Theory of Linear Models

Longhai Li

2026-04-25

Table of contents

Preface 7
- Key Features . . . 7
- **Overview** . . . 7
- **Audience** . . . 7
- **Prerequisites** . . . 8

1 Introduction 9
- 1.1 Multiple Linear Regression . . . 9
- 1.2 Examples . . . 9
  - 1.2.1 Polynomial Regression . . . 9
  - 1.2.2 Design Matrix Construction . . . 9
  - 1.2.3 One-Way ANOVA . . . 10
  - 1.2.4 Analysis of Covariance (ANCOVA) . . . 10
- 1.3 Least Squares Estimation . . . 10
- 1.4 Geometric Perspective of Least Square Estimation . . . 11

2 Projection in Vector Space 13
- 2.1 Vector and Projection onto a Line . . . 13
  - 2.1.1 Vectors and Operations . . . 13
  - 2.1.2 Scalar Multiplication and Distance . . . 13
  - 2.1.3 Angle and Inner Product . . . 14
  - 2.1.4 Coordinate (Scalar) Projection . . . 15
  - 2.1.5 Vector Projection Formula . . . 15
  - 2.1.6 Perpendicularity (Orthogonality) . . . 16
  - 2.1.7 Projection onto a Line (Subspace) . . . 16
  - 2.1.8 Projection Matrix (P_x) . . . 17
  - 2.1.9 Pythagorean Theorem . . . 19
  - 2.1.10 Least Square Property . . . 19
- 2.2 Vector Space . . . 20
  - 2.2.1 Spanned Vector Space . . . 21
  - 2.2.2 Column Space and Row Space . . . 21
  - 2.2.3 Linear Independence and Rank . . . 21
- 2.3 Rank of Matrices and Dim of Vector Space . . . 21
  - 2.3.1 Orthogonality to a Subspace . . . 23
  - 2.3.2 Kernel (Null Space) and Image . . . 23
  - 2.3.3 Nullity Theorem . . . 24
  - 2.3.4 Rank Inequalities . . . 25
  - 2.3.5 Rank of $X'X$ and XX' . . . 26

2.4 Orthogonal Projection onto a Subspace . . . 27
2.4.1 Equivalence to Least Squares . . . 27
2.4.2 Uniqueness of Projection . . . 29
2.5 Projection via Orthonormal Basis (Q) . . . 29
2.5.1 Orthonomal Basis . . . 29
2.5.2 Projection Matrix via Orthonomal Basis (Q) . . . 31
2.5.3 Gram-Schmidt Process . . . 34
2.6 Hat Matrix (Projection Matrix via X) . . . 34
2.6.1 Norm Equations . . . 34
2.6.2 Hat Matrix . . . 36
2.6.3 Equivalence of Hat Matrix and QQ' . . . 36
2.6.4 Properties of Hat Matrix . . . 36
2.7 Projection Defined with Orthogonal Projection Matrix . . . 37
2.7.1 Orthogonal Projection Matrix . . . 37
2.7.2 Projection onto Complement Space . . . 39
2.7.3 Projections onto Nested Subspaces . . . 40
2.7.4 Projection onto Three Multually Orthogonal Subspaces . . . 42
2.8 Projections onto More than Three Orthogonal Subspaces . . . 48

3 Matrix Algebra **54**
3.1 Eigenvalues and Eigenvectors . . . 54
3.2 Spectral Theory for Symmetric Matrices . . . 54
3.2.1 Spectral Decomposition . . . 54
3.2.2 Quadratic Form . . . 55
3.2.3 Positive and Non-Negative Definite Matrices . . . 55
3.2.4 Properties of Symmetric Matrices . . . 57
3.2.5 Spectral Representation of Projection Matrices . . . 57
3.3 Singular Value Decomposition (SVD) . . . 58
3.4 Cholesky Decomposition . . . 60
3.4.1 Matrix Representation of the Algorithm . . . 60
3.4.2 Applications in Statistics . . . 61

4 Multivariate Normal Distribution **63**
4.1 Motivation . . . 63
4.2 Random Vectors and Matrices . . . 63
4.2.1 Derivation of Covariance Matrix Structure . . . 64
4.3 Properties of Mean and Variance . . . 65
4.4 The Multivariate Normal Distribution . . . 66
4.4.1 Definition and Density . . . 66
4.4.2 Geometric Interpretation . . . 66
4.4.3 Probability Density Function . . . 66
4.4.4 Moment Generating Function . . . 67
4.5 Construction and Linear Transformations . . . 67
4.5.1 Important Corollaries of Theorem 4.2 . . . 68
4.6 Independence . . . 68

4.7 Signal-Noise Decomposition for Multivariate Normal Distribution . . . 69
4.7.1 Connections with Other Formulas . . . 70
4.8 Partial and Multiple Correlation . . . 72
4.9 Examples . . . 72

5 Distribution of Quadratic Forms **76**
5.1 Quadratic Forms . . . 76
5.2 Mean of Quadratic Forms . . . 76
5.3 Non-central χ^2 Distribution . . . 81
5.3.1 Visualizing χ^2 Distributions . . . 82
5.3.2 Mean, Variance, and MGF . . . 82
5.3.3 Additivity . . . 85
5.3.4 Poisson Mixture Representation . . . 86
5.4 Distribution of Quadratic Forms . . . 87
5.4.1 MGF of Quadratic Forms . . . 87
5.4.2 Distribution of the Sum Squares of Projected Spherical Normal . . . 87
5.4.3 Distribution of General Quadratic Forms . . . 90
5.4.4 Standardized Distance Distribution . . . 91
5.5 Distributions of Projections of Spherical Normal . . . 92
5.5.1 Independence of Forms . . . 92
5.5.2 Cochran's Theorem . . . 93
5.6 Non-central Distributions Derived from Non-central χ^2 . . . 93
5.6.1 The Non-central F-distribution $F(\mathrm{df}_1, \mathrm{df}_2, \lambda)$ . . . 93
5.6.2 Type I Non-central Beta $\mathrm{Beta}_1(\mathrm{df}_1/2, \mathrm{df}_2/2, \lambda)$ . . . 94
5.6.3 Type II Non-central Beta $\mathrm{Beta}_2(\mathrm{df}_2/2, \mathrm{df}_1/2, \lambda)$ . . . 95
5.6.4 Scaled Type II Beta Scaled-$\mathrm{Beta}_2(\mathrm{df}_2/2, \mathrm{df}_1/2, \lambda)$ . . . 95
5.6.5 The Non-central t-distribution $t(\mathrm{df}_2, \delta)$ . . . 96
5.7 Example: Inference of the Mean of Normal Sample . . . 96
5.7.1 Sum of Squares and Their Distributions . . . 98
5.7.2 Distributions of Equivalent Statistics . . . 98
5.7.3 Expectations Under M_1 and M_0 . . . 99

6 Inference for A Multiple Linear Regression Model **100**
6.1 Linear Models and Least Square Estimator . . . 100
6.1.1 Assumptions in Linear Models . . . 100
6.1.2 Matrix Formulation . . . 100
6.1.3 Least Squares Estimator of β and Fitted Value $\hat{Y}$ . . . 101
6.1.4 Properties of the Estimator $\hat{\beta}$ . . . 102
6.2 Best Linear Unbiased Estimator (BLUE) . . . 103
6.2.1 Notes on Gauss-markov . . . 104
6.2.2 Limitations: Restriction to Unbiased Estimators . . . 105
6.3 Unbiased Estimator of Error Variance . . . 105
6.3.1 Unbiasedness of s^2 . . . 106
6.4 Distributions Under Normality . . . 106
6.5 Maximum Likelihood Estimator (MLE) . . . 107
6.6 Linear Models in Centered Form . . . 108

6.7 Least Squares Estimates for Linear Models in Centered Form . . . 109
6.8 Decomposition of Sum of Squares and their Distributions . . . 111
6.8.1 3D Visualization of Decomposition of y . . . 112
6.8.2 A Diagram to Show Decomposition of Sum of Squares . . . 114
6.8.3 Distribution of Sum of Squares . . . 114
6.9 F-test for Testing Overall Regression Effect . . . 116
The F-statistic . . . 117
Understanding F via Expectations . . . 117
6.9.1 Distributional Theory . . . 117
6.9.2 Visualization of the Rejection Region . . . 118
6.10 Optimistic Bias in Raw Coefficient of Determination (R^2) . . . 118
6.11 Adjusted R-squared (R_a^2) and Population Proportion (ρ^2) . . . 121
6.11.1 Definition . . . 121
6.11.2 Expectation in terms of λ . . . 122
6.11.3 Definitions of Population Metrics for Predictivity . . . 122
6.11.4 Remarks on Variance Estimation and Effect Size . . . 123
6.11.5 Confidence Interval of Population ρ^2 . . . 124
6.12 An Animation for Illustrating R_a^2 Under H_0 and H_1 . . . 126
6.13 A Data Example with House Price Valuation . . . 127
6.13.1 Visualize the Data . . . 127
6.13.2 Fit the Model . . . 128
6.13.3 Visualization of Fitted Values vs Mean . . . 129
6.13.4 Computing Sums of Squares (SSE, SST, SSR) . . . 130
6.13.5 Analysis of Variance (ANOVA) . . . 131
6.13.6 Coefficient of Determination and Variance Decomposition . . . 133
6.13.7 Confidence Interval for Population R^2 (ρ^2) . . . 133
6.14 Underfitting and Overfitting . . . 134
6.14.1 Notation and Setup . . . 134
6.14.2 Case 1: Underfitting . . . 135
6.14.3 Case 2: Overfitting . . . 136

7 Hypothesis Testing in Linear Models 138
7.1 Testing Reduced Model vs Full Model (Partial F-test) . . . 138
7.1.1 Geometric Interpretation . . . 139
7.1.2 Distributional Properties . . . 139
7.1.3 The F-Test . . . 140
7.1.4 Overall Regression Test . . . 140
7.2 The General Linear Hypothesis . . . 141
7.2.1 Test Statistic for $C\beta = 0$ . . . 141
7.2.2 Nested Models Interpretation . . . 141
7.2.3 F-Test for Non-Zero General Linear Hypothesis . . . 143
7.2.4 Numerical Examples . . . 144
7.3 Speficic Tests for Linear Combinations of β . . . 151
7.3.1 Numerical Examples . . . 152

8 Generalized Inverses **157**
8.1 Motivation . . . 157
8.2 Definition of Generalized Inverse . . . 157
8.3 A Procedure to Find a Generalized Inverse . . . 159
8.4 Moore-Penrose Inverse . . . 160
8.5 Solving Linear Systems with Generalized Inverse . . . 160
8.6 Least Squares for Non-full-rank X with Generalized Inverse . . . 162
8.6.1 Projection Matrix with Generalized Inverse of $X'X$ . . . 162
8.6.2 Invariance and Uniqueness of "the" Projection Matrix . . . 162
Example: Projection with Linearly Dependent Columns and Generalized Inverses . . . 163
1. Define the Vectors . . . 164
2. Design Matrix and X^TX . . . 164
3. Compute Two Generalized Inverses . . . 164
4. The Projection Matrix P is Invariant . . . 165
5. Comparison to Projecting onto x_1 Individually . . . 165
6. R Implementation Verification . . . 166
8.7 The Left Inverse View: Recovering $\hat{\beta}$ from $\hat{y}$ . . . 166
8.7.1 The Generalized Left Inverse . . . 166
8.7.2 Verification of the Inverse Property . . . 167
8.7.3 Recovering the Estimator . . . 167
8.8 Non-full-rank Least Squares with QR Decomposition . . . 168
8.8.1 Constructing a Solution by Solving Normal Equations . . . 168
8.8.2 Constructing a Solution by Solving Reparametrized β . . . 169

9 Estimation and Inference with Non-full-rank Models **170**
9.1 Non-full-rank Models and Parameter Non-identifiability . . . 170
9.1.1 One-way ANOVA Model . . . 170
9.2 Least Square Estimation of β and $\mu = X\beta$ . . . 172
9.2.1 Distribution of $\hat{\beta}$ and s^2 . . . 172
9.3 Sum Squares and F-test . . . 173
9.3.1 Examples . . . 173
9.4 Inference for Estimable Parameters . . . 186
9.4.1 Estimability . . . 186
9.4.2 Example: Checking Estimability in One-Way ANOVA . . . 187
9.4.3 Properties of Estimators for Estimabile Parameters . . . 189
9.4.4 Properties of Estimators for Estimabile Parameters . . . 190
9.4.5 Testable Hypotheses in Non-Full Rank Models . . . 190
9.4.6 The General Linear Hypothesis F-Test . . . 190

Preface

Key Features

This text adopts a geometric approach to the statistical theory of linear models, aiming to provide a deeper understanding than standard algebraic treatments. Key features include:

- **Projection Perspective:** We prioritize the geometric interpretation of least squares, viewing estimation as a projection of the response vector onto a model subspace. This visual framework unifies diverse topics—from simple regression to complex ANOVA designs—under a single theoretical umbrella.
- **Interactive Visualizations:** Abstract concepts are brought to life through interactive 3D plots. Readers can rotate and inspect vector spaces, residual planes, and projection geometries to build a tangible intuition for high-dimensional operations.
- **Computational Integration:** Theory is seamlessly integrated with practice. The text provides implementation examples using R (and Python), demonstrating how theoretical matrix equations translate directly into computational code.
- **Rigorous Foundations:** While visually driven, the text maintains mathematical rigor, covering essential topics such as spectral theory, the generalized inverseand the multivariate normal distribution to ensure a solid theoretical grounding.

Overview

This course is a rigorous examination of the general linear models using vector space theory, in particular the approach of regarding least square as projection. The topics includes: vector space; projection; matrix algebra; generalized inverses; quadratic forms; theory for point estimation; theory for hypothesis test; theory for non-full-rank models.

Audience

This book is designed for graduate students and advanced undergraduate students in statistics, data science, and related quantitative fields. It serves as a bridge between applied regression analysis and the theoretical foundations of linear models. Researchers and practitioners seeking a deeper geometric and algebraic understanding of the statistical methods they use daily will also find this text valuable.

Prerequisites

To get the most out of this book, readers should have a comfortable grasp of the following topics:

Linear Algebra: An elementary understanding of matrix operations is essential. You should be familiar with matrix multiplication, determinants, inversion, and the basic concepts of vector spaces (such as linear independence, basis vectors, and subspaces). While we review key spectral theory concepts (like eigenvalues and the singular value decomposition) in the early chapters, prior exposure to these ideas is helpful.

Probability and Statistics: A standard introductory course in probability and mathematical statistics is required. Readers should be familiar with random variables, expectation, variance, covariance, common probability distributions (especially the Normal distribution), and fundamental concepts of hypothesis testing and estimation.

1 Introduction

1.1 Multiple Linear Regression

Suppose we have observations on Y and X_j. The data can be represented in matrix form.

$$\underset{n\times 1}{y} = \underset{n\times p}{X}\beta + \underset{n\times 1}{\epsilon} \tag{1.1}$$

where the error terms are distributed as:

$$\epsilon \sim N_n(0, \sigma^2 I_n), \tag{1.2}$$

in which I_n is the identity matrix:

$$I_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{1.3}$$

The scalar equation for a single observation is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i \tag{1.4}$$

1.2 Examples

1.2.1 Polynomial Regression

Polynomial regression fits a curved line to the data points but remains linear in the parameters (β).

The model equation is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{p-1} x_i^{p-1} \tag{1.5}$$

1.2.2 Design Matrix Construction

The design matrix X is constructed by taking powers of the input variable.

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^{p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{1.6}$$

1.2.3 One-Way ANOVA

ANOVA can be expressed as a linear model using categorical predictors (dummy variables).

Suppose we have 3 groups (G_1, G_2, G_3) with observations:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{1.7}$$

$$\begin{array}{ccc} G_1 & G_2 & G_3 \\ \boxed{\begin{matrix} Y_{11} \\ Y_{12} \end{matrix}} & \boxed{\begin{matrix} Y_{21} \\ Y_{22} \end{matrix}} & \boxed{\begin{matrix} Y_{31} \\ Y_{32} \end{matrix}} \end{array} \tag{1.8}$$

We construct the matrix X to select the group mean (μ) corresponding to the observation:

$$\underset{6\times 1}{y} = \underset{6\times 3}{X} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \epsilon \tag{1.9}$$

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \epsilon \tag{1.10}$$

1.2.4 Analysis of Covariance (ANCOVA)

ANCOVA combines continuous variables and categorical (dummy) variables in the same design matrix.

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{1,\text{cont}} & 1 & 0 \\ X_{2,\text{cont}} & 1 & 0 \\ \vdots & 0 & 1 \\ X_{n,\text{cont}} & 0 & 1 \end{bmatrix} \beta + \epsilon \tag{1.11}$$

1.3 Least Squares Estimation

For the general linear model $y = X\beta + \epsilon$, the Least Squares estimator is:

$$\hat{\beta} = (X'X)^{-1}X'y \tag{1.12}$$

The predicted values ($\hat{y}$) are obtained via the Projection Matrix (Hat Matrix) P_X:

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = P_X y \tag{1.13}$$

The residuals and Sum of Squared Errors are:

$$\hat{e} = y - \hat{y} \tag{1.14}$$

$$\text{SSE} = ||\hat{e}||^2 \tag{1.15}$$

The coefficient of determination is:

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} \tag{1.16}$$

where $\text{SST} = \sum(y_i - \bar{y})^2$.

1.4 Geometric Perspective of Least Square Estimation

We align the coordinate system to the models for clarity:

1. **Reduced Model (**M_0): Represented by the **X-axis** (labeled j_3).
    - $\hat{y}_0$ is the projection of y onto this axis.
2. **Full Model (**M_1): Represented by the **XY-plane** (the floor).
    - $\hat{y}_1$ is the projection of y onto this plane ($z = 0$).
3. **Observed Data (**y): A point in 3D space.

The "improvement" due to adding predictors is the distance between $\hat{y}_0$ and $\hat{y}_1$.

The geometric perspective is not merely for intuition, but as the most robust framework for mastering linear models. This approach offers three distinct advantages:

- **Statistical Clarity:** Geometry provides the most natural path to understanding the properties of estimators. By viewing least square estimation as an orthogonal projection, the decomposition of sums of squares into independent components becomes visually obvious, demystifying how degrees of freedom relate to subspace dimensions rather than abstract algebraic constants. The sampling distribution of the sum squares become straightforward.
- **Computational Stability:** A geometric understanding is essential for implementing efficient and numerically stable algorithms. While the algebraic "Normal Equations" ($(X'X)^{-1}X'y$) are theoretically valid, they are often computationally hazardous. The geometric approach leads directly to superior methods—such as QR and Singular Value Decompositions—that are the backbone of modern statistical software.
- **Generalizability:** The principles of projection and orthogonality extend far beyond the Gaussian linear model. These geometric insights provide the foundational intuition needed for tackling non-Gaussian optimization problems, including Generalized Linear Models (GLMs) and convex optimization, where solutions can often be viewed as projections onto convex sets.

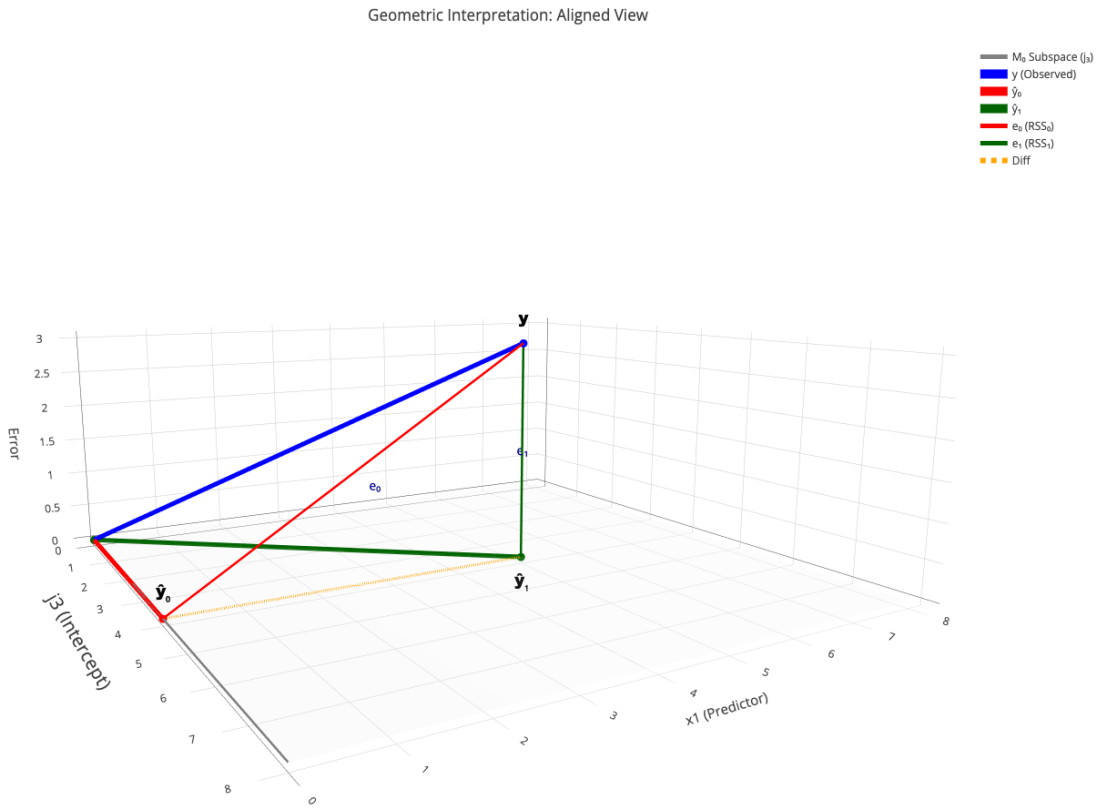


Figure 1.1: Geometric Interpretation: Projection onto Axis (M0) vs Plane (M1)

2 Projection in Vector Space

2.1 Vector and Projection onto a Line

2.1.1 Vectors and Operations

The concept of a vector is fundamental to linear algebra and linear models. We begin by formally defining what a vector is in the context of Euclidean space.

Definition 2.1 (Vector)**.** A **vector** x is defined as a point in n-dimensional space ($\mathbb{R}^n$). It is typically represented as a column vector containing n real-valued components:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{2.1}$$

Vectors are not just static points; they can be combined and manipulated. The two most basic geometric operations are addition and subtraction.

Vector Arithmetic: Vectors can be manipulated geometrically:

Definition 2.2 (Vector Addition)**.** The sum of two vectors x and y creates a new vector. The operation is performed component-wise, adding corresponding elements from each vector. Geometrically, this follows the "parallelogram rule" or the "head-to-tail" method, where you place the tail of y at the head of x.

$$x + y = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix} \tag{2.2}$$

Definition 2.3 (Vector Subtraction)**.** The difference $d = y - x$ is the vector that "closes the triangle" formed by x and y. It represents the displacement vector that connects the tip of x to the tip of y, such that $x + d = y$.

2.1.2 Scalar Multiplication and Distance

In addition to combining vectors with each other, we can modify a single vector using a real number, known as a scalar.

Definition 2.4 (Scalar Multiplication)**.** Multiplying a vector by a scalar c scales its magnitude (length) without changing its line of direction. If c is positive, the direction remains the same; if c is negative, the direction is reversed.

$$cx = \begin{pmatrix} cx_1 \\ \vdots \\ cx_n \end{pmatrix} \tag{2.3}$$

We often need to quantify the "size" of a vector. This is done using the concept of length, or norm.

Definition 2.5 (Euclidean Distance (Length))**.** The length (or norm) of a vector $x = (x_1, \dots, x_n)^T$ corresponds to the straight-line distance from the origin to the point defined by x. It is defined as the square root of the sum of squared components:

$$||x||^2 = \sum_{i=1}^{n} x_i^2 \tag{2.4}$$

$$||x|| = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{2.5}$$

2.1.3 Angle and Inner Product

To understand the relationship between two vectors x and y beyond just their lengths, we must look at the angle between them. Consider the triangle formed by the vectors x, y, and their difference $y - x$. By applying the classic **Law of Cosines** to this triangle, we can relate the geometric angle to the vector lengths.

Theorem 2.1 (Law of Cosines)**.** *For a triangle with sides a, b, c and angle θ opposite to side c:*

$$c^2 = a^2 + b^2 - 2ab\cos\theta \tag{2.6}$$

Translating this geometric theorem into vector notation where the side lengths correspond to the norms of the vectors, we get:

$$||y - x||^2 = ||x||^2 + ||y||^2 - 2||x|| \cdot ||y|| \cos\theta \tag{2.7}$$

This equation provides a critical link between the geometric angle θ and the algebraic norms of the vectors.

Derivation of Inner Product

We can express the squared distance term $||y - x||^2$ purely algebraically by expanding the components:

$$||y - x||^2 = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{2.8}$$

$$= \sum_{i=1}^{n} (x_i^2 + y_i^2 - 2x_i y_i) \tag{2.9}$$

$$= ||x||^2 + ||y||^2 - 2\sum_{i=1}^{n} x_i y_i \tag{2.10}$$

By comparing this expanded form with the result from the Law of Cosines derived previously, we can identify a corresponding interaction term. This term is so important that we give it a special name: the **Inner Product** (or dot product).

Definition 2.6 (Inner Product)**.** The inner product of two vectors x and y is defined as the sum of the products of their corresponding components:

$$x'y = \sum_{i=1}^{n} x_i y_i = \langle x, y \rangle \tag{2.11}$$

Thus, equating the geometric and algebraic forms yields the fundamental relationship:

$$x'y = ||x|| \cdot ||y|| \cos\theta \tag{2.12}$$

2.1.4 Coordinate (Scalar) Projection

The inner product allows us to calculate projections, which quantify how much of one vector "lies along" another. If we rearrange the cosine formula derived above, we can isolate the term that represents the length of the "shadow" cast by vector y onto vector x.

The length of this projection is given by:

$$||y|| \cos\theta = \frac{x'y}{||x||} \tag{2.13}$$

This expression can be interpreted as the inner product of y with the normalized (unit) vector in the direction of x:

$$\text{Scalar Projection} = \left\langle \frac{x}{||x||}, y \right\rangle \tag{2.14}$$

2.1.5 Vector Projection Formula

The scalar projection only gives us a magnitude (a number). To define the projection as a vector in the same space, we need to multiply this scalar magnitude by the direction of the vector we are projecting onto.

Definition 2.7 (Vector Projection)**.** The projection of vector y onto vector x, denoted $\hat{y}$, is calculated as:

$$\text{Projection Vector} = (\text{Length}) \cdot (\text{Direction}) \tag{2.15}$$

$$\hat{y} = \left(\frac{x'y}{||x||} \right) \cdot \frac{x}{||x||} \tag{2.16}$$

This is often written compactly by combining the denominators:

$$\hat{y} = \frac{x'y}{||x||^2}x \tag{2.17}$$

2.1.6 Perpendicularity (Orthogonality)

A special case of the angle between vectors arises when $\theta = 90°$. This geometric concept of perpendicularity is central to the theory of projections and least squares.

Definition 2.8 (Perpendicularity)**.** Two vectors are defined as **perpendicular** (or orthogonal) if the angle between them is $90°$ ($\pi/2$).
Since $\cos(90°) = 0$, the condition for orthogonality simplifies to the inner product being zero:

$$x'y = 0 \iff x \perp y \tag{2.18}$$

Example 2.1 (Orthogonal Vectors)**.** Consider two vectors in $\mathbb{R}^2$: $x = (1, 1)'$ and $y = (1, -1)'$.

$$x'y = 1(1) + 1(-1) = 1 - 1 = 0 \tag{2.19}$$

Since their inner product is zero, these vectors are orthogonal to each other.

2.1.7 Projection onto a Line (Subspace)

We can generalize the concept of projecting onto a single vector to projecting onto the entire line (a 1-dimensional subspace) defined by that vector.

Definition 2.9 (Line Spanned by a Vector)**.** The line space $L(x)$, or the space spanned by a vector x, is defined as the set of all scalar multiples of x:

$$L(x) = \{cx \mid c \in \mathbb{R}\} \tag{2.20}$$

The projection of y onto $L(x)$, denoted $\hat{y}$, is defined by the geometric property that it is the closest point on the line to y. This implies that the error vector (or residual) must be perpendicular to the line itself.

Definition 2.10 (Projection onto a Line)**.** A vector $\hat{y}$ is the projection of y onto the line $L(x)$ if:

1. $\hat{y}$ lies on the line $L(x)$ (i.e., $\hat{y} = cx$ for some scalar c).
2. The residual vector $(y - \hat{y})$ is perpendicular to the direction vector x.

Derivation: To find the value of the scalar c, we apply the orthogonality condition:

$$(y - \hat{y}) \perp x \implies x'(y - cx) = 0 \tag{2.21}$$

Expanding this inner product gives:

$$x'y - c(x'x) = 0 \tag{2.22}$$

Solving for c, we obtain:

$$c = \frac{x'y}{||x||^2} \tag{2.23}$$

This confirms the formula derived previously using the inner product geometry. It shows that the least squares principle (shortest distance) leads to the same result as the geometric projection.

Alternative Forms of the Projection Formula

We can express the projection vector $\hat{y}$ in several equivalent ways to highlight different geometric interpretations.

Definition 2.11 (Forms of Projection)**.** The projection of y onto the vector x is given by:

$$\hat{y} = \frac{x'y}{||x||^2}x = \left\langle y, \frac{x}{||x||} \right\rangle \frac{x}{||x||} \tag{2.24}$$

This second form separates the components into:

$$\text{Projection} = (\text{Scalar Projection}) \times (\text{Unit Direction}) \tag{2.25}$$

2.1.8 Projection Matrix (P_x)

In linear models, it is often more convenient to view projection as a linear transformation applied to the vector y. This allows us to define a **Projection Matrix**.

We can rewrite the formula for $\hat{y}$ by factoring out y:

$$\hat{y} = \text{proj}(y|x) = x\frac{x'y}{||x||^2} = \frac{xx'}{||x||^2}y \tag{2.26}$$

This leads to the definition of the projection matrix P_x.

Definition 2.12 (Projection Matrix onto a Single Vector)**.** The matrix P_x that projects any vector y onto the line spanned by x is defined as:

$$P_x = \frac{xx'}{||x||^2} \tag{2.27}$$

Using this matrix, the projection is simply:

$$\hat{y} = P_x y \tag{2.28}$$

If $x \in \mathbb{R}^n$, then P_x is a $n \times n$ symmetric matrix.

Let's apply these concepts to a concrete example.

Example 2.2 (Numerical Projection)**.** Let $y = (1, 3)'$ and $x = (1, 1)'$. We want to find the projection of y onto x.

Method 1: Using the Vector Formula First, calculate the inner products:

$$x'y = 1(1) + 1(3) = 4 \tag{2.29}$$

$$||x||^2 = 1^2 + 1^2 = 2 \tag{2.30}$$

Now, apply the formula:

$$\hat{y} = \frac{4}{2}\begin{pmatrix}1\\1\end{pmatrix} = 2\begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}2\\2\end{pmatrix} \tag{2.31}$$

Method 2: Using the Projection Matrix Construct the matrix P_x:

$$P_x = \frac{1}{2}\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1 & 1\end{pmatrix} = \frac{1}{2}\begin{pmatrix}1 & 1\\1 & 1\end{pmatrix} = \begin{pmatrix}0.5 & 0.5\\0.5 & 0.5\end{pmatrix} \tag{2.32}$$

Multiply by y:

$$\hat{y} = P_x y = \begin{pmatrix}0.5 & 0.5\\0.5 & 0.5\end{pmatrix}\begin{pmatrix}1\\3\end{pmatrix} = \begin{pmatrix}0.5(1) + 0.5(3)\\0.5(1) + 0.5(3)\end{pmatrix} = \begin{pmatrix}2\\2\end{pmatrix} \tag{2.33}$$

Example: Projection onto the Ones Vector (j_n)

A very common operation in statistics is calculating the sample mean. This can be viewed geometrically as a projection onto a specific vector.

Example 2.3 (Projection onto the Ones Vector)**.** Let $y = (y_1, \dots, y_n)'$ be a data vector. Let $j_n = (1, 1, \dots, 1)'$ be a vector of all ones.
The projection of y onto j_n is:

$$\text{proj}(y|j_n) = \frac{j_n' y}{||j_n||^2} j_n \tag{2.34}$$

Calculating the components:

$$j_n' y = \sum_{i=1}^{n} y_i \quad \text{(Sum of observations)} \tag{2.35}$$

$$||j_n||^2 = \sum_{i=1}^{n} 1^2 = n \tag{2.36}$$

Substituting these back:

$$\hat{y} = \frac{\sum y_i}{n} j_n = \bar{y} j_n = \begin{pmatrix}\bar{y}\\\vdots\\\bar{y}\end{pmatrix} \tag{2.37}$$

Thus, replacing a data vector with its mean vector is geometrically equivalent to projecting the data onto the line spanned by the vector of ones.

2.1.9 Pythagorean Theorem

The Pythagorean theorem generalizes from simple geometry to vector spaces using the concept of orthogonality defined by the inner product.

Theorem 2.2 (Pythagorean Theorem)**.** *If two vectors x and y are orthogonal (i.e., $x \perp y$ or $x'y = 0$), then the squared length of their sum is equal to the sum of their squared lengths:*

$$||x + y||^2 = ||x||^2 + ||y||^2 \tag{2.38}$$

Proof. We expand the squared norm using the inner product:

$$\begin{aligned} ||x + y||^2 &= (x + y)'(x + y) \\ &= x'x + x'y + y'x + y'y \\ &= ||x||^2 + 2x'y + ||y||^2 \end{aligned} \tag{2.39}$$

Since $x \perp y$, the inner product $x'y = 0$. Thus, the term $2x'y$ vanishes, leaving:

$$||x + y||^2 = ||x||^2 + ||y||^2 \tag{2.40}$$

□

The proof after defining inner product to represent $\cos(\theta)$ is trivival. Figure 2.1 shows a geometric proof of the fundamental Pythagorean Theorem.

2.1.10 Least Square Property

One of the most important properties of the orthogonal projection is that it minimizes the distance between the vector y and the subspace (or line) onto which it is projected.

Theorem 2.3 (Least Square Property)**.** *Let $\hat{y}$ be the projection of y onto the line $L(x)$. For any other vector y^* on the line $L(x)$, the distance from y to y^* is always greater than or equal to the distance from y to $\hat{y}$.*

$$||y - y^*|| \geq ||y - \hat{y}|| \tag{2.41}$$

Proof. Since both $\hat{y}$ and y^* lie on the line $L(x)$, their difference $(\hat{y} - y^*)$ also lies on $L(x)$. From the definition of projection, the residual $(y - \hat{y})$ is orthogonal to the line $L(x)$. Therefore:

$$(y - \hat{y}) \perp (\hat{y} - y^*) \tag{2.42}$$

We can write the vector $(y - y^*)$ as:

$$y - y^* = (y - \hat{y}) + (\hat{y} - y^*) \tag{2.43}$$

Applying the Pythagorean Theorem:

$$||y - y^*||^2 = ||y - \hat{y}||^2 + ||\hat{y} - y^*||^2 \tag{2.44}$$

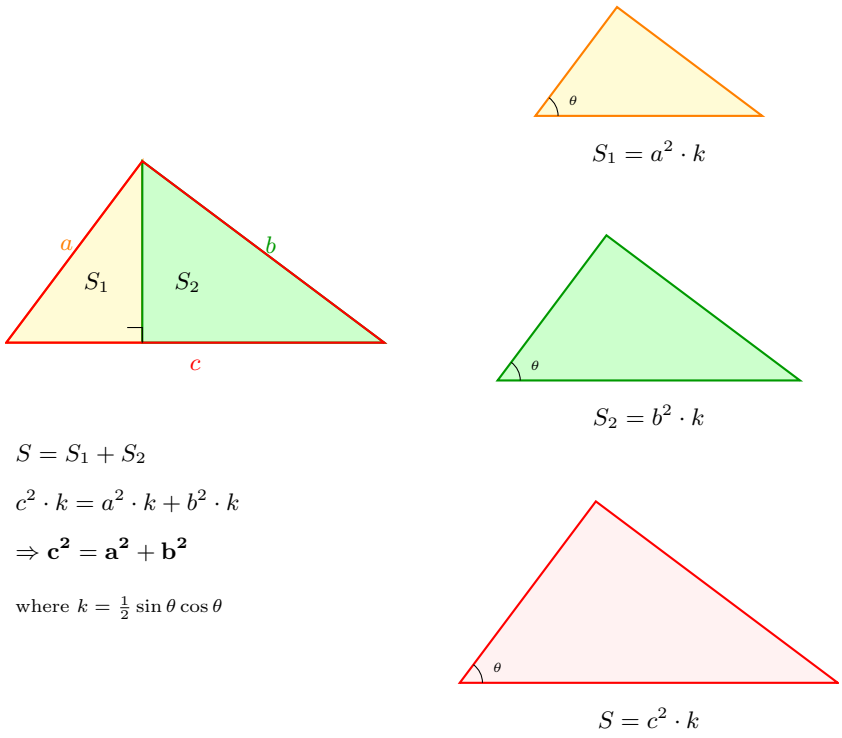


Figure 2.1: Proof of Pythagorean Theorem using Area Scaling

Since $||\hat{y} - y^*||^2 \geq 0$, it follows that:

$$||y - y^*||^2 \geq ||y - \hat{y}||^2 \tag{2.45}$$

□

2.2 Vector Space

We now generalize our discussion from lines to broader spaces.

Definition 2.13 (Vector Space)**.** A set $V \subseteq \mathbb{R}^n$ is called a **Vector Space** if it is closed under vector addition and scalar multiplication:

1. **Closed under Addition:** If $x_1 \in V$ and $x_2 \in V$, then $x_1 + x_2 \in V$.
2. **Closed under Scalar Multiplication:** If $x \in V$, then $cx \in V$ for any scalar $c \in \mathbb{R}$.

It follows that the zero vector 0 must belong to any subspace (by choosing $c = 0$).

2.2.1 Spanned Vector Space

The most common way to construct a vector space in linear models is by spanning it with a set of vectors.

Definition 2.14 (Spanned Vector Space)**.** Let $x_1, \dots, x_p$ be a set of vectors in $\mathbb{R}^n$. The space spanned by these vectors, denoted $L(x_1, \dots, x_p)$, is the set of all possible linear combinations of them:

$$L(x_1, \dots, x_p) = \{r \mid r = c_1 x_1 + \cdots + c_p x_p, \text{ for } c_i \in \mathbb{R}\} \tag{2.46}$$

2.2.2 Column Space and Row Space

When vectors are arranged into a matrix, we define specific spaces based on their columns and rows.

Definition 2.15 (Column Space)**.** For a matrix $X = (x_1, \dots, x_p)$, the **Column Space**, denoted $\text{Col}(X)$, is the vector space spanned by its columns:

$$\text{Col}(X) = L(x_1, \dots, x_p) \tag{2.47}$$

Definition 2.16 (Row Space)**.** The **Row Space**, denoted $\text{Row}(X)$, is the vector space spanned by the rows of the matrix X.

2.2.3 Linear Independence and Rank

Not all vectors in a spanning set contribute new dimensions to the space. This concept is captured by linear independence.

Definition 2.17 (Linear Independence)**.** A set of vectors $x_1, \dots, x_p$ is said to be **Linearly Independent** if the only solution to the linear combination equation equal to zero is the trivial solution:

$$\sum_{i=1}^{p} c_i x_i = 0 \implies c_1 = c_2 = \cdots = c_p = 0 \tag{2.48}$$

If there exist non-zero c_i's such that sum is zero, the vectors are **Linearly Dependent**.

2.3 Rank of Matrices and Dim of Vector Space

Definition 2.18 (Rank)**.** The **Rank** of a matrix X, denoted $\text{Rank}(X)$, is the maximum number of linearly independent columns in X. This is equivalent to the dimension of the column space:

$$\text{Rank}(X) = \text{Dim}(\text{Col}(X)) \tag{2.49}$$

There are several fundamental properties regarding the rank of a matrix.

Example 2.4 (Example of the Equality of Row and Col Rank)**.** Consider the following 3×4 matrix ($n = 3, p = 4$):

$$X = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \tag{2.50}$$

Notice that the third row is the sum of the first two ($r_3 = r_1 + r_2$).

1. Row Rank and Basis U The first two rows are linearly independent. We set the row rank $r = 2$ and use these rows as our basis matrix U (2×4):

$$U = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \tag{2.51}$$

2. Coefficient Matrix C We express every row of X as a linear combination of the rows of U:

   - Row 1: $1 \cdot u_1 + 0 \cdot u_2$
   - Row 2: $0 \cdot u_1 + 1 \cdot u_2$
   - Row 3: $1 \cdot u_1 + 1 \cdot u_2$

These coefficients form the matrix C (3×2):

$$C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \tag{2.52}$$

1. The Decomposition $X = CU$) We verify that X is the product of C and U:

$$\underbrace{\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}}_{X\ (3\times4)} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}}_{C\ (3\times2)} \underbrace{\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}}_{U\ (2\times4)} \tag{2.53}$$

2. Conclusion on Column Rank The columns of X are linear combinations of the columns of C.

$$\text{Col}(X) \subseteq \text{Col}(C) \tag{2.54}$$

   Since C has only 2 columns, the dimension of its column space (and thus X's column space) cannot exceed 2.

$$\text{Dim}(\text{Col}(X)) \leq 2 \tag{2.55}$$

   This confirms that Row Rank (2) $\geq$ Column Rank. (By symmetry, they are equal).

Theorem 2.4 (Row Rank equals Column Rank)**.**

1. ***Row Rank equals Column Rank:*** *The dimension of the column space is equal to the dimension of the row space.*

$$Dim(Col(X)) = Dim(Row(X)) \implies Rank(X) = Rank(X') \tag{2.56}$$

2. ***Bounds:*** *For an $n \times p$ matrix X:*

$$Rank(X) \leq \min(n, p) \tag{2.57}$$

2.3.1 Orthogonality to a Subspace

We can extend the concept of orthogonality from single vectors to entire subspaces.

Definition 2.19 (Orthogonality to a Subspace)**.** A vector y is orthogonal to a subspace V (denoted $y \perp V$) if y is orthogonal to **every** vector x in V.

$$y \perp V \iff y'x = 0 \quad \forall x \in V \tag{2.58}$$

Definition 2.20 (Orthogonal Complement)**.** The set of all vectors that are orthogonal to a subspace V is called the **Orthogonal Complement** of V, denoted $V^{\perp}$.

$$V^{\perp} = \{y \in \mathbb{R}^n \mid y \perp V\} \tag{2.59}$$

2.3.2 Kernel (Null Space) and Image

For a matrix transformation defined by X, we define two key spaces: the Image (Column Space) and the Kernel (Null Space).

Definition 2.21 (Image and Kernel)**.**

1. **Image (Column Space):** The set of all possible outputs.

$$\text{Im}(X) = \text{Col}(X) = \{X\beta \mid \beta \in \mathbb{R}^p\} \tag{2.60}$$

2. **Kernel (Null Space):** The set of all inputs mapped to the zero vector.

$$\text{Ker}(X) = \{\beta \in \mathbb{R}^p \mid X\beta = 0\} \tag{2.61}$$

Theorem 2.5 (Relationship between Kernel and Row Space)**.** *The kernel of X is the orthogonal complement of the row space of X:*

$$Ker(X) = [Row(X)]^{\perp} \tag{2.62}$$

Proof. Let $x \in \mathbb{R}^p$. $x \in \text{Ker}(X)$ if and only if $Xx = 0$. If we denote the rows of X as $r'_1, \dots, r'_n$, then the equation $Xx = 0$ is equivalent to the system of equations:

$$\begin{pmatrix} r'_1 \\ \vdots \\ r'_n \end{pmatrix} x = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \iff r'_i x = 0 \text{ for all } i = 1, \dots, n \tag{2.63}$$

This means x is orthogonal to every row of X. Since the rows span the row space $\text{Row}(X)$, being orthogonal to every generator r_i implies x is orthogonal to the entire space $\text{Row}(X)$. Thus, $\text{Ker}(X) = \{x \mid x \perp \text{Row}(X)\} = [\text{Row}(X)]^\perp$. □

2.3.3 Nullity Theorem

There is a fundamental relationship between the dimensions of these spaces.

Theorem 2.6 (Rank-Nullity Theorem)**.** *For an $n \times p$ matrix X:*

$$\textit{Rank}(X) + \textit{Nullity}(X) = p \tag{2.64}$$

where $\textit{Nullity}(X) = \textit{Dim}(\textit{Ker}(X))$.

Proof. From the previous theorem, we established that the kernel is the orthogonal complement of the row space:

$$\text{Ker}(X) = [\text{Row}(X)]^\perp \tag{2.65}$$

Since the row space is a subspace of $\mathbb{R}^p$, the entire space can be decomposed into the direct sum of the row space and its orthogonal complement:

$$\mathbb{R}^p = \text{Row}(X) \oplus [\text{Row}(X)]^\perp = \text{Row}(X) \oplus \text{Ker}(X) \tag{2.66}$$

Taking the dimensions of these spaces:

$$\text{Dim}(\mathbb{R}^p) = \text{Dim}(\text{Row}(X)) + \text{Dim}(\text{Ker}(X)) \tag{2.67}$$

Substituting the definitions of Rank (dimension of row/column space) and Nullity:

$$p = \text{Rank}(X) + \text{Nullity}(X) \tag{2.68}$$

□

Comparing Ranks via Kernel Containment

The Rank-Nullity Theorem provides a powerful and convenient tool for comparing the ranks of two matrices A and B (with the same number of columns) by inspecting their null spaces.

Theorem 2.7 (Kernel Containment and Rank Inequality)**.** *Let A and B be two matrices with p columns. If the*

kernel of A is contained within the kernel of B, then the rank of A is greater than or equal to the rank of B.

$$Ker(A) \subseteq Ker(B) \implies Rank(A) \geq Rank(B) \tag{2.69}$$

Proof. From the subspace inclusion $\text{Ker}(A) \subseteq \text{Ker}(B)$, it follows that the dimension of the smaller space cannot exceed the dimension of the larger space:

$$\text{Nullity}(A) \leq \text{Nullity}(B) \tag{2.70}$$

Using the Rank-Nullity Theorem (Rank $= p -$ Nullity), we reverse the inequality:

$$p - \text{Nullity}(A) \geq p - \text{Nullity}(B) \tag{2.71}$$

$$\text{Rank}(A) \geq \text{Rank}(B) \tag{2.72}$$

□

2.3.4 Rank Inequalities

Understanding the bounds of the rank of matrix products is crucial for deriving properties of linear estimators.

Theorem 2.8 (Rank of a Matrix Product)**.** *Let X be an $n \times p$ matrix and Z be a $p \times k$ matrix. The rank of their product XZ is bounded by the rank of the individual matrices:*

$$Rank(XZ) \leq \min(Rank(X), Rank(Z)) \tag{2.73}$$

Proof. The columns of XZ are linear combinations of the columns of X. Thus, the column space of XZ is a subspace of the column space of X:

$$\text{Col}(XZ) \subseteq \text{Col}(X) \implies \text{Rank}(XZ) \leq \text{Rank}(X) \tag{2.74}$$

Similarly, the rows of XZ are linear combinations of the rows of Z. Thus, the row space of XZ is a subspace of the row space of Z:

$$\text{Row}(XZ) \subseteq \text{Row}(Z) \implies \text{Rank}(XZ) \leq \text{Rank}(Z) \tag{2.75}$$

□

Rank and Invertible Matrices

Multiplying by an invertible (non-singular) matrix preserves the rank. This is a very useful property when manipulating linear equations.

Theorem 2.9 (Rank with Non-Singular Multiplication)**.** *Let A be an $n \times n$ invertible matrix (i.e., $Rank(A) = n$) and X be an $n \times p$ matrix. Then:*

$$Rank(AX) = Rank(X) \tag{2.76}$$

Similarly, if B is a $p \times p$ invertible matrix, then:

$$Rank(XB) = Rank(X) \tag{2.77}$$

Proof. From the previous theorem, we know $\text{Rank}(AX) \leq \text{Rank}(X)$. Since A is invertible, we can write $X = A^{-1}(AX)$. Applying the theorem again:

$$\text{Rank}(X) = \text{Rank}(A^{-1}(AX)) \leq \text{Rank}(AX) \tag{2.78}$$

Thus, $\text{Rank}(AX) = \text{Rank}(X)$. □

2.3.5 Rank of $X'X$ and XX'

The matrix $X'X$ (the Gram matrix) appears in the normal equations for least squares ($X'X\beta = X'y$). Its properties are closely tied to X.

Theorem 2.10 (Rank of Gram Matrix)**.** *For any real matrix X, the rank of $X'X$ and XX' is the same as the rank of X itself:*

$$Rank(X'X) = Rank(X) \tag{2.79}$$

$$Rank(XX') = Rank(X) \tag{2.80}$$

Proof. We first show that the null space (kernel) of X is the same as the null space of $X'X$. If $v \in \text{Ker}(X)$, then $Xv = 0 \implies X'Xv = 0 \implies v \in \text{Ker}(X'X)$. Conversely, if $v \in \text{Ker}(X'X)$, then $X'Xv = 0$. Multiply by v':

$$v'X'Xv = 0 \implies (Xv)'(Xv) = 0 \implies ||Xv||^2 = 0 \implies Xv = 0 \tag{2.81}$$

So $\text{Ker}(X) = \text{Ker}(X'X)$. By the Rank-Nullity Theorem, since they have the same number of columns and same nullity, they must have the same rank. □

Column Space of XX'

Beyond just the rank, the column spaces themselves are related.

Theorem 2.11 (Column Space Equivalence)**.** *The column space of XX' is identical to the column space of X:*

$$Col(XX') = Col(X) \tag{2.82}$$

Proof.

1. **Forward ($\subseteq$):** Let $z \in \text{Col}(XX')$. Then $z = XX'w$ for some vector w. We can rewrite this as $z = X(X'w)$. Since z is a linear combination of columns of X (with coefficients $X'w$), $z \in \text{Col}(X)$. Thus, $\text{Col}(XX') \subseteq \text{Col}(X)$.
2. **Equality via Rank:** From the previous theorem, we know that $\text{Rank}(XX') = \text{Rank}(X)$. Since $\text{Col}(XX')$ is a subspace of $\text{Col}(X)$ and they have the same finite dimension (Rank), the subspaces must be identical.

□

Implication: This property ensures that for any y, the projection of y onto $\text{Col}(X)$ lies in the same space as the projection onto $\text{Col}(XX')$. This is vital for the existence of solutions in generalized least squares.

2.4 Orthogonal Projection onto a Subspace

Let V be a subspace of $\mathbb{R}^n$. For any vector $y \in \mathbb{R}^n$, there exists a **unique** vector $\hat{y} \in V$ such that the residual is orthogonal to the subspace:

$$(y - \hat{y}) \perp V \tag{2.83}$$

Equivalently:

$$\langle y - \hat{y}, v \rangle = 0 \quad \forall v \in V \tag{2.84}$$

2.4.1 Equivalence to Least Squares

The geometric definition of projection (orthogonality) is mathematically equivalent to the optimization problem of minimizing distance (least squares).

Theorem 2.12 (Best Approximation Theorem (Least Squares Property))**.** *Let V be a subspace of $\mathbb{R}^n$ and $y \in \mathbb{R}^n$. Let $\hat{y}$ be the orthogonal projection of y onto V. Then $\hat{y}$ is the closest point in V to y. That is, for any vector $v \in V$ such that $v \neq \hat{y}$:*

$$\|y - \hat{y}\|^2 < \|y - v\|^2 \tag{2.85}$$

Proof. Let v be any vector in V. We can rewrite the difference vector $y - v$ by adding and subtracting the projection $\hat{y}$:

$$y - v = (y - \hat{y}) + (\hat{y} - v) \tag{2.86}$$

Observe the properties of the two terms on the right-hand side:

1. **Residual:** $(y - \hat{y})$ is orthogonal to V by definition.
2. **Difference in Subspace:** Since both $\hat{y} \in V$ and $v \in V$, their difference $(\hat{y} - v)$ is also in V.

Therefore, the two terms are orthogonal to each other:

$$(y - \hat{y}) \perp (\hat{y} - v) \tag{2.87}$$

Applying the Pythagorean Theorem:

$$\|y - v\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - v\|^2 \tag{2.88}$$

Since squared norms are non-negative, and $\|\hat{y} - v\|^2 > 0$ (because $v \neq \hat{y}$):

$$\|y - v\|^2 > \|y - \hat{y}\|^2 \tag{2.89}$$

The projection $\hat{y}$ minimizes the squared error distance (and error distance itself). □

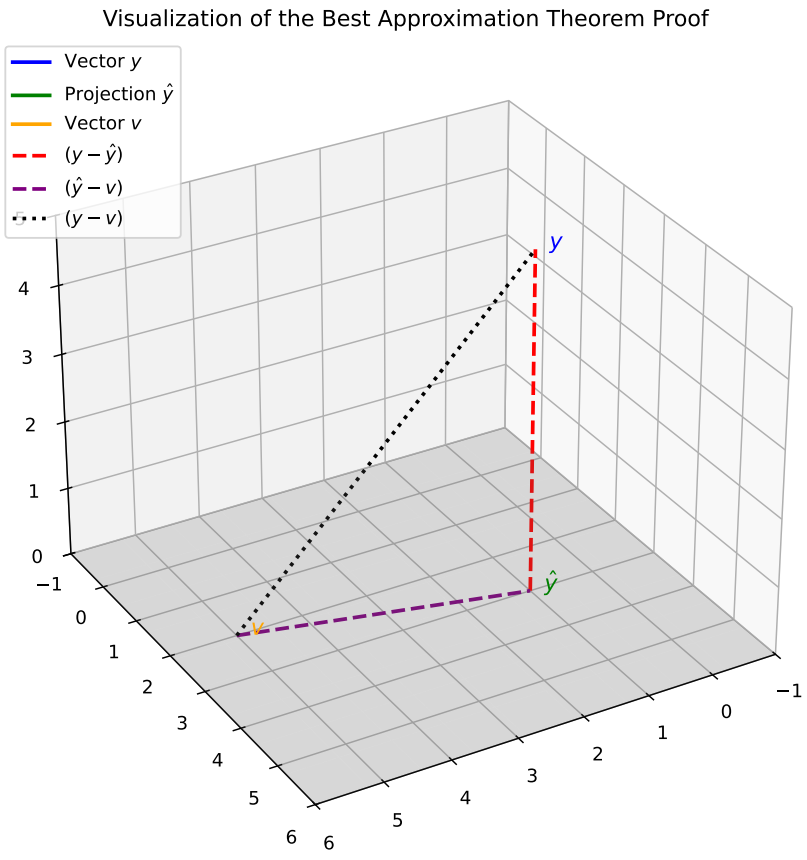


Figure 2.2: Visualization of the Best Approximation Theorem

2.4.2 Uniqueness of Projection

While the existence of a least-squares solution is guaranteed, we must also prove that there is only one such vector.

