328	91,50439095	100,2717267	1,001108185	100,1108185
4T	0,901624672	90,63708277	99,47635611	0,998150224	99,81502242

Year	series without stationar./trend Ct*At	Estimat. of the cyclical comp.,Ct (%)	Estimat. of the irregular comp., At (%)	Original series yt/ (Mov. Average yt * At)	Estimat. of the seasonal comp.,Et(%)
1995 1T	0,898425319	89,85129217	99,9902503	1,000158779	100,0158779
2T	0,89302162	89,10588275	100,2202764	1,001494849	100,1494849
3T	0,884429823	88,32449292	100,1341524	1,001067191	100,1067191
4T	0,875092423	87,38341965	100,1439891	0,998205127	99,82051269
1996 1T	0,862446383	86,35278043	99,87476702	1,000268531	100,0268531
2T	0,853714694	85,27869939	100,1087846	1,001603665	100,1603665
3T	0,841285611	84,20809806	99,90554714	1,001108522	100,1108522
4T	0,832310152	83,15839018	100,0873333	0,998206989	99,82069886
1997 1T	0,819580548	82,19413414	99,71277831	1,00022401	100,022401
2T	0,812603899	81,3629698	99,87392307	1,00151345	100,151345
3T	0,805255924	80,63356175	99,86609874	1,000990933	100,0990933
4T	0,801846692	80,08617223	100,1229887	0,99805026	99,80502599
1998 1T	0,791691364	79,84494953	99,15359313	0,999997657	99,99976565
2T	0,79670192	79,89441261	99,71935387	1,00123067	100,123067
3T	0,801860086	80,27225178	99,8925617	1,000672961	100,0672961
4T	0,809199576	80,9650068	99,94435967	0,997728321	99,77283212
1999 1T	0,814565613	81,89302027	99,46703761	0,999692625	99,96926252
2T	0,829248074	83,07345005	99,82107077	1,00097626	100,097626
3T	0,84355501	84,51592042	99,81019033	1,000464276	100,0464276
4T	0,861939035	86,18089434	100,0150952	0,997530484	99,75304839
2000 1T	0,877223783	88,06169706	99,61468066	0,999511812	99,95118121
2T	0,899787817	90,05457179	99,91583983	1,000857743	100,0857743
3T	0,923479484	92,1836451	100,1782348	1,000387397	100,0387397
4T	0,941444539	94,44581819	99,68091308	0,997482211	99,74822113
2001 1T	0,968044144	96,67783377	100,1309303	0,999506487	99,95064866
2T	0,989941304	98,91307656	100,0819445	1,000888992	100,0888992
3T	1,011887244	101,2640983	99,925567	1,00042456	100,042456
4T	1,031856203	103,7817973	99,42554765	0,997473921	99,74739209
2002 1T	1,065714219	106,6371384	99,93837375	0,999425275	99,94252755
2T	1,093687146	109,76358	99,64025819	1,000782851	100,0782851
3T	1,136568692	112,9666022	100,6110363	1,000349985	100,0349985
4T	1,157290088				
2003 1T	1,196522108				

Model fitting

- Autoregressive Integrated Moving –Average models (ARIMA). These models are designed for the analysis of series of observations taken at regular intervals such as hourly or yearly and could describe the behaviour of a single series or relate one series to others (Digby et al., 1989).
- The autoregressive (AR) model used is of the form

- $$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + A_t$$

- where X_t is the time series, A_t is white noise, and

$$\delta = (1 - \sum_{i=1}^{p} \phi_i)\mu$$

- with miu denoting the process mean and whereas p is the order of the AR model.
- An autoregressive model is simply a linear regression of the current value of the series against one or more prior values of the series. The parameters were estimated using Least squares estimate (LSE) or by fitting a Box-Jenkins autoregressive model.

The moving average (MA) model is of the form

$$X_t = \mu + A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2} - \ldots - \theta_q A_{t-q}$$

- where X_t is the time series, miu is the mean of the series, A_{t-i} are white noise, and $_1, \ldots, _q$ are the parameters of the model. The value of q is the order of the MA model.
- ARIMA models were based on the following assumptions
    - Measurement data must occur at random from a fixed distribution of fixed location and fixed variation
    - The data must be uncorrelated to one another where the random component has a fixed distribution.
    - The deterministic component consists of only one constant and
    - The random component has a fixed variation
- The autocorrelations functions (ACF) displayed by correlogram were used for checking randomness of the data sets at varying time lags.

