Econometrics

UTKAL UNIVERSITY

NAAC ACCREDITED A+
DDCE
Education for all

ଦୂରନିରନ୍ତର ଶିକ୍ଷା ନିର୍ଦ୍ଦେଶାଳୟ, ଉତ୍କଳ ବିଶ୍ୱବିଦ୍ୟାଳୟ

DIRECTORATE OF DISTANCE AND CONTINUING EDUCATION
Utkal University, Bhubaneswar - 7, Odisha
E-mail: helpline@ddceutkal.ac.in
Website: www.ddceutakal.ac.in

Name of Study Material: .Econometrics

ISBN: ....

Author Name’s: ....

Year of Print: ....

Printed and Published by:

Mrs. Meena Pandey
Himalaya Publishing House Pvt. Ltd.,
"Ramdoot", Dr. Bhalerao Marg, Girgaon,
Mumbai-400 004. Phones: 23860170 & 23863863, Fax: 022-23877178
Email: himpub@vsnl.com; **Website:** www.himpub.com

For:

Directorate of Distance and Continuing Education
Utkal University, Vani Vihar, Bhubaneswar - 751 007.
www.ddceutkal.ac.in

UTKAL UNIVERSITY
DDCE
Education for All

DIRECTORATE OF DISTANCE AND CONTINUING EDUCATION
UTKAL UNIVERSITY : VANI VIHAR
BHUBANESWAR – 751 007

From the Director’s Desk

The Directorate of Distance & Continuing Education, originally established as the University Evening College way back in 1962 has travelled a long way in the last 56 years. **‘EDUCATION FOR ALL’** is our motto. Increasingly the Open and Distance Learning institutions are aspiring to provide education for anyone, anytime and anywhere. DDCE, Utkal University has been constantly striving to rise up to the challenges of Open Distance Learning system. Nearly one lakh students have passed through the portals of this great temple of learning. We may not have numerous great tales of outstanding academic achievements but we have great tales of success in life, of recovering lost opportunities, tremendous satisfaction in life, turning points in career and those who feel that without us they would not be where they are today. There are also flashes when our students figure in best ten in their honours subjects. Our students must be free from despair and negative attitude. They must be enthusiastic, full of energy and confident of their future. To meet the needs of quality enhancement and to address the quality concerns of our stake holders over the years, we are switching over to self instructional material printed courseware. We are sure that students would go beyond the course ware provided by us. We are aware that most of you are working and have also family responsibility. Please remember that only a busy person has time for everything and a lazy person has none. We are sure, that you will be able to chalk out a well planned programme to study the courseware. By choosing to pursue a course in distance mode, you have made a commitment for self improvement and acquiring higher educational qualification. You should rise up to your commitment. Every student must go beyond the standard books and self instructional course material. You should read number of books and use ICT learning resources like the internet, television and radio programmes etc. As only limited number of classes will be held, a student should come to the personal contact programme well prepared. The PCP should be used for clarification of doubt and counseling. This can only happen if you read the course material before PCP. You can always mail your feedback on the course ware to us. It is very important that one should discuss the contents of the course materials with other fellow learners.

We wish you happy reading.

DIRECTOR

SYLLABUS

GROUP-A - ECO.3.2: ECONOMETRICS

UNIT I 1. The Econometric Approach (i) Meaning and objectives of Econometrics. (ii) The sources of hypothesis used in Econometrics. (iii) The raw materials of econometrics- Time series and Cross section data: the problem of their pooling together. 2. Elements of Statistical inferences. (i) Point and interval estimation- estimator and its properties, Method of Maximum Likelihood, interval Estimation- confidence interval. (ii) Test of Hypothesis- Simple and composite hypothesis, two types of errors, Neyman Pearson Lemma, Power Function of a test, Likelihood ratio, Test Exact Sampling Distributions, Z-statistics, Chi-square, t-statistics and F-statistics.

UNIT II 3. Classical Linear Regression (with one explanatory variable) (i) Assumption and their economic interpretation, Least square estimations of regression parameters, their properties, Gauss-Markov Theory, Theorem: Standard errors of estimates. Estimator of errors, Control limit theorem, Maximum likelihood estimator. (ii) Normality of errors, control limit theorem, Maximum Likelihood Estimator. (iii) Significance test and confidence intervals of estimates-z-test, t-test and f-ratio test. (iv) Prediction point and interval.

4. Extension of the two variable linear model:

(i) Three–variable linear model, the coefficient of multiple correlation, partial correlation coefficient. (ii) General Linear model (with K- Explanatory variable)- Assumptions, Least-square estimates and their properties, Variance- covariance matrix of estimates, Estimates of error variance, Multiple coefficient of determination- R2 and multiple correlation coefficient- R. Significance test and confidence intervals, prediction. (iii) Non-linear Models-Choice of functional forms, estimation.

UNIT III (i) Extensions of the general model: Dummy variables, Use of dummy variable in seasonal analysis, dummy dependant variable.

UNIT IV (i) Violations of the assumptions of the classical model. (ii) Errors in variables consequence, Methods of estimation-classical method of maximum likelihood, use of instrumental variable. (iii) Autocorrelation –Sources, Consequences, GLSM. Tests for autocorrelation, Remedial measures, Prediction. (iv) Heteroscedasticity- Nature and consequences, Heteroscedasticity structures, Tests for Heteroscedasticity, Remedial measures the methods of weighted least square. (v) Multicolinearity- Implications consequences, Tests for multicolinearity, Methods of estimation Multicolinearity and prediction, Remedial measures.

UNIT V 1. Distributive Lag Models: Lagged exogenous and endogenous methods, consequences of applying OLMS to lagged and generous model. Estimation of distribution log models KOYCK’s approach, Adaptive expectation, Use of instrumental variable, Almon’s approach. 2. Simultaneous Equations Methods. (i) Jointly dependent and predetermined variables, structural form reduced form, final form. (ii) The identification problem- Rank and order conditioned. (iii) Methods of Estimation- Method of Indirect least squares 2 LS, Method of instrumental variable MLIML, 3 SLS and FIMLM.

CONTENTS

1

Lesson

THE ECONOMETRIC APPROACH

Objectives

The objectives of this lesson are to:

- The Econometric Approach
- Meaning and objectives of Econometrics
- The sources of hypothesis used in Econometrics
- The raw materials of econometrics
- Time series and Cross section data: the problem of their pooling together
- Elements of Statistical inferences
- Point and interval estimation- estimator and its properties, Method of
- Maximum Likelihood, interval Estimation- confidence interval
- Test of Hypothesis - Simple and composite hypothesis, two types of errors, Neyman Pearson Lemma, Power Function of a test, Likelihood ratio, Test Exact Sampling Distributions, Z-statistics, Chi-square, t-statistics and F-statistics

Structure:

Notes

1.1 INTRODUCTION

Econometrics is the analysis and testing of economic theories to verify hypotheses and improve prediction of financial trends. Econometrics takes mathematical and statistical models proposed in economic theory and tests them. First, models are tested against statistical trials, followed by testing against real-world examples to support or disprove hypotheses. Econometrics uses an important statistical method called regression analysis, which assesses the connection among variables. Economists use the regression method since they cannot usually carry out controlled experiments, choosing to instead gather information from natural experiments.

It is an integration of economics, mathematical economics and statistics with an objective to provide numerical values to the parameters of economic relationships. The relationships of economic theories are usually expressed in mathematical forms and combined with empirical economics. The econometrics methods are used to obtain the values of parameters which are essentially the coefficients of mathematical form of the economic relationships. The statistical methods which help in explaining the economic phenomenon are adapted as econometric methods. The econometric relationships depict the random behaviour of economic relationships which are generally not considered in economics and mathematical formulations.

It may be pointed out that the econometric methods can be used in other areas like engineering sciences, biological sciences, medical sciences, geosciences, agricultural sciences etc. In simple words, whenever there is a need of finding the stochastic relationship in mathematical format, the econometric methods and tools help. The econometric tools are helpful in explaining the relationships among variables.

1.2 THE ECONOMETRIC APPROACHES OR METHODS

The Econometric Approaches that make use of statistical tools and economic theories in combination to estimate the economic variables and to forecast the intended variables. Various approaches of econometrics are:

1. Regression Approach

The regression approach is the most common method used to forecast the demand for a product. This method combines the economic theory with statistical tools of estimation. The economic theory is applied to specify the demand determinants and the nature of the relationship between product’s demand and its determinants. Thus, through an economic theory, a general form of a demand function is determined. The statistical techniques are applied to estimate the values of parameters in the projected equation.

Under the regression method, the first and the foremost thing is to determine the demand function. While specifying the demand functions for several commodities, one may come across many commodities whose demand depends by or large, on a single independent variable. For example, suppose in a city, the demand for items like tea and coffee is found to depend largely on the population of the city, then the demand functions of these items are said to be single-variable demand functions.

On the other hand, if it is found out that the demand for commodities like sweets, ice-creams, fruits, vegetables, etc., depends on a number of variables like commodity’s own price, the price of substitute goods, household incomes, population, etc. Then such demand functions are called as multi-variable demand functions.

Thus, for a single variable demand function, the simple regression equation is used while for multiple variable functions, a multi-variable equation is used for estimating the demand for a product.

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

2. Simultaneous Equations Approach

Under simultaneous equation model, demand forecasting involves the estimation of several simultaneous equations. These equations are often the behavioral equations, market-clearing equations, and mathematical identities.

The regression technique is based on the assumption of one-way causation, which means independent variables cause variations in the dependent variables, and not vice-versa. In simple terms, the independent variable is in no way affected by the dependent variable. For example, D = a – bP, which shows that price affects demand, but demand does not affect the price, which is an unrealistic assumption.

On the contrary, the simultaneous equations model enables a forecaster to study the simultaneous interaction between the dependent and independent variables. Thus, simultaneous equation model is a systematic and complete approach to forecasting. This method employs several mathematical and statistical tools of estimation.

The econometric methods are most widely used in forecasting the demand for a product, for a group of products and the economy as a whole. The forecast made through these methods is more reliable than the other forecasting methods.

1.3 MEANING OF ECONOMETRICS

Econometrics is the quantitative application of statistical and mathematical models using data to develop theories or test existing hypotheses in economics, and for forecasting future trends from historical data. It subjects real-world data to statistical trials and then compares and contrasts the results against the theory or theories being tested. Depending on if you are interested in testing an existing theory or using existing data to develop a new hypothesis based on those observations, econometrics can be subdivided into two major categories: theoretical and applied. Those who routinely engage in this practice are commonly known as econometricians.

1.4 OBJECTIVES OF ECONOMETRICS

The three main objectives of econometrics are as follows:

1. ***Formulation and specification of econometric models:*** The economic models are formulated in an empirically testable form. Several econometric models can be derived from an economic model. Such models differ due to different choice of functional form, specification of stochastic structure of the variables etc.

2. ***Estimation and testing of models:*** The models are estimated on the basis of observed set of data and are tested for their suitability. This is the part of statistical inference of the modeling. Various estimation procedures are used to know the numerical values of the unknown parameters of the model. Based on various formulations of statistical models, a suitable and appropriate model is selected.

3. ***Use of models:*** The obtained models are used for forecasting and policy formulation which is an essential part in any policy decision. Such forecasts help the policy makers to judge the goodness of fitted model and take necessary measures in order to re-adjust the relevant economic variables.

Econometrics and Statistics

Econometrics differs both from mathematical statistics and economic statistics. In economic statistics, the empirical data is collected recorded, tabulated and used in describing the pattern in their development over time. The economic statistics is a descriptive aspect of economics. It does not provide either the explanations of the development of various variables or measurement of the parameters of the relationships.

Statistical methods describe the methods of measurement which are developed on the basis of controlled experiments. Such methods may not be suitable for economic phenomenon as they don't fit in the framework of controlled experiments. For example, in real world experiments, the variables usually change continuously and simultaneously and so the setup of controlled experiments is not suitable.

Econometrics uses statistical methods after adapting them to the problems of economic life. These adopted statistical methods are usually termed as econometric methods. Such methods are adjusted so that they become appropriate for the measurement of stochastic relationships. These adjustments basically attempts to specify attempts to the stochastic element which operate in real world data and enters into the determination of observed data. This enables the data to be called as random sample which is needed for the application of statistical tools.

The theoretical econometrics includes the development of appropriate methods for the measurement of economic relationships which are not meant for controlled experiments conducted inside the laboratories.

The econometric methods are generally developed for the analysis of non-experimental data.

The applied econometrics includes the application of econometric methods to specific branches of econometric theory and problems like demand, supply, production, investment, consumption etc. The applied econometrics involves the application of the tools of econometric theory for the analysis of economic phenomenon and forecasting the economic behaviour.

Applications of Econometrics

Econometrics can be used in various areas which are as follows:

1. Forecasting macroeconomic indicators

Some macroeconomists are concerned with the expected effects of monetary and fiscal policy on the aggregate performance of the economy. Time-series models can be used to make predictions about these economic indicators.

Notes

2. Estimating the impact of immigration on native workers

Immigration increases the supply of workers, so standard economic theory predicts that equilibrium wages will decrease for all workers. However, since immigration can also have positive demand effects, econometric estimates are necessary to determine the net impact of immigration in the labor market.

3. Identifying the factors that affect a firm's entry and exit into a market

The microeconomic field of industrial organization, among many issues of interest, is concerned with firm concentration and market power. Theory suggests that many factors, including existing profit levels, fixed costs associated with entry/exit, and government regulations can influence market structure. Econometric estimation helps determine which factors are the most important for firm entry and exit.

4. Determining the influence of minimum-wage laws on employment levels

The minimum wage is an example of a price floor, so higher minimum wages are supposed to create a surplus of labor (higher levels of unemployment). However, the impact of price floors like the minimum wage depends on the shapes of the demand and supply curves. Therefore, labor economists use econometric techniques to estimate the actual effect of such policies.

5. Finding the relationship between management techniques and worker productivity

The use of high-performance work practices (such as worker autonomy, flexible work schedules and other policies designed to keep workers happy) has become more popular among managers. At some point, however, the cost of implementing these policies can exceed the productivity benefits. Econometric models can be used to determine which policies lead to the highest returns and improve managerial efficiency.

6. Measuring the association between insurance coverage and individual health outcomes

One of the arguments for increasing the availability (and affordability) of medical insurance coverage is that it should improve health outcomes and reduce overall medical expenditures. Health economists may use econometric models with aggregate data (from countries) on medical coverage rates and health outcomes or use individual-level data with qualitative measures of insurance coverage and health status.

7. Deriving the effect of dividend announcements on stock market prices and investor behavior

Dividends represent the distribution of company profits to its shareholders. Sometimes the announcement of a dividend payment can be viewed as good news when shareholders seek investment income, but sometimes they can be viewed as bad news when shareholders prefer reinvestment of firm profits through retained earnings. The net effect of dividend announcements can be estimated using econometric models and data of investor behavior.

8. Predicting revenue increases in response to a marketing campaign

The field of marketing has become increasingly dependent on empirical methods. A marketing or sales manager may want to determine the relationship between marketing efforts and sales. How much additional revenue is generated from an additional dollar spent on advertising? Which type of advertising (radio, TV, newspaper, and so on) yields the largest impact on sales? These types of questions can be addressed with econometric techniques.

9. Calculating the impact of a firm's tax credits on R&D expenditure

Tax credits for research and development (R&D) are designed to provide an incentive for firms to engage in activities related to product innovation and quality improvement. Econometric estimates can be used to determine how changes in the tax credits influence R&D expenditure and how distributional effects may produce tax-credit effects that vary by firm size.

10. Estimating the impact of cap-and-trade policies on pollution levels

Environmental economists have discovered that combining legal limits on emissions with the creation of a market that allows firms to purchase the "right to pollute" can reduce overall pollution levels. Econometric models can be used to determine the most efficient combination of state regulations, pollution permits, and taxes to improve environmental conditions and minimize the impact on firms.

1.5 THE SOURCES OF HYPOTHESIS USED IN ECONOMETRICS

A hypothesis (plural hypotheses) is a proposed explanation for a phenomenon. For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it. Scientists generally base scientific hypotheses on previous observations that cannot satisfactorily be explained with the available scientific theories. Even though the words "hypothesis" and "theory" are often used synonymously, a scientific hypothesis is not the same as a scientific theory. A working hypothesis is a provisionally accepted hypothesis proposed for further research, in a process beginning with an educated guess or thought.

In its ancient usage, hypothesis referred to a summary of the plot of a classical drama. The English word hypothesis comes from the ancient Greek word hypothesis, meaning "to put under" or "to suppose".

In Plato's Meno, Socrates dissects virtue with a method used by mathematicians, that of "investigating from a hypothesis." In this sense, 'hypothesis' refers to a clever idea or to a convenient mathematical approach that simplifies cumbersome calculations. Cardinal Bellarmine gave a famous example of this usage in the warning issued to Galileo in the early 17th century: that he must not treat the motion of the Earth as a reality, but merely as a hypothesis.

In common usage in the 21st century, a hypothesis refers to a provisional idea whose merit requires evaluation. For proper evaluation, the framer of a hypothesis needs to define specifics in operational terms. A hypothesis requires more work by the researcher in order to either confirm or disprove it. In due course, a confirmed hypothesis may become part of a theory or occasionally may grow to become a theory itself. Normally, scientific hypotheses have the form of a mathematical model. Sometimes, but not always, one can also formulate them as existential statements, stating that some particular instance of the phenomenon under examination has some characteristic and causal explanations, which have the general form of universal statements, stating that every instance of the phenomenon has a particular characteristic.

In entrepreneurial science, a hypothesis is used to formulate provisional ideas within a business setting. The formulated hypothesis is then evaluated where either the hypothesis is proven to be "true" or "false" through a verifiability or falsifiability-oriented experiment.

Any useful hypothesis will enable predictions by reasoning (including deductive reasoning). It might predict the outcome of an experiment in a laboratory setting or the observation of a phenomenon in nature. The prediction may also invoke statistics and only talk about probabilities. Karl Popper, following others, has argued that a hypothesis must be falsifiable, and that one cannot regard a proposition or theory as scientific if it does not admit the possibility of being shown false. Other philosophers of science have rejected the criterion of falsifiability or supplemented it with other criteria, such as verifiability (e.g., verifications) or coherence (e.g., confirmation holism). The scientific method involves experimentation, to test the ability of some hypothesis to adequately answer the question under investigation. In contrast, unfettered observation is not as likely to raise unexplained issues or open questions in science, as would the formulation of a crucial experiment to test the hypothesis. A thought experiment might also be used to test the hypothesis as well.

In framing a hypothesis, the investigator must not currently know the outcome of a test or that it remains reasonably under continuing investigation. Only in such cases does the experiment, test or study potentially increase the probability of showing the truth of a hypothesis. If the researcher already knows the outcome, it counts as a "consequence" and the researcher should have already considered this while formulating the hypothesis. If one cannot assess the predictions by observation or by experience, the hypothesis needs to be tested by others providing observations. For example, a new technology or theory might make the necessary experiments feasible.

There are diverse sources of hypothesis in research. First, an explorative research work might lead to the establishment of hypothesis. Second, the environment is a source of hypothesis, because environment portrays broad relationship across factors which form the basis for drawing an inference. Third, analogies are a source of hypothesis. The term analogies refer to parallelism. Though human system and animal system are different, there is some parallelism. That is why medicines are tried first on rats or monkeys then used for human consumption. So, hypothesis on animal behavior can be done based on proven behavior of human and vice versa. Similarly, between thermodynamics and group dynamics, biological system and social system, nervous system and central processing unit of a computer, parallelism can be thought of and spring hypotheses therefrom. Fourth, previous research studies are a great source of hypotheses. That is why review of literature is made. Fifth, assumptions of certain theories become a source of hypothesis in research. Similarly, exceptions to certain theory are ground for new hypotheses. Sixth, personal experiences and experiences of others are another source of hypotheses. Everyone encounters numerous experiences in day to day life in relation to one’s avocation. From these glimpses of hypothetical relations between events, variables, etc. emanate. These are, therefore, bases for establishment of possible hypotheses. Seventh, social, physical and other theories and laws provide for hypotheses. Newton’s laws of motion might be a source of hypotheses, in social science, say behavior and reward and the like. Finally, for the research mind, the whole universe is a source of hypotheses. The searching mind fathoms out new hypotheses from seemingly events of insignificance.

1. General Culture in which a Science Develops

A cultural pattern influences the thinking process of the people and the hypothesis may be formulated to test one or more of these ideas. Cultural values serve to direct research interests. The function of culture has been responsible for developing today’s science to a great dimension. In the words of Goode and Hatt, “to say that the hypotheses are the product of the cultural values does not make them scientifically less important than others, but it does at least indicate that attention has been called to them by the culture itself.

Notes

For example, in the Western society race is thought to be an important determinant of human behaviour. Such a proposition can be used to formulate a hypothesis. We may also cite metaphysical bias and metaphysical ideas of Indian culture to have been responsible for the formulation of certain types of hypotheses. It implies that cultural elements of common cultural pattern may form a source of the formulation of hypotheses.

2. Scientific Theory

A major source of hypothesis is theory. A theory binds a large body of facts by positing a consistent and lawful relationship among a set of general concepts representing those facts. Further generalizations are formed on the basis of the knowledge of theory. Corollaries are drawn from the theories.

These generalizations or corollaries constitute a part of hypothesis. Since theories deal with abstractions which cannot be directly observed and can only remain in the thought process, a scientific hypothesis which is concerned with observable facts and observable relationship between facts can only be used for the purpose of selecting some of the facts as concrete instances of the concepts and for making a tentative statement about the existence of a relation among the selected facts with the purpose of subjecting the relation to an empirical test."

A hypothesis emerges as a deduction from theory. Hence, hypotheses become "working instruments of theory" Every worthwhile theory provides for the formulation of additional hypothesis. "The hypothesis is the backbone of all scientific theory construction; without it, confirmation or rejection of theories would be impossible."

The hypotheses when tested are "either proved or disproved and in turn constitute further tests of the original theory." Thus the hypothetical type of verbal proposition forms the link between the empirical propositions or facts and the theories. The validity of a theory can be examined only by means of scientific predictions or experimental hypothesis.

3. Analogies

Observation of a similarity between two phenomena may be a source of formation of a hypothesis aimed at testing similarity in any other respect. Julian Huxley has pointed out that "casual observation in nature or in the framework of another science may be a fertile source of hypothesis. The success of a system in one discipline can be used in other discipline also. The theory of ecology is based on the observation of certain plants in certain geographical conditions. As such, it remains in the domain of Botany. On the basis of that the hypothesis of human ecology could be conceived.

Hypothesis of social physics is also based on analogy. "When the hypothesis was born out by social observation, the same term was taken into sociology. It has become an important idea in sociological theory". Although analogy is not always considered, at the time of formulation of hypothesis; it is generally satisfactory when it has some structural analogies to other well established theories. For the systematic simplicity of our knowledge, the analogy of a hypothesis becomes inversely helpful. Formulation of an analogous hypothesis is construed as an achievement because by doing so its interpretation is made easy.

4. Consequences of Personal, Idiosyncratic Experience as the Sources of Hypothesis

Not only culture, scientific theory and analogies provide the sources of hypothesis, but also the way in which the individual reacts to each of these is also a factor in the statement

Notes

of hypotheses. Certain facts are present, but every one of us is not able to observe them and formulate a hypothesis.

Referring to Fleming's discovery of penicillin, Backrach has maintained that such discovery is possible only when the scientist is prepared to be impressed by the 'unusual'. An unusual event struck Fleming when he noted that the dish containing bacteria had a green mould and the bacteria were dead. Usually he would have washed the dish and have attempted once again to culture the bacteria.

But normally, he was moved to bring the live bacteria in close contact with the green mould, resulting in the discovery of penicillin. The example of Sir Issac Newton, the discoverer of the theory of Gravitation, is another glaring example of this type of 'personal experience'. Although prior to Newton's observation, several persons had witnessed the falling of the apple, he was the right man to formulate the theory of gravitation on the basis of this phenomenon.

Thus, emergence of a hypothesis is a creative manner. To quote Mc Guigan, "to formulate a useful and valuable hypothesis, a scientist needs first sufficient experience in that area, and second the quality of the genius." In the field of social sciences, an illustration of individual perspective may be visualized in Veblen's work. Thorstein Veblen's own community background was replete with negative experiences concerning the functioning of economy and he was a 'marginal man', capable of looking at the capitalist system objectively.

Thus, he could be able to attack the fundamental concepts and postulates of classical economics and in real terms Veblen could experience differently to bear upon the economic world, resulting in the making of a penetrating analysis of our society. Such an excellent contribution of Veblen has, no doubt, influenced social science since those days.

Hypotheses, Concepts and Measurement

Concepts in Hempel's deductive-nomological model play a key role in the development and testing of hypotheses. Most formal hypotheses connect concepts by specifying the expected relationships between propositions. When a set of hypotheses are grouped together they become a type of conceptual framework. When a conceptual framework is complex and incorporates causality or explanation it is generally referred to as a theory. According to noted philosopher of science Carl Gustav Hempel "An adequate empirical interpretation turns a theoretical system into a testable theory: The hypotheses whose constituent terms have been interpreted become capable of test by reference to observable phenomena. Frequently the interpreted hypothesis will be derivative hypotheses of the theory; but their confirmation or disconfirmation by empirical data will then immediately strengthen or weaken also the primitive hypotheses from which they were derived."

Hempel provides a useful metaphor that describes the relationship between a conceptual framework and the framework as it is observed and perhaps tested (interpreted framework). "The whole system floats, as it were, above the plane of observation and is anchored to it by rules of interpretation. These might be viewed as strings which are not part of the network but link certain points of the latter with specific places in the plane of observation. By virtue of those interpretative connections, the network can function as a scientific theory." Hypotheses with concepts anchored in the plane of observation are ready to be tested. In "actual scientific practice the process of framing a theoretical structure and of interpreting it are not always sharply separated, since the intended interpretation usually guides the construction of the theoretician." It is, however, "possible and indeed desirable, for the purposes of logical clarification, to separate the two steps conceptually.

Statistical Hypothesis Testing

Notes

When a possible correlation or similar relation between phenomena is investigated, such as whether a proposed remedy is effective in treating a disease, the hypothesis that a relation exists cannot be examined the same way one might examine a proposed new law of nature. In such an investigation, if the tested remedy shows no effect in a few cases, these do not necessarily falsify the hypothesis. Instead, statistical tests are used to determine how likely it is that the overall effect would be observed if the hypothesized relation does not exist. If that likelihood is sufficiently small (e.g., less than 1%), the existence of a relation may be assumed. Otherwise, any observed effect may be due to pure chance.

In statistical hypothesis testing, two hypotheses are compared. These are called the null hypothesis and the alternative hypothesis. The null hypothesis is the hypothesis that states that there is no relation between the phenomena whose relation is under investigation, or at least not of the form given by the alternative hypothesis. The alternative hypothesis, as the name suggests, is the alternative to the null hypothesis: it states that there is some kind of relation. The alternative hypothesis may take several forms, depending on the nature of the hypothesized relation; in particular, it can be two-sided (for example: there is some effect, in a yet unknown direction) or one-sided (the direction of the hypothesized relation, positive or negative, is fixed in advance).

Conventional significance levels for testing hypotheses (acceptable probabilities of wrongly rejecting a true null hypothesis) are .10, .05, and .01. The significance level for deciding whether the null hypothesis is rejected and the alternative hypothesis is accepted must be determined in advance, before the observations are collected or inspected. If these criteria are determined later, when the data to be tested are already known, the test is invalid.

The above procedure is actually dependent on the number of the participants (units or sample size) that are included in the study. For instance, to avoid having the sample size be too small to reject a null hypothesis, it is recommended that one specify a sufficient sample size from the beginning. It is advisable to define a small, medium and large effect size for each of a number of important statistical tests which are used to test the hypotheses.

1.6 THE RAW MATERIALS OF ECONOMETRICS

Data are the raw material from which econometric analysis is constructed. Just as a building is no stronger than the wood or steel used in its framework, an econometric study is only as reliable as the data used in its analysis. Many econometricians over the years have written about problems with data. One of the most comprehensive and comprehensible is a chapter that noted econometrician Zvi Griliches wrote for the third volume of Elsevier's Handbook of Econometrics back in 1986.

Obviously much has changed in the world of data and econometrics in the last 30 years, but many of the points that Griliches made are still relevant today, and some are even more important. This document uses extensive quotes from Griliches's chapter to highlight some important issues that every practitioner of econometrics should consider.

Economists sometimes collect their own data from experiments or surveys, but most econometric analysis relies on "found data," often from government sources.

Econometrics involve the formulation of mathematical models to represent real-world economic systems, whether the whole economy or an industry, or an individual business.

Econometric modeling is used to analyze complex market trends (the demand function) to determine the variables driving the growth or shrinkage of demand for a product or service. Econometric models are used to decipher the economic forces that affect supply and costs (the supply function) within an industry. Few companies really understand the external forces that drive their industries, their companies, or their brands. Understanding these forces provides the foundation for strategy development and business planning.

Times-series analysis, cross-sectional time-series analysis, structural-equation modeling, input-output analysis, Markov-chain analysis, and multiple regression are some of the techniques used in econometric modeling. Many other statistical and mathematical tools are employed as well, depending on the nature of the econometric task, in the development of econometric models.

Marketing mix modeling is one application of econometric modeling, wherein all marketing inputs are modeled over time to arrive at an optimal allocation of marketing inputs. For example, what is the correct amount to spend on television advertising compared to the radio or the Web? Should a company invest money in more salespeople or in more advertising? What is the impact of promotional spending?

Demand forecasting is another econometric application. For example, econometric analyses reveal that the growth in the number of women working in the U.S. played a major role in the growth of the restaurant industry from 1950 to 2000. But other variables were at work too. Rising incomes made eating out more affordable. Rising car ownership, especially among teenagers and college students, translated into greater restaurant sales. Understanding the variables that underlie demand makes it possible to forecast an industry's future.

1.7 TIME SERIES AND CROSS SECTION DATA: THE PROBLEM OF THEIR POOLING TOGETHER

Time Series

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the Dow Jones Industrial Average.

Time series are very frequently plotted via line charts. Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communications engineering, and largely in any domain of applied science and engineering which involves temporal measurements.

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis", which focuses on comparing values of a single time series or multiple dependent time series at different points in time. Interrupted time series analysis is the analysis of interventions on a single time series.

Time series data have a natural temporal ordering. This makes time series analysis distinct from cross-sectional studies, in which there is no natural ordering of the observations (e.g. explaining people's wages by reference to their respective education levels, where the individuals' data could be entered in any order). Time series analysis is also distinct from spatial data analysis where the observations typically relate to geographical locations (e.g. accounting for house prices by the location as well as the intrinsic characteristics of the houses). A stochastic model for a time series will generally reflect the fact that observations close together in time will be more closely related than observations further apart. In addition, time series models will often make use of the natural one-way ordering of time so that values for a given period will be expressed as deriving in some way from past values, rather than from future values.

Methods for time series analysis may be divided into two classes: frequency-domain methods and time-domain methods. The former include spectral analysis and wavelet analysis; the latter include auto-correlation and cross-correlation analysis. In the time domain, correlation and analysis can be made in a filter-like manner using scaled correlation, thereby mitigating the need to operate in the frequency domain.

Additionally, time series analysis techniques may be divided into parametric and non-parametric methods. The parametric approaches assume that the underlying stationary stochastic process has a certain structure which can be described using a small number of parameters (for example, using an autoregressive or moving average model). In these approaches, the task is to estimate the parameters of the model that describes the stochastic process. By contrast, non-parametric approaches explicitly estimate the covariance or the spectrum of the process without assuming that the process has any particular structure.

The observed values of the variable studied, such as the price of a commodity are results of various influences. Discovering and measuring the effects of these influences are the primary purposes of a time series analysis. Although the effects cannot always be determined exactly, been made over a sufficiently long period.

Time series analysis is done primarily for the purpose of making forecasts for future and also for the purpose of evaluating past performances. An economist or a business man is very naturally interested in estimating the future figure of national income, population, prices and wages etc. So the success or failure of a businessman depends to a large extent on the accuracy of this future forecasts.

Meaning of Time Series

Time series analysis are basic to understanding past behaviour, evaluating current accomplishments, planning for future operations and comparing different time series. Thus a series of successive observations of the same phenomenon over a period of time are called “time series”.

Definitions of Time Series

According to **Patterson,** “A time series consists of statistical data which are collected, recorded or observed over successive increments.”

“A set of data depending on the time is called time series”. – **Kenny and Keeping**

“A time series is a set of statistical observations arranged in chronological order”. – **Morris Hamburg**

According to **Croxton and Cowden,** “A time series consists of data arranged chronologically.”

"A time series is a set of observations taken at specified time, usually at 'equal intervals'. Mathematically, a time series is defined by the values y1, y2 ...... of a variable Y (temperature, closing price of share, etc.) at the time t1, t2 .... Thus, Y is a function of t, symbolised by y = F(t)." **– Spiegel**

"A time series may be defined as a collection of magnetite belonging to different time periods, of some variable or different time periods, of some variable or composite of variables, such as production of steel, per capita income, gross national product, price of tobacco or index of industrial production." **– Ya–Lun–Chou**

Uses of Time Series

The uses of time series is of well significance to the economist, scientist, sociologist, biologist, researcher and businessman etc. The following uses are:

(i) It helps to understand past behaviour.

(ii) It helps in evaluating current accomplishment.

(iii) It helps in planing future operations.

(iv) It also helps in comparing the actual performance.

(v) It helps to study the factor which influence the changes in economic activities and predict the future variations in them with certain limitations.

Component of Time Series Data

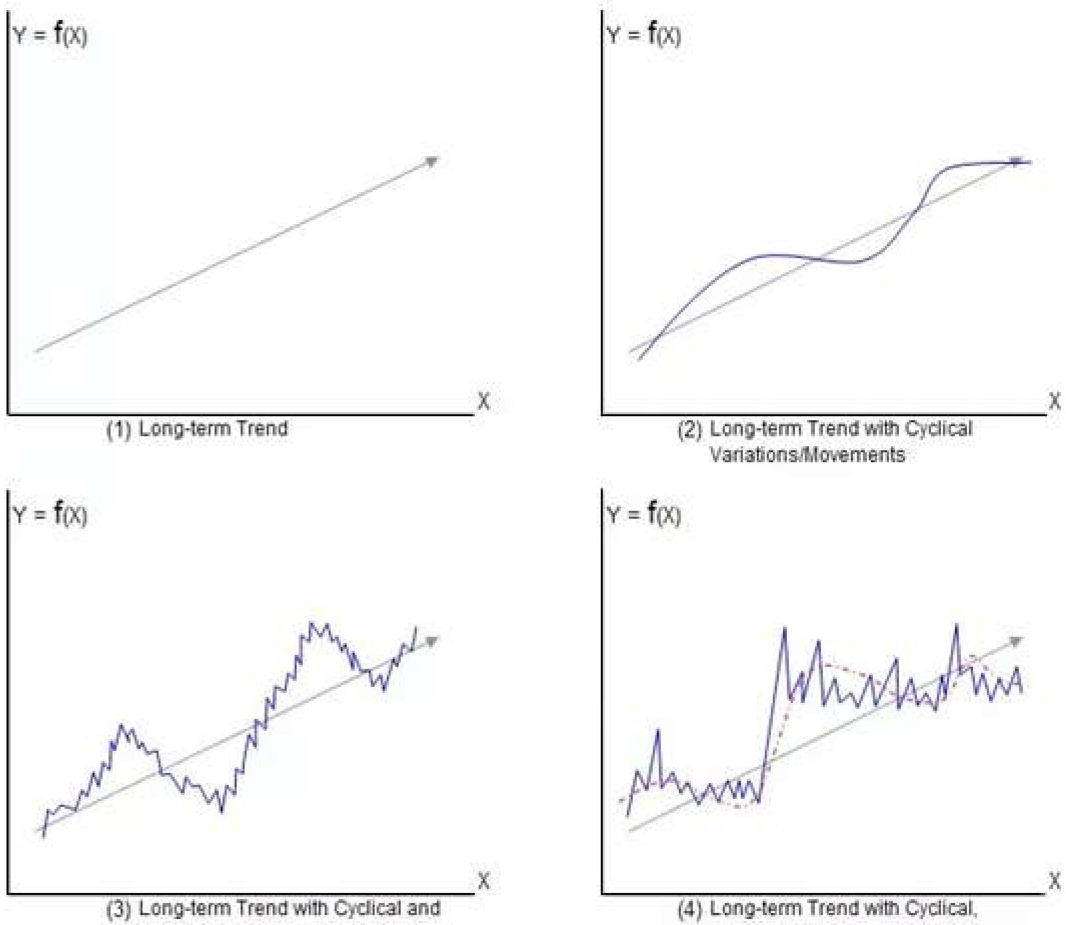
Traditional methods of time series analysis are concerned with decomposing of a series into a trend, a seasonal variation and other irregular fluctuations. The components, by which time series is composed of, are called component of time series data. There are four basic Components of time series data described below:

Seasonal effect (Seasonal Variation or Seasonal Fluctuations)

Many of the time series data exhibits a seasonal variation which is annual period, such as sales and temperature readings. This type of variation is easy to understand and can be easily measured or removed from the data to give de-seasonalized data. Seasonal Fluctuations describes any regular variation (fluctuation) with a period of less than one year for example cost of variation types of fruits and vegetables, cloths, unemployment figures, average daily rainfall, increase in sale of tea in winter, increase in sale of ice cream in summer etc., all show seasonal variations. The changes which repeat themselves within a fixed period, are also called seasonal variations, for example, traffic on roads in morning and evening hours, Sales at festivals like EID etc., increase in the number of passengers at weekend etc. Seasonal variations are caused by climate, social customs, religious activities etc.

Cyclical Variation or Cyclic Fluctuations

Time series exhibits Cyclical Variations at a fixed period due to some other physical cause, such as daily variation in temperature. Cyclical variation is a non-seasonal component which varies in recognizable cycle. Sometime series exhibits oscillation which does not have a fixed period but are predictable to some extent. For example, economic data affected by business cycles with a period varying between about 5 and 7 years. In weekly or monthly data, the cyclical component may describes any regular variation (fluctuations) in time series data. The cyclical variation are periodic in nature and repeat themselves like business cycle, which has four phases (i) Peak (ii) Recession (iii) Trough/Depression (iv) Expansion.



(1) Long-term Trend

(2) Long-term Trend with Cyclical Variations/Movements

(3) Long-term Trend with Cyclical and Seasonal Variations/Movements

(4) Long-term Trend with Cyclical, Seasonal, and Random Variations/Movements

Trend (Secular Trend or Long Term Variation)

It is a longer term change. Here we take into account the number of observations available and make a subjective assessment of what is long term. To understand the meaning of long term, let for example climate variables sometimes exhibit cyclic variation over a very long time period such as 50 years. If one just had 20 years data, this long term oscillation would appear to be a trend, but if several hundred years of data is available, then long term oscillations would be visible. These movements are systematic in nature where the movements are broad, steady, showing slow rise or fall in the same direction. The trend may be linear or non-linear (curvilinear). Some examples of secular trend are: Increase in prices, Increase in pollution, increase in the need of wheat, increase in literacy rate, decrease in deaths due to advances in science. Taking averages over a certain period is a simple way of detecting trend in seasonal data. Change in averages with time is evidence of a trend in the given series, though there are more formal tests for detecting trend in time series.

Irregular Fluctuations

When trend and cyclical variations are removed from a set of time series data, the residual left, which may or may not be random. Various techniques for analyzing series of this type examine to see "if irregular variation may be explained in terms of probability models such as moving average or autoregressive models, i.e. we can see if any cyclical variation is still left in the residuals. These variation occur due to sudden causes are called residual variation (irregular variation or accidental or erratic fluctuations) and are unpredictable, for example rise in prices of steel due to strike in the factory, accident due to failure of break, flood, earth quick, war etc.

Cross Section Data

Cross-sectional data or a cross section of a study population, in statistics and econometrics is a type of data collected by observing many subjects (such as individuals, firms, countries, or regions) at the same point of time, or without regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among the subjects.

For example, if we want to measure current obesity levels in a population, we could draw a sample of 1,000 people randomly from that population (also known as a cross section of that population), measure their weight and height, and calculate what percentage of that sample is categorized as obese. This cross-sectional sample provides us with a snapshot of that population, at that one point in time. Note that we do not know based on one cross-sectional sample if obesity is increasing or decreasing; we can only describe the current proportion.

Cross-sectional data differs from time series data, in which the same small-scale or aggregate entity is observed at various points in time. Another type of data, panel data (or longitudinal data), combines both cross-sectional and time series data ideas and looks at how the subjects (firms, individuals, etc.) change over time. Panel data differs from pooled cross-sectional data across time, because it deals with the observations on the same subjects in different times whereas the latter observes different subjects in different time periods. Panel analysis uses panel data to examine changes in variables over time and differences in variables between the subjects.

In a rolling cross-section, both the presence of an individual in the sample and the time at which the individual is included in the sample are determined randomly. For example, a political poll may decide to interview 1000 individuals. It first selects these individuals randomly from the entire population. It then assigns a random date to each individual. This is the random date that the individual will be interviewed, and thus included in the survey.

Cross-sectional data can be used in cross-sectional regression, which is regression analysis of cross-sectional data. *For example,* the consumption expenditures of various individuals in a fixed month could be regressed on their incomes, accumulated wealth levels, and their various demographic features to find out how differences in those features lead to differences in consumers' behavior.

Methods of Finding Trend

Estimation of trend values can be achieved in several methods:

(i) Graphic or Free hand Curve Method

(ii) Semi-Average Method

(iii) Moving Average Method

(iv) Least Square Method

(i) Graphic or Free hand Curve Method

Which consists of fitting a trend line or curve simply by looking at the graph, can be used to estimated trend.

The time is shown on the horizontal axis and the value of the variable on the vertical axis. The fitting of the trend may be straight line or a curved line and it may be done free hand by scale rules, spline, string or French curves of different shapes. Smooth out irregularities by drawing a free hand curve through the scatter points.

Merits of Graphic or Free hand Curve Method

(i) A free hand trend fitting enables an understanding of the character of time series.

(ii) These are its flexibility and simplicity.

(iii) This method only used to describe all types of trends – linear and non linear.

Demerits of Graphic or Free hand Curve Method

(i) This is depending too much on individual judgment.

(ii) It does not involved any complex mathematical techniques.

(iii) It does not enable us to measure trend in precise quantitative terms.

(iv) Different trend curves could be obtained by different persons for the same data. It is highly subjective.

(ii) Semi-Average Method

The method is used only when the trend is linear or almost linear. For non-linear trends this method is not applicable. It is used for the calculation of averages, and averages are affected by extreme values. Thus if there is some very large value or very small value in the time series, that extreme value should either be omitted or this method should not be applied.

Merits of Semi-Average Method

(i) It is simple and easy to understand.

(ii) It can compared with the moving average of the least squares method of measuring trend.

(iii) This method is objectivity in the sense that it does not depend of the personal judgment.

(iv) It is also applicable where the trend is linear or approximately linear.

Demerits of Semi-Average Method

(i) This method is based on the assumption that there is a linear trend which may not be true.

(ii) It is not suitable when time period represented by average is small.

(iii) The use of arithmetic mean for obtaining semi–average may be questioned because of limitation of the method.

Illustration - 1

Using the method of semi–average determine the trend of the following data:

Year :	1998	1999	2000	2001	2002	2003	2004	2005
Production:	48	42	60	54	38	66	70	66

Solution:

The number of observations are even i.e., 8. The two middle parts will be 1997 to 2001 and 2002 to 2006.

Notes

Year	Actual value	4 years total	semi–average and average value
1898	48		
1999	42	$\frac{48+42+60+54}{4}=\frac{204}{4}$	51 .... $\bar{x}_1$
2000	60		
2001	54		
2002	38	$\frac{38+66+70+66}{4}=\frac{240}{4}$	61 ...... $\bar{x}_2$
2003	66		
2004	70		
2005	66		

Here, the value 51 is plotted against the middle of the first four years i.e., 1998–2001 and the value 60 is plotted against the middle of the last four years i.e., 2002–2005. So both the point are joined by a straight line as under.

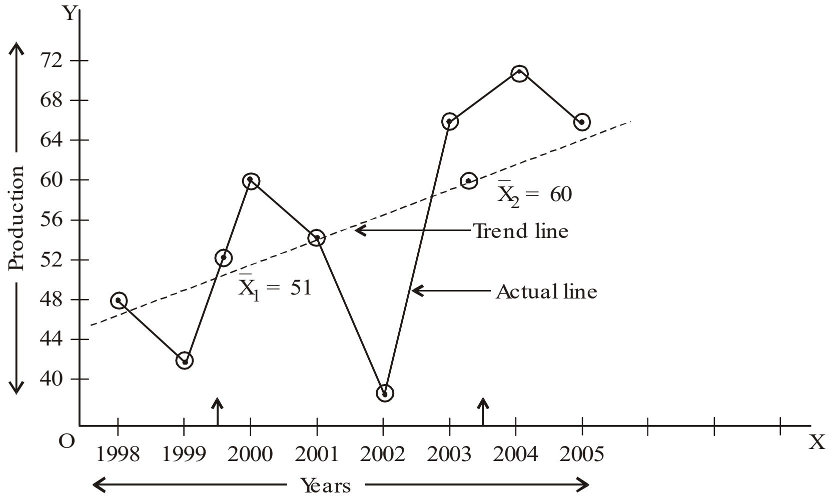


Illustration - 2

Calculate trend values from the following data by the method of semi–average:

Year:	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Sales:	30	35	37	24	42	36	27	45	40	42

Solution:

Year	Sales	Averages	Annual changes	Trend value
1996	30			29.2
1997	35			30
1998	39	$\frac{170}{5}=34$		30.8
1999	24			31.6
2000	42		0.8	32.4

2001	36			33.2
2002	27	....	34	
2003	45		34.8	
2004	40		35.6	
2005	42		36.4	

Let us have two periods of 5 years each 1996 to 2000 and 2001 to 2005. The averages for the two periods are:

$$\frac{30+35+39+24+42}{5}=\frac{170}{5}=34..........\bar{x}_1$$

and $$\frac{36+27+45+40+42}{5}=\frac{190}{5}=38..........\bar{x}_2$$

The increase of two averages is 38 – 34 = 4 which takes place in 5 years, therefore, annual change is .... = 0.8 (4/5)

"Graph showing Sales with Trend line"

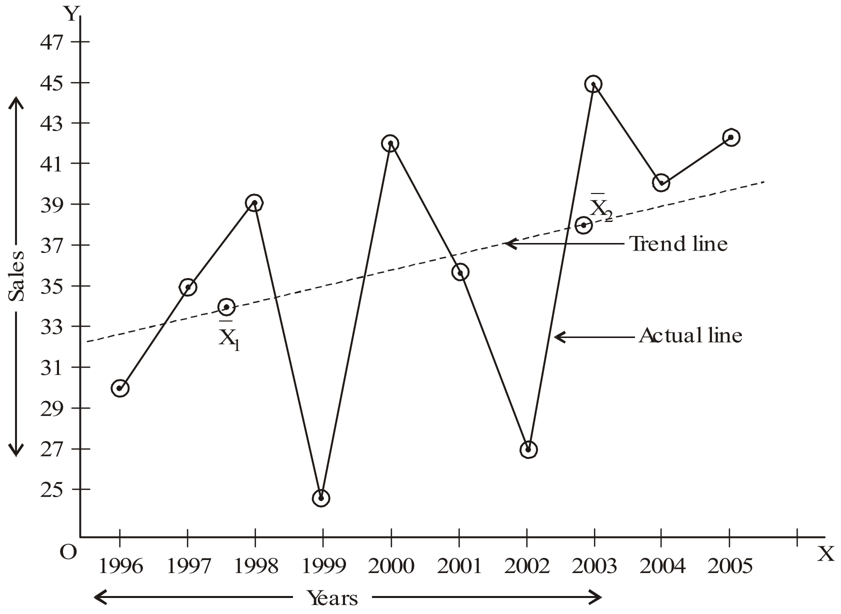


Illustration - 3

Draw a trend by the method of semi average from the following data:

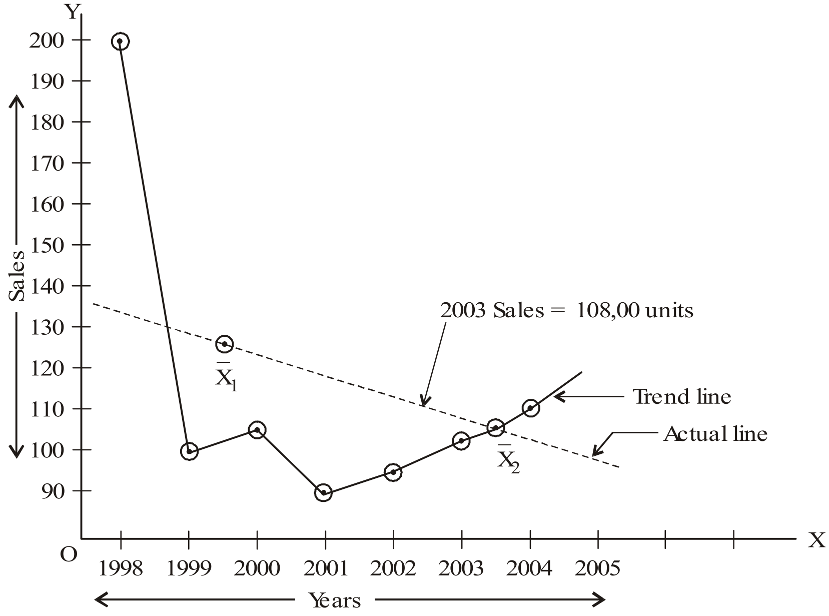
Year:	1998	1999	2000	2001	2002	2003	2004	2005
Sales ('000 units):	195	100	104	90	95	102	110	116

Also predict the sales for the year 2003 from the graph.

Solution:

Year	*Sales(x)*	*Semi Average*
1998	195	
1999	100	$\frac{489}{4}$ = 122.25 .... $\bar{x}_1$
2000	104	
2001	90	
2002	95	
2003	102	$\frac{423}{4}$ .... = 1.5.75 ..... $\bar{x}_2$
2004	110	
2005	116	

Graph showing sales "000"units



(iii) Moving Average Method

By using moving averages of appropriate orders, cyclical, seasonal and irregular patterns may be eliminated thus leaving only the trend movement. Here, trend values can be obtained by employing arithmetic means of the series except at the two ends of the series. So moving average consists of a series of arithmetic means calculated from overlapping groups of successive values of a time series. Moving average based on values covering a fixed time interval, called moving average period. It is shown against the centre of the period.

The moving average for period 't' is a series of successive averages of values at a time, starting with 1st, 2nd and 3rd to 't' terms. Here, the first average is the mean of the 1st to 't' terms, the second is the mean of the 't' terms from 2nd to $(t + 1)^{th}$ terms and third is the mean of the 3rd to $(t + 2)^{th}$ terms and so on.

Hence, the time series values X_1, X_2, X_3 .....

for different time periods, the moving average of period 't' is given by:

1st value moving average = $\frac{1}{t}$ $(X_1 + X_2 + ... X_t)$,

2nd value moving average = $\frac{1}{t}$ $(X_2 + X_3 + ... X_{t+1})$ and

3rd value moving average = $\frac{1}{t}(X_3 + X_4 + ... X_{t+2})$

(a) ***Odd period:*** When the period is odd, if the period '*t*' of the moving average is odd the successive value of the moving averages are placed against the middle value of concerned group of items. If $t = 5$, the first moving average value is placed against the middle period i.e. third value and the second moving average value is placed against the time period four and so on.

(b) ***Even period:*** When the period is even, if the period '*t*' of moving average is even there are two middle periods and the moving average value is placed between the two middle terms of the time intervals. In the particular period $t = 4$, the first moving average is placed against the middle of second and third values, the second moving average is placed in between third and fourth values and so on.

For odd: In the below table (a) have considered 5 years moving averages.

Year	*Data*	*5 year moving total*	*5 year moving average*
1995	50.0		
1996	36.5		
1997	43.0	212.9	42.6
1998	44.5	201.0	40.2
1999	38.9	197.1	39.4
2000	38.1	192.8	39.6
2001	32.6	190.0	38.0
2002	38.7	192.2	38.4
2003	41.7	187.9	37.6
2004	41.1		
2005	33.8		

Here, the first moving total 219.9 of column 3 is the sum of the 1st through 5th entries of column 2. The second moving total 201.0 is the sum of the 2nd through 6th entries in column 2 etc.

In practice, after obtaining the first moving total 212.9, the second moving total is easily obtained by subtracting 50.0 (1st entry of column 2), the result being 201.0. Succeeding moving totals are obtained similarly. Dividing each moving total by 5 yields the required moving average.

In the below table ... b have considered 4 years moving averages.

Notes

For even:

Year	*Data*	*4 years moving total*	*4 years moving total*
1995	50.0		
1996	36.5		
1997	43.0	174.0	43.5
1998	44.5	162.9	40.7
1999	38.9	164.5	41.1
2000	38.1	154.1	38.5
2001	32.6	148.3	37.1
2002	38.7	151.1	37.8
2003	41.7	154.1	38.5
2004	41.1	155.3	38.8
2005	33.8		

Here, the 4 year moving totals are obtained as in past (a), except that 4 entries of column 2 are added instead of 5. Note that the moving totals are centred between successive years, unlike part (a). This is always the case when an even number of years. is taken in the moving average. If we consider that 1996, for example, stands for July 1, 1996, the first 4 year moving total is centred at Jan 1, 1997 or Dec. 31, 1996. The 4 year moving averages are obtained by dividing the 4 year moving totals by 4.

Merits of Moving Average Methods

(i) It is very simple to understand and easy to calculate as compared to other methods.

(ii) It can be used in all facts of time series analysis, for the measurement of trends as well as in connection with seasonal cyclical irregular components.

(iii) It calculate is simple because no higher degree mathematical calculations.

(iv) It considers all the values in the series. The extreme values are included in the process of determining averages.

(v) It is a objective method. No personal judgment like freehand method.

Demerits of Moving Average Methods

(i) In this method, that data at the beginning and end of a series are lost.

(ii) It may generate cycles or other movements which were not present in the original data.

(iii) Its are strongly affected by extreme values of items. They are said to be sensitive to.... movement in the data.

(iv) It does not establish functional relationship between the period and the value of variables.

(v) It cannot determine irregular variations completely.

Illustration - 4

Find trend values by three yearly moving average method.

Year:	1999	2000	2001	2002	2003	2004	2005
Production: ('000' in unit)	112	138	146	154	170	183	190

Solution:

I *Year*	*II* *Production*	*III* *Column of differences*	*IV* *3 Year moving total*	*V* *3 Year moving average*
1999	112			
2000	138	154 – 112 = 42	396	$\frac{396}{3} = 132$
2001	146	170 – 138 = 32	 438	$\frac{438}{3} = 146$
2002	154	183 – 146 = 37	 470	$\frac{470}{3} = 157$
2003	170	190 – 154 = 36	 507	$\frac{507}{3} = 169$
2004	183		 543	$\frac{543}{3} = 181$
2005	190			

Calculating total production of first three years which is (112 + 138 + 146 = 396). This total is placed in column (IV) of three yearly moving total before the middle year. 2000 and in column (V) the three year moving average 396/3 = 132 is placed before the year 2000 as shown in the above table. And find the second moving total we find the difference of the production of a year 2002 and 1999 which is (154 – 112 = 42). This is written in column (III) in front of the year 2000 and the second moving total is (396 + 42 = 438). Such way it can complete columns (III) and (IV) of the above table.

Illustration - 5

Find trend values from the following data using three yearly moving averages and show the trend line on the graph.

Year	*Price (₹)*	*Year*	*Price (₹)*
1994	52	2000	75
1995	65	2001	70
1996	58	2002	64
1997	63	2003	78
1998	66	2004	80
1999	72	2005	73

Solution:

Computation of Trend Values

Year	Price (₹)	3 yearly moving total	3 yearly moving average
1994	52	–	
1995	65	175	58.33
1996	58	186	62.00
1997	63	187	62.33
1998	66	201	67.00
1999	72	213	71.00
2000	75	217	72.33
2001	70	209	69.67
2002	64	212	70.67
2003	78	222	74.00
2004	80	231	77.00
2005	73	–	

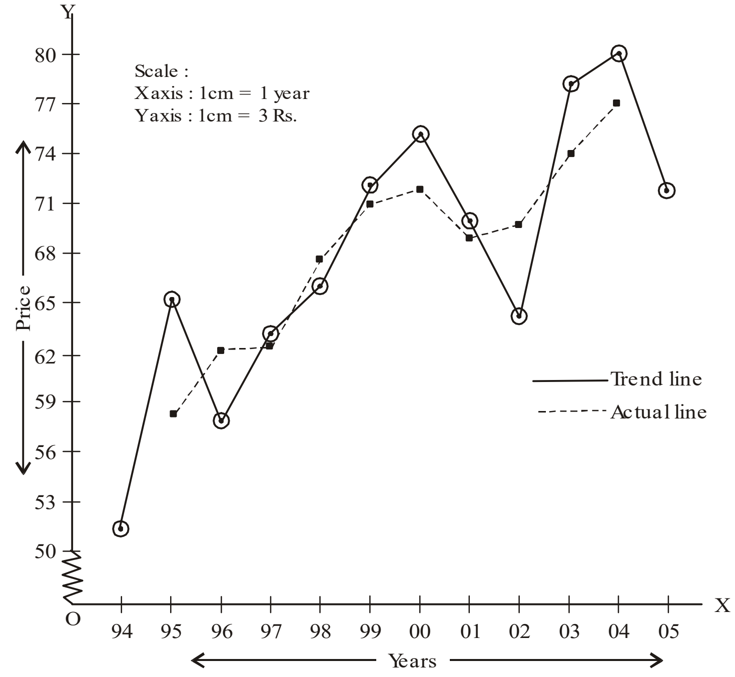


Illustration - 6

Using a 3 yearly moving averages determine the trend values.

Year:	1995	1996	1997	1998	1999	2000	2001	2002
Production:	21	22	23	24	25	26	27	26
(in ‘000 units)								

Solution:

Years	Production (in 000 units)	3 yearly Moving total	3 yearly Moving average
1995	21	-	-
1996	22	66	22
1997	23	69	23
1998	24	72	24
1999	25	75	25
2000	26	78	26
2001	27	79	26.33
2002	26	-	-

Graphical representation

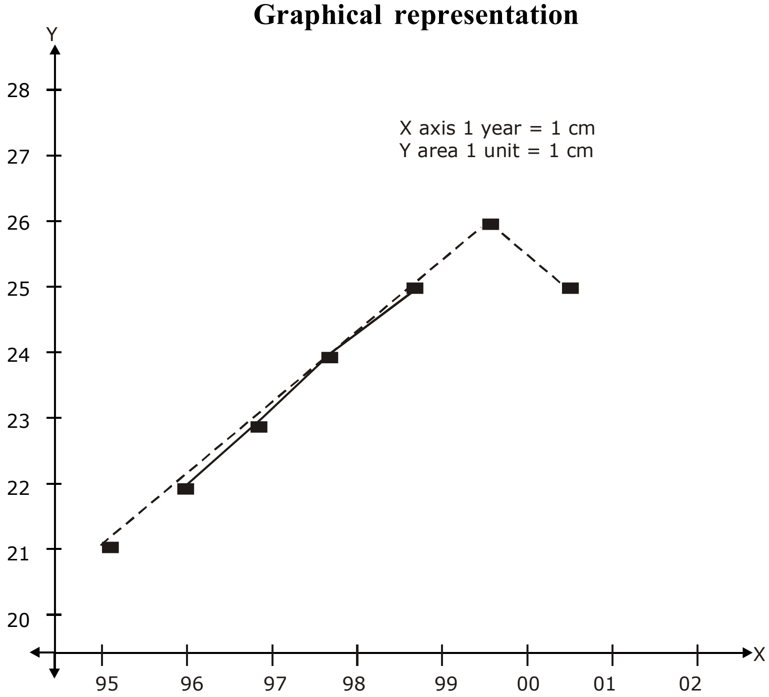


Illustration - 7

Find the four yearly moving averages for the following data:

Year:	1991	1992	1993	1994	1995	1996	1997	1998
Values:	30.1	45.4	39.3	41.4	42.2	46.4	46.6	49.2

Solution:

Year	*Values*	*4 yearly moving Total*	*4 yearly moving Average*	*Central value*
1991	30.1			
1992	45.4			
		156.2	39.05	

1993	39.3			40.56
		168.3	42.07	
1994	41.4			42.195
		169.3	42.32	
1995	42.2			43.235
		176.6	44.15	
1996	46.4			45.125
		184.4	46.1	
1997	46.6			
1998	49.2			

Illustration - 8

Using four yearly moving averages determine the trend values and also plot the original and trend values on a graph.

Year	*Production (1000 units)*	*Year*	*Production (1000 units)*
1994	75	2000	96
1985	62	2001	128
1996	76	2002	116
1997	78	2003	76
1998	94	2004	102
1999	84	2005	168

Solution:

Year	*Production*	*4yearly*	*4yearly*	*Centred*	*Year*
	Moving Average	*Moving Total*	*Moving Average*		
1994	75				
1995	62	291	72.75	75.125	1996
1996	76	310	77.50	80.25	1997
1997	78	332	83.00	85.50	1998
1998	94	352	88.00	94.25	1999
1999	84	402	100.50	103.25	2000
2000	96	424	106.00	105.00	2001
2001	128	461	104.00	104.75	2002
2002	116	422	105.50	110.50	2003
2003	76	462	115.50		
2004	102				
2005	168				

Graph showing production and trend value for the years 1994-2005

Scale: OX= 1 cm = 1year, OY = 1 cm = 1000 units.

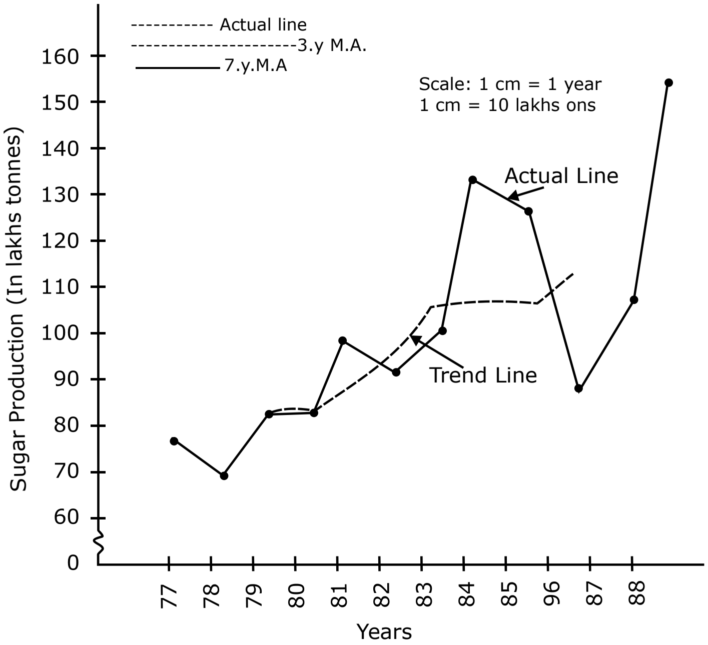


Illustration - 9

Calculate the trend values by five yearly moving average method and plot the same on a graph from the following:

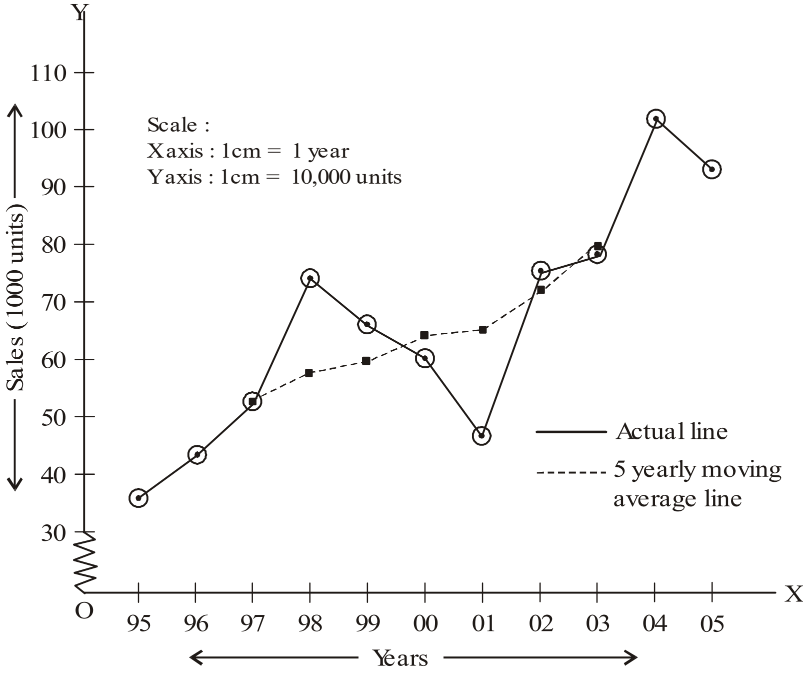
Year:	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Sales: ('000'units)	36	42	54	72	66	60	48	75	78	102	93

Solution:

Computation of Trend Values

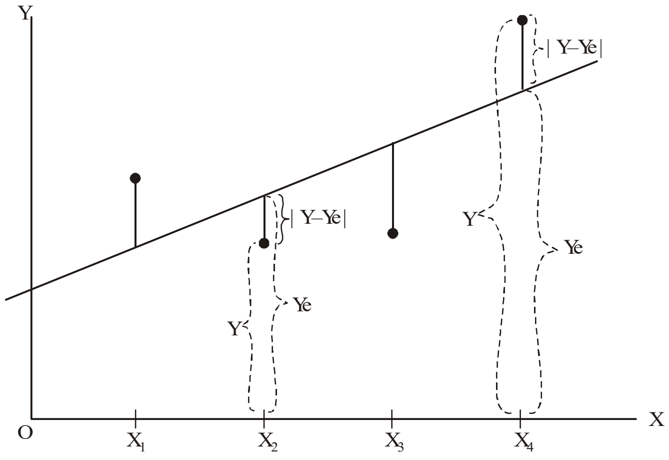
Year	Sales moving total	5 years moving total	5 years
1995	36	–	
1996	42	–	
1997	54	270	54
1998	72	294	58.8
1999	66	300	60
2000	60	321	64.2
2001	48	327	65.4
2002	75	363	72.5
2003	78	396	79.2
2004	102	–	
2005	93	–	

Graph showing sales of units for 11 years and 5 yearly trend lines



(iv) Least Square Method

A definition of the term, line of best fit, should give a unique line. Let us proceed by considering, for each value of X, the absolute value $|Y - Y_c|$ of the difference between the actual a value and the estimated Y value (see below figure). This difference represents the error committed by using the estimated Y value instead of the actual value. One way to determine the line of best fit might be to find that line for which the sum of these errors for all the given values of X, i.e., $\sum |Y - Y_c|$, has the smallest possible value. However, the procedure, while yielding a unique line, leads to mathematical difficulties customarily associated with the occurrence of absolute values. A better method and one which accomplishes the same aims, defines the line in such a way that $\sum (Y - Y_c)^2$ has the smallest value. This method, called the method of least squares, is the one most generally used in statistics for obtaining the line of best fit.



