LINEAR MODELS AND APPLIED REGRESSION ANALYSIS

M.Sc., STATISTICS First Year

SEMESTER-II, PAPER-IV

LESSON WRITERS

Prof. V. V. Haragopal
Professor of Statistics
BITS Pilani, Jawaharnagar
Hyderabad - 500078

Prof. G. V. S. R. Anjaneyulu
Professor of Statistics (Retd.)
Acharya Nagarjuna University

Dr. G. Madhu Sudan
Assistant Professor
Department of Statistics
University of Allahabad
Prayagraj -211002

Dr: U. Ramkiran
Department of Statistics
Acharya Nagarjuna University

EDITOR

Prof. V. V. Haragopal
Professor of Statistics
BITS Pilani, Jawaharnagar
Hyderabad - 500078

ACADEMIC ADVISOR

Prof. G. V. S. R. Anjaneyulu
Professor of Statistics (Retd.)
Acharya Nagarjuna University

DIRECTOR, I/c.

Prof. V. Venkateswarlu
M.A., M.P.S., M.S.W., M.Phil., Ph.D.

Centre for Distance Education
Acharya Nagarjuna University
Nagarjuna Nagar 522 510

Ph: 0863-2346222, 2346208
0863- 2346259 (Study Material)
Website www.anucde.info
E-mail: anucdedirector@gmail.com

M.Sc., STATISTICS : Linear Models and Applied Regression Analysis

First Edition : 2025

No. of Copies :

© Acharya Nagarjuna University

This book is exclusively prepared for the use of students of M.Sc., STATISTICS Centre for Distance Education, Acharya Nagarjuna University and this book is meant for limited circulation only.

Published by:

Prof. V. VENKATESWARLU
Director, I/c
Centre for Distance Education,
Acharya Nagarjuna University

Printed at:

FOREWORD

Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A+' grade from the NAAC in the year 2024, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over 221 affiliated colleges spread over the two districts of Guntur and Prakasam.

The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the door step of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.Sc., B.A., B.B.A., and B.Com courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.

To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.

It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson-writers of the Centre who have helped in these endeavors.

Prof. K. Gangadhara Rao
M.Tech., Ph.D.,
Vice-Chancellor I/c
Acharya Nagarjuna University.

M.Sc. – Statistics Syllabus
SEMESTER–II
204ST24: Linear Models and Applied Regression Analysis

UNIT-I

Gauss-Markov set-up, Normal equations and Least squares estimates, variances and covariances of least squares estimates, estimation of error variance.

UNIT-II

Estimation with correlated observations, least squares estimates with restriction on parameters, simultaneous estimates of linear parametric functions.

UNIT-III

Tests of hypotheses for one and more than one linear parametric functions, confidence intervals and regions. Analysis of Variance.

UNIT-IV

Simple linear regression, multiple regression, fit of polynomials and use of orthogonal polynomials.

UNIT-V

Multicollinearity, Ridge regression and principal component regression, subset selection of explanatory variables.

BOOKS FOR STUDY:

1) Graybill, F.A. (1983): Matrices with Applications in Statistics. Wadsworth.
2) Draper, N.R. and Smith, H. (1998): Applied Regression Analysis. 3rd Edition. Wiley-Blackwell.
3) Douglas C. Montgomery, Elizabeth A. Peck and G. Geoffrey Vining (2012): Introduction to Linear Regression Analysis – 5th Edition. Wiley.
4) Goon, Gupta and Das Gupta (2003): An outline of Statistical Theory. Volume II. The World Press Pvt. Ltd.

BOOKS FOR REFERENCES:

1) Bapat, R.B. (2012): Linear Algebra and Linear Models. 3rd Edition. Springer.
2) Cook, R.D. and Weisberg, S. (1983): Residual and Influence in Regression. 1st Edition. Chapman and Hall.
3) Johnson, J. (1996): Econometric Methods. 4th Edition. McGraw Hill.
4) Rao, C.R. (2002): Linear Statistical Inference and Its Applications. 2nd Edition. Wiley-Blackwell.
5) Weisberg, S. (2013): Applied Linear Regression. 4th Edition. Wiley.

SET-I **MODEL QUESTION PAPER** **(204ST24)**

M.Sc. DEGREE EXAMINATION

SECOND SEMESTER

STATISTICS

204ST24 :: LINEAR MODELS AND APPLIED REGRESSION ANALYSIS

(w.e.f. 2024-2025 Academic Year admitted batch)

Time: 3 hours **Maximum: 70 marks**

ANSWER ONE QUESTION FROM EACH UNIT

(Each question carries equal marks)

UNIT–I

1. (a) Explain the Gauss–Markov model along with underlying assumptions. Also give an example of the same.
   (b) Obtain the maximum likelihood estimator of error variance in the normal linear regression model. Derive its expectation and variance.
2. (a) Show that in a general linear model the least squares estimators are BLUE.
   (b) Derive an expression for the dispersion (variance–covariance) matrix of the BLUE for the parameter vector of the general linear model.

UNIT–II

3. (a) Explain restricted least squares estimation for a general linear model and derive the restricted LS estimator.
   (b) State and prove Aitken's theorem.
4. (a) Give the simultaneous estimates of linear parametric functions in a general linear model.
   (b) Define estimable function and explain how estimation is modified when observations are correlated.

UNIT–III

5. (a) Describe the test procedure to test the significance of a single parametric function.
   (b) Explain the general linear test procedure to test the significance of multiple hypotheses with an example.

(OR)

6. (a) Explain analysis of variance for two–way classification with multiple observations per cell. Obtain the ANOVA table.
   (b) Obtain confidence intervals for the least squares estimates in the case of a two-variable linear model.

UNIT–IV

7. (a) What is simple linear regression? Explain with a suitable example and obtain the partial correlation coefficient.
   (b) Explain the multiple regression in the three–variable case and derive the coefficient of multiple determination.

(OR)

8. (a) Define polynomial regression and explain the use of orthogonal polynomials with an example.
   (b) Write down and explain the sampling properties of regression coefficients.

UNIT–V

9. (a) Explain multicollinearity with suitable examples. What are the consequences of multicollinearity and how can it be detected?
   (b) What are ridge regression estimators? Discuss their properties and compare with ordinary least squares.

(OR)

10. (a) What are principal components? Explain their use in regression analysis by a suitable example.
    (b) Explain subset selection of explanatory variables and compare major subset selection procedures.

CONTENTS

LESSON-1
GAUSS-MARKOV SET-UP

OBJECTIVES:

After completing this lesson, students will be able to:

- Understand the Gauss–Markov linear model and its underlying assumptions.
- Formulate regression problems using matrix notation**.**
- Derive normal equations and obtain least squares estimates of regression parameters.
- Explain and apply the Gauss–Markov theorem and interpret the concept of BLUE**.**
- Compute and interpret variances and covariances of least squares estimators.
- Estimate the error variance and assess model adequacy.
- Apply linear regression techniques to real-world data in social sciences, engineering, economics, and health sciences.
- Use statistical software to fit and interpret applied regression models**.**

STRUCTURE:

1.1 INTRODUCTION:

The Gauss–Markov set-up forms the theoretical foundation of classical linear regression analysis. It provides a rigorous framework for estimation and inference in linear statistical models under minimal distributional assumptions. Central to this theory is the Gauss–Markov Theorem, which establishes the optimality of the Ordinary Least Squares (OLS) estimator within the class of linear unbiased estimators. This framework underpins much of modern econometrics, biostatistics, engineering analysis, and social science research.

Description: Gauss–Markov Set-up

- The Gauss–Markov set-up forms the theoretical basis of linear regression analysis. It considers the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,
- Where the error terms have zero mean, constant variance, and are uncorrelated.
- Under these assumptions, the ordinary least squares method is used to estimate the unknown parameters. The resulting estimator is shown to be unbiased, and its variance–covariance matrix can be explicitly derived. The central result of this framework is the Gauss–Markov Theorem, which states that the least squares estimator is the Best Linear Unbiased Estimator (BLUE) of the parameter vector.
- The Gauss–Markov set-up provides the foundation for hypothesis testing, confidence intervals, and analysis of variance in linear models and serves as the basis for more advanced regression methods.

1.2. LINEAR STATISTICAL MODEL:

Let $y_1, y_2, \ldots, y_n$ denote observed responses. The general linear statistical model is

$$\mathrm{y} = X\beta + \varepsilon,$$

where:

- yis an $n \times 1$ vector of observations,
- $X$is an $n \times p$ known design matrix of full column rank p,
- $\beta$is a $p \times 1$ vector of unknown parameters,
- $\varepsilon$is an $n \times 1$ vector of random errors.

The systematic component $X\beta$ represents the mean structure, while εcaptures unexplained variability.

1.3 ASSUMPTIONS OF THE GAUSS-MARKOV MODEL:

The Gauss–Markov set-up relies on the following assumptions:

1. Linearity: The model is linear in parameters, $E(\mathrm{y}) = X\beta$.
2. Full Rank: $\mathrm{rank}(X) = p$, ensuring identifiability of parameters.
3. Unbiased Errors: $E(\varepsilon) = 0$.
4. Homoscedasticity: $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$, where $\sigma^2 > 0$.
5. No Correlation: Errors are uncorrelated.

Normality of errors is not required for the Gauss–Markov Theorem, but is often imposed for exact inference.

1.3.1 GAUSS-MARKOV THEOREM:

Statement: Consider the linear statistical model

$$\mathrm{y} = X\beta + \varepsilon,$$

where

$$E(\varepsilon) = 0, \mathrm{Var}\,(\varepsilon) = \sigma^2 I_n,$$

and X is an $n \times p$ known matrix of full column rank p.
Then the ordinary least squares estimator

$$\hat{\beta} = (X'X)^{-1}X'\mathrm{y}$$

has minimum variance among all estimators of βthat are linear and unbiased. Hence, $\hat{\beta}$ is the Best Linear Unbiased Estimator (BLUE).

Proof:

Step 1: Consider the class of linear estimators

Let $\tilde{\beta} = Ay,$
be any linear estimator of β, where $A$is a $p \times n$non-stochastic matrix.

Step 2: Impose the unbiasedness condition

$$E(\tilde{\beta}) = \beta.$$

Now,

$$E(\tilde{\beta}) = AE(\mathrm{y}) = AX\beta.$$

For this to hold for all β,

$$AX = I_p. \qquad (1)$$

Step 3: Variance of a linear unbiased estimator

Since $\mathrm{Var}\ (\mathrm{y}) = \sigma^2 I_n$,

$$\mathrm{Var}\ (\tilde{\beta}) = A\ \mathrm{Var}\ (\mathrm{y})\ A' = \sigma^2 AA'. \qquad (2)$$

Step 4: Variance of the OLS estimator

The OLS estimator is

$$\hat{\beta} = (X'X)^{-1}X'\mathrm{y}.$$

Therefore,

$$\mathrm{Var}\ (\hat{\beta}) = \sigma^2 (X'X)^{-1}. \qquad (3)$$

Step 5: Difference of variance–covariance matrices

From (2) and (3),

$$\mathrm{Var}\ (\tilde{\beta}) - \mathrm{Var}\ (\hat{\beta}) = \sigma^2 [AA' - (X'X)^{-1}]. \qquad (4)$$

Step 6: Algebraic decomposition

Define

$$B = A - (X'X)^{-1}X'.$$

Then

$$BB' = AA' - A(X'X)^{-1}X' - (X'X)^{-1}XA' + (X'X)^{-1}X'X(X'X)^{-1}.$$

Using the unbiasedness condition $AX = I_p$,

$$BB' = AA' - (X'X)^{-1}. \qquad (5)$$

Step 7: Positive semi-definiteness

Substituting (5) into (4),

$$\mathrm{Var}\ (\tilde{\beta}) - \mathrm{Var}\ (\hat{\beta}) = \sigma^2 BB' = \sigma^2 (A - (X'X)^{-1}X')(A - (X'X)^{-1}X')'.$$

For any vector c,

$$c'BB'c = (B'c)'(B'c) \geq 0.$$

Hence,

$$\text{Var}\,(\tilde{\beta}) - \text{Var}\,(\hat{\beta}) \succeq 0$$

Step 8: Conclusion

The variance–covariance matrix of any linear unbiased estimator $\tilde{\beta}$is greater than or equal to that of $\hat{\beta}$in the Loewner sense. Therefore,

$$\boxed{\hat{\beta} = (X'X)^{-1}X'y \text{ is the BLUE.}}$$

1.3.2 APPLICATIONS OF GAUSS-MARKOV THEORY:

- Econometric modeling
- Industrial process optimization
- Experimental design
- Biostatistical dose–response analysis
- Signal processing and calibration problems

1.4 MATRIX FORMULATION OF THE MODEL:

The expectation and variance of yare

$$E(y) = X\beta, \text{Var}(y) = \sigma^2 I_n.$$

Define the projection (hat) matrix:

$$H = X(X'X)^{-1}X'.$$

This matrix projects yonto the column space of X.

1.5 ORDINARY LEAST SQUARES ESTIMATION:

The OLS estimator minimizes the residual sum of squares

$$S(\beta) = (y - X\beta)'(y - X\beta).$$

Normal Equations

Differentiating and equating to zero yields

$$X'X\hat{\beta} = X'y.$$

OLS Estimator

Provided $X'X$is nonsingular,

$$\boxed{\hat{\beta} = (X'X)^{-1}X'y}$$

1.6 PROPERTIES OF LEAST SQUARES ESTIMATORS:

Under the Gauss–Markov assumptions:

1. Unbiasedness:

$$E(\hat{\beta}) = \beta.$$

2. Variance–Covariance Matrix:

$$\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

3. Linearity: $\hat{\beta}$is a linear function of y.

1.6.1 BEST LINEAR UNBIASED ESTIMATOR (BLUE):

The concept of the Best Linear Unbiased Estimator (BLUE) is central to the Gauss–Markov theory and classical linear regression analysis. It provides a precise optimality criterion for estimating the unknown parameter vector βin a linear statistical model.

An estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is called BLUE if it satisfies the following three properties:

1. Linearity

The estimator must be a linear function of the observed data y. That is,

$$\widehat{\beta} = A\text{y},$$

where A is a fixed (non-random) $p \times n$ matrix.

This requirement restricts attention to estimators that depend linearly on the observations.

2. Unbiasedness

The estimator must satisfy $E(\hat{\beta}) = \beta$.

Unbiasedness ensures that, on average, the estimator correctly targets the true parameter vector and does not systematically overestimate or underestimate it.

3. Best (Minimum Variance)

Among all estimators that are linear and unbiased, the estimator must have the minimum variance–covariance matrix.

Formally, if $\tilde{\beta}$ is any other linear unbiased estimator, then

$$\text{Var}\,(\tilde{\beta}) - \text{Var}\,(\hat{\beta}) \succeq 0,$$

meaning the difference is positive semi-definite.

Thus, BLUE is the most precise estimator within the specified class.

Class of Estimators Considered

The Gauss–Markov theorem restricts attention to the class of estimators

$$\mathcal{L} = \{A\text{y} : AX = I_p\}$$

This condition arises from unbiasedness:

$$E(A\text{y}) = AX\beta = \beta$$

Important implications:

- Nonlinear estimators are excluded.
- Biased estimators are excluded.
- Optimality is defined only within this class.

Why Ordinary Least Squares is BLUE

Under the Gauss–Markov assumptions:

$$E(\mathrm{y}) = X\beta, \mathrm{Var}\ (\mathrm{y}) = \sigma^2 I_n,$$

the ordinary least squares estimator is

$$\hat{\beta} = (X'X)^{-1}X'\mathrm{y}$$

It satisfies all three BLUE conditions:

1. Linear:
   $\hat{\beta}$is a linear function of y.
2. Unbiased:

$$E(\hat{\beta}) = (X'X)^{-1}X'X\beta = \beta.$$

3. Minimum Variance:

$$\mathrm{Var}\ (\hat{\beta}) = \sigma^2(X'X)^{-1},$$

which is smaller than or equal to the variance of any other linear unbiased estimator.

Hence, OLS coincides with the BLUE.

Interpretation of the Word "Best"

The term *best* refers strictly to variance efficiency, not to closeness in any single sample.

Specifically:

- For any linear function $\mathrm{c}'\beta, \mathrm{c}'\hat{\beta}$ has the smallest possible variance among all linear unbiased estimators of $\mathrm{c}'\beta$.
- No other linear unbiased estimator can uniformly dominate OLS in terms of precision.

BLUE and Distributional Assumptions

A crucial feature of BLUE is that:

- Normality of errors is not required.
- Only the first two moments of $\varepsilon$are used.

If, in addition,

$$\varepsilon \sim N(0, \sigma^2 I_n),$$

then OLS is not only BLUE but also the minimum variance unbiased estimator (MVUE) among *all* unbiased estimators.

BLUE and Generalized Least Squares (GLS)

When the assumption of homoscedastic and uncorrelated errors is violated, i.e.,

$$\mathrm{Var}\ (\varepsilon) = \sigma^2 V,$$

the OLS estimator is no longer BLUE.

In this case, the Generalized Least Squares (GLS) estimator

$$\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}\mathrm{y}$$

becomes the BLUE under the new covariance structure.

An estimator $\hat{\beta}$is called BLUE if:

- it is linear in y,
- it is unbiased,
- it has minimum variance among all such estimators.

Under Gauss–Markov assumptions, OLS = BLUE.

1.6.2 VARIANCE–COVARIANCE MATRIX OF ESTIMATORS:

For any linear combination $\mathrm{c}'\beta$,

$$\mathrm{Var}(\mathrm{c}'\hat{\beta}) = \sigma^2\mathrm{c}'(X'X)^{-1}\mathrm{c}.$$

This result is fundamental for inference on contrasts and parametric functions.

1.7 ESTIMATION OF ERROR VARIANCE:

An unbiased estimator of $\sigma^2$is

$$\hat{\sigma}^2 = \frac{(\mathrm{y} - X\hat{\beta})'(\mathrm{y} - X\hat{\beta})}{n-p}.$$

Hypothesis Testing under Gauss–Markov Set-up

Consider the general linear hypothesis:

$$H_0: C\beta = \mathrm{d},$$

where C is $q \times p$.

Test Statistic

$$F = \frac{(C\hat{\beta} - \mathrm{d})'[C(X'X)^{-1}C']^{-1}(C\hat{\beta} - \mathrm{d})/q}{\hat{\sigma}^2}.$$

Under H_0, $F \sim F_{q,n-p}$.

Confidence Intervals and Regions

- Individual Confidence Interval for β_j:

$$\hat{\beta}_j \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2[(X'X)^{-1}]_{jj}}$$

- Joint Confidence Region:

$$(\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) \leq p\hat{\sigma}^2 F_{p,n-p}(\alpha).$$

Special Cases and Extensions

1. Simple Linear Regression: $p = 2$.
2. Regression Through Origin.
3. Generalized Least Squares (GLS): $\text{Var}(\varepsilon) = \sigma^2 V$.
4. Weighted Least Squares.
5. Random Effects and Mixed Models.

1.8 KEY WORDS:

- Linear model
- Gauss–Markov assumptions
- ordinary least squares
- normal equations
- BLUE, variance–covariance matrix
- error variance
- confidence interval
- confidence region

1.9 SUMMARY:

The **Gauss–Markov set-up** provides the fundamental theoretical framework for **linear statistical models and regression analysis.** It begins with the formulation of the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where the response variable is expressed as a linear function of unknown parameters and a random error term. The model relies on a set of assumptions regarding linearity, unbiasedness of errors, constant variance, absence of correlation among errors, and full rank of the design matrix.

Within this framework, the **ordinary least squares (OLS)** method is used to estimate the unknown parameters by minimizing the sum of squared residuals. The resulting estimators are linear functions of the observations and are unbiased. Their precision is quantified through the **variance–covariance matrix**, which depends on the error variance and the structure of the design matrix.

A central result of this theory is the **Gauss–Markov Theorem**, which establishes that the OLS estimator is the **Best Linear Unbiased Estimator (BLUE)** of the regression parameters. Estimation of the error variance and construction of confidence intervals and confidence regions enable statistical inference and assessment of model reliability.

The **Gauss–Markov set-up** forms the cornerstone of classical **linear regression analysis.** It justifies the use of ordinary least squares estimation under minimal assumptions, without requiring normality of errors. The optimality of OLS as the BLUE makes it a powerful and widely applicable estimation technique in both theoretical and applied contexts.

Overall, the Gauss–Markov theory provides a strong foundation for applied regression analysis and remains essential for modern statistical modeling and data analysis.

1.10 SELF-ASSESSMENT QUESTIONS:

1. Explain the Gauss–Markov model along with underlying assumptions. Also give an example of the same.
2. State and prove the Gauss–Markov Theorem.
3. Explain why normality is not required for BLUE.
4. Derive the variance of $\hat{\beta}$.
5. Distinguish between OLS and GLS.
6. Define a linear statistical model. State the assumptions of the Gauss–Markov model.
7. Obtain the variance–covariance matrix of OLS estimators.

1.11 SUGGESTED READING:

1. Draper, N. R. & Smith, H. Applied Regression Analysis
2. Montgomery, D. C., Peck, E. A., & Vining, G. G. Introduction to Linear Regression Analysis
3. Rao, C. R. Linear Statistical Inference and Its Applications
4. Kutner, M. H. et al. Applied Linear Regression Models

Prof. V. V. Haragopal

LESSON-2

NORMAL EQUATIONS AND LEAST SQUARES ESTIMATES

OBJECTIVES:

After studying this lesson, the student should be able to:

- Understand the theoretical basis of the least squares principle in linear models.
- Derive the normal equations for simple and multiple linear regression models.
- Obtain least squares estimates using scalar and matrix methods.
- Interpret fitted values, residuals, and error components in regression analysis.
- Examine the conditions for existence and uniqueness of least squares solutions.
- Analyze the statistical properties of least squares estimators.
- Apply least squares estimation techniques to practical and real-life data problems**.**
- Use statistical software for computation and interpretation of regression estimates.

STRUCTURE:

2.1 INTRODUCTION:

The method of least squares occupies a central position in statistical theory and practice, forming the foundation of linear regression analysis and the general linear model. Originating from the works of Gauss and Legendre, least squares provides a systematic and

optimal procedure for estimating unknown parameters in linear models by minimizing the discrepancy between observed and fitted values. The resulting estimating equations, known as normal equations, yield estimators with well-established optimality properties under standard assumptions.

This chapter develops the theory of normal equations and least squares estimators in a rigorous and unified manner suitable for university-level study. Emphasis is placed on matrix formulations, precise notation, formal derivations, existence and uniqueness conditions, geometrical interpretation, and statistical properties. Applications and illustrative examples are included to connect theory with practice.

Description

The method of least squares is a fundamental technique used in regression analysis to estimate unknown parameters in a linear model. It is based on minimizing the sum of squared differences between the observed values and the corresponding fitted values obtained from the model.

In this topic, the linear regression model is expressed in matrix form, which allows the estimation problem to be handled systematically. By applying the least squares criterion, a set of equations known as the **normal equations** is obtained. Solving these equations yields the least squares estimates of the regression parameters.

The existence and uniqueness of the least squares estimates depend on the linear independence of the columns of the design matrix. The fitted values and residuals are then defined using the estimated parameters, and the residuals are used to obtain an estimate of the error variance.

This topic provides the basic mathematical foundation required for further study of regression analysis, hypothesis testing, and analysis of variance.

2.2 LINEAR REGRESSION MODEL AND ASSUMPTIONS:

Consider a set of observations $(y_1, y_2, \dots, y_n)$ on a response variable and corresponding values of $p$explanatory variables.

The multiple linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, i = 1,2, \dots, n$$

Assumptions (Gauss–Markov Framework)

1. Linearity in parameters: The model is linear in $\beta_0, \beta_1, \dots, \beta_p$.
2. Full rank (no exact multicollinearity): The regressors are linearly independent.
3. Zero mean errors: $E(\varepsilon_i) = 0$.
4. Homoscedasticity: $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all i.
5. Uncorrelated errors: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.
6. (Optional for inference)Normality: $\varepsilon_i \sim N(0, \sigma^2)$.

Matrix Representation of the Linear Model

Let

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

and the *design matrix*

$$\mathrm{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Then the model is compactly written as

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon,$$

with $E(\varepsilon) = 0$ and $\mathrm{Var}\,(\varepsilon) = \sigma^2 \mathrm{I}_n$.

2.3 LEAST SQUARES PRINCIPLE:

Definition

The *least squares estimator* minimizes the residual sum of squares

$$S(\beta) = (\mathrm{y} - \mathrm{X}\beta)'(\mathrm{y} - \mathrm{X}\beta)$$

Objective

Find $\hat{\beta}$ such that

$$S(\hat{\beta}) = \min_{\beta \in \mathbb{R}^{p+1}} S(\beta)$$

2.4 DERIVATION OF NORMAL EQUATIONS:

Expanding the sum of squares:

$$S(\beta) = \mathrm{y}'\mathrm{y} - 2\beta'\mathrm{X}'\mathrm{y} + \beta'\mathrm{X}'\mathrm{X}\hat{\beta}$$

Taking the gradient with respect to β:

$$\frac{\partial S}{\partial \beta} = -2\mathrm{X}'\mathrm{y} + 2\mathrm{X}'\mathrm{X}\hat{\beta}$$

Setting the gradient equal to zero yields the *normal equations*

$$\mathrm{X}'\mathrm{X}\hat{\beta} = \mathrm{X}'\mathrm{Y}$$

Solution of Normal Equations

1. Theorem: Least Squares Solution

Statement:

Let $\mathrm{Y} = \mathrm{X}\beta + \varepsilon$, where X is an $n \times (p+1)$design matrix. If $\mathrm{rank}\,(\mathrm{X}) = p + 1$, then $\mathrm{X}'\mathrm{X}$ is nonsingular and the normal equations $\mathrm{X}'\mathrm{X}\hat{\beta} = \mathrm{X}'\mathrm{Y}$ have the unique solution $\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$.

Proof :

Step 1: Linear independence of the columns of X
Since rank $(X) = p + 1$,
the columns of Xare linearly independent.
Hence, for any nonzero vector $a \in \mathbb{R}^{p+1}$, $Xa \neq 0$

Step 2: Positive definiteness of $X'X$
Consider the quadratic form $a'X'Xa$
This can be written as $a'X'Xa = (Xa)'(Xa) = \| Xa \|^2$
Since $Xa \neq 0$for all $a \neq 0$, $a'X'Xa > 0$
Thus, $X'X$is symmetric and positive definite.

Step 3: Nonsingularity of $X'X$
Every real symmetric positive definite matrix is nonsingular.
Therefore, $(X'X)^{-1}$ exists.

Step 4: Solution of the normal equations
The normal equations are $X'X\beta = X'y$
Premultiplying both sides by $(X'X)^{-1}$, we obtain $\hat{\beta} = (X'X)^{-1}X'y$

Step 5: Uniqueness
Since $X'X$ is nonsingular, the above solution is unique.

$$\boxed{\hat{\beta} = (X'X)^{-1}X'y}$$

2.5 THEOREMS:

2.5.1 EXISTENCE AND UNIQUENESS OF LEAST SQUARES ESTIMATES

Theorem: Existence of Least Squares Estimates

Statement:

For the linear model $y = X\beta + \varepsilon$,
there exists at least one vector $\hat{\beta} \in \mathbb{R}^{p+1}$that minimizes the residual sum of squares

$$S(\beta) = (y - X\beta)'(y - X\beta),$$

irrespective of whether the matrix $X'X$is singular or nonsingular.

Proof:

The function $S(\beta)$is a quadratic form in $\beta$and can be expressed as

$$S(\beta) = y'y - 2\beta'X'y + \beta'X'X\beta$$

Since $X'X$ is symmetric and positive semidefinite, the quadratic form $S(\beta)$ is bounded below. Therefore, $S(\beta)$attains a minimum over $\mathbb{R}^{p+1}$.

Equivalently, the normal equations

$$X'X\hat{\beta} = X'Y$$

always admit at least one solution. When $X'X$ is singular, solutions exist in the sense of generalized inverses.

Hence, a least squares estimator always exists.

2.5.2 UNIQUENESS OF LEAST SQUARES ESTIMATES:

Statement

The least squares estimator $\hat{\beta}$is unique if and only if rank (X) $= p + 1$.

Proof:

(Sufficiency)
If rank (X) $= p + 1$, then the columns of Xare linearly independent. Consequently, the matrix X′Xis symmetric and positive definite, and hence nonsingular. The normal equations therefore have the unique solution

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

(Necessity)
Suppose rank (X) $< p + 1$. Then the columns of Xare linearly dependent and X′X is singular. In this case, the normal equations have infinitely many solutions. Specifically, if β_0 is a solution, then for any nonzero vector d satisfying $\mathrm{Xd} = 0$, the vector $\beta_0 + \mathrm{d}$ is also a solution.

Hence, the least squares estimator is not unique.

Therefore, the least squares estimator is unique if and only if rank (X) $= p + 1$.

2.6 EXPLICIT FORM OF LEAST SQUARES ESTIMATOR:

The estimator is linear in y: $\hat{\beta} = \mathrm{Cy}, \mathrm{C} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$
Expectation $E(\hat{\beta}) = \beta$
Variance-Covariance Matrix Var $(\hat{\beta}) = \sigma^2(\mathrm{X}'\mathrm{X})^{-1}$

2.6.1 GEOMETRICAL INTERPRETATION OF LEAST SQUARES:

The column space $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$is the space spanned by the columns of X

Key Result

The fitted vector $\hat{\mathrm{y}} = \mathrm{X}\hat{\beta}$ is the *orthogonal projection* of **y**onto $\mathcal{C}(\mathbf{X})$.

The projection matrix (hat matrix) is $\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$

Residuals satisfy $\mathrm{X}'\hat{\varepsilon} = 0$, showing orthogonality between residuals and fitted values.

Properties of Least Squares Estimates

Gauss–Markov Theorem

Statement

Under the linear model

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon, E(\varepsilon) = 0, \mathrm{Var}\,(\varepsilon) = \sigma^2\mathrm{I},$$

the least squares estimator $\hat{\beta}$is the Best Linear Unbiased Estimator (BLUE) of β.

Proof :

Step 1: General form of a linear estimator

Any linear estimator of βcan be written as

$$\tilde{\beta} = \mathrm{Ay},$$

whereAis a $(p+1) \times n$matrix.

Step 2: Condition for unbiasedness

$$E(\tilde{\beta}) = \mathrm{A}E(\mathrm{y}) = \mathrm{AX}\beta$$

For unbiasedness,

$$E(\tilde{\beta}) = \beta \Rightarrow \mathrm{AX} = \mathrm{I}_{p+1}$$

Step 3: Variance of a linear unbiased estimator

$$\mathrm{Var}\,(\tilde{\beta}) = \mathrm{A}\mathrm{Var}\,(\mathrm{y})\mathrm{A}' = \sigma^2\mathrm{AA}'$$

Step 4: Least squares estimator

The least squares estimator is

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}.$$

Its variance is

$$\mathrm{Var}\,(\hat{\beta}) = \sigma^2(\mathrm{X}'\mathrm{X})^{-1}.$$

Step 5: Variance comparison

Let

$$\mathrm{D} = \mathrm{A} - (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'.$$

Then

$$\mathrm{Var}\,(\tilde{\beta}) - \mathrm{Var}\,(\hat{\beta}) = \sigma^2\mathrm{DD}',$$

which is positive semidefinite.

Hence,$\mathrm{Var}\,(\tilde{\beta}) \geq \mathrm{Var}\,(\hat{\beta})$.

Therefore, $\hat{\beta}$ is BLUE.

Additional Properties of Least Squares Estimator

(i) Linearity $\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$ is a linear function of y.

(ii) Unbiasedness $E(\hat{\beta}) = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'E(\mathrm{y}) = \beta$.

(iii) Consistency as $n \to \infty$, $\hat{\beta} \xrightarrow{P} \beta$.

(iv) Normality, If $\varepsilon \sim N(0, \sigma^2\mathrm{I})$, then $\hat{\beta} \sim N(\beta, \sigma^2(\mathrm{X}'\mathrm{X})^{-1})$

2.6.3 ESTIMATION OF ERROR VARIANCE:

The residual sum of squares is

$$\text{RSS} = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}.$$

An unbiased estimator of $\sigma^2$is

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-(p+1)}$$

Fitted Values and Residuals

- Fitted values: $\hat{\text{y}} = \text{Hy}$
- Residuals: $\hat{\varepsilon} = (\text{I} - \text{H})\text{y}$

Properties:

- $E(\hat{\varepsilon}) = 0$
- $\text{Var}\ (\hat{\varepsilon}) = \sigma^2(\text{I} - \text{H})$

2.7 APPLICATIONS:

1. Simple Linear Regression: Explicit normal equations reduce to familiar scalar equations for slope and intercept.
2. Polynomial Regression: Linear in parameters despite nonlinear regressors.
3. Econometrics and Engineering: Parameter estimation, calibration, and prediction.

