Statistical Theory for Linear Models

Longhai Li

2025-12-23

1 Preface

1.1 Overview

This book has evolved from the lecture notes for *STAT 851: Statistical Linear Models*, a graduate-level course taught at the University of Saskatchewan. 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.

1.2 Audience and Prerequisites

1.3 Key Features

1.4 Structure of the Book

1.5 Acknowledgements

The lectures were built upon the lecture notes for stat 8260 by Danniel Hall and the textbook LINEAR MODELS IN STATISTICS, Second Edition, by Alvin C. Rencher and G. Bruce Schaalje, ISBN 978-0-471-75498-5 (cloth)

2 Introduction to Theory of Linear Models

2.1 Overview

This lecture covers the foundations of Linear Statistical Models, including Multiple Linear Regression and ANOVA.

2.1.1 Geometric Interpretation of Least Squares

We compare models based on the reduction of errors. Consider a full model and a reduced model (K_1).

Let:

- y be the observed vector.
- $\hat{y}_1$ be the prediction from the reduced model.
- $\hat{y}_2$ be the prediction from the full model.

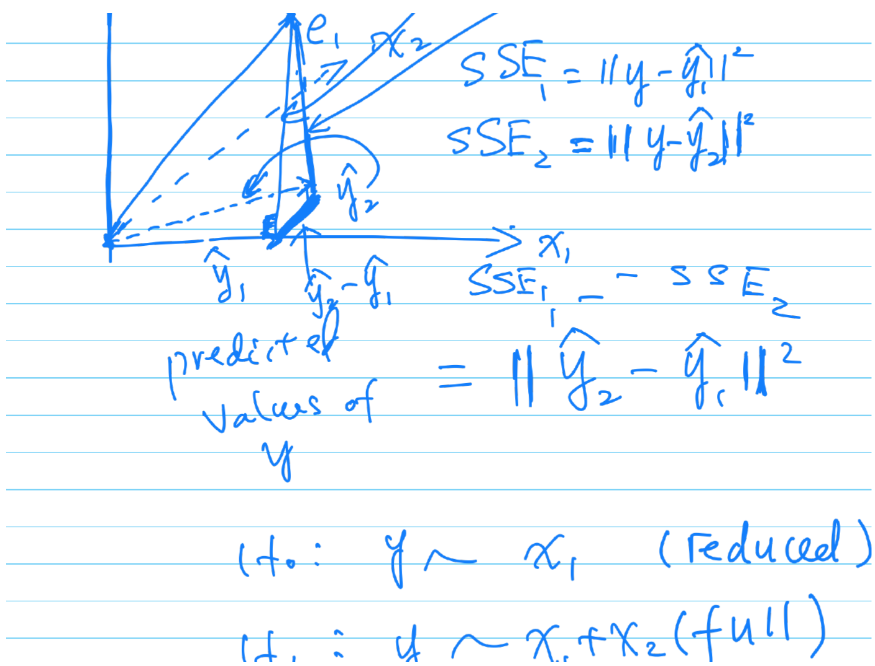

Figure 2.1: Geometric Interpretation of Least Squares

The errors (residuals) are defined as:

$$e_1 = y - \hat{y}_1$$
$$e_2 = y - \hat{y}_2$$

The Sum of Squared Errors (SSE) representing the distance is:

$$SSE_1 = ||y - \hat{y}_1||^2$$
$$SSE_2 = ||y - \hat{y}_2||^2$$

The statistical test is often based on the comparison of SSE_1 and SSE_2 (or the reduction in sums of squares $||\hat{y}_2 - \hat{y}_1||^2$).