The trend line, which is a "best fit" to a scatter diagram of n points, its employ a theorem from elementary mathematics.

From the below figure, the equation of any no–vertical line can be written in the form

$Y = a + bx,$

where, α is the Y–intercept

and b is the slop of the line.

The slope is given $b = \tan \alpha$, where a is the angle measured from the positive X – axis to the line and is positive or negative as α is obtuse.

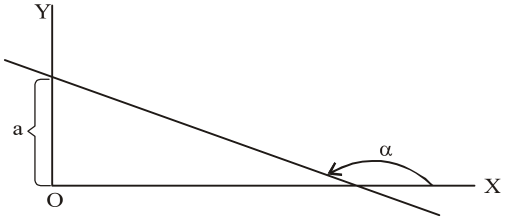


Fig: *Graph showing of straight line Y–intercept a and angle of inclination a.*

To obtain the values of '*a*' and '*b*' constants. Instead to following two secondary equation:

$na + b\Sigma x = \Sigma Y$ ....(i)

$a\Sigma x + b\Sigma\, x^2 = \Sigma XY$ ....(ii)

Here, x represents the number of years or any period for which the data is given.

Mid point in time is taken as the origin, so that negative values in the first half of the series balance out the positive values in the second half, i.e., Sx = 0.

The period indicating the higher values can be reduced to the minimum symbolically as under:

Year (Y)	*Deviation (x)*	*Year (y)*	*Deviation(x)*
1999	–3	1998	–7
2000	–2	1999	–5
2001	–1	2000	–3
2002	0	2001	–1
2003	+1	2002	+1
2004	+2	2003	+3
2005	+3	2004	+5
		2005	+7

So that Σx = 0 as the Deviation are calculated from the mean. The sum of deviations of the actual value from the computed values is equal to zero.

Here, Y = Actual values

$Y_c = na$ and $\Delta xy = b\Sigma x^2$

Notes

So the value of 'a' and 'b' determined as -

$$a = \frac{\Sigma y}{n} \; ; \qquad b = \frac{\Sigma xy}{\Sigma x^2}$$

Finally the trend line, the line of 'best fit' will be drawn under the least squares method. It is the line from which the sum of the squares of the items, measured parallel to the Y axis, is the least.

So equation represented by

$$Y = a + bx$$

Least square method for Trend values

$$Y_c = a + bx$$

$$a = \frac{\Sigma y}{n} \; ; \qquad b = \frac{\Sigma xy}{\Sigma x^2}$$

Here, x = independent variable (Deviation)

y = Dependent variable on x.

a = The 'y' intercept.

b = Indicate slope and signifies the changes in 'x' ...constant.

n = Number of observations.

xy = Product of x and y.

x^2 = Square of x.

Σy = Sum of 'y' values

Δxy = Sum of the product of x and y values.

Σx^2 = Sum of square value of x.

Merits of Least Squares Methods

(i) It is objective method which does not variations in the results.

(ii) It is very easy calculation because no higher degree calculation.

(iii) It determines the trend values and also reflects light on the seasonal, cyclical and irregular variations.

(iv) Since it is based on an algebraic method of calculating the trend, it is free from any bias, subjectivity does not enter into it.

(v) It is a flexible method, the trend values calculated any period and answer should be accurate.

Demerits of Least Squares Methods

(i) This method is in appropriate for a very short series and is unnecessary for along one.

(ii) It does not establish functional relationship between the period (x) and the values of variable 'y'.

(iii) It is a tedious method and involves more calculations compared to the methods discussed earlier.

(iv) It can estimate a value only immediate future and not for distant future.

Illustration - 1

Fit a straight line trend by the method of least squares, tabulate the trend values and show the values on a graph from the following figure of production of sugar factory:

Year:	1999	2000	2001	2002	2003	2004	2005
Production: ('000' tons)	80	90	92	83	94	99	92

Solution:

Computation of trend values

Year	*Production*	*x*	x^2	*xy*	*Trend values* (Y_c)
1999	80	–3	9	–240	90 + 2 (–3) = 84
2000	90	–2	4	–180	90 + 2 (–2) = 86
2001	92	–1	1	–92	90 + 2 (–1) = 88
2002	83	0	0	0	90 + 2 (0) = 90
2003	94	+1	1	94	90 + 2 (+1) = 92
2004	99	+2	4	198	90 + 2 (+2) = 94
2005	92	+3	9	276	90 + 2 (+3) = 96
	$\Sigma y = 630$	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma xy = 56$	$\Sigma xy = 56$

The straight line equation, $Y_c = a + bx$

$$\text{Here,} \quad a = \frac{\Sigma y}{n} = \frac{630}{7} = 90 \text{ and } b = \frac{\Sigma xy}{\Sigma x} = \frac{56}{28} = 2$$

Graph showing production for the year 2007 to 2013

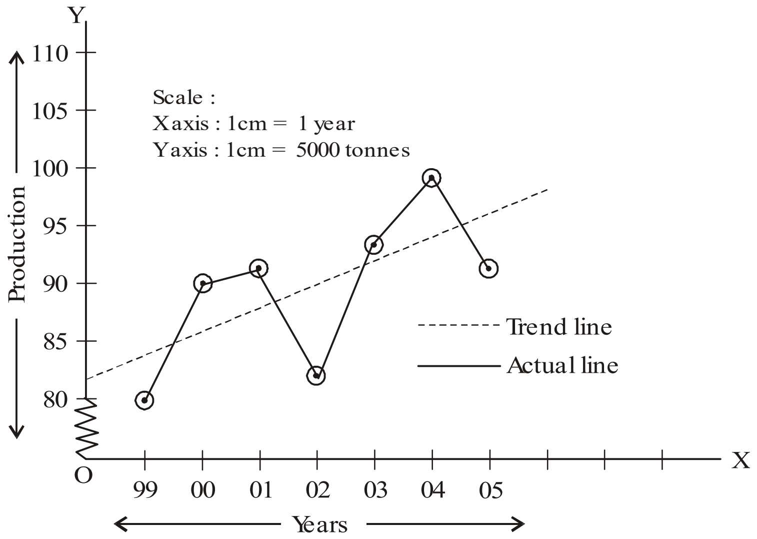


Illustration - 2

Fit a straight line trend to the following data taking *X* as the independent variable and prove S($Y - Y_c$) = 0.

X: 2000 2001 2002 2003 2004 2005

Y:	1	1.8	3.3	4.5	6.3	10

Solution:

Calculation of Trend by the method of least squares

Year	*y*	*x*	x^2	*xy*	*Trend value* $Y_c = a + bx$
2000	1	–5	25	–5	0.223
2001	1.8	–3	9	–5.4	1.927
2002	3.3	–1	1	–3.3	3.631
2003	4.5	+1	1	4.5	5.335
2004	6.3	+3	9	18.9	7.037
2005	10	+5	25	50	8.743
	$\Sigma y = 26.9$	$\Sigma x = 0$	$\Sigma x^2 = 70$	$\Sigma xy = 59.7$	

We know, trend value $Y_c = a + bx$

$$a = \frac{\Sigma y}{n} = \frac{26.9}{6} = 4.483 \qquad b = \frac{\Sigma xy}{\Sigma x^2} = \frac{59.7}{70} = 0.852$$

For the year $Y_c = a + bx$

2000 = 4.483 + 0.852 (–5) = 0.223

2001 = 4.483 + 0.852 (–3) = 1.927

2002 = 4.483 + 0.852 (–1) = 3.631

2003 = 4.483 + 0.852 (+1) = 5.335

2004 = 4.483 + 0.852 (+3) = 7.037

2005 = 4.483 + 0.852 (+5) = 8.743

Illustration - 3

The following figures represent the Annual sales of M/s. Suman Roy Company.

a) Find the trend values for each year, by adopting the method of least squares.

b) Estimate the Annual sales for the next five years.

Year:	1996	1997	1998	1999	2000
Sales in lakhs:	46	50	40	70	60

Solution:

Fitting a straight line trend by method of least square

Year	*Sales* *y*	*Deviation* *x*	*Square of deviation*(x^2)	*Derivation & sales (xy)*	*Trend Value Yc*
1996	46	–2	4	–92	43.6
1997	50	–1	1	–50	48.4
1998	40	0	0	0	53.2
1999	70	1	1	70	58
2000	60	2	4	120	62.8
$n = 5$	$\Sigma y = 266$	$\Sigma x = 0$	$\Sigma x^2 = 10$	$\Sigma xy = 48$	

$Y_c = a + bx$ $a = \frac{\Sigma y}{n}$ and $b = \frac{\Sigma xy}{\Sigma x^2}$

$a = \frac{266}{5}$; $b = \frac{48}{10}$; a = 53.2 and b = 4.8

For the year $Y_c = a + bx$

1996 = 53.2 + 4.8 (–2) = 43.6

1997 = 53.2 + 4.8 (–1) = 48.4

1998 = 53.2 + 4.8 (0) = 53.2

1999 = 53.2 + 4.8 (1) = 58

2000 = 53.2 + 4.8 (2) = 62.8

2001 = 53.2 + 4.8 (3) = 67.6

2002 = 53.2 + 4.8 (4) = 72.4

2003 = 53.2 + 4.8 (5) = 77.2

2004 = 53.2 + 4.8 (6) = 82

2005 = 53.2 + 4.8 (7) = 86.8

Illustration - 4

Fit a straight line by method of least squares from the data given below:

Find the trend values and predict the sales for the year 2000 also show the trend values on a graph.

Year:	1990	1991	1992	1993	1994	1995	1996	1997
Values:	15	18	20	30	39	40	44	50

Solution:

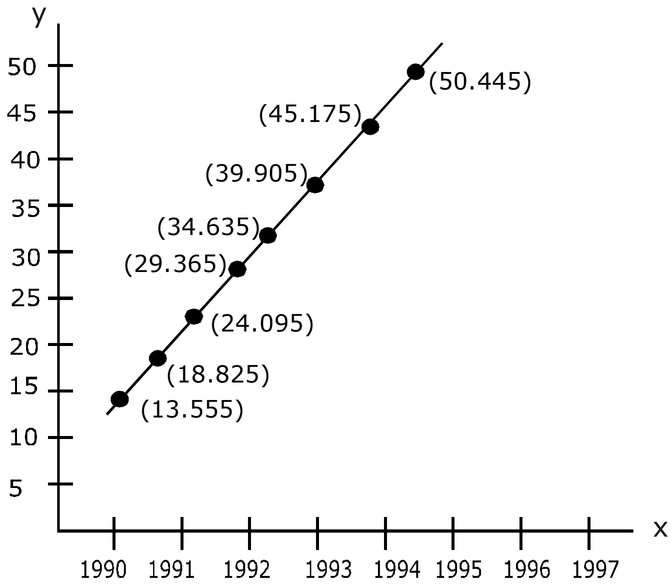
Year	*Sales (y)*	*x*	x^2	*xy*	*Trend value* *Yc = a + bx*
1990	15	-3.5	12.25	-52.5	13.555
1991	18	-2.5	6.25	-.45.0	18.825
1992	20	-1.5	2.25	-30.0	24.095
1993	30	-0.5	0.25	-15.0	29.365
1994	39	+0.5	0.25	19	34.645
1995	40	+1.5	2.25	60	39.905
1996	44	+2.5	6.25	110	45.175
1997	50	+3.5	12.25	175	50.445
$n = 8$	$\Sigma y = 256$	$\Sigma x = 0$	$\Sigma x^2 = 42$	$\Sigma xy = 221.5$	

$Yc = a + bx$

$$a = \frac{\Sigma y}{N} = \frac{256}{8} = 32 \text{ and } b = \frac{\Sigma xy}{\Sigma x^2} = \frac{221.5}{42} = 5.27$$

Notes

Sales for the year 2000 = 32 + 5.27(6.5) = 66.255



The Problem of their Pooling together

Pooled data occur when we have a "time series of cross sections," but the observations in each cross section do not necessarily refer to the same unit. Panel data refers to samples of the same cross-sectional units observed at multiple points in time.

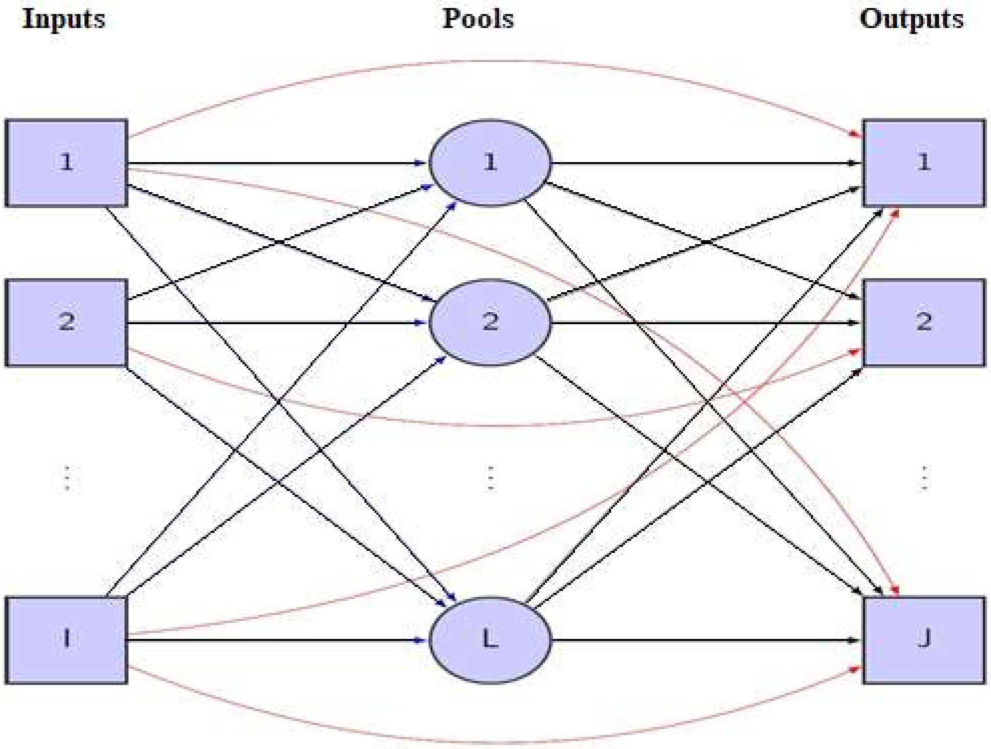
The pooling problem, which is a sub-problem of wastewater networks, crude oil refinery planning etc., is important because of the huge amount of money that can be saved by solving it to optimality. In a pooling problem, flow streams from different sources are mixed in intermediate tanks (pools) and blended again in the terminal points. At the pools and terminals, the quality of a mixture is given as the volume (weight) average of the qualities of the flow streams that go into them.

In the pooling problem, there are three types of tanks: inputs or sources, which are the tanks to store the raw materials, pools, which are the places to blend incoming flow streams and make new compositions and outputs or terminals, which are the tanks to store the final products. According to the links among different tanks, pooling problems can be classified into three classes:

a) ***Standard pooling problem:*** In this class there is no flow stream among the pools. It means that the flow streams are in the form of input-output, input-pool and pool-output; see Figure 1.

b) ***Generalized pooling problem:*** In this class, the complexity of the pooling problem is increased by allowing flow streams between the pools.

c) ***Extended pooling problem:*** In this class, the problem is to maximize the profit (minimize the cost) on a standard pooling problem network while complying with constraints on nonlinearly blending fuel qualities such as those in the Environmental Protection Agency.



There are many equivalent mathematical formulations for a pooling problem, such as P-, Q-, PQ- and HYB- formulations, and all of them are formulated as a non-convex (bilinear) problem, and consequently the problem can possibly have many local optima. These formulations vary in the way of representing specifications in the pools. For instance, in the Q-formulation the fraction of incoming flow to a pool that is contributed by an input is considered.

Role of Hypothesis in Social Research

In any scientific investigation, the role of hypothesis is indispensable as it always guides and gives direction to scientific research. Research remains unfocused without a hypothesis. Without it, the scientist is not in position to decide as to what to observe and how to observe. He may at best beat around the bush. In the words of Northrop, “The function of hypothesis is to direct our search for order among facts, the suggestions formulated in any hypothesis may be solution to the problem, whether they are, is the task of the enquiry”.

Several near consequences are provided in the process of deductive development of hypothesis. In the process of conducting experiments for confirming the hypothesis, scores of new facts develop and expand the horizon of knowledge of the scientist. Since h3rpothesis is concerned with explaining facts, the rejection of hypothesis is not futile.

Rather, it is worthwhile in the sense that it can be of great service in pointing out the way to true hypothesis. Even a false hypothesis is capable of showing the direction of inquiry. Realizing the indispensability of hypothesis in a scientific investigation, Cohen and Nagel observe, ‘Hypotheses are required at every stage of an inquiry. It must not be forgotten that what are called general principles or laws can be applied to a present, still un-terminated inquiry only with some risk. For they may not in fact be applicable.

The general laws of any science function as hypothesis, which guide the inquiry in all its phases”. Thus, there is little doubt that the importance of hypothesis in the field of scientific research is tremendous. At least five reasons may be advanced for justifying hypothesis as a significant device for scientific research.

First, it is an operating tool of theory. It can be deduced from other hypotheses and theories. If it is correctly drawn and scientifically formulated, it enables the researcher to proceed on correct line of study. Due to this progress, the investigator becomes capable of drawing proper conclusions.

In the words of Goode and Hatt, “without hypothesis the research is unfocussed, a random empirical wandering. The results cannot be studied as facts with clear meaning. Hypothesis is a necessary link between theory and investigation which leads to discovery and addition to knowledge.

Secondly, the hypothesis acts as a pointer to enquiry. Scientific research has to proceed in certain definite lines and through hypothesis the researcher becomes capable of knowing specifically what he has to find out by determining the direction provided by the hypothesis. Hypotheses acts like a pole star or a compass to a sailor with the help of which he is able to head in the proper direction.

Thirdly, the hypothesis enables us to select relevant and pertinent facts and makes our task easier. Once, the direction and points are identified, the researcher is in a position to eliminate the irrelevant facts and concentrate only on the relevant facts. Highlighting the role of hypothesis in providing pertinent facts, P.V. Young has stated, “The use of hypothesis prevents a blind research and indiscriminate gathering of masses of data which may later prove irrelevant to the problem under study”.

For example, if the researcher is interested in examining the relationship between broken home and juvenile delinquency, he can easily proceed in the proper direction and collect pertinent information succeeded only when he has succeed in formulating a useful hypothesis.

Fourthly, the hypothesis provides guidance by way of providing the direction, pointing to enquiry, enabling to select pertinent facts and helping to draw specific conclusions. It saves the researcher from the botheration of ‘trial and error’ which causes loss of money, energy and time.

Finally, the hypothesis plays a significant role in facilitating advancement of knowledge beyond one’s value and opinions. In real terms, the science is incomplete without hypotheses.

1.8 ELEMENTS OF STATISTICAL INFERENCES

Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution. Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population.

Any statistical inference requires some assumptions. A statistical model is a set of assumptions concerning the generation of the observed data and similar data. Descriptions of statistical models usually emphasize the role of population quantities of interest, about which we wish to draw inference. Descriptive statistics are typically used as a preliminary step before more formal inferences are drawn.

Degree of Models/Assumptions

Statisticians distinguish between three levels of modeling assumptions:

1. ***Fully parametric:*** The probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters. For example, one may assume that the distribution of population values is truly Normal, with unknown mean and variance, and that datasets are generated by 'simple' random sampling. The family of generalized linear models is a widely used and flexible class of parametric models.
2. ***Non-parametric:*** The assumptions made about the process generating the data are much less than in parametric statistics and may be minimal. For example, every continuous probability distribution has a median, which may be estimated using the sample median or the Hodges–Lehmann–Sen estimator, which has good properties when the data arise from simple random sampling.
3. ***Semi-parametric:*** This term typically implies assumptions 'in between' fully and non-parametric approaches. For example, one may assume that a population distribution has a finite mean. Furthermore, one may assume that the mean response level in the population depends in a truly linear manner on some covariate (a parametric assumption) but not make any parametric assumption describing the variance around that mean (i.e. about the presence or possible form of any heteroscedasticity). More generally, semi-parametric models can often be separated into 'structural' and 'random variation' components. One component is treated parametrically and the other non-parametrically. The well-known Cox model is a set of semi-parametric assumptions.

1.9 POINT AND INTERVAL ESTIMATION-ESTIMATOR AND ITS PROPERTIES

Estimation is the process of making inferences from a sample about an unknown population parameter. An estimator is a statistic that is used to infer the value of an unknown parameter.

A point estimate is the best estimate, in some sense, of the parameter based on a sample. It should be obvious that any point estimate is not absolutely accurate. It is an estimate based on only a single random sample. If repeated random samples were taken from the population, the point estimate would be expected to vary from sample to sample.

A confidence interval is an estimate constructed on the basis that a specified proportion of the confidence intervals include the true parameter in repeated sampling. How frequently the confidence interval contains the parameter is determined by the confidence level. 95% is commonly used and means that in repeated sampling 95% of the confidence intervals include the parameter. 99% is sometimes used when more confidence is needed and means that in repeated sampling 99% of the intervals include the true parameter. It is unusual to use a confidence level of less than 90% as too many intervals would fail to include the parameter. Likewise, confidence levels larger than 99% are not used often because the intervals become wider the higher the confidence level and therefore require large sample sizes to make usable intervals.

Many people misunderstand confidence intervals. A confidence interval does not predict with a given probability that the parameter lies within the interval. The problem arises because the word confidence is misinterpreted as implying probability. In frequentist statistics, probability statements cannot be made about parameters. Parameters are fixed, not random variables, and so a probability statement cannot be made about them. When a confidence interval has been constructed, it either does or does not include the parameter.

In recent years the use of confidence intervals has become more common. A confidence interval provides much more information than just a hypothesis test p-value. It indicates the uncertainty of an estimate of a parameter and allows you to consider the practical importance, rather than just statistical significance.

Confidence intervals and hypothesis tests are closely related. Most introductory textbooks discuss how confidence intervals are equivalent to a hypothesis test. When the 95% confidence interval contains the hypothesized value the hypothesis test is statistically significant at the 5% significance level, and when it does not contain the value the test is not significant. Likewise, with a 99% confidence interval and hypothesis test at the 1% significance level. A confidence interval can be considered as the set of parameter values consistent with the data at some specified level, as assessed by testing each possible value in turn.

The relationship between hypothesis tests and confidence interval only holds when the estimator and the hypothesis test both use the same underlying evidence function. When a software package uses different evidence functions for the confidence interval and hypothesis test, the results can be inconsistent. The hypothesis test may be statistically significant, but the confidence interval may include the hypothesized value suggesting the result is not significant. Where possible the same underlying evidence function should be used to form the confidence interval and test the hypotheses. Be aware that not many statistical software packages follow this rule.

Estimation in Statistics

In statistics, estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample.

Point Estimate vs. Interval Estimate

Statisticians use sample statistics to estimate population parameters. For example, sample means are used to estimate population means; sample proportions, to estimate population proportions.

An estimate of a population parameter may be expressed in two ways:

a) **Point estimate:** A point estimate of a population parameter is a single value of a statistic. For example, the sample mean x is a point estimate of the population mean μ. Similarly, the sample proportion p is a point estimate of the population proportion P.

b) **Interval estimate:** An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example, $a < x < b$ is an interval estimate of the population mean μ. It indicates that the population mean is greater than a but less than b.

1.10 METHOD OF MAXIMUM LIKELIHOOD, INTERVAL ESTIMATION-CONFIDENCE INTERVAL

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model, given observations. MLE attempts to find the parameter values that maximize the likelihood function, given the observations. The resulting estimate is called a maximum likelihood estimate, which is also abbreviated as MLE.

The method of maximum likelihood is used with a wide range of statistical analyses. As an example, suppose that we are interested in the heights of adult female penguins, but are unable to measure the height of every penguin in a population (due to cost or time constraints). Assuming that the heights are normally distributed with some unknown mean and variance, the mean and variance can be estimated with MLE while only knowing the heights of some sample of the overall population. MLE would accomplish that by taking the mean and variance as parameters and finding particular parametric values that make the observed results the most probable given the normal model.

From the point of view of Bayesian inference, MLE is a special case of maximum a posteriori estimation (MAP) that assumes a uniform prior distribution of the parameters. In frequentist inference, MLE is one of several methods to get estimates of parameters without using prior distributions. Priors are avoided by not making probability statements about the parameters, but only about their estimates, whose properties are fully defined by the observations and the statistical model.

Interval Estimation - Confidence Interval

Interval estimation is the use of sample data to calculate an interval of possible (or probable) values of an unknown population parameter, in contrast to point estimation, which is a single number.

Interval estimation, in statistics, the evaluation of a parameter for example, the mean (average) of a population by computing an interval, or range of values, within which the parameter is most likely to be located. Intervals are commonly chosen such that the parameter falls within with a 95 or 99 percent probability, called the confidence coefficient. Hence, the intervals are called confidence intervals; the end points of such an interval are called upper and lower confidence limits.

The interval containing a population parameter is established by calculating that statistic from values measured on a random sample taken from the population and by applying the knowledge (derived from probability theory) of the fidelity with which the properties of a sample represent those of the entire population.

The probability tells what percentage of the time the assignment of the interval will be correct but not what the chances are that it is true for any given sample. Of the intervals computed from many samples, a certain percentage will contain the true value of the parameter being sought.

Confidence Intervals

Statisticians use a confidence interval to express the precision and uncertainty associated with a particular sampling method. A confidence interval consists of three parts.

a) A confidence level.

b) A statistic.

c) A margin of error.

The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the sample statistic + margin of error.

For example, suppose we compute an interval estimate of a population parameter. We might describe this interval estimate as a 95% confidence interval. This means that if we used the same sampling method to select different samples and compute different interval estimates, the true population parameter would fall within a range defined by the sample statistic + margin of error 95% of the time.

Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

Confidence Level

The probability part of a confidence interval is called a confidence level. The confidence level describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter.

Here is how to interpret a confidence level. Suppose we collected all possible samples from a given population, and computed confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.

Margin of Error

In a confidence interval, the range of values above and below the sample statistic is called the margin of error.

For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.

Note: Many public opinion surveys report interval estimates, but not confidence intervals. They provide the margin of error, but not the confidence level. To clearly interpret survey results you need to know both! We are much more likely to accept survey findings if the confidence level is high (say, 95%) than if it is low (say, 50%).

Problem Statement

Suppose a student measuring the boiling temperature of a certain liquid observes the readings (in degrees Celsius) 102.5, 101.7, 103.1, 100.9, 100.5 and 102.2 on 6 different samples of the liquid. He calculates the sample mean to be 101.82. If he knows that the standard deviation for this procedure is 1.2 degrees, what is the interval estimation for the population mean at a 95% confidence level?

Solution:

The student calculated the sample mean of the boiling temperatures to be 101.82, with standard deviation s = 0.49. The critical value for a 95% confidence interval is 1.96, where 1-0.952 = 0.025. A 95% confidence interval for the unknown mean.

= ((101.82 - (1.96 × 0.49)), (101.82 + (1.96 × 0.49))) = (101.82 - 0.96, 101.82 + 0.96) = (100.86,102.78)

As the level of confidence decreases, the size of the corresponding interval will decrease. Suppose the student was interested in a 90% confidence interval for the boiling temperature. In this case, s = 0.90 and 1-0.902 = 0.05. The critical value for this level is equal to 1.645, so the 90% confidence interval is:

= ((101.82 - (1.645 × 0.49)),(101.82 + (1.645 × 0.49))) = (101.82-0.81,101.82 + 0.81) = (101.01,102.63)

An increase in sample size will decrease the length of the confidence interval without reducing the level of confidence. This is because the standard deviation decreases as n increases.

Margin of Error

The margin of error m of interval estimation is defined to be the value added or subtracted from the sample mean which determines the length of the interval:

Suppose in the example above, the student wishes to have a margin of error equal to 0.5 with 95% confidence. Substituting the appropriate values into the expression for m and solving for n gives the calculation.

n = (1.96 × 1.20.5)2 = 2.350.52 = (4.7)2 = 22.09

To achieve 95% interval estimation for the mean boiling point with total length less than 1 degree, the student will have to take 23 measurements.

1.11 TEST OF HYPOTHESIS

Hypothesis test is a method of making decisions using data from a scientific study. In statistics, a result is called statistically significant if it has been predicted as unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the significance level. The phrase "test of significance" was coined by statistician Ronald Fisher. These tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance; this can help to decide whether results contain enough information to cast doubt on conventional wisdom, given that conventional wisdom has been used to establish the null hypothesis. The critical region of a hypothesis test is the set of all outcomes which cause the null hypothesis to be rejected in favor of the alternative hypothesis. Statistical hypothesis testing is sometimes called confirmatory data analysis, in contrast to exploratory data analysis, which may not have pre-specified hypotheses. Statistical hypothesis testing is a key technique of frequents inference.

Statistical hypothesis tests define a procedure that controls (fixes) the probability of incorrectly deciding that a default position (null hypothesis) is incorrect based on how likely it would be for a set of observations to occur if the null hypothesis were true. Note that this probability of making an incorrect decision is not the probability that the null hypothesis is true, nor whether any specific alternative hypothesis is true. This contrasts with other possible techniques of decision theory in which the null and alternative hypothesis are treated on a more equal basis. One naive Bayesian approach to hypothesis testing is to base decisions on the posterior probability, but this fails when comparing point and continuous hypotheses. Other approaches to decision making, such as Bayesian decision theory, attempt to balance the consequences of incorrect decisions across all possibilities, rather than concentrating on a single null hypothesis. A number of other approaches to reaching a decision based on data are available via decision theory and optimal decisions, some of which have desirable properties, yet hypothesis testing is a dominant approach to data analysis in many fields of science. Extensions to the theory of hypothesis testing include the study of the power of tests, which refers to the probability of correctly rejecting the null hypothesis when a given state of nature exists. Such considerations can be used for the purpose of sample size determination prior to the collection of data.

Meaning of Hypothesis testing

Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses. It is an assumption about a population parameter. This assumption may or may not be true.

Explanation

The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

In doing so, one has to take the help of certain assumptions or hypothetical values about the characteristics of the population if some such information is available. Such hypothesis about the population is termed as statistical hypothesis and the hypothesis is tested on the basis of sample values. The procedure enables one to decide on a certain hypothesis and test its significance. “A claim or hypothesis about the population parameters is known as Null Hypothesis and is written as, H0.”

This hypothesis is then tested with available evidence and a decision is made whether to accept this hypothesis or reject it. If this hypothesis is rejected, then we accept the alternate hypothesis. This hypothesis is written as H1.

For testing hypothesis or test of significance we use both parametric tests and nonparametric or distribution free tests. Parametric tests assume within properties of the population, from which we draw samples. Such assumptions may be about population parameters, sample size, etc. In case of non-parametric tests, we do not make such assumptions. Here we assume only nominal or ordinal data.

1.12 TYPES OF STATISTICAL HYPOTHESIS

There are two types of statistical hypotheses:

1. ***Null hypothesis:*** The null hypothesis, denoted by H0, is usually the hypothesis that sample observations result purely from chance.
2. ***Alternative hypothesis:*** The alternative hypothesis, denoted by H1 or Ha, is the hypothesis that sample observations are influenced by some non-random cause.

For example, suppose we wanted to determine whether a coin was fair and balanced. A null hypothesis might be that half the flips would result in Heads and half, in Tails. The alternative hypothesis might be that the number of Heads and Tails would be very different. Symbolically, these hypotheses would be expressed as:

H0: P = 0.5

Ha: P ? 0.5

Suppose we flipped the coin 50 times, resulting in 40 Heads and 10 Tails. Given this result, we would be inclined to reject the null hypothesis. We would conclude, based on the evidence, that the coin was probably not fair and balanced.

1.13 STEPS IN HYPOTHESIS TESTS

Statisticians follow a formal process to determine whether to reject a null hypothesis, based on sample data. This process, called hypothesis testing, consists of four steps:

Step-1: State the hypotheses: This involves stating the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.

Step-2: Formulate an analysis plan: The analysis plan describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.

Step-3: Analyze sample data: Find the value of the test statistic (mean score, proportion, t-score, z-score, etc.) described in the analysis plan.

Step-4: Interpret results: Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.

Decision Rules

The analysis plan includes decision rules for rejecting the null hypothesis. In practice, statisticians describe these decision rules in two ways - with reference to a P-value or with reference to a region of acceptance.

1. ***P-value:*** The strength of evidence in support of a null hypothesis is measured by the P-value. Suppose the test statistic is equal to S. The P-value is the probability of observing a test statistic as extreme as S, assuming the null hypothesis is true. If the P-value is less than the significance level, we reject the null hypothesis.
2. ***Region of acceptance:*** The region of acceptance is a range of values. If the test statistic falls within the region of acceptance, the null hypothesis is not rejected. The region of acceptance is defined so that the chance of making a Type I error is equal to the significance level. The set of values outside the region of acceptance is called the region of rejection. If the test statistic falls within the region of rejection, the null hypothesis is rejected. In such cases, we say that the hypothesis has been rejected at a level of significance.

One-Tailed and Two-Tailed Tests

A test of a statistical hypothesis, where the region of rejection is on only one side of the sampling distribution, is called a one-tailed test. For example, suppose the null hypothesis states that the mean is less than or equal to 10. The alternative hypothesis would be that the mean is greater than 10. The region of rejection would consist of a range of numbers located on the right side of sampling distribution; that is, a set of numbers greater than 10.

A test of a statistical hypothesis, where the region of rejection is on both sides of the sampling distribution, is called a two-tailed test. For example, suppose the null hypothesis states that the mean is equal to 10. The alternative hypothesis would be that the mean is less than 10 or greater than 10. The region of rejection would consist of a range of numbers located on both sides of sampling distribution; that is, the region of rejection would consist partly of numbers that were less than 10 and partly of numbers that were greater than 10.

Notes

Procedure for Testing of Hypothesis

1. State the null hypothesis as well as the alternate hypothesis

For example, let us assume the population mean = 50 and set up the hypothesis μ = 50. This is called the null hypothesis and is denoted as;

Null hypothesis, H0: μ = 50

Alternative hypothesis H1: μ ? 50

Or μ > 50

μ < 50

2. Establish a level of significance

The level of significance signifies the probability of committing Type 1 error a and is generally taken as equal to 0.05. Sometimes, the value a is established as 0.01, but it is at the discretion of the investigator to select its value, depending upon the sensitivity of the study. To illustrate per cent level of significance indicates that a researcher is willing to take 5 per cent risk of rejecting the Null Hypothesis when it happens to be true.

3. Choosing a suitable test statistic

Now the researcher would choose amongst the various tests (i.e. z, t, χ^2 and f-tests). Actually, for the purpose of rejecting or accepting the null hypothesis, a suitable statistics called 'test statistics' is chosen. This means that H0 is assumed to be really true. Obviously due to sampling fluctuations, the observed value of the statistic based on random sample will differ from the expected value. If the difference is large enough, one suspects the validity of the assumption and rejects the null hypothesis (H0). On the other hand, if the difference may be assumed due to sampling (random) fluctuation, the null hypothesis (H0) is accepted.

4. Defining the critical rejection regions and making calculations for test statistics

If we select the value of α = Level of significance = 0.05, and use the standard normal distribution (z-test) as our test statistic for testing the population parameter u, then the value of the difference between the assumption of null hypothesis (assumed value of the population parameter) and the value obtained by the analysis of the sample results is not expected to be more than 1.96 σ at α = 0.05.

1.14 SIMPLE AND COMPOSITE HYPOTHESIS

If a random sample is taken from a distribution with parameter θ, a hypothesis is said to be a simple hypothesis if the hypothesis uniquely specifies the distribution of the population from which the sample is taken. Any hypothesis that is not a simple hypothesis is called a composite hypothesis.

Simple Hypothesis

It refers to the one in which all parameters associated with the distribution are stated. For instance, if the height of the students in a school is distributed normally with $\sigma^2 = 6$ and the hypothesis that the mean stands equivalent to 70 implying $H_o : \mu = 70$. This stands to be the simple hypothesis as variance and mean both completely specify the normal distribution.

In general, a simple hypothesis reflects that $\theta_0 = \theta$ where θ_0 shows the parameters' specified value, wherein θ reflects μ, p, $\mu_1 - \mu_2$ etc.

Composite Hypothesis

It refers to the hypothesis that does not stand to be simple. For instance, if it is assumed that $H_o : \mu \succ 70$ and $\sigma^2 \prec 6$ or $H_o : \mu = 70$ and $\sigma^2 \prec 6$ then such hypothesis tends to become a composite hypothesis. This is because, in either of the cases, the exact distribution is not known.

The form associated with the composite hypothesis that stands to be common is $\theta \leq \theta_0$ or $\theta \geq \theta_0$. It reflects that parameter does not fall short or does not exceed beyond the value that is being specified by θ_0. Such concept pertaining to composite and simple hypothesis stands applicable to both null and alternative hypothesis.

The main difference between simple hypothesis and the composite hypothesis is listed below:

1. A composite hypothesis does not specify the distribution completely, however simple hypothesis specifies the distribution completely.
2. When a simple null hypothesis is required to be tested in consultation with a composite alternative, the power associated with the test tends to be a function pertaining to a parameter of interest.
3. Variation in power can be experienced due to the size of the sample.

Probability distributions are widely used primarily in experiments which involve counting. The sampling errors which occur in counting experiments are called statistical errors. Statistical errors are one special kind of error in a class of errors which are known as random errors. You will find that what you learn in this laboratory is relevant not only in the natural and social sciences, but also in every day life. Please read the theory section that follows, and then the file on Error Analysis before proceeding to do the prelab. Bring the completed error analysis prelab with you.

This section will help the student with the prelab homework. You are probably familiar with polls conducted before a presidential election. If the sample of people who are polled is carefully chosen to represent the general population, then the error in the prediction depends on the number of people in the sample. The larger the number of people, the smaller the error. If the sample is not properly chosen, it would result in a bias (i.e. an additional systematic error).

If a fraction p of the population will vote Democratic and a fraction q = (1-p) will vote Republican, then one expects that in a sample of N people, one will find on average people who say that they will vote Democratic and N(1 - p) who say that they will vote Republican. If this poll is taken many times for different samples one will find that the distribution of the results for x (which is the number of people who say they will vote Democratic) follows a binomial distribution with the mean of x. The probability distribution B(x) for finding x in a sample of N is a function of the probabilities p and q, and is given by the binomial distribution as follows:

$$B(x) = \{{}_N C_x\} p^x q^{N-x} = \left(\frac{N!}{x!(N-x)!}\right) p^x q^{N-x}$$

Notes

1.15 TWO TYPES OF ERRORS

Two types of errors can result from a hypothesis test.

i) ***Type I error:*** A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the significance level. This probability is also called alpha, and is often denoted by a.

   A type I error occurs when the null hypothesis (H0) is true, but is rejected. It is asserting something that is absent, a false hit. A type I error may be likened to a so-called false positive (a result that indicates that a given condition is present when it actually is not present).

   In terms of folk tales, an investigator may see the wolf when there is none ("raising a false alarm"). Where the null hypothesis, H0, is: no wolf.

   The type I error rate or significance level is the probability of rejecting the null hypothesis given that it is true. It is denoted by the Greek letter a (alpha) and is also called the alpha level. Often, the significance level is set to 0.05 (5%), implying that it is acceptable to have a 5% probability of incorrectly rejecting the null hypothesis.

ii) ***Type II error:*** A Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called Beta, and is often denoted by ß. The probability of not committing a Type II error is called the Power of the test.

A type II error occurs when the null hypothesis is false, but erroneously fails to be rejected. It is failing to assert what is present, a miss. A type II error may be compared with a so-called false negative (where an actual 'hit' was disregarded by the test and seen as a 'miss') in a test checking for a single condition with a definitive result of true or false. A Type II error is committed when we fail to believe a true alternative hypothesis.

In terms of folk tales, an investigator may fail to see the wolf when it is present ("failing to raise an alarm"). Again, H0: no wolf.

The rate of the type II error is denoted by the Greek letter ß (beta) and related to the power of a test (which equals 1-ß).

Example 1:

Hypothesis: "Adding water to toothpaste protects against cavities."

Null hypothesis (H0): "Adding water does not make toothpaste more effective in fighting cavities."

This null hypothesis is tested against experimental data with a view to nullifying it with evidence to the contrary.

A type I error occurs when detecting an effect (adding water to toothpaste protects against cavities) that is not present. The null hypothesis is true (i.e., it is true that adding water to toothpaste does not make it more effective in protecting against cavities), but this null hypothesis is rejected based on bad experimental data or an extreme outcome of chance alone.

Example 2:

Hypothesis: "Adding fluoride to toothpaste protects against cavities."

Null hypothesis (H0): "Adding fluoride to toothpaste has no effect on cavities."

This null hypothesis is tested against experimental data with a view to nullifying it with evidence to the contrary.

A type II error occurs when failing to detect an effect (adding fluoride to toothpaste protects against cavities) that is present. The null hypothesis is false (i.e., adding fluoride is actually effective against cavities), but the experimental data is such that the null hypothesis cannot be rejected.

Example 3;

Hypothesis: "The evidence produced before the court proves that this man is guilty."

Null hypothesis (H0): "This man is innocent."

A type I error occurs when convicting an innocent person (a miscarriage of justice). A type II error occurs when letting a guilty person go free (an error of impunity).

A positive correct outcome occurs when convicting a guilty person. A negative correct outcome occurs when letting an innocent person go free.

Example 4:

Hypothesis: "A patient's symptoms improve after treatment A more rapidly than after a placebo treatment."

Null hypothesis (H0): "A patient's symptoms after treatment A are indistinguishable from a placebo."

A Type I error would falsely indicate that treatment A is more effective than the placebo, whereas a Type II error would be a failure to demonstrate that treatment A is more effective than placebo even though it actually is more effective.

1.16 NEYMAN PEARSON LEMMA

The Neyman-Pearson Lemma is a way to find out if the hypothesis test you are using is the one with the greatest statistical power. The power of a hypothesis test is the probability that test correctly rejects the null hypothesis when the alternate hypothesis is true. The goal would be to maximize this power, so that the null hypothesis is rejected as much as possible when the alternate is true. The lemma basically tells us that good hypothesis tests are likelihood ratio tests.

The lemma is named after Jerzy Neyman and Egon Sharpe Pearson, who described it in 1933. It is considered by many to the theoretical foundation of hypothesis testing theory, from which all hypothesis tests are built.

The Neyman-Pearson lemma is based on a simple hypothesis test. A "simple" hypothesis test is one where the unknown parameters are specified as single values. For example:

- H0: $\mu = 0$ is simple because the population mean is specified as 0 for the null hypothesis.
- H0: $\mu = 0$; H_A: $\mu = 1$ is also simple because the population mean for the null hypothesis and alternate hypothesis are specified, single values. Basically, you're assuming that the parameters for this test can only be 0, or 1 (which is theoretically possible if the test was binomial).

In contrast, the hypothesis $\sigma^2 > 7$ isn't simple; it's a composite hypothesis test that doesn't state a specific value for σ^2.

Notes

Simple hypothesis tests even optimized ones have limited practical value. However, they are important hypothetical tools; the simple hypothesis test is the one that all others are built on.

Alpha and Beta Levels

A Type I error under the null hypothesis is defined as:

$P_0(X \in R \mid H0 \text{ is true})$,

Where:

- R = the rejection region and
- $\in$ = is the set membership.

A Type II error under the null hypothesis is defined as:

$P_0(X \in R_c \mid H0 \text{ is false})$,

Where:

- R_c = the complement of R.

Usually, an alpha level is set (e.g. 0.05) to restrict the probability of making a Type I error (α) to a certain percentage (in this case, 5%). Next, a test is chosen which minimizes Type II errors (β).

1.17 POWER FUNCTION OF A TEST

Tests with a certain alpha level α can be written as:

Size a tests: $\sup ß(\theta) = \alpha(\theta \in \Theta_0)$

Where:

- Θ_0 = set of all possible values for θ under the null hypothesis

A level a test is one that has the largest power function. Mathematically, it is written as:

Level α test: $\sup ß(\theta) \leq \alpha(\Theta \in \Theta_0)$

1.18 LIKELIHOOD RATIO

Likelihood functions for reliability data are described in Section 4. Two ways we use likelihood functions to choose models or verify/validate assumptions are:

1. Calculate the maximum likelihood of the sample data based on an assumed distribution model (the maximum occurs when unknown parameters are replaced by their maximum likelihood estimates). Repeat this calculation for other candidate distribution models that also appear to fit the data (based on probability plots). If all the models have the same number of unknown parameters, and there is no convincing reason to choose one particular model over another based on the failure mechanism or previous successful analyses, then pick the model with the largest likelihood value.
2. Many model assumptions can be viewed as putting restrictions on the parameters in a likelihood expression that effectively reduce the total number of unknown parameters. Some common examples are:

Examples: where assumptions can be tested by the Likelihood Ratio Test.

i) It is suspected that a type of data, typically modeled by a Weibull distribution, can be fit adequately by an exponential model. The exponential distribution is a special case of the Weibull, with the shape parameter γ set to 1. If we write the Weibull likelihood function for the data, the exponential model likelihood function is obtained by setting γ to 1, and the number of unknown parameters has been reduced from two to one.

ii) Assume we have n cells of data from an acceleration test, with each cell having a different operating temperature. We assume a lognormal population model applies in every cell. Without an acceleration model assumption, the likelihood of the experimental data would be the product of the likelihoods from each cell and there would be 2n unknown parameters (a different T50 and s for each cell). If we assume an Arrhenius model applies, the total number of parameters drops from 2n to just 3, the single common s and the Arrhenius A and ?Hparameters. This acceleration assumption "saves" (2n-3) parameters.

iii) We life test samples of product from two vendors. The product is known to have a failure mechanism modeled by the Weibull distribution, and we want to know whether there is a difference in reliability between the vendors. The unrestricted likelihood of the data is the product of the two likelihoods, with 4 unknown parameters (the shape and characteristic life for each vendor population). If, however, we assume no difference between vendors, the likelihood reduces to having only two unknown parameters (the common shape and the common characteristic life). Two parameters are "lost" by the assumption of "no difference".

Clearly, we could come up with many more examples like these three, for which an important assumption can be restated as a reduction or restriction on the number of parameters used to formulate the likelihood function of the data. In all these cases, there is a simple and very useful way to test whether the assumption is consistent with the data.

The Likelihood Ratio Test Procedure

Details of the Likelihood Ratio Test procedure

In general, calculations are difficult and need to be built into the software you use Let L1 be the maximum value of the likelihood of the data without the additional assumption. In other words, L1 is the likelihood of the data with all the parameters unrestricted and maximum likelihood estimates substituted for these parameters.

Let L0 be the maximum value of the likelihood when the parameters are restricted (and reduced in number) based on the assumption. Assume k parameters were lost (i.e., L0 has kless parameters than L1).

Form the ratio λ = L0/L1. This ratio is always between 0 and 1 and the less likely the assumption is, the smaller λ will be. This can be quantified at a given confidence level as follows:

1. Calculate $\chi^2 = -2 \ln \lambda$. The smaller λ is, the larger χ^2 will be.
2. We can tell when χ^2 is significantly large by comparing it to the 100(1-a) percentile point of a Chi-Square distribution with degrees of freedom. χ^2 has an approximate Chi-Square distribution with k degrees of freedom and the approximation is usually good, even for small sample sizes.

Notes

3. The likelihood ratio test computes χ^2 and rejects the assumption if χ^2 is larger than a Chi-Square percentile with k degrees of freedom, where the percentile corresponds to the confidence level chosen by the analyst.

1.19 TEST EXACT

An exact (significance) test is a test where all assumptions, upon which the derivation of the distribution of the test statistic is based, are met as opposed to an approximate test (in which the approximation may be made as close as desired by making the sample size big enough). This will result in a significance test that will have a false rejection rate always equal to the significance level of the test. For example an exact test at significance level 5% will in the long run reject true null hypotheses exactly 5% of the time.

Parametric tests, such as those described in exact statistics, are exact tests when the parametric assumptions are fully met, but in practice the use of the term exact (significance) test is reserved for those tests that do not rest on parametric assumptions – non-parametric tests. However, in practice most implementations of non-parametric test software use asymptotical algorithms for obtaining the significance value, which makes the implementation of the test non-exact.

So when the result of a statistical analysis is said to be an “exact test” or an “exact p-value”, it ought to imply that the test is defined without parametric assumptions and evaluated without using approximate algorithms. In principle however it could also mean that a parametric test has been employed in a situation where all parametric assumptions are fully met, but it is in most cases impossible to prove this completely in a real world situation. Exceptions when it is certain that parametric tests are exact include tests based on the binomial or Poisson distributions. Sometimes permutation test is used as a synonym for exact test, but although all permutation tests are exact tests, not all exact tests are permutation tests.

Biologist and statistician Ronald Fisher

Fisher's exact test is a statistical significance test used in the analysis of contingency tables. Although in practice it is employed when sample sizes are small, it is valid for all sample sizes. It is named after its inventor, Ronald Fisher, and is one of a class of exact tests, so called because the significance of the deviation from a null hypothesis (e.g., P-value) can be calculated exactly, rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity, as with many statistical tests.

Fisher is said to have devised the test following a comment from Muriel Bristol, who claimed to be able to detect whether the tea or the milk was added first to her cup. He tested her claim in the "lady tasting tea" experiment.

Purpose and Scope

The test is useful for categorical data that result from classifying objects in two different ways; it is used to examine the significance of the association (contingency) between the two kinds of classification. So in Fisher's original example, one criterion of classification could be whether milk or tea was put in the cup first; the other could be whether Bristol thinks that the milk or tea was put in first. We want to know whether these two classifications are associated that is, whether Bristol really can tell whether milk or tea was poured in first. Most uses of the Fisher test involve, like this example, a 2 × 2 contingency table. The p-value from the test is computed as if the margins of the table are fixed, i.e. as if, in the tea-tasting example, Bristol knows the number of cups with each treatment (milk or tea first) and will therefore provide guesses with the correct number in each category. As pointed out by Fisher, this leads under a null hypothesis of independence to a hypergeometric distribution of the numbers in the cells of the table.

With large samples, a chi-squared test (or better yet, a G-test) can be used in this situation. However, the significance value it provides is only an approximation, because the sampling distribution of the test statistic that is calculated is only approximately equal to the theoretical chi-squared distribution. The approximation is inadequate when sample sizes are small, or the data are very unequally distributed among the cells of the table, resulting in the cell counts predicted on the null hypothesis (the “expected values”) being low. The usual rule of thumb for deciding whether the chi-squared approximation is good enough is that the chi-squared test is not suitable when the expected values in any of the cells of a contingency table are below 5 or below 10 when there is only one degree of freedom (this rule is now known to be overly conservative). In fact, for small, sparse, or unbalanced data, the exact and asymptotic p-values can be quite different and may lead to opposite conclusions concerning the hypothesis of interest. In contrast the Fisher exact test is, as its name states, exact as long as the experimental procedure keeps the row and column totals fixed, and it can therefore be used regardless of the sample characteristics. It becomes difficult to calculate with large samples or well-balanced tables, but fortunately these are exactly the conditions where the chi-squared test is appropriate.

For hand calculations, the test is only feasible in the case of a 2 × 2 contingency table. However the principle of the test can be extended to the general case of an m × n table, and some statistical packages provide a calculation (sometimes using a Monte Carlo method to obtain an approximation) for the more general case.

1.20 SAMPLING DISTRIBUTIONS

A sampling distribution or finite-sample distribution is the probability distribution of a given random-sample-based statistic. If an arbitrarily large number of samples, each involving multiple observations (data points), were separately used in order to compute one value of a statistic (such as, for example, the sample mean or sample variance) for each sample, then the sampling distribution is the probability distribution of the values that the statistic takes on. In many contexts, only one sample is observed, but the sampling distribution can be found theoretically.

Sampling distributions are important in statistics because they provide a major simplification en route to statistical inference. More specifically, they allow analytical considerations to be based on the probability distribution of a statistic, rather than on the joint probability distribution of all the individual sample values.

Sampling distributions are important in the understanding of statistical inference. Probability distributions permit us to answer questions about sampling and they provide the foundation for statistical inference procedures.

Suppose that we draw all possible samples of size n from a given population. Suppose further that we compute a statistic (e.g., a mean, proportion, standard deviation) for each sample. The probability distribution of this statistic is called a sampling distribution. And the standard deviation of this statistic is called the standard error.

Variability of a Sampling Distribution

The variability of a sampling distribution is measured by its variance or its standard deviation. The variability of a sampling distribution depends on three factors:

N: The number of observations in the population.

n: The number of observations in the sample.

The way that the random sample is chosen.

If the population size is much larger than the sample size, then the sampling distribution has roughly the same standard error, whether we sample with or without replacement. On the other hand, if the sample represents a significant fraction (say, 1/20) of the population size, the standard error will be meaningfully smaller, when we sample without replacement.

Meaning of Sampling

Sampling refers to the process of using statistical analysis in which a predetermined number of observations will be taken from a larger population.

Characteristics of Good Sample

The characteristics of good sample are as follows:

1. Representatives

A sample is a subset of the population or universe. The sample must be representatives of the universe. Therefore, the researcher must select the sample members who have the characteristics of the universe. For example, when a research is undertaken to study job satisfaction in police force; then the sample members must be the police persons belonging to different levels in the police force.

2. Focus on Objectives

The sample size must be selected depending upon the research objectives. For instance, if a research is undertaken 'to find out the impact of inflation on the poor' then the sample size would be larger, as there are more poor households in India.

3. Flexibility

The sample size should not be rigidly followed. The sample size can be modified depending upon the circumstances. For instance, the sample size may be reduced, if sufficient information is already available or if there is a limitation of time and funds. However, sample size may be increased, if proper information is not available from the current sample.

4. Method of Sampling

The researcher must select proper method of sampling. The sampling methods are broadly divided into two groups – probability methods and non-probability methods. Certain methods require less time to complete data collection. For instance, convenience sampling requires less time to collect data. Therefore, the researcher may select convenience method, if there is limitation of time.

5. Proper Selection of Sample Unit

The sample unit must be appropriate. The universe comprises of the elements, and each element can be further divided into units. For instance, if a study is conducted to study job satisfaction among bank employees, then bank employees comprise the universe. The element of universe may comprise of bank employees/manager in rural banks, and in urban banks. The sample unit may include male and female employees, junior or senior employees depending upon the type of research, researcher must select proper sampling of unit(s) to conduct the research activity.

6. Proper Sampling Frame

The researcher should select proper sampling frame to collect information .sampling frame is an instrument to obtain addresses or such other information about various element of the universe. The sampling frame may include telephone directories, register of member in an organization, etc.

7. Geographic Area of the Study

The researcher must consider the size of the area for selecting the sample size. For instance, if the area coverage is large such as the entire state or country, then the size of the sample would be large. In such situation, the researcher may adopt multi-stage cluster sampling.

8. Suitability

The sample size should be suitability to collect the relevant data. For instance, if a research is conducted to find out reading habits of college student in the city of Mumbai, then the sample would be the students from the colleges of Mumbai city, and the sample size may be smaller. But if the research is conducted to find out the reading habits of college students in India, then the sample would consist of students from various colleges across India, and the sample size would be larger.

9. Economy

The sample size must be economical. The sample size must be cost-effective.it should not put extra burden on the resources .At the same time, the sample size should be such that it facilitates proper collection of data. Normally, the researcher must first consider the availability of resources and then plan for the sample size. For instance, the sample size can be large; funds are available for research activity, and vice-versa.

Importance of Sampling

The importance of sampling can be summarized as follows:

i) A large number of units can be studied.

ii) It saves a lot of time, energy and money.

iii) Homogeneous universe sampling is very useful.

iv) Intensive study is possible.

v) When the data are unlimited – highly useful.

vi) When cent per cent accuracy is not required.

vii) It makes easier for tabulation and analysis.

Error in Sampling

A sample is expected to represent the population from which it comes; however, there is no guarantee that any sample will be precisely representative of the population from which it comes. In practice, it is rarely known when a sample is unrepresentative and should be discarded. The sample may be unrepresentative because of sampling or non-sampling errors.

1. Sampling Errors

Sampling error comprises the differences between the sample and the population that are due solely to the particular units that happen to have been selected. For example, suppose that a sample of 100 females from Haryana is taken and all are found to be taller than six feet. It is very clear even without any statistical prove that this would be a highly unrepresentative sample leading to invalid conclusions. This is a very unlikely occurrence because naturally such rare cases are widely distributed among the population. But it can occur. Sampling error may be committed due to the chance factor. Unusual units in a population do exist and-there is always a possibility that an abnormally large number of them will be chosen. Sampling error may also be committed due to sampling bias. Sampling bias is a tendency to favour the selection of units that have particular characteristics. Sampling bias is usually the result of a poor sampling plan. The most notable is the bias of non-response when for some reason some units have no chance of appearing in the sample. As an example we would like to know the average income of some community and we decide to use the telephone numbers to select a sample of the total population in a locality where only the rich and middle class households have telephone lines. We will end up with high average income, which will lead to the wrong policy decisions.

2. Non Sampling Errors

Non-sampling errors occur whether a census or a sample is being used. A non-sampling error is an error that results solely from the manner in which the observations are made. The simplest example of non-sampling error is inaccurate measurements due to malfunctioning instruments or poor procedures. For example, if persons are asked to state their own weights themselves, no two answers will be of equal reliability. An individual's weight fluctuates during the day and so the time of weighing will also affect the answer.

Sampling Design

Sampling design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sampling design is determined before any data are collected.

Guidelines for Sampling Design

While developing a sampling strategy, the researcher must pay attention to the following points:

i) The first step in developing any sample design is to clearly define the population to be sampled.

ii) A decision has to be taken concerning a sampling unit before selecting sample. Sampling unit may be of some geographical area such as a state, district, village, etc., or construction units such as house, flat, etc. It may be a social unit such as family, club, school, etc., or an individual.

iii) Frame should be comprehensive, correct, reliable and appropriate. It is extremely important for the frame to be as representative of the population as possible.

iv) The size of sample should neither be excessively large, nor too small. It should be optimum. An optimum sample is one which fulfils the requirements of efficiency, representativeness, reliability and flexibility.

v) In determining the sample design, one must take into consideration the specific population parameters, which are of interest.

vi) Cost considerations, from practical point of view, have a major, impact upon decisions relating to not only the size of the sample but also to the sample design. Cost constraint can even lead to the use of a non-probability sample.

Characteristics of a Good Sample Design

A good sampling design is said to posses the following characteristics:

i) It should result in a truly representative sample.

ii) It should result in a small sampling error

iii) It should be within budget available for the research study.

iv) It should be such so that systematic bias can be controlled in a better way.

Steps in Sampling Design

The researcher must keep in mind the following points while preparing a sample design:

(i) ***Universe:*** While preparing a sample design, it is foremost required to define the set of objects to be studied. Technically, it is also known as the Universe, which can be finite or infinite. In case of a finite universe, the number of items is limited. Whereas an infinite universe the number of items is limitless.

(ii) ***Sampling unit:*** It is necessary to decide a sampling unit before selecting a sample. It can be a geographical one (state, district, village, etc.), a construction unit (house, flat, etc.), a social unit (family, club, school, etc.), or an individual.

(iii) ***Source list:*** In other words, it is called the 'sampling frame' from which the sample is drawn. It comprises the names of all items of a universe (finite universe only). If source list/sampling frame is unavailable, the researcher has to prepare it by himself.

(iv) ***Sample size:*** This is the number of items, selected from the universe, constituting a sample. The sample size should not be too large or too small, but optimum. In other words, an optimum sample accomplishes the requirements of efficiency, representativeness, reliability and flexibility.

(v) ***Parameters of interest:*** While determining a sample design, it is required to consider the question of the specific population parameters of interest. For example, we may like to estimate the proportion of persons with some specific attributes in the population, or we may also like to know some average or other measure concerning the population.

(vi) ***Budgetary constraint:*** Practically, cost considerations have a major impact upon the decisions concerning not only the sample size but also the sample type. In fact, this can even lead to the use of a non-probability sample.

(vii) ***Sampling procedure:*** The researcher, at last, decides the techniques to be used in selecting the items for the sample. In fact, this technique/procedure stands for the sample design itself. Apparently, such a design should be selected, which for a provided sample size and cost, has a smaller sampling error.

Representative Sample

Representative Sample is a subset of a statistical population that accurately reflects the members of the entire population. A representative sample should be an unbiased indication of what the population is like. In a classroom of 30 students in which half the students are male and half are female, a representative sample might include six students: three males and three females.

When a sample is not representative, the result is known as a sampling error. Using the classroom example again, a sample that included six students, all of whom were male, would not be a representative sample. Whatever conclusions were drawn from studying the six male students would not be likely to translate to the entire group since no female students were studied.

Sampling Designs or Techniques of Sampling

The various types of sampling design method are as follows:

A. Probability Sampling

B. Non-probability Sampling

A. Probability Sampling

A probability sampling method is any method of sampling that utilizes some form of random selection. In order to have a random selection method, you must set up some process or procedure that assures that the different units in your population have equal probabilities of being chosen. Humans have long practiced various forms of random selection, such as picking a name out of a hat or choosing the short straw. These days, we tend to use computers as the mechanism for generating random numbers as the basis for random selection.

1. Simple random sampling
2. Systematic random sampling
3. Stratified random sampling
4. Cluster sampling

1. Simple Random Sampling

Simple random sample (SRS) is a special case of a random sample. A sample is called simple random sample if each unit of the population has an equal chance of being selected for the sample. Whenever a unit is selected for the sample, the units of the population are equally likely to be selected. It must be noted that the probability of selecting the first element is not to be compared with the probability of selecting the second unit. When the first unit is selected, all the units of the population have the equal chance of selection which is 1/N. When the second unit is selected, all the remaining (N - 1) nits of the population have 1/(N-1) chance of selection.

Selection of Sample Random Sample

A simple random sample is usually selected by without replacement. The following methods are used for the selection of a simple random sample:

i) ***Lottery Method:*** This is an old classical method but it is a powerful technique and modern methods of selection are very close to this method. All the units of the population are numbered from1 to N. This is called sampling frame. These numbers are written on the small slips of paper or the small round metallic balls. The paper slips or the metallic balls should be of the same size otherwise the selected sample will not be truly random. The slips or the balls are thoroughly mixed and a slip or ball is picked up. Again the population of slips is mixed and the next unit is selected. In this manner, the number of slips equal to the sample size n is selected. The units of the population which appear on the selected slips make the simple random

Notes

sample. This method of selection is commonly used when size of the population is small. For a large population there is a big heap of paper slips and it is difficult to mix the slips properly

ii) ***Using a Random Number Table:*** All the units of the population are numbered from 1 to N or from 0 toN-1. We consult the random number table to take a simple random sample. Suppose the size of the population is 80 and we have to select a random sample of 8 units. The units of the population are numbered from 01 to 80. We read two-digit numbers from the table of random numbers. We can take a start from any columns or rows of the table. Let us consult random number table given in this content. Two-digit numbers are taken from the table. Any number above 80 will be ignored and if any number is repeated, we shall not record it if sampling is done without replacement. Let us read the first two columns of the table. The random number from the table is 10, 37, 08, 12, 66, 31, 63 and 73. The two numbers 99 and 85 have not been recorded because the population does not contain these numbers. The units of the population whose numbers have been selected constitute the simple random sample. Let us suppose that the size of the population is 100. If the units are numbered from 001 to 100, we shall have to read 3-digit random numbers. From the first 3 columns of the random number table, the random numbers are 100, 375, 084, 990 and 128 and so on. We find that most of the numbers are above 100 and we are wasting our time while reading the table. We can avoid it by numbering the units of the population from 00 to 99. In this way, we shall read 2-digit numbers from the table. Thus if N is 100, 1000 or 10000, the numbering is done from 00 to 99, 000 to 999 or 0000 to 9999.

iii) ***Using the Computer:*** The facility of selecting a simple random sample is available on the computers. The computer is used for selecting a sample of prize-bond winners, a sample of Hajj applicants, and a sample of applicants for residential plots and for various other purposes.

