Directorate of Distance and Continuing Education

Manonmaniam Sundaranar University

Tirunelveli – 627 012, Tamil Nadu.

B.A. ECONOMICS

(Third Year)

Basic Econometrics

(JEEC61)

Compiled by

Dr. T. Shammy

Assistant Professor of Economics

Manonmaniam Sundaranar University

Tirunelveli – 627 012.

Semester	Course	Title of the Course	Course Code	Credits
VI	Elective – VII	Basic Econometrics	JEEC61	3

SYLLABUS

Unit I: Introduction

Definition – Scope – Divisions – Objectives – Use of Econometrics – Econometrics and Mathematical Economics – Econometrics and Statistics – Methodology of Econometrics – Basic ideas of Linear Regression Model – Two Variable Model, Population Regression Function (PRF), Sample Regression Function (SRF) – Error Term U – Significance – Stochastic form of PRF and SRF.

Unit II: Estimation

Classical Linear Regression Model – Assumptions – Method of Ordinary Least Squares (OLS) – Derivation of OLS Estimators – Derivation of variance and standard error of OLS estimators (Simple Linear Regression) – Properties of OLS estimators – Gauss-Markov Theorem – Proof – Three Variable Model Estimation (Basic Idea Only) – Hypothesis Testing (t and F test) – Test of Goodness of Fit. R^2 and Adjusted R^2.

Unit III: Violation of Assumptions

Multi-collinearity: Nature, Consequences, Detection and Remedial measures – Heteroscedasticity: Nature – Consequences, Detection and Remedial Measures – Autocorrelation: Nature, Consequences, Detection and Remedial Measures.

Unit IV: Functional Forms and Dummy Variables

Regression through the origin – Double Log Model – Measurement of Elasticity – Semi Log Model – Measurement of Growth.

Dummy Variables – ANOVA and ANCOVA Models – Dummy Variable Trap – Uses – Interaction Effects – Structural Changes – Autoregressive and Distributed Lag Model – Ad Hoc Method of Estimation – Koyck Transformation – Mean and Median Lag.

Unit V: Simultaneous Equation Model

Simultaneous Equation Model: Definition and Examples – Simultaneous Equation Bias – Structural and Reduced Form Equations – Identification – Rank and Order Condition – Indirect Least Square Estimation – Two Stage Least Square Estimation.

Unit I: Introduction

Definition – Scope – Divisions – Objectives – Use of Econometrics – Econometrics and Mathematical Economics – Econometrics and Statistics – Methodology of Econometrics – Basic ideas of Linear Regression Model – Two Variable Model, Population Regression Function (PRF), Sample Regression Function (SRF) – Error Term U – Significance – Stochastic form of PRF and SRF.

Introduction

Econometrics refers to the application of economic theory and statistical techniques for the purpose of testing hypothesis and estimating and forecasting economic phenomenon. Literally interpreted, econometrics means "economic measurement." **Prof. Ragnar Frisch**, a Norwegian economist and statistician first of all named this science as "Econometrics" in 1926. Econometrics emerged as an independent discipline studying economics phenomena. But it recognized and got attention after the world war. In 1931, the realization of the necessity of econometric work had become so evident, which made to form "Econometric Society". This International association includes practically all the worker in the field. The society published a periodical called "*Econometrica*" which disseminates the result of econometric research work. The electronic gadgets like computers have stimulated the utilization of econometrics in recent days.

Econometrics means economic measurement. Econometrics deals with the measurement of economic relationships. It's an amalgamation of economic theory with mathematics and statistics. It is a science which combines economic theory with economic statistics and tries by mathematical and statistical methods to investigate the empirical support of general economic law established by economic theory. The term econometrics is formed from two words of Greek origin, "*oukovouia"* meaning economy and "*uetpov"* meaning measure.

Definition

The book „Econometric Theory" was authored by **Arthur S Goldberger**, and defined econometrics in that book as "Econometrics may be defined as the social science in which the tools of economic theory, mathematics and statistical inference are applied to the analysis of economic phenomena".

Gerhard Tinbergen points out that "Econometrics, as a result of certain outlook on the role of economics, consists of application of mathematical statistics to economic data to

lend empirical support to the models constructed by mathematical economics and to obtain numerical results".

H Theil "Econometrics is concerned with the empirical determination of economic laws"

In the words **of Ragnar Frisch** "The mutual penetration of quantitative econometric theory and statistical observation is the essence of econometrics".

Thus, econometrics may be considered as the integration of economics, mathematics and statistics for the purpose of providing numerical values for the parameters of economic relationships and verifying economic theories. It is a special type of economic analysis and research in which the general economic theory, formulated in mathematical terms, is combined with empirical measurement of economic phenomena.

Scope

Econometrics is concerned with applying statistical and mathematical tools to economic data in order to give empirical content to economic theories. The **scope** of econometrics refers to the **range of areas, activities, and applications** where econometric techniques are used. It shows how econometrics connects economic theory with real-world data.

1. Testing Economic Theories

Econometrics helps verify whether an economic theory actually holds true in the real world.

Examples:

- Testing whether **demand falls when price rises**.
- Verifying the **Phillips Curve** (inflation vs unemployment).
- Checking if **money supply affects GDP growth**.

Econometrics thus serves as a "bridge" between **abstract theory** and **observed behaviour**.

2. Estimation of Economic Relationships

Econometrics quantifies economic relationships numerically.

Examples:

- Estimating the **price elasticity of demand**.
- Measuring the **marginal propensity to consume (MPC)**.
- Estimating the impact of **education on wages**.

This helps policymakers understand the strength and direction of economic influences.

3. Forecasting Economic Variables

Econometric models are crucial for predicting future economic outcomes.

Examples:

- Forecasting **GDP**, **inflation**, **unemployment**, **exports**, **interest rates**.
- Predicting **agricultural yields** or **tax revenue**.

Governments, businesses, and financial institutions rely heavily on econometric forecasts for planning.

4. Policy Formulation and Evaluation

Econometrics helps governments and institutions design, implement, and assess policies.

Examples:

- Evaluating whether **fiscal stimulus** increases GDP.
- Measuring the effect of **subsidies**, **scholarship schemes**, or **MGNREGA** outcomes.
- Assessing the impact of **GST** on tax revenue.

Thus, econometrics supports **evidence-based policymaking**.

5. Testing Economic Hypotheses (Statistical Inference)

Econometrics allows formal testing using statistical tools such as **t-test**, **F-test**, **chi-square test**.

Examples:

- Testing if the coefficient of price in a demand function is statistically significant.
- Testing whether returns on education are different for males and females.

It helps distinguish between random noise and real economic effects.

6. Studying Structural Changes

Econometrics identifies changes in relationships over time.

Examples:

- Whether the **consumption function** changed after liberalisation (post-1991).
- Whether **export responsiveness** changed after GST.
- Detecting **breaks** before and after economic crises or policy reforms.

Dummy variables and structural break tests are used here.

7. Dealing with Real-World Economic Problems

Real-world data often suffers from issues like: **Multicollinearity**, **Heteroscedasticity**, **Autocorrelation**, **Simultaneity** and **Measurement error.**

Econometrics studies these problems and gives solutions (remedial measures), making the conclusions reliable.

8. Building Empirical Models

Econometrics develops models that describe real-life behaviour and predict future trends.

These include: Consumption models, Investment functions, Demand & supply models, Production functions, Labour market models, Growth models and Financial econometric models. These are used in research, policy, and industry.

9. Business and Financial Applications

Econometrics has a wide scope in business decision-making:

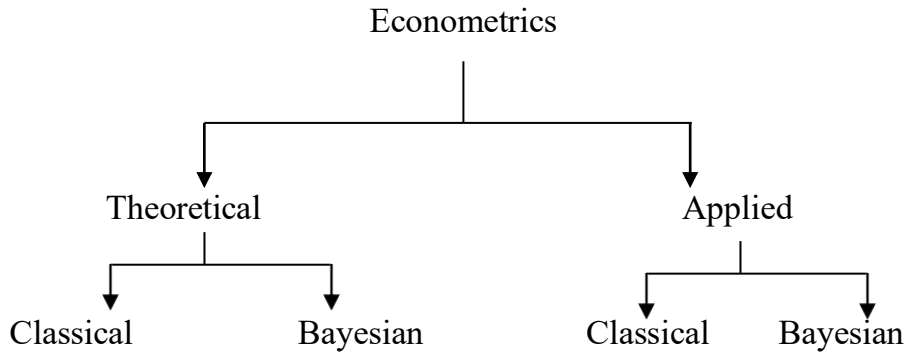
- Price forecasting for products
- Predicting sales, demand, and market share
- Stock market and financial return modelling
- Risk measurement (e.g., Value at Risk using time series models)

Conclusion

The scope of econometrics is extensive and growing. It covers **testing theories**, **estimating relationships**, **forecasting**, **policy evaluation**, and **solving real-world data problems**. By connecting theory with data, econometrics plays a central role in modern economic analysis and decision-making.

Divisions

Econometrics is divided into two broad categories as (a) Theoretical Econometrics and (b) Applied Econometrics.



Theoretical Econometrics is concerned with the development of appropriate methods for measuring economic relationship specified by econometric models. While Applied Econometrics use the tools of theoretical econometrics to study the field like economics and business specifically production function, investment function, demand and supply function etc.

Objectives

Econometrics aims to combine **economic theory**, **mathematics**, and **statistics** to understand real-world economic relationships. Its objectives show *why* econometrics is needed and *what* it tries to accomplish.

1. To Give Empirical Content to Economic Theory

Economic theory is often abstract and qualitative. Econometrics converts these theories into **quantifiable**, **measurable** relationships.

Examples:

- From the theory "Demand falls when price rises" → econometrics estimates how *much* demand changes when price rises.
- From consumption theory → it estimates MPC (Marginal Propensity to Consume).

2. To Estimate Economic Relationships

One of the primary objectives is to provide **numerical estimates** of economic parameters.

Examples:

- ✓ Price elasticity of demand

- ✓ Income elasticity of consumption
- ✓ Return to education
- ✓ Impact of interest rate on investment

This helps in understanding the strength of relationships among variables.

3. To Test Economic Hypotheses

Econometrics provides tools to statistically test whether theoretical assumptions hold true.

Examples:

- ➢ Does price actually influence demand significantly?
- ➢ Is there a long-run relationship between money supply and inflation?
- ➢ Does gender affect wage differences?

Tools used: **t-test**, **F-test**, **Chi-square test**, etc.

4. To Predict and Forecast Economic Variables

Econometric models are widely used to forecast future trends.

Examples:

- ✓ Predicting GDP growth
- ✓ Forecasting inflation or unemployment
- ✓ Projecting agricultural output
- ✓ Forecasting stock returns

These predictions help governments and businesses plan better.

5. To Evaluate Economic Policies

A key objective of econometrics is to determine whether a policy works.

Examples:

- Did GST increase tax revenue?
- Did subsidies raise agricultural productivity?
- Did PMJDY (Jan Dhan) improve financial inclusion?
- Is MGNREGA reducing rural unemployment?

Econometrics identifies the **causal effect** of policies.

6. To Provide a Basis for Decision-Making

Businesses, governments, banks, and industries use econometric results to make **data-based**

decisions, not assumptions.

Examples:

- Firms estimate demand before fixing prices.
- RBI uses econometric models to set interest rates.
- Government uses forecasts for budget planning.

7. To Analyse and Solve Real-World Data Problems

Economic data is imperfect—affected by noise, errors, non-linearity, and violations of assumptions. Econometrics aims to **detect**, **measure**, and **correct** these problems.

Examples: Multicollinearity**,** Heteroscedasticity**,** Autocorrelation**,** Simultaneity bias**,** Measurement error

This ensures results are reliable and scientifically valid.

8. To Build and Improve Econometric Models

Econometrics constantly works to develop new models for various economic phenomena.

Examples: Consumption models**,** Investment models**,** Labour market models**,** Growth models**,** Time-series models (ARIMA, VAR)**,** Panel data models.

This expands the scope of economics research.

Conclusion

The main objectives of econometrics are to **estimate economic relationships**, **test theories**, **forecast future outcomes**, and **evaluate policies** using real-world data. By combining theory with statistical evidence, econometrics helps make economics a **scientific discipline with measurable results**.

Use of Econometrics

Econometrics plays a central role in modern economic analysis because it connects **economic theory** with **real-world data**. Its uses go far beyond simple estimation—they include testing, forecasting, policy evaluation, and solving real-life economic problems.

1. To Quantify Economic Relationships

Econometrics converts qualitative economic theories into **numerical, measurable** relationships.

Examples:

- Estimating how much demand changes when price rises.
- Measuring the effect of education on income.
- Calculating the MPC in a consumption function.

This gives **precision** to economic theory.

2. To Test Economic Theories and Hypotheses

Econometrics provides statistical tools such as **t-test**, **F-test**, and **χ^2-test** to verify whether theoretical predictions hold in reality.

Examples:

- Testing the Phillips Curve.
- Testing if money supply significantly affects inflation.
- Testing whether savings depend on income.

Thus, econometrics gives **empirical validity** to theories.

3. To Forecast Economic Variables

One of the most important uses is **prediction**. Econometric models help forecast: GDP, Inflation, Unemployment, Agricultural output, Demand for products, Stock market trends. Etc. Forecasting helps governments and businesses plan future decisions.

4. For Policy Formulation and Evaluation

Econometrics helps evaluate the **impact of government policies** and design better interventions.