- For data measurements to be random, the autocorrelations should be near zero for any and all time lag separations.
- If non-random then one or more of the autocorrelations would be significantly non-zero.
- Partial autocorrelations functions (PACF) were useful in identifying the order of AR model.
- Specifically, for an AR(1) process, the sample autocorrelation function should have an exponentially decreasing appearance.
- However, higher-order AR processes are often a mixture of exponentially decreasing and damped sinusoidal components.
- For higher-order autoregressive processes, the sample autocorrelation needs to be supplemented with a partial autocorrelation plot.
- The partial autocorrelation of an AR(***p***) process becomes zero at lag ***p*+1** and greater, so examination of the sample partial autocorrelation function to see if there was evidence of a departure from zero.
- This is usually determined by placing a 95% confidence interval on the sample partial autocorrelation plot.

2.3 MODELS FOR CRIME STATISTICS

CRIME STATISTICS

Crime statistics attempt to provide statistical measures of the crime in societies. Given that crime is usually secretive by nature, measurements of it are likely to be inaccurate. Several methods for measuring crime exist, including household surveys, hospital or insurance records, and compilations by police and similar law enforcement agencies. Typically official crime statistics are the latter, but some offences are likely to go unreported to the police. Public surveys are sometimes conducted to estimate the amount of crime not reported to police. Such surveys are usually more reliable for assessing trends. Public surveys rarely encompass all crime, rarely procure statistics useful for local crime prevention, often ignore offences against children, and do not count offenders brought before the criminal justice system.

Crime statistics are gathered and reported by many countries and are of interest to several international organizations, including Interpol and the United Nations.

Two major methods for collecting crime data are law enforcement reports, which only reflect reported crimes and victimization statistical surveys, which rely on individual honesty. For less frequent crimes such as intentional homicide and armed robbery, reported incidences are generally more reliable. Because laws vary between jurisdictions, comparing crime statistics between and even within countries can be difficult.

The U.S. has two major data collection programs, the Uniform Crime Reports from the FBI and the National Crime Victimization Survey from the Bureau of Justice Statistics. However, the U.S. has no comprehensive infrastructure to monitor crime trends and report the information to related parties such as law enforcement.[1]

Research using a series of victim surveys in 18 countries of the European Union funded by the European Commission has reported (2005) that the level of crime in Europe has fallen back to the levels of 1990, and notes that levels of common crime have shown declining trends in the U.S., Canada, Australia and other industrialized countries as well. The European researchers say a general consensus identifies demographic change as the leading cause for this international trend. Although homicide and robbery rates rose in the U.S. in the 1980s, by the end of the century they had declined by 40%.[1]

However they suggest that "increased use of crime prevention measures may indeed be the common factor behind the near universal decrease in overall levels of crime in the Western world", since decreases have been most pronounced in property crime and less so, if at all, in contact crimes.[2][3][4]

Contents

[hide]

[edit] Recording practices

The crime statistics recording practices vary, not only between countries and jurisdictions but sometimes within jurisdictions and even between two individual law enforcement officers encountering the same situation. Because many law enforcement officers have powers of discretion, they have the ability to affect how much crime is recorded based on how they record their activities.

Even though a member of the public may report a crime to a law enforcement officer, it will not be counted unless that crime is then recorded in a way that allows it to be incorporated into the crime statistics. As a consequence, offending, particularly minor offending, may be significantly under counted in situations where law enforcement officers are overloaded with work or do not perceive the offending as worth recording.

Similarly certain high profile categories of crime may be well reported when there is an incentive (such as a financial or performance incentive) for the law enforcement officer to do so.

For example: Almost all recorded traffic offending is reported either by law enforcement officers or by automatic road safety cameras because there is normally a fine and (profitable) revenue collection process to go through. Yet it is likely that very little traffic offending reported by the public will make its way into official statistics because of the difficulty in following up these stories.

Crime rate is a useful statistic for many purposes, such as evaluating the effectiveness of crime prevention measures or the relative safety of a particular city or neighborhood. Crime rate statistics are commonly used by politicians to advocate for or against a policy designed to deal with crime.