2. Systematic Random Sampling

Systematic sampling is a random sampling technique which is frequently chosen by researchers for its simplicity and its periodic quality. In systematic random sampling, the researcher first randomly picks the first item or subject from the population. Then, the researcher will select each n th subject from the list.

Systematic sampling is a statistical method involving the selection of elements from an ordered sampling frame. The most common form of systematic sampling is an equal-probability method, in which every kth element in the frame is selected, where k, the sampling interval (sometimes known as the skip), is calculated as: K = N/n, where n is the sample size, and N is the population size. This is one of the methods that have been used.

Using this procedure each element in the population has a known and equal probability of selection. This makes systematic sampling functionally similar to simple random sampling. It is however, much more efficient if variance within systematic sample is more than variance of population. The researcher must ensure that the chosen sampling interval does not hide a pattern. Any pattern would threaten randomness. A random starting point must also be selected.

Systematic sampling is to be applied only if the given population is logically homogeneous, because systematic sample units are uniformly distributed over the population. Example: Suppose a supermarket wants to study buying habits of their customers, then using systematic sampling they can choose every 10th or 15th customer entering the supermarket and conduct the study on this sample.

This is random sampling with a system. From the sampling frame, a starting point is chosen at random, and choices thereafter are at regular intervals. For example, suppose you want to sample 8 houses from a street of 120 houses. 120/8=15, so every 15th house is chosen after a random starting point between 1 and 15. If the random starting point is 11, then the houses selected are 11, 26, 41, 56, 71, 86, 101, and 116.

If, as more frequently, the population is not evenly divisible (suppose you want to sample 8 houses out of 125, where 125/8=15.625), should you take every 15th house or every 16th house? If you take every 16th house, 8*16=128, so there is a risk that the last house chosen does not exist. On the other hand, if you take every 15th house, 8*15=120, so the last five houses will never be selected. The random starting point should instead be selected as a non-integer between 0 and 15.625 (inclusive on one endpoint only) to ensure that every house has equal chance of being selected; the interval should now be non-integral (15.625); and each non-integer selected should be rounded up to the next integer. If the random starting point is 3.6, then the houses selected are 4, 19, 35, 51, 66, 82, 98, and 113, where there are 3 cyclic intervals of 15 and 5 intervals of 16.

To illustrate the danger of systematic skip concealing a pattern, suppose we were to sample a planned neighborhood where each street has ten houses on each block. This places houses #1, 10, 11, 20, 21, 30... on block corners; corner blocks may be less valuable, since more of their area is taken up by street front etc. that is unavailable for building purposes. If we then sample every 10th household, our sample will either be made up only of corner houses (if we start at 1 or 10) or have no corner houses (any other start); either way, it will not be representative.

Systematic sampling may also be used with non-equal selection probabilities. In this case, rather than simply counting through elements of the population and selecting every kth unit, we allocate each element a space along a number line according to its selection probability. We then generate a random start from a uniform distribution between 0 and 1, and move along the number line in steps of 1.

Example: We have a population of 5 units (A to E). We want to give unit A a 20% probability of selection, unit B a 40% probability, and so on up to unit E (100%). Assuming we maintain alphabetical order, we allocate each unit to the following interval: qwewqeqe qw eqw e qw e qw e wq e qwe

The process of obtaining the systematic sample is much like an arithmetic progression.

i) ***Starting number:*** The researcher selects an integer that must be less than the total number of individuals in the population. This integer will correspond to the first subject.

ii) ***Interval:*** The researcher picks another integer which will serve as the constant difference between any two consecutive numbers in the progression.

The integer is typically selected so that the researcher obtains the correct sample size.

For example, the researcher has a population total of 100 individuals and need 12 subjects. He first picks his starting number, 5. Then the researcher picks his interval, 8. The members of his sample will be individuals 5, 13, 21, 29, 37, 45, 53, 61, 69, 77, 85, 97.

Other researchers use a modified systematic random sampling technique wherein they first identify the needed sample size. Then, they divide the total number of the population with the sample size to obtain the sampling fraction. The sampling fraction is then used as the constant difference between subjects.

Advantages of Systematic Sampling

i) The main advantage of using systematic sampling over simple random sampling is its simplicity.

ii) It allows the researcher to add a degree of system or process into the random selection of subjects.

iii) Another advantage of systematic random sampling over simple random sampling is the assurance that the population will be evenly sampled.

iv) There exists a chance in simple random sampling that allows a clustered selection of subjects.

v) This is systematically eliminated in systematic sampling.

Disadvantage of Systematic Sampling

i) The process of selection can interact with a hidden periodic trait within the population.

ii) If the sampling technique coincides with the periodicity of the trait, the sampling technique will no longer be random and representativeness of the sample is compromised.

3. Stratified Random Sampling

Stratified Random Sampling is a method of sampling that involves the division of a population into smaller groups known as strata. In stratified random sampling, the strata are formed based on members' shared attributes or characteristics. A random sample from each stratum is taken in a number proportional to the stratum's size when compared to the population.

These subsets of the strata are then pooled to form a random sample. The main advantage with stratified sampling is how it captures key population characteristics in the sample. Similar to a weighted average, this method of sampling produces characteristics in the sample that are proportional to the overall population. Stratified sampling works well for populations with a variety of attributes, but is otherwise ineffective, as subgroups cannot be formed.

Stratified random sampling is a sampling method that has the following properties:

i) The population consists of N elements.

ii) The population is divided into H groups, called strata.

iii) Each element of the population can be assigned to one, and only one, stratum.

iv) The number of observations within each stratum Nh is known, and N = N1 + N2 + N3 + ... + NH-1 + NH.

v) The researcher obtains a probability sample from each stratum.

Advantages of Stratified Random Sampling

i) Stratified sampling offers several advantages over simple random sampling.

ii) A stratified sample can provide greater precision than a simple random sample of the same size.

iii) Because it provides greater precision, a stratified sample often requires a smaller sample, which saves money.

iv) A stratified sample can guard against an "unrepresentative" sample (e.g., an all-male sample from a mixed-gender population).

v) We can ensure that we obtain sufficient sample points to support a separate analysis of any subgroup.

Disadvantages of Stratified Random Sampling

i) The main disadvantage of a stratified sample is that it may require more administrative effort than a simple random sample.

ii) Stratified sampling is not useful when the population cannot be exhaustively partitioned into disjoint subgroups.

iii) It would be a misapplication of the technique to make subgroups' sample sizes proportional to the amount of data available from the subgroups, rather than scaling sample sizes to subgroup sizes.

iv) Data representing each subgroup are taken to be of equal importance if suspected variation among them warrants stratified sampling. If, on the other hand, the very variances vary so much, among subgroups, that the data need to be stratified by variance, there is no way to make the subgroup sample sizes proportional to the subgroups' sizes within the total population.

Proportionate Versus Disproportionate Stratification

All stratified sampling designs fall into one of two categories, each of which has strengths and weaknesses as described below:

i) Proportionate stratification: With proportionate stratification, the sample size of each stratum is proportionate to the population size of the stratum. This means that each stratum has the same sampling fraction. Proportionate stratification provides equal or better precision than a simple random sample of the same size. Gains in precision are greatest when values within strata are homogeneous.

ii) Disproportionate stratification: With disproportionate stratification, the sampling fraction may vary from one stratum to the next. The precision of the design may be very good or very poor, depending on how sample points are allocated to strata. The way to maximize precision through disproportionate stratification is discussed in a subsequent lesson. If variances differ across strata, disproportionate stratification can provide better precision than proportionate stratification, when sample points are correctly allocated to strata. With disproportionate stratification, the researcher can maximize precision for a single important survey measure. However, gains in precision may not accrue to other survey measures.

4. Multi Stages or Cluster Sampling

Cluster Sampling is a sampling technique used when "natural" groupings are marked in a statistical population. It is often used in marketing research. In this technique, the total population is divided into these groups (or clusters) and a sample of the groups is selected. Then the required information is collected from the elements within each selected group. This may be done for every element in these groups or a subsample of elements may be selected within each of these groups. A common motivation for cluster sampling is to reduce the average cost per interview. Given a fixed budget, this can allow an increased sample size. Assuming a fixed sample size, the technique gives more accurate results when most of the variation in the population is within the groups, not between them.

Cluster elements

Elements within a cluster should ideally be as heterogeneous as possible, but there should be homogeneity between cluster means. Each cluster should be a small scale representation of the total population. The clusters should be mutually exclusive and collectively exhaustive. A random sampling technique is then used on any relevant clusters to choose which clusters to include in the study. In single-stage cluster sampling, all the elements from each of the selected clusters are used. In two-stage cluster sampling, a random sampling technique is applied to the elements from each of the selected clusters.

The main difference between cluster sampling and stratified sampling is that in cluster sampling the cluster is treated as the sampling unit so analysis is done on a population of clusters (at least in the first stage). In stratified sampling, the analysis is done on elements within strata. In stratified sampling, a random sample is drawn from each of the strata, whereas in cluster sampling only the selected clusters are studied. The main objective of cluster sampling is to reduce costs by increasing sampling efficiency. This contrasts with stratified sampling where the main objective is to increase precision. There also exists multistage sampling, where more than two steps are taken in selecting clusters from clusters.

Aspects of Cluster Sampling

One version of cluster sampling is area sampling or geographical cluster sampling. Clusters consist of geographical areas. Because a geographically dispersed population can be expensive to survey, greater economy than simple random sampling can be achieved by treating several respondents within a local area as a cluster. It is usually necessary to increase the total sample size to achieve equivalent precision in the estimators, but cost savings may make that feasible.

In some situations, cluster analysis is only appropriate when the clusters are approximately the same size. This can be achieved by combining clusters. If this is not possible, probability proportionate to size sampling is used. In this method, the probability of selecting any cluster varies with the size of the cluster, giving larger clusters a greater probability of selection and smaller clusters a lower probability. However, if clusters are selected with probability proportionate to size, the same number of interviews should be carried out in each sampled cluster so that each unit sampled has the same probability of selection.

B. Non-Probability Sampling

Non-probability sampling represents a group of sampling techniques that help researchers to select units from a population that they are interested in studying. Collectively, these units form the sample that the researcher studies. A core characteristic of non-probability sampling techniques is that samples are selected based on the subjective judgement of the researcher, rather than random selection (i.e. probabilistic methods), which is the cornerstone of probability sampling techniques. At the same time as some researchers may view non-probability sampling techniques as inferior to probability sampling techniques, there are strong theoretical and practical reasons for their use.

1. Convenience sampling
2. Judgment sampling

Notes

3. Quota sampling

4. Snowball sampling

1. Convenience Sampling

Convenience sampling is a non-probability sampling technique where subjects are selected because of their convenient accessibility and proximity to the researcher.

The subjects are selected just because they are easiest to recruit for the study and the researcher did not consider selecting subjects that are representative of the entire population.

In all forms of research, it would be ideal to test the entire population, but in most cases, the population is just too large that it is impossible to include every individual. This is the reason why most researchers rely on sampling techniques like convenience sampling, the most common of all sampling techniques. Many researchers prefer this sampling technique because it is fast, inexpensive, easy and the subjects are readily available.

Examples

One of the most common examples of convenience sampling is using student volunteers as subjects for the research. Another example is using subjects that are selected from a clinic, a class or an institution that is easily accessible to the researcher. A more concrete example is choosing five people from a class or choosing the first five names from the list of patients.

In these examples, the researcher inadvertently excludes a great proportion of the population. A convenience sample is either a collection of subjects that are accessible or a self selection of individuals willing to participate which is exemplified by your volunteers.

Uses

Researchers use convenience sampling not just because it is easy to use, but because it also has other research advantages.

In pilot studies, convenience sample is usually used because it allows the researcher to obtain basic data and trends regarding his study without the complications of using a randomized sample.

This sampling technique is also useful in documenting that a particular quality of a substance or phenomenon occurs within a given sample. Such studies are also very useful for detecting relationships among different phenomena.

Criticisms

The most obvious criticism about convenience sampling is sampling bias and that the sample is not representative of the entire population. This may be the biggest disadvantage when using a convenience sample because it leads to more problems and criticisms.

Systematic bias stems from sampling bias. This refers to a constant difference between the results from the sample and the theoretical results from the entire population. It is not rare that the results from a study that uses a convenience sample differ significantly with the results from the entire population. A consequence of having systematic bias is obtaining skewed results.

Another significant criticism about using a convenience sample is the limitation in generalization and inference making about the entire population. Since the sample is not representative of the population, the results of the study cannot speak for the entire population. This results to a low external validity of the study.

2. Judgment Sampling

Judgement sampling involves the choice of subjects who are most advantageously placed or in the best position to provide the information required. They could reasonably be expected to have expert knowledge by virtue of having gone through the experience and processes themselves and might perhaps be able to provide good data or information to the researcher. Thus the judgement sampling design is used when a limited number or category of people have the information that is sought. In such cases any type of probability sampling across a cross-section of the entire population is purposeless and not useful.

Judgement sampling may curtail the generalizability of the finding due to the fact that we are using a sample of experts who are conveniently available to us. However it is the only viable sampling method for obtaining the type of information that is required from very specific pockets of people who are very knowledgeable are included in the sample.

Researchers, scientists, or farm managers may be called in when a crop shows a certain growing pattern or when surface differences are observed for a soil. For example, differences may occur in soil color which may be the result of many factors. The researcher judges the color differences: e.g. he may judge a particular shade of color to be typical for a sample at certain sites. Then from these sites, samples are drawn. The accuracy of these samples depends totally on the judgment of the researcher - which may or may not be good. Some persons, in order to include as many extremes as possible, commit the error of over sampling. Probably less desirable is the person who takes the opposite approach, excluding the extremes and ending up with a sample which is not representative. In either case, the judgment of the sampler determines the accuracy of the results.

In certain situations, sample choices based on judgment are accurate enough. For example, if small sites are involved and no estimate of accuracy is needed, judgment sampling might be satisfactory. As the sample site becomes larger, and the selection of representative samples becomes more difficult and time consuming, judgment sampling is inaccurate and other sampling methods must be used.

3. Quota Sampling

Quota sampling is a method for selecting survey participants. In quota sampling, a population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion. For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60. This means that individuals can put a demand on who they want to sample.

This second step makes the technique non-probability sampling. In quota sampling, the selection of the sample is non-random sample and can be unreliable. For example, interviewers might be tempted to interview those people in the street who look most helpful, or may choose to use accidental sampling to question those closest to them, for time-keeping sake. The problem is that these samples may be biased because not everyone gets a chance of selection. This non-random element is a source of uncertainty about the nature of the actual sample and quota versus probability has been a matter of controversy for many years.

Quota sampling is useful when time is limited, a sampling frame is not available, the research budget is very tight or when detailed accuracy is not important. Subsets are chosen and then either convenience or judgment sampling is used to choose people from each subset. The researcher decides how many of each category is selected.

Notes

Quota sampling is the non probability version of stratified sampling. In stratified sampling, subsets of the population are created so that each subset has a common characteristic, such as gender. Random sampling chooses a number of subjects from each subset with, unlike a quota sample, each potential subject having a known probability of being selected.

4. Snowball Sampling

Snowball Sampling is a method used to obtain research and knowledge, from extended associations, through previous acquaintances. "Snowball sampling uses recommendations to find people with the specific range of skills that has been determined as being useful." An individual or a group receives information from different places through a mutual intermediary. This is referred to metaphorically as snowball sampling because as more relationships are built through mutual association, more connections can be made through those new relationships and a plethora of information can be shared and collected, much like a snowball that rolls and increases in size as it collects more snow. Snowball sampling is a useful tool for building networks and increasing the number of participants. However, the success of this technique depends greatly on the initial contacts and connections made. Thus it is important to correlate with those that are popular and honorable to create more opportunities to grow, but also to create a credible and dependable reputation.

Method of Snowball Sampling

i) Draft up a participation program.

ii) Approach stakeholders and ask for contacts.

iii) Gain contacts and ask them to participate.

iv) Community issues groups may emerge that can be included in the participation program.

v) Continue the snowballing with contacts to gain more stakeholders if necessary.

vi) Ensure a diversity of contacts by widening the profile of persons involved in the snowballing exercise.

Uses of Snowball Sampling

There are many reasons why an individual may want to use snowball sampling across any industry, research, job, etc. Specific to business and marketing, however, snowball sampling can be used to things such as identify experts in a certain field, product, manufacturing processes, customer relation methods, etc. 3M did this when they were trying to identify experts in different fields of work in order to become the lead user for surgical drapes, the small plastic covering that is applied at the incision site of a surgery. To do this, 3M called in specialist from all fields that related to how a surgical drape could be applied to the body. For example, they called in a veterinarian, who specializes with surgeries on creatures with a lot of hair, and a Broadway make-up artist who specialized in applying foreign materials to human skin in a non-irritating manner. In order to successfully identify these people, 3m used snowball sampling. They called "experts" that they had contacts and after gathering information, asked them to suggest another expert that they may know who could offer more information. They repeated this process until they were satisfied with their experts and felt that they had found the most knowledgeable individuals in a specific field. Thus, snowball sampling can be used to gather expert information.

Notes

Advantages of Snowball Sampling

There are many different kinds of sampling, each with their own advantages and disadvantages. Snowball sampling has a lot of advantages as opposed to other sampling methods. It is possible for the surveyors to include people in the survey that they would not have known. It is also very good for locating people of a specific population if they are difficult to locate. The advantage of this is that you can quickly find people who are experts in their fields, because people often know someone who is better at their job than them. This leads to only having the most well known experts for your sampling group, and also can help you find lead users more simply.

Disadvantages of Snowball Sampling

Snowball sampling is inexact, and can produce varied and inaccurate results. The method is heavily reliant on the skill of the individual conducting the actual sampling, and that individual's ability to vertically network and find an appropriate sample. To be successful requires previous contacts within the target areas, and the ability to keep the information flow going throughout the target group. Identifying the appropriate person to conduct the sampling, as well as locating the correct targets is a time consuming process which renders the benefits only slightly outweighing the costs. Another disadvantage of snowball sampling is the lack of definite knowledge as to whether or not the sample is an accurate reading of the target population. By targeting only a few select people, it is not always indicative of the actual trends within the result group. To help mitigate these risks, it is important to not rely on any one single method of sampling to gather data about a target sector. In order to most accurately obtain information, a company must do everything it possibly can to ensure that the sampling is controlled. Also, it is imperative that the correct personnel are used to execute the actual sampling, because one missed opportunity could skew the results.

Examples of Snowball Sampling

i) ***Positive:*** When attempting to gather information about a particular topic, and a limited number of participants or test subjects are available, snowball sampling would increase the efficiency of the study. It is cost efficient to use this method because locating respondents to acquire information may take time and finances. In order to acquire more participants, snowball sampling relies on referrals and by word of mouth. The more effort that goes into the preliminary rounds of the study, contacting people and spreading the word of the main goals of the study, etc., will pay dividends in the long run due to the increase in size of the overall study sample. Bias plays a major role within every study, and increasing the amount of participants will only help the accuracy of the information. A positive example of snowball sampling would be if a researcher is having trouble reaching individuals within its target market. For instance, if someone was attempting to do a research sample involving football players because they were trying to sell a customized piece of equipment, they would need to meet with some players to get their point of view about the product. If the researcher only knew a few players, they would have to go out and personally introduce themselves to other players to expand their study. They could contact the player or players that they already know and ask them to refer them to a few others. They could offer a small incentive to quicken the process, and maybe this perk would attract other players to participate in the study. They could also gain access to the roster from the school's website and try and contact players via email or telephone. The more relationships they create, the more information they will receive. If they put the effort in to meet with a few kids from a few different teams, they would have the opportunity to be referred to by every kid on the team. The snowball effect would occur as more and more referrals are acquired. If I was attempting this study I would try and meet with the captain or seniors on the team and offer incentives to them. If you attract the "best" players to be involved within your study, it is a safe assumption to say that others will follow. Another example would be drug dealers. Although the topic is inappropriate, it is a good example to explain the essence of snowball sampling. As the dealer brings in product, they need to find customers to move their product. Everyone they sell their products to can refer them to other potential customers which will increase their business and continue to make them revenue.

ii) ***Negative:*** Snowball sampling can be a strenuous process at times if not planned out properly. A number of issues can arise when using snowball sampling as a method for gathering information. For instance, if a marketing team is trying to gather information that will result in a new, innovative product that can spur the business's success and develop a competitive advantage. As the marketing team contacts people throughout their respective customer base and other important individuals in their industry, a number of challenges or barriers may develop. Certain individuals may become resistant and not want to provide referrals based upon the people they know who may possibly help the firm's efforts. If this information cannot be obtained, the targeted individuals the team is seeking may not be complete and vital ideas generated from such individuals will not be taken into consideration. Resistance to providing referrals will cause the team to waste time having to research new contacts to get in touch with. This inability to gather appropriate information from select participants and loss of time will possibly jeopardize the opportunity to develop an innovative product in time and allow competitors to take advantage of this setback. This negative example of snowball sampling illustrates some of the difficulties associated with utilizing snowball sampling as a method for gathering information from select individuals.

Definition of a Sampling Distribution

The sampling distribution of a statistic is the distribution of all possible values of the statistic, computed from samples of the same size randomly drawn from the same population. When sampling a discrete, finite population, a sampling distribution can be constructed. Note that this construction is difficult with a large population and impossible with an infinite population.

Construction of Sampling Distributions

1. From a population of size N, randomly draw all possible samples of size n.
2. Compute the statistic of interest for each sample.
3. Create a frequency distribution of the statistic.

Importance of Sampling Distribution in Research Methodology

Some important sampling distributions, which are commonly used, are:

1. sampling distribution of mean;
2. sampling distribution of proportion;
3. student's 't' distribution;

4. F distribution; and
5. Chi-square distribution.

A brief mention of each one of this sampling distribution will be helpful.

1. ***Sampling distribution of mean:*** Sampling distribution of mean refers to the probability distribution of all the possible means of random samples of a given size that we take from a population. If samples are taken from a normal population, N dm,s p i, the sampling distribution of mean would also be normal with mean mx = m and standard deviation = s p n, where m is the mean of the population, s p is the standard deviation of the population and n means the number of items in a sample. But when sampling is from a population which is not normal (may be positively or negatively skewed), even then, as per the central limit theorem, the sampling distribution of mean tends quite closer to the normal distribution, provided the number of sample items is large i.e., more than 30. In case we want to reduce the sampling distribution of mean to unit normal distribution i.e., N (0,1), we can write the normal variate Formula for the sampling distribution of mean. This characteristic of the sampling distribution of mean is very useful in several decision situations for accepting or rejection of hypotheses.

2. ***Sampling distribution of proportion:*** Like sampling distribution of mean, we can as well have a sampling distribution of proportion. This happens in case of statistics of attributes. Assume that we have worked out the proportion of defective parts in large number of samples, each with say 100 items, that have been taken from an infinite population and plot a probability distribution of the said proportions, we obtain what is known as the sampling distribution of the said proportions, we obtain what is known as the sampling distribution of proportion. Usually the statistics of attributes correspond to the conditions of a binomial distribution that tends to become normal distribution as n becomes larger and larger. If p represents the proportion of defectives i.e., of successes and q the proportion of non-defectives i.e., of failures (or q = 1 – p) and if p is treated as a random variable, then the sampling distribution of proportion of successes has a mean = p with standard deviation = Formula where n is the sample size. Presuming the binomial distribution approximating the normal distribution for large n, the normal variate of the sampling distribution of proportion z = Formula where $p (pronounced as p-hat) is the sample proportion of successes, can be used for testing of hypotheses.

3. ***Student's t-distribution:*** When population standard deviation Formula is not known and the sample is of a small size bi.e., n < 30 g , we use t distribution for the sampling distribution of mean and workout t variable as:

$$t=\left(\bar{X}-\mu\right)\Big/\left(\sigma_s/\sqrt{n}\right)$$

where $\sigma_s=\sqrt{\dfrac{\Sigma\left(X_i-\bar{X}\right)^2}{n}-1}$

i.e., the sample standard deviation. t-distribution is also symmetrical and is very close to the distribution of standard normal variate, z, except for small values of n. The variable t differs from z in the sense that we use sample standard deviation s s b g in the calculation of t, whereas we use standard deviation of population s p d i in the calculation of z. There is a different t distribution for every possible sample size i.e., for different degrees of freedom. The degrees of freedom for a sample of size n is n – 1. As the sample size gets larger, the shape of the t distribution becomes apporximately equal to the normal distribution. In fact for sample sizes of more than 30, the t distribution is so close to the normal distribution that we can use the normal to approximate the t-distribution. But when n is small, the t-distribution is far from normal but when n a , t-distribution is identical with normal distribution. The t-distribution tables are available which give the critical values of t for different degrees of freedom at various levels of significance. The table value of t for given degrees of freedom at a certain level of significance is compared with the calculated value of t from the sample data, and if the latter is either equal to or exceeds, we infer that the null hypothesis cannot be accepted.

4. ***F distribution:*** F ratio is computed in a way that the larger variance is always in the numerator. Tables have been prepared for F distribution that give critical values of F for various values of degrees of freedom for larger as well as smaller variances. The calculated value of F from the sample data is compared with the corresponding table value of F and if the former is equal to or exceeds the latter, then we infer that the null hypothesis of the variances being equal cannot be accepted. We shall make use of the F ratio in the context of hypothesis testing and also in the context of ANOVA technique.

5. ***Chi-square Formula distribution:*** Chi-square distribution is encountered when we deal with collections of values that involve adding up squares. Variances of samples require us to add a collection of squared quantities and thus have distributions that are related to chi-square distribution. If we take each one of a collection of sample variances, divide them by the known population variance and multiply these quotients by (n – 1), where n means the number of items in the sample, we shall obtain a chi-square distribution. Thus, Formula would have the same distribution as chi-square distribution with (n – 1) degrees of freedom. Chi-square distribution is not symmetrical and all the values are positive. One must know the degrees of freedom for using chi-square distribution. This distribution may also be used for judging the significance of difference between observed and expected frequencies and also as a test of goodness of fit. The generalised shape of c 2 distribution depends upon the d.f. and the c 2 value is worked out as under:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Tables are there that give the value of c 2 for given d.f. which may be used with calculated value of c 2 for relevant d.f. at a desired level of significance for testing hypotheses.

Types of Sampling Distributions

A) Distribution of the sample mean

The Sampling Distribution of the Sample Mean. If repeated random samples of a given size n are taken from a population of values for a quantitative variable, where the population mean is μ (mu) and the population standard deviation is s (sigma) then the mean of all sample means (x-bars) is population mean μ (mu).

B) Distribution of the difference between two means

Statistical analyses are very often concerned with the difference between means. A typical example is an experiment designed to compare the mean of a control group with the mean of an experimental group. Inferential statistics used in the analysis of this type of experiment depend on the sampling distribution of the difference between means.

The sampling distribution of the difference between means can be thought of as the distribution that would result if we repeated the following three steps over and over again: (1) sample n1 scores from Population 1 and n2 scores from Population 2, (2) compute the means of the two samples (M1 and M2), and (3) compute the difference between means, M1 - M2. The distribution of the differences between means is the sampling distribution of the difference between means.

As you might expect, the mean of the sampling distribution of the difference between means is:

which says that the mean of the distribution of differences between sample means is equal to the difference between population means. For example, say that the mean test score of all 12-year-olds in a population is 34 and the mean of 10-year-olds is 25. If numerous samples were taken from each age group and the mean difference computed each time, the mean of these numerous differences between sample means would be 34 - 25 = 9.

From the variance sum law, we know that:

which says that the variance of the sampling distribution of the difference between means is equal to the variance of the sampling distribution of the mean for Population 1 plus the variance of the sampling distribution of the mean for Population 2. Recall the formula for the variance of the sampling distribution of the mean:

Since we have two populations and two samples sizes, we need to distinguish between the two variances and sample sizes. We do this by using the subscripts 1 and 2. Using this convention, we can write the formula for the variance of the sampling distribution of the difference between means as:

Since the standard error of a sampling distribution is the standard deviation of the sampling distribution, the standard error of the difference between means is:

Just to review the notation, the symbol on the left contains a sigma (s), which means it is a standard deviation. The subscripts M1 - M2 indicate that it is the standard deviation of the sampling distribution of M1 - M2.

Now let's look at an application of this formula. Assume there are two species of green beings on Mars. The mean height of Species 1 is 32 while the mean height of Species 2 is 22. The variances of the two species are 60 and 70, respectively and the heights of both species are normally distributed. You randomly sample 10 members of Species 1 and 14 members of Species 2. What is the probability that the mean of the 10 members of Species 1 will exceed the mean of the 14 members of Species 2 by 5 or more? Without doing any calculations, you probably know that the probability is pretty high since the difference in population means is 10. But what exactly is the probability?

First, let's determine the sampling distribution of the difference between means. Using the formulas above, the mean is

The standard error is:

C) Distribution of the sample proportion

The Sampling Distribution of the Sample Proportion. If repeated random samples of a given size n are taken from a population of values for a categorical variable, where the proportion in the category of interest is p, then the mean of all sample proportions (p-hat) is the population proportion (p).

If repeated random samples of a given size n are taken from a population of values for a categorical variable, where the proportion in the category of interest is p, then the mean of all sample proportions (p-hat) is the population proportion (p).

As for the spread of all sample proportions, theory dictates the behavior much more precisely than saying that there is less spread for larger samples. In fact, the standard deviation of all sample proportions is directly related to the sample size, n as indicated below.

Since the sample size n appears in the denominator of the square root, the standard deviation does decrease as sample size increases. Finally, the shape of the distribution of p-hat will be approximately normal as long as the sample size n is large enough. The convention is to require both np and n(1 – p) to be at least 10.

We can summarize all of the above by the following:

Let's apply this result to our example and see how it compares with our simulation.

In our example, n = 25 (sample size) and p = 0.6. Note that np = 15 = 10 and n(1 – p) = 10 = 10. Therefore we can conclude that p-hat is approximately a normal distribution with mean p = 0.6 and standard deviation

(which is very close to what we saw in our simulation).

D) Distribution of the difference between two proportions

1.21 Z-STATISTICS

A Z-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Because of the central limit theorem, many test statistics are approximately normally distributed for large samples. For each significance level, the Z-test has a single critical value (for example, 1.96 for 5% two tailed) which makes it more convenient than the Student's t-test which has separate critical values for each sample size. Therefore, many statistical tests can be conveniently performed as approximate Z-tests if the sample size is large or the population variance known. If the population variance is unknown (and therefore has to be estimated from the sample itself) and the sample size is not large, the Student t-test may be more appropriate.

General form

The most general way to obtain a Z-test is to define a numerical test statistic that can be calculated from a collection of data, such that the sampling distribution of the statistic is approximately normal under the null hypothesis. Statistics that are averages of approximately independent data values are generally well-approximated by a normal distribution. An example of a statistic that would not be well-approximated by a normal distribution would be an extreme value such as the sample maximum.

If T is a statistic that is approximately normally distributed under the null hypothesis, the next step in performing a Z-test is to determine the expected value 0 of T under the null hypothesis, and then obtain an estimate s of the standard deviation of T. Then calculate the standard score Z = (T- 0) / s, from which one-tailed and two-tailed p-values can be calculated as Φ (- |Z|) and 2Φ (- |Z|), respectively, where F is the standard normal cumulative distribution function.

Use in Location Testing

The term Z-test is often used to refer specifically to the one-sample location test comparing the mean of a set of measurements to a given constant. If the observed data X1, ..., Xn are (i) uncorrelated, (ii) have a common mean m and (iii) have a common variance s^2, then the sample average X has mean ì and variance s^2 / n. If our null hypothesis is that the mean value of the population is a given number m 0, it can use X - m0 as a test-statistic, rejecting the null hypothesis if X - m0 is large.

To calculate the standardized statistic Z = (X - m0) / s, we need to either know or have an approximate value for s^2, from which we can calculate $S^2 = s^2$ / n. In some applications, s^2 is known, but this is uncommon. If the sample size is moderate or large, can substitute the sample variance for s^2, giving a plug-in test. The resulting test will not be an exact Z-test since the uncertainty in the sample variance is not accounted for however, it will be a good approximation unless the sample size is small. A t-test can be used to account for the uncertainty in the sample variance when the sample size is small and the data are exactly normal. There is no universal constant at which the sample size is generally considered large enough to justify use of the plug-in test. Typical rules of thumb range from 20 to 50 samples. For larger sample sizes, the t-test procedure gives almost identical p-values as the Z-test procedure.

Conditions

For the Z-test to be applicable, certain conditions must be met:

i) Nuisance parameters should be known, or estimated with high accuracy (an example of a nuisance parameter would be the standard deviation in a one-sample location test). Z-tests focus on a single parameter and treat all other unknown parameters as being fixed at their true values. In practice, due to Slutsky's theorem, "plugging in" consistent estimates of nuisance parameters can be justified. However if the sample size is not large enough for these estimates to be reasonably accurate, the Z-test may not perform well.

ii) The test statistic should follow a normal distribution. Generally, one appeals to the central limit theorem to justify assuming that a test statistic varies normally. There is a great deal of statistical research on the question of when a test statistic varies approximately normally. If the variation of the test statistic is strongly non-normal, a Z-test should not be used.

iii) If estimates of nuisance parameters are plugged in as discussed above, it is important to use estimates appropriate for the way the data were sampled. In the special case of Z-tests for the one or two sample location problem, the usual sample standard deviation is only appropriate if the data were collected as an independent sample.

iv) In some situations, it is possible to devise a test that properly accounts for the variation in plug-in estimates of nuisance parameters. In the case of one and two sample location problems, a t-test does this.

Z-tests other than location tests

Location tests are the most familiar t-tests. Another class of Z-tests arises in maximum likelihood estimation of the parameters in a parametric statistical model. Maximum likelihood estimates are approximately normal under certain conditions, and their asymptotic variance can be calculated in terms of the Fisher information. The maximum likelihood estimate divided by its standard error can be used as a test statistic for the null hypothesis that the population value of the parameter equals zero.

When using a Z-test for maximum likelihood estimates, it is important to be aware that the normal approximation may be poor if the sample size is not sufficiently large. Although there is no simple, universal rule stating how large the sample size must be to use a Z-test, simulation can give a good idea as to whether a Z-test is appropriate in a given situation.

Z-tests are employed whenever it can be argued that a test statistic follows a normal distribution under the null hypothesis of interest. Many non-parametric test statistics, such as U statistics, are approximately normal for large enough sample sizes, and hence are often performed as Z-tests.

1.22 CHI-SQUARE

χ^2 test is a test that uses the chi-square statistic to test the fit between a theoretical frequency distribution and a frequency distribution of observed data for which each observation may fall into one of several classes.

Conditions of Chi-square (χ^2) Test

A chi-square (χ^2) test can be used when the data satisfies four conditions:

i) There must be two observed sets of data or one observed set of data and one expected set of data (generally, there are n-rows and c-columns of data)

ii) The two sets of data must be based on the same sample size.

iii) Each cell in the data contains the observed or expected count of five or large?

iv) The different cells in a row of column must have categorical variables (male, female or younger than 25 years of age, 25 year of age, older than 40 years of age etc.)

Assumptions of Chi-square Test

The chi-squared test, when used with the standard approximation that a chi-squared distribution is applicable, has the following assumptions:

i) ***Simple random sample:*** The sample data is a random sampling from a fixed distribution or population where each member of the population has an equal probability of selection. Variants of the test have been developed for complex samples, such as where the data is weighted.

ii) ***Sample size (whole table):*** A sample with a sufficiently large size is assumed. If a chi squared test is conducted on a sample with a smaller size, then the chi squared test will yield an inaccurate inference. The researcher, by using chi squared test on small samples, might end up committing a Type II error.

iii) ***Expected cell count:*** Adequate expected cell counts. Some require 5 or more, and others require 10 or more. A common rule is 5 or more in all cells of a 2-by-2 table and 5 or more in 80% of cells in larger tables, but no cells with zero expected count. When this assumption is not met, Yates's correction is applied.

iv) ***Independence:*** The observations are always assumed to be independent of each other. This means chi-squared cannot be used to test correlated data (like matched pairs or panel data). In those cases you might want to turn to McNamara's test.

Application areas of Chi-square test

The χ^2 distribution typically looks like a normal distribution, which is skewed to the right with a long tail to the right. It is a continuous distribution with only positive values. It has following applications:

i) To test whether the sample differences among various sample proportions are significant or can they be attributed to chance.

ii) To test the independence of two variables in a contingency table.

iii) To use it as a test of goodness of fit.

Degrees of Freedom (d.f)

The degree of freedom, abbreviated as d.f, denotes the extent of independence (freedom) enjoyed by a given set of observed frequencies. Degrees of freedom are usually denoted by the letter 'v' of the Greek alphabet.

Suppose, if we are given a set of 'n' observed frequencies which are subjected to 'k' independent constraints (restrictions). Then

Degrees of Freedom = No. of frequencies – No. of independent constraints v = n – k

Formula of Chi-square text:

$$x^2 = \sum\left(\frac{(O_i - E_i)^2}{E_i}\right)$$

Table value of x^2 for d.f and α

$X^2_{cal} < X^2_{table}$, accept H_0

1.23 T-STATISTICS

A statistical examination of two population means. A two-sample t-test examines whether two samples are different and is commonly used when the variances of two normal distributions are unknown and when an experiment uses a small sample size.

Formula: $$t = \frac{X - \mu}{\frac{S}{\sqrt{N}}}$$

Where, is the sample mean, Ä is a specified value to be tested, s is the sample standard deviation, and n is the size of the sample. Look up the significance level of the z-value in the standard normal table.

When the standard deviation of the sample is substituted for the standard deviation of the population, the statistic does not have a normal distribution; it has what is called the t-distribution. Because there is a different t-distribution for each sample size, it is not practical to list a separate area of the curve table for each one. Instead, critical t-values for common alpha levels (0.10, 0.05, 0.01, and so forth) are usually given in a single table for a range of sample sizes. For very large samples, the t-distribution approximates the standard normal (z) distribution. In practice, it is best to use t-distributions any time the population standard deviation is not known.

Values in the t-table are not actually listed by sample size but by degrees of freedom (df). The number of degrees of freedom for a problem involving the t-distribution for sample size n is simply n – 1 for a one-sample mean problem.

Uses of T Test

Among the most frequently used t-tests are:

i) A one-sample location test of whether the mean of a normally distributed population has a value specified in a null hypothesis.

ii) A two sample location test of the null hypothesis that the means of two normally distributed populations are equal. All such tests are usually called Student's t-tests, though strictly speaking that name should only be used if the variances of the two populations are also assumed to be equal; the form of the test used when this assumption is dropped is sometimes called Welch's t-test. These tests are often referred to as "unpaired" or "independent samples" t-tests, as they are typically applied when the statistical units underlying the two samples being compared are non-overlapping.

iii) A test of the null hypothesis that the difference between two responses measured on the same statistical unit has a mean value of zero. For example, suppose we measure the size of a cancer patient's tumor before and after a treatment. If the treatment is effective, we expect the tumor size for many of the patients to be smaller following the treatment. This is often referred to as the "paired" or "repeated measures" t-test: A test of whether the slope of a regression line differs significantly from 0.

Assumptions

Most *t*-test statistics have the form $\mathrm{T} = \frac{\mathrm{Z}}{\mathrm{S}}$, where *Z* and *s* are functions of the data. Typically, *Z* is designed to be sensitive to the alternative hypothesis (i.e. its magnitude tends to be larger when the alternative hypothesis is true), whereas *s* is a scaling parameter that allows the distribution of *T* to be determined.

As an example, in the one-sample *t*-test $\mathrm{Z} = \frac{\overline{\mathrm{X}}}{\frac{\sigma}{\sqrt{\mathrm{n}}}}$, where $\overline{\mathrm{X}}$ is the sample mean of the data, *n* is the sample size, and ó is the population standard deviation of the data; *S* in the one-sample *t*-test is $\hat{\sigma} / \sigma$, where $\hat{\sigma}$ is the sample standard deviation.

The assumptions underlying a t-test are that:

i) Z follows a standard normal distribution under the null hypothesis

ii) ps2 follows a χ2 distribution with p degrees of freedom under the null hypothesis, where p is a positive constant

iii) Z and S are independent.

Unpaired and paired two-sample t-tests

Two-sample t-tests for a difference in mean can be either unpaired or paired. Paired t-tests are a form of blocking, and have greater power than unpaired tests when the paired units are similar with respect to "noise factors" that are independent of membership in the two groups being compared. In a different context, paired t-tests can be used to reduce the effects of confounding factors in an observational study.

Unpaired

The unpaired or "independent samples" t-test is used when two separate sets of independent and identically distributed samples are obtained, one from each of the two populations being compared. For example, suppose we are evaluating the effect of a medical treatment and we enroll 100 subjects into our study, and then randomize 50 subjects to the treatment group and 50 subjects to the control group. In this case, we have two independent samples and would use the unpaired form of the t-test. The randomization is not essential here if we contacted 100 people by phone and obtained each person's age and gender, and then used a two-sample t-test to see whether the mean ages differ by gender, this would also be an independent samples t-test, even though the data are observational.

Paired

Dependent samples (or "paired") t-tests typically consist of a sample of matched pairs of similar units or one group of units that has been tested twice (a "repeated measures" t-test). A typical example of the repeated measures t-test would be where subjects are tested prior to a treatment, say for high blood pressure, and the same subjects are tested again after treatment with a blood-pressure lowering medication.

A dependent t-test based on a "matched-pairs sample" results from an unpaired sample that is subsequently used to form a paired sample, by using additional variables that were measured along with the variable of interest. The matching is carried out by identifying pairs of values consisting of one observation from each of the two samples, where the pair is similar in terms of other measured variables. This approach is often used in observational studies to reduce or eliminate the effects of confounding factors.

Calculations

Explicit expressions that can be used to carry out various t-tests are given below. In each case, the formula for a test statistic that either exactly follows or closely approximates a t-distribution under the null hypothesis is given. Also, the appropriate degrees of freedom are given in each case. Each of these statistics can be used to carry out either a one-tailed test or a two-tailed test.

Once a t value is determined, a p-value can be found using a table of values from Student's t-distribution. If the calculated p-value is below the threshold chosen for statistical significance (usually the 0.10, the 0.05 or 0.01 level), the null hypothesis is rejected in favor of the alternative hypothesis.

One-sample t*-test*

In testing the null hypothesis that the population means is equal to a specified value μ_0 one uses the statistic.

$$t = \frac{\overline{x} - \mu o}{s/\sqrt{n}}$$

Where is the sample mean, S is the sample standard deviation of the sample and n is the sample size. The degrees of freedom used in this test is n - 1.

Slope of a regression line

Suppose one is fitting the model -

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

where x_i, $i = 1, ..., n$ are known, α and β are unknown, and e_i are independent identically normally distributed random errors with expected value 0 and unknown variance s^2, and Y_i, $i = 1, ..., n$ are observed. It is desired to test the null hypothesis that the slope β is equal to some specified value β_0 (often taken to be 0, in which case the hypothesis is that x and y are unrelated).

1.24 F-STATISTICS

F-tests are named after its test statistic, F, which was named in honor of Sir Ronald Fisher. The F-statistic is simply a ratio of two variances. Variances are a measure of dispersion, or how far the data are scattered from the mean. Larger values represent greater dispersion.

F is for F-test

Variance is the square of the standard deviation. For us humans, standard deviations are easier to understand than variances because they're in the same units as the data rather than squared units. However, many analyses actually use variances in the calculations.

F-statistics are based on the ratio of mean squares. The term "mean squares" may sound confusing but it is simply an estimate of population variance that accounts for the degrees of freedom (DF) used to calculate that estimate.

Despite being a ratio of variances, you can use F-tests in a wide variety of situations. Unsurprisingly, the F-test can assess the equality of variances. However, by changing the variances that are included in the ratio, the F-test becomes a very flexible test. For example, you can use F-statistics and F-tests to test the overall significance for a regression model, to compare the fits of different models, to test specific regression terms, and to test the equality of means.

Using the F-test in One-Way ANOVA

To use the F-test to determine whether group means are equal, it's just a matter of including the correct variances in the ratio. In one-way ANOVA, the F-statistic is this ratio:

F = variation between sample means / variation within the samples

The best way to understand this ratio is to walk through a one-way ANOVA example.

Analysis of variance (ANOVA)

Analysis of Variance (ANOVA) is a collection of statistical models and their associated procedures, in which the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes t-test to more than two groups. Doing multiple two-sample t-tests would result in an increased chance of committing a type I error. For this reason, ANOVAs are useful in comparing two, three or more means.

An important technique for analyzing the effect of categorical factors on a response is to perform an Analysis of Variance. An ANOVA decomposes the variability in the response variable amongst the different factors. Depending upon the type of analysis, it may be important to determine: (a) which factors have a significant effect on the response, and/or (b) how much of the variability in the response variable is attributable to each factor.

Notes

Statgraphics Centurion provides several procedures for performing an analysis of variance:

1. ***One-Way ANOVA -*** used when there is only a single categorical factor. This is equivalent to comparing multiple groups of data.
2. ***Multifactor ANOVA -*** used when there is more than one categorical factor, arranged in a crossed pattern. When factors are crossed, the levels of one factor appear at more than one level of the other factors.
3. ***Variance Components Analysis -*** used when there are multiple factors, arranged in a hierarchical manner. In such a design, each factor is nested in the factor above it.
4. ***General Linear Models -*** used whenever there are both crossed and nested factors, when some factors are fixed and some are random, and when both categorical and quantitative factors are present.

One-Way ANOVA

A one-way analysis of variance is used when the data are divided into groups according to only one factor. The questions of interest are usually: (a) Is there a significant difference between the groups and (b) If so, which groups are significantly different from which others? Statistical tests are provided to compare group means, group medians, and group standard deviations. When comparing means, multiple range tests are used, the most popular of which is Tukey's HSD procedure. For equal size samples, significant group differences can be determined by examining the means plot and identifying those intervals that do not overlap.

Multifactor ANOVA

When more than one factor is present and the factors are crossed, a multifactor ANOVA is appropriate. Both main effects and interactions between the factors may be estimated. The output includes an ANOVA table and a new graphical ANOVA from the latest edition of Statistics for Experimenters by Box, Hunter and Hunter (Wiley, 2005). In a graphical ANOVA, the points are scaled so that any levels that differ by more than exhibited in the distribution of the residuals are significantly different.

Variance Components Analysis

A Variance Components Analysis is most commonly used to determine the level at which variability is being introduced into a product. A typical experiment might select several batches, several samples from each batch and then run replicates tests on each sample. The goal is to determine the relative percentages of the overall process variability that is being introduced at each level.

Assumptions of ANOVA

The analysis of variance has been studied from several approaches, the most common of which use a linear model that relates the response to the treatments and blocks. Even when the statistical model is nonlinear, it can be approximated by a linear model for which an analysis of variance may be appropriate.

Notes

(1) The model is correctly specified.

(2) The εij's are normally distributed.

(3) The εij's have mean zero and a common variance, .

(4) The εij's are independent across observations.

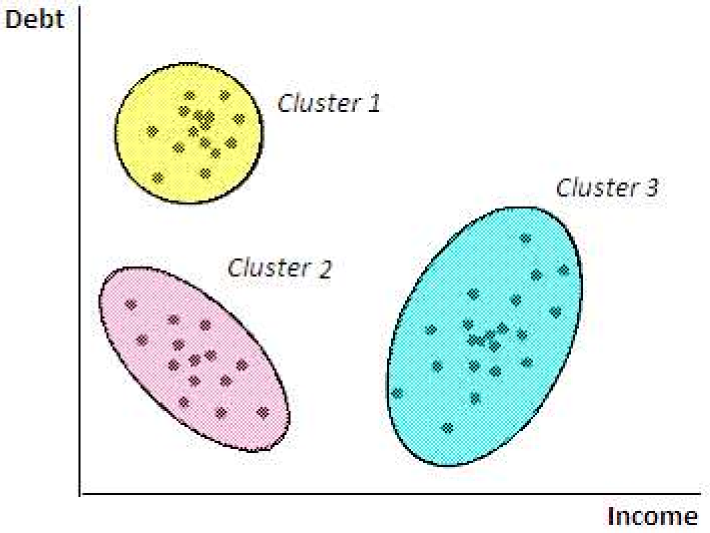
With multiple populations, detection of violations of these assumptions requires examining the residuals rather than the Y-values themselves.

Cluster Analysis

Cluster analysis is a collection of statistical methods, which identifies groups of samples that behave similarly or show similar characteristics. In common parlance it is also called look-a-like groups. The simplest mechanism is to partition the samples using measurements that capture similarity or distance between samples. In this way, clusters and groups are interchangeable words. Often in market research studies, cluster analysis is also referred to as a segmentation method. In neural network concepts, clustering method is called unsupervised learning (refers to discovery as against prediction – even discovery in loose sense may be called prediction, but it does not have predefined learning sets to validate the knowledge). Typically in clustering methods, all the samples with in a cluster is considered to be equally belonging to the cluster. If each observation has its unique probability of belonging to a group (cluster) and the application is interested more about these probabilities than we have to use (binomial) multinomial models.

Cluster analysis is a class of statistical techniques that can be applied to data that exhibit "natural" groupings. Cluster analysis sorts through the raw data and groups them into clusters. A cluster is a group of relatively homogeneous cases or observations. Objects in a cluster are similar to each other. They are also dissimilar to objects outside the cluster, particularly objects in other clusters.

In an unsupervised learning environment the system has to discover its own classes and one way in which it does this is to cluster the data in the database as shown in the following diagram. The first step is to discover subsets of related objects and then find descriptions e.g. D1, D2, D3 etc. which describe each of these subsets.



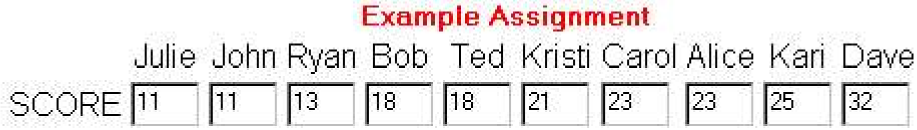
Clustering and segmentation basically partition the database so that each partition or group is similar according to some criteria or metric. Clustering according to similarity is a concept which appears in many disciplines. If a measure of similarity is available there are a number of techniques for forming clusters. Membership of groups can be based on the level of similarity between members and from this the rules of membership can be defined. Another approach is to build set functions that measure some property of partitions i.e., groups or subsets as functions of some parameter of the partition. This latter approach achieves what is known as optimal partitioning.

Many data mining applications make use of clustering according to similarity for example to segment a client/customer base. Clustering according to optimization of set functions is used in data analysis e.g. when setting insurance tariffs the customers can be segmented according to a number of parameters and the optimal tariff segmentation achieved.

Clustering/segmentation in databases are the processes of separating a data set into components that reflect a consistent pattern of nalyzin. Once the patterns have been established they can then be used to "deconstruct" data into more understandable subsets and also they provide sub-groups of a population for further analysis or action which is important when dealing with very large databases. For example a database could be used for profile generation for target marketing where previous response to mailing campaigns can be used to generate a profile of people who responded and this can be used to predict response and filter mailing lists to achieve the best response.

Simple Cluster Analysis

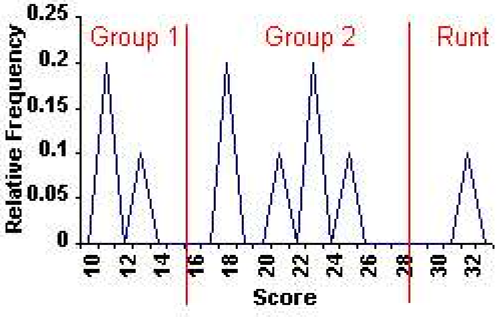
In cases of one or two measures, a visual inspection of the data using a frequency polygon or scatter plot often provides a clear picture of grouping possibilities. *For example,* the following is the data from the "Example Assignment" of the cluster analysis homework assignment.



Example Assignment

	Julie	John	Ryan	Bob	Ted	Kristi	Carol	Alice	Kari	Dave
SCORE	11	11	13	18	18	21	23	23	25	32

The relative frequency polygon appears as follows:



It is fairly clear from this picture that two subgroups, the first including X, Y, and Z and the second including everyone else except describe the data fairly well. When faced with complex multivariate data, such visualization procedures are not available and computer programs assist in assigning objects to groups. The following text describes the logic involved in cluster analysis algorithms.

Steps in Doing a Cluster Analysis

A common approach to doing a cluster analysis is to first create a table of relative similarities or differences between all objects and second to use this information to combine

Notes

the objects into groups. The table of relative similarities is called a proximities matrix. The method of combining objects into groups is called a clustering algorithm. The idea is to combine objects that are similar to one another into separate groups.

The Proximities Matrix

Cluster analysis starts with a data matrix, where objects are rows and observations are columns. From this beginning, a table is constructed where objects are both rows and columns and the numbers in the table are measures of similarity or differences between the two observations. *For example,* given the following data matrix:

$$\begin{array}{cccccc} & X_1 & X_2 & X_3 & X_4 & X_5 \\ O_1 & & & & & \\ O_2 & & & & & \\ O_3 & & & & & \\ O_4 & & & & & \end{array}$$

A proximities matrix would appear as follows:

$$\begin{array}{ccccc} & O_1 & O_2 & O_3 & O_4 \\ O_1 & & & & \\ O_2 & & & & \\ O_3 & & & & \\ O_4 & & & & \end{array}$$

The difference between a proximities matrix in cluster analysis and a correlation matrix is that a correlation matrix contains similarities between variables (X_1, X_2) while the proximities matrix contains similarities between observations (O_1, O_2).

The researcher has dual problems at this point. The first is a decision about what variables to collect and include in the analysis. Selection of irrelevant measures will not aid in classification. For example, including the number of legs an animal has would not help in differentiating cats and dogs, although it would be very valuable in differentiating between spiders and insects.

The second problem is how to combine multiple measures into a single number, the similarity between the two observations. This is the point where univariate and multivariate cluster analysis separate. Univariate cluster analysis groups are based on a single measure, while multivariate cluster analysis is based on multiple measures.

Univariate Measures

A simpler version of the problem of how to combine multiple measures into a measure of difference between objects is how to combine a single observation into a measure of difference between objects. Consider the following scores on a test for four students:

Student	Score
X	11
Y	11
Z	13
A	18

The proximities matrix for these four students would appear as follows:

	X	Y	Z	A
X				
Y				
Z				
A				

The entries of this matrix will be described using a capital "D", for distance with a subscript describing which row and column. *For example,* D_{34} would describe the entry in row 3, column 4, or in this case, the intersection of Z and A.

One means of filling in the proximities matrix is to compute the absolute value of the difference between scores. *For example,* the distance, D, between Z and A would be |13-18| or 5. Completing the proximities matrix using the example data would result in the following:

	X	Y	Z	A
X	0	0	2	7
Y	0	0	2	7
Z	2	2	0	5
A	7	7	5	0

A second means of completing the proximities matrix is to use the squared difference between the two measures. Using the example above D_{34}, the distance between Z and A, would be (13-18)2 or 25. This distance measure has the advantage of being consistent with many other statistical measures, such as variance and the least squares criterion and will be used in the examples that follow. The example proximities matrix using squared differences as the distance measure is presented below.

	X	Y	Z	A
X	0	0	4	49
Y	0	0	4	49
Z	4	4	0	25
A	49	49	25	0

Note that both example proximities matrices are symmetrical. Symmetrical means that row and column entries can be interchanged or that the numbers are the same on each half of the matrix defined by a diagonal running from top left to bottom right.

Other distance measures have been proposed and are available with statistical packages. For example, SPSS/WIN provides the following options for distance measures.

Some of these options themselves contain options. *For example,* Minkowski and Customized are really many different possible measures of distance.

Multivariate Measures

When more than one measure is obtained for each observation, then some method of combining the proximities matrices for different measures must be found. Usually the matrices are summed in a combined matrix. *For example:* given the following scores.

	X1	X2
O1	25	11
O2	33	11
O3	34	13
O4	35	18

The two proximities matrices resulting from squared Euclidean distance that result could be summed to produce a combined distance matrix.

	O1	O2	O3	O4
O1	0	64	81	100
O2	64	0	1	4
O3	81	1	0	1
O4	100	4	1	0

+

	O1	O2	O3	O4
O1	0	0	4	49
O2	0	0	4	49
O3	4	4	0	25
O4	49	49	25	0

=

	O1	O2	O3	O4
O1	0	64	85	149
O2	64	0	5	53
O3	85	5	0	26
O4	149	53	26	0

Note that each corresponding cell is added. With more measures there are more matrices to be added together.

This system works reasonably well if the measures share similar scales. One measure can overwhelm the other if the measures use different scales. Consider the following scores.

	X1	X2
O1	25	11
O2	33	21
O3	34	33
O4	35	48

The two proximities matrices resulting from squared Euclidean distance that result could be summed to produce a combined distance matrix.

	O1	O2	O3	O4
O1	0	64	81	100
O2	64	0	1	4

O3	81	1	0	1
O4	100	4	1	0

+

	O1	O2	O3	O4
O1	0	100	484	49
O2	100	0	144	729
O3	484	144	0	225
O4	1369	729	225	0

=

	O1	O2	O3	O4
O1	0	164	485	153
O2	164	0	145	733
O3	565	145	0	226
O4	1469	733	226	0

It can be seen that the second measure overwhelms the first in the combined matrix.

For this reason the measures are optionally transformed before they are combined. For example, the previous data matrix might be converted to standard scores before computing the separated distance matrices.

	X1	X2	Z1	Z2
O1	25	11	-1.48	-1.08
O2	33	21	.27	-.45
O3	34	33	.49	.30
O4	35	48	.71	1.24

The two proximities matrices resulting from squared Euclidean distance that result from the standard scores could be summed to produce a combined distance matrix.

	O1	O2	O3	O4
O1	0	3.06	3.88	4.80
O2	3.06	0	.05	.19
O3	3.88	.05	0	.05
O4	4.80	.19	.05	0

+

	O1	O2	O3	O4
O1	0	.40	1.90	5.38
O2	.40	0	.56	2.86
O3	1.9	.56	0	.88
O4	5.38	2.86	.88	0

=

Notes

	O1	O2	O3	O4
O1	0	3.46	5.78	10.18
O2	3.46	0	.61	3.05
O3	5.78	.61	0	.93
O4	10.18	3.05	.93	0

The point is that the choice of whether to transform the data and the choice of distance metric can result in vastly different proximities matrices.

Using Distances to Group Objects

After the distances between objects have been found, the next step in the cluster analysis procedure is to divide the objects into groups based on the distances. Again, any number of options is available to do this.

If the number of groups is known beforehand, a "flat" method might be preferable. Using this method, the objects are assigned to a given group at the first step based on some initial criterion. The means for each group are calculated. The next step reshuffles the objects into groups, assigning objects to groups based on the object's similarity to the current mean of that group. The means of the groups are recalculated at the end of this step. This process continues recursively until no objects change groups. This idea is the basis for "k-means cluster analysis" available on SPSS/WIN and other statistical packages. This method works well if the number of groups matches the data and the initial solution is reasonably close to the final solution.

Hierarchical clustering methods do not require preset knowledge of the number of groups. Two general methods of hierarchical clustering methods are available: divisive and agglomerative.

The divisive techniques start by assuming a single group, partitioning that group into subgroups, partitioning these subgroups further into subgroups and so on until each object forms its own subgroup. The agglomerative techniques start with each object describing a subgroup, and then combine like subgroups into more inclusive subgroups until only one group remains. The agglomerative techniques will be described further in this chapter, although many of the procedures described would hold for either method.

In either case, the results of the application of the clustering technique are best described using a dendogram or binary tree. The objects are represented as nodes in the dendogram and the branches illustrate when the cluster method joins subgroups containing that object. The length of the branch indicates the distance between the subgroups when they are joined. The interpretation of a dendogram is fairly straightforward. In the above dendogram, for example, X, Y, and Z form a group, A, Ted, Kristi, Carol, Alice, and Kari form a second group, and Dave is called a "runt" because he doesn't enter any group until near the end of the procedure.

Different methods exist for computing the distance between subgroups at each step in the clustering algorithm. Again, statistical packages give options for which procedure to use. For example, SPSS/WIN optionally allows the following methods.

The following methods will now be discussed: single linkage or nearest neighbor, complete linkage or furthest neighbor, and average linkage.

Single Linkage

Single linkage (nearest neighbor in SPSS/WIN) computes the distance between two subgroups as the minimum distance between any two members of opposite groups. For example, consider the following proximities matrix.

	X	Y	Z	A	Ted	Kristi
X	0	4	36	81	196	225
Y	4	0	16	49	144	169
Z	36	16	0	9	64	81
A	81	49	9	0	25	36
Ted	196	144	64	25	0	1
Kristi	225	169	81	36	1	0

Based on this data, Ted and Kristi would be joined on the first step.

	X	Y	Z	A	Ted and Kristi
X	0	4	36	81	196
Y	4	0	16	49	144
Z	36	16	0	9	64
A	81	49	9	0	25
Ted and Kristi	196	144	64	25	0

X and Y would be joined on the second step.

	X and Y	Z	A	Ted and Kristi
X and Y	0	16	49	144
Z	16	0	9	64
A	49	9	0	25
Ted and Kristi	144	64	25	0

Z and A would be joined on the third step. At that step four would be three groups of two each and the proximities matrix at that point would appear as follows.

	X and Y	Z and A	Ted and Kristi
X and Y	0	16	144
Z and A	16	0	25
Ted and Kristi	144	25	0

Using single linkage, X and Y would be grouped with Z and A at the third step. The distance of this linkage would be 16. The last step would join all into a single group with a distance of 25.

Complete Linkage

Complete linkage (furthest neighbor in SPSS/WIN) computes the distance between subgroups in each step as the maximum distance between any two members of the different groups. Using the example proximities matrix, complete linkage would join similar members at steps 1, 2 and 3 in the procedure. At step for the proximities matrix for the three groups would appear as follows.

	X and Y	Z and A	Ted and Kristi
X and Y	0	81	225
Z and A	81	0	81
Ted and Kristi	225	81	0

Step four would combine all three subgroups into a single group with a distance of 81.

Average Linkage

Average linkage computes the distance between subgroups at each step as the average of the distances between the two subgroups.

Using the example proximities matrix, average linkage would join similar members at steps 1, 2, and 3 in the procedure. At step for the proximities matrix for the three groups would appear as follows.

	X and Y	Z and A	Ted and Kristi
X and Y	0	45.5	51.5
Z and A	45.5	0	81
Ted and Kristi	183.5	51.5	0

Thus, X and Y would be grouped with Z and A with a distance of 51.5. The final step would join this group with Ted and Kristi with an average distance of 117.5. A program generating a cluster analysis homework assignment permits the student to view a dendogram for single, complete, and average linkage for a random proximities matrix. The student should verify that different dendograms result when different linkage methods are used.

The entire gamut of statistical techniques can be broadly classified into univariate and multivariate, based on the nature of the problem.

Univariate techniques are appropriate when there is a single measurement of each of the n sample objects, or when there are several measurements of each of the n observations but each variable is analysed in isolation. On the other hand, multivariate techniques are appropriate for nalyzing data when there are two or more measurements of each observation and the variables are to be analysed simultaneously. Based on the type of data, univariate techniques can be further classified into non-metric or metric. The non-metric data are measured on a nominal or ordinal scale, whereas metric data are measured on an interval or ratio scale. Non-parametric statistical tests can be used to analyse non-metric data. Non-parametric tests do not require any assumptions regarding the distribution of data.

For both non-metric and metric data, the next level of classification involves determining whether a single sample or multiple samples are involved. Further, in the case of multiple samples, the appropriate statistical test depends on whether the samples are independent or dependent. For metric data, t-tests and z-tests can be used for one or two samples. For more than two samples, the analysis of variance (ANOVA) is used. For non-metric data, with a single sample, chi-square, Kolmogorov-Smirnov (K-S), and RUNS tests can be used. For two or more independent samples, chi-square, rank sum tests, K-S, and ANOVA (Kruskal-Wallis ANOVA) should be used. For two or more dependent samples, sign test, Wilcoxon test, McNemar and Cochran Q-tests can be used. A detailed discussion of non-parametric statistics is beyond the scope of this lesson.

1.25 SUMMARY

Econometrics is the analysis and testing of economic theories to verify hypotheses and improve prediction of financial trends. Econometrics takes mathematical and statistical models proposed in economic theory and tests them. First, models are tested against statistical trials, followed by testing against real-world examples to support or disprove hypotheses. Econometrics uses an important statistical method called regression analysis, which assesses the connection among variables. Economists use the regression method since they cannot usually carry out controlled experiments, choosing to instead gather information from natural experiments.

The Econometric Approaches that make use of statistical tools and economic theories in combination to estimate the economic variables and to forecast the intended variables.

The regression approach is the most common method used to forecast the demand for a product. This method combines the economic theory with statistical tools of estimation. The economic theory is applied to specify the demand determinants and the nature of the relationship between product's demand and its determinants. Thus, through an economic theory, a general form of a demand function is determined. The statistical techniques are applied to estimate the values of parameters in the projected equation.

Under simultaneous equation model, demand forecasting involves the estimation of several simultaneous equations. These equations are often the behavioral equations, market-clearing equations, and mathematical identities.

The regression technique is based on the assumption of one-way causation, which means independent variables cause variations in the dependent variables, and not vice-versa. In simple terms, the independent variable is in no way affected by the dependent variable. For example, $D = a - bP$, which shows that price affects demand, but demand does not affect the price, which is an unrealistic assumption.

A hypothesis (plural hypotheses) is a proposed explanation for a phenomenon. For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it. Scientists generally base scientific hypotheses on previous observations that cannot satisfactorily be explained with the available scientific theories. Even though the words "hypothesis" and "theory" are often used synonymously, a scientific hypothesis is not the same as a scientific theory. A working hypothesis is a provisionally accepted hypothesis proposed for further research, in a process beginning with an educated guess or thought.