Examples:

- Did GST increase revenue?
- Did MNREGA reduce rural unemployment?
- Does a subsidy program raise agricultural productivity?
- Does monetary tightening reduce inflation?

Thus, econometrics supports **evidence-based policymaking**.

5. To Make Business and Financial Decisions

Businesses use econometric models for practical decision-making. Uses include: Sales

forecasting, Price optimization, Risk analysis, Market demand estimation, Credit scoring (in banks) and Stock price modelling. In financial markets, econometrics forms the base of **quantitative finance**.

6. To Understand Causal Relationships

Econometrics helps determine **cause-and-effect**, not just correlation.

Examples:

- Does education cause higher earnings?
- Does advertisement spending increase sales?
- Do interest rates affect investment?

Causality is essential for making informed decisions.

7. To Detect and Correct Data Problems

Real-world data suffers from many issues; econometrics helps identify and fix them. Problems include: Heteroscedasticity, Multicollinearity, Autocorrelation, Endogeneity, Simultaneity and Measurement error. Correcting these improves the **accuracy** and **reliability** of results.

8. For Building Economic Models

Econometrics is widely used to develop various macro and microeconomic models such as:

- Consumption and investment models
- Demand and supply models
- Growth models
- Labour market models
- Time-series models (ARIMA, VAR)
- Panel data models

These models explain behaviour and provide a framework for further research.

9. To Improve Decision-Making in Public and Private Sectors

Econometrics converts raw data into **actionable insights** for: Governments, Businesses, Financial institutions, International organisations and Research institutions. Better decisions lead to better outcomes.

Conclusion

The use of econometrics lies in estimating economic relationships, testing theories, forecasting future trends, evaluating policies, building models, and solving practical economic problems.

It makes economics **scientific, measurable, and applicable** to real-world decisions.

Econometrics and Mathematical Economics

Econometrics and **Mathematical Economics** are two important branches of economic analysis. Although they are related, each has a different focus and method.

Meaning of Mathematical Economics

Mathematical Economics uses **mathematical symbols, equations, and models** to express economic theories. It helps to state economic relationships clearly, for example:

- Demand function: $Q_d = f(P)$
- Production function: $Q = f(K, L)$
- Mathematics makes economic theories more precise, logical, and easier to manipulate.

Meaning of Econometrics

Econometrics uses **statistical and mathematical tools** to test, estimate, and verify economic theories using **real-world data**. It connects theory with actual evidence.

Example: If theory says demand falls when price rises, econometrics checks it using real data on price and quantity.

Relation Between Econometrics and Mathematical Economics

1. **Mathematical Economics provides the model**, on the other hand, Econometrics **estimates the model using data**.
2. Mathematical Economics gives **functional forms** (e.g., linear, quadratic). Where, Econometrics gives **numerical values** to the parameters (like elasticity, coefficients).
3. Mathematical Economics is **theoretical**, while, Econometrics is **empirical**.
4. Econometric models are often built using equations derived from Mathematical Economics.

Differences Between the Two

Mathematical Economics	**Econometrics**
Deals with symbolic and theoretical models.	Deals with estimation and testing of models.
Uses mathematical logic.	Uses statistics + mathematics + data.

Does not verify theories.	Verifies theories using real data.
Concerned with relationships between variables.	Measures the strength and direction of relationships.

Conclusion

Mathematical Economics provides the **theoretical backbone**, while Econometrics provides the **empirical validation**. Together, they make economics both logically strong and practically relevant.

Methodology of Econometrics

Broadly speaking, traditional or classical econometric methodology consists of the following steps.

1) Statement of the theory or hypothesis
2) Specification of the mathematical model of the theory
3) Specification of the econometric model of the theory
4) Obtaining the data
5) Estimation of the parameters of the econometric model
6) Hypothesis testing
7) Forecasting or prediction
8) Using the model for control or policy purposes.

1. Statement of theory or hypothesis

Keynes stated "the fundamental psychological law.....is that men (women) are disposed, as a rule and on average, to increase their consumption as their income increases, but not as much as the increase in their income". In short, Keynes postulated that the marginal propensity to consume (MPC), that is, the rate of change in consumption as a result of change in income, is greater than zero, but less than one. That is $0<MPC<1$.

2. Specification on the mathematical model of consumption

Mathematical model is specifying mathematical equations that describe the relationships between economic variables as proposed by the economic theory. Although Keynes postulated a positive relationship between consumption and income, he did not specify the precise form of functional relationship between the two. For simplicity, a mathematical economist might suggest the following form of the Keynesian consumption function:

$$\mathbf{Y_i = \beta_1 + \beta_2 X_i} \qquad \mathbf{0 < \beta_2 < 1} \qquad \mathbf{(1.1)}$$

Where Y_i = consumption expenditure, X_i= income and $\beta_1$and β_2, known as parameters of the model are intercept and slope coefficients respectively. The slope coefficient β_2 measures the MPC.

In the above equation (1.1), the variable appearing on the left side of the equality sign is called the dependent variable and the variables on the right side are called the independent or explanatory variables. Thus, in the Keynesian consumption function, consumption expenditure is the dependent variable and income is the explanatory variable.

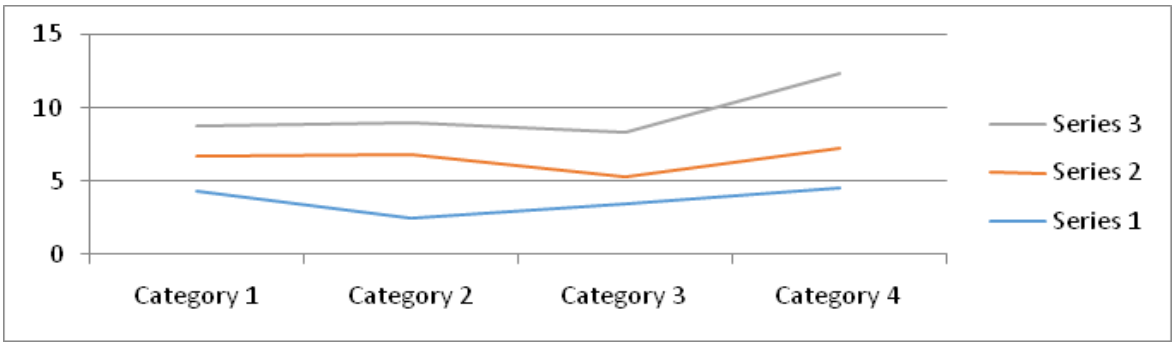
3.Specification of the econometric model of consumption

The purely mathematical model of the consumption function as in equation (1.1) is an exact or deterministic relationship between consumption and income. But relationships between economic variables are generally inexact. This is because of the fact that in addition to income other variables affect consumption expenditure. For example, size of family, ages of the members in the family, family religion etc are likely to exert some influence on consumption.

To allow for the in exact relationship between economic variables, the econometrician would modify the deterministic consumption function as follows

$$\mathbf{Y_i = \beta_1 + \beta_2 X_i + U_i} \qquad \mathbf{(1.2)}$$

Where U_i is known as the disturbance or error term, which is a random or stochastic variable. The disturbance term $U_i$represents all those factors that affect consumption but are not taken into account explicitly. This equation is an example of an econometric model. More technically, it is an example of linear regression model. The econometric consumption function hypotheses that the dependent variable Y (consumption) is linearly related to the explanatory variable X (income) but that the relationship between the two is not exact; it is subject to individual variation. The econometric model of consumption function is shown in the following diagram:



4.Obtaining Data

To estimate the econometric model given in equation (1.2), that is, to obtain the numerical values of $\beta_1$and β_2, one need data. Consumption expenditure, Income are collected from 5 respondents which are as follows

Income and expenditure of household in Mumbai

S.No	Income	Consumption Expenditure
1	3000	2500
2	3500	2750
3	2750	1800
4	6500	5400
5	1780	1470
4	4000	3700
5	2570	3500

5.Estimation of the econometric model

After the model has been specified and data has been collected, the econometrician must proceed with its estimation. The task is to estimate the parameters of the consumption function, that is, β_1 and β_2. The numerical estimates of the parameters gives empirical content to the consumption function. Choice of the appropriate econometric technique for the estimation of the function and critical examination of the assumptions of the chosen technique is a crucial step.

6.Hypothesis Testing

A hypothesis is a theoretical proposition that is capable of empirical verification or disproof. It may be viewed as an explanation of some event or events, and which may be true or false explanation. Confirmation or refutation of economic theories on the basis of sample evidence is based on a branch of statistical theory known as statistical inference or hypothesis testing. The rate of change in consumption as a result of change in income is greater than zero, but less than one. That is 0<MPC<1 will be the hypothesis.

7.Forecasting or Prediction

To predict the future values of the dependent or forecast variable Y, on the basis of known value or expected values of the explanatory, or predictor, variable X.

8. Use the model for control or policy purposes

Suppose the estimated Keynesian consumption function, and then the government can use it for control or policy purposes such as to determine the level of income that will guarantee the target amount of consumption expenditure. In other words, an estimated model may be used for control or policy purposes. By appropriate fiscal and monetary policy mix, the government can manipulate the control variable X to produce the desired level the target variable Y.

Basic ideas of Linear Regression Model

The **Linear Regression Model** is one of the most important tools in econometrics. It studies the relationship between a **dependent variable (Y)** and one or more **independent variables (X)**. It helps us understand how changes in X affect Y.

1. Meaning of Linear Regression

Linear regression explains how a dependent variable can be expressed as a **linear function** of one or more independent variables.

Example: $Y = a + bX + u$

- a = intercept
- b = slope (effect of X on Y)
- u = error term

2. Dependent and Independent Variables

- **Dependent variable (Y):** the variable we want to explain (e.g., income, demand).
- **Independent variable (X):** the variable that influences Y (e.g., education, price).

3. Intercept and Slope

- **Intercept (a):** value of Y when X = 0.
- **Slope (b):** how much Y changes when X increases by one unit.
  Example: If b=0.5, then a one-unit increase in X raises Y by 0.5 units.

4. Error Term (u)

The error term represents all other factors affecting Y that are **not included** in the model. It reflects: measurement errors, omitted variables and random shocks.

5. Simple vs. Multiple Regression

- **Simple Linear Regression:** one X variable.
  $Y = a + bX + u$
- **Multiple Linear Regression:** more than one X variable.
  $Y = a + b_1X_1 + b_2X_2 + ... + u$

6. Estimation Using OLS

The model is usually estimated using the **Ordinary Least Squares (OLS)** method. OLS chooses values of a and b such that the **sum of squared errors is minimum**. This gives the *best-fitting line* through the data.

7. Assumptions of Linear Regression (Basic)

Some basic assumptions include:

- ✓ Relationship between X and Y is linear
- ✓ Error term has zero mean
- ✓ Error term has constant variance

- ✓ No perfect multicollinearity (in multiple regression)

These ensure that OLS estimates are reliable.

8. Uses of Linear Regression

- Forecasting
- Measuring relationships between variables
- Testing economic theories
- Policy analysis

Conclusion

The **Linear Regression Model** provides a simple but powerful way to study economic relationships. It helps quantify the effect of one variable on another and is the foundation of modern econometric analysis.

Two Variable Model

Two variable or bivariate Means regression in which the dependent variable (the regressand) is related to a single explanatory variable (the regression).

When mean values depend upon conditioning (variable X) is called conditional expected value. Regression analysis is largely concerned with estimating and/or predicting the (population) mean value of the dependent variable on the basis of the known or fixed values of the explanatory variable (s).

To understand this, consider the data given in the below table. The data in the table refer to a total population of 60 families in a hypothetical community & their weekly income (X) and weekly consumption expenditure (Y), both in dollars.

The 60 families are divided into 10 income groups (from $80 to $260) and the weekly expenditures of each family in the various groups are as shown in the table. Therefore, we have 10 fixed values of X and the corresponding Y values against each of the X values; and hence there are 10 Y subpopulations. There is considerable variation in weekly consumption expenditure in each income group, which can be seen clearly but the general picture that one gets is that, despite the variability of weekly consumption expenditure within each income bracket, on the average, weekly consumption expenditure increases as income increases. To see this clearly, in the given table we have given the mean, or average, weekly consumption expenditure corresponding to each of the 10 levels of income. Thus, corresponding to the weekly income level of $80, the mean consumption expenditure is $65, while corresponding to the income level of $200, it is $137. In all we have 10 mean values for the 10 subpopulations

of *Y*. We call these mean values conditional expected values, as they depend on the given values of the (conditioning) variable *X*. Symbolically, we denote them as $E(Y \mid X)$, which is read as the expected value of *Y* given the value of *X*.

