BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title : Econometrics

Course Code: 23ST03DEC

Unit-I

Econometrics

Dr. T. Jai Sankar
Associate Professor and Head
Department of Statistics

Ms. E. Devi
Guest Faculty
Department of Statistics

Unit : 1

Introduction of Econometrics

- Econometrics was coined in 1926 by Norwegian economist Ragnar Frisch. He shared the first Nobel Prize in Economics in 1969 with Jan Tinbergen. The term highlighted a distinct approach to interpreting economic data. Though calculations existed earlier, econometrics formalized data usage in economics. Today, it's a dynamic field evolving with new statistical and computational tools. If you'd like, I can also structure this into a flashcard or note-style format for easy reference. Econometrics applies economic theory and statistical techniques to test hypotheses and forecast economic phenomena. Its literal meaning is "economic measurement," though its scope extends far beyond mere quantification. It involves using mathematical statistics to validate economic models and generate empirical results. Econometrics blends theory and observation through quantitative analysis of real-world data. It leverages tools from economics, mathematics, and statistical inference for rigorous analysis. Ultimately, econometrics seeks to uncover and empirically determine economic laws. Econometricians craft realistic assumptions to best utilize available data for analysis. They help improve the public perception of economics, which often suffers from abstract interpretations. Their work bridges economic theory and empirical measurement using statistical inference. The essence of econometric research lies in grounding theory with quantifiable evidence.

  "The application of statistical and mathematical methods to the analysis of economic data, with a purpose of giving empirical content to economic theories and verifying them or refuting them."

Nature and Scope of Econometrics

Nature :

- **Interdisciplinary**: Combines economics, mathematics, and statistics to analyze economic relationships.
- **Empirical Focus**: Translates theoretical models into testable hypotheses using real-world data.
- **Quantitative Analysis**: Uses tools like regression, time-series, and panel data methods to estimate and forecast.
- **Decision-Oriented**: Supports evidence-based policymaking and business strategy.

Scopes :

- **Theory Testing**: Validates economic theories with statistical evidence.
- **Policy Evaluation**: Assesses the impact of fiscal, monetary, and social policies.
- **Forecasting**: Predicts future trends in GDP, inflation, employment, etc.
- **Model Building**: Constructs models for consumer behavior, investment, and market dynamics.
- **Applied Fields**: Used in finance, healthcare, agriculture, labor economics, and environmental studies.

Methodology of Econometrics

Broadly speaking, traditional econometric methodology proceeds along the following lines:

1. Statement of theory or hypothesis
2. Specification of the mathematical model of the theory
3. Specification of the statistical, or econometric, model
4. Obtaining the data
5. Estimation of the parameters of the econometric model

6. Hypothesis testing

7. Forecasting or prediction

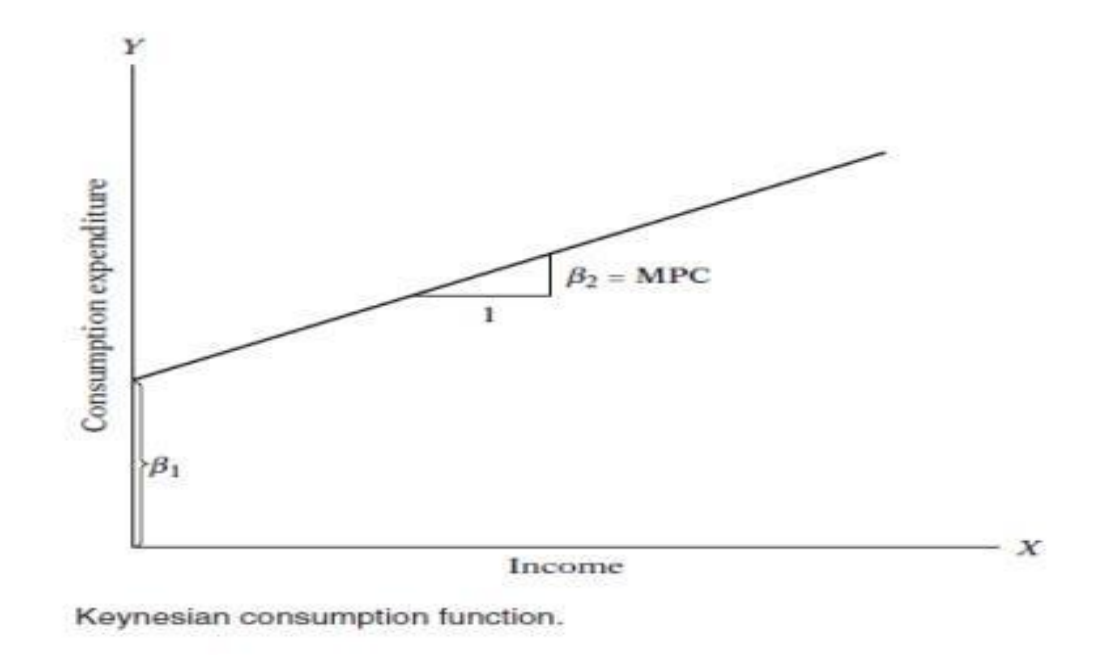
8. Using the model for control or policy purposes.

Statement of theory or Hypothesis :

Keynes postulated that Marginal propensity to consume (MPC), the rate of change of consumption for a unit, change in income, is greater than zero but less than one. i.e., 0 < MPC < 1 .

Specification of the Mathematical Model of Consumption :

Keynes postulated a positive relationship between consumption and income.



Keynesian consumption function.

Keynesian Consumption Function

The function is given by: $\mathbf{Y = \beta_1 + \beta_2 X}$

Where:

- **Y** represents consumption expenditure
- **X** represents income
- $\boldsymbol{\beta_1}$ is the intercept (autonomous consumption)
- $\boldsymbol{\beta_2}$ is the slope coefficient, indicating the marginal propensity to consume (MPC), with the condition $\mathbf{0 < \beta_2 < 1}$

β_1 and β_2 are known as the **parameters of the model**, representing the intercept and slope, respectively. This equation expresses a **linear and deterministic relationship** between income and consumption.

Remark :

- The Keynesian consumption function is expressed as **$Y = \beta_1 + \beta_2 X$**, where **Y** is consumption and **X** is income.

- The coefficient **β_2** represents the **marginal propensity to consume (MPC)**.
- Parameters **β_1** and **β_2** indicate the intercept and the slope, respectively.
- The condition **$0 < \beta_2 < 1$** ensures consumption increases with income, but not proportionally.
- This function defines a **linear relationship** between consumption and income.
- It's a **mathematical model** illustrating economic behavior.
- A **single equation** forms a single-equation model; multiple equations form a multiple-equation model.