Theorem 2.13 (Uniqueness of Orthogonal Projection)**.** *For a given vector y and subspace V, the projection vector $\hat{y}$ satisfying $(y - \hat{y}) \perp V$ is unique.*

Proof. Assume there are two vectors $\hat{y}_1 \in V$ and $\hat{y}_2 \in V$ that both satisfy the orthogonality condition.

$$(y - \hat{y}_1) \perp V \quad \text{and} \quad (y - \hat{y}_2) \perp V \tag{2.90}$$

This means that for any $v \in V$, both inner products are zero:

$$\langle y - \hat{y}_1, v \rangle = 0 \tag{2.91}$$

$$\langle y - \hat{y}_2, v \rangle = 0 \tag{2.92}$$

Subtracting the second equation from the first:

$$\langle y - \hat{y}_1, v \rangle - \langle y - \hat{y}_2, v \rangle = 0 \tag{2.93}$$

Using the linearity of the inner product:

$$\langle (y - \hat{y}_1) - (y - \hat{y}_2), v \rangle = 0 \tag{2.94}$$

$$\langle \hat{y}_2 - \hat{y}_1, v \rangle = 0 \tag{2.95}$$

This equation holds for **all** $v \in V$. Since $\hat{y}_1$ and $\hat{y}_2$ are both in V, their difference $d = \hat{y}_2 - \hat{y}_1$ must also be in V. We can therefore choose $v = d = \hat{y}_2 - \hat{y}_1$.

$$\langle \hat{y}_2 - \hat{y}_1, \hat{y}_2 - \hat{y}_1 \rangle = 0 \implies \|\hat{y}_2 - \hat{y}_1\|^2 = 0 \tag{2.96}$$

The only vector with a norm of zero is the zero vector itself.

$$\hat{y}_2 - \hat{y}_1 = 0 \implies \hat{y}_1 = \hat{y}_2 \tag{2.97}$$

Thus, the projection is unique. □

2.5 Projection via Orthonormal Basis (Q)

2.5.1 Orthonomal Basis

Before discussing projections onto general subspaces, we must formally define the coordinate system of a subspace, known as a basis.

Definition 2.22 (Basis)**.** A set of vectors $\{x_1, \dots, x_k\}$ is a **Basis** for a vector space V if:

1. The vectors span the space: $V = L(x_1, \dots, x_k)$.
2. The vectors are linearly independent.

The number of vectors in a basis is unique and is defined as the **Dimension** of V.

Calculations become significantly simpler if we choose a basis with special geometric properties.

Definition 2.23 (Orthonormal Basis)**.** A basis $\{q_1, \dots, q_k\}$ is called an **Orthonormal Basis** if:

1. **Orthogonal:** Each pair of vectors is perpendicular.

$$q_i' q_j = 0 \quad \text{for } i \neq j \tag{2.98}$$

2. **Normalized:** Each vector has unit length.

$$||q_i||^2 = q_i' q_i = 1 \tag{2.99}$$

Combining these, we write $q_i' q_j = \delta_{ij}$ (Kronecker delta).

We now generalize the projection problem. Instead of projecting y onto a single line, we project it onto a subspace V of dimension k.

If we have an orthonormal basis $\{q_1, \dots, q_k\}$ for V, the projection $\hat{y}$ is simply the sum of the projections onto the individual basis vectors.

Definition 2.24 (Projection Defined with Orthonormal Basis)**.** The projection of y onto the subspace $V = L(q_1, \dots, q_k)$ is:

$$\hat{y} = \sum_{i=1}^{k} \text{proj}(y|q_i) = \sum_{i=1}^{k} (q_i' y) q_i \tag{2.100}$$

Since the basis vectors are normalized, we do not need to divide by $||q_i||^2$.

Theorem 2.14 (Projection via Orthonormal Basis)**.** *Let $\{q_1, \dots, q_k\}$ be an orthonormal basis for the subspace $V \subseteq \mathbb{R}^n$. The vector defined by the sum of individual projections:*

$$\hat{y} = \sum_{i=1}^{k} \langle y, q_i \rangle q_i \tag{2.101}$$

is indeed the orthogonal projection of y onto V. That is, it satisfies $(y - \hat{y}) \perp V$.

Proof. To prove this, we must check two conditions:

1. $\hat{y} \in V$: This is immediate because $\hat{y}$ is a linear combination of the basis vectors $\{q_1, \dots, q_k\}$.

2. $(y - \hat{y}) \perp V$: It suffices to show that the error vector $e = y - \hat{y}$ is orthogonal to every basis vector q_j (for $j = 1, \dots, k$).

   Let's calculate the inner product $\langle y - \hat{y}, q_j \rangle$:

$$\begin{aligned} \langle y - \hat{y}, q_j \rangle &= \langle y, q_j \rangle - \langle \hat{y}, q_j \rangle \\ &= \langle y, q_j \rangle - \left\langle \sum_{i=1}^{k} \langle y, q_i \rangle q_i, q_j \right\rangle \\ &= \langle y, q_j \rangle - \sum_{i=1}^{k} \langle y, q_i \rangle \underbrace{\langle q_i, q_j \rangle}_{\delta_{ij}} \end{aligned} \tag{2.102}$$

   Since the basis is orthonormal, $\langle q_i, q_j \rangle$ is 1 if $i = j$ and 0 otherwise. Thus, the summation collapses to a single term where $i = j$:

$$\begin{aligned} \langle y - \hat{y}, q_j \rangle &= \langle y, q_j \rangle - \langle y, q_j \rangle \cdot 1 \\ &= 0 \end{aligned} \tag{2.103}$$

   Since $(y - \hat{y})$ is orthogonal to every basis vector q_j, it is orthogonal to the entire subspace V. Thus, $\hat{y}$ is the unique orthogonal projection.

□

2.5.2 Projection Matrix via Orthonomal Basis (Q)

Matrix Form with Orthonormal Basis

We can express the summation formula for $\hat{y}$ compactly using matrix notation.

Let Q be an $n \times k$ matrix whose columns are the orthonormal basis vectors $q_1, \dots, q_k$.

$$Q = \begin{pmatrix} q_1 & q_2 & \dots & q_k \end{pmatrix} \tag{2.104}$$

Properties of Q:

- $Q'Q = I_k$ (Identity matrix of size $k \times k$).
- QQ' is **not** necessarily I_n (unless $k = n$).

Definition 2.25 (Projection Matrix in Terms of Q)**.** The projection $\hat{y}$ can be written as:

$$\hat{y} = \begin{pmatrix} q_1 & \dots & q_k \end{pmatrix} \begin{pmatrix} q_1'y \\ \vdots \\ q_k'y \end{pmatrix} = Q(Q'y) = (QQ')y \tag{2.105}$$

Thus, the projection matrix P onto the subspace V is:

$$P = QQ' \tag{2.106}$$

Properties of Projection Matrices

We have defined the projection matrix as $P = X(X'X)^{-1}X'$ (or $P = QQ'$ for orthonormal bases). All orthogonal projection matrices share two fundamental algebraic properties.

Theorem 2.15 (Symmeticity and Idempotence)**.** *A square matrix P represents an orthogonal projection onto some subspace if and only if it satisfies:*

1. ***Idempotence:*** $P^2 = P$ *(Applying the projection twice is the same as applying it once).*
2. ***Symmetry:*** $P' = P$.

Proof. If $\hat{y} = Py$ is already in the subspace Col(X), then projecting it again should not change it.

$$P(Py) = Py \implies P^2y = Py \quad \forall y \tag{2.107}$$

Thus, $P^2 = P$. □

Example: ANOVA (Analysis of Variance)

One of the most common applications of projection is in Analysis of Variance (ANOVA). We can view the calculation of group means as a projection onto a subspace defined by group indicator variables.

Example 2.5 (Finding Projection for One-way ANOVA)**.** Consider a one-way ANOVA model with k groups:

$$y_{ij} = \mu_i + \epsilon_{ij} \tag{2.108}$$

where $i \in \{1, \dots, k\}$ represents the group and $j \in \{1, \dots, n_i\}$ represents the observation within the group. Let $N = \sum_{i=1}^{k} n_i$ be the total number of observations.

1. **Matrix Definitions**

   We define the data vector y and the design matrix X as follows:

   - **Data Vector** (y): An $N \times 1$ vector containing all observations by group:

$$y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{kn_k} \end{pmatrix} \tag{2.109}$$

   - **Design Matrix** (X): An $N \times k$ matrix constructed from k column vectors, $X = (x_1, x_2, \dots, x_k)$. Each vector x_g is an **indicator** (dummy variable) for group g:

$$x_g = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \quad \leftarrow \text{Entries are 1 if observation belongs to group } g \tag{2.110}$$

2. **Orthogonality**

   These column vectors $x_1, \ldots, x_k$ are mutually orthogonal because no observation can belong to two groups at once. The dot product of any two distinct columns is zero:

$$\langle x_g, x_h \rangle = 0 \quad \text{for } g \neq h \tag{2.111}$$

   This allows us to find the projection onto the column space of X by simply summing the projections onto each column individually.

3. **Calculating Individual Projections**

   For a specific group vector x_g, the projection is:

$$\text{proj}(y|x_g) = \frac{\langle y, x_g \rangle}{\langle x_g, x_g \rangle} x_g \tag{2.112}$$

   We calculate the two scalar terms:

   - **Denominator** ($\langle x_g, x_g \rangle$): The sum of squared elements of x_g. Since x_g contains n_g ones and zeros elsewhere:

$$\langle x_g, x_g \rangle = \sum \mathbb{1}^2_{\{i=g\}} = n_g \tag{2.113}$$

   - **Numerator** ($\langle y, x_g \rangle$): The dot product sums only the y values belonging to group g:

$$\langle y, x_g \rangle = \sum_{i,j} y_{ij} \cdot \mathbb{1}_{\{i=g\}} = \sum_{j=1}^{n_g} y_{gj} = y_{g.} \quad \text{(Group Total)} \tag{2.114}$$

4. **The Resulting Projection**

   Substituting these back into the formula gives the coefficient for the vector x_g:

$$\text{proj}(y|x_g) = \frac{y_{g.}}{n_g} x_g = \bar{y}_{g.} x_g \tag{2.115}$$

   The total projection $\hat{y}$ is the sum over all groups:

$$\hat{y} = \sum_{g=1}^{k} \bar{y}_{g.} x_g \tag{2.116}$$

   This confirms that the fitted value for any specific observation y_{ij} is simply its group mean $\bar{y}_{i.}$.

2.5.3 Gram-Schmidt Process

To use the simplified formula $P = QQ'$, we need an orthonormal basis. The Gram-Schmidt process provides a method to construct such a basis from any set of linearly independent vectors.

Algorithm

Gram-Schmidt Process Given linearly independent vectors $x_1, \dots, x_p$:

1. **Step 1:** Normalize the first vector.

$$q_1 = \frac{x_1}{||x_1||} \tag{2.117}$$

2. **Step 2:** Project x_2 onto q_1 and subtract it to find the orthogonal component.

$$v_2 = x_2 - (x_2' q_1) q_1 \tag{2.118}$$

Then normalize:

$$q_2 = \frac{v_2}{||v_2||} \tag{2.119}$$

3. **Step k:** Subtract the projections onto all previous q vectors.

$$v_k = x_k - \sum_{j=1}^{k-1} (x_k' q_j) q_j \tag{2.120}$$

$$q_k = \frac{v_k}{||v_k||} \tag{2.121}$$

This process leads to the **QR Decomposition** of a matrix: $X = QR$, where Q is orthogonal and R is upper triangular.

2.6 Hat Matrix (Projection Matrix via X)

2.6.1 Norm Equations

Let $X = (x_1, \dots, x_p)$ be an $n \times p$ matrix, where each column x_j is a predictor vector.

We want to project the target vector y onto the column space $\text{Col}(X)$. This is equivalent to finding a coefficient vector $\beta \in \mathbb{R}^p$ such that the error vector (residual) is orthogonal to the entire subspace $\text{Col}(X)$.

$$y - X\beta \perp \text{Col}(X) \tag{2.122}$$

Since the columns of X span the subspace, the residual must be orthogonal to **every** column vector x_j individually:

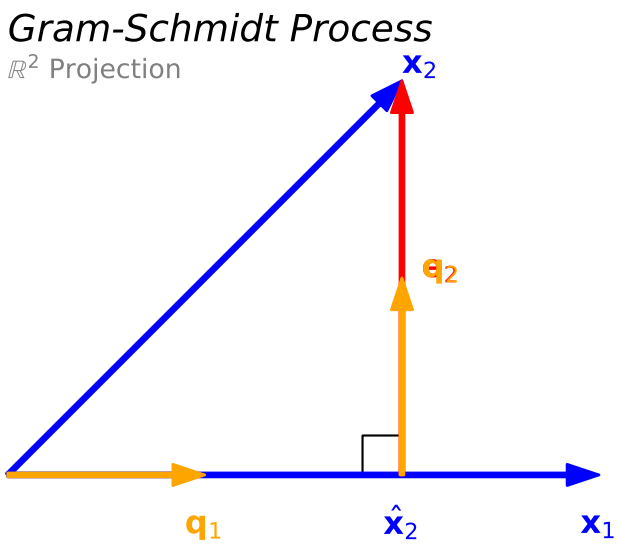


Figure 2.3: Gram-Schmidt Process: Projecting x_2 onto x_1

$$y - X\beta \perp x_j \quad \text{for } j = 1, \dots, p \tag{2.123}$$

Writing this geometric condition as an algebraic dot product (where x_j' denotes the transpose):

$$x_j'(y - X\beta) = 0 \quad \text{for each } j \tag{2.124}$$

We can stack these p separate linear equations into a single matrix equation. Since the rows of X' are the columns of X, this becomes:

$$\begin{pmatrix} x_1' \\ \vdots \\ x_p' \end{pmatrix} (y - X\beta) = \mathbf{0} \implies X'(y - X\beta) = 0 \tag{2.125}$$

Finally, we distribute the matrix transpose and rearrange terms to solve for β:

$$\begin{aligned} X'y - X'X\beta &= 0 \\ X'X\beta &= X'y \end{aligned} \tag{2.126}$$

This system is known as the **Normal Equations**.

Theorem 2.16 (Least Squares Estimator)**.** *If $X'X$ is invertible (i.e., X has full column rank), the unique solution for β is:*

$$\hat{\beta} = (X'X)^{-1}X'y \tag{2.127}$$

2.6.2 Hat Matrix

Substituting the estimator $\hat{\beta}$ back into the equation for $\hat{y}$ gives us the projection matrix.

Definition 2.26 (Hat Matrix)**.** The projection of y onto $\text{Col}(X)$ is given by:

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y \tag{2.128}$$

Thus, the hat matrix H is defined as:

$$H = X(X'X)^{-1}X' \tag{2.129}$$

2.6.3 Equivalence of Hat Matrix and QQ'

If we use the QR decomposition such that $X = QR$, where the columns of Q form an orthonormal basis for $\text{Col}(X)$, the formula simplifies significantly.

Recall that for orthonormal columns, $Q'Q = I$. Substituting $X = QR$ into the general formula:

$$\begin{aligned}
H &= QR((QR)'(QR))^{-1}(QR)' \\
&= QR(R'Q'QR)^{-1}R'Q' \\
&= QR(R' \underbrace{Q'Q}_{I} R)^{-1}R'Q' \\
&= QR(R'R)^{-1}R'Q' \\
&= QRR^{-1}(R')^{-1}R'Q' \\
&= Q \underbrace{RR^{-1}}_{I} \underbrace{(R')^{-1}R'}_{I} Q' \\
&= QQ'
\end{aligned} \tag{2.130}$$

This confirms that $H = QQ'$ is consistent with the general formula $H = X(X'X)^{-1}X'$.

2.6.4 Properties of Hat Matrix

We revisit the properties of projection matrices in this general context.

Theorem 2.17 (Properties of Hat Matrix)**.** *The matrix $H = X(X'X)^{-1}X'$ satisfies:*

1. ***Symmetric:*** $H' = H$

2. ***Idempotent:*** $H^2 = H$
3. ***Trace:*** *The trace of a projection matrix equals the dimension of the subspace it projects onto.*

$$tr(H) = tr(X(X'X)^{-1}X') = tr((X'X)^{-1}X'X) = tr(I_p) = p \tag{2.131}$$

2.7 Projection Defined with Orthogonal Projection Matrix

Projection don't have to be defined with a subspace or a matrix X as we discussed before. Projection matrix is a self-contained definition of the subspace it projects onto.

2.7.1 Orthogonal Projection Matrix

Definition 2.27 (Orthogonal Projection Matrix)**.** A square matrix P is called an **orthogonal projection matrix** if it satisfies two conditions:

1. **Symmetry:** $P^\top = P$
2. **Idempotency:** $P^2 = P$

Theorem 2.18 (Projection onto Column Space)**.** *Let P be a $p \times p$ symmetric ($P^\top = P$) and idempotent ($P^2 = P$) matrix in $\mathbb{R}^p$. Then P represents the orthogonal projection onto its column space, $Col(P)$. Specifically, for any vector $y \in \mathbb{R}^p$, the vector $\hat{y} = Py$ satisfies the definition of orthogonal projection:*

1. $\hat{y} \in Col(P)$
2. $y - \hat{y} \perp Col(P)$

Proof. To prove that P is the orthogonal projector onto $\text{Col}(P)$, we verify the two conditions for an arbitrary vector $y \in \mathbb{R}^p$.

1. Condition: $\hat{y} \in \text{Col}(P)$

   By the definition of matrix-vector multiplication, $\hat{y} = Py$ is a linear combination of the columns of P. Therefore, $\hat{y}$ is, by definition, an element of $\text{Col}(P)$.

2. Condition: $y - \hat{y} \perp \text{Col}(P)$

   Let $e = y - \hat{y} = (I_n - P)y$. To verify that e is orthogonal to $\text{Col}(P)$, it suffices to show that e is orthogonal to every column of P. In matrix notation, this is equivalent to showing $e^\top P = 0$. We compute this directly:

$$
\begin{aligned}
e^\top P &= [(I_p - P)y]^\top P \\
&= y^\top (I_p - P)^\top P \\
&= y^\top (I_p - P)P \qquad \text{(Symmetry: } P^\top = P\text{)} \\
&= y^\top (P - P^2) \\
&= y^\top (P - P) \qquad \text{(Idempotency: } P^2 = P\text{)} \\
&= 0
\end{aligned}
\tag{2.132}
$$

Since $e^\top P = 0$, the residual e is orthogonal to every column of P. Consequently, e is orthogonal to the space spanned by those columns, $\text{Col}(P)$. □

Lemma 2.1 (0-1 Projection)**.** *Let P be a $n \times n$ matrix. P is the orthogonal projection matrix onto $Col(P)$ if and only if:*

1) $Pv = v$ *for all* $v \in Col(P)$.
2) $Pv = 0$ *for all* $v \perp Col(P)$.

Proof. **Forward Implication ($\implies$):** Given P is an orthogonal projection ($P^2 = P, P^\top = P$).

(1) **Proof of (1):** Let $v \in \text{Col}(P)$. Then $v = Px$ for some x.

$$Pv = P(Px) = P^2x = Px = v \tag{2.133}$$

(2) **Proof of (2):** Let $v \perp \text{Col}(P)$. Then v is orthogonal to every column of P, so $v^\top P = 0$. Since P is symmetric:

$$Pv = (v^\top P^\top)^\top = (v^\top P)^\top = 0^\top = 0 \tag{2.134}$$

Reverse Implication ($\impliedby$): Given conditions (1) and (2) hold.
We must show that P is idempotent ($P^2 = P$) and symmetric ($P^\top = P$).

(1) **Proof of Idempotence ($P^2 = P$):** For any vector $x \in \mathbb{R}^n$, let $y = Px$. By definition, $y \in \text{Col}(P)$. Applying condition (1) to the vector y:

$$Py = y \implies P(Px) = Px \implies P^2x = Px \tag{2.135}$$

Since this holds for all x, $P^2 = P$.

(2) **Proof of Symmetry ($P^\top = P$):** We decompose any two vectors $x, y \in \mathbb{R}^n$ into components inside and orthogonal to $\text{Col}(P)$. Let $x = x_1 + x_2$ and $y = y_1 + y_2$, where $x_1, y_1 \in \text{Col}(P)$ and $x_2, y_2 \perp \text{Col}(P)$. Using conditions (1) and (2):

$$Px = P(x_1 + x_2) = Px_1 + Px_2 \overset{(1),(2)}{=} x_1 + 0 = x_1 \tag{2.136}$$

$$Py = P(y_1 + y_2) = Py_1 + Py_2 \overset{(1),(2)}{=} y_1 + 0 = y_1 \tag{2.137}$$

Now we compare the inner products $\langle Px, y\rangle$ and $\langle x, Py\rangle$:

$$\langle Px, y\rangle = \langle x_1, y_1 + y_2\rangle = \langle x_1, y_1\rangle + \underbrace{\langle x_1, y_2\rangle}_{0} = \langle x_1, y_1\rangle \tag{2.138}$$

$$\langle x, Py\rangle = \langle x_1 + x_2, y_1\rangle = \langle x_1, y_1\rangle + \underbrace{\langle x_2, y_1\rangle}_{0} = \langle x_1, y_1\rangle \tag{2.139}$$

Since $\langle Px, y\rangle = \langle x, Py\rangle$ implies $x^\top P^\top y = x^\top Py$ for all x, y, we conclude $P^\top = P$.
Since P is symmetric and idempotent, it is the orthogonal projection matrix. □

2.7.2 Projection onto Complement Space

Theorem 2.19 (Projection onto Orthogonal Complement)**.** *Let P be a $n \times n$ orthogonal projection matrix operating in the space $\mathbb{R}^n$. The matrix M defined as:*

$$M = I_p - P \tag{2.140}$$

is the orthogonal projection matrix onto the orthogonal complement of the column space of P, denoted $Col(P)^\perp \subseteq \mathbb{R}^n$.

Proof.

(1) **Symmetry and Idempotency** Since P is a projection matrix, $P^\top = P$ and $P^2 = P$. We verify these properties for M:

$$M^\top = (I_p - P)^\top = I_p - P^\top = I_p - P = M \tag{2.141}$$

$$M^2 = (I_p - P)(I_p - P) = I_p - 2P + P^2 = I_p - 2P + P = I_p - P = M \tag{2.142}$$

By Equation 2.141 and Equation 2.142, M is symmetric and idempotent, so it is an orthogonal projection matrix.

(2) **Identifying the Subspace** We now show that $\text{Col}(M) = \text{Col}(P)^\perp$ by mutual inclusion.

(1) **Direction 1**: $\text{Col}(M) \subseteq \text{Col}(P)^\perp$ Let $v \in \text{Col}(M)$. Then $v = Mx$ for some vector x. Multiplying by P:

$$Pv = P(I_p - P)x = (P - P^2)x = 0 \tag{2.143}$$

Since P is symmetric ($P = P'$), taking the transpose of $Pv = 0$ gives $v'P = 0$. This means v is orthogonal to every column of P. Therefore, $v \in \text{Col}(P)^\perp$.

(2) **Direction 2**: $\text{Col}(P)^\perp \subseteq \text{Col}(M)$ Let $v \in \text{Col}(P)^\perp$. By definition, v is orthogonal to the columns of P, so $v'P = 0$. Taking the transpose and using symmetry ($P' = P$), we get $Pv = 0$.
Now applying M to v:

$$Mv = (I_p - P)v = v - Pv = v \tag{2.144}$$

Since $Mv = v$, v lies in the column space of M. Therefore, $v \in \text{Col}(M)$.

Since both inclusions hold, $\text{Col}(M) = \text{Col}(P)^\perp$. □

2.7.3 Projections onto Nested Subspaces

2.7.3.1 Iterative Projections

Theorem 2.20 (Iterative Projections)**.** *Let P_0 and P_1 be $n \times n$ orthogonal projection matrices such that $Col(P_0) \subseteq Col(P_1)$. Then:*

(1) $P_1 P_0 = P_0$
(2) $P_0 P_1 = P_0$

Proof. **Method 1**:
Proof of $P_1 P_0 = P_0$:
Let $y \in \mathbb{R}^n$ be an arbitrary vector. By definition, the vector $v = P_0 y$ lies in $\text{Col}(P_0)$. Given $\text{Col}(P_0) \subseteq \text{Col}(P_1)$, it follows that $v \in \text{Col}(P_1)$.
Using **Lemma Lemma 2.1**, since $v \in \text{Col}(P_1)$, P_1 acts as the identity on v, so $P_1 v = v$. Substituting $v = P_0 y$:

$$P_1(P_0 y) = P_0 y \tag{2.145}$$

Since $P_1 P_0 y = P_0 y$ holds for all $y \in \mathbb{R}^n$, we conclude $P_1 P_0 = P_0$.
Proof of $P_0 P_1 = P_0$:
Taking the transpose of the result from part 1 and applying the symmetry property ($P' = P$):

$$(P_1 P_0)' = P_0' \implies P_0' P_1' = P_0' \implies P_0 P_1 = P_0 \tag{2.146}$$

Method 2:
To prove $P_0 P_1 = P_0$, for any $y \in \mathbb{R}^n$, let $\hat{y}_1 = P_1 y$, $\hat{y}_0 = P_0 y$, $e_1 = y - \hat{y}_1$, and $e_0 = y - \hat{y}_0$. Note that both e_0 and e_1 are orthogonal to $\text{Col}(P_0)$ (since $\text{Col}(P_0) \subseteq \text{Col}(P_1)$).
We have:

$$P_0(P_1 - P_0)y = P_0(\hat{y}_1 - \hat{y}_0) = P_0(e_0 - e_1) = 0 \tag{2.147}$$

This implies $P_0 P_1 - P_0 = 0$, so $P_0 P_1 = P_0$. □

2.7.3.2 Difference of Projections

Theorem 2.21 (Difference Projection)**.** *The matrix $P_\Delta = P_1 - P_0$ is an orthogonal projection matrix onto the subspace $Col(P_1) \cap Col(P_0)^\perp$. This subspace represents the "extra" information in the full model that is orthogonal to the reduced model. Additionally, the following column space relationship holds:*

$$Col(P_1 - P_0) = Col(P_0)^\perp \cap Col(P_1) \tag{2.148}$$

Proof.

1. **Symmetry:** Since P_1 and P_0 are symmetric:

$$(P_1 - P_0)' = P_1' - P_0' = P_1 - P_0 \tag{2.149}$$

2. **Idempotency:**

$$\begin{aligned}(P_1 - P_0)^2 &= (P_1 - P_0)(P_1 - P_0) \\ &= P_1^2 - P_1P_0 - P_0P_1 + P_0^2\end{aligned} \tag{2.150}$$

Using the projection property ($P^2 = P$) and the nested property ($P_1P_0 = P_0$ and $P_0P_1 = P_0$):

$$= P_1 - P_0 - P_0 + P_0 = P_1 - P_0 \tag{2.151}$$

3. **Orthogonality to P_0:**

$$(P_1 - P_0)P_0 = P_1P_0 - P_0^2 = P_0 - P_0 = 0 \tag{2.152}$$

4. **Column Space Identity:** We show $\text{Col}(P_1 - P_0) = \text{Col}(P_0)^\perp \cap \text{Col}(P_1)$ via double containment.

   ($\subseteq$) **Forward Containment:** Let $y \in \text{Col}(P_1 - P_0)$. By definition, $y = (P_1 - P_0)x$ for some x.

   - Check $y \in \text{Col}(P_1)$: $P_1y = P_1(P_1 - P_0)x = (P_1 - P_0)x = y$. Thus $y \in \text{Col}(P_1)$.
   - Check $y \in \text{Col}(P_0)^\perp$: $P_0y = P_0(P_1 - P_0)x = (P_0 - P_0)x = 0$. Thus $y \in \text{Col}(P_0)^\perp$.
   - Therefore, $\text{Col}(P_1 - P_0) \subseteq \text{Col}(P_0)^\perp \cap \text{Col}(P_1)$.

   ($\supseteq$) **Reverse Containment**: Let $y \in \text{Col}(P_0)^\perp \cap \text{Col}(P_1)$.

   - Since $y \in \text{Col}(P_1)$, $P_1y = y$.
   - Since $y \in \text{Col}(P_0)^\perp$, $P_0y = 0$.
   - Observe $(P_1 - P_0)y = P_1y - P_0y = y - 0 = y$.
   - This implies y is in the range of $(P_1 - P_0)$. Therefore, $\text{Col}(P_0)^\perp \cap \text{Col}(P_1) \subseteq \text{Col}(P_1 - P_0)$.

□

> **!** Important
>
> This is important as we can use $P_2 - P_1$ to construct the projection matrix and the space that it projects onto.

Hat Matrix of Incremental Space

Theorem 2.22 (Hat Matrix of Incremental Space)**.** *Let X_1 be a design matrix of dimension $n \times k_1$ and X_2 be a design matrix of dimension $n \times k_2$, such that the combined matrix $X = [X_1, X_2]$ has full column rank. Let $V_1 = Col(X_1)$ and $V_2 = Col([X_1, X_2])$. Let P_1 and P_2 be the orthogonal projection matrices onto V_1 and V_2, respectively.*
Define the matrix of residuals $\tilde{X}_2$ as:

$$\tilde{X}_2 = (I - P_1)X_2 \tag{2.153}$$

Let $W = Col(\tilde{X}_2)$. Let P_W be the $n \times n$ projection matrix onto W, which is the hat matrix constructed from $\tilde{X}_2$:

$$P_W = \tilde{X}_2(\tilde{X}_2^T\tilde{X}_2)^{-1}\tilde{X}_2^T \tag{2.154}$$

(a) Let $X^* = [X_1, \tilde{X}_2]$. Prove that the column space of the original design matrix X is identical to the column space of the modified design matrix X^*:

$$Col([X_1, X_2]) = Col([X_1, \tilde{X}_2]) \tag{2.155}$$

(b) Using the result from Part (a) and the definition of the Hat Matrix, prove that:

$$P_W = P_2 - P_1 \tag{2.156}$$

Proof. Assignment question. □

2.7.4 Projection onto Three Multually Orthogonal Subspaces

Theorem 2.23 (Orthogonal Decomposition)**.** *Let $M_0 \subset M_1$ be two nested linear models associated with orthogonal projection matrices P_0 and P_1, such that $Col(P_0) \subset Col(P_1)$.*
For any observation vector y, we have the decomposition:

$$y = \underbrace{P_0 y}_{\hat{y}_0} + \underbrace{(P_1 - P_0)y}_{\hat{y}_1 - \hat{y}_0} + \underbrace{(I - P_1)y}_{y - \hat{y}_1} \tag{2.157}$$

Geometric Interpretation:

1. *$\hat{y}_0 \in Col(P_0)$: The fit of the reduced model.*
2. *$(\hat{y}_1 - \hat{y}_0) \in Col(P_0)^\perp \cap Col(P_1)$: The additional fit provided by M_1 over M_0.*
3. *$(y - \hat{y}_1) \in Col(P_1)^\perp$: The projection of y onto the **orthogonal complement** of $Col(P_1)$.*

The three component vectors are mutually orthogonal. Consequently, their squared norms sum to the total squared norm:

$$\|y\|^2 = \|\hat{y}_0\|^2 + \|\hat{y}_1 - \hat{y}_0\|^2 + \|y - \hat{y}_1\|^2 \tag{2.158}$$

Theorem 2.24 (Orthogonal Decomposition)**.** *Let $M_0 \subset M_1$ be two nested linear models associated with orthogonal projection matrices P_0 and P_1, such that $Col(P_0) \subset Col(P_1)$.*
For any observation vector y, we have the decomposition:

$$y = \underbrace{P_0 y}_{\hat{y}_0} + \underbrace{(P_1 - P_0)y}_{\hat{y}_1 - \hat{y}_0} + \underbrace{(I - P_1)y}_{y - \hat{y}_1} \tag{2.159}$$

Geometric Interpretation:

1. $\hat{y}_0 \in Col(P_0)$*: The fit of the reduced model.*
2. $(\hat{y}_1 - \hat{y}_0) \in Col(P_0)^\perp \cap Col(P_1)$*: The additional fit provided by* M_1 *over* M_0*.*
3. $(y - \hat{y}_1) \in Col(P_1)^\perp$*: The projection of* y *onto the* ***orthogonal complement*** *of* $Col(P_1)$*.*

The three component vectors are mutually orthogonal. Consequently, their squared norms sum to the total squared norm:

$$\|y\|^2 = \|\hat{y}_0\|^2 + \|\hat{y}_1 - \hat{y}_0\|^2 + \|y - \hat{y}_1\|^2 \tag{2.160}$$

Proof.

1. **Definition of Vectors and Nested Spaces**

   Let I be the identity matrix, which is the orthogonal projection onto the entire space $\mathbb{R}^n$. We effectively have a three-level nested sequence of subspaces:

   $$\text{Col}(P_0) \subset \text{Col}(P_1) \subset \mathbb{R}^n \tag{2.161}$$

   We define the components of the decomposition using successive difference projections:

   - $v_0 = P_0 y$
   - $v_1 = (P_1 - P_0)y$
   - $v_2 = (I - P_1)y$

   Summing these gives the identity: $y = v_0 + v_1 + v_2$.

2. **Sequential Orthogonality via Theorem 2.21**

   We apply the Difference Projection Theorem (Theorem 2.21) to each successive pair of nested spaces to establish orthogonality.

   - **Step 1: Verify** $v_1 \perp v_0$
     - Consider the nested pair P_0 and P_1.
     - By Theorem 2.21, the matrix $(P_1 - P_0)$ projects onto $\text{Col}(P_0)^\perp \cap \text{Col}(P_1)$.
     - Since $v_0 \in \text{Col}(P_0)$ and v_1 lies in the orthogonal complement $\text{Col}(P_0)^\perp$, we have $v_1 \perp v_0$.
   - **Step 2: Verify** $v_2 \perp \{v_0, v_1\}$
     - Consider the nested pair P_1 and I (where I projects onto $\mathbb{R}^n$).
     - By Theorem 2.21, the matrix $(I - P_1)$ projects onto $\text{Col}(P_1)^\perp \cap \mathbb{R}^n = \text{Col}(P_1)^\perp$.
     - Since both v_0 and v_1 reside within $\text{Col}(P_1)$ (as shown in Step 1), and v_2 lies in the orthogonal complement $\text{Col}(P_1)^\perp$, it follows that v_2 is orthogonal to the entire subspace $\text{Col}(P_1)$.
     - Therefore, $v_2 \perp v_0$ and $v_2 \perp v_1$.

3. **Conclusion**

   Since $\{v_0, v_1, v_2\}$ are mutually orthogonal, the Pythagorean theorem applies:

$$\|y\|^2 = \|v_0\|^2 + \|v_1\|^2 + \|v_2\|^2 \tag{2.162}$$

   Substituting the original definitions back in:

$$\|y\|^2 = \|\hat{y}_0\|^2 + \|\hat{y}_1 - \hat{y}_0\|^2 + \|y - \hat{y}_1\|^2 \tag{2.163}$$

□

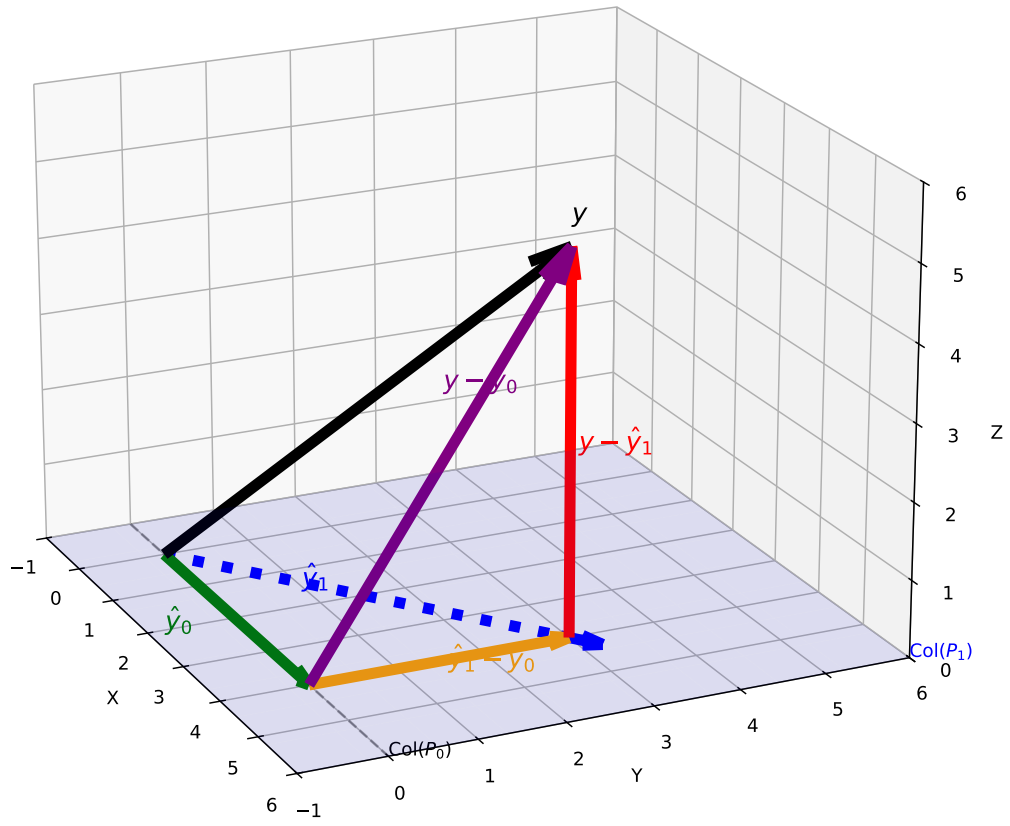


Figure 2.4: Illustration of Projections onto Nested Subspaces

Example 2.6 (ANOVA Sum Squares)**.** We apply the **Nested Model Theorem** ($M_0 \subset M_1$) to the One-way ANOVA setting.

1. Notation and Definitions

   Consider a dataset with k groups. Let $i = 1, \ldots, k$ index the groups, and $j = 1, \ldots, n_i$ index the

observations within group i.

- N: Total number of observations, $N = \sum_{i=1}^{k} n_i$.
- y_{ij}: The j-th observation in the i-th group.
- $\bar{y}_{i.}$: The sample mean of group i.

$$\bar{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \tag{2.164}$$

- $\bar{y}_{..}$: The grand mean of all observations.

$$\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \tag{2.165}$$

2. The Data and Projection Vectors

Table 2.1: ANOVA Vectors: Data, Null Model, and Full Model

Observation (y)	Null Projection ($\widehat{y}_0$)	Full Projection ($\widehat{y}_1$)
$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \hline \vdots \\ \hline y_{k1} \\ \vdots \\ y_{kn_k} \end{pmatrix}$	$\begin{pmatrix} \bar{y}_{\cdot\cdot} \\ \vdots \\ \bar{y}_{\cdot\cdot} \\ \hline \vdots \\ \hline \bar{y}_{\cdot\cdot} \\ \vdots \\ \bar{y}_{\cdot\cdot} \end{pmatrix}$	$\begin{pmatrix} \bar{y}_{1\cdot} \\ \vdots \\ \bar{y}_{1\cdot} \\ \hline \vdots \\ \hline \bar{y}_{k\cdot} \\ \vdots \\ \bar{y}_{k\cdot} \end{pmatrix}$

3. Decomposition and Sum of Squares

Component	Notation	Definition	Vector Elements	Squared Norm (Sum of Squares)
Null Proj.	$\hat{y}_0$	$P_0 y$	Grand Mean ($\bar{y}_{\cdot\cdot}$)	$\|\hat{y}_0\|^2 = N\bar{y}_{\cdot\cdot}^2$
Full Proj.	$\hat{y}_1$	$P_1 y$	Group Means ($\bar{y}_{i\cdot}$)	$\|\hat{y}_1\|^2 = \sum_{i=1}^{k} n_i \bar{y}_{i\cdot}^2$

3. Geometric Justification of Shortcut Formulas

A. Total Sum of Squares (SST) Since $\hat{y}_0 \perp (y - \hat{y}_0)$, we have $\|y\|^2 = \|\hat{y}_0\|^2 + \|y - \hat{y}_0\|^2$:

$$\text{SST} = \|y - \hat{y}_0\|^2 = \|y\|^2 - \|\hat{y}_0\|^2 \tag{2.166}$$

$$\text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - N\bar{y}_{\cdot\cdot}^2 \tag{2.167}$$

B. Between Group Sum of Squares (SSB) Since $\hat{y}_0 \perp (\hat{y}_1 - \hat{y}_0)$, we have $\|\hat{y}_1\|^2 = \|\hat{y}_0\|^2 + \|\hat{y}_1 - \hat{y}_0\|^2$:

$$\text{SSB} = \|\hat{y}_1 - \hat{y}_0\|^2 = \|\hat{y}_1\|^2 - \|\hat{y}_0\|^2 \tag{2.168}$$

$$\text{SSB} = \sum_{i=1}^{k} n_i \bar{y}_{i\cdot}^2 - N\bar{y}_{\cdot\cdot}^2 \tag{2.169}$$

C. Within Group Sum of Squares (SSW) Since $\hat{y}_1 \perp (y - \hat{y}_1)$, we have $\|y\|^2 = \|\hat{y}_1\|^2 + \|y - \hat{y}_1\|^2$:

$$\text{SSW} = \|y - \hat{y}_1\|^2 = \|y\|^2 - \|\hat{y}_1\|^2 \tag{2.170}$$

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k} n_i \bar{y}_{i\cdot}^2 \tag{2.171}$$

Conclusion:

$$\underbrace{\|y\|^2 - N\bar{y}_{\cdot\cdot}^2}_{\text{SST}} = \underbrace{(\sum n_i \bar{y}_{i\cdot}^2 - N\bar{y}_{\cdot\cdot}^2)}_{\text{SSB}} + \underbrace{(\sum\sum y_{ij}^2 - \sum n_i \bar{y}_{i\cdot}^2)}_{\text{SSW}} \tag{2.172}$$

4. Visualizing ANOVA Components in Data Space

2.8 Projections onto More than Three Orthogonal Subspaces

Finally, we consider the case where the entire space $\mathbb{R}^n$ is decomposed into mutually orthogonal subspaces.

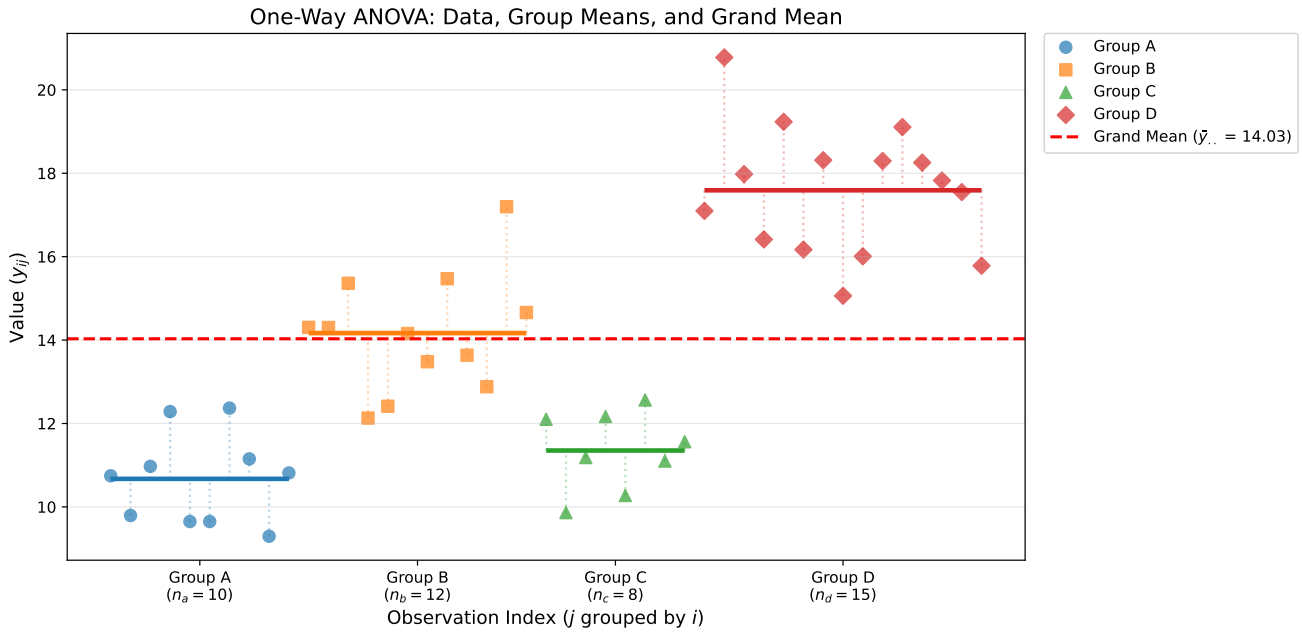


Figure 2.5: Visualization of Group Means vs. Grand Mean

Theorem 2.25 (General Orthogonal Projections). *If $\mathbb{R}^n$ is the direct sum of orthogonal subspaces $V_1, V_2, \dots, V_k$:*

$$\mathbb{R}^n = V_1 \oplus V_2 \oplus \cdots \oplus V_k \tag{2.173}$$

where $V_i \perp V_j$ for all $i \neq j$.
Then any vector y can be uniquely written as:

$$y = \hat{y}_1 + \hat{y}_2 + \cdots + \hat{y}_k \tag{2.174}$$

where $\hat{y}_i \in V_i$.
Furthermore, each component $\hat{y}_i$ is simply the projection of y onto the subspace V_i:

$$\hat{y}_i = P_i y \tag{2.175}$$

Proof.

1. Existence: Since $\mathbb{R}^n$ is the direct sum of $V_1, \dots, V_k$, by definition, any vector $y \in \mathbb{R}^n$ can be written as a sum $y = v_1 + \cdots + v_k$ where $v_i \in V_i$.

2. Uniqueness: Suppose there are two such representations: $y = \sum v_i = \sum w_i$, with $v_i, w_i \in V_i$. Then $\sum(v_i - w_i) = 0$. Since subspaces in a direct sum are independent, the only way for the sum of elements to be zero is if each individual element is zero. Thus, $v_i - w_i = 0 \implies v_i = w_i$. The representation is unique. Let $\hat{y}_i = v_i$.

3. Projection Property: We claim that the i-th component $\hat{y}_i$ is the orthogonal projection of y onto V_i. We

must show that the residual $(y - \hat{y}_i)$ is orthogonal to V_i.

$$y - \hat{y}_i = \sum_{j \neq i} \hat{y}_j \tag{2.176}$$

Let z be any vector in V_i. We calculate the inner product:

$$\langle y - \hat{y}_i, z \rangle = \left\langle \sum_{j \neq i} \hat{y}_j, z \right\rangle = \sum_{j \neq i} \langle \hat{y}_j, z \rangle \tag{2.177}$$

Since $\hat{y}_j \in V_j$ and $z \in V_i$, and the subspaces are mutually orthogonal ($V_j \perp V_i$ for $j \neq i$), every term in the sum is zero. Therefore, $(y - \hat{y}_i) \perp V_i$. By the definition of orthogonal projection, $\hat{y}_i = P_i y$.

□

This implies that the identity matrix can be decomposed into a sum of projection matrices:

$$I_n = P_1 + P_2 + \cdots + P_k \tag{2.178}$$

Theorem 2.26 (Complete Orthogonal Decomposition of $\mathbb{R}^n$)**.** *Let $P_0, P_1, \ldots, P_k$ be a sequence of orthogonal projection matrices with nested column spaces:*

$$Col(P_0) \subseteq Col(P_1) \subseteq \cdots \subseteq Col(P_k) \tag{2.179}$$

Define the sequence of difference matrices ΔP_i and their column spaces V_i as follows:

$$\begin{aligned} \Delta P_0 &= P_0, & V_0 &= Col(\Delta P_0) \\ \Delta P_i &= P_i - P_{i-1} \quad (1 \leq i \leq k), & V_i &= Col(\Delta P_i) \\ \Delta P_{k+1} &= I - P_k, & V_{k+1} &= Col(\Delta P_{k+1}) \end{aligned}$$

Conclusion:

1. ***Projection Property:*** *Each ΔP_i is the orthogonal projection matrix onto V_i for $i = 0, \ldots, k+1$.*
2. ***Mutual Orthogonality:*** *The collection $\{\Delta P_i\}$ are mutually orthogonal operators:*

$$\Delta P_i \Delta P_j = 0 \quad \textit{for all } i \neq j \tag{2.180}$$

3. ***Direct Sum Decomposition:*** *The vector space $\mathbb{R}^n$ is the direct sum of these orthogonal subspaces:*

$$\mathbb{R}^n = V_0 \oplus V_1 \oplus \cdots \oplus V_{k+1} \tag{2.181}$$

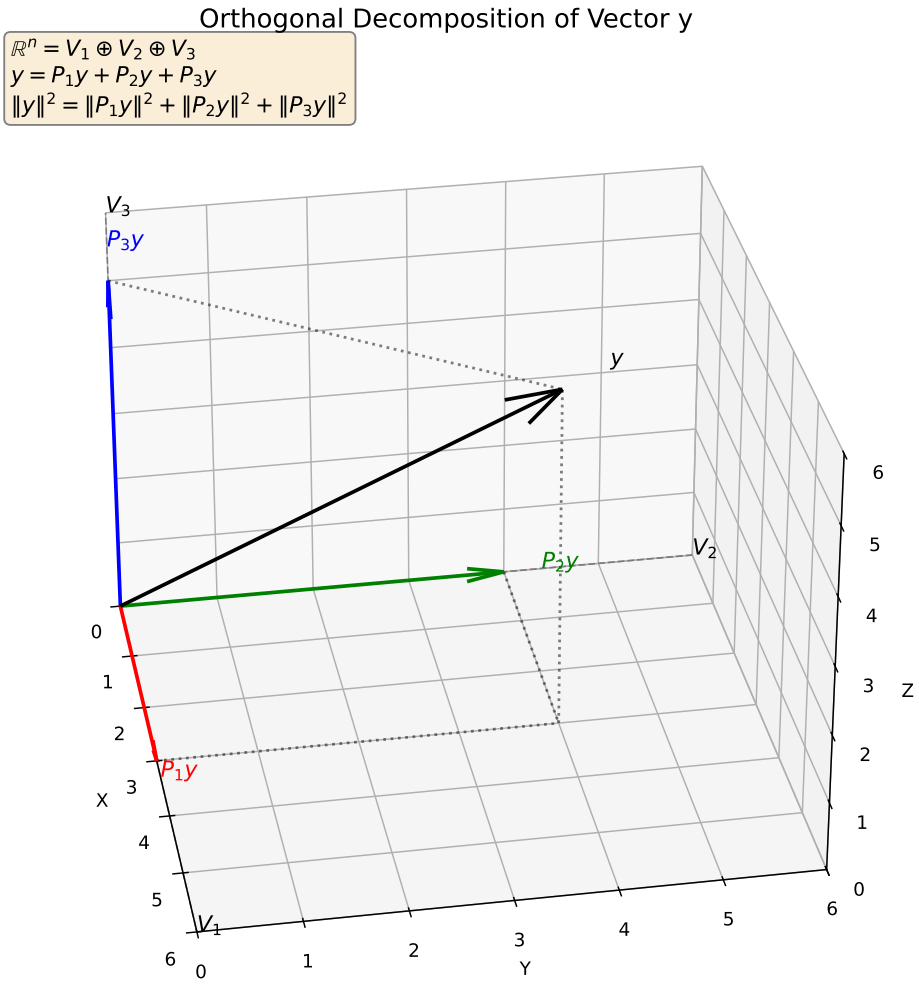


Figure 2.6: Orthogonal decomposition of vector y into subspaces

Proof.

1. Proof that ΔP_i is the Projection onto V_i

   We must show each ΔP_i is symmetric and idempotent.

   - For $\Delta P_0 = P_0$: True by definition.
   - For ΔP_i $(1 \leq i \leq k)$:
     - **Symmetry:** Difference of symmetric matrices ($P_i, P_{i-1} $) is symmetric.
     - **Idempotency:** $(\Delta P_i)^2 = (P_i - P_{i-1})^2 = P_i^2 - P_i P_{i-1} - P_{i-1} P_i + P_{i-1}^2$. Using nested properties ($P_i P_{i-1} = P_{i-1}$), this simplifies to $P_i - P_{i-1} = \Delta P_i$.
   - For $\Delta P_{k+1} = I - P_k$:
     - **Symmetry:** $(I - P_k)' = I - P_k$.
     - **Idempotency:** $(I - P_k)^2 = I - 2P_k + P_k^2 = I - P_k$.

2. Proof of Mutual Orthogonality

   We show $\Delta P_j \Delta P_i = 0$ for $i < j$.

   - **Case 1: Both indices** $\leq k$ (i.e., $1 \leq i < j \leq k$):

   $$(P_j - P_{j-1})(P_i - P_{i-1}) = P_j P_i - P_j P_{i-1} - P_{j-1} P_i + P_{j-1} P_{i-1} \tag{2.182}$$

   Since $\text{Col}(P_i) \subseteq \text{Col}(P_{j-1})$, all terms reduce to $P_i - P_{i-1} - P_i + P_{i-1} = 0$.

   - **Case 2: One index is the residual** $(j = k + 1)$: We check $\Delta P_{k+1} \Delta P_i = (I - P_k)\Delta P_i$ for any $i \leq k$. Since $V_i \subseteq \text{Col}(P_k)$, we have $P_k \Delta P_i = \Delta P_i$.

   $$(I - P_k)\Delta P_i = \Delta P_i - P_k \Delta P_i = \Delta P_i - \Delta P_i = 0 \tag{2.183}$$

3. Proof of Direct Sum

   The sum of the difference matrices forms a telescoping series:

   $$\sum_{j=0}^{k+1} \Delta P_j = P_0 + \sum_{i=1}^{k} (P_i - P_{i-1}) + (I - P_k) \tag{2.184}$$

   $$= P_k + (I - P_k) = I \tag{2.185}$$

   Since the identity operator I (which maps $\mathbb{R}^n$ to itself) is the sum of mutually orthogonal projection operators, the space $\mathbb{R}^n$ decomposes into the direct sum of their respective image subspaces V_i.