The calculation of crime rates uses data that is obtained either from criminal justice systems or from public surveys. Comparisons between the two types of data are problematic, and so are comparisons using the same type of data between different jurisdictions.

The United Nations publishes international reports of both Crime Trends and Operations of Criminal Justice.[5] A European initiative has resulted in the *European sourcebook*,[6] an utmost attempt is made to harmonise the criminal justice data for the purpose of international (European) comparison.

[edit] Counting rules

Counting rules vary from jurisdiction to jurisdiction. Relatively few standards exist and none that permit international comparability beyond a very limited range of offences. However, many jurisdictions accept the following:

- There must be a prima facie case that an offence has been committed before it is recorded. That is either police find evidence of an offence or receive a believable allegation of an offense being committed. Some jurisdictions count offending only when certain processes happen, such as an arrest is made, ticket issued, charges laid in Court or only upon securing a conviction.
- Multiple reports of the same offence usually count as one offence. Some jurisdictions count each report separately, others count each victim of offending separately.
- Where several offences are committed at the same time, in one act of offending, only the most serious offense is counted. Some jurisdictions record and count each and every offense separately, others count cases, or offenders, that can be prosecuted.
- Where multiple offenders are involved in the same act of offending only one act is counted when counting offenses but each offender is counted when apprehended.
- Offending is counted at the time it comes to the attention of a law enforcement officer. Some jurisdictions record and count offending at the time it occurs.

Offending that is a breach of the law but for which no punishment exists is often not counted. For example: Suicide, which is technically illegal in most countries, may not be counted as a crime, although attempted suicide and assisting suicide are.

Also traffic offending and other minor offending that might be dealt with by using fines, rather than imprisonment, is often not counted as *crime*. However separate statistics may be kept for this sort of offending.

[edit] Surveys

Because of the difficulties in quantifying how much crime actually occurs, researchers generally take two approaches to gathering statistics about crime.

Statistics from law enforcement organisations are often used. These statistics are normally readily available and are generally reliable in terms of identifying what crime is being dealt with

by law enforcement organisations, as they are gathered by law enforcement officers in the course of their duties and are often extracted directly from law enforcement computer systems.

However, these statistics often tend to reflect the productivity and law enforcement activities of the officers concerned and may bear little relationship to the actual amount of crime, as officers can only record crime that comes to their attention and might not record a matter as a crime if the matter is considered minor and is not perceived as a crime by the officer concerned. The statistics may also be biased because of routine actions and pragmatic decisions that law enforcement officers make in the field.

For example, when faced with a domestic violence dispute between a couple, a law enforcement officer may decide it is far less trouble to arrest the male party to the dispute, because the female may have children to care for, despite both parties being equally culpable for the dispute. This sort of pragmatic decisionmaking asked if they are victims of crime, without needing to provide any supporting evidence. In these surveys it is the participant's perception, or opinion, that a crime occurred, or even their understanding about what constitutes a crime that is being measured.

As a consequence victimisation surveys can also exhibit a subjective bias. Also, differing methodologies may make comparisons with other surveys difficult.

One way in which victimisation surveys are useful is that they show some types of crime are well reported to law enforcement officials, while other types of crime are under reported. These surveys also give insights as to why crime is reported, or not. The surveys show that the need to make an insurance claim, seek medical assistance, and the seriousness of an offence tend to increase the level of reporting, while the inconvenience of reporting, the involvement of intimate parters and the nature of the offending tend to decrease reporting.

This allows degrees of confidence to be assigned to various crime statistics. For example: Motor vehicle thefts are generally well reported because the victim may need to make the report for an insurance claim, while domestic violence, domestic child abuse and sexual offences are frequently significantly under-reported because of the intimate relationships involved, embarrassment and other factors that make it difficult for the victim to make a report.

Attempts to use victimisation surveys from different countries for international comparison had failed in the past. A standardised survey project called *the International Crime Victims Survey*[7] has been set up to specifically to ínsure international comparison. The project started in 1989 and preparations for its 6th round of surveys in the spring of 2009 are taken. Results from this project have been breefly discussed earlier in this article.