Data are the raw material from which econometric analysis is constructed. Just as a building is no stronger than the wood or steel used in its framework, an econometric study is only as reliable as the data used in its analysis. Many econometricians over the years have written about problems with data. One of the most comprehensive and comprehensible is a chapter that noted econometrician Zvi Griliches wrote for the third volume of Elsevier's Handbook of Econometrics back in 1986.

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the Dow Jones Industrial Average.

Time series are very frequently plotted via line charts. Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather

Notes

forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communications engineering, and largely in any domain of applied science and engineering which involves temporal measurements.

Cross-sectional data or a cross section of a study population, in statistics and econometrics is a type of data collected by observing many subjects (such as individuals, firms, countries, or regions) at the same point of time, or without regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among the subjects.

Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution. Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population.

Maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model, given observations. MLE attempts to find the parameter values that maximize the likelihood function, given the observations. The resulting estimate is called a maximum likelihood estimate, which is also abbreviated as MLE.

The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the sample statistic + margin of error.

The probability part of a confidence interval is called a confidence level. The confidence level describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter.

The margin of error m of interval estimation is defined to be the value added or subtracted from the sample mean which determines the length of the interval:

Hypothesis test is a method of making decisions using data from a scientific study. In statistics, a result is called statistically significant if it has been predicted as unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the significance level. The phrase "test of significance" was coined by statistician Ronald Fisher.

Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses. It is an assumption about a population parameter. This assumption may or may not be true.

The null hypothesis, denoted by H0, is usually the hypothesis that sample observations result purely from chance.

The alternative hypothesis, denoted by H1 or Ha, is the hypothesis that sample observations are influenced by some non-random cause.

A test of a statistical hypothesis, where the region of rejection is on only one side of the sampling distribution, is called a one-tailed test. For example, suppose the null hypothesis states that the mean is less than or equal to 10. The alternative hypothesis would be that the mean is greater than 10. The region of rejection would consist of a range of numbers located on the right side of sampling distribution; that is, a set of numbers greater than 10.

The Neyman-Pearson Lemma is a way to find out if the hypothesis test you are using is the one with the greatest statistical power. The power of a hypothesis test is the probability that test correctly rejects the null hypothesis when the alternate hypothesis is true. The goal would be to maximize this power, so that the null hypothesis is rejected as much as possible when the alternate is true. The lemma basically tells us that good hypothesis tests are likelihood ratio tests.

A sampling distribution or finite-sample distribution is the probability distribution of a given random-sample-based statistic. If an arbitrarily large number of samples, each involving multiple observations (data points), were separately used in order to compute one value of a statistic (such as, for example, the sample mean or sample variance) for each sample, then the sampling distribution is the probability distribution of the values that the statistic takes on. In many contexts, only one sample is observed, but the sampling distribution can be found theoretically.

Sampling design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sampling design is determined before any data are collected.

A Z-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Because of the central limit theorem, many test statistics are approximately normally distributed for large samples.

χ^2 test is a test that uses the chi-square statistic to test the fit between a theoretical frequency distribution and a frequency distribution of observed data for which each observation may fall into one of several classes.

A statistical examination of two population means. A two-sample t-test examines whether two samples are different and is commonly used when the variances of two normal distributions are unknown and when an experiment uses a small sample size.

1.26 SELF ASSESSMENT QUESTIONS

1. What are Econometric Approaches? Discuss the Econometric Approaches.
2. What is Econometrics? Explain various objectives of Econometrics.
3. Discuss the sources of hypothesis used in Econometrics.
4. Explain about the raw materials of econometrics.
5. Define the term Time Series? Discuss applications of Time series and Cross section data.
6. Explain the problem of their pooling together.
7. What is Statistical Inference? Discuss various elements of Statistical inferences.
8. What is interval estimation? Explain about point and interval estimation- estimator and its properties.
9. What is Likelihood ratio? Discuss various method of Maximum Likelihood, interval Estimatio.
10. Define the term Hypothesis Testing. Explain about the Test of Hypothesis.

Notes

11. Explain Simple and composite hypothesis.
12. Discuss about Neyman Pearson Lemma.
13. What is Sampling Distribution? Explain about Sampling Distributions.
14. Briefly explain Z-statistics, Chi-square, t-statistics and F-statistics.

*****

2 Lesson

CLASSICAL LINEAR REGRESSION

Objectives

The objectives of this lesson are to:

- Classical Linear Regression (with one explanatory variable)
- Assumption and their economic interpretation,
- Least square estimations of regression parameters, their properties, Gauss-Markov Theory, Theorem: Standard errors of estimates. Estimator of errors, Control limit theorem,
- Maximum likelihood estimator
- Normality of errors, control limit theorem
- Maximum Likelihood Estimator
- Significance test and confidence intervals of estimates-z-test, t-test and f-ratio test
- Prediction point and interval
- Extension of the two variable linear model
- Three variable linear model, the coefficient of multiple correlation, partial correlation coefficient
- General Linear model (with K- Explanatory variable)- Assumptions
- Least-square estimates and their properties
- Variance - covariance matrix of estimates, Estimates of error variance
- Multiple coefficient of determination
- R2 and multiple correlation co-efficient- R
- Significance test and confidence intervals, prediction
- Non-linear Models-Choice of functional forms, estimation

Structure:

Notes

2.1 INTRODUCTION

Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regressions. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.

If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

2.2 CLASSICAL LINEAR REGRESSION (WITH ONE EXPLANATORY VARIABLE)

Model statistical-tool used in predicting future values of a target (dependent) variable on the basis of the behavior of a set of explanatory factors (independent variables). A type

Notes

of regression analysis model, it assumes the target variable is predictable, not chaotic or random.

1. Linear and Additive

If you fit a linear model to a non-linear, non-additive data set, the regression algorithm would fail to capture the trend mathematically, thus resulting in an inefficient model. Also, this will result in erroneous predictions on an unseen data set.

How to check: Look for residual vs fitted value plots (explained below). Also, you can include polynomial terms (X, X^2, X^3) in your model to capture the non-linear effect.

2. Autocorrelation

The presence of correlation in error terms drastically reduces model's accuracy. This usually occurs in time series models where the next instant is dependent on previous instant. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error.

If this happens, it causes confidence intervals and prediction intervals to be narrower. Narrower confidence interval means that a 95% confidence interval would have lesser probability than 0.95 that it would contain the actual value of coefficients. Let's understand narrow prediction intervals with an example:

For example, the least square coefficient of X^1 is 15.02 and its standard error is 2.08 (without autocorrelation). But in presence of autocorrelation, the standard error reduces to 1.20. As a result, the prediction interval narrows down to (13.82, 16.22) from (12.94, 17.10).

Also, lower standard errors would cause the associated p-values to be lower than actual. This will make us incorrectly conclude a parameter to be statistically significant.

How to check: Look for Durbin – Watson (DW) statistic. It must lie between 0 and 4. If $DW = 2$, implies no autocorrelation, $0 < DW < 2$ implies positive autocorrelation while $2 < DW < 4$ indicates negative autocorrelation. Also, you can see residual vs time plot and look for the seasonal or correlated pattern in residual values.

3. Multi-collinearity

This phenomenon exists when the independent variables are found to be moderately or highly correlated. In a model with correlated variables, it becomes a tough task to figure out the true relationship of predictors with response variable. In other words, it becomes difficult to find out which variable is actually contributing to predict the response variable.

Another point, with presence of correlated predictors, the standard errors tend to increase. And, with large standard errors, the confidence interval becomes wider leading to less precise estimates of slope parameters.

Also, when predictors are correlated, the estimated regression coefficient of a correlated variable depends on which other predictors are available in the model. If this happens, you'll end up with an incorrect conclusion that a variable strongly/weakly affects target variable. Since, even if you drop one correlated variable from the model, its estimated regression coefficients would change. That's not good!

How to check: You can use scatter plot to visualize correlation effect among variables. Also, you can also use VIF factor. VIF value <= 4 suggests no multicollinearity whereas a value of >= 10 implies serious multicollinearity. Above all, a correlation table should also solve the purpose.

4. Heteroskedasticity

The presence of non-constant variance in the error terms results in heteroskedasticity. Generally, non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influences the model's performance. When this phenomenon occurs, the confidence interval for out of sample prediction tends to be unrealistically wide or narrow.

How to check: You can look at residual vs fitted values plot. If heteroskedasticity exists, the plot would exhibit a funnel shape pattern (shown in next section). Also, you can use Breusch-Pagan / Cook – Weisberg test or White general test to detect this phenomenon.

5. Normal Distribution of error terms

If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. Presence of non – normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.

How to check: You can look at QQ plot (shown below). You can also perform statistical tests of normality such as Kolmogorov-Smirnov test, Shapiro-Wilk test.

2.3 ASSUMPTION AND THEIR ECONOMIC INTERPRETATION

The important assumptions in regression analysis:

a) There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.

b) There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.

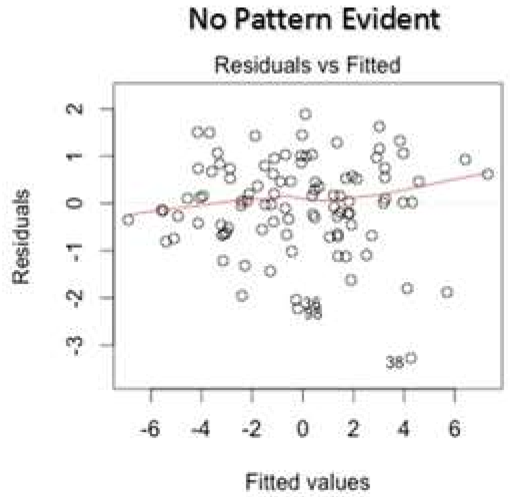
c) The independent variables should not be correlated. Absence of this phenomenon is known as multi-collinearity.

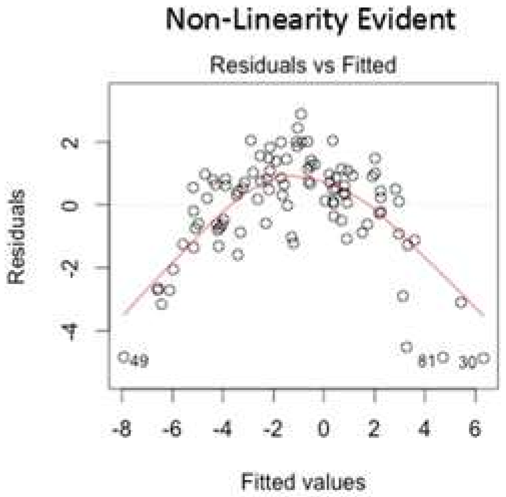
d) The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.

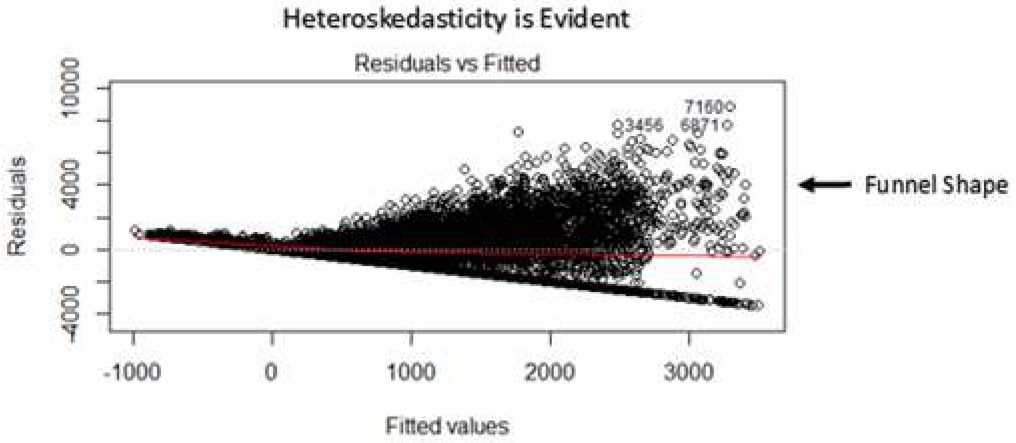
e) The error terms must be normally distributed.

1. Residual vs Fitted Values

This scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values). It is one of the most important plot which everyone must learn. It reveals various useful insights including outliers. The outliers in this plot are labeled by their observation number which make them easy to detect.







There are two major things which you should learn:

If there exist any pattern (may be, a parabolic shape) in this plot, consider it as signs of non-linearity in the data. It means that the model doesn’t capture non-linear effects.

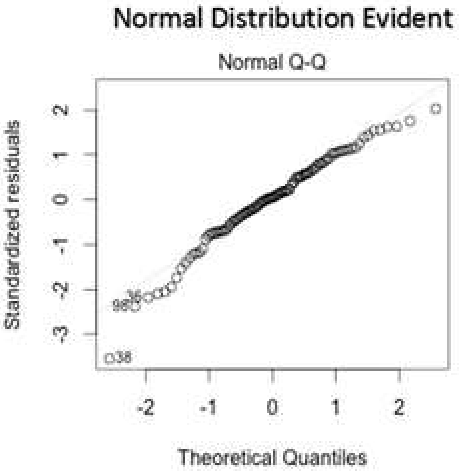
If a funnel shape is evident in the plot, consider it as the signs of non constant variance i.e. heteroskedasticity.

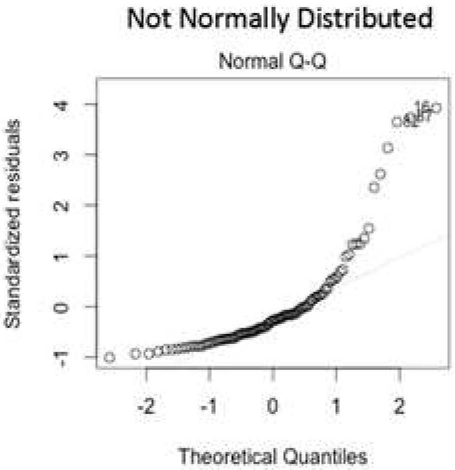
Solution:

To overcome the issue of non-linearity, you can do a non linear transformation of predictors such as log (X), vX or X^2 transform the dependent variable. To overcome heteroskedasticity, a possible way is to transform the response variable such as log(Y) or vY. Also, you can use weighted least square method to tackle heteroskedasticity.

2. Normal Q-Q Plot

This q-q or quantile-quantile is a scatter plot which helps us validate the assumption of normal distribution in a data set. Using this plot we can infer if the data comes from a normal distribution. If yes, the plot would show fairly straight line. Absence of normality in the errors can be seen with deviation in the straight line.



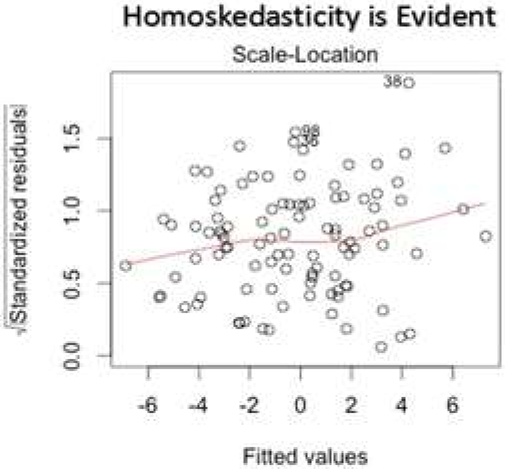


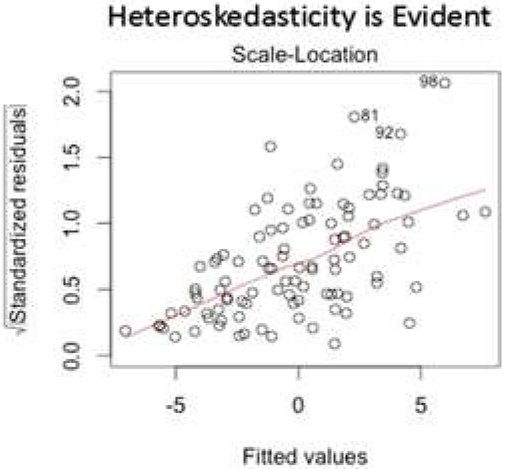
If you are wondering what is a ‘quantile’, here’s a simple definition: Think of quantiles as points in your data below which a certain proportion of data falls. Quantile is often referred to as percentiles. For example: when we say the value of 50th percentile is 120, it means half of the data lies below 120.

Solution:

If the errors are not normally distributed, non – linear transformation of the variables (response or predictors) can bring improvement in the model.

3. Scale Location Plot





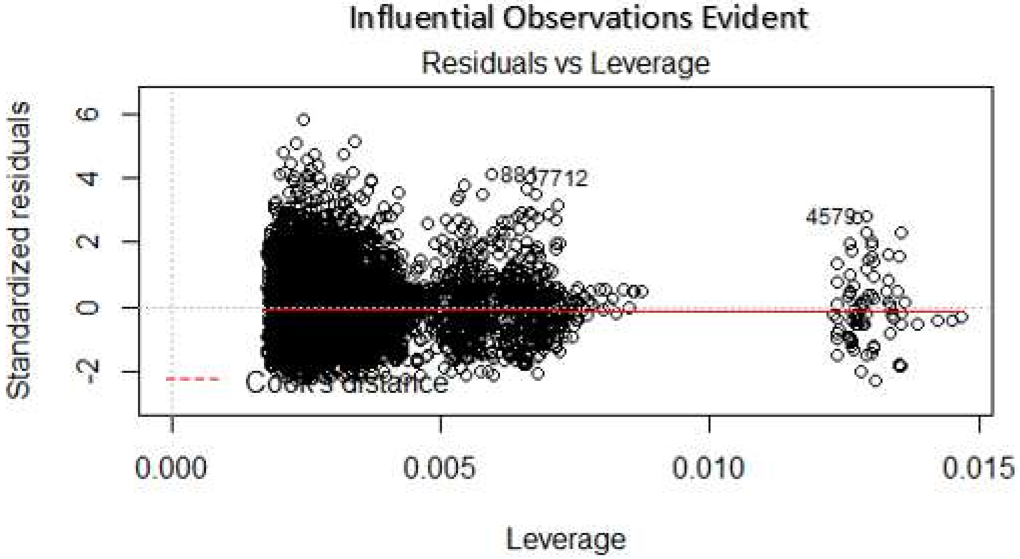
This plot is also used to detect homoskedasticity (assumption of equal variance). It shows how the residual are spread along the range of predictors. It’s similar to residual vs fitted value plot except it uses standardized residual values. Ideally, there should be no discernible pattern in the plot. This would imply that errors are normally distributed. But, in case, if the plot shows any discernible pattern (probably a funnel shape), it would imply non-normal distribution of errors.

Solution:

Follow the solution for heteroskedasticity given in plot 1.

Notes

4. Residuals vs Leverage Plot



It is also known as Cook's Distance plot. Cook's distance attempts to identify the points which have more influence than other points. Such influential points tends to have a sizable impact of the regression line. In other words, adding or removing such points from the model can completely change the model statistics.

But, can these influential observations be treated as outliers? This question can only be answered after looking at the data. Therefore, in this plot, the large values marked by cook's distance might require further investigation.

Solution:

For influential observations which are nothing but outliers, if not many, you can remove those rows. Alternatively, you can scale down the outlier observation with maximum value in data or else treat those values as missing values.

2.4 LEAST SQUARE ESTIMATIONS OF REGRESSION PARAMETERS THEIR PROPERTIES

In the previous chapters, several models used in stock assessment were analysed, the respective parameters having been defined. In the corresponding exercises, it was not necessary to estimate the values of the parameters because they were given. In this chapter, several methods of estimating parameters will be analysed. In order to estimate the parameters, it is necessary to know the sampling theory and statistical inference.

This manual will use one of the general methods most commonly used in the estimation of parameters - the least squares method. In many cases this method uses iterative processes, which require the adoption of initial values. Therefore, particular methods will also be presented, which obtain estimates close to the real values of the parameters. In many situations, these initial estimates also have a practical interest. These methods will be illustrated with the estimation of the growth parameters and the S-R stock-recruitment relation.

The least squares method is presented under the forms of Simple linear Regression, multiple linear model and non-linear models (method of Gauss-Newton).

Subjects like residual analysis, sampling distribution of the estimators (asymptotic or empiric Bookstrap and jacknife), confidence limits and intervals, etc., are important. However, these matters would need a more extensive course.

Simple Linear Regression - Least Squares Method

Model

Consider the following variables and parameters:

Response or dependent variable	= Y
Auxiliary or independent variable	= X
Parameters	= A, B

The response variable is linear with the parameters

Y = A + BX

Objective

The objective of the method is to estimate the parameters of the model, based on the observed pairs of values and applying a certain criterium function (the observed pairs of values are constituted by selected values of the auxiliary variable and by the corresponding observed values of the response variable), that is:

Observed values x_i and y_i for each pair i, where i = 1, 2, ..., i, ... n

Values to be estimated A and B and (Y1, Y2,..., Yi,...,Yn) for the n observed pairs of values

Estimates values: $\hat{A}$ and $\hat{B}$ (or a and b) and $(\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_i, ..., \hat{Y}_n)$

Object function (or criterium function)

$$\Phi = \sum_{i=1}^{n}(y_i - Y_i)^2$$

Estimation method

In the least squares method the estimators are the values of A and B which minimize the object function. Thus, one has to calculate the derivatives $\partial\Phi/\partial A$ e $\partial\Phi/\partial B$, equate them to zero and solve the system of equations in A and B.

The solution of the system can be presented as:

$\bar{x} = (1/n).\sum x$ $\bar{y} = (1/n).\sum y$

$Sxx = \sum(x - \bar{x})(x - \bar{x})$ $Sxy = \sum(x - \bar{x})(y - \bar{y})$

b = Sxy/Sxx $a = \bar{y} - b.\bar{x}$

Notice that the observed values y, for the same set of selected values of X, depend on the collected sample. For this reason, the problem of the simple linear regression is usually presented in the form:

$y = A + BX + \varepsilon$

where e is a random variable with expected value equal to zero and variance equal to σ^2.

Notes

So, the expected value of y will be Y or A + BX and the variance of y will be equal to the variance of ε .

The terms deviation and residual will be used in the following ways:

Deviation is the difference between y_{observed} and y_{mean} ($\bar{y}$) i.e., deviation = (y - $\bar{y}$)

while

Residual is the difference between y_{observed} and $Y_{\text{estimated}}$ ($\hat{Y}_i$), i.e., residual = $y_i - \hat{Y}_i$.

To analyse the adjustment of the model to the observed data, it is necessary to consider the following characteristics:

Sum of squares of the residuals:

$$SQ_{\text{residual}} = \sum \left(y - \hat{Y} \right)^2$$

This quantity indicates the residual variation of the observed values in relation to the estimated values of the response variable of the model, which can be considered as the variation of the observed values that is not explained by the model.

Sum of squares of the deviations of the estimated values of the response variable of the model:

$$SQ_{\text{model}} = \sum \left(\hat{Y} - \bar{y} \right)^2$$

This quantity indicates the variation of the estimated values of the response variable of the model in relation to its mean, that is the variation of the response variable explained by the model.

Total sum of squares of the deviations of the observed values equal to:

$$SQ_{\text{residual}} = \sum \left(y - \bar{y} \right)^2$$

This quantity indicates the total variation of the observed values in relation to the mean

It is easy to verify the following relation:

$$SQ_{\text{total}} = SQ_{\text{model}} + SQ_{\text{residual}}$$

or

$$1 = \frac{SQ_{\text{model}}}{SQ_{\text{total}}} + \frac{SQ_{\text{residual}}}{SQ_{\text{total}}}$$

or

$$1 = r^2 + (1 - r^2)$$

Where,

r^2 (co-efficient of determination) is the percentage of the total variation that is explained by the model and

$1 - r^2$ is the percentage of the total variation that is not explained by the model.

Multiple Linear Regression - Least Squares Method

Model

Consider the following variables and parameters:

Response or dependent variable	= Y
Auxiliary or independent variables	= X1, X2,..., Xj,..., Xk
Parameters	= B1, B2,..., Bj,..., Bk

The response variable is linear with the parameters

Y = B1X1+B2X2+... + BkXk = $\sum$ BjXj

Objective

The objective of the method is to estimate the parameters of the model, based on the observed n sets of values and by applying a certain criterium function (the observed sets of values are constituted by selected values of the auxiliary variable and by the corresponding observed values of the response variable), that is:

Observed values x1,i x2,i,..., xj,i,.., xk,i and yi for each set i, where i = 1,2,...,i,...n

Values to be estimated B1,B2,...,Bj,...,Bk et (Y1,Y2,..., Yi,..., Yn)

The estimated values can be represented by:

(B1,B2,...,Bj,...,Bk) (ou b1,b2,...,bj,...,bk) et (Y1,Y2,..., Yi,..., Yn)

Object function (or criterium function)

$$\Phi = \sum_{i=1}^{n}(y_i - Y_i)^2$$

Estimation Method

In the least squares method the estimators are the values of Bj which minimize the object function.

As with the simple linear model, the procedure of minimization requires equating the partial derivatives of F to zero in order to each parameter, Bj, where j=1, 2,..., k. The system is preferably solved using matrix calculus.

Matrix version

Matrix X(n,k) = Matrix of the n observed values of each of the k auxiliary variables

Vector y(n,1) = Vector of the n observed values of the response variable

Vector Y(n,1) = Vector of the values of the response variable given by the model (unknown)

Vector B(k,1) = Vector of the parameters

Vector b(k,1) = Vector of the estimators of the parameters

Model

Y(n,1) = X(n,k). B(k,1) ou Y = X.B+ ε

Object function

Φ (1,1) = (y-Y)T.(y-Y) ou Φ (1,1) = (y-X.B)T.(y-X.B)

To calculate the least squares estimators it will suffice to put the derivative $d\,\Phi\,/dB$ of Φ in order to vector B, equal to zero. $d\,\Phi\,/dB$ is a vector with components $\partial\,\Phi\,/\,\partial\,B1$, $\partial\,\Phi\,/\,\partial\,B2$,..., $\partial\,\Phi\,/\,\partial\,Bk$. Thus:

$d\,\Phi\,/dB(k,1)$ = -2.XT.(y-X.B) = 0

or XTy - (XT.X). B = 0

and b = B = (XT.X)-1. XTy

The results can be written as:

b(k,1) = (XT.X)-1.XTy

= X.b or = X (XT.X)-1.XT y

residuals(n,1) = (y-)

Comments

In statistical analysis it is convenient to write the estimators and the sums of the squares using idempotent matrices. Then the idempotent matrices L, (I - L) and (I - M) with L(n,n) = X (XT. X)-1. XT, I = unity matrix and M(n,n) = mean(n,1) matrix = 1/n [1] where [1] is a matrix with all its elements equal to one, are used.

It is also important to consider the sampling distributions of the estimators assuming that the variables ei are independent and have a normal distribution.

Non-Linear Model - Method of Gauss-Newton - Least Squares Method

Model

Consider the following variables and parameters:

Response or dependent variable	= Y
Auxiliary or independent variable	= X
Parameters	= B1,B2,...,Bj,...,Bk

The response variable is non-linear with the parameters

Y = f(X;B) where B is a vector with the components B1,B2,...,Bj,...,Bk

Objective

The objective of the method is to estimate the parameters of the model, based on the n observed pairs of values and by applying a certain criterium function (the observed sets of values are constituted by selected values of the auxiliary variable and by the corresponding observed values of the response variable), that is:

Observed values xi and yi for each pair i, where i = 1,2,...,i,...n

Values to be estimated B1,B2,...,Bj,..,Bk and (Y1,Y2,...,Yi,...,Yn) form the n pairs of observed values.

(Estimates = B1,B2,...,Bj,...,Bk or b1,b2,...,bj,...,bk and Y1,Y2,...,Yi,...,Yn)

Object function or criterium function

$$\Phi = \sum_{i=1}^{n}(y_i . Y_i)^2$$

Estimation Criterium

The estimators will be the values of Bj for which the object function is minimum.

(This criterium is called the least squares method).

Matrix Version

It is convenient to present the problem using matrices.

So:

Vector X(n,1) = Vector of the observed values of the auxiliary variable

Vector y(n,1) = Vector of the observed values of the response variable

Vector Y(n,1) = Vector of the values of the response variable given by the model

Vector B(k,1) = Vector of the parameters

Vector b(k,1) = Vector of the estimators of the parameters

Model

$Y_{(n,1)}$ = f(X; B)

Object function

Φ (1,1) = (y-Y)T.(y-Y)

In the case of the non linear model, it is not easy to solve the system of equations resulting from equating the derivative of the function F in order to the vector B, to zero. Estimation by the least squares method can, based on the Taylor series expansion of function Y, use iterative methods.

Revision of the Taylor series expansion of a function

Here is an example of the expansion of a function in the Taylor series in the case of a function with one variable.

The approximation of Taylor means to expand a function Y = f(x) around a selected point, x0, in a power series of x:

Y = f(x) = f(x0) +(x-x0).f'(x0)/1! + (x-x0)2f''(x0)/2! +... + (x- x0)i f(i)(x0)/i!+...

where

f(i)(x0) = ith derivatives of f(x) in order to x, at the point x0.

The expansion can be approximated to the desired power of x. When the expansion is approximated to the power 1 it is called a linear approximation, that is,

Y $\cong$ f(x0) + (x-x0).f'(x0)

The Taylor expansion can be applied to functions with more than one variable. For example, for a function Y = f(x1,x2) of two variables, the linear expansion would be:

$$Y \approx f(X_{1(0)}, X_{2(0)}) + (X_1 - X_{1(0)}).\frac{\delta f(X_{1(0)}, X_{2(0)})}{\delta X_1} + (X_2 - X_{2(0)}).\frac{\delta f(X_{1(0)}, X_{2(0)})}{\delta X_2}$$

which may be written, in matrix notation, as

$Y = Y_{(0)} + A_{(0)}.(x-x_{(0)})$

where Y(0) is the value of the function at the point x(0),with components x1(0) and x2(0) and A(0) is the matrix of derivatives whose elements are equal to the partial derivatives of f(x1,x2) in order to x1,x2 at the point (x1(0), x2(0)).

Notes

To estimate the parameters, the Taylor series expansion of function Y is made in order to the parameters B and not to the vector X.

For example, the linear expansion of Y = f(x,B) in B1, B2,..., Bk, would be:

Y = f(x;B) = f(x; B(0)) + (B1-B1(0)) f /B1 (x;B(0)) +..... +

(B2-B2(0))f /B2 (x;B(0)) +...... +.........+ (Bk-Bk(0)) f /Bk (x;B(0))

or, in matrix notation, it would be:

Y(n,1) = Y(0) (n,1) + A(0) (n,k). Δ B(0) (k,1)

where

A = matrix of order (n,k) of the partial derivatives of the matrix f(x;B) in order to the vector B at the point B(0) and

Δ B(0) = vector (B - B(0)).

Then, the object function will be:

Φ = (y-Y)T.(y-Y) = (y-Y(0) - A(0). Δ B(0))T(y-Y(0) - A(0). Δ B(0))

To obtain the minimum of this function it is more convenient to differentiate F in order to the vector Δ B than in relation to vector B and put it equal to zero. Thus:

$$0 = -2(A_{(0)})^T (y - Y_{(0)} - A_{(0)}.\Delta B_{(0)}) = -2A_{(0)}{}^T (y- Y_{(0)}) + 2A_{(0)}{}^T A_{(0)}.\Delta B_{(0)}$$

or

$$A^T{}_{(0)} A_{(0)}.\Delta B_{(0)} = A^T{}_{(0)} (y - Y_{(0)})$$

Therefore:

$$\Delta B_{(0)} = (A^T{}_{(0)}.A_{(0)})^{-1}.A^T{}_{(0)} (y - Y_{(0)})$$

If Δ B(0) is "equal to zero" then the estimate of B is equal to B(0).

(In practice, when we say "equal to zero" in this process, we really mean smaller than the approximation vector one has to define beforehand).

If Δ B(0) is not "equal to zero" then the vector B(0) will be replaced by:

B(1) = B(0) + Δ B(0)

And the process will be repeated, that is, there will be another iteration with B(0) replaced by B(1) (and A(0) replaced by A(1)). The iterative process will go on until the convergence at the desired level of approximation is reached.

Comments

1. It is not guaranteed that the process always converges. Sometimes it does not, some other times it is too slow (even for computers!) and some other times it converges to another limit!!
2. The above described method is the Gauss-Newton method which is the basis of many other methods. Some of those methods introduce modifications in order to obtain a faster convergence like the Marquardt method (1963), which is frequently used in fisheries research. Other methods use the second order Taylor expansion (Newton-Raphson method), looking for a better approximation. Some others, combine the two modifications.
3. These methods need the calculation of the derivatives of the functions. Some computer programs require the introduction of the mathematical expressions of

Notes

the derivatives, while others use sub-routines with numerical approximations of the derivatives.

4. In fisheries research, there are methods to calculate the initial values of the parameters, for example in growth, mortality, selectivity or maturity analyses.
5. It is important to point out that the convergence of the iterative methods is faster and more likely to approach the true limit when the initial value of the vector B(0) is close to the real value.

Estimation of Growth Parameters

The least squares method (non-linear regression) allows the estimation of the parameters K, L_∞ and to of the individual growth equations.

The starting values of K, L_∞ and t_0 for the iterative process of estimation can be obtained by simple linear regression using the following methods:

Ford-Walford (1933-1946) and Gulland and Holt (1959) Methods

The Ford-Walford and Gulland and Holt expressions, which were presented in Section 3.4, are already in their linear form, allowing the estimation of K and L_∞ with methods of simple linear regression on observed Li and Ti. The Gulland and Holt expression allows the estimation of K and L_∞ even when the intervals of time Ti are not constant. In this case, it is convenient to re-write the expression as:

$$\Delta L/T_1 = KL_\infty - K\overline{L}$$

Stamatopoulos and Caddy Method (1989)

These authors also present a method to estimate K, L_∞ and to (or Lo) using the simple linear regression. In this case the von Bertalanffy equation should be expressed as a linear relation of Lt against e-Kt.

Consider n pairs of values ti, Li where ti is the age and Li the length of the individual i where i = 1, 2,...., n.

The von Bertalanffy equation, in its general form is (as previously seen):

L_∞ - Lt = (L_∞ - La). e-K(t-ta)

It can be written as:

Lt = L_∞ - (L_∞ - La). e+Kta. e-Kt

The equation has the simple linear form, y = a + bx, where:

y = Lt a = L_∞ b = - (L_∞ - La). e+Kta

x = e-Kt

If one takes La = 0, then ta=to, but, if one considers ta = 0, then La = Lo.

The parameters to estimate from a and b will be L_∞, to or Lo.

The authors propose adopting an initial value K(0), of K, and estimating a(0), b(0) and r2(0) by simple linear regression between y (= Lt) and x(=ek(0)). The procedure may be repeated for several values of K, that is, K(1) K(2),.... One can then adopt the regression that results in the larger value of r2, to which Kmax, amax and bmax correspond. From the values of amax, bmax and Kmax one can obtain the values of the remaining parameters.

Notes

One practical process towards finding Kmax can be:

(i) To select two extreme values of K which include the required value, for example K = 0 and K = 2 (for practical difficulties, use K = 0.00001 instead of K = 0).

(ii) Calculate the 10 regressions for equally-spaced values of K between those two values in regular intervals.

(iii) The corresponding 10 values of r2 will allow one to select two new values of K which determine another interval, smaller than the one in (i), containing another maximum value of r2.

(iv) The steps (ii) and (iii) can be repeated until an interval of values of K with the desired approximation is obtained. Generally, the steps do not need many repetitions.

Estimation Of M - Natural Mortality Co-efficient

Several methods were proposed to estimate M and they are based on the association of M with other biological parameters of the resource. These methods can produce approximate results.

Relation of m with the longevity,

Longevity: Maximum mean age t_λ of the individuals in a non-exploited population.

Duration of the exploitable life: $t_\lambda - t_r = \lambda$ (Figure)

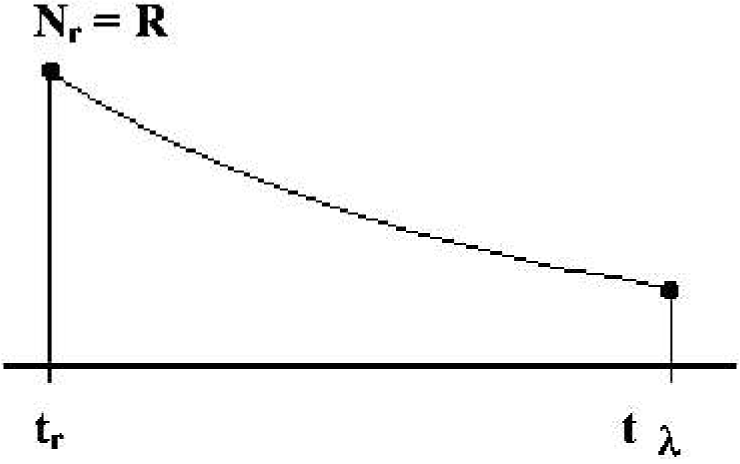


Figure: *Duration of the exploitable life*

Tanaka (1960) proposes "NATURAL" Survival Curves (Figure) to obtain the values of M from longevity.

A cohort practically vanishes when only a fraction, p, of the recruited individuals survives. In that case, $N_\lambda = R \cdot e^{-M.T_\lambda}$ and it can be written:

$$\frac{N_L}{R} = e^{-M.T_m} \text{ and so } M = -(1/\lambda).\ln p$$

Different values of the survival fraction produce different survival curves of M in function of λ.

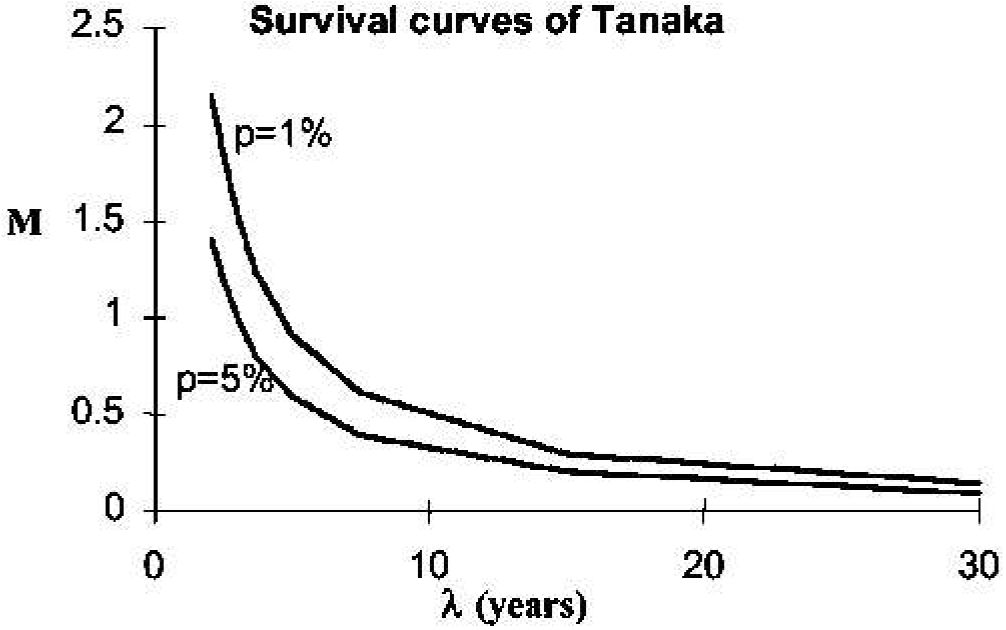


Figure: *Survival curves by Tanaka*

Any value of p can be chosen, for instance, p = 5%, (i.e. one in each twenty recruits survives until the age t_λ) as variable value of the survival curves.

Relation between M and Growth

Beverton and Holt Method (1959)

Gulland (1969) mentions that Beverton and Holt verified that species with a larger mortality rate M also presented larger values of K. Looking for a simple relation between these two parameters, they concluded approximately that:

$1 \leq \frac{M}{K} \leq 2$ for small pelagic fishes

$2 \leq \frac{M}{K} \leq 3$ for demersal fishes

Pauly Method (1980)

Based on the following considerations:

1. Resources with a high mortality rate cannot have a very big maximum size;
2. In warmer waters, the metabolism is accelerated, so the individuals can grow up to a larger size and reach the maximum size faster than in colder waters.

Based on data of 175 species, Pauly adjusted multiple linear regressions of transformed values of M against the corresponding transformed values of K, L_∞ and temperature, T, and selected one that was considered to have a better adjustment, that is, the following empirical relation:

$$\text{InM} = -\ 0.0152 - 0.0279.\text{InL}_\infty + 0.6543.\text{Ink} + 0.463.\text{InT}^\text{o}$$

with the parameters expressed in the following units:

M = year-1

L_∞ = cm of total length

K = year-1

T° = surface temperature of the waters in °C

Pauly highlights the application of this expression to small pelagic fishes and crustaceans. The Pauly relation uses decimal logarithms to present the first coefficient different from the value -0.0152 which was given in the previous expression, written with natural logarithms.

Relation between M and Reproduction

Rikhter and Efanov Method (1976)

These authors analysed the dependency between M and the age of first (or 50 percent) maturity. They used data from short, mean and long life species and suggested the following relation of M with the, t_{mat}, age of 1st maturity:

(Units)

$$M = \frac{1.521}{(t_{mat50\%})^{0.720}} - 0.155 \begin{pmatrix} t_{mat50\%} \rightarrow \text{year} \\ M \rightarrow \text{year}^{-1} \end{pmatrix}$$

Gundersson Method (1980)

Based on the assumption that the natural mortality rate should be related to the investment of the fish in reproduction, beyond the influence of other factors, Gundersson established several relations between M and those factors.

He proposed, however, the following simple empirical relation, using the Gonadosomatic Index (GSI) (estimated for mature females in the spawning period) in order to calculate M:

M = 4.64 x GSI - 0.37

Knowing the Stock Age Structure, at Beginning and End of Year, and Catches in Number, by Age, during the Year

The natural mortality co-efficients M_i, at age i can be calculated from the catch, C_i, in numbers and the survival numbers, N_i and N_{i+1} at the beginning and end of a year, by following the steps:

calculate $E_i = \frac{C_i}{N_i - N_{i+1}}$

calculate $Z_i = \text{In}N_i - \text{In}N_{i+1}$

calculate $M_i = Z_i .(1 - E_i)$

The several values of M obtained in each age could be combined to calculate a constant value, M, for all ages.

Paloheimo Method (1961)

Let us consider the supposition that F_i is proportional to f_i for several years i, that is

$F_i = q.\frac{f_i}{T_i}$ for T_i = 1 year, $F_i = q \cdot f_i$,

then:

$Z_i = q \cdot f_i + M$

So, the linear regression between Z_i and fi has a slope b = q and an intercept a = M.

Estimation of Z - Total Mortality Co-efficient

There are several methods of estimating the total mortality co-efficient, Z, assumed to be constant during a certain interval of ages or years.

It is convenient to group the methods, according to the basic data, into those using ages or those using lengths.

Methods Using Age Data

The different methods are based on the general expression of the number of survivors of a cohort, at the instant t, submitted to the total mortality, Z, during an interval of time, that is:

$$N_t = N_a.e^{-z(t-t_a)}$$

Z is supposed to be constant in the interval of time (ta,tb).

Taking logarithms and re-arranging the terms, the expression will be:

$\ln N_t = C_{te} - Z.t$

where Cte is a constant (= ln N_a + Z_{ta}).

This expression shows that the logarithm of the number of survivors is linear with the age, being the slope equal to -Z.

Any constant expression which does not affect the determination of Z will be referred to as Cte.

1. If Z can be considered constant inside the interval (ta,tb) and having available abundance data, Ni or indices of abundance in number, Ui in several ages, i, then, the application of the simple linear regression allows one to estimate the total mortality coefficient Z.

   In fact

   $$\overline{N}_i = N_i.\frac{1 - e^{-zt_i}}{ZT_i} \quad \text{so } \overline{N}_i = N_i.\text{Consant}$$

   and as

   $$N_i = N_a.e^{-z(t_i - t_{a)}}$$

   then, by substitution:

   $$\overline{N}_i = \text{Cte}.e^{-Zt_i} \quad (T_i = \text{const} = 1 \text{ year})$$

   and also

   $$\text{In}\overline{N}_i = \text{Cte} - Zt_i$$

   The simple linear regression between $\text{In}\overline{N}_i$ and ti allows the estimation of Z (notice that the constant, Cte is different from the previous one. In this case only the slope matters to estimate Z).

2. If ages are not at constant intervals, the expression could be approximated and expressed in terms of the tcentrali. For T_i variable, it will be:

   Ni ˜ Ni. e-ZTi/2

and, as Ni = Na. e -Z.(ti-ta)

it will be Ni ˜ Cte. e-Ztcentrali

and finally: ln Ni ˜ Cte - Z. tcentrali

3. When using indices Ui, the situation is similar because Ui = q. Ni, with q constant, and then, also:

$$\text{In}\overline{\text{U}}_{\text{i}} = \text{Cte - Zt}_{\text{i}}$$

The simple linear regression between $\text{In}\overline{\text{U}}_{\text{i}}$ and ti allows one to estimate Z.

4. If the intervals are not constant, the expression should be modified to:

$$\text{In}\overline{\text{U}}_{\text{i}} \approx \text{Cte - t}_{\text{centrali}}$$

Simple linear regression can be applied to obtain Z, from catches, Ci, and ages, ti, supposing that Fi is constant.

$\text{C}_{\text{i}} = \text{F}_{\text{i}}\overline{\text{N}}_{\text{i}}\text{T}_{\text{i}}$ and so, $\text{InC}_{\text{i}} = \text{Cte} + \text{In}\overline{\text{N}}_{\text{i}}$ when Ti is constant. So:

lnCi = Cte - Z. ti

5. If the intervals are not constant, the expression should be modified to:

lnCi/Ti ˜ Cte - Z. tcentrali

6. Let Vi be the cumulative catch from ti until the end of the life, then:

Vi = ∑ Ck = ∑ Fk Nkcum,

Where the sum goes from the last age until age i,

As Fk and Zk are supposed to be constant ∑ Nkcum = Ni/Z and so:

Vi = FN/Z and lnVi = Cte + lnNi

Therefore:

ln Vi = Cte - Z. ti

7. Following Beverton and Holt (1956), Z can be expressed as:

$$\text{Z} = \frac{1}{\bar{\text{t}} - \text{t}_{\text{a}}}$$

Then, it is possible to estimate Z from the mean age t

This expression was derived, considering the interval (ta, tb) as (ta, 8).

Methods Using Length Data

When one has available data by length classes instead of by age, the methods previously referred to can still be applied. For that purpose, it is convenient to define the relative age.

Using the von Bertalanffy equation one can obtain the age t in function of the length, as:

(the expression is written in the general form in relation to ta and not to t0)

$$\text{t} = \text{t}_a - \frac{1}{\text{K}}.\text{In}\left(1 - \frac{\text{L}_\infty - \text{L}_t}{\text{L}_\infty - \text{L}_a}\right)$$

or

$$t = t_a - \frac{1}{K}.\text{In}\left(1 - \frac{L_t - L_a}{L_\infty - L_a}\right)$$

(This equation is referred to by some authors as the inverse von Bertalanffy equation).

The difference t-ta is called relative age, t*,.

So: t* =-(1/K).ln[(L8- Lt)/(L8- La)] or t* =-(1/K)ln[1-(Lt-La)/ (L8- La)]

For ta = to, La = 0 and:

$$t^* = \frac{1}{K}.\text{In}\left(1 - \frac{L_t}{L_\infty}\right)$$

t* is called a relative age because the absolute ages, t, are related to a constant age, ta.

In this way, the duration of the interval Ti can either be calculated by the difference of the absolute ages or by the difference of the relative ages at the extremes of the interval:

Ti = ti+1 -ti = t*i +1 - t*i

Also:

t*centrali = tcentrali + Cte

$$\bar{t}^* = \bar{t} + Cte$$

So, the previous expressions still hold when the absolute ages are replaced by the relative ages:

ln Ni = Cte - Z. t*centrali

ln Ui = Cte - Z. t*centrali

ln Vi = Cte - Z. t*i

ln Ci/Ti = Cte - Z. t*centrali

Finally, the expression would also be:

$$Z = \frac{1}{t^*}$$

Beverton and Holt (1957) proved that:

$$Z = K.\frac{L_\infty - \bar{L}}{\bar{L} - L_a}$$

$\bar{L}$ must be calculated as the mean of the lengths weighted with abundances (or their indices) or with the catches in numbers.

Comments

1. The application of any of these methods must be preceeded by the graphical representation of the corresponding data, in order to verify if the assumptions of the methods are acceptable or not and also to determine the adequate interval, (ta, tb).
2. These formulas are proved with the indications that were presented, but it is a good exercise to develop the demonstrations as they clarify the methods.
3. It is useful to estimate a constant Z, even when it is not acceptable, because it gives a general orientation about the size of the values one can expect.

Notes

4. The methods are sometimes referred to by the names of the authors. For example, the expression ln Vi = Cte - Z.t*i is called the Jones and van Zalinge method (1981).
5. The mean age as well as the mean length in the catch can be calculated from the following expressions:

$$\bar{t} = \frac{\sum(t_{\text{centrali}}.C_i)}{\sum C_i}$$

with Ci = catch in number in the age class i

$$\overline{L} = \frac{\sum(L_{\text{centrali}}.C_i)}{\sum C_i}$$

where Ci = catch in number in the length class i

$$\bar{t} = \frac{\sum(t^*_{\text{centrali}}.C_i)}{\sum C_i}$$

with Ci = catch in number in the age class.

The relative age should be t* = - (1/K).ln[(L8- Lt)/(L8- La)]

Estimation of the Parameters of the Stock-Recruitment (S-R) Relation

The least squares method (non-linear model) can be used to estimate the parameters, a and k, of any of the S-R models.

The initial values of the Beverton and Holt model (1957) can be obtained by re-writing the equation as:

$$\frac{R}{S} - 1 \text{ or} \frac{S}{R} = \frac{1}{\alpha} + \frac{1}{\alpha K}.S$$

and estimating the simple linear regression between y (= S/R) and x (=S) which will give the estimations of 1/a and 1/(ak). From these values, it will then be possible to estimate the parameters a and k. These values can be considered as the initial values in the application of the non-linear model.

In the Ricker model (1954) the parameters can be obtained by re-writing the equation as:

$$\text{In}\frac{R}{S} = \text{In}\alpha - \frac{1}{K}.S$$

and applying the simple linear regression between y (= ln R/S) and x (=S) to estimate ln a and (-1/k). From these values, it will be possible to estimate the parameters (a and k) of the model, which can be considered as the initial values in the application of the non-linear model.

It is useful to represent the graph of y against x in order to verify if the marked points are adjustable to a straight line before applying the linear regression in any of these models.

In the models with the flexible parameter, c, like for example, the Deriso model (1980), the equation can be re-written as:

$$\left(\frac{R}{S}\right)^{c} = \alpha^{c} - c.\alpha^{c}.\frac{S}{K}$$

For a given value of c the linear regression between y (= (R/S)c) and x (=S) allows the estimation of the parameters a and k.

One can try several values of c to verify which one will have a better adjustment with the line y against x; for example, values of c between -1 and 1.

The values thus obtained for a, k and c, can be considered as initial values in the application of the iterative method, to estimate the parameters a, k and c of the non-linear Deriso model.

Estimation of the Matrix [F] and of the Matrix [N] - Cohort Analysis - AC and LCA

Cohort Analysis by Age - (AC)

The cohort analysis is a method to estimate the fishing mortality coefficients, Fi and the number of survivors, Ni, at the beginning of each age, from the annual structures of the stock catches, in number, over a period of years.

More specifically, consider a stock where the following is known:

Data

age, i, where i = 1, 2,...,k

year, j, where j = 1, 2,...,n

Matrix of catches [C] with

$C_{i,j}$ = Annual catch, in number, of the individuals with the age i and during the year j

Matrix of natural mortality [M] with

$M_{i,j}$ = natural mortality coefficient, at the age i and in the year j.

Vector [T] where

T_i = Size of the age interval i (in general, Ti = T = 1 year)

Objective

To estimate

matrix [F]

and

matrix [N].

In the resolution of this problem, it is convenient to consider these estimations separately; one interval of age i (part 1); all the ages during the life of a cohort (part 2); and finally, all the ages and years (part 3).

Part 1 (Interval Ti)

Consider that the following characteristics of a cohort, in an interval Ti are known:

C_i = Catch in number

M_i = Natural mortality co-efficient

T_i = Size of the interval

Adopting a value of Fi, it is then possible to estimate the number of survivors at the beginning, Ni, and at the end, Ni+1, of the interval.

Notes

In fact, from the expression:

$$C_i = \frac{F_i}{(F_i + M_i)}.N_i.\left(1 - e^{-(F_i + M_i).T_i}\right)$$

one can calculate Ni which is the only unknown variable in the expression.

To calculate Ni+1 one can use the expression $N_{i+1} = N_i.e^{-(F_i + M_i).T_i}$ where the values Ni, Fi and Mi were previously obtained.

Part 2 (During the Life)

Suppose now that the catches Ci of each age i, of a cohort during its life, the values of Mi and the sizes of the interval Ti are known.

Adopting a certain value, Ffinal, for the Fishing Mortality Coefficient in the last class of ages, it is possible, as mentioned in part 1, to estimate all the parameters (related to numbers) in that last age group. In this way, one will know the number of survivors at the beginning and end of the last age.

The number at the beginning of that last class of ages, is also the number Nlast at the end of the previous class, that is, Nfinal is the initial number of survivors of the class before last.

Using the Ci expression, resulting from the combination of the two expressions above:

$$C_i = \frac{F_i}{(F_i + M_i)}.N_{\text{final}}.\left(e^{(F_i + M_i).T_i} - 1\right)$$

one can estimate Fi in the previous class, which is the only unknown variable in the expression. The estimation may require iterative methods or trial and error methods.

Finally, to estimate the number Ni of survivors at the beginning of the class i, the following expression can be used:

$$N_i = N_{\text{final}}.e^{(F_i + M_i).T_i}$$

Repeating this process for all previous classes, one will successively obtain the parameters in all ages, until the first age.

In the case of a completely caught cohort, the number at the end of the last class is zero and the catch C has to be expressed as:

$$C_{\text{final}} = \frac{F_{\text{final}}}{(F_{\text{final}} + M_i)}.N_{\text{final}}$$

Pope Method

Pope (1972) presented a simple method to estimate the number of survivors at the beginning of each age of the cohort life, starting from the last age.

It is enough to apply successively in a backward way, the expression:

Ni ˜ (Ni+1 e MT/2 + Ci).e MT/2

Pope indicates that the approximation is good when MT = 0.6

Pope’s expression is obtained, supposing that the catch is made exactly at the central point of the interval Ti (Figure).

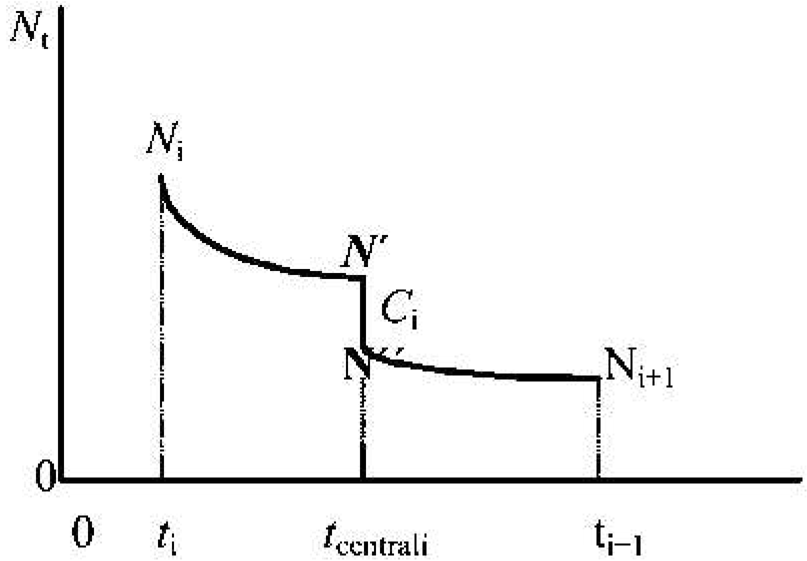


Figure: *Number of survivors during the interval Ti = ti+1 - ti with the catch extracted at the central point of the interval*

Proceeding from the end to the beginning one calculates successively:

N" = Ni+1e+MTi/2

N’ = N" + Ci

Ni = N’.e+MT/2

substituting N’ by N"+Ci, the expression will be:

Ni = (N" + Ci).e MT/2

Finally, substituting N" by Ni+1.e +MTi/2, it will be:

$$N_i \approx \left(N_{i+1}e^{MT/2} + C_i\right).e^{MT/2}$$

Part 3 (Period of Years)

Let us suppose now that the Catch matrix [C], the natural mortality [M] matrix and the vector size of the intervals [T], are known for a period of years.

Let us also assume that the values of F in the last age of all the years represented in the matrices and the values of F of all the ages of the last year were adopted. These values will be designated by Fterminal (Figure).

Figure: *Matrix of catch, [C], with Fterminal in the last line and in the last column of the matrix C. The shadowed zones exemplify the catches of a cohort*

		Years			
Ages	2000	2001	2002	2003	
1	C	C	C	C	Fterminal
2	C	C	C	C	Fterminal
3	C	C	C	C	Fterminal
	F*terminal*	F*terminal*	F*terminal*	F*terminal*	

Notice that in this matrix the elements of the diagonal correspond to values of the same cohort, because one element of a certain age and a certain year will be followed, in the diagonal, by the element that is a year older.

From parts 1 and 2 it will then be possible to estimate successively Fs and Ns for all the cohorts present in the catch matrix.

Comments

1. The values of $M_{i,j}$ are considered constant and equal to M, when there is no information to adopt other values.
2. When data is referred to ages, the values T_i will be equal to 1 year.
3. The last age group of each year is, sometimes grouped ages(+). The corresponding catches are composed of individuals caught during those years, with several ages. So, the cumulative values do not belong to the same cohorts, but are survivors of several previous cohorts with different recruitments and submitted to different fishing patterns. It would not be appropriate to use the catch of a group (+) and to apply cohort analysis. Despite this fact, the group (+) is important in order to calculate the annual totals of the catches in weight, Y, of total biomasses, B and the spawning stock biomass. So, it is usual to start with the cohort analysis on the age immediately before the group (+) and use the group (+) only to calculate the annuals Y, B and (SP). The value of F in that group (+) in each year, can be estimated as being the same fishing mortality coefficient as the previous age or, in some cases, as being a reasonable value in relation to the values of Fi in the year that is being considered.
4. A difficulty in the technical application appears when the number of ages is small or when the years are few. In fact, in those cases, the cohorts have few age classes represented in the Matrix [C] and the estimations will be very dependent on the adopted values of $F_{terminals}$.
5. The cohort analysis (CA) has also been designated as: VPA (Virtual Population Analysis), Derzhavin method, Murphy method, Gulland method, Pope method, Sequential Analysis, etc. Sometimes, CA is referred to when the Pope formula and the VPA are used in other cases. Megrey (1989) presents a very complete revision about the cohort analyses.
6. It is also possible to estimate the remaining parameters in an age i, related to numbers, that is, N_{cumi}, Ni, Di, Zi and Ei. When the information on initial individual or mean weights matrices [w] or [w] are available, one can also calculate the matrices of annual catch in weight [Y], of biomasses at the beginning of the years, [B], and of mean biomasses during the years [B]. If one has information on maturity ogives in each year, for example at the beginning of the year, spawning biomasses [SP] can also be calculated. Usually, only the total catches Y, the stock biomasses (total and spawning) at the beginning and the mean biomasses of the stock (total and spawning) in each year are estimated.
7. The elements on the first line of the matrix [N] can be considered estimates of the recruitment to the fishery in each year.
8. The fact that the $F_{terminals}$ are adopted and that these values have influence on the resulting matrix [F] and matrix [N], forces the selection of values of $F_{terminals}$ to be near the real ones. The agreement between the estimations of the parameters mentioned in the points 6 and 7. and other independent data or indices (for example, estimations by acoustic methods of recruitment or biomasses, estimations of abundance indices or cpue´s, of fishing efforts, etc) must be analysed.

9. The hypothesis that the exploitation pattern is constant from year to year, means that the fishing level and the exploitation pattern can be separated or Fsepi = Fj x si. This hypothesis can be tested based on the matrix [F] obtained from the cohort analysis.

   It is usual to call this separation VPA-Separable (SVPA).

   We have $\sum_{i} F_{ij} = F_{tot_i}$

   and $\sum_{j} F_{ij} = s_{tot_i}$

   and $\sum_{ij} F_{ij} = F_{tot}$

   Then, if Fij = Fj.si one can prove that $F_j.s_i = \left(F_{tot_j}.s_{tot_i}\right)/F_{tot}$.

   If the estimated values of Fij are the same as the previous Fsepij = Fj.si then the hypothesis is verified. This comparison can be carried out in two different ways, the simplest is to calculate the quotients Fsepij /Fij. If the hypothesis is true this quotient is equal to one. If the hypothesis is not verified it is always possible to consider other hypotheses with the annual vector [s] constant in some years only, mainly the last years.

10. It is usual to consider an interval of ages, where it can be assumed that the individuals caught are "completely recruited". In that case, the interval of ages corresponds to exploitation pattern constant (for the remaining ages, not completely recruited, the exploitation pattern should be smaller). For that interval of ages, the means of the values of Fi,j in each year are then calculated. Those means, Fj, are considered as fishing levels in the respective years. The exploitation pattern in each cell, would then be the ratio Fi,j / Fj. The si, for the period of years considered, can be taken as the mean of the relative pattern of exploitation calculated before. Alternatively, they can also be taken as referring to si of an age chosen for reference.

Length Cohort Analysis - (LCA)

The technique of the cohort ansalysis, applied to the structure of the catches of a cohort during its life, can be made with non constant intervals of time, Ti,. This means that the length classes structure of the catches of a cohort during its life, can also be analysed.

The methods of analysis of the cohort in those cases is called the LCA (Length Cohort Analysis). The same techniques; Pope method, iterative method, etc., of the CA for the ages, can be applied to the LCA analysis (the intervals Ti´s can be calculated from the relative ages).

One way to apply the LCA to the length annual catch compositions, will be: to group the catches of length classes belonging to the same age interval in each year. The technique CA can then be applied directly to the resulting age composition of the catches by age of the matrix [C]. This technique is known as "slicing" the length compositions. To "slice", one usually inverts the von Bertalanffy length growth equation and estimates the age ti for each length Li (sometimes using the relative ages t*i). It is possible that when grouping the length classes of the respective age interval, there are length classes composed by elements that belong to two consecutive age groups. In these cases, it will be necessary to "break" the catch of these extreme classes into two parts and distribute them to each of those ages. In the example of Figure, the catches of the length class (24-26] belong to age 0 and to age 1.

Notes

So, it is necessary to distribute that catch to the two ages. One simple method is to attribute to age 0 the fraction (1.00 - 0.98)/(1.06 - 0.98) = 0.25 of the annual catch of that length class and to age 1 the fraction (1.06 - 1.00)/(1.06 - 0.98) = 0.75. The method may not be the most appropriate one, because it is based on the assumption that, in the length classes, the distribution of the individuals by length is uniform. So, it is necessary to use the smallest possible interval of length classes, when applying this distribution technique.

Another way to do the length cohort analysis is to use the catches in the length classes of the same age group. It is possible to follow the cohorts in the matrix [C], through the length classes belonging to a same age, in a certain year, with the length classes of the next age, in the following year, etc. In this way, the different cohorts existing in the matrix will be separated and the evolution of each one of them will be by length classes, not by age (see Figure).

Figure: *Example of a matrix [C] with the catches of the cohort shadowed, written in bold, recruited at year 2000, "sliced" by length classes,*

Group			*Years*			
Age	*Relative*	*Length*				
	age	*Classes*	*2000*	*2001*	*2002*	*2003*
0	1.03	20-	41	30	17	49
	1.54	22-	400	292	166	472
	1.98	24-	952	699	400	1127
1	2.06	26-	1766	1317	757	2108
	2.30	28-	2222	1702	985	2688
	2.74	30-	2357	1872	1093	2902
	2.88	32-	2175	1091	1067	2739
2	3.00	34-	1817	948	1416	1445
	3.42	36-	1529	812	1270	1250
	3.64	38-	1251	684	980	1053
	3.83	40-	1003	560	702	710
	3.96	42-	787	290	310	558
3	4.01	44-	595	226	179	834
	4.25	46-	168	70	71	112

Cohort of the year 2000

The LCA R. Jones method (1961), of analysing a length composition during the life of a cohort can then be applied. The different values of Ti are calculated as Ti = ti+1*-ti*, where ti* e ti+1* are the relative ages corresponding to the extremes of the length interval i. The vector [N] can also be obtained as the number of initial survivors in each length class of the cohort and in each age class.

Comments on Cohort Analyses

1. Certain models, called integrated models, combine all the available information (catches, data collected on research cruises, effort and cpue data, etc) with the matrix [C], and integrate in a unique model, in order to optimize the previously

defined criterium function. A model integrating CA and the hypothesis of constant exploitation pattern was developed and called SVPA, separable VPA, because the Fishing level and Exploitation pattern are "separable".

2. Fry (1949) considered the cumulative catches of a cohort by age during its life, from the end to the beginning, as an image of the number of survivors at the beginning of each age (which the author designated as "virtual population"):

$$N_i = \sum_{k=\text{final}}^{i} C_k = V_i = N_i \text{virtual}$$

In the fishery that Fry studied, M was practically equal to zero.

If M is different from zero it can also be said that the number Ni of survivors at the beginning of the interval i will be

$$N_i = \sum_{k=\text{final}}^{i} D_k$$

where, Dk represents the number of total deaths at the interval k.