2.8 KEY WORDS:

- Least Squares Estimator
- Normal Equations
- Design Matrix
- Hat Matrix
- Residuals
- Gauss–Markov Theorem
- Projection Matrix.

2.9 SUMMARY:

The method of least squares provides a fundamental approach for estimating unknown parameters in linear regression models. It is based on minimizing the sum of squared deviations between observed values and the values predicted by the model. This optimization leads to a system of equations known as the normal equations, which form the basis for obtaining least squares estimates.

Using matrix notation, the normal equations are expressed as $\text{X}'\text{X}\hat{\beta} = \text{X}'Y$ and, under appropriate conditions, yield a unique solution for the parameter vector. The explicit form of the least squares estimator highlights its dependence on the design matrix and observed data.

Theoretical results establish the existence, uniqueness, and unbiasedness of least squares estimates when the design matrix has full rank. The geometrical interpretation further clarifies that least squares estimation corresponds to the orthogonal projection of the observation vector onto the column space of the design matrix. Residuals are shown to be orthogonal to the fitted values and explanatory variables, reinforcing the optimality of the fitted model.

Estimation of the error variance using residuals provides a basis for statistical inference, including confidence intervals and hypothesis testing. These concepts are essential for evaluating model adequacy and reliability in applied regression analysis.

Overall, the least squares framework remains a cornerstone of applied statistics, offering both theoretical rigor and practical relevance across diverse fields such as economics, engineering, medicine, and social sciences.

2.10 SELF-ASSESSMENT QUESTIONS:

1. Derive the normal equations for a multiple linear regression model using matrix notation.
2. Prove the existence and unbiasedness of least squares estimates.
3. Discuss the uniqueness of least squares estimates and the role of the rank of the design matrix.
4. Explain the geometrical interpretation of least squares estimation and the concept of projection.
5. Define residuals and prove the orthogonality property of residuals with the design matrix.
6. Discuss the estimation of error variance in linear regression and its importance in inference.

2.11 SUGGESTED READING:

1. Draper, N. R. & Smith, H. Applied Regression Analysis
2. Montgomery, D. C., Peck, E. A., & Vining, G. G. Introduction to Linear Regression Analysis
3. Rao, C. R. Linear Statistical Inference and its Applications
4. Kutner, M. H. et al. Applied Linear Regression Models
5. Seber, G. A. F. and Lee, A. J., Linear Regression Analysis.

Prof. V. V. Haragopal

LESSON-3

VARIANCE AND COVARIANCE OF LEAST SQUARES ESTIMATES

OBJECTIVES:

After studying this lesson, the student should be able to:

- Understand the need for measuring variability in least squares estimates.
- Define the variance of a least squares estimator and the covariance between two least squares estimators.
- Obtain the variance-covariance matrix of least squares estimates.Understand the role of error variance in determining variances of estimators.
- Compute standard errors of least squares estimates.
- Interpret variances and covariances in regression analysis.
- Apply these concepts to assess the precision of parameter estimates.

STRUCTURE:

3.1 INTRODUCTION:

An essential aspect of linear regression theory concerns the variability of the least squares estimators. While point estimates provide fitted values of the regression parameters, meaningful statistical inference requires an explicit understanding of their variances and covariances. These quantities quantify estimation uncertainty, determine the precision of

individual regression coefficients, and form the basis for hypothesis testing, confidence intervals, and diagnostics such as multicollinearity assessment.

This lesson presents a systematic and rigorous treatment of the variances and covariances of least squares estimates within the classical linear model framework. Emphasis is placed on matrix-based derivations, formal theorems with proofs, interpretation of the variance–covariance structure, and practical implications in regression analysis.

Description:

In regression analysis, least squares estimates provide point estimates of unknown parameters. To assess the reliability and precision of these estimates, it is necessary to study their variances and covariances.

This topic examines the variability of least squares estimates under the standard assumptions of the linear regression model. The variance of an estimator measures the spread of its sampling distribution, while the covariance between two estimators indicates the degree of linear association between them.

Using the matrix formulation of the regression model, the variance-covariance matrix of the least squares estimates is derived. This matrix plays a central role in regression inference, as its diagonal elements represent variances and its off-diagonal elements represent covariances.

The error variance is estimated using residuals, and this estimate is used to obtain standard errors of the regression coefficients. These results form the basis for hypothesis testing, confidence intervals, and interpretation of regression parameters.

3.2 LINEAR REGRESSION MODEL AND ASSUMPTIONS:

Consider the linear regression model $\mathrm{y} = \mathrm{X}\beta + \varepsilon$,
Where

- y is an $n \times 1$vector of observations,
- X is a known $n \times (p+1)$ design matrix of full column rank,
- β is a $(p+1) \times 1$vector of unknown parameters,
- ε is an $n \times 1$ vector of random errors.

Standard Assumptions

1. $E(\varepsilon) = 0$
2. $\mathrm{Var}\ (\varepsilon) = \sigma^2 \mathrm{I}_n$
3. $\mathrm{rank}\ (\mathrm{X}) = p + 1$
4. (For exact distributional results) $\varepsilon \sim N(0, \sigma^2 \mathrm{I}_n)$.

3.3 LEAST SQUARES ESTIMATOR:

The least squares estimator of $\beta$is defined as the minimizer of

$$S(\beta) = (\mathrm{y} - \mathrm{X}\beta)'(\mathrm{y} - \mathrm{X}\beta)$$

Under the full-rank assumption,

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

3.3.1 VARIANCE OF LEAST SQUARES ESTIMATES:

Theorem -1(Variance of the Least Squares Estimator)
Under the classical linear model assumptions, $\text{Var}\,(\hat{\beta}) = \sigma^2(\text{X}'\text{X})^{-1}$

Proof:
Since $\hat{\beta}$is a linear function of y,

$$\hat{\beta} = (\text{X}'\text{X})^{-1}\text{X}'(\text{X}\beta + \varepsilon) = \beta + (\text{X}'\text{X})^{-1}\text{X}'\varepsilon$$

Taking variances,

$$\text{Var}\,(\hat{\beta}) = (\text{X}'\text{X})^{-1}\text{X}'\text{Var}\,(\varepsilon)\text{X}(\text{X}'\text{X})^{-1}$$

Using $\text{Var}\,(\varepsilon) = \sigma^2\text{I}_n$,

$$\text{Var}\,(\hat{\beta}) = \sigma^2(\text{X}'\text{X})^{-1}\text{X}'\text{X}(\text{X}'\text{X})^{-1} = \sigma^2(\text{X}'\text{X})^{-1}$$

3.3.2 COVARIANCE BETWEEN LEAST SQUARES ESTIMATES:

Let $\hat{\beta}_j$and $\hat{\beta}_k$denote two components of $\hat{\beta}$.

Definition:
The covariance between $\hat{\beta}_j$and $\hat{\beta}_k$is given by the (j, k)th element of $\text{Var}\,(\hat{\beta})$.
That is,
$\text{Cov}\,(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2[(\text{X}'\text{X})^{-1}]_{jk}$
Nonzero covariances indicate linear dependence among the estimators, often arising from correlation among regressors.

Theorem-2 Covariance Between Least Squares Estimates

Statement:
Let $\hat{\beta}_j$and $\hat{\beta}_k$be the jth and kth components of $\hat{\beta}$. Then$\text{Cov}\,(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2[(\text{X}'\text{X})^{-1}]_{jk}$

Proof:
From Theorem 1, $\text{Var}\,(\hat{\beta}) = \sigma^2(\text{X}'\text{X})^{-1}$
By definition, the (j, k)th element of the variance–covariance matrix equals $\text{Cov}\,(\hat{\beta}_j, \hat{\beta}_k)$
Hence, $\text{Cov}\,(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2[(\text{X}'\text{X})^{-1}]_{jk}$

3.3.3 VARIANCE-COVARIANCE MATRIX OF LEAST SQUARES ESTIMATES:

The full variance–covariance matrix is

$$\text{Var}\,(\hat{\beta}) = \sigma^2(\text{X}'\text{X})^{-1}$$

Structure

- Diagonal elements: variances of individual parameter estimates
- Off diagonal elements: covariances between parameter estimates

The matrix is symmetric and positive definite under the full-rank assumption.

Theorem-3 (Variance-Covariance Matrix of Least Squares Estimates)
Statement:
The full variance–covariance matrix of $\hat{\beta}$is $V_{\hat{\beta}} = \sigma^2(X'X)^{-1}$,
which is symmetric and positive definite.

Proof:
- Symmetry follows since $(X'X)^{-1}$is symmetric.
- Positive definiteness follows from the positive definiteness of $X'X$when rank $(X) = p + 1$.

3.3.4 PROPERTIES OF VARIANCES AND COVARIANCES:

1. Dependence on design: Variances depend solely on Xand σ^2.
2. Effect of multicollinearity: Near-linear dependence among regressors inflates variances.
3. Orthogonality: If columns of Xare orthogonal, covariances vanish.
4. Scale sensitivity: Rescaling regressors alters variances of corresponding coefficients.

3.4 ESTIMATION OF ERROR VARIANCE:

Since $\sigma^2$is unknown, it is estimated by

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-(p+1)},$$

Where $\hat{\varepsilon} = y - X\hat{\beta}$

Theorem-4 (Unbiased Estimator of Error Variance)
Statement:
The estimator$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-(p+1)}$ is an unbiased estimator of σ^2.
Proof :
Step 1: Express residuals

$$\hat{\varepsilon} = (I - H)y, H = X(X'X)^{-1}X'$$

Step 2: Residual sum of squares

$$\hat{\varepsilon}'\hat{\varepsilon} = y'(I - mathbfH)y$$

Step 3: Take expectation
Using properties of quadratic forms,

$$E(\hat{\varepsilon}'\hat{\varepsilon}) = \sigma^2 \text{tr}\,(I - H)$$

Step 4: Evaluate the trace

$$\text{tr}\,(H) = \text{rank}\,(H) = p + 1, \text{tr}\,(I - H) = n - (p + 1)$$

Step 5: Conclude

$$E(\hat{\sigma}^2) = \frac{E(\hat{\varepsilon}'\hat{\varepsilon})}{n - (p + 1)} = \sigma^2.$$

3.5 STANDARD ERRORS OF LEAST SQUARES ESTIMATES:

The standard error of $\hat{\beta}_j$is defined as

$$\text{SE}\,(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2[(X'X)^{-1}]_{jj}}$$

Standard errors quantify the precision of individual regression coefficients and are fundamental to inference procedures.

3.6 INTERPRETATION OF VARIANCES AND COVARIANCES:

- Large variances indicate imprecise estimation
- Large covariances suggest strong dependence among coefficient estimates
- Correlation coefficients between estimates can be computed as

$$\rho_{jk} = \frac{\text{Cov}\,(\hat{\beta}_j, \hat{\beta}_k)}{\sqrt{\text{Var}\,(\hat{\beta}_j)\text{Var}\,(\hat{\beta}_k)}}$$

3.7 APPLICATIONS:

- Simple Linear Regression: Closed-form expressions for $\text{Var}\,(\hat{\beta}_0)$ and $\text{Var}\,(\hat{\beta}_1)$.
- Multicollinearity Diagnosis: Variance inflation factors derived from diagonal elements.
- Experimental Design: Choice of Xto minimize estimator variances.
- Construction of confidence intervals for regression coefficients
- Hypothesis testing in regression models
- Interpretation of regression output from software packages

3.8 KEY WORDS:

- Least squares estimator
- Variance
- Covariance
- Variance-Covariance matrix
- Error Variance
- Standard Error
- Multicollinearity
- Regression Inference

3.9 SUMMARY:

In linear models and applied regression analysis, the study of the variance and covariance of least squares estimates is essential for understanding the reliability and precision of estimated regression coefficients. While least squares estimation provides point estimates of parameters, their usefulness depends on how much these estimates vary across repeated samples.

Under the standard linear regression assumptions, the least squares estimator is linear and unbiased, and its variability is fully described by the variance-covariance matrix**.** The variances of individual regression coefficients measure their precision, while the covariances describe the degree of association between different parameter estimates. These quantities depend on the error variance and the structure of the design matrix**,** highlighting the role of data configuration and multicollinearity.

Estimation of the error variance using residuals enables practical computation of variances, covariances, and standard errors**.** These measures form the basis for constructing

confidence intervals, conducting hypothesis tests, and interpreting regression output in applied work.

The variance and covariance of least squares estimates provide the statistical foundation for inference in linear regression models. They allow researchers to assess the accuracy and stability of parameter estimates and to understand the relationships among estimated coefficients.

A clear understanding of these concepts enables:

- Evaluation of the precision and significance of regression coefficients
- Detection of issues such as multicollinearity and inefficient model design
- Proper interpretation of regression results in real-world applications

Overall, the analysis of variances and covariances complements least squares estimation by transforming point estimates into meaningful inferential statements. It is a crucial component of linear models and applied regression analysis across diverse scientific and practical domains.

3.10 SELF-ASSESSMENT QUESTIONS:

1. Derive the variance–covariance matrix of least squares estimators in a multiple linear regression model.
2. Explain the properties of variances and covariances of least squares estimates.
3. Discuss the estimation of error variance and its role in regression inference.
4. Explain how variances and covariances of regression coefficients are used in constructing confidence intervals and hypothesis tests.
5. Discuss the interpretation of variance and covariance of least squares estimates with suitable examples.
6. For a regression model, explain how multicollinearity affects the variances and covariances of least squares estimates. Illustrate with a suitable example.

3.11 SUGGESTED READING:

1. Draper, N. R. & Smith, H. Applied Regression Analysis
2. Montgomery, D. C., Peck, E. A., & Vining, G. G. Introduction to Linear Regression Analysis
3. Rao, C. R. Linear Statistical Inference and its Applications
4. Kutner, M. H. et al. Applied Linear Regression Models
5. Seber, G. A. F. and Lee, A. J., Linear Regression Analysis.

Prof. V. V. Haragopal

LESSON – 4

ESTIMATION OF ERROR VARIANCE

OBJECTIVES:

- Understand the role of error variance in linear regression models
- Derive the unbiased estimator of error variance under the linear model framework
- Explain the relationship between residuals and error variance estimation
- Analyze the effect of model assumptions on the estimation of error variance
- Use error variance estimates to compute standard errors of regression coefficients
- Apply error variance estimation in constructing confidence intervals and hypothesis tests
- Interpret error variance estimates in applied regression problems
- Implement error variance estimation using statistical software

STRUCTURE:

4.1 INTRODUCTION:

In linear regression analysis, the variability observed in the response variable cannot be completely explained by the systematic component of the model. This unexplained variation is attributed to random disturbances, collectively represented by the error term. Quantifying this variability is fundamental to statistical inference in regression, as it governs the precision of parameter estimates, the construction of confidence intervals, and the validity of hypothesis tests.

The estimation of error variance, denoted by σ^2, is therefore a central problem in regression theory. This chapter develops the theoretical framework for estimating σ^2under the classical linear regression model, derives its estimators, examines their sampling properties, and highlights their role in statistical inference.

Description: Estimation of error variance
Estimation of error variance is a core statistical topic focused on quantifying the inherent, unexplained variability (σ^2) in a model (like regression), crucial for hypothesis testing, confidence intervals, and model evaluation, often achieved by dividing the residual sum of squares (SSE) by its degrees of freedom (n-p) to get the Mean Squared Error (MSE), an

unbiased estimator, though complex high-dimensional or small-sample scenarios require advanced methods like adaptive lasso or cross-validation for accurate results.

4.2 ERROR TERM AND ITS ASSUMPTIONS:

Error Term ($\boldsymbol{\varepsilon}$)
Consider the multiple linear regression model

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon$$

Where

- y is an $n \times 1$ vector of observations
- X is an $n \times p$ design matrix of full rank p
- β is a $p \times 1$ vector of unknown parameters
- ε is an $n \times 1$ vector of random errors.

The error term ε_i represents the combined effect of omitted variables, measurement error, and inherent randomness.

Variance of the Error Term ($\boldsymbol{\sigma^2}$)
It is assumed that

$$\mathrm{Var}\ (\varepsilon) = \sigma^2 \mathrm{I}_n$$

where $\sigma^2 > 0$ is an unknown constant representing the common variance of the errors.
Assumptions about Error Variance
The classical regression model imposes the following assumptions on the error structure:

1. $E(\varepsilon) = 0$
2. $\mathrm{Var}\ (\varepsilon) = \sigma^2 \mathrm{I}_n$
3. ε_i and ε_j are uncorrelated for $i \neq j$
4. (For exact inference) $\varepsilon \sim N(0, \sigma^2 \mathrm{I}_n)$

Homoscedasticity
Homoscedasticity refers to the assumption that the variance of the error term is constant across all observations:

$$\mathrm{Var}\ (\varepsilon_i) = \sigma^2 \forall i$$

Violation of this assumption (heteroscedasticity) leads to biased variance estimates and invalid inference.

Constant Error Variance Assumption
The assumption $\mathrm{Var}(\varepsilon) = \sigma^2 \mathrm{I}$ ensures that ordinary least squares estimators are efficient under the Gauss–Markov theorem and permits unbiased estimation of σ^2.

4.3 RESIDUALS AND ERROR DECOMPOSITION:

Residuals
The ordinary least squares (OLS) estimator of β is

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

The vector of residuals is defined as

$$\mathrm{e} = \mathrm{y} - \hat{\mathrm{y}} = \mathrm{y} - \mathrm{X}\hat{\beta}$$

Fitted Residuals
Residuals are observable quantities and serve as estimates of the unobservable errors $\boldsymbol{\varepsilon}$. They satisfy

$$\mathrm{e} = (\mathrm{I} - \mathrm{H})\mathrm{y}$$

where $\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ is the hat matrix.

Relationship between Errors and Residuals
Residuals differ from true errors due to parameter estimation. Specifically,

$$E(\mathrm{e}) = 0, \mathrm{Var}\ (\mathrm{e}) = \sigma^2(\mathrm{I} - \mathrm{H}).$$

Thus, residuals are correlated and have unequal variances, even when errors are homoscedastic.

4.4 SUMS OF SQUARES FOR ERROR:

Residual Sum of Squares (RSS)
The residual sum of squares is defined as

$$\mathrm{RSS} = \mathrm{e}'\mathrm{e}$$

Error Sum of Squares (SSE)
In regression analysis, RSS and SSE are used interchangeably:

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2$$

Relationship between SSE and $\boldsymbol{\sigma^2}$
Using properties of quadratic forms,

$$E(\mathrm{SSE}) = (n - p)\sigma^2$$

This result forms the basis for unbiased estimation of the error variance.

4.5 DEGREES OF FREEDOM AND MEAN SQUARE ERROR:

Degrees of Freedom for Error ($\boldsymbol{n - p}$)
The number of independent pieces of information available to estimate $\sigma^2$is reduced by the estimation of $p$regression parameters.

Mean Square Error (MSE)
The mean square error is defined as

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n - p}$$

Estimated Standard Deviation of Errors
The estimator of the standard deviation of the errors is

$$\hat{\sigma} = \sqrt{\mathrm{MSE}}$$

4.6 ESTIMATION OF ERROR VARIANCE:

Estimator of Error Variance
An estimator of $\sigma^2$is any statistic based on the sample that approximates the true variance of the errors.

Unbiased Estimator of σ^2

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-p}$$

Proof of Unbiasedness

$$E(\hat{\sigma}^2) = \frac{1}{n-p} E(\text{SSE}) = \sigma^2$$

Maximum Likelihood Estimator (MLE) of σ^2

Under normality,

$$\tilde{\sigma}^2 = \frac{\text{SSE}}{n}$$

Comparison: MLE vs Unbiased Estimator

Property	MLE	Unbiased Estimator
Bias	Biased downward	Unbiased
Variance	Smaller	Larger
Used in inference	No	Yes

4.7 SAMPLING PROPERTIES OF THE VARIANCE ESTIMATOR:

Sampling Distribution of MSE

If $\varepsilon \sim N(0, \sigma^2 \text{I})$, then

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$$

Properties of Error Variance Estimator

- Unbiased
- Consistent
- Scaled chi-square distribution

Consistency of σ^2Estimator

$$\hat{\sigma}^2 \xrightarrow{p} \sigma^2 \text{as } n \to \infty$$

Efficiency of Estimator

Among unbiased estimators based on residuals, $\hat{\sigma}^2$has minimum variance under normality.

Effect of Model Degrees of Freedom

As $p$increases, $n-p$decreases, inflating the variance of $\hat{\sigma}^2$.

4.7.1 ROLE OF ERROR VARIANCE IN STATISTICAL INFERENCE:

Use of MSE in Inference

MSE is used to estimate the covariance matrix of $\hat{\beta}$:

$$\text{Var}\,(\hat{\beta}) = \sigma^2(\text{X}'\text{X})^{-1}$$

Role in Confidence Intervals

Confidence intervals for regression coefficients depend directly on $\hat{\sigma}^2$.

Role in Hypothesis Testing

Test statistics (t and F) use MSE as the denominator, making accurate variance estimation essential.

4.8 SUMMARY:

In linear models and applied regression analysis, the error variance represents the portion of variability in the response variable that is not explained by the regression model. Accurate estimation of this variance is essential because it quantifies random error and underpins all procedures of statistical inference.

The estimation of error variance is based on residuals**,** which measure the differences between observed and fitted values. By using the residual sum of squares (RSS) and adjusting for the loss of degrees of freedom due to parameter estimation, an unbiased estimator of the error variance is obtained. This adjustment ensures that the estimator correctly reflects the true variability of the error term.

Under standard model assumptions, particularly normality of errors, the estimator of error variance has well-defined sampling properties**,** including a chi-square distribution. These properties allow for interval estimation and hypothesis testing related to both model parameters and overall model adequacy.

A sound understanding of error variance estimation enables analysts to:

- Evaluate model fit and unexplained variability
- Perform valid statistical inference on regression parameters
- Compare competing models and assess predictive performance

In practice, careful estimation and interpretation of error variance are essential for meaningful and reliable regression analysis across scientific, economic, and engineering applications.

4.9 KEY WORDS:

- Error term
- Residuals
- Error sum of squares
- Mean square error
- Degrees of freedom
- Error variance
- Sampling distribution
- Chi-square distribution
- Statistical inference

4.10 SELF-ASSESSMENT QUESTIONS:

1. Define error variance and explain its importance.
2. State the assumptions of the error term in a linear regression model.
3. Derive the estimator of error variance.
4. Explain the role of degrees of freedom in variance estimation.
5. State the sampling distribution of the error variance estimator.
6. How is error variance used in hypothesis testing?

4.11 SUGGESTED READING:

1. Montgomery, D.C., Peck, E.A., Vining, G.G. Introduction to Linear Regression Analysis, Wiley
2. Weisberg, S. Applied Linear Regression, Wiley
3. Graybill, F.A. Matrices with Applications in Statistics, Wadsworth
4. Rao, C.R. Linear Statistical Inference and Its Applications, Wiley
5. Draper, N.R., Smith, H. Applied Regression Analysis, Wiley.

Prof. V. V. Haragopal

LESSON-5
ESTIMATED WITH CORELATED OBSERVATIONS

OBJECTIVES:

After studying this Lesson, the learner will be able to:

- Understand the concept of correlated observations arising in regression and other statistical models.
- Identify the consequences of violating the independence assumption.
- Explain autocorrelation and correlated error structures, including common models (AR(1), MA(1), compound symmetry).
- Apply Generalized Least Squares (GLS) for parameter estimation when observations are correlated.
- Use Maximum Likelihood Estimation (MLE) for models with correlated errors.
- Estimate variance and construct confidence intervals under correlated observations.
- Compare estimation efficiency between independent-error models and correlated-error models.
- Apply methods to real situations such as time-series and repeated-measures data.

STRUCTURE:

5.9 Self-Assessment Questions

5.10 Suggested Reading

5.1 INTRODUCTION:

In many statistical modelling situations-especially in regression analysis, time-series data, longitudinal designs, and clustered sampling-the assumption of independence of errors is unrealistic. Observations collected sequentially or within similar groups often exhibit correlation. When correlation among errors is ignored, the classical ordinary least squares (OLS) estimator becomes inefficient and its variance estimates become biased, leading to invalid statistical inference.

The study of estimation under correlated observations therefore extends classical linear model theory by allowing the error vector to follow a general variance-covariance structure. This framework is essential in econometrics, biostatistics, engineering, and the analysis of repeated measures.

Description
In some data sets, observations are correlated due to time effects, spatial relationships, or repeated measurements. In such cases, the assumption of independent errors in the regression model is violated. Although the ordinary least squares estimator remains unbiased, it is not efficient and gives incorrect standard errors. To obtain efficient estimates and valid inference, Generalized Least Squares (GLS) is used, which accounts for the correlation among observations.

5.2 CONCEPT OF CORRELATED OBSERVATIONS:

Let

$$Y = (Y_1, \dots, Y_n)'$$

be the observed response vector. Observations are said to be correlated if

$$\text{Cov}\,(Y_i, Y_j) \neq 0 \text{for some } i \neq j.$$

Correlation arises from:

- Time dependence (autocorrelation),
- Cluster-wise dependence,
- Spatial dependence,
- Repeated measurements on the same subject,
- Measurement error propagation.

In such cases the classical assumption$\text{Cov}(\varepsilon) = \sigma^2 I_n$ is violated.

5.3 NOTATIONS AND DEFINITIONS:

Consider the general linear model

$$Y = X\beta + \varepsilon$$

where

- Y is an $n \times 1$ vector
- X is an $n \times p$ design matrix of full rank p
- β is a $p \times 1$ vector of unknown parameters
- ε is an $n \times 1$ error vector

Definition (Correlated Error Model):
Errors follow

$$E(\varepsilon) = 0, \text{Cov}\,(\varepsilon) = \Sigma$$

where Σ is a known or specified positive definite matrix, not necessarily proportional to the identity matrix.

Definition (Generalized Least Squares Problem)
Given correlated errors, find an estimator of βminimizing the *generalized* quadratic form

$$Q(\beta) = (Y - X\beta)'\Sigma^{-1}(Y - X\beta)$$

5.3.1 Variance-Covariance Matrix Of Errors:

The structure of $\Sigma = \text{Var}(\varepsilon)$ determines the nature of correlation.
General form:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

Properties:

- symmetric
- positive definite
- invertible

Examples: Time-series Σmatrices are Toeplitz, while cluster-based matrices show block-diagonal structure.

5.3.2 Common Correlation Structures:

(i) Compound Symmetry (CS)

$$\Sigma = \sigma^2[(1-\rho)I_n + \rho J_n]$$

(ii) Autoregressive of order 1 (AR(1))

$$\Sigma_{ij} = \sigma^2\rho^{|i-j|}$$

(iii) Moving Average MA (1)

$$\Sigma_{ij} = \begin{cases} \sigma^2(1+\theta^2), & i = j, \\ \sigma^2\theta, & |\, i-j \,| = 1, \\ 0, & \text{otherwise.} \end{cases}$$

(iv) Block-diagonal (Cluster correlation)

$$\Sigma = \text{diag}\,(\Sigma_1, \Sigma_2, \ldots, \Sigma_k)$$

5.3.3 Problems Caused By Correlated Errors:

If OLS is applied:

1. Estimator remains unbiased:
$$E(\hat{\beta}_{OLS}) = \beta$$
2. Variance is no longer minimal:
$$\text{Var}\,(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}X'\Sigma X(X'X)^{-1}$$
3. Inefficient estimators (not BLUE)
4. Standard errors are incorrect, leading to:
    - invalid t-tests
    - invalid F-tests

- wrong confidence intervals

5.3.4 Generalized Least Squares (Gls) Estimator:

Theorem - 1 (Generalized Least Squares Estimator)
Statement: Consider the linear model $Y = X\beta + \varepsilon$, where

- Y is $n \times 1$
- X is $n \times p$ with rank p
- $E(\varepsilon) = 0$
- Var $(\varepsilon) = \sigma^2 V$
  where V is a known positive definite $n \times n$ matrix.

The GLS estimator of $\beta$is

$$\boxed{\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}Y}$$

Proof :
Step 1: Define the GLS objective function
GLS minimizes the generalized sum of squares

$$Q(\beta) = (Y - X\beta)'V^{-1}(Y - X\beta)$$

This is a quadratic function in β.

Step 2: Expand $Q(\beta)$Multiply the terms:

$$Q(\beta) = Y'V^{-1}Y - Y'V^{-1}X\beta - \beta'X'V^{-1}Y + \beta'X'V^{-1}X\beta$$

Since all terms are scalars,

$$Y'V^{-1}X\beta = \beta'X'V^{-1}Y$$

Thus,

$$Q(\beta) = Y'V^{-1}Y - 2\beta'X'V^{-1}Y + \beta'X'V^{-1}X\beta$$

Step 3: Differentiate w.r.t. β
Use matrix derivative:

- $\frac{\partial}{\partial\beta}(\beta'A\beta) = 2A\beta$when $A$is symmetric
- $\frac{\partial}{\partial\beta}(b'\beta) = b$

Here $X'V^{-1}X$is symmetric.
Derivative:

$$\frac{\partial Q}{\partial\beta} = -2X'V^{-1}Y + 2X'V^{-1}X\beta$$

Step 4: Equate derivative to zero

$$-2X'V^{-1}Y + 2X'V^{-1}X\beta = 0$$

Divide by 2:

$$X'V^{-1}X\beta = X'V^{-1}Y$$

Step 5: Solve for β
Since $X'V^{-1}X$is nonsingular (rank p):

$$\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}Y$$

Conclusion, $\boxed{\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}Y}$
This completes the proof.

Theorem 2. (Unbiasedness and Variance of GLS Estimator)
Statement: Under the model

$$Y = X\beta + \varepsilon, E(\varepsilon) = 0, \text{Var}\,(\varepsilon) = \sigma^2 V$$

the GLS estimator satisfies:

1. Unbiasedness $\boxed{E(\hat{\beta}_{GLS}) = \beta}$

Variance $\boxed{\text{Var}\,(\hat{\beta}_{GLS}) = \sigma^2 (X'V^{-1}X)^{-1}}$

Proof :
Step 1: Write GLS estimator

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y$$

Part A: Unbiasedness

Step 2: Take expectation

$$E(\hat{\beta}) = (X'V^{-1}X)^{-1}X'V^{-1}E(Y)$$

Since
$E(Y) = X\beta,$ substitute:

$$E(\hat{\beta}) = (X'V^{-1}X)^{-1}X'V^{-1}(X\beta)$$

Step 3: Simplify using associativity

$$E(\hat{\beta}) = (X'V^{-1}X)^{-1}(X'V^{-1}X)\beta$$

Step 4: Use inverse property

$$(X'V^{-1}X)^{-1}(X'V^{-1}X) = I_p$$

Thus, $E(\hat{\beta}) = I_p\beta = \beta$

Conclusion (Unbiasedness) $\boxed{E(\hat{\beta}_{GLS}) = \beta}$

Part B: Variance

Step 5: Use formula

$$\text{Var}\,(\hat{\beta}) = (X'V^{-1}X)^{-1}X'V^{-1}\text{Var}\,(Y)V^{-1}X(X'V^{-1}X)^{-1}$$

Step 6: Substitute $\text{Var}\,(Y) = \sigma^2 V$

$$= (X'V^{-1}X)^{-1}X'V^{-1}(\sigma^2 V)V^{-1}X(X'V^{-1}X)^{-1}$$

Step 7: Simplify $V^{-1}VV^{-1} = V^{-1}$

$$= \sigma^2(X'V^{-1}X)^{-1}X'V^{-1}X(X'V^{-1}X)^{-1}$$

Step 8: Collapse the middle terms

$$X'V^{-1}$$

cancels with one of its inverses:

$$= \sigma^2(X'V^{-1}X)^{-1}$$

Conclusion (Variance)

$$\boxed{\text{Var}\,(\hat{\beta}_{GLS}) = \sigma^2(X'V^{-1}X)^{-1}}$$

Theorem - 3 (GLS is BLUE - Generalized Gauss-Markov)
Statement: Among all linear unbiased estimators of the form $\tilde{\beta} = CY$ that satisfy $E(\tilde{\beta}) = \beta,$ the GLS estimator has minimum variance.

Proof :
Step 1: General linear estimator
Let$\tilde{\beta} = CY$

where C is any $p \times n$ matrix.