WEEKLY FAMILY INCOME *X*, $

Y ↓ \ *X* →	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure *Y*, $	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1043	966	1211
Conditional means of *Y*, $E(Y \mid X)$	65	77	89	101	113	125	137	149	161	173

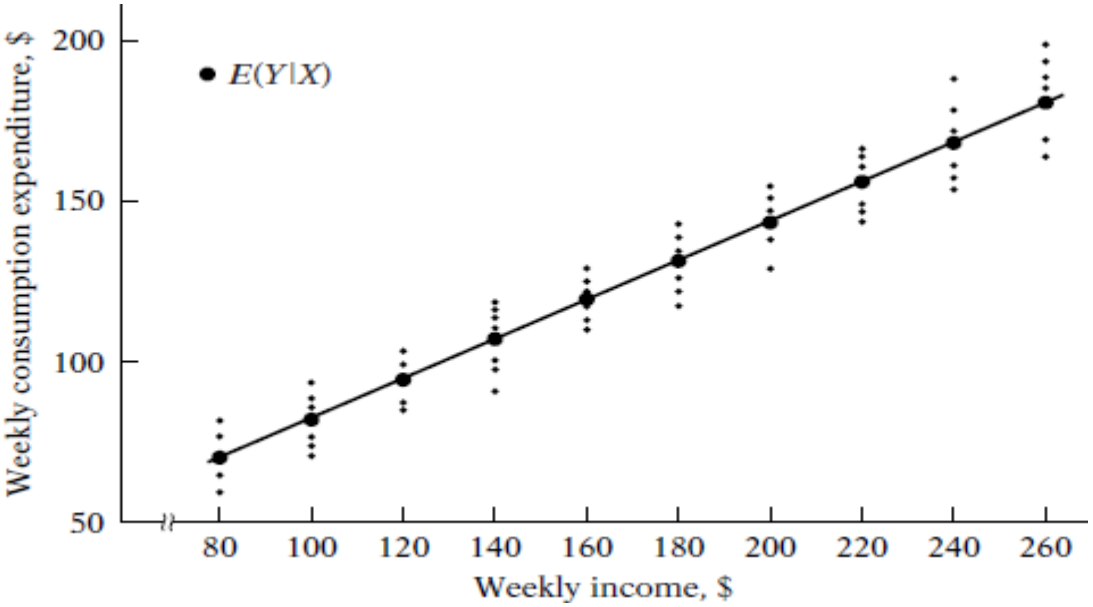


fig.: Conditional distribution of expenditure for various levels of income

It is important to distinguish these conditional expected values from the unconditional expected value of weekly consumption expenditure, $E(Y)$. If we add the weekly consumption expenditures for all the 60 families in the population and divide this number by 60, we get the number \$121.20 (\$7272/60), which is the unconditional

mean, or expected, value of weekly consumption expenditure, $E(Y)$; it is unconditional in the sense that in arriving at this number we have disregarded the income levels of the various families. Obviously, the various conditional expected values of Y given in given table are different from the unconditional expected value of Y of \$121.20. When we ask the question, "What is the expected value of weekly consumption expenditure of a family," we get the answer \$121.20 (the unconditional mean). But if we ask the question, "What is the expected value of weekly consumption expenditure of a family whose monthly income is, differently, if we ask the question, "What is the best (mean) prediction of weekly expenditure of families with a weekly income of \$140," the answer would be \$101. Thus the knowledge of the income level may enable us to better predict the mean value of consumption expenditure than if we do not have that knowledge.

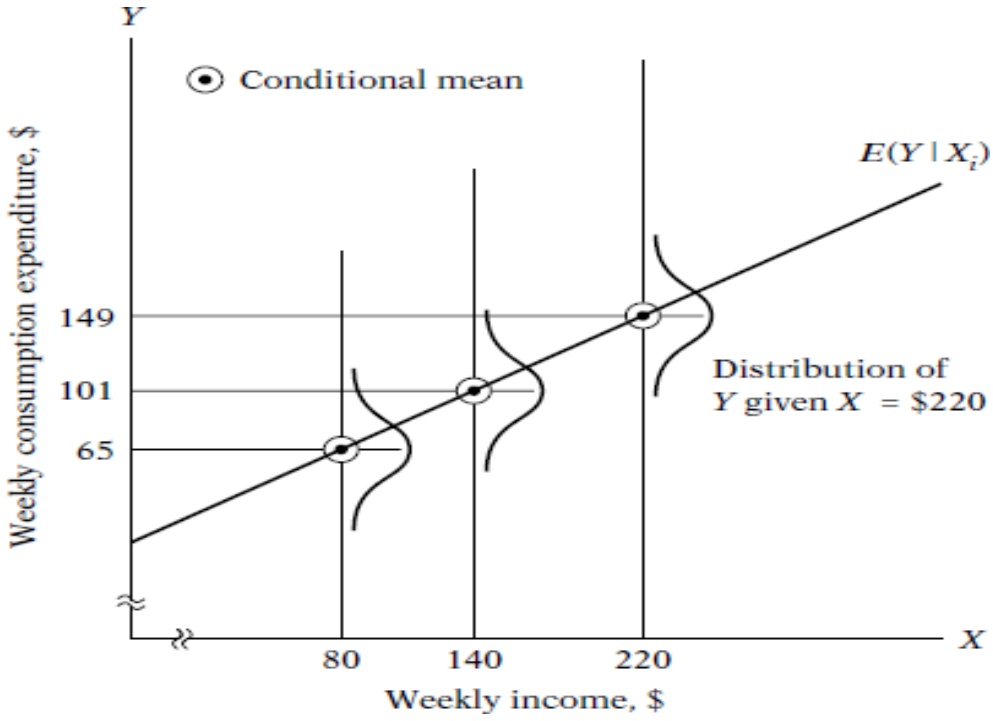


Fig.: Population Regression line.

This figure shows that for each X (i.e., income level) there is a population of Y values (weekly consumption expenditures) that are spread around the (conditional) mean of those Y values.

Population Regression Function (PRF)

The **Population Regression Function (PRF)** is a fundamental concept in econometrics. It shows the true or theoretical relationship between a dependent variable (Y) and one or more independent variables (X) at the **population level**.

1. Meaning of PRF

The PRF explains how the **average value of Y** changes with X in the entire population.

It is written as: $Y_i=\beta_0+\beta_1X_i+u_i$

Where:

- β_0 = population intercept
- β_1 = population slope
- u_i = error term (disturbance term)

These parameters represent the **true** values that we want to estimate.

2. Expected Value Form

Economists define PRF in terms of **expected value**:

$E(Y|X) = \beta0 + \beta_1X$

This means: Given a particular value of X, the **average** value of Y in the population is determined by a linear relationship.

3. Role of the Error Term

The error term u_i captures all factors that affect Y but are **not included** in the model, such as: individual differences, measurement errors, omitted variables, random shocks. Thus, PRF acknowledges that Y is not perfectly determined by X alone.

4. Characteristics of PRF

- PRF is **theoretical**, not directly observable.
- It describes the **true relationship** between variables.
- It uses **population parameters** $\beta0$ and $\beta1$, not estimates.
- It is assumed to be **linear in parameters**.

5. Purpose of PRF

The purpose of PRF is to provide a **benchmark** for understanding how Y behaves in the population.

We use **sample data** to estimate this function because we cannot observe the entire population.

6. PRF vs. Sample Regression Function (SRF)

- **PRF:** Refers to the *true* relationship using population parameters.
- **SRF:** Refers to the *estimated* relationship using sample data.

Thus, SRF attempts to approximate the PRF.

Conclusion

The Population Regression Function is a key concept in econometrics. It represents the true, average relationship between Y and X in the population. Although we cannot compute PRF directly, it guides us in estimating the relationship using sample data and regression techniques.

Sample Regression Function (SRF)

The SRF shows how Y **actually varies** with X in the given sample. It is obtained by applying a statistical method—usually **Ordinary Least Squares (OLS)**—to sample observations.

Mathematically: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Where:

- $\hat{\beta}_0$ = estimated intercept
- $\hat{\beta}_1$ = estimated slope
- $\hat{y}$ = predicted value of Y

These are not true (population) parameters but **estimates** based on the sample.

2. Purpose of SRF

- To **estimate** the true Population Regression Function (PRF).
- To measure the relationship between variables using sample data.
- To predict or forecast the value of Y for given values of X.

3. Derivation of SRF

The SRF is obtained by the **Ordinary Least Squares (OLS)** method. OLS chooses the line that **minimizes the sum of squared residuals**.

Residual: $e_i = y_i - \hat{y}_i$.

The OLS method ensures the best-fitting line for the given sample.

4. Properties of SRF

- It is **sample-specific** (varies from sample to sample).
- It gives **unbiased estimates** of the true parameters under CLRM assumptions.
- Residuals have zero mean.
- The SRF line is the line of "best fit" for the given sample.

Conclusion:

The Sample Regression Function is the **estimated** linear relationship derived from sample data. Because the true PRF is unknown, the SRF becomes essential in practical econometric work. It helps analyse, interpret, and predict economic behaviour.

Error Term U

In a regression model, the **Error Term (u)** represents all the factors that influence the dependent variable (Y) but are **not included** in the model. Since no economic model can include every possible variable, the error term captures the effect of all omitted or unobserved influences.

The regression model can be written as: $Y_i = \beta_0 + \beta_1 X_i + u_i$

Here, u_i explains the difference between the **actual value of Y** and the **value predicted** by the model.

Functions of the Error Term

1. **Captures omitted variables**: There may be many factors affecting Y that we do not include in the model. These are absorbed by u.
2. **Represents random shocks**: Unexpected events like weather changes, strikes, or sudden price changes are covered in u.
3. **Accounts for measurement errors**: If Y or X is measured inaccurately, the error term reflects this.
4. **Explains individual differences**: No two individuals or firms behave exactly alike; u accounts for this natural variation.

5. **Ensures realistic modelling**: Without u, the model would wrongly assume that X perfectly determines Y, which is unrealistic in economics.

Conclusion

The Error Term (u) is a vital part of the regression model. It captures all unobserved, omitted, or random factors that affect the dependent variable, making the regression model realistic and meaningful.

Significance

The **error term (u)** is an essential part of every econometric regression model. It ensures that the model remains realistic and statistically valid. Its significance can be explained as follows:

1. Captures All Unexplained Factors

In real life, many factors affect the dependent variable (Y). Since we cannot include all variables, the error term captures the influence of **omitted variables** such as taste, habits, weather, motivation, etc.

2. Makes the Model Realistic

A model without an error term assumes that X **perfectly determines** Y, which is impossible in economics.
The error term acknowledges that real-world data always has **randomness and uncertainty**.

3. Helps in Obtaining Unbiased Estimates

For OLS to give correct (unbiased) results, the error term must satisfy assumptions like:

- ✓ mean of u = 0
- ✓ no correlation between X and u.

These ensure that the estimated coefficients truly reflect the relationship between X and Y.

4. Represents Random Disturbances

Sudden events like strikes, rainfall changes, policy shocks, or measurement errors affect Y. These unpredictable factors are captured by u.

5. Basis for Statistical Testing

Hypothesis tests (t-test, F-test), confidence intervals, and model accuracy (R^2) all depend on the **variance of the error term**. If the error term behaves well, the model becomes statistically reliable.

Stochastic form of PRF and SRF

In econometrics, regression models are written in **stochastic (random)** form because economic variables are influenced by many random and unobservable factors. Therefore, both the **Population Regression Function (PRF)** and **Sample Regression Function (SRF)** include a **random error term**.

1. Stochastic Form of PRF

The **Population Regression Function (PRF)** explains the true relationship between the dependent variable (Y) and the independent variable (X) **in the entire population**.

Stochastic Form: $Y_i=\beta_0+\beta_1X_i+u_i$

Where:

- β_0,β_1 = true population parameters
- u_i= random error term
- Y_i= actual value of Y for the ith individual

Why "stochastic"?

Because PRF includes an error term representing all unobserved factors such as: omitted variables, measurement errors, random shocks, individual differences. Thus, the PRF shows **expected (average)** value of Y for a given X: $E(Y|X) = \beta_0 + \beta_1X$. This is the **true but unobservable** relationship.

2. Stochastic Form of SRF

The **Sample Regression Function (SRF)** is the estimated version of the PRF obtained from sample data.

Stochastic Form: $\widehat{\boldsymbol{y}}_{\boldsymbol{i}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\boldsymbol{x}_{\boldsymbol{i}}$

Here:

- $\hat{\beta}_0,\hat{\beta}_1$ = estimated coefficients
- $\hat{y}i$ = predicted value of Y
- Residual (error):$e_i = y_i - \hat{y}_i$

Unlike the PRF error term u_i , which is unknown, residuals e_i are **observable**.

Unit II: Estimation

Classical Linear Regression Model – Assumptions – Method of Ordinary Least Squares (OLS) – Derivation of OLS Estimators – Derivation of variance and standard error of OLS estimators (Simple Linear Regression) – Properties of OLS estimators – Gauss-Markov Theorem – Proof – Three Variable Model Estimation (Basic Idea Only) – Hypothesis Testing (t and F test) – Test of Goodness of Fit. R^2 and Adjusted R

Classical Linear Regression Model – Assumptions

The Classical Linear Regression Model (CLRM) is a fundamental framework in econometrics used to study the relationship between a dependent variable and one or more independent variables. It is based on the method of **Ordinary Least Squares (OLS)**, which provides estimates for the unknown parameters of the regression equation. CLRM helps in understanding how changes in explanatory variables affect the dependent variable and allows economists to make predictions, test hypotheses, and measure economic relationships.

To ensure that the OLS estimators are reliable, valid, and efficient, the CLRM operates under a set of key assumptions. When these assumptions hold true, the OLS estimators become **Best Linear Unbiased Estimators (BLUE)** according to the **Gauss–Markov Theorem**. Therefore, understanding the assumptions of CLRM is essential for accurate interpretation of regression results.