Adopting the initial values, Ek(0), for the exploitation rates, E, in all the classes, one can calculate the total deaths:

$D_{k(0)} = C_k/E_{k(0)}$.

Ni(0) can be calculated as the cumulative total deaths from the last class up to the ith class, that is:

$$N_{i(0)} = \sum_{k=\text{ult}}^{i} D_{k(o)}$$

Then the expression will be:

Zi(1).Ti = ln(Ni+1(0)/Ni(0))

and:

Fi(1).Ti = Ei(0).Zi(1).Ti

Comparing Ei(1) with Ei(0), the new values of E will be:

Ei(1) = Fi(1).Ti/ Fi(1).Ti + Mi.Ti

One can then estimate the values of E with the desired approximation by an iterative method, repeating the five calculations (of Di, Ni, ZiTi, FiTi and Ei,) using Ei(1) instead of Ei(0).

In the last class, the number, N_{last}, can be taken as equal to the number of deaths, D_{last} and in this case, N_{last} will be calculated as:

$N_{last} = D_{last} = C_{last} / E_{last}$

3. Finally, the results of CA and of LCA give a perspective view of the stock in the previous years. That information is useful for the short and long-term projections. Usually, data concerning the catches is not available for the year in which the assessment is done and so it is necessary to project the catches and the biomasses to the beginning of the present year before calculating the short-term projection.

4. When the relative ages are calculated, it is usual to adopt zero as the age ta corresponding to the value of La, taken as the lower limit of the first length class represented in the catches.

2.5 GAUSS-MARKOV THEORY

The Gauss–Markov theorem, named after Carl Friedrich Gauss and Andrey Markov, states that in a linear regression model in which the errors have expectation zero, are uncorrelated and have equal variances, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists. Here "best" means giving the lowest variance of the estimate, as compared to other unbiased, linear estimators. The errors do not need to be normal, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance). The requirement that the estimator be unbiased cannot be dropped, since biased estimators exist with lower variance. See, for example, the James–Stein estimator (which also drops linearity) or ridge regression.

The Gauss Markov theorem tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the best linear unbiased estimate (BLUE) possible.

Gauss Markov Assumptions

There are five Gauss Markov assumptions (also called conditions):

Linearity: the parameters we are estimating using the OLS method must be themselves linear.

Random: our data must have been randomly sampled from the population.

Non-Collinearity: the regressors being calculated aren’t perfectly correlated with each other.

Exogeneity: the regressors aren’t correlated with the error term.

Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

Purpose of the Assumptions

The Gauss Markov assumptions guarantee the validity of ordinary least squares for estimating regression coefficients.

Checking how well our data matches these assumptions is an important part of estimating regression co-efficients. When you know where these conditions are violated, you may be able to plan ways to change your experiment setup to help your situation fit the ideal Gauss Markov situation more closely.

In practice, the Gauss Markov assumptions are rarely all met perfectly, but they are still useful as a benchmark, and because they show us what ‘ideal’ conditions would be. They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.

The Gauss-Markov Assumptions in Algebra

We can summarize the Gauss-Markov Assumptions succinctly in algebra, by saying that a linear regression model represented by

yi = xi‘ ß + ei

and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if

E{ei} = 0, i = 1, . . . , N

{e1……en} and {x1…..,xN} are independent

cov{ei, ej} = 0, i, j = 1,…., N I ? j.

V{e1 = s2, i= 1, ….N

The first of these assumptions can be read as “The expected value of the error term is zero.”. The second assumption is collinearity, the third is exogeneity, and the fourth is homoscedasticity.

In many estimation problems, the MVUE or a sufficient statistics cannot be found or indeed the PDF of the data is itself unknown (only the second-order statistics are known in the sense that they can be estimated from data). In such cases, one solution is to assume a functional model of the estimator, as being linear in the data, and find the linear estimator which is unbiased and has minimum variance. This estimator is referred to as the best linear unbiased estimator (BLUE).

Consider the general vector parameter case θ, the estimator is required to be a linear function of the data, i.e.,

$$\theta = Ax$$

The first requirement is that the estimator should be unbiased, i.e.,

$$E(\theta) = AE(x) = \theta$$

which can be only satisfied if:

$$E(x) = H\theta \Rightarrow AH = I$$

The BLUE is derived by finding the A which minimizes the variance, $C_\theta = A^TCA$ subject to the constraint AH = I, where C is the covariance matrix of the data x. Carrying out the minimization yields the following form for the BLUE:

$$\theta = Ax = (H^TC^{-1}H)^{-1}H^TC^{-1}(x)$$

where, $C_\theta = (H^TC^{-1}H)^{-1}$

Salient Attributes of BLUE:

- For the general linear model, the BLUE is identical in form to the MVUE.
- The BLUE only assumes only up to 2nd-order statistics and not the complete PDF of the data unlike the MVUE which was derived assuming Gaussian PDF.
- If the data is truly Gaussian then the BLUE is also the MVUE.

The BLUE for the general linear model can be stated in terms of following theorem.

Gauss-Markov Theorem: Consider a general data model of the form:

$x = H\theta + w$

where H is known and w is noise with covariance C (the PDF of w otherwise arbitrary).

Then the BLUE of θ is:

$$\theta = (H^TC^{-1}H)^{-1}H^TC^{-1}x$$

where, $C_\theta = (H^TC^{-1}H)^{-1}$ is the minimum covariance matrix.

2.6 STANDARD ERRORS

The standard error (SE) of a statistic (usually an estimate of a parameter) is the standard deviation of its sampling distribution or an estimate of that standard deviation. If the parameter or the statistic is the mean, it is called the standard error of the mean (SEM).

The sampling distribution of a population mean is generated by repeated sampling and recording of the means obtained. This forms a distribution of different means and this distribution has its own mean and variance. Mathematically, the variance of the sampling distribution obtained is equal to the variance of the population divided by the sample size. This is because as the sample size increases, sample means cluster more closely around the population mean.

Therefore, the relationship between the standard error and the standard deviation is such that, for a given sample size, the standard error equals the standard deviation divided by the square root of the sample size. In other words, the standard error of the mean is a measure of the dispersion of sample means around the population mean.

In regression analysis, the term "standard error" refers either to the square root of the reduced chi-squared statistic or the standard error for a particular regression coefficient (as used in, e.g., confidence intervals).

Standard error plays a very crucial role in the large sample theory. It also may form the basis for the testing of a hypothesis. The statistical inference involved in the construction of the confidence interval is mainly based on standard error.

The magnitude of the standard error gives an index of the precision of the estimate of the parameter. It is inversely proportional to the sample size, meaning that smaller samples tend to produce greater standard errors.

The standard deviation of a sample is generally designated by the Greek letter sigma (s). It can also be defined as the square root of the variance present in the sample.

Standard Errors of Estimates

The term "standard error" is used to refer to the standard deviation of various sample statistics such as the mean or median. For example, the "standard error of the mean" refers to the standard deviation of the distribution of sample means taken from a population. The smaller the standard error, the more representative the sample will be of the overall population.

The standard error is also inversely proportional to the sample size; the larger the sample size, the smaller the standard error because the statistic will approach the actual value.

The standard error is considered part of descriptive statistics. It represents the standard deviation of the mean within a dataset. This serves as a measure of variation for random variables, providing a measurement for the spread. The smaller the spread, the more accurate the dataset.

Standard Error and Population Sampling

When a population is sampled, the mean, or average, is generally calculated. The standard error can include the variation between the calculated mean of the population and one which is considered known, or accepted as accurate. This helps compensate for any incidental inaccuracies related to the gathering of the sample.

In cases where multiple samples are collected, the mean of each sample may vary slightly from the others, creating a spread among the variables. This spread is most often measured as the standard error, accounting for the differences between the means across the datasets.

The more data points involved in the calculations of the mean, the smaller the standard error tends to be. When the standard error is small, the data is said to be more representative of the true mean. In cases where the standard error is large, the data may have some notable irregularities.

Standard Deviation and Standard Error

The standard deviation is a representation of the spread of each of the data points. The standard deviation is used to help determine the validity of the data based the number of data points displayed at each level of standard deviation. Standard error functions more as a way to determine the accuracy of the sample or the accuracy of multiple samples by analyzing deviation within the means.

2.7 ESTIMATOR OF ERRORS

The mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

The MSE is a measure of the quality of an estimator it is always non-negative, and values closer to zero are better.

The MSE is the second moment (about the origin) of the error, and thus incorporates both the variance of the estimator (how widely spread the estimates are from one data sample to another) and its bias (how far off the average estimated value is from the truth). For an unbiased estimator, the MSE is the variance of the estimator. Like the variance, MSE has the same units of measurement as the square of the quantity being estimated. In an analogy to standard deviation, taking the square root of MSE yields the root-mean-square error or root-mean-square deviation (RMSE or RMSD), which has the same units as the quantity being estimated; for an unbiased estimator the RMSE is the square root of the variance, known as the standard error.

2.8 CONTROL LIMIT THEOREM

In probability theory, the central limit theorem (CLT) establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed. The theorem is a key ("central") concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.

For example, suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on

Notes

the values of the other observations and that the arithmetic average of the observed values is computed. If this procedure is performed many times, the central limit theorem says that the distribution of the average will be closely approximated by a normal distribution. A simple example of this is that if one flips a coin many times the probability of getting a given number of heads in a series of flips will approach a normal curve, with mean equal to half the total number of flips in each series. (In the limit of an infinite number of flips, it will equal a normal curve.)

The central limit theorem has a number of variants. In its common form, the random variables must be identically distributed. In variants, convergence of the mean to the normal distribution also occurs for non-identical distributions or for non-independent observations, given that they comply with certain conditions.

The earliest version of this theorem, that the normal distribution may be used as an approximation to the binomial distribution, is now known as the de Moivre–Laplace theorem.

In more general usage, a central limit theorem is any of a set of weak-convergence theorems in probability theory. They all express the fact that a sum of many independent and identically distributed (i.i.d.) random variables, or alternatively, random variables with specific types of dependence, will tend to be distributed according to one of a small set of attractor distributions. When the variance of the i.i.d. variables is finite, the attractor distribution is the normal distribution. In contrast, the sum of a number of i.i.d. random variableswith power law tail distributions decreasing as $|x|^{-\alpha-1}$ where $0 < \alpha < 2$ (and therefore having infinite variance) will tend to an alpha-stable distribution with stability parameter (or index of stability) of a as the number of variables grows.

2.9 MAXIMUM LIKELIHOOD ESTIMATOR

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model, given observations. MLE attempts to find the parameter values that maximize the likelihood function, given the observations. The resulting estimate is called a maximum likelihood estimate, which is also abbreviated as MLE.

The method of maximum likelihood is used with a wide range of statistical analyses. As an example, suppose that we are interested in the heights of adult female penguins, but are unable to measure the height of every penguin in a population (due to cost or time constraints). Assuming that the heights are normally distributed with some unknown mean and variance, the mean and variance can be estimated with MLE while only knowing the heights of some sample of the overall population. MLE would accomplish that by taking the mean and variance as parameters and finding particular parametric values that make the observed results the most probable given the normal model.

From the point of view of Bayesian inference, MLE is a special case of maximum a posteriori estimation (MAP) that assumes a uniform prior distribution of the parameters. In frequentist inference, MLE is one of several methods to get estimates of parameters without using prior distributions. Priors are avoided by not making probability statements about the parameters, but only about their estimates, whose properties are fully defined by the observations and the statistical model.

2.10 NORMALITY OF ERRORS, CONTROL LIMIT THEOREM, MAXIMUM LIKELIHOOD ESTIMATOR

A large p-value and hence failure to reject this null hypothesis is a good result. It means that it is reasonable to assume that the errors have a normal distribution. Typically, assessment of the appropriate residual plots is sufficient to diagnose deviations from normality.

To complement the graphical methods just considered for assessing residual normality, we can perform a hypothesis test in which the null hypothesis is that the errors have a normal distribution. A large p-value and hence failure to reject this null hypothesis is a good result. It means that it is reasonable to assume that the errors have a normal distribution. Typically, assessment of the appropriate residual plots is sufficient to diagnose deviations from normality. However, more rigorous and formal quantification of normality may be requested. So this section provides a discussion of some common testing procedures (of which there are many) for normality. For each test discussed below, the formal hypothesis test is written as:

H0: the errors follow a normal distribution HA: the errors do not follow a normal distribution.H0: the errors follow a normal distribution HA: the errors do not follow a normal distribution.

While hypothesis tests are usually constructed to reject the null hypothesis, this is a case where we actually hope we fail to reject the null hypothesis as this would mean that the errors follow a normal distribution.

Anderson-Darling Test

The Anderson-Darling Test measures the area between a fitted line (based on the chosen distribution) and a nonparametric step function (based on the plot points). The statistic is a squared distance that is weighted more heavily in the tails of the distribution. Smaller Anderson-Darling values indicate that the distribution fits the data better. The test statistic is given by:

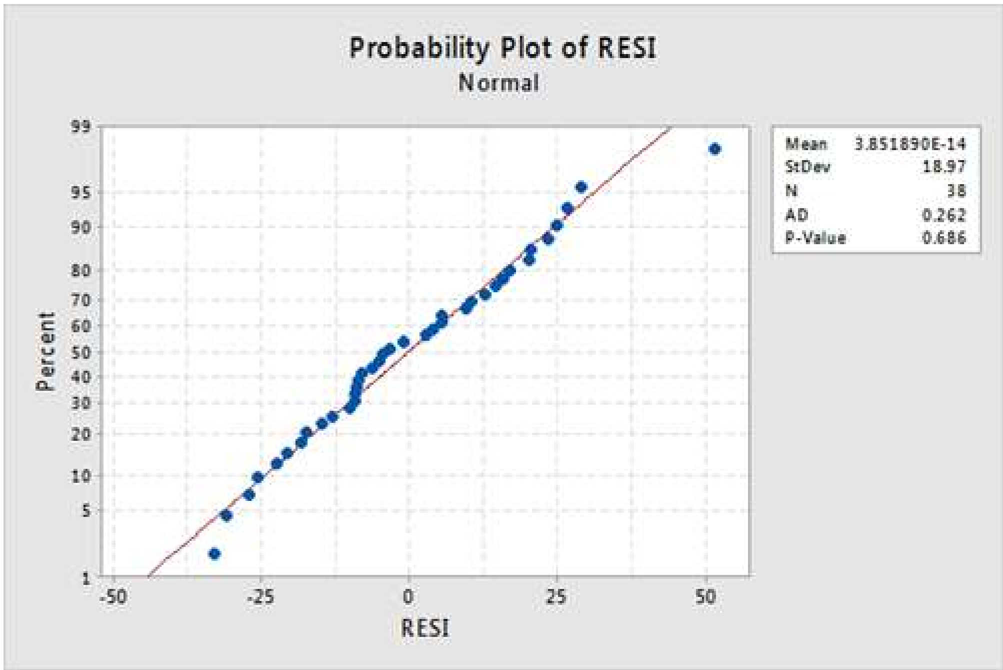
$$A2 = -n - n \sum i = 12i - 1n[\log F(ei) + \log(1 - F(en + 1 - i))],$$

$$A2 = -n - \sum i = 1n2i - 1n[\log F\left(ei\right) + \log\left(1 - F\left(en + 1 - i\right)\right)]$$

where $F(\cdot)F(\cdot)$ is the cumulative distribution of the normal distribution. The test statistic is compared against the critical values from a normal distribution in order to determine the p-value.

The Anderson-Darling test is available in some statistical software. To illustrate, here's statistical software output for the example on *IQ* and physical characteristics from Lesson 5 (iqsize.txt), where we've fit a model with *PIQ* as the response and Brain and Height as the predictors:

Notes



Since the Anderson-Darling test statistic is 0.262 with an associated p-value of 0.686, we fail to reject the null hypothesis and conclude that it is reasonable to assume that the errors have a normal distribution

Shapiro-Wilk Test

The Shapiro-Wilk Test uses the test statistic

$$W = (\sum ni = 1aie(i))2 \sum ni = 1(ei - {}^{-}e)2, W = \left(\sum i = 1naie\left(i\right)\right)2 \sum i = 1n\left(ei - e^{-}\right)2,$$

where the aiai values are calculated using the means, variances and covariances of the e(i)e(i). W is compared against tabulated values of this statistic's distribution. Small values of W will lead to rejection of the null hypothesis.

The Shapiro-Wilk test is available in some statistical software. For the IQ and physical characteristics model with PIQ as the response and Brain and Height as the predictors, the value of the test statistic is 0.976 with an associated p-value of 0.576, which leads to the same conclusion as for the Anderson-Darling test.

Ryan-Joiner Test

The Ryan-Joiner Test is a simpler alternative to the Shapiro-Wilk test. The test statistic is actually a correlation coefficient calculated by

$$Rp = \sum ni = 1e(i)z(i)\sqrt{}s2(n-1)\sum ni = 1z2(i), Rp = \sum i = 1ne\left(i\right)z\left(i\right)s2\left(n-1\right)\sum i = 1nz\left(i\right)2,$$

where the z(i)z(i) values are the z-score values (i.e., normal values) of the corresponding e(i)e(i) value and s2s2 is the sample variance. Values of RpRpcloser to 1 indicate that the errors are normally distributed.

The Ryan-Joiner test is available in some statistical software. For the IQ and physical characteristics model with PIQ as the response and Brain and Height as the predictors, the value of the test statistic is 0.988 with an associated p-value > 0.1, which leads to the same conclusion as for the Anderson-Darling test.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov Test (also known as the Lilliefors Test) compares the empirical cumulative distribution function of sample data with the distribution expected if the data were normal. If this observed difference is sufficiently large, the test will reject the null hypothesis of population normality. The test statistic is given by:

D = max(D+,D-), D = max(D+,D-),

where

D+ = maxi(i/n-F(e(i)))D- = maxi(F(e(i))-(i-1)/n),D+ = maxi(i/n-F(e(i)))D- = maxi(F(e(i))-(i-1)/n),

where e(i)e(i) pertains to the ithith largest value of the error terms. The test statistic is compared against the critical values from a normal distribution in order to determine the p-value.

The Kolmogorov-Smirnov test is available in some statistical software. For the IQ and physical characteristics model with PIQ as the response and Brain and Height as the predictors, the value of the test statistic is 0.097 with an associated p-value of 0.490, which leads to the same conclusion as for the Anderson-Darling test.

Central Limit Theorem

The central limit theorem states that when samples from a data set with a known variance are aggregated their mean roughly equals the population mean. Said another way, CLT is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population. Furthermore, all the samples will follow an approximate normal distribution pattern, with all variances being approximately equal to the variance of the population divided by each sample's size.

According to the central limit theorem, the mean of a sample of data will be closer to the mean of the overall population in question as the sample size increases, notwithstanding the actual distribution of the data, and whether it is normal or non-normal. As a general rule, sample sizes equal to or greater than 30 are considered sufficient for the CLT to hold, meaning the distribution of the sample means is fairly normally distributed.

The Central Limit Theorem in Finance

The CLT is useful when examining returns for a stock or index because it simplifies many analysis procedures. An appropriate sample size depends on the data available, but generally, having a sample size of at least 50 observations is sufficient. Due to the relative ease of generating financial data, it is often easy to produce much larger sample sizes. The CLT is the basis for sampling in statistics, so it holds the foundation for sampling and statistical analysis in finance too. Investors of all types rely on the CLT to analyze stock returns, construct portfolios and manage risk.

Example of Central Limit Theorem

If an investor is looking to analyze the overall return for a stock index made up of 1,000 stocks, he or she can take random samples of stocks from the index to get an estimate for the return of the total index. The samples must be random, and he or she must evaluate at least 30 stocks in each sample for the central limit theorem to hold. Random samples ensure a broad range of stock across industries and sectors is represented in the sample. Stocks previously selected must also be replaced for selection in other samples to avoid bias. The average returns from these samples approximates the return for the whole index and are approximately normally distributed. The approximation holds even if the actual returns for the whole index are not normally distributed.

A very general result concerning the weak consistency and uniform asymp- totic normality of the maximum likelihood estimator is presented. The result proves to be of particular value in establishing uniform asymptotic normality of randomly normalized maximum likelihood estimators of parameters in stochastic processes.

Maximum likelihood estimation

Maximum likelihood estimation (MLE) is a technique used for estimating the parameters of a given distribution, using some observed data. For example, if a population is known to follow a normal distribution but the mean and variance are unknown, MLE can be used to estimate them using a limited sample of the population, by finding particular values of the mean and variance so that the observation is the most likely result to have occurred.

MLE is useful in a variety of contexts, ranging from econometrics to MRIs to satellite imaging. It is also related to Bayesian statistics.

2.11 SIGNIFICANCE TEST

Once sample data has been gathered through an observational study or experiment, statistical inference allows analysts to assess evidence in favor or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as tests of significance.

Every test of significance begins with a null hypothesis H0. H0 represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug. We would write H0: there is no difference between the two drugs on average.

The alternative hypothesis, Ha, is a statement of what a statistical hypothesis test is set up to establish. For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect, on average, compared to that of the current drug. We would write Ha: the two drugs have different effects, on average. The alternative hypothesis might also be that the new drug is better, on average, than the current drug. In this case we would write Ha: the new drug is better than the current drug, on average.

The final conclusion once the test has been carried out is always given in terms of the null hypothesis. We either "reject H0 in favor of Ha" or "do not reject H0"; we never conclude "reject Ha", or even "accept Ha".

If we conclude "do not reject H0", this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against H0 in favor of Ha; rejecting the null hypothesis then, suggests that the alternative hypothesis may be true.

Hypotheses are always stated in terms of population parameter, such as the mean μ . An alternative hypothesis may be one-sided or two-sided. A one-sided hypothesis claims that a parameter is either larger or smaller than the value given by the null hypothesis. A two-sided hypothesis claims that a parameter is simply not equal to the value given by the null hypothesis - the direction does not matter.

Hypotheses for a one-sided test for a population mean take the following form:

H0: $\mu = k$

Ha: $\mu > k$

or

H0: $\mu = k$

Ha: $\mu < k$.

Hypotheses for a two-sided test for a population mean take the following form:

H0: $\mu = k$

Ha: $\mu \neq k$.

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data.

2.12 CONFIDENCE INTERVALS OF ESTIMATES-Z-TEST, T-TEST AND F-RATIO TEST

There are two types of estimates for each population parameter: the point estimate and confidence interval (CI) estimate. For both continuous variables (e.g., population mean) and dichotomous variables (e.g., population proportion) one first computes the point estimate from a sample. Recall that sample means and sample proportions are unbiased estimates of the corresponding population parameters.

For both continuous and dichotomous variables, the confidence interval estimate (CI) is a range of likely values for the population parameter based on:

- The point estimate, e.g., the sample mean
- The investigator's desired level of confidence (most commonly 95%, but any level between 0-100% can be selected) and
- The sampling variability or the standard error of the point estimate.

Strictly speaking a 95% confidence interval means that if we were to take 100 different samples and compute a 95% confidence interval for each sample, then approximately 95 of the 100 confidence intervals will contain the true mean value (μ). In practice, however, we select one random sample and generate one confidence interval, which may or may not contain the true mean. The observed interval may over- or underestimate μ. Consequently, the 95% CI is the likely range of the true, unknown parameter. The confidence interval does not reflect the variability in the unknown parameter. Rather, it reflects the amount of random error in the sample and provides a range of values that are likely to include the unknown parameter. Another way of thinking about a confidence interval is that it is the range of likely values of the parameter (defined as the point estimate + margin of error) with a specified level of confidence (which is similar to a probability).

Suppose we want to generate a 95% confidence interval estimate for an unknown population mean. This means that there is a 95% probability that the confidence interval will contain the true population mean. Thus, P([sample mean] - margin of error < μ < [sample mean] + margin of error) = 0.95.

Notes

The Central Limit Theorem introduced in the module on Probability stated that, for large samples, the distribution of the sample means is approximately normally distributed with a mean:

$$\mu_{\bar{x}} = \mu$$

and a standard deviation (also called the standard error):

$$\sigma_{\bar{x}} = \sigma/\sqrt{n}$$

For the standard normal distribution, P(-1.96 < Z < 1.96) = 0.95, i.e., there is a 95% probability that a standard normal variable, Z, will fall between -1.96 and 1.96. The Central Limit Theorem states that for large samples:

$$z = \frac{\overline{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

By substituting the expression on the right side of the equation:

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

Using algebra, we can rework this inequality such that the mean (μ) is the middle term, as shown below:

$$P\left(-1.96\ \sigma/\sqrt{n} < \mu < \overline{X} + 1.96\sigma/\sqrt{n}\right) = 0.95$$

then

$$P\left(-1.96\ \sigma/\sqrt{n} < \mu < \overline{X} + 1.96\sigma/\sqrt{n}\right) = 0.95$$

and finally

$$P\left(-1.96\ \sigma/\sqrt{n} < \mu < \overline{X} + 1.96\sigma/\sqrt{n}\right) = 0.95$$

This last expression, then, provides the 95% confidence interval for the population mean, and this can also be expressed as:

Thus, the margin of error is 1.96 times the standard error (the standard deviation of the point estimate from the sample), and 1.96 reflects the fact that a 95% confidence level was selected. So, the general form of a confidence interval is:

$$\overline{X} \pm 1.96\sigma/\sqrt{n}$$

Point estimate $\underline{+}$ Z SE (point estimate)

where Z is the value from the standard normal distribution for the selected confidence level (e.g., for a 95% confidence level, Z = 1.96).

In practice, we often do not know the value of the population standard deviation (σ). However, if the sample size is large (n > 30), then the sample standard deviations can be used to estimate the population standard deviation.

Confidence intervals of estimates t-test

The computation of t-value involves the following steps:

(i) Null Hypothesis

First of all, it is presumed that there is no difference between the mean of small sample and the population means (μ) or hypothetical mean. Thus,

Null hypothesis (Ho): sample mean ($\overline{X}$) = population mean (μ) or Ho = $\overline{X}$ = μ

(ii) Test statistics

T-value is calculated by the following formula:

$$t = \frac{(\overline{X} - \mu)}{S/\sqrt{n}}$$

where, $\overline{X}$ = Sample mean, μ = population mean

n = number of units in the sample

s = standard deviation of sample

For small samples

$$s = \sqrt{\frac{1}{n-1}\sum(X - \overline{X})^2} \quad \text{or} \quad \sqrt{\frac{1}{n-1}\sum d^2}$$

(iii) Degree of freedom

It is one less than the number of units in the sample.

Degree of freedom (d.f. or v) = No. of units in the sample – 1

= n – 1

(iv) Level of significance

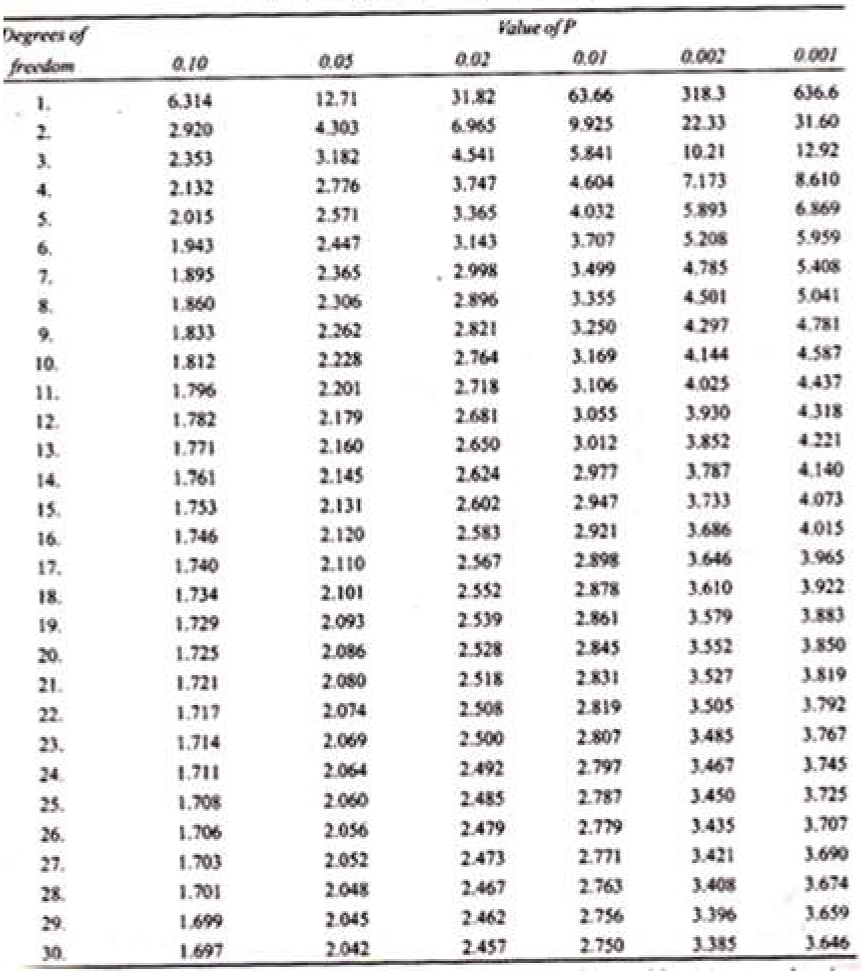
Any level of significance can be considered to test the hypothesis but generally 1 % (=0.01) or 5% (= 0.05) levels of probability is considered for testing the hypothesis.

(v) Table value of t

After calculating the value of (= t – cal) with the help of above formula, the value of t is noted from Fishers and Yates Mable at given degree of freedom and 5% level of significance. Then after the calculated value of t is compared with the table value of t.

Notes

Table 1. STUDENT'S *t*-DISTRIBUTION



Degrees of			*Value of P*			
freedom	*0.10*	*0.05*	*0.02*	*0.01*	*0.002*	*0.001*
1.	6.314	12.71	31.82	63.66	318.3	636.6
2.	2.920	4.303	6.965	9.925	22.33	31.60
3.	2.353	3.182	4.541	5.841	10.21	12.92
4.	2.132	2.776	3.747	4.604	7.173	8.610
5.	2.015	2.571	3.365	4.032	5.893	6.869
6.	1.943	2.447	3.143	3.707	5.208	5.959
7.	1.895	2.365	2.998	3.499	4.785	5.408
8.	1.860	2.306	2.896	3.355	4.501	5.041
9.	1.833	2.262	2.821	3.250	4.297	4.781
10.	1.812	2.228	2.764	3.169	4.144	4.587
11.	1.796	2.201	2.718	3.106	4.025	4.437
12.	1.782	2.179	2.681	3.055	3.930	4.318
13.	1.771	2.160	2.650	3.012	3.852	4.221
14.	1.761	2.145	2.624	2.977	3.787	4.140
15.	1.753	2.131	2.602	2.947	3.733	4.073
16.	1.746	2.120	2.583	2.921	3.686	4.015
17.	1.740	2.110	2.567	2.898	3.646	3.965
18.	1.734	2.101	2.552	2.878	3.610	3.922
19.	1.729	2.093	2.539	2.861	3.579	3.883
20.	1.725	2.086	2.528	2.845	3.552	3.850
21.	1.721	2.080	2.518	2.831	3.527	3.819
22.	1.717	2.074	2.508	2.819	3.505	3.792
23.	1.714	2.069	2.500	2.807	3.485	3.767
24.	1.711	2.064	2.492	2.797	3.467	3.745
25.	1.708	2.060	2.485	2.787	3.450	3.725
26.	1.706	2.056	2.479	2.779	3.435	3.707
27.	1.703	2.052	2.473	2.771	3.421	3.690
28.	1.701	2.048	2.467	2.763	3.408	3.674
29.	1.699	2.045	2.462	2.756	3.396	3.659
30.	1.697	2.042	2.457	2.750	3.385	3.646

(vi) Test criterion

If the calculated value of t (t-cal) irrespective of the positive (+) or negative sign (-) is less than the tabulated value off at respective degree of freedom and at 5% or 0.05 level of probability, then the Null hypothesis is correct, i.e., difference between the sample mean and population mean is insignificant and so the hypothesis is acceptable.

But, if the calculated value of t (- cal) is greater than the tabulated value of t (t-tab) at given degree of freedom and at 5% level of probability, then the observed difference is considered statistically non-significant and so the null hypothesis is incorrect and rejected.

In such a ease the observed data are not according to the expected data or in other words, the sample under test does not represent the population with μ as its mean.

(II) T-test for assessing the significance of the difference between the means of two samples drawn from the same population:

t- test is also applied to test the significance of the difference between the arithmetic means off two samples drawn from the same population.

The procedure of the test is as follows:

(i) Null hypothesis

In this, first of all it is presumed that there is no difference in the standard deviations of the two samples under test, i.e.

HO = µ1 = µ2

Notes

where µ1 and µ2are the standard deviations of the sample I and sample II respectively.

(ii) Test statistics

Next, the value of t is calculated by the following formula:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where, $\overline{X}_1$ = arithmetic mean of sample I

$\overline{X}_2$ = arithmetic mean of sample II

n_1 = No. of observations in sample I

n_2 = No. of observations in sample II

S_1 = Standard deviation of sample I

S_2 = Standard deviation of sample II

Since according to hypothesis $\mu_1 = \mu_2$

therefore, $$t = \frac{X_1 - X_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\text{difference between the sample means}}{\text{standard error of difference between the means}}$$

$$S = \frac{\sqrt{n_1 S_1^2 - n_2 S_2^2}}{n_1 + n_2 - 2}$$

or $$\frac{\sqrt{\sum(X_1 - \overline{X}_1) + \sum(X_2 - \overline{X}_2)}}{n_1 + n_2 - 2}$$

(iii) Degree of freedom

(d.f.) = n1+ n2 – 2)

(iv) Level of significance:

The level of significance is generally considered at 5% (0.05) level «f probability.

(v) Tabulated value of t

Value of t is recorded from the Fisher and Yates table at the given degree of freedom and at 5% level of significance.

(vi) Test criterion

At last, the calculated value of t (|t|- cal) is compared with the table value of t at the given degree of freedom and 5% level of significance.

(a) If the calculated value of r-exceeds the table value, the observed difference is considered statistically significant and hence null hypothesis is rejected.

(b) If the calculated value of (|t|- cal) is less than the table value of t, the differences are not significant. Therefore, Ho hypothesis is accepted and the samples represent the population well.

Notes

Confidence intervals of estimates f-ratio test

F-test was first worked out by G.W. Snedecore. This is based on F distribution and is used to test the significance of difference between the standard deviations of two samples. Snedecore calculated the variance ratio of the two samples and named this ratio after R. F. Fisher. Therefore, the test is referred to as f-test or variance ratio test.

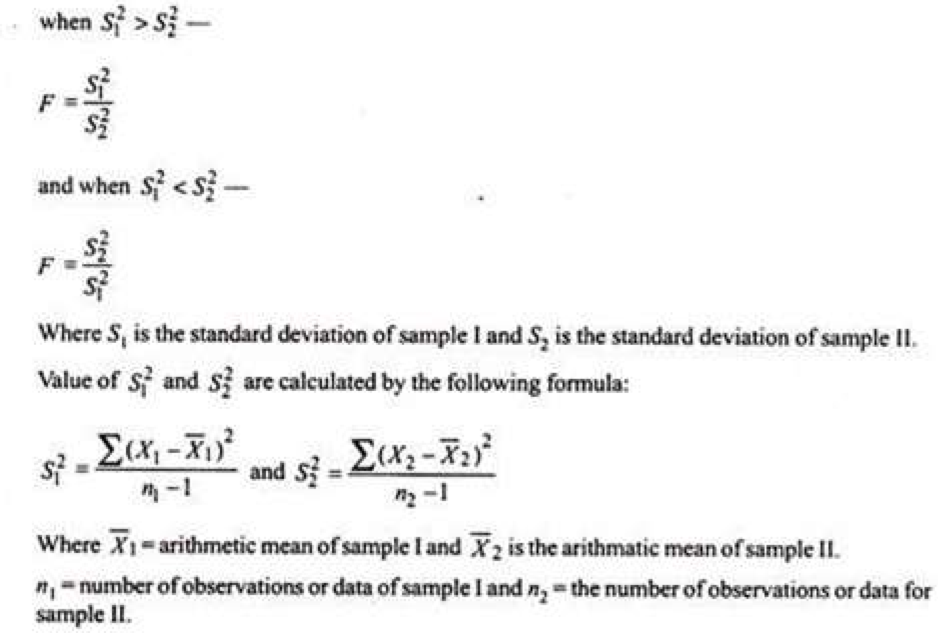
The procedure F-test is as follows:

(i) Null hypothesis

In this, it is presumed that the ratio of variance of two samples is not significant.

(ii) Test statistics

Value of F is determined by the following formula:



when $S_1^2 > S_2^2$ —

$$F = \frac{S_1^2}{S_2^2}$$

and when $S_1^2 < S_2^2$ —

$$F = \frac{S_2^2}{S_1^2}$$

Where S_1 is the standard deviation of sample I and S_2 is the standard deviation of sample II.

Value of S_1^2 and S_2^2 are calculated by the following formula:

$$S_1^2 = \frac{\sum(X_1 - \overline{X}_1)^2}{n_1 - 1} \text{ and } S_2^2 = \frac{\sum(X_2 - \overline{X}_2)^2}{n_2 - 1}$$

Where $\overline{X}_1$ = arithmetic mean of sample I and $\overline{X}_2$ is the arithmatic mean of sample II.

n_1 = number of observations or data of sample I and n_2 = the number of observations or data for sample II.

Degree of freedom:

(d.f.) for sample I (v1) = n1-1 and d.f. for sample II (v2) = n2 – 1

The level of significance:

Generally the level of significance is considered either at 1 % (0.01) or 5% (0.05).

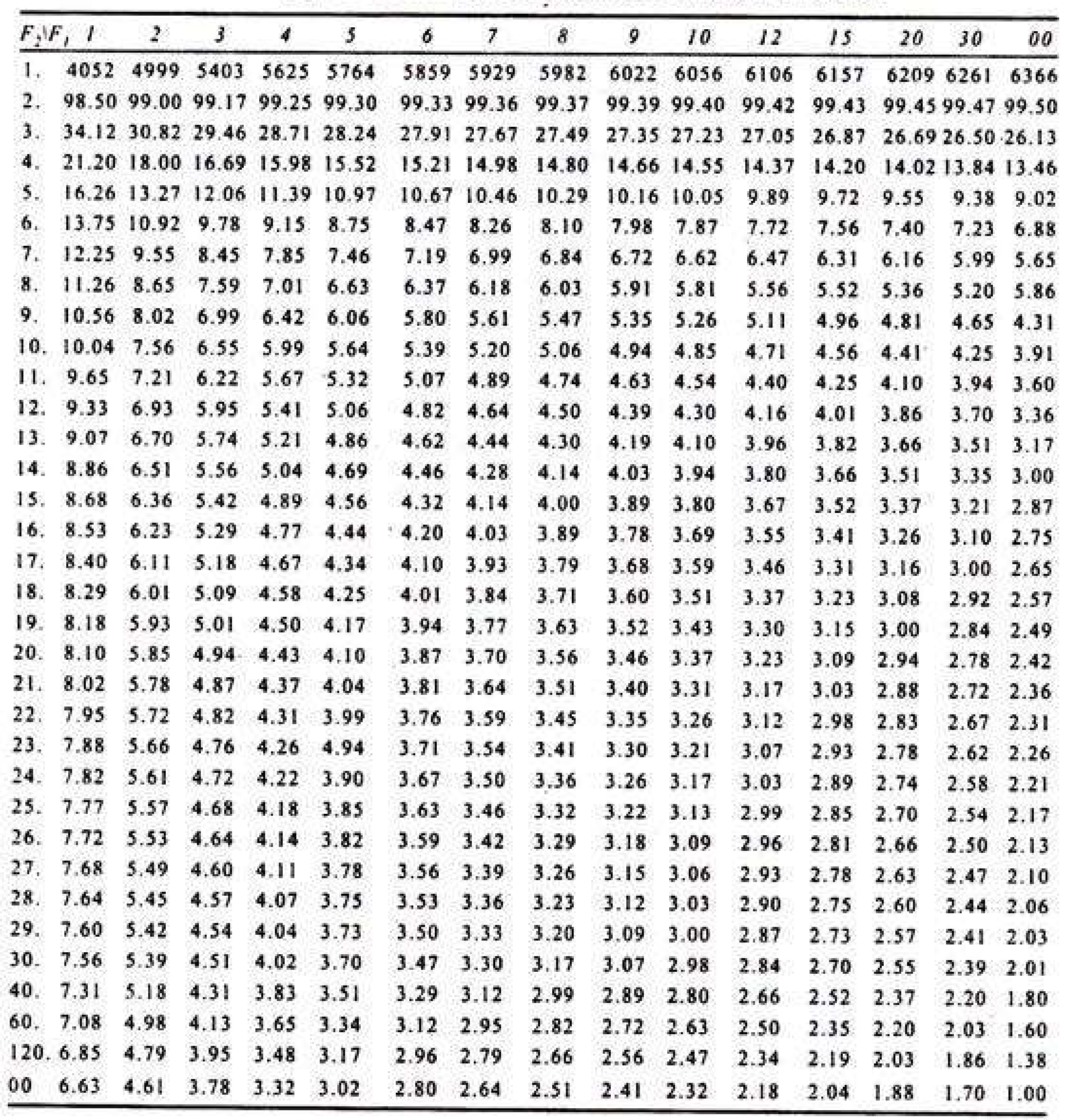
The tabulated value of F:

The value of F for a sample with greater spread at given d.f. is located in the Table from left to right and for the other sample with low spread at respective degree of freedom is located in the same table from top downward.

Where the two meet each other that value of given level of significance is F-value.

Now if the calculated value of F is less than the tabulated value of F then null hypothesis is true and accepted, i.e., the difference between the standard deviations of two samples is not significant and so the two samples under test might have been drawn from the same population.

Table 2(a). 1 Per cent Points of Variance-Ratio (F) Distribution



$F_2 \backslash F_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	30	00
1.	4052	4999	5403	5625	5764	5859	5929	5982	6022	6056	6106	6157	6209	6261	6366
2.	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.47	99.50
3.	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.50	26.13
4.	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.84	13.46
5.	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.38	9.02
6.	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	6.88
7.	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99	5.65
8.	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.56	5.52	5.36	5.20	5.86
9.	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65	4.31
10.	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	3.91
11.	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	3.94	3.60
12.	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.70	3.36
13.	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.51	3.17
14.	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.35	3.00
15.	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21	2.87
16.	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.10	2.75
17.	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.00	2.65
18.	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.92	2.57
19.	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.84	2.49
20.	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.42
21.	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.72	2.36
22.	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.67	2.31
23.	7.88	5.66	4.76	4.26	4.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.62	2.26
24.	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.58	2.21
25.	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.54	2.17
26.	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.50	2.13
27.	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.47	2.10
28.	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.44	2.06
29.	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.41	2.03
30.	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.01
40.	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.20	1.80
60.	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.60
120.	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86	1.38
00	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.70	1.00

If the calculated value of F is, however, greater than the table value of F (F – cal > F-tab.) the null hypothesis is rejected and the difference between the standard deviations is significant which means that the two samples under test cannot be supposed to be parts of the same population.

2.13 PREDICTION POINT AND INTERVAL

In statistical inference, specifically predictive inference, a prediction interval is an estimate of an interval in which a future observation will fall, with a certain probability, given what has already been observed. Prediction intervals are often used in regression analysis.

Prediction intervals are used in both frequentist statistics and Bayesian statistics: a prediction interval bears the same relationship to a future observation that a frequentist confidence interval or Bayesian credible interval bears to an unobservable population parameter: prediction intervals predict the distribution of individual future points, whereas confidence intervals and credible intervals of parameters predict the distribution of estimates of the true population mean or other quantity of interest that cannot be observed.

Point estimation gives us a particular value as an estimate of the population parameter. Interval estimation gives us a range of values which is likely to contain the population parameter.

Point Prediction uses the models fit during analysis and the factor settings specified on the factors tool to compute the point predictions and interval estimates. The predicted values are updated as the levels are changed. Prediction intervals (PI) are found under the Confirmation node.

Math Details

(1-α)*100% Confidence Interval

y^±t(1-a2,n-p)·SEpredictiony^±t(1-a2,n-p)·SEprediction

where,SEprediction=S·x0(XTX)-1xT0------------vwhere,SEprediction=S·x0(XTX)-1x0T

(1-α)*100% Prediction Interval

y^±t(1-a2,n-p)·SEpredictiony^±t(1-a2,n-p)·SEprediction

where, SEprediction = S·1Nx0(XTX)-1xT0-------------$\surd$ where, SEprediction = S·1Nx0(XTX)-1x0T

(1-α)*100% Tolerance Interval for P * 100% of the population

y^±s·TIMultipliery^±s·TIMultiplier

where, TIMultiplier = t(1-α, n-p)·xo(XTX)-1xT0------------$\surd$+ Φ-1(0.5+P2)·n-px2(a,n-p)----- $\surd$where,TIMultiplier = t(1-α, n-p)·xo(XTX)-1x0T + Φ-1(0.5+P2)·n-px(a,n-p)2

The TI uses only alpha rather than alpha/2 to compute the two-tailed interval.

Where:

y^y^ = predicted value at x0

s = estimated standard deviation

t = student's t critical value

α = acceptable type I error rate (1 - confidence level)

n = number of runs in the design

N = number of observation in the future sample

p = The number of terms in the model including the intercept

P = proportion of the population contained in the tolerance interval

X = expanded model matrix [*]

x_0 = expanded point vector [†]

ϕ = inverse normal function to convert the proportion to a normal score

X^2 = Chi-Square critical value

n-p is also the residual degrees of freedom (df) from the ANOVA. [‡]

The superscript T indicates the previous matrix is transposed.

The superscript -1 indicates the previous matrix is inverted.

2.14 EXTENSION OF THE TWO VARIABLE LINEAR MODELS

Suppose we have two random variables X and Y. We assume that Y depends on X. i.e., when variable X takes a specific value, we expect a response in the random variable Y. In other words, the value taken by X influences the value of Y.

So X is the independent variable and Y is the dependent variable. If we plot the data, we find that Y depends on X. i.e., higher expenditures is seen to be associated with higher incomes. However, since we do not know the precise form of the dependence or relation, what we can do is, to assume the simplest possible relation - a linear relation. In other words, we fit a straight line which most closely represent the plot. The points on the straight line gives us the expected values of Y for every possible value of X.

Since the fitted line is a straight line, we can represent the line by the following equation.

E(Y|X = xi) = a + bxi

where a stands for the y-intercept and b is the slope.

a = the value of Y when X = 0

b = the change in Y brought on by a unit change in X

2.15 THREE – VARIABLE LINEAR MODEL

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. The population regression line for p explanatory variables x1, x2, ... , xp is defined to be $\mu_{\mathrm{y}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{\mathrm{p}} x_{\mathrm{p}}$. This line describes how the mean response μ_{y} changes with the explanatory variables. The observed values for y vary about their means μ_{y} and are assumed to have the same standard deviation σ. The fitted values b0, b1, ..., bp estimate the parameters 0, 1, ..., p of the population regression line.

Since the observed values for y vary about their means μ_{y}, the multiple regression model includes a term for this variation. In words, the model is expressed as DATA = FIT + RESIDUAL, where the "FIT" term represents the expression $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{\mathrm{p}} x_{\mathrm{p}}$. The "RESIDUAL" term represents the deviations of the observed values y from their means μ_{y}, which are normally distributed with mean 0 and variance σ. The notation for the model deviations is ε.

Formally, the model for multiple linear regression, given n observations, is

$$y_i = \beta_0 + \beta_1 x_{\mathrm{i1}} + \beta_2 x_{\mathrm{i2}} + \beta_{\mathrm{p}} x_{\mathrm{ip}} + \varepsilon_i \; for \; i = 1,2,......n$$

In the least-squares model, the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values. The least-squares estimates b0, b1, ... bp are usually computed by statistical software.

The values fit by the equation b0 + b1xi1 + ... + bpxip are denoted i, and the residuals ei are equal to yi - i, the difference between the observed and fitted values. The sum of the residuals is equal to zero.

The variance ^{2} may be estimated by $s^2 =$, also known as the mean-squared error (or MSE).

The estimate of the standard error s is the square root of the MSE.

2.16 THE CO-EFFICIENT OF MULTIPLE CORRELATIONS

The co-efficient of multiple correlations is a measure of how well a given variable can be predicted using a linear function of a set of other variables. It is the correlation between the variable's values and the best predictions that can be computed linearly from the predictive variables.

The co-efficient of multiple correlation takes values between 0 and 1; a higher value indicates a better predictability of the dependent variable from the independent variables, with a value of 1 indicating that the predictions are exactly correct and a value of 0 indicating that no linear combination of the independent variables is a better predictor than is the fixed mean of the dependent variable.

The co-efficient of multiple correlation is computed as the square root of the co-efficient of determination, but under the particular assumptions that an intercept is included and that the best possible linear predictors are used, whereas the coefficient of determination is defined for more general cases, including those of nonlinear prediction and those in which the predicted values have not been derived from a model-fitting procedure.

Multiple Correlation

We can also calculate the correlation between more than two variables.

Definition 1: Given variables x, y and z, we define the multiple correlation co-efficient

$$R_{z,xy} = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2r_{xz}r_{yz}r_{xy}}{1 - r_{xy}^2}}$$

where r_{xz}, r_{yz}, r_{xy} are as defined in Definition 2 of Basic Concepts of Correlation. Here xand y are viewed as the independent variables and z is the dependent variable.

We also define the multiple coefficient of determination to be the square of the multiple correlation coefficient.

Often the subscripts are dropped and the multiple correlation coefficient and multiple coefficient of determination are written simply as R and R2 respectively. These definitions may also be expanded to more than two independent variables. With just one independent variable the multiple correlation coefficient is simply r.

Unfortunately R is not an unbiased estimate of the population multiple correlation coefficient, which is evident for small samples. A relatively unbiased version of R is given by R adjusted.

Definition 2: If R is Rz,xy as defined above (or similarly for more variables) then the adjusted multiple coefficient of determination is

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

where k = the number of independent variables and n = the number of data elements in the sample for z (which should be the same as the samples for x and y).

Excel Data Analysis Tools: In addition to the various correlation functions described elsewhere, Excel provides the Covariance and Correlation data analysis tools. The Covariance tool calculates the pairwise population covariances for all the variables in the data set. Similarly the Correlation tool calculates the various correlation co-efficients as described in the following example.

2.17 PARTIAL CORRELATION CO-EFFICIENT

Partial correlation is the measure of association between two variables, while controlling or adjusting the effect of one or more additional variables. Partial correlations can be used in many cases that assess for relationship, like whether or not the sale value of a particular commodity is related to the expenditure on advertising when the effect of price is controlled.

Assumptions

Useful in only small models like the models which involve three or four variables.

Used in only those models which assume a linear relationship.

The data is supposed to be interval in nature.

The residual variables or unmeasured variables are not correlated with any of the variables in the model, except for the one for which these residuals have occurred.

Partial correlation is a method used to describe the relationship between two variables whilst taking away the effects of another variable, or several other variables, on this relationship.

Partial correlation is best thought of in terms of multiple regressions; Stats Direct shows the partial correlation coefficient r with its main results from multiple linear regression.

A different way to calculate partial correlation coefficients, which does not require a full multiple regression, is show below for the sake of further explanation of the principles:

Consider a correlation matrix for variables A, B and C (note that the multiple line regression function in StatsDirect will output correlation matrices for you as one of its options):

	A	**B**	**C**
A:	*		
B:	r(AB)	*	
C:	r(AC)	r(BC)	*

The partial correlation of A and B adjusted for C is:

The same can be done using Spearman's rank correlation co-efficient.

The hypothesis test for the partial correlation co-efficient is performed in the same way as for the usual correlation co-efficient but it is based upon n-3 degrees of freedom.

Please note that this sort of relationship between three or more variables is more usefully investigated using the multiple regression itself (Altman, 1991).

Notes

The general form of partial correlation from a multiple regression is as follows:

- where t_k is the Student t statistic for the kth term in the linear model.

2.18 GENERAL LINEAR MODEL WITH K- EXPLANATORY VARIABLE

The term general linear model (GLM) usually refers to conventional linear regression models for a continuous response variable given continuous and/or categorical predictors. It includes multiple linear regression, as well as ANOVA and ANCOVA (with fixed effects only). The form is yi~N(xTiß,s2),yi~N(xiTß,s2), where xixicontains known covariates and ßß contains the coefficients to be estimated. These models are fit by least squares and weighted least squares using, for example: SAS Proc GLM or R functions lsfit() (older, uses matrices) and lm() (newer, uses data frames).

The term generalized linear model (GLIM or GLM) refers to a larger class of models popularized by McCullagh and Nelder (1982, 2nd edition 1989). In these models, the response variable yiyi is assumed to follow an exponential family distribution with mean µiµi, which is assumed to be some (often nonlinear) function of xTißxiTß. Some would call these "nonlinear" because µiµi is often a nonlinear function of the covariates, but McCullagh and Nelder consider them to be linear, because the covariates affect the distribution of yiyionly through the linear combination xTißxiTß. The first widely used software package for fitting these models was called GLIM. Because of this program, "GLIM" became a well-accepted abbreviation for generalized linear models, as opposed to "GLM" which often is used for general linear models. Today, GLIM's are fit by many packages, including SAS Proc Genmod and R function glm(). Notice, however, that Agresti uses GLM instead of GLIM short-hand and we will use GLM.

The generalized linear models (GLMs) are a broad class of models that include linear regression, ANOVA, Poisson regression, log-linear models etc. The table below provides a good summary of GLMs following Agresti (ch. 4, 2013):

Model	***Random***	***Link***	***Systematic***
Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed

There are three components to any GLM:

1. Random Component – refers to the probability distribution of the response variable (Y); e.g. normal distribution for Y in the linear regression, or binomial distribution for Y in the binary logistic regression. Also called a noise model or error model. How is random error added to the prediction that comes out of the link function?
2. Systematic Component - specifies the explanatory variables (X1, X2, ... Xk) in the model, more specifically their linear combination in creating the so called linear predictor; e.g., ß0 + ß1x1 + ß2x2 as we have seen in a linear regression or as we will see in a logistic regression in this lesson.

3. Link Function, η or g(μ) - specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables; e.g., η = g(E(Yi)) = E(Yi) for linear regression, or η = logit(p) for logistic regression.

Assumptions:

1. The data Y1, Y2, ..., Yn are independently distributed, i.e., cases are independent.
2. The dependent variable Yi does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal,...)
3. GLM does NOT assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the transformed response in terms of the link function and the explanatory variables; e.g., for binary logistic regression logit(p) = ß0 + ßX.
4. Independent (explanatory) variables can be even the power terms or some other nonlinear transformations of the original independent variables.
5. The homogeneity of variance does NOT need to be satisfied. In fact, it is not even possible in many cases given the model structure and overdispersion (when the observed variance is larger than what the model assumes) maybe present.
6. Errors need to be independent but NOT normally distributed.
7. It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.
8. Goodness-of-fit measures rely on sufficiently large samples, where a heuristic rule is that not more than 20% of the expected cells counts are less than 5.

2.19 LEAST-SQUARE ESTIMATES AND THEIR PROPERTIES

The method of least squares is a standard approach in regression analysis to approximate the solution of over determined systems, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the residuals made in the results of every single equation.

The most important application is in data fitting. The best fit in the least-squares sense minimizes the sum of squared residuals (a residual being: the difference between an observed value, and the fitted value provided by a model). When the problem has substantial uncertainties in the independent variable (the x variable), then simple regression and least-squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares.

Least-squares problems fall into two categories: linear or ordinary least squares and nonlinear least squares, depending on whether or not the residuals are linear in all unknowns. The linear least-squares problem occurs in statistical regression analysis; it has a closed-form solution. The nonlinear problem is usually solved by iterative refinement; at each iteration the system is approximated by a linear one, and thus the core calculation is similar in both cases.

Polynomial least squares describes the variance in a prediction of the dependent variable as a function of the independent variable and the deviations from the fitted curve.

When the observations come from an exponential family and mild conditions are satisfied, least-squares estimates and maximum-likelihood estimates are identical. The method of least squares can also be derived as a method of moments estimator.

The following discussion is mostly presented in terms of linear functions but the use of least squares is valid and practical for more general families of functions. Also, by iteratively applying local quadratic approximation to the likelihood (through the Fisher information), the least-squares method may be used to fit a generalized linear model.

Carl Friedrich Gauss

The first clear and concise exposition of the method of least squares was published by Legendre in 1805. The technique is described as an algebraic procedure for fitting linear equations to data and Legendre demonstrates the new method by analyzing the same data as Laplace for the shape of the earth. The value of Legendre's method of least squares was immediately recognized by leading astronomers and geodesists of the time.

In 1809 Carl Friedrich Gauss published his method of calculating the orbits of celestial bodies. In that work he claimed to have been in possession of the method of least squares since 1795. This naturally led to a priority dispute with Legendre. However, to Gauss's credit, he went beyond Legendre and succeeded in connecting the method of least squares with the principles of probability and to the normal distribution. He had managed to complete Laplace's program of specifying a mathematical form of the probability density for the observations, depending on a finite number of unknown parameters, and define a method of estimation that minimizes the error of estimation. Gauss showed that the arithmetic mean is indeed the best estimate of the location parameter by changing both the probability density and the method of estimation. He then turned the problem around by asking what form the density should have and what method of estimation should be used to get the arithmetic mean as estimate of the location parameter. In this attempt, he invented the normal distribution.

An early demonstration of the strength of Gauss' method came when it was used to predict the future location of the newly discovered asteroid Ceres. On 1 January 1801, the Italian astronomer Giuseppe Piazzi discovered Ceres and was able to track its path for 40 days before it was lost in the glare of the sun. Based on these data, astronomers desired to determine the location of Ceres after it emerged from behind the sun without solving Kepler's complicated nonlinear equations of planetary motion. The only predictions that successfully allowed Hungarian astronomer Franz Xaver von Zach to relocate Ceres were those performed by the 24-year-old Gauss using least-squares analysis.

In 1810, after reading Gauss's work, Laplace, after proving the central limit theorem, used it to give a large sample justification for the method of least squares and the normal distribution. In 1822, Gauss was able to state that the least-squares approach to regression analysis is optimal in the sense that in a linear model where the errors have a mean of zero, are uncorrelated, and have equal variances, the best linear unbiased estimator of the coefficients is the least-squares estimator. This result is known as the Gauss–Markov theorem.

The idea of least-squares analysis was also independently formulated by the American Robert Adrain in 1808. In the next two centuries workers in the theory of errors and in statistics found many different ways of implementing least squares.

2.20 VARIANCE- COVARIANCE MATRIX OF ESTIMATES

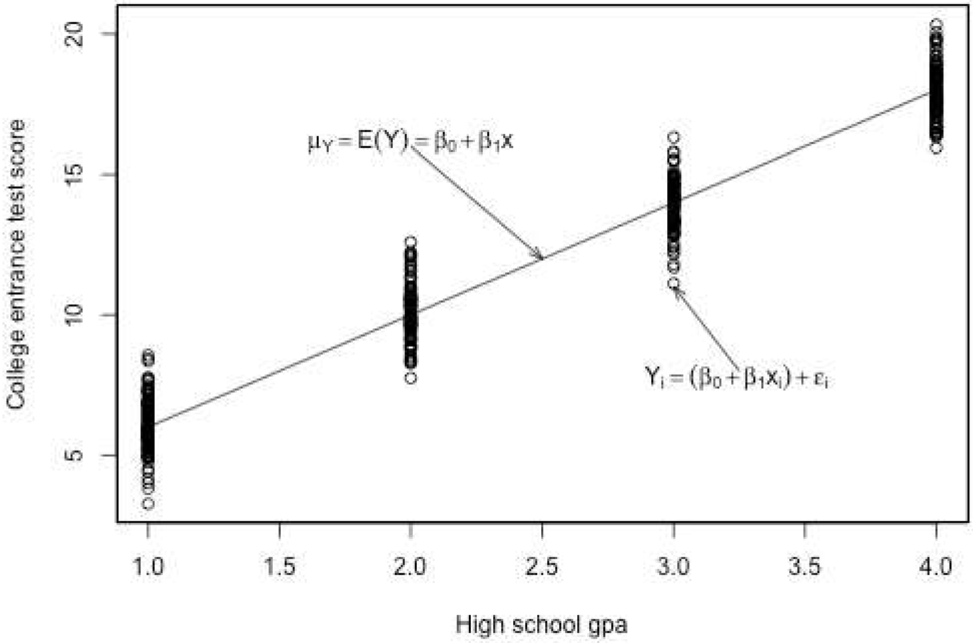
In statistics, sometimes the covariance matrix of a multivariate random variable is not known but has to be estimated. Estimation of covariance matrices then deals with the question of how to approximate the actual covariance matrix on the basis of a sample from the multivariate distribution. Simple cases, where observations are complete, can be dealt with by using the sample covariance matrix. The sample covariance matrix (SCM) is an unbiased and efficient estimator of the covariance matrix if the space of covariance matrices is viewed as an extrinsic convex cone in Rp×p; however, measured using the intrinsic geometry of positive-definite matrices, the SCM is a biased and inefficient estimator. In addition, if the random variable has normal distribution, the sample covariance matrix has Wishart distribution and a slightly differently scaled version of it is the maximum likelihood estimate. Cases involving missing data require deeper considerations. Another issue is the robustness to outliers, to which sample covariance matrices are highly sensitive. Statistical analyses of multivariate data often involve exploratory studies of the way in which the variables change in relation to one another and this may be followed up by explicit statistical models involving the covariance matrix of the variables. Thus the estimation of covariance matrices directly from observational data plays two roles:

- to provide initial estimates that can be used to study the inter-relationships;
- to provide sample estimates that can be used for model checking.

Estimates of covariance matrices are required at the initial stages of principal component analysis and factor analysis, and are also involved in versions of regression analysis that treat the dependent variables in a data-set, jointly with the independent variable as the outcome of a random sample.

2.21 ESTIMATES OF ERROR VARIANCE

The plot of our population of data suggests that the college entrance test scores for each subpopulation have equal variance. We denote the value of this common variance as σ^2.



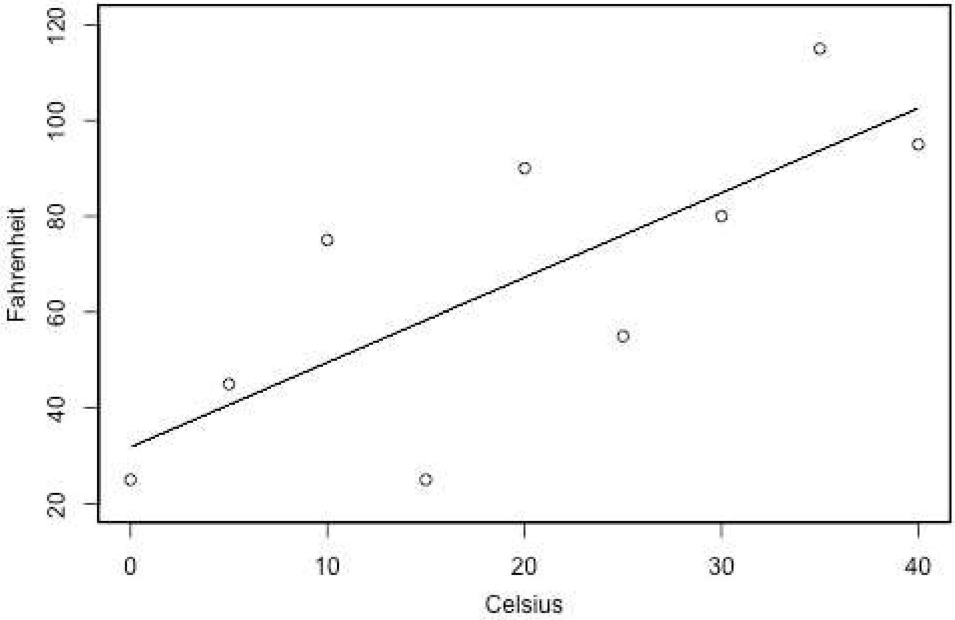
Notes

That is, σ^2 quantifies how much the responses (y) vary around the (unknown) mean population regression line μ_Y = E(Y) = ß0 + ß1xμY = E(Y) = ß0 + ß1x.

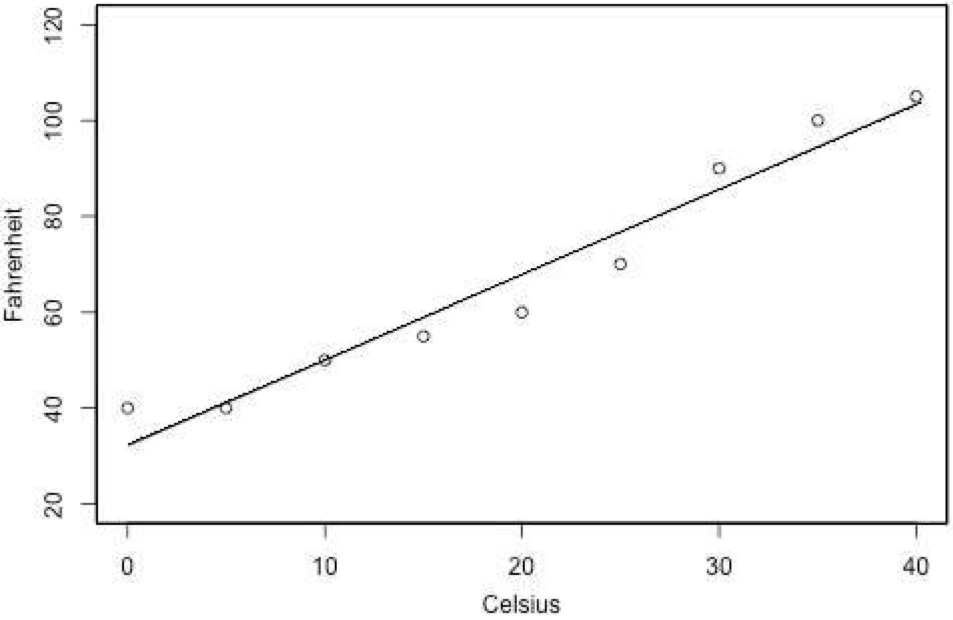
Why should we care about σ^2? The answer to this question pertains to the most common use of an estimated regression line, namely predicting some future response.