Step 2: Impose unbiasedness

$$E(\tilde{\beta}) = CE(Y) = CX\beta = \beta$$

Thus,

$$\boxed{CX = I_p} \quad (1)$$

This is the unbiasedness condition.
Step 3: Variance of any linear unbiased estimator

$$\text{Var}\,(\tilde{\beta}) = C(\sigma^2 V)C' = \sigma^2 CVC'$$

Step 4: Write GLS estimator

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y$$

Thus GLS corresponds to

$$C_0 = (X'V^{-1}X)^{-1}X'V^{-1}$$

Step 5: Compare variance matrices
We need to prove:

$$CVC' - C_0VC_0' \text{ is positive semi} - \text{definite}$$

Equivalent to showing:

$$\text{Var}\,(\tilde{\beta}) - \text{Var}\,(\hat{\beta}_{GLS}) \geq 0$$

Step 6: Use unbiasedness constraint
Let $C = C_0 + D$
Using condition (1):

$$CX = I_p \Rightarrow (C_0 + D)X = I_p$$

But $C_0X = I_p$(can be shown by substitution).
Thus: $DX = 0$ (2)

Step 7: Expand variance of $\tilde{\beta}$

$$\text{Var}\,(\tilde{\beta}) = \sigma^2(C_0 + D)V(C_0 + D)'$$

Expand:$= \sigma^2(C_0VC_0' + C_0VD' + DVC_0' + DVD')$

Step 8: Show cross-terms vanish
We show:

$$C_0VD' = 0 \text{and} DVC_0' = 0$$

Using (2):

$$DX = 0$$

One can show that $C_0V = (X'V^{-1}X)^{-1}X'$
Thus,$C_0VD' = (X'V^{-1}X)^{-1}X'D' = (X'V^{-1}X)^{-1}(DX)' = 0$
Similarly for the transpose.
Thus variance reduces to:Var $(\tilde{\beta}) = \sigma^2(C_0VC_0' + DVD')$

Step 9: Subtract variance of GLS

$$\text{Var}\,(\tilde{\beta}) - \text{Var}\,(\hat{\beta}_{GLS}) = \sigma^2 DVD'$$

Step 10: Show DVD' is positive semidefinite
For any vector a,

$$a'(DVD')a = (D'a)'V(D'a) \geq 0$$

since V is positive definite.
Thus: $DVD' \geq 0$
Conclusion

$$\boxed{\text{Var}\,(\tilde{\beta}) \;\geq\; \text{Var}\,(\hat{\beta}_{GLS})}$$

Hence GLS is the Best Linear Unbiased Estimator (BLUE)
This completes the proof.

Corollary - Reduction to OLS
If $\Sigma = \sigma^2 I_n$, then $X'\Sigma^{-1}X = \frac{1}{\sigma^2}X'X$ and

$$\hat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y = (X'X)^{-1}X'Y = \hat{\beta}_{OLS}$$

- **Properties of GLS:**
  1. Unbiased: $E(\hat{\beta}_{GLS}) = \beta$
  2. Variance: $\text{Var}\,(\hat{\beta}_{GLS}) = (X'\Sigma^{-1}X)^{-1}$
  3. BLUE (Gauss–Markov Theorem Extension): Among all linear unbiased estimators, GLS has minimum variance.

5.3.5 Maximum Likelihood Estimation With Correlated Errors:

Assume $\varepsilon \sim N(0, \Sigma)$
Likelihood

$$L(\beta, \sigma^2) = \frac{1}{(2\pi)^{n/2} \mid \Sigma \mid^{1/2}} \exp\left[-\frac{1}{2}(Y - X\beta)'\Sigma^{-1}(Y - X\beta)\right]$$

MLE of β
Maximization w.r.t. β yields exactly the same estimator:

$$\hat{\beta}_{MLE} = \hat{\beta}_{GLS}$$

MLE of σ^2

$$\hat{\sigma}^2 = \frac{1}{n}(Y - X\hat{\beta}_{GLS})'\Sigma^{-1}(Y - X\hat{\beta}_{GLS})$$

5.4 BIAS AND MEAN SQUARE ERROR (MSE):

Bias

$$\text{Bias}\,(\hat{\beta}_{GLS}) = 0$$

MSE

$$\text{MSE}\,(\hat{\beta}_{GLS}) = \text{Var}\,(\hat{\beta}_{GLS}) = (X'\Sigma^{-1}X)^{-1}$$

GLS has strictly smaller MSE than OLS when errors are correlated.

5.4.1 Estimation of Variance and Confidence Intervals:

Estimator of Variance

$$\hat{\sigma}^2 = \frac{(Y - X\hat{\beta}_{GLS})'\Sigma^{-1}(Y - X\hat{\beta}_{GLS})}{n - p}$$

Confidence Interval
For the j-th component of β:

$$\hat{\beta}_j \pm t_{n-p,\,\alpha/2}\sqrt{(X'\Sigma^{-1}X)^{-1}_{jj}}$$

5.4.2 Auto Correlation Models In Regression:

When errors follow

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

the covariance structure becomes Toeplitz.

5.4.3 Ar (1) Model:

Error Structure

$$\Sigma_{ij} = \sigma^2\rho^{|i-j|}, |\ \rho\ | < 1$$

Transformation Method
Let Tbe the Cholesky factor such that

$$T'\,T = \Sigma^{-1}$$

Define

$$Y^* = TY, X^* = TX$$

Then apply OLS to the transformed model

$$Y^* = X^*\beta + \varepsilon^*, \text{Var}\,(\varepsilon^*) = I$$

This yields GLS.

5.4.4 OTHER TIME-SERIES ERROR STRUCTURES:

- AR(p)
- ARMA(p,q)
- Random walk errors
- State-space errors

For each case, GLS remains valid with appropriate Σ^{-1}

5.5 ESTIMATION IN REPEATED-MEASURE OR CLUSTERED DATA:

If observations belong to groups with random subject effects:

$$Y_{ij} = x'_{ij}\beta + b_i + \varepsilon_{ij},$$

where

- $b_i$introduces within-subject correlation,
- $\text{Var}\,(b_i) = \tau^2$
- $\text{Var}\,(\varepsilon_{ij}) = \sigma^2$

Then

$$\Sigma_i = \sigma^2 I + \tau^2 J$$

GLS or Mixed Model Estimation (REML) is applied.

5.6 COMPARISON WITH ORDINARY LEAST SQUARES (OLS):

Feature	**OLS**	**GLS**
Error assumption	$\Sigma = \sigma^2 I$	Arbitrary positive definite Σ
Efficiency	Not efficient under correlation	Efficient
Variance formula	Incorrect if $\Sigma \neq \sigma^2 I$	Correct
BLUE?	Only under independence	Always under known Σ
Computational complexity	Simple	More complex (needs Σ^{-1})

5.7 SUMMARY:

In this unit, the concept of estimation with correlated observations has been systematically developed within the framework of linear models and applied regression analysis. Unlike the classical regression setting, where error terms are assumed to be independent, many real-world data structures such as time-series, longitudinal, repeated-measure, and clustered data exhibit correlation among observations. Ignoring such correlation leads to inefficient estimators and invalid statistical inference.

The unit began with an introduction to the nature and sources of correlated observations**,** highlighting situations where the independence assumption of errors is violated. Appropriate notations and definitions**,** particularly the variance-covariance matrix of the error vector**,** were introduced to formally represent correlation among errors. Common correlation structures, including compound symmetry, autoregressive, and block-diagonal forms**,** were discussed along with the practical problems caused by correlated errors, such as biased standard errors and misleading hypothesis tests.

To address these issues, the Generalized Least Squares (GLS) estimator was introduced as a natural extension of Ordinary Least Squares. The GLS estimator accounts for the known variance–covariance structure of the errors and was shown to be the Best Linear Unbiased Estimator (BLUE) under correlated error assumptions. When the error covariance matrix is unknown, the Maximum Likelihood Estimation (MLE) approach and feasible estimation procedures provide practical solutions.

The unit further examined bias, mean square error, and efficiency of estimators under correlated errors, emphasizing the superiority of GLS over OLS in terms of variance reduction. Methods for estimating error variance and constructing confidence intervals under correlation were also discussed. Special attention was given to autocorrelation models**,** particularly the AR(1) process**,** and other time-series error structures commonly encountered in regression analysis.

Estimation techniques for repeated-measure and clustered data were explored, illustrating how correlation within clusters affects estimation and inference. A detailed comparison between OLS and GLS highlighted that while OLS estimators remain unbiased under correlated errors, they are no longer efficient, and their associated inferential procedures become unreliable.

In conclusion, this unit establishes that proper modeling of correlation in regression errors is essential for valid estimation and inference. The use of GLS and related methods ensures

efficiency and correctness of results in the presence of correlated observations. These concepts form a crucial foundation for advanced topics in econometrics, biostatistics, longitudinal data analysis, and applied statistical modeling.

5.8 KEY WORDS:

Correlated observations
Variance–covariance matrix
Generalized Least Squares (GLS)
Autocorrelation
AR(1) process
Multivariate normality
Toeplitz matrix
Cholesky transformation
Compound symmetry
Clustered data
Linear model
Gauss–Markov theorem

5.9 SELF-ASSESSMENT QUESTIONS:

1. Define correlated observations. Give two real-life examples. State and prove the GLS estimator for β.
2. Explain estimation under linear restrictions. Derive the restricted LS estimator for a general linear constraint.
3. Why is OLS inefficient when errors are correlated? Explain the structure of the AR(1) covariance matrix.
4. Derive the variance of the GLS estimator. Show that GLS reduces to OLS when $\Sigma = \sigma^2 I$.
5. State the likelihood function under correlated normal errors. Derive the MLE of σ^2under correlated errors.
6. What is compound symmetry? Give its covariance matrix. Explain the Cholesky transformation method for GLS estimation.
7. Compare OLS and GLS in terms of variance and efficiency. Describe the covariance structure in repeated-measure models.

5.10 SUGGESTED READINGS:

1. Graybill, F.A. (1983): Matrices with Applications in Statistics. Wadsworth.
2. Draper, N.R. & Smith, H. (1998): Applied Regression Analysis, 3rd Ed. Wiley.
3. Montgomery, Peck & Vining (2012): Introduction to Linear Regression Analysis, 5th Ed. Wiley.
4. Goon, Gupta & Das Gupta (2003): An Outline of Statistical Theory, Vol. II, World Press.
5. Weisberg, S. (2013): Applied Linear Regression, 4th Ed. Wiley.

Prof. G. V. S. R. Anjaneyulu

LESSON-6

LEAST SQUARES ESTIMATES WITH RESTRICTION ON PARAMETERS

OBJECTIVES :

After completing this unit, students will be able to:

- ❖ Understand the need for parameter restrictions
- ❖ Formulate linear models with restricted parameters
- ❖ Derive the Restricted Least Squares (RLS) estimator
- ❖ Study the statistical properties of restricted estimators
- ❖ Compare restricted and unrestricted estimators
- ❖ Perform hypothesis testing using parameter restrictions
- ❖ Apply restricted least squares in practical situations.

STRUCTURE:

6.1 INTRODUCTION:

In many statistical modeling situations, particularly in regression and ANOVA, one may encounter **a** priori restrictions on the regression parameters. These restrictions can arise from:

- theoretical considerations,
- structural constraints (e.g., sum-to-zero),
- economic identities (e.g., budget constraints), or
- identifiability requirements in coded models.

The ordinary least squares (OLS) estimator does not directly incorporate such constraints. Therefore, Restricted Least Squares (RLS) provides a framework to estimate parameters subject to known linear restrictions.

This chapter develops theory, methods, proofs, and applications of least squares estimation with linear restrictions of the form: $R\beta = r$

Description:
Least Squares Estimation (LSE) is the most fundamental method used in regression analysis to estimate the unknown parameters of a linear model. In the standard linear model

$$\boldsymbol{Y = X\beta + \varepsilon}$$

the goal is to obtain estimates of the regression coefficients β that describe how the response variable varies with one or more predictor variables. The method of least squares selects parameter estimates by minimizing the sum of squared deviations between observed responses and the fitted values.

Both references—Montgomery, Peck & Vining (2012) and Weisberg (Applied Linear Regression)—emphasize that least squares methods are grounded in geometry, optimization, and probability theory. The fitted regression vector

$$\hat{Y} = X\hat{\beta}$$

represents the orthogonal projection of the observed data onto the column space of the design matrix X.

6.2 REVIEW OF ORDINARY LEAST SQUARES (OLS):

Consider the classical linear model:

$$Y = X\beta + \varepsilon$$

where

- Y is an $n \times 1$ vector of observations
- X is an $n \times p$ full column-rank design matrix
- β is a $p \times 1$ vector of unknown parameters
- ε is an $n \times 1$ vector of random errors

Assumptions:

$$\mathbb{E}(\varepsilon) = 0, \mathrm{Var}(\varepsilon) = \sigma^2 I_n$$

OLS Estimator
The OLS estimator minimizes the residual sum of squares:

$$S(\beta) = (Y - X\beta)'(Y - X\beta)$$

Theorem 1 (OLS estimator)
Claim. The vector $\hat{\beta}$that minimizes the residual sum of squares

$$S(\beta) = (Y - X\beta)'(Y - X\beta)$$

Is $\boxed{\hat{\beta} = (X'X)^{-1}X'Y}$

Proof:

1. Write the objective in expanded form

$$S(\beta) = (Y - X\beta)'(Y - X\beta) = Y'Y - 2Y'X\beta + \beta'X'X\beta$$

This is a scalar quadratic function in the vector β.

2. Compute the gradient with respect to β.
   Use standard matrix derivatives (derivative of $\beta'A\beta$is $(A + A')\beta$and derivative of a linear form $c'\beta$is c). Since $X'X$is symmetric,

$$\frac{\partial S(\beta)}{\partial \beta} = -2X'Y + 2X'X\beta$$

3. Set the gradient to zero (first-order necessary condition).

$$-2X'Y + 2X'X\hat{\beta} = 0 \Longrightarrow X'X\hat{\beta} = X'Y$$

These are the normal equations.

4. Solve the normal equations (existence and uniqueness)
   Because Xhas full column rank, $X'X$is invertible. Multiply both sides by $(X'X)^{-1}$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

This is the unique solution because the objective $S(\beta)$is strictly convex (its Hessian is $2X'X$, positive definite when $X'X$is positive definite).

5. Second-order condition (confirm minimum)
   The Hessian of $S(\beta)$is $2X'X$. Since $X'X$is positive definite, the stationary point is a strict global minimum.

Remarks:
The derivation did not require probabilistic assumptions — only matrix algebra and full rank of X. Under usual stochastic assumptions (e.g. $\mathbb{E}\varepsilon = 0$), $\hat{\beta}$has the familiar sampling properties.

Properties of OLS

1. Unbiasedness

$$\mathbb{E}(\hat{\beta}) = \beta$$

2. Variance

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

3. Orthogonality Condition
   Residuals are orthogonal to the column space of X:

$$X'(Y - X\hat{\beta}) = 0$$

4. Gauss–Markov Theorem
   $\widehat{\beta}$ is BLUE (Best Linear Unbiased Estimator).

6.3 INTRODUCTION TO PARAMETER RESTRICTIONS

Why Restrictions Occur

- Economic constraints (e.g., sum of shares equals 1)
- ANOVA coding (sum of treatment effects = 0)
- Identifiability in dummy variable regression
- Theoretical structure in time series or econometric models

Types of Restrictions

1. Equality constraints

$$\beta_1 + \beta_2 = 1$$

2. Linear constraints General form:

$$R\beta = r$$

where

- R is a $q \times p$ known restriction matrix of rank q,
- r is a $q \times 1$ vector of constants.

6.4 LEAST SQUARES ESTIMATION WITH RESTRICTIONS:

Objective
Minimize $S(\beta) = (Y - X\beta)'(Y - X\beta)$
subject to $R\beta = r$
This is a constrained optimization problem.

Method 1: Lagrange Multiplier Approach
Define the Lagrangian:

$$L(\beta, \lambda) = (Y - X\beta)'(Y - X\beta) + 2\lambda'(R\beta - r)$$

where $\lambda$is a $q \times 1$vector of Lagrange multipliers.

First-Order Conditions

1. Derivative w.r.t. β:

$$-2X'(Y - X\beta) + 2R'\lambda = 0$$
$$\Rightarrow X'X\beta - X'Y + R'\lambda = 0 \qquad (1)$$

2. Derivative w.r.t. λ:

$$R\beta = r \qquad (2)$$

Solving the System
From (1):

$$X'X\beta = X'Y - R'\lambda$$

Thus$\beta = \hat{\beta} - (X'X)^{-1}R'\lambda$
Substitute into (2):

$$R\hat{\beta} - R(X'X)^{-1}R'\lambda = r$$

Hence,

$$\lambda = [R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$$

Substitute back:

Theorem 2 (Restricted Least Squares estimator) - detailed proof

Minimize $S(\beta) = (Y - X\beta)'(Y - X\beta)$subject to linear constraints $R\beta = r$, where $R$is $q \times p$of rank $q$and $r$is $q \times 1$.
Claim. The unique constrained minimizer is

$$\boxed{\hat{\beta}_R = \hat{\beta} - (X'X)^{-1}R'[\,R(X'X)^{-1}R'\,]^{-1}(R\hat{\beta} - r)}$$

where $\hat{\beta} = (X'X)^{-1}X'Y$is the unrestricted OLS estimator and $R(X'X)^{-1}R'$is invertible because Rhas rank q.

Proof:
Constrained quadratic minimization over an affine subspace is classically handled by Lagrange multipliers. Form the Lagrangian

$$\mathcal{L}(\beta, \lambda) = (Y - X\beta)'(Y - X\beta) + 2\lambda'(R\beta - r)$$

with $\lambda \in \mathbb{R}^q$ the Lagrange multiplier vector (the factor 2 is conventional and simplifies expressions). Stationarity with respect to β yields

$$-2X'Y + 2X'X\beta + 2R'\lambda = 0$$

i.e.

$$X'X\beta + R'\lambda = X'Y \qquad (A)$$

Stationarity w.r.t. λ just returns the constraint $R\beta = r$. Solve (A) for βin terms of λ:

$$\beta = (X'X)^{-1}X'Y - (X'X)^{-1}R'\lambda = \hat{\beta} - (X'X)^{-1}R'\lambda$$

Substituting this expression into the constraint $R\beta = r$ gives the linear system for λ:

$$R\hat{\beta} - R(X'X)^{-1}R'\lambda = r$$

Because $R(X'X)^{-1}R'$is $q \times q$and full-rank, it is invertible; hence

$$\lambda = [R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$$

Substituting this λback into $\beta = \hat{\beta} - (X'X)^{-1}R'\lambda$ yields the closed-form expression for $\hat{\beta}_R$claimed above.

Interpretation. The formula shows $\hat{\beta}_R$equals the unconstrained OLS $\hat{\beta}$minus a correction that enforces the restriction $R\beta = r$. If the unrestricted $\hat{\beta}$already satisfies the restriction, i.e. $R\hat{\beta} = r$, the correction vanishes and $\hat{\beta}_R = \hat{\beta}$. Algebraically the correction is the projection (in the metric induced by $X'X$) of $\hat{\beta}$onto the affine subspace $\{\beta: R\beta = r\}$.

6.5 PROPERTIES OF RESTRICTED ESTIMATORS:

1. Bias $\mathbb{E}(\hat{\beta}_R) = \beta - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\beta - r)$

Thus:

- If true restrictions hold ($R\beta = r$)
  $\hat{\beta}_R$is unbiased.
- If restrictions are false, estimator becomes biased, but may have lower variance.

2. Variance
$$\text{Var}(\hat{\beta}_R) = \sigma^2[(X'X)^{-1} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}].$$
3. Efficiency
   - When restrictions are correct → RLS is more efficient than OLS.
   - When restrictions are incorrect → OLS may be preferable.
4. Distribution

Under normality:
$$\hat{\beta}_R \sim N(\beta_R, \text{Var}(\hat{\beta}_R))$$

6.6 TESTING THE VALIDITY OF RESTRICTIONS:

Goal: Test the hypothesis
$$H_0: R\beta = r \text{vs.} H_1: R\beta \neq r$$
Let

- RSS_0= Restricted residual sum of squares
- RSS_1= Unrestricted residual sum of squares

Theorem 3 (F-Test for Linear Restrictions)
$$F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-p)} \sim F_{q,n-p}$$
Proof:

Step 1 : Notation and projection matrices

Define the usual projection (hat) matrix and residual projector for the unrestricted model:
$$P = X(X'X)^{-1}X', M = I_n - P$$
For any estimator $\tilde{\beta}$with fitted values $X\tilde{\beta}$, its residual sum of squares is
$$RSS(\tilde{\beta}) = \| Y - X\tilde{\beta} \|^2$$
In particular, for the OLS (unrestricted) fit $\hat{\beta}$
$$RSS_1 = \| Y - X\hat{\beta} \|^2 = Y'MY$$

Step 2 : Restricted estimator and its properties

Let $\hat{\beta}_R$denote the restricted least squares estimator (minimizer of $\| Y - X\beta \|^2$subject to $R\beta = r$). From Lagrange multiplier solution we have the explicit formula
$$\hat{\beta}_R = \hat{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$$

Define the $q \times q$matrix $\Delta = R(X'X)^{-1}R'$(invertible since rank $(R) = q$)

Step 3 : Identity for RSS difference (key algebraic identity)
We claim and will use the identity

$$\boxed{RSS_0 - RSS_1 \;=\; (\hat{\beta} - \hat{\beta}_R)'X'X(\hat{\beta} - \hat{\beta}_R) \;=\; (R\hat{\beta} - r)'\,\Delta^{-1}\,(R\hat{\beta} - r)} \qquad (\bigstar)$$

Derivation (short): start from the expression of RSS as $Y'Y - 2Y'X\tilde{\beta} + \tilde{\beta}'X'X\tilde{\beta}$for $\tilde{\beta} = \hat{\beta}_R, \hat{\beta}$. Subtract and use $X'X\hat{\beta} = X'Y$(normal equations) to reduce the difference to $(\hat{\beta} - \hat{\beta}_R)'X'X(\hat{\beta} - \hat{\beta}_R)$. Substituting $\hat{\beta} - \hat{\beta}_R = (X'X)^{-1}R'\Delta^{-1}\big(R\hat{\beta} - r\big)$ yields the second equality. (You may expand the algebra step if you want it inserted verbatim; the identity is standard and follows by direct substitution.)

Step 4 : Sampling distribution of $R\hat{\beta}$
Since $\hat{\beta} = (X'X)^{-1}X'Y$is linear in $Y$and Var $(Y) = \sigma^2 I_n$

$$\text{Var}\,(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

Therefore the linear transform $R\hat{\beta}$has

$$R\hat{\beta} \sim N(R\beta,\, \sigma^2\Delta)$$

Under H_0: $R\beta = r$, we get

$$R\hat{\beta} - r \sim N(0,\, \sigma^2\Delta)$$

Step 5 :Numerator is chi-square
Using (★) and the previous distribution

$$\frac{RSS_0 - RSS_1}{\sigma^2} = \frac{(R\hat{\beta} - r)'\Delta^{-1}(R\hat{\beta} - r)}{\sigma^2} \sim \chi^2_q$$

(Quadratic form of $N(0, \sigma^2\Delta)$with Δ^{-1}yields χ^2_q)
So

$$\frac{RSS_0 - RSS_1}{q\sigma^2} \sim \frac{1}{q}\chi^2_q$$

Step 6 : Denominator is chi-square
The unrestricted residuals $e = Y - X\hat{\beta} = MY$ satisfy $e \sim N(0, \sigma^2 M)$since M is idempotent of rank $n - p$. Therefore

$$\frac{RSS_1}{\sigma^2} = \frac{e'e}{\sigma^2} = \frac{Y'MY}{\sigma^2} \sim \chi^2_{n-p}$$

So

$$\frac{RSS_1/(n-p)}{\sigma^2/(n-p)} = \frac{RSS_1}{(n-p)\sigma^2} \sim \frac{1}{n-p}\chi^2_{n-p}$$

Step 7 : Independence of numerator and denominator
To form an F statistic we require independence of the two chi-square quantities. This follows because the two quadratic forms are based on orthogonal projection matrices.

Reason: The quadratic for the numerator can be written as $Y'AY$ where $A$is symmetric idempotent of rank $q$projecting onto the subspace of variation captured by releasing the q constraints (equivalently $A = H_0 - H_1$, difference between the projection matrices of restricted and unrestricted fits). The denominator quadratic is $Y'MY$ where M is the residual projector of rank $n - p$. One can show $AM = 0$ (the projection subspaces are orthogonal).

For a Gaussian vector $Y \sim N(X\beta, \sigma^2 I)$, quadratic forms $Y'AY$and $Y'MY$ with $AM = 0$ are independent. Hence the two chi-square variables are independent.

(If you prefer: expand both as functions of ε; they are quadratic forms in $\varepsilon$with coefficient matrices that multiply to zero, which implies independence under normality.)

Step 8 : Construct the F statistic and its distribution
Given independence and the chi-square results:

- $(RSS_0 - RSS_1)/\sigma^2 \sim \chi^2_q$,
- $RSS_1/\sigma^2 \sim \chi^2_{n-p}$,
- independent,

the ratio

$$\frac{((RSS_0 - RSS_1)/\sigma^2)/q}{(RSS_1/\sigma^2)/(n-p)} = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-p)}$$

is the ratio of an independent $\chi^2_q/q$and $\chi^2_{n-p}/(n-p)$, hence has the $F_{q,n-p}$ distribution. Thus

$$\boxed{F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-p)} \sim F_{q,\,n-p}}$$

under H_0

Step 9 : Decision rule For significance level α:

- Reject H_0if $F_{\text{cal}} > F_{q,n-p;\,1-\alpha}$, where $F_{q,n-p;\,1-\alpha}$denotes the $1-\alpha$quantile of the $F_{q,n-p}$distribution.
- Otherwise do not reject H_0.
- Equivalently compute the p-value $P(F_{q,n-p} \geq F_{\text{cal}})$and reject when p-value $\leq \alpha$.
- Step 10 : Interpretation
- If $F_{\text{cal}}$is large (reject H_0), the increase in residual sum of squares produced by imposing $R\beta = r$is too big relative to sampling variability; thus the restrictions are inconsistent with the data.
- If $F_{\text{cal}}$is not large (fail to reject H_0), the data are compatible with the restrictions; enforcing them does not degrade fit more than can be explained by sampling variability.
- Step 11 : Special cases and connections
- If $q = 1$(single linear restriction), the numerator reduces to a squared t-statistic and $F = t^2$.
- The same F–test is the general linear hypothesis test for nested linear models: unrestricted model (larger) vs restricted (nested) model.
- Step 12 : Caveats
- Exact validity of the F-distribution requires the normality assumption $\varepsilon \sim N(0, \sigma^2 I)$. Without normality, the test is approximately valid asymptotically (large n) under mild conditions.
- Ensure $n - p > 0$and $R(X'X)^{-1}R'$invertible (independent restrictions).

$$\boxed{F_{\text{cal}} = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-p)} \sim F_{q,\,n-p} \text{under } H_0: R\beta = r}$$

Decision: Reject H_0iff $F_{\text{cal}} > F_{q,n-p;1-\alpha}$.

6.7 APPLICATIONS:

1. Economic Models with Budget Constraints
Demand share equations:

$$\beta_1 + \beta_2 + \beta_3 = 1$$

Thus,

$$R = [111], r = [1]$$

2. ANOVA Models with Sum-to-Zero Constraints
For treatment effects:

$$\sum_{i=1}^{k} \alpha_i = 0$$

This ensures identifiability of parameters.
3. Regression Models with Equality Restrictions
Example: Parallel-line regression with equal slopes:

$$\beta_2 = \beta_3$$
$$R = [01 - 10], r = [0]$$

6.8 KEY WORDS:

- Ordinary Least Squares (OLS)
- Restricted Least Squares (RLS)
- Linear Restrictions
- Hypothesis Testing
- General Linear Hypothesis
- F-Test
- Efficiency
- ANOVA Constraints
- Econometric Restrictions
- Lagrange Multiplier
- Bias, Variance, Mean Square Error

6.9 SUMMARY:

In this lesson, the concept of restricted least squares estimation has been studied as an extension of the ordinary least squares (OLS) method when prior information or theoretical constraints on regression parameters are available. Such restrictions commonly arise in economics, experimental design, and applied regression problems where parameters are known to satisfy linear relationships.

The lesson began with a review of the ordinary least squares estimator, highlighting its optimal properties under the standard linear model assumptions. The need for parameter restrictions was then motivated by situations where exact linear constraints of the form $R\beta = r$
are imposed on the regression coefficients.
The restricted least squares estimator (RLSE) was derived using the method of Lagrange multipliers, and its explicit matrix form was obtained. The estimator was shown to incorporate both the sample information and the imposed restrictions, thereby modifying the OLS estimator to satisfy the given constraints.

The statistical properties of the restricted estimator were discussed in detail. When the restrictions are correct, the restricted estimator remains unbiased and has a smaller variance than the unrestricted OLS estimator, leading to improved efficiency. However, incorrect restrictions may introduce bias, emphasizing the importance of testing the validity of restrictions before their adoption.

Methods for testing linear restrictions using appropriate test statistics were presented, allowing formal comparison between restricted and unrestricted models. Practical applications illustrated how restricted least squares estimation can simplify models, improve precision, and enhance interpretability.

- Restricted least squares provides a systematic way to incorporate prior information into regression estimation.
- When restrictions are correct, RLS estimators are more efficient than OLS estimators.
- Incorrect restrictions can lead to biased estimates, making testing of restrictions essential**.**
- Restricted estimation plays a crucial role in model validation, econometric analysis, and experimental studies**.**
-

Overall, least squares estimation with restrictions enhances both the theoretical rigor and practical applicability of linear regression models.

6.10 SELF-ASSESSMENT QUESTIONS:

1. State the restricted least squares estimator and derive it using the Lagrange multiplier method.
2. Explain the difference between OLS and RLS estimators in terms of bias and variance.
3. Give two practical situations where linear restrictions arise.
4. Prove that if $R\beta = r$is true, then $\hat{\beta}_R$is unbiased. Derive the variance of the restricted estimator.
5. How do you test whether restrictions $R\beta = r$are valid?
6. In ANOVA, why is $\sum \alpha_i = 0$imposed? Show that the restricted estimator has smaller variance than OLS when restrictions are valid.
7. Derive the distribution of $\hat{\beta}_R$under normality assumptions.

6.11 SUGGESTED READING:

1. Rao, C. R. Linear Statistical Inference and Its Applications
2. Seber & Lee, Linear Regression Analysis
3. Graybill, F.A. (1983): Matrices with Applications in Statistics. Wadsworth
4. Draper, N.R. & Smith, H. (1998): Applied Regression Analysis, 3rd Ed. Wiley
5. Montgomery, Peck & Vining (2012): Introduction to Linear Regression Analysis, 5th Ed. Wiley
6. Goon, Gupta & Das Gupta (2003): An Outline of Statistical Theory, Vol. II, World Press
7. Weisberg, S. (2013): Applied Linear Regression, 4th Ed. Wiley.

Prof. G. V. S. R. Anjaneyulu

LESSON-7

SIMULTANEOUS ESTIMATES OF LINEAR PARAMETRIC FUNCTIONS

OBJECTIVES :

After completing this lesson, students will be able to:

- Understand linear parametric functions **-** Define and interpret linear functions of regression parameters.
- Formulate simultaneous estimation problems - Express multiple linear parametric functions in matrix form.
- Derive estimators for linear parametric functions - Obtain estimators using least squares principles.
- Study variance–covariance structure - Compute and interpret the joint variance–covariance matrix of simultaneous estimators.
- Apply simultaneous inference techniques - Construct confidence regions and perform joint hypothesis tests.
- Understand efficiency and optimality - Identify conditions under which estimators are unbiased and minimum variance.
- Relate simultaneous estimation to hypothesis testing - Connect estimation of parametric functions with general linear hypotheses.
- Apply concepts to practical problems - Use simultaneous estimation in ANOVA, regression contrasts, and applied data analysis.

STRUCTURE:

7.1. INTRODUCTION:

In regression analysis, interest often lies not in individual regression coefficients, but in linear parametric functions of these coefficients. Common examples include treatment contrasts, rates of change, and combined effects of predictors. When several such linear functions are estimated simultaneously, classical single-parameter confidence intervals are inadequate, because the overall probability of making at least one incorrect inference increases with the number of intervals constructed.

To remedy this, statistical theory provides simultaneous inference procedures, such as the joint confidence region, Bonferroni method, and Scheffé method, that control the overall family-wise error rate and provide valid inference for multiple linear functions. These methods form a central part of higher-level regression analysis and multivariate statistical inference. They are extensively presented in standard sources such as Montgomery et al. (2012), Draper and Smith (1998), Weisberg (2005), and Rao (2002).