Specification of the econometric model of consumption :

To account for the inexact relationship between economic variables, econometricians modify the deterministic consumption function as:

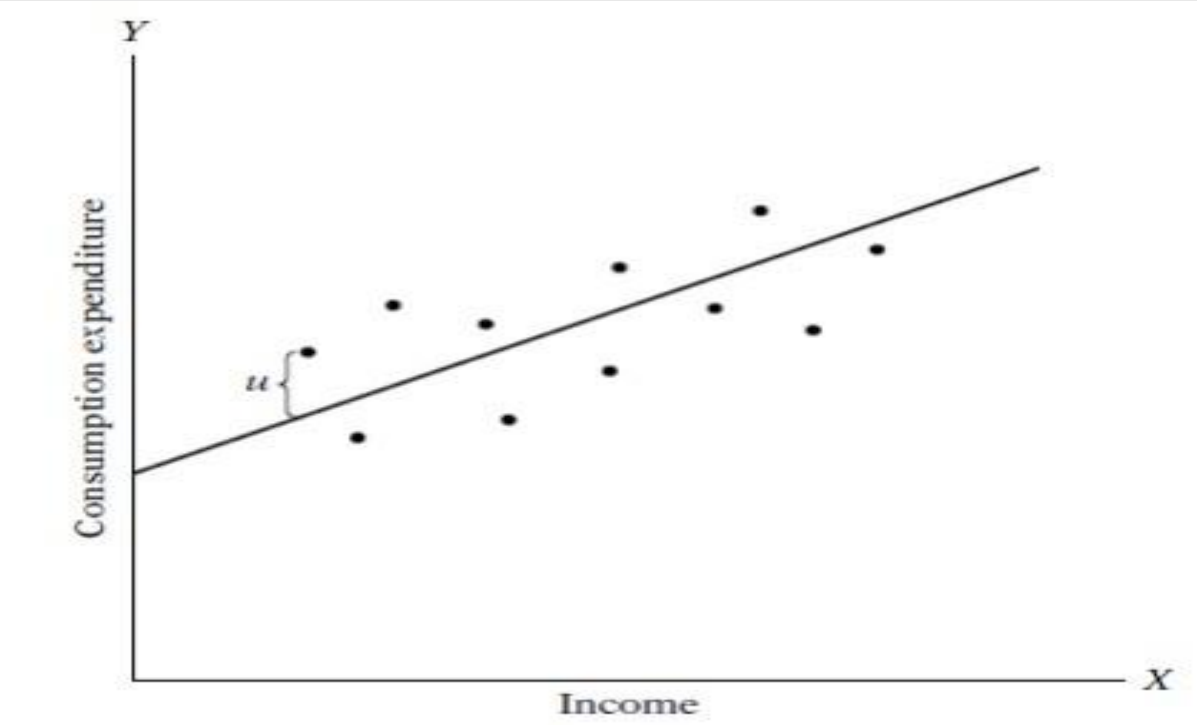
$$Y = \beta_1 + \beta_2 X + U$$

Where:

- **Y** is consumption expenditure (dependent variable)
- **X** is income (explanatory variable)
- **β_1** is the intercept, and **β_2** is the slope coefficient (MPC)
- **U** represents the error term, capturing the effects of all other factors not explicitly included in the model

This equation exemplifies an **econometric model**, more specifically, a **linear regression model**. It assumes a **linear but not exact** relationship between consumption and income

acknowledging individual variability and external influences.



Econometric model of the Keynesian consumption function.

Original Data :

Estimation of Parameters β_1 and β_2

To determine the numerical values of **β_1** and **β_2**, we require relevant economic data.

- **Y (Dependent Variable):** Represents aggregate *Personal Consumption Expenditure (PCE)*
- **X (Independent Variable):** Denotes *Gross Domestic Product (GDP)*, used as a measure of aggregate income
- The model: **$Y = \beta_1 + \beta_2 X + U$**, where **U** accounts for unexplained variations

Note:

The **Marginal Propensity to Consume (MPC)** is calculated as the *average change in consumption* in response to a *change in real income*.

Estimation of the Econometric Model :

Regression Analysis and the Estimated Consumption Function

Regression analysis is the primary statistical technique used to estimate the parameters of the consumption function.

- The estimated form of the function is: $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_i$
- Where $\hat{Y}$ is the estimated value of consumption
- X_i is the observed income
- $\hat{\beta}_1$ and $\hat{\beta}_2$ are estimated parameters (intercept and slope)

This equation is known as the **regression line**, representing the best-fitting linear relationship between income and consumption based on observed data.

Hypothesis Testing :

Keynes expected the MPC is positive but less than 1. Confirmation or refulation of economic theories on the basis of sample evidence is based on a branch of statistical theory known as statistical inference (hypothesis testing).

Forecasting or Prediction :

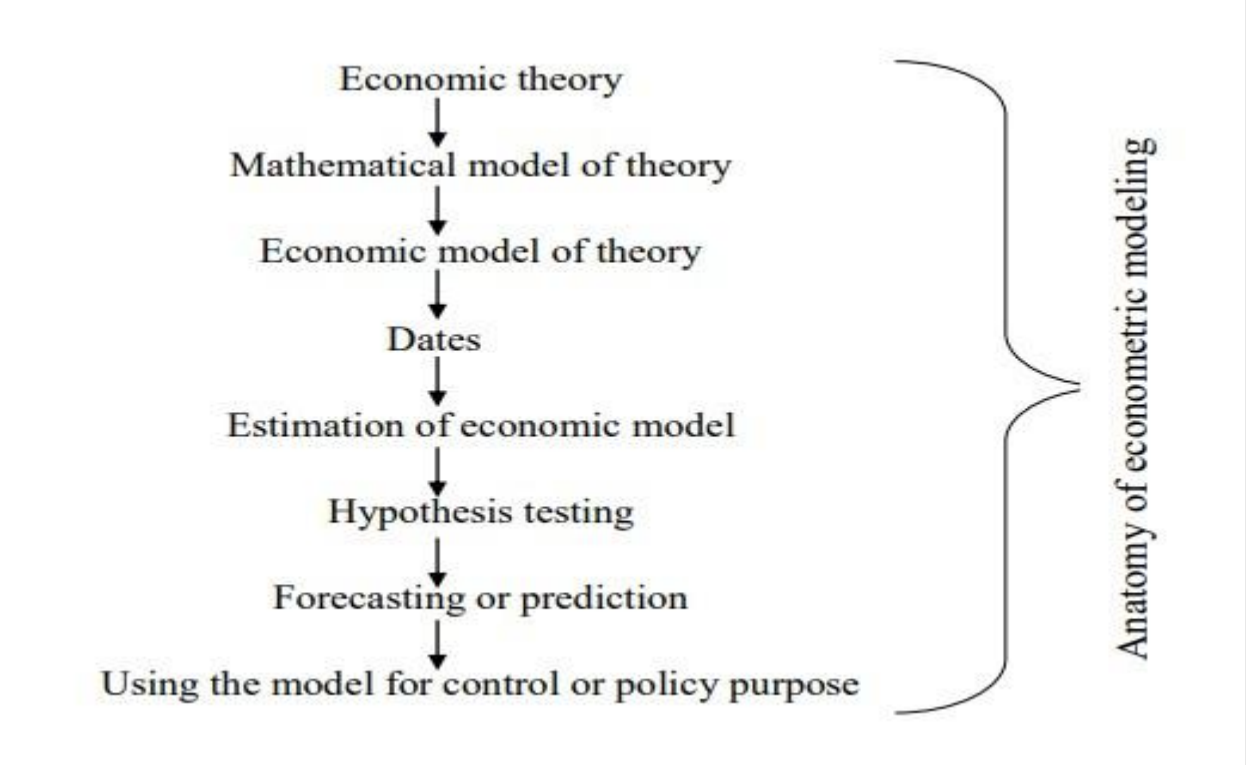
Model-Based Forecasting s Income Multiplier in Macroeconomics