□

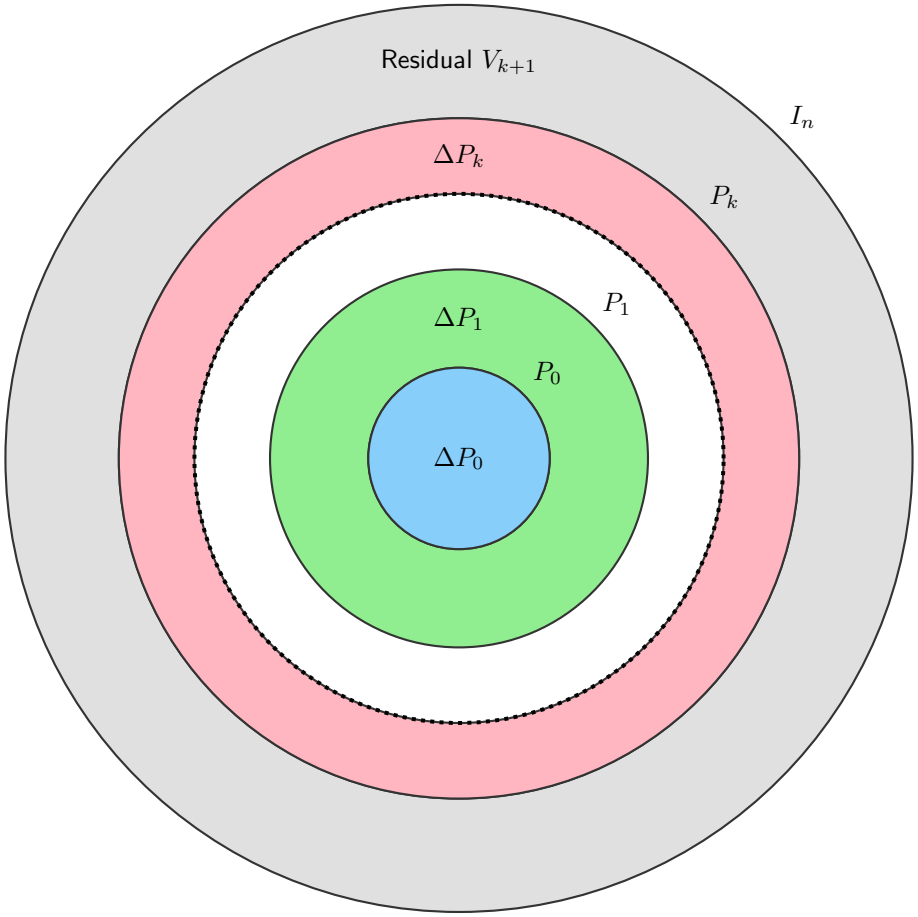


Figure 2.7: Venn Diagram of Nested Projections with Colored Increments

3 Matrix Algebra

This chapter covers a review of matrix algebra concepts essential for linear models, including eigenvalues, spectral decomposition, singular value decomposition.

3.1 Eigenvalues and Eigenvectors

Definition 3.1 (Eigenvalues and Eigenvectors)**.** For a square matrix A $(n \times n)$, a scalar λ is an **eigenvalue** and a non-zero vector x is the corresponding **eigenvector** if:

$$Ax = \lambda x \iff (A - \lambda I_n)x = 0 \tag{3.1}$$

The eigenvalues are found by solving the characteristic equation:

$$|A - \lambda I_n| = 0 \tag{3.2}$$

3.2 Spectral Theory for Symmetric Matrices

3.2.1 Spectral Decomposition

For symmetric matrices, we have a powerful decomposition theorem.

Theorem 3.1 (Spectral Decomposition)**.** *If A is a symmetric $n \times n$ matrix, all its eigenvalues $\lambda_1, \dots, \lambda_n$ are real. Furthermore, there exists an orthogonal matrix Q such that:*

$$A = Q\Lambda Q' = \sum_{i=1}^{n} \lambda_i q_i q_i' \tag{3.3}$$

where:

- $\Lambda = diag(\lambda_1, \dots, \lambda_n)$ *contains the eigenvalues.*
- $Q = (q_1, \dots, q_n)$ *contains the corresponding orthonormal eigenvectors ($q_i' q_j = \delta_{ij}$).*

Explantion: This allows us to view the transformation Ax as a rotation (Q'), a scaling (Λ), and a rotation back (Q). For a symmetric matrix A, we can write the spectral decomposition as a product of the eigenvector matrix Q and eigenvalue matrix Λ:

$$
\begin{aligned}
A &= Q\Lambda Q' \\
&= \begin{pmatrix} q_1 & q_2 & \cdots & q_n \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \begin{pmatrix} q_1' \\ q_2' \\ \vdots \\ q_n' \end{pmatrix} \\
&= \begin{pmatrix} \lambda_1 q_1 & \lambda_2 q_2 & \cdots & \lambda_n q_n \end{pmatrix} \begin{pmatrix} q_1' \\ q_2' \\ \vdots \\ q_n' \end{pmatrix} \\
&= \lambda_1 q_1 q_1' + \lambda_2 q_2 q_2' + \cdots + \lambda_n q_n q_n' \\
&= \sum_{i=1}^{n} \lambda_i q_i q_i'
\end{aligned}
\tag{3.4}
$$

where the eigenvectors q_i satisfy the orthogonality conditions:

$$
q_i' q_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{3.5}
$$

And Q is an orthogonal matrix: $Q'Q = QQ' = I_n$.

3.2.2 Quadratic Form

Definition 3.2. A **quadratic form** in n variables $x_1, x_2, \ldots, x_n$ is a scalar function defined by a symmetric matrix A:

$$
Q(x) = x'Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \tag{3.6}
$$

3.2.3 Positive and Non-Negative Definite Matrices

Definition 3.3 (Positive and Non-Negative Definite Matrices)**.** A symmetric matrix A is **positive definite (p.d.)** if:

$$
x'Ax > 0 \quad \forall x \neq 0 \tag{3.7}
$$

It is **non-negative definite (n.n.d.)** if:

$$
x'Ax \geq 0 \quad \forall x \tag{3.8}
$$

Theorem 3.2 (Properties of Definite Matrices)**.** *Let A be a symmetric $n \times n$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$.*

1. ***Eigenvalue Characterization:***
   - *A is p.d. $\iff$ all $\lambda_i > 0$.*
   - *A is n.n.d. $\iff$ all $\lambda_i \geq 0$.*

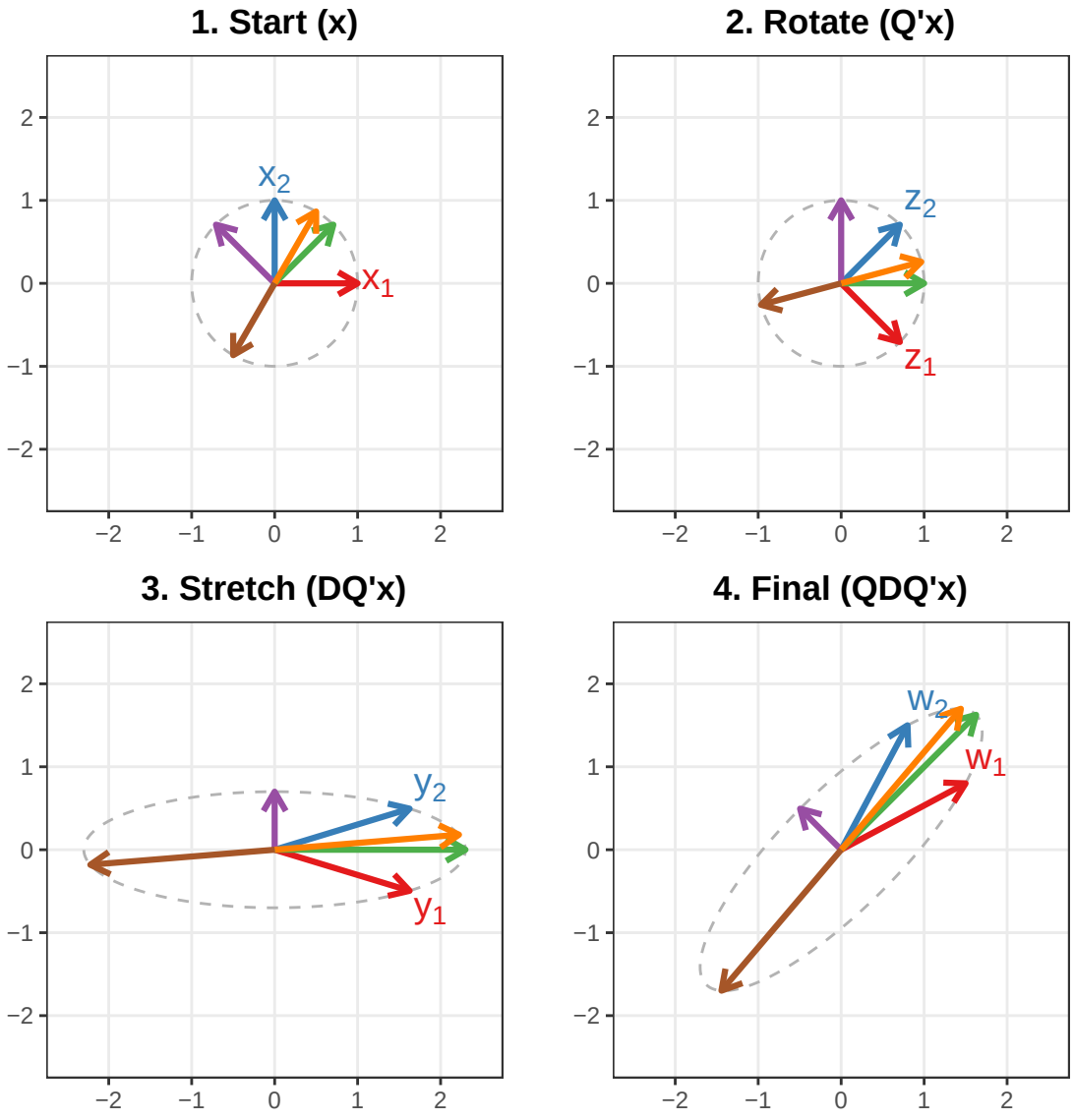

1. Start (x)
2. Rotate (Q'x)
3. Stretch (DQ'x)
4. Final (QDQ'x)
x2
x1
z2
z1
y2
y1
w2
w1

Figure 3.1

2. ***Determinant and Inverse:***
    - *If A is p.d., then $|A| > 0$ and A^{-1} exists.*
    - *If A is n.n.d. and singular, then $|A| = 0$ (at least one $\lambda_i = 0$).*
3. ***Gram Matrices ($B'B$):*** *Let B be an $n \times p$ matrix.*
    - *If rank$(B) = p$, then $B'B$ is p.d.*
    - *If rank$(B) < p$, then $B'B$ is n.n.d.*

3.2.4 Properties of Symmetric Matrices

Theorem 3.3 (Properties of Symmetric Matrices)**.** *Let A be a symmetric matrix with spectral decomposition $A = Q\Lambda Q'$. The following properties hold:*

1. ***Trace:*** *tr$(A) = \sum \lambda_i$.*
2. ***Determinant:*** *$|A| = \prod \lambda_i$.*
3. ***Singularity:*** *A is singular if and only if at least one $\lambda_i = 0$.*
4. ***Inverse:*** *If A is non-singular ($\lambda_i \neq 0$), then $A^{-1} = Q\Lambda^{-1}Q'$.*
5. ***Powers:*** *$A^k = Q\Lambda^k Q'$.*
    - Square Root: *$A^{1/2} = Q\Lambda^{1/2}Q'$ (if $\lambda_i \geq 0$).*
6. ***Spectral Representation of Quadratic Forms:*** *The quadratic form $x'Ax$ can be diagonalized using the eigenvectors of A:*
$$x'Ax = x'Q\Lambda Q'x = y'\Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2 \tag{3.9}$$
where $y = Q'x$ represents a rotation of the coordinate system.

3.2.5 Spectral Representation of Projection Matrices

We revisit projection matrices in the context of eigenvalues.

Theorem 3.4 (Eigenvalues of Projection Matrices)**.** *A symmetric matrix P is a projection matrix (idempotent, $P^2 = P$) if and only if its eigenvalues are either 0 or 1.*

$$P^2x = \lambda^2 x \quad \textit{and} \quad Px = \lambda x \implies \lambda^2 = \lambda \implies \lambda \in \{0, 1\} \tag{3.10}$$

For a projection matrix P:

- If $x \in \text{Col}(P)$, $Px = x$ (Eigenvalue 1).
- If $x \perp \text{Col}(P)$, $Px = 0$ (Eigenvalue 0).
- rank$(P) = \text{tr}(P) = \sum \lambda_i$ (Count of 1s).

Example 3.1. For $P = \frac{1}{n} J_n J_n'$, the rank is $\text{tr}(P) = 1$.

3.3 Singular Value Decomposition (SVD)

Theorem 3.5 (Singular Value Decomposition (SVD))**.** *Let X be an $n \times p$ matrix with rank $r \leq \min(n, p)$. X can be decomposed into the product of three matrices:*

$$X = U\mathbf{D}V' \tag{3.11}$$

1. *Partitioned Matrix Form*

$$X = \underset{n\times n}{(U_1, U_2)} \begin{pmatrix} \Lambda_r & O_{r\times(p-r)} \\ O_{(n-r)\times r} & O_{(n-r)\times(p-r)} \end{pmatrix} \underset{p\times p}{\begin{pmatrix} V_1' \\ V_2' \end{pmatrix}} \tag{3.12}$$

2. *Detailed Matrix Form*

Expanding the diagonal matrix explicitly:

$$X = \underset{n\times n}{(u_1, \ldots, u_n)} \left(\begin{array}{cccc|c} \lambda_1 & 0 & \ldots & 0 & \\ 0 & \lambda_2 & \ldots & 0 & O_{12} \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \ldots & \lambda_r & \\ \hline \multicolumn{4}{c|}{O_{21}} & O_{22} \end{array}\right) \underset{p\times p}{\begin{pmatrix} v_1' \\ \vdots \\ v_p' \end{pmatrix}} \tag{3.13}$$

3. *Reduced Form*

$$X = U_1 \Lambda_r V_1' = \sum_{i=1}^{r} \lambda_i u_i v_i' \tag{3.14}$$

Properties:

1. ***Singular Values (Λ_r):*** *$\Lambda_r = diag(\lambda_1, \ldots, \lambda_r)$ contains the singular values ($\lambda_i > 0$), which are the square roots of the non-zero eigenvalues of $X'X$.*
2. ***Orthogonality:***
   - *U is $n \times n$ orthogonal ($U'U = I_n$).*
   - *V is $p \times p$ orthogonal ($V'V = I_p$).*

3.3.0.1 Connection to Gram Matrices

The matrices U and V provide the basis vectors (eigenvectors) for the Gram matrices of X.

1. **Right Singular Vectors (V):** The columns of V are the eigenvectors of the Gram matrix $X'X$.

$$X'X = (U\Lambda V')'(U\Lambda V') = V\Lambda U'U\Lambda V' = V\Lambda^2 V' \tag{3.15}$$

   - The eigenvalues of $X'X$ are the squared singular values λ_i^2.

2. **Left Singular Vectors (U):** The columns of U are the eigenvectors of the Gram matrix XX'.

$$XX' = (U\Lambda V')(U\Lambda V')' = U\Lambda V'V\Lambda U' = U\Lambda^2 U' \tag{3.16}$$

   - The eigenvalues of XX' are also λ_i^2 (for non-zero values).

Example 3.2 (Example of SVD)**.** Consider the matrix $X = \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}$.

1. **Compute $X'X$ and find V:**

$$X'X = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 5 \\ 5 & 5 \end{pmatrix} \tag{3.17}$$

   - Eigenvalues of $X'X$: Trace is 10, Determinant is 0. Thus, $\mu_1 = 10, \mu_2 = 0$.
   - **Singular Values:** $\lambda_1 = \sqrt{10}, \lambda_2 = 0$.
   - Eigenvector for $\mu_1 = 10$: Normalized $v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.
   - Eigenvector for $\mu_2 = 0$: Normalized $v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$.
   - Therefore, $V = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$.

2. **Compute XX' and find U:**

$$XX' = \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 4 \\ 4 & 8 \end{pmatrix} \tag{3.18}$$

   - Eigenvalues are again 10 and 0.
   - Eigenvector for $\mu_1 = 10$: Normalized $u_1 = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ 2 \end{pmatrix}$.
   - Eigenvector for $\mu_2 = 0$: Normalized $u_2 = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 \\ -1 \end{pmatrix}$.
   - Therefore, $U = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 2 \\ 2 & -1 \end{pmatrix}$.

3. **Verification:**

$$X = \sqrt{10}u_1 v_1' = \sqrt{10}\begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix} \tag{3.19}$$

3.4 Cholesky Decomposition

A symmetric matrix A has a Cholesky decomposition if and only if it is **non-negative definite** (i.e., $x'Ax \geq 0$ for all x).

$$A = B'B \tag{3.20}$$

where B is an **upper triangular** matrix with non-negative diagonal entries.

3.4.1 Matrix Representation of the Algorithm

To derive the algorithm, we equate the elements of A with the product of the lower triangular matrix B' and the upper triangular matrix B.

For a 3×3 matrix, this looks like:

$$\underbrace{\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} b_{11} & 0 & 0 \\ b_{12} & b_{22} & 0 \\ b_{13} & b_{23} & b_{33} \end{pmatrix}}_{B'} \underbrace{\begin{pmatrix} b_{11} & b_{12} & b_{13} \\ 0 & b_{22} & b_{23} \\ 0 & 0 & b_{33} \end{pmatrix}}_{B} \tag{3.21}$$

Multiplying the matrices on the right yields the system of equations:

$$A = \begin{pmatrix} \mathbf{b_{11}^2} & b_{11}b_{12} & b_{11}b_{13} \\ b_{12}b_{11} & \mathbf{b_{12}^2 + b_{22}^2} & b_{12}b_{13} + b_{22}b_{23} \\ b_{13}b_{11} & b_{13}b_{12} + b_{23}b_{22} & \mathbf{b_{13}^2 + b_{23}^2 + b_{33}^2} \end{pmatrix} \tag{3.22}$$

By solving for the bolded diagonal terms and substituting known values from previous rows, we get the recursive algorithm.

Algorithm

1. **Row 1:** Solve for b_{11} using a_{11}, then solve the rest of the row (b_{1j}) by division.
   - $b_{11} = \sqrt{a_{11}}$
   - $b_{1j} = a_{1j}/b_{11}$
2. **Row 2:** Solve for b_{22} using a_{22} and the known b_{12}, then solve b_{2j}.
   - $b_{22} = \sqrt{a_{22} - b_{12}^2}$
   - $b_{2j} = (a_{2j} - b_{12}b_{1j})/b_{22}$
3. **Row 3:** Solve for b_{33} using a_{33} and the known b_{13}, b_{23}.
   - $b_{33} = \sqrt{a_{33} - b_{13}^2 - b_{23}^2}$

Remark. **Handling the Singular Case**

If A is positive semi-definite (singular), a diagonal element b_{ii} may evaluate to 0 (or a very small number close to 0 due to floating-point error). Standard algorithms often crash here because calculating off-diagonal terms involves division by b_{ii}.

To handle this robustly without pivoting:

- If $b_{ii} \approx 0$, it implies that the entire remaining row $b_{i,i:n}$ must be 0 for the matrix to remain consistent with being positive semi-definite.
- The algorithm should explicitly set $b_{ij} = 0$ for all $j \geq i$ and proceed to the next row, rather than attempting division.

Example 3.3 (Example of Cholesky Decomposition)**.** Consider the positive definite matrix A:

$$A = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 10 & 2 \\ -2 & 2 & 6 \end{pmatrix} \tag{3.23}$$

We find B such that $A = B'B$:

1. **First Row of B (b_{11}, b_{12}, b_{13}):**
   - $b_{11} = \sqrt{4} = 2$
   - $b_{12} = 2/2 = 1$
   - $b_{13} = -2/2 = -1$
2. **Second Row of B (b_{22}, b_{23}):**
   - $b_{22} = \sqrt{10 - (1)^2} = \sqrt{9} = 3$
   - $b_{23} = (2 - (1)(-1))/3 = 3/3 = 1$
3. **Third Row of B (b_{33}):**
   - $b_{33} = \sqrt{6 - (-1)^2 - (1)^2} = \sqrt{4} = 2$

Result:

$$B = \begin{pmatrix} 2 & 1 & -1 \\ 0 & 3 & 1 \\ 0 & 0 & 2 \end{pmatrix} \tag{3.24}$$

3.4.2 Applications in Statistics

Cholesky decomposition is preferred over other methods (like LU or SVD) for symmetric positive-definite matrices because it is numerically stable and roughly twice as fast.

1. Solving Linear Equations

In linear regression, we solve the normal equations $(X'X)\beta = X'y$. Since $X'X$ is symmetric and positive definite, we can decompose it as $B'B$. The system becomes:

$$B'B\beta = X'y \tag{3.25}$$

This allows us to solve for β using two efficient triangular substitutions (first solving $B'z = X'y$ for z, then $B\beta = z$ for β) without explicitly inverting the matrix, which is computationally expensive and unstable.

2. Computing the Determinant

   The determinant of a triangular matrix is simply the product of its diagonal entries. Therefore, the determinant of A can be computed instantly from B:

$$\det(A) = \det(B'B) = \det(B')\det(B) = \left(\prod_{i=1}^{n} b_{ii}\right)^2 \tag{3.26}$$

   This is widely used in Maximum Likelihood Estimation (e.g., REML in mixed models) where log-determinants of large covariance matrices are required.

3. Generating Multivariate Normal Random Variables

   To generate a random vector $Y \sim N(\mu, \Sigma)$, we first generate a vector of independent standard normal variables $Z \sim N(0, I)$. Using the Cholesky decomposition $\Sigma = B'B$:

$$Y = \mu + B'Z \tag{3.27}$$

   The covariance of Y is confirmed by $\text{Cov}(Y) = B'\text{Cov}(Z)B = B'IB = B'B = \Sigma$. This is the standard method used by functions like `mvrnorm` in R.

4 Multivariate Normal Distribution

4.1 Motivation

Consider the linear model:

$$y = X\beta + \epsilon, \quad \epsilon_i \sim N(0, \sigma^2) \tag{4.1}$$

We are often interested in the distributional properties of the response vector y and the residuals. Specifically, if $y = (y_1, \dots, y_n)'$, we need to understand its multivariate distribution.

$$\hat{y} = Py, \quad e = y - \hat{y} = (I_n - P)y \tag{4.2}$$

4.2 Random Vectors and Matrices

Definition 4.1 (Random Vector and Matrix)**.** A **Random Vector** is a vector whose elements are random variables. E.g.,

$$x_{k\times 1} = (x_1, x_2, \dots, x_k)^T \tag{4.3}$$

where $x_1, \dots, x_k$ are each random variables.
A **Random Matrix** is a matrix whose elements are random variables. E.g., $X_{n\times k} = (x_{ij})$, where $x_{11}, \dots, x_{nk}$ are each random variables.

Definition 4.2 (Expected Value)**.** The expected value (population mean) of a random matrix (or vector) is the matrix (or vector) of expected values of its elements.
For $X_{n\times k}$:

$$E(X) = \begin{pmatrix} E(x_{11}) & \dots & E(x_{1k}) \\ \vdots & \ddots & \vdots \\ E(x_{n1}) & \dots & E(x_{nk}) \end{pmatrix} \tag{4.4}$$

$$E\left(\begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}\right) = \begin{pmatrix} E(x_1) \\ \vdots \\ E(x_k) \end{pmatrix} \tag{4.5}$$

Definition 4.3 (Variance-Covariance Matrix)**.** For a random vector $x_{k\times 1} = (x_1, \dots, x_k)^T$, the matrix is:

$$\text{Var}(x) = \Sigma_x = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kk} \end{pmatrix} \tag{4.6}$$

Where:

- $\sigma_{ij} = \text{Cov}(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)]$
- $\sigma_{ii} = \text{Var}(x_i) = E[(x_i - \mu_i)^2]$

In matrix notation:

$$\text{Var}(x) = E[(x - \mu_x)(x - \mu_x)^T] \tag{4.7}$$

Note: $\text{Var}(x)$ is symmetric.

4.2.1 Derivation of Covariance Matrix Structure

Expanding the vector multiplication for variance:

$$(x - \mu_x)(x - \mu_x)' \quad \text{where } \mu_x = (\mu_1, \dots, \mu_n)' \tag{4.8}$$

$$= \begin{pmatrix} x_1 - \mu_1 \\ \vdots \\ x_n - \mu_n \end{pmatrix} (x_1 - \mu_1, \dots, x_n - \mu_n) \tag{4.9}$$

This results in the matrix $A = (a_{ij})$ where $a_{ij} = (x_i - \mu_i)(x_j - \mu_j)$. Taking expectations yields the covariance matrix elements σ_{ij}.

Definition 4.4 (Covariance Matrix (Two Vectors))**.** For random vectors $x_{k\times 1}$ and $y_{n\times 1}$, the covariance matrix is:

$$\text{Cov}(x, y) = E[(x - \mu_x)(y - \mu_y)^T] = \begin{pmatrix} \text{Cov}(x_1, y_1) & \dots & \text{Cov}(x_1, y_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(x_k, y_1) & \dots & \text{Cov}(x_k, y_n) \end{pmatrix} \tag{4.10}$$

Note that $\text{Cov}(x, x) = \text{Var}(x)$.

Definition 4.5 (Correlation Matrix)**.** The correlation matrix of a random vector x is:

$$\text{corr}(x) = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ \vdots & \ddots & \vdots & \\ \rho_{k1} & \rho_{k2} & \dots & 1 \end{pmatrix} \tag{4.11}$$

where $\rho_{ij} = \text{corr}(x_i, x_j)$.
Relationships: Let $V_x = \text{diag}(\text{Var}(x_1), \dots, \text{Var}(x_k))$.

$$\Sigma_x = V_x^{1/2} \rho_x V_x^{1/2} \quad \text{and} \quad \rho_x = (V_x^{1/2})^{-1} \Sigma_x (V_x^{1/2})^{-1} \tag{4.12}$$

Similarly for two vectors:

$$\Sigma_{xy} = V_x^{1/2} \rho_{xy} V_y^{1/2} \tag{4.13}$$

4.3 Properties of Mean and Variance

We can derive several key algebraic properties for operations on random vectors.

1. $E(X+Y) = E(X) + E(Y)$
2. $E(AXB) = AE(X)B$ (In particular, $E(AX) = A\mu_x$)
3. $\text{Cov}(x, y) = \text{Cov}(y, x)^T$
4. $\text{Cov}(x + c, y + d) = \text{Cov}(x, y)$
5. $\text{Cov}(Ax, By) = A\text{Cov}(x, y)B^T$
    - Special case for scalars: $\text{Cov}(ax, by) = ab \cdot \text{Cov}(x, y)$
6. $\text{Cov}(x_1 + x_2, y_1) = \text{Cov}(x_1, y_1) + \text{Cov}(x_2, y_1)$
7. $\text{Var}(x + c) = \text{Var}(x)$
8. $\text{Var}(Ax) = A\text{Var}(x)A^T$
9. $\text{Var}(x_1 + x_2) = \text{Var}(x_1) + \text{Cov}(x_1, x_2) + \text{Cov}(x_2, x_1) + \text{Var}(x_2)$
10. $\text{Var}(\sum x_i) = \sum \text{Var}(x_i)$ if independent.

Proof. **Property 5 (Covariance of Linear Transformation):**

$$\begin{aligned} \text{Cov}(Ax, By) &= E[(Ax - A\mu_x)(By - B\mu_y)^T] \\ &= AE[(x - \mu_x)(y - \mu_y)^T]B^T \\ &= A\text{Cov}(x, y)B^T \end{aligned} \tag{4.14}$$

Property 2 (Expectation of Linear Transformation):
To prove $E(AXB) = AE(X)B$: First consider $E(Ax_j)$ where x_j is a column of X.

$$E(Ax_j) = E\begin{pmatrix} a_1'x_j \\ \vdots \\ a_n'x_j \end{pmatrix} = \begin{pmatrix} E(a_1'x_j) \\ \vdots \\ E(a_n'x_j) \end{pmatrix} \tag{4.15}$$

Since a_i are constants:

$$E(a_i'x_j) = E\left(\sum_{k=1}^{p} a_{ik}x_{kj}\right) = \sum_{k=1}^{p} a_{ik}E(x_{kj}) = a_i'E(x_j) \tag{4.16}$$

Thus $E(Ax_j) = AE(x_j)$. Applying this to all columns of X:

$$E(AX) = [E(Ax_1), \ldots, E(Ax_m)] = [AE(x_1), \ldots, AE(x_m)] = AE(X) \tag{4.17}$$

Similarly, $E(XB) = E(X)B$.
Proof of Property 9 (Variance of Sum):

$$\text{Var}(x_1 + x_2) = E[(x_1 + x_2 - \mu_1 - \mu_2)(x_1 + x_2 - \mu_1 - \mu_2)^T] \tag{4.18}$$

Let centered variables be denoted by differences.

$$= E[((x_1 - \mu_1) + (x_2 - \mu_2))((x_1 - \mu_1) + (x_2 - \mu_2))^T] \tag{4.19}$$

Expanding terms:

$$= E[(x_1-\mu_1)(x_1-\mu_1)^T + (x_1-\mu_1)(x_2-\mu_2)^T + (x_2-\mu_2)(x_1-\mu_1)^T + (x_2-\mu_2)(x_2-\mu_2)^T] \quad (4.20)$$

$$= \text{Var}(x_1) + \text{Cov}(x_1, x_2) + \text{Cov}(x_2, x_1) + \text{Var}(x_2) \quad (4.21)$$

□

4.4 The Multivariate Normal Distribution

4.4.1 Definition and Density

Definition 4.6 (Independent Standard Normal)**.** Let $z = (z_1, \dots, z_n)'$ where $z_i \sim N(0,1)$ are independent. We say $z \sim N_n(0, I_n)$. The joint PDF is the product of marginals:

$$f(z) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2} z^T z} \quad (4.22)$$

Properties: $E(z) = 0$ and $\text{Var}(z) = I_n$ (Covariance is 0 for $i \neq j$, Variance is 1).

Definition 4.7 (Multivariate Normal Distribution)**.** A random vector x ($n \times 1$) has a **multivariate normal distribution** if it has the same distribution as:

$$x = A_{n\times p} z_{p\times 1} + \mu_{n\times 1} \quad (4.23)$$

where $z \sim N_p(0, I_p)$, A is a matrix of constants, and μ is a vector of constants. The moments are:

- $E(x) = \mu$
- $\text{Var}(x) = AA^T = \Sigma$

4.4.2 Geometric Interpretation

Using Spectral Decomposition, $\Sigma = Q\Lambda Q'$. We can view the transformation $x = Az + \mu$ as:

1. Scaling by eigenvalues ($\Lambda^{1/2}$).
2. Rotation by eigenvectors (Q).
3. Shift by mean (μ).

4.4.3 Probability Density Function

If Σ is positive definite, the PDF exists. We use the change of variable formula for $x = Az + \mu$:

$$f_x(x) = f_z(g^{-1}(x)) \cdot |J| \quad (4.24)$$

where $z = A^{-1}(x-\mu)$ and $J = \det(A^{-1}) = |A|^{-1}$.

$$f_x(x) = (2\pi)^{-p/2}|A|^{-1}\exp\left\{-\frac{1}{2}(A^{-1}(x-\mu))^T(A^{-1}(x-\mu))\right\} \tag{4.25}$$

Using $|\Sigma| = |AA^T| = |A|^2$ and $\Sigma^{-1} = (AA^T)^{-1}$, we get:

$$f_x(x) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \tag{4.26}$$

4.4.4 Moment Generating Function

Definition 4.8 (Moment Generating Function (MGF))**.** The MGF of a random vector x is $M_x(t) = E(e^{t^T x})$. For $x = Az + \mu$:

$$M_x(t) = E[e^{t^T(Az+\mu)}] = e^{t^T\mu}E[e^{(A^Tt)^Tz}] = e^{t^T\mu}M_z(A^Tt) \tag{4.27}$$

Since $M_z(u) = e^{u^Tu/2}$:

$$M_x(t) = e^{t^T\mu}\exp\left(\frac{1}{2}t^T(AA^T)t\right) = \exp\left(t^T\mu + \frac{1}{2}t^T\Sigma t\right) \tag{4.28}$$

Key Properties:

1. **Uniqueness:** Two random vectors with the same MGF have the same distribution.
2. **Independence:** y_1 and y_2 are independent iff $M_y(t) = M_{y_1}(t_1)M_{y_2}(t_2)$.

4.5 Construction and Linear Transformations

Theorem 4.1 (Constructing MVN Random Vector)**.** *Let $\mu \in \mathbb{R}^n$ and Σ be an $n \times n$ symmetric non-negative definitive (n.n.d) matrix. Then there exists a multivariate normal distribution with mean μ and covariance Σ.*

Proof. Since Σ is n.n.d., there exists B such that $\Sigma = BB^T$ (e.g., via Cholesky or Spetral Decomposition). Let $z \sim N_n(0, I)$ and define $x = Bz + \mu$. □

Theorem 4.2 (Linear Transformation Theorem)**.** *Let $x \sim N_n(\mu, \Sigma)$. Let $y = Cx + d$ where C is $r \times n$ and d is $r \times 1$. Then:*

$$y \sim N_r(C\mu + d, C\Sigma C^T) \tag{4.29}$$

Proof. $x = Az + \mu$ where $AA^T = \Sigma$.

$$y = C(Az + \mu) + d = (CA)z + (C\mu + d) \tag{4.30}$$

This fits the definition of MVN with mean $C\mu + d$ and variance $C\Sigma C^T$. □

4.5.1 Important Corollaries of Theorem 4.2

Corollary 4.1 (Marginals). *Any subvector of a multivariate normal vector is also multivariate normal.*

Proof. If we partition $x = (x_1', x_2')'$, we can use $C = (I_r, 0)$ to show $x_1 \sim N(\mu_1, \Sigma_{11})$. □

Corollary 4.2 (Univariate Combinations). *Any linear combination $a^T x$ is univariate normal:*

$$a^T x \sim N(a^T \mu, a^T \Sigma a) \tag{4.31}$$

Corollary 4.3 (Orthogonal Transformations). *If $x \sim N(0, I_n)$ and Q is orthogonal ($Q'Q = I$), then $y = Q'x \sim N(0, I_n)$.*

Corollary 4.4 (Standardization). *If $y \sim N_n(\mu, \Sigma)$ and Σ is positive definite:*

$$\Sigma^{-1/2}(y - \mu) \sim N_n(0, I_n) \tag{4.32}$$

Proof. Let $z = \Sigma^{-1/2}(y - \mu)$. Then $\text{Var}(z) = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_n$. □

4.6 Independence

Theorem 4.3 (Independence in MVN). *Let $y \sim N(\mu, \Sigma)$ be partitioned into y_1 and y_2.*

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{4.33}$$

Then y_1 and y_2 are independent if and only if $\Sigma_{12} = 0$ (zero covariance).

Proof.

1. Independence $\implies$ Covariance is 0: This holds generally for any distribution.

$$\text{Cov}(y_1, y_2) = E[(y_1 - \mu_1)(y_2 - \mu_2)'] = 0 \tag{4.34}$$

2. Covariance is 0 $\implies$ Independence: This is specific to MVN. We use MGFs. If $\Sigma_{12} = 0$, the quadratic form in the MGF splits:

$$t^T \Sigma t = t_1^T \Sigma_{11} t_1 + t_2^T \Sigma_{22} t_2 \tag{4.35}$$

The MGF becomes:

$$M_y(t) = \exp(t_1^T \mu_1 + \frac{1}{2} t_1^T \Sigma_{11} t_1) \times \exp(t_2^T \mu_2 + \frac{1}{2} t_2^T \Sigma_{22} t_2) \tag{4.36}$$

$$M_y(t) = M_{y_1}(t_1)M_{y_2}(t_2) \tag{4.37}$$

Thus, they are independent.

□

4.7 Signal-Noise Decomposition for Multivariate Normal Distribution

We can formalize the relationship between two random vectors y and x through a decomposition theorem that separates the systematic signal from the stochastic noise.

Theorem 4.4 (Regression Decomposition Theorem)**.** *Let the random vector V of dimension $p \times 1$ be partitioned into two subvectors y ($p_1 \times 1$) and x ($p_2 \times 1$). Assume V follows a multivariate normal distribution:*

$$\begin{pmatrix} y \\ x \end{pmatrix} \sim N_p \left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} \right) \tag{4.38}$$

The response vector y can be uniquely decomposed into a systematic component and a stochastic error:

$$y = m(x) + e \tag{4.39}$$

*where we define the **Regression Coefficient Matrix** B and the components as:*

$$B = \Sigma_{yx}\Sigma_{xx}^{-1} \tag{4.40}$$

$$m(x) = \mu_y + B(x - \mu_x) \tag{4.41}$$

$$e = y - m(x) \tag{4.42}$$

Properties:

1. ***Independence:*** *The noise vector e is statistically independent of the predictor x (and consequently independent of $m(x)$).*
2. ***Marginal Distributions:***
   - $m(x) \sim N_{p_1}(\mu_y,\ B\Sigma_{xx}B^T)$
   - $e \sim N_{p_1}(0,\ \Sigma_{yy} - B\Sigma_{xx}B^T)$
3. ***Conditional Distribution:*** *Since $y = m(x) + e$, and e is independent of x, the conditional distribution is:*

$$y|x \sim N_{p_1}(m(x), \Sigma_{y|x}) \tag{4.43}$$

where:

$$m(x) = \mu_y + B(x - \mu_x) = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x) \tag{4.44}$$

$$\Sigma_{y|x} = \Sigma_{yy} - B\Sigma_{xx}B^T = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} \tag{4.45}$$

Proof. We define a transformation from the input vector $V = \begin{pmatrix} y \\ x \end{pmatrix}$ to the target vector $W = \begin{pmatrix} m(x) \\ e \end{pmatrix}$.
Using the linear transformation $W = \text{CV} + d$:

$$\underbrace{\begin{pmatrix} m(x) \\ e \end{pmatrix}}_{W} = \underbrace{\begin{pmatrix} 0 & B \\ I & -B \end{pmatrix}}_{C} \underbrace{\begin{pmatrix} y \\ x \end{pmatrix}}_{V} + \underbrace{\begin{pmatrix} \mu_y - B\mu_x \\ -(\mu_y - B\mu_x) \end{pmatrix}}_{d} \tag{4.46}$$

1. Mean Vector

$$E[W] = CE[V] + d = \begin{pmatrix} 0 & B \\ I & -B \end{pmatrix} \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} + \begin{pmatrix} \mu_y - B\mu_x \\ -\mu_y + B\mu_x \end{pmatrix} = \begin{pmatrix} B\mu_x \\ \mu_y - B\mu_x \end{pmatrix} + \begin{pmatrix} \mu_y - B\mu_x \\ -\mu_y + B\mu_x \end{pmatrix} = \begin{pmatrix} \mu_y \\ 0 \end{pmatrix} \tag{4.47}$$

2. Covariance Matrix

We compute $\text{Var}(W) = C\Sigma C^T$ directly:

$$\begin{aligned} C\Sigma C^T &= \begin{pmatrix} 0 & B \\ I & -B \end{pmatrix} \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} \begin{pmatrix} 0 & I \\ B^T & -B^T \end{pmatrix} \\ &= \begin{pmatrix} B\Sigma_{xy} & B\Sigma_{xx} \\ \Sigma_{yy} - B\Sigma_{xy} & \Sigma_{yx} - B\Sigma_{xx} \end{pmatrix} \begin{pmatrix} 0 & I \\ B^T & -B^T \end{pmatrix} \\ &= \begin{pmatrix} B\Sigma_{xx}B^T & B\Sigma_{xy} - B\Sigma_{xx}B^T \\ \Sigma_{yx}B^T - B\Sigma_{xx}B^T & (\Sigma_{yy} - B\Sigma_{xy}) - (\Sigma_{yx} - B\Sigma_{xx})B^T \end{pmatrix} \\ &= \begin{pmatrix} B\Sigma_{xx}B^T & 0 \\ 0 & \Sigma_{yy} - B\Sigma_{xx}B^T \end{pmatrix} \end{aligned} \tag{4.48}$$

3. Conditional Distribution

We have established that $y = m(x) + e$ where e is independent of x. To find the distribution of y conditional on x, we observe that $m(x)$ becomes a constant vector when x is fixed, and the randomness comes solely from e:

$$E[y|x] = m(x) + E[e|x] = m(x) + 0 = m(x) \tag{4.49}$$

$$\text{Var}(y|x) = \text{Var}(m(x)|x) + \text{Var}(e|x) = 0 + \text{Var}(e) = \Sigma_{y|x} \tag{4.50}$$

Thus, $y|x \sim N(m(x), \Sigma_{y|x})$. □

4.7.1 Connections with Other Formulas

4.7.1.1 Rao-Blackwell Decomposition of Variance

The Law of Total Variance (Rao-Blackwell theorem) allows us to decompose the total variance of y into two orthogonal components based on the predictor x:

$$\text{Var}(y) = \underbrace{E[\text{Var}(y|x)]}_{\text{Unexplained (Noise)}} + \underbrace{\text{Var}[E(y|x)]}_{\text{Explained (Signal)}} \tag{4.51}$$

In the Multivariate Normal case, this decomposition perfectly aligns with our regression model $y = m(x) + e$.

Variance of Noise

This term represents the average variance remaining in y after accounting for x. It corresponds to the variance of the error term e:

$$E[\text{Var}(y|x)] = \text{Var}(e) = \Sigma_{yy} - B\Sigma_{xx}B^T \tag{4.52}$$

Variance of Signal

This term represents the variability of the conditional mean $m(x)$ itself. Using the matrix B, this takes the quadratic form:

$$\text{Var}[E(y|x)] = \text{Var}[m(x)] = B\Sigma_{xx}B^T \tag{4.53}$$

Total Variance

Summing the Signal and Noise components recovers the total marginal variance of y:

$$\Sigma_{yy} = \underbrace{\Sigma_{yy} - B\Sigma_{xx}B^T}_{\text{Unexplained (Noise)}} + \underbrace{B\Sigma_{xx}B^T}_{\text{Explained (Signal)}} \tag{4.54}$$

4.7.1.2 Connection to OLS Regression Estimators

In OLS regression, centering the data allows us to separate the intercept from the slopes. Let $\mathbf{y}_c$ and $\mathbf{X}_c$ be the centered response and design matrices (where $\mathbf{X}_c$ **excludes the column of 1s**). Using this centered form, the total sum of squares decomposes exactly like the population variance:

$$\text{SST} = \text{SSR} + \text{SSE} \tag{4.55}$$

Comparing the sample quantities to their population counterparts:

1. **Regression Coefficients:**
$$\hat{\beta}^T = (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y}_c \approx B \tag{4.56}$$
Note: $\hat{\beta}$ here represents only the slope coefficients, matching the dimensions of the covariance matrix Σ_{xx}.

2. **Explained Variation (Signal):**

$$\text{SSR} = \hat{\beta}^T(\mathbf{X}_c^T\mathbf{X}_c)\hat{\beta} \quad \approx \quad (n-1)B\Sigma_{xx}B^T \tag{4.57}$$

3. **Unexplained Variation (Noise):**

$$\text{SSE} = \mathbf{y}_c^T\mathbf{y}_c - \hat{\beta}^T(\mathbf{X}_c^T\mathbf{X}_c)\hat{\beta} \quad \approx \quad (n-1)(\Sigma_{yy} - B\Sigma_{xx}B^T) \tag{4.58}$$

4.8 Partial and Multiple Correlation

Definition 4.9 (Partial Correlation)**.** The partial correlation between elements y_i and y_j given a set of variables x is derived from the conditional covariance matrix $\Sigma_{y|x}$:

$$\rho_{ij|x} = \frac{\sigma_{ij|x}}{\sqrt{\sigma_{ii|x}\sigma_{jj|x}}} \tag{4.59}$$

where $\sigma_{ij|x}$ are elements of $\Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$.

Definition 4.10 (Multiple Correlation (R^2))**.** For a scalar y and vector x, the squared multiple correlation is the proportion of variance of y explained by the conditional mean:

$$R^2_{y|x} = \frac{\text{Var}(E(y|x))}{\text{Var}(y)} = \frac{\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}}{\sigma_y^2} \tag{4.60}$$

Note: this definition is the population or theretical R^2, which is estimated by adjusted R^2 using sample in linear regression.

4.9 Examples

Example 4.1 (Bivariate Normal)**.** Let the random vector $\begin{pmatrix} y \\ x \end{pmatrix}$ follow a bivariate normal distribution:

$$\begin{pmatrix} y \\ x \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix}\right) \tag{4.61}$$

Here, $\mu_y = 1, \mu_x = 2, \Sigma_{yy} = 2, \Sigma_{xx} = 4$, and $\Sigma_{yx} = 2$.

1. Finding the Regression Coefficient Matrix B Using the population formula:

$$B = \Sigma_{yx}\Sigma_{xx}^{-1} = 2(4)^{-1} = 0.5 \tag{4.62}$$

2. Finding the Conditional Mean $m(x)$ (The Signal) The systematic component represents the projection of

y onto x:

$$\begin{aligned} m(x) &= \mu_y + B(x - \mu_x) \\ &= 1 + 0.5(x - 2) = 0.5x \end{aligned} \tag{4.63}$$

3. Variance of the Signal $\text{Var}(m(x))$ Using the quadratic form established in the theorem:

$$\text{Var}(m(x)) = B\Sigma_{xx}B^T = 0.5(4)(0.5) = 1 \tag{4.64}$$

4. Variance of the Noise $\text{Var}(y|x)$ (The Residual) By the Signal-Noise Decomposition:

$$\begin{aligned} \text{Var}(y|x) &= \Sigma_{yy} - \text{Var}(m(x)) \\ &= 2 - 1 = 1 \end{aligned} \tag{4.65}$$

Thus, $y|x \sim N(m(x), 1)$. The total variance (2) is split equally between signal (1) and noise (1).

5. Multiple Correlation Coefficient (R^2)

$$R^2 = \frac{\text{Var}(m(x))}{\Sigma_{yy}} = \frac{1}{2} = 0.5 \tag{4.66}$$

Example 4.2 (Trivariate Normal with 2 Predictors)**.** Let $V = (y, x_1, x_2)' \sim N_3(\mu, \Sigma)$ with:

$$\mu = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 10 & 3 & 4 \\ 3 & 2 & 1 \\ 4 & 1 & 4 \end{pmatrix} \tag{4.67}$$

We partition these into $\Sigma_{yy} = 10$, $\Sigma_{yx} = \begin{pmatrix} 3 & 4 \end{pmatrix}$, and $\Sigma_{xx} = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$.

1. Finding the Regression Coefficient Matrix B

$$\Sigma_{xx}^{-1} = \frac{1}{7}\begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix} \implies B = \Sigma_{yx}\Sigma_{xx}^{-1} = \begin{pmatrix} \frac{8}{7} & \frac{5}{7} \end{pmatrix} \tag{4.68}$$

2. Finding the Conditional Mean $m(x)$ (The Signal)

$$m(x) = 1 + \frac{8}{7}(x_1 - 2) + \frac{5}{7}(x_2 - 3) \tag{4.69}$$

3. Variance of the Signal $\text{Var}(m(x))$

$$\text{Var}(m(x)) = B\Sigma_{xx}B^T = \begin{pmatrix} \frac{8}{7} & \frac{5}{7} \end{pmatrix}\begin{pmatrix} 3 \\ 4 \end{pmatrix} = \frac{44}{7} \approx 6.29 \tag{4.70}$$

4. Variance of the Noise $\text{Var}(y|x)$ (The Residual) Using the Signal-Noise Decomposition:

$$\Sigma_{y|x} = \Sigma_{yy} - \text{Var}(m(x)) = 10 - 6.29 = 3.71 \tag{4.71}$$

5. Multiple Correlation Coefficient (R^2)

$$R^2 = \frac{6.29}{10} = 0.629 \tag{4.72}$$

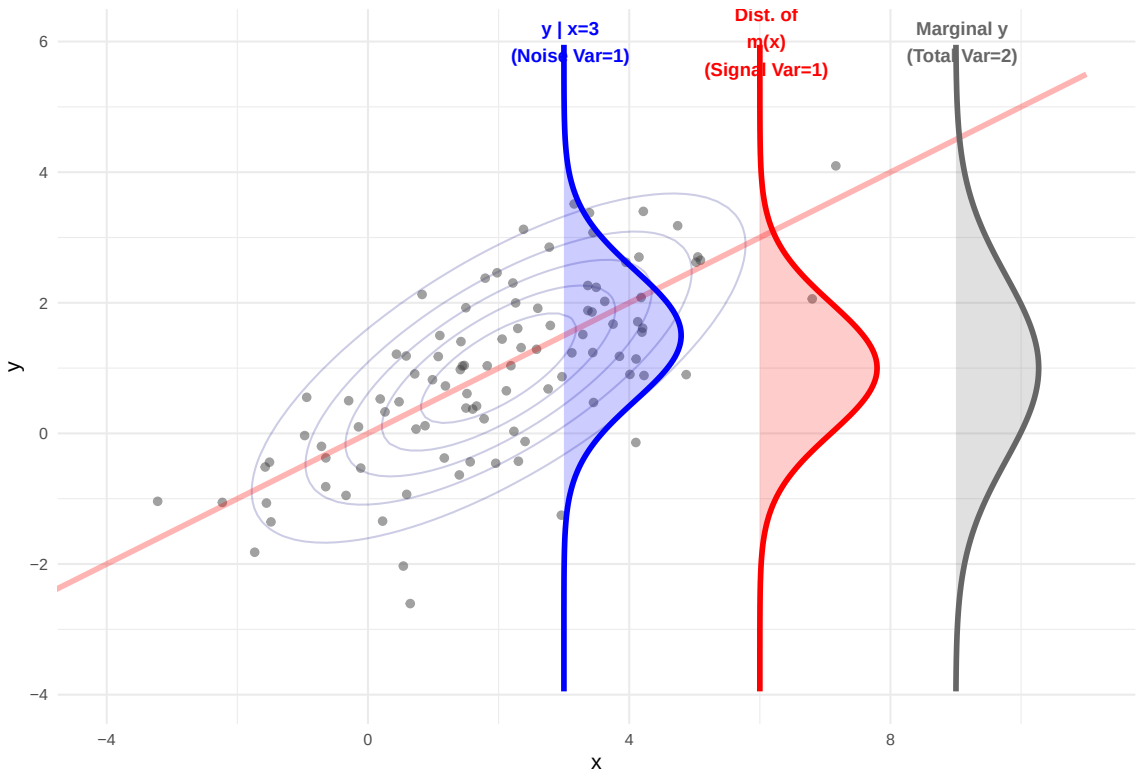


Figure 4.1: Illustration of Rao-Blackwell Variance Decomposition in Bivariate Normal

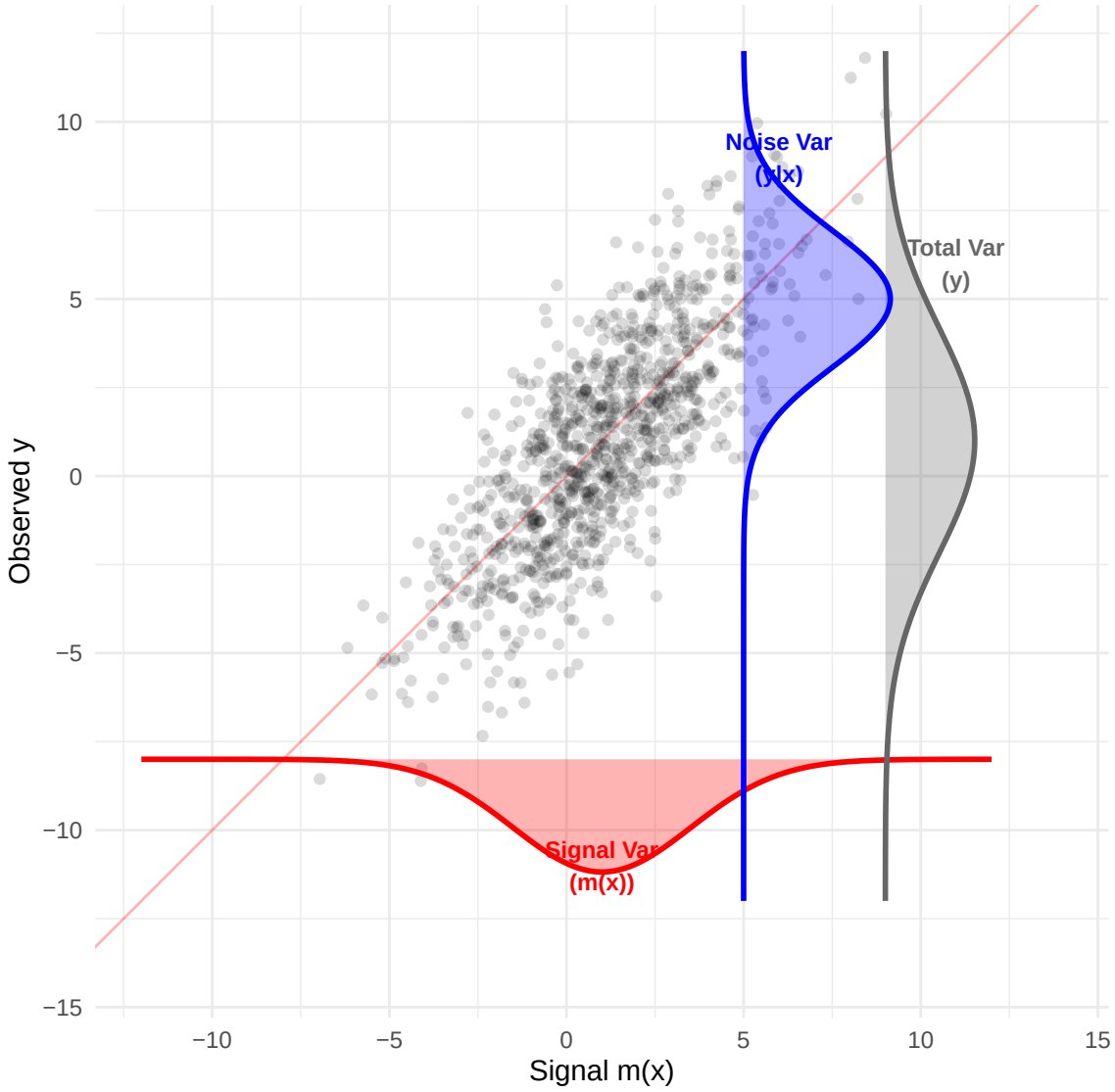


Figure 4.2: Signal-Noise Variance Decomposition in Multivariate Normal

5 Distribution of Quadratic Forms

This chapter covers the distribution of quadratic forms (sums of squares), which is crucial for hypothesis testing in linear models.