Description:
In linear regression, we often need to estimate several functions of model parameters at the same time. These quantities, called linear parametric functions (such as $a^T\beta$), include regression coefficients, contrasts, and predicted responses. Constructing separate confidence intervals for each function can lead to an inflated overall error rate. Simultaneous estimation provides methods to control this error and maintain a specified joint confidence level for all functions considered together.

Three major approaches are used:

- Joint confidence regions, based on the multivariate normal distribution of least-squares estimators.
- Bonferroni simultaneous intervals, which are simple to apply and guarantee overall confidence.
- Scheffé's method, which gives valid intervals for all possible linear combinations of parameters.

Simultaneous estimation is essential for reliable inference when studying multiple parameter relationships in regression and experimental design.

7.2. Linear Parametric Functions:

Consider the classical linear regression model

$$Y = X\beta + \varepsilon$$

where

Y is an $n \times 1$ vector of observations,

X is an $n \times p$full-rank design matrix,

β is a $p \times 1$vector of unknown regression parameters, and $\varepsilon \sim N(0, \sigma^2 I_n)$.

A linear parametric function of the regression parameters is defined as

$$L = a'\beta$$

where a is a known $p \times 1$ constant vector.

Examples include:

- Contrasts: $\beta_1 - \beta_2$
- Weighted combinations: $2\beta_1 + 5\beta_3$

- A predicted value at a design point x_0, where $a = x_0$

The estimator of $L$is given by

$$\hat{L} = a'\hat{\beta}$$

with sampling variance

$$Var\,(\hat{L}) = \sigma^2 a'(X'X)^{-1}a$$

This result follows from linear properties of least squares estimators (see Montgomery et al., 2012, Ch. 3; and Weisberg, 2005, Sec. 3.4.4).

7.3. SIMULTANEOUS ESTIMATION PROBLEM:

Suppose we wish to estimate $m$linear parametric functions

$$L_i = a_i'\beta, i = 1,2,\dots,m$$

with simultaneous confidence levels.

If each function is estimated separately using a$1-\alpha$ confidence interval, the joint confidence that all intervals are correct is less than $1-\alpha$. Specifically,

$$P(all\ intervals\ correct) \leq 1 - m\alpha$$

Thus, special procedures are required to maintain a prescribed family-wise confidence level $1-\alpha$. The three most widely used are:

1. Joint confidence region
2. Bonferroni simultaneous intervals
3. Scheffé simultaneous intervals.

7.4. NOTATIONS & DEFINITIONS:

- $\hat{\beta} = (X'X)^{-1}X'Y$: least squares estimator
- $s^2 = \frac{(Y-X\hat{\beta})'(Y-X\hat{\beta})}{n-p}$: unbiased estimator of σ^2
- $t_{v,\alpha}$: t-distribution quantile with $v$degrees of freedom,
- $F_{p,n-p;\alpha}$: upper α-point of F-distribution with $(p, n-p)$degrees of freedom
- Family-wise error rate (FWER): probability of at least one false rejection.

7. 5. THEOREMS:

7. 5.1 joint confidence region:

Notation / model reminders (used throughout)

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

X is $n \times p$ of full column rank p. Ordinary least squares estimator:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

and the unbiased estimator of σ^2

$$s^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n-p}.$$

Key facts used below:

- $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$
- Residual sum of squares $RSS = (Y - X\hat{\beta})'\big(Y - X\hat{\beta}\big)$satisfies $\frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$ and is independent of $\hat{\beta}$.

Theorem -1 :

Joint confidence region for β Claim. A $100(1-\alpha)\%$ joint confidence region for β is

$$(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta) \ \leq \ p\, s^2\, F_{p,\,n-p;\,\alpha}$$

Proof:

1. Distribution of the centered estimator in quadratic form.
   From $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$, set
   $$u = \sigma^{-1}(X'X)^{1/2}(\hat{\beta} - \beta)$$
   Then $u \sim N(0, I_p)$. Hence the sum of squares
   $$u'u = \frac{(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta)}{\sigma^2}$$
   has the χ^2_pdistribution:
   $$\frac{(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta)}{\sigma^2} \sim \chi^2_p$$
2. Distribution of $RSS/\sigma^2$and independence
   As noted, $RSS/\sigma^2 \sim \chi^2_{n-p}$and it is independent of $\hat{\beta}$(thus independent of the quadratic form above).

3. Form an F ratio
   The standard construction using two independent chi-square variates gives
   $$\frac{\frac{1}{p} \cdot \frac{(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)}{\sigma^2}}{\frac{1}{n-p} \cdot \frac{RSS}{\sigma^2}} = \frac{(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta)/p}{RSS/(n-p)} \sim F_{p,n-p}.$$
4. Replace unknown σ^2 by $s^2 = \frac{RSS}{n-p}$ and invert the F inequality.
   For the upper α - point $F_{p,n-p;\alpha}$, we have
   $$P\left(\frac{(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta)/p}{s^2} \leq F_{p,n-p;\alpha}\right) = 1 - \alpha$$

Multiply both sides by $p\, s^2$to obtain the confidence region:

$$P\left((\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta) \leq p\, s^2 F_{p,n-p;\alpha}\right) = 1 - \alpha$$

5. Interpretation
   The set of all βsatisfying the displayed inequality is an ellipsoid centered at $\hat{\beta}$. This is the joint $100(1-\alpha)\%$ confidence region.

7.5.2 bonferroni simultaneous confidence intervals:

Theorem - 2 : Bonferroni simultaneous confidence intervals

Claim. For $m$specified linear functions $L_i = a_i'\beta (i = 1, \dots, m)$, the intervals

$$\hat{L}_i \pm t_{n-p;\, \alpha/(2m)}\; s \sqrt{a_i'(X'X)^{-1}a_i} (i = 1, \dots, m)$$

form a $100(1-\alpha)\%$ simultaneous confidence system; i.e.

$$P(\forall i,\ L_i \ lies \ in \ its \ interval) \geq 1 - \alpha$$

Proof:

1. Sampling distribution of each standardized estimate.
   For a fixed i,
   $$\hat{L}_i = a_i'\hat{\beta}$$
   and
   $$Var\,(\hat{L}_i) = \sigma^2 a_i'(X'X)^{-1}a_i$$
   Therefore

$$T_i = \frac{\hat{L}_i - L_i}{s\sqrt{a_i'(X'X)^{-1}a_i}}$$

has a Student t-distribution with $n-p$ degrees of freedom (because the numerator is normal and independent of s^2).

2. Individual $1-\alpha_i$interval.
   For any chosen α_i, an individual two-sided $100(1-\alpha_i)\%$ CI is
   $$\hat{L}_i \pm t_{n-p;\,\alpha_i/2}\; s\sqrt{a_i'(X'X)^{-1}a_i}$$
3. Use Bonferroni inequality to control family-wise error.
   Let $A_i$be the event "interval $i$contains L_i". We want $P(\bigcap_{i=1}^{m}A_i)$. By Bonferroni (union bound)
   $$P(\bigcap_{i=1}^{m}A_i) = 1 - P(\bigcup_{i=1}^{m}A_i^c) \geq 1 - \sum_{i=1}^{m}P(A_i^c)$$
   If we choose each interval to have coverage $1-\alpha/m$(i.e. $\alpha_i = \alpha/m$), then $P(A_i^c) = \alpha/m$ and therefore
   $$P(\bigcap_{i=1}^{m}A_i) \geq 1 - \sum_{i=1}^{m}\frac{\alpha}{m} = 1-\alpha$$
4. Conclusion
   Thus the $m$intervals constructed with individual tail probability $\alpha/(2m)$(two-sided) guarantee overall coverage at least $1-\alpha$. This is the Bonferroni simultaneous CI construction. (Reference and discussion: Montgomery / Draper & Smith, and standard texts on multiple comparisons.)

Remarks: The Bonferroni bound is conservative (inequality), since it ignores correlations among the tests. When the $\hat{L}_i$ are positively correlated the bound may be loose; but it is simple and guaranteed.

7.5.3 scheffe simultaneous confidence intervals:

Theorem-3 : Scheffé simultaneous confidence intervals

Claim. For any linear parametric function $L = a'\beta$(with arbitrary a), the interval

$$\hat{L} \pm \sqrt{p\,F_{p,n-p;\,\alpha}}\; s\,\sqrt{a'(X'X)^{-1}a}$$

is a $100(1-\alpha)\%$ simultaneous confidence interval valid simultaneously for all a. In other words, the stated multiplier guarantees that the stated interval contains L for every linear combination $a$with probability at least $1-\alpha$.

Proof:

1. Start from the joint confidence ellipsoid.
   By Theorem 1 (joint region), with probability $1-\alpha$,
   $$(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta) \leq p\,s^2\,F_{p,n-p;\alpha}$$
   This describes an ellipsoid of admissible β around $\hat{\beta}$.
2. Relate scalar deviations to the quadratic form (Cauchy–Schwarz).
   For any fixed$a$,

$$(a'(\hat{\beta}-\beta))^2 = ([(X'X)^{-1/2}a]'\,[(X'X)^{1/2}(\hat{\beta}-\beta)])^2 \le \| (X'X)^{-1/2}a \|^2 \| (X'X)^{1/2}(\hat{\beta}-\beta) \|^2 \; (Cauchy-Schwarz)$$
$$= a'(X'X)^{-1}a \cdot (\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)$$

Thus the squared scalar error $(\hat{L}-L)^2$ is bounded by the quadratic form times the factor $a'(X'X)^{-1}a$.

3. Insert the joint region bound
   On the event that the joint ellipsoid inequality holds
   $$(\hat{L}-L)^2 \le a'(X'X)^{-1}a \cdot p\,s^2\,F_{p,n-p;\alpha}$$
4. Take square roots and rearrange to an interval
   Therefore, with probability at least $1-\alpha$, for every a:
   $$|\,\hat{L}-L\,| \le \sqrt{p\,F_{p,n-p;\alpha}}\; s\; \sqrt{a'(X'X)^{-1}a}$$

which is exactly the Scheffé interval claim:

$$L \in [\hat{L} \pm \sqrt{p\,F_{p,n-p;\alpha}}\; s\; \sqrt{a'(X'X)^{-1}a}]$$

simultaneously for all a.

5. Interpretation/justification
Scheffé's multiplier arises because the worst-case scalar error across all directions a is controlled by the largest possible projection of the joint ellipsoid along that direction; Cauchy–Schwarz gives the necessary inequality. This yields an interval valid for infinitely many linear combinations (all a), not only a finite prespecified set. See Draper & Smith, Rao, and Montgomery for details and geometric discussion.

7.5.4 example:

1. Consider a three-parameter regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

We wish to estimate simultaneously:

$$L_1 = \beta_1 + \beta_2, L_2 = 2\beta_1 - \beta_2$$

1. Construct Bonferroni intervals using significance level $\alpha/4$
2. Construct Scheffé interval using $p = 3$
3. Compare widths and discuss efficiency

(Instructor can insert numerical data and matrix computations.)

7.6 COMPARISON OF METHODS:

Method	Purpose	Protection	Interval Width
Bonferroni	Finite set of $m$functions	Good	Narrower
Scheffé	All possible linear functions	Very strong	Usually wider
Joint region	Geometric interpretation	Global	Not interval-based

Scheffé is recommended when the number of potential linear combinations is large or not predetermined (e.g., contrasts in ANOVA). Bonferroni is preferred when there are only a few specific linear functions to be tested (e.g., pairwise comparisons).

7.7 KEY WORDS:

- Linear model
- Contrast
- Confidence ellipsoid
- F-distribution
- Bonferroni procedure
- Scheffé method
- Family-wise error rate
- Quadratic form
- Least squares estimation

7.8 SUMMARY:

This unit focused on the theory and application of simultaneous estimation of linear parametric functions within the framework of linear models. A linear parametric function is any function of the regression parameters that can be expressed as a linear combination, such as contrasts, differences of means, or predicted values.

The concept of simultaneous estimation arises when interest lies not in a single parameter or function, but in several linear functions of the parameters considered jointly**.** Estimating these functions simultaneously allows proper assessment of their joint variability and correlation, leading to valid statistical inference.

The unit demonstrated that linear parametric functions can be conveniently expressed in matrix form, and their estimators are obtained directly from the least squares estimator of the parameter vector. The variance–covariance matrix of the estimators plays a central role in understanding the precision and dependence among estimated functions.

Emphasis was placed on simultaneous confidence regions and joint hypothesis testing**,** highlighting the inadequacy of separate (individual) confidence intervals when multiple inferences are made. Methods such as F-tests and Wald tests were shown to provide appropriate tools for testing sets of linear hypotheses.

Theoretical results established that simultaneous estimators retain desirable properties such as unbiasedness and minimum variance**,** provided the underlying linear model assumptions are satisfied. The unit also illustrated how these methods are applied in practice, particularly in analysis of variance, regression contrasts, and prediction problems**.**

In conclusion, simultaneous estimation of linear parametric functions provides a **powerful and unified framework** for making joint inferences in linear models. It enhances the reliability of conclusions drawn from regression and experimental data and forms a critical link between estimation theory and applied statistical analysis.

7.9 SELF-ASSESSMENT QUESTIONS:

1. Define a linear parametric function and provide two examples. Explain why simultaneous estimation is necessary in regression inference.
2. Derive the variance of $\hat{L} = a'\hat{\beta}$ and prove the Scheffé simultaneous confidence theorem.

3. Discuss estimation when observations are correlated and provide an example.
4. Compare Bonferroni and Scheffé methods in terms of family-wise error control.
5. Provide a numerical example of simultaneous intervals using real or simulated data.

7.10 SUGGESTED READING:

1. Graybill, F.A. (1983): Matrices with Applications in Statistics. Wadsworth.
2. Draper, N.R. and Smith, H. (1998): Applied Regression Analysis. Wiley-Blackwell.
3. Montgomery, D.C., Peck, E.A. and Vining, G.G. (2012): Introduction to Linear Regression Analysis, 5th Ed. Wiley.
4. Bapat, R.B. (2012): Linear Algebra and Linear Models. Springer.
5. Rao, C.R. (2002): Linear Statistical Inference and Its Applications. 2nd Ed. Wiley-Blackwell.
6. Weisberg, S. (2013): Applied Linear Regression, 4th Ed. Wiley.

Prof. G. V. S. R. Anjaneyulu

LESSON-8

TEST OF HYPOTHESES FOR ONE AND MORE THAN ONE LINEAR PARAMETRIC FUCTIONS

OBJECTIVES:

After studying this lesson, the learner will be able to:

- Understand the role of linear parametric functions in statistical inference under the general linear model.
- Identify and verify the estimability of single and multiple linear parametric functions using matrix rank and row space conditions.
- Formulate null and alternative hypotheses involving one or more linear parametric functions of regression parameters.
- Derive and apply appropriate test statistics for testing hypotheses on linear parametric functions.
- Distinguish between tests based on single restrictions (t-tests) and multiple linear restrictions (F-tests).

STRUCTURE:

8.1. INTRODUCTION:

In the theory of linear models, statistical inference concerning unknown parameters occupies a central role. After estimation, the next fundamental task is testing statistical hypotheses about model parameters or functions thereof. In many practical situations, interest does not lie directly in the individual regression coefficients but rather in linear parametric functions of the form $L\beta$, where β denotes the vector of unknown parameters. Examples include testing

the equality of regression coefficients, testing the significance of subsets of regressors, and testing linear restrictions arising from scientific or economic theory.

This chapter develops hypothesis testing procedures for:

- a single linear parametric function, and
- several linear parametric functions simultaneously

within the framework of the General Linear Model (GLM). The treatment emphasizes matrix formulation, distributional results, exact tests (t- and F-tests), and connections with estimation theory. The exposition follows classical developments found in Graybill, Rao, Draper and Smith, Montgomery-Peck-Vining, and related standard references.

Description

This lesson presents a systematic treatment of hypothesis testing for one and more than one linear parametric function under the general linear model. Emphasis is placed on the formulation of linear hypotheses, estimability conditions, and the derivation of exact t-and F-tests based on least squares estimation. The equivalence between the general linear hypothesis and the extra sum of squares principle is highlighted, along with the construction of confidence intervals and confidence regions. The chapter provides a theoretical foundation for testing linear restrictions commonly encountered in regression analysis and related statistical applications.

8.2. CONCEPT OF LINEAR PARAMETRIC FUNCTION:

8.2.1 general linear model:

Consider the general linear model

where:

- Y is an $n \times 1$ random vector of observations
- X is a known $n \times p$ design matrix of rank $r \leq p$
- β is a $p \times 1$ vector of unknown parameters
- $\varepsilon \sim N_n(0, \sigma^2 I_n)$

8.2.2 definition:

A linear parametric function is any function of the form

$$\theta = L\beta$$

where L is a known $q \times p$ matrix of constants.

Special cases:

- $q = 1$: one linear parametric function
- $q > 1$: several linear parametric functions

8.2.3 estimability:

A linear parametric function $L\beta$ is said to be estimable if there exists a linear estimator $a'Y$ such that $E(a'Y) = L\beta$

Theorem 1: Estimability of a Linear Parametric Function

Statement

A linear parametric function $L\beta$ is estimable under the general linear model

$$Y = X\beta + \varepsilon$$

if and only if each row of L belongs to the row space of the design matrix X

Proof

Step 1: Definition of estimability
A parametric function $L\beta$ is estimable if there exists a linear estimator $a'Y$ such that

$$E(a'Y) = L\beta$$

Step 2: Evaluate the expectation
Since

$$E(Y) = X\beta$$
$$E(a'Y) = a'X\beta$$

Step 3: Necessary condition
For

$$a'X\beta = L\beta$$

to hold for all β, we must have

$$a'X = L$$

Step 4: Interpretation
The equality $a'X = L$ implies that each row of L is a linear combination of the rows of X. Hence, the rows of L lie in the row space of X.

Step 5: Sufficiency
Conversely, if the rows of L lie in the row space of X, then there exists a vector a such that $a'X = L$. Therefore,

$$E(a'Y) = L\beta$$

and $L\beta$ is estimable.

Hence proved.

Theorem 2: Unbiasedness of the Least Squares Estimator of an Estimable Function
Statement:
If $L\beta$ is estimable, then the least squares estimator $L\hat{\beta}$ is an unbiased estimator of $L\beta$.

Proof:
Step 1: Least squares estimator
The least squares estimator of β is

$$\hat{\beta} = (X'X)^{-}X'Y$$

where $(X'X)^{-}$ is a generalized inverse

Step 2: Take expectation

$$E(\hat{\beta}) = (X'X)^{-}X'E(Y) = (X'X)^{-}X'X\beta$$

Step 3: Use estimability condition
Since $L\beta$ is estimable, there exists a matrix C such that

$$L = CX$$

Step 4: Evaluate expectation of $L\hat{\beta}$

$$E(L\hat{\beta}) = LE(\hat{\beta}) = CX(X'X)^{-}X'X\beta = CX\beta = L\beta$$

Thus,

$$\text{Bias}(L\hat{\beta}) = 0$$

Hence proved.

Theorem 3: Variance of an Estimable Linear Parametric Function
Statement:

$$\mathbf{Var}\,(\boldsymbol{L\hat{\beta}}) = \boldsymbol{\sigma^2 L(X'X)^- L'}.$$

Proof:

Step 1: Variance of the least squares estimator

$$\text{Var}\,(\hat{\beta}) = \sigma^2 (X'X)^-.$$

Step 2: Linear transformation rule

For any constant matrix L,

$$\text{Var}\,(L\hat{\beta}) = L\text{Var}\,(\hat{\beta})L'.$$

Step 3: Substitution

$$\text{Var}\,(L\hat{\beta}) = \sigma^2 L(X'X)^- L'.$$

Hence proved.

Theorem 5: F-Distribution for Multiple Linear Parametric Functions

Statement

Under the hypothesis $H_0: L\beta = c$,

$$\frac{(L\hat{\beta} - c)'[L(X'X)^- L']^{-1}(L\hat{\beta} - c)}{q\sigma^2} \sim F_{q,\, n-r}.$$

Proof:

Step 1: Distribution of $L\hat{\beta}$

$L\hat{\beta}$ follows a multivariate normal distribution with mean $L\beta$ and covariance matrix $\sigma^2 L(X'X)^- L'$

Step 2: Quadratic form

Under H_0, the quadratic form

$$(L\hat{\beta} - c)'[L(X'X)^- L']^{-1}(L\hat{\beta} - c)$$

follows a $\sigma^2 \chi_q^2$distribution.

Step 3: Error sum of squares

Independently,

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-r}$$

Step 4: Ratio of independent chi-square variables

The ratio of two independent chi-square variables divided by their degrees of freedom follows an F-distribution.

Hence proved.

8.3 HYPOTHESIS TESTING FOR ONE LINEAR PARAMETRIC FUNCTION

8.3.1 statement of hypothesis

Let $\theta = l'\beta$be an estimable linear parametric function. Consider testing

$$H_0: l'\beta = l'\beta_0 \text{vs.} H_1: l'\beta \neq l'\beta_0$$

8.3.2 estimator and distribution

Let $\hat{\beta}$ be the least squares estimator. Then

$\hat{\theta} = l'\hat{\beta}$

is unbiased with variance

$$\text{Var}\,(\hat{\theta}) = \sigma^2 l'(X'X)^{-}l$$

where $(X'X)^{-}$ denotes a generalized inverse.

Moreover,

$$\frac{\hat{\theta} - l'\beta}{\sigma\sqrt{l'(X'X)^{-}l}} \sim N(0,1)$$

8.3.3 test statistic:

Replacing $\sigma^2$by its unbiased estimator $\hat{\sigma}^2 = \text{SSE}/(n-r)$, the test statistic is

$$T = \frac{\hat{\theta} - l'\beta_0}{\hat{\sigma}\sqrt{l'(X'X)^{-}l}} \sim t_{n-r}$$

8.3.4 decision rule

Reject H_0 at level α if

8.4 HYPOTHESIS TESTING FOR MORE THAN ONE LINEAR PARAMETRIC FUNCTION:

8.4.1 general linear hypothesis

Let L be a $q \times p$ matrix of rank q. Consider testing

$$H_0: L\beta = c \text{vs.} H_1: L\beta \neq c$$

This is known as the general linear hypothesis Quadratic Form

Define $Q = (L\hat{\beta} - c)'[L(X'X)^{-}L']^{-1}(L\hat{\beta} - c)$

Then $\frac{Q}{q\sigma^2} \sim F_{q,\,n-r}$

8.4.3 f-test statistic

Replacing σ^2 by $\hat{\sigma}^2$, the test statistic becomes

$$F = \frac{1}{q\hat{\sigma}^2}(L\hat{\beta} - c)'[L(X'X)^{-}L']^{-1}(L\hat{\beta} - c)$$

Reject H_0 if

$$F > F_{\alpha;q,\,n-r}$$

8.4.4 connection with extra sum of squares

The general linear hypothesis test is equivalent to the extra sum of squares principle:

$$F = \frac{(\text{SSE}_R - \text{SSE}_F)/q}{\text{SSE}_F/(n-r)}$$

where R and F denote the reduced and full models, respectively.

8.5 BIAS AND MEAN SQUARE ERROR (MSE):

8.5.1 bias

An estimator $\hat{\theta}$of θhas bias

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

For least squares estimators of estimable linear parametric functions

$$\text{Bias}(l'\hat{\beta}) = 0$$

8.5.2 mean square error

$$\text{MSE}(\hat{\theta}) = \text{Var}\,(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$$

Thus, for unbiased estimators

$$\text{MSE}(l'\hat{\beta}) = \sigma^2 l'(X'X)^{-}l$$

8.6 APPLICATIONS AND EXAMPLES:

Example 1: Testing a Single Regression Coefficient
In a multiple regression model, test $H_0: \beta_j = 0$. This is a special case with $l' = (0, \dots, 1, \dots, 0)$.

Example 2: Equality of Two Coefficients
Test $H_0: \beta_1 = \beta_2$, equivalently $H_0: (1, -1, 0, \dots, 0)\beta = 0$.

Example 3: Joint Significance of Predictors
Test $H_0: \beta_2 = \beta_3 = \beta_4 = 0$, a multivariate linear hypothesis commonly used in model adequacy assessment.

8.7 CONFIDENCE REGIONS FOR MULTIPLE PARAMETERS:

A $(1-\alpha)$confidence region for $L\beta$is given by

$$(L\hat{\beta} - L\beta)'[L(X'X)^{-}L']^{-1}(L\hat{\beta} - L\beta) \leq q\hat{\sigma}^2 F_{\alpha;q,n-r}$$

This region is an ellipsoid in $\mathbb{R}^q$.

8.8 KEY WORDS:

- General Linear Model
- Linear Parametric Function
- Estimability
- Least Squares Estimator
- General Linear Hypothesis
- t-test
- F-test
- Extra Sum of Squares
- Mean Square Error
- Confidence Region

8.9 SUMMARY:

This lesson focused on hypothesis testing for linear parametric functions within the framework of the General Linear Model (GLM)**.** A linear parametric function, expressed as a

linear combination of regression parameters, plays a central role in statistical inference in linear models.

The concept of estimability was emphasized as a prerequisite for meaningful inference. Conditions and theorems ensuring estimability were discussed, highlighting that only estimable functions of parameters can be unbiasedly estimated and tested.

For a single linear parametric function**,** hypothesis testing procedures were developed using t-tests**,** relying on the least squares estimator and its variance. These tests allow researchers to assess the significance of specific linear combinations of parameters.

For multiple linear parametric functions**,** joint hypothesis testing was introduced using F-tests (or equivalently Wald tests). This approach enables simultaneous testing of several restrictions, ensuring control over the overall error rate and providing a unified framework for multivariate inference.

The role of bias and mean square error (MSE) in hypothesis testing was examined, demonstrating how estimator efficiency affects test performance. The construction of confidence intervals and confidence regions further complemented hypothesis testing by providing interval-based inference for one or more parametric functions.

Applications illustrated how these tests are widely used in regression analysis, analysis of variance, econometrics, and experimental design, where practical decisions often depend on testing single or multiple parameter functions simultaneously.

In conclusion, hypothesis testing for linear parametric functions forms a fundamental bridge between estimation and inference in linear models. Mastery of these methods equips students with the tools required for rigorous statistical analysis and sound decision-making in applied research.

8.10 SELF ASSESSMENT QUESTIONS:

1. Explain the concept of linear parametric functions with suitable examples.
2. Discuss the estimability of linear functions and state the relevant theorems.
3. Derive the test statistic for testing a single linear parametric function in the general linear model.
4. Explain the procedure for testing more than one linear parametric function simultaneously**.**
5. Describe the F-test for general linear hypotheses**.**
6. Explain how bias and mean square error (MSE) are related to hypothesis testing in linear models.
7. Discuss the construction of confidence regions for multiple linear parametric functions.
8. Explain the importance of testing linear parametric functions in applied regression analysis.

8.10 SUGGESTED READING:

1. Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, 2nd Edition, Wiley.
   (Classic reference for estimability and general linear hypotheses)
2. Graybill, F. A. (1976), *Theory and Application of the Linear Model*, Duxbury Press.
3. Draper, N. R. & Smith, H. (1998), *Applied Regression Analysis*, 3rd Edition, Wiley.
4. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005), *Applied Linear Statistical Models*, McGraw-Hill.
5. Seber, G. A. F. & Lee, A. J. (2003), *Linear Regression Analysis*, 2nd Edition, Wiley.
6. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012), *Introduction to Linear Regression Analysis*, Wiley.

Dr. U. Ramkiran

LESSON-9

CONFIDENCE INTERVALS AND CONFIDENCE REGIONS

OBJECTIVES:

After completing this lesson, learners will be able to:

- Understand the fundamental ideas of interval estimation, including the interpretation of confidence levels.
- Construct confidence intervals for key population parameters such as the mean, proportion, variance, and regression coefficients.
- Apply sampling distributions (normal, t, chi-square, F) in deriving interval estimators.
- Explain the concept of confidence regions and develop joint confidence regions for multiple parameters using matrix notation and multivariate distributions.
- Interpret the geometric meaning of elliptical confidence regions in regression and multivariate analysis.
- Compare marginal and joint confidence intervals and evaluate the precision of estimates based on interval width and coverage.

STRUCTURE:

9.1 INTRODUCTION:

Statistical estimation involves providing plausible values for unknown population parameters. Point estimators give single values but lack information on their reliability. Confidence Intervals (CIs) extend this idea by providing an interval estimate with a specified probability of containing the true parameter.

In regression analysis (Montgomery et al., 2012; Weisberg, 2005), confidence intervals and confidence regions form a crucial part of statistical inference-helping quantify uncertainty associated with estimated regression coefficients, mean responses, and predictions.

Description:

- Confidence intervals and confidence regions form a central part of statistical inference, providing a range of plausible values for unknown population parameters rather than a single-point estimate. A confidence interval uses the sampling distribution of an estimator to specify an interval that contains the true parameter with a stated probability (commonly 95% or 99%). These intervals are developed using distributions such as the normal, t, chi-square, and F distributions depending on the parameter of interest and the assumptions involved.
- In many practical applications—especially in regression analysis and multivariate statistics-multiple parameters must be estimated simultaneously. In such cases, confidence regions extend the idea of interval estimation to higher dimensions. These regions, often taking the form of ellipses or ellipsoids, account for the correlation between parameter estimates and provide a more accurate joint assessment of uncertainty.
- This topic introduces the theoretical foundation, mathematical formulation, and practical interpretation of both confidence intervals and confidence regions. It emphasizes the role of sampling distributions, matrix algebra, and geometric representation, enabling students to apply these tools rigorously in statistical modeling and data analysis.

9.2 CONCEPT OF CONFIDENCE INTERVALS:

A confidence interval is an interval of plausible parameter values constructed in such a way that it will contain the true parameter with a pre-specified probability (confidence level).

A 95% CI means: If the procedure is repeated many times, then 95% of the constructed intervals will contain the true parameter.

It does not mean a 95% probability that the specific interval contains the parameter.
CIs are derived using sampling distributions of estimators.

9.2.1 basic definitions and notation:

- Population mean: μ
- Population variance: σ^2
- Population proportion: p
- Sample mean: $\bar{X}$
- Sample variance: s^2
- Sample proportion: $\hat{p}$
- Regression coefficients: $\beta_0, \beta_1, \dots, \beta_p$

- Estimated coefficients: $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$
- Design matrix: **X**
- Error variance: σ^2

9.2.2 confidence interval for mean:

Case 1: Population variance known

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

A 100(1–α)% CI is:

$$\mu \in [\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

Case 2: Population variance unknown
Use Student's t-distribution:

$$\mu \in [\bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}]$$

9.2.3 confidence interval for proportion:

For sample proportion $\hat{p} = \frac{x}{n}$:
Approximate CI:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

For small samples, exact binomial CIs are recommended (Clopper–Pearson)

9.3 CONFIDENCE INTERVAL FOR VARIANCE / STANDARD DEVIATION

Using chi-square distribution:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Thus CI is:

$$[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}]$$

CI for $\sigma$is obtained by square root.

9.4 CONFIDENCE INTERVALS IN REGRESSION MODELS:

Consider multiple regression:

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I)$$

Least squares estimator:

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$$

Variance:

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathrm{X}'\mathrm{X})^{-1}$$

9.4.1 ci for regression coefficients

For the jth coefficient:

$$\hat{\beta}_j \pm t_{\alpha/2,\, n-p}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}$$

9.4.2 ci for mean response:

At predictor vector x_0:

$$\hat{Y}_0 = x_0'\hat{\beta}$$

CI:

$$\hat{Y}_0 \pm t_{\alpha/2,\, n-p}\sqrt{\sigma^2 x_0'(\mathrm{X}'\mathrm{X})^{-1}x_0}$$

9.4.3 ci for prediction of new observation:

Prediction variance includes error variance:

$$\hat{Y}_0 \pm t_{\alpha/2,\, n-p}\sqrt{\sigma^2(1 + x_0'(\mathrm{X}'\mathrm{X})^{-1}x_0)}$$

9.5 CONCEPT OF CONFIDENCE REGIONS:

A confidence region generalizes confidence intervals to multiple parameters simultaneously.
For regression:

$$(\hat{\beta} - \beta)'[\sigma^2(\mathrm{X}'\mathrm{X})^{-1}]^{-1}(\hat{\beta} - \beta) \le pF_{p,\, n-p}(1-\alpha)$$

This defines an **ellipsoidal region** in p-dimensional space.

9.5.1 ellipsoidal confidence regions (multivariate case):

The joint distribution of $\hat{\beta}$is:

$$\hat{\beta} \sim N_p(\beta, \sigma^2(\mathrm{X}'\mathrm{X})^{-1})$$

Using Hotelling's T^2:
Confidence region:

$$(\hat{\beta} - \beta)'(\mathrm{X}'\mathrm{X})(\hat{\beta} - \beta) \le p\, s^2\, F_{p,n-p}(1-\alpha)$$

This geometrically corresponds to a p-dimensional ellipsoid centered at $\hat{\beta}$.