- If a model doesn't reject the underlying economic theory, it can be used to **predict future values of the dependent variable (Y)** using known or expected values of the **explanatory variable (X)**.
- According to macroeconomic theory, a change in **investment expenditure** leads to a change in **income**, governed by the **income multiplier (M)**.
- The multiplier is defined as:

M = 1 / (1 – MPC), where **MPC** is the **Marginal Propensity to Consume**

- Estimating **MPC** provides key insights for economic policy.
- With MPC, one can forecast future **income**, **consumption**, and **employment**, especially in response to fiscal changes

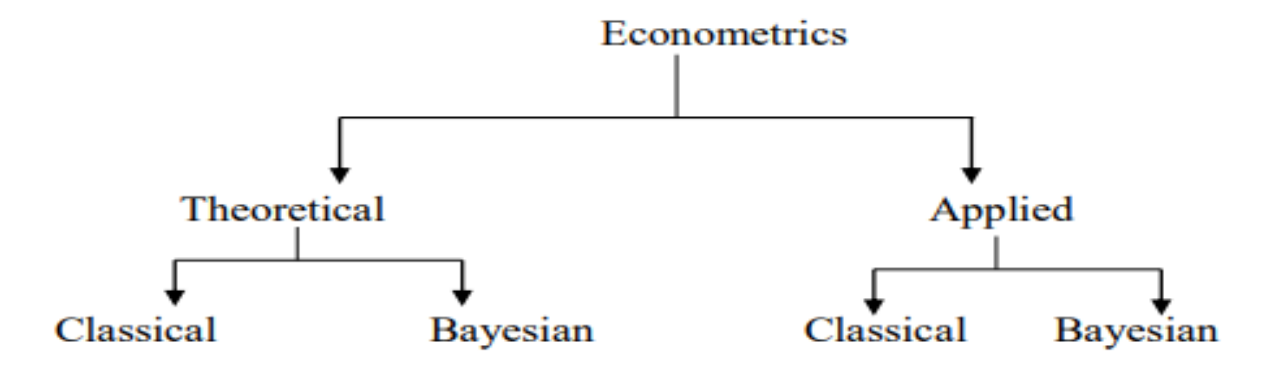
Use of the Model for control or Policy purpose :



Note :

1. Milton Friedmen has developed a model of consumption theory permanent income hypothesis.
2. Robert Hall has developed a model of consumption as life cycle permanent income hypothesis.

Types of Econometrics :



1. Theoretical econ is concerned with the development of appropriate methods of measuring economic relationship specified by economic models.

2. Applied econ uses the tool of theoretical econ to study some special fields of eco and business, such as production function etc.

Illustrative Examples Production and Cost Analysis :

The **production function** represents the functional relationship between a firm's physical inputs and its output during a specific time period. Mathematically, it's expressed as:

Q = f(A, B, C, D)

Where:

- **Q** = Quantity of output
- **A, B, C, D** = Inputs like land, labor, capital, and organization
- Output (**Q**) is the **dependent variable**
- Inputs (**A, B, C, D**) are **independent variables**

Mathematical Expression

To capture the **quantitative impact** of inputs on output, a simplified linear equation is used:

Y = a + bX

Where:

- **Y** = Output
- **X** = A single input factor
- **a** = Intercept
- **b** = Rate of change in output per unit of input (constant relationship)

Importance of the Production Function

1. **Estimation of Output**: Helps calculate output when inputs are given in physical units.

2. **Input Combinations**: Identifies combinations of inputs that produce the same output.
3. **Substitution Guidance**: Shows how one input can replace another without changing output.
4. **Cost Efficiency**: Aids in selecting the least-cost input mix for a desired output level (when prices are considered).
5. **Input–Output Laws**: Explains:
    a. **Law of Variable Proportions**: Short-run behavior with varying variable inputs
    b. **Law of Returns to Scale**: Long-run behavior with proportional changes in all inputs
6. **Efficiency Insights**: Determines max output from given inputs or minimum inputs needed for target output

Flexibility and Application

- Can be tailored to a **specific firm**, **industry**, or the **overall economy**
- Evolves with **technological advancements**, affecting input–output relationships