The Classical Linear Regression Model (CLRM) is based on several assumptions that ensure the **Ordinary Least Squares (OLS)** estimators are **Best Linear Unbiased Estimators (BLUE)**. The major assumptions are:

1. Linearity in Parameters

The regression model must be linear **in the coefficients**, not necessarily in variables. Example: $Y = \beta_0 + \beta_1 X + u$

Non-linear equations can also be transformed into linear form.

2. Fixed (Non-Random) Independent Variables

The values of the explanatory variables X are assumed to be fixed in repeated samples. This assumption simplifies analysis and helps in deriving properties of estimators.

3. Zero Mean of Error Term

The error term must have an expected value of zero. $E(u_i)=0$. This means the model is correctly specified; no systematic error exists.

4. Constant Variance (Homoscedasticity)

The variance of the error term is constant for all observations. $Var(u_i) = \sigma^2$. If variance differs across observations, it becomes **heteroscedastic**, violating CLRM.

5. No Autocorrelation

The error terms must be independent of each other. $Cov(u_i,u_j)=0$. This assumption is especially important for time-series data.

6. No Perfect Multicollinearity

The explanatory variables should not have perfect linear relationships among them. If one variable can be exactly predicted from another, OLS cannot be estimated.

7. Error Term is Normally Distributed (for Inference)

For hypothesis testing and constructing confidence intervals, the error term is assumed to follow a normal distribution. This ensures that t-tests and F-tests are valid.

8. Correct Model Specification

The model must include all relevant variables and exclude irrelevant ones. No measurement errors in variables. Functional form should be correct.

Method of Ordinary Least Squares (OLS) – Derivation of OLS Estimators

---

The Ordinary Least Squares (OLS) method is the most widely used technique in econometrics for estimating the parameters of a linear regression model. The main idea of OLS is to find the line that **best fits the data** by minimizing the **sum of the squared differences** between the actual values of the dependent variable and the predicted values. These differences are called **residuals**.

OLS is popular because it is simple, easy to compute, and—under classical assumptions—gives estimators that are **Best Linear Unbiased Estimators (BLUE)**. The OLS method allows economists to quantify relationships, make predictions, and conduct hypothesis testing.

Derivation of OLS Estimators

Consider the simple linear regression model:

$Y_i = \beta_0 + \beta_1 X_i + u_i$

Here:

Y_i = dependent variable

X_i = independent variable

β_0, β_1 = parameters to be estimated

u_i = error term

OLS aims to estimate β_0 and β_1 by minimizing the **Sum of Squared Residuals (SSR)**:

$SSR = \sum(Yi - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$

Step 1: Form the normal equations

Minimize SSR with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$:

$\frac{\partial SSR}{\partial \hat{\beta}_0} = 0$ and $\frac{\partial SSR}{\partial \hat{\beta}_1} = 0$

Solving the partial derivatives gives two **normal equations**:

1. $\sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i$
2. $\sum X_i Y_i = \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2$

Step 2: Solve for $\hat{\beta}_1$

Using algebra: $\hat{\beta}_1 = \frac{\sum(Xi - \bar{X})(Yi - \bar{Y})}{\sum(X_i - \bar{X})^2}$

This is the formula for the **slope estimator**.

Another equivalent form: $\hat{\beta}_1 = \frac{Cov(X,Y)}{\mathrm{Var(X)}}$

Step 3: Solve for $\hat{\beta}_0$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

This is the **intercept estimator**.

Final OLS Estimators

- **Slope estimator:** $\hat{\beta}_1 = \frac{\sum(Xi-\bar{X})(Yi-\bar{Y})}{\sum(X_i-\bar{X})^2}$
- **Intercept estimator:** $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$

These estimators minimize the sum of squared residuals and provide the best-fitting regression line.

Properties of OLS estimators

The Ordinary Least Squares (OLS) method is widely used because the estimators it produces have several desirable statistical properties. These properties are derived under the assumptions of the **Classical Linear Regression Model (CLRM)**. The important properties are:

1. Linearity

The OLS estimators are **linear functions** of the dependent variable Y. This means:

$\hat{\beta}_0, \hat{\beta}_1$ = =linear combinations of Y_i .

Because of this linearity, OLS becomes simple to compute and easy to interpret.

2. Unbiasedness

An estimator is unbiased if its expected value equals the true parameter value. Under CLRM assumptions: $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$

This means OLS estimators, on average, hit the true population values—they do not systematically overestimate or underestimate.

3. Minimum Variance (Efficiency)

Among all **linear and unbiased** estimators, OLS estimators have the **lowest variance**. This property is guaranteed by the **Gauss–Markov Theorem**. Thus, OLS estimators are **BLUE**: **B**est, **L**inear, **U**nbiased, **E**stimators. They are "best" because they provide the most precise (minimum variance) estimates.

4. Consistency

As the sample size increases: $\hat{\beta}_0 \rightarrow \beta 0$, $\hat{\beta}_1 \rightarrow \beta 1$

OLS estimators converge to the true parameter values. This means OLS remains reliable even in large samples, even if some assumptions (like normality) are relaxed.

5. Normality

If the error term u is normally distributed, then: $\hat{\beta}_0$ and $\hat{\beta}_1$ are also normally distributed. This property allows us to conduct **t-tests**, **F-tests**, and construct **confidence intervals**, which are essential for hypothesis testing.

6. Sufficient Statistics (in simple regression)

In simple linear regression, the OLS estimators capture all relevant sample information needed to estimate the model. So, $\hat{\beta}_0$ and $\hat{\beta}_1$ are **sufficient statistics** for β_0 and β_1 .

Gauss-Markov Theorem – Proof

The Gauss–Markov Theorem states that under the assumptions of the Classical Linear Regression Model—such as linearity of the model, zero mean of error term, constant variance of errors, absence of autocorrelation, and no perfect multicollinearity—the Ordinary Least Squares (OLS) estimators are the Best Linear Unbiased Estimators (BLUE). The proof is based on three ideas. First, OLS estimators are *linear* because they are obtained as linear combinations of the observed values of the dependent variable. Second, they are *unbiased* since, under the classical assumptions, their expected values are equal to the true population parameters. Finally, among all linear and unbiased estimators, OLS has the *minimum variance*, meaning it provides the most precise estimates. Any alternative linear unbiased estimator will always have a variance equal to or higher than that of OLS. Therefore, OLS satisfies the conditions of being linear, unbiased, and most efficient, and hence is termed the Best Linear Unbiased Estimator, which completes the proof of the Gauss–Markov Theorem.

Three Variable Model Estimation (Basic Idea Only)

A **three-variable model** is a multiple regression model with one dependent variable and **two independent variables**. The general form is: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

Here:

Y = dependent variable

X_1, X_2 = explanatory variables

β_0, β_1, β_2 = parameters

u = error term

Basic Idea

The purpose of estimating this model is to study how **each independent variable affects the dependent variable while holding the other variable constant** (ceteris paribus effect).

OLS (Ordinary Least Squares) is used to estimate the coefficients. The computer/software or manual calculation provides estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

Interpretation:

- $\hat{\beta}_1$: change in Y due to one-unit change in X_1, keeping X_2 constant.
- $\hat{\beta}_2$: change in Y due to one-unit change in X_2, keeping X_1 constant.

The three-variable model helps avoid **omitted variable bias** and gives more realistic results than simple regression.

Hypothesis Testing (t and F test)

After estimating coefficients, we test whether the independent variables significantly influence the dependent variable. We use:

✓ **t-test** → to test individual coefficients
✓ **F-test** → to test overall significance of the model

(a) t-Test (Individual Significance Test)

The **t-test** examines whether a particular coefficient (e.g., β_1) is statistically significant.

Purpose

To check if an **individual explanatory variable** has a meaningful effect on Y.

Null Hypothesis (H_0)

H0: $\beta_i = 0$

Meaning the variable has no effect.

Decision Rule

If calculated t > table t, then we have to reject H_0, hence the variable is significant.

Interpretation

If significant → that variable contributes to explaining changes in Y.
If not significant → variable does not have a strong independent influence.

(b) F-Test (Overall Significance Test)

The F-test checks whether the **entire regression model** is statistically meaningful.

Purpose

To test whether *all independent variables together* explain variations in the dependent variable.

Null Hypothesis (H_0)

H0: $\beta_1 = \beta_2 = 0$

Meaning the model has no explanatory power.

Decision Rule

If calculated F > table F → reject H_0. Model is statistically significant.

Interpretation

A significant F-value means the set of variables X_1 and X_2 together significantly explain Y.

Test of Goodness of Fit

The **Goodness of Fit** tells how well the regression line fits the data. The commonly used measure is:

(a) R-Squared (R^2)

Measures the proportion of variation in the dependent variable explained by the independent variables. It ranges from **0 to 1**. Higher R^2= better fit.

Example: If $R^2 = 0.70$, then 70% of the variation in Y is explained by $X_1$and X_2.

(b) Adjusted R-Squared

Used in multiple regression because it adjusts for the number of variables. It increases only if the added variable improves the model. A better measure than simple R-square for three-variable or multi-variable models.

Unit III: Violation of Assumptions

Multi-collinearity: Nature, Consequences, Detection and Remedial measures – Heteroscedasticity: Nature – Consequences, Detection and Remedial Measures – Autocorrelation: Nature, Consequences, Detection and Remedial Measures.

Multi-collinearity

---

1. Nature of Multicollinearity

Multicollinearity occurs in a multiple regression model when **two or more independent variables are highly correlated** with each other. This means they move together and provide overlapping information about the dependent variable.

Examples: Income and consumption expenditure, Height and weight, Education and experience

Because of multicollinearity, it becomes difficult for the model to separate the individual effect of each variable. Multicollinearity does **not** violate any CLRM assumptions completely, but it creates practical problems in estimation and interpretation.

2. Consequences of Multicollinearity

a) Coefficient estimates become unstable

Small changes in data can lead to very different estimated values of coefficients.

b) High standard errors

Standard errors of the affected coefficients become large, making t-values small and coefficients insignificant.

c) Difficulty in judging the importance of variables

Individually, the variables may appear insignificant (due to high standard errors), but jointly they may be very important.

d) Signs of coefficients may become illogical

Positive relationships may appear negative or vice versa.

e) Reduced precision of estimates

OLS estimates remain unbiased but become less precise.

f) Model interpretation becomes difficult

It becomes hard to understand the individual contribution of each variable.

3. Detection of Multicollinearity

a) High correlation between independent variables

If two regressors have a correlation above 0.8 or 0.9, multicollinearity is suspected.

b) High R^2 but low t-values

Overall model is significant (high R^2, high F-value), but individual coefficients are not significant. This is the sign of multicollinearity.

c) Variance Inflation Factor (VIF)

VIF values greater than 10 indicate serious multicollinearity.

d) Large standard errors

Coefficients have unusually large standard errors.

e) Condition index

High condition index (> 30) suggests multicollinearity.

f) Coefficients change drastically when adding or removing a variable

Instability of coefficients indicates high interdependence.

4. Remedial Measures for Multicollinearity

a) Drop one of the highly correlated variables

If two variables provide similar information, remove one to reduce overlap.

b) Combine correlated variables

Create an index or composite variable (example: socio-economic index).

c) Increase sample size

More data may reduce the degree of multicollinearity.

d) Transform variables

Using logarithms or ratios sometimes reduces correlation.

e) Use Principal Component Analysis (PCA) or Ridge Regression

These advanced methods reduce multicollinearity by adjusting or combining variables.

f) Accept multicollinearity if the purpose is prediction

If interpretation is not required, and the model predicts well, the problem may be ignored.

Conclusion

Multicollinearity refers to high correlation among independent variables in a regression model. It does not bias the OLS estimates but makes them unstable and imprecise. Detecting it through correlations, VIF, or unexpected t-values is essential. Remedies include removing variables, combining them, transforming data, or using specialized estimation techniques.

Heteroscedasticity

---

1. Meaning / Nature

In the **Classical Linear Regression Model (CLRM)**, one key assumption is that **$\mathbf{Var(u_i) = \sigma^2}$ (constant for all observations).** This is called **homoscedasticity**.

Heteroscedasticity means: **$\mathbf{Var(u_i) \neq constant}$**. The variance of the error term **changes** across observations.

In simple words: The scatter of residuals is **not uniform** — sometimes wide, sometimes narrow.

It commonly occurs when: Income data varies widely, Cross-section data contain rich and poor groups, Firms of different sizes are in the same sample, Growth rate differs heavily across observations.

2. Consequences of Heteroscedasticity

Even though OLS estimates remain **unbiased**, heteroscedasticity causes several problems:

1. OLS becomes inefficient

OLS is no longer the **Best Linear Unbiased Estimator (BLUE)**. It does not use minimum variance among unbiased estimators.

2. Standard errors become wrong

The estimated standard errors are **biased**. This leads to misleading test statistics.

3. t-tests and F-tests become unreliable

Because of wrong standard errors, **t, F, and χ^2 tests cannot be trusted**. You may accept a false hypothesis or reject a true one.

4. Confidence intervals become invalid

They may become too wide or too narrow.

5. Model interpretation becomes misleading

Coefficients may look statistically insignificant when they are actually significant, and vice versa.

3. Detection of Heteroscedasticity

A. Graphical Methods

Residual Plot ($\hat{u}$ vs X or $\hat{u}$ vs $\hat{Y}$): If the scatter is **funnel-shaped**, curved, or uneven, heteroscedasticity exists.

B. Formal Statistical Tests

1. Breusch–Pagan Test

Regress squared residuals on explanatory variables. Significant result → heteroscedasticity.