5.1 Quadratic Forms

A quadratic form is a polynomial with terms all of degree two.

Definition 5.1 (Quadratic Form)**.** Let $y = (y_1, \dots, y_n)'$ be a random vector and A be a symmetric $n \times n$ matrix. The scalar quantity $y'Ay$ is called a **quadratic form** in y.

$$y'Ay = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}y_iy_j \tag{5.1}$$

Examples:

- **Squared Norm:** If $A = I_n$, then $y'I_ny = y'y = \sum y_i^2 = ||y||^2$.
- **Weighted Sum of Squares:** If A is diagonal with elements λ_i, then $y'Ay = \sum \lambda_i y_i^2$.
- **Projection Sum of Squares:** If P is a projection matrix, $||Py||^2 = (Py)'(Py) = y'P'Py = y'Py$ (since P is symmetric and idempotent).

5.2 Mean of Quadratic Forms

We can find the expected value of a quadratic form without assuming normality.

Lemma 5.1 (Mean of Simplified Quadratic Form)**.** *If y is a random vector with mean $E(y) = \mu$ and covariance matrix $Var(y) = I_n$, then:*

$$E(y'y) = tr(I_n) + \mu'\mu = n + \mu'\mu \tag{5.2}$$

Proof. Let us decompose y into its mean and a stochastic component: $y = \mu + z$, where $E(z) = 0$ and $\text{Var}(z) = E(zz') = I_n$. Substituting this into the quadratic form:

$$\begin{aligned} y'y &= (\mu + z)'(\mu + z) \\ &= \mu'\mu + \mu'z + z'\mu + z'z \\ &= \mu'\mu + 2\mu'z + z'z \end{aligned} \tag{5.3}$$

Taking the expectation:

$$\begin{aligned} E(y'y) &= \mu'\mu + 2\mu' E(z) + E(z'z) \\ &= \mu'\mu + 0 + E\left(\sum_{i=1}^{n} z_i^2\right) \end{aligned} \tag{5.4}$$

Since $\text{Var}(z_i) = E(z_i^2) - (E(z_i))^2 = 1 - 0 = 1$, we have $E(\sum z_i^2) = \sum 1 = n$. Thus, $E(y'y) = n + \mu'\mu$. □

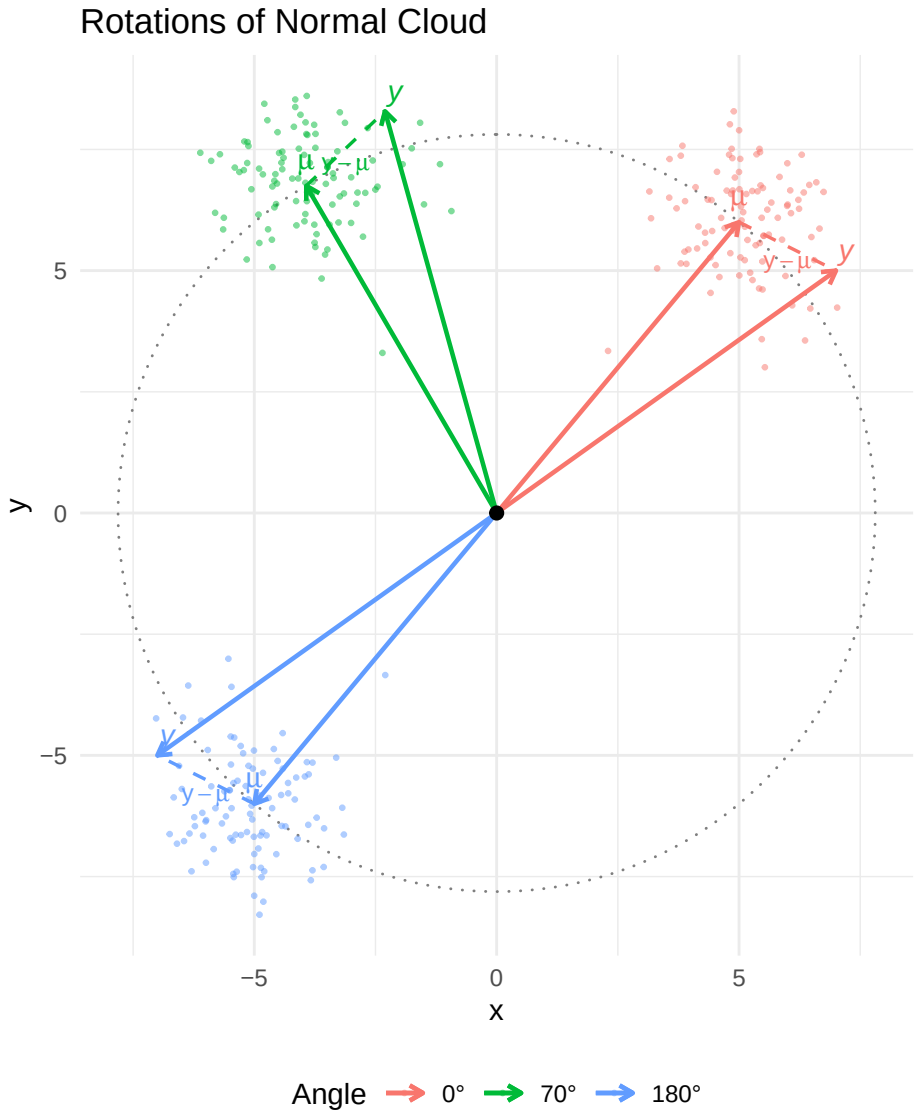


Figure 5.1: Illustration of the Mean and Distribution of Quadratic Forms

Theorem 5.1 (Mean of Quadratic Form)**.** *If y is a random vector with mean $E(y) = \mu$ and covariance matrix $Var(y) = \Sigma$, and A is a symmetric matrix of constants, then:*

$$E(y'Ay) = tr(A\Sigma) + \mu' A\mu \tag{5.5}$$

Proof. We present three methods to derive the expectation of the quadratic form.

Method 1: Using the Trace Trick

Using the fact that a scalar is equal to its own trace ($\text{tr}(c) = c$) and the linearity of expectation:

$$\begin{aligned} E(y'Ay) &= E[\text{tr}(y'Ay)] \\ &= E[\text{tr}(Ayy')] \quad \text{(cyclic property of trace)} \\ &= \text{tr}(AE[yy']) \quad \text{(linearity of expectation)} \end{aligned} \tag{5.6}$$

Recall that the covariance matrix is defined as $\Sigma = E[(y-\mu)(y-\mu)'] = E(yy') - \mu\mu'$. Rearranging this gives the second moment: $E(yy') = \Sigma + \mu\mu'$. Substituting this back:

$$\begin{aligned} E(y'Ay) &= \text{tr}(A(\Sigma + \mu\mu')) \\ &= \text{tr}(A\Sigma) + \text{tr}(A\mu\mu') \\ &= \text{tr}(A\Sigma) + \text{tr}(\mu' A\mu) \quad \text{(cyclic property on second term)} \\ &= \text{tr}(A\Sigma) + \mu' A\mu \end{aligned} \tag{5.7}$$

Method 2: Using Scalar Summation

We can express the quadratic form in scalar notation using the entries of $A = (a_{ij})$, $\Sigma = (\sigma_{ij})$, and $\mu = (\mu_i)$:

$$\begin{aligned} E(y'Ay) &= E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}y_i y_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}E(y_i y_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(\sigma_{ij} + \mu_i\mu_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\sigma_{ji} + \sum_{i=1}^{n}\sum_{j=1}^{n} \mu_i a_{ij}\mu_j \quad (\text{since } \Sigma \text{ is symmetric, } \sigma_{ij} = \sigma_{ji}) \\ &= \text{tr}(A\Sigma) + \mu' A\mu \end{aligned} \tag{5.8}$$

Method 3: Using Spectral Decomposition of A

Since A is symmetric, we use its spectral decomposition $A = \sum_{i=1}^{n} \lambda_i q_i q_i'$. Substituting this into the quadratic form:

$$y'Ay = y'\left(\sum_{i=1}^{n} \lambda_i q_i q_i'\right) y = \sum_{i=1}^{n} \lambda_i (q_i' y)^2 \tag{5.9}$$

Let $w_i = q_i' y$. This is a scalar random variable which is a linear transformation of y. Its properties are:

1. **Mean:** $E(w_i) = q_i' E(y) = q_i'\mu$.

2. **Variance:** $\text{Var}(w_i) = \text{Var}(q_i'y) = q_i'\text{Var}(y)q_i = q_i'\Sigma q_i$.

Using the relation $E(w_i^2) = \text{Var}(w_i) + [E(w_i)]^2$, we have:

$$E[(q_i'y)^2] = q_i'\Sigma q_i + (q_i'\mu)^2 \tag{5.10}$$

Summing over all i weighted by λ_i:

$$\begin{aligned} E(y'Ay) &= \sum_{i=1}^{n} \lambda_i \left[q_i'\Sigma q_i + (q_i'\mu)^2\right] \\ &= \sum_{i=1}^{n} \text{tr}(\lambda_i q_i'\Sigma q_i) + \mu' \left(\sum_{i=1}^{n} \lambda_i q_i q_i'\right) \mu \\ &= \text{tr}\left(\Sigma \sum_{i=1}^{n} \lambda_i q_i q_i'\right) + \mu' A\mu \\ &= \text{tr}(\Sigma A) + \mu' A\mu \end{aligned} \tag{5.11}$$

□

Remark (Geometric Interpretation via Sigma). If we further decompose $\Sigma = \sum_{j=1}^{n} \gamma_j v_j v_j'$ (where γ_j, v_j are eigenvalues/vectors of Σ), the trace term becomes:

$$\text{tr}(A\Sigma) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \gamma_j (q_i'v_j)^2 \tag{5.12}$$

Here, $(q_i'v_j)^2 = \cos^2(\theta_{ij})$ represents the alignment between the axes of the quadratic form (A) and the axes of the data covariance (Σ). The expectation is maximized when the eigenspaces of A and Σ align.

Corollary 5.1 (Expectation with Projection Matrix)**.** *Consider the special case where:*

1. *P is a **projection matrix** (symmetric and idempotent, $P^2 = P$).*
2. *The covariance is **spherical**: $\Sigma = \sigma^2 I_n$.*

Then the expectation simplifies to:

$$E(y'Py) = \sigma^2 r + ||P\mu||^2 \tag{5.13}$$

where $r = rank(P) = tr(P)$.
Proof: *Using Theorem 5.1 with $A = P$ and $\Sigma = \sigma^2 I_n$:*

1. ***Trace Term:*** *$tr(P\Sigma) = tr(P(\sigma^2 I_n)) = \sigma^2 tr(P)$. Since P is idempotent, its eigenvalues are either 0 or 1, so $tr(P) = rank(P) = r$.*
2. ***Mean Term:*** *Since P is symmetric and idempotent ($P'P = P^2 = P$), we can rewrite the quadratic form:*

$$\mu' P\mu = \mu' P' P\mu = (P\mu)'(P\mu) = ||P\mu||^2 \tag{5.14}$$

Example 5.1 (Expectation of Sum of Squares Decomposition (i.i.d. Case))**.** Consider a random vector $y = (y_1, \dots, y_n)'$ with mean vector $\mu_y = \mu j_n$ and covariance $\Sigma = \sigma^2 I_n$. We analyze the two components of the total sum of squares by projecting y onto the mean space (P_{j_n}) and the residual space ($I - P_{j_n}$).

1. The Projection Vectors

First, we write the explicit forms of the projected vectors using $P_{j_n} = \frac{1}{n} j_n j_n'$:

- **Mean Vector ($P_{j_n} y$):** Projecting y onto the column space of j_n replaces every element with the sample mean $\bar{y}$.

$$P_{j_n} y = \bar{y} j_n = \begin{pmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix} \tag{5.15}$$

- **Residual Vector ($(I - P_{j_n})y$):** Subtracting the mean projection from y yields the deviations.

$$(I - P_{j_n})y = y - \bar{y} j_n = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} \tag{5.16}$$

2. Expectations of Squared Norms

We now find the expectation of the squared length of these vectors using Corollary 5.1.
Part A: Sum of Squares for Mean The quadratic form is the squared norm of the projected mean vector:

$$y' P_{j_n} y = ||P_{j_n} y||^2 = \sum_{i=1}^{n} \bar{y}^2 = n\bar{y}^2 \tag{5.17}$$

Applying the corollary with $P = P_{j_n}$:

- **Rank:** $\text{tr}(P_{j_n}) = 1$.
- **Mean:** $P_{j_n} \mu_y = P_{j_n}(\mu j_n) = \mu j_n$. The squared norm is $n\mu^2$.

$$E[||P_{j_n} y||^2] = \sigma^2(1) + n\mu^2 \tag{5.18}$$

Part B: Sum of Squared Errors (SSE) The quadratic form is the squared norm of the residual vector:

$$y'(I - P_{j_n})y = ||(I - P_{j_n})y||^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{5.19}$$

Applying the corollary with $P = I - P_{j_n}$:

- **Rank:** $\text{tr}(I - P_{j_n}) = n - 1$.
- **Mean:** $(I - P_{j_n})\mu_y = \mu_y - P_{j_n}\mu_y = \mu j_n - \mu j_n = 0$. The squared norm is 0.

$$E[||(I - P_{j_n})y||^2] = \sigma^2(n-1) + 0 \tag{5.20}$$

Conclusion These results confirm the standard properties: $E(\bar{y}^2) = \frac{\sigma^2}{n} + \mu^2$ and $E(S^2) = \sigma^2$.

Example 5.2 (Expectation of Total Sum of Squares (Regression Case))**.** Consider now a regression setting where the mean of y depends on covariates (e.g., $\mu_i = \beta_0 + \beta_1 x_i$). The mean vector μ_y is **not** proportional to j_n. We are interested in the expectation of the **Total Sum of Squares (SST)**.

1. Identification The SST measures the variation of y around the *global sample mean* $\bar{y}$, ignoring the covariates:

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = y'(I - P_{j_n})y \tag{5.21}$$

   This is the same quadratic form as Part B in the previous example, but the underlying mean μ_y has changed.

2. Calculation We apply Corollary 5.1 with $P = I - P_{j_n}$ and general μ_y:

- **Rank Term:** Same as before, $\text{tr}(I - P_{j_n}) = n - 1$.
- **Mean Term:** The projection of the mean vector is no longer zero.

$$(I - P_{j_n})\mu_y = \mu_y - \bar{\mu}j_n = \begin{pmatrix} \mu_1 - \bar{\mu} \\ \vdots \\ \mu_n - \bar{\mu} \end{pmatrix} \tag{5.22}$$

  where $\bar{\mu} = \frac{1}{n}\sum \mu_i$ is the average of the true means. The squared norm is the sum of squared deviations of the true means:

$$||(I - P_{j_n})\mu_y||^2 = \sum_{i=1}^{n}(\mu_i - \bar{\mu})^2 \tag{5.23}$$

Conclusion

$$E(\text{SST}) = (n-1)\sigma^2 + \sum_{i=1}^{n}(\mu_i - \bar{\mu})^2 \tag{5.24}$$

This shows that in regression, the SST estimates $(n-1)\sigma^2$ *plus* the variability introduced by the regression signal (the spread of the true means μ_i).

5.3 Non-central χ^2 Distribution

To understand the distribution of quadratic forms under normality, we introduce the non-central chi-square distribution.

Definition 5.2 (Non-central χ^2 Distribution)**.** Let $y \sim N_n(\mu, I_n)$. The random variable $V = y'y = \sum y_i^2$ follows a **non-central chi-square distribution** with n degrees of freedom and non-centrality parameter λ.

$$V \sim \chi^2(n, \lambda) \quad \text{where } \lambda = \mu'\mu = ||\mu||^2 \tag{5.25}$$

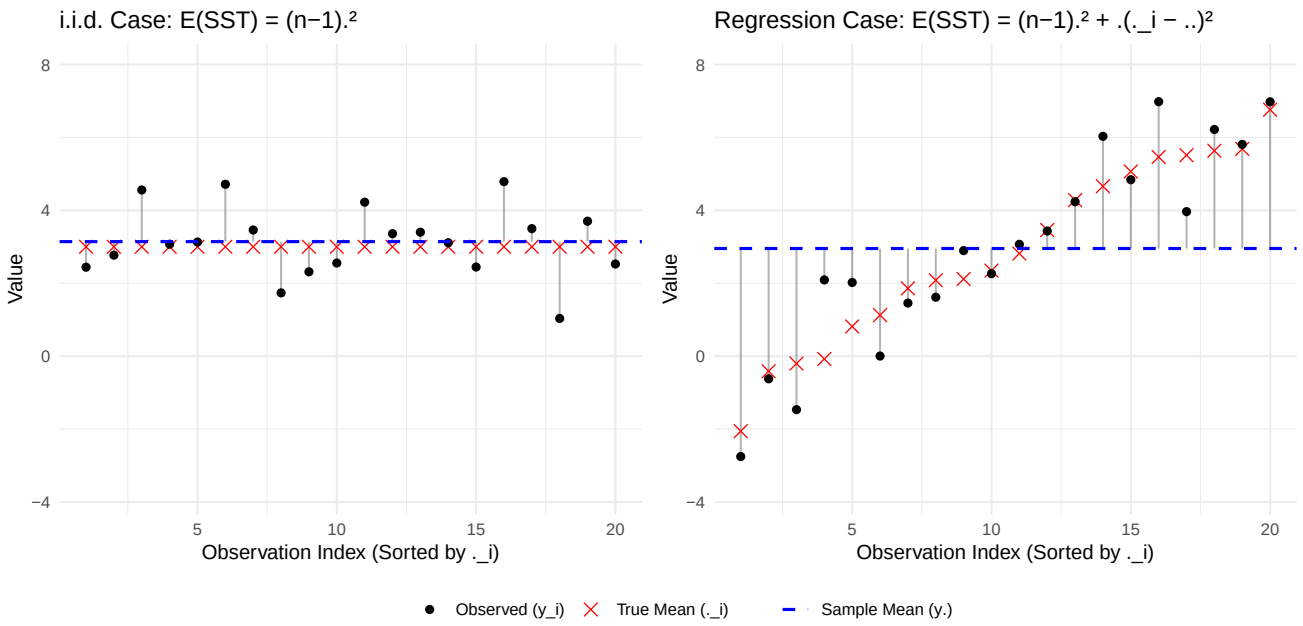


Figure 5.2: Comparison of SST components with increased variation in the true means. The vertical lines represent the deviations $(y_i - \bar{y})$. With $\text{sd}(\mu_i) = 3$, the regression case (right) shows significantly larger deviations, illustrating how the systematic spread of the means dominates the Total Sum of Squares.

> **!** Important
>
> **Note on NCP Definition:** Some definitions of non-central χ^2 use $\lambda = \frac{1}{2}\mu'\mu$. In this course, we use $\lambda = \mu'\mu$. With this convention, the Poisson-mixture representation below uses Poisson$(\lambda/2)$ weights.

5.3.1 Visualizing χ^2 Distributions

Here is a plot visualizing the difference between central and non-central Chi-square distributions.

The density of the non-central chi-square distribution shifts to the right and becomes flatter as the non-centrality parameter λ increases.

5.3.2 Mean, Variance, and MGF

We summarize the key properties of the non-central chi-square distribution.

Theorem 5.2 (Properties of Non-central Chi-square)**.** *Let* $V \sim \chi^2(n, \lambda)$*. Then:*

1. ***Mean:*** $E(V) = n + \lambda$
2. ***Variance:*** $Var(V) = 2n + 4\lambda$

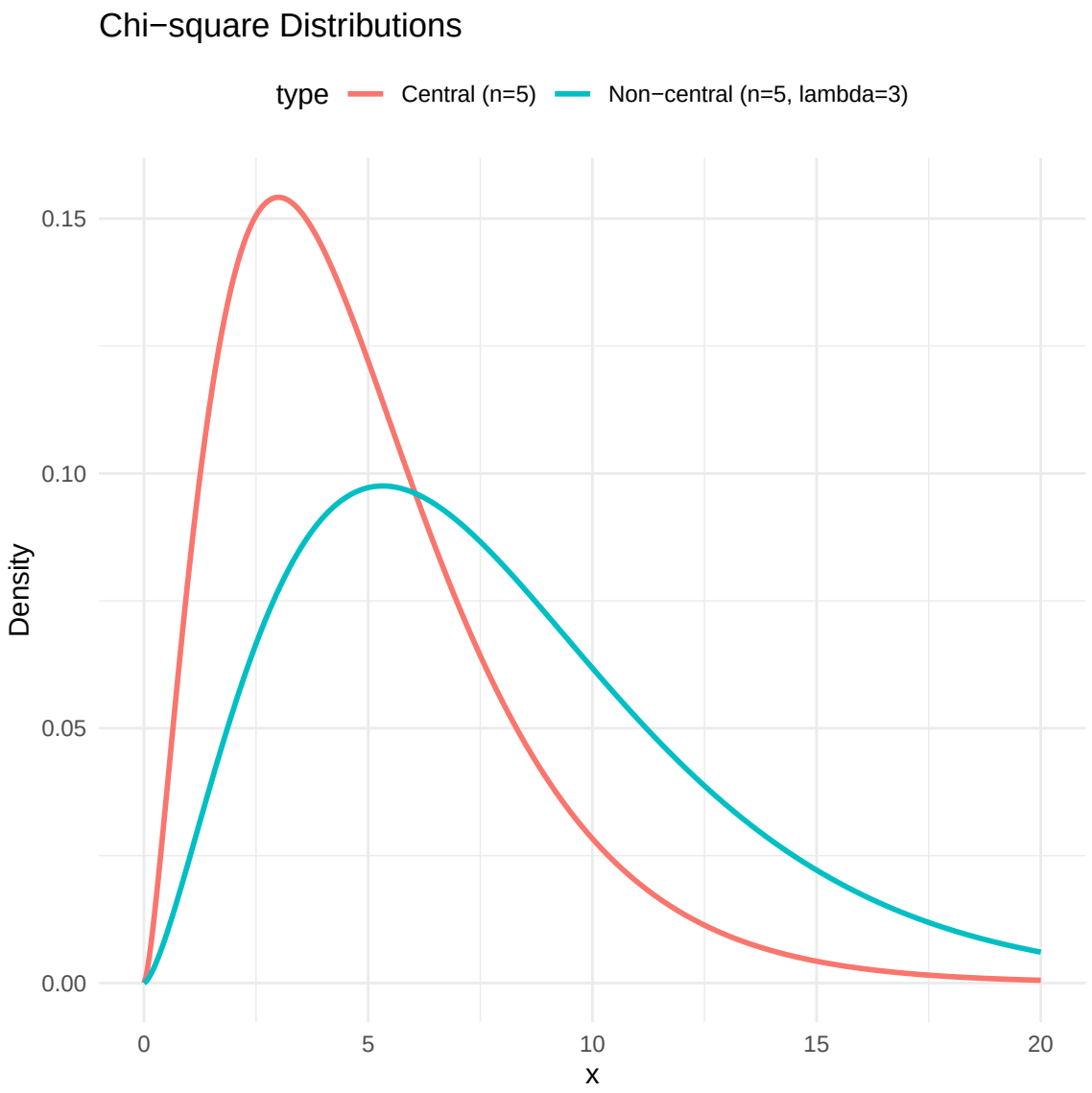


Figure 5.3: Central vs Non-central Chi-square Distribution

3. ***Moment Generating Function (MGF):***

$$m_V(t) = \frac{\exp\left[-\frac{\lambda}{2}\left\{1 - \frac{1}{1-2t}\right\}\right]}{(1-2t)^{n/2}} \quad \textit{for } t < 1/2 \tag{5.26}$$

Mean. By definition, $V \sim \chi^2(n, \lambda)$ is the distribution of $y'y$ where $y \sim N_n(\mu, I_n)$ and the non-centrality parameter is $\lambda = \mu'\mu = ||\mu||^2$. Applying Lemma 5.1 to the random vector y:

$$E(V) = E(y'y) = n + \mu'\mu = n + \lambda \tag{5.27}$$

□

MGF. Since the components y_i of the vector y are independent $N(\mu_i, 1)$, and $V = \sum_{i=1}^n y_i^2$, the MGF of V is the product of the MGFs of each y_i^2:

$$m_V(t) = E[e^{t\sum y_i^2}] = \prod_{i=1}^n E[e^{ty_i^2}] \tag{5.28}$$

Consider a single component $y_i \sim N(\mu_i, 1)$. Its squared expectation is:

$$\begin{aligned} E[e^{ty_i^2}] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{ty^2} e^{-\frac{1}{2}(y-\mu_i)^2} dy \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[(1-2t)y^2 - 2\mu_i y + \mu_i^2\right]\right\} dy \end{aligned} \tag{5.29}$$

Completing the square in the exponent for y (assuming $t < 1/2$):

$$(1-2t)y^2 - 2\mu_i y + \mu_i^2 = (1-2t)\left(y - \frac{\mu_i}{1-2t}\right)^2 + \mu_i^2 - \frac{\mu_i^2}{1-2t} \tag{5.30}$$

The integral of the Gaussian kernel $\exp\{-\frac{1}{2}(1-2t)(y-\ldots)^2\}$ yields $\sqrt{\frac{2\pi}{1-2t}}$. The remaining constant term is:

$$\exp\left\{-\frac{1}{2}\left(\mu_i^2 - \frac{\mu_i^2}{1-2t}\right)\right\} = \exp\left\{\frac{\mu_i^2}{2}\left(\frac{1}{1-2t} - 1\right)\right\} = \exp\left\{\frac{\mu_i^2 t}{1-2t}\right\} \tag{5.31}$$

Thus, for a single component:

$$m_{y_i^2}(t) = (1-2t)^{-1/2} \exp\left(\frac{\mu_i^2 t}{1-2t}\right) \tag{5.32}$$

Multiplying the MGFs for all n components:

$$\begin{aligned} m_V(t) &= \prod_{i=1}^n (1-2t)^{-1/2} \exp\left(\frac{\mu_i^2 t}{1-2t}\right) \\ &= (1-2t)^{-n/2} \exp\left(\frac{t\sum \mu_i^2}{1-2t}\right) \end{aligned} \tag{5.33}$$

Substituting $\lambda = \sum \mu_i^2$ (so $\sum \mu_i^2 = \lambda$):

$$m_V(t) = (1-2t)^{-n/2} \exp\left(\frac{\lambda t}{1-2t}\right) \tag{5.34}$$

Note that $\frac{\lambda t}{1-2t} = -\frac{\lambda}{2}\left(1 - \frac{1}{1-2t}\right)$, which leads to the Poisson-mixture representation with $J \sim \text{Poisson}(\lambda/2)$. □

Variance. We use the **Cumulant Generating Function**, $K_V(t) = \ln m_V(t)$, as its derivatives yield the mean and variance directly:

$$K_V(t) = -\frac{n}{2}\ln(1-2t) + \frac{\lambda t}{1-2t} \tag{5.35}$$

First derivative (Mean):

$$\begin{aligned} K_V'(t) &= -\frac{n}{2}\left(\frac{-2}{1-2t}\right) + \lambda\left[\frac{1(1-2t) - t(-2)}{(1-2t)^2}\right] \\ &= \frac{n}{1-2t} + \frac{\lambda}{(1-2t)^2} \end{aligned} \tag{5.36}$$

Second derivative (Variance):

$$\begin{aligned} K_V''(t) &= n(-1)(1-2t)^{-2}(-2) + \lambda(-2)(1-2t)^{-3}(-2) \\ &= \frac{2n}{(1-2t)^2} + \frac{4\lambda}{(1-2t)^3} \end{aligned} \tag{5.37}$$

Evaluating at $t = 0$:

$$\text{Var}(V) = K_V''(0) = 2n + 4\lambda \tag{5.38}$$

□

5.3.3 Additivity

Theorem 5.3 (Additivity of Chi-square)**.** *If $v_1, \ldots, v_k$ are independent random variables distributed as $\chi^2(n_i, \lambda_i)$, then their sum follows a chi-square distribution:*

$$\sum_{i=1}^{k} v_i \sim \chi^2\left(\sum_{i=1}^{k} n_i, \sum_{i=1}^{k} \lambda_i\right) \tag{5.39}$$

Proof. **Method 1: Using MGFs**
The moment generating function of $v_i \sim \chi^2(n_i, \lambda_i)$ is:

$$M_{v_i}(t) = \frac{\exp\left[-\frac{\lambda_i}{2}\left(1 - \frac{1}{1-2t}\right)\right]}{(1-2t)^{n_i/2}} \tag{5.40}$$

Since $v_1, \ldots, v_k$ are independent, the MGF of their sum $V = \sum v_i$ is the product of their individual MGFs:

$$\begin{aligned} M_V(t) &= \prod_{i=1}^{k} M_{v_i}(t) \\ &= \prod_{i=1}^{k} \frac{\exp\left[-\frac{\lambda_i}{2}\left(1 - \frac{1}{1-2t}\right)\right]}{(1-2t)^{n_i/2}} \\ &= \frac{\exp\left[-\frac{\sum \lambda_i}{2}\left(1 - \frac{1}{1-2t}\right)\right]}{(1-2t)^{\sum n_i/2}} \end{aligned} \tag{5.41}$$

This is the MGF of a non-central chi-square distribution with degrees of freedom $\sum n_i$ and non-centrality parameter $\sum \lambda_i$.

Method 2: Geometric Interpretation

Let $v_i = ||y_i||^2$ where $y_i \sim N_{n_i}(\mu_i, I_{n_i})$. Since the vectors y_i are independent, we can stack them into a larger vector $y = (y_1', \ldots, y_k')'$.

$$y \sim N_{\sum n_i}(\mu, I_{\sum n_i}) \quad \text{where } \mu = (\mu_1', \ldots, \mu_k')' \tag{5.42}$$

The sum of squares is:

$$\sum v_i = \sum ||y_i||^2 = ||y||^2 \tag{5.43}$$

By definition, $||y||^2$ follows a non-central chi-square distribution with degrees of freedom equal to the dimension of y ($\sum n_i$) and non-centrality parameter $\lambda = ||\mu||^2$.

$$\lambda = \sum_{i=1}^{k} ||\mu_i||^2 = \sum_{i=1}^{k} \lambda_i \tag{5.44}$$

□

5.3.4 Poisson Mixture Representation

Theorem 5.4 (Poisson Mixture Representation)**.** *Let $v \sim \chi^2(n, \lambda)$ be a non-central chi-square random variable. Its probability density function can be represented as a Poisson-weighted sum of central chi-square density functions:*

$$f(v; n, \lambda) = \sum_{j=0}^{\infty} \left(\frac{e^{-\lambda/2}(\lambda/2)^j}{j!} \right) f(v; n + 2j, 0) \tag{5.45}$$

where $f(v; \nu, 0)$ is the density of a central chi-square distribution with ν degrees of freedom.

Proof. We use the Moment Generating Function (MGF) approach. The MGF of a non-central chi-square distribution $v \sim \chi^2(n, \lambda)$ is:

$$M_v(t) = (1-2t)^{-n/2} \exp\left(\frac{\lambda}{2} \left[\frac{1}{1-2t} - 1 \right] \right) \tag{5.46}$$

We can expand the exponential term using the power series $e^x = \sum_{j=0}^{\infty} \frac{x^j}{j!}$:

$$\begin{aligned} M_v(t) &= (1-2t)^{-n/2} e^{-\lambda/2} \exp\left(\frac{\lambda/2}{1-2t}\right) \\ &= e^{-\lambda/2}(1-2t)^{-n/2} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{\lambda/2}{1-2t}\right)^j \\ &= \sum_{j=0}^{\infty} \left(\frac{e^{-\lambda/2}(\lambda/2)^j}{j!}\right) (1-2t)^{-(n+2j)/2} \end{aligned} \tag{5.47}$$

Recognizing the terms:

1. The term in parentheses, $P(J=j) = \frac{e^{-\lambda/2}(\lambda/2)^j}{j!}$, is the probability mass function of a **Poisson** random variable $J \sim \text{Poisson}(\lambda/2)$.
2. The term $(1-2t)^{-(n+2j)/2}$ is the MGF of a **central chi-square** distribution with $n+2j$ degrees of freedom.

Since the MGF of the mixture is the sum of the MGFs of the components weighted by the mixture probabilities, the density must follow the same mixture structure. □

Remark. This theorem implies a hierarchical model for generating a non-central chi-square variable:

1. Sample $J \sim \text{Poisson}(\lambda/2)$.
2. Given $J = j$, sample $V \sim \chi^2(n+2j, 0)$.

This is particularly useful for numerical computation, as it allows the non-central CDF to be approximated by a finite sum of central chi-square CDFs.

5.4 Distribution of Quadratic Forms

5.4.1 MGF of Quadratic Forms

To determine the distribution of general quadratic forms $y'Ay$, we look at their MGF.

Theorem 5.5 (MGF of Quadratic Form)**.** *If $y \sim N_p(\mu, \Sigma)$, then the MGF of $Q = y'Ay$ is:*

$$M_Q(t) = |I - 2tA\Sigma|^{-1/2} \exp\left(-\frac{1}{2}\mu'[I - (I-2tA\Sigma)^{-1}]\Sigma^{-1}\mu\right) \tag{5.48}$$

5.4.2 Distribution of the Sum Squares of Projected Spherical Normal

We will prove a simplified version of Theorem 5.7 first.

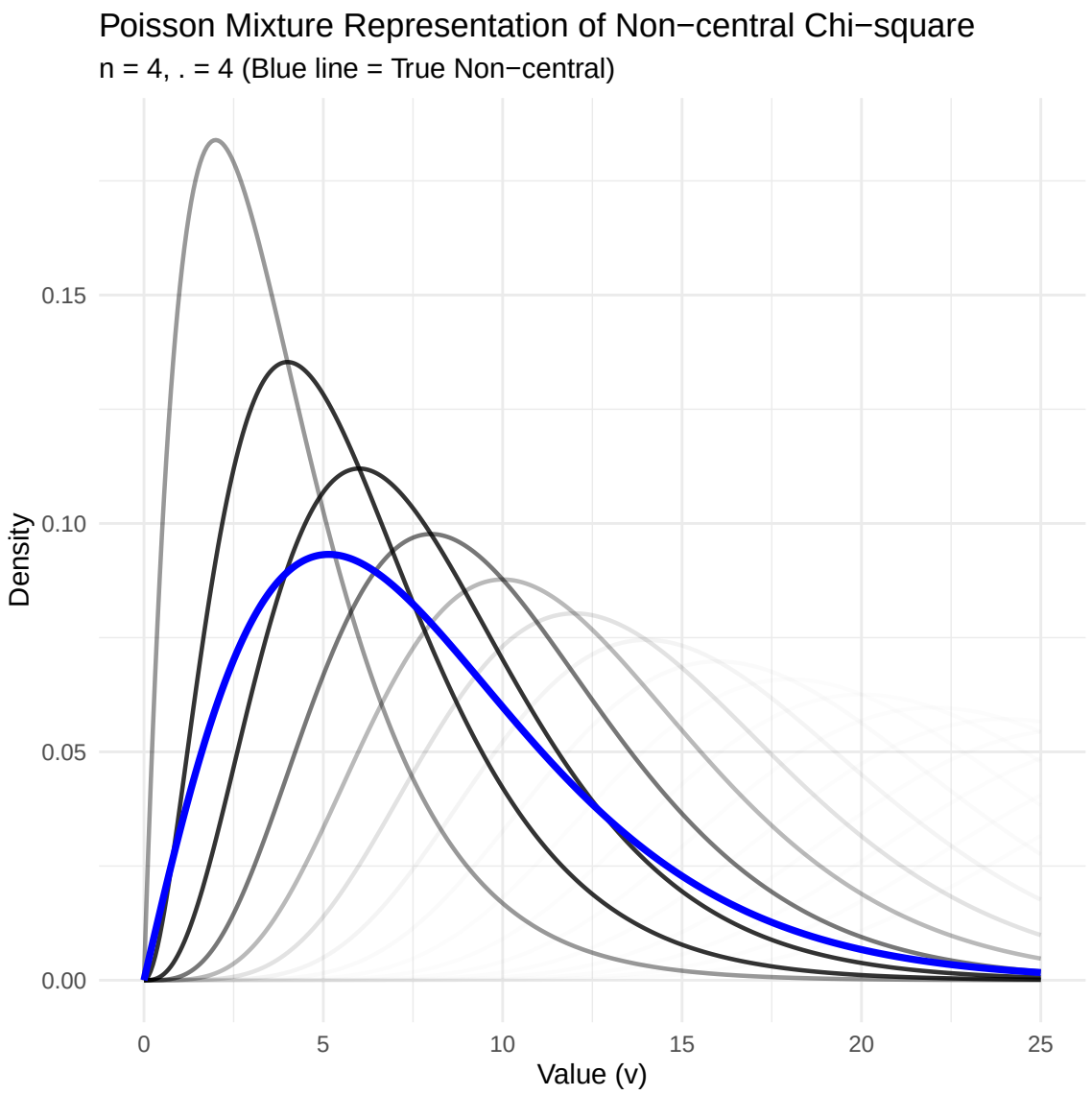


Figure 5.4: The non-central chi-square distribution as a Poisson mixture. The black curves represent central chi-square densities with $df = n + 2j$, with transparency (alpha) proportional to the Poisson weight $P(J = j)$. The solid blue line is the true non-central chi-square density.

Theorem 5.6 (Distribution of Projected Spherical Normal)**.** *If $y \sim N_n(\mu, \sigma^2 I_n)$ and P_V is a projection matrix onto a subspace V of dimension r, then:*

$$\frac{1}{\sigma^2} y' P_V y = \frac{||P_V y||^2}{\sigma^2} \sim \chi^2 \left(r, \frac{||P_V \mu||^2}{\sigma^2} \right) \tag{5.49}$$

This holds because $\frac{1}{\sigma^2} P_V (\sigma^2 I) = P_V$, which is idempotent.

> **!** Crucial Theorem
>
> This is one of the most important theorems in the course, establishing the fundamental conditions under which a quadratic form follows a chi-square distribution.

Proof. **When $\sigma^2 = 1$**
Let P_V be the projection matrix. We know $P_V = QQ'$ where $Q = (q_1, \dots, q_r)$ is an $n \times r$ matrix with orthonormal columns ($Q'Q = I_r$).
The projection of vector y onto the subspace V can be expressed using the orthonormal basis vectors:

$$P_V y = QQ'y = (q_1, \dots, q_r) \begin{pmatrix} q_1' y \\ \vdots \\ q_r' y \end{pmatrix} = \sum_{i=1}^{r} (q_i' y) q_i \tag{5.50}$$

The squared norm of the projection is:

$$y' P_V y = y' QQ' y = (Q'y)'(Q'y) = ||Q'y||^2 \tag{5.51}$$

Since $y \sim N(\mu, I_n)$, the linear transformation $w = Q'y$ follows:

$$w \sim N(Q'\mu, Q' I_n Q) = N(Q'\mu, I_r) \tag{5.52}$$

Thus, w is a vector of r independent normal variables with variance 1. The sum of squares $||w||^2$ is by definition non-central chi-square:

$$||w||^2 \sim \chi^2(r, \lambda) \tag{5.53}$$

where the non-centrality parameter is:

$$\lambda = ||E(w)||^2 = ||Q'\mu||^2 \tag{5.54}$$

Note that $||Q'\mu||^2 = \mu' QQ' \mu = \mu' P_V \mu = ||P_V \mu||^2$.
Thus, $y' P_V y \sim \chi^2(r, ||P_V \mu||^2)$.
When $\sigma^2 \neq 1$
If $y \sim N(\mu, \sigma^2 I_n)$, we standardize by dividing by σ.
Let $w = y/\sigma$. Then $w \sim N(\mu/\sigma, I_n)$. Applying the previous result to w:

$$w' P_V w = \frac{y' P_V y}{\sigma^2} \sim \chi^2 \left(r, \left|\left| P_V \frac{\mu}{\sigma} \right|\right|^2 \right) \tag{5.55}$$

which simplifies to:

$$\frac{||P_V y||^2}{\sigma^2} \sim \chi^2 \left(r, \frac{||P_V \mu||^2}{\sigma^2} \right) \tag{5.56}$$

□

> **!** Important
>
> The term $\|P_V y\|^2$ itself is **not** a standard chi-square variable; it is a scaled chi-square variable. Its mean is:
>
> $$E(\|P_V y\|^2) = \sigma^2 \left(r + \frac{\|P_V \mu\|^2}{\sigma^2} \right) = r\sigma^2 + \|P_V \mu\|^2 \tag{5.57}$$

5.4.3 Distribution of General Quadratic Forms

Lemma 5.2 (Idempotent Matrix Property)**.** *Let Σ be a positive definite matrix such that $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$. The matrix $A\Sigma$ is idempotent if and only if $\Sigma^{1/2} A \Sigma^{1/2}$ is idempotent.*

Proof. ($\Rightarrow$) Assume $A\Sigma$ is idempotent, so $A\Sigma A\Sigma = A\Sigma$. Then:

$$\begin{aligned} (\Sigma^{1/2} A \Sigma^{1/2})^2 &= \Sigma^{1/2} A (\Sigma^{1/2}\Sigma^{1/2}) A \Sigma^{1/2} \\ &= \Sigma^{1/2} (A\Sigma A) \Sigma^{1/2} \end{aligned} \tag{5.58}$$

From the assumption $A\Sigma A\Sigma = A\Sigma$, post-multiplying by Σ^{-1} gives $A\Sigma A = A$. Substituting this back:

$$\Sigma^{1/2}(A)\Sigma^{1/2} = \Sigma^{1/2} A \Sigma^{1/2} \tag{5.59}$$

($\Leftarrow$) Assume $\Sigma^{1/2} A \Sigma^{1/2}$ is idempotent. Then:

$$(\Sigma^{1/2} A \Sigma^{1/2})(\Sigma^{1/2} A \Sigma^{1/2}) = \Sigma^{1/2} A \Sigma^{1/2} \tag{5.60}$$

Expanding the left side:

$$\Sigma^{1/2} A (\Sigma^{1/2}\Sigma^{1/2}) A \Sigma^{1/2} = \Sigma^{1/2} A \Sigma A \Sigma^{1/2} \tag{5.61}$$

Equating this to the right side:

$$\Sigma^{1/2} A \Sigma A \Sigma^{1/2} = \Sigma^{1/2} A \Sigma^{1/2} \tag{5.62}$$

Pre-multiply by $\Sigma^{-1/2}$ and post-multiply by $\Sigma^{1/2}$ (which exist since Σ is positive definite):

$$\begin{aligned} \Sigma^{-1/2}(\Sigma^{1/2} A \Sigma A \Sigma^{1/2})\Sigma^{1/2} &= \Sigma^{-1/2}(\Sigma^{1/2} A \Sigma^{1/2})\Sigma^{1/2} \\ I(A\Sigma A)\Sigma &= I(A)\Sigma \\ A\Sigma A\Sigma &= A\Sigma \end{aligned} \tag{5.63}$$

□

Lemma 5.3 (Rank Invariance)**.** *Under the conditions of Lemma 5.2, if $A\Sigma$ is idempotent, then:*

$$\mathit{rank}(A\Sigma) = \mathit{rank}(\Sigma^{1/2} A \Sigma^{1/2}) = \mathit{tr}(A\Sigma) \tag{5.64}$$

Proof. Since $A\Sigma$ and $\Sigma^{1/2}A\Sigma^{1/2}$ are both idempotent (by Lemma 5.2), their ranks are equal to their traces. Using the cyclic property of the trace operator ($\text{tr}(XYZ) = \text{tr}(ZXY)$):

$$\begin{aligned} \text{rank}(A\Sigma) &= \text{tr}(A\Sigma) \\ &= \text{tr}(A\Sigma^{1/2}\Sigma^{1/2}) \\ &= \text{tr}(\Sigma^{1/2}A\Sigma^{1/2}) \\ &= \text{rank}(\Sigma^{1/2}A\Sigma^{1/2}) \end{aligned} \tag{5.65}$$

Alternatively, notice that $A\Sigma$ is similar to $\Sigma^{1/2}A\Sigma^{1/2}$:

$$A\Sigma = \Sigma^{-1/2}(\Sigma^{1/2}A\Sigma^{1/2})\Sigma^{1/2} \tag{5.66}$$

Since similar matrices have the same rank, the equality holds. □

Theorem 5.7 (Distribution of y'Ay)**.** *Let $y \sim N_p(\mu, \Sigma)$. Let A be a symmetric matrix of rank r. Then $y'Ay \sim \chi^2(r, \lambda)$ with $\lambda = \mu'A\mu$* ***if and only if*** *$A\Sigma$ is idempotent ($A\Sigma A\Sigma = A\Sigma$).*
Special Case ($\Sigma = I$): *If $\Sigma = I$, the condition simplifies to A being idempotent ($A^2 = A$).*

Proof. Let $y^* = \Sigma^{-1/2}y$, so $y^* \sim N_n(\Sigma^{-1/2}\mu, I_n)$. We rewrite the quadratic form:

$$y'Ay = y'\Sigma^{-1/2}(\Sigma^{1/2}A\Sigma^{1/2})\Sigma^{-1/2}y = (y^*)'P_V y^* = \|P_V y^*\|^2 \tag{5.67}$$

Since $A\Sigma$ is idempotent, $P_V = \Sigma^{1/2}A\Sigma^{1/2}$ is a projection matrix with rank r. By the definition of the non-central chi-square, $y'Ay \sim \chi^2(r, \|P_V\Sigma^{-1/2}\mu\|^2)$. The non-centrality parameter simplifies to $\lambda = \mu'A\mu$. □

5.4.4 Standardized Distance Distribution

Corollary 5.2 (Standardized Distance Distribution)**.** *Suppose $y \sim N_n(\mu, \Sigma)$. Then the quadratic form representing the standardized distance from a constant vector μ_0 follows a non-central chi-square distribution:*

$$(y - \mu_0)'\Sigma^{-1}(y - \mu_0) \sim \chi^2(n, \lambda = (\mu - \mu_0)'\Sigma^{-1}(\mu - \mu_0)) \tag{5.68}$$

Proof. Let $A = \Sigma^{-1}$. Then $A\Sigma = \Sigma^{-1}\Sigma = I_n$, which is clearly idempotent. Alternatively, let $w = \Sigma^{-1/2}(y - \mu_0)$, then $w \sim N_n(\Sigma^{-1/2}(\mu - \mu_0), I_n)$. By the definition of chi-square, $\|w\|^2 = (y - \mu_0)'\Sigma^{-1}(y - \mu_0)$ follows the stated distribution. □

> **!** Crucial Theorem
>
> This is an important theorem we will use later.

5.5 Distributions of Projections of Spherical Normal

Theorem 5.8 (Distribution of Projections)**.** *Let V be a k-dimensional subspace of $\mathcal{R}^n$ with projection matrix P_V, and let y be a random vector in $\mathcal{R}^n$ with mean $E(y) = \mu$. Then:*

1. *$E(P_V y) = P_V \mu$.*
2. *If $Var(y) = \sigma^2 I_n$, then $Var(P_V y) = \sigma^2 P_V$ and $E(\|P_V y\|^2) = \sigma^2 k + \|P_V \mu\|^2$.*
3. *If $y \sim N_n(\mu, \sigma^2 I_n)$, then $\frac{1}{\sigma^2}\|P_V y\|^2 = \frac{1}{\sigma^2} y' P_V y \sim \chi^2(k, \frac{1}{\sigma^2}\|P_V \mu\|^2)$.*

Proof.

1. Since the projection operation is linear, $E(P_V y) = P_V E(y) = P_V \mu$.
2. $\text{Var}(P_V y) = P_V \text{Var}(y) P_V^T = P_V \sigma^2 I_n P_V = \sigma^2 P_V$. The expectation of the squared norm follows from the mean of a quadratic form: $E(y' P_V y) = \text{tr}(P_V \sigma^2 I) + \mu' P_V \mu = \sigma^2 k + \|P_V \mu\|^2$.
3. This is a special case of the general quadratic distribution theorem where $A = \frac{1}{\sigma^2} P_V$ and $A(\sigma^2 I) = P_V$, which is idempotent.

□

Theorem 5.9 (Orthogonal Projections)**.** *Let $V_1, \ldots, V_k$ be mutually orthogonal subspaces with dimensions d_i and projection matrices P_i. If $y \sim N_n(\mu, \sigma^2 I_n)$, then:*

1. *The projections $\hat{y}_i = P_i y$ are independent with $\hat{y}_i \sim N(P_i \mu, \sigma^2 P_i)$.*
2. *The squared norms $\|\hat{y}_i\|^2$ are mutually independent.*
3. *$\frac{1}{\sigma^2}\|\hat{y}_i\|^2 \sim \chi^2(d_i, \frac{1}{\sigma^2}\|P_i \mu\|^2)$.*

Proof.

1. For $i \neq j$, $\text{Cov}(P_i y, P_j y) = \sigma^2 P_i P_j = 0$ because orthogonal projection matrices satisfy $P_i P_j = 0$. Under normality, zero covariance implies independence.
2. Since $\hat{y}_i$ are independent, any measurable functions of them, such as their squared norms, are also independent.
3. This follows directly from applying the projection distribution theorem to each independent subspace.

□

5.5.1 Independence of Forms

Theorem 5.10 (Independence Conditions)**.** *Suppose $y \sim N_n(\mu, \Sigma)$.*

- ***Linear and Quadratic:*** *By and $y' Ay$ (where A is symmetric) are independent if and only if $B\Sigma A = 0$.*
- ***Quadratic and Quadratic:*** *$y' Ay$ and $y' By$ (where A, B are symmetric) are independent if and only if $A\Sigma B = 0$.*

Proof. If $B\Sigma A = 0$, the normal vectors By and Ay have zero covariance and are independent. Because By is independent of Ay, it is also independent of any measurable function of Ay, specifically $y'Ay = \|Ay\|^2$ (if A is idempotent). □

5.5.2 Cochran's Theorem

Theorem 5.11 (Cochran's Result)**.** *Let $y \sim N_n(\mu, \sigma^2 I)$ and $y'y = \sum y' A_i y$. The quadratic forms $y^T A_i y/\sigma^2$ are mutually independent $\chi^2(r_i, \lambda_i)$ if and only if any one of the following holds:*

- *Each A_i is idempotent.*
- *$A_i A_j = 0$ for all $i \neq j$.*
- *$n = \sum r_i$.*

5.6 Non-central Distributions Derived from Non-central χ^2

We begin by defining two independent Chi-squared random variables that form the building blocks for statistical power analysis.

- **Non-central Component (X_1):** $X_1 \sim \chi^2(\text{df}_1, \lambda)$. Here, λ is the non-centrality parameter, defined as the sum of squared means, $\lambda = ||\mu||^2$. This is consistent with the definition used throughout this chapter. *(Note: This definition is also used by R's `ncp` argument.)*
- **Central Component (X_2):** $X_2 \sim \chi^2(\text{df}_2)$. X_2 often represents the **Noise Sum of Squares**, SSE_1 of an adequate model, which is assume to follow a central χ^2,

We visualize these components as using the follow diagram.

Figure 5.5: A diagram of two independent χ^2 random variables

5.6.1 The Non-central F-distribution $F(\text{df}_1, \text{df}_2, \lambda)$

Definition 5.3 (Non-central F)**.** Let $X_1 \sim \chi^2(\text{df}_1, \lambda)$ and $X_2 \sim \chi^2(\text{df}_2)$ be independent. The random variable F follows a **non-central F-distribution**:

$$F = \frac{X_1/\text{df}_1}{X_2/\text{df}_2} \sim F(\text{df}_1, \text{df}_2, \lambda) \tag{5.69}$$

- **Expectation:**

- **Under H_0 ($\lambda = 0$):** Exact mean is $\frac{\text{df}_2}{\text{df}_2 - 2}$ (for $\text{df}_2 > 2$).
- **Under H_1 ($\lambda \neq 0$):** Approximate mean is $1 + \frac{\lambda}{\text{df}_1}$.

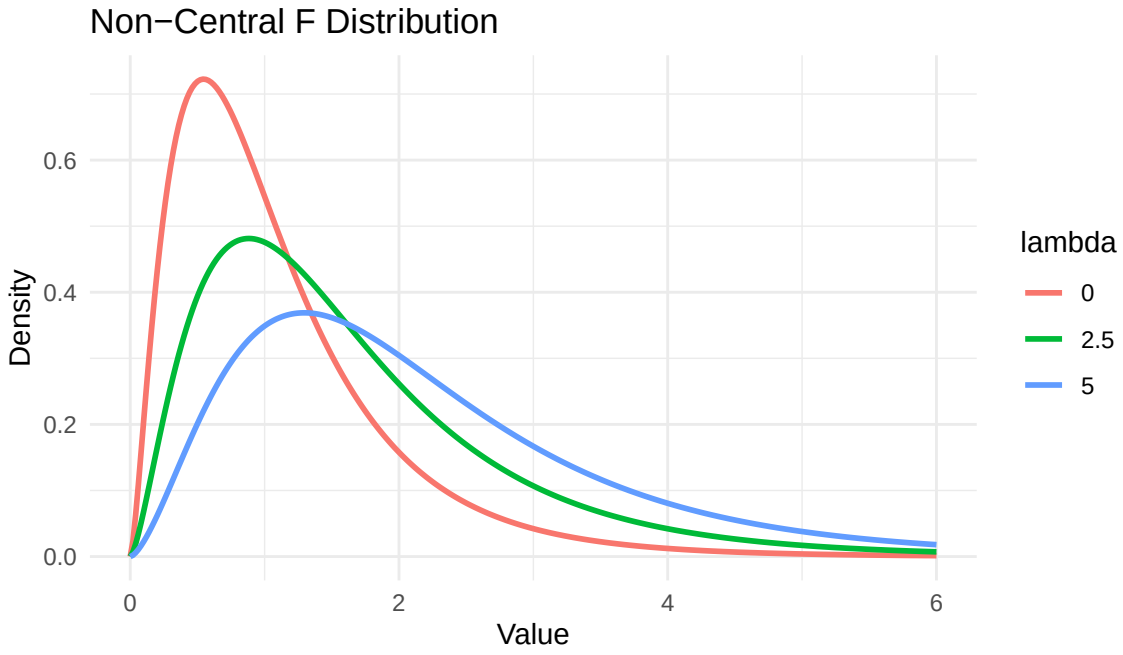


Figure 5.6: Densities of Non-Central F (λ defined as sum of squares).