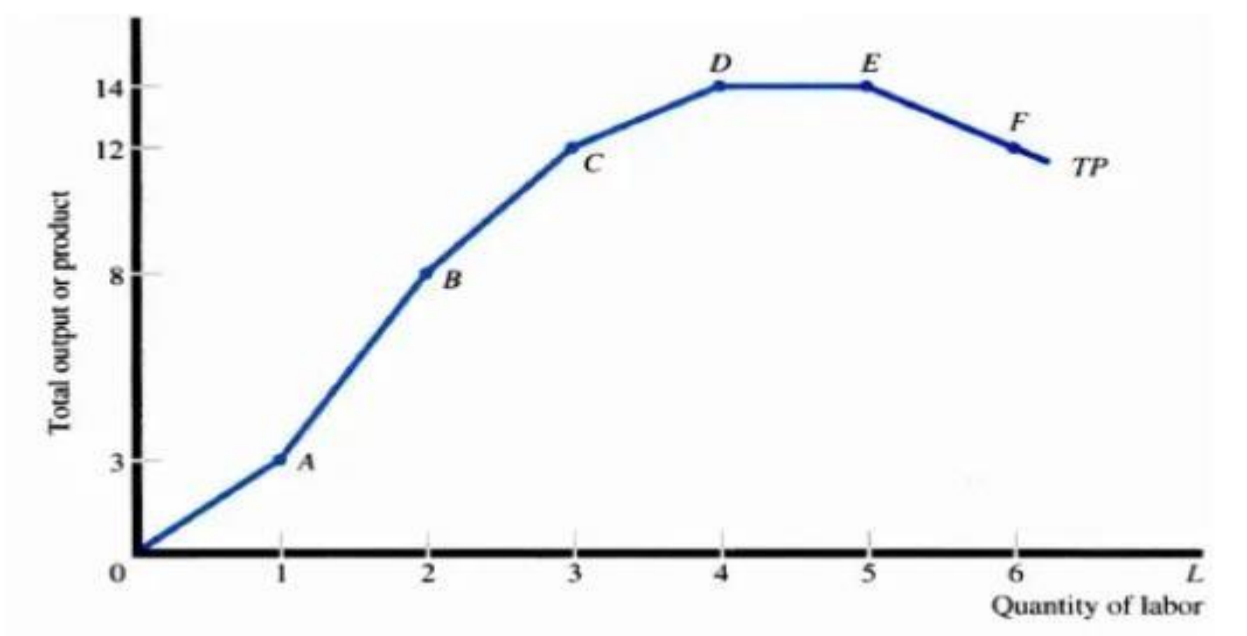
Key Assumptions

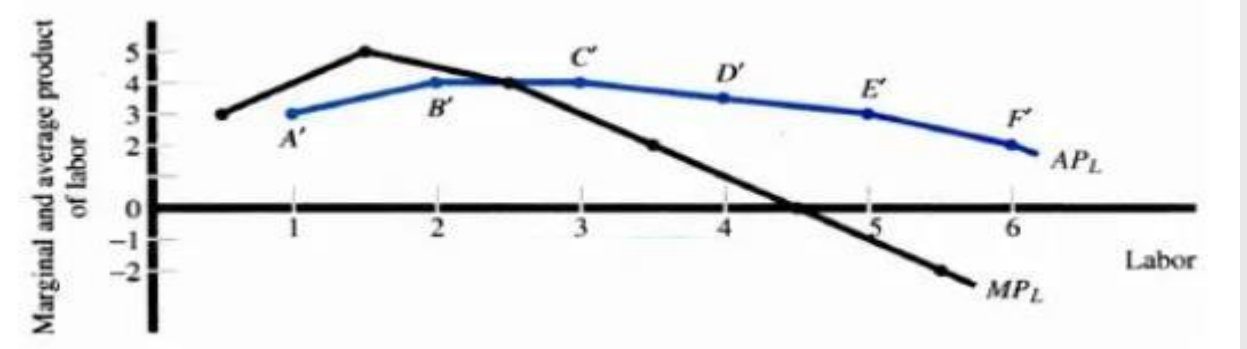
1. Applicable to a specific time frame
2. Technology remains constant
3. Firm uses optimal techniques
4. Inputs are divisible
5. Can model both short-run and long-run scenarios.

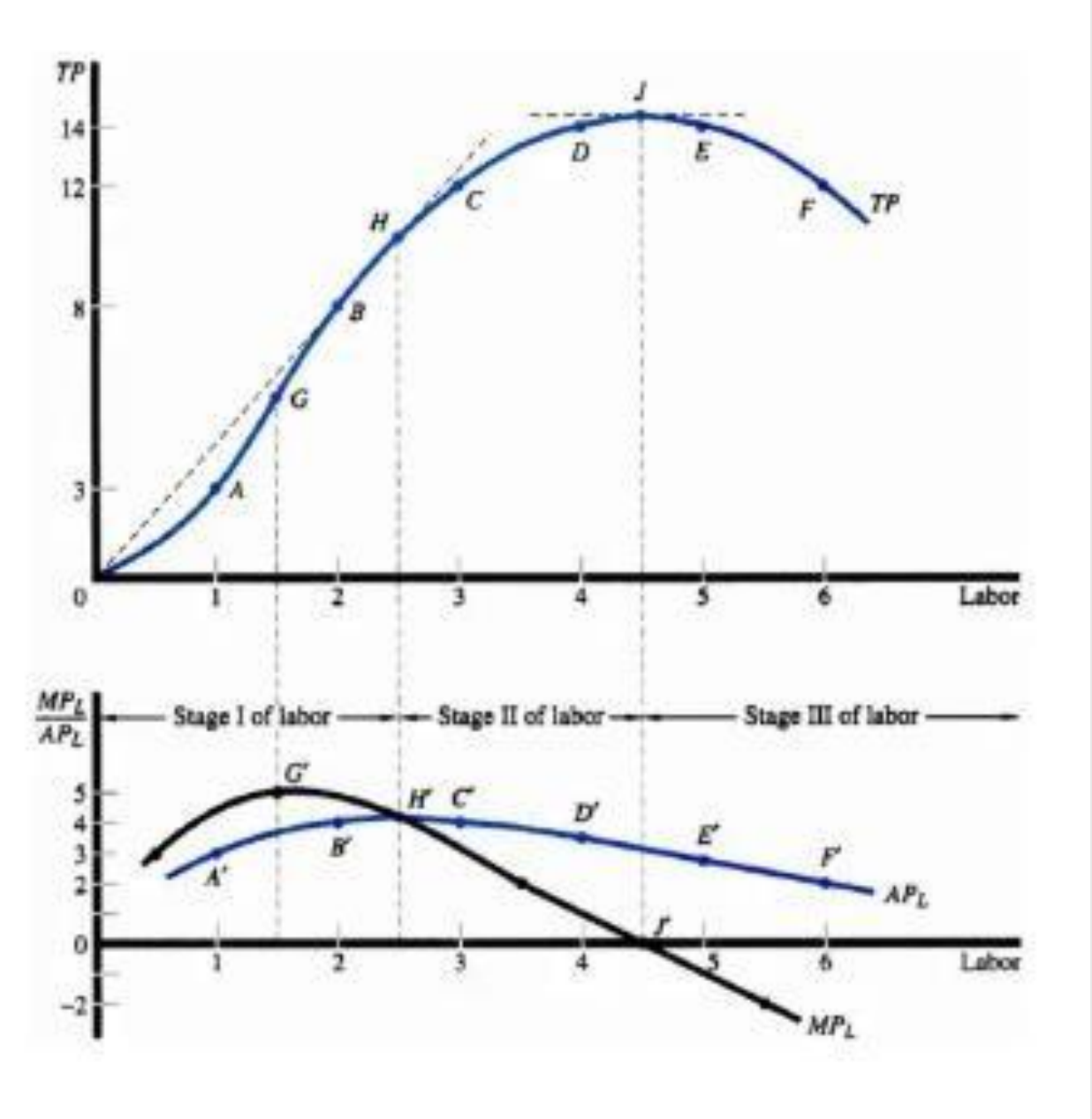
LAW OF DIMINISHING RETURNS / STAGES OF PRODUCTION :

As more and more units of a variable input are applied to a fixed input, the marginal product of the variable input will eventually decline.