2. White Test

General test; does not require specifying the form of heteroscedasticity.

3. Goldfeld–Quandt Test

Split sample into two groups. Then compare variances. Useful when heteroscedasticity is suspected to increase with a variable.

4. Park Test / Glejser Test

Regress $|\hat{u}|$ or $\hat{u}^2$ on suspected variables (e.g., Y, X). Significant results → heteroscedasticity.

4. Remedial Measures

A. Transformations

1. Log Transformation

Convert variables to logarithmic form. Example: log(Y), log(X). This often stabilizes variance.

2. Weighted Least Squares (WLS)

If the form of heteroscedasticity is known, give **less weight** to observations with high variance and **more weight** to those with low variance. WLS becomes BLUE under heteroscedasticity.

B. Robust Standard Errors

1. White's Heteroscedasticity-Consistent Standard Errors

Keep OLS coefficients but **correct standard errors**. Useful when heteroscedasticity's structure is unknown.

C. Model Specification

Sometimes heteroscedasticity happens due to **omitted variables** or incorrect functional form. Correcting model specification may solve the issue.

D. Data Cleaning

Remove extreme outliers that cause variance instability.

Autocorrelation

---

1. Meaning / Nature

In CLRM, one assumption is: **$Cov(u_i , u_j) = 0$ for all $i \neq j$** means the error terms must be **independent**.

Autocorrelation means: *Error terms are correlated with each other.*

Common in **time-series data** such as GDP, prices, sales, rainfall, inflation, stock prices. The causes of it is because of Persistence in economic variables, Wrong functional form, Omitted variables, Measurement errors and Lagged dependent variable models.

2. Consequences of Autocorrelation

1. OLS estimates remain unbiased: Coefficients (β's) are still **correct on average**.

2. But OLS becomes inefficient: OLS is no longer **BLUE**, and there exist better estimators with smaller variance.

3. Standard errors become biased: Leads to misleading t-statistics.

4. t-tests and F-tests become unreliable: Hypothesis testing becomes invalid. Also, you may wrongly reject or accept hypotheses.

5. Confidence intervals are misleading: Too narrow or too wide.

6. R^2 becomes exaggerated: Model may look "good" when it is not.

3. Detection of Autocorrelation

A. Graphical Method

- ✓ Plot residuals against time.
- ✓ If residuals show a **pattern** (waves, cycles, trend), autocorrelation is present.

B. Formal Tests

1. Durbin–Watson (DW) Test

Most widely used. Where,

- DW value ≈ **2** → No autocorrelation
- DW < **2** → Positive autocorrelation
- DW > **2** → Negative autocorrelation

2. Breusch–Godfrey (BG) Test

Used for higher-order autocorrelation (AR(1), AR(2), AR(p)).

3. Runs Test

- Checks if residuals occur in runs (groups).
- Too many or too few runs → autocorrelation.

4. Correlogram (ACF/PACF)

Shows pattern in correlation across time lags.

4. Remedial Measures

A. Transformations

1. First Difference Transformation

- Convert data to $\Delta Y_t = Y_t - Y_{t-1}$
- Useful when autocorrelation comes from trend or persistence.

B. Change Model Specification

- Include omitted variables.
- Add lagged variables if needed.

C. Generalized Least Squares (GLS)

If the structure of autocorrelation is known (e.g., AR(1)), GLS or **Cochrane–Orcutt Method** is used.

D. Newey–West Standard Errors

Keep OLS estimates but adjust standard errors for autocorrelation (and heteroscedasticity).

Unit IV: Functional Forms and Dummy Variables

Regression through the origin – Double Log Model – Measurement of Elasticity – Semi Log Model – Measurement of Growth.

Dummy Variables – ANOVA and ANCOVA Models – Dummy Variable Trap – Uses – Interaction Effects – Structural Changes – Autoregressive and Distributed Lag Model – Ad Hoc Method of Estimation – Koyck Transformation – Mean and Median Lag.

Regression through the origin

In a normal regression model, we estimate: $\mathbf{Y = \beta_0 + \beta_1 X + u}$, where β_0 is the intercept.

But in some situations, theory tells us that **when X = 0, Y must also be 0**.
Examples:

- Cost is zero when output is zero (in some ideal cases).
- Distance travelled is zero when speed is zero.
- Consumption of a good is zero when income is zero (in some models).

In such cases, we **force the regression line to pass through the origin** (0,0).

So we estimate the model: $\mathbf{Y = \beta_1 X + u}$ (no intercept term).

This is called **Regression Through the Origin**.

Why do we do this? (Justification)

1. **Theoretical reasoning:** Some variables have a natural zero point.
2. **Economic logic:** Including an intercept may not make sense (e.g., cost cannot be positive when quantity is zero).
3. **Improves accuracy** when theory clearly supports zero-intercept.

Estimation (Basic Idea Only)

Since there is no intercept, OLS simply finds the slope that minimises: $\sum(Y - \beta_1 X)^2$.

The OLS formula becomes:
$\beta_1 = \sum XY / \sum X^2$
(You don't need to derive this for UG unless asked.)

Properties

- **Unbiased:** β_1 is unbiased if the model truly passes through origin.
- **Efficient:** If theory is correct, it yields a more precise estimate.
- **But risky:** If the true model actually has an intercept, forcing it to zero **creates bias**.

Advantages

1. **Simpler model** (only one parameter).
2. **Useful when theory demands no intercept**.
3. **Better fit** when the relationship naturally begins at (0,0).

Disadvantages

1. If the true intercept is not zero, the whole model becomes **biased**.
2. R^2 interpretation becomes difficult because the model is forced through the origin.
3. It may give **misleading results** if used without theoretical justification.
4. Standard OLS assumptions may no longer hold properly.

When to Use Regression Through the Origin

Use it **only when economic theory, physical laws, or prior knowledge strongly suggest Y must be zero when X is zero**.

Examples:

- Productivity per worker when workers = 0 is meaningless → origin assumption valid.
- Distance = Speed × Time (when Time = 0, Distance = 0).
- Electricity consumption when number of bulbs = 0.

Double Log Model

A Double Log Model (also called Log–Log Model) is a regression model in which both the dependent variable (Y) and the independent variable (X) are transformed into their natural logarithms. The model is written as:

$$\ln Y = \beta_0 + \beta_1 \ln X + u$$

This model is very popular in economics because it allows us to interpret the regression coefficient directly as an **elasticity**.

Why Use a Double Log Model?

1. **Elasticity measurement:**
   β_1 shows the *percentage* change in Y for a *percentage* change in X. where, β_1 = elasticity of Y with respect to X.
2. **Linearising a non-linear relationship:**
   When Y and X have a curvilinear pattern (like a power function), taking logs makes it linear.
3. **Reduces skewness:**
   Taking logs makes data more normally distributed.
4. **Reduces heteroscedasticity:**
   Variance becomes more stable.
5. **Makes interpretation easy:**
   β_1 = "% change in Y when X increases by 1%".

Interpretation of Coefficient

If the model is: $\ln Y = \beta_0 + \beta_1 \ln X$

Then: **β_1 = elasticity**

Example: If $\beta_1 = 0.8$ → "A 1% increase in X leads to a 0.8% increase in Y."

✓ If $\beta_1 > 1$ → Y changes more than proportionately
✓ If $\beta_1 < 1$ → Y changes less than proportionately
✓ If $\beta_1 = 1$ → unit elasticity

When is Double Log Model Appropriate?

Use it when:

- The relationship between variables is multiplicative.
- Economic theory indicates percentage effects.
- Data show exponential or power-type growth patterns.
- Behavioural variables like consumption, demand, production follow elasticity concepts.

Common examples: Demand elasticity, Production functions, Income elasticity of consumption and Population growth models.

Advantages

- Coefficients are easy to interpret (elasticity).
- Helps solve heteroscedasticity.
- Linearises non-linear relationships.
- Works well for economic data with large ranges.

Disadvantages

- Cannot use when Y or X takes zero or negative values (log is undefined).
- Interpretation becomes difficult if theory doesn't support elasticity.
- Too much transformation may lose original meaning.

Measurement of Elasticity

Elasticity refers to the degree of responsiveness of one variable to changes in another variable. In economics, we commonly measure **Price Elasticity of Demand**, **Income Elasticity**, **Cross Elasticity**, and **Price Elasticity of Supply**. The measurement of elasticity helps understand how consumers and producers react to changes in prices, income, and related goods.

There are **four main methods** for measuring elasticity:

1. Percentage (Proportionate) Method

Also called the *Elasticity Formula Method*. Elasticity is measured as the **percentage change in quantity** divided by the **percentage change in price** or income.

$$E = \frac{\%\Delta Q}{\%\Delta P}$$

If price changes by 10% and quantity falls by 20%, elasticity = 2. This method is easy and commonly used but requires accurate percentage data.

2. Point Elasticity

Used when elasticity is measured at a **specific point** on a demand curve. Applicable for infinitesimally small changes. Formula (conceptually): $E = \frac{dQ}{dP} \times \frac{P}{Q}$

It is useful for economists but difficult for practical application because it requires calculus.

3. Arc Elasticity

Used when elasticity is measured **between two points** on a demand curve (finite changes). Useful when price changes are not small.

Concept: $E = \frac{\Delta Q}{\Delta P} \times \frac{P1+P2}{Q1+Q2}$

This method gives a better average elasticity between two points.

4. Total Expenditure (Outlay) Method – By Marshall

This method measures elasticity by observing **what happens to total expenditure** when price changes.

- If price falls and total expenditure **rises** → demand is **elastic** (E > 1)
- If price falls and total expenditure **falls** → demand is **inelastic** (E < 1)
- If price falls and total expenditure **remains the same** → **unit elasticity** (E = 1)

This method is simple and used when only price and expenditure data are available.

5. Geometric (Graphical) Method

Demand elasticity is measured using the **shape of the demand curve**.

On a straight-line demand curve:

$$E = \frac{Lower\ segment}{Upper\ segmen}$$

Elasticity is infinite at the top of the curve, unitary at the midpoint, and zero at the bottom.

Semi Log Model

A **Semi-Log Model** is a regression model in which **either the dependent variable (Y)** or the **independent variable (X)** is converted into logarithms, but **not both**. Only **one side** of the model is in log form. There are **two types** of Semi-Log Models:

1. Log–Linear Model (Log Y, Linear X)

$$\ln Y = \beta_0 + \beta_1 X + u$$

β_1 shows the percentage change in Y for a one-unit absolute change in X.

Interpretation: If X increases by 1 unit, Y increases by $\beta_1 \times 100$ percent.

Example: If $\beta_1 = 0.02$ → 1 unit increase in X raises Y by 2%.

2. Linear–Log Model (Linear Y, Log X)

$$Y = \beta_0 + \beta_1 \ln X + u$$

β□ shows the absolute change in Y for a 1% change in X.

Interpretation: If X increases by 1%, Y changes by $\beta_1/100$ units.

Example: If $\beta_1 = 5$ → 1% increase in X raises Y by 0.05 units.

Semi-log models are useful when the relationship between variables is **non-linear**, but only one variable grows proportionately.

When to Use Semi-Log Models

- When the effect is partly proportional and partly absolute.
- When the relationship is curved but not fully exponential.
- When economic theory suggests diminishing or increasing returns.
- For modelling: Wage equations, Growth rates, Cost functions, Demand studies (when only one variable is large in range)

Advantages

- Handles data with wide variation.
- Reduces heteroscedasticity.
- Linearises mildly non-linear relationships.
- Easy interpretation (percentage or unit changes).

Disadvantages

- Cannot use zero or negative values for log variables.
- Interpretation must follow the correct type of semi-log model.
- Too much transformation may hide the original economic meaning.

Measurement of Growth

Economic growth refers to the **increase in the output or income of an economy over time**. To study how fast an economy, sector, firm, or variable is growing, economists use different methods to **measure growth rates**. Growth is usually measured as the **percentage change** in a variable (like GDP, population, income, production) over a period.

1. Simple Growth Rate (Percentage Change Method)

This is the most common method.

$$\text{Growth Rate} = \frac{Y_t - Y_{t-1}}{Y_{t-1}} \times 100$$

Where:

Y_t= value in the current year

Y_{t-1}= value in the previous year

Example: If GDP increases from 100 to 110 → growth rate = 10%. This method is easy and used for annual growth calculations.

2. Compound Annual Growth Rate (CAGR)

CAGR measures the **average yearly growth** over a multi-year period, assuming growth compounds every year.

$$\text{CAGR} = \left(\frac{Y_t}{Y_0}\right)^{\frac{1}{n}} - 1$$

Where n = number of years.

This is used for **long-term trends** like population growth, investment returns, and economic planning.

3. Trend Growth Rate (Using Time Trend)

Economists sometimes use a **trend line** to measure long-term growth, especially when annual values fluctuate.

Trend equation:

$$Y = a + bt$$

Here, **b** represents the **average annual growth**.

This method smoothens short-term ups and downs and shows *long-term steady growth*.

4. Logarithmic Growth Rate (Semi-log method)

When the relationship is exponential (GDP, population), a log transformation helps.

Growth rate $\approx b \times 100$
where b is the coefficient of time in:

$$\ln Y = a + bt$$

This method is useful for **economic time series**, especially when the data show compounding.