2.2 Multiple Linear Regression

2.2.1 Matrix Formulation

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

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

Where the error terms are distributed as:

$$\epsilon \sim N_n(0, \sigma^2 I_n)$$

And I_n is the identity matrix:

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

The scalar equation for a single observation is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip} + \epsilon_i$$

2.3 Polynomial Regression

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

The model equation is:

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

2.3.1 Design Matrix Construction

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

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

2.4 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:

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

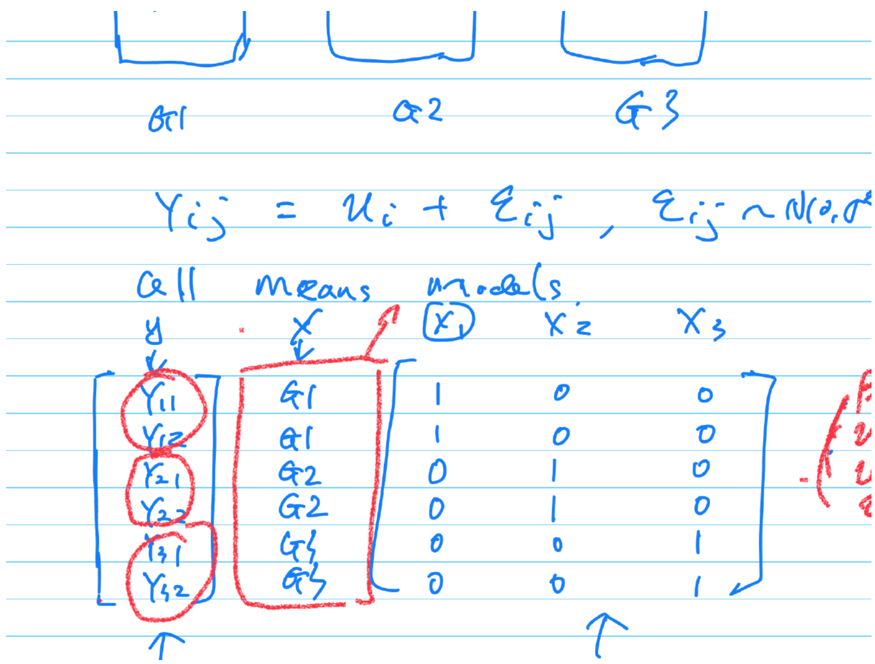

Figure 2.2: One-Way ANOVA Diagram

2.4.1 Dummy Variable Matrix

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

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

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

2.5 Analysis of Covariance (ANCOVA)

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

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

2.5.1 Least Squares Estimation

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

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

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

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

The residuals and Sum of Squared Errors are:

$$\hat{e} = y - \hat{y}$$
$$SSE = ||\hat{e}||^2$$

The coefficient of determination is:

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

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

3 Vector Space and Projection

3.1 Vector and Projection onto a Line

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 3.1 (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:

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

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 3.2 (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.

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

Definition 3.3 (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 $x + d = y$.

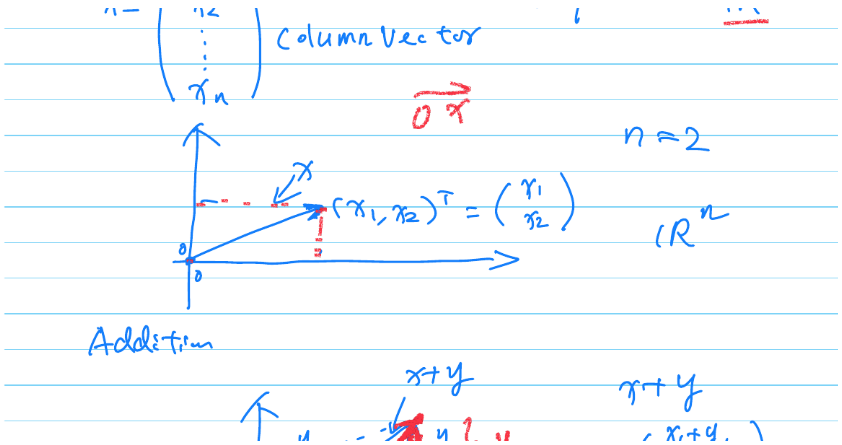

Figure 3.1: Vector Addition and Subtraction

Scalar Multiplication and Length

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

Definition 3.4 (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.