Suppose you have two brands (A and B) of thermometers, and each brand offers a Celsius thermometer and a Fahrenheit thermometer. You measure the temperature in Celsius and Fahrenheit using each brand of thermometer on ten different days. Based on the resulting data, you obtain two estimated regression lines - one for brand A and one for brand B. You plan to use the estimated regression lines to predict the temperature in Fahrenheit based on the temperature in Celsius.

Will this thermometer brand (A) yield more precise future predictions …?



… or this one (B)?

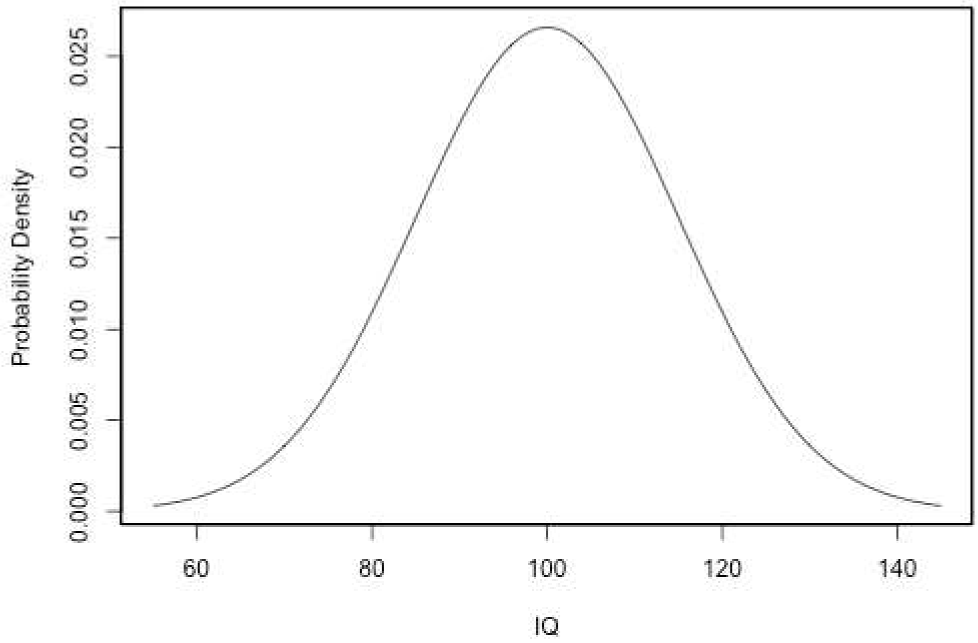


As the two plots illustrate, the Fahrenheit responses for the brand B thermometer don't deviate as far from the estimated regression equation as they do for the brand A thermometer. If we use the brand B estimated line to predict the Fahrenheit temperature, our prediction should never really be too far off from the actual observed Fahrenheit temperature. On the other hand, predictions of the Fahrenheit temperatures using the brand A thermometer can deviate quite a bit from the actual observed Fahrenheit temperature. Therefore, the brand B thermometer should yield more precise future predictions than the brand A thermometer.

To get an idea, therefore, of how precise future predictions would be, we need to know how much the responses (y) vary around the (unknown) mean population regression line μ_Y = E(Y) = ß0 + ß1xμY = E(Y) = ß0 + ß1x. As stated earlier, σ^2 quantifies this variance in the responses. Will we ever know this value σ^2? No! Because σ^2 is a population parameter, we will rarely know its true value. The best we can do is estimate it!

To understand the formula for the estimate of σ^2 in the simple linear regression setting, it is helpful to recall the formula for the estimate of the variance of the responses, σ^2, when there is only one population.

The following is a plot of the (one) population of IQ measurements. As the plot suggests, the average of the IQ measurements in the population is 100. But, how much do the IQ measurements vary from the mean? That is, how "spread out" are the IQs?

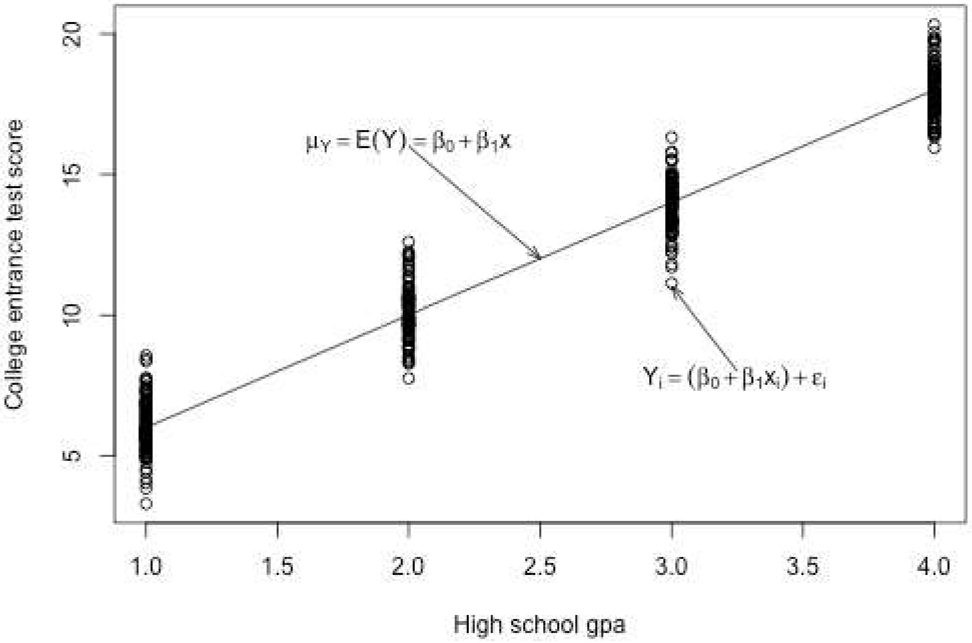


The sample variance:

$$s2=\sum ni=1(yi-{}^{-}y)2n-1s2=\sum i=1n\left(yi-y^{-}\right)2n-1$$

estimates σ^2, the variance of the one population. The estimate is really close to being like an average. The numerator adds up how far each response yi is from the estimated mean ¯yy¯ in squared units, and the denominator divides the sum by n-1, not n as you would expect for an average. What we would really like is for the numerator to add up, in squared units, how far each response yi is from the unknown population mean μ. But, we don't know the population mean μ, so we estimate it with ¯yy¯. Doing so "costs us one degree of freedom". That is, we have to divide by n-1 and not n, because we estimated the unknown population mean µ.

Now let's extend this thinking to arrive at an estimate for the population variance σ^2 in the simple linear regression setting. Recall that we assume that σ^2 is the same for each of the subpopulations. For our example on college entrance test scores and grade point averages, how many subpopulations do we have?



There are four subpopulations depicted in this plot. In general, there are as many subpopulations as there are distinct x values in the population. Each subpopulation has its own mean µY, which depends on x through µY = E(Y) = ß0 + ß1xµY = E(Y) = ß0 + ß1x. And, each subpopulation mean can be estimated using the estimated regression equation ^yi = b0+b1xiy^i=b0+b1xi.

The mean square error:

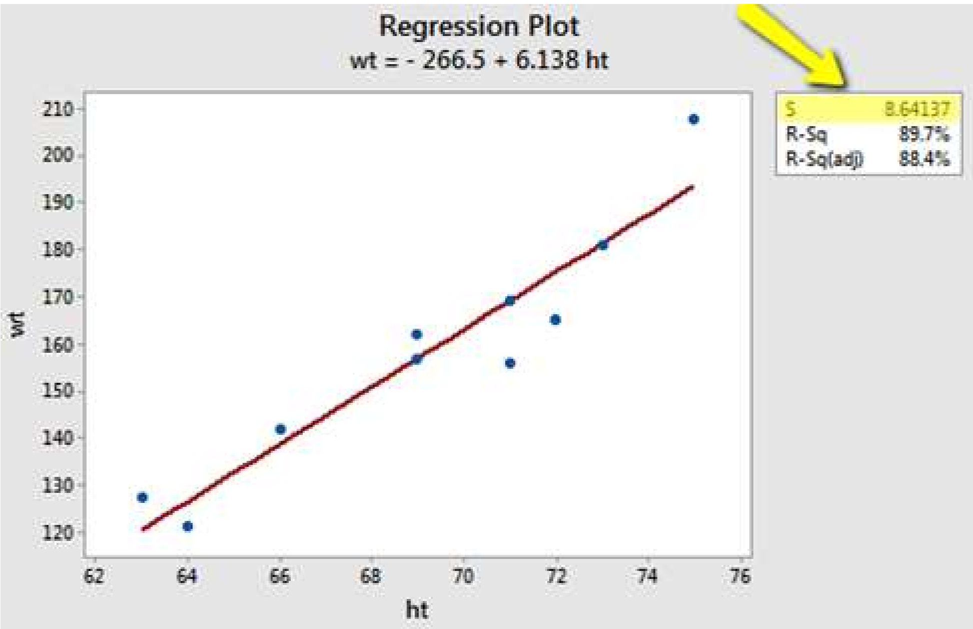
$$MSE = \sum ni = 1(Yi - Yi)2n - 2MSE = \sum i = 1n(Yi - Y^{i})2$$

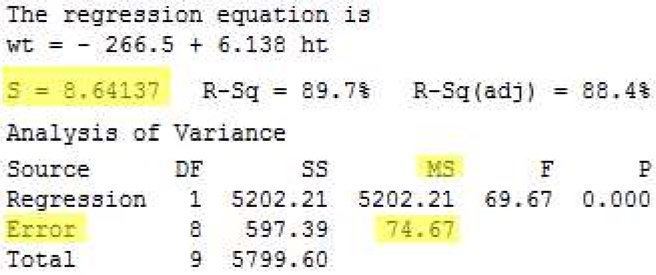
estimates σ^2, the common variance of the many subpopulations.

How does the mean square error formula differ from the sample variance formula? The similarities are more striking than the differences. The numerator again adds up, in squared units, how far each response yi is from its estimated mean. In the regression setting, though, the estimated mean is ^yiy^i. And, the denominator divides the sum by n-2, not n-1, because in using ^yiy^i to estimate µY, we effectively estimate two parameters - the population intercept ß0 and the population slope ß1. That is, we lose two degrees of freedom.

In practice, we will let statistical software, such as Minitab, calculate the mean square error (MSE) for us. The estimate of σ^2 shows up indirectly on Minitab's "fitted line plot." For example, for the student height and weight data (student_height_weight.txt), the quantity emphasized in the box, S = 8.64137, is the square root of MSE. That is, in general, S= MSES = MSE, which estimates s and is known as the regression standard error or the residual standard error. The fitted line plot here indirectly tells us, therefore, that MSE = 8.641372 = 74.67.

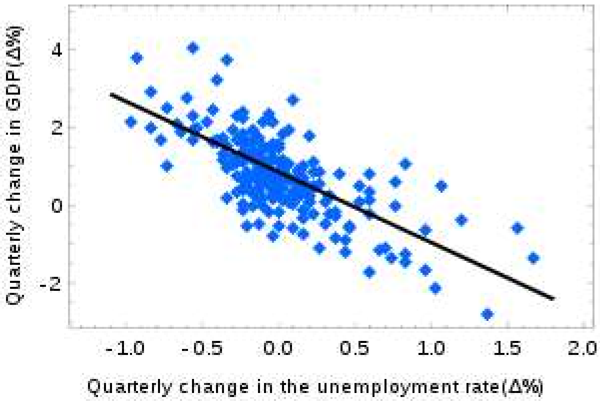
The estimate of σ^2 shows up directly in Minitab's standard regression analysis output. Again, the quantity S = 8.64137 is the square root of MSE. In the Analysis of Variance table, the value of MSE, 74.67, appears appropriately under the column labeled MS (for Mean Square) and in the row labeled Residual Error (for Error).





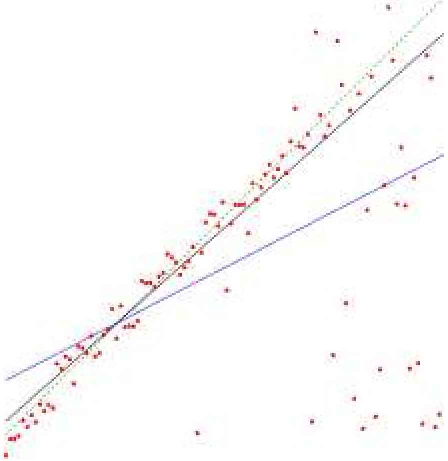
```
The regression equation is
wt = - 266.5 + 6.138 ht

S = 8.64137   R-Sq = 89.7%   R-Sq(adj) = 88.4%

Analysis of Variance

Source      DF       SS       MS      F      P
Regression   1  5202.21  5202.21  69.67  0.000
Error        8   597.39    74.67
Total        9  5799.60
```

2.22 MULTIPLE COEFFICIENTS OF DETERMINATION R2 AND MULTIPLE CORRELATION COEFFICIENT- R



Ordinary least squares regression of Okun's law. Since the regression line does not miss any of the points by very much, the R2 of the regression is relatively high.

Notes



Comparison of the Theil–Sen estimator (black) and simple linear regression (blue) for a set of points with outliers. Because of the many outliers, neither of the regression lines fits the data well, as measured by the fact that neither gives a very high R2.

In statistics, the coefficient of determination, denoted R2 or r2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

There are several definitions of R2 that are only sometimes equivalent. One class of such cases includes that of simple linear regression where r2 is used instead of R2. When an intercept is included, then r2 is simply the square of the sample correlation coefficient (i.e., r) between the observed outcomes and the observed predictor values. If additional regressors are included, R2is the square of the coefficient of multiple correlation. In both such cases, the coefficient of determination ranges from 0 to 1.

There are cases where the computational definition of R2 can yield negative values, depending on the definition used. This can arise when the predictions that are being compared to the corresponding outcomes have not been derived from a model-fitting procedure using those data. Even if a model-fitting procedure has been used, R2 may still be negative, for example when linear regression is conducted without including an intercept, or when a non-linear function is used to fit the data. In cases where negative values arise, the mean of the data provides a better fit to the outcomes than do the fitted function values, according to this particular criterion. Since the most general definition of the coefficient of determination is also known as the Nash–Sutcliffe model efficiency coefficient, this last notation is preferred in many fields, because denoting a goodness-of-fit indicator that can vary from -8 to 1 (i.e., it can yield negative values) with a squared letter is confusing.

When evaluating the goodness-of-fit of simulated (Y_{pred}) vs. measured (Y_{obs}) values, it is not appropriate to base this on the R2 of the linear regression (i.e., $Y_{obs} = m \cdot Y_{pred} + b$). The R2 quantifies the degree of any linear correlation between Yobs and Y_{pred}, while for the goodness-of-fit evaluation only one specific linear correlation should be taken into consideration: $Y_{obs} = 1 \cdot Y_{pred} + 0$ (i.e., the 1:1 line).

2.23 SIGNIFICANCE TEST AND CONFIDENCE INTERVALS, PREDICTION

A confidence interval is a range of values that is likely to contain an unknown population parameter. If you draw a random sample many times, a certain percentage of the confidence intervals will contain the population mean. This percentage is the confidence level.

Most frequently, you'll use confidence intervals to bound the mean or standard deviation, but you can also obtain them for regression coefficients, proportions, rates of occurrence (Poisson), and for the differences between populations.

Just as there is a common misconception of how to interpret P values, there's a common misconception of how to interpret confidence intervals. In this case, the confidence level is not the probability that a specific confidence interval contains the population parameter.

The confidence level represents the theoretical ability of the analysis to produce accurate intervals if you are able to assess many intervals and you know the value of the population parameter. For a specific confidence interval from one study, the interval either contains the population value or it does not there's no room for probabilities other than 0 or 1. And you can't choose between these two possibilities because you don't know the value of the population parameter.

"The parameter is an unknown constant and no probability statement concerning its value may be made."

Jerzy Neyman, original developer of confidence intervals.

This will be easier to understand after we discuss the graph below . . .

With this in mind, how do you interpret confidence intervals?

Confidence intervals serve as good estimates of the population parameter because the procedure tends to produce intervals that contain the parameter. Confidence intervals are comprised of the point estimate (the most likely value) and a margin of error around that point estimate. The margin of error indicates the amount of uncertainty that surrounds the sample estimate of the population parameter.

2.24 NON-LINEAR MODELS-CHOICE OF FUNCTIONAL FORMS, ESTIMATION

In statistics, nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations.

Although the linear relationship assumed so far in this chapter is often adequate, there are many cases in which a nonlinear functional form is more suitable. To keep things simple in this section we assume that we only have one predictor xx.

The simplest way of modelling a nonlinear relationship is to transform the forecast variable yyand/or the predictor variable xx before estimating a regression model. While this provides a non-linear functional form, the model is still linear in the parameters. The most commonly used transformation is the (natural) logarithm.

A log-log functional form is specified as

logy = ß0 + ß1logx + e.logy = ß0 + ß1 logx + e.

Notes

In this model, the slope ß1ß1 can be interpreted as an elasticity: ß1ß1 is the average percentage change in yy resulting from a 1%1% increase in xx. Other useful forms can also be specified. The log-linear form is specified by only transforming the forecast variable and the linear-log form is obtained by transforming the predictor.

Recall that in order to perform a logarithmic transformation to a variable, all of its observed values must be greater than zero. In the case that variable xx contains zeros, we use the transformation log(x+1)log(x+1); i.e., we add one to the value of the variable and then take logarithms. This has a similar effect to taking logarithms but avoids the problem of zeros. It also has the neat side-effect of zeros on the original scale remaining zeros on the transformed scale.

There are cases for which simply transforming the data will not be adequate and a more general specification may be required. Then the model we use is

y = f(x) + ey = f(x) + e

where ff is a nonlinear function. In standard (linear) regression, f(x) = ß0 + ß1 x f(x) = ß0 + ß1x.

In the specification of nonlinear regression that follows, we allow ff to be a more flexible nonlinear function of xx, compared to simply a logarithmic or other transformation.

One of the simplest specifications is to make ff piecewise linear. That is, we introduce points where the slope of ff can change. These points are called knots. This can be achieved by letting

x1, t = xx1, t = x and introducing variable x2, tx2, t

such that

x2, t = (x - c) + = {0x<c(x-c)x = cx2, t = (x-c) + = {0x<c(x-c)x = c

The notation (x-c) + (x-c) + means the value x-cx-c if it is positive and 0 otherwise. This forces the slope to bend at point cc. Additional bends can be included in the relationship by adding further variables of the above form.

An example of this follows by considering x = tx = t and fitting a piecewise linear trend to a time series.

Piece wise linear relationships constructed in this way are a special case of regression splines. In general, a linear regression spline is obtained using

x1 = xx2 = (x-c1) + …xk = (x-ck-1) + x1 = xx2 = (x-c1) + …xk = (x-ck-1)

where c1, …, ck-1c1, …, ck-1

are the knots (the points at which the line can bend). Selecting the number of knots (k-1k-1) and where they should be positioned can be difficult and somewhat arbitrary. Some automatic knot selection algorithms are available in some software, but are not yet widely used.

A smoother result can be obtained using piecewise cubics rather than piecewise lines. These are constrained to be continuous (they join up) and smooth (so that there are no sudden changes of direction, as we see with piecewise linear splines). In general, a cubic regression spline is written as

x1 = xx2 = x2x3 = x3x4 = (x-c1) +…xk = (x-ck-3) + .x1 = xx2 = x2x3 = x3x4 = (x-c1) +…xk = (x-ck-3).

Cubic splines usually give a better fit to the data. However, forecasts of yy become unreliable when xx is outside the range of the historical data.

2.25 SUMMARY

Notes

Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regressions. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

Model statistical-tool used in predicting future values of a target (dependent) variable on the basis of the behavior of a set of explanatory factors (independent variables). A type of regression analysis model, it assumes the target variable is predictable, not chaotic or random.

In the previous chapters, several models used in stock assessment were analysed, the respective parameters having been defined. In the corresponding exercises, it was not necessary to estimate the values of the parameters because they were given. In this chapter, several methods of estimating parameters will be analysed. In order to estimate the parameters, it is necessary to know the sampling theory and statistical inference.

The technique of the cohort ansalysis, applied to the structure of the catches of a cohort during its life, can be made with non constant intervals of time, Ti,. This means that the length classes structure of the catches of a cohort during its life, can also be analysed.

The Gauss–Markov theorem, named after Carl Friedrich Gauss and Andrey Markov, states that in a linear regression model in which the errors have expectation zero, are uncorrelated and have equal variances, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists. Here "best" means giving the lowest variance of the estimate, as compared to other unbiased, linear estimators.

The standard error (SE) of a statistic (usually an estimate of a parameter) is the standard deviation of its sampling distribution or an estimate of that standard deviation. If the parameter or the statistic is the mean, it is called the standard error of the mean (SEM).

The sampling distribution of a population mean is generated by repeated sampling and recording of the means obtained. This forms a distribution of different means, and this distribution has its own mean and variance. Mathematically, the variance of the sampling distribution obtained is equal to the variance of the population divided by the sample size. This is because as the sample size increases, sample means cluster more closely around the population mean.

Therefore, the relationship between the standard error and the standard deviation is such that, for a given sample size, the standard error equals the standard deviation divided by the square root of the sample size. In other words, the standard error of the mean is a measure of the dispersion of sample means around the population mean.

In regression analysis, the term "standard error" refers either to the square root of the reduced chi-squared statistic or the standard error for a particular regression coefficient (as used in, e.g., confidence intervals).

Standard error plays a very crucial role in the large sample theory. It also may form the basis for the testing of a hypothesis. The statistical inference involved in the construction of the confidence interval is mainly based on standard error.

Notes

The magnitude of the standard error gives an index of the precision of the estimate of the parameter. It is inversely proportional to the sample size, meaning that smaller samples tend to produce greater standard errors.

The standard deviation of a sample is generally designated by the Greek letter sigma (s). It can also be defined as the square root of the variance present in the sample.

The term "standard error" is used to refer to the standard deviation of various sample statistics such as the mean or median. For example, the "standard error of the mean" refers to the standard deviation of the distribution of sample means taken from a population. The smaller the standard error, the more representative the sample will be of the overall population.

The mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

In probability theory, the central limit theorem (CLT) establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed. The theorem is a key ("central") concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.

A large p-value and hence failure to reject this null hypothesis is a good result. It means that it is reasonable to assume that the errors have a normal distribution. Typically, assessment of the appropriate residual plots is sufficient to diagnose deviations from normality.

The Anderson-Darling Test measures the area between a fitted line (based on the chosen distribution) and a nonparametric step function (based on the plot points). The statistic is a squared distance that is weighted more heavily in the tails of the distribution.

Maximum likelihood estimation (MLE) is a technique used for estimating the parameters of a given distribution, using some observed data. For example, if a population is known to follow a normal distribution but the mean and variance are unknown, MLE can be used to estimate them using a limited sample of the population, by finding particular values of the mean and variance so that the observation is the most likely result to have occurred.

MLE is useful in a variety of contexts, ranging from econometrics to MRIs to satellite imaging. It is also related to Bayesian statistics.

Once sample data has been gathered through an observational study or experiment, statistical inference allows analysts to assess evidence in favor or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as tests of significance.

Prediction intervals are used in both frequentist statistics and Bayesian statistics: a prediction interval bears the same relationship to a future observation that a frequentist confidence interval or Bayesian credible interval bears to an unobservable population parameter: prediction intervals predict the distribution of individual future points, whereas confidence intervals and credible intervals of parameters predict the distribution of estimates of the true population mean or other quantity of interest that cannot be observed.

Point estimation gives us a particular value as an estimate of the population parameter. . Interval estimation gives us a range of values which is likely to contain the population parameter.

Point Prediction uses the models fit during analysis and the factor settings specified on the factors tool to compute the point predictions and interval estimates. The predicted values are updated as the levels are changed. Prediction intervals (PI) are found under the Confirmation node.

Partial correlation is the measure of association between two variables, while controlling or adjusting the effect of one or more additional variables. Partial correlations can be used in many cases that assess for relationship, like whether or not the sale value of a particular commodity is related to the expenditure on advertising when the effect of price is controlled.

The method of least squares is a standard approach in regression analysis to approximate the solution of over determined systems, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the residuals made in the results of every single equation.

Estimation of covariance matrices then deals with the question of how to approximate the actual covariance matrix on the basis of a sample from the multivariate distribution. Simple cases, where observations are complete, can be dealt with by using the sample covariance matrix.

A confidence interval is a range of values that is likely to contain an unknown population parameter. If you draw a random sample many times, a certain percentage of the confidence intervals will contain the population mean. This percentage is the confidence level.

Nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations.

2.26 SELF-ASSESSMENT QUESTIONS

1. What is Linear Regression? Discuss Classical Linear Regression with one explanatory variable.
2. Explain the least square estimations of regression parameters their properties.
3. What is Gauss-Markov Theory? Discuss about Gauss-Markov Theory.
4. Give the meaning of Standard Error. Explain about Standard Errors and Estimator of Errors.
5. What is Control limit theorem? Discuss about Control limit theorem.
6. What is Maximum Likelihood Estimator? Explain about Maximum Likelihood Estimator.
7. What is normality of error? Discuss Normality of errors and control limit theorem.
8. Explain about Maximum Likelihood Estimator.
9. What is Significance Test? Briefly explain about Significance Test.
10. Discuss about confidence intervals of estimates-z-test, t-test and f-ratio test.
11. Explain in details about extension of the two variable linear models.

Notes

12. Discuss three – variable linear model.

13. What is Co-efficient of Multiple Correlation? Explain about the Co-efficient of Multiple Correlations.

14. Discuss Partial Correlation Coefficient.

15. Explain about General Linear model with K- Explanatory variable.

16. Discuss Least-square estimates and their properties.

*****

3 Lesson

EXTENSIONS OF THE GENERAL MODEL

Objectives

The objectives of this lesson are to:

- Extensions of the General Model
- Dummy Variables
- Use of Dummy Variable in Seasonal Analysis
- Dummy Dependent Variable

Structure:

3.1 EXTENSIONS OF THE GENERAL MODEL OF ECONOMETRICS

Econometric techniques are used to estimate economic models, which ultimately allow you to explain how various factors affect some outcome of interest or to forecast future events. The ordinary least squares (OLS) technique is the most popular method of performing regression analysis and estimating econometric models, because in standard situations (meaning the model satisfies a series of statistical assumptions) it produces optimal (the best possible) results.

The proof that OLS generates the best results is known as the Gauss-Markov theorem, but the proof requires several assumptions. These assumptions, known as the classical linear regression model (CLRM) assumptions, are the following:

a) The model parameters are linear, meaning the regression coefficients don't enter the function being estimated as exponents (although the variables can have exponents).

b) The values for the independent variables are derived from a random sample of the population, and they contain variability.

c) The explanatory variables don't have perfect collinearity (that is, no independent variable can be expressed as a linear function of any other independent variables).

d) The error term has zero conditional mean, meaning that the average error is zero at any specific value of the independent variable(s).

e) The model has no heteroskedasticity (meaning the variance of the error is the same regardless of the independent variable's value).

f) The model has no autocorrelation (the error term doesn't exhibit a systematic relationship over time).

If the classical linear regression model (CLRM) doesn't work for your data because one of its assumptions doesn't hold, then you have to address the problem before you can finalize your analysis. Fortunately, one of the primary contributions of econometrics is the development of techniques to address such problems or other complications with the data that make standard model estimation difficult or unreliable. If historical data is available, forecasting typically involves the use of one or more quantitative techniques. If historical data isn't available, or if it contains significant gaps or is unreliable, then forecasting can actually be qualitative. Quantitative approaches to forecasting in econometrics involve the use of causal and/or smoothing models, whereas qualitative forecasting uses expert consensus and/or scenario analysis.

3.2 THE CONCEPT OF VARIABLES

Variables have defined by considering the technical words of Research Title and various Objectives. Variables represent the measurable traits that can change over the course of a scientific experiment.

Types of Variables

In all there are six basic variable types: dependent, independent, intervening, moderator, controlled and extraneous variables.

Independent and Dependent Variables

The variables were categorized into independent variables and dependent variables. Independent variables are the variables that are changed or controlled in a scientific experiment to test the effects on the dependent variable. Dependent variables are the variables being tested and measured in a scientific experiment.

In general, experiments purposefully change one variable, which is the independent variable. But a variable that changes in direct response to the independent variable is the dependent variable. Say there’s an experiment to test whether changing the position of an ice cube affects its ability to melt. The change in an ice cube's position represents the independent variable. The result of whether the ice cube melts or not is the dependent variable.

Intervening and Moderator Variables

Intervening variables link the independent and dependent variables, but as abstract processes, they are not directly observable during the experiment. For example, if studying the use of a specific teaching technique for its effectiveness, the technique represents the independent variable, while the completion of the technique's objectives by the study participants represents the dependent variable, while the actual processes used internally by the students to learn the subject matter represents the intervening variables.

By modifying the effect of the intervening variables - the unseen processes - moderator variables influence the relationship between the independent and dependent variables. Researchers measure moderator variables and take them into consideration during the experiment.

Constant or Controllable Variable

Sometimes certain characteristics of the objects under scrutiny are deliberately left unchanged. These are known as constant or controlled variables. In the ice cube experiment, one constant or controllable variable could be the size and shape of the cube. By keeping the ice cubes' sizes and shapes the same, it's easier to measure the differences between the cubes as they melt after shifting their positions, as they all started out as the same size.

Extraneous Variables

A well-designed experiment eliminates as many unmeasured extraneous variables as possible. This makes it easier to observe the relationship between the independent and dependent variables. These extraneous variables, also known as unforeseen factors, can affect the interpretation of experimental results. Lurking variables, as a subset of extraneous variables represent the unforeseen factors in the experiment.

Another type of lurking variable includes the confounding variable, which can render the results of the experiment useless or invalid. Sometimes a confounding variable could be a variable not previously considered. Not being aware of the confounding variable’s influence skews the experimental results. For example, say the surface chosen to conduct the ice-cube experiment was on a salted road, but the experimenters did not realize the salt was there and sprinkled unevenly, causing some ice cubes to melt faster. Because the salt affected the experiment's results, it's both a lurking variable and a confounding variable.

3.3 RELATIONSHIP BETWEEN DEPENDENT AND INDEPENDENT VARIABLES

The relationship between independent variable and dependent variables is those independent variables causes a change in dependent variable and that it not possible that a dependent variable could cause a change in an independent variable. For example, time spent studying causes a change in test score and it is not possible that test scores could cause a change in time spent studying.

Therefore, "Time Spent Studying" must be the independent variable and "Test Score" must be the dependent variable because the sentence doesn't make sense the other way around.

The other relationship can be traced from its terms independent and dependent referring to the relationship between these two types of variables. The terms have meaning only with respect to each other. In the case of the dependent variable, its value or behavior is considered reliant, to an extent, upon the value of the independent variable but not the other way around. That is why it is considered "dependent." The independent variable, on the other hand, is truly independent from the dependent variable. Its value does not change according to the value of the dependent variable.

In many researches, the major task for researchers is to be able to determine the relationship between the independent and dependent variables, such that if the independent variable is changed, then the researcher will be able to accurately predict how the dependent variable will change. When this correlation is determined, a further question is whether varying the independent variable caused the independent variable to change. This adds complexity and debate to the situation.

3.4 IMPORTANCE OF DEPENDENT AND INDEPENDENT VARIABLES

The importance of dependent and independent variables is that they guide the researchers to per sue their studies with maximum curiosity. Dependent and independent variables are important because they drive the research process. As defined earlier, a variable as opposed to a constant is simply anything that can vary and that many researchers consistently look at the relationship between these two variables. While the variation of an independent variable will influence the variation of dependent variable, both variables give the study a focus. If we were to study the effects of work experience on college performance, we might look at the grades of students who have worked prior to starting college and the grades of students who did not work prior to starting college. In this study, you may notice that both groups are students so student status remains constant between the two groups. You may also notice that work experience is not the same between the two groups, therefore work experience varies and is considered a variable. If we choose students for each group who are of similar age or similar background, we are holding these aspects constant and therefore, they too will not vary within our study.

Dependent and independent variables are also important because they determine the cause and effects in research. Although not all independent and dependent variables are causal related variables, the notion of cause and effect can help clarify the idea of "independence" in the independent variable and "dependence" in the dependent variable. In the studying and scores example, cause-effect is fairly obvious, and therefore, it is relatively easy to understand what the independent dependent variables are. However, as Kalof, Dan and Dietz (2008:36) state, “It can be hard to understand what variables are independent (causes) and what variables are dependent (effects) when we are reading research or thinking about the implications of theory”.

3.5 THE CONCEPT OF DUMMY VARIABLES

Dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study. In research design, a dummy variable is often used to distinguish different treatment groups.

Dummy variables are "proxy" variables or numeric stand-ins for qualitative facts in a regression model. In regression analysis, the dependent variables may be influenced not only by quantitative variables (income, output, prices, etc.), but also by qualitative variables (gender, religion, geographic region, etc.). A dummy independent variable (also called a dummy explanatory variable) which for some observation has a value of 0 will cause that variable's coefficient to have no role in influencing the dependent variable, while when the dummy takes on a value 1 its coefficient acts to alter the intercept. For example, suppose membership in a group is one of the qualitative variables relevant to a regression. If group membership is arbitrarily assigned the value of 1, then all others would get the value 0. Then the intercept (the value of the dependent variable if all other explanatory variables hypothetically took on the value zero) would be the constant term for non-members but would be the constant term plus the coefficient of the membership dummy in the case of group members.

Dummy variables are used frequently in time series analysis with regime switching, seasonal analysis and qualitative data applications. Dummy variables are involved in studies for economic forecasting, bio-medical studies, credit scoring, response modelling, etc. Dummy variables may be incorporated in traditional regression methods or newly developed modeling paradigms.

A model with a dummy dependent variable (also known as a qualitative dependent variable) is one in which the dependent variable, as influenced by the explanatory variables, is qualitative in nature. Some decisions regarding 'how much' of an act must be performed involve a prior decision making on whether to perform the act or not. For example, the amount of output to produce, the cost to be incurred, etc. involve prior decisions on whether to produce or not, whether to spend or not, etc. Such "prior decisions" become dependent dummies in the regression model.

For example, the decision of a worker to be a part of the labour force becomes a dummy dependent variable. The decision is dichotomous, i.e., the decision has two possible outcomes: yes and no. So the dependent dummy variable Participation would take on the value 1 if participating, 0 if not participating. Some other examples of dichotomous dependent dummies are cited below:

Decision: Choice of Occupation. Dependent Dummy: Supervisory = 1 if supervisor, 0 if not supervisor.

Decision: Affiliation to a Political Party. Dependent Dummy: Affiliation = 1 if affiliated to the party, 0 if not affiliated.

Decision: Retirement. Dependent Dummy: Retired = 1 if retired, 0 if not retired.

When the qualitative dependent dummy variable has more than two values (such as affiliation to many political parties), it becomes a multiresponse or a multinomial or polychotomous model.

3.6 SEASONALITY

Seasonality is a characteristic of a time series in which the data experiences regular and predictable changes that recur every calendar year. Any predictable change or pattern in a time series that recurs or repeats over a one-year period can be said to be seasonal. Seasonal effects are different from cyclical effects, as seasonal cycles are observed within one calendar year, while cyclical effects, such as boosted sales due to low unemployment rates, can span time periods shorter or longer than one calendar year.

Seasonality refers to periodic fluctuations in certain business areas that occur regularly based on a particular season. A season may refer to a time period as denoted by the calendar seasons, such as summer or winter, as well as commercial seasons, such as the holiday season. Companies that understand the seasonality of their business can time inventories, staffing, and other decisions to coincide with the expected seasonality of the associated activities.

It is important to consider the effects of seasonality when analyzing stocks from a fundamental point of view. A business that experiences higher sales in certain seasons appears to be making significant gains during peak seasons and significant losses during off-peak seasons. If this is not taken into consideration, an investor may choose to buy or sell securities based on the activity at hand without accounting for the seasonal change that subsequently occurs as part of the company's seasonal business cycle.

Examples of Seasonality

Seasonality can be observed in a variety of predictable changes in costs or sales as it relates to the regular transition through the times of year. For example, if you live in a climate with cold winters and warm summers, your home's heating costs probably rise in the winter and fall in the summer. You reasonably expect the seasonality of your heating costs to recur every year. Similarly, a company that sells sunscreen and tanning products within the United States sees sales jump up in the summer but drop in the winter.

Temporary Workers

Large retailers, such as Wal-Mart, may hire temporary workers in response to the higher demands associated with the holiday season. In 2014, Wal-Mart anticipated hiring approximately 60,000 employees to help offset the increased activity expected in stores. This determination was made by examining traffic patterns from previous holiday seasons and using that information to extrapolate what may be expected in the coming season. Once the season is over, a number of the temporary employees will be released as they are no longer needed based on the post-season traffic expectations.

Adjusting Data for Seasonality

A lot of data is affected by the time of the year, and adjusting for the seasonality means that more accurate relative comparisons can be drawn between different time periods. Adjusting data for seasonality evens out periodic swings in statistics or movements in supply and demand related to changing seasons. By using a tool known as Seasonally Adjusted Annual Rate (SAAR), seasonal variations in the data can be removed. For example, homes tend to sell more quickly and at higher prices in the summer than in the winter. As a result, if a person compares summer real estate sales prices to median prices from the previous year, he may get a false impression that prices are rising. However, if he adjusts the initial data based on the season, he can see whether values are truly rising or just being momentarily increased by the warm weather.

3.7 SEASONAL ANALYSIS

Seasonal Analysis is a technique aimed at analyzing economic data with the purpose of removing fluctuations that take place as a result of seasonal factors.

Description: Seasonal analysis of economic/time data plays a crucial role analyzing/ judging the general trend. In the world of finance, comparison of economic data is of immense importance in order to ascertain the growth and performance of a company. In this process, seasonal factors might create big fluctuations in the pattern.

For example, sales of air conditioners are at their peak during summers and quite less during winters. Thus, to study the general trend of sales of air conditioners, the data needs to be seasonally adjusted.

3.8 USE OF DUMMY VARIABLE IN SEASONAL ANALYSIS

1. Dummy variables are used frequently in time series analysis with regime switching, seasonal analysis and qualitative data applications.
2. Dummy variables are involved in studies for economic forecasting, bio-medical studies, credit scoring, response modelling, etc.
3. Dummy variable plays a crucial role analyzing/judging the general trend.
4. In the world of finance, comparison of economic data is of immense importance in order to ascertain the growth and performance of a company.
5. In this process, seasonal factors might create big fluctuations in the pattern.
6. These are used for sales of air conditioners are at their peak during summers and quite less during winters.
7. Thus, to study the general trend of sales of air conditioners, the data needs to be seasonally adjusted.

3.9 DUMMY DEPENDENT VARIABLES

Analysis of dependent dummy variable models can be done through different methods. One such method is the usual OLS method, which in this context is called the linear probability model. An alternative method is to assume that there is an unobservable continuous latent variable Y^* and that the observed dichotomous variable $Y = 1$ if $Y^* > 0$, 0 otherwise. This is the underlying concept of the logit and probit models. These models are discussed in brief below.

A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study. In research design, a dummy variable is often used to distinguish different treatment groups. In the simplest case, we would use a 0,1 dummy variable where a person is given a value of 0 if they are in the control group or a 1 if they are in the treated group. Dummy variables are useful because they enable us to use a single regression equation to represent multiple groups. This means that we don't need to write out separate equation models for each subgroup. The dummy variables act like 'switches' that turn various parameters on and off in an equation. Another advantage of a 0,1 dummy-coded variable is that even though it is a nominal-level variable you can treat it statistically like an interval-level variable (if this made no sense to you, you probably should refresh your memory on levels of measurement). For instance, if you take an average of a 0,1 variable, the result is the proportion of 1s in the distribution.

$$Y_i = \beta_0 + \beta_1 Z_i + e_i$$

Where,

Y_i = outcome score for the ith unit

β_0 = co-efficient for the intercept

β_1 = co-efficient for the slope

Z_i = 1 if ith unit is in the treatment group

0 if ith unit is in the control group

e_i = residual for the ith unit

To illustrate dummy variables, consider the simple regression model for a posttest-only two-group randomized experiment. This model is essentially the same as conducting a t-test on the posttest means for two groups or conducting a one-way Analysis of Variance (ANOVA). The key term in the model is b1, the estimate of the difference between the groups. To see how dummy variables work, we'll use this simple model to show you how to use them to pull out the separate sub-equations for each subgroup. Then we'll show how you estimate the difference between the subgroups by subtracting their respective equations. You'll see that we can pack an enormous amount of information into a single equation using dummy variables. All I want to show you here is that b1 is the difference between the treatment and control groups.

To see this, the first step is to compute what the equation would be for each of our two groups separately. For the control group, Z = 0. When we substitute that into the equation, and recognize that by assumption the error term averages to 0, we find that the predicted value for the control group is b0, the intercept. Now, to figure out the treatment group line, we substitute the value of 1 for Z, again recognizing that by assumption the error term averages to 0. The equation for the treatment group indicates that the treatment group value is the sum of the two beta values.

$$Y_i = \beta_0 + \beta_1 Z_i + e_i$$

First, determine effect for each group:

For control group ($Z_i = 0$):

$$Y_c = \beta_0 + \beta_1(0) + 0$$

$$Y_c = \beta_0$$

e_i averages to 0 across the group

For treatment group ($Z_i = 1$):

$$Y_T = \beta_0 + \beta_1(1) + 0$$

$$Y_T = \beta_0 + \beta_1$$

Now, we're ready to move on to the second step - computing the difference between the groups. How do we determine that? Well, the difference must be the difference between the equations for the two groups that we worked out above. In other word, to find the difference between the groups we just find the difference between the equations for the two groups! It should be obvious from the figure that the difference is b1. Think about what this means. The difference between the groups is b1. OK, one more time just for the sheer heck of it. The difference between the groups in this model is b1!

Then, find the difference between the two groups:

Treatment	**Control**
$Y_T = \beta_0 + \beta_1$	$Y_C = \beta_0$

$$Y_T - Y_C = (\beta_0 + \beta_1) - \beta_0$$
$$Y_T - Y_C = \beta_0 + \beta_1 - \beta_0$$
$$Y_T - Y_C = \beta_1$$

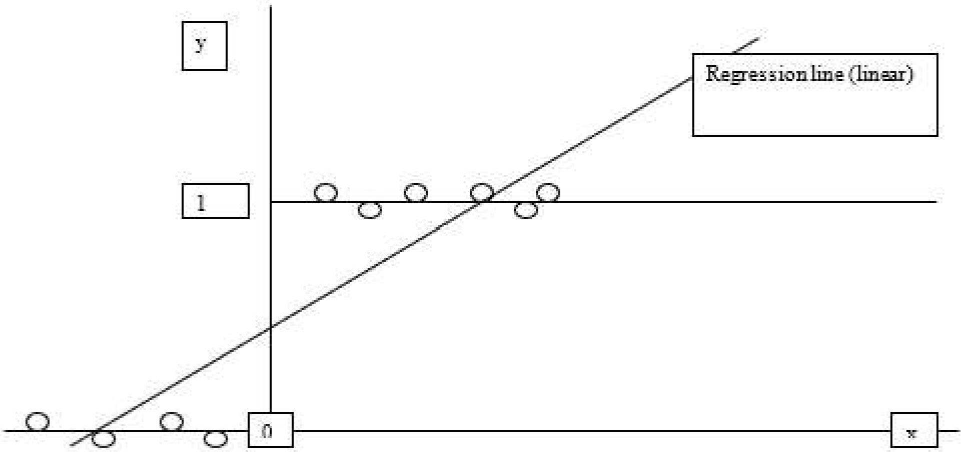
Whenever you have a regression model with dummy variables, you can always see how the variables are being used to represent multiple subgroup equations by following the two steps described above:

a) Create separate equations for each subgroup by substituting the dummy values.

b) Find the difference between groups by finding the difference between their equations.

3.10 DUMMY DEPENDENT VARIABLE MODELS

The dependent variable can also take the form of a dummy variable, where the variable consists of 1s and 0s. If it takes the value of 1, it can be interpreted as a success. Examples might include home ownership or mortgage approvals, where the dummy variable takes the value of a 1 of someone owns a home and 0 if they do not. This can then be regressed on a mix of variables including both other dummy variables and the usual continuous variables. The scatterplot of such a model would appear as follows:



There are a number of problems with the above approach, usually called the Linear Probability Model (LPM), which is estimated in the usual way using OLS. The regression line is not a good fit of the data so the usual measures of this, such as the R2 statistic are not reliable. There are other problems with this approach:

1) There will be heteroskedasticity in any model estimated using the LPM approach.

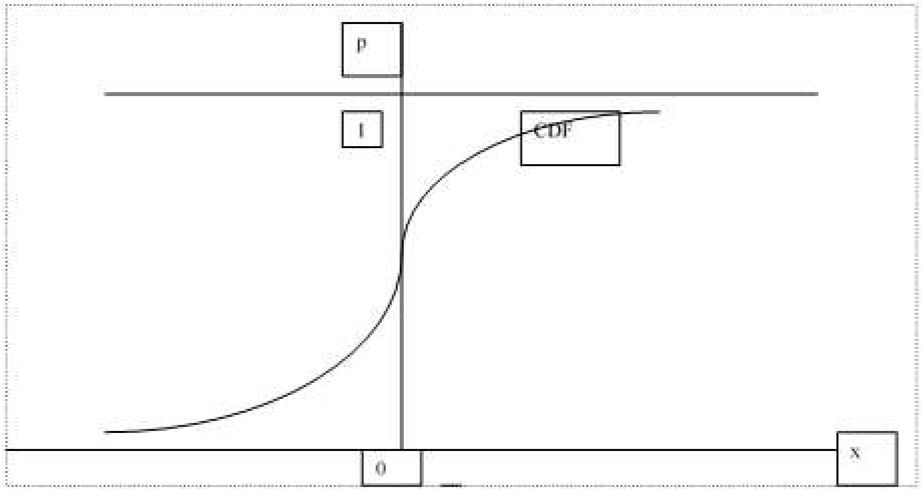
2) It is possible the LPM will produce estimates that are greater than 1 and less than 0, which is difficult to interpret as the estimates are probabilities and a probability of more than 1 does not exist.

3) The error term in such a model is likely to be non-normal, as it will follow the distribution in the above diagram.

4) The largest problem is that the relationship between the variables in this model is likely to be non-linear. This suggests we need a different type of regression line, that will fit the data more accurately, such as a 'S' shaped curve.

Logit Model

The Logit model is a significant improvement on the LPM model. It is based on the cumulative distribution function (CDF) of a random variable, with the Logit model following the Logistic CDF, giving the following relationship:



As is evident the above regression line is non-linear, giving a more realistic description of the data, with very little change in the probability at the extreme value that the explanatory variable can take. In the Logit model the probability of the dependent variable taking the value of 1, for a given value of the explanatory variable can be expressed as:

$$P_i = \frac{1}{1 + e^{-Z_i}}$$

$$\text{Where: } Z_i = \alpha_0 + \alpha_1 \chi_i$$

The model is then expressed as the odds ratio, which is simply the probability of an event occurring relative to the probability that it will not occur. Then by taking the natural log of the odds ratio we produce the Logit (Li), as follows:

$$L_i = \ln\left(\frac{P_i}{1 - P_i}\right) = Z_i = \alpha_0 + \alpha_1 \chi_i$$

The above relationship shows that L is linear in x, the probabilities (p) are not linear. In general the Logit model is estimated using the Maximum Likelihood approach. (We will cover this in more detail later, but in large samples it gives the same results as OLS)

Probit Model

The Probit model is very similar to the Logit model, with the normal CDF used instead of the Logistic CDF, in most other respects it follows the Logit approach. However it can be interpreted in a slightly different way. McFadden for instance interprets the model using utility theory, where the probability of an event occurring can be expressed as a latent variable, determined by the explanatory variables in the model.

3.11 SUMMARY

Econometric techniques are used to estimate economic models, which ultimately allow you to explain how various factors affect some outcome of interest or to forecast future events. The ordinary least squares (OLS) technique is the most popular method of performing regression analysis and estimating econometric models, because in standard situations (meaning the model satisfies a series of statistical assumptions) it produces optimal (the best possible) results.

If the classical linear regression model (CLRM) doesn't work for your data because one of its assumptions doesn't hold, then you have to address the problem before you can finalize your analysis. Fortunately, one of the primary contributions of econometrics is the development of techniques to address such problems or other complications with the data that make standard model estimation difficult or unreliable. If historical data is available, forecasting typically involves the use of one or more quantitative techniques. If historical data isn't available, or if it contains significant gaps or is unreliable, then forecasting can actually be qualitative.

Variables have defined by considering the technical words of Research Title and various Objectives. Variables represent the measurable traits that can change over the course of a scientific experiment.

The variables were categorized into independent variables and dependent variables. Independent variables are the variables that are changed or controlled in a scientific experiment to test the effects on the dependent variable. Dependent variables are the variables being tested and measured in a scientific experiment.

In general, experiments purposefully change one variable, which is the independent variable. But a variable that changes in direct response to the independent variable is the dependent variable. Say there's an experiment to test whether changing the position of an ice cube affects its ability to melt. The change in an ice cube's position represents the independent variable. The result of whether the ice cube melts or not is the dependent variable.

Intervening variables link the independent and dependent variables, but as abstract processes, they are not directly observable during the experiment. For example, if studying the use of a specific teaching technique for its effectiveness, the technique represents the independent variable, while the completion of the technique's objectives by the study participants represents the dependent variable, while the actual processes used internally by the students to learn the subject matter represents the intervening variables.

Dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study. In research design, a dummy variable is often used to distinguish different treatment groups.

Dummy variables are "proxy" variables or numeric stand-ins for qualitative facts in a regression model. In regression analysis, the dependent variables may be influenced not only by quantitative variables (income, output, prices, etc.), but also by qualitative variables (gender, religion, geographic region, etc.). A dummy independent variable (also called a dummy explanatory variable) which for some observation has a value of 0 will cause that variable's coefficient to have no role in influencing the dependent variable, while when the dummy takes on a value 1 its coefficient acts to alter the intercept.

Dummy variables are used frequently in time series analysis with regime switching, seasonal analysis and qualitative data applications. Dummy variables are involved in studies for economic forecasting, bio-medical studies, credit scoring, response modelling, etc. Dummy variables may be incorporated in traditional regression methods or newly developed modeling paradigms.

A model with a dummy dependent variable (also known as a qualitative dependent variable) is one in which the dependent variable, as influenced by the explanatory variables, is qualitative in nature. Some decisions regarding 'how much' of an act must be performed involve a prior decision making on whether to perform the act or not. For example, the amount of output to produce, the cost to be incurred, etc. involve prior decisions on whether to produce or not, whether to spend or not, etc. Such "prior decisions" become dependent dummies in the regression model.

Seasonality is a characteristic of a time series in which the data experiences regular and predictable changes that recur every calendar year. Any predictable change or pattern in a time series that recurs or repeats over a one-year period can be said to be seasonal. Seasonal effects are different from cyclical effects, as seasonal cycles are observed within one calendar year, while cyclical effects, such as boosted sales due to low unemployment rates, can span time periods shorter or longer than one calendar year.

Seasonal Analysis is a technique aimed at analyzing economic data with the purpose of removing fluctuations that take place as a result of seasonal factors.

Seasonal analysis of economic/time data plays a crucial role analyzing/judging the general trend. In the world of finance, comparison of economic data is of immense importance in order to ascertain the growth and performance of a company. In this process, seasonal factors might create big fluctuations in the pattern. For example, sales of air conditioners are at their peak during summers and quite less during winters. Thus, to study the general trend of sales of air conditioners, the data needs to be seasonally adjusted.

3.12 SELF-ASSESSMENT QUESTIONS

1. Discuss about the Extensions of the general model of Econometrics.
2. What is Variable? Explain the Concept of Variables.
3. Discuss the relationship between dependent and independent Variables.
4. Give the meaning of independent variable? Explain about importance of dependent and independent variables.
5. What is Dummy Variable? Discuss the concept of Dummy Variables.
6. Give the meaning of Seasonality. Write note on: Seasonality.
7. What is Seasonal Analysis? Discuss applications of Seasonal Analysis.
8. Explain various use of Dummy variable in Seasonal Analysis.
9. What is Dummy Dependent Variable? Discuss in details about Dummy Dependent Variables.
10. Explain about Dummy Dependent Variable Models.

*****

4 Lesson

VIOLATIONS OF THE ASSUMPTIONS OF THE CLASSICAL MODEL

Objectives

The objectives of this lesson are to:

- Violations of the assumptions of the classical model
- Errors in variables consequence, Methods of estimation-classical method of maximum likelihood, use of instrumental variable
- Autocorrelation –Sources, Consequences, GLSM. Tests for autocorrelation, Remedial measures, Prediction
- Heteroscedasticity- Nature and consequences, Heteroscedasticity structures, Tests for Heteroscedasticity, Remedial measures the methods of weighted least square
- Multicolinearity - Implications consequences, Tests for multicolinearity, Methods of estimation Multicolinearity and prediction, Remedial measures

Structure:

Notes

4.1 VIOLATIONS OF THE ASSUMPTIONS OF THE CLASSICAL MODEL

The classical general equilibrium model aims to describe the economy by aggregating the behavior of individuals and firms. Note that the classical general equilibrium model is unrelated to classical economics, and was instead developed within neoclassical economics beginning in the late 19th century. An econometric model specifies the statistical relationship that is believed to hold between the various economic quantities pertaining to a particular economic phenomenon.

Classical economics can be traced to the pioneering work of Adam Smith (often referred to as the father of economics). The specific event launching the modern study of economics, as well as classical economics, was the publication by Adam Smith of An Inquiry into the Nature and Causes of the Wealth of Nations in 1776. In this book, Smith contended that the "wealth of a nation" was attributed to the production of goods and services (rather than stockpiles of gold in the royal vault, which was the prevailing view at the time). And this production was best achieved by unrestricted market exchanges with buyers and sellers motivated by the pursuit of self-interests.

The work by Smith was refined and enhanced by scores of others over the ensuing 150 years, including Jean-Baptiste Say, John Stuart Mill, David Ricardo, Thomas Robert Malthus, and Alfred Marshall, to name just a few. Their work led to the creation of a sophisticated body of principles and analyses that offered insight into a wide range of economic phenomena--both microeconomic and macroeconomic. Many of these principles remain essential to the modern microeconomic theory. And while classical economics was largely discredited by John Maynard Keynes and advocates of Keynesian economics from the 1930s through the 1970s (due in large part to the Great Depression), it has reemerged (albeit with modifications) in recent decades.

Flexible Prices

The first assumption of classical economics is that prices are flexible. Price flexibility means that markets are able to adjust quickly and efficiently to equilibrium. While this assumption does not mean that every market in the economy is in equilibrium at all times, any imbalance (shortage or surplus) is short lived. Moreover, the adjust to equilibrium is accomplished automatically through the market forces of demand and supply without the need for government action.

The most important macroeconomic dimension of this assumption applies to resource markets, especially labor markets. The unemployment of labor, particularly involuntary unemployment, arises if a surplus exists in labor markets. With a surplus, the quantity of labor supplied exceeds the quantity of labor demanded--at the exist price of labor (wages). With flexible prices, any surplus is temporary. Wages fall to eliminate the surplus imbalance and restore equilibrium and achieve full employment.

If, for example, aggregate demand in the economy takes a bit of a drop (perhaps due to fewer exports of goods to other countries), then production also declines (temporarily) and so too does the demand for labor, creating a surplus of labor and involuntarily unemployed workers. However, flexible prices mean that wages decline to eliminate the surplus.

Say's Law

The second assumption of classical economics is that the aggregate production of good and services in the economy generates enough income to exactly purchase all output. This notion commonly summarized by the phrase "supply creates its own demand" is attributed

Notes

to the Jean-Baptiste Say, a French economist who helped to popularize the work of Adam Smith in the early 1800s. Say's law was a cornerstone of classical economics, and although it was subject to intense criticism by Keynesian economists, it remains relevant in modern times and is reflected in the circular flow model.

Say's law is occasionally misinterpreted as applying to a single good, that is, the production of a good is ensured to be purchased by waiting buyers. That law actually applies to aggregate, economy-wide supply and demand. A more accurate phrase is "aggregate supply creates its own aggregate demand." This interpretation means that the act of production adds to the overall pool of aggregate income, which is then used to buy a corresponding value of production - although most likely not the original production.

This law, first and foremost, directed attention to the production or supply-side of the economy. That is, focus on production and the rest of the economy will fall in line. Say's law further implied that extended periods of excess production and limited demand, the sort of thing that might cause an economic downturn, were unlikely. Economic downturns could occur, but not due to the lack of aggregate demand.

Saving-Investment Equality

The last assumption of classical economics is that saving by the household sector exactly matches investment expenditures on capital goods by the business sector. A potential problem with Say's law is that not all income generated by the production of goods is necessarily spent by the household sector on consumption demand--some income is saved. In other words, while the production of $100 million of output generates $100 million of income, the household sector might choose to spend only $90 million, directing the remaining $10 million to saving. If so, then supply does NOT create its own demand. Supply falls $10 million short of creating enough demand.

If this happens, then producers reduce production and lay off workers, which causes a drop in income and induces a decline in consumption, which then triggers further reductions in production, employment, income, and consumption in a contractionary downward spiral.

However, if this $10 million of saving is matched by an equal amount of investment, then no drop off in aggregate demand occurs. Such a match between saving and investment is assured in classical economics through flexible prices. However, in this case price flexibility applies to interest rates. Should saving not match investment, then interest rates adjust to restore balance. In particular, if saving exceeds investment, then interest rates fall, which stimulates investment and curtails saving until the two are once again equal.

4.2 CLASSICAL LINEAR REGRESSION ASSUMPTIONS

Assumptions form the foundation upon which theories, models, and analyses are constructed. They simplify and highlight the problem or topic under study. Even though assumptions often appear to be "unrealistic," when properly used they make it possible to analyze an exceedingly complex set of events.

Abstraction

Assumptions are inherently abstract and seemingly unrealistic. However, they make it possible to identify a specific cause-and-effect relation by assuming other influences are not involved. For example, the law of demand is the relation between demand price and quantity demanded. Demand, however, is also affected by factors other than demand price, such as buyers' income, the prices of other goods, or buyers' preferences. When working with the law of demand, it is essential to assume that these other factors do not influence demand when identifying the law of demand.

However, without an abstract assumption holding these other influences unchanged, the law of demand relation is lost in the confusion.

Assumptions are used for two primary reasons--to simplify a complex analysis and to control an analysis.

Simplification: One important use of assumptions is to simplify an analysis. The world is complex. The economy is complex. A multitude of forces are perpetually at work. Making sense of the apparent chaos is the goal of science. This goal is pursued by focusing on one topic, one problem, and one segment of the world at a time. In so doing, it is essential to assume that other aspects of the world are unchanged or irrelevant. Simplifying assumptions often establish ideal benchmarks that can be used to evaluate real world conditions.

For example, a study of short-run production that is designed to identify the law of diminishing marginal returns, is likely to ignore (that is, assume as irrelevant) the government sector and government regulation of business. Is this totally realistic? No. But it does simplify the analysis. It enables the analysis of those aspects of the complex world that are MOST relevant to the law of diminishing marginal returns.

Control: Assumptions are also commonly used as control variables. The use of seemingly unrealistic assumptions makes it possible to control an analysis and to identify individual cause-and-effect relations. That is, at first a factor might be assumed constant (implementing that ceteris paribus assumption) merely to see what happens when the factor changes.

For example, the standard market analysis employs the ceteris paribus assumption to hold demand and supply determinants constant when deriving the market equilibrium. They are not REALLY constant. But, by holding them constant initially, each can be changed separately (that is, the ceteris paribus assumption is relaxed) to analyze how each affects equilibrium.

Misuse and Politics

Unfortunately, economic analysis occasionally makes excessive use of unrealistic assumptions, assumptions that not only define the problem but ensure particular conclusions. For example, the study of pollution externalities might begin with the assumption of a competitive market, free of market failures. In so doing, the problem of pollution is effectively assumed away, which is not only unrealistic, but defeats the purpose of the analysis. However, if the analysis is intended to "prove" pollution is not a problem, then the goal has been achieved.

Much like chemists occasionally blow up their laboratories, economists do misuse assumptions. Whether they realize it or not, economists are inclined to use economic theories that conform to preconceived political philosophies and world views. Liberals work with liberal economic theories and conservatives work with conservative economic theories. This, by itself, is no crime. The ongoing debate and competition of ideas brings out the best of both and enables a better overall understanding of the economy. However, the temptation to use unrealistic and unjustified assumptions that produce conclusions and support policies consistent with preconceived beliefs is always present. Doing so is not necessarily good.

Notes

4.3 VIOLATIONS OF CLASSICAL LINEAR REGRESSION ASSUMPTIONS

Mis-Specification

Assumption 1: $Y = X\beta + \varepsilon$

a) What if the true specification is $Y = X\beta + Z\gamma + \varepsilon$ but we leave out the relevant variable Z?

Then the error in the estimated equation is really the sum $Z\beta + \varepsilon$. Multiply the true regression by X' to get the mis-specified OLS:

$X'Y = X'X\beta + X'Z\gamma + X'\varepsilon$.

The OLS estimator is $b = (X'X)^{-1} X'Y = (X'X)^{-1} X'X\beta + (X'X)^{-1} X'Z\gamma + (X'X)^{-1} X'\varepsilon$.

The last term is on average going to vanish, so we get $b = \beta + (X'X)^{-1} X'Z\gamma$. Unless $\gamma = 0$ or in the data, the regression of X on Z is zero, the OLS b is biased.

b) What if the true specification is $Y = X\beta + \varepsilon$ but we include the irrelevant variable Z: $Y = X\beta + Z\gamma + (\varepsilon - Z\gamma)$. The error is $\varepsilon^* = \varepsilon - Z\gamma$. $\mathrm{Var}(\varepsilon^*) = \mathrm{var}(\varepsilon) + \gamma'\mathrm{var}(Z)\gamma$.

The estimator of [β γ]' is

$$\begin{bmatrix} b \\ g \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix}^{-1} \begin{bmatrix} X'Y \\ Z'Y \end{bmatrix}$$

The expected value of this is $E\begin{bmatrix} b \\ g \end{bmatrix} = \begin{bmatrix} \beta + (X'X)^{-1}X'Z0 \\ 0 \end{bmatrix} = \begin{bmatrix} \beta \\ 0 \end{bmatrix}$.

Thus the OLS produces an unbiased estimate of the truth when irrelevant variables are added. However, the standard error of the estimate is enlarged in general by g'Z'Zg/(n-k) (since e*'e* = e'e - 2e'Zg + g'Z'Zg). This could easily lead to the conclusion that $\beta = 0$ when in fact it is not.

c) What if the coefficients change within the sample, so β is not a constant? Suppose that $\beta_i = \beta + Z_i\gamma$. Then the proper model is $Y = X(\beta + Z\gamma) + \varepsilon = X\beta + XZ\gamma + \varepsilon$. Thus we need to include the interaction term XZ. If we do not, then we are in the situation (a) above, and the OLS estimates of the co-efficients of X will be biased. On the other hand, if we include the interaction term when it is not really appropriate, the estimators are unbiased but not minimum variance. We can get fooled about the true value of β.

How do you test whether the interactions belong or not. Run an unconstrained regression (which includes interactions) and then run a constrained regression (set interaction co-efficients equal to zero). $[(SSE_{const} - SSE_{unconst})/q]/[SSE_{unconst}/(n-k)] \sim F_{q,n-k}$ where q = number of interaction terms.

d) Many researchers do a "search" for the proper specification. This can lead to spurious results and we will look at this is some detail in a lecture to follow.

Censored Data and Frontier Regression

Assumption 2: $E[\varepsilon|X]=0$.

Suppose that $E[\varepsilon_i|X] = \mu \neq 0$. Note: this is the same for all i.

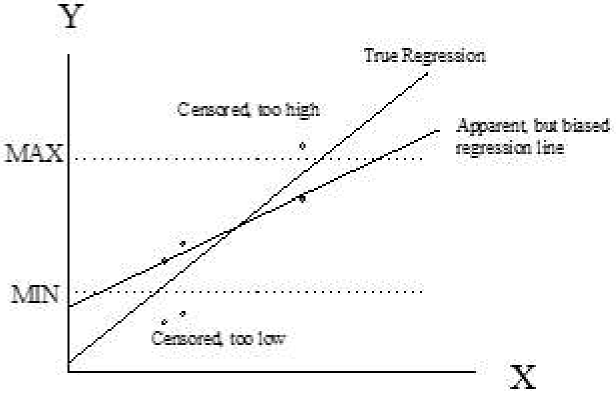
Notes

$$b = \left(X'X\right)^{-1}X'Y = \left(X'X\right)^{-1}X'(X\beta+\varepsilon) = b + \left(X'X\right)^{-1}X'\varepsilon.$$

Thus, $E\left[b\right] = \beta + m\left(X'X\right)^{-1}X'1$. The term $\left(X'X\right)^{-1}X'1$ is the regression of 1 on X, but the first column of X is 1 so the resulting regression co-efficients must be [1 0 0…0]'. As a result $E\left[b\right] = \beta + [\mu 0\ 0\ \ldots\ 0]'$. Only the intercept is biased.