9.5.2 interpretation and geometric meaning

- Single-parameter CI → line segment
- Two parameters → ellipse
- p parameters → ellipsoid

The shape reflects:

- Variances (lengths of axes)
- Covariances (tilt of the ellipsoid)

Overlap of two confidence regions indicates similarity of parameter sets.

9.6 THEOREMS WITH EXAMPLES :

Notation & standing assumptions (used in many proofs)

- Scalars: n= sample size, p= number of regression parameters (including intercept).
- For IID sample $X_1, \dots, X_n$from $N(\mu, \sigma^2)$we use $\bar{X}$and s^2as usual.
- For regression: model $Y = X\beta + \varepsilon$with $Yn \times 1$, $Xn \times p$full column rank, $\varepsilon \sim N(0, \sigma^2 I_n)$. OLS estimator $\hat{\beta} = (X'X)^{-1}X'Y$. Residual sum of squares $RSS = (Y - X\hat{\beta})'(Y - X\hat{\beta})$. Estimate $\hat{\sigma}^2 = RSS/(n-p)$.
- z_γ= normal quantile, $t_{\gamma,\nu}$= Student-t quantile with νdf, $\chi^2_{\gamma,\nu}$= chi-square quantile with νdf, $F_{\nu_1,\nu_2}(\gamma)$= F-quantile.

Cochran's theorem / standard normal quadratic-form results will be invoked where appropriate (the independence of certain quadratic forms and chi-square results).

THEOREM 1 — Student's *t* Confidence Interval for a Mean (unknown variance)

Statement: Let $X_1, \dots, X_n$be iid $N(\mu, \sigma^2)$. Then a $100(1-\alpha)\%$confidence interval for $\mu$is

$$\boxed{\bar{X} \pm t_{\alpha/2,\, n-1}\, \frac{s}{\sqrt{n}}}$$

where $\bar{X} = \frac{1}{n}\sum X_i$ and $s^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$

Assumptions. IID normal observations; μ, σ^2 unknown

Proof : Distribution of the sample mean. For normal samples

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Therefore

$$Z \equiv \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad (1)$$

1. Distribution of the sample variance (chi-square). Define

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

By Cochran's theorem (or standard normal orthogonal decomposition),

$$U \equiv \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \qquad (2)$$

2. Independence. For normal samples $\bar{X}$and $s^2$are independent (again from Cochran's theorem or properties of the multivariate normal). Thus $Z$and $U$are independent.
3. Form the studentized pivot. Consider

$$T \equiv \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{Z}{\sqrt{U/(n-1)}}$$

Because $Z \sim N(0,1)$ and $U \sim \chi^2_{n-1}$ independent, the ratio has a Student-t distribution with $n-1$ degrees of freedom:

$$T \sim t_{n-1} \qquad (3)$$

1. Inversion to CI. By symmetry of the t-distribution

$$P(-t_{\alpha/2,n-1} \le T \le t_{\alpha/2,n-1}) = 1 - \alpha$$

Substituting the definition of $T$and solving for μgives

$$P\left(\bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Hence the stated interval is a $100(1-\alpha)\%$confidence interval for μ.

Remark. Exact under normality. For large n, $t_{\alpha/2,n-1} \to z_{\alpha/2}$.

THEOREM 2 — Chi-square Confidence Interval for Variance

Statement: Under the same normal model, a $100(1-\alpha)\%$CI for σ^2 is

$$\boxed{\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,\,n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,\,n-1}}\right)}$$

Assumptions. $X_i \sim iidN(\mu, \sigma^2)$.

Proof :

1. Start with the known pivot:

$$U = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}. \qquad \text{(A)}$$

Use chi-square quantiles. By definition of quantiles,

$$P\left(\chi^2_{\alpha/2,n-1} \leq U \leq \chi^2_{1-\alpha/2,n-1}\right) = 1-\alpha.$$

Substitute $U$and solve for σ^2:

$$P\left(\chi^2_{\alpha/2,n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{1-\alpha/2,n-1}\right) = 1-\alpha.$$

Invert each side (inequalities reverse when dividing by positive quantities appropriately) to get

$$P\ \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}) = 1-\alpha.$$

This is the desired interval.

Remark. CI for σ(std dev) is the square-root of the endpoints above.

THEOREM 3 — *t*-Confidence Interval for a Regression Coefficient

Statement: In the linear model $Y = X\beta + \varepsilon$with $\varepsilon \sim N(0, \sigma^2 I_n)$, the OLS estimator $\hat{\beta}$satisfies

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$$

and for the j-th component β_ja $100(1-\alpha)\%$CI is

$$\boxed{\hat{\beta}_j \pm t_{\alpha/2,\,n-p}\ \hat{\sigma}\sqrt{v_{jj}}}\ v_{jj} \equiv [(X'X)^{-1}]_{jj}$$

where $\hat{\sigma}^2 = RSS/(n-p)$

Assumptions: Linear model with Gaussian errors; Xfull column rank.

Proof :

1. Distribution of $\hat{\beta}$. Standard OLS theory (or properties of the multivariate normal) gives

$$\hat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon,$$

so

$$\hat{\beta} \sim N_p\ (\beta, \sigma^2(X'X)^{-1}). \qquad \text{(B)}$$

Marginal for $\hat{\beta}_j$. From (B), the marginal distribution of component $j$is normal:

$$\hat{\beta}_j \sim N\,(\beta_j, \sigma^2 v_{jj}).$$

Residual variance chi-square and independence. The residual vector $e = Y - X\hat{\beta} = MY$with $M = I - P$(projection onto orthogonal complement) has $e'\,e = RSS$and

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-p}.$$

By Cochran's theorem (or properties of normal projections), $\hat{\beta}$(hence $\hat{\beta}_j$) is independent of RSS(hence independent of $\hat{\sigma}^2$).

2. Form the studentized statistic. Define

$$T_j \equiv \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_{jj}}}.$$

Writing $\hat{\beta}_j - \beta_j = Z$where $Z \sim N(0, \sigma^2 v_{jj})$and using independence with $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$, we get

$$T_j = \frac{Z/(\sigma\sqrt{v_{jj}})}{\sqrt{\frac{(n-p)\hat{\sigma}^2}{\sigma^2}/(n-p)}} = \frac{N(0,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} \sim t_{n-p}$$

3. Invert to obtain CI. By t_{n-p}quantiles

$$P\,(-t_{\alpha/2,n-p} \le T_j \le t_{\alpha/2,n-p}) = 1 - \alpha$$

which after algebra yields

$$P(\hat{\beta}_j - t_{\alpha/2,n-p}\hat{\sigma}\sqrt{v_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\hat{\sigma}\sqrt{v_{jj}}) = 1 - \alpha$$

Thus the stated interval is the $100(1-\alpha)$%CI for β_j

Remark: This reduces to Theorem 1 in the simple regression intercept/slope special cases; degrees of freedom are $n - p$.

THEOREM 4 — Hotelling's T^2(Ellipsoidal) Confidence Region for β

Statement: Under the linear model with normal errors, a $100(1-\alpha)$%joint confidence region for the vector $\beta$is the ellipsoid

$$\boxed{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \le p\,\hat{\sigma}^2\,F_{p,\,n-p}(1-\alpha)}$$

Equivalently,

$$(\hat{\beta} - \beta)'[\hat{\sigma}^2(X'X)^{-1}]^{-1}(\hat{\beta} - \beta) \le p\,F_{p,n-p}(1-\alpha)$$

Assumptions. Linear model $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$, Xfull rank.

Proof :

1. Distribution of the quadratic form. From Theorem 3 (B),

$$\hat{\beta} - \beta \sim N_p(0, \sigma^2 (X'X)^{-1}).$$

Consider the quadratic form

$$Q \equiv \frac{1}{\sigma^2} (\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta)$$

Because $(\hat{\beta} - \beta) = \sigma (X'X)^{-1/2} Z$for $Z \sim N_p(0, I)$, it follows that $Q \sim \chi^2_p$ (This is the usual fact: if $W \sim N_p(0, I)$then $W'W \sim \chi^2_p$)

2. Replace $\sigma^2$by $\hat{\sigma}^2$and form an F pivot. The residual sum of squares gives

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$$

and this is independent of $\hat{\beta} - \beta$(Cochran). Hence the ratio

$$\frac{(Q/p)}{((n-p)\hat{\sigma}^2/\sigma^2)/(n-p)} = \frac{Q/p}{\hat{\sigma}^2/\sigma^2} \sim F_{p, n-p}$$

Multiplying both numerator and denominator by σ^2yields

$$\frac{(\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta)/(p\, \hat{\sigma}^2)}{1} \sim F_{p,n-p}$$

So

$$P\left(\frac{(\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta)}{p\, \hat{\sigma}^2} \leq F_{p,n-p}(1-\alpha)\right) = 1 - \alpha$$

Rearrange to the ellipsoid form. Multiply both sides by $p\, \hat{\sigma}^2$to obtain

$$P\left((\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta) \leq p\, \hat{\sigma}^2\, F_{p,n-p}(1-\alpha)\right) = 1 - \alpha$$

which is the stated ellipsoidal confidence region for β

Remark: (geometry). The set $\{\beta: (\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta) \leq c\}$is an ellipsoid centered at $\hat{\beta}$. Axis directions are eigenvectors of $(X'X)^{-1}$, axis lengths scale with $\hat{\sigma}$and the F-quantile. This region gives simultaneous coverage for all components or linear combinations of β.

Regression Coefficient CI

Simple regression:$Y = \beta_0 + \beta_1 x + \varepsilon$

Given:

- $\hat{\beta}_1 = 2.5$
- $se(\hat{\beta}_1) = 0.4$
- $n = 20 \Rightarrow df = 18$
- 95% level: $t_{0.025,18} = 2.101$

CI:$2.5 \pm 2.101(0.4) = [1.659, 3.341]$

Interpretation: The slope is significantly positive

9.7 KEY WORDS:

- Confidence Interval
- Confidence Level
- Margin of Error
- Sampling Distribution
- Regression Coefficient
- Mean Response
- Prediction Interval
- Confidence Region
- Ellipsoidal Region
- Variance–Covariance Matrix
- Multivariate Normal Distribution
- Linear Model

9.8 SUMMARY:

This lesson focused on the construction, interpretation, and application of confidence intervals and confidence regions in statistical inference, particularly within the framework of linear models and regression analysis**.** The concept of confidence intervals was introduced as a range of plausible values for an unknown population parameter, reflecting sampling variability and uncertainty.

Beginning with basic definitions and notation, confidence intervals for means, proportions, and variances were discussed using appropriate sampling distributions such as the normal, t, chi-square, and F distributions. These ideas were then extended to regression models, where confidence intervals were developed for regression coefficients, the mean response**,** and the prediction of a new observation**,** highlighting the distinction between estimation and prediction.
The concept of confidence regions was introduced to handle simultaneous inference for multiple parameters**.** Ellipsoidal confidence regions in the multivariate case were derived and interpreted geometrically, emphasizing their dependence on the variance–covariance structure of the estimators. Relevant theorems supporting the validity of these intervals and regions were presented with illustrative examples.

Overall, this unit provides a comprehensive framework for quantifying uncertainty and making statistically valid inferences in both univariate and multivariate settings within linear models.

9.9 SELF-ASSESSMENT QUESTIONS:

1. Define a confidence interval and explain its interpretation. Derive the CI for a population mean with known variance.
2. Obtain confidence intervals for the least squares estimates in the case of a two-variable linear model.
3. Explain why Student's t-distribution is used when variance is unknown. Construct a 95% CI for a proportion with example data.
4. Derive the chi-square CI for variance. State the matrix form of the variance of $\hat{\beta}$.
5. Derive CI for a regression coefficient using least squares theory.
6. Distinguish between CI for mean response and prediction interval.
7. Define a confidence region and explain why it is ellipsoidal.
8. Explain the role of the F-distribution in constructing multivariate confidence regions.

9.10 SUGGESTED READING:

1. raper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, Wiley.
2. Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*, Wiley.
3. Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Wiley.
4. Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*, Wiley.
5. Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*, McGraw-Hill.

Dr. U. Ramkiran

LESSON-10

ANALYSIS OF VARIANCE

OBJECTIVES:

After studying this lesson, the student will be able to:

- Understand the fundamental concept of ANOVA - Explain the rationale of analysis of variance and its role in comparing multiple population means.
- Relate ANOVA to linear statistical models - Interpret ANOVA as a special case of the general linear model.
- Identify sources of variation - Decompose total variation into between-group and within-group components.
- Apply one-way and two-way ANOVA models - Formulate and analyze fixed-effects ANOVA models.
- Derive and interpret ANOVA test statistics - Compute sums of squares, mean squares, and F-statistics.
- Test hypotheses using ANOVA - Perform hypothesis testing for equality of means under various experimental designs.
- Interpret ANOVA results in applied contexts - Draw meaningful conclusions from ANOVA tables in real-life data analysis.
- Develop analytical skills for experimental data - Apply ANOVA techniques to problems in agriculture, industry, social sciences, and biomedical research.

STRUCTURE:

10.1 INTRODUCTION:

Analysis of Variance (ANOVA) is the framework for testing whether several population means are equal. It partitions total observed variation into components attributable to factors (treatments, groups) and random error, and uses the F-ratio (treatment mean square over error mean square) to test hypotheses. ANOVA is both a modeling and an inferential device closely related to the general linear model (regression) formulation of mean functions. Classic treatments and matrix derivations appear in Montgomery et al. and Weisberg.

Description

This lesson introduces Analysis of Variance (ANOVA) as a fundamental statistical tool used to analyze experiments and regression models by decomposing total variation into meaningful components. Drawing from the structure presented in *Montgomery, Peck & Vining* and the matrix-based treatment of *Weisberg*, the chapter develops both the classical ANOVA framework and its natural extension within the general linear model.

The chapter begins with the model

$$Y = X\beta + \varepsilon$$

which serves as the mathematical foundation for all ANOVA procedures. Using this model, the total variation in the response is partitioned into components explained by the fitted model (regression or treatment effects) and unexplained random variation (error). The derivation of sums of squares, degrees of freedom, and mean squares follows the same approach as in Montgomery's regression ANOVA chapters.

The lesson then addresses assumption diagnostics, including graphical and analytical methods for assessing normality, homogeneity of variance, independence, leverage, and influence. These align with the standard residual analysis presented in both PDFs and your DOCX lesson file.

Finally, practical applications and interpretation are emphasized. Examples include completely randomized designs, randomized block designs, and factorial experiments, consistent with the style of examples in your provided materials.

10.2 GENERAL LINEAR MODEL FRAMEWORK:

We view ANOVA as a special case of the general linear model

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

Each ANOVA design has a specific X. For one-way ANOVA with g groups, a common parameterization is

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \sum_{i=1}^{g} \tau_i = 0$$

or equivalently the cell means parameterization $y_{ij} = \mu_i + \varepsilon_{ij}$. The linear model view allows use of projection matrices, sums of squares as quadratic forms, derivation of distributions (chi-square, F) via Cochran's theorem, and extensions to unbalanced designs and mixed models.

10.3 PARTITIONING OF TOTAL VARIATION:

Fundamental identity (one-way case):

$$SS_T = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\right)^2 = SS_B + SS_E$$

where$SS_B = \sum_i n_i\,(\bar{y}_i - \bar{y})^2$and $SS_E = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\right)^2$

Degrees of freedom partition:

$$n - 1 = (g - 1) + (n - g)$$

Mean squares: $MS_B = SS_B/(g-1)$, $MS_E = SS_E/(n-g)$

Interpretation: SS_Bmeasures variation due to differences among group means; SS_Emeasures within-group (random) variation.

10.3.1 ONE-WAY ANALYSIS OF VARIANCE:

Model and Hypotheses

Model:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \varepsilon_{ij} \sim iidN(0, \sigma^2), \sum_i \tau_i = 0$$

Test:

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_g = 0\text{(all group means equal)vs}H_A\text{:not all } \tau_i = 0$$

Derivation of the F-test

1. Compute $SS_B$and SS_Eas above. Under H_0, the between-group variation arises solely from sampling error.
2. Under the model with normal errors and the constraint $\sum\tau_i = 0$, Cochran's theorem gives:
   - $\frac{SS_E}{\sigma^2} \sim \chi^2_{n-g}$
   - $\frac{SS_B}{\sigma^2} \sim \chi^2_{g-1} if\ H_0 is\ true$ (or more generally has noncentral chi-square under alternatives), and
   - SS_B and $SS_E$are independent (This follows because $SS_B$and $SS_E$are quadratic forms in normal variables corresponding to orthogonal projections onto complementary subspaces.)
3. Form the ratio

   $$F = \frac{MS_B}{MS_E} = \frac{SS_B/(g-1)}{SS_E/(n-g)}$$

   Under H_0this has an exact $F_{g-1,\,n-g}$distribution
4. Decision rule: reject H_0if $F_{\text{cal}} > F_{g-1,n-g;\,\alpha}$. Equivalently compute p-value $P(F_{g-1,n-g} \geq F_{\text{cal}})$.

Remarks: For balanced designs SS_Bhas the simple algebraic form above. For unbalanced designs, $SS_B$and $SS_E$are defined similarly but the correctness of the test uses the appropriate model matrix $X$and projections (Type I/II/III sums of squares discussions). See Montgomery/Weisberg for details on unbalanced cases and choice of sum-of-squares type.

TWO-WAY ANOVA (FIXED EFFECTS) — WITH / WITHOUT INTERACTION:

Consider two factors A(levels $i = 1, \dots, a$) and B(levels $j = 1, \dots, b$). Observations y_{ijk}at cell (i, j), $k = 1, \dots, n_{ij}$. Fixed effects model (no replication or with replication):

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0.$$

Objectives: test main effects α_i, β_j, and interaction $(\alpha\beta)_{ij}$.
ANOVA table (balanced replication, $n_{ij} = n_0$):

Source	**df**	**SS**	**MS**	F
A	$a - 1$	SS_A	$MS_A = SS_A/(a-1)$	MS_A/MS_E
B	$b - 1$	SS_B	MS_B	MS_B/MS_E
AxB	$(a-1)(b-1)$	SS_{AB}	MS_{AB}	MS_{AB}/MS_E
Error	$ab(n_0 - 1)$	SS_E	MS_E	
Total	$abn_0 - 1$	SS_T		

Derivations use projection matrices and Cochran's theorem to establish null distributions of mean squares (see Montgomery for derivations). For unbalanced designs use the general linear model and appropriate contrasts.

10.4 ANOVA THROUGH REGRESSION (INDICATOR VARIABLES)

Equivalently, represent factor levels by dummy (indicator) variables in $X$and fit $Y = X\beta + \varepsilon$. For one-way ANOVA with $g$groups, one can encode $g - 1$dummies (with intercept) and test the joint significance of the dummy coefficients using an F-test equivalent to the ANOVA F. This perspective:

- unifies ANOVA with regression
- handles covariates (ANCOVA)
- allows complex contrasts and hypothesis testing for linear functions of β. See Weisberg §6 and Montgomery §3 for worked examples and the mapping between sums of squares and regression projections.

10.5 MATRIX APPROACH TO ANOVA (CONCISE DERIVATION)

Let $Y \in \mathbb{R}^n$, model $Y = X\beta + \varepsilon$. Define projection matrices:

- $P_X = X(X'X)^{-1}X'$(fitted space)
- $M_X = I - P_X$(residual projector)

Sums of squares are quadratic forms:

- $SS_{Reg} = \| P_X Y - \bar{Y}\mathbf{1} \|^2$(model/treatment SS)
- $SS_{Res} = Y' M_X Y$(residual SS)

By Cochran's theorem (and properties of idempotent matrices) if $\varepsilon \sim N(0, \sigma^2 I)$:

- $\frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-p}$
- $\frac{SS_{Reg}}{\sigma^2} \sim \chi^2_{p-1}$under null of no effect (or noncentral otherwise)
- independence holds between orthogonal quadratic forms. Hence the F-test arises as ratio of scaled chi-squares:

$$F = \frac{(SS_{Reg}/(p-1))}{(SS_{Res}/(n-p))} \sim F_{p-1,n-p}$$

This derivation is the most general and covers unbalanced/complex designs. See Appendix C (Montgomery) for details on SSR/SSRes relationships and proofs.

10.5.1 APPLICATIONS OF ANOVA:

- Experimental comparisons (agriculture, industry) — compare treatment means.
- Factorial experiments (study main effects and interactions).
- Blocked designs (randomized block ANOVA) to remove nuisance variation.
- ANCOVA — ANOVA with covariates (combine regression and ANOVA).
- Random effects models and variance component estimation (mixed models). See examples and practice problems in both Montgomery and Weisberg.

10.6 ASSUMPTIONS & DIAGNOSTICS:

Classical ANOVA assumptions (same as linear model):

5. Linearity (mean structure correct)
6. Errors $\varepsilon_{ij}$are independent
7. Homoscedasticity: Var $(\varepsilon_{ij}) = \sigma^2$(constant variance across cells)
8. Normality: $\varepsilon_{ij} \sim N(0, \sigma^2)$(for exact small-sample inference)

Diagnostics: residual plots vs fitted values, normal probability plots of residuals, Levene/Bartlett tests for homogeneity, interaction plots for factorials, influence diagnostics for outliers. See Weisberg Chap. 8 and Montgomery Chap. 4 for detailed diagnostic procedures and graphical examples.

10.6.1 remedies for assumption violations

- Nonconstant variance: transform the response (Box–Cox), use weighted least squares; consider variance-stabilizing transforms (square root, log).
- Nonnormality / outliers: robust methods, trimmed means, or bootstrap inference.
- Unbalanced designs / missing cells: use general linear model, Type II/III sums of squares, or mixed models (REML) for random effects. Montgomery recommends REML for unbalanced random/mixed models.

10.7 THEOREMS & PROOFS (KEY RESULTS YOU MUST INCLUDE):

THEOREM I. Partitioning identity (one–way ANOVA)
Claim.

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 \;=\; \sum_i n_i(\bar{y}_i - \bar{y})^2 \;+\; \sum_i \sum_j (y_{ij} - \bar{y}_i)^2\,(y_{ij} - \bar{y}_i)^2.$$

Interpretation. Total corrected sum of squares = between–groups SS + within–groups (error) SS.

Proof (algebraic).

1. Start with the left–hand side, expand each term by inserting $\bar{y}_i$:

$$y_{ij} - \bar{y} \;=\; (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})$$

2. Square and sum over all i, j:

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i \sum_j [(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})]^2$$

3. Expand the square:

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 \qquad = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$$
$$+2 \sum_i \sum_j (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y})$$
$$+ \sum_i \sum_j (\bar{y}_i - \bar{y})^2 .$$

4. Simplify each term:

- The first term is $\sum_i \sum_j (y_{ij} - \bar{y}_i)^2$(that will become SS_E).
- The third term: $\sum_i \sum_j (\bar{y}_i - \bar{y})^2 = \sum_i n_i(\bar{y}_i - \bar{y})^2$because $\bar{y}_i - \bar{y}$does not depend on $j$and there are $n_i$observations in group i.
- The middle term: use that for each i,

$$\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = 0$$

(by definition of group mean), therefore

$$\sum_i \sum_j (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) = \sum_i (\bar{y}_i - \bar{y}) \sum_j (y_{ij} - \bar{y}_i) = 0.$$

5. Putting pieces together,

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i n_i (\bar{y}_i - \bar{y})^2,$$

which is the desired identity.

THEOREM II Distribution of sums–of–squares under normality (via Cochran's theorem)
One–way ANOVA model

$$y_{ij} = \mu_i + \varepsilon_{ij}, \varepsilon_{ij} \sim iidN(0, \sigma^2)$$

Let $SS_E = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$ and $SS_B = \sum_i n_i (\bar{y}_i - \bar{y})^2$

Under H_0: $\mu_1 = \cdots = \mu_g = \mu$, we have the standard null model with common mean.
Claims

- Under H_0, $\frac{SS_E}{\sigma^2} \sim \chi^2_{n-g}$
- Under H_0, $\frac{SS_B}{\sigma^2} \sim \chi^2_{g-1}$
- Moreover $SS_E$and $SS_B$are independent.

Proof :

Vector form and projections. Stack observations into $Y \in \mathbb{R}^n$. Under $H_0$the mean vector is $\mu 1$and the errors vector $\varepsilon \sim N(0, \sigma^2 I_n)$. Consider subspaces:

- $\mathcal{S}_1 = \text{span } (\mathbf{1})$(overall mean subspace, dimension 1)
- $\mathcal{S}_2 = \{$vectors constant on each group, with sum zero across groups$\}$ the between–group subspace orthogonal to $\mathbf{1}$ describing deviations of group means from the grand mean; $\dim(\mathcal{S}_2) = g - 1$
- $\mathcal{S}_3$ =the within–group residual subspace (vectors with each group sum zero); $\dim(\mathcal{S}_3) = n - g$

These three subspaces are pairwise orthogonal and their direct sum is $\mathbb{R}^n$

1. Quadratic forms as projections. Let P_3denote the orthogonal projector onto $\mathcal{S}_3$. Then

$$SS_E = \| P_3 Y \|^2 = Y' P_3 Y$$

Similarly, P_2(projector onto $\mathcal{S}_2$) gives

$$SS_B = \| P_2 Y \|^2 = Y' P_2 Y$$

And the total corrected SS is $Y'(P_1 + P_2 + P_3)Y$with P_1projector onto $\mathcal{S}_1$

2. Cochran's theorem application. Cochran's theorem states: if $Y \sim N(0, \sigma^2 I_n)$and $A_1, \ldots, A_k$are symmetric idempotent matrices with ranks $r_1, \ldots, r_k$such that

$\sum_{i=1}^{k} A_i = I_n$, then the quadratic forms $Y'A_iY/\sigma^2$ are independent and $Y'A_iY/\sigma^2 \sim \chi^2_{r_i}$. In our ANOVA decomposition we have projectors P_1, P_2, P_3 summing to I_n on the centered data space, with ranks $1, (g-1), (n-g)$ respectively. Hence Cochran's theorem applies.

3. Conclude distributions and independence. Under H_0 (so the mean structure is in $\mathcal{S}_1$), with $Y - \mu 1$ playing the role of random normal vector centered at 0, we obtain:
   - $SS_E/\sigma^2 = Y'P_3Y/\sigma^2 \sim \chi^2_{n-g}$,
   - $SS_B/\sigma^2 = Y'P_2Y/\sigma^2 \sim \chi^2_{g-1}$,
   - and SS_E and SS_B are independent because $P_2P_3 = 0$ (orthogonal projectors onto orthogonal subspaces).

Remarks.

- If an alternative (nonnull) model holds, the distribution of SS_B/σ^2 becomes a noncentral chi–square with noncentrality determined by the true group means; SS_E/σ^2 remains central chi–square if homoscedastic normal errors hold.
- The geometric / projector viewpoint and Cochran's theorem are the standard rigorous route — see Montgomery for this exposition.

THEOREM III F–test (ratio of scaled chi–squares)

Claim. Under H_0, the statistic

$$F = \frac{MS_B}{MS_E} = \frac{SS_B/(g-1)}{SS_E/(n-g)}$$

has an F-distribution with $(g-1, n-g)$ degrees of freedom:

$$F \sim F_{g-1,\, n-g}$$

Proof (straightforward from II)

1. By II, under H_0,
$$\frac{SS_B}{\sigma^2} \sim \chi^2_{g-1}, \frac{SS_E}{\sigma^2} \sim \chi^2_{n-g}$$
and the two are independent.
2. The definition of an F-distributed variable: if $U \sim \chi^2_{r_1}$ and $V \sim \chi^2_{r_2}$ are independent, then
$$\frac{(U/r_1)}{(V/r_2)} \sim F_{r_1, r_2}$$
3. Apply the definition with $U = SS_B/\sigma^2$, $V = SS_E/\sigma^2$, $r_1 = g-1$, $r_2 = n-g$. Then
$$\frac{(SS_B/\sigma^2)/(g-1)}{(SS_E/\sigma^2)/(n-g)} \sim F_{g-1,n-g}$$
4. Cancel σ^2 in numerator and denominator to get
$$\frac{SS_B/(g-1)}{SS_E/(n-g)} \sim F_{g-1,n-g}$$

which is the desired result.

Decision rule. Reject H_0 at level α if $F_{\text{obs}} > F_{g-1,n-g;1-\alpha}$

THEOREM IV Equivalence of ANOVA F and regression F (SSR = SS_B and orthogonality)

Claim.Fitting the one–way ANOVA model by ordinary least squares using dummy (indicator) variables yields the same model sum of squares as the classical ANOVA between–groups SS. Thus the ANOVA F-test is algebraically identical to the regression F-test for testing the joint significance of the dummy coefficients.

Proof

1. Regression encoding. Let there be $g$groups. One standard regression parameterization is:
$$y_{ij} = \beta_0 + \beta_2 d_{i2} + \beta_3 d_{i3} + \cdots + \beta_g d_{ig} + \varepsilon_{ij}$$
where$d_{ik} = 1$ if observation is in group k, else 0; group 1 serves as baseline (so there are $g - 1$dummies plus intercept). The design matrix X has columns: a column of ones (intercept) and $g - 1$dummy columns. Full rank $p = g$.

2. Model fitted values and group means. For this design, the fitted value for any observation in group $i$equals the estimated group mean $\hat{\mu}_i$. (Because the OLS solution sets the fitted value constant within each group equal to the group sample mean when using cell means parameterization or equivalent dummy coding.) Concretely, the fitted vector $\hat{Y} = P_X Y$is piecewise constant on the groups, with value $\bar{y}_i$for observations in group i.
3. Regression SSR equals ANOVA between–groups SS. The regression sum of squares (SSR or SSReg) is
$$SS_{Reg} = \sum_{k=1}^{n} (\hat{y}_k - \bar{y})^2 = \sum_i n_i (\bar{y}_i - \bar{y})^2$$
because$\hat{y}_k = \bar{y}_i$for observations in group i. But the right side is exactly $SS_B$from the ANOVA algebraic identity. Thus $SS_{Reg} = SS_B$
4. Residual (error) SS equality**.** The residual sum of squares from regression is
$$SS_{Res} = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 = \sum_j \sum_i (y_{ij} - \bar{y}_i)^2 = SS_E$$Type equation here.

This also follows from the partition $SS_T = SS_{Reg} + SS_{Res}$ and the partition identity in I.

5. Orthogonality / independence viewpoint. In matrix language, the model subspace spanned by the columns of X equals the span of group–indicator vectors (which is the sum of the grand–mean subspace and the between–group subspace). The residual projector $M_X = I - P_X$ projects onto the within–group subspace; hence regression residuals are orthogonal to the columns of X, which yields orthogonality of SSR and SSE quadratic forms. This is the same geometry underlying Cochran's theorem used in II.
6. Equivalence of Ftests. In regression one tests the null that the $g - 1$ dummy coefficients are all zero (i.e. only intercept remains). The standard regression $F$for the model with $p = g$parameters is
$$F = \frac{SS_{Reg}/(p-1)}{SS_{Res}/(n-p)} = \frac{SS_B/(g-1)}{SS_E/(n-g)}$$

which is exactly the ANOVA F derived earlier. Thus the tests are algebraically identical.

Conclusion:
ANOVA can be viewed as a special case of the general linear model; sums of squares and tests coincide when factors are encoded by dummy variables. For elaboration and examples see Weisberg and Montgomery.

10.8 KEY WORDS:

- Analysis of Variance (ANOVA)
- Total Sum of Squares (TSS)
- Treatment Sum of Squares (TrSS)
- Error Sum of Squares (ESS)
- Mean Square
- F-statistic
- One-Way ANOVA
- Two-Way ANOVA
- Fixed Effects Model
- Random Effects Model
- General Linear Model
- ANOVA Table
- Degrees of Freedom
- Null Hypothesis
- Homogeneity of Variance
- Additivity
- Interaction Effects
- Matrix Approach to ANOVA
- ANOVA through Regression

10.9 SUMMARY:

Analysis of Variance (ANOVA) is a fundamental statistical technique used to compare means of two or more populations by decomposing total variability in the data into meaningful components attributable to different sources. Within the general linear model framework**,** ANOVA is shown to be a special case of regression analysis, thereby unifying regression and experimental data analysis.

The unit begins with the formulation of ANOVA under the general linear model**,** emphasizing the role of design matrices and parameter interpretation. The partitioning of total variation into treatment (explained) and error (unexplained) components forms the basis for hypothesis testing using the F-statistic**.** One-way and two-way ANOVA models illustrate how factor effects and interactions influence response variability.