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

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

Definition 3.5 (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:

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

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

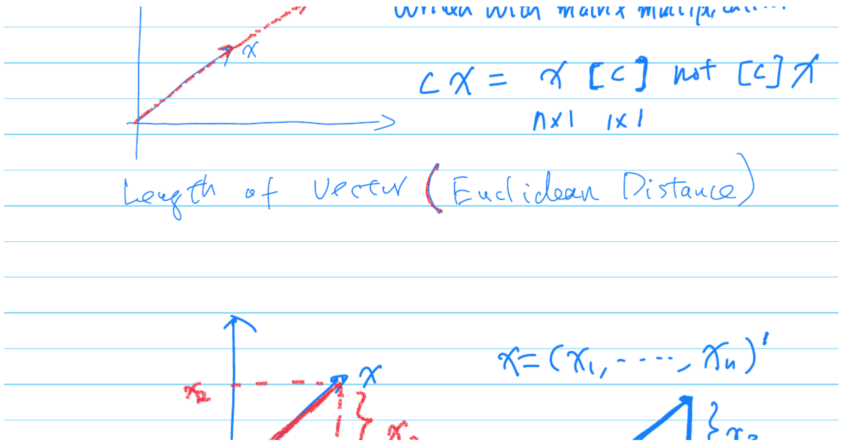

Figure 3.2: Scalar Multiplication and Length

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

$$c^2 = a^2 + b^2 - 2ab\cos\theta$$

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

$$||y - x||^2 = ||x||^2 + ||y||^2 - 2||x|| \cdot ||y|| \cos\theta$$

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

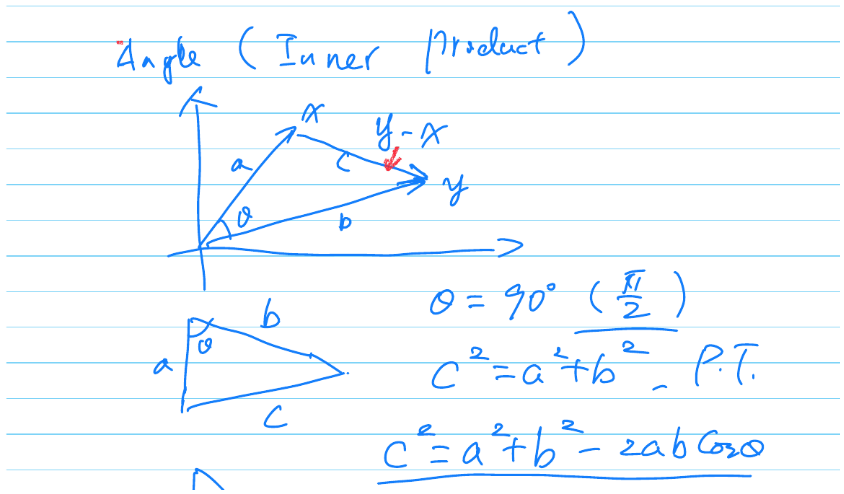

Figure 3.3: Geometry of Inner Product

Derivation of Inner Product

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

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

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

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

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

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

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

$$x'y = ||x|| \cdot ||y|| \cos\theta$$

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:

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

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

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

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

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

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

This is often written compactly by combining the denominators:

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

Perpendicularity (Orthogonality)

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

Definition 3.8 (Perpendicularity)**.** Two vectors are defined as **perpendicular** (or orthogonal) if the angle between them is 90° ($\pi/2$).

Since $\cos(90^\circ) = 0$, the condition for orthogonality simplifies to the inner product being zero:

$$x'y = 0 \iff x \perp y$$

Example 3.1 (Orthogonal Vectors)**.** Consider two vectors in $\mathbb{R}^2$: $x = (1, 1)'$ and $y = (1, -1)'$.

$$x'y = 1(1) + 1(-1) = 1 - 1 = 0$$

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

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 3.9 (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:

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

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 3.10 (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:

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

Expanding this inner product gives:

$$x'y - c(x'x) = 0$$

Solving for c, we obtain:

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

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.