Now suppose that $E[\varepsilon_i \mid X] = \mu_i$ but this varies with i. That is, $\mu \neq \mu 1$. By reasoning like the above, $E\left[b\right] = \beta + \left(X'X\right)^{-1}X'\mu$. The regression of μ on X will in general have non-zero co-efficients everywhere and the estimate of b will be biased in all ways.

In particular, what if the data was censored in the sense that only observations of Y that are not too small nor too large are included in the sample: MIN $\leq Y_i \leq$ MAX. Hence for values of X_i such that $X_i\beta$ are very small or vary large, only errors that are high and low respectively will lead to observations in the dataset. This can lead to the type of bias discussed above for all the co-efficients, not just the intercept. See the graph below where the slope is also biased.



Frontier Regression: Stochastic Frontier Analysis

Cost Regression: $C_i = a + bQ_i + \varepsilon_i + \phi_i$

The term $a + bQ + \varepsilon$ represents the minimum cost measured with a slight measurement error ε. Given this, the actual costs must be above the minimum so the inefficiency term ϕ must be positive. Suppose that ϕ has an exponential distribution:

$$f(\phi) = e^{-\phi/\lambda} / \lambda \text{for} \phi \geq 0.$$

[Note: $E[\phi] = \lambda$ and $Var[\phi] = \lambda^2$.] Suppose that the measurement error $\varepsilon \sim N(0, \sigma^2)$ and is independent of the inefficiency ϕ. The joint probability of ε and ϕ is

$f(\varepsilon,\varphi) = \dfrac{1}{\sqrt{2\pi\sigma\lambda}} e^{-\varphi/\lambda - \frac{1}{2}\varepsilon^2/\sigma^2}$. Let the total error be denoted $\theta = \varepsilon + \phi$. [Note: $E[\theta] = \lambda$ and $Var[\theta] = \sigma^2 + \lambda^2$.] Then the joint probability of the inefficiency and total error is $f(\theta,\varphi) = \dfrac{1}{\sqrt{2\pi\sigma\lambda}} e^{-\varphi/\lambda - \frac{1}{2}(\theta-\varphi)^2/\sigma^2}$. The marginal distribution of the total error is found by integrating the $f(\theta, \phi)$ with respect to ϕ over the range $[0, \infty)$. Using "complete-the-square" this can be seen to equal $f(\theta) = \frac{1}{\lambda}\Phi(\theta / \sigma - \sigma / \lambda)e^{-\theta/\lambda + \frac{1}{2}\sigma^2/\lambda^2}$, where Φ is the cumulative standard normal.

To fit the model to n data-points, we would select a, b , λ and σ to maximize log-likelihood:

$$\ln(L) = -n\ln(\lambda) + n(\sigma^2 / \lambda^2 2) + \sum_i \ln\Phi((C_i - a - bQ_i) / \sigma - \sigma / \lambda) - \sum_i (C_i - a - bQ_i) / \lambda.$$

Once we have estimated the parameters, we can measure the amount of inefficiency for each observation, ϕ_i. The conditional pdf $f(\phi_i \mid \theta_i)$ is computed for $\theta_i = C_i - a - bQ_i$:

$$f(\varphi_i \mid \theta_i) = \frac{1}{\sqrt{2\pi}\sigma\Phi(\theta_i / \sigma - \sigma / \lambda)} e^{-\frac{1}{2}\left(\frac{\varphi_i - (\theta_i - \sigma^2/\lambda)}{\sigma}\right)^2}$$

This is a half-normal distribution and has a mode of ?i-?2/?, assuming this is positive. The degree of cost inefficiency is defined as IEi = e^{φ_i} ; this is a number greater than 1, and the bigger it is the more inefficiently large is the cost. Of course, we do not know ?i, but if we evaluate IEi at the posterior mode $\theta_i - \sigma^2 / \lambda$ it equals IEi $\approx e^{C_i - \sigma^2/\lambda - a - bQ_i}$. Note that the term σ^2 / λ captures the idea that we do not precisely know what the minimum cost equals, so we slightly discount the measured cost to account for our uncertainty about the frontier.

Non-Spherical Errors

Assumption 3: $\text{var}(Y \mid X) = \text{var}(e \mid X) = \sigma^2 I$

Suppose that $\text{var}(e \mid X) = \sigma^2 W$, where W is a symmetric, positive definite matrix but $W \neq I$. What are the consequences for OLS?

a) $E[b] = E[(X'X)^{-1} X'(X\beta + \varepsilon)] = \beta + (X'X)^{-1} X'E[\varepsilon] = \beta$, so OLS is still unbiased even if $W \neq I$.

b) $\text{Var}[b] = E[(b - \beta)(b - \beta)'] = (X'X)^{-1} X'E[\varepsilon\varepsilon']X(X'X)^{-1}$

$$= \sigma^2 (X'X)^{-1} X'WX(X'X)^{-1} \neq \sigma^2 (X'X)^{-1}$$

Hence, the OLS computed standard errors and t-stats are wrong. The OLS estimator will not be BLUE.

Generalized Least-Squares

Suppose we find a matrix P (n x n) such that PWP'= I or equivalently $W = P^{-1}P'^{-1}$ or $W^{-1} = P'P$ (use spectral demcomposition). Multiply the regression model $(Y = X\beta + \varepsilon)$ on left by P: $PY = PX\beta + P\varepsilon$ Write PY = Y*, PX = X* and Pε = ε*, so in the transformed variables Y*=X*β+ε*. Why do this? Look at the variance of ε*: $\text{Var}(\varepsilon^*) = E[\varepsilon^* \varepsilon^{*\prime}] = E[P\varepsilon\varepsilon'P'] = PE[\varepsilon\varepsilon']P' = \sigma^2 PWP' = \sigma^2 I$. The error ε* is spherical; that's why.

GLS estimator: b* = (X*'X*)-1X*'Y*=(X'P'PX)-1X'P'PY=(X'W-1X)-1X'W-1Y.

Analysis of the transformed data equation says that GLS b* is BLUE. So it has lower variance that the OLS b.

$$\mathrm{Var}\left[b\,*\right] = \sigma^2\left(X^{*\prime}X\,*\right)^{-1} = \sigma^2\left(X'W^{-1}X\right)^{-1}$$

How do we estimate σ^2?

[Note: from OLS $E\left[e'e\right]/\left(n-k\right) = E[e'Me\left]/\right(n-k) = E\left[\,\mathrm{tr}(e'Me)\right]/$

$/\left(n-k\right) = E[\mathrm{tr}(Mee')\left]/\right(n-k) = \mathrm{tr}(ME\left[\,ee'\right])/\left(n-k\right) = s^2\mathrm{tr}\left(MW\right)/\left(n-k\right)$. Since $W \neq I$, $\mathrm{tr}\left(MW\right) \neq n-k$, so $E\left[e'e\right]/\left(n-k\right)\ \neq \sigma^2$.] Hence, to estimate σ^2 we need to use the errors from the transformed equation Y* = X*b* + e*.

$$s\,*^2 = \left(e^{*\prime}e\,*\right)/\left(n-k\right)$$

$$E\left[s\,*^2\right] = \mathrm{tr}(M * E[e * e^{*\prime}])/\left(n-k\right) = \sigma^2\mathrm{tr}\left(M * PWP'\right)/\left(n-k\right) = \sigma^2\mathrm{tr}\left(M\,*\right)/\left(n-k\right) = \sigma^2.$$

Hence, s*2 is an unbiased estimator of σ^2.

Important Note: All of the above assumes that W is known and that it can be factored into $P^{-1}P'^{-1}$. How do we know W? Two special cases are autocorrelation and heteroskedasticity.

Autocorrelated Errors

Suppose that $Y_t = X_t b + u_t$ (notice the subscript t denotes time since this problem occurs most frequently with time-series data). Instead of assuming that the errors ut are iid, let us assume they are autocorrelated (also called serially correlated errors) according to the lagged formula

$$u_t = ru_{t-1} + e_t,$$

where $\varepsilon_t$is iid . Successively lagging and substituting for ut gives the equivalent formula

$$u_t = e_t + re_{t-1} + r^2e_{t-2} + \ldots$$

Using this, we can see that

$E\left[u_t u_t\right] = s^2(1 + r^2 + r^4 + \ldots) = s^2/(1-r^2)$, $E\left[u_t u_{t-1}\right] = r\,s^2/(1-r^2)$,

$E\left[u_t u_{t-2}\right] = r^2 s^2/(1-r^2)$, … $E\left[u_t u_{t-m}\right] = r^m s^2/(1-r^2)$. Therefore, the variance matrix of u is

$$\mathrm{var}(u) = E[uu'] = \sigma^2\frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix} = \sigma^2 W,$$

where
$$W = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$$

Notes

and
$$W^{-1} = \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}$$

It is possible to show that W-1 can be factored into P'P where

$$P = \begin{bmatrix} \sqrt{1 - \rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}$$

Given this P, the transformed data for GLS is

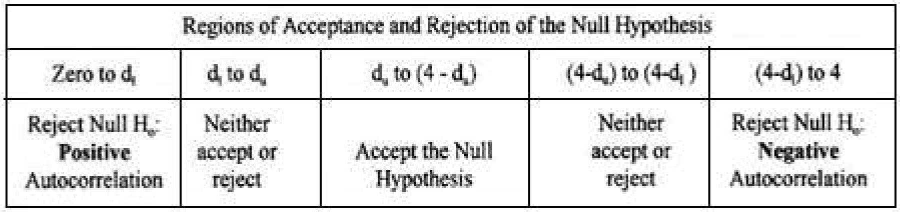
$$Y^* = PY = \begin{bmatrix} \sqrt{1 - \rho^2}\, y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_n - \rho y_{n-1} \end{bmatrix}, \; X^* = \begin{bmatrix} \sqrt{1 - \rho^2} & \sqrt{1 - \rho^2}x_{11} & \sqrt{1 - \rho^2}x_{1p} \\ 1 - \rho & x_{21} - \rho x_{11} & xp - \rho x_{1p} \\ & & \\ 1 - \rho & x_{n1} - \rho x_{n-1,1} & x_{n1p} - \rho x_{n-1,p} \end{bmatrix}$$

Notice that only the first element is unique. The rest just involves subtracting a fraction ρ of the lagged value from the current value. Many modelers drop the first observation and use only the last n-1 because it is easier, but this throws away information and I would not recommend doing it unless you had a very large n. The Cochrane-Orcutt technique successively estimates of ρ from the errors and re-estimating based upon new transformed data (Y*, X*).

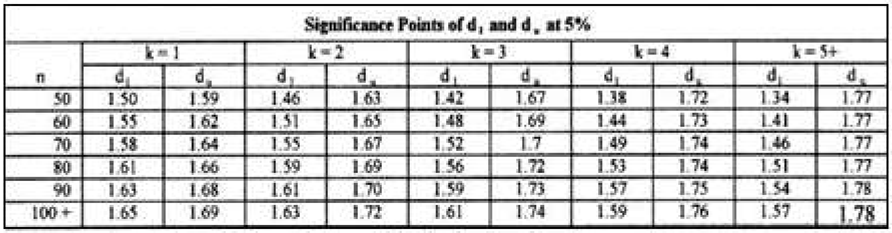
1. Guess a starting ρ 0.
2. At stage m, estimate β in model $Y_t - \rho_m Y_{t-1} = (X_t - \rho_m X_{t-1})\beta + \varepsilon_t$ using OLS. If the estimate bm is not different from the previous bm-1, then stop. Otherwise, compute error vector $e_m = \left(Y^* - X^* b_m\right)$.
3. Estimate ρ in $e_{mt} = \rho e_{m,t-1} + \varepsilon_t$ via OLS. This estimate becomes the new ρ_{m+1}. Go back to 2.

Durbin-Watson test for $\rho \neq 0$ in $u_t = \rho u_{t-1} + \varepsilon_t$

1. Compute OLS errors e.
2. Calculate $d = \dfrac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$.
3. $d<2 \Rightarrow \rho>0, d>2 \Rightarrow \rho<0, d=2 \Rightarrow \rho=0.$



Regions of Acceptance and Rejection of the Null Hypothesis				
Zero to d_l	d_l to d_u	d_u to (4 - d_u)	(4-d_u) to (4-d_l)	(4-d_l) to 4
Reject Null H_o: **Positive** Autocorrelation	Neither accept or reject	Accept the Null Hypothesis	Neither accept or reject	Reject Null H_o: **Negative** Autocorrelation



Significance Points of d_l and d_u at 5%										
	k = 1		k = 2		k = 3		k = 4		k = 5+	
n	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
70	1.58	1.64	1.55	1.67	1.52	1.7	1.49	1.74	1.46	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
100 +	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

where: k = the number of independent variables in the equation.

Heteroskedasticity

Here we assume that the errors are independent, but not necessarily identically distributed. That is the matrix W is diagonal, but not the identity matrix. The most common way for this to occur is because Yi is the average response of a group i that has a number of members mi. Larger groups have smaller variance in the average response: var $(\varepsilon_i) = \sigma$ mi. Hence the variance matrix would be:

$$\text{Var}(\varepsilon) = \sigma^2 \begin{bmatrix} \frac{1}{m_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{m_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{m_n} \end{bmatrix}$$

An related example of this would be that Y is the sum across the members of many similar elements, so that the var $(\varepsilon_i) = \sigma$ mi and

$$\text{Var}(\varepsilon) = \sigma^2 \begin{bmatrix} m_1 & 0 & \cdots & 0 \\ 0 & m_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_n \end{bmatrix}$$

If we knew how big the groups where and whether we had the average or total response, we could substitute for mi in the above matrix W.

More generally, we think that the variance of ε I depends upon some variable Z. We can do a Glessjer Test of this as follows.

1. Compute OLS estimate of b,e
2. Regress $|e_i|$ on Z_i^{η}, where $h = 1, -1$ and ½.
3. If the co-efficient of Z^{η} is 0 then the model is homoscedastic, but if it is not zero, then the model has heteroskedastic errors.

In SPSS, you can correct for heteroskedasticity by using Analyze/Regression/Weight Estimation rather than Analyze/Regression/Linear. You have to know the variable Z, of course.

Notes

Trick: Suppose that $\sigma_t^2 = \sigma^2 Z_t^2$. Notice Z is squared. Divide both sides of equation by Z to get $Y_t / Z_t = (X_t / Z_t) b + e_t / Z_t$. This new equation has homoscedastic errors and so the OLS estimate of this transformed model is BLUE.

Simultaneous Equations

Assumption 4: X is fixed

Later in the semester will return to the problem that X is often determined by actors in the play we are studying rather than by us scientists. This is a serious problem in simultaneous equation models.

Multicollinearity

Assumption 5: X has full column rank.

What is the problem if you have multicollinearity? In X'X there will be some portions that look like a little square $\begin{bmatrix} x'x & x'x \\ x'x & x'x \end{bmatrix}$ and this has a determinant equal to zero, so its reciprocal will be near infinity. OLS is still BLUE, but estimated $\text{var}[b] = (X'X)^{-1} Y'(I - X(X'X)^{-1} X')Y / (n-k)$ can be very large.

If there is collinearity, then there exists a weighting vector α such that Xα is close to the 0 vector. Of course, we cannot just allow α to be zero. Hence let's look for the value of α that minimizes ||Xα||2 subject to α'α = 1. The Lagrangian for this constrained optimization is L = α'X'Xα + γ(1-α'α) and the first order conditions are X'Xα - γ α = 0 This is the equation for the eigenvalue and eigenvector of X'X. Multiply the first order condition by α' and use the fact that eigenvectors have a length of 1 to see that α'X'Xα = γ, so we are looking at the smallest of the eigenvalues when we seek collinearity. When is this eigenvalue "small" enough to measure serious collinearity? We compute a Condition Index as the square root of the ratio largest eigenvalue to the smallest eigenvalue: $CI \equiv \sqrt{\frac{\lambda_{\text{largest}}}{\lambda_{\text{smallest}}}}$. When the condition index is greater than 20 or 30, we have serious collinearity. In SPSS Regression/Linear/Statistics click "Collinearity Diagnostics."

Warning: Many people use the Variance Inflation Factor to identify collinearity. This should be avoided (see Chennamaneni, Echambadi, Hess and Syam 2009). The problem is that VIF confuses "collinearity" with "correlation" as follows. Let R be the correlation matrix of X:

R = D-½X'HXD - ½/(n-1) where the standard deviation matrix D½ = sqrt(diag(X'HX)/(n-1)). Compute R-1. For example,

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{1-\rho^2} & \frac{\rho}{1-\rho^2} \\ \frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{bmatrix}$$

and along the diagonal is 1/(1-ρ^2) which is called the Variance Inflation Factor (VIF). More generally $\text{VIF}_i = (1 - R_i^2)^{-1}$ where R_i^2 is the R-square from regressing xi on the k-1 other variables in X. The problem with VIF is that it starts with a mean-centered data HX, when collinearity is a problem of the raw data X. In OLS we compute $\left(X'X\right)^{-1}$, not $\left(X'HX\right)^{-1}$. Chennamani et al. provide a variant of VIF that does not suffer from these problems.

4.4 ERRORS IN VARIABLES CONSEQUENCE

In statistics, errors-in-variables models or measurement error models are regression models that account for measurement errors in the independent variables. In contrast, standard regression models assume that those regressors have been measured exactly, or observed without error; as such, those models account only for errors in the dependent variables, or responses.

In the case when some regressors have been measured with errors, estimation based on the standard assumption leads to inconsistent estimates, meaning that the parameter estimates do not tend to the true values even in very large samples. For simple linear regression the effect is an underestimate of the coefficient, known as the attenuation bias. In non-linear models the direction of the bias is likely to be more complicated.

In traditional asymmetric regression, the value of Y is assumed to depend on the value of X and the scientist is interested in the value of Y for a given value of X. Ordinary least squares regression assumes that X (the independent variable) is measured without error, and that all error is on the Y variable. There are two sources of errors - measurement error (d) and intrinsic or equation error (e). These error terms are usually assumed to be random with a mean of zero (in other words no bias). By definition all equation error in asymmetric regression is assumed to be on the Y-variable since one is interested in the value of Y for a given value of X. But there may be substantial measurement error on the X variable. This does not matter if values of X are fixed by the experimenter, as is commonly the case in an experiment - in this situation the estimate of the slope is still unbiased. But if values of X are random and X is measured with error, then the estimate of the slope of the regression relationship is attenuated or closer to zero than it should be. One type of errors-in variables regression (the method of moments) enables one to correct the slope of an asymmetric regression for measurement error of the X-variable. The other type of regression is symmetrical or orthogonal regression. Here there is no question of a dependent or independent variable (hence sometimes Y1 and Y2 are used to denote the variables, rather than X and Y). We simply want to model the relationship between two (random) variables, each of which may be subject to both measurement and equation error.

Errors-in-variables regression is used much less than ordinary least squares regression apart from in certain specialized areas such as comparison of methods studies and allometry/ isometry assessments. Its lack of use may seem surprising given how often the X-variable is not measured without error. Nevertheless, in situations when one is using regression purely for descriptive purposes or for prediction, ordinary least squares regression is still the best option, albeit one accepts that the slope is probably attenuated. We give two examples of where ordinary least squares (OLS) linear regression could have been used rather than the more complex errors-in-variable regression.

The wrong type of errors-in-variables regression is often used when dealing with an asymmetric relationship - in other words where there is a clear independent and dependent variable. In this situation, orthogonal regression (including major axis and reduced major axis) is inappropriate. Instead, if there is substantial measurement error on the X-axis, the slope of the OLS regression should be corrected for attenuation using the method of moments.

Notes

It is hard to find examples of the use of this method in the literature, but we do give several examples (such as relating survival time to HIV load and relating phytophage species richness to tree abundance) where it should have been used rather than orthogonal regression.

We give four examples of where orthogonal regression (and its variants) are used in comparison of methods studies - for example a comparison of several techniques for estimation of cardiac output, and a comparison of two methods for making egg counts of nematodes. However, all these examples of symmetrical regression make simplifying assumptions about the errors by using major axis regression or reduced major axis regression. Providing measurement error really was the only factor causing points to deviate from a 1:1 relationship, a better approach would have been to assess measurement error for each method by repeatedly measuring the same sampling unit. This is especially the case when one method is known to produce more variable results than the other. There is a strong case for using the Bland-Altman method instead or in addition to errors-in-variables regression in such studies, and we give one example (comparing the results of a portable clinical analyzer with those obtained using a traditional analyzer) where this is done.

When testing for allometry (our example looks at the shape of a species of spider) both equation error and measurement error are present on both axes - and most authors again 'opt out' by using reduced major axis regression rather than attempting to estimate the different error terms. It would be better to guesstimate upper and lower limits of likely error on each variable, and then estimate the range of slopes that might be possible. We give an example of the use of errors-in-variables regression to obtain mass/length residuals, which are then used as measures of body condition. This is a controversial issue - should one for example use analysis of covariance instead - and we consider both sides of the argument. Lastly we look at a study which uses errors-in-variables regression to test Taylor's power law of the relationship between log variance and log mean.

Many of the issues considered previously for ordinary least squares re-emerge in this example. Bivariate observations must be independent and not (for example) obtained on the same units in a time series as is done in a trapping study. Relationships must be linear - a questionable assumption in some cases. Quite often the spread of values of the Y-variables appears to increase with the X values indicating heteroscedasticity. We also note that if variables are log transformed, the estimation of Y is biased after detransformation, and must be corrected appropriately.

4.5 METHODS OF ESTIMATION-CLASSICAL METHOD OF MAXIMUM LIKELIHOOD

The term parameter estimation refers to the process of using sample data (in reliability engineering, usually times-to-failure or success data) to estimate the parameters of the selected distribution. Several parameter estimation methods are available. This section presents an overview of the available methods used in life data analysis. More specifically, we start with the relatively simple method of Probability Plotting and continue with the more sophisticated methods of Rank Regression (or Least Squares), Maximum Likelihood Estimation and Bayesian Estimation Methods.

Maximum likelihood estimation (ML estimation) is another estimation method. In the case of the linear model with errors distributed as N(0, σ^2), the ML and least-squares estimators are the same. However, in the general case, this is not true. Often the variance of the ML estimator (MLE) is less than the variance of other estimators (including the least-squares estimator) and hence is the preferable estimation method. MLEs also have many other nice properties, such as the fact that ß^ gets close to ß (the true parameter) with high probability as the sample size gets large. For these reasons, we will focus on ML estimation in this course. The likelihood function is algebraically the same as the probability distribution of the observed data. The joint distribution of Y1, . . . , Yn, fY1,...,Yn (y1, . . . , yn; θ), is seen as a function of y1, . . . , yn with fixed parameters θ . In contrast, the likelihood, L(θ ; y1, . . . , yn) is seen as a function of θ for a given set of data points. The idea behind ML estimation is that we try to find the values of the parameters θ that seem most likely, given our observed data. To do this, we locate θ ^, the value which maximizes L(θ ; y). The value θ ^ also maximizes the function log L(θ ; y), since log is a monotonically increasing function. Often, it is easier to maximize the log-likelihood than the likelihood itself.

4.6 USE OF INSTRUMENTAL VARIABLE

In statistics, econometrics, epidemiology and related disciplines, the method of instrumental variables (IV) is used to estimate causal relationships when controlled experiments are not feasible or when a treatment is not successfully delivered to every unit in a randomized experiment. Intuitively, IV is used when an explanatory variable of interest is correlated with the error term, in which case ordinary least squares and ANOVA gives biased results. A valid instrument induces changes in the explanatory variable but has no independent effect on the dependent variable, allowing a researcher to uncover the causal effect of the explanatory variable on the dependent variable.

Instrumental variable methods allow for consistent estimation when the explanatory variables (covariates) are correlated with the error terms in a regression model. Such correlation may occur 1) when changes in the dependent variable change the value of at least one of the covariates ("reverse" causation), 2) when there are omitted variables that affect both the dependent and independent variables or 3) when the covariates are subject to non-random measurement error. Explanatory variables which suffer from one or more of these issues in the context of a regression are sometimes referred to as endogenous. In this situation, ordinary least squares produces biased and inconsistent estimates. However, if an instrument is available, consistent estimates may still be obtained. An instrument is a variable that does not itself belong in the explanatory equation but is correlated with the endogenous explanatory variables, conditional on the value of other covariates.

In linear models, there are two main requirements for using IV:

The instrument must be correlated with the endogenous explanatory variables, conditional on the other covariates. If this correlation is strong, then the instrument is said to have a strong first stage. A weak correlation may provide misleading inferences about parameter estimates and standard errors.

The instrument cannot be correlated with the error term in the explanatory equation, conditional on the other covariates. In other words, the instrument cannot suffer from the same problem as the original predicting variable. If this condition is met, then the instrument is said to satisfy the exclusion restriction.

Informally, in attempting to estimate the causal effect of some variable X on another Y, an instrument is a third variable Z which affects Y only through its effect on X. For example, suppose a researcher wishes to estimate the causal effect of smoking on general health. Correlation between health and smoking does not imply that smoking causes poor health because other variables, such as depression, may affect both health and smoking, or because health may affect smoking. It is at best difficult and expensive to conduct controlled experiments on smoking status in the general population. The researcher may attempt to estimate the causal effect of smoking on health from observational data by using the tax rate for tobacco products as an instrument for smoking. The tax rate for tobacco products is a reasonable choice for an instrument because the researcher assumes that it can only be correlated with health through its effect on smoking. If the researcher then finds tobacco taxes and state of health to be correlated, this may be viewed as evidence that smoking causes changes in health.

An instrumental variable (sometimes called an "instrument" variable) is a third variable, Z, used in regression analysis when you have endogenous variables variables that are influenced by other variables in the model. In other words, you use it to account for unexpected behavior between variables. Using an instrumental variable to identify the hidden (unobserved) correlation allows you to see the true correlation between the explanatory variable and response variable, Y.

Z is correlated with the explanatory variable (X) and uncorrelated with the error term, ε, (What is ε?) in the equation:

$$Y = X\beta + \varepsilon .$$

Instrumental variables are widely used in econometrics, a branch of economics that uses statistics to describe economic systems, and is sometimes seen in other fields like health sciences and epidemiology.

Example of an Instrumental Variable

Let's say you had two correlated variables that you wanted to regress: X and Y. Their correlation might be described by a third variable Z, which is associated with X in some way. Z is also associated with Y but only through Y's direct association with X. For example, let's say you wanted to investigate the link between depression (X) and smoking (Y). Lack of job opportunities (Z) could lead to depression, but it is only associated with smoking through it's association with depression (i.e. there isn't a direct correlation between lack of job opportunities and smoking). This third variable, Z (lack of job opportunities), can generally be used as an instrumental variable if it can be measured and it's behavior can be accounted for.

4.7 INSTRUMENTAL VARIABLES REGRESSION

Instrumental Variables regression (IV) basically splits your explanatory variable into two parts: one part that could be correlated with e and one part that probably isn't. By isolating the part with no correlation, it's possible to estimate ß in the regression equation:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

This type of regression can control for threats to internal validity, like:

- Confounding variables,
- Measurement error,
- Omitted variable bias (sometimes called spuriousness),
- simultaneity,
- Reverse Causality.

In essence, IV is used when your variables are related in some way; If you have some type of correlation going on between variables (e.g. bidirectional correlation), then you can't use the more common methods like ordinary least squares, because one requirement of those methods is that variables are not correlated.

4.8 FINDING INSTRUMENTAL VARIABLES

IV regression isn't an easy fix for confounding or other issues; In real life, instrumental variables can be difficult to find and in fact, may not exist at all. You cannot use the actual data to find IVs (e.g. you can't perform a regression to identify any) - you must rely on your knowledge about the model's structure and the theory behind your experiment (e.g. economic theory). When looking for IVs, keep in mind that Z should be:

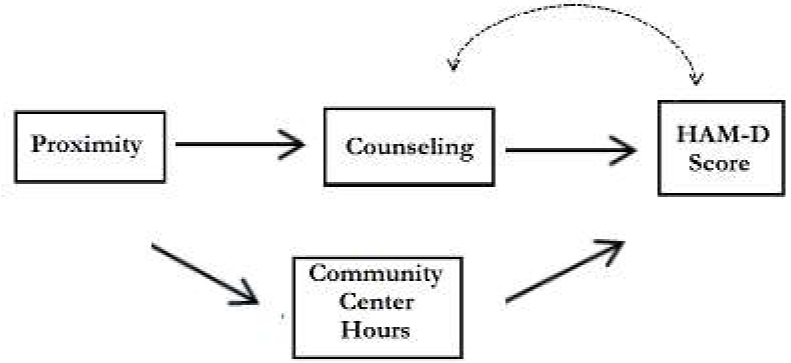
- **Exogenous -** not affected by other variables in the system (i.e. Cov(z, ε) = 0). This can't be directly tested; you have to use your knowledge of the system to determine if your system has exogenous variables or not.
- **Correlated with X,** an endogenous explanatory variable (i.e. Cov(Z,X) $\neq$ 0). A very significant correlation is called a strong first stage. Weak correlations can lead to misleading estimates for parameters and standard errors.

A couple of ideas for finding IVs: if available you could use two different data sources for your instrumental variables or you could collect longitudinal data and use that. If you know that a mediating variable is causing the effect of X and Y, you can use it as an instrumental variable.

Causal Graphs

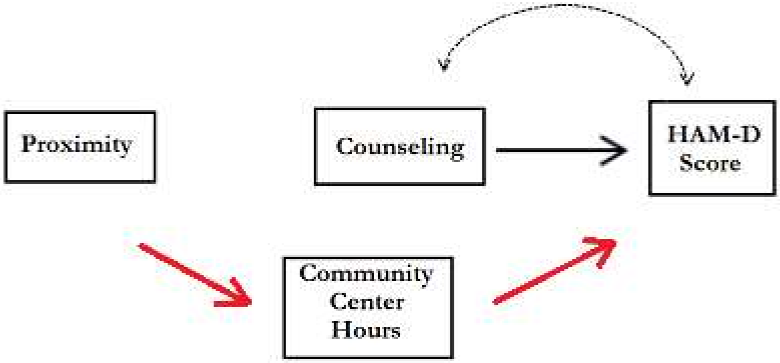
Causal graphs can be used to outline your model structure and identify possible IVs.

Suppose that you want to estimate the effect of a counseling program on senior depression (measured by a rating scale like the HAM-D). The relationship between attending counseling and score on the HAM-D may be confounded by various factors. For example, people who attend counseling sessions might care more about improving their health, or they may have a support network encouraging them to go to counseling. The proximity of a patient's home to the counseling program is a potential instrumental variable.



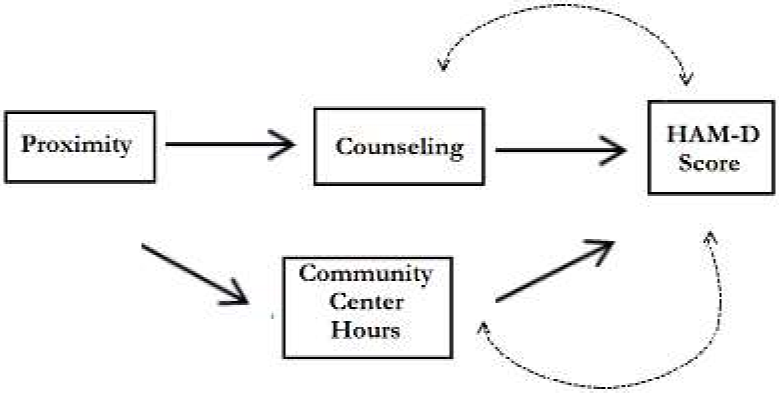
Proximity is a potential IV in this model.

However, what if the counseling center is located within a senior community center? Proximity may then cause seniors to spend time socializing or taking up a hobby, which could improve their HAM-D scores. The causal graph in Figure shows that Proximity cannot be used as an IV because it is connected to depression scoring through the path Proximity → Community Center Hours → HAM-D Score.



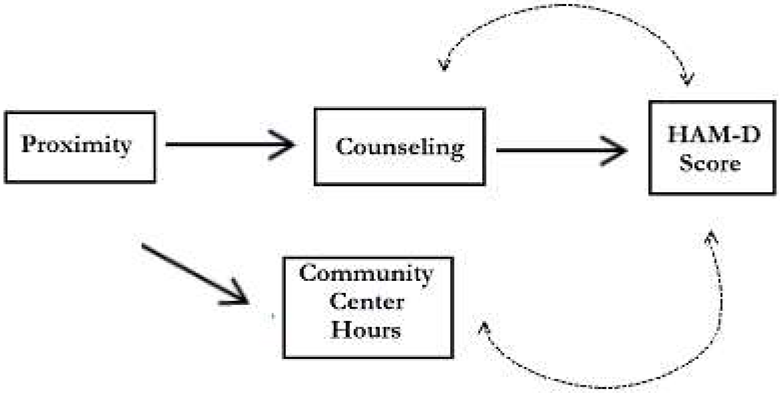
However, you can control for Community Center Hours by adding it as a covariate ; If you do that, then Proximity can be used as an IV, since Proximity is separated from HAM-D score, given community center hours.

Next, suppose that extroverts are more likely to spend time in the community center and are generally happier than introverts. This is shown in the following graph:



Community center hours is a collider variable; conditioning on it opens up a part-bidirectional path Proximity → Community Center Hours → HAM-D. This means that Proximity can't be used as an IV.

As a final step for this example, let's say you find that community center hours doesn't affect HAM-D Scores because people who don't socialize in the community center actually socialize in other places. This is depicted on the following graph:



If you don’t control for community center hours and remove it as a covariate, then you can use Proximity again as an IV.

4.9 AUTOCORRELATION

Autocorrelation is the correlation of a signal with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them. The analysis of autocorrelation is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies. It is often used in signal processing for analyzing functions or series of values, such as time domain signals.

Sources of Autocorrelation

Autocorrelation is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals. It is the same as calculating the correlation between two different time series, except that the same time series is actually used twice: once in its original form and once lagged one or more time periods.

Autocorrelation can also be referred to as lagged correlation or serial correlation, as it measures the relationship between a variable's current value and its past values. When computing autocorrelation, the resulting output can range from 1.0 to negative 1.0, in line with the traditional correlation statistic. An autocorrelation of +1.0 represents a perfect positive correlation (an increase seen in one time series leads to a proportionate increase in the other time series). An autocorrelation of negative 1.0, on the other hand, represents perfect negative correlation (an increase seen in one time series results in a proportionate decrease in the other time series). Autocorrelation measures linear relationships; even if the autocorrelation is minuscule, there may still be a nonlinear relationship between a time series and a lagged version of itself.

4.10 AUTOCORRELATION IN TECHNICAL ANALYSIS

Autocorrelation can be useful for technical analysis, which is most concerned with the trends of, and relationships between, security prices using charting techniques in lieu of a company's financial health or management. Technical analysts can use autocorrelation to see how much of an impact past prices for a security has on its future price.

Autocorrelation can show if there is a momentum factor associated with a stock. For example, if you know that a stock historically has a high positive autocorrelation value and you witnessed the stock making solid gains over the past several days, then you might reasonably expect the movements over the upcoming several days (the leading time series) to match those of the lagging time series and to move upward.

Example of Autocorrelation

Assume an investor is looking to discern if a stock's returns in her portfolio exhibit autocorrelation; the stock's returns are related to its returns in previous trading sessions. If the returns do exhibit autocorrelation, the stock could be characterized as a momentum stock; its past returns seem to influence its future returns. The investor runs a regression with two prior trading sessions' returns as the independent variables and the current return as the dependent variable. She finds that returns one day prior have a positive autocorrelation of 0.7, while the returns two days prior have a positive autocorrelation of 0.3. Past returns seem to influence future returns, and she can adjust her portfolio to take advantage of the autocorrelation and resulting momentum.

4.11 CONSEQUENCES OF AUTOCORRELATION

Autocorrelation is a characteristic of data in which the correlation between the values of the same variables is based on related objects. It violates the assumption of instance independence, which underlies most of the conventional models. It generally exists in those types of data-sets in which the data, instead of being randomly selected, is from the same source.

(i) Presence

The presence of autocorrelation is generally unexpected by the researcher. It occurs mostly due to dependencies within the data. Its presence is a strong motivation for those researchers who are interested in relational learning and inference.

Examples: In order to understand autocorrelation, we can discuss some instances that are based upon cross sectional and time series data. In cross sectional data, if the change in the income of a person A affects the savings of person B (a person other than person A), then autocorrelation is present. In the case of time series data, if the observations show inter-correlation, specifically in those cases where the time intervals are small, then these inter-correlations are given the term of autocorrelation.

In time series data, autocorrelation is defined as the delayed correlation of a given series. Autocorrelation is a delayed correlation by itself, and is delayed by some specific number of time units. On the other hand, serial autocorrelation is that type which defines the lag correlation between the two series in time series data.

(ii) Patterns

Autocorrelation depicts various types of curves which show certain kinds of patterns, for example, a curve that shows a discernible pattern among the residual errors, a curve that shows a cyclical pattern of upward or downward movement, and so on.

In time series, it generally occurs due to sluggishness or inertia within the data. If a non-expert researcher is working on time series data, then he might use an incorrect functional form, and this again can cause autocorrelation.

The handling of the data by the researcher, when it involves extrapolation and interpolation, can also give rise to autocorrelation. Thus, one should make the data stationary in order to remove autocorrelation in the handling of time series data.

Autocorrelation is a matter of degree, so it can be positive as well as negative. If the series (like an economic series) depicts an upward or downward pattern, then the series is considered to exhibit positive autocorrelation. If, on the other hand, the series depicts a constant upward and downward pattern, then the series is considered to exhibit negative autocorrelation.

When a researcher has applied ordinary least square over an estimator in the presence of autocorrelation, then the estimator is incompetent.

(iii) Detecting the Presence

There is a very popular test called the Durbin Watson test that detects the presence of autocorrelation. If the researcher detects autocorrelation in the data, then the first thing the researcher should do is to try to find whether or not it is pure. If it is pure, then one can transform it into the original model that is free from pure autocorrelation

GLSM

GLSM stands for Geometric Least Squares Mean. This is a mean estimated from a linear model. In contrast, a raw or arithmetic mean is a simple average of your values, using no model. Least squares means are adjusted for other terms in the model (like covariates), and are less sensitive to missing data. Theoretically, they are better estimates of the true population mean.

As a simple example, suppose you have a treatment applied to 3 trees (experimental unit), and 2 observations (samples) are collected on each. However, one observation is missing, giving values of (45, 36), (56) and (37, 41), where parentheses are around each tree. The raw average is simply (45 + 36 + 56 + 37 + 41)/5 = 43, and note the reduced influence of the second tree since it has fewer values. The least squares mean would be based on a model u + T + S(T), resulting in an average of the tree averages, as follows.

Least squares mean = [(45+36)/2 + 56 + (37 + 41)/2] / 3 = 45.17. This more accurately reflects the average of the 3 trees, and is less affected by the missing value.

4.12 TESTS FOR AUTOCORRELATION

Here we present some formal tests and remedial measures for dealing with error autocorrelation.

Durbin-Watson Test

We usually assume that the error terms are independent unless there is a specific reason to think that this is not the case. Usually violation of this assumption occurs because there is a known temporal component for how the observations were drawn. The easiest way to assess if there is dependency is by producing a scatterplot of the residuals versus the time measurement for that observation (assuming you have the data arranged according to a time sequence order). If the data are independent, then the residuals should look randomly scattered about 0. However, if a noticeable pattern emerges (particularly one that is cyclical) then dependency is likely an issue.

Recall that if we have a first-order autocorrelation with the errors, then the errors are modeled as:

$$\epsilon t = \rho \epsilon t - 1 + \omega t, \epsilon t = \rho \epsilon t - 1 + \omega t,$$

where $|\rho|<1|\rho|<1$ and the $\omega t \sim \text{iidN}(0, \sigma 2) \omega t \sim \text{iidN}\left(0, \sigma 2\right)$. If we suspect first-order autocorrelation with the errors, then a formal test does exist regarding the parameter $\rho\rho$. In particular, the Durbin-Watson test is constructed as:

$$\text{H}0 : \rho = 0 \text{HA} : \rho \neq 0 . \text{H}0 : \rho = 0 \text{HA} : \rho \neq 0 .$$

So the null hypothesis of $\rho = 0 \rho = 0$ means that $\epsilon t = \omega t \epsilon t = \omega t$ or that the error term in one period is not correlated with the error term in the previous period, while the alternative hypothesis of means the error term in one period is either positively or negatively correlated with the error term in the previous period. Often times, a researcher will already have an indication of whether the errors are positively or negatively correlated. For example, a regression of oil prices (in dollars per barrel) versus the gas price index will surely have positively correlated errors. When the researcher has an indication of the direction of the correlation, then the Durbin-Watson test also accommodates the one-sided alternatives $\mathrm{HA}: \rho < 0\mathrm{HA}: \rho < 0$ for negative correlations or $\mathrm{HA}: \rho > 0\mathrm{HA}: \rho > 0$ for positive correlations (as in the oil example).

The test statistic for the Durbin-Watson test on a data set of size n is given by:

$$\mathrm{D} = \sum \mathrm{nt} = 2(\mathrm{et} - \mathrm{et} - 1)2 \sum \mathrm{nt} = 1\mathrm{e}2\mathrm{t}, \mathrm{D} = \sum \mathrm{t} = 2\mathrm{n}\left(\mathrm{et} - \mathrm{et} - 1\right)2 \sum \mathrm{t} = 1\mathrm{net}2,$$

The DW test statistic varies from 0 to 4, with values between 0 and 2 indicating positive autocorrelation, 2 indicating zero autocorrelation, and values between 2 and 4 indicating negative autocorrelation. Exact critical values are difficult to obtain, but tables (for certain significance values) can be used to make a decision (e.g., see the tables at this link, where T represents the sample size, n, and K represents the number of regression parameters, p). The tables provide a lower and upper bound, called dLdL and dUdU, respectively. In testing for positive autocorrelation, if D<dLD<dL then reject H0H0, if D>dUD>dU then fail to reject H0H0 or if dL = D = dUdL = D = dU, then the test is inconclusive. While the prospect of having an inconclusive test result is less than desirable, there are some programs which use exact and approximate procedures for calculating a *p*-value. These procedures require certain assumptions on the data which we will not discuss. One "exact" method is based on the beta distribution for obtaining p-values.

To illustrate, consider the Blaisdell Company example from page 489 of Applied Linear Regression Models (4th ed) by Kutner, Nachtsheim, and Neter. If we fit a simple linear regression model with response comsales (company sales in $ millions) and predictor indsales(industry sales in $ millions) and click the "Results" button in the Regression Dialog and check "Durbin-Watson statistic" we obtain the following output:

```
Term           Coef        SE Coef      T-Value     P-Value          VIF
Constant     -1.455          0.214        -6.79       0.000
indsales    0.17628        0.00144       122.02       0.000         1.00
Durbin-Watson Statistic =  0.734726
```

Since the value of the Durbin-Watson Statistic falls below the lower bound at a 0.01 significance level (obtained from a table of Durbin-Watson test bounds), there is strong evidence the error terms are positively correlated.

Ljung-Box Q Test

The Ljung-Box Q test (sometimes called the Portmanteau test) is used to test whether or not observations over time are random and independent. In particular, for a given k, it tests the following:

H0 : the autocorrelations up to lag k are all 0HA:the autocorrelations of one or more lags differ from 0.H0:the autocorrelations up to lag k are all 0HA:the autocorrelations of one or more lags differ from 0.

The test statistic is calculated as:

$$\mathrm{Qk} = \mathrm{n}(\mathrm{n} + 2)\mathrm{k} \sum \mathrm{j} = 1\mathrm{r}2\mathrm{jn} - \mathrm{j}, \mathrm{Qk} = \mathrm{n}\left(\mathrm{n} + 2\right) \sum \mathrm{j} = 1\mathrm{krj}2\mathrm{n} - \mathrm{j},$$

which is approximately $\chi 2\mathrm{k} \chi \mathrm{k}2$ – distributed.

To illustrate how the test works for $k = 1$, consider the Blaisdell Company example from above. If we store the residuals from a simple linear regression model with response comsales and predictor indsales and then find the autocorrelation function for the residuals (select Stat > Time Series > Autocorrelation), we obtain the following output:

```
Autocorrelation Function: RESI1

Lag        ACF          T            LBQ
 1       0.626005      2.80          9.08
```

The Ljung-Box Q test statistic of 9.08 corresponds to a $\chi^{21}\chi^{12}$ p-value of 0.0026, so there is strong evidence the lag-1 autocorrelation is non-zero.

Cochrane-Orcutt Procedure

The first of the three transformation methods we discuss is called the Cochrane-Orcutt procedure, which involves an iterative process (after identifying the need for an AR(1) process):

Estimate $\rho\rho$ for $\in t = \rho \in t\text{-}1 + \omega t \in t = \rho \in t - 1 + \omega t$ by performing a regression through the origin. Call this estimate *r*.

Transform the variables from the multiple regression modelyt=ß0+ß1xt,1+…+ßp-1xt,p-1+?tyt=ß0+ß1xt,1+…+ßp-1xt,p-1+?tby setting y*t=yt-ryt-1yt*=yt-ryt-1 and x*t,j=xt,j-rxt-1,jxt,j*=xt,j-rxt-1,j for j=1,…,p-1j=1,…,p-1.

Regress y*tyt* on the transformed predictors using ordinary least squares to obtain estimates ^ß*0,…,^ß*p-1ß^0*,…,ß^p-1*. Look at the error terms for this fit and determine if autocorrelation is still present (such as using the Durbin-Watson test). If autocorrelation is still present, then iterate this procedure. If it appears to be corrected, then transform the estimates back to their original scale by setting ^ß0=^ß*0/(1-r)ß^0=ß^0*/(1-r) and ^ßj=^ß*jß^j=ß^j* for j=1,…,p-1j=1,…,p-1. Notice that only the intercept parameter requires a transformation. Furthermore, the standard errors of the regression estimates for the original scale can also be obtained by setting s.e.(^ß0)=s.e.(^ß*0)/(1-r)s.e.(ß^0)=s.e.(ß^0*)/(1-r) and s.e.(^ßj)=s.e.(^ß*j)s.e.(ß^j)=s.e.(ß^j*) for j=1,…,p-1j=1,…,p-1.

To illustrate the Cochrane-Orcutt prcedure, consider the Blaisdell Company example from above:

- Store the residuals, RESI1, from a simple linear regression model with response comsales and predictor indsales.
- Use Minitab's Calculator to define a lagged residual variable, lagRESI1 = LAG(RESI1,1).
- Fit a simple linear regression model with response RESI1 and predictor lagRESI1 and no intercept. Use the Storage button to store the Coefficients. We find the estimated slope from this regression to be 0.631164, which is the estimate of the autocorrelation parameter $\rho\rho$.
- Use Minitab's Calculator to define a transformed response variable, Y_co = comsales-0.631164*LAG(comsales,1).
- Use Minitab's Calculator to define a transformed predictor variable, X_co = indsales-0.631164*LAG(indsales,1).
- Fit a simple linear regression model with response Y_co and predictor X_co to obtain the following output:

```
Term        Coef       SE Coef     T-Value   P-Value   VIF
Constant    -0.394     0.167        -2.36     0.031
X_co        0.17376    0.00296      58.77     0.000    1.00
Durbin-Watson Statistic =  1.65025
```

- Since the value of the Durbin-Watson Statistic falls above the upper bound at a 0.01 significance level (obtained from a table of Durbin-Watson test bounds), there is no evidence the error terms are positively correlated in the model with the transformed variables.
- Transform the intercept parameter, -0.394/(1-0.631164) = -1.068, and its standard error, 0.167/(1-0.631164) = 0.453 (the slope estimate and standard error don't require transforming).
- The fitted regression function for the original variables is predicted comsales = -1.068 + 0.17376 indsales.

One thing to note about the Cochrane-Orcutt approach is that it does not always work properly. This occurs primarily because if the errors are positively autocorrelated, then *r* tends to underestimate ρρ. When this bias is serious, then it can seriously reduce the effectiveness of the Cochrane-Orcutt procedure.

Hildreth-Lu Procedure

The Hildreth-Lu procedure is a more direct method for estimating ρρ. After establishing that the errors have an AR(1) structure, follow these steps:

1. Select a series of candidate values for ρρ (presumably values that would make sense after you assessed the pattern of the errors).
2. For each candidate value, regress y*tyt* on the transformed predictors using the transformations established in the Cochrane-Orcutt procedure. Retain the SSEs for each of these regressions.
3. Select the value which minimizes the SSE as an estimate of ρρ.

Notice that this procedure is similar to the Box-Cox transformation discussed previously and that it is not iterative like the Cochrane-Orcutt procedure.

To illustrate the Hildreth-Lu procedure, consider the Blaisdell Company example from above:

- Use Minitab's Calculator to define a transformed response variable, Y_hl.1 = comsales-0.1*LAG(comsales,1).
- Use Minitab's Calculator to define a transformed predictor variable, X_hl.1 = indsales-0.1*LAG(indsales,1).
- Fit a simple linear regression model with response Y_hl.1 and predictor X_hl.1 and record the SSE.
- Repeat steps 1-3 for a series of estimates of ?? to find when SSE is minimized (0.96 leads to the minimum in this case).
- The output for this model is:

```
Analysis of Variance
Source          DF     Adj SS      Adj MS    F-Value    P-Value
Regression       1    2.31988     2.31988     550.26      0.000
```

```
 X_hl.96        1    2.31988     2.31988    550.26      0.000
Error          17    0.07167     0.00422
Total          18    2.39155
Term         Coef          SE Coef      T-Value        P-Value   VIF
Constant     0.0712        0.0580        1.23          0.236
X_hl.96      0.16045       0.00684       23.46         0.000     1.00
Durbin-Watson Statistic =  1.72544
```

- Since the value of the Durbin-Watson Statistic falls above the upper bound at a 0.01 significance level (obtained from a table of Durbin-Watson test bounds), there is no evidence the error terms are positively correlated in the model with the transformed variables.
- Transform the intercept parameter, 0.0712/(1-0.96) = 1.78, and its standard error, 0.0580/(1-0.96) = 1.45 (the slope estimate and standard error don't require transforming).
- The fitted regression function for the original variables is predicted comsales = 1.78 + 0.16045 indsales.

First Differences Procedure

Since ρρ is frequently large for AR(1) errors (especially in economics data), many have suggested just setting ρρ = 1 in the transformed model of the previous two procedures. This procedure is called the first differences procedure and simply regresses y*t = yt-yt-1yt*=yt-yt-1 on the x*t,j=xt,j-xt-1,jxt,j*=xt,j-xt-1,j for j=1,…,p-1j=1,…,p-1 using regression through the origin. The estimates from this regression are then transformed back, setting ^ßj=^ß*jß^j=ß^j* for j=1,…,p-1j=1,…,p-1 and ^ß0=¯y-(^ß1¯x1+…+^ßp-1¯xp-1)ß^0=y¯-(ß^1x¯1+…+ß^p-1x¯p-1).

To illustrate the first differences procedure, consider the Blaisdell Company example from above:

- Use Minitab's Calculator to define a transformed response variable, Y_fd = comsales-LAG(comsales,1).
- Use Minitab's Calculator to define a transformed predictor variable, X_fd = indsales-LAG(indsales,1).
- Fit a simple linear regression model with response Y_fd and predictor X_fd and use the "Results" button to select the Durbin-Watson statistic:

  Durbin-Watson Statistic = 1.74883
- Since the value of the Durbin-Watson Statistic falls above the upper bound at a 0.01 significance level (obtained from a table of Durbin-Watson test bounds), there is no evidence the error terms are correlated in the model with the transformed variables.
- Fit a simple linear regression model with response Y_fd and predictor X_fd and no intercept. The output for this model is:

```
Term          Coef          SE Coef       T-Value       P-Value   VIF
X_fd          0.16849       0.00510        33.06         0.000    1.00
```

- Find the sample mean of comsales and indsales using Stat > Basic Statistics > Display Descriptive Statistics:

```
Variable  N N*   Mean SE Mean StDev Minimum    Q1 Median    Q3 Maximum
comsales 20  0 24.569   0.539 2.410  20.960 22.483 24.200 26.825  28.780
indsales 20  0 147.62    3.06 13.67  127.30 135.53 145.95 159.85  171.70
```

- Calculate the intercept parameter, 24.569 - 0.16849(147.62) = -0.303.
- The fitted regression function for the original variables is predicted comsales = -0.303 + 0.16849 indsales.

Remedial Measures of Autocorrelation

When autocorrelated error terms are found to be present, then one of the first remedial measures should be to investigate the omission of a key predictor variable. If such a predictor does not aid in reducing/eliminating autocorrelation of the error terms, then certain transformations on the variables can be performed. We discuss three transformations which are designed for AR(1) errors. Methods for dealing with errors from an AR(k) process do exist in the literature, but are much more technical in nature.

4.13 PREDICTION OF AUTOCORRELATION

Autocorrelation Prediction (AP) has been shown to be an effective technique for Pole-Zero modeling. This paper develops a new linear method for identifying a stable Pole-Zero model whose spectrum matches the envelope of a given spectrum. All the operations are performed in Autocorrelation domain, using no Fourier transformations. At one extreme, Autocorrelation Prediction reduces to a linear method for all-Zero modeling. At the other extreme, AP becomes the well-known Linear Prediction (LP). AP can also automatically determine the lowest denominator and numerator orders required to model efficiently the given spectral envelope. Spectra whose envelopes have deep valleys are shown to be matched more accurately at the valleys using AP rather than LP.

Linear prediction is a mathematical operation where future values of a discrete-time signal are estimated as a linear function of previous samples.

In digital signal processing, linear prediction is often called linear predictive coding (LPC) and can thus be viewed as a subset of filter theory. In system analysis (a subfield of mathematics), linear prediction can be viewed as a part of mathematical modelling or optimization.

4.14 HETEROSCEDASTICITY

Heteroscedasticity is a hard word to pronounce, but it doesn't need to be a difficult concept to understand. Heteroscedasticity refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

Nature of Heteroscedasticity

In statistics, a collection of random variables is heteroscedastic (or heteroskedastic;[a] from Ancient Greek hetero "different" and skedasis "dispersion") if there are sub-populations that have different variabilities from others. Here "variability" could be quantified by the

Notes

variance or any other measure of statistical dispersion. Thus heteroscedasticity is the absence of homoscedasticity.

The existence of heteroscedasticity is a major concern in the application of regression analysis, including the analysis of variance, as it can invalidate statistical tests of significance that assume that the modelling errors are uncorrelated and uniform hence that their variances do not vary with the effects being modeled. For instance, while the ordinary least squares estimator is still unbiased in the presence of heteroscedasticity, it is inefficient because the true variance and covariance are underestimated. Similarly, in testing for differences between sub-populations using a location test, some standard tests assume that variances within groups are equal.

Because heteroscedasticity concerns expectations of the second moment of the errors, its presence is referred to as misspecification of the second order.

Suppose there are a sequence of random variables and a sequence of vectors of random variables. In dealing with conditional expectations of Yt given Xt, the sequence {Yt}t=1n is said to be heteroscedastic if the conditional variance of Yt given Xt, changes with t. Some authors refer to this as conditional heteroscedasticity to emphasize the fact that it is the sequence of conditional variances that changes and not the unconditional variance. In fact, it is possible to observe conditional heteroscedasticity even when dealing with a sequence of unconditional homoscedastic random variables; however, the opposite does not hold. If the variance changes only because of changes in value of X and not because of a dependence on the index t, the changing variance might be described using a scedastic function.

When using some statistical techniques, such as ordinary least squares (OLS), a number of assumptions are typically made. One of these is that the error term has a constantvariance. This might not be true even if the error term is assumed to be drawn from identical distributions.

For example, the error term could vary or increase with each observation, something that is often the case with cross-sectional or time series measurements. Heteroscedasticity is often studied as part of econometrics, which frequently deals with data exhibiting it. While the influential 1980 paper by Halbert White used the term "heteroskedasticity" rather than "heteroscedasticity", the latter spelling has been employed more frequently in later works.

The econometrician Robert Engle won the 2003 Nobel Memorial Prize for Economics for his studies on regression analysis in the presence of heteroscedasticity, which led to his formulation of the autoregressive conditional heteroscedasticity (ARCH) modeling technique.

Consequences

One of the assumptions of the classical linear regression model is that there is no heteroscedasticity. Breaking this assumption means that the Gauss–Markov theorem does not apply, meaning that OLS estimators are not the Best Linear Unbiased Estimators (BLUE) and their variance is not the lowest of all other unbiased estimators. Heteroscedasticity does not cause ordinary least squares coefficient estimates to be biased, although it can cause ordinary least squares estimates of the variance (and thus, standard errors) of the coefficients to be biased, possibly above or below the true or population variance. Thus, regression analysis using heteroscedastic data will still provide an unbiased estimate for the relationship between the predictor variable and the outcome, but standard errors and therefore inferences obtained from data analysis are suspect. Biased standard errors lead to biased inference, so results of hypothesis tests are possibly wrong. For example, if OLS is performed on a heteroscedastic data set, yielding biased standard error estimation, a researcher might fail to reject a null hypothesis at a given significance level, when that null hypothesis was actually uncharacteristic of the actual population (making a type II error).

Under certain assumptions, the OLS estimator has a normal asymptotic distribution when properly normalized and centered (even when the data does not come from a normal distribution). This result is used to justify using a normal distribution, or a chi square distribution (depending on how the test statistic is calculated), when conducting a hypothesis test. This holds even under heteroscedasticity. More precisely, the OLS estimator in the presence of heteroscedasticity is asymptotically normal, when properly normalized and centered, with a variance-covariance matrix that differs from the case of homoscedasticity. In 1980, White proposed a consistent estimator for the variance-covariance matrix of the asymptotic distribution of the OLS estimator. This validates the use of hypothesis testing using OLS estimators and White's variance-covariance estimator under heteroscedasticity.

However, it has been said that students in econometrics should not overreact to heteroscedasticity. One author wrote, "unequal error variance is worth correcting only when the problem is severe." In addition, another word of caution was in the form, "heteroscedasticity has never been a reason to throw out an otherwise good model." With the advent of heteroscedasticity-consistent standard errors allowing for inference without specifying the conditional second moment of error term, testing conditional homoscedasticity is not as important as in the past.

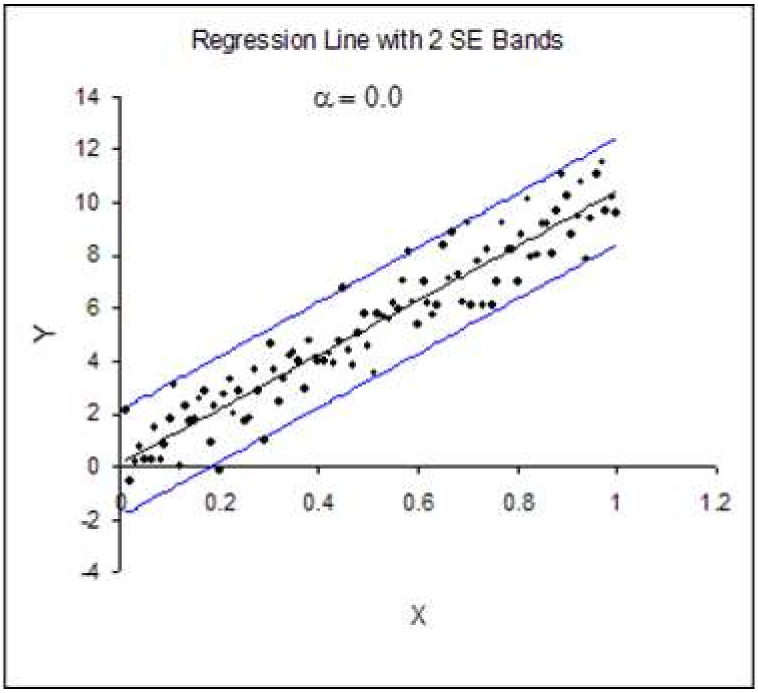
For any non-linear model (for instance Logit and Probit models), however, heteroscedasticity has more severe consequences: the maximum likelihood estimates (MLE) of the parameters will be biased, as well as inconsistent (unless the likelihood function is modified to correctly take into account the precise form of heteroscedasticity). Yet, in the context of binary choice models (Logit or Probit), heteroscedasticity will only result in a positive scaling effect on the asymptotic mean of the misspecified MLE (i.e. the model that ignores heteroscedasticity). As a result, the predictions which are based on the misspecified MLE will remain correct. In addition, the misspecified Probit and Logit MLE will be asymptotically normally distributed which allows performing the usual significance tests (with the appropriate variance-covariance matrix). However, regarding the general hypothesis testing, as pointed out by Greene, "simply computing a robust covariance matrix for an otherwise inconsistent estimator does not give it redemption. Consequently, the virtue of a robust covariance matrix in this setting is unclear."

4.15 HETEROSCEDASTICITY STRUCTURES

When this condition holds, the error terms are homoskedastic, which means the errors have the same scatter regardless of the value of X. When the scatter of the errors is different, varying depending on the value of one or more of the independent variables, the error terms are heteroskedastic.

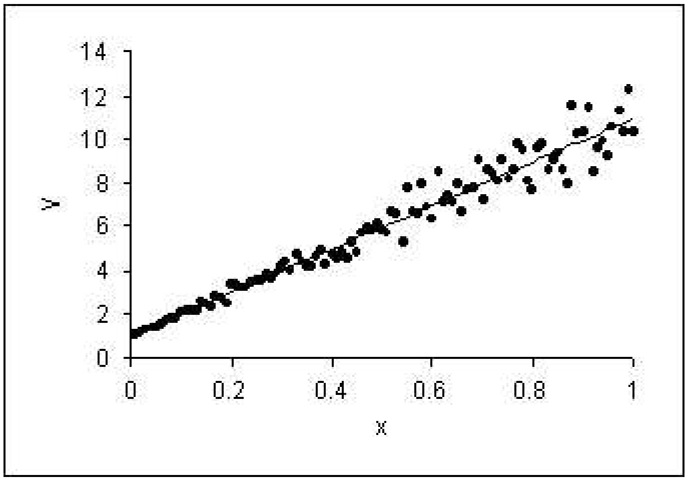
Heteroskedasticity has serious consequences for the OLS estimator. Although the OLS estimator remains unbiased, the estimated SE is wrong. Because of this, confidence intervals and hypotheses tests cannot be relied on. In addition, the OLS estimator is no longer BLUE. If the form of the heteroskedasticity is known, it can be corrected (via appropriate transformation of the data) and the resulting estimator, generalized least squares (GLS), can be shown to be BLUE. This chapter is devoted to explaining these points.

Heteroskedasticity can best be understood visually. Figure depicts a classic picture of a homoskedastic situation. We have drawn a regression line estimated via OLS in a simple, bivariate model. The vertical spread of the data around the predicted line appears to be fairly constant as X changes. In contrast, Figure shows the same model with heteroskedasticity. The vertical spread of the data around the predicted line is clearly increasing as X increases.



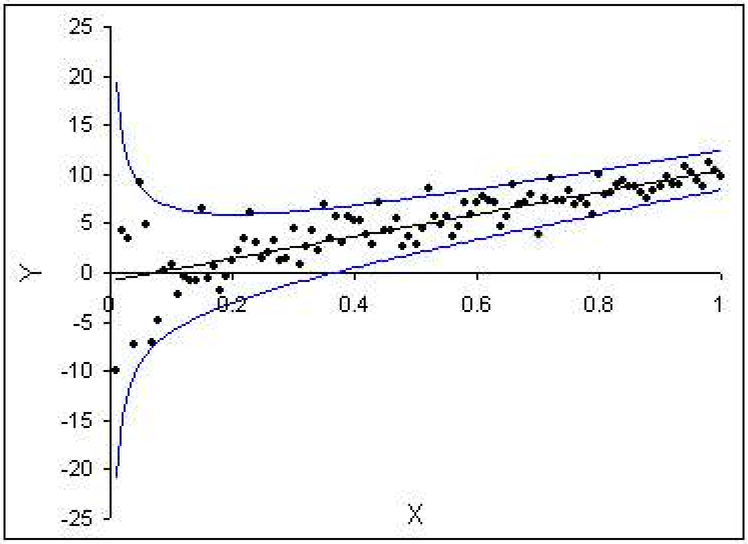
Homoskedasticity in a Simple, Bivariate Model

One of the most difficult parts of handling heteroskedasticity is that it can take many different forms. Figure shows another example of heteroskedasticity. In this case, the spread of the errors is large for small values of X and then gets smaller as X rises. If the spread of the errors is not constant across the X values, heteroskedasticity is present.



Heteroskedasticity in a Simple, Bivariate Model

Notes



Another Form of Heteroskedasticity

This is organized around four basic issues:

- Understanding the violation itself
- Appreciating the consequences of the violation
- Diagnosing the presence of the violation
- Correcting the problem.

The next two sections describe heteroskedasticity and its consequences in two simple, contrived examples. Although heteroskedasticity can sometimes be identified by eye, Section presents a formal hypothesis test to detect heteroskedasticity. Section describes the most common way in which econometricians handle the problem of heteroskedasticity – using a modified computation of the estimated SE that yields correct reported SEs. Section discusses a more aggressive method for dealing with heteroskedasticity comparable to the approaches commonly employed in dealing with autocorrelation in which data transformation is applied to obtain the best linear unbiased estimator. Finally, Section offers an extended discussion of heteroskedasticity in an actual data set.

4.16 TESTS FOR HETEROSCEDASTICITY

The concept of heteroscedasticity - the opposite being homoscedasticity - is used in statistics, especially in the context of linear regression or for time series analysis, to describe the case where the variance of errors or the model is not the same for all observations, while often one of the basic assumption in modeling is that the variances are homogeneous and that the errors of the model are identically distributed.