5.6.2 Type I Non-central Beta $\text{Beta}_1(\text{df}_1/2, \text{df}_2/2, \lambda)$

Definition 5.4 (Type I Non-central Beta)**.** The random variable B_I follows a **Type I non-central Beta distribution**, defined as the signal's proportion of the total sum (R^2):

$$B_I = \frac{X_1}{X_1 + X_2} \sim \text{Beta}_1\left(\frac{\text{df}_1}{2}, \frac{\text{df}_2}{2}, \lambda\right) \tag{5.70}$$

- **Relationship to F:** $B_I = \frac{(\text{df}_1/\text{df}_2)F}{1+(\text{df}_1/\text{df}_2)F}$
- **Expectation:**
  - **Under H_0 ($\lambda = 0$):** Exact mean is $\frac{\text{df}_1}{\text{df}_1+\text{df}_2}$.
  - **Under H_1 ($\lambda \neq 0$):** Approximate mean is $\frac{\text{df}_1+\lambda}{\text{df}_1+\text{df}_2+\lambda}$.

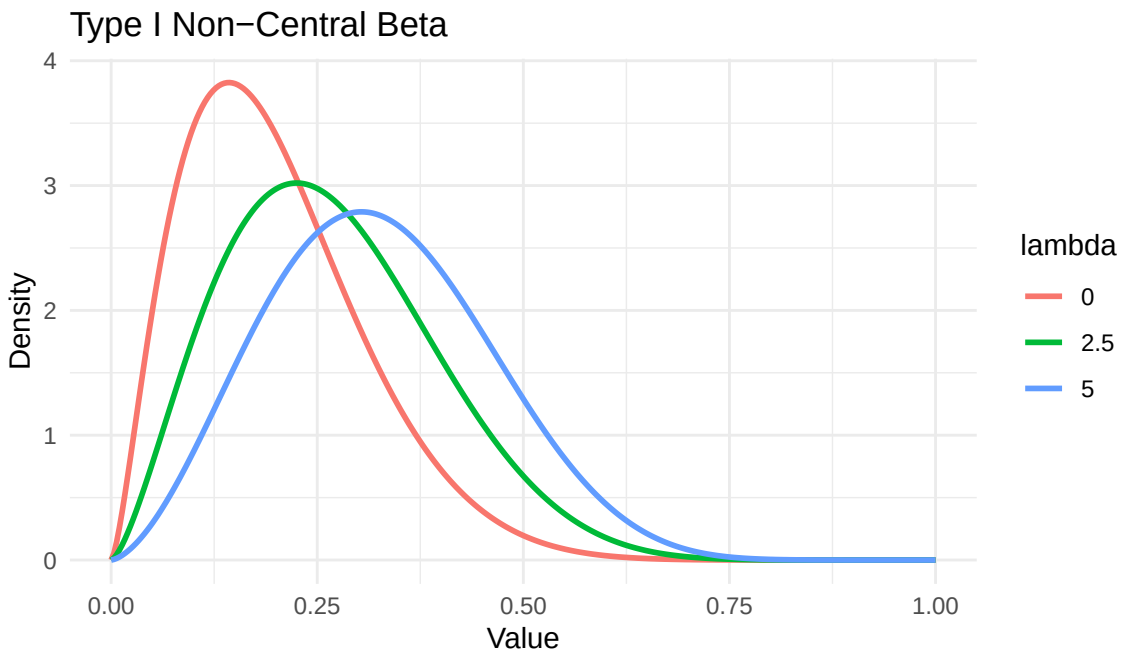


Figure 5.7: Densities of Type I Beta (R^2).

5.6.3 Type II Non-central Beta $\text{Beta}_2(\text{df}_2/2, \text{df}_1/2, \lambda)$

Definition 5.5 (Type II Non-central Beta)**.**

$$B_{II} = \frac{X_2}{X_1 + X_2} = 1 - B_I \sim \text{Beta}_2\left(\frac{\text{df}_2}{2}, \frac{\text{df}_1}{2}, \lambda\right) \tag{5.71}$$

- **Relationship to F:** $B_{II} = \frac{1}{1+(\text{df}_1/\text{df}_2)F}$
- **Expectation:**
  - **Under H_0 ($\lambda = 0$):** Exact mean is $\frac{\text{df}_2}{\text{df}_1+\text{df}_2}$.
  - **Under H_1 ($\lambda \neq 0$):** Approximate mean is $\frac{\text{df}_2}{\text{df}_1+\text{df}_2+\lambda}$.

5.6.4 Scaled Type II Beta $\text{Scaled-Beta}_2(\text{df}_2/2, \text{df}_1/2, \lambda)$

Definition 5.6 (Scaled Type II Beta)**.**

$$S = \frac{X_2/\text{df}_2}{(X_1 + X_2)/(\text{df}_1 + \text{df}_2)} \sim \text{Scaled-Beta}_2 \tag{5.72}$$

- **Relationship to F:** $S = \frac{\text{df}_1+\text{df}_2}{\text{df}_2+\text{df}_1 F}$
- **Expectation:**

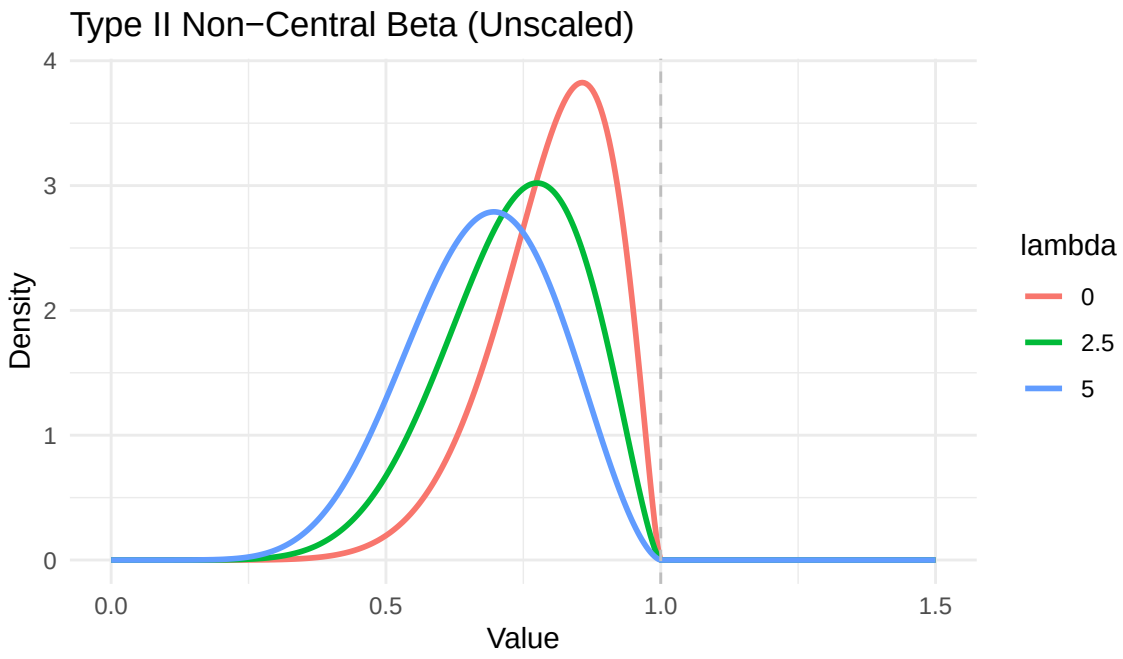


Figure 5.8: Densities of Type II Beta (SSE/SST). Support is [0, 1].

- **Under H_0 ($\lambda = 0$):** Exact mean is 1.
- **Under H_1 ($\lambda \neq 0$):** Approximate mean is $\frac{\text{df}_1+\text{df}_2}{\text{df}_1+\text{df}_2+\lambda}$.

5.6.5 The Non-central t-distribution $t(\text{df}_2, \delta)$

Definition 5.7 (Non-central t)**.** Let $Z \sim N(\delta, 1)$ and $X_2 \sim \chi^2(\text{df}_2)$ be independent. The random variable T follows a **non-central t-distribution**:

$$T = \frac{Z}{\sqrt{X_2/\text{df}_2}} \sim t(\text{df}_2, \delta) \tag{5.73}$$

- **Relationship to F:** $F = T^2$ (when $\text{df}_1 = 1$). Note $\delta^2 = \lambda$.
- **Expectation:**
  - **Under H_0 ($\delta = 0$):** Exact mean is 0.
  - **Under H_1 ($\delta \neq 0$):** Approximate mean is δ.

5.7 Example: Inference of the Mean of Normal Sample

Consider a random sample $y \sim N_n(\mu j_n, \sigma^2 I_n)$. We wish to test:

- M_1 **(Full Model):** μ is unknown.

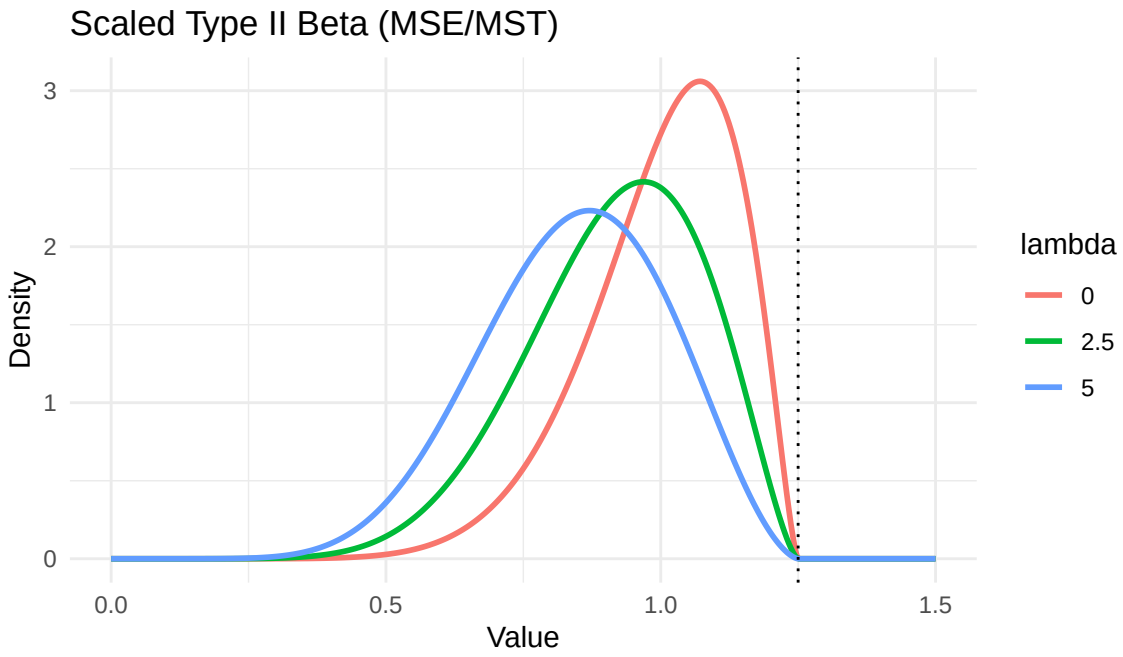


Figure 5.9: Densities of Scaled Type II Beta (MSE/MST).

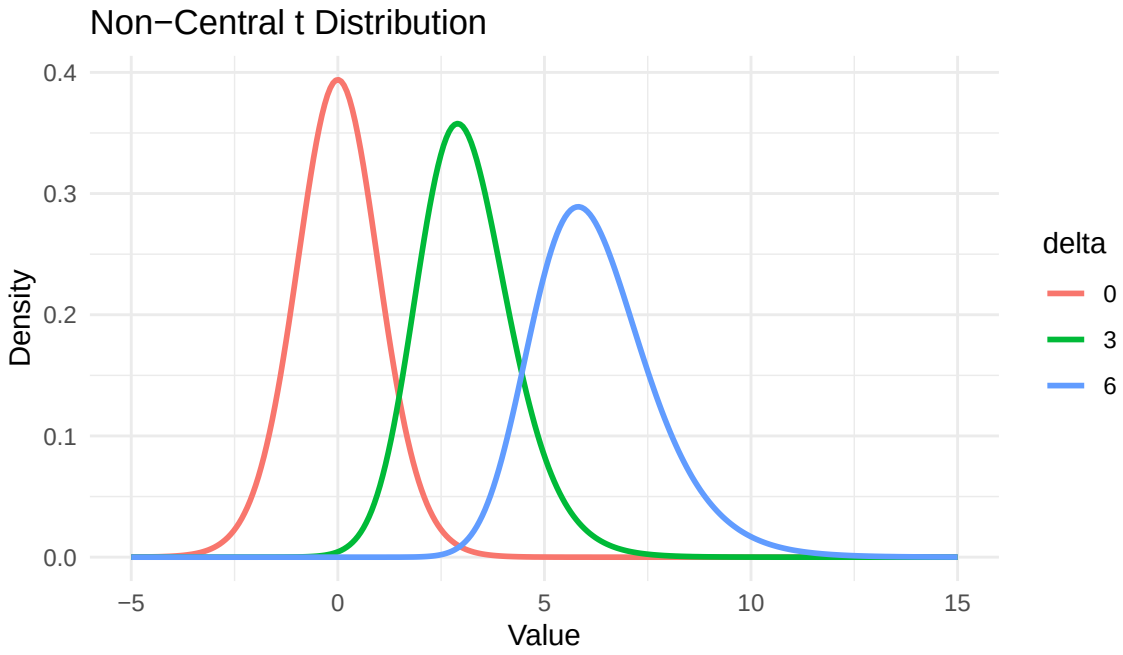


Figure 5.10: Densities of Non-Central t ($df = 20$).

- M_0 **(Reduced Model):** $\mu = \mu_0$.

Let's define the transformed vector $y^* = y - \mu_0 j_n$. Note that $y^* \sim N_n((\mu - \mu_0)j_n, \sigma^2 I_n)$.

5.7.1 Sum of Squares and Their Distributions

We use the projection matrix $P_{j_n} = \frac{1}{n} j_n j_n'$ and its complement $(I_n - P_{j_n})$ to partition the transformed vector.

- **Total SSE (SSE_0 for M_0):**

$$SSE_0 = \|I_n y^*\|^2 = \sum_{i=1}^{n} (Y_i - \mu_0)^2 \tag{5.74}$$

  This follows a non-central distribution with $\text{df}_{\text{total}} = n$:

$$\frac{SSE_0}{\sigma^2} \sim \chi^2(n, \lambda) \quad \text{where } \lambda = \frac{n(\mu - \mu_0)^2}{\sigma^2} \tag{5.75}$$

- **Residual SSE (SSE_1 for M_1):**

$$SSE_1 = \|(I_n - P_{j_n})y^*\|^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{5.76}$$

  This captures the random noise (central component) with $\text{df}_2 = n - 1$:

$$\frac{SSE_1}{\sigma^2} \sim \chi^2(n-1) \tag{5.77}$$

- **Difference SS (SS_{diff}):**

$$SS_{\text{diff}} = \|P_{j_n} y^*\|^2 = n(\bar{Y} - \mu_0)^2 \tag{5.78}$$

  This captures the signal (non-central component) with $\text{df}_1 = 1$:

$$\frac{SS_{\text{diff}}}{\sigma^2} \sim \chi^2(1, \lambda) \tag{5.79}$$

5.7.2 Distributions of Equivalent Statistics

We can construct five equivalent statistics to compare M_0 and M_1.

- **The t-statistic (T):**

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \tag{5.80}$$

- **The F-statistic (F):**

$$F = \frac{n(\bar{Y} - \mu_0)^2}{S^2} = T^2 \tag{5.81}$$

- **The Type I Beta statistic (B_I):**

$$B_I = \frac{SS_{\text{diff}}}{SSE_0} = \frac{n(\bar{Y} - \mu_0)^2}{\sum(Y_i - \mu_0)^2} \tag{5.82}$$

- **The Type II Beta statistic (B_{II}):**

$$B_{II} = \frac{SSE_1}{SSE_0} = \frac{\sum(Y_i - \bar{Y})^2}{\sum(Y_i - \mu_0)^2} = 1 - B_I \tag{5.83}$$

- **The Scaled Type II Beta statistic (S_{scaled}):**

$$S_{\text{scaled}} = \frac{SSE_1/(n-1)}{SSE_0/n} = \left(\frac{n}{n-1}\right) B_{II} \tag{5.84}$$

5.7.3 Expectations Under M_1 and M_0

The table below contrasts the distributions and expected values of these statistics. We assume the sample size n is large enough for the mean of F to exist ($n > 3$).

- **Degrees of Freedom:** $\text{df}_1 = 1, \text{df}_2 = n - 1$.
- **Non-centrality:** $\delta = \frac{\sqrt{n}(\mu - \mu_0)}{\sigma}$ and $\lambda = \delta^2 = \frac{n(\mu - \mu_0)^2}{\sigma^2}$.

Table 5.1: Expected Values of Test Statistics Under Null and Alternative Hypotheses

Statistic	Distribution under H_1 ($\mu \neq \mu_0$)	Exact Mean under H_0 ($\mu = \mu_0$)	Approximate Mean under H_1
T	$t(n-1, \delta)$	0	$\frac{\sqrt{n}(\mu - \mu_0)}{\sigma}$
F	$F(1, n-1, \lambda)$	$\frac{n-1}{n-3} \approx 1$	$1 + \frac{n(\mu - \mu_0)^2}{\sigma^2}$
B_I	$\text{Beta}_1\left(\frac{1}{2}, \frac{n-1}{2}, \lambda\right)$	$\frac{1}{n}$	$\frac{1/n + \frac{(\mu - \mu_0)^2}{\sigma^2}}{1 + \frac{(\mu - \mu_0)^2}{\sigma^2}}$
B_{II}	$\text{Beta}_2\left(\frac{n-1}{2}, \frac{1}{2}, \lambda\right)$	$\frac{n-1}{n}$	$\frac{(n-1)/n}{1 + \frac{(\mu - \mu_0)^2}{\sigma^2}}$
S_{scaled}	$\text{Scaled-Beta}_2\left(\frac{n-1}{2}, \frac{1}{2}, \lambda\right)$	1	$\frac{1}{1 + \frac{(\mu - \mu_0)^2}{\sigma^2}}$

Key Interpretation: All statistics are functionally driven by the signal energy. Notably, for S_{scaled}, the sample size n cancels out in the approximate mean. This makes it a direct measure of the ratio between Noise Variance and Total Variance (Noise + Signal) in the population distributions, connected to the Rao-Blackwell decomposition of variances.

6 Inference for A Multiple Linear Regression Model

6.1 Linear Models and Least Square Estimator

6.1.1 Assumptions in Linear Models

Suppose that on a random sample of n units (patients, animals, trees, etc.) we observe a response variable Y and explanatory variables $X_1, ..., X_k$. Our data are then $(y_i, x_{i1}, ..., x_{ik})$, $i = 1, ..., n$, or in vector/matrix form $y, x_1, ..., x_k$ where $y = (y_1, ..., y_n)$ and $x_j = (x_{1j}, ..., x_{nj})^T$ or y, X where $X = (x_1, ..., x_k)$.

Either by design or by conditioning on their observed values, $x_1, ..., x_k$ are regarded as vectors of known constants. The linear model in its classical form makes the following assumptions:

Assumptions on Linear Models

- **A1. (Additive Error)** $y = \mu + e$ where $e = (e_1, ..., e_n)^T$ is an unobserved random vector with $E(e) = 0$. This implies that $\mu = E(y)$ is the unknown mean of y.
- **A2. (Linearity)** $\mu = \beta_1 x_1 + \cdots + \beta_k x_k = X\beta$ where $\beta_1, ..., \beta_k$ are unknown parameters. This assumption says that $E(y) = \mu \in \text{Col}(X)$ (lies in the column space of X); i.e., it is a linear combination of explanatory vectors $x_1, ..., x_k$ with coefficients the unknown parameters in $\beta = (\beta_1, ..., \beta_k)^T$. Note that it is linear in $\beta_1, ..., \beta_k$, not necessarily in the x's.
- **A3. (Independence)** $e_1, ..., e_n$ are independent random variables (and therefore so are $y_1, ..., y_n$).
- **A4. (Homoscedasticity)** $e_1, ..., e_n$ all have the same variance σ^2; that is, $\text{Var}(e_1) = \cdots = \text{Var}(e_n) = \sigma^2$ which implies $\text{Var}(y_1) = \cdots = \text{Var}(y_n) = \sigma^2$.
- **A5. (Normality)** $e \sim N_n(0, \sigma^2 I_n)$.

6.1.2 Matrix Formulation

The model can be written algebraically as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \quad i = 1, ..., n \tag{6.1}$$

Or in matrix notation:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \tag{6.2}$$

This is expressed compactly as:

$$y = X\beta + e \tag{6.3}$$

where X is the design matrix, and $e \sim N_n(0, \sigma^2 I)$. Alternatively:

$$y = \beta_0 j_n + \beta_1 x_1 + \cdots + \beta_k x_k + e \tag{6.4}$$

Taken together, all five assumptions can be stated more succinctly as:

$$y \sim N_n(X\beta, \sigma^2 I) \tag{6.5}$$

with the mean vector $\mu_y = X\beta \in \text{Col}(X)$.

! Coefficients and Variance of Reduced Models

The effect of a parameter and the magnitude of the error variance depend upon what other explanatory variables are present in the model. For example, the coefficients β_0, β_1 and error standard deviation σ in the model:

$$y = \beta_0 j_n + \beta_1 x_1 + \beta_2 x_2 + e, \quad \text{Var}(e) = \sigma^2 I \tag{6.6}$$

will typically be different than β_0^*, β_1^* and σ^* in the model:

$$y = \beta_0^* j_n + \beta_1^* x_1 + e^*, \quad \text{Var}(e^*) = (\sigma^*)^2 I \tag{6.7}$$

In this context, β_0^* and β_1^* are the population-projected coefficients of the full model. Furthermore, σ^* will typically be larger than σ, as the error term e^* absorbs the variation previously explained by x_2.

! Important

We will first consider the case that $\text{rank}(X) = k + 1$.

6.1.3 Least Squares Estimator of β and Fitted Value $\hat{Y}$

Definition 6.1 (Least Squares Estimator)**.** The **Least Squares Estimator (LSE)** of β, denoted as $\hat{\beta}$, is the vector that minimizes the Sum of Squared Errors (SSE), which measures the discrepancy between the observed responses y and the fitted values $X\hat{\beta}$.

$$Q(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = (y - X\beta)'(y - X\beta) \tag{6.8}$$

Theorem 6.1 (Least Squares Estimator)**.** *Consider the linear model $y = X\beta + e$, where X is of full column rank. The Ordinary Least Squares (OLS) estimator $\hat{\beta}$ is given by the closed-form solution:*

$$\hat{\beta} = (X'X)^{-1}X'y \tag{6.9}$$

Consequently, the vector of fitted values $\hat{y}$ is the orthogonal projection of y onto Col(X):

$$\hat{y} = X\hat{\beta} = Hy \tag{6.10}$$

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

Proof. The derivation relies on the geometry of orthogonal projections.

1. Obtaining the Fitted Values $\hat{y}$

In the linear model, the systematic component $E[y]$ is constrained to lie in the column space of X, denoted as $\text{Col}(X)$. We seek the vector in $\text{Col}(X)$ that is "closest" to the observed data y. This vector is the **orthogonal projection** of y onto $\text{Col}(X)$, denoted as $\hat{y}$. Using the projection matrix $H = X(X'X)^{-1}X'$, we have:

$$\hat{y} = Hy = X(X'X)^{-1}X'y \tag{6.11}$$

2. Obtaining $\hat{\beta}$ by Solving $X\beta = \hat{y}$

Since $\hat{y}$ is a projection onto $\text{Col}(X)$, the system $X\hat{\beta} = \hat{y}$ is consistent. To isolate $\hat{\beta}$, we pre-multiply both sides by $(X'X)^{-1}X'$:

$$\begin{aligned} (X'X)^{-1}X'(X\hat{\beta}) &= (X'X)^{-1}X'\hat{y} \\ \underbrace{(X'X)^{-1}(X'X)}_{I}\hat{\beta} &= (X'X)^{-1}X'\hat{y} \\ \hat{\beta} &= (X'X)^{-1}X'\hat{y} \end{aligned} \tag{6.12}$$

Finally, we express the estimator in terms of the observed y. Because $\hat{y}$ is an orthogonal projection, the residual $y - \hat{y}$ is orthogonal to the columns of X, implying $X'\hat{y} = X'y$. Substituting this into the equation above yields the result:

$$\hat{\beta} = (X'X)^{-1}X'y \tag{6.13}$$

□

6.1.4 Properties of the Estimator $\hat{\beta}$

Theorem 6.2 (Unbiasedness of $\hat{\beta}$)**.** *If $E(y) = X\beta$, then $\hat{\beta}$ is an unbiased estimator for β.*

Proof.

$$\begin{aligned} E(\hat{\beta}) &= E[(X'X)^{-1}X'y] \\ &= (X'X)^{-1}X'E(y) \quad \text{[using linearity of expectation]} \\ &= (X'X)^{-1}X'X\beta \\ &= \beta \end{aligned} \tag{6.14}$$

□

Theorem 6.3 (Variance of $\hat{\beta}$)**.** *If $Var(y) = \sigma^2 I$, the covariance matrix for $\hat{\beta}$ is given by $\sigma^2(X'X)^{-1}$.*

Proof.

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \text{Var}[(X'X)^{-1}X'y] \\ &= (X'X)^{-1}X'\text{Var}(y)[(X'X)^{-1}X']' \quad [\text{using Var}(Ay) = A\text{Var}(y)A'] \\ &= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \tag{6.15}$$

□

Note: These theorems require no assumption of normality.

6.2 Best Linear Unbiased Estimator (BLUE)

Theorem 6.4 (Gauss-Markov Theorem)**.** *If $E(y) = X\beta$ and $Var(y) = \sigma^2 I$, the least-squares estimators $\hat{\beta}_j, j = 0, 1, ..., k$ have minimum variance among all linear unbiased estimators.*

Proof. We consider a linear estimator Ay of β and seek the matrix A for which Ay is a minimum variance unbiased estimator.

1. Unbiasedness Condition:

In order for Ay to be an unbiased estimator of β, we must have $E(Ay) = \beta$. Using the assumption $E(y) = X\beta$, this is expressed as:

$$E(Ay) = AE(y) = AX\beta = \beta \tag{6.16}$$

which implies the condition $AX = I_{k+1}$ since the relationship must hold for any β.

2. Minimizing Variance:

The covariance matrix for the estimator Ay is:

$$\text{Var}(Ay) = A\text{Var}(y)A' = A(\sigma^2 I)A' = \sigma^2 AA' \tag{6.17}$$

We need to choose A (subject to $AX = I$) so that the diagonal elements of AA' are minimized.
To relate Ay to $\hat{\beta} = (X'X)^{-1}X'y$, we define $\hat{A} = (X'X)^{-1}X'$ and write $A = (A - \hat{A}) + \hat{A}$. Then:

$$AA' = [(A - \hat{A}) + \hat{A}][(A - \hat{A}) + \hat{A}]' \tag{6.18}$$

Expanding this, the cross terms vanish because $(A - \hat{A})\hat{A}' = A\hat{A}' - \hat{A}\hat{A}'$. Note that $\hat{A}\hat{A}' = (X'X)^{-1}X'X(X'X)^{-1} = (X'X)^{-1}$. Also, $A\hat{A}' = AX(X'X)^{-1} = I(X'X)^{-1} = (X'X)^{-1}$ (since $AX = I$). Thus, $(A - \hat{A})\hat{A}' = 0$.

The expansion simplifies to:

$$AA' = (A - \hat{A})(A - \hat{A})' + \hat{A}\hat{A}' \tag{6.19}$$

The matrix $(A - \hat{A})(A - \hat{A})'$ is positive semidefinite, meaning its diagonal elements are non-negative. To minimize the diagonal of AA', we must set $A - \hat{A} = 0$, which implies $A = \hat{A}$.
Thus, the minimum variance estimator is:

$$Ay = (X'X)^{-1}X'y = \hat{\beta} \tag{6.20}$$

□

6.2.1 Notes on Gauss-markov

1. **Distributional Generality:** The remarkable feature of the Gauss-Markov theorem is that it holds for *any* distribution of y; normality is not required. The only assumptions used are linearity ($E(y) = X\beta$) and homoscedasticity ($\text{Var}(y) = \sigma^2 I$).
2. **Extension to All Linear Combinations:** The theorem extends beyond just the parameter vector β to any linear combination of the parameters.
3. **Scaling Invariance:** The predictions made by the model are invariant to the scaling of the explanatory variables.

Corollary 6.1 (BLUE for All Linear Combinations)**.** *If $E(y) = X\beta$ and $Var(y) = \sigma^2 I$, the best linear unbiased estimator of the scalar $a'\beta$ is $a'\hat{\beta}$, where $\hat{\beta}$ is the least-squares estimator.*

Proof. Let $\tilde{\beta} = Ay$ be any other linear unbiased estimator of β. The variance of the linear combination $a'\tilde{\beta}$ is:

$$\frac{1}{\sigma^2}\text{Var}(a'\tilde{\beta}) = \frac{1}{\sigma^2}\text{Var}(a'Ay) = a'AA'a \tag{6.21}$$

From the proof of the Gauss-Markov theorem, we established that $AA' = (A - \hat{A})(A - \hat{A})' + (X'X)^{-1}$ where $\hat{A} = (X'X)^{-1}X'$. Substituting this into the variance equation:

$$a'AA'a = a'(A - \hat{A})(A - \hat{A})'a + a'(X'X)^{-1}a \tag{6.22}$$

The term $a'(A - \hat{A})(A - \hat{A})'a$ is a quadratic form with a positive semidefinite matrix, so it is always non-negative. Therefore:

$$a'AA'a \geq a'(X'X)^{-1}a = \frac{1}{\sigma^2}\text{Var}(a'\hat{\beta}) \tag{6.23}$$

The variance is minimized when $A = \hat{A}$ (specifically when the first term is zero), proving that $a'\hat{\beta}$ has the minimum variance among all linear unbiased estimators. □

Theorem 6.5 (Scaling Explanatory Variables)**.** *If $x = (1, x_1, ..., x_k)'$ and $z = (1, c_1x_1, ..., c_kx_k)'$, then the fitted values are identical: $\hat{y} = \hat{\beta}'x = \hat{\beta}_z'z$.*

Proof. Let $D = \text{diag}(1, c_1, ..., c_k)$ such that the design matrix is transformed to $Z = XD$. The LSE for the transformed data is:

$$\begin{aligned}\hat{\beta}_z &= (Z'Z)^{-1}Z'y = [(XD)'(XD)]^{-1}(XD)'y \\ &= D^{-1}(X'X)^{-1}(D')^{-1}D'X'y \\ &= D^{-1}(X'X)^{-1}X'y = D^{-1}\hat{\beta}\end{aligned} \tag{6.24}$$

. Then, the prediction is:

$$\hat{\beta}_z'z = (D^{-1}\hat{\beta})'(Dx) = \hat{\beta}'(D^{-1})'Dx = \hat{\beta}'x \tag{6.25}$$

. □

6.2.2 Limitations: Restriction to Unbiased Estimators

It is crucial to recognize that the Gauss-Markov theorem only guarantees optimality within the class of **linear** and **unbiased** estimators.

- **Assumption Sensitivity:** If the assumptions of linearity ($E(y) = X\beta$) and homoscedasticity ($\text{Var}(y) = \sigma^2 I$) do not hold, $\hat{\beta}$ may be biased or may have a larger variance than other estimators.
- **Unbiasedness Constraint:** The theorem does not compare $\hat{\beta}$ to biased estimators. It is possible for a biased estimator (e.g., shrinkage estimators) to have a smaller Mean Squared Error (MSE) than the BLUE by accepting some bias to significantly reduce variance. The LSE is only "best" (minimum variance) among those estimators that satisfy the unbiasedness constraint.

6.3 Unbiased Estimator of Error Variance

We estimate σ^2 by the residual mean square:

Definition 6.2 (Residual Variance Estimator)**.**

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 = \frac{\text{SSE}}{n-k-1} \tag{6.26}$$

where $\text{SSE} = (y - X\hat{\beta})'(y - X\hat{\beta})$.

Alternatively, SSE can be written as:

$$\text{SSE} = y'y - \hat{\beta}'X'y \tag{6.27}$$

This is often useful for computation ($y'y$ is the total sum of squares of the raw data).

6.3.1 Unbiasedness of s^2

Theorem 6.6 (Unbiasedness of s-squared)**.** *If s^2 is defined as above, and if $E(y) = X\beta$ and $Var(y) = \sigma^2 I$, then $E(s^2) = \sigma^2$.*

Proof. We use the Hat Matrix $H = X(X'X)^{-1}X'$, which projects y onto Col(X). Thus, $\hat{y} = Hy$. The residuals are $y - \hat{y} = (I - H)y$. The Sum of Squared Errors is:

$$\text{SSE} = \|(I - H)y\|^2 = y'(I - H)'(I - H)y \tag{6.28}$$

Since H is symmetric and idempotent, $(I - H)$ is also symmetric and idempotent. Thus:

$$\text{SSE} = y'(I - H)y \tag{6.29}$$

To find the expectation, we use the trace trick for quadratic forms: $E[y'Ay] = \text{tr}(A\text{Var}(y)) + E[y]'AE[y]$.

$$\begin{aligned} E(\text{SSE}) &= E[y'(I - H)y] \\ &= \text{tr}((I - H)\sigma^2 I) + (X\beta)'(I - H)(X\beta) \\ &= \sigma^2 \text{tr}(I - H) + \beta' X'(I - H)X\beta \end{aligned} \tag{6.30}$$

Trace Term: $\text{tr}(I_n - H) = \text{tr}(I_n) - \text{tr}(H) = n - (k + 1)$, since $\text{tr}(H) = \text{tr}(X(X'X)^{-1}X') = \text{tr}((X'X)^{-1}X'X) = \text{tr}(I_{k+1}) = k + 1$.
Non-centrality Term: Since $HX = X$, we have $(I - H)X = 0$. Therefore, the second term vanishes: $\beta' X'(I - H)X\beta = 0$.
Combining these:

$$E(\text{SSE}) = \sigma^2(n - k - 1) \tag{6.31}$$

Dividing by the degrees of freedom $(n - k - 1)$, we get $E(s^2) = \sigma^2$. □

6.4 Distributions Under Normality

If we add Assumption A5 ($y \sim N_n(X\beta, \sigma^2 I)$), we can derive the exact sampling distributions.

Corollary 6.2 (Estimated Covariance of Beta)**.** *An unbiased estimator of Cov$(\hat{\beta})$ is given by:*

$$\widehat{Cov}(\hat{\beta}) = s^2(X'X)^{-1} \tag{6.32}$$

Theorem 6.7 (Sampling Distributions)**.** *Under assumptions A1-A5:*

1. $\hat{\beta} \sim N_{k+1}(\beta, \sigma^2(X'X)^{-1})$.
2. $(n - k - 1)s^2/\sigma^2 \sim \chi^2(n - k - 1)$.
3. $\hat{\beta}$ *and* s^2 *are independent.*

Proof. **Part (i):** Since $\hat{\beta} = (X'X)^{-1}X'y$ is a linear transformation of the normal vector y, it is also normally distributed. We already established its mean and variance in Theorem 6.2 and Theorem 6.3.
Part (ii): We showed SSE $= y'(I-H)y$. Since $(I-H)$ is idempotent with rank $n-k-1$, and $(I-H)X\beta = 0$, by the theory of quadratic forms in normal variables, SSE$/\sigma^2 \sim \chi^2(n-k-1)$.
Part (iii): $\hat{\beta}$ depends on Hy (or $X'y$), while s^2 depends on $(I-H)y$. Since $H(I-H) = H - H^2 = 0$, the linear forms defining the estimator and the residuals are orthogonal. For normal vectors, zero covariance implies independence. □

6.5 Maximum Likelihood Estimator (MLE)

Theorem 6.8 (MLE for Linear Regression)**.** *If $y \sim N_n(X\beta, \sigma^2 I)$, the Maximum Likelihood Estimators (MLE) for the coefficients and the error variance are:*

$$\hat{\beta}_{MLE} = (X'X)^{-1}X'y \tag{6.33}$$

$$\hat{\sigma}^2_{MLE,e} = \frac{SSE}{n} \tag{6.34}$$

Similarly, under the Null Model ($y \sim N(\mu, \sigma_y^2)$), the MLE for the total variance of y is:

$$\hat{\sigma}^2_{MLE,y} = \frac{SST}{n} \tag{6.35}$$

Proof.

1. Derivation of Error Variance ($\hat{\sigma}^2_{\text{MLE},e}$)

The probability density function for the multivariate normal distribution $y \sim N(X\beta, \sigma^2 I)$ is:

$$f(y) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right) \tag{6.36}$$

The log-likelihood function is $\ln L = \ln f(y)$:

$$\ln L(\beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta) \tag{6.37}$$

First, we maximize with respect to β. Since only the last term involves β, maximizing the likelihood is equivalent to minimizing the sum of squared errors:

$$\text{SSE}(\beta) = (y - X\beta)'(y - X\beta) \tag{6.38}$$

This yields the standard Least Squares estimator $\hat{\beta} = (X'X)^{-1}X'y$. Substituting this back into the SSE term gives the minimized sum of squares, SSE.
Next, we maximize with respect to the variance σ^2. Let $v = \sigma^2$. The log-likelihood becomes:

$$\ln L(v) = C - \frac{n}{2}\ln(v) - \frac{\text{SSE}}{2v} \tag{6.39}$$

Differentiating with respect to v:

$$\frac{\partial \ln L}{\partial v} = -\frac{n}{2v} + \frac{\text{SSE}}{2v^2} \tag{6.40}$$

Setting the derivative to zero to find the critical point:

$$-\frac{n}{2\hat{v}} + \frac{\text{SSE}}{2\hat{v}^2} = 0 \tag{6.41}$$

$$\frac{n}{2\hat{v}} = \frac{\text{SSE}}{2\hat{v}^2} \tag{6.42}$$

$$n = \frac{\text{SSE}}{\hat{v}} \implies \hat{v} = \frac{\text{SSE}}{n} \tag{6.43}$$

Thus, the MLE for the error variance is:

$$\hat{\sigma}^2_{\text{MLE},e} = \frac{\text{SSE}}{n} \tag{6.44}$$

2. Derivation of Total Variance ($\hat{\sigma}^2_{\text{MLE},y}$)

Under the Null Model (intercept only), we assume $y_i \sim N(\mu, \sigma_y^2)$. The design matrix X is simply a column of ones (j_n). Maximizing the likelihood for μ yields the sample mean $\hat{\mu} = \bar{y}$.
The term $(y - X\beta)'(y - X\beta)$ simplifies to the Total Sum of Squares:

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{6.45}$$

Following the exact same differentiation steps as above (replacing SSE with SST), the MLE for the variance of y is:

$$\hat{\sigma}^2_{\text{MLE},y} = \frac{\text{SST}}{n} \tag{6.46}$$

□

Note: The MLEs for variance are biased. They divide by the sample size n, whereas the unbiased estimators (used in standard ANOVA tables) divide by the degrees of freedom ($n - k - 1$ for error, $n - 1$ for total).

6.6 Linear Models in Centered Form

Starting with the original model, let the design matrix be $X^* = [j_n, X]$, where X is the $n \times p$ matrix of predictors excluding the intercept, and let the original coefficients be $\beta^* = [\beta_0^*, (\beta_1^*)^\top]^\top$. The mean vector $\mu_y = E(y)$ is:

$$\mu_y = X^*\beta^* = [j_n, X] \begin{pmatrix} \beta_0^* \\ \beta_1^* \end{pmatrix} = \beta_0^* j_n + X\beta_1^* \tag{6.47}$$

We define the centered design matrix X_c as the projection of X onto the orthogonal complement of the intercept space:

$$X_c = (I - P_{j_n})X = X - P_{j_n}X \tag{6.48}$$

Because $P_{j_n} = j_n(j_n^\top j_n)^{-1} j_n^\top$, the term $P_{j_n} X$ computes the column means, which we can write as $j_n \bar{x}^\top$, where $\bar{x}^\top$ is the row vector of means. Rearranging the definition gives:

$$X = j_n \bar{x}^\top + X_c \tag{6.49}$$

Substituting this expression for X back into our mean vector μ_y:

$$\begin{aligned} \mu_y &= \beta_0^* j_n + (j_n \bar{x}^\top + X_c)\beta_1^* \\ &= \beta_0^* j_n + j_n \bar{x}^\top \beta_1^* + X_c \beta_1^* \\ &= j_n(\beta_0^* + \bar{x}^\top \beta_1^*) + X_c \beta_1^* \end{aligned} \tag{6.50}$$

By defining the parameters of the centered model as $\alpha = \beta_0^* + \bar{x}^\top \beta_1^*$ and $\beta_1 = \beta_1^*$, the equation simplifies cleanly to:

$$\mu_y = \alpha j_n + X_c \beta_1 = [j_n, X_c] \begin{pmatrix} \alpha \\ \beta_1 \end{pmatrix} \tag{6.51}$$

Adding the error term, the full centered model is:

$$y = \mu_y + \epsilon = j_n \alpha + X_c \beta_1 + \epsilon \tag{6.52}$$

Orthogonality of j_n and X_c

By construction, X_c is orthogonal to j_n. This is quickly proven using the properties of the idempotent projection matrix P_{j_n}:

$$j_n^\top X_c = j_n^\top (I - P_{j_n}) X = (j_n^\top - j_n^\top P_{j_n}) X = (j_n^\top - j_n^\top) X = 0 \tag{6.53}$$

Orthogonal Projections of the Mean Vector

Because j_n and X_c are strictly orthogonal, projecting μ_y onto their respective column spaces completely isolates their components:

$$\boxed{P_{j_n} \mu_y = P_{j_n}(j_n \alpha + X_c \beta_1) = j_n \alpha + 0 = \alpha j_n} \tag{6.54}$$

$$\boxed{P_{X_c} \mu_y = P_{X_c}(j_n \alpha + X_c \beta_1) = 0 + X_c \beta_1 = X_c \beta_1} \tag{6.55}$$

6.7 Least Squares Estimates for Linear Models in Centered Form

Because the column space of the intercept j_n is strictly orthogonal to the column space of X_c, the total projection of y onto the combined column space spanned by $[j_n, X_c]$ decomposes into the sum of the independent orthogonal projections onto these two subspaces.

Let $\hat{y}_0$ denote the projection of y onto j_n, and $\hat{y}_1$ denote the projection of y onto X_c:

$$\hat{y}_0 = P_{j_n} y \tag{6.56}$$

$$\hat{y}_1 = P_{X_c} y \tag{6.57}$$

We can express the total fitted values $\hat{y}$ as the sum of these orthogonal projections:

$$\boxed{\hat{y} = \hat{y}_0 + \hat{y}_1 = P_{j_n} y + P_{X_c} y} \tag{6.58}$$

Applying the hat matrix based on j_n and X_c we obtain that

$$\boxed{\hat{y}_0 = P_{j_n} y = j_n (j_n^T j_n)^{-1} j_n^T y = j_n \hat{\alpha}} \tag{6.59}$$

$$\boxed{\hat{y}_1 = P_{X_c} y = X_c (X_c^T X_c)^{-1} X_c^T y = X_c \hat{\beta}_1} \tag{6.60}$$

By looking at the expressions of $\hat{y}_0$ and $\hat{y}_1$, we easily isolate the least squares estimators:

$$\boxed{\hat{\alpha} = (j_n^\top j_n)^{-1} j_n^\top y = \bar{y}} \tag{6.61}$$

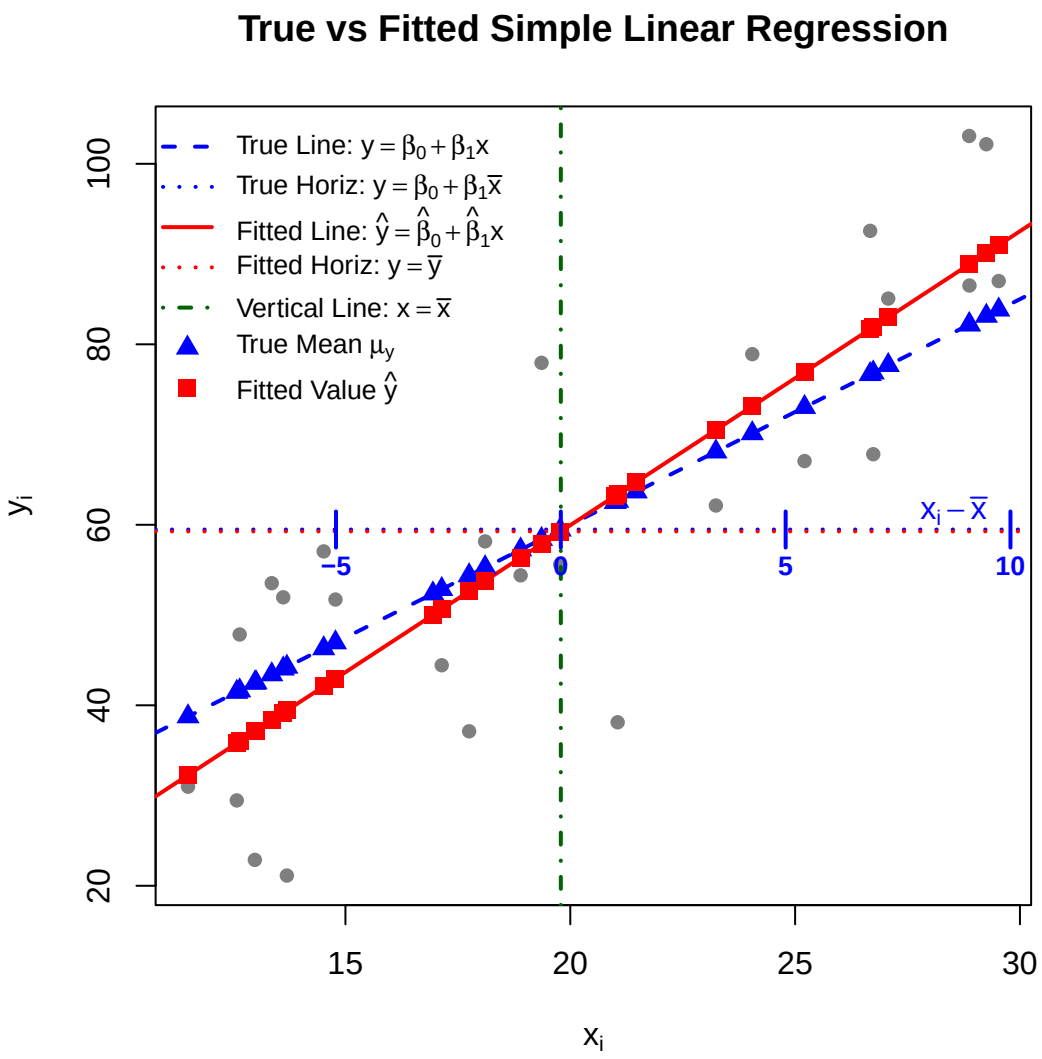
(The sample mean of y)

$$\boxed{\hat{\beta}_1 = (X_c^\top X_c)^{-1} X_c^\top y = S_{xx}^{-1} S_{xy}} \tag{6.62}$$

(Using the sample covariance matrix notations)

Recovering the original intercept:

$$\hat{\beta}_0^* = \hat{\alpha} - \bar{x}^\top \hat{\beta}_1 \tag{6.63}$$



6.8 Decomposition of Sum of Squares and their Distributions

We partition the total variation based on the orthogonal subspaces.

Definition 6.3 (Sum of Squares Components)**.** The total variation is decomposed as SST $=$ SSR $+$ SSE.

1. **Total Sum of Squares (SST):** The squared length of the centered response vector.

$$\text{SST} = \|y - \bar{y} j_n\|^2 = \|(I - P_{j_n})y\|^2 \tag{6.64}$$

2. **Regression Sum of Squares (SSR):** The variation explained by the regressors X_c.

$$\text{SSR} = \|\hat{y} - \bar{y}j_n\|^2 = \|P_{X_c}y\|^2 = \hat{\beta}_1' X_c' X_c \hat{\beta}_1 \tag{6.65}$$

3. **Sum of Squared Errors (SSE):** The residual variation.

$$\text{SSE} = \|y - \hat{y}\|^2 = \|(I - H)y\|^2 \tag{6.66}$$

6.8.1 3D Visualization of Decomposition of y

We partition the total variation in y based on the orthogonal subspaces.

1. **Space of the Mean:** $L(j_n)$, spanned by the intercept vector j_n.
2. **Space of the Regressors:** $L(X_c)$, spanned by the centered predictors X_c.
3. **Error Space:** $\text{Col}(X)^{\perp}$, orthogonal to the model space.

The vector y can be decomposed into three orthogonal components:

$$y = \bar{y}j_n + P_{X_c}y + (y - \hat{y}) \tag{6.67}$$

Visually, this corresponds to projecting the vector y onto three orthogonal axes.

Interactive Visualization:

We generate a cloud of 100 observations of y from $N(\mu, \sigma = 1)$ where $\mu = (5, 5, 0)$. The projections onto the Model Plane ($z = 0$) are highlighted in **red**, and the projections onto the error axis (z) are in **yellow**.

6.8.1.1 Effect Exists (signal)

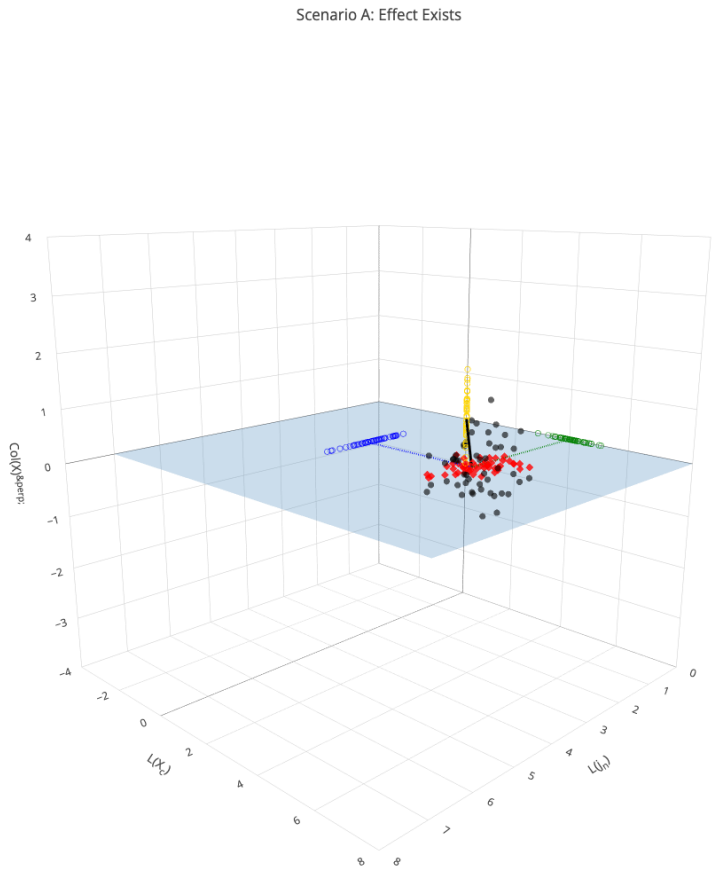


Figure 6.1: Scenario 1: Significant regression effect ($\beta_1 \neq 0$). The mean vector projects significantly onto the predictor space.

6.8.1.2 No Effect (noise)

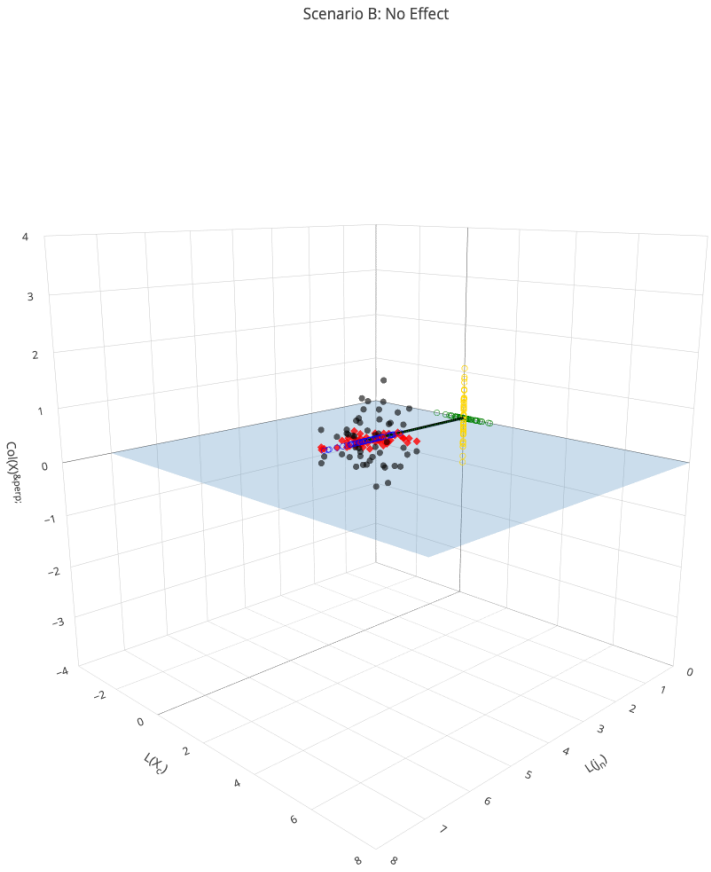


Figure 6.2: Scenario 2: No regression effect ($\beta_1 = 0$). The mean vector lies purely on the intercept axis.

6.8.2 A Diagram to Show Decomposition of Sum of Squares

The decomposition of the total variation is visualized below. The total deviation (Orange) is the vector sum of the regression deviation (Green) and the residual error (Red).

6.8.3 Distribution of Sum of Squares

We apply the general theory of projections to the specific components defined in Definition 6.3.