Variable Input (X)	Total Product (TP)	Marginal Product (MP)	Average Product (AP)
0	0		
1	8	8	8
2	18	10	9
3	29	11	9.7
4	39	10	9.8
5	47	8	9.4
6	52	5	8.7
7	56	4	8
8	52	-4	6.5







- Diminishing returns are illustrated using both a numerical example (table) and a graphical representation (figure).

- The **marginal product of the first unit** of input is **8**.
- Marginal product reaches its **maximum of 11** between the **second and third units** of input.
- At **2.5 units of input**, the law of diminishing returns begins to take effect.
- The law states: When additional units of a **variable input** are combined with a **fixed input**, the **marginal product** will eventually start to **decline**.

- Managers typically identify the point of diminishing returns through **experience** and **trial-and-error**.

Reasons for the Occurrence of Diminishing Returns

1. In the earlier numerical example:
    - **a.** When **no workers** are employed, **total product = 0**
    - **b.** The **first worker** produces **8 units**
    - c. His **marginal product (MP)** = 8
    - d. His **average product (AP)** = 8
2. When **two workers** are employed:
    - a. Their **combined efforts** increase output
    - b. **MP of the second worker** is **greater** than that of the first
3. This situation reflects:
    - a. **Teamwork and specialization** among workers
    - **b.** Leading to **increasing returns**
4. As more workers are added:
    - **a.** Opportunities for specialization **decrease**
    - **b.** MP starts to **decline**
5. When MP begins to fall:
    - a. **Law of Diminishing Returns** sets in
    - **b.** Every additional unit of input contributes **less output**

Summary of Theory of Production and Costs

Basic Concepts of Production Theory

- **Production**: The process of converting inputs (labor, capital, land, etc.) into outputs (goods/services). Examples include manufacturing cars, refining gasoline, or providing services like education and healthcare.
- **Inputs**: Classified as fixed (e.g., machinery, buildings) or variable (e.g., labor, raw materials). Fixed inputs cannot be easily adjusted in the short run, while variable inputs can.

- **Short-run vs. Long-run**: In the short run, some inputs are fixed; in the long run, all inputs are variable, allowing firms to adjust production capacity.

Production Function

- **Definition**: A mathematical relationship showing the maximum output achievable from given inputs (e.g., $Q = f(L, K)$).
- **Efficiency**:
  - **Technical Efficiency**: Maximizing output for a given input combination.
  - **Economic Efficiency**: Minimizing cost for a given output level, considering input prices.

Short-Run Production Analysis

- **Key Metrics**:
  - **Total Product (TP)**: Total output produced.
  - **Average Product (AP)**: Output per unit of input (e.g., $AP = Q/L$).
  - **Marginal Product (MP)**: Additional output from one more unit of input (e.g., $MP = \text{Delta } Q/\text{Delta } L$).
- **Law of Diminishing Marginal Returns (LDMR)**: As more variable inputs are added to fixed inputs, MP eventually declines.

Long-Run Production Analysis

- **Isoquants**: Curves showing input combinations that produce the same output level. Characteristics include:
  - Downward slope (input substitutability).
  - Convex shape (diminishing Marginal Rate of Technical Substitution, MRTS).
- **Isocosts**: Lines showing input combinations with the same total cost. Slope reflects input price ratios ($-w/r$).
- **Optimal Input Combination**: Achieved where isoquant is tangent to isocost ($MRTS = w/r$).

Returns to Scale

- **Increasing Returns to Scale (IRS)**: Output increases more than proportionally to input increases.
- **Constant Returns to Scale (CRS)**: Output increases proportionally with inputs.

- **Decreasing Returns to Scale (DRS)**: Output increases less than proportionally.

Theory of Costs

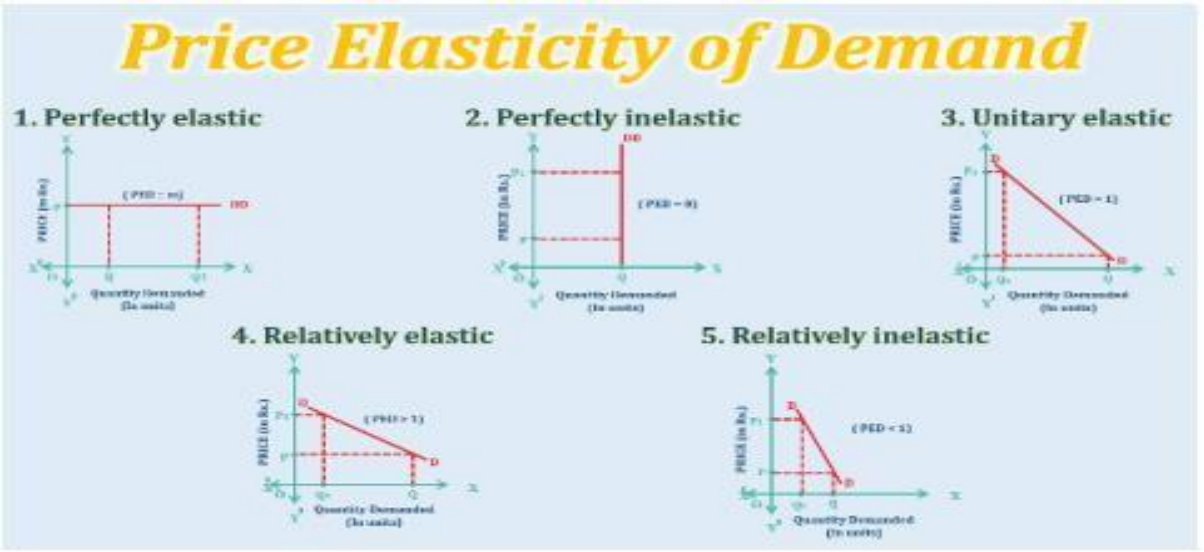
- **Cost Types**:
  - **Actual vs. Opportunity Cost**: Out-of-pocket expenses vs. forgone alternatives.
  - **Explicit vs. Implicit Costs**: Direct payments vs. imputed costs of owned resources.
  - **Short-run Costs**: Include fixed (TFC) and variable costs (TVC). Total Cost (TC) = TFC + TVC.
  - **Long-run Costs**: All costs are variable; firms plan optimal plant sizes.
- **Cost Curves**:
  - **Short-run**: U-shaped AVC, ATC, and MC curves. MC intersects AVC and ATC at their minima.
  - **Long-run**: U-shaped LAC curve, derived from short-run curves, representing economies/diseconomies of scale.
- **Economies of Scope**: Cost savings from joint production of multiple goods.