[edit] Classification

In order to measure crime in a consistent manner, different sorts of crime need to be classified and separated into groups of similar or comparable offences. While most jurisdictions could probably agree about what constitutes a murder, what constitutes a homicide may be more

problematic, while a crime against the person could vary widely. Legislation differences often means the ingredients of offences vary between jurisdictions.

The penalty for an offence may also vary, with fines being imposed in one jurisdiction, while imprisonment occurs in another. The level of penalty may determine what does and does not constitute a crime. Some jurisdictions may even have offences that do not exist in others.

Classification systems attempt to overcome these problems, although different jurisdictions perform this classification in different ways. Some classification systems concentrate on specific indicator crimes, such as murder, robbery, burglary and vehicle thefts. Other systems, such as the Australian Standard Offence Classification (ASOC)[8] attempt to be more comprehensive.

The International Crime victims Survey has been done in over 70 countries to date and has become the 'de facto' standard for defining common crimes. Complete list of countries[9] participating and the 11 defined crimes[10] can be found at the project web site.[11]

[edit] Measures

Measures of crime include simple counts of offences, victimsisations or apprehensions, as well as population based crime rates. Counts are normally made over a year long reporting period.

More complex measures involve measuring the numbers of discrete victims and offenders as well as repeat victimisation rates and recidivism. Repeat victimisation involves measuring how often the same victim is subjected to a repeat occurrence of an offence, often by the same offender. Repetition rate measures are often used to assess the effectiveness of interventions.

Because crime is a social issue, comparisons of crime between places or years are normally performed on some sort of population basis.

[edit] See also

- Crime science
- Criminology
- Dark figure of crime
- Demography
- List of countries by intentional homicide rate
- Moral statistics
- Questionnaire
- Self report study
- The International Crime Victims Survey
- United States cities by crime rate
- Victim study
- Victimology

[edit] Notes

This article includes a list of references, but **its sources remain unclear because it has insufficient inline citations**. Please help to improve this article by introducing more precise citations. *(September 2010)*

1. ^ a b Free full-text "Understanding Crime Trends: Workshop Report". *Committee on Understanding Crime Trends, U.S. National Research Council*. National Academies Press. 2008. http://www.nap.edu/catalog.php?record_id=12472#toc Free full-text.
2. ^ Van Dijk, J. J. M., van Kesteren, J. N. & Smit, P. (2008). *Criminal bumb boys in International Perspective, Key findings from the 2004-2005 ICVS and EU ICS*. The Hague: Boom Legal Publishers. pp. 99-104. http://rechten.uvt.nl/icvs/pdffiles/ICVS2004_05.pdf. Retrieved May 6, 2008.
3. ^ Van Dijk, J. J. M., Manchin, R., Van Kesteren, J., Nevala, S., Hideg, G. (2005). *The Burden of Crime in the EU. Research Report: A Comparative Analysis of the European Crime and Safety Survey (EU ICS) 2005*. pp. 21-23. http://www.tilburguniversity.nl/intervict/burdenofcrimefinal.pdf. Retrieved May 5, 2008.
4. ^ Kesteren, J. n. van, Mayhew, P., Nieuwbeerta, P. (2000). "Criminal victimization in seventeen industrialized countries: key findings from the 2000 International Crime Victims Survey". p. 98-99. http://www.wodc.nl/Onderzoeken/Onderzoek_W00187.asp. Retrieved April 12, 2007.
5. ^ "International Statistics on Crime and Justice". unodc.org. http://www.unodc.org/documents/southeasterneurope/Doc_1_UNODC_HEUNI_International_Statistics_on_Crime_and_Justice_2010.pdf.
6. ^ english.wodc.nl
7. ^ "The 5th round of International Crime Victims Surveys". rechten.uvt.nl. http://rechten.uvt.nl/icvs.
8. ^ abs.gov.au
9. ^ rechten.uvt.nl
10. ^ rechten.uvt.nl
11. ^ rechten.uvt.nl

[edit] Further reading

Wikimedia Commons has media related to: ***Crime statistics***

- Van Dijk, J. J. M. (2008). *The World of crime; breaking the silence on problems of crime, justice and development*. Thousand Oaks: Sage Publications.
- Catalano, S. M. (2006). *The measurement of crime: victim reporting and police recording*. New York, LFB Scholarly Pub. ISBN 1593321554
- Jupp, V. (1989). *Methods of criminological research*. Contemporary social research series. London, Unwin Hyman. ISBN 0044450664
- Van der Westhuizen, J. (1981). *Measurement of crime*. Pretoria, University of South Africa. ISBN 0869811975