Theorem 6.9 (Distribution of Sum of Squares)**.** *Let $y \sim N(\mu, \sigma^2 I_n)$, where $\mu \in Col(X)$. Consider the decomposition defined by the projection matrices P_{X_c} and $M = I - H$.*

1. ***Independence***

   The quadratic forms SSR and SSE are statistically independent because the subspaces $L(X_c)$ and $Col(X)^{\perp}$ are orthogonal.

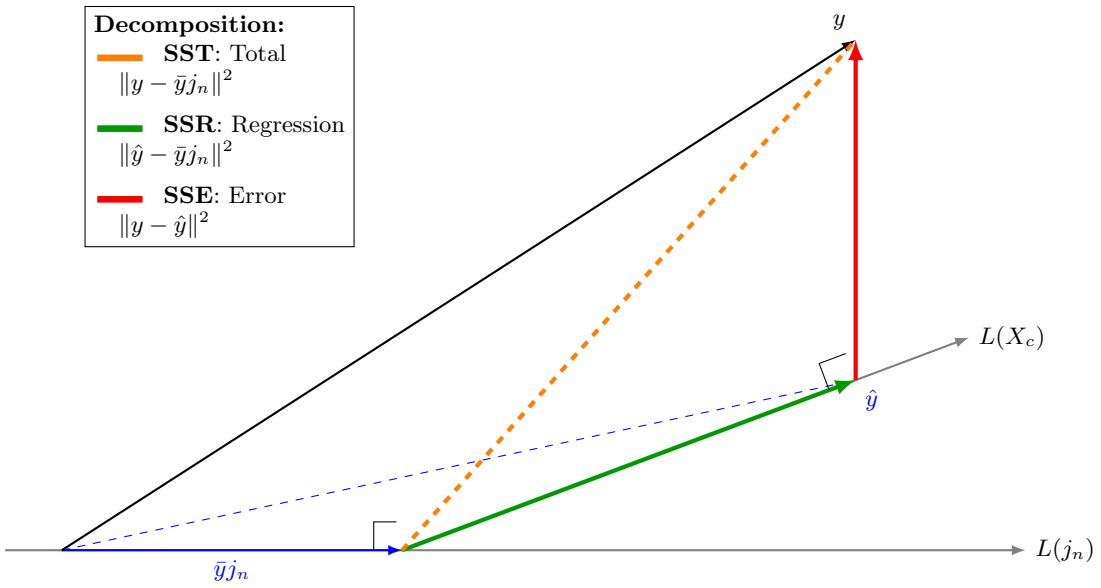


Figure 6.3: Geometric Decomposition: SST = SSR + SSE

2. ***Distribution of SSE***

   The scaled sum of squared errors follows a central Chi-squared distribution:

$$\frac{SSE}{\sigma^2} = \frac{\|(I-H)y\|^2}{\sigma^2} \sim \chi^2(n-k-1) \tag{6.68}$$

   Mean:

$$E[SSE] = \sigma^2(n-k-1) \tag{6.69}$$

3. ***Distribution of SSR***

   The scaled regression sum of squares follows a ***non-central*** *Chi-squared distribution:*

$$\frac{SSR}{\sigma^2} = \frac{\|P_{X_c}y\|^2}{\sigma^2} \sim \chi^2(k,\lambda) \tag{6.70}$$

   Mean:

$$E[SSR] = \sigma^2 k + \|P_{X_c}\mu_y|^2 \tag{6.71}$$

   Non-centrality Parameter (λ):

$$\lambda = \frac{1}{\sigma^2}\|P_{X_c}\mu_y\|^2 \tag{6.72}$$

   where

$$\boxed{\|P_{X_c}\mu_y\|^2 = \|X_c\beta_1\|^2 = (X_c\beta_1)'(X_c\beta_1) = \beta_1' X_c' X_c \beta_1} \tag{6.73}$$

Proof. We apply Theorem 5.8 to the specific projection matrices identified in the definitions.

- **For SSE (Error Space):** SSE is defined by the projection matrix $P_V = I - H$.
  - **Dimension:** The rank of $(I - H)$ is $n - \text{rank}(X) = n - (k + 1) = n - k - 1$.
  - **Non-centrality:** Since $\mu \in \text{Col}(X)$, the projection onto the orthogonal complement is zero: $\|(I - H)\mu\|^2 = 0$. Thus, $\lambda = 0$.
  - **Expectation:** Using Part 2 of Theorem 5.8 ($E(\|P_V y\|^2) = \sigma^2 \text{rank}(P_V) + \|P_V \mu\|^2$):

$$E[\text{SSE}] = \sigma^2(n - k - 1) + 0 = \sigma^2(n - k - 1) \tag{6.74}$$

- **For SSR (Regression Space):** SSR is defined by the projection matrix $P_V = P_{X_c}$.
  - **Dimension:** The rank of P_{X_c} is $(k + 1) - 1 = k$.
  - **Non-centrality:** The projection of μ onto $L(X_c)$ is $P_{X_c}\mu_y$.

$$\lambda = \frac{\|P_{X_c}\mu_y\|^2}{\sigma^2} \tag{6.75}$$

  - **Expectation:** Using Part 2 of Theorem 5.8:

$$E[\text{SSR}] = \sigma^2 k + \|P_{X_c}\mu_y\|^2 \tag{6.76}$$

This shows that while $E[\text{SSE}]$ depends only on the noise variance and sample size, $E[\text{SSR}]$ is inflated by the magnitude of the true regression signal $\|P_{X_c}\mu_y\|^2$.

□

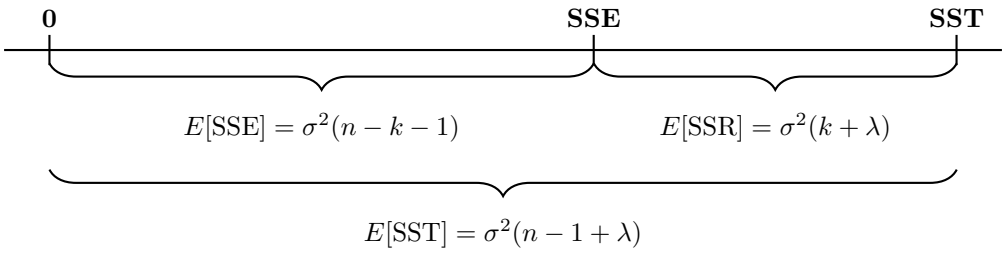


Figure 6.4: Stick Diagram of Mean of SSE, SSR, and SST

6.9 F-test for Testing Overall Regression Effect

We wish to test whether the regression model provides any explanatory power beyond the simple intercept-only model.

Hypotheses:

- **Null Hypothesis (H_0):** $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (No regression effect). This implies $\mu \in \text{span}(j_n)$ and the true signal variance $\|X_c\beta_1\|^2 = 0$.
- **Alternative Hypothesis (H_1):** At least one $\beta_j \neq 0$.

The F-statistic

We construct the test statistic using the ratio of the Mean Squares defined previously:

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} \tag{6.77}$$

Understanding F via Expectations

The logic of the F-test is transparent when we examine the expected values of the numerator and denominator:

$$\begin{aligned} E[\text{MSE}] &= \sigma^2 \\ E[\text{MSR}] &= \sigma^2 + \frac{\|X_c\beta_1\|^2}{k} \end{aligned} \tag{6.78}$$

- **If H_0 is true:** The signal term is zero. Both Mean Squares estimate σ^2 unbiasedly. We expect $F \approx 1$.
- **If H_1 is true:** The numerator includes the positive term $\frac{\|X_c\beta_1\|^2}{k}$. We expect $F > 1$.

Therefore, we reject H_0 for sufficiently large values of F. Specifically, we reject at level α if $F_{obs} > F_\alpha(k, n-k-1)$.

6.9.1 Distributional Theory

To derive the exact sampling distribution, we rely on the independence of the sums of squares (from Theorem 6.9) and the definition of the non-central F-distribution given in **Definition 5.3**.

Theorem 6.10 (Distribution of Regression F-Statistic)**.** *Under the assumption of normality, the regression F-statistic follows a* ***non-central F-distribution****:*

$$F \sim F(k, n-k-1, \lambda) \tag{6.79}$$

The non-centrality parameter λ is determined by the ratio of the signal sum of squares to the error variance:

$$\lambda = \frac{\|X_c\beta_1\|^2}{\sigma^2} \tag{6.80}$$

Special Cases:

1. ***Under H_1 (Signal exists):*** *$\lambda > 0$, so F follows the non-central distribution.*
2. ***Under H_0 (No signal):*** *$\beta_1 = 0 \implies \lambda = 0$. The distribution collapses to the* ***central F-distribution****:*

$$F \sim F(k, n-k-1) \tag{6.81}$$

Proof. We identify the components from Definition 5.3:

1. **Numerator (X_1):** Let $X_1 = \text{SSR}/\sigma^2$. From Theorem 6.9, $X_1 \sim \chi^2(k, \lambda)$.
2. **Denominator (X_2):** Let $X_2 = \text{SSE}/\sigma^2$. From Theorem 6.9, $X_2 \sim \chi^2(n-k-1)$.
3. **Independence:** X_1 and X_2 are independent.

Substituting these into the F-statistic:

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{(\text{SSR}/\sigma^2)/k}{(\text{SSE}/\sigma^2)/(n-k-1)} = \frac{X_1/k}{X_2/(n-k-1)} \tag{6.82}$$

By definition Definition 5.3, this ratio follows $F(k, n-k-1, \lambda)$. □

6.9.2 Visualization of the Rejection Region

The following plot illustrates the central F-distribution (valid under H_0) for $k = 3$ predictors and $n = 20$ observations ($df_1 = 3, df_2 = 16$). An observed statistic of $F = 2$ is marked, with the p-value represented by the shaded tail area.

6.10 Optimistic Bias in Raw Coefficient of Determination (R^2)

Definition

The R^2 statistic measures the proportion of total variation explained by the regression model. It is formally defined as the ratio of the Regression Sum of Squares to the Total Sum of Squares.

Definition 6.4 (R-Squared)**.**

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{6.83}$$

Since $0 \leq \text{SSE} \leq \text{SST}$, it follows that $0 \leq R^2 \leq 1$.

Relationship to MLE Variances

A crucial insight is that the unexplained variance term $(1-R^2)$ is simply the ratio of the **biased** Maximum Likelihood Estimators for the error variance and the total variance.

Recall that:

$$\hat{\sigma}^2_{\text{MLE},e} = \frac{\text{SSE}}{n} \quad \text{and} \quad \hat{\sigma}^2_{\text{MLE},y} = \frac{\text{SST}}{n} \tag{6.84}$$

Therefore:

$$1 - R^2 = \frac{\text{SSE}}{\text{SST}} = \frac{n \cdot \hat{\sigma}^2_{\text{MLE},e}}{n \cdot \hat{\sigma}^2_{\text{MLE},y}} = \frac{\hat{\sigma}^2_{\text{MLE},e}}{\hat{\sigma}^2_{\text{MLE},y}} \tag{6.85}$$

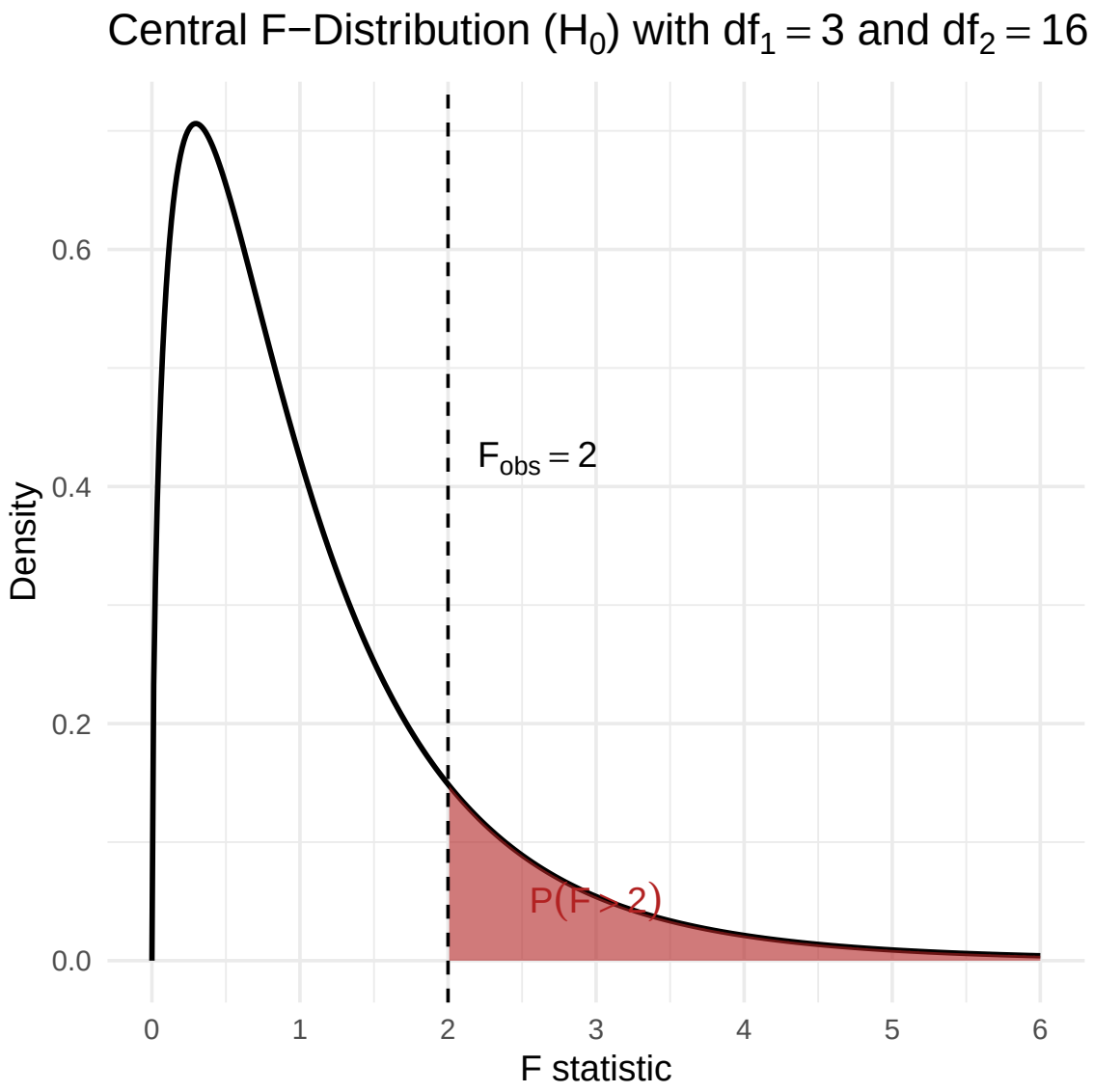


Figure 6.5: Probability Density Function of F(3, 16) under H0. The shaded region represents the p-value.

This highlights that R^2 is constructed from estimators that divide by n, ignoring degrees of freedom.

Exact Distribution

The R^2 statistic follows the Type I Non-central Beta distribution derived from the ratio of independent Chi-squared variables.

Theorem 6.11 (Distribution of R-Squared)**.**

$$R^2 \sim Beta_1\left(\frac{k}{2}, \frac{n-k-1}{2}, \lambda\right) \tag{6.86}$$

where the shape parameters correspond to half the degrees of freedom: $\alpha = k/2$ *and* $\beta = (n-k-1)/2$.

Expectation and Bias

To understand the bias in R^2, we analyze the expectation of this ratio.

1. General Approximation:

Using the first-order approximation $E[X/Y] \approx E[X]/E[Y]$:

$$E[1-R^2] \approx \frac{E[\text{SSE}]}{E[\text{SST}]} = \frac{\sigma^2(n-k-1)}{\sigma^2(n-1+\lambda)} = \frac{n-k-1}{n-1+\lambda} \tag{6.87}$$

2. Exact Behavior under Null Hypothesis (H_0):

When there is no true signal ($\beta_1 = 0$), the non-centrality parameter λ vanishes. In this specific case, R^2 follows a central Beta distribution with shape parameters $\alpha = k/2$ and $\beta = (n-k-1)/2$.

Using the standard mean of a Beta distribution ($E[X] = \frac{\alpha}{\alpha+\beta}$), we find that the expectation is **exact**:

$$E[R^2|H_0] = \frac{k/2}{k/2 + (n-k-1)/2} = \frac{k}{n-1} \tag{6.88}$$

Consequently, the expected unexplained variance is:

$$E[1-R^2|H_0] = \frac{n-k-1}{n-1} \tag{6.89}$$

(This is an exact equality, explained by the properties of the Beta distribution).

! Source of Bias

The expectation is **strictly less than 1**. This confirms that R^2 is **positively biased** (it inflates the perceived fit). The model "eats up" k degrees of freedom to fit noise, reducing the SSE artificially relative to the SST. This specific bias factor $\frac{n-k-1}{n-1}$ is the exact inverse of the correction applied by **Adjusted R-squared (R^2_a)**.

A Visualization of Bias

The figure below aligns these three scales. Note how the Unbiased Estimator for error (s_e^2) shifts significantly to the right of the MLE ($\hat{\sigma}_e^2$) due to the loss of $k+1$ degrees of freedom, whereas the Total Variance estimator (s_y^2) shifts only slightly.

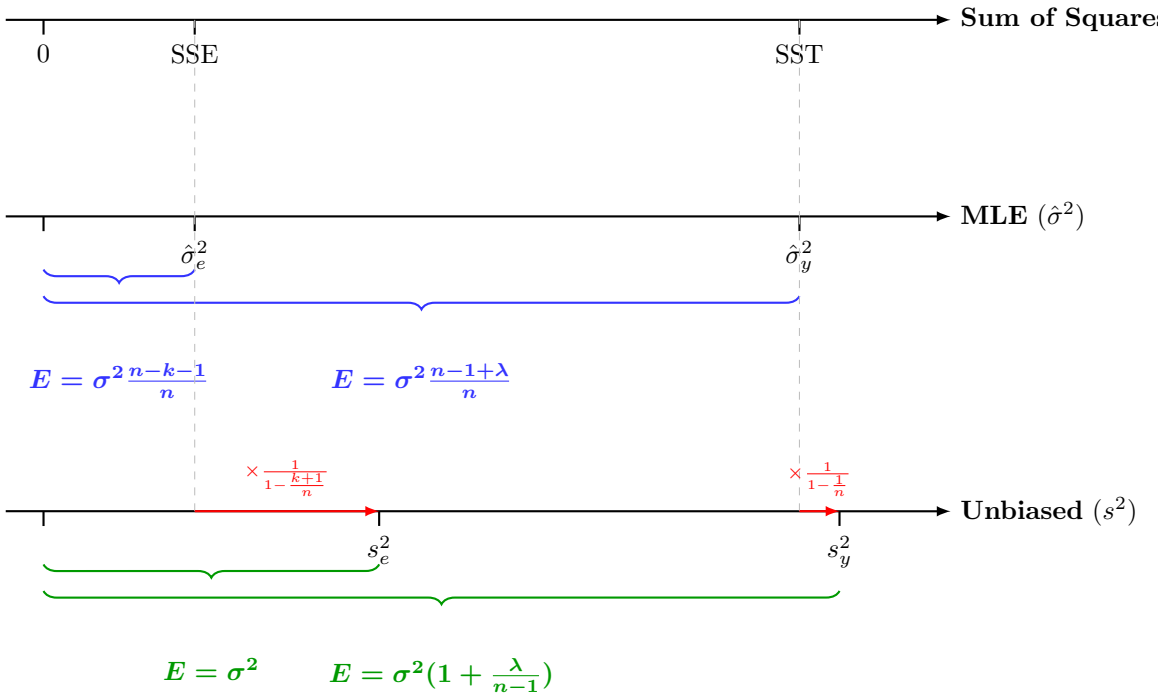


Figure 6.6: Comparison of Variance Estimators.

6.11 Adjusted R-squared (R_a^2) and Population Proportion (ρ^2)

To correct for the inflation of R^2 due to model complexity, we introduce the Adjusted R^2.

6.11.1 Definition

The Adjusted R^2 is defined not as a ratio of Sums of Squares (which are biased by sample size), but as the ratio of the **unbiased variance estimators**: the Mean Squared Error (s_e^2) and the Sample Variance of Y (s_y^2).

Definition 6.5 (Adjusted R-squared)**.**

$$R_a^2 = 1 - \frac{s_e^2}{s_y^2} = 1 - \frac{\text{MSE}}{\text{MST}} \tag{6.90}$$

This definition naturally incorporates the degrees of freedom correction:

$$R_a^2 = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k-1} \tag{6.91}$$

6.11.2 Expectation in terms of λ

We can derive the expectation of R_a^2 using the first-order approximation $E[f/g] \approx E[f]/E[g]$.

Recall the expectations of the mean squares:

1. **Error Mean Square:**

$$E[\text{MSE}] = \sigma^2 \tag{6.92}$$

2. **Total Mean Square:**

$$E[\text{MST}] = \frac{E[\text{SST}]}{n-1} = \frac{\sigma^2(n-1+\lambda)}{n-1} = \sigma^2\left(1 + \frac{\lambda}{n-1}\right) \tag{6.93}$$

Substituting these into the expectation formula:

$$\begin{aligned} E[R_a^2] &\approx 1 - \frac{E[\text{MSE}]}{E[\text{MST}]} \\ &= 1 - \frac{\sigma^2}{\sigma^2\left(1 + \frac{\lambda}{n-1}\right)} \\ &= 1 - \frac{1}{1 + \frac{\|X_c\beta\|^2}{(n-1)\sigma^2}} \\ &= \frac{\frac{\|X_c\beta\|^2}{n-1}}{\sigma^2 + \frac{\|X_c\beta\|^2}{n-1}} \end{aligned} \tag{6.94}$$

6.11.3 Definitions of Population Metrics for Predictivity

The term in the numerator relies on the squared norm of the centered true means. We can expand the centered signal vector $X_c\beta$ to see this explicitly. Since $\mu \in \text{Col}(X)$, we know $H\mu = \mu$:

$$X_c\beta_1 = P_{X_c}\mu_y = (H - P_{j_n})\mu = \mu - \bar{\mu}j_n = \begin{pmatrix} \mu_1 - \bar{\mu} \\ \mu_2 - \bar{\mu} \\ \vdots \\ \mu_n - \bar{\mu} \end{pmatrix} \tag{6.95}$$

This vector represents the deviation of each observation's true mean from the grand mean. Consequently, we define the **Signal Variance** (σ_μ^2) as the average squared deviation.

Definition 6.6 (Population Signal Variance)**.**

$$\sigma_\mu^2 = \frac{\|X_c\beta\|^2}{n} = \frac{\sum_{i=1}^n (\mu_i - \bar{\mu})^2}{n} \tag{6.96}$$

Using this definition, we can define two key population parameters that characterize the strength of the relationship.

Population Coefficient of Determination (ρ^2)

The expected Adjusted R^2 estimates the proportion of the **total variance** ($\sigma_Y^2 = \sigma_\mu^2 + \sigma^2$) that is attributable to the signal.

Definition 6.7 (Population ρ^2)**.**

$$\rho^2 = \frac{\text{Signal Variance}}{\text{Total Variance}} = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma^2} \tag{6.97}$$

Cohen's Effect Size (f^2)

In power analysis, it is often more useful to look at the **Signal-to-Noise Ratio (SNR)** directly. This is known as Cohen's f^2.

Definition 6.8 (Cohen's f^2 (Signal-to-Noise Ratio))**.**

$$f^2 = \frac{\text{Signal Variance}}{\text{Noise Variance}} = \frac{\sigma_\mu^2}{\sigma^2} \tag{6.98}$$

Relationships and Non-Centrality Parameter (λ)

There is a simple functional relationship between ρ^2 and f^2:

$$f^2 = \frac{\rho^2}{1 - \rho^2} \tag{6.99}$$

Finally, we can express the non-centrality parameter λ directly in terms of these parameters. Since $\lambda = \frac{\|X_c\beta\|^2}{\sigma^2}$ and $\|X_c\beta\|^2 = (n-1)\sigma_\mu^2$:

$$\lambda = n\frac{\sigma_\mu^2}{\sigma^2} = nf^2 \tag{6.100}$$

Substituting f^2 with ρ^2, we obtain the mapping used to invert the F-test for confidence intervals:

$$\boxed{\lambda = nf^2 = n\frac{\rho^2}{1 - \rho^2}} \tag{6.101}$$

6.11.4 Remarks on Variance Estimation and Effect Size

1. **Fixed vs. Random Covariates** In the fixed covariate framework, the "parameter" ρ^2 is a function of the specific design matrix X, the coefficients β, and the sample size n. If we assume the x_i are random draws from a population, then as $n \to \infty$, σ_μ^2 converges to $\text{Var}(x^T\beta)$ (where x is a random vector), and ρ^2 converges to the true population proportion of variance explained.

2. **MSR Is Not a Variance Estimator** It is critical to distinguish between hypothesis testing statistics and variance estimators:

   - **Estimating Signal Variance (σ_μ^2):** Observing that $E[\text{MST}] = \sigma^2 + \sigma_\mu^2$ and $E[\text{MSE}] = \sigma^2$, the difference MST $-$ MSE provides a direct method-of-moments estimator for the variance of the signal itself.
   - **Testing for Signal Existence (MSR):** The commonly used **Mean Square Regression (MSR)**, defined as SSR$/k$, is **not** an estimator of the signal variance. Because $E[\text{MSR}] = \sigma^2 + \frac{n-1}{k}\sigma_\mu^2$, it scales linearly with the sample size n. MSR is designed to explode as $n \to \infty$ to ensure power for hypothesis testing, not to estimate the magnitude of the signal.

3. **Significance vs. Predictive Effect Size** The F-test p-value measures **statistical significance**—the strength of evidence against the null hypothesis—rather than the magnitude of the effect. In large datasets, even a negligible predictive effect (a very small ρ^2) can produce a highly significant p-value. Conversely, ρ^2 and R_a^2 provide measures of **predictive effect size**, indicating the practical utility of the model regardless of the sample size.

6.11.5 Confidence Interval of Population ρ^2

While R_a^2 provides a point estimate, we can construct an exact confidence interval for ρ^2 by exploiting the distribution of the F-statistic.

1. The link between λ, f^2, and ρ^2

   Recall that the F-statistic follows a non-central distribution $F(k, n-k-1, \lambda)$. The non-centrality parameter λ is directly related to the population parameters. Using our definition of signal variance σ_μ^2 (with the $n-1$ divisor):

   $$\lambda = \frac{\|X_c\beta\|^2}{\sigma^2} = n\frac{\sigma_\mu^2}{\sigma^2} \tag{6.102}$$

   Substituting Cohen's effect size $f^2 = \frac{\sigma_\mu^2}{\sigma^2}$ and the relationship $f^2 = \frac{\rho^2}{1-\rho^2}$, we obtain the following mapping:

   $$\boxed{\lambda = nf^2 = n\left(\frac{\rho^2}{1-\rho^2}\right)} \tag{6.103}$$

   To recover ρ^2 from λ, we invert the mapping:

   $$\boxed{\rho^2(\lambda) = \frac{\lambda}{\lambda + n}} \tag{6.104}$$

   Remark 6.1 (Remark on Divisor Conventions). It is important to note that different authors and software packages use different conventions for the divisor in the definition of signal variance. For example, the R package `MBESS` for fixed predictors defines signal variance as $\sigma_\mu^2 = \|X_c\beta\|^2/n$. With this definiton, $\lambda = n\frac{\rho^2}{1-\rho^2}$.

2. Inverting the Test Statistic

   We find a confidence interval $[\lambda_L, \lambda_U]$ for λ by "inverting" the observed F-statistic (F_{obs}). We search for two specific non-central F-distributions: one where F_{obs} cuts off the upper $\alpha/2$ tail, and one where it cuts off the lower $\alpha/2$ tail.

   This concept is illustrated in the figure below.

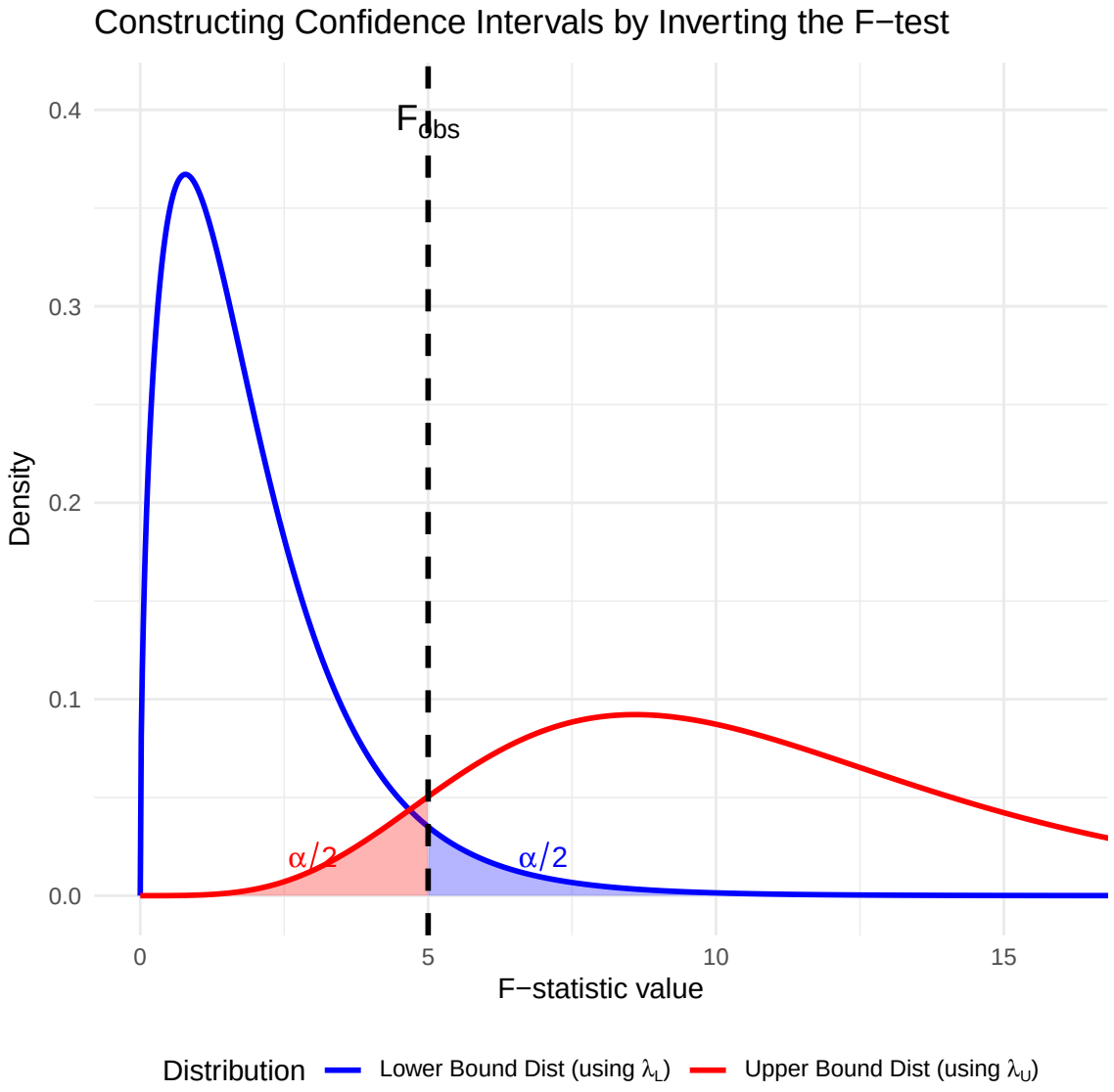


Figure 6.7: Illustration of constructing a confidence interval for the non-centrality parameter λ by inverting the F-test. The observed F_{obs} (dashed line) is the 97.5^{th} percentile of the distribution defined by the lower bound λ_L (blue), and the 2.5^{th} percentile of the distribution defined by the upper bound λ_U (red).

3. The Interval for ρ^2

Once $[\lambda_L, \lambda_U]$ are found numerically, we map them back to the population R^2 scale using the updated inverse relationship:

$$\rho^2 = \frac{\lambda}{\lambda + n} \tag{6.105}$$

This produces an exact confidence interval $[\rho_L^2, \rho_U^2]$ for the proportion of variance explained by the model in the population.

4. R function for finding the CI with F quantiles

6.12 An Animation for Illustrating R_a^2 Under H_0 and H_1

We simulate a dataset with $n = 30$ observations and consider a sequence of nested models adding groups of predictors.

Predictor Groups:

1. **Group 1 ($k = 1$):** Add x_1. (Signal under H_1).
2. **Group 2 ($k = 6$):** Add $x_2, \ldots, x_6$ (Noise).
3. **Group 3 ($k = 11$):** Add $x_7, \ldots, x_{11}$ (Noise).
4. **Group 4 ($k = 20$):** Add $x_{12}, \ldots, x_{20}$ (Noise).

6.12.0.1 Null Hypothesis (H_0)

Under H_0, the true coefficient for x_1 is $\beta_1 = 0$. All predictors are noise.

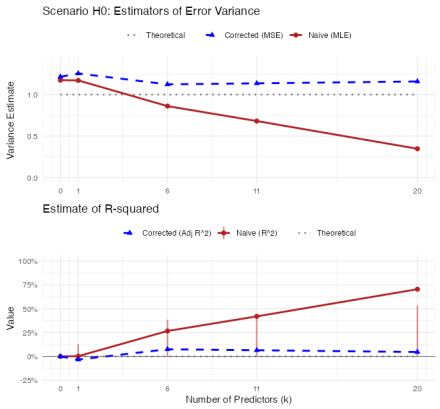


Figure 6.8: Simulation under H0: As predictors are added (pure noise), standard R-squared increases while Adjusted R-squared and MSE remain stable.

6.12.0.2 Alternative Hypothesis (H_1)

Under H_1, x_1 is a true predictor ($\beta_1 = 2$). The subsequent groups ($x_2 \dots x_{20}$) remain noise.

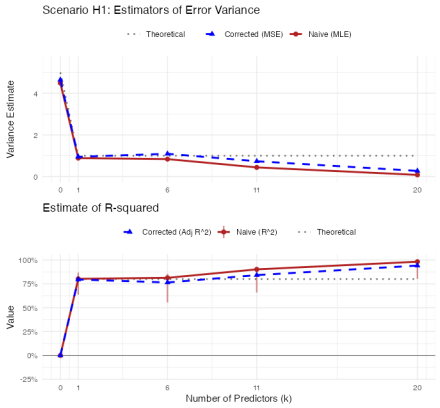


Figure 6.9: Simulation under H1: Adjusted R-squared correctly identifies the signal at k=1, then penalizes the subsequent noise predictors.

6.13 A Data Example with House Price Valuation

A real estate agency wants to refine their pricing model. They regress the selling price of houses (y) on five predictors (X): Size, Age, Bedrooms, Garage Capacity, and Lawn Size.

We assume the data has been collected and saved to `house_prices_5pred.csv`.

6.13.1 Visualize the Data

First, we load the dataset. We display the first 10 rows for PDF output, or a full paged table for HTML.

Table 6.1: First 10 rows of House Prices

Price	Size	Age	Beds	Garage	Lawn
497808	3092	4	3	2	426
364297	1802	26	5	0	88
610217	2701	22	4	1	403
536122	2745	38	4	0	437
347259	2143	18	2	1	141
343784	2754	49	5	1	186
379522	2039	53	4	0	451
341432	1758	43	5	1	832
515913	3191	19	4	0	276
292732	1298	17	2	2	804

6.13.2 Fit the Model

We will solve for the coefficients $\hat{\beta}$ using three distinct methods.

1. **Method 1: Naive Matrix Formula**

   This method solves the normal equations directly on the raw data: $\hat{\beta} = (X'X)^{-1}X'y$.

```
Matrix X'X (Cross-products of predictors):
```

	Intercept	Size	Age	Beds	Garage	Lawn
Intercept	60	136483	1674	206	80	29392
Size	136483	343078981	3738402	469757	177877	63939128
Age	1674	3738402	63528	5874	2353	827130
Beds	206	469757	5874	776	281	98738
Garage	80	177877	2353	281	196	41915
Lawn	29392	63939128	827130	98738	41915	19306096

```
Matrix X'y (Cross-products with response):
```

	[,1]
Intercept	25884407
Size	63115001244
Age	694594579
Beds	89683035
Garage	34067413
Lawn	12402228016

```
Solved Coefficients (Beta):
```

	Intercept	Size	Age	Beds	Garage	Lawn
[1,]	113186	129.3434	-1218.352	12664.16	875.1155	27.2443

2. **Method 2: Centralized Formula**

   This method reduces multicollinearity issues. Formula: $\hat{\beta}_{\text{slope}} = (X_c'X_c)^{-1}X_c'y_c$.

```
Matrix X_c'X_c (Centered Sum of Squares):
```

	Size	Age	Beds	Garage	Lawn
Size	32618826	-69474	1165	-4100	-2919344
Age	-69474	16823	127	121	7093
Beds	1165	127	69	6	-2175
Garage	-4100	121	6	89	2726
Lawn	-2919344	7093	-2175	2726	4907935

```
Matrix X_c'y_c (Centered Cross-products):
```

```
             [,1]
Size   4235309234
Age     -27580376
Beds       813238
Garage    -445130
Lawn   -277680160


Solved Coefficients (Beta):

     Intercept     Size       Age     Beds   Garage    Lawn
[1,]    113186 129.3434 -1218.352 12664.16 875.1155 27.2443
```

3. **Method 3: Using R's `lm` Function**

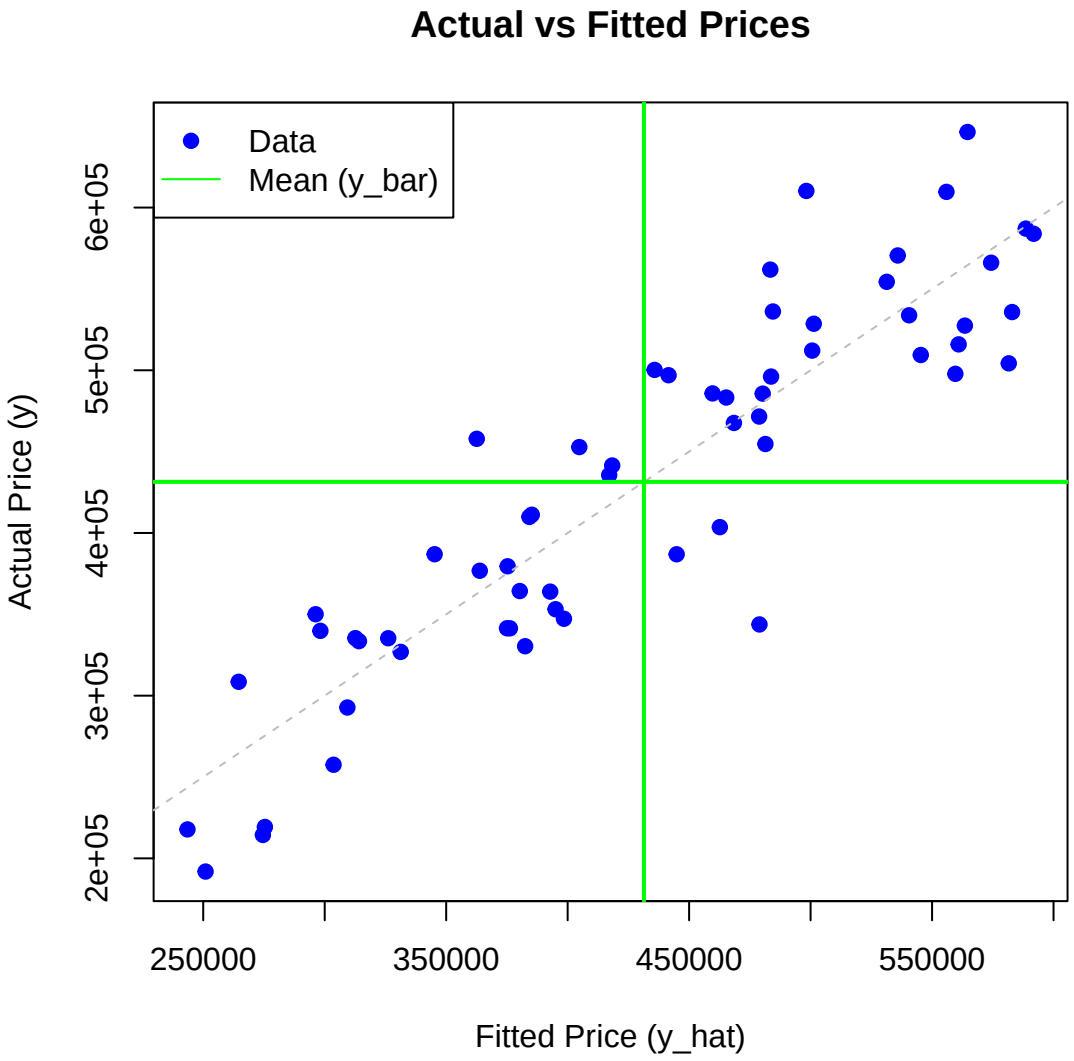
   This is the standard approach for practitioners.

```
Call:
lm(formula = Price ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-135178  -36006    1710   26401  111967

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 113185.971  35675.435   3.173  0.00249 **
Size           129.343      8.927  14.490  < 2e-16 ***
Age          -1218.352    386.414  -3.153  0.00264 **
Beds         12664.157   6064.435   2.088  0.04150 *
Garage         875.115   5316.490   0.165  0.86987
Lawn            27.244     23.243   1.172  0.24629
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49360 on 54 degrees of freedom
Multiple R-squared:  0.8161,    Adjusted R-squared:  0.799
F-statistic: 47.92 on 5 and 54 DF,  p-value: < 2.2e-16
```

6.13.3 Visualization of Fitted Values vs Mean

We define $\hat{y}_0$ as the vector of the mean of y ($\bar{y}$). We plot the actual y against our fitted model $\hat{y}$, using a green line to represent the "Null Model" ($\hat{y}_0$).

Note: Axes have been set so that X = Predicted Value and Y = Actual Value.



Question:

$$\bar{y} = \bar{\hat{y}}? \tag{6.106}$$

6.13.4 Computing Sums of Squares (SSE, SST, SSR)

We compare different methods to calculate the sources of variation.

1. **Naive Sum of Squared Errors**

   This uses the standard summation definitions: $\sum(Difference)^2$.

   - **SST (Total):** Variation of y around $\hat{y}_0$ (Mean).

- **SSR (Regression):** Variation of $\hat{y}$ around $\hat{y}_0$ (Mean).
- **SSE (Error):** Variation of y around $\hat{y}$ (Model).

```
Naive Calculation:

SST: 715333529746  SSR: 583756306788  SSE: 131577222958
```

2. **Pythagorean Shortcut (Vector Lengths)**

   Based on the geometry of least squares, we can treat the variables as vectors. Because the vectors are orthogonal, we can use squared lengths (dot products with themselves).

   Formula: SSR $= ||\hat{y}||^2 - ||\hat{y}_0||^2$

```
Pythagorean Calculation:

SST: 715333529746  SSR: 583756306788  SSE: 131577222958
```

3. **Matrix Algebra Shortcuts**

   These formulas use the β and X matrices directly. This is computationally efficient for large datasets.

   - Formula A (Centered with y_c): SSR $= \hat{\beta}_c' X_c' y_c$
   - Formula B (Alternative with y): SSR $= \hat{\beta}_c' X_c' y$
   - Formula C (Uncentered): SSR $= \hat{\beta}' X' y - n\bar{y}^2$

Table 6.2: Demonstration of SSR Formula Equivalence

Metric	Formula	Value
SSR (Centered X_c, y_c)	$\hat{\beta}_c' X_c' y_c$	583756306788
SSR (Centered X_c)	$\hat{\beta}_c' X_c' y$	583756306788
SSR (Uncentered)	$\hat{\beta}' X' y - n\bar{y}^2$	583756306788

6.13.5 Analysis of Variance (ANOVA)

We now evaluate the sources of variation to test the overall model significance.

1. **Computing Sums of Squares**

   We calculate the following components:

   - Total Sum of Squares: SST $= \sum (y_i - \bar{y})^2$
   - Regression Sum of Squares: SSR $= \sum (\hat{y}_i - \bar{y})^2$
   - Sum of Squared Errors: SSE $= \sum (y_i - \hat{y}_i)^2$

```
SST: 715333529746  SSR: 583756306788  SSE: 131577222958
```

2. **Manual ANOVA Construction**

   We build the table manually using the sums of squares and degrees of freedom. We calculate the Mean Squares and the F-statistic:

   - MSR $= \text{SSR}/k$
   - MSE $= \text{SSE}/(n-k-1)$
   - MST $= \text{SST}/(n-1)$
   - $F = \text{MSR}/\text{MSE}$

Table 6.3: Manual ANOVA Table

Source	DF	SS	MS	F_Statistic	P_Value
Regression (Model)	5	583756306788	116751261358	47.9153	0
Error (Residual)	54	131577222958	2436615240	NA	NA
Total	59	715333529746	12124297114	NA	NA

3. **Standard R Output (`anova`)**

   We display the standard `summary()` which provides the coefficients, t-tests, and the overall F-statistic found at the bottom. We also show `anova()` which gives the sequential sum of squares.

```
ANOVA comparing intercept-only model to fitted model:

Analysis of Variance Table

Model 1: Price ~ 1
Model 2: Price ~ Size + Age + Beds + Garage + Lawn
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1     59 7.1533e+11
2     54 1.3158e+11  5 5.8376e+11 47.915 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of Variance Table

Response: Price
          Df     Sum Sq    Mean Sq  F value    Pr(>F)
Size       1 5.4992e+11 5.4992e+11 225.6914 < 2.2e-16 ***
Age        1 2.0657e+10 2.0657e+10   8.4777  0.005216 **
Beds       1 9.5872e+09 9.5872e+09   3.9346  0.052396 .
Garage     1 2.4151e+08 2.4151e+08   0.0991  0.754107
Lawn       1 3.3476e+09 3.3476e+09   1.3739  0.246291
Residuals 54 1.3158e+11 2.4366e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.13.6 Coefficient of Determination and Variance Decomposition

We calculate R^2 and Adjusted R^2, and then present them in a **Variance Decomposition Table**.

1. **Calculation**

   We calculate the coefficients of determination:

   - Standard $R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$
   - Adjusted $R^2_a = 1 - \frac{\text{MSE}}{\text{MST}}$

```
Standard R^2:   0.8161

Adjusted R^2:   0.799
```

2. **Variance Decomposition Table**

   This table extends standard ANOVA. While ANOVA focuses on **Mean Squares (MS)** for hypothesis testing (is MSR $>$ MSE?), this table focuses on **Variance Components ($\hat{\sigma}^2$)** for estimation (how much variance is Signal vs. Noise?). We estimate the variance components as follows:

   - Signal Variance: $\hat{\sigma}^2_\mu = \text{MST} - \text{MSE}$
   - Noise Variance: $\hat{\sigma}^2 = \text{MSE}$
   - Total Variance: $\hat{\sigma}^2_Y = \text{MST}$
   - **Signal Variance ($\hat{\sigma}^2_\mu$):** Estimated by MST $-$ MSE. (Note: MSR is biased and overestimates signal).
   - **Noise Variance ($\hat{\sigma}^2$):** Estimated by MSE.
   - **Total Variance ($\hat{\sigma}^2_Y$):** Estimated by MST.

Table 6.4: Variance Decomposition Table: Estimating Signal vs. Noise

Component	DF	SS	MS	Value ($\hat{\sigma}^2$)	Proportion
Signal (Model)	5	583756306788	NA	9687681874	0.799
Noise (Error)	54	131577222958	2436615240	2436615240	0.201
Total (Y)	59	715333529746	12124297114	12124297114	1.000

6.13.7 Confidence Interval for Population R^2 (ρ^2)

We construct a 95% confidence interval for the population proportion of variance explained (ρ^2).

1. **Manual Inversion Method**

   We solve for the non-centrality parameters λ_L and λ_U such that our observed F_{obs} corresponds to the appropriate quantiles.

```
Manual Calculation using ci_R2_F:
```

```
95% CI for Population Rho^2: [ 0.6982 ,  0.8556 ]
```

2. **Using R Package `MBESS`**

   The `MBESS` package automates this procedure. We use `Random.Predictors = FALSE` to match the fixed-predictor assumption used in our manual calculation.