Key Takeaways

- Firms optimize production by balancing input usage (short-run) and scaling operations (long-run).
- Cost minimization involves understanding marginal products, input prices, and efficiency.
- Long-run planning considers returns to scale and cost structures to achieve competitive advantage.

Price and Income Elastic Ties of demand :

Elasticity of demand measures the responsiveness of the quantity demanded of a good to changes in factors such as price, income, or the price of related goods. Unlike the law of demand, which only indicates the direction of change, elasticity quantifies the magnitude of the response.



.

Concept of Elasticity of Demand

Elasticity is calculated as the percentage change in quantity demanded divided by the percentage change in an independent variable (price, income, or price of related goods). There are three main types:

1. Price Elasticity of Demand (PED)

Measures how demand responds to price changes.

Formula:

$$P_{ed} = \frac{\Delta D}{\Delta P} \times \frac{P}{D}$$

Interpretation:

- $P_{ed} = 0$: **Perfectly inelastic** (no change in demand).
- $P_{ed} = \infty$: **Perfectly elastic** (infinite change in demand).
- $P_{ed} = 1$: **Unitary elastic** (proportional change).
- $P_{ed} > 1$: **Relatively elastic** (demand changes more than price).
- $0 < P_{ed} < 1$: **Relatively inelastic** (demand changes less than price).

2. Income Elasticity of Demand (YED)

Measures how demand responds to income changes.

Formula:

$$Y_{ed} = \frac{\Delta D}{\Delta Y} \times \frac{Y}{D}$$

Interpretation:

- $Y_{ed} > 0$: **Normal good** (demand rises with income).
- $Y_{ed} < 0$: **Inferior good** (demand falls with income).
- $Y_{ed} = 0$: **No relationship** (e.g., necessities).

Cross-Elasticity of Demand (CED)

Measures how demand responds to the price of related goods.

Formula:

$$C_{ed} = \frac{\Delta D_x}{\Delta P_y} \times \frac{P_y}{D_x}$$

Interpretation:

- $C_{ed} > 0$: **Substitutes** (e.g., tea and coffee).
- $C_{ed} < 0$: **Complements** (e.g., cars and petrol).

Measurement Methods

1. Point Method

Used for small changes in price and quantity.

2. Outlay Method

Determines elasticity by observing changes in total expenditure:

Outlay constant: Unitary elasticity.

Outlay increases: Elastic demand.

Outlay decreases: Inelastic demand.

3. Geometrical Method

Measures elasticity at specific points on a demand curve using tangents.

Determinants of Price Elasticity

Nature of the Commodity: Necessities are less elastic; luxuries are more elastic.

Substitutes: More substitutes → higher elasticity.

Number of Uses: More uses → higher elasticity.

Price Level: Extremely high or low prices → lower elasticity.

Time Period: Longer periods → higher elasticity.

Importance of Elasticity

1. Pricing Decisions

Monopolists set prices based on elasticity to maximize revenue.

2. Government Policies

Price Support: Prevents prices from falling too low (e.g., agricultural products).

Ceiling Price: Caps prices to protect consumers (e.g., essential goods).

3. Tax Incidence

Determines how tax burdens are shared between consumers and producers. Higher demand elasticity shifts more burden to producers.

Key

- Elasticity quantifies demand sensitivity to price, income, or related goods.
- Measurement methods include point, outlay, and geometrical approaches.
- Factors like substitutes, necessity, and price level influence elasticity.
- Applications include pricing strategies, policy-making, and tax analysis.

Notes :

- Perfectly Elastic/Inelastic: Extreme responsiveness or no responsiveness.
- Substitutes/Complements: Goods with positive/negative cross-elasticity.
- Rectangular Hyperbola: Demand curve with constant unitary elasticity.

Single Equation – Ordinary Least Square (OLS) Method, Maximum Likelihood Estimate (MLE)

Single Equation Methods in Econometrics

Single equation methods are foundational techniques in econometrics used to estimate relationships where one dependent variable is explained by one or more independent variables. Unlike system-wide approaches (e.g., simultaneous equations), these methods focus on one equation at a time, making them simpler and computationally efficient. However, their applicability depends on the absence of endogeneity,when explanatory variables are uncorrelated with the error term. Below are key single-equation methods and their roles in econometric analysis.

Ordinary Least Squares (OLS)

OLS is the most basic and widely used method for estimating linear relationships. It minimizes the sum of squared residuals to derive coefficients that best fit the data.

- **Assumptions**:
    - Linearity
    - Exogeneity (no correlation between regressors and error term)
    - Homoskedasticity (constant error variance)
    - No autocorrelation
- **Limitations**:

    - Biased and inconsistent if regressors are endogenous (e.g., due to omitted variables, measurement error, or simultaneity).

Generalized Least Squares (GLS)

GLS extends OLS to handle heteroskedasticity or autocorrelation by weighting observations to improve efficiency.

- **Use Case**: When error terms violate OLS assumptions (e.g., time-series data with serial correlation).
- **Transformation**: Applies a weighting matrix to "whiten" the errors.

Instrumental Variables (IV) Regression

IV addresses endogeneity by replacing correlated regressors with instruments—variables that are:

- **Relevant**: Strongly correlated with endogenous regressors.
- **Exogenous**: Uncorrelated with the error term.
- **Example**: Using rainfall as an instrument for agricultural productivity when studying its impact on GDP.
- **Challenges**: Weak instruments can lead to biased estimates.

Two-Stage Least Squares (2SLS)

A special case of IV, 2SLS is used for over-identified systems (more instruments than endogenous variables).