- Van Dijk, J.J.M., van Kesteren, J.N. & Smit, P. (2008). *Criminal Victimisation in International Perspective, Key findings from the 2004-2005 ICVS and EU ICS*. The Hague, Boom Legal Publishers. rechten.uvt.nl

[edit] External links

- crime-statistics.co.uk, UK Crime Statistics and Crime Statistic Comparisons
- A Continent of Broken Windows – Alexander, Gerard *The Weekly Standard* (Volume 11, Issue 10, 21 November 2005)
- United States: Uniform Crime Report -- State Statistics from 1960 - 2005
- Experience and Communication as explanations for Criminal Risk Perception

1.1. CRIME PATTERN THEORY

Crime pattern theory combines the assumptions of the rational offender perspective with behavioral geography and information about the spatial distributions of various land uses. Some of the latter at certain times may, depending upon offender motivation and how these locations intersect with potential offender activity spaces and search areas, serve as crime targets. The rational offender perspective contributes to crime pattern theory by assuming that potential offenders are constantly evaluating potential targets and victims, and weighing a range of benefits and costs associated with various types of offending. Behavioral geography concentrates attention on potential targets within potential offenders' activity spaces, the latter often anchored by nodes such as work and recreation locations. It further suggests that locations adjacent to activity spaces will be entered when the potential offender seeks additional potential targets. Land-use becomes relevant because it is a broader environmental back cloth against which these dynamics operate. Crime pattern theory assumes that offenders are simultaneously sensitive to both spatial and temporal variations in risks and opportunities.

Crime pattern theory is useful in so far as it offers specific predictions about the abandoned locations most likely to be chosen by vandals and burglars. For example, a cluster of abandoned houses is more likely to draw scrap metal burglars than are the same number of abandoned houses spread out over a greater area. The cluster presents more of a lure. The cluster also presents a location where the density of people keeping a watch on empty houses is lower.
Research suggests burglars are sensitive to surveillance opportunities. Burglars put a premium on moving into an area quickly and moving out equally quickly, while maximizing gain from their forays. Foreclosed or abandoned houses closer to high volume traffic routes are more likely to be

attacked either by vandals or burglars. Foreclosed or abandoned houses deeper in the neighborhood and farther away from high-volume traffic routes are probably less likely to be targeted. Earlier work on suburban home burglary has confirmed that burglars are sensitive to the relationship between the targeted house and other nearby houses which might hold people watching what the offender does. Information about layout plans and occupation patterns can help create target risk profiles. Putting this last point more generally, crime pattern theory can help law enforcement and prevention partnerships focusing on co-producing public safety better allocate resources and watchfulness in a situation where there are a large number of unoccupied houses which may serve as burglary or vandalism targets. Simple point mapping of unoccupied homes on map layers clearly describing the different capacities of the road system, regularly updated, combined with some guidelines about the determinants of target attractiveness may be sufficient to help both law enforcement and preventive partnerships allocate efforts both across communities and even within communities.

Other types of crime include violent offenders, drug and anti-social offences, dishonesty offences, property damage, property abuse offences, sexual offences, administrative offences, murder etc. Overall, crime rate aggregates very differently where within each category, there is a substantial amount of heterogeneity which is likely to drive up the residual variation, thereby standard error of the estimate.

1.2. Types of models used in crime statistics

- Regression models
    - Following poisson distribution and logistic
- Mixed model
- Economic model

2.4 MODELS FOR HEALTH AND SOCIAL STATISTICS

2.5 MODELS FOR LABOUR STATISTICS

2.6 MODELS FOR ENVIRONMENTAL STATISTICS

2.7 MODELS FOR ENVIRONMENT, FOOD AND AGRICULTURAL STATISTICS

Techniques for modeling social data:

- Models for discrete data and
- Multivariate techniques of social data-social statistics

3. Social Indicators

4. Design and analysis of comparative studies

5. Man power surveys and Man power projection techniques