```
$Lower.Conf.Limit.R2
[1] 0.6982442

$Prob.Less.Lower
[1] 0.025

$Upper.Conf.Limit.R2
[1] 0.8555948

$Prob.Greater.Upper
[1] 0.025
```

6.14 Underfitting and Overfitting

We compare the properties of two competing estimators for the mean response vector $\mu = E[y]$.

6.14.1 Notation and Setup

We consider the general linear model:

$$y = X\beta + e = X_1\beta_1 + X_2\beta_2 + e \tag{6.107}$$

where X_1 is $n \times p_1$, X_2 is $n \times p_2$, and $\text{Var}(e) = \sigma^2 I$.

We distinguish between two estimation approaches based on this model:

1. Full Model (M_1)

We estimate β without restrictions. The estimator projects y onto the full column space $\text{Col}(X)$.

$$\begin{aligned} P_1 &= X(X^TX)^{-1}X^T \quad \text{(Projection onto Col}(X)\text{)} \\ \hat{y}_1 &= P_1 y \qquad\qquad\qquad \text{(Unrestricted Estimator)} \end{aligned} \tag{6.108}$$

2. Reduced Model (M_0)

We estimate β subject to the constraint:

$$M_0 : \beta_2 = 0 \tag{6.109}$$

This effectively reduces the model to $y = X_1\beta_1 + e$, projecting y onto the subspace $\text{Col}(X_1)$.

$$\begin{aligned} P_0 &= X_1(X_1^TX_1)^{-1}X_1^T \quad \text{(Projection onto Col}(X_1)\text{)} \\ \hat{y}_0 &= P_0 y \qquad\qquad\qquad\quad \text{(Restricted Estimator)} \end{aligned} \tag{6.110}$$

Key Geometric Property: Since the constraint $\beta_2 = 0$ restricts the estimation to a subspace ($\text{Col}(X_1) \subset \text{Col}(X)$), we have the nesting property:

$$P_1 P_0 = P_0 \quad \text{and} \quad P_1 - P_0 \text{ is a projection matrix.} \tag{6.111}$$

6.14.2 Case 1: Underfitting

The Truth: The Full Model (M_1) is correct.

$$y = X_1\beta_1 + X_2\beta_2 + e, \quad \beta_2 \neq 0 \tag{6.112}$$

The true mean is $\mu = X_1\beta_1 + X_2\beta_2$.

We analyze the properties of the **Reduced Estimator** $\hat{y}_0$ (from M_0) compared to the correct Full Estimator $\hat{y}_1$ (from M_1).

Theorem 6.12 (Bias-Variance Tradeoff in Underfitting)**.** *When M_1 is true:*

1. ***Bias:*** *The estimator $\hat{y}_0$ is **biased**, while $\hat{y}_1$ is unbiased.*

$$\mathit{Bias}(\hat{y}_0) = -(I - P_0)X_2\beta_2 \tag{6.113}$$

2. ***Variance:*** *The estimator $\hat{y}_0$ has **smaller variance** (matrix difference is positive semidefinite).*

$$\mathit{Var}(\hat{y}_1) - \mathit{Var}(\hat{y}_0) = \sigma^2(P_1 - P_0) \geq 0 \tag{6.114}$$

Proof. **Part 1 (Bias):**

$$\begin{aligned} E[\hat{y}_0] &= P_0 E[y] = P_0(X_1\beta_1 + X_2\beta_2) \\ &= X_1\beta_1 + P_0X_2\beta_2 \quad (\text{Since } P_0X_1 = X_1) \end{aligned} \tag{6.115}$$

The bias is:

$$\text{Bias} = E[\hat{y}_0] - \mu = (X_1\beta_1 + P_0X_2\beta_2) - (X_1\beta_1 + X_2\beta_2) = -(I - P_0)X_2\beta_2 \tag{6.116}$$

Part 2 (Variance):

$$\text{Var}(\hat{y}_1) = \sigma^2 P_1, \quad \text{Var}(\hat{y}_0) = \sigma^2 P_0 \tag{6.117}$$

The difference is $\sigma^2(P_1 - P_0)$. Since $\text{Col}(X_1) \subset \text{Col}(X)$, the difference $P_1 - P_0$ projects onto the orthogonal complement of $\text{Col}(X_1)$ within $\text{Col}(X)$. It is idempotent and positive semidefinite. □

Remark: Scalar Variance and Coefficients

From the matrix inequality above, we can state that for any arbitrary vector a, the scalar variance of the linear combination $a^T\hat{y}$ is always smaller in the reduced model:

$$\text{Var}(a^T\hat{y}_0) \leq \text{Var}(a^T\hat{y}_1) \tag{6.118}$$

We can extend this property to the regression coefficients $\hat{\beta}$. Since $\hat{y} = X\hat{\beta}$, we can recover the coefficients from the fitted values using the left pseudo-inverse:

$$\begin{aligned} (X^T X)^{-1} X^T (X\hat{\beta}) &= (X^T X)^{-1} X^T \hat{y} \\ \underbrace{(X^T X)^{-1}(X^T X)}_{I} \hat{\beta} &= (X^T X)^{-1} X^T \hat{y} \end{aligned} \tag{6.119}$$

Corollary 6.3 (Variance of Coefficients)**.** *Because $\hat{\beta}$ is a linear transformation of $\hat{y}$, the variance reduction in $\hat{y}_0$ propagates to the coefficients.*
For any specific coefficient β_j included in the reduced model (i.e., $\beta_j \in \beta_1$), the variance of the estimator is smaller in the reduced model than in the full model:

$$Var(\hat{\beta}_{j,reduced}) \leq Var(\hat{\beta}_{j,full}) \tag{6.120}$$

Conclusion: Using M_0 when M_1 is true introduces bias but reduces variance for both the fitted values and the estimated coefficients.

6.14.3 Case 2: Overfitting

The Truth: The Reduced Model (M_0) is correct.

$$y = X_1\beta_1 + e \quad \text{(i.e., } \beta_2 = 0\text{)} \tag{6.121}$$

The true mean is $\mu = X_1\beta_1$.

We analyze the properties of the **Full Estimator** $\hat{y}_1$ (from M_1) compared to the correct Reduced Estimator $\hat{y}_0$ (from M_0).

Theorem 6.13 (Variance Inflation in Overfitting)**.** *When M_0 is true:*

1. ***Bias:*** *Both estimators are **unbiased**.*

$$E[\hat{y}_1] = \mu \quad \textit{and} \quad E[\hat{y}_0] = \mu \tag{6.122}$$

2. ***Variance:*** *The estimator $\hat{y}_1$ has **unnecessarily higher variance**.*

$$Var(\hat{y}_1) \geq Var(\hat{y}_0) \tag{6.123}$$

Proof. **Part 1 (Bias):** Since $\mu = X_1\beta_1$:

$$E[\hat{y}_1] = P_1 X_1 \beta_1 = X_1\beta_1 = \mu \quad (\text{Since } X_1 \in \text{Col}(X)) \tag{6.124}$$

$$E[\hat{y}_0] = P_0 X_1 \beta_1 = X_1\beta_1 = \mu \quad (\text{Since } X_1 \in \text{Col}(X_1)) \tag{6.125}$$

Part 2 (Variance): As shown in Case 1, the difference is $\sigma^2(P_1 - P_0)$. The cost of overfitting is purely variance

inflation. The total variance (trace) increases by the number of unnecessary parameters (p_2):

$$\text{tr}(\text{Var}(\hat{y}_1)) - \text{tr}(\text{Var}(\hat{y}_0)) = \sigma^2(\text{tr}(P_1) - \text{tr}(P_0)) = \sigma^2(p_{full} - p_{reduced}) = \sigma^2 p_2 \tag{6.126}$$

□

Conclusion: Using M_1 when M_0 is true offers no benefit in bias but strictly increases estimation variance.

7 Hypothesis Testing in Linear Models

7.1 Testing Reduced Model vs Full Model (Partial F-test)

Consider the general linear model partitioned as follows:

$$y = X\beta + e = (X_1, X_2)\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix} + e = X_1\beta_1 + X_2\beta_2 + e \tag{7.1}$$

where X is an $n \times (k+1)$ matrix, X_1 corresponds to the first set of predictors, and X_2 corresponds to the remaining h predictors . We assume $e \sim N(0, \sigma^2 I)$.

We are often interested in testing the hypothesis that the second set of coefficients is zero:

$$H_0 : \beta_2 = 0 \tag{7.2}$$

This leads to a comparison between two models:

1. **Full Model (FM)**: The maintained hypothesis where $\mu \in C(X) = C([X_1, X_2])$.

$$y = X_1\beta_1 + X_2\beta_2 + e \tag{7.3}$$

2. **Reduced Model (RM)**: The model under H_0 where $\mu \in C(X_1)$.

$$y = X_1\beta_1 + e^* \tag{7.4}$$

The testing problem is to test $H_0 : \mu \in C(X_1)$ versus $H_1 : \mu \in C(X)$ but $\mu \notin C(X_1)$.

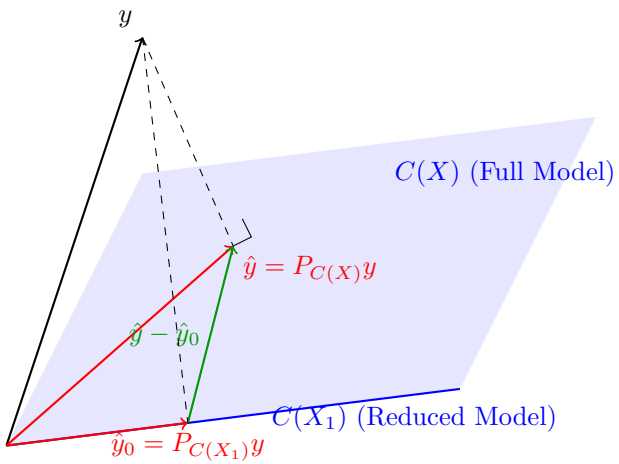


Figure 7.1: Geometric interpretation of the Full vs Reduced Model. The vector y is projected onto the Full Model space C(X) to get y_hat, and onto the Reduced Model space C(X1) to get y_hat0.

7.1.1 Geometric Interpretation

Since $C(X_1) \subset C(X)$, if the Reduced Model is true, the Full Model must also be true. Under H_0, the least squares estimates of the mean μ from both models, $P_{C(X_1)}y$ and $P_{C(X)}y$, estimate the same quantity.

This suggests that the difference vector :

$$P_{C(X)}y - P_{C(X_1)}y = (P_{C(X)} - P_{C(X_1)})y = \hat{y} - \hat{y}_0 \tag{7.5}$$

should be "small" under H_0. The matrix $P_{C(X)} - P_{C(X_1)}$ is the projection matrix onto $C(X_1)^\perp \cap C(X)$, which is the orthogonal complement of $C(X_1)$ with respect to $C(X)$.

The squared length of this difference is used as a test statistic:

$$\text{SSH} = ||\hat{y} - \hat{y}_0||^2 = y^T(P_{C(X)} - P_{C(X_1)})y \tag{7.6}$$

7.1.2 Distributional Properties

To derive a test, we analyze the distribution of the sum of squares.

Theorem 7.1 (Distribution of Quadratic Forms)**.** *Suppose $y \sim N(X\beta, \sigma^2 I)$ where X is $n \times (k+1)$ of full rank, $X\beta = X_1\beta_1 + X_2\beta_2$, and X_2 is $n \times h$. Let $\hat{y} = P_{C(X)}y$, $\hat{y}_0 = P_{C(X_1)}y$, and $\mu_0 = P_{C(X_1)}\mu$. Then:*

1. $\frac{1}{\sigma^2}||y - \hat{y}||^2 = \frac{1}{\sigma^2}y^T(I - P_{C(X)})y \sim \chi^2(n-k-1)$.
2. $\frac{1}{\sigma^2}||\hat{y} - \hat{y}_0||^2 = \frac{1}{\sigma^2}y^T(P_{C(X)} - P_{C(X_1)})y \sim \chi^2(h, \lambda_1)$.
   where the non-centrality parameter is $\lambda_1 = \frac{1}{\sigma^2}||(P_{C(X)} - P_{C(X_1)})\mu||^2 = \frac{1}{\sigma^2}||\mu - \mu_0||^2$.

3. *The two quadratic forms are independent.*

Under H_0, $\mu \in C(X_1)$, so $\mu = \mu_0$ and $\lambda_1 = 0$. Thus, the numerator sum of squares follows a central χ^2 distribution.

7.1.3 The F-Test

Based on the independence and distribution of the quadratic forms, we construct the F-statistic.

Theorem 7.2 (F-Test for Reduced Model)**.** *Under the conditions of Theorem 7.1, the statistic*

$$F = \frac{||\hat{y} - \hat{y}_0||^2/h}{||y - \hat{y}||^2/(n-k-1)} = \frac{SSH/h}{SSE_{FM}/(n-k-1)} \tag{7.7}$$

follows a non-central F-distribution $F(h, n-k-1, \lambda_1)$. Under $H_0 : \beta_2 = 0$, it follows a central F-distribution $F(h, n-k-1)$.

Using the Pythagorean Theorem, we can express the numerator in terms of Sum of Squares for Error (SSE) or Regression (SSR):

$$||\hat{y} - \hat{y}_0||^2 = SSE_{\text{RM}} - \text{SSE}_{\text{FM}} \tag{7.8}$$

$$||\hat{y} - \hat{y}_0||^2 = \text{SSR}_{\text{FM}} - \text{SSR}_{\text{RM}} \tag{7.9}$$

This quantity is often denoted as $\text{SS}(\beta_2|\beta_1)$, the "extra" regression sum of squares due to β_2 after accounting for β_1.

The ANOVA table for this comparison is constructed as follows:

Source	SS	MS	F
Hypothesis	$\text{SSR}_{\text{FM}} - \text{SSR}_{\text{RM}}$	SSH/h	$\frac{\text{SSH}/h}{\text{MSE}_{\text{FM}}}$
Error	SSE_{FM}	$\text{SSE}_{\text{FM}}/(n-k-1)$	
Total	SST		

7.1.4 Overall Regression Test

A special case of the test $H_0 : \beta_2 = 0$ is when β_1 contains only the intercept β_0, and β_2 contains all slope coefficients. The model is:

$$y = \beta_0 j_n + X_2\beta_2 + e \tag{7.10}$$

The hypothesis $H_0 : \beta_2 = 0$ is equivalent to $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. In this case:

1. The reduced model estimates $\hat{y}_0 = \bar{y} j_n$.
2. The degrees of freedom $h = k$.
3. The numerator becomes $\text{SSR}/k \equiv \text{MSR}$.

Theorem 7.3 (Overall Regression F-Statistic)**.** *The test statistic for overall regression is given by:*

$$F = \frac{SSR/k}{SSE_{FM}/(n-k-1)} = \frac{MSR}{MSE} \tag{7.11}$$

Under H_0, $F \sim F(k, n-k-1)$.

This statistic can be expressed in terms of the coefficient of determination R^2:

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} \tag{7.12}$$

7.2 The General Linear Hypothesis

We can generalize these tests to the hypothesis $H_0 : C\beta = t$, where C is a $q \times (k+1)$ matrix of rank $q \leq k+1$.

Common examples include :

- Testing a subset of coefficients: $C = (0, I_h)$.
- Overall regression: $C = (0, I_k)$.
- Equality of coefficients (e.g., $\beta_1 = \beta_2$).

7.2.1 Test Statistic for $C\beta = 0$

The test compares $C\hat{\beta}$ to its null value 0 using a squared statistical distance. Since $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$, it follows that $C\hat{\beta} \sim N(C\beta, \sigma^2 C(X^TX)^{-1}C^T)$.

Theorem 7.4 (F-Test for General Linear Hypothesis)**.** *If* $y \sim N(X\beta, \sigma^2 I)$ *and* C *has rank* q*, then under* $H_0 : C\beta = 0$ *:*

$$F = \frac{(C\hat{\beta})^T\{C(X^TX)^{-1}C^T\}^{-1}C\hat{\beta}/q}{SSE_{FM}/(n-k-1)} \sim F(q, n-k-1) \tag{7.13}$$

If H_0 *is false, it follows a non-central F distribution with parameter* $\lambda = \frac{(C\beta)^T[C(X^TX)^{-1}C^T]^{-1}C\beta}{\sigma^2}$.

7.2.2 Nested Models Interpretation

The general linear hypothesis test is fundamentally a test of nested models.

Theorem 7.5 (General Linear Hypothesis as Nested Models)**.** *The F-test for the general linear hypothesis $H_0 : C\beta = 0$ is equivalent to a full-and-reduced model test. Specifically, testing H_0 is equivalent to testing whether the mean vector μ lies in a subspace $V_0 \subset C(X)$. The test statistic satisfies:*

$$SSH = (C\hat{\beta})^T \{C(X^TX)^{-1}C^T\}^{-1} C\hat{\beta} = ||P_{C(Z)}y||^2 \tag{7.14}$$

where $C(Z)$ is the orthogonal complement of the reduced model space V_0 with respect to the full model space $C(X)$.

Proof. **Geometric Proof (Projections)**
Under the null hypothesis $H_0 : C\beta = 0$, we have a constrained model. We observe that:

$$C\beta = 0 \implies C(X^TX)^{-1}X^T(X\beta) = 0 \tag{7.15}$$

Let $Z^T = C(X^TX)^{-1}X^T$. Then the condition becomes $Z^T\mu = 0$ (since $\mu = X\beta$). This implies that under H_0, μ must be orthogonal to the column space of Z, denoted $\text{Col}(Z)$.
Since μ must also lie in the full model space $\text{Col}(X)$, under H_0, μ belongs to the intersection:

$$V_0 = \text{Col}(Z)^\perp \cap \text{Col}(X) \tag{7.16}$$

This V_0 represents the subspace for the Reduced Model (RM), while $\text{Col}(X)$ is the subspace for the Full Model (FM). The hypotheses correspond to nested models since $V_0 \subset \text{Col}(X)$.
The numerator sum of squares for comparing these models is the squared length of the projection of y onto the difference space. Since V_0 is the orthogonal complement of $\text{Col}(Z)$ relative to $\text{Col}(X)$, the difference space is exactly $\text{Col}(Z)$. Note that $Z = X(X^TX)^{-1}C^T$, and since C has rank q, Z has rank q.
The sum of squares for the hypothesis is:

$$\text{SSH} = ||(P_{\text{Col}(X)} - P_{V_0})y||^2 = ||P_{\text{Col}(Z)}y||^2 \tag{7.17}$$

Expanding the projection matrix $P_{\text{Col}(Z)} = Z(Z^TZ)^{-1}Z^T$:

$$y^T P_{\text{Col}(Z)} y = y^T Z(Z^TZ)^{-1}Z^T y \tag{7.18}$$

Substituting $Z = X(X^TX)^{-1}C^T$:

1. **Term Z^Ty**:

$$Z^Ty = C(X^TX)^{-1}X^Ty = C\hat{\beta} \tag{7.19}$$

2. **Term Z^TZ**:

$$Z^TZ = C(X^TX)^{-1}X^TX(X^TX)^{-1}C^T = C(X^TX)^{-1}C^T \tag{7.20}$$

Thus, the quadratic form becomes:

$$y^T P_{\text{Col}(Z)} y = (C\hat{\beta})^T [C(X^TX)^{-1}C^T]^{-1} (C\hat{\beta}) \tag{7.21}$$

This confirms that the Wald-type statistic derived for the General Linear Hypothesis is algebraically identical to the difference in sums of squares between the implied full and reduced models. □

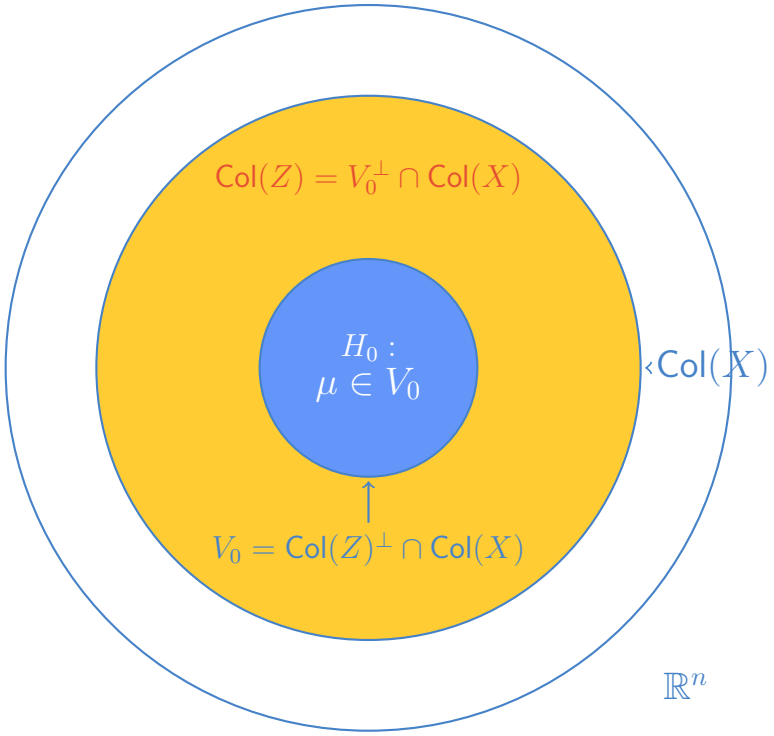


Figure 7.2: Geometric Interpretation of the General Linear Hypothesis

7.2.3 F-Test for Non-Zero General Linear Hypothesis

Theorem 7.6 (F-Test for Non-Zero General Linear Hypothesis)**.** *Consider the general linear model $y = X\beta + \epsilon$ with $\epsilon \sim N(0, \sigma^2 I)$. To test the non-homogeneous hypothesis $H_0 : C\beta = t$ where t is a known vector and C has rank q, the test statistic is:*

$$F = \frac{(C\hat{\beta} - t)^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\beta} - t)/q}{SSE_{FM}/(n - k - 1)} \tag{7.22}$$

Under H_0, this statistic follows an F distribution with q and $n - k - 1$ degrees of freedom:

$$F \sim F(q, n - k - 1) \tag{7.23}$$

Proof. **Distributional Proof (Alternative)**
This proof derives the F-statistic directly from the multivariate normal distribution of the coefficients, without explicitly invoking the geometry of projections.
Step 1: Distribution of the Linear Combination Recall that $\hat{\beta} \sim N_{k+1}(\beta, \sigma^2 (X^T X)^{-1})$. Under the null hypothesis $H_0 : C\beta = t$, we define the random vector $\theta = C\hat{\beta} - t$. Its expected value is:

$$E[\theta] = CE[\hat{\beta}] - t = C\beta - t = 0 \tag{7.24}$$

Its variance-covariance matrix is:

$$\text{Var}(\theta) = C\text{Var}(\hat{\beta})C^T = C[\sigma^2(X^TX)^{-1}]C^T = \sigma^2[C(X^TX)^{-1}C^T] \tag{7.25}$$

Let $V = C(X^TX)^{-1}C^T$. Since C has full row rank q, V is positive definite. Thus:

$$\theta \sim N_q(0, \sigma^2 V) \tag{7.26}$$

Step 2: Formation of the Chi-Squared Variable A standard result in multivariate statistics states that if $x \sim N_q(0, \Sigma)$, then the quadratic form $x^T\Sigma^{-1}x \sim \chi^2_q$. Applying this to θ:

$$\theta^T(\sigma^2 V)^{-1}\theta = \frac{(C\hat{\beta} - t)^T[C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta} - t)}{\sigma^2} \sim \chi^2_q \tag{7.27}$$

Let $Q_H = (C\hat{\beta} - t)^T[C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta} - t)$. We have established that $Q_H/\sigma^2 \sim \chi^2_q$.

Step 3: Independence and the F-Ratio From the properties of Least Squares, $\hat{\beta}$ is independent of the residuals (and thus independent of SSE). Consequently, the numerator Q_H is independent of the denominator SSE. We know that $\text{SSE}/\sigma^2 \sim \chi^2_{n-k-1}$.

The F-statistic is constructed as the ratio of two independent Chi-squared variables divided by their respective degrees of freedom:

$$F = \frac{(Q_H/\sigma^2)/q}{(\text{SSE}/\sigma^2)/(n-k-1)} = \frac{Q_H/q}{\text{SSE}/(n-k-1)} \tag{7.28}$$

The σ^2 terms cancel out, leaving the standard General Linear Test statistic. □

7.2.4 Numerical Examples

Example 7.1 (Chemical Reaction Data)**.** Consider an experiment designed to optimize the yield of a chemical reaction. The explanatory variables are:

- x_1: Temperature ($^\circ C$)
- x_2: Concentration of a reagent (%)
- x_3: Time of reaction (hours)

The response variable y_1 is the percent of unchanged starting material. The data is provided below :

y_1	x_1	x_2	x_3
41.5	162	23	3
33.8	162	23	8
27.7	162	30	5
21.7	162	30	8
19.9	172	25	5
15.0	172	25	8
12.2	172	30	5

4.3	172	30	8
19.3	167	27.5	6.5
6.4	177	27.5	6.5
37.6	157	27.5	6.5
18.0	167	32.5	6.5
26.3	167	22.5	6.5
9.9	167	27.5	9.5
25.0	167	27.5	3.5
14.1	177	20	6.5
15.2	177	20	6.5
15.9	160	34	7.5
19.6	160	34	7.5

We fit the full model:

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{7.29}$$

Suppose we wish to test the hypothesis that the effect of temperature and concentration are equal (scaled by 2) and related to time as follows: $H_0 : 2\beta_1 = 2\beta_2 = \beta_3$. This can be broken down into two independent linear constraints:

1. $2\beta_1 - 2\beta_2 = 0$
2. $2\beta_2 - \beta_3 = 0$

We formulate this as a general linear hypothesis $C\beta = 0$, where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ and C is a 2×4 matrix :

$$C = \begin{pmatrix} 0 & 2 & -2 & 0 \\ 0 & 0 & 2 & -1 \end{pmatrix} \tag{7.30}$$

Using the data, the estimated coefficients $\hat{\beta}$ yield $C\hat{\beta} = \begin{pmatrix} -0.1214 \\ -0.6118 \end{pmatrix}$. The variance term involving the design matrix is calculated as:

$$C(X^T X)^{-1} C^T = \begin{pmatrix} 0.003366 & -0.006943 \\ -0.006943 & 0.044974 \end{pmatrix} \tag{7.31}$$

The test statistic is computed using the quadratic form:

$$F = \frac{(C\hat{\beta})^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\beta})/2}{s^2} = \frac{28.62301/2}{5.3449} = 2.6776 \tag{7.32}$$

where $s^2 = \text{MSE}_{\text{FM}} = 5.3449$. This statistic follows an $F(2, 15)$ distribution. The corresponding p-value is 0.101, suggesting that we fail to reject the null hypothesis at the $\alpha = 0.05$ level.

Method 1: General Linear Hypothesis (Manual Matrix & car Package)

This approach tests the hypothesis $H_0 : C\beta = 0$. We calculate the Sum of Squares for the Hypothesis (SSH) using the general linear hypothesis formula and then structure the results into a standard ANOVA table format.

Method 1: General Linear Hypothesis (Manual Matrix & car Package)

This approach tests the hypothesis $H_0 : C\beta = 0$. We calculate the Sum of Squares for the Hypothesis (SSH) using the general linear hypothesis formula and then structure the results into a standard ANOVA table format.

```r
library(knitr)
library(car)

# Load Data
chemical_data <- data.frame(
  y1 = c(41.5, 33.8, 27.7, 21.7, 19.9, 15.0, 12.2, 4.3, 19.3, 6.4,
         37.6, 18.0, 26.3, 9.9, 25.0, 14.1, 15.2, 15.9, 19.6),
  x1 = c(162, 162, 162, 162, 172, 172, 172, 172, 167, 177,
         157, 167, 167, 167, 167, 177, 177, 160, 160),
  x2 = c(23, 23, 30, 30, 25, 25, 30, 30, 27.5, 27.5,
         27.5, 32.5, 22.5, 27.5, 27.5, 20, 20, 34, 34),
  x3 = c(3, 8, 5, 8, 5, 8, 5, 8, 6.5, 6.5,
         6.5, 6.5, 6.5, 9.5, 3.5, 6.5, 6.5, 7.5, 7.5)
)

# Fit Full Model
full_model <- lm(y1 ~ x1 + x2 + x3, data = chemical_data)
beta_hat <- coef(full_model)
XtX_inv  <- summary(full_model)$cov.unscaled
SSE_FM   <- sum(residuals(full_model)^2)
n <- nrow(chemical_data)
k <- 3

# --- Manual Matrix Implementation ---
C <- matrix(c(0, 2, -2,  0,
              0, 0,  2, -1), nrow = 2, byrow = TRUE)
q <- nrow(C)

Cb <- C %*% beta_hat
middle_term <- solve(C %*% XtX_inv %*% t(C))
SSH <- as.numeric(t(Cb) %*% middle_term %*% Cb)

# Constructing ANOVA components
df_error <- n - k - 1
MSE <- SSE_FM / df_error
F_stat <- (SSH / q) / MSE
p_val <- pf(F_stat, q, df_error, lower.tail = FALSE)

# Display as an ANOVA table
anova_manual <- data.frame(
  Df = c(q, df_error),
  `Sum Sq` = c(SSH, SSE_FM),
  `Mean Sq` = c(SSH/q, MSE),
  `F value` = c(F_stat, NA),
  `Pr(>F)` = c(p_val, NA),
  row.names = c("Hypothesis", "Residuals")
)

print(as.table(as.matrix(anova_manual)))
```

```
                   Df     Sum.Sq    Mean.Sq    F.value     Pr..F.
Hypothesis  2.0000000 28.6232434 14.3116217  2.6776207  0.1013007
Residuals  15.0000000 80.1735397  5.3449026
```

```r
# --- Verification with car package ---
hypotheses <- c("2*x1 - 2*x2 = 0", "2*x2 - x3 = 0")
print(linearHypothesis(full_model, hypotheses))
```

```
Linear hypothesis test:
2 x1 - 2 x2 = 0
2 x2 - x3 = 0

Model 1: restricted model
Model 2: y1 ~ x1 + x2 + x3

  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     17 108.797
2     15  80.174  2    28.623 2.6776 0.1013
```

Method 2: Nested Models Approach

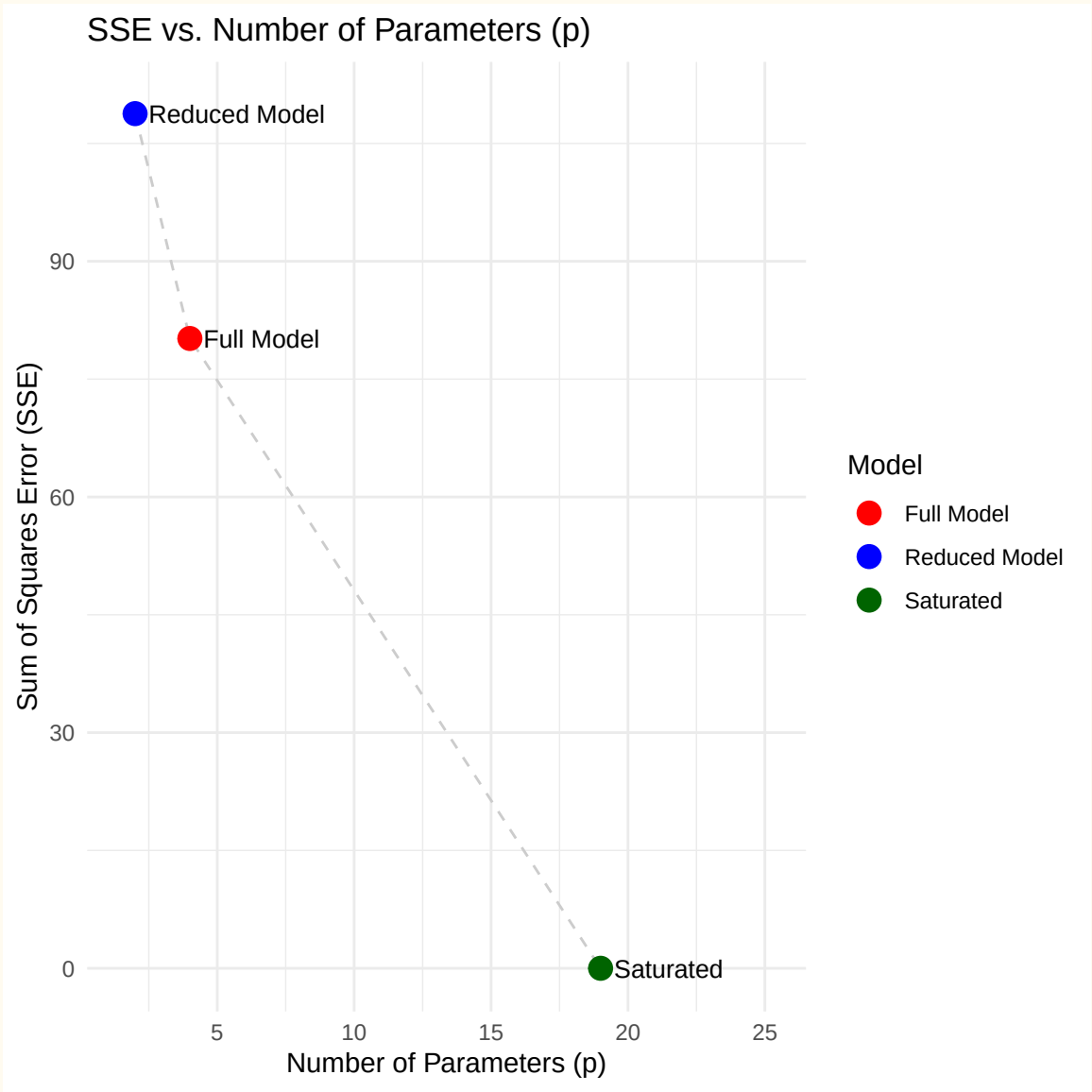
The null hypothesis implies relationships between the parameters that allow us to rewrite the model. If $2\beta_1 = \beta_3$ and $2\beta_2 = \beta_3$, then $\beta_1 = 0.5\beta_3$ and $\beta_2 = 0.5\beta_3$. Substituting these into the full model:

$$y_1 = \beta_0 + 0.5\beta_3 x_1 + 0.5\beta_3 x_2 + \beta_3 x_3 + \epsilon \tag{7.33}$$

$$y_1 = \beta_0 + \beta_3(0.5x_1 + 0.5x_2 + x_3) + \epsilon \tag{7.34}$$

This reduced model has only two parameters: an intercept β_0 and a slope β_3 for the constructed variable $z = 0.5x_1 + 0.5x_2 + x_3$. We could then calculate F using the difference in SSE between the full model and this reduced model. We analyze the change in error when moving from the Full Model to a Reduced Model. The plot illustrates SSE against the number of parameters p. Labels are positioned to the right of each data point for clarity.

```r
library(ggplot2)

# Create the reduced variable z based on constraints
chemical_data$z <- 0.5 * chemical_data$x1 + 0.5 * chemical_data$x2 + chemical_data$x3

# Fit Models
reduced_model <- lm(y1 ~ z, data = chemical_data)

# Data for Plotting
n_obs <- nrow(chemical_data)
p_full <- length(coef(full_model))
p_red  <- length(coef(reduced_model))
p_sat  <- n_obs

viz_data <- data.frame(
  p = c(p_sat, p_full, p_red),
  sse = c(0, sum(residuals(full_model)^2), sum(residuals(reduced_model)^2)),
  Model = c("Saturated", "Full Model", "Reduced Model")
)

ggplot(viz_data, aes(x = p, y = sse)) +
  geom_line(color = "gray80", linetype = "dashed") +
  geom_point(aes(color = Model), size = 4) +
  # Use nudge_x and hjust=0 to place text on the right
  geom_text(aes(label = Model), hjust = 0, nudge_x = 0.5, size = 3.5) +
  labs(
    title = "SSE vs. Number of Parameters (p)",
    x = "Number of Parameters (p)",
    y = "Sum of Squares Error (SSE)"
  ) +
  theme_minimal() +
  scale_color_manual(values = c("Full Model" = "red",
                                "Reduced Model" = "blue",
                                "Saturated" = "darkgreen")) +
  # Expand x-axis slightly to make room for text on the right
  scale_x_continuous(expand = expansion(mult = c(0.1, 0.4))) +
  ylim(0, max(viz_data$sse) * 1.01)
```




```
# Display raw text output from anova for nested models
anova(reduced_model, full_model)
```

```
Analysis of Variance Table

Model 1: y1 ~ z
Model 2: y1 ~ x1 + x2 + x3
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     17 108.797
2     15  80.174  2    28.623 2.6776 0.1013
```

7.3 Speficic Tests for Linear Combinations of β

Test for $a^T\beta = 0$

To test a linear combination $a^T\beta = 0$, the F-statistic is:

$$F = \frac{(a^T\hat{\beta})^2}{s^2 a^T(X^TX)^{-1}a} \sim F(1, n-k-1) \tag{7.35}$$

This is equivalent to the t-test:

$$t = \frac{a^T\hat{\beta}}{s\sqrt{a^T(X^TX)^{-1}a}} \sim t(n-k-1) \tag{7.36}$$

Test for $\beta_j = 0$

For a single coefficient β_j, we set a to extract the j-th element. The test statistic becomes:

$$t = \frac{\hat{\beta}_j}{s\sqrt{g_{jj}}} = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)} \tag{7.37}$$

where g_{jj} is the j-th diagonal element of $(X^TX)^{-1}$.

Confidence Region for β

A $100(1-\alpha)\%$ confidence region for the vector β is the set of all vectors satisfying:

$$(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta) \leq (k+1)s^2 F_{1-\alpha}(k+1, n-k-1) \tag{7.38}$$

This region forms an ellipsoid centered at $\hat{\beta}$.

Confidence Interval for $a^T\beta$ and $E(y_0)$

A $100(1-\alpha)\%$ confidence interval for a linear combination $a^T\beta$ is given by:

$$a^T\hat{\beta} \pm t_{1-\alpha/2}(n-k-1)s\sqrt{a^T(X^TX)^{-1}a} \tag{7.39}$$

To estimate the mean response $E(y_0) = x_0^T\beta$ for a given x_0, we set $a = x_0$:

$$x_0^T\hat{\beta} \pm t_{1-\alpha/2}(n-k-1)s\sqrt{x_0^T(X^TX)^{-1}x_0} \tag{7.40}$$

Prediction Interval for y_0

When predicting a new random observation $y_0 = x_0^T\beta + e_0$, we must account for both the variance of the estimator and the variance of the new error term.

$$\text{Var}(y_0 - \hat{y}_0) = \text{Var}(e_0) + \text{Var}(x_0^T\hat{\beta}) = \sigma^2(1 + x_0^T(X^TX)^{-1}x_0) \tag{7.41}$$

The $100(1-\alpha)\%$ prediction interval is:

$$\hat{y}_0 \pm t_{1-\alpha/2}(n-k-1)s\sqrt{1 + x_0^T(X^TX)^{-1}x_0} \tag{7.42}$$

7.3.1 Numerical Examples

Example 7.2. Specific Tests and Intervals Implementation
This example demonstrates how to:

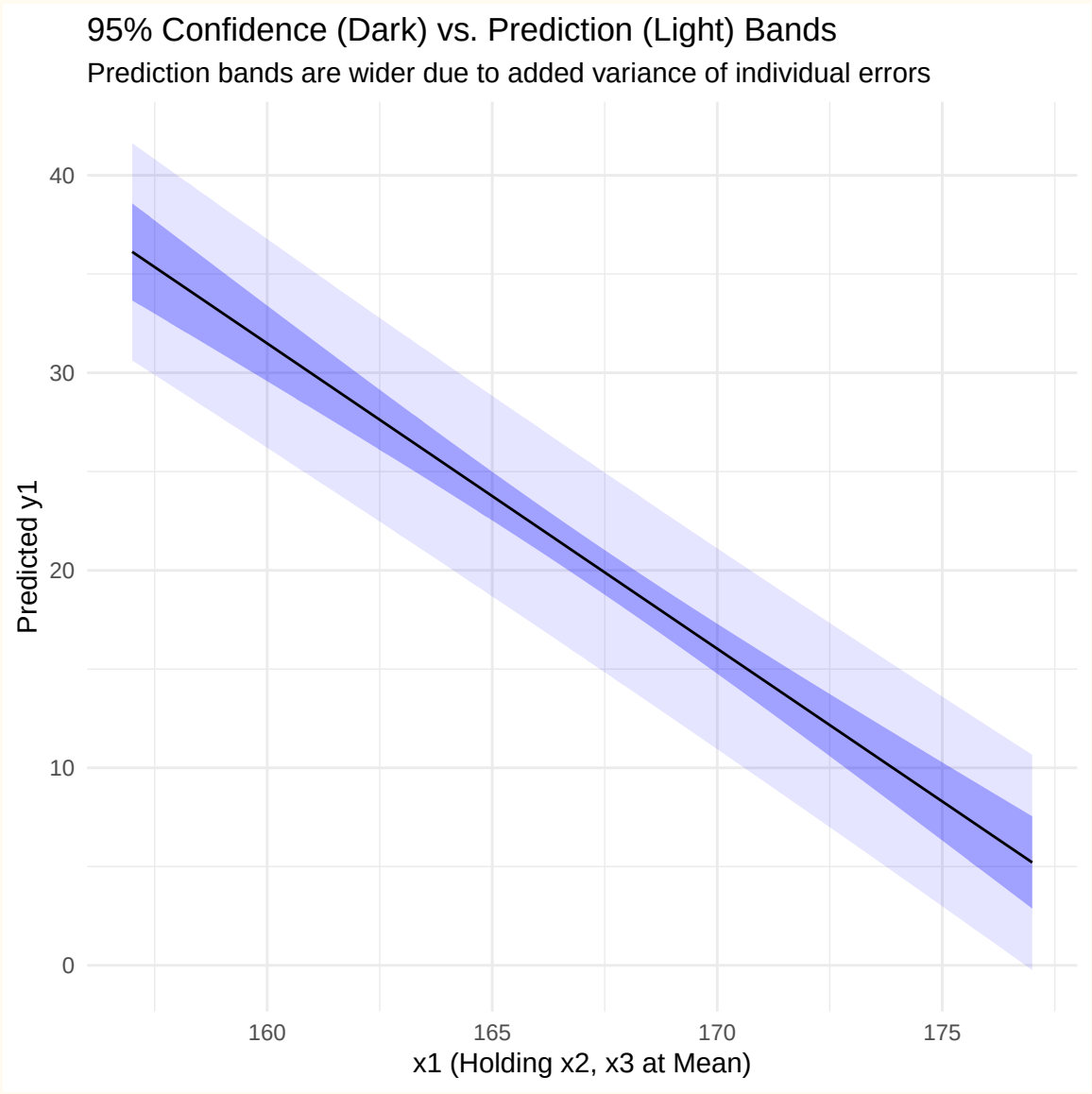
1. Test a specific linear combination of coefficients (Manually).
2. Calculate Confidence and Prediction intervals using R's `predict()` function.
3. Verify those intervals using manual matrix algebra.

```r
library(knitr)
library(ggplot2)

# --- 0. Setup: Load Data and Fit Model ---
chemical_data <- data.frame(
  y1 = c(41.5, 33.8, 27.7, 21.7, 19.9, 15.0, 12.2, 4.3, 19.3, 6.4,
         37.6, 18.0, 26.3, 9.9, 25.0, 14.1, 15.2, 15.9, 19.6),
  x1 = c(162, 162, 162, 162, 172, 172, 172, 172, 167, 177,
         157, 167, 167, 167, 167, 177, 177, 160, 160),
  x2 = c(23, 23, 30, 30, 25, 25, 30, 30, 27.5, 27.5,
         27.5, 32.5, 22.5, 27.5, 27.5, 20, 20, 34, 34),
  x3 = c(3, 8, 5, 8, 5, 8, 5, 8, 6.5, 6.5,
         6.5, 6.5, 6.5, 9.5, 3.5, 6.5, 6.5, 7.5, 7.5)
)

full_model <- lm(y1 ~ x1 + x2 + x3, data = chemical_data)
beta_hat <- coef(full_model)
s_sq <- summary(full_model)$sigma^2
XtX_inv <- summary(full_model)$cov.unscaled

# --- 1. Test for Linear Combination: H0: beta1 + beta2 = 50 ---

# Define 'a' vector (Intercept, x1, x2, x3)
a <- c(0, 1, 1, 0)

# Manual t-statistic calculation
est <- sum(a * beta_hat)
se_est <- sqrt(s_sq * t(a) %*% XtX_inv %*% a)
t_stat <- (est - 50) / se_est
p_val_t <- 2 * pt(abs(t_stat), df = full_model$df.residual, lower.tail = FALSE)

# --- 2. Confidence and Prediction Intervals using predict() ---

# Point of interest: x0 = (165, 25, 5)
new_pt <- data.frame(x1=165, x2=25, x3=5)

# Confidence Interval (for the Mean Response)
ci_r <- predict(full_model, newdata = new_pt, interval = "confidence", level = 0.95)

# Prediction Interval (for a New Observation)
pi_r <- predict(full_model, newdata = new_pt, interval = "prediction", level = 0.95)

# --- 3. Manual Verification of Intervals ---
x0_vec <- matrix(c(1, 165, 25, 5), ncol = 1) # Note the 1 for intercept
y_hat  <- as.numeric(t(x0_vec) %*% beta_hat)
t_crit <- qt(0.975, full_model$df.residual)

# Standard Errors
se_mean <- sqrt(as.numeric(s_sq * (t(x0_vec) %*% XtX_inv %*% x0_vec)))
se_pred <- sqrt(as.numeric(s_sq * (1 + t(x0_vec) %*% XtX_inv %*% x0_vec)))

# Manual Intervals
ci_man <- c(y_hat - t_crit * se_mean, y_hat + t_crit * se_mean)
```

Table 7.1: Comparison of R functions vs Manual Matrix Algebra

Method	Type	Lower	Upper
R predict()	Confidence (Mean)	28.4576	31.9957
Manual Calc	Confidence (Mean)	28.4576	31.9957
R predict()	Prediction (New)	24.9910	35.4623
Manual Calc	Prediction (New)	24.9910	35.4623

```r
# --- 5. Visualization ---

# Varying x1 while holding x2 and x3 constant at their means
x_range <- data.frame(
  x1 = seq(min(chemical_data$x1), max(chemical_data$x1), length.out = 100),
  x2 = mean(chemical_data$x2),
  x3 = mean(chemical_data$x3)
)

c_bands <- predict(full_model, newdata = x_range, interval = "confidence")
p_bands <- predict(full_model, newdata = x_range, interval = "prediction")

plot_df <- data.frame(x_range, c_bands)
plot_df$p_lwr <- p_bands[,2]
plot_df$p_upr <- p_bands[,3]

ggplot(plot_df, aes(x = x1, y = fit)) +
  geom_ribbon(aes(ymin = p_lwr, ymax = p_upr), fill = "blue", alpha = 0.1) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.3) +
  geom_line(color = "black") +
  labs(title = "95% Confidence (Dark) vs. Prediction (Light) Bands",
       subtitle = "Prediction bands are wider due to added variance of individual errors",
       x = "x1 (Holding x2, x3 at Mean)", y = "Predicted y1") +
  theme_minimal()
```


95% Confidence (Dark) vs. Prediction (Light) Bands
Prediction bands are wider due to added variance of individual errors
Predicted y1
40
30
20
10
0
160
165
170
175
x1 (Holding x2, x3 at Mean)

8 Generalized Inverses

8.1 Motivation

Consider the linear system $X\beta = y$. In $\mathbb{R}^2$, if $X = [x_1, x_2]$ is invertible, the solution is unique: $\beta = X^{-1}y$. This satisfies $X(X^{-1}y) = y$.However, if X is not square or not invertible (e.g., X is 2×3), $X\beta = y$ does not have a unique solution. We seek a matrix G such that $\beta = Gy$ provides a solution whenever $y \in C(X)$ (the column space of X). Substituting $\beta = Gy$ into the equation $X\beta = y$:

$$X(Gy) = y \quad \forall y \in C(X) \tag{8.1}$$

Since any $y \in C(X)$ can be written as Xw for some vector w:

$$XGXw = Xw \quad \forall w \tag{8.2}$$

This implies the defining condition:

$$XGX = X \tag{8.3}$$

8.2 Definition of Generalized Inverse

Definition 8.1 (Generalized Inverse)**.** Let X be an $n \times p$ matrix. A matrix X^- of size $p \times n$ is called a **generalized inverse** of X if it satisfies:

$$XX^-X = X \tag{8.4}$$

Example 8.1 (Examples of Generalized Inverse)**.**

- **Example 1: Diagonal Matrix** If $X = \text{diag}(\lambda_1, \lambda_2, 0, 0)$, we can write it in matrix form as:

$$X = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{8.5}$$

  A generalized inverse is obtained by inverting the non-zero elements:

$$X^- = \begin{pmatrix} \lambda_1^{-1} & 0 & 0 & 0 \\ 0 & \lambda_2^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{8.6}$$

- **Example 2: Row Vector** Let $X = (1, 2, 3)$. One possible generalized inverse is a column vector where the first element is the reciprocal of the first non-zero element of X (which is 1), and others are zero:

$$X^- = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \tag{8.7}$$

Verification:

$$XX^-X = (1, 2, 3) \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} (1, 2, 3) = (1) \cdot (1, 2, 3) = (1, 2, 3) = X \tag{8.8}$$

Other valid generalized inverses include $\begin{pmatrix} 0 \\ 1/2 \\ 0 \end{pmatrix}$ or $\begin{pmatrix} 0 \\ 0 \\ 1/3 \end{pmatrix}$.

- **Example 3: Rank Deficient Matrix** Let $A = \begin{pmatrix} 2 & 2 & 3 \\ 1 & 0 & 1 \\ 3 & 2 & 4 \end{pmatrix}$. Note that Row 3 = Row 1 + Row 2, so Rank$(A) = 2$.

Solution: A generalized inverse can be found by locating a non-singular 2×2 submatrix, inverting it, and padding the rest with zeros. Let's take the top-left minor $M = \begin{pmatrix} 2 & 2 \\ 1 & 0 \end{pmatrix}$. The inverse is $M^{-1} = \frac{1}{-2} \begin{pmatrix} 0 & -2 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0.5 & -1 \end{pmatrix}$.

Placing this in the corresponding position in A^- and setting the rest to 0:

$$A^- = \begin{pmatrix} 0 & 1 & 0 \\ 0.5 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{8.9}$$

Verification ($AA^-A = A$): First, compute AA^-:

$$AA^- = \begin{pmatrix} 2 & 2 & 3 \\ 1 & 0 & 1 \\ 3 & 2 & 4 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 0.5 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \tag{8.10}$$

Then multiply by A:

$$(AA^-)A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 2 & 2 & 3 \\ 1 & 0 & 1 \\ 3 & 2 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 2 & 3 \\ 1 & 0 & 1 \\ 3 & 2 & 4 \end{pmatrix} = A \tag{8.11}$$

8.3 A Procedure to Find a Generalized Inverse

If we can partition X (possibly after permuting rows/columns) such that R_{11} is a non-singular rank r submatrix:

$$X = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \tag{8.12}$$

Then a generalized inverse is:

$$X^{-} = \begin{pmatrix} R_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \tag{8.13}$$

Verification:

$$\begin{aligned} XX^{-}X &= \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \begin{pmatrix} R_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \\ &= \begin{pmatrix} I_r & 0 \\ R_{21}R_{11}^{-1} & 0 \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \\ &= \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{21}R_{11}^{-1}R_{12} \end{pmatrix} \end{aligned} \tag{8.14}$$

Note that since $\text{rank}(X) = \text{rank}(R_{11})$, the rows of $[R_{21}, R_{22}]$ are linear combinations of $[R_{11}, R_{12}]$, implying $R_{22} = R_{21}R_{11}^{-1}R_{12}$. Thus, $XX^{-}X = X$.

An Algorithm for Finding a Generalized Inverse

A systematic procedure to find a generalized inverse A^{-} for any matrix A:

1. Find any non-singular $r \times r$ submatrix C, where r is the rank of A. It is not necessary for the elements of C to occupy adjacent rows and columns in A.
2. Find C^{-1} and $(C^{-1})'$.
3. Replace the elements of C in A with the elements of $(C^{-1})'$.
4. Replace all other elements in A with zeros.
5. Transpose the resulting matrix.

Matrix Visual Representation

$$\underset{\text{Original } A}{\begin{pmatrix} \times & \otimes & \times & \otimes \\ \times & \otimes & \times & \otimes \\ \times & \times & \times & \times \end{pmatrix}} \xrightarrow[\text{with } (C^{-1})']{\text{Replace } C} \underset{\text{Intermediate}}{\begin{pmatrix} \times & \triangle & \times & \triangle \\ \times & \triangle & \times & \triangle \\ \times & \times & \times & \times \end{pmatrix}} \xrightarrow[\text{Result}]{\text{Transpose}} \underset{\text{Final } A^{-}}{\begin{pmatrix} \times & \times & \times \\ \square & \square & \times \\ \times & \times & \times \\ \square & \square & \times \end{pmatrix}} \tag{8.15}$$

Legend:

- $\otimes$: Elements of submatrix C
- $\triangle$: Elements of $(C^{-1})'$
- $\square$: Elements of C^{-1} (after transposition)
- $\times$: Other elements (replaced by 0 in the final calculation)

8.4 Moore-Penrose Inverse

The Moore-Penrose inverse (denoted X^+) is a unique generalized inverse defined via Singular Value Decomposition (SVD).

If X has SVD:

$$X = U \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} V' \tag{8.16}$$

Then the Moore-Penrose inverse is:

$$X^+ = V \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U' \tag{8.17}$$

where $\Lambda_r = \text{diag}(\lambda_1, \ldots, \lambda_r)$ contains the singular values. Unlike standard generalized inverses, X^+ is unique.

Verification:

We verify that X^+ satisfies the condition $XX^+X = X$.

1. **Substitute definitions:**

$$XX^+X = \left[U \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} V'\right] \left[V \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U'\right] \left[U \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} V'\right] \tag{8.18}$$

2. **Apply orthogonality:** Recall that $V'V = I$ and $U'U = I$.

$$= U \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} \underbrace{(V'V)}_{I} \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} \underbrace{(U'U)}_{I} \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} V' \tag{8.19}$$

3. **Multiply diagonal matrices:**

$$= U \left[\begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix}\right] V' \tag{8.20}$$

Since $\Lambda_r \Lambda_r^{-1} \Lambda_r = I \cdot \Lambda_r = \Lambda_r$:

$$= U \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} V' = X \tag{8.21}$$

8.5 Solving Linear Systems with Generalized Inverse

We apply generalized inverses to solve systems of linear equations $X\beta = c$ where X is $n \times p$.

Definition 8.2 (Consistency and Solution)**.** The system $X\beta = c$ is consistent if and only if $c \in \text{Col}(X)$ (the column space of X). If consistent, $\beta = X^- c$ is a solution.

Proof: If the system is consistent, there exists some b such that $Xb = c$. Using the definition $XX^-X = X$:

$$X(X^-c) = X(X^-Xb) = (XX^-X)b = Xb = c \tag{8.22}$$

Thus, X^-c is a solution. Note that the solution is not unique if X is not full rank.

Example 8.2 (Examples of Solutions of Linear System with Generalized Inverse)**.**

- **Example 1: Underdetermined System**

  Let $X = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$ and we want to solve $X\beta = 4$.

  Solution 1: Using the generalized inverse $X^- = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$:

$$\beta = X^- \cdot 4 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} 4 = \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix} \tag{8.23}$$

  Verification:

$$X\beta = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix} = 1(4) + 2(0) + 3(0) = 4 \quad \checkmark \tag{8.24}$$

  Solution 2: Using another generalized inverse $X^- = \begin{pmatrix} 0 \\ 0 \\ 1/3 \end{pmatrix}$:

$$\beta = X^- \cdot 4 = \begin{pmatrix} 0 \\ 0 \\ 1/3 \end{pmatrix} 4 = \begin{pmatrix} 0 \\ 0 \\ 4/3 \end{pmatrix} \tag{8.25}$$

  Verification:

$$X\beta = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 4/3 \end{pmatrix} = 0 + 0 + 3(4/3) = 4 \quad \checkmark \tag{8.26}$$

- **Example 2: Overdetermined System**

  Let $X = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$. Solve $X\beta = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix} = c$. Here $c = 2X$, so the system is consistent. Since X is a column vector, β is a scalar.

  Solution: Using the generalized inverse $X^- = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$:

$$\beta = X^- c = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix} = 1(2) + 0(4) + 0(6) = 2 \tag{8.27}$$

  Verification:

$$X\beta = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} (2) = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix} = c \quad \checkmark \tag{8.28}$$

8.6 Least Squares for Non-full-rank X with Generalized Inverse

8.6.1 Projection Matrix with Generalized Inverse of $X'X$

For the normal equations $(X'X)\beta = X'y$, a solution is given by:

$$\hat{\beta} = (X'X)^{-}X'y \tag{8.29}$$

The fitted values are

$$\hat{y} = X\hat{\beta} = X(X'X)^{-}X'y. \tag{8.30}$$

This $\hat{y}$ represents the unique orthogonal projection of y onto Col(X).

8.6.2 Invariance and Uniqueness of "the" Projection Matrix

Theorem 8.1 (Transpose Property of Generalized Inverses)**.** *$(X^{-})'$ is a version of $(X')^{-}$. That is, $(X^{-})'$ is a generalized inverse of X'.*

Proof. By definition, a generalized inverse X^{-} satisfies the property:

$$XX^{-}X = X \tag{8.31}$$

To verify that $(X^{-})'$ is a generalized inverse of X', we need to show that it satisfies the condition $AGA = A$ where $A = X'$ and $G = (X^{-})'$.

1. Start with the fundamental definition:
$$XX^{-}X = X \tag{8.32}$$
2. Take the transpose of both sides of the equation:
$$(XX^{-}X)' = X' \tag{8.33}$$
3. Apply the reverse order law for transposes, $(ABC)' = C'B'A'$:
$$X'(X^{-})'X' = X' \tag{8.34}$$

Since substituting $(X^{-})'$ into the generalized inverse equation for X' yields X', $(X^{-})'$ is a valid generalized inverse of X'. □

Lemma 8.1 (Invariance of Generalized Least Squares)**.** *For any version of the generalized inverse $(X'X)^{-}$, the matrix $X'(X'X)^{-}X'$ is invariant and equals X'.*

$$X'X(X'X)^{-}X' = X' \tag{8.35}$$

Proof (using Projection): Let $P = X(X'X)^{-}X'$. This is the projection matrix onto Col(X). By definition of projection, $Px = x$ for any $x \in$ Col(X). Since columns of X are in Col(X), $PX = X$. Taking the transpose:

$(PX)' = X' \implies X'P' = X'$. Since projection matrices are symmetric ($P = P'$), $X'P = X'$. Substituting P: $X'X(X'X)^{-}X' = X'$.

Proof (Direct Matrix Manipulation): Decompose $y = X\beta + e$ where $e \perp \text{Col}(X)$ (i.e., $X'e = 0$).

$$\begin{aligned} X'X(X'X)^{-}X'y &= X'X(X'X)^{-}X'(X\beta + e) \\ &= X'X(X'X)^{-}X'X\beta + X'X(X'X)^{-}X'e \end{aligned} \tag{8.36}$$

Using the property $AA^{-}A = A$ (where $A = X'X$), the first term becomes $X'X\beta$. The second term is 0 because $X'e = 0$. Thus, the expression simplifies to $X'X\beta = X'(X\beta) = X'\hat{y}_{\text{proj}}$. This implies the operator acts as X'.

Theorem 8.2 (Properties of Projection Matrix P)**.** *Let $P = X(X'X)^{-}X'$. This matrix has the following properties:*

1. ***Symmetry:*** $P = P'$.
2. ***Idempotence:*** $P^2 = P$.

$$P^2 = X(X'X)^{-}X'X(X'X)^{-}X' = X(X'X)^{-}(X'X(X'X)^{-}X') \tag{8.37}$$

Using the identity from Lemma 8.1 ($X'X(X'X)^{-}X' = X'$), this simplifies to:

$$X(X'X)^{-}X' = P \tag{8.38}$$

3. ***Uniqueness:*** *P is unique and invariant to the choice of the generalized inverse $(X'X)^{-}$.*

Proof. **Proof of Uniqueness:**
Let A and B be two different generalized inverses of $X'X$. Define $P_A = XAX'$ and $P_B = XBX'$. From Lemma 8.1, we know that $X'P_A = X'$ and $X'P_B = X'$.
Subtracting these two equations:

$$X'(P_A - P_B) = 0 \tag{8.39}$$

Taking the transpose, we get $(P_A - P_B)X = 0$. This implies that the columns of the difference matrix $D = P_A - P_B$ are orthogonal to the columns of X (i.e., $D \perp \text{Col}(X)$).
However, by definition, the columns of P_A and P_B (and thus D) are linear combinations of the columns of X (i.e., $D \in \text{Col}(X)$).
The only matrix that lies *in* the column space of X but is also *orthogonal* to the column space of X is the zero matrix. Therefore:

$$P_A - P_B = 0 \implies P_A = P_B \tag{8.40}$$

□

Example: Projection with Linearly Dependent Columns and Generalized Inverses

In this example, we project a vector $y \in \mathbb{R}^4$ onto the column space of a design matrix $X = [x_1, x_2]$. We examine the specific case where the predictors are perfectly dependent such that $x_2 = 2x_1$.

We will: 1. Construct the design matrix X and compute $X^T X$. 2. Find two different generalized inverses of $X^T X$. 3. Show that the projection matrix P is invariant to the choice of generalized inverse. 4. Demonstrate that the projection of y onto $L(x_1, x_2)$ is exactly the same as projecting onto $L(x_1)$ individually.

1. Define the Vectors

Let's define our vectors in $\mathbb{R}^4$. We set x_2 to be exactly twice x_1 to enforce linear dependence.

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 1 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 2 \\ 4 \\ 2 \\ 2 \end{bmatrix}, \quad y = \begin{bmatrix} 4 \\ 2 \\ 1 \\ 3 \end{bmatrix} \tag{8.41}$$

2. Design Matrix and $X^T X$

The design matrix $X = [x_1, x_2]$ is:

$$X = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 1 & 2 \\ 1 & 2 \end{bmatrix} \tag{8.42}$$

To find the projection, we first compute the matrix $X^T X$:

$$X^T X = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 2 & 4 & 2 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 1 & 2 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 7 & 14 \\ 14 & 28 \end{bmatrix} \tag{8.43}$$

The determinant of this matrix is $(7)(28) - (14)(14) = 196 - 196 = 0$. Because the determinant is 0, the matrix $X^T X$ is singular and cannot be inverted normally. We must use a generalized inverse.

3. Compute Two Generalized Inverses

A generalized inverse G satisfies the condition $AGA = A$. Because $X^T X$ has a rank of 1, we can easily find G-inverses by taking the reciprocal of a single non-zero diagonal element and setting all other entries to zero.

Choice 1 (G_1): Invert the top-left element ($7 \to 1/7$).

$$G_1 = \begin{bmatrix} 1/7 & 0 \\ 0 & 0 \end{bmatrix} \tag{8.44}$$

Choice 2 (G_2): Invert the bottom-right element ($28 \to 1/28$).

$$G_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1/28 \end{bmatrix} \tag{8.45}$$

4. The Projection Matrix P is Invariant

The projection matrix is calculated as $P = XGX^T$. Let's compute it using G_1:

$$XG_1 = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 1 & 2 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1/7 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1/7 & 0 \\ 2/7 & 0 \\ 1/7 & 0 \\ 1/7 & 0 \end{bmatrix} \tag{8.46}$$

$$P_1 = (XG_1)X^T = \begin{bmatrix} 1/7 & 0 \\ 2/7 & 0 \\ 1/7 & 0 \\ 1/7 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 & 1 \\ 2 & 4 & 2 & 2 \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 1 & 2 & 1 & 1 \\ 2 & 4 & 2 & 2 \\ 1 & 2 & 1 & 1 \\ 1 & 2 & 1 & 1 \end{bmatrix} \tag{8.47}$$

If you perform the same calculation using G_2, you will yield the exact same 4×4 projection matrix. Thus, $P_1 = P_2 = P$.

The projection of y onto $L(x_1, x_2)$ is:

$$\hat{y} = Py = \frac{1}{7} \begin{bmatrix} 1 & 2 & 1 & 1 \\ 2 & 4 & 2 & 2 \\ 1 & 2 & 1 & 1 \\ 1 & 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \\ 1 \\ 3 \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 12 \\ 24 \\ 12 \\ 12 \end{bmatrix} \approx \begin{bmatrix} 1.714 \\ 3.429 \\ 1.714 \\ 1.714 \end{bmatrix} \tag{8.48}$$

5. Comparison to Projecting onto x_1 Individually

For a single vector x_1, the projection matrix is simply $P_{x_1} = x_1(x_1^T x_1)^{-1} x_1^T$.

First, we find the scalar denominator:

$$x_1^T x_1 = 1^2 + 2^2 + 1^2 + 1^2 = 7 \tag{8.49}$$

Next, we construct the matrix:

$$P_{x_1} = \frac{1}{7} \begin{bmatrix} 1 \\ 2 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 & 1 \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 1 & 2 & 1 & 1 \\ 2 & 4 & 2 & 2 \\ 1 & 2 & 1 & 1 \\ 1 & 2 & 1 & 1 \end{bmatrix} \tag{8.50}$$

Applying this to y:

$$P_{x_1} y = \frac{1}{7} \begin{bmatrix} 1 & 2 & 1 & 1 \\ 2 & 4 & 2 & 2 \\ 1 & 2 & 1 & 1 \\ 1 & 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \\ 1 \\ 3 \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 12 \\ 24 \\ 12 \\ 12 \end{bmatrix} \tag{8.51}$$

We have shown that $Py = P_{x_1} y = \hat{y}$. Because x_2 is merely a scaled version of x_1, it exists entirely within the 1-dimensional space already spanned by x_1. Incorporating x_2 into the model adds no new geometric dimensions, and the resulting projection is identical.

6. R Implementation Verification