5. Index Number Growth

Growth can also be measured using **index numbers** like GDP index, industrial production index (IIP), price index.

$$\text{Growth Rate} = \frac{\text{Index}_t - \text{Index}_{t-1}}{\text{Index}_{t-1}} \times 100$$

Useful for comparing growth across sectors.

Dummy Variables

A **dummy variable** is a variable that takes only **two values: 0 or 1**, used in regression analysis to represent **qualitative (categorical) factors** such as gender, region, season, policy changes, education levels, etc.

Since regression requires numerical values, dummy variables convert **non-numeric information** into a format usable in a regression model.

Example:
Gender Dummy

- Male = 1
- Female = 0

Purpose of Dummy Variables

Dummy variables help to **compare groups**, **measure shifts**, and **capture qualitative effects** in regression models.

They allow us to include: Gender effects, Regional effects, Before/after policy changes, Seasonal variations, Education categories, Structural breaks.

Types of Dummy Variable Uses

1. Intercept Dummy (Shift Dummy)

Used to measure whether the **intercept** changes between groups.

Example:

$$Y = \beta_0 + \beta_1 X + \beta_2 D + u$$

If D = 1 → group A
If D = 0 → group B

β_2 tells how much the intercept changes between the two groups.

2. Slope Dummy (Interaction Dummy)

Used to check whether **slopes differ** between groups.

Example:

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (DX) + u$$

β_3 shows how the slope differs between groups.

3. Seasonal Dummy Variables

Used in quarterly data:

Q1 = 1, others = 0

Q2 = 1, others = 0

Q3 = 1, others = 0

Q4 is left out (to avoid dummy trap)

These dummies measure seasonal effects.

4. Policy Dummy / Before–After Dummy

Used to measure impact of a policy change or event.

Example:

Before GST = 0

After GST = 1

β_2 captures the shift due to the policy.

Dummy Variable Trap

If we include **all categories** as dummies, perfect multicollinearity occurs.

Example:
Male = 1, Female = 0
and
Female = 1, Male = 0

They add up to 1, creating collinearity.

Solution: **Always drop one category** (called the reference group).

Interpretation of Dummy Coefficient

If model is:

$$Y = \beta_0 + \beta_1 D + u$$

- For D = 0 → Y = β_0
- For D = 1 → Y = $\beta_0 + \beta_1$

Thus, **β_1 shows the difference in mean values** between two groups.

Example: If $\beta_1 = 5$ → Group 1 has Y higher by 5 units than Group 0.

Advantages

- Converts qualitative data into quantitative form
- Easy interpretation
- Used to compare groups
- Useful for structural break analysis
- Essential in time-series and cross-sectional analysis

Disadvantages

- Too many dummies increase the number of variables
- Interpretation becomes difficult when many categories exist
- Dummy trap must be avoided
- Arbitrary coding (0/1) may affect interpretation if not explained clearly

---

ANOVA Models

ANOVA is a statistical technique used to check whether **three or more groups have the same mean**. It helps to test if differences in sample means are statistically significant.

Econometricians use ANOVA when the **dependent variable is continuous** but the **explanatory variables are categorical** (e.g., gender, type of school, region).

Purpose

- To identify whether group differences exist.
- To test hypotheses like:
  "Is there any difference in average income among different education levels?"

Basic Idea

Total variation in data is divided into:

1. **Between-group variation** (explained variation)
2. **Within-group variation** (unexplained variation)

A high ratio of between-group to within-group variation means group means are significantly different.

ANOVA in Regression Form

ANOVA can be represented using **dummy variables**:

Example:

$$Y_i = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + u_i$$

Where D_1, D_2 are dummy variables for categories.

F-Test

ANOVA mainly uses the **F-statistic** to test:

- H_0: All group means are equal
- H_1: At least one group mean is different

ANCOVA Models

ANCOVA is an extension of ANOVA. It tests group differences **after controlling for the effects of continuous variables** (called covariates).

In simple words: **ANCOVA = ANOVA + Regression**

It adjusts group means by removing the influence of other continuous variables.

Example

To study the effect of **school type** on student test scores, controlling for:

- IQ
- Hours studied

We use:

$$Y_i = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 X_1 + \beta_4 X_2 + u_i$$

Where

- Dummy variables = categories (e.g., school type)
- Covariates X_1, X_2= continuous variables

Purpose

- To compare adjusted group means
- To control for confounding factors
- To increase accuracy of group comparisons

When to Use ANCOVA

- When groups differ on some background characteristics
- When you want to isolate the effect of a treatment

F-Test in ANCOVA

ANCOVA also uses F-tests:

- To test for significance of dummy variables after controlling covariates
- To test for the effect of covariates

Dummy Variable Trap

The **Dummy Variable Trap** occurs when **dummy variables are perfectly multicollinear**. This means the dummy variables **provide redundant information**, leading to a situation where the regression model cannot estimate coefficients.

In simple words: **It happens when you use too many dummy variables for the same categorical variable.**

Why Does It Happen?

If a category has **k groups**, and we create **k dummy variables**, their sum will always be 1.

Example: For Gender

Male = 1, Female = 0

Female = 1, Male = 0

Then:

$$\text{Male} + \text{Female} = 1$$

This creates **perfect multicollinearity**, and the regression cannot run.

Example

Suppose a variable has **3 categories**:

1. Rural
2. Urban
3. Semi-urban

If we create 3 dummies:

D_1 = 1 if Rural

D_2 = 1 if Urban

D_3 = 1 if Semi-urban

Then:

$$D_1 + D_2 + D_3 = 1 \text{for every observation}$$

This makes them perfectly correlated → regression fails.

How to Avoid the Dummy Variable Trap?

Use **k – 1 dummy variables** for **k categories**.

Example: If 3 categories → use 2 dummy variables.

Let:

D_1 = 1 if Rural (0 otherwise)

D_2 = 1 if Urban (0 otherwise)

Then **Semi-urban** becomes the **reference (base) category**.

Interpretation After Avoiding the Trap

The coefficient of each dummy variable shows the **difference** from the **base category**.

Example:
If Semi-urban is base,

- Coefficient of Rural = difference between Rural & Semi-urban
- Coefficient of Urban = difference between Urban & Semi-urban

Why Leaving One Out Works?

Because leaving one category out removes the exact linear relationship and eliminates multicollinearity.

Mathematically: Avoids $D_1 + D_2 + D_3 = 1$

Interaction Effects

An **interaction effect** occurs when the impact of one independent variable on the dependent variable **depends on the level of another independent variable**.
In other words, the variables **do not act independently**; instead, they work **together** to influence the outcome.

Interaction effects are used when we believe that the relationship between $X$and Ychanges depending on another variable Z.

Why Interaction Effects Are Used

- To capture **combined effects** of variables
- To understand **more realistic relationships**
- To avoid misleading conclusions from simple additive models
- To see whether two factors **strengthen**, **weaken**, or **modify** each other's impact

General Form

In a regression, an interaction term is usually written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + u$$

Here,

- $X_1 \times X_2$ is the **interaction term**
- β_3 measures how the relationship between X_1 and Y changes when X_2 changes

Example

Suppose we study the effect of:

- **Education (X_1)**
- **Work Experience (X_2)**
  on **Income (Y)**.

Education and experience may not act separately. Higher experience may make education more valuable.

So, an interaction term $X_1 X_2$ helps show this combined effect.

If β_3 is positive → education boosts income more for experienced workers.
If β_3 is negative → the benefit of education decreases as experience increases.

Interaction With Dummy Variables

Interaction effects can also be used between:

- A **continuous variable** and a **dummy variable**, or
- Two **dummy variables**

Example: A dummy for gender × years of education → shows whether the return to education differs for men and women.

Interpretation

If the interaction term is **significant**, it means:

- The effect of one variable **changes depending on** another
- The two variables have a **combined impact** on the dependent variable
- The relationship is **not constant** across groups or levels

Why Interaction Effects Are Important

- They make models more realistic
- They reveal relationships that simple models miss
- They allow researchers to study **conditional effects**
- They help in policy decisions (e.g., education programs work differently across income groups)

Structural Changes

A **structural change** refers to a situation where the underlying relationship between the dependent variable and the independent variables **changes over time**. This means the **regression parameters (intercept or slope)** are not stable for the entire period. Something has happened—like a policy shift, technological change, economic crisis, new program, or behavioural change—that causes the relationship to alter.

In simple words: Structural change means the model before and after a particular point behaves differently.

When Do Structural Changes Occur?

Structural breaks may happen due to:

- Government policy changes (GST introduction, liberalisation)
- Economic shocks (recession, COVID-19)
- Technological progress (automation)
- Social or behavioural changes
- Institutional reforms

These events can cause a sudden or gradual change in the economic relationship.

Why Structural Changes Matter?

- If structural breaks are ignored, regression results become **biased** and **misleading**.
- Forecasts become **inaccurate** because old relationships no longer hold.
- Hypothesis testing loses reliability.
- Policy conclusions may be wrong.

Econometric models assume **parameter stability**, so detecting changes is important.

Types of Structural Changes

1. **Change in Intercept Only**
   Relationship shifts up or down (e.g., incomes rise suddenly due to policy).

2. **Change in Slope Only**
   Strength of relationship changes (e.g., effect of education on income becomes stronger over time).
3. **Change in Both Intercept and Slope**
   Entire regression line changes.
4. **Sudden Structural Break**
   Happens at a known point (e.g., demonetisation).
5. **Gradual Structural Change**
   Relationship changes slowly over time.

Detection of Structural Changes

Economists use statistical tests to check if coefficients differ over time.

1. Chow Test (Most Common)

Used when the break point is known.
Compares the regression before and after the suspected break.

2. Recursive Residual Tests

Used for gradually changing relationships.

3. CUSUM and CUSUMSQ Tests

Graphical tests to detect instability in parameters.

How to Handle Structural Changes (Remedial Measures)

1. **Use Dummy Variables**
   Add a dummy variable to represent the period after the structural break.
2. **Run Separate Regressions**
   Estimate the model separately for each time period.
3. **Use Models that Allow Changing Parameters**
   Like rolling regressions or dynamic models.
4. **In case of gradual changes**, use time-varying parameter models.

Autoregressive Lag Model

An **Autoregressive (AR) Lag Model** is a time-series model in which the current value of a variable depends on its **own past (lagged) values**. It captures the idea that past behaviour influences present behaviour.

In simple words: Y today depends on Y yesterday, Y before yesterday, and so on.

General Form

An **AR(1)** model (first-order autoregressive) is:

$$Y_t = \alpha + \beta Y_{t-1} + u_t$$

- Y_t= value of the variable at time t
- Y_{t-1}= value of the same variable in the previous period
- u_t= error term

Higher-order models include more lags, such as:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + u_t$$

Why AR Lag Models Are Used

- Many economic variables show **persistence over time**.
- Past values strongly influence present values (e.g., GDP, inflation, prices).
- To capture time-dependent patterns.
- To improve forecasting accuracy.

Interpretation

If β is:

- **Positive and large** → strong influence of past values
- **Between –1 and 1** → process is stable
- **Greater than 1** → explosive series (unstable)
- **Negative** → cycles or oscillations in the data

Examples

1. **Income depends on past income**
   People tend to maintain their consumption patterns, so past income influences current income.
2. **Inflation depends on previous inflation rates**
   Inflation usually carries forward due to price stickiness.
3. **Stock prices depend on past prices**
   Financial markets show momentum or reversal patterns.

Advantages

- Simple to estimate
- Good for short-term forecasting

- Captures persistence and inertia in economic time series
- Useful for stationary data

Disadvantages

- Cannot handle structural breaks
- Needs stationarity; otherwise results are misleading
- Ignores influence of external variables (uses only past Y)

Use in Econometrics

Autoregressive lag models are used:

- As part of **ARIMA models**
- In **dynamic regression models**
- In studying **business cycles**, **inflation**, **GDP growth**, etc.

They help understand how a variable evolves over time and how strongly it is tied to its own past.

Distributed Lag Model

A **Distributed Lag Model** is a regression model where the dependent variable depends not only on the current value of an explanatory variable but also on **its past (lagged) values**. This means the effect of an independent variable is **spread out over several time periods**.

In simple words: X affects Y not only today but also in future periods.

General Form

A basic distributed lag model looks like:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + u_t$$

Where:

- X_t= current value of the independent variable
- X_{t-1}, X_{t-2}= lagged values (1-period, 2-period, etc.)
- $\beta_0, \beta_1, \beta_2$= short-run, delayed, or long-run effects

Why Distributed Lag Models Are Used

Many economic variables do not adjust immediately. Their effects appear gradually over time (lag effect).

Examples:

- Advertising today increases sales **over several months**.
- Monetary policy changes affect GDP **after a delay**.
- Rainfall influences agricultural output **in future seasons**.

Thus, the model captures both **immediate** and **delayed** impacts.

Interpretation of Coefficients

β_0: Immediate (short-run) effect of X on Y

β_1: Effect after 1 time period

β_2: Effect after 2 time periods

The total or long-run effect is:

$$\beta_0 + \beta_1 + \beta_2 + \cdots$$

Examples

1. Sales depend on current and past advertising expenditure.
2. Agricultural output depends on rainfall of current and past years.
3. Consumption depends on current income and past incomes.