- **Stages**:
    - Regress endogenous variables on instruments (first stage).
    - Replace endogenous variables with their predicted values and re-estimate via OLS (second stage).
- **Advantage**: More efficient than simple IV when multiple instruments are available.

Limited-Information Maximum Likelihood (LIML)

An alternative to 2SLS, LIML minimizes the variance ratio of residuals from restricted and unrestricted models.

- **Advantage**: Less sensitive to weak instruments than 2SLS.
- **Disadvantage**: Computationally intensive.

k-Class Estimators

A generalization of 2SLS and OLS, where estimates are weighted between OLS ((k=0)) and 2SLS ((k=1)).

- **Use Case**: When instrument strength is uncertain.

When to Use Single-Equation Methods?

- **OLS**: When regressors are exogenous (e.g., controlled experiments).
- **IV/2SLS**: When endogeneity is present (e.g., demand-supply models).
- **GLS**: For heteroskedastic or autocorrelated data.
- **LIML/k-Class**: For weak-instrument robustness.

Limitations

- Single-equation methods ignore cross-equation dependencies, making them unsuitable for simultaneous systems unless properly adjusted (e.g., via IV).
- Reliance on strong assumptions (e.g., exogeneity) that may not hold in real-world data.

Conclusion

Single-equation methods provide a flexible toolkit for econometric analysis, balancing simplicity and rigor. While OLS is the workhorse, IV and 2SLS are essential for causal inference in the presence of endogeneity. Choosing the right method depends on the data structure, identification conditions, and the strength of available instruments.

Remark

Estimation of Simultaneous Equation Models

Simultaneous equation models present unique challenges due to the interdependence between variables, where endogenous variables influence each other within the system. Traditional single-equation methods like Ordinary Least Squares (OLS) fail here because

they ignore this simultaneity, leading to biased and inconsistent estimates. To address this, specialized techniques such as Indirect Least Squares (ILS), Instrumental Variables (IV), and Two-Stage Least Squares (2SLS) are employed. These methods account for endogeneity by leveraging reduced-form equations or exogenous instruments, ensuring consistent parameter estimates.

The **Indirect Least Squares (ILS) method** is suitable for exactly identified equations. It first estimates reduced-form equations via OLS, then derives structural parameters from these estimates. While ILS is straightforward, its reliance on exact identification limits its applicability. For over-identified systems, ILS fails to provide unique solutions, necessitating alternative approaches like the **Instrumental Variables (IV) method**. IV replaces endogenous variables with exogenous instruments uncorrelated with the error term, yielding consistent estimates. However, IV's effectiveness hinges on the quality of instruments, which must be strongly correlated with endogenous variables but uncorrelated with errors.

The **Two-Stage Least Squares (2SLS) method** extends IV by addressing over-identification. In the first stage, endogenous variables are regressed on all exogenous variables to obtain predicted values. These predicted values, which are purged of endogeneity, replace the original endogenous variables in the second-stage OLS regression. 2SLS is widely favored for its balance of simplicity and robustness, particularly in over-identified systems. Its performance, however, depends on the strength of the first-stage regressions; high (R^2) values indicate reliable instruments, while low values render the estimates meaningless.

For more complex scenarios, the **Limited-Information Maximum Likelihood (LIML) method** offers an alternative. LIML minimizes the variance ratio of residuals from restricted and unrestricted reduced-form regressions, providing estimates invariant to normalization. Though computationally intensive, LIML is asymptotically equivalent to 2SLS and is preferred when exact identification or normalization issues arise. Its reliance on full system information makes it less sensitive to specification errors compared to single-equation methods.

The **k-Class estimator** bridges OLS and 2SLS by adjusting endogenous variables with a scalar (k). When (k = 1), it replicates 2SLS, and when (k = 0), it reduces to OLS. This flexibility allows researchers to balance bias and efficiency, though the choice of (k) requires careful consideration. The k-Class estimator is particularly useful in scenarios where instrument strength is uncertain, offering a middle ground between biased OLS and consistent but potentially inefficient 2SLS estimates.

Choosing the right method depends on the identification status of the equations. For **exactly identified systems**, ILS, IV, and 2SLS yield identical results, with ILS being the simplest. For **over-identified systems**, 2SLS is preferred due to its robustness and single-solution property. LIML is reserved for cases requiring invariance to normalization or when system-wide efficiency is critical. The k-Class estimator provides flexibility but is less commonly used due to its dependency on the arbitrary choice of (k).

In practice, 2SLS dominates empirical work due to its computational ease and reliability. However, researchers must validate instrument strength and ensure model correctness, as weak instruments or misspecification can undermine results. Understanding these methods' assumptions and properties is crucial for accurate estimation in simultaneous equation models, enabling robust policy analysis and economic forecasting.

General linear model (GLM) and its extensions

The General Linear Model (GLM) is a foundational statistical framework used to analyze the relationship between a continuous dependent variable and one or more independent variables, which can be either continuous or categorical. It serves as a unifying approach for many common statistical techniques, including linear regression, analysis of variance (ANOVA), and analysis of covariance (ANCOVA). The GLM is highly flexible, allowing researchers to test hypotheses, predict outcomes, and assess the significance of predictor variables in a structured manner.

Core Structure of the GLM

The GLM is expressed by the equation:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

where:

- $\mathbf{y}$ is the vector of observed dependent variable values,
- $\mathbf{X}$ is the design matrix containing independent variables,
- $\mathbf{b}$ is the vector of coefficients (regression weights),
- $\mathbf{e}$ is the vector of residuals (prediction errors).

The model assumes that the relationship between predictors and the outcome is linear, errors (ee) are normally distributed with a mean of zero, and observations are independent with constant variance (homoscedasticity).

Estimation and Hypothesis Testing

The coefficients bb are typically estimated using the ordinary least squares (OLS) method, minimizing the sum of squared residuals:

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Once estimated, the model's fit is evaluated using **sums of squares** (SS):

- **Total SS** – Total variability in $\mathbf{y}$,
- **Model SS** – Variability explained by predictors,
- **Residual SS** – Unexplained variability.

An **F-test** compares the explained and unexplained variance to determine if the model is statistically significant:

$$F = \frac{SS_{\text{Model}}/K}{SS_{\text{Residual}}/(N-K-1)}$$

where K is the number of predictors and N is the sample size.

Applications and Extensions

The GLM can be adapted for various analyses:

- Regression – Predicting a continuous outcome from continuous predictors.
- ANOVA – Comparing group means using categorical predictors.
- ANCOVA – Combining continuous and categorical predictors.

The GLM has limitations:

Fixed Effects Only – It does not handle random effects, requiring mixed-effects models for hierarchical data.