```
[1] "Are the generalized inverses valid?"
```

```
G1_Valid G2_Valid
    TRUE     TRUE
```

```
[1] "Projection Results:"
```

```
  Target_y Proj_G1 Proj_G2 Proj_x1
1        4  1.7143  1.7143  1.7143
2        2  3.4286  3.4286  3.4286
3        1  1.7143  1.7143  1.7143
4        3  1.7143  1.7143  1.7143
```

8.7 The Left Inverse View: Recovering $\widehat{\beta}$ from $\widehat{y}$

While the geometric properties of the linear model are most naturally established via the unique orthogonal projection $\widehat{y}$, we require a functional mapping—a statistical "bridge"—to translate the distribution of these fitted values back into the parameter space of $\widehat{\beta}$. This bridge is provided by the generalized left inverse.

8.7.1 The Generalized Left Inverse

To recover the parameter estimates directly from the fitted values, we define the generalized left inverse, denoted as X_{left}^{-}, such that:

$$\widehat{\beta} = X_{\text{left}}^{-}\widehat{y} \tag{8.52}$$

A standard choice for this operator, derived from the normal equations, is:

$$X_{\text{left}}^{-} = (X'X)^{-}X' \tag{8.53}$$

When X is full-rank, the X_{left}^{-} is unique, which is given by

$$X_{\text{left}}^{-} = (X'X)^{-1}X' \tag{8.54}$$

8.7.2 Verification of the Inverse Property

To verify that X_{left}^{-} acts as a valid generalized inverse of X, it must satisfy the condition $XX_{\text{left}}^{-}X = X$. Substituting our definition:

$$X \underbrace{[(X'X)^{-}X']}_{X_{\text{left}}^{-}} X = X(X'X)^{-}(X'X) \tag{8.55}$$

Using the property of generalized inverses for symmetric matrices where $(X'X)(X'X)^{-}X' = X'$, the transpose of this identity gives $X(X'X)^{-}(X'X) = X$. Thus, the condition holds:

$$XX_{\text{left}}^{-}X = X \tag{8.56}$$

8.7.3 Recovering the Estimator

We can now demonstrate that applying this left inverse to the fitted values $\hat{y}$ yields the standard solution to the normal equations.

Substituting the projection formula $\hat{y} = X(X'X)^{-}X'y$:

$$\begin{aligned} X_{\text{left}}^{-}\hat{y} &= [(X'X)^{-}X'] \, [X(X'X)^{-}X'y] \\ &= (X'X)^{-} \underbrace{(X'X)(X'X)^{-}(X'X)}_{\text{Property } AA^{-}A=A}(X'X)^{-}X'y \end{aligned} \tag{8.57}$$

Simplifying using the generalized inverse property $A^{-}AA^{-} = A^{-}$ (where $A = X'X$):

$$\begin{aligned} X_{\text{left}}^{-}\hat{y} &= \underbrace{(X'X)^{-}(X'X)(X'X)^{-}}_{(X'X)^{-}} X'y \\ &= (X'X)^{-}X'y \end{aligned} \tag{8.58}$$

Thus, we recover the standard estimator used in the normal equations:

$$= (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} \tag{8.59}$$

8.8 Non-full-rank Least Squares with QR Decomposition

When X has rank $r < p$ (where X is $n \times p$), we can derive the least squares estimator using partitioned matrices.

Assume the first r columns of X are linearly independent. We can partition X as:

$$X = Q(R_1, R_2) \tag{8.60}$$

where Q is an $n \times r$ matrix with orthogonal columns ($Q'Q = I_r$), R_1 is an $r \times r$ non-singular matrix, and R_2 is $r \times (p - r)$.

The normal equations are:

$$X'X\beta = X'y \implies \begin{pmatrix} R_1' \\ R_2' \end{pmatrix} Q'Q(R_1, R_2)\beta = \begin{pmatrix} R_1' \\ R_2' \end{pmatrix} Q'y \tag{8.61}$$

Simplifying ($Q'Q = I_r$):

$$\begin{pmatrix} R_1'R_1 & R_1'R_2 \\ R_2'R_1 & R_2'R_2 \end{pmatrix} \beta = \begin{pmatrix} R_1'Q'y \\ R_2'Q'y \end{pmatrix} \tag{8.62}$$

8.8.1 Constructing a Solution by Solving Normal Equations

One specific generalized inverse of $X'X$ can be found by focusing on the non-singular block $R_1'R_1$:

$$(X'X)^- = \begin{pmatrix} (R_1'R_1)^{-1} & 0 \\ 0 & 0 \end{pmatrix} \tag{8.63}$$

Using this generalized inverse, the estimator $\hat{\beta}$ becomes:

$$\hat{\beta} = (X'X)^- X'y = \begin{pmatrix} (R_1'R_1)^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} R_1'Q'y \\ R_2'Q'y \end{pmatrix} \tag{8.64}$$

$$\hat{\beta} = \begin{pmatrix} (R_1'R_1)^{-1}R_1'Q'y \\ 0 \end{pmatrix} = \begin{pmatrix} R_1^{-1}Q'y \\ 0 \end{pmatrix} \tag{8.65}$$

The fitted values are:

$$\hat{y} = X\hat{\beta} = Q(R_1, R_2) \begin{pmatrix} R_1^{-1}Q'y \\ 0 \end{pmatrix} = QR_1R_1^{-1}Q'y = QQ'y \tag{8.66}$$

This confirms that $\hat{y}$ is the projection of y onto the column space of Q (which is the same as the column space of X).

8.8.2 Constructing a Solution by Solving Reparametrized β

We can view the model as:

$$y = Q(R_1, R_2)\beta + \epsilon = Qb + \epsilon \tag{8.67}$$

where $b = R_1\beta_1 + R_2\beta_2$.

Since the columns of Q are orthogonal, the least squares estimate for b is simply:

$$\hat{b} = (Q'Q)^{-1}Q'y = Q'y \tag{8.68}$$

To find β, we solve the underdetermined system:

$$R_1\beta_1 + R_2\beta_2 = \hat{b} = Q'y \tag{8.69}$$

Solution 1: Set $\beta_2 = 0$. Then:

$$R_1\beta_1 = Q'y \implies \hat{\beta}_1 = R_1^{-1}Q'y \tag{8.70}$$

This yields the same result as the generalized inverse method above: $\hat{\beta} = \begin{pmatrix} R_1^{-1}Q'y \\ 0 \end{pmatrix}$.

Solution 2: Using the generalized inverse of $R = (R_1, R_2)$:

$$R^{-} = \begin{pmatrix} R_1^{-1} \\ 0 \end{pmatrix} \tag{8.71}$$

$$\hat{\beta} = R^{-}Q'y = \begin{pmatrix} R_1^{-1}Q'y \\ 0 \end{pmatrix} \tag{8.72}$$

This demonstrates that finding a solution to the normal equations using $(X'X)^{-}$ is equivalent to solving the reparameterized system $b = R\beta$.

9 Estimation and Inference with Non-full-rank Models

9.1 Non-full-rank Models and Parameter Non-identifiability

9.1.1 One-way ANOVA Model

Consider the balanced one-way layout model for y_{ij} a response on the j^{th} unit in the i^{th} treatment group. Suppose that there are a treatments and n units in the i^{th} treatment group, with total sample size $N = an$. Let x_i be an indicator vector for treatment i, and j_N be a vector of ones.

1. The Cell-Means Model

The **cell-means model** represents the response strictly in terms of the mean of its respective treatment group.

Equation and Vector Form:

$$y_{ij} = \mu_i + e_{ij}, \quad i = 1, \dots, a, \quad j = 1, \dots, n \tag{9.1}$$

where the e_{ij} are i.i.d. $N(0, \sigma^2)$. In vector notation, this is:

$$y = \mu_1 x_1 + \mu_2 x_2 + \cdots + \mu_a x_a + e, \quad e \sim N(0, \sigma^2 I) \tag{9.2}$$

Matrix Formulation ($n = 2$ observations per level): Let the parameter vector be $\mu = (\mu_1, \mu_2, \dots, \mu_a)^T$. Arranging the response and error vectors by treatment group, the model $\mathbf{y} = X_1\mu + \mathbf{e}$ is written as:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{a1} \\ y_{a2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_a \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{a1} \\ e_{a2} \end{pmatrix} \tag{9.3}$$

Identifiability: The $2a \times a$ model matrix X_1 is of full column rank (rank a). Therefore, the parameters μ_i are uniquely identifiable.

---

2. The Effects Model

An alternative, but equivalent, linear model is the **effects model**, which decomposes the cell mean (μ_i) into a baseline value (μ) and a treatment-specific deviation (α_i). The explicit relationship between the parameters of the two models is:

$$\mu_i = \mu + \alpha_i \tag{9.4}$$

Equation and Vector Form: Substituting this relationship into the cell-means model yields:

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \dots, a, \quad j = 1, \dots, n \tag{9.5}$$

with the same assumptions on the errors. In vector notation, this is:

$$y = \mu j_N + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_a x_a + e, \quad e \sim N(0, \sigma^2 I) \tag{9.6}$$

Matrix Formulation ($n = 2$ observations per level): The parameter vector is expanded to include the baseline/grand mean μ, so $\beta = (\mu, \alpha_1, \alpha_2, \dots, \alpha_a)^T$. The model $\mathbf{y} = X_2\beta + \mathbf{e}$ is written as:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{a1} \\ y_{a2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{a1} \\ e_{a2} \end{pmatrix} \tag{9.7}$$

Non-identifiability and Constraints: The effects model has the same model matrix as the cell-means model, but with one extra column of ones in the first position (j_N). Notice that $\sum_i x_i = j_N$ and $\text{Col}(X_1) = \text{Col}(X_2)$. Therefore, the columns of the model matrix X_2 are linearly dependent. In this $2a \times (a+1)$ matrix, the first column is exactly equal to the sum of the remaining a columns.

Because X_2 is not full rank (it has rank a, but $a+1$ columns), the parameters are not uniquely identified. μ is in no sense the grand mean; it is just an arbitrary baseline value. However, subject to the constraint $\sum_i \alpha_i = 0$, the parameters gain specific interpretations:

1. **Grand Mean (μ)**: The grand mean response across all treatments.
2. **Treatment Effect (α_i)**: The deviation from the grand mean.

Adding the constraint $\sum_i \alpha_i = 0$ essentially reparameterizes the overparameterized (non-full rank) effects model to a just-parameterized (full rank) model equivalent to the cell means model.

Definition 9.1 (Equivalent Linear Models)**.** In general, two linear models $y = X_1\beta_1 + e_1$ and $y = X_2\beta_2 + e_2$ with the same assumptions on e_1 and e_2 are equivalent linear models if $\text{Col}(X_1) = \text{Col}(X_2)$.

Different Strategies for Non-Full Rank Models

When faced with a non-full rank model, we have three ways to proceed:

1. **Reparameterize to a full rank model** (e.g., cell-means model).
2. **Add constraints (side-conditions)** to the model parameters (e.g., $\sum \alpha_i = 0$).
3. **Analyze the model as a non-full rank model**, limiting inference to **estimable** functions.

9.2 Least Square Estimation of β and $\mu = X\beta$

Even if X is not of full rank, the least-squares criterion is still reasonable and leads to the normal equation:

$$X^T X\hat{\beta} = X^T y \tag{9.8}$$

Theorem 9.1 (Consistency of Normal Equations)**.** *For X an $n \times p$ matrix of rank $k < p \leq n$, the normal equations form a consistent system of equations because Col$(X^T X) =$ Col(X^T).*

Since the system is consistent, it has a (non-unique) solution given by:

$$\hat{\beta} = (X^T X)^- X^T y \tag{9.9}$$

where $(X^T X)^-$ is some (any) generalized inverse of $X^T X$.

Theorem 9.2 (Uniqueness of Projection)**.** *$\hat{\beta} = (X^T X)^- X^T y$ is a solution to $(X^T X)\beta = X^T y$. Then $\hat{y} = X\hat{\beta} = X(X^T X)^- X^T y$ is the projection of y onto Col(X).*
Since the projection onto a subspace is unique, $\hat{y}$ is unique regardless of the choice of generalized inverse.

9.2.1 Distribution of $\hat{\beta}$ and s^2

In the model $y = X\beta + e$, where $E(e) = 0$, Var$(e) = \sigma^2 I$, and X is $n \times p$ of rank $k \leq p \leq n$:

1. **Unbiasedness**: $E(s^2) = \sigma^2$.
2. **Uniqueness**: s^2 is invariant to the choice of $\hat{\beta}$ (i.e., to the choice of $(X^T X)^-$).

$$s^2 = \frac{SSE}{n-k} = \frac{||y - X\hat{\beta}||^2}{n-k} \tag{9.10}$$

where $k =$ rank(X).

Theorem 9.3 (Distributions in Non-Full Rank Models)**.** *In the normal-errors model $y \sim N(X\beta, \sigma^2 I)$:*

1. *For any given choice of $(X^T X)^-$:*

$$\hat{\beta} \sim N_p[(X^T X)^- X^T X\beta, \sigma^2 (X^T X)^- X^T X\{(X^T X)^-\}^T] \tag{9.11}$$

2. *$(n-k)s^2/\sigma^2 \sim \chi^2(n-k)$ and $SSE/\sigma^2 \sim \chi^2(n-k)$.*
3. *For any given choice of $(X^T X)^-$, $\hat{\beta}$ and s^2 are independent.*

9.3 Sum Squares and F-test

Suppose $y \sim N_n(X\beta, \sigma^2 I)$ where X is a matrix with rank $k+1$. Let $X = [X_1, X_2]$ where rank$(X) = k+1$ and assume we are testing a reduced model involving X_1.

Let:

- $\hat{y} = P_{\text{Col}(X)} y$ (Full Model Projection)
- $\hat{y}_0 = P_{\text{Col}(X_1)} y$ (Reduced Model Projection)

Theorem 9.4 (Partitioning of Sum of Squares)**.**

1. ***Full Model Residual Sum of Squares****:*

$$\frac{1}{\sigma^2}||y - \hat{y}||^2 = \frac{1}{\sigma^2} y^T (I - P_{Col(X)}) y \sim \chi^2(n-k-1) \tag{9.12}$$

2. ***Difference in Sum of Squares (Hypothesis)****:*

$$\frac{1}{\sigma^2}||\hat{y} - \hat{y}_0||^2 = \frac{1}{\sigma^2} y^T (P_{Col(X)} - P_{Col(X_1)}) y \sim \chi^2(h, \lambda_1) \tag{9.13}$$

where h is the difference in rank (degrees of freedom). The non-centrality parameter is:

$$\lambda_1 = \frac{1}{2\sigma^2}||(P_{Col(X)} - P_{Col(X_1)})\mu||^2 = \frac{1}{2\sigma^2}||\mu - \mu_0||^2 \tag{9.14}$$

3. ***Independence****: The two quadratic forms above are independent.*

Theorem 9.5 (F-Statistic for Nested Models)**.** *Under the conditions of the previous theorem:*

$$F = \frac{||\hat{y} - \hat{y}_0||^2/h}{s^2} = \frac{y^T (P_{Col(X)} - P_{Col(X_1)}) y/h}{y^T (I - P_{Col(X)}) y/(n-k-1)} \tag{9.15}$$

$$\sim \begin{cases} F(h, n-k-1), & \textit{under } H_0 \\ F(h, n-k-1, \lambda_1), & \textit{under } H_1 \end{cases} \tag{9.16}$$

9.3.1 Examples

9.3.1.1 Numerical Example: One-way ANOVA with Non-Full Rank Matrix

Example 9.1. We generate a balanced dataset with $a = 3$ groups and $n = 4$ observations per group. We then explicitly construct the non-full-rank design matrix $X = [j_N, x_1, x_2, x_3]$ and compute projections. We add a column for $\hat{y} - \hat{y}_0$ to explicitly show the vector whose squared length equals the Between Group Sum of Squares.

```r
library(MASS)  # For ginv()
library(knitr) # For kable()

# 1. Data Generation (3 Groups, 4 obs per group)
group_means_true <- c(10, 20, 15)
n_per_group <- 4
groups <- factor(rep(1:3, each = n_per_group))

# Adding some random noise
set.seed(123)
y <- c(rnorm(n_per_group, group_means_true[1], 2),
       rnorm(n_per_group, group_means_true[2], 2),
       rnorm(n_per_group, group_means_true[3], 2))
n <- length(y)

# 2. Construct Non-full-rank Design Matrix X
j_N <- rep(1, n)
X_ind <- model.matrix(~ groups - 1)
colnames(X_ind) <- c("x_1", "x_2", "x_3")
X <- cbind(Intercept = j_N, X_ind)

# 3. Compute Projections
# A. Projection onto j_N (Grand Mean)
P0 <- j_N %*% t(j_N) / as.numeric(t(j_N) %*% j_N)
y_grand_mean <- P0 %*% y

# B. Projection onto [j_N, X] (Full Column Space)
XtX <- t(X) %*% X
X_inv_gen <- ginv(XtX)
P_X <- X %*% X_inv_gen %*% t(X)
y_proj_X <- P_X %*% y

# C. Difference Vector and Residuals
y_diff <- y_proj_X - y_grand_mean
e <- y - y_proj_X

# --- Compute Arithmetic Group Means for Verification ---
group_means_vec <- ave(y, groups)

# Combine into a data frame
results_df <- data.frame(
  y = y,
  j_N = j_N,
  x1 = X_ind[,1],
  x2 = X_ind[,2],
  x3 = X_ind[,3],
  Proj_Grand = as.vector(y_grand_mean),
  Proj_FullX = as.vector(y_proj_X),
  Group_Means = group_means_vec,
  Diff = as.vector(y_diff),
  Residuals = as.vector(e)
)
```

Table 9.1: Data, Projections, and Verification of Group Means

	y	j_N	x_1	x_2	x_3	$\hat{y}_0$	$\hat{y}$	$\bar{y}_{i\cdot}$	$\hat{y} - \hat{y}_0$	e
1	8.88	1	1	0	0	15.39	10.42	10.42	-4.97	-1.54
2	9.54	1	1	0	0	15.39	10.42	10.42	-4.97	-0.88
3	13.12	1	1	0	0	15.39	10.42	10.42	-4.97	2.70
4	10.14	1	1	0	0	15.39	10.42	10.42	-4.97	-0.28
5	20.26	1	0	1	0	15.39	20.52	20.52	5.13	-0.26
6	23.43	1	0	1	0	15.39	20.52	20.52	5.13	2.91
7	20.92	1	0	1	0	15.39	20.52	20.52	5.13	0.40
8	17.47	1	0	1	0	15.39	20.52	20.52	5.13	-3.05
9	13.63	1	0	0	1	15.39	15.23	15.23	-0.16	-1.60
10	14.11	1	0	0	1	15.39	15.23	15.23	-0.16	-1.12
11	17.45	1	0	0	1	15.39	15.23	15.23	-0.16	2.22
12	15.72	1	0	0	1	15.39	15.23	15.23	-0.16	0.49
Sum of Squares	3083.33	12	4	4	4	2841.62	3045.83	3045.83	204.21	37.49

```r
# 4. Compute F and p-value
SSH <- sum(y_diff^2)
SSE <- sum(e^2)
SST <- SSH + SSE  # Total Sum of Squares

rank_X <- 3
df_hyp <- rank_X - 1
df_err <- n - rank_X
df_tot <- n - 1

MSH <- SSH / df_hyp
MSE <- SSE / df_err
MST <- SST / df_tot # Mean Square Total

F_stat <- MSH / MSE
p_val <- pf(F_stat, df_hyp, df_err, lower.tail = FALSE)

# 5. Adjusted R-squared
R2_adj <- 1 - (MSE / MST)

# 6. Output Extended ANOVA Table
anova_tab <- data.frame(
  Source = c("Groups (Model)", "Residuals (Error)", "Total"),
  DF = c(df_hyp, df_err, df_tot),
  SS = c(SSH, SSE, SST),
  MS = c(MSH, MSE, MST),
  F = c(F_stat, NA, NA),
  P_value = c(p_val, NA, NA),

  # Custom Variance Estimation Column
  Sigma_Hat_Sq = c(MST - MSE,  # Group Row: MST - MSE
                   MSE,        # Error Row: MSE
                   MST),       # Total Row: MST

  R_sq_adj = c(R2_adj, NA, NA)
)

kable(anova_tab, digits = 4,
      col.names = c("Source", "DF", "SS", "MS", "F", "P-value",
                    "$\\hat{\\sigma}^2$", "$R_a^2$"),
      caption = "ANOVA Table with Variance Estimates and Adjusted R-squared",
      escape = FALSE)
```

Table 9.2: ANOVA Table with Variance Estimates and Adjusted R-squared

Source	DF	SS	MS	F	P-value	$\hat{\sigma}^2$	R_a^2
Groups (Model)	2	204.2120	102.1060	24.509	2e-04	17.8073	0.8104
Residuals (Error)	9	37.4945	4.1661	NA	NA	4.1661	NA
Total	11	241.7065	21.9733	NA	NA	21.9733	NA

Key Observation from the Table

- Note that the projection onto the group means ($\hat{y}_{groups}$) and the projection onto the overparameterized space ($\hat{y} = P_{[j_N, X]}y$) are identical (Columns 7 and 8). This confirms numerically that adding the linearly dependent intercept column does not change the column space of the model; $\text{Col}(X) = \text{Col}(\text{Indicators})$.
- The Sum of Squares for the new column $\hat{y} - \hat{y}_0$ (Column 8) is exactly equal to the **Between Group Sum of Squares (SSH)** in the ANOVA table. This numerically demonstrates that $SSH = ||\hat{y} - \hat{y}_0||^2$.

9.3.1.2 Numerical Example: Regression with Linear Dependency

Example 9.2. We generate a dataset with $n = 12$ observations and 4 predictors where $x_4 = 0.1x_1 + 0.5x_2 + x_3$. This makes the design matrix rank-deficient. We use the generalized inverse to compute projections onto the full model space and compare them with projections onto the subspace of just x_1, x_2, x_3.

1. Fitting Models and F-test

```r
library(MASS)  # For ginv()
library(knitr) # For kable()
library(gt)    # For HTML tables

# 1. Data Generation
set.seed(42)
n <- 12
x1 <- rnorm(n, mean=5, sd=2)
x2 <- rnorm(n, mean=10, sd=2)
x3 <- rnorm(n, mean=0, sd=1)
x4 <- 0.1*x1 + 0.5*x2 + 0.1*x3
beta_true <- c(10, 1, -1, 1)
y <- beta_true[1] + beta_true[2]*x1 + beta_true[3]*x2 + beta_true[4]*x3 + rnorm(n, sd=3)

# 2. Projections (Condensed for brevity, same as before)
j_N <- rep(1, n)
X_full <- cbind(Intercept = j_N, x1, x2, x3, x4)
X_sub  <- cbind(Intercept = j_N, x1, x2, x3)
P0 <- j_N %*% t(j_N) / as.numeric(t(j_N) %*% j_N)
y_0 <- P0 %*% y

# B. Projection onto Full Model Space (Rank Deficient)
XtX_full <- t(X_full) %*% X_full
X_inv_full <- ginv(XtX_full)
beta_hat_full <- X_inv_full %*% t(X_full) %*% y
y_hat_full <- X_full %*% beta_hat_full

# C. Projection onto Subspace (Full Rank)
XtX_sub <- t(X_sub) %*% X_sub
X_inv_sub <- solve(XtX_sub)
beta_hat_sub <- X_inv_sub %*% t(X_sub) %*% y
y_hat_sub <- X_sub %*% beta_hat_sub

# D. Difference Vector and Residuals
y_diff <- y_hat_full - y_0
e <- y - y_hat_full

# 4. Create Detailed Results Table
results_df <- data.frame(
y = y,
x1 = x1,
x2 = x2,
x3 = x3,
x4 = x4,
y_hat_0 = as.vector(y_0),
y_hat_sub = as.vector(y_hat_sub),   # New Column
y_hat_full = as.vector(y_hat_full),
y_diff = as.vector(y_diff),
e = as.vector(e)
)

# Calculate Sum of Squares for each column
SS_row <- colSums(results_df^2)
```

Table 9.3: Comparison of Projections: Full (Rank-Deficient) vs Subspace

	y	x_1	x_2	x_3	x_4	$\hat{y}_0$	$\hat{y}_{sub}$	$\hat{y}_{full}$	$\hat{y}_{full} - \hat{y}_0$	e
1	10.06	7.74	7.22	1.90	4.57	6.42	9.65	9.65	3.24	0.41
2	1.44	3.87	9.44	-0.43	5.07	6.42	1.72	1.72	-4.70	-0.27
3	-1.51	5.73	9.73	-0.26	5.41	6.42	4.26	4.26	-2.16	-5.76
4	3.34	6.27	11.27	-1.76	6.09	6.42	3.92	3.92	-2.50	-0.58
5	7.46	5.81	9.43	0.46	5.34	6.42	4.41	4.41	-2.01	3.05
6	8.38	4.79	4.69	-0.64	2.76	6.42	9.46	9.46	3.04	-1.08
7	15.63	8.02	5.12	0.46	3.41	6.42	13.57	13.57	7.16	2.06
8	0.70	4.81	12.64	0.70	6.87	6.42	-1.48	-1.48	-7.90	2.18
9	6.58	9.04	9.39	1.04	5.70	6.42	9.40	9.40	2.98	-2.82
10	9.13	4.87	6.44	-0.61	3.65	6.42	7.32	7.32	0.90	1.81
11	6.02	7.61	9.66	0.50	5.64	6.42	7.01	7.01	0.60	-0.99
12	9.76	9.57	12.43	-1.72	7.00	6.42	7.76	7.76	1.34	2.00
Sum of Squares	745.54	546.12	1037.30	12.92	334.51	493.97	676.12	676.12	182.15	69.42

```r
# 5. ANOVA Statistics
SSH <- sum(y_diff^2)
SSE <- sum(e^2)
rank_X <- qr(X_full)$rank
df_hyp <- rank_X - 1
df_err <- n - rank_X
df_tol <- df_hyp + df_err
MSH <- SSH / df_hyp
MSE <- SSE / df_err
F_stat <- MSH / MSE
p_val <- pf(F_stat, df_hyp, df_err, lower.tail = FALSE)
SST <- SSH + SSE
MST <- SST/df_tol
R2_adj <- 1-MSE/MST

# 4. Create ANOVA Data Frame
anova_df <- data.frame(
Source = c("Regression (Model)", "Residuals (Error)", "Total"),
DF = c(df_hyp, df_err, df_tot),
SS = c(SSH, SSE, SST),
MS = c(MSH, MSE, MST),
F = c(F_stat, NA, NA),
P_value = c(p_val, NA, NA),
Sigma_Hat_Sq = c(MST - MSE, MSE, MST),
R_sq_adj = c(R2_adj, NA, NA)
)

# 5. Output with LaTeX Headers in gt
if (knitr::is_html_output()) {
anova_df |>
   gt() |>
   # Use md() to enable LaTeX parsing in headers
   cols_label(
      Source = "Source",
      DF = "DF",
      SS = "SS",
      MS = "MS",
      F = "F",
      P_value = "P-value",
      Sigma_Hat_Sq = md("$\\hat{\\sigma}^2$"),
      R_sq_adj = md("$R_a^2$")
   ) |>
   fmt_number(
      columns = c(SS, MS, F, P_value, Sigma_Hat_Sq, R_sq_adj),
      decimals = 4
   ) |>
   sub_missing(columns = everything(), missing_text = "") |>
   tab_style(
      style = cell_text(color = "red", weight = "bold"),
      locations = cells_body(columns = DF, rows = Source == "Regression (Model)")
   ) |>
   tab_header(title = "ANOVA Table")
} else {
```

Table 9.4: ANOVA Table

Source	DF	SS	MS	F	P-value	$\hat{\sigma}^2$	R_a^2
Regression (Model)	3	182.1536	60.7179	6.9973	0.0126	14.1929	0.6206
Residuals (Error)	8	69.4190	8.6774	NA	NA	8.6774	NA
Total	11	251.5726	22.8702	NA	NA	22.8702	NA

Key Observation

Notice that the columns $\hat{y}_{sub}$ (projection onto x_1, x_2, x_3) and $\hat{y}_{full}$ (projection onto $x_1, \ldots, x_4$) are identical. This confirms that adding the linearly dependent predictor x_4 does not change the column space of the design matrix or the fitted values, even though it makes individual coefficients non-unique.

2. Model Performance and Inference for ρ^2

In a rank-deficient model where $r = \text{rank}(\mathbf{X})$, we define the following metrics based on the projection onto the model space $\mathcal{C}(\mathbf{X})$.

Mathematical Formulas

The **Coefficient of Determination** (R^2) and its adjusted version (R^2_a) are defined as:

$$R^2 = \frac{SS_{Reg}}{SS_{Total}} = \frac{\|\hat{\mathbf{y}}_{full} - \bar{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2} \tag{9.17}$$

$$R^2_a = 1 - (1 - R^2)\frac{n-1}{n-r} \tag{9.18}$$

For the **Confidence Interval of** ρ^2, we utilize the non-central F-distribution. The non-centrality parameter λ is related to the population squared multiple correlation ρ^2 by:

$$\lambda = n\frac{\rho^2}{1-\rho^2} \iff \rho^2 = \frac{\lambda}{\lambda + n} \tag{9.19}$$

We find the confidence limits $[\lambda_L, \lambda_U]$ by inverting the non-central F cumulative distribution function such that:

$$P(F \leq F_{obs} \mid \lambda_U) = \alpha/2 \quad \text{and} \quad P(F \leq F_{obs} \mid \lambda_L) = 1 - \alpha/2 \tag{9.20}$$

R Implementation

```r
# 7. R-Squared and Adjusted R-Squared
SS_Total <- sum((y - mean(y))^2)
r_sq <- SSH / SS_Total

# Using the rank computed earlier (rank_X = 4)
adj_r_sq <- 1 - ((1 - r_sq) * (n - 1) / (n - rank_X))

# 8. Confidence Interval for rho^2 (95%)
ci_rho2_F <- function(F_stat, df1, df2, n, conf_level = 0.95) {

   # Helper function to find the non-centrality parameter (lambda)
   get_ncp <- function(f_val, df1, df2, prob) {
      # If observed F is less than the critical value, the lower bound is 0
      if (f_val <= qf(prob, df1, df2)) return(0)

      # Objective function: find lambda such that P(F < f_val | lambda) = prob
      obj <- function(lambda) pf(f_val, df1, df2, ncp = lambda) - prob

      # Attempt to solve using uniroot
      result <- try(uniroot(obj, interval = c(0, 10000))$root, silent = TRUE)

      if (inherits(result, "try-error")) return(NA) else return(result)
   }

   # Calculate alpha tail probabilities
   alpha <- 1 - conf_level
   prob_lower <- 1 - (alpha / 2) # Upper quantile for lower limit of lambda
   prob_upper <- alpha / 2       # Lower quantile for upper limit of lambda

   # Calculate Non-centrality Parameters (Lambdas)
   # Note: The "Upper" probability yields the "Lower" lambda bound and vice versa
   lambda_low  <- get_ncp(F_stat, df1, df2, prob_lower)
   lambda_high <- get_ncp(F_stat, df1, df2, prob_upper)

   # Convert Lambda to Rho-squared
   # Formula: rho^2 = lambda / (lambda + n)
   rho2_low  <- lambda_low / (lambda_low + n)
   rho2_high <- lambda_high / (lambda_high + n)

   return(c(lower = rho2_low, upper = rho2_high))
}

ci_rho2 <- ci_rho2_F(F_stat, df_hyp, df_err, n)
rho2_low  <- ci_rho2[1]
rho2_high <- ci_rho2[2]

# 9. Summary Table
metrics_df <- data.frame(
Metric = c("R-Squared", "Adjusted R-Squared", "95% CI for $\\rho^2$"),
Value = c(
   round(r_sq, 4),
   round(adj_r_sq, 4),
   paste0("[", round(rho2_low, 4), ", ", round(rho2_high, 4), "]")
```

Table 9.5: Model Fit Metrics and Population Correlation Interval

Metric	Value
R-Squared	0.7241
Adjusted R-Squared	0.6206
95% CI for ρ^2	[0.0721, 0.8171]

Example 9.3 (Rank-Deficient Model Fitting and ANOVA Analysis)**.** This example demonstrates the end-to-end process of fitting a model with linear dependencies using R's `lm()` function and generating a formatted ANOVA table. We explicitly handle the potential for fitting errors and highlight the reduced degrees of freedom resulting from the singularity.

1. **Data Generation and Model Fitting**

   We first generate a dataset where x_4 is a perfect linear combination of x_1, x_2, and x_3. We then attempt to fit the full model using `lm()` within a `tryCatch` block to ensure robustness.

```
library(MASS)  # For ginv()
library(knitr) # For kable()
library(gt)    # For HTML tables

# 1. Generate Data with linear dependency: x4 = 0.1*x1 + 0.5*x2 + x3
set.seed(42)
n <- 12
x1 <- rnorm(n, mean=5, sd=2)
x2 <- rnorm(n, mean=10, sd=2)
x3 <- rnorm(n, mean=0, sd=1)
x4 <- 0.1*x1 + 0.5*x2 + x3
y <- 10 + 2*x1 - 1*x2 + 3*x3 + rnorm(n, sd=1)
df <- data.frame(y=y, x1=x1, x2=x2, x3=x3, x4=x4)

# 2. Fit model using lm() with tryCatch
lm_fit <- tryCatch({
  model <- lm(y ~ x1 + x2 + x3 + x4, data = df)
  model
}, error = function(e) {
  message("Error in fitting: ", e$message)
  return(NULL)
})

# 3. Analyze Coefficients and ANOVA
# Note: lm() sets the coefficient of x4 to NA due to the singularity
print("Estimated Coefficients:")
```

```
[1] "Estimated Coefficients:"
```

```r
print(coef(lm_fit))
```

```
(Intercept)          x1          x2          x3          x4
   9.143931    2.206439   -1.097512    2.494560          NA
```

```r
anova_results <- anova(lm_fit)
anova_df <- as.data.frame(anova_results)
anova_df$Term <- rownames(anova_results)
anova_df <- anova_df[, c("Term", "Df", "Sum Sq", "Mean Sq", "F value", "Pr(>F)")]

# 4. Combined Formatting and Output
if (knitr::is_html_output()) {
  # HTML Output with gt highlighting
  anova_df |>
    gt() |>
    fmt_number(columns = c("Sum Sq", "Mean Sq", "F value", "Pr(>F)"), decimals = 4) |>
    sub_missing(columns = everything(), missing_text = "") |>
    tab_style(
      style = cell_text(color = "red", weight = "bold"),
      locations = cells_body(columns = Df, rows = Term %in% c("x1", "x2", "x3"))
    ) |>
    tab_header(title = "ANOVA Table")

} else {
  # LaTeX Output with color highlighting
  anova_df$Df <- as.character(anova_df$Df)
  # Identify rows for model components and color them
  model_rows <- which(anova_df$Term %in% c("x1", "x2", "x3"))
  anova_df$Df[model_rows] <- paste0("\\textcolor{red}{", anova_df$Df[model_rows], "}")

  kable(anova_df, digits = 4, escape = FALSE,
        caption = "Sequential ANOVA Table")
}
```

Table 9.6: Sequential ANOVA Table

	Term	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	x1	1	184.6483	184.6483	191.5136	0
x2	x2	1	139.5381	139.5381	144.7261	0
x3	x3	1	72.8565	72.8565	75.5654	0
Residuals	Residuals	8	7.7132	0.9642	NA	NA

2. **Interpretation of Results**

   The `lm()` function automatically identifies that x_4 is aliased with the other predictors and assigns it `NA`. Consequently, in the ANOVA table, the variation is partitioned among the first three predictors, and x_4 contributes no additional degrees of freedom or sum of squares. This demonstrates that for non-full rank models, inference is limited to the estimable part of the parameter space.

9.4 Inference for Estimable Parameters

9.4.1 Estimability

Definition 9.2 (Estimability)**.** Let $\lambda = (\lambda_1, ..., \lambda_p)^T$ be a vector of constants. The parameter $\lambda^T\beta = \sum_j \lambda_j\beta_j$ is said to be estimable if there exists a vector a in R^n such that:

$$E(a^T y) = \lambda^T\beta \tag{9.21}$$

for all $\beta \in \mathcal{R}^p$.

This condition is equivalent to $\lambda^T\beta$ being estimable if and only if there exists a such that $X^T a = \lambda$ (i.e., iff λ lies in the row space of X).

Theorem 9.6 (Equivalent Conditions for Estimability)**.** *In the model $y = X\beta + e$, the linear function $\lambda^T\beta$ is estimable if and only if any one of the following equivalent conditions holds:*

1. *λ^T is a linear combination of the rows of X; that is, there exists a vector a such that $a^T X = \lambda^T$. Equivalently, this means $\lambda \in Col(X^T) = Row(X)$.*
2. *λ is a linear combination of the columns of $X^T X$. Because $Col(X^T) = Col(X^T X)$, there exists a vector b such that $X^T Xb = \lambda$.*
3. *λ satisfies the consistency condition using a generalized inverse of $X^T X$, such that $X^T X(X^T X)^- \lambda = \lambda$.*

Proof. We prove the equivalence of the conditions for estimability by following this logical progression:

1. **Definition of Estimability $\iff$ Condition (1)** By definition, $\lambda^T\beta$ is estimable if there exists a vector a such that $E(a^T y) = \lambda^T\beta$ for all β. Since $E(a^T y) = a^T E(y) = a^T X\beta$, the requirement $a^T X\beta = \lambda^T\beta$ must hold for all $\beta \in \mathbb{R}^p$. This is true if and only if $a^T X = \lambda^T$. Taking the transpose, $X^T a = \lambda$, which explicitly means $\lambda \in \text{Col}(X^T) = \text{Row}(X)$.

2. **Condition (1) $\iff$ Condition (2)** From Condition 1, we established that $\lambda \in \text{Col}(X^T)$. A fundamental result in matrix algebra states that the column space of X^T is identical to the column space of $X^T X$, meaning $\text{Col}(X^T) = \text{Col}(X^T X)$. Therefore, $\lambda \in \text{Col}(X^T X)$. By the definition of a column space, this implies there must exist a vector b such that $X^T Xb = \lambda$.

3. **Condition (2) $\iff$ Condition (3)** The existence of a solution b to the linear system $X^T Xb = \lambda$ means the system is consistent. A standard property of generalized inverses is that a linear system $Ax = y$

is consistent if and only if $AA^{-}y = y$. Applying this to our system, it is consistent if and only if $X^TX(X^TX)^{-}\lambda = \lambda$. We can verify this directly; if such a b exists, then:

$$X^TX(X^TX)^{-}(X^TXb) = X^TXb = \lambda \tag{9.22}$$

using the defining property of generalized inverses, $AA^{-}A = A$.

□

9.4.2 Example: Checking Estimability in One-Way ANOVA

Example 9.4 (Checking Estimability in One-Way ANOVA)**.** We use the balanced one-way ANOVA effects model matrix X ($a = 3$ groups, $n = 4$ observations per group) to verify different linear functions $\lambda^T\beta$. In this model, the parameter vector is $\beta = (\mu, \alpha_1, \alpha_2, \alpha_3)^T$.

1. **Defining the Vectors and Explicit Matrix Form**

   First, we observe the explicit form of the X^TX matrix. For a balanced one-way layout with $n = 4$ per group, the matrix is as follows :

   $$X^TX = \begin{pmatrix} 12 & 4 & 4 & 4 \\ 4 & 4 & 0 & 0 \\ 4 & 0 & 4 & 0 \\ 4 & 0 & 0 & 4 \end{pmatrix} \tag{9.23}$$

   We examine the following functions of the parameters:

   - $\lambda_1 = (1, 1, 0, 0)^T$: The cell mean for group 1 ($\mu + \alpha_1$).
   - $\lambda_2 = (0, 1, 0, 0)^T$: The individual treatment effect α_1.
   - $\lambda_3 = (0, 1, -1, 0)^T$: The contrast between group 1 and group 2 ($\alpha_1 - \alpha_2$).
   - $\lambda_4 = (1, 0, 0, 1)^T$: The cell mean for group 3 ($\mu + \alpha_3$).
   - $\lambda_5 = (12, 4, 4, 4)^T$: This corresponds to $12\mu + 4\sum \alpha_i$, which is the first column of X^TX.

2. **Numerical Consistency Check (Condition 3)**

   A linear function $\lambda^T\beta$ is estimable if and only if $X^TX(X^TX)^{-}\lambda = \lambda$. We perform this check for all vectors using the Moore-Penrose generalized inverse.

```r
library(MASS)
library(knitr)

# Setup X'X for balanced one-way layout (n=4 per group)
XtX <- matrix(c(
  12,  4,  4,  4,
   4,  4,  0,  0,
   4,  0,  4,  0,
   4,  0,  0,  4
), nrow = 4, byrow = TRUE)

XtX_inv <- ginv(XtX)

# Define lambda vectors
L1 <- c(1, 1, 0, 0)      # mu + alpha1
L2 <- c(0, 1, 0, 0)      # alpha1
L3 <- c(0, 1, -1, 0)     # alpha1-alpha2
L4 <- c(1, 0, 0, 1)      # mu + alpha3
L5 <- c(12, 4, 4, 4)     # 12mu + 4(alpha1 + alpha2 + alpha3)

# Compute (X'X)(X'X)^- * lambda for each
check_L1 <- XtX %*% XtX_inv %*% L1
check_L2 <- XtX %*% XtX_inv %*% L2
check_L3 <- XtX %*% XtX_inv %*% L3
check_L4 <- XtX %*% XtX_inv %*% L4
check_L5 <- XtX %*% XtX_inv %*% L5

# Combine results for display
final_comp <- data.frame(
  Param = c("mu", "alpha1", "alpha2", "alpha3"),
  L1_Target = L1, L1_Res = as.vector(check_L1),
  L2_Target = L2, L2_Res = as.vector(check_L2),
  L3_Target = L3, L3_Res = as.vector(check_L3),
  L4_Target = L4, L4_Res = as.vector(check_L4),
  L5_Target = L5, L5_Res = as.vector(check_L5)
)

kable(final_comp, digits = 3,
      col.names = c("Param", "$\\lambda_1$", "$L_1$",
                    "$\\lambda_2$", "$L_2$",
                    "$\\lambda_3$", "$L_3$",
                    "$\\lambda_4$", "$L_4$",
                    "$\\lambda_5$", "$L_5$"),
      caption = "Estimability Check for Multiple Parameter Functions")
```

Table 9.7: Estimability Check for Multiple Parameter Functions

Param	λ_1	L_1	λ_2	L_2	λ_3	L_3	λ_4	L_4	λ_5	L_5
mu	1	1	0	0.25	0	0	1	1	12	12
alpha1	1	1	1	0.75	1	1	0	0	4	4
alpha2	0	0	0	-0.25	-1	-1	0	0	4	4
alpha3	0	0	0	-0.25	0	0	1	1	4	4

3. **Summary of Results**

   - **Estimable Functions**: $\lambda_1, \lambda_3, \lambda_4$, and λ_5 all satisfy the condition because their result columns (L_i) match their target columns (λ_i). These represent cell means, contrasts, and linear combinations of the rows of X^TX.
   - **Non-Estimable Function**: λ_2 (individual α_1) fails the condition as the target and result do not match.
   - **The First Column (λ_5)**: This vector is estimable because it is explicitly the first row/column of X^TX. Since any linear combination of rows of X^TX is estimable, $\lambda_5^T\beta = 12\mu + 4\alpha_1 + 4\alpha_2 + 4\alpha_3$ is estimable. Note that this is equal to $\sum_{i,j} E(y_{ij})$, the sum of all expected cell means.

4. **Theoretical Conclusion**

   In non-full-rank models, we cannot estimate individual parameters that are not uniquely identified by the data. However, linear functions that lie in the row space of X (or the column space of X^TX)—such as cell means or treatment contrasts—are invariant to the choice of solution $\hat{\beta}$ and remain uniquely estimable.

Theorem 9.7 (Number of Estimable Functions)**.** *In the non-full-rank model $y = X\beta + e$, the number of linearly independent estimable functions of β is the rank of X.*

9.4.3 Properties of Estimators for Estimabile Parameters

Let $\lambda^T\beta$ be an estimable function. Let $\hat{\beta}$ be any solution to the normal equations. Then the estimator $\lambda^T\hat{\beta}$ has the following properties:

1. **Unbiasedness**: $E(\lambda^T\hat{\beta}) = \lambda^T\beta$.
2. **Uniqueness**: $\lambda^T\hat{\beta}$ is invariant to the choice of $\hat{\beta}$ (to the choice of generalized inverse).

Theorem 9.8 (Variance and Covariance)**.**

1. ***Variance**: The variance of $\lambda^T\hat{\beta}$ is unique and is given by:*

$$Var(\lambda^T\hat{\beta}) = \sigma^2\lambda^T(X^TX)^-\lambda \tag{9.24}$$

2. ***Covariance**: For two estimable functions $\lambda_1^T\beta$ and $\lambda_2^T\beta$:*

$$Cov(\lambda_1^T\hat{\beta}, \lambda_2^T\hat{\beta}) = \sigma^2\lambda_1^T(X^TX)^-\lambda_2 \tag{9.25}$$

9.4.4 Properties of Estimators for Estimabile Parameters

Let $\lambda^T\beta$ be an estimable function. Let $\hat{\beta}$ be any solution to the normal equations. Then the estimator $\lambda^T\hat{\beta}$ has the following properties:

1. **Unbiasedness**: $E(\lambda^T\hat{\beta}) = \lambda^T\beta$.
2. **Uniqueness**: $\lambda^T\hat{\beta}$ is invariant to the choice of $\hat{\beta}$ (to the choice of generalized inverse).

Theorem 9.9 (Variance and Covariance)**.**

1. ***Variance**: The variance of $\lambda^T\hat{\beta}$ is unique and is given by:*

$$Var(\lambda^T\hat{\beta}) = \sigma^2\lambda^T(X^TX)^-\lambda \tag{9.26}$$

2. ***Covariance**: For two estimable functions $\lambda_1^T\beta$ and $\lambda_2^T\beta$:*

$$Cov(\lambda_1^T\hat{\beta}, \lambda_2^T\hat{\beta}) = \sigma^2\lambda_1^T(X^TX)^-\lambda_2 \tag{9.27}$$

Theorem 9.10 (Gauss-Markov in Non-Full Rank Case)**.** *If $\lambda^T\beta$ is estimable in the spherical errors non-full rank linear model, then $\lambda^T\hat{\beta}$ is its Best Linear Unbiased Estimator (BLUE).*

9.4.5 Testable Hypotheses in Non-Full Rank Models

In non-full rank models, we cannot test hypotheses about individual parameters that are not uniquely identified by the data. Instead, we must restrict our inference to **testable hypotheses**, which are formulated using estimable functions.

Definition 9.3 (Formal Description of a Testable Hypothesis)**.** A linear hypothesis $H_0 : C\beta = 0$ is said to be **testable** if and only if every row of the $m \times p$ matrix C is an estimable function of β.
This implies that there exists a matrix A such that $C = AX$. Physically, this means H_0 can be expressed entirely in terms of the expected values of the observations.

9.4.6 The General Linear Hypothesis F-Test

If $y \sim N_n(X\beta, \sigma^2 I)$, where X is $n \times p$ with $\text{rank}(X) = k < p$, and $C\beta$ is a set of m linearly independent estimable functions (where $m \leq k$), the following results hold:

1. **Distribution of the Estimator**

   The least-squares estimator $C\hat{\beta}$ follows a multivariate normal distribution:

$$C\hat{\beta} \sim N_m \left[C\beta, \sigma^2 C(X^T X)^- C^T\right] \tag{9.28}$$

2. **The F-Statistic**

   The test statistic for $H_0 : C\beta = 0$ is defined as:

$$F = \frac{SSH/m}{SSE/(n-k)} \tag{9.29}$$

   where:

   - $SSH = (C\hat{\beta})^T [C(X^T X)^- C^T]^{-1} C\hat{\beta}$ is the Hypothesis Sum of Squares.
   - $SSE = y^T [I - X(X^T X)^- X^T] y$ is the Error Sum of Squares.
   - k is the **rank** of the matrix X, not the number of columns p.

3. **Sampling Distribution**

   The statistic F follows an F-distribution:

$$F \sim F(m, n-k) \tag{9.30}$$

   Under the alternative hypothesis $H_1 : C\beta eq 0$, the statistic follows a non-central F-distribution $F(m, n-k, \lambda)$, with non-centrality parameter:

$$\lambda = \frac{(C\beta)^T [C(X^T X)^- C^T]^{-1} C\beta}{2\sigma^2} \tag{9.31}$$

Summary of Degrees of Freedom

Component	Degrees of Freedom
Numerator (Model/Hypothesis)	m (Number of linearly independent estimable constraints)
Denominator (Error)	$n - k$ (Total observations minus **rank** of X)