In linear regression analysis, the fact that the errors of the model (also named residuals) are not homoskedastic has the consequence that the model coefficients estimated using ordinary least squares (OLS) are neither unbiased nor those with minimum variance. The estimation of their variance is not reliable.

If it is suspected that the variances are not homogeneous (a representation of the residuals against the explanatory variables may reveal heteroscedasticity), it is therefore necessary to perform a test for heteroscedasticity. Several tests have been developed, with the following null and alternative hypotheses:

H0 : The residuals are homoscedastic

Ha : The residuals are heteroscedastic

XLSTAT now includes:

Breusch-Pagan test

White test and modified White test (Wooldridge)

4.17 REMEDIAL MEASURES THE METHODS OF WEIGHTED LEAST SQUARE

The OLS estimators remains unbiased and consistent in the presence of Heteroscedasticity, but they are no longer efficient not even asymptotically. This lack of efficiency makes the usual hypothesis testing procedure of dubious value. Therefore remedial measures may be called. There are two approaches for remedial measures for heteroscedasticity

(i) $\sigma 2\mathrm{i}\sigma \mathrm{i}2$ is known

Consider the simple linear regression model $Y_i = \alpha + \beta X_i + \mu_i$

If $\mathrm{V}(\mu \mathrm{i}) = \sigma 2\mathrm{iV}\left(\mu \mathrm{i}\right) = \sigma \mathrm{i}2$ then heteroscedasticity is present. Given the values of s2isi2 heteroscedasticity can be corrected by using weighted least squares (WLS) as a special case of Generalized Least Square (GLS). Weighted least squares is the OLS method of estimation applied to the transformed model.

When heteroscedasticity is detected by any appropriate statistical test, then appropriate solution is transform the original model in such a way that the transformed disturbance term has constant variance. The transformed model reduces to the adjustment of the original data. The transformed error term µi has a constant variance i.e. homoscedastic. Mathematically

$$\mathrm{V}(\mu * \mathrm{i}) === \mathrm{V}(\mu \mathrm{i}\sigma \mathrm{i})1\sigma 2\mathrm{iVar}(\mu \mathrm{i})1\sigma 2\mathrm{i}\sigma 2\mathrm{i} = 1\mathrm{V}(\mu \mathrm{i}*) = \mathrm{V}\left(\mu \mathrm{i}\sigma \mathrm{i}\right) = 1\sigma \mathrm{i}2\mathrm{Var}\left(\mu \mathrm{i}\right) = 1\sigma \mathrm{i}2\sigma \mathrm{i}2 = 1$$

This approach has its limited use as the individual error variance are not always known a priori. In case of significant sample information, reasonable guesses of the true error variances can be made and be used for $\sigma 2\mathrm{i}\sigma \mathrm{i}2$.

(ii) When s2isi2 is unknown

If $\sigma 2\mathrm{i}\sigma \mathrm{i}2$ is not known a priori, then heteroscedasticity is corrected by hypothesizing a relationship between the error variance and one of the explanatory variables. There can be several versions of the hypothesized relationship. Suppose hypothesized relationship is $\mathrm{Var}(\mu) = \sigma 2\mathrm{X}2\mathrm{iVar}\left(\mu\right) = \sigma 2\mathrm{Xi}2$ (error variance is proportional to $\mathrm{X}2\mathrm{iXi}2$). For this hypothesized relation we will use the following transformation to correct for heteroscedasticity for the following simple linear regression model $Y_i = \alpha + \beta X_i + \mu_i$.

$$\mathrm{YiXi} \Rightarrow \mathrm{Y} * \mathrm{iwhere\,Y} * \mathrm{i} = \alpha \mathrm{Xi} + \beta + \mu \mathrm{iXi}\beta + \alpha * \mathrm{i} + \mu * \mathrm{iYiXi}, \alpha * \mathrm{I} = 1\mathrm{Xi}$$

$$\mathrm{and}\,\mu * \mathrm{i} = \mu \mathrm{XiYiXi} = \alpha \mathrm{Xi} + \beta + \mu \mathrm{iXi} \Rightarrow \mathrm{Yi}* = \beta + \alpha \mathrm{i} * + \mu \mathrm{i} *$$

where $Yi* = YiXi, \alpha I* = 1Xi\text{and}\,\mu i* = \mu Xi$

Now the OLS estimation of the above transformed model will yield the efficient parameter estimates as μ*iμi*'s have constant variance. i.e.

$V(\mu * i) = V(\mu iXi)1X2iV(\mu 2i)1X2i\sigma 2X2i\sigma 2$

$= \text{Constant}V(\mu i*) = V\left(\mu iXi\right) = 1Xi2V\left(\mu i2\right) = 1Xi2\sigma 2Xi2 = \sigma 2 = \text{Constant}$

For correction of heteroscedasticity some other hypothesized relations are

- Error variance is proportional to Xi (Square root transformation) i.e E(μ2i) = s2XiE(μi2) = s2Xi

  The transformed model is

  YiXi--v = aXi--v + ßXi--v + μiXi--vYiXi = aXi + ßXi + μiXi

  It (transformed model) has no intercept term. Therefore we have to use the regression through the origin model to estimate aa and ß.To get original model, multiply Xi--vXi with transformed model.

- **Error Variance is proportional to the square of the mean value of Y.** i.e. E(μ2i)= σ^2 [E(Yi)]2E(μi2) = σ^2 [E(Yi)]2

  Here the variance of μiμi is proportional to the square of the expected value of Y, and E(Yi) = a + ßxi.

  The transformed model will be

  YiE(Yi) = aE(Yi) + ßXiE(Yi) + μiE(Yi)YiE(Yi) = aE(Yi) + ßXiE(Yi) + μiE(Yi)

  This transformation is not appropriate because E(Yi) depends upon aa and ß which are unknown parameters. Yi^=a^+ß^Yi^=a^+ß^ is an estimator of E(Yi), so we will proceed in two steps

1. We run the usual OLS regression dis-regarding the heteroscedasticity problem and obtain Yi^Yi^
2. We will transform the model by using estimated Yi^Yi^ i.e. YiYi^ = a1Yi^+ß1XiYi^+μiYi^YiYi^ = a1Yi^+ß1XiYi^+μiYi^ and run the regression on transformed model.

   This transformation will perform satisfactory results only if the sample size is reasonably large.

- **Log transformation such as ln** $Y_i = \alpha + \beta\ ln\ X_i + \mu_i$

  Log transformation compresses the scales in which the variables are measured. But this transformation is not applicable if some of the Y and X values are zero or negative.

4.18 MULTICOLINEARITY

Multicollinearity is a state of very high intercorrelations or inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable.

There are certain reasons why multicollinearity occurs:

a) It is caused by an inaccurate use of dummy variables.

b) It is caused by the inclusion of a variable which is computed from other variables in the data set.

c) Multicollinearity can also result from the repetition of the same kind of variable.

d) Generally occurs when the variables are highly correlated to each other.

Multicollinearity can result in several problems. These problems are as follows:

The partial regression coefficient due to multicollinearity may not be estimated precisely. The standard errors are likely to be high.

Multicollinearity results in a change in the signs as well as in the magnitudes of the partial regression coefficients from one sample to another sample.

Multicollinearity makes it tedious to assess the relative importance of the independent variables in explaining the variation caused by the dependent variable.

In the presence of high multicollinearity, the confidence intervals of the coefficients tend to become very wide and the statistics tend to be very small. It becomes difficult to reject the null hypothesis of any study when multicollinearity is present in the data under study.

There are certain signals which help the researcher to detect the degree of multicollinearity.

One such signal is if the individual outcome of a statistic is not significant but the overall outcome of the statistic is significant. In this instance, the researcher might get a mix of significant and insignificant results that show the presence of multicollinearity.Suppose the researcher, after dividing the sample into two parts, finds that the co-efficients of the sample differ drastically. This indicates the presence of multicollinearity. This means that the co-efficients are unstable due to the presence of multicollinearity. Suppose the researcher observes drastic change in the model by simply adding or dropping some variable. This also indicates that multicollinearity is present in the data.

Multicollinearity can also be detected with the help of tolerance and its reciprocal, called variance inflation factor (VIF). If the value of tolerance is less than 0.2 or 0.1 and, simultaneously, the value of VIF 10 and above, then the multicollinearity is problematic.

4.19 IMPLICATIONS OF MULTICOLINEARITY

Multicollinearity is problem that you can run into when you're fitting a regression model or other linear model. It refers to predictors that are correlated with other predictors in the model. Unfortunately, the effects of multicollinearity can feel murky and intangible, which makes it unclear whether it's important to fix.

My goal in this blog post is to bring the effects of multicollinearity to life with real data! Along the way, I'll show you a simple tool that can remove multicollinearity in some cases.

Moderate multicollinearity may not be problematic. However, severe multicollinearity is a problem because it can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable and difficult to interpret. Multicollinearity saps the statistical power of the analysis, can cause the co-efficients to switch signs, and makes it more difficult to specify the correct model.

Notes

The symptoms sound serious, but the answer is both yes and no depending on your goals. (Don’t worry, the example we'll go through next makes it more concrete.) In short, multicollinearity:

a) Can make choosing the correct predictors to include more difficult.

b) Interferes in determining the precise effect of each predictor,

c) Doesn’t affect the overall fit of the model or produce bad predictions.

Depending on your goals, multicollinearity isn’t always a problem. However, because of the difficulty in choosing the correct model when severe multicollinearity is present, it’s always worth exploring.

The Regression Scenario: Predicting Bone Density

I’ll use a subset of real data that I collected for an experiment to illustrate the detection, effects and removal of multicollinearity. You can read about the actual experiment here and the worksheet is here. (If you're not already using it, please download the free 30-day trial of Minitab and play along!)

We’ll use Regression to assess how the predictors of physical activity, percent body fat, weight and the interaction between body fat and weight are collectively associated with the bone density of the femoral neck.

Given the potential for correlation among the predictors, we’ll have Minitab display the variance inflation factors (VIF), which indicate the extent to which multicollinearity is present in a regression analysis. A VIF of 5 or greater indicates a reason to be concerned about multicollinearity.

Multicollinearity refers to a situation with a high correlation among the explanatory variables within a multiple regression model. For the obvious reason it could never appear in the simple regression model, since it only has one explanatory variable. In chapter 8 we shortly described the consequences of including the full exhaustive set of dummy variables created from a categorical variable with several categories. We referred to that as to fall in the dummy variable trap. By including the full set of dummy variables, one end up with a perfect linear relation between the set of dummies and the constant term. When that happens we have what is called perfect multicollinearity. In this chapter we will in more detail discuss the issue of multicollinearity and focus on what sometimes is called imperfect multicollinearity which referrers to the case where a set of variables are highly correlated but not perfect.

4.20 CONSEQUENCES OF MULTICOLINEARITY

The lack of independence among the explanatory variables in a data set. It is a sample problem and a state of nature that results in relatively large standard errors for the estimated regression co-efficients, but not biased estimates.

The consequences of perfect correlation among the explanatory variables is easiest explained by an example. Assume that we would like to estimate the parameters of the following model:

$$Y = B_0 + B_1X_1 + B_2X_2 + U$$

where X1 is assumed to be a linear combination of X2 in the following way:

$$X_2 = a + bX_1$$

Notes

and where a and b are two arbitrary constants. If we substitute into we receive:

$$Y = B_0 + B_1X_1 + B_2(a + bX_2) + U$$

$$Y = (B_0 + aB_2) + (B_1 + bB_2)X_1 + U$$

It implies we can only receive estimates of [B0 + aB2) and (Bx + bB2). But since these two expressions contain three unknown parameters there is no way we can receive estimates for all three parameters. We simply need more information, which is not available. Hence, with perfect multicollinearity it is impossible to receive an estimate of the intercept and the slope coefficients.

This was an example of the extreme case of perfect multicollinearity, which is not very likely to happen in practice, other than when we end up in a dummy variable trap or a similar situation. More interesting is to investigate the consequences on the parameters and their standard errors when high correlation is present. We will start this discussion with the sample estimator of the slope coefficient B1 in the assumption that X1 and X2 is highly correlated but not perfect. The situation for the sample estimator of B2 is identical to that of B1 so it is not necessary to look at both. The sample estimator for B1 is given by:

$$b_1 = \frac{(r_{\gamma 1} - r_{\gamma 12}r_{\gamma 2})}{(1 - r_{12}^2)}\frac{S_\gamma}{S_1}$$

The estimator b1 is a function of r which is the correlation between Y and X1, r the correlation between X1 and X2, rY2 the correlation between Y and X2, SY and S1 which are the standard deviations for Y and X1 respectively.

The first thing to observe is that r appears in both the numerator and the denominator, but that it is squared in the denominator and makes the denominator zero in case of perfect correlation. In case of a strong correlation, the denominator has an increasing effect on the size of the expression but since the correlation coefficient appears in the numerator as well with a negative sign, it is difficult to say how the size of the parameter will change, without any further assumptions. However, it can be shown that the OLS estimators remain unbiased and consistent, which means that estimated coefficients in repeated sampling still will center around the population coefficient. On the other hand, this property says nothing about how the estimator will behave in a specific sample. Therefore we will go through an example in order to shed some light on this issue.

Example:

Consider the following regression model:

$$Y = B_0 + B_1X_1 + U$$

We would like to know how the estimate of B1 changes when we include another variable X2 that is highly correlated with X1. Using a random sample of 20 observations we calculate the following statistics.

$$S_\gamma = 5.1 \qquad r_{\gamma 1} = 0.843$$

$$S_1 = 5.0 \qquad r_{\gamma 2} = 0.878$$

$$r_{12} = 0.924$$

For the simple regression case we receive:

$$b_1 = r_{\gamma 1}\frac{S_\gamma}{S_1} = 0.843 \times \frac{5.1}{5.0} = 0.86$$

Notes

For the multiple regression case when including both X1 and X2 we receive:

$$b_1^* = \frac{0.843 - 0.924 \times 0.878}{1 - 0.924^2} = \frac{5.1}{5.0} = 0.211$$

Hence, when including an additional variable the estimated coefficient decreased in size as a result of the correlation between the two variables. Is it possible to find an example where the estimator is increasing in size in absolute terms? Well, consider the case where X2 is even more correlated with X , lets say that r12 = 0.99. That would generate a negative estimate and the small number in the denominator will make the estimate larger in absolute terms. It is also possible to make up an examples where the estimator moves in the other direction. Hence, the estimated slope coefficient could move in any direction as a result of multicollinearity.

In order to analyze how the variance of the parameter estimates change it is informative to look at the equation for the variance. The variance is given by the following expression.

$$V(b_1) = \frac{\sum_{i=1}^{n} e_i^2}{n-3} \frac{1}{(1 - r_{i2}^2 \sum_{i=1}^{n} (x_{1i} - \overline{x_1})^2}$$

When the correlation between x and x^2 equals zero, will the variance of the multiple regression coefficient coincide with the variance for the coefficient of the simple regression model. However, when the correlation equals 1 or-S the variance given by undefined just as the estimated slope coefficient. In sum, the greSte0 the degree of the multicollinearity, the less precise will be the estimates of the parameters, which means that the estimated coefficients will - vary a lot from sample to sample. But make no mistakes; does not destroy the nice property of minimum variance among lineal" unbiased estimator. It still has a minimum 'variance, but minimum variance does not mean that the variance will be small.

It seems like the level of both the estimated parameter and its standard error are affected by multicollinearity. But how will this affect the ratio between them; the f-value. It can be shown that the computed f-value in general will decrease since the standard error is affected more strongly compared to the coefficient. This will usually result in non-significant parameter estimates.

Another problem with multicollinearity is that the estimates will be very sensitive to changes in specification. This is a consequence from the fact that there is very little unique variation left to explain the dependent variable since most of the variation is in common between the two explanatory variables. Hence, the parameter estimates are very unstable and sometimes it can even result in wrong signs for the regression coefficient, despite the fact that it is unbiased. A wrong sign is referred to a sign that is unexpected according to the underlying theoretical model, or the prior believes based on common sense. However, sometimes we are dealing with inferior goods which means that we have to be careful with what we call "wrong" sign. Unexpected signs usually require more analysis to understand where it comes from.

4.21 TESTS FOR MULTICOLINEARITY

Fortunately, there is a very simple test to assess multicollinearity in your regression model. The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation. Statistical software calculates a VIF for each independent variable.

Some of the common methods used for detecting multicollinearity include:

- The analysis exhibits the signs of multicollinearity - such as, estimates of the coefficients vary from model to model.
- The t-tests for each of the individual slopes are non-significant ($P > 0.05$), but the overall F-test for testing all of the slopes are simultaneously 0 is significant ($P < 0.05$).
- The correlations among pairs of predictor variables are large.

Looking at correlations only among pairs of predictors, however, is limiting. It is possible that the pairwise correlations are small and yet a linear dependence exists among three or even more variables, for example, if $X_3 = 2X_1 + 5X_2$ + error, say. That's why many regression analysts often rely on what are called variance inflation factors (VIF) to help detect multicollinearity.

Recall that we learned previously that the standard errors and hence the variances - of the estimated co-efficients are inflated when multicollinearity exists. So, the variance inflation factor for the estimated coefficient b_k denoted VIF_k is just the factor by which the variance is inflated.

Let's be a little more concrete. For the model in which xk is the only predictor:

yi = ß0 + ßkxik + ∈iyi = ß0 + ßkxik + ∈i

it can be shown that the variance of the estimated coefficient bk is:

Var(bk)min = σ^2 ∑ ni = 1(xik-¯xk)2Var(bk)min = σ^2 ∑ = 1n(xik-x¯k)2

Note that we add the subscript "min" in order to denote that it is the smallest the variance can be. Don't worry about how this variance is derived — we just need to keep track of this baseline variance, so we can see how much the variance of bk is inflated when we add correlated predictors to our regression model.

Let's consider such a model with correlated predictors:

yi = ß0+ß1xi1+....+ßkxik+....+ßp-1xi,p-1+∈iyi = ß0+ß1xi1+....+ßkxik+.....+ßp-1xi,p-1+∈i

Now, again, if some of the predictors are correlated with the predictor xk, then the variance of bk is inflated. It can be shown that the variance of bk is:

Var(bk) = σ^2 ∑ ni = 1(xik-¯xk)2 × 11 - R2kVar(bk) = σ^2 ∑ i = 1n(xik-x¯k)2 × 11-Rk2

where R2kRk2 is the R^2- value obtained by regressing the k^{th} predictor on the remaining predictors. Of course, the greater the linear dependence among the predictor xk and the other predictors, the larger the R2kRk2 value. And, as the above formula suggests, the larger the R2kRk2 value, the larger the variance of bk.

To answer this question, all we need to do is take the ratio of the two variances. Doing so, we obtain:

Var(bk)Var(bk)min = (σ^2 ∑ (xik-¯xk)2 × 11-R2k) (σ^2 ∑ (xik-¯xk)2) = 11-R2kVar(bk) Var(bk)min = (σ^2 ∑ (xik-x¯k)2 × 11-Rk2) (σ^2 ∑ (xik-x¯k)2) = 11-Rk2

The above quantity is what is deemed the variance inflation factor for the kth predictor. That is:

VIFk = 11 - R2k VIFk = 11 - Rk2

where R_2kRk_2 is the R^2 - value obtained by regressing the k^{th} predictor on the remaining predictors. Note that a variance inflation factor exists for each of the k predictors in a multiple regression model.

How do we interpret the variance inflation factors for a regression model? Again, it is a measure of how much the variance of the estimated regression co-efficient *bk* is "inflated" by the existence of correlation among the predictor variables in the model. A VIF of 1 means that there is no correlation among the kth predictor and the remaining predictor variables, and hence the variance of bk is not inflated at all. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

4.22 METHODS OF ESTIMATION OF MULTICOLINEARITY

- The easiest way for the detection of multicollinearity is to examine the correlation between each pair of explanatory variables. If two of the variables are highly correlated, then this may the possible source of multicollinearity. However, pair-wise correlation between the explanatory variables may be considered as the sufficient, but not the necessary condition for the multicollinearity.
- The second easy way for detecting the multicollinearity is to estimate the multiple regression and then examine the output carefully. The rule of thumb to doubt about the presence of multicollinearity is very high R_2R_2 but most of the coefficients are not significant according to their p-values. However, this cannot be considered as an acid test for detecting multicollinearity. It will provide an apparent idea for the presence of multicollinearity.
- As, the co-efficient of determination in the regression of regressor X_jX_j on the remaining regressors in the model, increases toward unity, that is, as the collinearity of X_jX_j with the other regressors increases, VIFVIF also increases and in the limit it can be infinite. Therefore, we can use the VIFVIF as an indicator of multicollinearity. The larger the value of VIF_jVIF_j, the more “troublesome” or collinear the variable X_jX_j. As a rule of thumb, if the VIFVIF of a variable exceeds 10, which will happen if multiple correlation co-efficient for j-th variable $R2_jR_j2$ exceeds 0.90, that variable is said to be highly collinear.
- The Farrar-Glauber test (F-G test) for multicollinearity is the best way to deal with the problem of multicollinearity.

4.23 MULTICOLINEARITYAND PREDICTION

Multicollinearity means that some of the regressors (Independent variables) are highly correlated with each other. It will make the estimate highly in-stable. This instability will increase the variance of estimates. It means that if there is a small change in X, produces large changes in estimate.

Effects of Multicollinearity

1. It will be difficult to find the correct predictors from the set of predictors.
2. It will be difficult to find out precise effect of each predictor.

If the estimates are not reliable, then it will perform poorly on test data because the estimated function mightn't have generalised it properly for the data. Then, the prediction accuracy for test data will be bad.

If you want to know the intuition behind the disadvantages of multicollinearity, please read the long answer.

Long Answer:

The least square estimate is given by

ß^~N(ß, σ^2 (XTX)-1)(1)(1)ß^~N(ß, σ^2 (XTX) - 1)

From the above equation, The variance-covariance matrix of estimate is given by

Var(ß^) = σ^2 (XTX)-1(2)(2)Var(ß^) = σ^2 (XTX) - 1

The variance of any single estimate is given by

Var(ßj^) = σ^2 ∑j(xj-x¯)2(3)(3)Var(ßj^) = σ^2 ∑j(xj-x¯)2

These formulas don't work if Multicollinearity exists between the regressors (Balaji Pitchai Kannu's answer to What is an intuitive explanation of the multiple linear regression assumptions?). Multicollinearity means that some of the regressors(Independent variables) are highly correlated with each other. If the regressors are highly correlated (Features of X are not linearly independent), then the rank of matrix X is less than p+1 (where p is number of regressors). So, the inverse of XTXXTX matrix doesn't exist. But, we already knew that the closed form equation of regression estimate need (XTX)-1(XTX)-1

ß^ = (XTX) - 1XTY()()ß^ = (XTX) - 1XTY

Even though multicollinearity exist between the regressors, still we can able to find the regression estimate by some other methods (Pseudo-Inverse, etc.,). But, that estimate won't be unique. It means that ß'sß's are highly in stable. This instability will increase the variance of estimates. We can able to measure the increased variance of an estimate due to multicollinearity using the formula given below.

Var(ßj^) = σ^2 ∑j(xj-x¯)2*VIF(4)(4)Var(ßj^) = σ^2 ∑j(xj-x¯)2*VIF

Var(ßj^) = σ^2 ∑j(xj-x¯)2*11-R2j(5)(5)Var(ßj^) = σ^2 ∑j(xj-x¯)2*11-Rj2

VIF = Variance Inflation Factor.

RjRj is a measure of how much variance of feature xjxj can be explained by other features. If all the features are orthogonal to each other, then RjRj will be zero and the equation 3 and 4 will be same. If 90% of variance of xjxj is possible to explain by other features, then the variance of ßj^ßj^ will be inflated by 10 times. As I said already, the inflation of estimate's variance happens due to instability of estimates. ßß is highly in-stable means that the small changes in the X produces high impact on the solution of XTXß^=XTYXTXß^=XTY. Usually, the stability of an estimate is measures by condition number.

Condition number = Largest eigen value of X^{T}XSmallest eigen value of X^{T}X()()Condition number = Largest eigen value of X^{T}XSmallest eigen value of X^{T}X

If condition number is very large, then ßj^ßj^will be highly in-stable. If the regressors are highly correlated with each other, then the columns of XTXXTXare not linearly independent which turns out to be the smallest eigen value of XTXXTX will be zero.

4.24 REMEDIAL MEASURES OF MULTICOLINEARITY

Multicollinearity does not actually bias results; it just produces large standard errors in the related independent variables. With enough data, these errors will be reduced.

In a pure statistical sense multicollinearity does not bias the results, but if there are any other problems which could introduce bias multicollinearity can multiply (by orders of magnitude) the effects of that bias. More importantly, the usual use of regression is to take coefficients from the model and then apply them to other data. If the new data differs in any way from the data that was fitted we may introduce large errors in predictions because the pattern of multicollinearity between the independent variables is different in new data from the data used for your estimates. We try seeing what happens if we use independent subsets of your data for estimation and apply those estimates to the whole data set.

1) Leave the model as is, despite multicollinearity. The presence of multicollinearity doesn’t affect the fitted model provided that the predictor variables follow the same pattern of multicollinearity as the data on which the regression model is based.
2) Drop one of the variables. An explanatory variable may be dropped to produce a model with significant coefficients. However, you lose information (because you’ve dropped a variable). Omission of a relevant variable results in biased coefficient estimates for the remaining explanatory variables.
3) Obtain more data. This is the preferred solution. More data can produce more precise parameter estimates (with lower standard errors).
4) Mean-center the predictor variables. Mathematically this has no effect on the results from a regression. However, it can be useful in overcoming problems arising from rounding and other computational steps if a carefully designed computer program is not used.
5) Standardize your independent variables. This may help reduce a false flagging of a condition index above 30.

4.25 SUMMARY

The classical general equilibrium model aims to describe the economy by aggregating the behavior of individuals and firms. Note that the classical general equilibrium model is unrelated to classical economics, and was instead developed within neoclassical economics beginning in the late 19th century. An econometric model specifies the statistical relationship that is believed to hold between the various economic quantities pertaining to a particular economic phenomenon.

Assumptions form the foundation upon which theories, models, and analyses are constructed. They simplify and highlight the problem or topic under study. Even though assumptions often appear to be "unrealistic," when properly used they make it possible to analyze an exceedingly complex set of events.

Assumptions are inherently abstract and seemingly unrealistic. However, they make it possible to identify a specific cause-and-effect relation by assuming other influences are not involved. For example, the law of demand is the relation between demand price and quantity demanded. Demand, however, is also affected by factors other than demand price, such as buyers' income, the prices of other goods, or buyers' preferences. When working with the law of demand, it is essential to assume that these other factors do not influence demand when identifying the law of demand.

In statistics, errors-in-variables models or measurement error models are regression models that account for measurement errors in the independent variables. In contrast, standard regression models assume that those regressors have been measured exactly, or observed without error; as such, those models account only for errors in the dependent variables, or responses.

In the case when some regressors have been measured with errors, estimation based on the standard assumption leads to inconsistent estimates, meaning that the parameter estimates do not tend to the true values even in very large samples. For simple linear regression the effect is an underestimate of the coefficient, known as the attenuation bias. In non-linear models the direction of the bias is likely to be more complicated.

Errors-in-variables regression is used much less than ordinary least squares regression apart from in certain specialized areas such as comparison of methods studies and allometry/ isometry assessments. Its lack of use may seem surprising given how often the X-variable is not measured without error. Nevertheless, in situations when one is using regression purely for descriptive purposes or for prediction, ordinary least squares regression is still the best option, albeit one accepts that the slope is probably attenuated. We give two examples of where ordinary least squares (OLS) linear regression could have been used rather than the more complex errors-in-variable regression.

The wrong type of errors-in-variables regression is often used when dealing with an asymmetric relationship - in other words where there is a clear independent and dependent variable. In this situation, orthogonal regression (including major axis and reduced major axis) is inappropriate. Instead, if there is substantial measurement error on the X-axis, the slope of the OLS regression should be corrected for attenuation using the method of moments. It is hard to find examples of the use of this method in the literature, but we do give several examples (such as relating survival time to HIV load, and relating phytophage species richness to tree abundance) where it should have been used rather than orthogonal regression.

The term parameter estimation refers to the process of using sample data (in reliability engineering, usually times-to-failure or success data) to estimate the parameters of the selected distribution. Several parameter estimation methods are available. This section presents an overview of the available methods used in life data analysis. More specifically, we start with the relatively simple method of Probability Plotting and continue with the more sophisticated methods of Rank Regression (or Least Squares), Maximum Likelihood Estimation and Bayesian Estimation Methods.

Maximum likelihood estimation (ML estimation) is another estimation method. In the case of the linear model with errors distributed as N(0, s2), the ML and least-squares estimators are the same. However, in the general case, this is not true. Often the variance of the ML estimator (MLE) is less than the variance of other estimators (including the least-squares estimator), and hence is the preferable estimation method.

In statistics, econometrics, epidemiology and related disciplines, the method of instrumental variables (IV) is used to estimate causal relationships when controlled experiments are not feasible or when a treatment is not successfully delivered to every unit in a randomized experiment. Intuitively, IV is used when an explanatory variable of interest is correlated with the error term, in which case ordinary least squares and ANOVA gives biased results. A valid instrument induces changes in the explanatory variable but has no independent effect on the dependent variable, allowing a researcher to uncover the causal effect of the explanatory variable on the dependent variable.

Instrumental variable methods allow for consistent estimation when the explanatory variables (covariates) are correlated with the error terms in a regression model. Such correlation may occur 1) when changes in the dependent variable change the value of at least one of the covariates ("reverse" causation), 2) when there are omitted variables that affect both the dependent and independent variables, or 3) when the covariates are subject to non-random measurement error. Explanatory variables which suffer from one or more of these issues in the context of a regression are sometimes referred to as endogenous. In this situation, ordinary least squares produces biased and inconsistent estimates. However, if an instrument is available, consistent estimates may still be obtained. An instrument is a variable that does not itself belong in the explanatory equation but is correlated with the endogenous explanatory variables, conditional on the value of other covariates.

Instrumental Variables regression (IV) basically splits your explanatory variable into two parts: one part that could be correlated with e and one part that probably isn’t. By isolating the part with no correlation, it’s possible to estimate ß in the regression equation:

Yi = ß0 + ß1Xi + ei.

Autocorrelation is the correlation of a signal with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them. The analysis of autocorrelation is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies. It is often used in signal processing for analyzing functions or series of values, such as time domain signals.

Autocorrelation is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals. It is the same as calculating the correlation between two different time series, except that the same time series is actually used twice: once in its original form and once lagged one or more time periods.

Autocorrelation can be useful for technical analysis, which is most concerned with the trends of, and relationships between, security prices using charting techniques in lieu of a company's financial health or management. Technical analysts can use autocorrelation to see how much of an impact past prices for a security has on its future price.

Notes

Autocorrelation can show if there is a momentum factor associated with a stock. For example, if you know that a stock historically has a high positive autocorrelation value and you witnessed the stock making solid gains over the past several days, then you might reasonably expect the movements over the upcoming several days (the leading time series) to match those of the lagging time series and to move upward.

Autocorrelation is a characteristic of data in which the correlation between the values of the same variables is based on related objects. It violates the assumption of instance independence, which underlies most of the conventional models. It generally exists in those types of data-sets in which the data, instead of being randomly selected, is from the same source.

GLSM stands for Geometric Least Squares Mean. This is a mean estimated from a linear model. In contrast, a raw or arithmetic mean is a simple average of your values, using no model. Least squares means are adjusted for other terms in the model (like covariates), and are less sensitive to missing data. Theoretically, they are better estimates of the true population mean.

Autocorrelation Prediction (AP) has been shown to be an effective technique for Pole-Zero modeling. This paper develops a new linear method for identifying a stable Pole-Zero model whose spectrum matches the envelope of a given spectrum. All the operations are performed in Autocorrelation domain, using no Fourier transformations. At one extreme, Autocorrelation Prediction reduces to a linear method for all-Zero modeling. At the other extreme, AP becomes the well-known Linear Prediction (LP). AP can also automatically determine the lowest denominator and numerator orders required to model efficiently the given spectral envelope. Spectra whose envelopes have deep valleys are shown to be matched more accurately at the valleys using AP rather than LP.

Linear prediction is a mathematical operation where future values of a discrete-time signal are estimated as a linear function of previous samples.

In digital signal processing, linear prediction is often called linear predictive coding (LPC) and can thus be viewed as a subset of filter theory. In system analysis (a subfield of mathematics), linear prediction can be viewed as a part of mathematical modelling or optimization.

Heteroscedasticity is a hard word to pronounce, but it doesn't need to be a difficult concept to understand. Heteroscedasticity refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

In statistics, a collection of random variables is heteroscedastic (or heteroskedastic;[a] from Ancient Greek hetero “different” and skedasis “dispersion”) if there are sub-populations that have different variabilities from others. Here "variability" could be quantified by the variance or any other measure of statistical dispersion. Thus heteroscedasticity is the absence of homoscedasticity.

Multicollinearity is a state of very high intercorrelations or inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable.

Notes

4.26 SELF-ASSESSMENT QUESTIONS

1. Discuss about the violations of the assumptions of the classical model.
2. What is Classical Linear Regression? Explain the Classical Linear Regression Assumptions.
3. Discuss Errors in Variables Consequence.
4. Explain methods of estimation-classical method of maximum likelihood.
5. What is Instrumental Variables Regression? Discuss various uses of Instrumental Variable.
6. Explain about Autocorrelation in Technical Analysis.
7. Give the meaning of Autocorrelation. Discuss Consequences of Autocorrelation.
8. What is Test for Autocorrelation? Explain Tests for Autocorrelation and Prediction of Autocorrelation.
9. What is Heteroscedasticity? Discuss about Heteroscedasticity Structures and Tests for Heteroscedasticity.
10. Discuss Remedial measures the methods of weighted least square.

*****

5

Lesson

DISTRIBUTIVE LAG MODELS

Objectives

The objectives of this lesson are to:

- Distributive Lag Models
- Lagged exogenous and endogenous methods
- Consequences of applying OLMS to lagged and generous model
- Estimation of distribution log models KOYCK's approach
- Adaptive expectation
- Use of instrumental variable
- Almon's Approach
- Simultaneous Equations Methods
- Structural form reduced form, final form
- Methods of Estimation
- Method of Indirect least squares 2 LS
- Method of instrumental variable MLIML, 3 SLS and FIMLM

Structure:

5.1 Distributive Lag Models

5.2 Lagged exogenous and endogenous methods

5.3 Consequences of applying OLMS to lagged and generous model

5.4 Estimation of distribution log models KOYCK's approach

5.5 Use of instrumental variable

5.6 Simultaneous Equations Methods

5.7 Methods of Estimation- Method of Indirect least squares 2 LS

5.8 Method of Indirect least squares 2 LS

5.9 Method of instrumental variable MLIML, 3 SLS and FIMLM

5.10 Summary

5.11 Self-Assessment Questions

Notes

5.1 DISTRIBUTIVE LAG MODELS

Distributed lag model is a model for time series data in which a regression equation is used to predict current values of a dependent variable based on both the current values of an explanatory variable and the lagged (past period) values of this explanatory variable.

In an infinite distributed lag model, an infinite number of lag weights need to be estimated; clearly this can be done only if some structure is assumed for the relation between the various lag weights, with the entire infinitude of them expressible in terms of a finite number of assumed underlying parameters. In a finite distributed lag model, the parameters could be directly estimated by ordinary least squares (assuming the number of data points sufficiently exceeds the number of lag weights); nevertheless, such estimation may give very imprecise results due to extreme multicollinearity among the various lagged values of the independent variable, so again it may be necessary to assume some structure for the relation between the various lag weights.

The concept of distributed lag models easily generalizes to the context of more than one right-side explanatory variable.

The simplest way to estimate parameters associated with distributed lags is by ordinary least squares, assuming a fixed maximum lag {\displaystyle p} p, assuming independently and identically distributed errors, and imposing no structure on the relationship of the coefficients of the lagged explanators with each other. However, multicollinearity among the lagged explanators often arises, leading to high variance of the coefficient estimates.

Structured Estimation

Structured distributed lag models come in two types: finite and infinite. Infinite distributed lags allow the value of the independent variable at a particular time to influence the dependent variable infinitely far into the future, or to put it another way, they allow the current value of the dependent variable to be influenced by values of the independent variable that occurred infinitely long ago; but beyond some lag length the effects taper off toward zero. Finite distributed lags allow for the independent variable at a particular time to influence the dependent variable for only a finite number of periods.

5.2 LAGGED EXOGENOUS AND ENDOGENOUS METHODS

The variables which are explained by the functioning of system and values of which are determined by the simultaneous interaction of the relations in the model are endogenous variables or jointly determined variables.

Exogenous variables (Predetermined variables)

The variables that contribute to provide explanations for the endogenous variables and values of which are determined from outside the model are exogenous variables or predetermined variables.

Exogenous variables help is explaining the variations in endogenous variables. It is customary to include past values of endogenous variables in the predetermined group. Since exogenous variables are predetermined, so they are independent of disturbance term in the model. They satisfy those assumptions which explanatory variables satisfy in the usual regression model. Exogenous variables influence the endogenous variables but are not themselves influenced by them. One variable which is endogenous for one model can be exogenous variable for the other model.

Note that in linear regression model, the explanatory variables influence study variable but not vice versa.

So relationship is one sided.

The classification of variables as endogenous and exogenous is important because a necessary condition for uniquely estimating all the parameters is that the number of endogenous variables is equal to the number of independent equations in the system. Moreover, the main distinction of predetermined variable in estimation of parameters is that they are uncorrelated with disturbance term in the equations in which they appear.

Simultaneous equation systems:

A model constitutes a system of simultaneous equations if all the relationships involved are needed for determining the value of at least one of the endogenous variables included in the model. This implies that at least one of the relationships includes more them one endogenous variable.

5.3 CONSEQUENCES OF APPLYING OLMS TO LAGGED AND GENEROUS MODEL

Ohm's law states that the current through a conductor between two points is directly proportional to the voltage across the two points. Introducing the constant of proportionality, the resistance, one arrives at the usual mathematical equation that describes this relationship:

{\displaystyle I={\frac {V}{R}},} I={\frac {V}{R}},

where I is the current through the conductor in units of amperes, V is the voltage measured across the conductor in units of volts, and R is the resistance of the conductor in units of ohms. More specifically, Ohm's law states that the R in this relation is constant, independent of the current. Ohm's law is an empirical relation which accurately describes the conductivity of the vast majority of electrically conductive materials over many orders of magnitude of current. However some materials do not obey Ohm's law, these are called non-ohmic.

The law was named after the German physicist Georg Ohm, who, in a treatise published in 1827, described measurements of applied voltage and current through simple electrical circuits containing various lengths of wire.

In physics, the term Ohm's law is also used to refer to various generalizations of the law; for example the vector form of the law used in electromagnetics and material science:

{\displaystyle \mathbf {J} =\sigma \mathbf {E} ,} \mathbf {J} =\sigma \mathbf {E} ,

where J is the current density at a given location in a resistive material, E is the electric field at that location, and s (sigma) is a material-dependent parameter called the conductivity. This reformulation of Ohm's law is due to Gustav Kirchhoff.

In January 1781, before Georg Ohm's work, Henry Cavendish experimented with Leyden jars and glass tubes of varying diameter and length filled with salt solution. He measured the current by noting how strong a shock he felt as he completed the circuit with his body. Cavendish wrote that the "velocity" (current) varied directly as the "degree of electrification" (voltage). He did not communicate his results to other scientists at the time and his results were unknown until Maxwell published them in 1879.

Notes

Francis Ronalds delineated “intensity” (voltage) and “quantity” (current) for the dry pile – a high voltage source – in 1814 using a gold-leaf electrometer. He found for a dry pile that the relationship between the two parameters was not proportional under certain meteorological conditions.

Ohm's law in Georg Ohm's lab book

Ohm did his work on resistance in the years 1825 and 1826, and published his results in 1827 as the book Die galvanische Kette, mathematisch bearbeitet ("The galvanic circuit investigated mathematically"). He drew considerable inspiration from Fourier's work on heat conduction in the theoretical explanation of his work. For experiments, he initially used voltaic piles, but later used a thermocouple as this provided a more stable voltage source in terms of internal resistance and constant voltage. He used a galvanometer to measure current, and knew that the voltage between the thermocouple terminals was proportional to the junction temperature. He then added test wires of varying length, diameter, and material to complete the circuit. He found that his data could be modeled through the equation

{\displaystyle x={\frac {a}{b+l}},} x={\frac {a}{b+l}},

where x was the reading from the galvanometer, l was the length of the test conductor, a depended on the thermocouple junction temperature, and b was a constant of the entire setup. From this, Ohm determined his law of proportionality and published his results.

Ohm's law was probably the most important of the early quantitative descriptions of the physics of electricity. We consider it almost obvious today. When Ohm first published his work, this was not the case; critics reacted to his treatment of the subject with hostility. They called his work a "web of naked fancies" and the German Minister of Education proclaimed that "a professor who preached such heresies was unworthy to teach science." The prevailing scientific philosophy in Germany at the time asserted that experiments need not be performed to develop an understanding of nature because nature is so well ordered, and that scientific truths may be deduced through reasoning alone. Also, Ohm's brother Martin, a mathematician, was battling the German educational system. These factors hindered the acceptance of Ohm's work, and his work did not become widely accepted until the 1840s. However, Ohm received recognition for his contributions to science well before he died.

In the 1850s, Ohm's law was known as such and was widely considered proved, and alternatives, such as "Barlow's law", were discredited, in terms of real applications to telegraph system design, as discussed by Samuel F. B. Morse in 1855.

The electron was discovered in 1897 by J. J. Thomson, and it was quickly realized that it is the particle (charge carrier) that carries electric currents in electric circuits. In 1900 the first (classical) model of electrical conduction, the Drude model, was proposed by Paul Drude, which finally gave a scientific explanation for Ohm's law. In this model, a solid conductor consists of a stationary lattice of atoms (ions), with conduction electrons moving randomly in it. A voltage across a conductor causes an electric field, which accelerates the electrons in the direction of the electric field, causing a drift of electrons which is the electric current. However the electrons collide with and scatter off of the atoms, which randomize their motion, thus converting the kinetic energy added to the electron by the field to heat (thermal energy). Using statistical distributions, it can be shown that the average drift velocity of the electrons and thus the current, is proportional to the electric field, and thus the voltage, over a wide range of voltages.

The development of quantum mechanics in the 1920s modified this picture somewhat, but in modern theories the average drift velocity of electrons can still be shown to be proportional to the electric field, thus deriving Ohm's law. In 1927 Arnold Sommerfeld applied the quantum Fermi-Dirac distribution of electron energies to the Drude model, resulting in the free electron model. A year later, Felix Bloch showed that electrons move in waves (Bloch waves) through a solid crystal lattice, so scattering off the lattice atoms as postulated in the Drude model is not a major process; the electrons scatter off impurity atoms and defects in the material. The final successor, the modern quantum band theory of solids, showed that the electrons in a solid cannot take on any energy as assumed in the Drude model but are restricted to energy bands, with gaps between them of energies that electrons are forbidden to have. The size of the band gap is a characteristic of a particular substance which has a great deal to do with its electrical resistivity, explaining why some substances are electrical conductors, some semiconductors, and some insulators.

While the old term for electrical conductance, the mho (the inverse of the resistance unit ohm), is still used, a new name, the siemens, was adopted in 1971, honoring Ernst Werner von Siemens. The siemens is preferred in formal papers.

In the 1920s, it was discovered that the current through a practical resistor actually has statistical fluctuations, which depend on temperature, even when voltage and resistance are exactly constant; this fluctuation, now known as Johnson–Nyquist noise, is due to the discrete nature of charge. This thermal effect implies that measurements of current and voltage that are taken over sufficiently short periods of time will yield ratios of V/I that fluctuate from the value of R implied by the time average or ensemble average of the measured current; Ohm's law remains correct for the average current, in the case of ordinary resistive materials.

Ohm's work long preceded Maxwell's equations and any understanding of frequency-dependent effects in AC circuits. Modern developments in electromagnetic theory and circuit theory do not contradict Ohm's law when they are evaluated within the appropriate limits.

Scope

Ohm's law is an empirical law, a generalization from many experiments that have shown that current is approximately proportional to electric field for most materials. It is less fundamental than Maxwell's equations and is not always obeyed. Any given material will break down under a strong-enough electric field, and some materials of interest in electrical engineering are "non-ohmic" under weak fields.

Ohm's law has been observed on a wide range of length scales. In the early 20th century, it was thought that Ohm's law would fail at the atomic scale, but experiments have not borne out this expectation. As of 2012, researchers have demonstrated that Ohm's law works for silicon wires as small as four atoms wide and one atom high.

Microscopic origins

The dependence of the current density on the applied electric field is essentially quantum mechanical in nature; (see Classical and quantum conductivity.) A qualitative description leading to Ohm's law can be based upon classical mechanics using the Drude model developed by Paul Drude in 1900.

The Drude model treats electrons (or other charge carriers) like pinballs bouncing among the ions that make up the structure of the material. Electrons will be accelerated in the opposite direction to the electric field by the average electric field at their location. With each collision, though, the electron is deflected in a random direction with a velocity that is much larger than the velocity gained by the electric field. The net result is that electrons take a zigzag path due to the collisions, but generally drift in a direction opposing the electric field.

The drift velocity then determines the electric current density and its relationship to E and is independent of the collisions. Drude calculated the average drift velocity from p = - eEt where p is the average momentum, - e is the charge of the electron and t is the average time between the collisions. Since both the momentum and the current density are proportional to the drift velocity, the current density becomes proportional to the applied electric field; this leads to Ohm's law.

Hydraulic Analogy

A hydraulic analogy is sometimes used to describe Ohm's law. Water pressure, measured by pascals (or PSI), is the analog of voltage because establishing a water pressure difference between two points along a (horizontal) pipe causes water to flow. Water flow rate, as in liters per second, is the analog of current, as in coulombs per second. Finally, flow restrictors such as apertures placed in pipes between points where the water pressure is measured are the analog of resistors. We say that the rate of water flow through an aperture restrictor is proportional to the difference in water pressure across the restrictor. Similarly, the rate of flow of electrical charge, that is, the electric current, through an electrical resistor is proportional to the difference in voltage measured across the resistor.

Flow and pressure variables can be calculated in fluid flow network with the use of the hydraulic ohm analogy. The method can be applied to both steady and transient flow situations. In the linear laminar flow region, Poiseuille's law describes the hydraulic resistance of a pipe, but in the turbulent flow region the pressure–flow relations become nonlinear. The hydraulic analogy to Ohm's law has been used, for example, to approximate blood flow through the circulatory system.

5.4 ESTIMATION OF DISTRIBUTION LOG MODELS KOYCK'S APPROACH

The geometric distributed lag model, after application of the so-called Koyck transformation, is often used to establish the dynamic link between sales and advertising. This year, the Koyck model celebrates its 50th anniversary.In this paper we focus on the econometrics of this popular model,and we show that this seemingly simple model is a little more complicated than we always tend to think. First, the Koyck transformation entails a parameter restriction, which should not be overlooked for efficiency reasons. Second, the t-statistic for the parameter for direct advertising effects has a non-standard distribution. We provide solutions to these two issues. For the monthly Lydia Pinkham data, it is shown that various practical decisions lead to very different conclusions.

Adaptive Expectation

Once a forecasting error is made by agents, due to a stochastic shock, they will be unable to correctly forecast the price level again even if the price level experiences no further shocks since they only ever incorporate part of their errors. The backward nature of expectation formulation and the resultant systematic errors made by agents (see Cobweb model) was unsatisfactory to economists such as John Muth, who was pivotal in the development of an alternative model of how expectations are formed, called rational expectations. This has largely replaced adaptive expectations in macroeconomic theory since its assumption of optimality of expectations is consistent with economic theory. However, it must be stressed that confronting adaptivity and rationality is not necessarily justified, in other words, there are situations in which following the adaptive scheme is a rational response.

Adaptive expectations were instrumental in the Phillips curve outlined by Milton Friedman. For Friedman, workers form adaptive expectations, so the government can easily surprise them through unexpected monetary policy changes. As agents are trapped by the money illusion, they are unable to correctly perceive price and wage dynamics, so, for Friedman, unemployment can always be reduced through monetary expansions. The result is an increasing level of inflation if the government chooses to fix unemployment at a low rate for an extended period of time. However, in this framework it is clear why and how adaptive expectations are problematic. Agents are arbitrarily supposed to ignore sources of information which, otherwise, would affect their expectations. For example, government announcements are such sources: agents are expected to modify their expectations and break with the former trends when changes in economic policy necessitate it. This is the reason why the theory of adaptive expectations is often regarded as a deviation from the rational tradition of economics.

5.5 USE OF INSTRUMENTAL VARIABLE

Instrumental Variables (IV) is a method of estimation that is widely used in many economic applications when correlation between the explanatory variables and the error term is suspected - for example, due to omitted variables, measurement error, or other sources of simultaneity bias

Almon’s Approach

The polynomial distributed lag model is based on the assumption that a distributed lag model: shows polynomial dependence between ßj and the number j, that can be expressed as: To apply the Almon's method, first define the number of lags q.

That post drew quite a number of email requests for more information about the Almon estimator, and how it fits into the overall scheme of things. In addition, Almon's approach to modelling distributed lags has been used very effectively more recently in the estimation of the so-called MIDAS model. The MIDAS model (developed by Eric Ghysels and his colleagues - e.g., see Ghysels et al., 2004) is designed to handle regression analysis using data with different observation frequencies. The acronym, "MIDAS", stands for "Mixed-Data Sampling". The MIDAS model can be implemented in R, for instance (e.g., see here), as well as in EViews. (I discussed this in this earlier post.)

Suppose that we want to estimate the coefficients of the following DL model:

yt = ß0 xt + ß1 xt-1 + ß2 xt-2 + ... + ßn x t-n + ut;   t = 1, 2, ..., T.   (1)

This is called a "finite" DL model if the value of n is finite.

We could add an intercept into the model, and/or add other regressors, but that won't alter the basic ideas in the following discussion. So let's keep the model as simple as possible. We'll presume that the error term, ut, satisfies all of the usual assumptions - but that can be relaxed too.

If the maximum lag length in the model, n, is much less than T, then we could just apply OLS to estimate the regression coefficients. However, even if this is feasible, in the sense that there are positive degrees of freedom, this may not be the smartest way in which to proceed. For most economic time-series, x, the successive lags of the variable are likely to be highly correlated with each other. Inevitably, this will result in quite severe multicollinearity.

In response, Shirley Almon (1965) suggested a pretty neat way of re-formulating the model prior to its estimation. She made use of Weierstrass's Approximation Theorem, which tells us (roughly) that: "Every continuous function defined on a closed interval [a, b] can be uniformly approximated, arbitrarily closely, by a polynomial function of finite degree, P."

Notice that the theorem doesn't tell us what the value of P will be. This presents a type of model-selection problem that we have to solve. The flip-side of this is that if we select a value for P, and get it wrong, then there will be model mis-specification issues that we have to face. In fact, we can re-cast these issues in terms of those associated with the incorrect imposition of linear restrictions on the parameters of our model.

5.6 SIMULTANEOUS EQUATIONS METHODS

Simultaneous equations and linear equations, after studying this section, you will be able to:

a) Solve simultaneous linear equations by substitution.

b) Solve simultaneous linear equations by elimination.

c) Solve simultaneous linear equations using straight line graphs.

If an equation has two unknowns, such as 2y + x = 20, it cannot have unique solutions. Two unknowns require two equations which are solved at the sametime (simultaneously) - but even then two equations involving two unknowns do not always give unique solutions.

Example:

Solve the two simultaneous equations:

$2y + x = 8$ [1]

$1 + y = 2x$ [2]

from [2] $y = 2x - 1$ .... subtract 1 from each side

Substituting this value for y into [1] gives:

$2(2x - 1) + x = 8$

$4x - 2 + x = 8$ .... expand the brackets

$5x - 2 = 8$ .... tidy up

$5x = 10$ .... Add 2 to each side

$x = 2$ .... By dividing both sides by 5 the value of x is found.

Substitute the value of x into $y = 2x - 1$ gives

$y = 4 - 1 = 3$

So $x = 2$ and $y = 3$

Note:

- It is a good idea to label each equation. It helps you explain what you are doing - and may gain you method marks.
- This value of x can be substituted into equation [1] or [2], or into the expression for y: y = 2x - 1.
- Choose the one that is easiest!
- As a check, substitute the values back into each of the two starting equations.
- The second method is called solution by elimination.

Note:

The method is not quite as hard as it first seems, but it helps if you know why it works.

It works because of two properties of equations:

- Multiplying (or dividing) the expression on each side by the same number does not alter the equation.
- Adding two equations produces another valid equation:

e.g. 2x = x + 10 (x = 10) and x - 3 = 7 (x also = 10).

Adding the equations gives 2x + x - 3 = x + 10 + 7 (x also = 10).

The object is to manipulate the two equations so that, when combined, either the x term or the y term is eliminated (hence the name) - the resulting equation with just one unknown can then be solved:

Here we will manipulate one of the equations so that when it is combined with the other equation either the x or y terms will drop out. In this example the x term will drop out giving a solution for y. This is then substituted into one of the otiginal equations.

Label your equations so you know which one your are working with at each stage.

Equation [1] is 2y + x = 8

Equation [2] is 1 + y = 2x

Rearrange one equation so it is similar to the other.

[2] y – 2x = -1

also 2 x [1] gives 4y + 2x = 16 which we call [3]

[2] y – 2x = -1

[3] 4y +2x = 16

[2] + [3] gives 5y = 15

so y = 3

substituting y = 3 into [1] gives 1 + (3) = 2x

so 2x = 4, giving x = 2 and y = 3

5.7 METHODS OF ESTIMATION- METHOD OF INDIRECT LEAST SQUARES 2 LS

Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured empirical data that has a random component. The parameters describe an underlying physical setting in such a way that their value affects the distribution of the measured data. An estimator attempts to approximate the unknown parameters using the measurements. When the data consist of multiple variables and one is estimating the relationship between them, estimation is known as regression analysis.

In estimation theory, two approaches are generally considered.

The probabilistic approach (described in this article) assumes that the measured data is random with probability distribution dependent on the parameters of interest

The set-membership approach assumes that the measured data vector belongs to a set which depends on the parameter vector.

For example, it is desired to estimate the proportion of a population of voters who will vote for a particular candidate. That proportion is the parameter sought; the estimate is based on a small random sample of voters. Alternatively, it is desired to estimate the probability of a voter voting for a particular candidate, based on some demographic features, such as age; this estimates a relationship, and thus is a regression question.

Or, for example, in radar the aim is to find the range of objects (airplanes, boats, etc.) by analyzing the two-way transit timing of received echoes of transmitted pulses. Since the reflected pulses are unavoidably embedded in electrical noise, their measured values are randomly distributed, so that the transit time must be estimated.

As another example, in electrical communication theory, the measurements which contain information regarding the parameters of interest are often associated with a noisy signal.

5.8 METHOD OF INDIRECT LEAST SQUARES 2 LS

Once we have confirmed that our model is identified we can proceed with the estimation of the parameters of the structural coefficients. In this section we will present two methods of estimation that can be used to estimate coefficients of a simultaneous equation system.

Indirect Least Squares (ILS)

When all the equations are exactly identified one can use the method of Indirect Least Square to estimate the coefficients of the structural equations. It is done by the following three steps:

1) Form the reduced form equations

2) Estimate the coefficients of the reduced form using OLS

3) Use the estimated coefficients of the reduced form to derive the structural coefficients.

Example: (ILS)

Consider the following simple macro economic model:

$$Y_t = C_t + I_t$$
$$C_t = B_0 + B_1 Y_t + U_t$$

This model has two endogenous variables (y and Ct) and one exogenous variable (it), and we would like to estimate the coefficients of the behavioral equation. Since one of the variables of the model is excluded from the consumption function it is identified according to the order condition. The two structural equations could be used to form the reduced form equations for consumption. If we do that we receive:

$$C_t = \pi_0 + \pi_1 l_t + V_t$$

The equations show how the reduced form coefficients are related to the structural coefficients. By using the estimated values of the reduced form coefficients we can solve for the structural coefficients. We have:

$$\pi_0 = \frac{B_0}{1 - B_1} \text{ and } \pi_1 = \frac{B_1}{1 - B_1}$$

The above equations can now be used to solve for B0 and B1. Since is an equation with only one unknown we solve for B1 first (remember that *n1* is an estimate and therefore a number in this expression). Once we receive the value of B1 we can use it in to solve for B0. Hence we receive:

$$B_0 = \frac{\pi_0}{1 - \pi_1} \text{ and } B_1 = \frac{\pi_1}{1 - \pi_1}$$

In order to determine the standard errors for Ƃ0 and Ƃ1 we can use linear approximations to their expression based on the standard errors and covariance of the reduced form estimated coefficients. It can be shown that the corresponding variance for Ƃ0 and Ƃ1 is:

$$V(B_0) \approx a^2\sigma_0^2 + b^2\sigma_1^2 + 2aba^2\sigma_{01}$$

$$V(B_0) \approx b^2\sigma_1^2$$

with a = - and Ƃ = - and where *cr0* is the variance of ж 0, *ax* the variance for *nx* and cr12

the covariance between *n0* and *nx*.

ILS will result in consistent estimates but will still be biased in small samples. When using larger systems with more variables and equations it is often burdensome to find the estimates, and in those cases the equations are often over identified, which means that ILS cannot be used. For that reason ILS is not used very often in practice. Instead a much more popular method called 2SLS is used.

5.9 METHOD OF INSTRUMENTAL VARIABLE MLIML, 3 SLS AND FIMLM

Instrumental variables (IVs) are used to control for confounding and measurement error in observational studies. They allow for the possibility of making causal inferences with observational data. Like propensity scores, Ivs can adjust for both observed and unobserved confounding effects. Other methods of adjusting for confounding effects, which include stratification, matching and multiple regression methods, can only adjust for observed confounders. Ivs have primarily been used in economics research, but have recently begun to appear in epidemiological studies.

Observational studies are often implemented as a substitute for or complement to clinical trials, although clinical trials are the gold standard for making causal inference. The main concern with using observational data to make causal inferences is that an individual may be more likely to receive a treatment because that individual has one or more co-morbid conditions. The outcome may be influenced by the fact that some individuals received the treatment because of their personal or health characteristics.

Notes

Z is referred to as the instrumental variable because it satisfies the following conditions: (i) Z has a casual effect on X; (ii) Z affects the outcome variable Y only through X (Z does not have a direct influence on Y which is referred to as the exclusion restriction); (iii) There is no confounding for the effect of Z on Y. There are two main criteria for defining an IV: (i) It causes variation in the treatment variable; (ii) It does not have a direct effect on the outcome variable, only indirectly through the treatment variable. A reliable implementation of an IV must satisfy these two criteria and utilize a sufficient sample size to allow for reasonable estimation of the treatment effect. If the first assumption is not satisfied, implying that the IV is associated with the outcome, then estimation of the IV effect may be biased. If the second assumption is not satisfied, implying that the IV does not affect the treatment variable then the random error will tend to have the same effect as the treatment.

Although I vs can control for confounding and measurement error in observational studies they have some limitations. We must be careful when dealing with many confounders and also if the correlation between the IV and the exposure variables is small. Both weak instruments and confounders produce large standard error which results in imprecise and biased results. Even when the two key assumptions are satisfied and the sample size is large, Ivs cannot be used as a substitute for the use of clinical trials to make causal inference, although they are often useful in answering questions that an observational study cannot. In general, instrumental variables are most suitable for studies in which there are only moderate to small confounding effects. They are least useful when there are strong confounding effects.

While a good deal of research in simultaneous equation models has been conducted to examine the small sample properties of coefficient estimators there has not been a corresponding interest in the properties of estimators for the associated variances. Kiviet and Phillips (2000) and explore the biases in variance estimators. This is done for the 2SLS and the MLIML estimators. The approximations to the bias are then used to develop less biased estimators whose properties are examined and compared in a number of simulation experiments. In addition, a bootstrap estimator is included which is found to perform especially well. The experiments also consider coverage probabilities/test sizes and test powers of the t-tests where it is shown that tests based on 2SLS are generally oversized while test sizes based on MLIML are closer to nominal levels. In both cases test statistics based on the corrected variance estimates generally have a higher power than standard procedures.

The three-stage least squares estimator was introduced by Zellner & Theil (1962). It can be seen as a special case of multi-equation GMM where the set of instrumental variables is common to all equations. If all regressors are in fact predetermined, then 3SLS reduces to seemingly unrelated regressions (SUR). Thus it may also be seen as a combination of two-stage least squares (2SLS) with SUR.

Across fields and disciplines simultaneous equation models are applied to various observational phenomena. These equations are applied when phenomena are assumed to be reciprocally causal. The classic example is supply and demand in economics. In other disciplines there are examples such as candidate evaluations and party identification or public opinion and social policy in political science; road investment and travel demand in geography; and educational attainment and parenthood entry in sociology or demography. The simultaneous equation model requires a theory of reciprocal causality that includes special features if the causal effects are to be estimated as simultaneous feedback as opposed to one-sided 'blocks' of an equation where a researcher is interested in the causal effect of X on Y while holding the causal effect of Y on X constant or when the researcher knows the exact amount of time it takes for each causal effect to take place, i.e., the length of the causal lags. Instead of lagged effects, simultaneous feedback means estimating the simultaneous and perpetual impact of X and Y on each other. This requires a theory that causal effects are simultaneous in time, or so complex that they appear to behave simultaneously; a common example are the moods of roommates. To estimate simultaneous feedback models a theory of equilibrium is also necessary that X and Y are in relatively steady states or are part of a system (society, market, classroom) that is in a relatively stable state.

5.10 SUMMARY

Distributed lag model is a model for time series data in which a regression equation is used to predict current values of a dependent variable based on both the current values of an explanatory variable and the lagged (past period) values of this explanatory variable.

In an infinite distributed lag model, an infinite number of lag weights need to be estimated; clearly this can be done only if some structure is assumed for the relation between the various lag weights, with the entire infinitude of them expressible in terms of a finite number of assumed underlying parameters.

Structured distributed lag models come in two types: finite and infinite. Infinite distributed lags allow the value of the independent variable at a particular time to influence the dependent variable infinitely far into the future, or to put it another way, they allow the current value of the dependent variable to be influenced by values of the independent variable that occurred infinitely long ago; but beyond some lag length the effects taper off toward zero. Finite distributed lags allow for the independent variable at a particular time to influence the dependent variable for only a finite number of periods.

The variables which are explained by the functioning of system and values of which are determined by the simultaneous interaction of the relations in the model are endogenous variables or jointly determined variables.

The variables that contribute to provide explanations for the endogenous variables and values of which are determined from outside the model are exogenous variables or predetermined variables.

Exogenous variables help is explaining the variations in endogenous variables. It is customary to include past values of endogenous variables in the predetermined group. Since exogenous variables are predetermined, so they are independent of disturbance term in the model. They satisfy those assumptions which explanatory variables satisfy in the usual regression model. Exogenous variables influence the endogenous variables but are not themselves influenced by them. One variable which is endogenous for one model can be exogenous variable for the other model.

Ohm's law is an empirical law, a generalization from many experiments that have shown that current is approximately proportional to electric field for most materials. It is less fundamental than Maxwell's equations and is not always obeyed. Any given material will break down under a strong-enough electric field, and some materials of interest in electrical engineering are "non-ohmic" under weak fields.

A hydraulic analogy is sometimes used to describe Ohm's law. Water pressure, measured by pascals (or PSI), is the analog of voltage because establishing a water pressure difference between two points along a (horizontal) pipe causes water to flow. Water flow rate, as in liters per second, is the analog of current, as in coulombs per second. Finally, flow restrictors such as apertures placed in pipes between points where the water pressure is measured are the analog of resistors. We say that the rate of water flow through an aperture restrictor is proportional to the difference in water pressure across the restrictor. Similarly, the rate of flow of electrical charge, that is, the electric current, through an electrical resistor is proportional to the difference in voltage measured across the resistor.

Instrumental Variables (IV) is a method of estimation that is widely used in many economic applications when correlation between the explanatory variables and the error term is suspected for example, due to omitted variables, measurement error, or other sources of simultaneity bias

The polynomial distributed lag model is based on the assumption that a distributed lag model: shows polynomial dependence between ßj and the number j, that can be expressed as: To apply the Almon's method, first define the number of lags q.

Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured empirical data that has a random component. The parameters describe an underlying physical setting in such a way that their value affects the distribution of the measured data. An estimator attempts to approximate the unknown parameters using the measurements. When the data consist of multiple variables and one is estimating the relationship between them, estimation is known as regression analysis.

Observational studies are often implemented as a substitute for or complement to clinical trials, although clinical trials are the gold standard for making causal inference. The main concern with using observational data to make causal inferences is that an individual may be more likely to receive a treatment because that individual has one or more co-morbid conditions. The outcome may be influenced by the fact that some individuals received the treatment because of their personal or health characteristics.

5.12 SELF-ASSESSMENT QUESTIONS

1. What is Lag Model and Distributive Lag Model? Explain in details about Distributive Lag Models.
2. What is Lagged exogenous model? Discuss Lagged exogenous and endogenous methods.
3. What is OLMS to lagged and generous model? Explain about Consequences of applying OLMS to lagged and generous model.
4. Discuss about Estimation of distribution log models KOYCK's approach.
5. What is instrumental variable? Explain about Use of instrumental variable.
6. What is Simultaneous Equations Method? Discuss about Simultaneous Equations Methods.
7. Explain about Methods of Estimation.
8. Discuss about Method of Indirect least squares 2 LS.
9. Explain Method of instrumental variable MLIML, 3 SLS and FIMLM.

******