ANOVA is further developed through a regression and matrix approach**,** which provides a compact and powerful representation for estimation, testing, and interpretation. This approach highlights connections between sums of squares, projections, and estimability of

effects. Practical applications of ANOVA demonstrate its usefulness in agriculture, industry, economics, medicine, and social sciences.

The unit also stresses the importance of model assumptions-normality, independence, and homoscedasticity-and introduces diagnostic tools for detecting violations. Appropriate remedial measures**,** such as transformations and alternative modeling strategies, are discussed to ensure valid inference.

In conclusion, ANOVA serves as a cornerstone of linear models by:

- Providing a systematic method for comparing multiple means,
- Linking experimental design with regression analysis,
- Offering a matrix-based framework for estimation and testing,
- Supporting sound statistical inference through diagnostics and assumptions.

A solid understanding of ANOVA equips students with essential tools for analyzing structured data and lays the foundation for advanced topics in linear and mixed models.

10.10 SELF-ASSESSMENT QUESTIONS:

1. Derive the one-way ANOVA partition $SST = SSB + SSE$ from first principles.
2. Show step-by-step that under the normal error model $SS_E/\sigma^2 \sim \chi^2_{n-g}$. (Use Cochran's theorem.)
3. Explain analysis of variance for two–way classification with multiple observations per cell. Obtain the ANOVA table.
4. Show equivalence between the ANOVA F-test and the regression Ftest for group indicators.
5. Given an unbalanced one-way design, explain differences among Type I/II/III sums of squares and when each is appropriate.
6. Given residual diagnostics showing increasing variance with fitted values, propose remedial steps and justify them.

10.11 SUGGESTED READING:

1. Graybill, F.A. (1983): Matrices with Applications in Statistics. Wadsworth.
2. Draper, N.R. and Smith, H. (1998): Applied Regression Analysis. Wiley-Blackwell.
3. Montgomery, D.C., Peck, E.A. and Vining, G.G. (2012): Introduction to Linear Regression Analysis, 5th Ed. Wiley.
4. Bapat, R.B. (2012): Linear Algebra and Linear Models. Springer.
5. Rao, C.R. (2002): Linear Statistical Inference and Its Applications. 2nd Ed. Wiley-Blackwell.
6. Weisberg, S. (2013): Applied Linear Regression, 4th Ed. Wiley.

Dr. U. Ramkiran

LESSON -11

SIMPLE LINEAR REGRESSION

OBJECTIVES:

After completing this lesson, students will be able to:

- Understand the structure and assumptions of the simple linear regression model
- Derive and compute least squares estimators of regression parameters
- Interpret regression coefficients in practical contexts
- Perform hypothesis testing and construct confidence intervals for parameters
- Evaluate model adequacy using coefficient of determination (R^2)

STRUCTURE:

11.1 INTRODUCTION

In many real-world situations, understanding how one measurable quantity changes in response to another is essential for scientific analysis and decision-making. Such situations frequently arise in economics, agriculture, engineering, medicine, environmental studies, and social sciences. When the relationship between two quantitative variables is of interest, statistical modeling provides a systematic approach to describe, analyze, and interpret this relationship. Among the various statistical tools available, **simple linear regression** is one of the most fundamental and widely applied techniques.

Simple linear regression focuses on studying the relationship between **two variables**, where one variable depends on the other. The variable whose value is to be explained or predicted is called the **response** or **dependent variable**, while the variable used to explain or predict changes is known as the **explanatory** or **independent variable**. The primary objective of regression analysis is not only to identify whether a relationship exists but also to quantify the nature and strength of that relationship.

The fundamental idea of simple linear regression is to represent the dependence of the response variable on the explanatory variable through a **straight-line relationship**. This linear form is chosen for its simplicity, interpretability, and usefulness in practical applications. Although many real-world relationships may be complex, linear regression often serves as an effective first approximation that captures the overall trend in the data. Once such a model is established, it can be used to predict future values of the response variable for given values of the explanatory variable.

An important feature of regression analysis is the recognition of **random variation**. In practice, observed data rarely follow a perfect deterministic relationship. Various unobserved factors, measurement errors, and natural variability introduce randomness into the observed values. Simple linear regression accounts for this uncertainty by incorporating a random error term into the model. This allows the analyst to separate the systematic component of the relationship from random fluctuations and to make probabilistic statements about model parameters and predictions.

Regression analysis differs from correlation analysis in its objective. While correlation measures the degree of association between two variables, regression aims to establish a functional relationship that enables explanation and prediction. In simple linear regression, the direction of dependence is clearly defined: changes in the explanatory variable are assumed to influence the response variable, not vice versa. This distinction is crucial in applications such as forecasting, policy analysis, and experimental studies.

The simplicity of the linear regression model also allows for meaningful interpretation of its parameters. The slope of the regression line indicates the average rate of change of the response variable with respect to the explanatory variable, while the intercept provides a baseline level of the response variable under specific conditions. These interpretations make the model particularly useful for conveying results to practitioners and decision-makers who may not have a strong background in statistics.

Another important role of simple linear regression is its function as a **foundation for more advanced models**. Concepts such as least squares estimation, hypothesis testing, confidence interval construction, and model diagnostics are first introduced in the context of the simple linear regression model. These ideas are later extended to multiple regression, generalized linear models, and other advanced statistical techniques. Therefore, a clear understanding of simple linear regression is essential for further study in statistical modeling and data analysis.
In addition, simple linear regression plays a vital role in empirical research. Researchers use it to test theoretical relationships, validate assumptions, and generate insights from data. By quantifying relationships and assessing their statistical significance, regression analysis supports evidence-based conclusions across a wide range of disciplines.

In summary, simple linear regression provides a powerful yet accessible framework for studying relationships between variables. Its practical relevance, conceptual clarity, and

methodological importance make it an indispensable tool in statistics and applied research. A thorough understanding of this technique enables students and practitioners to model real-world phenomena effectively and to progress toward more sophisticated analytical methods.

11.2 SIMPLE LINEAR REGRESSION MODEL

In regression analysis, the objective is to study the relationship between a **response (dependent) variable**, whose value is unknown, and one or more **explanatory (independent) variables**, whose values are known or observed. When the model involves **a single response variable** Y and **only one explanatory variable** X, the resulting linear relationship

$Y = \beta_0 + \beta_1 X + \varepsilon$

where:

β0 = intercept of the regression line

β1 = slope of the regression line

ε = random error term

The model assumes that the mean value of Y changes linearly with X, while random disturbances are captured by the error term.

Examples

1. Predicting sales of the product based on advertisement;

$Y(Sales) = \beta_0 + \beta_1 X(Advertisements) + \varepsilon$

2. Estimating exam scores based on study hours of PG Students;

$Y(Exam\ Score) = \beta_0 + \beta_1 X(Study\ hour) + \varepsilon$ **11.2.1 Assumptions of Simple Linear Regression**

For the simple linear regression model to produce reliable estimates and valid statistical inferences, certain underlying assumptions must be satisfied. These assumptions describe the behavior of the relationship between the variables and the random error component of the model.

Linearity

The expected value of the response variable is assumed to be a linear function of the explanatory variable. This means that changes in the independent variable lead to proportional changes in the mean of the dependent variable. The relationship between the two variables can therefore be adequately represented by a straight line.

Independence of Errors

The random error terms associated with different observations are assumed to be independent of one another. This implies that the error corresponding to one observation does not influence or provide information about the error of another observation. Independence is particularly important when data are collected over time or across units.

Zero Mean of Error Terms

The error term is assumed to have an expected value of zero for all values of the independent variable. This condition ensures that the regression line represents the average relationship between the variables and that the model does not systematically overestimate or underestimate the response variable.

Constant Variance (Homoscedasticity)
The variability of the error term is assumed to remain constant for all levels of the independent variable. In other words, the spread of the residuals around the regression line should be approximately the same across the entire range of the explanatory variable. This assumption ensures efficiency and reliability of the parameter estimates.

Normality of Error Terms
For purposes of hypothesis testing and interval estimation, the error terms are assumed to follow a normal distribution. While this assumption is not strictly necessary for parameter estimation, it is essential for constructing confidence intervals and performing significance tests using standard statistical methods.

1. Modelling temperature and cool drinks sales in the city;

$Y(Cool\ drink\ Sales) = \beta_0 + \beta_1 X(Temperature) + \varepsilon$

To complete the model in (1), we make the following assumptions:

1. $E(\varepsilon) = 0$ or equivalently, $E(Y) = \beta_0 + \beta_1 X$
2. $V(\varepsilon) = \sigma^2$ or equivalently, $V(Y) = \sigma^2$
3. $Cov(\varepsilon_i \varepsilon_j) = 0 \ \ \forall i \neq j$ or equivalently, $Cov(Y_i Y_j) = 0$

11.2.1 Assumptions of Simple Linear Regression

For the simple linear regression model to produce reliable estimates and valid statistical inferences, certain underlying assumptions must be satisfied. These assumptions describe the behavior of the relationship between the variables and the random error component of the model.

Linearity
The expected value of the response variable is assumed to be a linear function of the explanatory variable. This means that changes in the independent variable lead to proportional changes in the mean of the dependent variable. The relationship between the two variables can therefore be adequately represented by a straight line.

Independence of Errors
The random error terms associated with different observations are assumed to be independent of one another. This implies that the error corresponding to one observation does not influence or provide information about the error of another observation. Independence is particularly important when data are collected over time or across units.

Zero Mean of Error Terms
The error term is assumed to have an expected value of zero for all values of the independent variable. This condition ensures that the regression line represents the average relationship between the variables and that the model does not systematically overestimate or underestimate the response variable.

Constant Variance (Homoscedasticity)
The variability of the error term is assumed to remain constant for all levels of the independent variable. In other words, the spread of the residuals around the regression line

should be approximately the same across the entire range of the explanatory variable. This assumption ensures efficiency and reliability of the parameter estimates.

Normality of Error Terms
For purposes of hypothesis testing and interval estimation, the error terms are assumed to follow a normal distribution. While this assumption is not strictly necessary for parameter estimation, it is essential for constructing confidence intervals and performing significance tests using standard statistical methods.

11.2.2 Interpretation of Regression Parameters

- **Intercept (β_0)**: Represents the expected value of Y when X=0.
- **Slope (β_1)**: Measures the average change in Y for a one-unit increase in X.

A positive slope indicates a direct relationship, while a negative slope indicates an inverse relationship.

11.3 LEAST SQUARES ESTIMATION

The **method of least squares** is used to estimate the unknown regression parameters. The principle is to choose estimators that minimize the sum of squared deviations between observed values and fitted values.

11.3.1 Estimation of Regression Coefficients

The estimators of β0 and β1 are given by:

$$\hat{\beta}_1 = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}$$

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

The fitted regression line is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

11.3.2 Properties of Least Squares Estimators

Under the standard assumptions, the least squares estimators have the following properties:

- **Unbiasedness**: $E(\hat{\beta}_0) = \beta 0$ and $E(\hat{\beta}_1) = \beta_1$
- **Minimum Variance**: They have the smallest variance among all linear unbiased estimators
- **Consistency**: Estimators converge to true parameter values as sample size increases
- **Efficiency**: They achieve the Gauss–Markov optimality condition

11.4 STATISTICAL INFERENCE

Statistical inference in regression analysis focuses on drawing conclusions about the unknown parameters of a regression model based on sample data. Once the regression coefficients are estimated, inferential procedures are used to determine whether the estimated relationships are statistically meaningful and to assess the precision of these estimates.

One important aspect of statistical inference is **hypothesis testing**. Hypotheses are formulated to test assumptions about regression parameters, particularly to examine whether an explanatory variable has a significant effect on the response variable. A common null hypothesis states that a regression coefficient is equal to zero, implying no linear relationship between the variables. Test statistics are computed using the estimated coefficients and their standard errors, and decisions are made by comparing these values with appropriate critical values.

Another key component of regression inference is the construction of **confidence intervals** for the model parameters. Confidence intervals provide a range of plausible values for the true regression coefficients and indicate the level of uncertainty associated with the estimates. A wider interval reflects greater uncertainty, while a narrower interval suggests more precise estimation.

Statistical inference also allows for assessing the overall adequacy of the regression model. By combining hypothesis tests and confidence intervals, researchers can evaluate the reliability of parameter estimates and the strength of the relationship between variables. These inferential tools support informed conclusions and enable effective prediction and decision-making based on the regression model.

11.4.1 Tests of Hypotheses on Regression Parameters

A commonly tested hypothesis is:
$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$
The test statistic is:

$$t = \frac{\hat{\beta}1}{SE\,(\hat{\beta}1)}$$

If the calculated t-value exceeds the critical value, the null hypothesis is rejected, indicating a significant linear relationship.

11.4.2 Confidence Intervals for Regression Coefficients

A $(1-\alpha)\times 100\%$ confidence interval for β_1 is:

$$\hat{\beta}1 \pm t_{\alpha/2,\, n-2} \times SE\,(\hat{\beta}1)$$

Similarly, confidence intervals can be constructed for β_0.

11.5 GOODNESS OF FIT

Goodness of fit refers to the extent to which a regression model is able to explain the observed variation in the response variable using the explanatory variable. It provides a quantitative measure of how accurately the fitted regression equation represents the underlying data. A model with a good fit closely follows the observed data points, while a poor fit indicates that the model does not adequately capture the relationship between the variables.

In regression analysis, the total variation in the response variable can be separated into two components: the variation explained by the regression model and the variation due to random error. Goodness-of-fit measures evaluate the proportion of total variation that is explained by

the fitted model. A higher proportion of explained variation indicates that the model successfully captures the systematic relationship between the variables.

Assessing goodness of fit is essential for understanding the usefulness and reliability of a regression model. Even if the estimated regression coefficients are statistically significant, the model may still perform poorly if it explains only a small portion of the variability in the data. Therefore, goodness-of-fit measures complement statistical inference by providing insight into the practical effectiveness of the model.

Graphical methods such as residual plots are often used alongside numerical measures of goodness of fit. These plots help detect patterns, outliers, or deviations from model assumptions that may reduce the quality of the fit. A well-fitted model typically shows residuals that are randomly scattered without any systematic structure.

Overall, goodness of fit plays a crucial role in evaluating regression models. It helps determine whether the model is suitable for interpretation, prediction, and decision-making. By examining goodness-of-fit measures, analysts can compare competing models and select the one that best represents the relationship between the variables while maintaining simplicity and accuracy.

11.5.1 Coefficient of Determination (R^2)

The coefficient of determination is defined as:

$$\mathbf{R^2} = \frac{\text{Explained Sum of Squares}}{\text{Total Sum of Squares}}$$

Its value lies between 0 and 1. A higher value of $\mathbf{R^2}$ indicates that a greater proportion of variation in Y is explained by X.

11.6 CONCLUSION

Simple linear regression is one of the most important and widely used techniques in statistical analysis for examining the relationship between two quantitative variables. By expressing this relationship through a linear model, it enables researchers and practitioners to describe patterns, quantify associations, and make predictions based on observed data. Its simplicity, interpretability, and broad applicability make it a foundational tool in both theoretical and applied statistics.

A major strength of simple linear regression lies in its ability to estimate unknown model parameters using observed sample data. These estimates provide meaningful numerical summaries that describe how the response variable changes with respect to the explanatory variable. In addition to parameter estimation, the method offers a structured approach for testing hypotheses, allowing analysts to determine whether the observed relationship is statistically significant or likely due to random variation.

Another important aspect of simple linear regression is the evaluation of model fit. Measures of goodness of fit help assess how effectively the model explains variability in the data, while diagnostic tools highlight potential limitations or violations of assumptions. Together, these techniques ensure that conclusions drawn from the model are not only statistically valid but also practically relevant.

The reliability of regression analysis depends heavily on the validity of its underlying assumptions. Assumptions such as linearity, independence, constant variance, and normality of errors form the basis for accurate estimation and inference. Careful examination and validation of these assumptions enhance the credibility of the results and reduce the risk of misleading interpretations.

In summary, simple linear regression provides a comprehensive framework for understanding relationships between variables, making predictions, and supporting data-driven decisions. When properly applied and interpreted, it serves as a powerful analytical tool that lays the groundwork for more advanced regression techniques and statistical modeling approaches.

11.7 SELF ASSESSMENT QUESTIONS

- Define the simple linear regression model and explain its components.
- State and explain the assumptions of simple linear regression.
- Derive the least squares estimators of regression coefficients.
- Explain the significance test for the regression slope.
- What is the importance of the coefficient of determination?

11.8 FURTHER READINGS

- Draper, N.R. and Smith, H., *Applied Regression Analysis*, Wiley.
- Montgomery, D.C., Peck, E.A., and Vining, G.G., *Introduction to Linear Regression Analysis*, Wiley.
- Rao, C.R., *Linear Statistical Inference and Its Applications*, Wiley.
- Weisberg, S., *Applied Linear Regression*, Wiley.

Dr. U. Ramkiran

LESSON -12

MULTIPLE REGRESSION

OBJECTIVES:

By the end of this lesson, students will be able to:

- Formulate and analyze multiple linear regression models
- Apply matrix methods to estimate regression coefficients
- Interpret partial regression coefficients
- Conduct t-tests and F-tests for model and parameter significance
- Assess goodness of fit and analyze residuals

STRUCTURE

12.1 INTRODUCTION

In many real-life situations, the behavior of a response variable cannot be explained adequately by a single influencing factor. Instead, outcomes are usually determined by the combined effect of several explanatory variables acting simultaneously. For example, agricultural yield may depend on rainfall, soil quality, fertilizer usage, and temperature; economic growth may be influenced by investment, labor, inflation, and government policies; and patient recovery in medical studies may depend on age, treatment type, dosage, and health conditions. In such cases, analyzing the effect of one variable at a time may lead to incomplete or misleading conclusions. Multiple linear regression provides a systematic statistical framework to study such complex relationships.

Multiple linear regression extends the concept of simple linear regression by allowing the response variable to be expressed as a linear function of **more than one independent variable**. This extension enables the model to account for the simultaneous influence of several predictors on an outcome. By incorporating multiple explanatory variables within a single model, it captures the combined and individual effects of predictors more effectively. As a result, it offers a more realistic representation of real-world phenomena where outcomes are rarely driven by a single factor.

A key advantage of multiple linear regression lies in its ability to **isolate the effect of each explanatory variable** while holding other variables constant. This characteristic is especially valuable in observational studies where controlled experiments may not be feasible. By controlling for the influence of other predictors, the model allows researchers to determine the unique contribution of each variable to the response. This helps in identifying important predictors, understanding cause-and-effect relationships, and making informed decisions based on statistical evidence.

Multiple linear regression also plays an important role in **prediction and forecasting**. When several relevant explanatory variables are available, a multiple regression model typically provides more accurate predictions than models based on a single predictor. The inclusion of additional meaningful variables reduces unexplained variability and improves the precision of predicted values. This makes multiple regression particularly useful in applications such as demand forecasting, risk assessment, quality control, and policy analysis.

Another significant feature of multiple linear regression is its flexibility. The model can accommodate both quantitative and categorical variables through appropriate coding techniques. This allows analysts to study a wide range of practical problems involving diverse types of data. Furthermore, multiple regression forms the foundation for many advanced statistical methods, including analysis of covariance, logistic regression, and machine learning regression techniques. Understanding multiple linear regression is therefore essential for further study in applied statistics and data science.

The use of multiple linear regression requires careful attention to model assumptions and diagnostics. Assumptions regarding linearity, independence of errors, constant variance, and the absence of strong multicollinearity must be examined to ensure reliable results. Diagnostic tools such as residual analysis help assess the validity of the model and guide improvements when assumptions are violated. Proper model evaluation enhances the credibility and interpretability of regression results.

In summary, multiple linear regression is a powerful and widely used statistical technique for analyzing relationships involving several explanatory variables. It provides deeper insights into complex data structures, improves predictive accuracy, and supports evidence-based decision-making across various disciplines. By enabling the study of multiple factors simultaneously, it serves as an indispensable tool for researchers, analysts, and practitioners dealing with real-world data.

12.2 MULTIPLE LINEAR REGRESSION MODEL

The multiple linear regression model expresses the response variable Y as a linear function of several explanatory variables X1,X2,…,Xk:

$Yi = \beta 0 + \beta 1 X_{i1} + \beta 2 X_{i2} + \cdots + \beta k X_{ik} + \varepsilon_i, i=1,2,\ldots,n$

where
$\beta 0$ is the intercept,
$\beta 1, \beta 2, \ldots, \beta k$ are regression coefficients,
εi is the random error term.
Each coefficient measures the effect of its corresponding explanatory variable on the response variable, assuming other variables remain fixed.

12.2.1 Assumptions of Multiple Regression

The multiple regression model relies on the following assumptions:
Assumptions of Multiple Linear Regression
For multiple linear regression to yield reliable parameter estimates and valid statistical inferences, certain fundamental assumptions must be satisfied. These assumptions describe the nature of the relationship between the response variable and the explanatory variables, as well as the behavior of the random error component of the model. Careful verification of these assumptions is essential for correct interpretation and effective application of regression results.

1. Linearity
The multiple linear regression model assumes that the **mean of the response variable** is a linear function of the explanatory variables. This does not require the variables themselves to be linearly related, but it assumes that the expected value of the response can be expressed as a weighted sum of the predictors plus a constant term. Each explanatory variable contributes to the response in an additive manner, and the effect of a variable is measured through its regression coefficient.

Linearity simplifies both interpretation and estimation, allowing each coefficient to represent the average change in the response for a unit change in the corresponding predictor, assuming other variables remain constant. Although real-world relationships may be complex, linear models often provide an adequate approximation over limited ranges of the data.

2. Independence of Observations and Error Terms
Another important assumption is that the observations, and consequently the error terms, are **statistically independent**. This means that the value of the error associated with one observation does not influence the error of another observation. Independence is particularly relevant in data collected over time or space, where patterns such as autocorrelation may occur.

Violations of this assumption can result in biased estimates of standard errors, leading to incorrect conclusions in hypothesis testing. Ensuring independence improves the reliability of statistical inference and the validity of conclusions drawn from the model.

3. Zero Mean of Error Terms
The regression model assumes that the **expected value of the error term is zero** for all combinations of the explanatory variables. This condition implies that the model is correctly specified in the sense that it does not systematically overestimate or underestimate the response variable.

A zero mean error ensures that the regression line passes through the center of the data and that the estimated coefficients are unbiased. If this assumption is violated, it indicates that

important variables may be missing from the model or that the functional form of the relationship is not properly specified.

4. Constant Variance (Homoscedasticity)
Homoscedasticity refers to the assumption that the **variance of the error term remains constant** across all observations. In other words, the spread of the residuals around the regression surface should be approximately the same for all values of the explanatory variables.

When this assumption holds, the least squares estimators are efficient and have minimum variance. If the variance of errors changes with the level of the predictors, a condition known as heteroscedasticity arises. This can distort standard errors and reduce the effectiveness of statistical tests, making it essential to diagnose and address when present.

5. No Perfect Multicollinearity
Multiple linear regression requires that the explanatory variables are **not exact linear combinations of one another**. This assumption ensures that each regression coefficient can be uniquely estimated. Perfect multicollinearity occurs when one predictor can be expressed exactly as a linear combination of others, making the estimation of coefficients impossible.
Although perfect multicollinearity is rare in practice, high levels of correlation among predictors can still cause instability in coefficient estimates. Avoiding or addressing multicollinearity improves interpretability and numerical reliability of the regression results.
Importance of Assumptions

The validity of regression results depends strongly on how well these assumptions are satisfied. When the assumptions hold, the estimated coefficients are unbiased, consistent, and efficient, and standard hypothesis tests and confidence intervals are valid. Diagnostic tools such as residual plots and variance measures help assess these assumptions in applied work.
Conclusion

The assumptions of multiple linear regression form the foundation for reliable estimation and meaningful inference. Linearity, independence, zero mean errors, constant variance, and absence of perfect multicollinearity collectively ensure that the model accurately represents the underlying data-generating process. Careful examination and validation of these assumptions are crucial for drawing sound conclusions and making effective predictions using multiple linear regression.

12.2.2 Interpretation of Regression Coefficients

In multiple linear regression analysis, regression coefficients play a central role in explaining the relationship between the response variable and the explanatory variables. Each coefficient provides a quantitative measure of how the response variable is expected to change as a specific explanatory variable changes, while all other variables in the model are held constant. This interpretation allows researchers to study the individual contribution of each predictor within a multivariable framework.

The coefficient associated with a particular explanatory variable represents the **marginal effect** of that variable on the response. Specifically, it measures the expected change in the

response variable resulting from a one-unit increase in the explanatory variable, assuming that all remaining variables in the model remain unchanged. This "holding other variables constant" condition is essential, as it allows the effect of one variable to be isolated from the influence of others, which is particularly important when explanatory variables are correlated. The **sign of a regression coefficient** indicates the direction of the relationship between the explanatory variable and the response variable. A positive coefficient suggests that, on average, an increase in the explanatory variable leads to an increase in the response variable, provided other variables are fixed. Conversely, a negative coefficient implies an inverse relationship, where an increase in the explanatory variable is associated with a decrease in the response variable. The sign therefore provides immediate qualitative insight into the nature of the relationship.

The **magnitude of a regression coefficient** reflects the strength of the relationship. Larger absolute values indicate a stronger effect of the explanatory variable on the response variable, while smaller values suggest a weaker influence. However, the magnitude must be interpreted carefully, as it depends on the scale and units of measurement of the variables involved. For meaningful comparisons among coefficients, variables often need to be standardized or appropriately transformed.

Regression coefficients also carry important **contextual meaning** depending on the variables used in the model. For continuous explanatory variables, the coefficient describes the expected change in the response per unit change of the predictor. When categorical variables are included through indicator or dummy variables, the coefficients represent differences in the mean response relative to a reference category. Thus, correct interpretation requires an understanding of how each variable is defined and measured.

Another important aspect of interpreting regression coefficients is the distinction between **statistical significance and practical significance**. A coefficient may be statistically significant, indicating strong evidence of an association, yet have a small magnitude that limits its practical importance. Conversely, a coefficient with a large magnitude may not be statistically significant if the data exhibit substantial variability. Therefore, both the size of the coefficient and its statistical significance must be considered together.

Regression coefficients are also influenced by the presence of other variables in the model. Adding or removing explanatory variables can change coefficient estimates, particularly when predictors are correlated. This highlights the importance of careful model specification and awareness of multicollinearity, which can inflate standard errors and make coefficient estimates unstable.

In applied analysis, interpretation of regression coefficients supports decision-making and policy formulation. By quantifying the effect of individual variables while controlling for others, multiple regression enables analysts to identify key drivers of an outcome and assess potential impacts of changes in explanatory variables.

Regression coefficients provide meaningful and interpretable measures of the relationship between explanatory variables and the response variable in multiple linear regression. Their sign indicates the direction of influence, their magnitude reflects the strength of the effect, and their interpretation depends on both the model structure and the measurement scale. Proper understanding of regression coefficients is essential for drawing valid conclusions and making informed decisions based on regression analysis.

12.3 Estimation of parameters

In multiple linear regression analysis, the unknown parameters of the model are typically estimated using the **method of least squares**. This method provides a systematic and objective approach to determine the values of regression coefficients that best represent the relationship between the response variable and the set of explanatory variables. The central idea of least squares estimation is to minimize the discrepancy between the observed values of the response variable and the values predicted by the regression model.

The discrepancy between an observed value and its corresponding predicted value is known as a **residual**. For each observation, the residual represents the portion of the response variable that is not explained by the regression model. Least squares estimation seeks to determine the regression coefficients such that the **sum of the squared residuals** across all observations is as small as possible. Squaring the residuals ensures that both positive and negative deviations are treated equally and gives greater weight to larger errors.

Mathematically, the multiple linear regression model can be expressed as a linear combination of explanatory variables along with a random error term. The fitted model generates predicted values for the response variable, and the difference between the observed and predicted values forms the residuals. The least squares criterion minimizes the aggregate of these squared residuals, thereby producing parameter estimates that provide the closest possible fit to the observed data in the sense of minimizing overall error.

One of the important features of least squares estimation is its **analytical convenience**. Under standard regression assumptions, the minimization problem yields a set of normal equations that can be solved to obtain explicit expressions for the regression coefficients. These equations ensure that the resulting estimates balance the deviations in the data and satisfy optimality conditions. The least squares estimators are therefore well-defined and computationally efficient, even when multiple explanatory variables are involved.

Another key advantage of the least squares method is its **statistical optimality**. When the assumptions of linearity, independence, zero mean of errors, and constant variance are satisfied, least squares estimators possess desirable properties. They are unbiased, meaning that their expected values equal the true parameter values. They are also efficient in the sense that they have the smallest variance among all linear unbiased estimators. As sample size increases, the estimators become more precise, making them reliable for large datasets.

Least squares estimation also serves as the foundation for **statistical inference** in regression analysis. Once parameter estimates are obtained, their sampling distributions can be studied to conduct hypothesis tests and construct confidence intervals. This enables analysts to assess whether individual explanatory variables have significant effects on the response variable and to quantify the uncertainty associated with the estimated coefficients.

In practice, least squares estimation is closely tied to **model evaluation and diagnostics**. The residuals obtained from the fitted model are used to assess the validity of model assumptions and to identify potential issues such as outliers, non-linearity, or unequal variance. Thus, the estimation process not only provides parameter estimates but also supports model refinement and validation.

In summary, the method of least squares is a fundamental and widely used technique for estimating the parameters of the multiple linear regression model. By minimizing the sum of squared residuals, it produces efficient and interpretable estimates that form the basis for inference, prediction, and decision-making. Its mathematical simplicity, optimal properties, and practical usefulness make it an essential tool in regression analysis.

12.3.1 Least Squares Estimation

A multiple linear regression model predicts the value of a dependent variable (Y) using multiple independent variables

($X_1 X_2 X_3 X_4 X_5 ..... X_k$)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon$$

Whereas β0 = Intercept;

Regression coefficients (parameters) and

ε = Random error term

To estimate the unknown coefficients $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4 ..... \beta_k$ from observed data using least squares method**.**

Method of Least Squares Estimation:

We find estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4 ..... \hat{\beta}_k$ that minimize the sum of squared errors (SSE)

$$SSE = \sum (Y - \hat{Y})^2 = \sum (Y - \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k)^2$$

Matrix Form for Estimation; The model can be written in matrix form: $Y = X\beta + \varepsilon$

Whereas:

- Y = (n×1) vector of responses
- X = (n×k) matrix of predictors (with a column of 1s for intercept)
- β = (k×1) vector of parameters
- ε = (n×1) Error vector

Least Squares Estimator: $\hat{\beta} = \left(X^1 X \right)^{-1} X^1 y$; It gives the best linear unbiased estimator (BLUE) under classical regression assumptions.

12.3.2 Matrix Approach to Multiple Regression

The multiple regression model can be expressed in matrix form as:

Least Squares Estimator: $\hat{\beta} = \left(X^1 X \right)^{-1} X^1 y$; It gives the best linear unbiased estimator (BLUE) under classical regression assumptions.

The matrix formulation simplifies computation and forms the basis for theoretical analysis and extension to advanced regression models.

12.4 TESTS OF SIGNIFICANCE

Statistical tests play a crucial role in multiple linear regression analysis by determining whether the explanatory variables have a meaningful influence on the response variable. While parameter estimation provides numerical values for regression coefficients, statistical

testing helps assess whether these estimated effects are statistically significant or could have arisen due to random variation in the data.

In regression analysis, significance testing is primarily concerned with evaluating hypotheses about the regression parameters. The most common approach is to test whether the coefficient of an explanatory variable is equal to zero. A zero coefficient implies that the variable has no linear effect on the response variable when other variables in the model are held constant. Statistical tests allow researchers to decide whether there is sufficient evidence to reject this assumption.

Two main types of tests are widely used in multiple linear regression. **Individual significance tests**, such as the t-test, examine the contribution of each explanatory variable separately. These tests help identify which variables are important predictors of the response variable and which may be excluded from the model without substantially reducing its explanatory power. This is especially useful in models with many predictors, where some variables may not have a significant impact.

In addition to individual tests, **overall model significance tests**, such as the F-test, are used to evaluate the regression model as a whole. The F-test determines whether the set of explanatory variables collectively provides a better explanation of the response variable than a model with no predictors. A significant result indicates that at least one explanatory variable has a non-zero effect on the response variable.

Statistical tests also support **model building and validation**. By examining significance levels, analysts can refine models, compare competing models, and avoid including unnecessary variables. This leads to simpler, more interpretable models without compromising predictive performance.

It is important to note that statistical significance does not always imply practical importance. A variable may be statistically significant but have a small effect size that is of limited practical relevance. Therefore, significance tests should be interpreted alongside regression coefficients, confidence intervals, and subject-matter knowledge.