Advantages

- Captures delayed responses in economic behaviour
- More realistic than models using only current values
- Useful for policy analysis where effects are not immediate

Disadvantages

- Including too many lagged variables reduces degrees of freedom
- High multicollinearity among lagged X values
- Choosing the correct number of lags is difficult

Types of Distributed Lag Models

1. **Finite Distributed Lag Model (FDL)**
   Limited number of lags (e.g., 2 or 3).
2. **Infinite Distributed Lag Model (IDL)**
   Effects continue indefinitely with declining weights. (e.g., Koyck and Almon Lag models)

Ad Hoc Method of Estimation

The **Ad Hoc Method of Estimation** refers to an informal, non-theoretical, and approximate way of estimating parameters in an econometric model when standard techniques like OLS, Maximum Likelihood, or GLS cannot be applied directly. "Ad hoc" means **"for a specific purpose"**. So, this method uses **practical, experience-based, or convenient rules** to obtain estimates, even if they do not come from a formal statistical procedure.

In simple words: Ad hoc estimates are rough, practical estimates used when exact methods are difficult or impossible.

Why Ad Hoc Methods Are Needed

Economists use ad hoc methods when:

- Data is insufficient for standard estimation
- Relationships are complex
- Lag structure is unknown
- Variables cannot be perfectly measured
- Theoretical models are too difficult to estimate

Thus, ad hoc estimation provides **quick, workable solutions** even if they are not theoretically perfect.

Common Situations Where Ad Hoc Methods Are Used

1. Estimating Lag Models

Lag lengths are often chosen using:

- Experience
- Trial and error
- Plots of data
- Simple heuristics (e.g., stopping when coefficients become insignificant)

This is an ad hoc approach.

2. Ad Hoc Weights

In distributed lag models, instead of using complex lag structures, economists sometimes assign:

- Equal weights

- Declining weights
- Arbitrary weights
  based on judgement rather than theory.

3. Ad Hoc Adjustments

Economists may sometimes:

- Smooth time series
- Average data
- Adjust for inflation, without following strict statistical rules.

Advantages

- Easy to understand
- Quick and practical
- Useful when standard methods are impossible
- Can give reasonable estimates in real-world situations
- Less computational effort

Disadvantages

- Not based on strong theory
- May be biased or inconsistent
- No guarantee of efficiency
- Results depend heavily on the researcher's judgement
- Hard to generalize or repeat

Examples

1. **Estimating demand using moving averages** instead of OLS.
2. **Choosing lag lengths based purely on past studies** rather than statistical tests.
3. **Assigning weights "by intuition"** in distributed lag models.
4. **Dropping variables without formal tests** because they "look" insignificant.

Koyck Transformation

The **Koyck Transformation** is a technique used to convert an **Infinite Distributed Lag (IDL) model** into a simpler, **estimable** regression model.
It assumes that the effect of an independent variable declines **geometrically** over time.

Because estimating an infinite number of lagged variables is impossible, the Koyck method helps express the model in a **compact, workable form**.

In simple words: It converts a model with many lags into a model with only one lag of the dependent variable**.**

Starting Point: Infinite Distributed Lag Model

Consider a model where the dependent variable Y_t depends on current and all past values of X_t:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + u_t$$

This cannot be estimated directly because there are **infinitely many coefficients**.

Key Assumption of Koyck

The lag coefficients decline geometrically:

$$\beta_1 = \lambda\beta_0, \beta_2 = \lambda^2\beta_0, \beta_3 = \lambda^3\beta_0, \ldots$$

where

- $0 < \lambda < 1$
- λ is called the **lag decay factor**

This means the effect of X gets smaller and smaller over time.

Koyck Transformation – Basic Idea

Using the geometric decline assumption, the infinite lag model can be rewritten as:

$$Y_t = \alpha + \beta_0 X_t + \lambda Y_{t-1} + v_t$$

Thus, instead of infinite lags of X, we now have:

- the current value of X, and
- **one lag of Y**

This is called the **Koyck (or geometrically declining lag) model**.

Advantages of the Koyck Transformation

- Converts an infinite lag model into a **simple dynamic model**
- Easy to estimate using OLS
- Requires only **current X** and **lagged Y**
- Reduces the number of parameters drastically

- Useful for forecasting

Disadvantages

- Based on a strong assumption (geometric decline)
- Errors become **autocorrelated** (presence of serial correlation)
- May not fit well if actual lag structure is different
- Interpretation becomes slightly complex

Economic Example

Advertising and sales:
Advertising today creates awareness immediately, but its effect continues in future months.

Koyck model helps capture this **gradual, declining impact** of advertising on sales.

Interpretation

- β_0: Short-run (immediate) effect
- λ: Speed of decay of lag effect
- Long-run effect:

$$\frac{\beta_0}{1-\lambda}$$

This shows the total impact of X over time.

Mean Lag

In distributed lag models, the effect of an independent variable (X) on the dependent variable (Y) usually spreads over several periods. The **Mean Lag** measures the **average delay** with which the effect of X influences Y.

In simple words: Mean lag tells us, on average, how many periods it takes for X to fully affect Y. It is a summary measure of how the lagged effects are distributed across time.

Why Mean Lag Is Important

- It helps understand the **timing** of economic responses.
- Useful in policy analysis (e.g., How long will it take for monetary policy to affect GDP?)
- Helps compare different lag structures.

- Shows whether effects are immediate or delayed.

Formula (Conceptual, UG Level)

Mean lag is calculated as a **weighted average**, where each lag length is weighted by its corresponding coefficient.

If the effects are spread over 0, 1, 2, … periods with coefficients:

$$\beta_0, \beta_1, \beta_2, \ldots$$

Then the **mean lag** is:

$$\text{Mean Lag} = \frac{1 \cdot \beta_1 + 2 \cdot \beta_2 + 3 \cdot \beta_3 + \cdots}{\beta_0 + \beta_1 + \beta_2 + \cdots}$$

Interpretation

Mean lag = 0 → Effects are immediate

Mean lag = 1 → Most of the effect appears after one period

Higher mean lag → Longer delay in response

Example:
If mean lag = 3, then the effect of X on Y spreads out and peaks around **3 periods later**.

Example (Conceptual)

Suppose advertising has effects:

- 50% this month
- 30% next month
- 20% the month after

The mean lag will be around **1 month**, showing a moderate delay in response.

Mean Lag in Koyck (Geometric Lag) Model

In the Koyck model, where effects decline geometrically, mean lag is:

$$\text{Mean Lag} = \frac{\lambda}{1 - \lambda}$$

If λ is high hence it is the long mean lag
If λ is low hence it is the short mean lag

Median Lag

In a distributed lag model, the effect of an independent variable (X) on a dependent variable (Y) is spread across several future periods. The **Median Lag** tells us the time period by which **half of the total effect** of X on Y has already occurred.

In simple words: Median lag is the time it takes for 50% of the total impact of X to be felt on Y.

It is similar to a "midpoint in time" of the lagged responses.

Why Median Lag Is Useful

- Shows how quickly a system responds to changes
- Helps understand short-run vs. long-run behaviour
- Useful in policy analysis (e.g., when will half the impact of a policy be visible?)
- Helps compare different lag structures (shorter vs. longer response periods)

How It Works

You list the lag coefficients:

$$\beta_0, \beta_1, \beta_2, \beta_3, \ldots$$

Then you calculate the **cumulative effect** until it reaches **50% of the total effect**.

The time period where this happens is the **Median Lag**.

For example:

- If 20% effect happens in period 0
- 30% more in period 1
  → After period 1, 50% of total effect is completed
  → So **median lag = 1**.

Interpretation

- **Median lag = 0** → half the effect is immediate
- **Median lag = 1** → half the effect happens after one period
- **Higher median lag** → slower adjustment, more delayed effect

It tells us **how fast the main impact takes place**.

Example (Conceptual)

Suppose the distributed impact of advertising is:

- 30% in the current month

- 40% next month
- 30% the month after

Cumulative effects:

- Month 0: 30%
- Month 1: 70%

Since 50% is crossed in **Month 1**, → **Median lag = 1 month**.

Unit V: Simultaneous Equation Model

Simultaneous Equation Model: Definition and Examples – Simultaneous Equation Bias – Structural and Reduced Form Equations – Identification – Rank and Order Condition – Indirect Least Square Estimation – Two Stage Least Square Estimation.

Simultaneous Equation Model: Definition and Examples

Definition

A **Simultaneous Equation Model** is a system of two or more equations in which the variables are **jointly determined**. In these models, one variable may be the dependent variable in one equation and an explanatory variable in another. Because the variables influence each other at the same time, their relationships are *simultaneous*, and therefore ordinary least squares (OLS) cannot be directly used.

Examples

1. Demand–Supply Model

Demand equation: $Q_d = a - bP$

Supply equation: $Q_s = c + dP$

Here, **Price (P)** and **Quantity (Q)** are determined together by both equations. Price affects demand and supply, and demand–supply together determine price → **simultaneous determination**.

2. Keynesian Income–Consumption Model

Consumption function: $C = a + bY$

Income identity: $Y = C + I + G$

Consumption depends on income, but income itself depends on consumption. Both **Y** and **C** determine each other → **simultaneous relationship**.

3. Money Market Model

Money demand: $Md = f(i, Y)$

Money market equilibrium: $Ms = Md$

Interest rate (**i**) affects money demand, but the interest rate is determined by the balance of demand and supply of money. Hence, interest and money demand are jointly determined.

Simultaneous Equation Bias

In econometrics, **simultaneous equation bias** refers to the bias that arises when two or more economic variables are determined together and influence each other at the same time, but the researcher incorrectly estimates one equation as if it were independent. In such systems, the dependent variable of one equation often appears as an explanatory variable in another equation. Because of this mutual dependence, the explanatory variables become correlated with the error term. This violates a key assumption of the Classical Linear Regression Model—that the regressors must be independent of the disturbance term. As a result, applying Ordinary Least Squares (OLS) to a single structural equation within a simultaneous system produces **biased and inconsistent estimates**, meaning that even with a large sample size, the OLS estimates do not approach the true parameter values.

The root cause of simultaneous equation bias is **endogeneity**. When a regressor is jointly determined with the dependent variable, changes in the error term of one equation spill over to other equations in the system. For example, in a simple **demand and supply model**, price and quantity are determined simultaneously. If we estimate the demand equation by regressing quantity on price using OLS, the estimated coefficient of price will be biased because price is influenced not only by demand factors but also by supply factors captured in the error term. Hence, price is not a purely exogenous variable, making the OLS assumption invalid. This situation is common in economics where variables such as income, consumption, investment, interest rates, and output influence each other.

Simultaneous equation bias leads to serious **consequences** for empirical research. First, the coefficient estimates will be misleading, making policy conclusions unreliable. Second, hypothesis testing becomes invalid because the standard errors are incorrect. Third, predictions made from biased estimates tend to be inaccurate. Therefore, simultaneous equation models require special estimation techniques that can handle endogeneity and account for the interdependence among variables.

To solve the problem of simultaneous equation bias, econometricians use **identification** and specialized estimation methods. Identification ensures that enough information is available to estimate each structural equation. Once an equation is identified, the appropriate method can be applied. The commonly used methods include **Indirect Least Squares (ILS)**, **Two-Stage Least Squares (2SLS)**, and **Limited Information Maximum Likelihood (LIML)**. These methods break the correlation between the explanatory variable and the error term by using **instrumental variables** or by transforming equations. As a result, the estimated coefficients become unbiased and consistent.

In summary, simultaneous equation bias occurs because economic relationships operate jointly rather than separately, causing explanatory variables to become endogenous. OLS fails in such cases, giving biased and inconsistent results. Proper identification and the use of instrumental variable methods like ILS and 2SLS are essential for obtaining reliable estimates in simultaneous equation systems.

Structural and Reduced Form Equations

1. Structural Form Equations

Structural form refers to the **original economic relationships** derived directly from economic theory. Each structural equation shows how one variable depends on other variables that influence economic behaviour, such as preferences, technology, income, cost, etc.

Characteristics of Structural Equations

1. They contain **endogenous variables** (jointly determined within the system).
2. They also include **exogenous variables** (determined outside the model).
3. Each equation has its **own error term**, representing unobserved influences.
4. The coefficients have **clear economic interpretation**.
5. Equations often contain **simultaneous interactions**.

Example

Consider a basic demand– supply model:

Demand equation (structural):

$$Q_d = a - bP + u_1$$

Supply equation (structural):

$$Q_s = c + dP + u_2$$

Here, **price (P)** and **quantity (Q)** are endogenous. Income, cost, or weather could be exogenous.

These equations are "structural" because they reflect the actual behaviour of consumers and producers.

2. Reduced Form Equations

Reduced form equations are obtained by **solving the entire system of structural equations algebraically**. In reduced form, each endogenous variable is expressed **entirely in terms of exogenous variables and disturbances**, with no endogenous variable appearing on the right-hand side.

Characteristics of Reduced Form

1. Endogenous variables are on the **left-hand side**.
2. Only **exogenous variables and error terms** appear on the right-hand side.
3. Coefficients do **not have direct economic meaning**.
4. They represent the **equilibrium solutions** of the system.
5. Useful for **estimation**, especially using OLS.

Example (Reduced Form of Demand–Supply System)

Solving the two structural equations gives:

$$P = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + v_1$$
$$Q = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + v_2$$

Where Z_1, Z_2 are exogenous variables (income, cost, etc.).

These reduced-form equations show how equilibrium price and quantity depend on outside factors.

3. Relationship Between Structural and Reduced Forms

Structural form shows **how economic agents behave**, while reduced form shows **how the final equilibrium values are determined**.