Linearity Assumption – Nonlinear relationships need generalized linear models (GLMs) or polynomial terms.

Full-Rank Design Matrix – Collinearity can be addressed using regularization or generalized inverses.

Conclusion

The GLM is a powerful and versatile tool for statistical modeling, providing a unified approach to regression, ANOVA, and related techniques. While it has some constraints, extensions like mixed-effects models and generalized linear models broaden its applicability. Understanding the GLM's assumptions, estimation methods, and testing procedures is essential for accurate data analysis in research across psychology, medicine, economics, and beyond.

Generalized least squares (GLS) Estimation and Prediction

The Generalized Least Squares (GLS) method extends the classical linear regression framework to handle cases where the error terms exhibit heteroskedasticity or serial correlation. Unlike the Ordinary Least Squares (OLS) estimator, which assumes a scalar covariance matrix for the errors, GLS accounts for more general covariance structures, improving efficiency and ensuring valid statistical inferences. The GLS estimator is derived by transforming the original model so that the transformed errors meet the classical assumptions, allowing OLS to be applied to the transformed data. However, GLS requires knowledge of the true covariance matrix, which is rarely available in practice, leading to the use of Feasible GLS (FGLS) estimators that rely on estimated covariance structures.

Key Concepts and Methodology

The GLS approach begins with the linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{var}(\mathbf{e}) = \boldsymbol{\Sigma}_o$, a positive definite but non-scalar matrix. The GLS estimator is given by:

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}^\top \boldsymbol{\Sigma}_o^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Sigma}_o^{-1} \mathbf{y},$$

which minimizes the generalized sum of squared errors. When $\boldsymbol{\Sigma}_o$ is unknown, FGLS substitutes an estimate $\hat{\boldsymbol{\Sigma}}_o$, often derived from OLS residuals. However, FGLS introduces complexities, such as the need for correctly specified covariance structures and the lack of finite-sample properties, making asymptotic theory essential for inference.

Applications and Limitations

GLS is particularly useful in two scenarios:

Heteroskedasticity: When error variances differ across observations, GLS weights data inversely to their variances, yielding more efficient estimates. Tests like the Goldfeld-Quandt test help detect heteroskedasticity.

Serial Correlation: In time-series data, errors may correlate over time (e.g., AR(1) disturbances). The Durbin-Watson test detects serial correlation, while FGLS adjusts for it using transformations like Cochrane-Orcutt.

Despite its advantages, GLS has limitations:

Implementation Challenges: FGLS requires plausible assumptions about ΣoΣo , which may be misspecified.

Finite-Sample Uncertainty: The properties of FGLS estimators are typically only known asymptotically.

Dependence on Correct Specification: Incorrect covariance structures can render FGLS less efficient than OLS.

Extensions and Related Models

The GLS framework extends to various advanced models:

Seemingly Unrelated Regressions (SUR): For systems of equations with correlated errors across equations.

Panel Data Models: Including fixed-effects and random-effects models to account for unobserved heterogeneity.

Linear Probability Models: Though GLS can address heteroskedasticity in binary outcomes, logistic regression is often more appropriate.

Remark

GLS and FGLS provide powerful tools for addressing non-spherical error structures in linear regression, enhancing estimation efficiency and inference validity. However, their practical application requires careful consideration of covariance specifications and reliance on large-sample properties. Understanding these methods' assumptions and limitations is crucial for their effective use in econometric analysis. For further exploration, consult advanced texts like Amemiya (1985) and Greene (2000), which delve deeper into GLS theory and applications.

Heteroscedasticity Disturbances, Pure and Mixed Estimation

Heteroscedasticity occurs when the variance of error terms in a regression model is not constant across observations, a common issue in cross-sectional data. For example, in consumer budget studies, residual variance often increases with income, while in firm-level analyses, it may rise with firm size. The standard Ordinary Least Squares (OLS) estimator becomes inefficient under heteroscedasticity, as it assumes homoscedastic errors. To address this, the **Generalized Least Squares (GLS)** method is employed. GLS transforms the model by weighting observations inversely to their variances, yielding the estimator:

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Omega}^{-1} \mathbf{y},$$

where $\mathbf{\Omega}$ is a diagonal matrix of known weights λ_i. This estimator is **Best Linear Unbiased (BLUE)** and has a variance-covariance matrix:

$$\text{Var}(\mathbf{b}) = \sigma^2 (\mathbf{X}^\top \mathbf{\Omega}^{-1} \mathbf{X})^{-1}.$$

When λ_i are unknown, **Feasible GLS (FGLS)** substitutes estimates, often derived from OLS residuals. For instance, if variance scales with X_i^2, weights $\lambda_i = 1/X_i^2$ are used. Simulations show GLS can significantly outperform OLS in efficiency—e.g., achieving 56% lower variance for slope estimates in specific cases.

Pure and Mixed Estimation in Econometrics:

Pure and Mixed Estimation: Incorporating Prior Information

In econometrics, prior information—such as estimates from previous studies or theoretical constraints—can enhance parameter estimation. Theil and Goldberger formalized this by combining sample data with prior knowledge in a **mixed estimation** framework. The prior is expressed as:

$$\mathbf{r} = \mathbf{R}\boldsymbol{\beta} + \mathbf{v}, \quad \text{where} \quad \mathbf{v} \sim (0, \mathbf{\Psi}),$$

with $\mathbf{r}$ (known vector), $\mathbf{R}$ (known matrix), and $\mathbf{\Psi}$ (known covariance). For example, prior estimates of income elasticity (β_3) with mean 0.5 and variance $1/16$ can be encoded as:

$$\mathbf{r} = 0.5, \quad \mathbf{R} = [0\ 0\ 1\ 0\ 0], \quad \mathbf{\Psi} = 1/16.$$

Combining this with the sample model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, the mixed estimator becomes:

$$\mathbf{b} = (\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \mathbf{R}^\top\mathbf{\Psi}^{-1}\mathbf{R})^{-1}(\sigma^{-2}\mathbf{X}^\top\mathbf{y} + \mathbf{R}^\top\mathbf{\Psi}^{-1}\mathbf{r}),$$

Key Insight:

Mixed estimation balances sample evidence with prior knowledge, offering a flexible tool for models with limited data or strong theoretical foundations.

Practical Implications and Limitations

1. GLS/FGLS: Requires correct specification of ΩΩ. Misspecification can worsen efficiency.
2. Mixed Estimation: Effective when priors are reliable. Poor priors may introduce bias.
3. Computational Complexity: Both methods involve matrix inversions, demanding larger samples for stability.