In summary, statistical tests in multiple linear regression provide a formal framework for evaluating the influence of explanatory variables on the response variable. They help distinguish genuine relationships from random noise, support sound model selection, and enhance the reliability of conclusions drawn from regression analysis.

12.4.1 t-test for Individual Regression Coefficients

The t-test examines whether a particular regression coefficient differs significantly from zero. The null hypothesis is: $H0: \beta j = 0$
If the calculated t-value exceeds the critical value, the null hypothesis is rejected, indicating that the variable has a significant effect on the response.

12.4.2 F-test for Overall Model Significance

The F-test assesses whether the regression model as a whole is statistically significant. It tests whether at least one explanatory variable has a non-zero effect on the response variable. A

significant F-value implies that the model provides a better fit than a model without explanatory variables.

12.5 MODEL ADEQUACY AND DIAGNOSTICS

Model adequacy refers to the extent to which a regression model appropriately represents the underlying data and satisfies the assumptions on which the regression analysis is based. An adequate model not only fits the observed data reasonably well but also provides reliable estimates, valid inferences, and meaningful predictions. Evaluating model adequacy is therefore a critical step in regression analysis.

A regression model is considered adequate when it captures the systematic relationship between the response variable and the explanatory variables without leaving important patterns unexplained. If the model is poorly specified, estimates of regression coefficients may be biased or inefficient, leading to incorrect conclusions. Thus, assessing model adequacy helps determine whether the chosen model structure is suitable for the data under study.

One important aspect of model adequacy is the verification of **regression assumptions**. Assumptions such as linearity, independence of errors, constant variance, normality of errors, and absence of strong multicollinearity must be reasonably satisfied. Violations of these assumptions can affect the accuracy of parameter estimates and the validity of hypothesis tests. Diagnostic checks help identify such problems and guide necessary model improvements.

Another key component of model adequacy is the analysis of **residuals**. Residuals represent the differences between observed and fitted values and provide valuable information about model performance. Patterns in residual plots may indicate issues such as non-linearity, unequal variance, or outliers. A well-fitted model typically shows residuals that are randomly scattered around zero without any systematic structure.

Numerical measures also play an important role in assessing model adequacy. Measures such as the coefficient of determination indicate how much of the variability in the response variable is explained by the model. While a higher value generally suggests a better fit, it must be interpreted carefully and in conjunction with other diagnostic tools.

In summary, model adequacy ensures that a regression model is both statistically sound and practically useful. By examining residuals, checking assumptions, and evaluating goodness-of-fit measures, analysts can confirm whether the model provides a reliable representation of the data. Careful assessment of model adequacy strengthens confidence in the conclusions drawn from regression analysis and improves the quality of predictions.

12.5.1 Coefficient of Multiple Determination

The coefficient of multiple determination, denoted by R2R^2R2, measures the proportion of total variation in the response variable explained by all explanatory variables together. A higher value of R2R^2R2 indicates better explanatory power of the model.

12.5.2 Residual Analysis

Residual analysis involves studying the residuals to detect violations of model assumptions. Plots of residuals are used to identify non-linearity, unequal variances, outliers, and influential observations. Proper residual analysis enhances the reliability of regression results.

12.6 CONCLUSION

Multiple linear regression is a powerful extension of simple regression that allows the study of relationships involving several explanatory variables. By estimating parameters, testing statistical significance, and assessing model adequacy, it provides deeper insight into complex data structures. Proper application and validation of assumptions ensure meaningful interpretation and effective prediction.

12.7 SELF ASSESSMENT QUESTIONS

- Define multiple linear regression and explain its importance.
- State the assumptions of the multiple regression model.
- Explain the interpretation of regression coefficients.
- Describe the t-test and F-test used in multiple regression.
- Discuss the role of residual analysis in model diagnostics.

12.8 FURTHER READINGS

- Draper, N. R. and Smith, H., *Applied Regression Analysis*, Wiley.
- Montgomery, D. C., Peck, E. A., and Vining, G. G., *Introduction to Linear Regression Analysis*, Wiley.
- Rao, C. R., *Linear Statistical Inference and Its Applications*, Wiley.
- Weisberg, S., *Applied Linear Regression*, Wiley.

Dr. G V S R Anjaneyulu

LESSON -13

POLYNOMIAL REGRESSION AND ORTHOGONAL POLYNOMIALS

OBJECTIVES:

By the end of this lesson, students will be able to:

- Polynomial Regression Model
- Fitting of Polynomial Regression Models
- Problems of Multicollinearity in Polynomial Regression
- Introduction to Orthogonal Polynomials
- Construction of Orthogonal Polynomials
- Advantages of Orthogonal Polynomials

STRUCTURE:

13.1 INTRODUCTION

In many practical situations, the relationship between a response variable and an explanatory variable cannot be adequately represented by a straight line. Although linear and multiple linear regression models are effective for describing simple trends, real-world data often display more complex patterns such as curvature, increasing or decreasing rates of change, and turning points. These nonlinear patterns commonly arise in fields such as agriculture, economics, engineering, environmental studies, and the biological sciences. When such behavior is present, linear models may fail to provide an accurate or meaningful description of the underlying relationship.

To address this limitation, **polynomial regression** serves as a valuable extension of linear regression. Polynomial regression allows the inclusion of higher-degree powers of the explanatory variable, enabling the model to capture curvature and more flexible trends in the data. Despite involving nonlinear functions of the explanatory variable, the model remains linear in its parameters. This important property allows the use of standard estimation techniques while enhancing the model's ability to represent complex relationships.

Polynomial regression is particularly useful when exploratory analysis or scatter plots suggest that the effect of the explanatory variable on the response changes at different levels. For example, growth processes may accelerate or decelerate, demand may rise at a decreasing rate, or physical systems may exhibit peak or saturation effects. By incorporating squared, cubic, or higher-order terms, polynomial regression can model such behavior more accurately than simple linear regression.

Another advantage of polynomial regression is its interpretability within a familiar regression framework. The fitted model can be analyzed using established tools such as least squares estimation, hypothesis testing, confidence intervals, and goodness-of-fit measures. This makes polynomial regression both accessible and practical for analysts who are already familiar with linear regression techniques.

However, the inclusion of higher-order polynomial terms introduces certain challenges. One of the major difficulties is **multicollinearity**, which arises because the powers of the explanatory variable are often highly correlated with one another. This correlation can inflate the variances of estimated coefficients, leading to unstable estimates and difficulties in interpreting individual effects. Additionally, polynomial regression may suffer from **numerical instability**, particularly when high-degree polynomials are fitted or when the range of the explanatory variable is large. Small changes in the data may result in large variations in coefficient estimates, reducing the reliability of the model.

To overcome these computational and interpretational issues, **orthogonal polynomials** are commonly employed in polynomial regression. Orthogonal polynomials are constructed in such a way that each polynomial term is uncorrelated with the others over the observed data. This property effectively eliminates multicollinearity among polynomial terms and leads to more stable and efficient parameter estimates.

The use of orthogonal polynomials preserves the fitted values of the regression model while improving numerical behavior and simplifying model assessment. By ensuring that each polynomial term contributes independently to the model, orthogonal polynomials allow clearer identification of the degree of the polynomial that best fits the data. This makes it easier to determine whether additional higher-order terms significantly improve the model or merely add unnecessary complexity.

In practical applications, polynomial regression combined with orthogonal polynomials provides a balanced approach to modeling nonlinear relationships. It offers greater flexibility than simple linear regression while maintaining the interpretability and analytical strengths of linear models. When properly applied, this approach supports accurate representation of complex trends, reliable parameter estimation, and meaningful inference.

In summary, polynomial regression extends the scope of regression analysis by enabling the modeling of nonlinear relationships within a linear framework. While higher-order

polynomial terms can introduce multicollinearity and numerical instability, the use of orthogonal polynomials addresses these issues effectively. Together, they form a powerful and reliable methodology for analyzing curved trends and complex patterns in real-world data.

13.2 POLYNOMIAL REGRESSION

Polynomial regression is an extension of simple linear regression that is used when the relationship between the response variable and an explanatory variable cannot be adequately described by a straight line. Instead of restricting the model to a single linear term, polynomial regression includes higher-order powers of the explanatory variable, allowing the model to capture curvature and more complex trends in the data.

In polynomial regression, the response variable is expressed as a function of the explanatory variable raised to different powers, such as squared or cubic terms. This enables the model to represent nonlinear patterns like increasing or decreasing rates of change and turning points. Although the fitted curve may appear nonlinear when plotted, the model is linear in its parameters. This means that the coefficients enter the model in a linear manner and can be estimated using standard least squares techniques.

The linearity in parameters is an important advantage of polynomial regression. It allows the use of well-established methods for estimation, hypothesis testing, and model diagnostics that are commonly applied in linear regression. As a result, polynomial regression combines flexibility in modeling nonlinear relationships with the simplicity and interpretability of linear regression models.

Polynomial regression is particularly useful in situations where exploratory analysis suggests that a straight-line model is inadequate but the underlying relationship can still be approximated smoothly. By choosing an appropriate degree for the polynomial, the model can achieve a good balance between capturing the true pattern in the data and avoiding unnecessary complexity.

In summary, polynomial regression provides a practical and effective approach for modeling nonlinear relationships. By incorporating powers of the explanatory variable while remaining linear in parameters, it extends the capabilities of linear regression without sacrificing analytical convenience and statistical rigor.

13.2.1 Polynomial Regression Model

Modelling the relationship between a dependent variable (Y) and an independent variable (X) using powers of X is known as polynomial fitting. Polynomial models for one variable may consists in

1. Orthogonal polynomials
2. Piecewise Polynomial.

Examples:

1. This is useful when the relationship between X and Y is **non-linear**. Assume we aim to model the relationship between crop yield and the quantity of fertilizer applied. The relationship may be curved, small amounts help, but too much fertilizer reduces yield.

A **quadratic polynomial fit** $Y=\beta_0+\beta_1X+\beta_2X^2+\varepsilon$ like can model this better than a straight line.

2. We fit a model like: $Y=\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+\varepsilon$ is a polynomial regression model in one variable and is called a cubic model. The coefficients β_1 , β_2 and β_3 are called the linear effect parameter and cubic effect parameters respectively.
3. The k^{th} order polynomial model in one variable is given by $Y=\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+.....+\beta_kX^k+\varepsilon$; If X = X^i for all i= 1, 2, ..., k then the model is multiple linear regression model in k explanatory variables. So the linear regression model includes $Y=X\beta+\varepsilon$ the polynomial regression model. Thus the techniques for fitting linear regression model can be used for fitting the polynomial regression model.

Polynomial Models

A. Order of the polynomial model: $k \le 2$

B. Strategy for polynomial Model building: forward selection: start with linear models

$Y=\beta_0+\beta_1X+\varepsilon$

$Y=\beta_0+\beta_1X+\beta_2X^2+\varepsilon$

$Y=\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+\varepsilon$

$Y=\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+.....+\beta_kX^k+\varepsilon$; Successive fit model of increasing order unit the t-test for the highest order term is non-significant.

C. Ill-Conditioning: as the order of the polynomial increases, the $\left(X^1X\right)$ matrix becomes ill-conditioned, that is $\left(X^1X\right)^{-1}$ calculation becomes inaccurate. Then $\hat{\beta}=\left(X^1X\right)^{-1}X^1y$ does not exists. If the value of X are limited to a narrow range in columns of X.

Example: Let us consider polynomial model

$Y=\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+.....+\beta_kX^k+\varepsilon$

$$\begin{bmatrix} 1 & x & x^2 & .. \\ 1 & 0.11 & 0.0121 & .. \\ 1 & 0.12 & 0.0144 & .. \\ 1 & 0.13 & 0.0169 & .. \end{bmatrix}$$

Centring the data may remove ill-conditioning. We fit the model

$y-\bar{y}=\beta_0+\beta_1(x-\bar{x})+\beta_2(x-\bar{x})^2+\varepsilon$

Instead of $Y=\beta_0+\beta_1X+\beta_2X^2+\varepsilon$

Orthogonal Polynomials:

suppose we wish to fit the polynomial regression model $Y=\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+.....+\beta_kX^k+\varepsilon$; If we wish to add another term $\beta_{k+1}X^{k+1}$ that is $\begin{bmatrix}1 & x & x^2 & .....x^k & x_{k+1}\\ & & & & \\ & & & & \end{bmatrix}$ we must recalculate $\left(X^1X\right)^{-1}$ and estimates of lower order parameters $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\beta}_4.....\hat{\beta}_k$ will change. For this kind problems we use orthogonal polynomials. If we construct polynomials $P_0(X),P_1(X),.....P_K(X)$ with the property that they are orthogonal polynomials. $\sum_{i=1}^{n}P_n(X_i)P_s(X_i)=0, r\neq s=1(1)k$; we can rewrite the model as $Y_i=\beta_0+\beta_1X_i+\beta_2X_i^2+\beta_3X_i^3+.....+\beta_kX_i^k+\varepsilon_i$; Where $P_n(X_i)$ is the r[th] ordered orthogonal polynomial.

13.2.2 Estimation of Polynomial Regression Coefficients

Consider the polynomial model of order k is one variable as

$Y_i=\beta_0+\beta_1X_i+\beta_2X_i^2+\beta_3X_i^3+.....+\beta_kX_i^k+\varepsilon_i;\ \forall i=1,2,.....k$

When writing this model as $Y=X\beta+\varepsilon$ the columns of X will not be orthogonal. If we add another term $\beta_{k+1}X^{k+1}$ then the matrix $\left[X^1X\right]^{-1}$ has to be recomputed and consequently, the lower order parameters $\beta_0,\beta_1,\beta_2,.....\beta_k$ will also change.

Consider the fitting of the following model:

$Y_i=\theta_0P_0(X_i)+\theta_1P_1(X_i)+\theta_2P_2(X_i)+.....+\theta_kP_k(X_i)+\varepsilon_i$; for all i= 1,2…….n

In the context of $Y=X\beta+\varepsilon$, the X-matrix, in this case, is given by

$$X=\begin{bmatrix}P_0(X_1) & P_1(X_1) & P_2(X_1)..... & P_k(X_1)\\ P_0(X_2) & P_1(X_2) & P_2(X_2)..... & P_k(X_2)\\ --- & --- & --- & ---\\ P_0(X_n) & P_1(X_n) & P_2(X_n)..... & P_k(X_n)\end{bmatrix}$$

Since this X-matrix has orthogonal columns, so X^1X matrix becomes

$$X^1X=\begin{bmatrix}\sum P_0^2(X_i) & 0 & ..... & 0\\ 0 & \sum P_1^2(X_i) & ..... & 0\\ --- & --- & --- & ---\\ 0 & 0 & ..... & \sum P_k^2(X_i)\end{bmatrix}$$

The ordinary least squares estimator is $\hat{\theta}=\left(X^1X\right)^{-1}X^1y$ and its variance is obtained from $V\left(\hat{\theta}\right)=\sigma^2\left(X^1X\right)^{-1}$ When σ^2 is unknown, it can be estimated from the analysis of variance table.

13.2.2 Estimation of Polynomial Regression Coefficients

The coefficients of a polynomial regression model are estimated using the **method of least squares**. The objective is to minimize the sum of squared deviations between observed values and fitted values of the response variable.
Although the model includes nonlinear terms of X, it is linear in parameters and therefore standard least squares techniques apply. The normal equations derived from minimizing the residual sum of squares yield estimates of the regression coefficients.

13.3 PROBLEMS IN POLYNOMIAL REGRESSION

While polynomial regression can model complex relationships, it may lead to certain practical difficulties.

13.3.1 Multicollinearity

Multicollinearity arises when explanatory variables are highly correlated with one another. In polynomial regression, this problem is common because higher powers of X are often strongly correlated.
High multicollinearity can result in:

- Unstable coefficient estimates
- Large standard errors
- Difficulty in interpreting individual regression coefficients

13.3.2 Numerical Instability

Numerical instability occurs when computations become sensitive to small changes in data, particularly when high-degree polynomials are used. Large powers of X can cause rounding errors and lead to unreliable estimates.
This issue becomes more severe when X values are large or unevenly spaced, making it difficult to obtain accurate and stable regression coefficients.

13.4 ORTHOGONAL POLYNOMIALS

To overcome the problems of multicollinearity and numerical instability in polynomial regression, **orthogonal polynomials** are employed.

13.4.1 Concept and Construction of Orthogonal Polynomials

Orthogonal polynomials are a set of polynomial functions that are mutually uncorrelated with respect to a given inner product or weighting scheme. This means that the cross-products of different polynomial terms sum to zero over the observed data.

These polynomials are usually constructed by applying orthogonalization techniques to ordinary polynomial terms. As a result, each term contributes independently to the regression model.

13.4.2 Use of Orthogonal Polynomials in Regression

When orthogonal polynomials are used as regressors:

- Multicollinearity among polynomial terms is eliminated
- Parameter estimates become more stable
- Numerical computation becomes more reliable

The fitted values of the model remain unchanged, but the interpretation and estimation of coefficients improve significantly.

13.5 ADVANTAGES OF ORTHOGONAL POLYNOMIALS

Orthogonal polynomials offer several advantages in regression analysis:

- Reduction of multicollinearity
- Improved numerical stability
- Independent contribution of each polynomial term
- Easier identification of the effective degree of the polynomial
- Reliable parameter estimation for higher-degree models

These advantages make orthogonal polynomials particularly useful in practical data analysis involving polynomial relationships.

13.6 CONCLUSION

Polynomial regression extends linear regression by allowing curved relationships between variables while maintaining linearity in parameters. However, the use of higher-degree polynomial terms may introduce multicollinearity and numerical instability. Orthogonal polynomials provide an effective solution to these problems by producing stable and reliable estimates. Together, polynomial regression and orthogonal polynomials form important tools for modeling complex nonlinear trends in data.

13.7 SELF ASSESSMENT QUESTIONS

- What is polynomial regression and how does it differ from simple linear regression?
- Write the general form of a polynomial regression model.
- Explain the problem of multicollinearity in polynomial regression.
- What is numerical instability and why does it occur?
- Define orthogonal polynomials and explain their use in regression.

13.8 FURTHER READINGS

- Draper, N. R. and Smith, H., *Applied Regression Analysis*, Wiley.
- Montgomery, D. C., Peck, E. A., and Vining, G. G., *Introduction to Linear Regression Analysis*, Wiley.
- Rao, C. R., *Linear Statistical Inference and Its Applications*, Wiley.
- Weisberg, S., *Applied Linear Regression*, Wiley.

Dr. G V S R Anjaneyulu

LESSON -14
MULTICOLLINEARITY

OBJECTIVES:

By the end of this lesson, students will be able to:

- Multicollinearity
- Introduction to Multicollinearity
- Causes of Multicollinearity
- Effects of Multicollinearity on Regression Estimates
- Detection of Multicollinearity
- Correlation Matrix
- Variance Inflation Factor (VIF)
- Condition Index
- Remedies for Multicollinearity

STRUCTURE

14.1 INTRODUCTION

Multiple linear regression is widely used to analyze relationships in which a response variable depends on several explanatory variables simultaneously. This approach allows researchers and analysts to understand how different factors jointly influence an outcome while controlling for the effect of other variables. For multiple linear regression to function

effectively, certain assumptions must be satisfied. One of the most critical among these is the requirement that the explanatory variables should not be highly correlated with one another.

When this assumption is violated, a situation known as **multicollinearity** arises. Multicollinearity refers to the presence of strong linear relationships among two or more explanatory variables included in a regression model. In such cases, it becomes difficult to separate the individual effect of each variable on the response. This problem is especially common in applied data analysis, where explanatory variables often originate from related measurements, shared underlying processes, or constructed variables.

Multicollinearity frequently occurs when variables are related by nature. For example, economic indicators such as income, savings, and expenditure are often interrelated. Similarly, in agricultural or biological studies, measurements like rainfall, humidity, and soil moisture may show strong association. Additionally, multicollinearity commonly arises in models that include **polynomial terms or interaction variables**, as higher powers or combined terms of a variable are naturally correlated with the original variable.

Another important cause of multicollinearity is poor study design or data limitations. When data lack sufficient variation or when too many explanatory variables are included relative to the sample size, the regression model may exhibit near-linear dependence among predictors. This situation results in unstable estimation and reduces the clarity of interpretation.

A key feature of multicollinearity is that, despite its adverse effects on parameter estimation, it **does not necessarily reduce the overall predictive ability** of the regression model. The fitted values and predictions may remain accurate, and overall goodness-of-fit measures such as the coefficient of determination may still appear satisfactory. However, the reliability of individual regression coefficients is significantly affected. Coefficients may have large standard errors, unexpected signs, or magnitudes that are inconsistent with theoretical expectations.

Because of inflated standard errors, explanatory variables that are genuinely important may appear statistically insignificant in hypothesis tests. This can lead to incorrect conclusions about which variables influence the response. Moreover, small changes in the data or model specification may cause large fluctuations in the estimated coefficients, making the model unstable and difficult to interpret.

Understanding multicollinearity is therefore essential for proper regression analysis. Analysts must be able to identify its presence, assess its severity, and decide on appropriate corrective measures. Diagnostic tools such as correlation matrices, variance inflation factors, and condition indices are commonly used to detect multicollinearity. Once identified, various strategies can be employed, including variable selection, transformation, or the use of alternative estimation techniques such as ridge regression and principal component regression.

In summary, multicollinearity is a common and important issue in multiple linear regression analysis. While it does not impair the model's ability to predict the response variable, it undermines the reliability and interpretability of individual regression coefficients. Therefore, careful attention to the detection and treatment of multicollinearity is essential for drawing meaningful conclusions and ensuring the effectiveness of regression-based decision-making.

14.2 MULTICOLLINEARITY

Multicollinearity refers to a situation in multiple linear regression where two or more explanatory variables exhibit strong linear relationships among themselves. In such cases, one explanatory variable can be approximately expressed as a linear combination of one or more other explanatory variables. This lack of independence among the predictors creates difficulties in estimating and interpreting regression coefficients accurately.

When explanatory variables are highly correlated, the regression model faces challenges in distinguishing the individual effect of each variable on the response. Although the combined effect of these variables may be meaningful, their separate contributions become uncertain. As a result, the estimated regression coefficients may become unstable and sensitive to small changes in the data.

Multicollinearity can exist in two forms: **perfect multicollinearity** and **imperfect multicollinearity**. Perfect multicollinearity occurs when one explanatory variable is an exact linear combination of other variables, making it impossible to estimate unique regression coefficients. Imperfect multicollinearity, which is more common in real-world data, involves strong but not exact linear relationships among explanatory variables. While estimation is still possible in this case, the results may lack precision and reliability.

It is important to note that multicollinearity does not necessarily affect the overall fit or predictive performance of the regression model. Measures such as the coefficient of determination may remain high even in the presence of multicollinearity. However, hypothesis testing and interpretation of individual regression coefficients become problematic due to inflated standard errors and reduced statistical significance.

In practical data analysis, multicollinearity often arises due to the inclusion of related variables, polynomial terms, or interaction effects, as well as limitations in data collection. Recognizing the presence of multicollinearity is therefore essential for conducting meaningful regression analysis.

14.2.1 Meaning and Nature of Multicollinearity

Multicollinearity occurs when two or more explanatory variables are highly correlated, meaning that one variable can be approximately expressed as a linear combination of others. Multicollinearity arises in multiple linear regression when two or more explanatory variables exhibit a strong linear relationship among themselves. In such situations, one explanatory variable can be closely approximated by a linear combination of the other variables included in the model. This lack of independence among the predictors reduces the model's ability to clearly distinguish the individual contribution of each explanatory variable to the response variable.

In extreme situations, known as **perfect multicollinearity**, an explanatory variable is an exact linear combination of other variables, making it impossible to obtain unique estimates of the regression coefficients. In practice, multicollinearity is usually **imperfect**, but it may still be strong enough to introduce considerable instability into the estimation process. As a consequence, the estimated regression coefficients may exhibit unreasonable magnitudes or incorrect signs and may change substantially with small variations in the data, even though overall goodness-of-fit measures indicate that the model fits the data satisfactorily.

14.2.2 Causes of Multicollinearity

- Multicollinearity can arise due to several reasons, including:
- **Inherent relationship among variables** – Variables measuring similar characteristics tend to be correlated.
- **Use of polynomial or interaction terms** – Higher-order terms highly correlated with X.
- **Data collection methods** – Poor experimental design or lack of variation in data.
- **Over-specification of the model** – Including too many related explanatory variables.
- **Dummy variable trap** – Inclusion of all categories of a categorical variable without omitting a reference level.

14.3 EFFECTS OF MULTICOLLINEARITY

Multicollinearity significantly affects the stability and interpretability of results obtained from a multiple linear regression model. Although the presence of multicollinearity does not violate the basic form of the regression model or necessarily reduce its predictive power, it introduces several practical difficulties that can undermine the usefulness of regression analysis for interpretation and inference.

One of the primary effects of multicollinearity is the **instability of regression coefficient estimates**. When explanatory variables are highly correlated, small changes in the data or in the model specification can lead to large changes in the estimated coefficients. This sensitivity makes the regression coefficients unreliable, as they may vary considerably from one sample to another. Consequently, the estimated coefficients may not accurately reflect the true relationship between individual explanatory variables and the response variable.

Another important consequence of multicollinearity is the **inflation of standard errors** associated with the regression coefficients. High correlations among explanatory variables increase the variability of coefficient estimates, leading to larger standard errors. As a result, the calculated t-statistics for individual regression coefficients may be smaller in absolute value, making statistically significant variables appear insignificant. This can lead to incorrect conclusions regarding the importance of explanatory variables.

Multicollinearity also complicates the **interpretation of regression coefficients**. In the presence of strong correlations among predictors, it becomes difficult to interpret the effect of one variable while holding the others constant. This is because changes in one variable are often associated with changes in another, violating the practical meaning of ceteris paribus interpretation. Even when coefficients are statistically significant, their practical interpretation may be unclear or misleading.

Despite these issues, it is important to note that multicollinearity does not necessarily reduce the **overall goodness of fit** of the model. Measures such as the coefficient of determination may remain high, indicating that the model explains a large proportion of variability in the response variable. However, a good overall fit can be deceptive if individual coefficients are unstable or unreliable.

Another effect of multicollinearity is its impact on **model selection and inference**. When explanatory variables are strongly correlated, it becomes difficult to determine which variables should be retained or removed from the model. Different subset selection methods

may lead to different models, and conclusions drawn from hypothesis tests may lack robustness. This uncertainty weakens the confidence in regression-based decisions.

Multicollinearity can also obscure the underlying relationships in the data. Variables that are theoretically important may be excluded due to insignificance caused by inflated standard errors, while less relevant variables may appear important due to chance correlations. This distortion can misguide researchers and practitioners who rely on regression results for policy formulation or scientific interpretation.

In summary, while multicollinearity does not bias the least squares estimates or invalidate the regression model as a whole, it adversely affects the precision, stability, and interpretability of individual regression coefficients. Careful diagnosis and appropriate remedial measures are therefore essential when multicollinearity is present. Addressing this problem improves the reliability of regression analysis and enhances the clarity of conclusions drawn from the model.

14.3.1 Impact on Regression Coefficients

When multicollinearity is present:

- Regression coefficients may become **unstable and sensitive** to small changes in data.
- Estimated coefficients may have **unexpected signs or magnitudes**.
- Individual coefficients become difficult to interpret, even if they are theoretically important.
- Although the fitted values of the model may remain accurate, the individual regression coefficients lose their reliability.

14.3.2 Effect on Standard Errors and Tests of Significance

Multicollinearity leads to:

- **Inflated standard errors** of regression coefficients.
- Reduced **t-statistics**, causing important variables to appear statistically insignificant.
- Difficulty in identifying truly influential explanatory variables.
- As a result, hypothesis tests and confidence intervals become unreliable.

14.4 DETECTION OF MULTICOLLINEARITY

Several diagnostic tools are available to detect multicollinearity in a regression model.

14.4.1 Correlation Matrix

A simple method of detecting multicollinearity is examining the correlation matrix of explanatory variables. High pairwise correlations indicate potential multicollinearity. However, this method may fail to detect complex multivariate relationships.

14.4.2 Variance Inflation Factor (VIF)

The Variance Inflation Factor measures the extent to which the variance of a regression coefficient is inflated due to multicollinearity.

$$VIFj = \frac{1}{1-Rj^2}$$

where R_j^2 is the coefficient of determination obtained by regressing the j^{th} explanatory variable on all other explanatory variables.
VIF =1: No multicollinearity
VIF > 10: Serious multicollinearity

14.4.3 Condition Index

Condition Index
The **condition index** is an important diagnostic measure used to detect the presence and severity of multicollinearity in a multiple linear regression model. It is based on the **eigenvalues of the correlation matrix (or equivalently, the cross-product matrix) of the explanatory variables**. By examining how the explanatory variables relate to one another at a multivariate level, the condition index provides deeper insight than simple pairwise correlations.

The condition index is calculated by taking the square root of the ratio of the largest eigenvalue of the correlation matrix to each individual eigenvalue. Mathematically, it is expressed as

$$\text{Condition Index}_i = \sqrt{\frac{\lambda_{max}}{\lambda_i}}$$

where λ_{max} is the largest eigenvalue and λ_i is the i-th eigenvalue of the correlation matrix. A small eigenvalue indicates that some linear combination of the explanatory variables contributes very little independent information, which is a sign of multicollinearity.

Large values of the condition index suggest **strong dependencies among explanatory variables**. Generally, condition index values below 10 indicate weak or no multicollinearity, values between 10 and 30 suggest moderate multicollinearity, and values exceeding 30 are taken as evidence of severe multicollinearity. Very high condition index values imply that the regression coefficients may be highly unstable and sensitive to small changes in the data.

Unlike simple correlation measures, the condition index is capable of detecting complex multicollinearity involving more than two explanatory variables. It therefore provides a comprehensive diagnostic tool when explanatory variables are related in a multivariate manner rather than just in pairs.

In summary, the condition index is a valuable method for diagnosing multicollinearity in multiple regression analysis. By relying on eigenvalues of the explanatory variable correlation matrix, it identifies hidden linear dependencies and helps assess the reliability and stability of regression coefficient estimates.

- Values less than 10 indicate weak dependence.
- Values above 30 suggest severe multicollinearity.

14.5 REMEDIES FOR MULTICOLLINEARITY

- Possible remedial measures include:
- Removing or combining highly correlated variables
- Increasing sample size
- Centering variables in polynomial regression
- Using **ridge regression**

Applying **principal component regression**

- Selecting an appropriate subset of explanatory variables
- The choice of remedy depends on the objective of the analysis and the nature of the data.

14.6 CONCLUSION

Multicollinearity represents a significant challenge in multiple regression analysis because it directly affects the stability, precision, and interpretability of regression coefficients. When explanatory variables are highly correlated, the regression model encounters difficulty in isolating the individual effects of each predictor on the response variable. As a result, estimated coefficients may become unstable, exhibit unexpected signs or magnitudes, and change substantially with small modifications in the data. Although multicollinearity does not necessarily reduce the predictive accuracy or overall goodness of fit of the regression model, it greatly undermines the reliability of individual parameter estimates. Inflated standard errors lead to unreliable hypothesis tests, often causing statistically important variables to appear insignificant. This weakens the confidence in inferential conclusions and complicates decision-making based on the regression results.

Effective regression analysis therefore requires careful diagnosis of multicollinearity using appropriate diagnostic tools such as correlation matrices, variance inflation factors, and condition indices. Once detected, suitable corrective measures—such as variable selection, transformation of variables, or alternative estimation techniques—should be employed. By properly addressing multicollinearity, researchers and analysts can ensure that regression models yield meaningful, stable, and dependable results for interpretation and practical application.

14.7 SELF ASSESSMENT QUESTIONS

- Define multicollinearity and explain its nature.
- List the causes of multicollinearity.
- Discuss the effects of multicollinearity on regression coefficients.
- Explain the Variance Inflation Factor (VIF).
- Suggest remedies for multicollinearity.

14.8 FURTHER READINGS

- Draper, N. R. and Smith, H., *Applied Regression Analysis*, Wiley.
- Montgomery, D. C., Peck, E. A., and Vining, G. G., *Introduction to Linear Regression Analysis*, Wiley.
- Rao, C. R., *Linear Statistical Inference and Its Applications*, Wiley.
- Weisberg, S., *Applied Linear Regression*, Wiley.

Dr. G. Madhu Sudan

LESSON -15

RIDGE REGRESSION AND PRINCIPAL COMPONENT REGRESSION

OBJECTIVES:

By the end of this lesson, students will be able to:

- Understand the limitations of ordinary least squares estimation under multicollinearity.
- Explain the concept and motivation behind ridge regression.
- Derive and interpret the ridge regression estimator.
- Analyze the role and selection of the ridge parameter in regression modeling.
- Describe the concept of principal components and their use in regression analysis.
- Construct a principal component regression model.
- Compare ridge regression and principal component regression in terms of methodology, advantages, and limitations.
- Apply ridge regression and PCR techniques to improve model stability and predictive performance in the presence of multicollinearity.