- Structural form → directly meaningful economically.
- Reduced form → directly useful for econometric estimation.

Reduced-form coefficients are often used as the basis for estimating the structural coefficients through methods like **Indirect Least Squares (ILS)** and **Two Stage Least Squares (2SLS)**.

4. Why Economists Need Both Forms

Structural Form

- Helps understand the underlying economic mechanism.
- Used to test economic theories.
- Necessary for policy analysis (e.g., effect of taxes on supply).

Reduced Form

- Necessary for statistical estimation.
- Used to identify whether the model can be estimated (identification).
- Helps detect whether an equation is **under-identified**, **exactly identified**, or **over-identified**.

Thus, structural form explains **why** relationships exist, while reduced form explains **what** the outcome will be.

5. Example Showing Connection

Suppose the structural equations are:

$$Q = a_0 + a_1 P + u_1$$
$$P = b_0 + b_1 Z + u_2$$

Solve for Q:

Substitute for P:

$$Q = a_0 + a_1(b_0 + b_1 Z + u_2) + u_1$$

This reduced form is:

$$Q = \alpha_0 + \alpha_1 Z + v$$

Where:

$$\alpha_0 = a_0 + a_1 b_0, \alpha_1 = a_1 b_1, v = u_1 + a_1 u_2$$

Here, coefficients α_0, α_1 do not have direct interpretation but are crucial for estimation.

6. Importance in Econometrics

1. Structural form helps identify **economic parameters**, like elasticity, marginal effects, etc.
2. Reduced form provides the **OLS-estimable version** of the model.
3. Helps in determining identification using **order and rank conditions**.
4. Supports estimation using ILS, 2SLS, LIML, and 3SLS.
5. Useful in macroeconomic modelling (IS-LM, consumption–investment models, etc.).

Conclusion

Structural and reduced form equations are central to the study of simultaneous equation models. The structural form represents the underlying theoretical relationships, while the reduced form provides empirically estimable equations that express endogenous variables as functions of exogenous variables. Both forms are essential because they serve different purposes—structural form for theory and policy analysis, and reduced form for econometric estimation and identification. Understanding the connection between the two strengthens the foundation for advanced methods like Indirect Least Squares and Two Stage Least Squares.

Identification

In simultaneous equation models, **identification** refers to the problem of determining whether the structural parameters of an equation can be uniquely estimated from the available data. Since several endogenous variables are jointly determined, it is not always possible to isolate the true economic relationship of a single equation. For an equation to be **identified**, there must be enough information to distinguish it from the other equations in the system. Without identification, the coefficients of the structural equation cannot be estimated, and the economic interpretation becomes impossible.

There are three types of identification. An equation is **exactly identified** when there is just enough information to estimate its parameters uniquely. It is **over-identified** when more than one set of instruments or restrictions is available to estimate the parameters. In this case, advanced estimation techniques like Two-Stage Least Squares (2SLS) are used. An equation is **under-identified** when there is insufficient information to estimate the parameters at all. In such cases, no econometric method can recover the structural coefficients, making the equation useless for empirical analysis.

The identification status is usually checked using two rules. The **order condition** states that the number of excluded exogenous variables from an equation must be at least equal to the number of endogenous variables in that equation minus one. The **rank condition**, which is more technical, ensures that the excluded variables have a real influence on the endogenous variables and are not redundant. Only when both conditions are satisfied is the equation considered identified. Identification is important because without it, the estimated coefficients will suffer from simultaneous equation bias, leading to incorrect policy conclusions and theoretical misinterpretations. Thus, identification ensures that each structural equation in a simultaneous system can be meaningfully and accurately estimated.

Rank and Order Condition

In simultaneous equation models, the **Rank and Order Conditions** are used to determine whether a structural equation is **identified**, meaning whether its parameters can be uniquely estimated. Identification is essential because, without it, we cannot separate one economic relationship from another in a system where variables are jointly determined.

1. Order Condition of Identification

The **Order Condition** is a *necessary but not sufficient* condition. It states that:

The number of exogenous variables excluded from a structural equation must be at least equal to the number of endogenous variables in that equation minus one.

Symbolically:
Excluded exogenous variables ≥ (Endogenous variables – 1)

If this requirement is not met, the equation is **under-identified**, meaning its parameters cannot be estimated. If the condition is met, the equation may be exactly identified or over-identified, but further verification (rank condition) is needed.

2. Rank Condition of Identification

The **Rank Condition** is a *necessary and sufficient* condition. It states that:

There must be at least one non-zero determinant in the matrix formed by the coefficients of the excluded exogenous variables that affect the endogenous variables in the equation.

In simpler words, the excluded variables must have a **real and independent influence** on the endogenous variables of the system. If the matrix lacks this influence, the equation cannot be uniquely identified even if the order condition is satisfied.

Relation Between the Two Conditions

- **Order Condition** → gives a quick check of whether identification is possible.
- **Rank Condition** → confirms identification with certainty. Thus, a structural equation is considered **identified** only when the *rank condition is satisfied*.

Conclusion

The Rank and Order Conditions together help determine whether a structural equation is under-identified, exactly identified, or over-identified, ensuring that the econometric estimation of simultaneous equation models is valid and meaningful.

Indirect Least Square Estimation

Indirect Least Squares (ILS) is an important method used in the estimation of **simultaneous equation models**, especially when one of the equations in the system is **exactly identified**. In simultaneous equation models, Ordinary Least Squares (OLS) cannot be used directly because the explanatory variables are jointly determined and are correlated with the error term, leading

to simultaneous equation bias. ILS helps solve this problem by transforming the structural equation into its **reduced form** and then estimating the structural parameters using the reduced-form coefficients. The method is called "indirect" because we do not estimate the structural equation directly; instead, we estimate the reduced-form equations and then use algebra to derive the structural coefficients.

Concept of ILS

In a simultaneous equation model, structural equations represent theoretical economic relationships. However, because endogenous variables appear on both sides of the equations, these equations cannot be estimated directly. Therefore, the system is first rewritten into **reduced form**, where each endogenous variable is expressed only in terms of exogenous variables and error terms. These reduced-form equations can be estimated using OLS because exogenous variables satisfy the OLS assumption of being uncorrelated with the error term. Once the reduced-form coefficients are estimated, the structural coefficients are recovered using algebraic substitution. This process of solving structural parameters from reduced-form estimates is known as Indirect Least Squares.

When is ILS Used?

ILS is mainly used when:

1. **The structural equation is exactly identified** (i.e., number of unknowns equals number of independent reduced-form coefficients available).
2. The model contains **recursive or two-equation systems** where closed-form algebraic solutions exist.
3. The model is simple enough to solve by hand without requiring advanced methods like 2SLS or 3SLS.

ILS is not suitable for over-identified equations because those require instrumental variable techniques.

Steps in Indirect Least Squares

Step 1: Write the structural equations

Example: A simple demand–supply model

- Demand: $Q = a - bP + u_1$
- Supply: $Q = c + dP + u_2$

These cannot be directly estimated using OLS because price (P) is endogenous.

Step 2: Convert the structural system into reduced form

Solve the two equations simultaneously to express **P** and **Q** only in terms of exogenous variables and error terms.

For example:

$$P = \alpha_0 + \alpha_1 Z + v_1$$
$$Q = \beta_0 + \beta_1 Z + v_2$$

Here, **Z** stands for exogenous variables such as income, cost, weather, etc.

Step 3: Estimate the reduced-form equations using OLS

Since reduced forms involve only exogenous variables, OLS gives unbiased and consistent estimates of
$\alpha_0, \alpha_1, \beta_0, \beta_1$.

Step 4: Recover the structural parameters

After estimating the reduced-form coefficients, the original structural coefficients (a, b, c, d) are obtained by solving algebraic relationships.

For example,

$$b = -\frac{\alpha_1}{\beta_1}$$
$$a = \beta_0 + b\alpha_0$$

This final step gives the required structural parameters indirectly, hence the name **Indirect Least Squares**.

Numerical Illustration (Simple Form)

Suppose reduced-form estimates from OLS are:

$$P = 2 + 0.5Z$$
$$Q = 10 + 1.5Z$$

Then:
$b = -\frac{0.5}{1.5} = -0.33$(price coefficient in demand)
$a = 10 + (-0.33)(2) = 9.34$(intercept of demand)

Thus, demand equation becomes:
Q = 9.34 – 0.33P

This shows how structural parameters are obtained indirectly.

Advantages of ILS

1. **Simple and easy to apply** for small simultaneous systems.
2. **Requires only OLS**, so no advanced software is needed.
3. Gives **consistent estimates** for exactly identified equations.
4. Useful for teaching and introductory econometrics.

Limitations of ILS

1. Works **only when the equation is exactly identified**.
2. Cannot be used when the system is **over-identified**.
3. Deriving reduced forms may become complicated for large models.
4. Errors in reduced-form estimation affect structural coefficients.
5. Not efficient compared to methods like 2SLS or GLS.

Conclusion

Indirect Least Squares is a fundamental estimation method used in simultaneous equation models when structural equations are exactly identified. It allows economists to obtain structural parameters by first estimating reduced-form equations using OLS and then solving for the structural coefficients. Though simple and useful for basic models, the method has limitations for complex or over-identified systems. Despite this, ILS remains a valuable tool for understanding the connection between structural and reduced-form models and forms the basis for more advanced estimation techniques.

Two Stage Least Square Estimation

Two-Stage Least Squares (2SLS) is one of the most widely used estimation techniques in simultaneous equation models, especially when an equation is **over-identified**. In simultaneous systems, explanatory variables are not truly independent because they are jointly determined with the dependent variable. This creates the problem of **simultaneous equation bias**, making Ordinary Least Squares (OLS) estimates biased and inconsistent. 2SLS solves this problem by using **instrumental variables (IV)**—variables that are correlated with the endogenous explanatory variables but uncorrelated with the error term. The method works in two stages: first predicting the endogenous variables using only exogenous variables, and then using these predicted values in the structural equation. This ensures unbiased and consistent estimates.

Why 2SLS Is Needed

Simultaneous equation models contain variables that influence each other. Because of this two-way causation, endogenous variables are correlated with the disturbance term. OLS fails because one of its key assumptions— "No correlation between explanatory variables and error term"— is violated. 2SLS breaks this correlation by replacing actual endogenous values with **fitted (predicted)** values that come only from exogenous variables. This makes the estimation statistically valid.

Basic Idea Behind 2SLS

2SLS uses the information contained in all the exogenous variables of the entire system to "clean" the endogenous variables. It does this by:

1. Generating predicted values of endogenous variables using exogenous variables.
2. Substituting these predicted values into the structural equation.
3. Estimating the equation using OLS in the second stage.

By using fitted values instead of actual endogenous variables, 2SLS eliminates the source of simultaneous equation bias.

Stage 1: Predicting Endogenous Variables

In the first stage, each endogenous explanatory variable in the structural equation is regressed on **all exogenous variables in the system**, whether they belong to that equation or not.

Example:
For the demand–supply model, price (P) is endogenous.

Stage 1 regression:

$$P = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \cdots + v$$

Where $Z_1, Z_2$are exogenous variables (income, wage, cost, weather, etc.).

From this, we obtain the **predicted value** $\hat{P}$.

This predicted value is uncorrelated with the error term.

Stage 2: Estimating the Structural Equation

In the second stage, replace the endogenous variable with its predicted value and then apply OLS.

Original structural equation:

$$Q = a + bP + u$$

Replace actual $P$with predicted $\hat{P}$:

$$Q = a + b\hat{P} + e$$

Now, OLS gives **consistent** estimates of $a$and b.

Illustrative Example (Simple Form)

Structural equations:

Demand:

$$Q = a - bP + u_1$$

Supply:

$$Q = c + dP + u_2$$

Price (P) is endogenous.

Stage 1:

Predict P using exogenous variables such as income (Y) and cost (C):

$$P = \alpha_0 + \alpha_1 Y + \alpha_2 C$$

Obtain predicted price: $\hat{P}$.

Stage 2:

Substitute $\hat{P}$in the demand equation:

$$Q = a - b\hat{P}$$

Estimate using OLS → unbiased and consistent values of **a** and **b**.

Advantages of 2SLS

1. **Eliminates simultaneous equation bias**, giving consistent estimates.
2. Works for **over-identified** and **exactly identified** equations.
3. Uses information from **all exogenous variables**, improving efficiency.
4. Easy to implement with modern statistical software.
5. Forms the foundation for more advanced methods like 3SLS and GMM.

Limitations of 2SLS

1. **Less efficient** than Maximum Likelihood methods when sample size is small.
2. Requires good **instrumental variables**; weak instruments give poor results.
3. Standard errors of 2SLS estimates are larger than OLS.
4. Method can become complicated for large systems.
5. Wrong instrument choice leads to misleading results.

When to Use 2SLS

- When the structural equation is **over-identified**.
- When an explanatory variable is **endogenous**.
- When OLS gives biased and inconsistent estimates.
- When proper instruments are available.

Conclusion

Two-Stage Least Squares is one of the most important estimation techniques in econometrics for simultaneous equation systems. By replacing endogenous variables with predicted values generated from exogenous instruments, 2SLS produces unbiased and consistent structural estimates. It is simple, powerful, and widely used in applied economic research, especially when dealing with demand-supply models, consumption-income systems, investment functions, and macroeconomic policy models. Despite some limitations, 2SLS remains a cornerstone of modern econometric practice.