STRUCTURE

15.1 INTRODUCTION

In multiple linear regression analysis, the method of least squares is widely used to estimate regression coefficients because of its simplicity and desirable statistical properties. Under standard assumptions, least squares estimators are unbiased and have minimum variance

among all linear unbiased estimators. However, these favorable properties rely heavily on the assumption that explanatory variables are not highly correlated with one another. When this assumption is violated, serious practical difficulties arise in the estimation and interpretation of regression coefficients.

A common problem encountered in applied regression analysis is **multicollinearity**, which occurs when two or more explanatory variables exhibit strong linear relationships among themselves. In the presence of multicollinearity, the matrix involved in least squares estimation becomes nearly singular. As a result, the least squares estimators tend to have large variances, making them highly sensitive to small changes in the data. This instability leads to coefficient estimates that may fluctuate widely across different samples and may even possess signs or magnitudes that are inconsistent with theoretical expectations.

Although multicollinearity does not necessarily reduce the overall goodness of fit of the regression model or its predictive ability, it substantially weakens statistical inference. Inflated variances lead to large standard errors, causing important explanatory variables to appear statistically insignificant in hypothesis tests. Consequently, interpretation of individual regression coefficients becomes unreliable, and decision-making based on such results may be misleading.

To overcome these limitations of the least squares method under multicollinearity, several alternative estimation techniques have been developed. Among these, **Ridge Regression** and **Principal Component Regression (PCR)** are two of the most important and widely used approaches. These methods modify the estimation process in different ways to reduce the harmful effects of multicollinearity, with the objective of producing more stable and reliable regression estimates.

Ridge regression addresses the multicollinearity problem by introducing a small amount of bias into the estimation process. This is achieved by adding a penalty term to the least squares objective function, which shrinks the regression coefficients toward zero. Although the resulting estimators are biased, their variances are substantially reduced, often leading to a lower mean squared error compared to ordinary least squares estimators. Ridge regression therefore represents a trade-off between bias and variance, emphasizing stability and predictive accuracy over strict unbiasedness.

Principal Component Regression takes a different approach by transforming the original explanatory variables into a new set of uncorrelated variables called principal components. These components are obtained as linear combinations of the original variables and are ordered according to the amount of variation they explain. By selecting only a subset of principal components for regression, PCR effectively eliminates multicollinearity and reduces dimensionality. While this method improves numerical stability, it may reduce interpretability because the principal components may not have direct physical or practical meaning.

Both ridge regression and principal component regression aim to stabilize regression estimates while retaining good predictive performance. They are particularly valuable in situations where explanatory variables are highly correlated and where reliable estimation of individual regression coefficients is important. The choice between these methods depends on the objectives of the analysis, the importance of interpretability, and the nature of the data.

In summary, when multicollinearity undermines the reliability of least squares estimators, alternative methods such as ridge regression and principal component regression provide effective solutions. By modifying the estimation process, these techniques enhance stability, reduce variance, and support meaningful inference, making them essential tools in advanced regression analysis.

15.2 RIDGE REGRESSION

Ridge regression is a biased estimation technique developed to overcome the difficulties caused by multicollinearity in multiple linear regression models. In ordinary least squares estimation, when explanatory variables are highly correlated, the estimated regression coefficients tend to have large variances and become highly unstable. This instability makes the coefficients sensitive to small changes in the data, leading to unreliable estimation and weak statistical inference. Ridge regression addresses this issue by deliberately introducing a small amount of bias into the estimation process in order to achieve a substantial reduction in variance.

The basic idea behind ridge regression is to modify the least squares estimation procedure so that extreme coefficient values are discouraged. This is done by adding a penalty term to the least squares objective function. The penalty restricts the size of the regression coefficients by shrinking them toward zero. Although this shrinkage introduces bias, it reduces the variability of the estimates, resulting in more stable and reliable coefficient values. This trade-off between bias and variance is central to the motivation of ridge regression.

From a practical perspective, ridge regression is particularly useful in situations where multicollinearity makes ordinary least squares estimates unreliable, even though the overall regression model fits the data well. In such cases, ridge regression improves numerical stability and produces coefficient estimates that are less sensitive to sampling fluctuations. As a result, predictions obtained from ridge regression are often more accurate than those from ordinary least squares, especially in datasets with highly correlated predictors.

Another important feature of ridge regression is that it retains all explanatory variables in the model. Unlike variable selection techniques that remove predictors, ridge regression keeps all variables but controls their influence through shrinkage. This is advantageous when all variables are considered theoretically important and should be included in the model, despite being correlated.

The effectiveness of ridge regression depends on the appropriate selection of the ridge parameter, which determines the strength of the penalty applied to the coefficients. A small value of the ridge parameter introduces little bias and closely resembles ordinary least squares estimation, while a larger value increases shrinkage, reducing variance at the cost of greater bias. Choosing an optimal ridge parameter is essential to balance stability and accuracy.

In summary, ridge regression provides a practical solution to the problem of multicollinearity by introducing bias deliberately to reduce variance. By stabilizing coefficient estimates and improving predictive performance, ridge regression enhances the reliability and usefulness of regression analysis in the presence of highly correlated explanatory variables. It represents an important extension of ordinary least squares estimation and plays a key role in modern regression methodology.

15.2.1 Need for Ridge Regression

When explanatory variables are highly correlated, the matrix involved in least squares estimation becomes nearly singular. This results in large variances of regression coefficients and unstable estimates. Ridge regression was introduced to overcome this problem by shrinking regression coefficients toward zero.
The primary motivation for ridge regression is to:

- Reduce variance of regression coefficients
- Improve numerical stability
- Enhance predictive performance in the presence of multicollinearity

15.2.2 Ridge Regression Estimator

In ridge regression, the least squares objective function is modified by adding a penalty term proportional to the square of the regression coefficients. The ridge estimator is given by:
$\hat{\beta}$ ridge$=(X^{\top}X+kI)^{-1}X^{\top}Y$
where
$K > 0$ is the **ridge parameter**,
I is the identity matrix.
The addition of kI ensures that the matrix is well-conditioned, allowing stable estimation even when multicollinearity is present.

15.2.3 Choice of Ridge Parameter

The value of the ridge parameter k controls the degree of shrinkage applied to the coefficients.

- When k=0, ridge regression reduces to ordinary least squares.
- As k increases, coefficients are increasingly shrunk toward zero.

The optimal value of k is usually chosen using techniques such as:

- Ridge trace plots
- Cross-validation
- Mean squared error minimization

15.2.4 Properties of Ridge Regression

Important properties of ridge regression include:

- Ridge estimators are **biased but have smaller variance**
- Mean squared error may be lower than least squares estimates
- Regression coefficients are more stable under multicollinearity
- Improved prediction accuracy compared to OLS in collinear data

15.3 PRINCIPAL COMPONENT REGRESSION (PCR)

Principal Component Regression (PCR) is a statistical technique developed to overcome the limitations of ordinary least squares regression when multicollinearity is present among explanatory variables. Multicollinearity arises when two or more predictors are strongly correlated, leading to unstable regression coefficient estimates and unreliable statistical

inference. Principal Component Regression addresses this issue by combining **principal component analysis (PCA)** with regression modeling.

The key idea behind PCR is to transform the original set of correlated explanatory variables into a new set of uncorrelated variables known as **principal components**. These components are constructed as linear combinations of the original variables and are ordered according to the amount of variation they explain in the data. The first principal component explains the maximum possible variance, followed by the second principal component, which explains the maximum remaining variance subject to being uncorrelated with the first, and so on.

Once the principal components are obtained, regression is performed using a selected subset of these components rather than the original explanatory variables. By excluding components associated with small eigenvalues, PCR removes directions in the data that contribute little information and are often responsible for multicollinearity. As a result, regression coefficient estimates become more stable, and the effects of correlated predictors are effectively eliminated.

An important advantage of principal component regression is that it improves numerical stability without requiring the removal of original explanatory variables. Instead, PCR replaces the original correlated predictors with a smaller number of uncorrelated components. This leads to a reduction in dimensionality and simplifies the regression problem while retaining most of the important information contained in the data.

Another benefit of PCR is its ability to reduce variance in regression estimates. Although PCR introduces bias by discarding some components, the overall mean squared error of the estimates may be reduced due to the substantial decrease in variance. This bias–variance trade-off is particularly beneficial in situations where multicollinearity is severe and least squares estimation performs poorly.

Despite its advantages, PCR has certain limitations. One major drawback is the loss of interpretability. Since principal components are linear combinations of the original variables, they often lack a clear physical or practical meaning. Consequently, it may be difficult to interpret the relationship between individual explanatory variables and the response variable. Additionally, the principal components are constructed solely based on variability in the explanatory variables and do not take into account the response variable. As a result, components that explain large variance in predictors may not necessarily be the most relevant for predicting the response.

The selection of the number of principal components to include in the regression model is a crucial step in PCR. Choosing too few components may exclude important information, leading to poor predictions, while choosing too many components may reintroduce noise and reduce the benefits of dimensionality reduction. Techniques such as scree plots, cumulative variance criteria, and cross-validation are commonly used to determine the appropriate number of components.

In practical applications, principal component regression is particularly useful when the primary objective is prediction rather than interpretation. It is widely used in fields such as chemometrics, economics, engineering, and bioinformatics, where datasets often contain a large number of highly correlated predictors. In such contexts, PCR provides a reliable and effective alternative to ordinary least squares regression.

In summary, principal component regression is a powerful approach that combines PCA with regression analysis to handle multicollinearity. By transforming correlated explanatory variables into uncorrelated principal components, PCR enhances the stability and reliability of regression estimates. While it may sacrifice interpretability, its ability to improve prediction accuracy and numerical robustness makes it an important tool in advanced regression analysis.

15.3.1 Concept of Principal Components

Principal components are new variables obtained as linear combinations of the original explanatory variables. These components:

- Are mutually uncorrelated
- Capture maximum variance in descending order
- Reduce dimensionality while retaining essential information

15.3.2 Construction of PCR Model

- The PCR procedure involves the following steps:
- Standardize explanatory variables
- Perform principal component analysis
- Select a subset of principal components
- Regress the response variable on selected components
- By excluding components associated with small eigenvalues, PCR eliminates multicollinearity and stabilizes regression estimates.

15.3.3 Advantages and Limitations of PCR

Eliminates Multicollinearity
One of the major advantages of Principal Component Regression is its ability to eliminate multicollinearity among explanatory variables. Since principal components are constructed to be mutually uncorrelated, the problem of strong linear dependence among predictors is completely removed. As a result, regression coefficients obtained from PCR are not affected by instability arising from correlated variables, leading to more reliable estimation.

Reduces Dimensionality
Principal Component Regression effectively reduces the dimensionality of the regression problem by replacing a large set of explanatory variables with a smaller number of principal components. These components retain most of the variability present in the original data while discarding redundant or less informative information. Dimensionality reduction simplifies the regression model, improves computational efficiency, and is particularly useful when dealing with large datasets or many predictors.

Improves Numerical Stability
By removing near-linear dependencies among explanatory variables, PCR improves the numerical stability of the regression estimation process. The transformation of correlated predictors into orthogonal components ensures that matrix inversion required for estimation is well-conditioned. This leads to stable coefficient estimates that are less sensitive to small changes in data and reduces the risk of numerical errors.

Enhances Prediction Accuracy
Although PCR introduces bias by excluding some components, it often reduces the variance of estimates substantially. This trade-off frequently results in lower mean squared error compared to ordinary least squares estimation under multicollinearity. Consequently, PCR often provides better prediction performance, especially when predictors are highly correlated.

Useful for Complex and High-Dimensional Data
PCR is particularly advantageous in applications involving many explanatory variables, such as econometrics, engineering, chemometrics, and bioinformatics. It allows analysts to handle complex datasets efficiently while retaining the most important structural information.

Lack of Interpretability of Principal Components
One of the major limitations of Principal Component Regression is that the principal components used as predictors often lack direct interpretability. Each principal component is a linear combination of several original explanatory variables, making it difficult to associate the regression results with specific predictors. This reduces the usefulness of the model in situations where understanding the individual effect of explanatory variables is important for decision-making or policy analysis.

Possible Exclusion of Important Predictors
In PCR, principal components are selected based on the amount of variation they explain in the explanatory variables, not on their relationship with the response variable. As a result, components that explain relatively little variance in the predictors—but may still be strongly related to the response—can be excluded from the model. This may lead to the omission of important predictive information and reduce model effectiveness.

Need for Careful Selection of Components
The performance of PCR depends critically on the number of principal components included in the regression model. Selecting too few components may result in loss of important information, leading to poor predictions, while including too many components may reintroduce noise and reduce the advantages of dimensionality reduction. Therefore, careful and informed selection of components using appropriate criteria is essential.

Bias in Estimation
Since PCR discards some principal components, the resulting estimators are biased. Although this bias may be acceptable when it leads to a reduction in variance, it must be considered when interpreting regression results, particularly in inferential studies.

Limited Suitability for Interpretive Analysis
Because of the emphasis on variance rather than explanatory power, PCR is more suitable for prediction-focused applications than for models aimed at interpretation of individual variables.

15.4 COMPARISON OF RIDGE REGRESSION AND PCR

Aspect	Ridge Regression	Principal Component Regression
Approach	Penalized regression	Dimension reduction
Use of predictors	Uses all predictors	Uses selected components
Bias	Biased estimation	Biased estimation
Interpretability	Moderate	Low
Handling multicollinearity	Reduces effect	Eliminates effect

15.5 CONCLUSION

Ridge regression and principal component regression provide effective alternatives to ordinary least squares estimation when multicollinearity is present in multiple linear regression models. Multicollinearity undermines the stability and reliability of least squares estimates by inflating variances and producing unstable regression coefficients. Both ridge regression and PCR address this problem, though they adopt different strategies to achieve stability and improved performance.

Ridge regression controls multicollinearity by shrinking regression coefficients toward zero through the introduction of a penalty term in the estimation process. This shrinkage reduces the variance of the estimates at the cost of introducing a small amount of bias. As a result, ridge regression produces more stable and reliable coefficients while retaining all explanatory variables in the model. It is particularly useful when interpretability of predictors is still important and when all variables are theoretically relevant.

Principal component regression, on the other hand, removes multicollinearity by transforming the original correlated explanatory variables into a new set of uncorrelated principal components. Regression is then performed using a selected subset of these components. By eliminating linear dependence among predictors and reducing dimensionality, PCR improves numerical stability and often enhances predictive accuracy. However, the use of principal components may reduce interpretability, as these components are linear combinations of the original variables.

The choice between ridge regression and principal component regression depends on the objectives of the analysis. If the primary goal is prediction and numerical stability in the presence of severe multicollinearity, both methods are suitable. Ridge regression is generally preferred when maintaining the original explanatory variables is important, while PCR is advantageous when dimensionality reduction is desired. Careful consideration of interpretability, predictive performance, and model objectives is essential when selecting the appropriate method.

In summary, ridge regression and principal component regression are powerful tools in advanced regression analysis. By effectively addressing the challenges posed by multicollinearity, they enhance the stability, reliability, and usefulness of regression models in practical applications.

15.6 SELF ASSESSMENT QUESTIONS

- Explain the need for ridge regression.
- Derive the ridge regression estimator.
- What is the role of the ridge parameter?
- Describe the steps involved in principal component regression.
- Compare ridge regression and PCR.

15.7 FURTHER READINGS

- Draper, N. R. and Smith, H., *Applied Regression Analysis*, Wiley.
- Montgomery, D. C., Peck, E. A., and Vining, G. G., *Introduction to Linear Regression Analysis*, Wiley.
- Rao, C. R., *Linear Statistical Inference and Its Applications*, Wiley.
- Weisberg, S., *Applied Linear Regression*, Wiley.

Dr. G. Madhu Sudan

LESSON -16

SUBSET SELECTION OF EXPLANATORY VARIABLES

OBJECTIVES:

By the end of this lesson, students will be able to:

- After completing this lesson, students will be able to:
- Understand the need for variable selection in multiple linear regression models.
- Explain the concept and importance of subset selection of explanatory variables.
- Describe various subset selection methods such as best subset selection, forward selection, backward elimination, and stepwise regression

STRUCTURE :

16.1 INTRODUCTION

In multiple linear regression analysis, it is common to encounter situations where a large number of explanatory variables are available to model a response variable. Advances in data collection and storage have made it easier to gather many potential predictors, but the presence of a large number of variables does not necessarily improve the quality of a regression model. In fact, including all available variables may lead to several statistical and practical difficulties, making the regression model less reliable and harder to interpret.

One important issue arising from the inclusion of many explanatory variables is the introduction of **unnecessary complexity**. A complex model with many predictors may fit the observed data well, but it often lacks interpretability. Understanding the role and contribution of each variable becomes difficult, especially when predictors are correlated or when their effects overlap. Such models may also perform poorly when applied to new data, a phenomenon known as overfitting.

Another major concern is **multicollinearity**, which occurs when explanatory variables are highly correlated with one another. Including several related variables in the same regression model can inflate the variances of regression coefficient estimates, resulting in unstable and unreliable coefficients. This instability weakens statistical inference and makes it difficult to identify which variables are truly influential.

Some explanatory variables may contribute little or no useful information in explaining the response variable. These variables act mainly as noise and can obscure the effects of more important predictors. Including such irrelevant variables increases estimation variance without providing meaningful improvement in model performance. Consequently, hypothesis tests may become less powerful, and confidence intervals may become wider than necessary.

To address these challenges, **subset selection of explanatory variables** is used as an essential tool in multiple regression analysis. Subset selection involves identifying a smaller group of explanatory variables that adequately explains the response variable while excluding redundant or unimportant predictors. The purpose is not merely to reduce the number of variables, but to improve the overall effectiveness, stability, and interpretability of the regression model.

The central objective of subset selection is to achieve an appropriate **balance between model fit and model complexity**. A good regression model should explain the data well, but it should also be as simple as possible. Simpler models are easier to interpret, often more stable, and tend to generalize better to new data. By excluding irrelevant or redundant predictors, subset selection reduces the risk of overfitting and enhances the predictive performance of the model.
Subset selection also plays an important role in improving the **precision of parameter estimation**. When fewer but more relevant explanatory variables are included, regression coefficients tend to have smaller variances and greater stability. This leads to more reliable hypothesis testing and more meaningful confidence intervals. As a result, conclusions drawn from the regression analysis become more trustworthy.

From a practical standpoint, subset selection can also reduce **data collection and computational costs**. In many applications, obtaining measurements for certain variables may be expensive, time-consuming, or difficult. Identifying a subset of important predictors allows analysts to focus resources efficiently while still maintaining acceptable model performance. This is particularly valuable in fields such as economics, engineering, medicine, and environmental studies.

It is important to note that subset selection is not a purely mechanical process. While statistical criteria and automated procedures provide useful guidance, subject-matter knowledge and practical considerations must also be incorporated. A variable that appears statistically insignificant in one dataset may still be important from a theoretical or practical

perspective. Therefore, variable selection should be carried out carefully, combining statistical evidence with expert judgment.

In summary, subset selection of explanatory variables is a crucial step in multiple linear regression analysis when many potential predictors are available. By identifying a smaller and more relevant set of variables, subset selection enhances model interpretability, reduces multicollinearity, improves estimation accuracy, and supports better prediction. The ultimate goal is to construct a regression model that is both statistically sound and practically useful, achieving an effective balance between simplicity and explanatory power.

16.2 NEED FOR VARIABLE SELECTION

The need for variable selection arises due to the following reasons:

- **Improved interpretability**: Models with fewer variables are easier to understand and explain.
- **Reduction of multicollinearity**: Removing redundant variables helps reduce correlation among predictors.
- **Improved estimation accuracy**: Eliminating irrelevant variables reduces variance of estimates.
- **Better prediction performance**: Simpler models often generalize better to new data.
- **Cost and efficiency**: Collecting and processing fewer variables saves time and resources.

Therefore, selecting an appropriate subset of explanatory variables is essential for reliable and meaningful regression analysis.

16.3 SUBSET SELECTION METHODS

Several systematic procedures are available for selecting appropriate subsets of explanatory variables in multiple linear regression. These procedures are designed to identify a set of predictors that provides a good balance between model accuracy and simplicity. Since different explanatory variables may contribute differently to explaining the response variable, subset selection methods offer structured approaches to determine which variables should be included in the regression model.

One widely used approach is **best subset selection**, which involves fitting regression models for all possible combinations of explanatory variables. For each subset size, the best-performing model is selected based on predefined evaluation criteria. Although this method is comprehensive and often yields optimal models, it becomes computationally infeasible when the number of explanatory variables is large.

Another commonly applied procedure is **forward selection**. This method begins with an empty model and sequentially adds explanatory variables that contribute most significantly to improving the model fit. Forward selection is computationally efficient and easy to implement, but it may fail to identify the optimal subset when important combinations of variables are overlooked.

Backward elimination takes the opposite approach by starting with the full model containing all explanatory variables. Variables are then removed one at a time based on statistical insignificance until a satisfactory model is obtained. This method is effective when

the sample size is large enough to support estimation of the full model, but it cannot be applied if the number of variables exceeds the number of observations.

Stepwise regression combines features of both forward selection and backward elimination. At each stage, variables may be added or removed depending on their contribution to model performance. Stepwise regression offers flexibility and adaptability but may yield different results depending on the chosen selection criteria.

These systematic procedures provide practical tools for variable selection, but their results should be interpreted carefully. Selection methods often depend on the data and criteria used, and different procedures may lead to different subsets. Therefore, statistical techniques should be complemented with subject-matter knowledge to ensure meaningful and reliable regression models.

16.3.1 Best Subset Selection

Best subset selection involves fitting regression models for **all possible combinations** of explanatory variables. For each subset size, the best-performing model is identified based on a chosen criterion.

Features:

- Evaluates all possible subsets
- Provides optimal subsets for each size
- Computationally expensive for large numbers of variables
- Best subset selection is ideal when the number of predictors is small.

16.3.2 Forward Selection

Forward selection begins with an empty model containing no explanatory variables. Variables are added one at a time based on their contribution to improving model fit.

Procedure:

- Start with no predictors
- Add the variable that gives the greatest improvement
- Continue until no significant improvement is possible
- This method is computationally efficient but may miss the best overall model.

16.3.3 Backward Elimination

Backward elimination starts with the full model that includes all explanatory variables. Variables are removed one by one based on lack of statistical significance.

Procedure:

- Fit the full model
- Remove the least significant variable
- Continue until all remaining variables are significant
- Backward elimination requires a reasonably large sample size.

16.3.4 Stepwise Regression

Stepwise regression combines features of forward selection and backward elimination.

Characteristics:

- Variables can be added or removed at each step
- Dynamic and flexible approach
- Popular in applied data analysis
- However, results may depend on the chosen significance levels.

16.4 MODEL SELECTION CRITERIA

In multiple linear regression analysis, it is common to obtain several competing models through different subset selection procedures. Each model may differ in the number of explanatory variables included as well as in its goodness of fit. To select the most appropriate model among these alternatives, **model selection criteria** are used. These criteria provide objective measures to compare models and help identify a model that offers an optimal balance between accuracy and simplicity.

One of the most widely used criteria is the **adjusted coefficient of determination**. Unlike the ordinary coefficient of determination, which always increases when more variables are added, the adjusted measure accounts for the number of explanatory variables in the model. It increases only when a newly added variable improves the model more than would be expected by chance. This property makes it useful for comparing models with different numbers of predictors.

Another important criterion is the **Akaike Information Criterion (AIC)**, which evaluates models based on a trade-off between goodness of fit and model complexity. AIC penalizes the inclusion of additional variables and helps prevent overfitting. Models with smaller AIC values are preferred, as they are considered to provide a better balance between fit and complexity.

The **Bayesian Information Criterion (BIC)** is similar in spirit to AIC but imposes a stronger penalty for model complexity, particularly when the sample size is large. As a result, BIC generally favors more parsimonious models. It is especially useful when the primary goal is to identify the most relevant set of explanatory variables rather than to maximize predictive accuracy.

Mallows' Cp is another useful criterion that compares the bias and variance of competing models. It evaluates how well a subset model approximates the full model. Models with Cp values close to the number of explanatory variables are usually considered desirable, as they indicate a good balance between model fit and parsimony.

In summary, model selection criteria play a crucial role in evaluating competing regression models. By considering both goodness of fit and model complexity, these criteria help in selecting models that are not only statistically sound but also simple, stable, and suitable for practical application.

16.4.1 Adjusted Coefficient of Determination

Adjusted R^2 accounts for the number of explanatory variables in the model.The **Adjusted Coefficient of Determination**, denoted asR^2, is an important statistical measure used to evaluate the goodness of fit of a multiple regression model while accounting for the number

of explanatory variables included in the model. It is a modified version of the ordinary coefficient of determination R^2, which measures the proportion of total variation in the response variable explained by the regression model.

Adjusted $R^2 = \frac{1-(RSS/(n-p-1))}{TSS/(n-1)}$

It increases only when a new variable improves the model beyond chance.

16.4.2 Akaike Information Criterion (AIC)

The **Akaike Information Criterion (AIC)** is one of the most widely used statistical measures for comparing and selecting regression models. It is designed to balance **model fit** and **model complexity**, helping to identify a model that explains the data well without including unnecessary explanatory variables. The fundamental idea behind AIC is that a good model should achieve high explanatory power while remaining as simple as possible.

AIC is based on the concept of information loss. When a statistical model is used to represent the true data-generating process, some amount of information is inevitably lost. The AIC provides an estimate of this information loss, and models with smaller AIC values are considered to be closer to the true underlying process.

AIC balances model fit and complexity:
AIC = nln(RSS/n)+2p
where

- n is the sample size,
- RSS is the residual sum of squares, and
- p is the number of estimated parameters in the model.

The first term measures the lack of fit of the model, while the second term introduces a penalty for model complexity. As the number of explanatory variables increases, the penalty term increases, discouraging the inclusion of unnecessary variables.

An important feature of AIC is that it allows **comparison among competing models**, even when they are not nested. The model with the **lowest AIC value** is preferred. However, AIC does not test a model in an absolute sense; it only ranks models relative to one another. AIC is particularly useful in subset selection problems, where many candidate models are available. By penalizing excessive complexity, it helps reduce the risk of overfitting and encourages the selection of models that are more likely to perform well on new data.

In summary, the Akaike Information Criterion is a powerful and practical tool for model selection. By balancing goodness of fit with simplicity, AIC supports the construction of regression models that are both efficient and reliable for inference and prediction. Smaller AIC values indicate better models.

16.4.3 Bayesian Information Criterion (BIC)

The **Bayesian Information Criterion (BIC)** is a widely used statistical criterion for model selection in regression analysis. Like the Akaike Information Criterion (AIC), BIC aims to evaluate competing models by balancing model fit and model complexity. However, BIC imposes a **stronger penalty for model complexity**, especially as the sample size increases, and therefore tends to favor simpler and more parsimonious models.

BIC is derived from Bayesian principles and provides an approximate measure for selecting the most probable model among a set of candidate models, given the observed data.
BIC introduces a heavier penalty for model complexity:
BIC = nln(RSS/n) +p
where

- n is the sample size,
- RSS is the residual sum of squares, and
- p is the number of estimated parameters in the model.

The first term measures how well the model fits the data, while the second term penalizes the inclusion of additional parameters. The penalty term in BIC, pln$f_0$np \ln nplnn, increases more rapidly than the penalty term in AIC as the sample size grows. As a result, BIC places greater emphasis on simplicity.

An important property of BIC is that it tends to select the true model, assuming it is among the candidate models and certain regularity conditions are satisfied, as the sample size becomes large. This property makes BIC especially attractive in problems where the goal is to identify a parsimonious model rather than maximize predictive accuracy.

Like AIC, BIC is used for **relative comparison** of models. The model with the **smallest BIC value** is preferred. Differences in BIC values can be interpreted as evidence in favor of one model over another, with larger differences indicating stronger support.

In summary, the Bayesian Information Criterion is an effective tool for model selection that emphasizes simplicity and interpretability. By imposing a stronger penalty for complexity, BIC helps prevent overfitting and supports the selection of regression models that are stable, efficient, and theoretically sound.

16.4.4 Mallows' Cp

Mallows' Cp is a widely used statistical criterion for model selection in multiple linear regression, particularly in the context of subset selection of explanatory variables. It is designed to assess the trade-off between **bias and variance** in a regression model and to determine how well a subset model approximates the full regression model.

The basic idea behind Mallows' Cp is to evaluate whether a regression model with a selected subset of explanatory variables provides an adequate fit without unnecessary complexity. Unlike criteria that focus only on goodness of fit, Mallows' Cp explicitly accounts for the number of variables included in the model, thereby helping to identify parsimonious models.
Mallows' Cp assesses the trade-off between bias and variance:

$$Cp = \frac{RSSp}{\hat{\sigma}^2} - (n-2p)$$

Models with Cp ≈ p are preferred.
where

- RSSp is the residual sum of squares for the model containing ppp explanatory variables,
- $\hat{\sigma}^2$ is an estimate of the error variance obtained from the full model, and
- n is the sample size.

A desirable property of Mallows' Cp is that it provides guidance on both model adequacy and simplicity. Models with Cp values close to p (the number of explanatory variables) are

generally considered satisfactory. Such models achieve a good balance between bias and variance and are likely to provide reliable estimates.

If the value of Cp is much larger than p, it suggests that the model may be missing important explanatory variables, leading to bias. On the other hand, a very small value of Cp may indicate overfitting, where unnecessary variables have been included in the model.

Mallows' Cp is particularly useful in best subset selection procedures, where many competing models are evaluated simultaneously. By comparing Cp values across models, analysts can identify subsets that approximate the full model closely while using fewer variables.

In summary, Mallows' Cp is an effective criterion for selecting regression models that balance accuracy and simplicity. By considering both the goodness of fit and the number of explanatory variables, it supports the identification of models that are stable, efficient, and well-suited for inference and prediction.

16.5 PRACTICAL ISSUES IN VARIABLE SELECTION

While performing variable selection, the following issues must be considered:

- Risk of **overfitting**
- Instability due to correlated predictors
- Dependence on sample size
- Ignoring domain knowledge
- Different methods may yield different subsets
- Sound judgment and subject-matter expertise are essential.

16.6 CONCLUSION

Subset selection of explanatory variables is a crucial step in multiple regression analysis, particularly when a large number of potential predictors are available. Including all explanatory variables in a regression model often leads to unnecessary complexity, reduced interpretability, and potential statistical issues such as multicollinearity and overfitting. By carefully identifying an appropriate subset of predictors, a regression model can be made more efficient, stable, and meaningful.

Selecting a suitable subset of explanatory variables enhances **interpretability** by focusing attention on the most influential predictors and clarifying their relationship with the response variable. Simpler models are easier to understand and communicate, especially in applied fields where practical interpretation is as important as statistical accuracy. In addition, reducing the number of variables often leads to more stable parameter estimates with smaller variances, thereby improving the reliability of statistical inference.

Subset selection also plays an important role in improving **predictive performance**. Models that avoid unnecessary variables tend to generalize better to new data, as they are less prone to overfitting. By balancing model fit with model simplicity, subset selection methods help construct regression models that perform well both on observed data and in future predictions.

A variety of systematic selection methods—such as best subset selection, forward selection, backward elimination, and stepwise regression—along with model selection criteria like adjusted coefficient of determination, AIC, BIC, and Mallows' Cp, provide valuable guidance in identifying appropriate models. However, no single method is universally optimal, and different approaches may lead to different subsets of variables.

Therefore, thoughtful application of subset selection techniques is essential. Statistical criteria should be used in conjunction with subject-matter knowledge and practical considerations to ensure that selected models are not only statistically sound but also meaningful in real-world contexts. When applied carefully, subset selection contributes significantly to the development of reliable, interpretable, and effective multiple regression models.

16.7 SELF ASSESSMENT QUESTIONS

- Why is variable selection important in multiple regression?
- Explain best subset selection and its limitations.
- Differentiate between forward selection and backward elimination.
- Explain AIC and BIC.
- What is Mallows' Cp?

16.8 FURTHER READINGS

- Draper, N. R. and Smith, H., *Applied Regression Analysis*, Wiley.
- Montgomery, D. C., Peck, E. A., and Vining, G. G., *Introduction to Linear Regression Analysis*, Wiley.
- Rao, C. R., *Linear Statistical Inference and Its Applications*, Wiley.
- Weisberg, S., *Applied Linear Regression*, Wiley.

Dr. G. Madhu Sudan