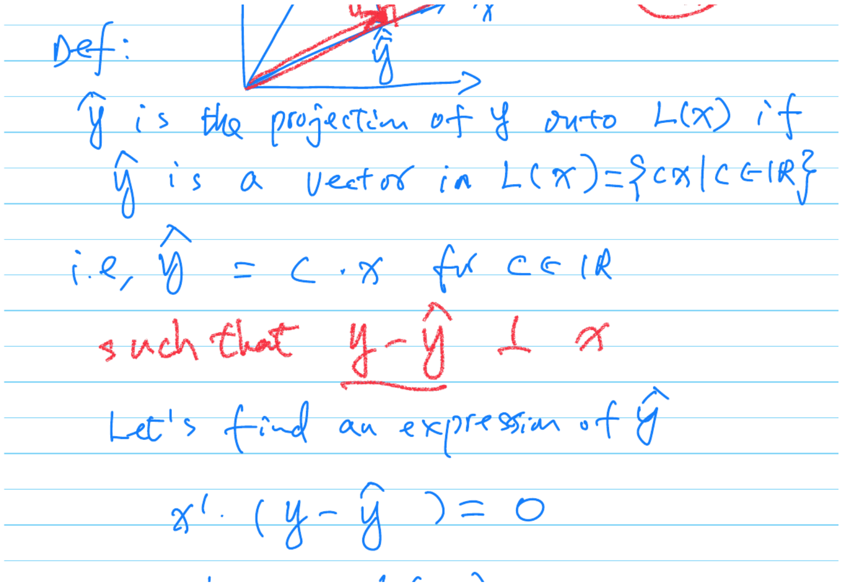

Figure 3.4: Projection Definition Diagram

Alternative Forms of the Projection Formula

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

Definition 3.11 (Forms of Projection)**.** The projection of y onto the vector x is given by:

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

This second form separates the components into:

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

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:

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

This leads to the definition of the projection matrix P_x.

Definition 3.12 (Projection Matrix onto a Single Vector)**.** The matrix P_x that projects any vector y onto the line spanned by x is defined as:

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

Using this matrix, the projection is simply:

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

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

Example: Projection in $\mathbb{R}^2$

Let's apply these concepts to a concrete example.

Example 3.2 (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:

$$x'y = 1(1) + 1(3) = 4$$

$$||x||^2 = 1^2 + 1^2 = 2$$

Now, apply the formula:

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

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

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

Multiply by y:

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

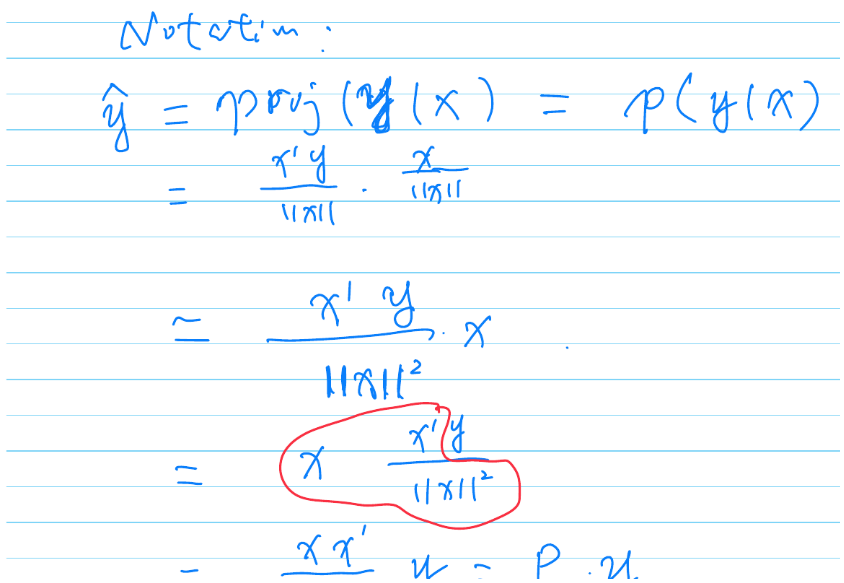

Figure 3.5: Example Calculation

Example: Projection onto the Mean Vector

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

Example 3.3 (Projection onto the "One" 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:

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

Calculating the components:

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

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

Substituting these back:

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

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.

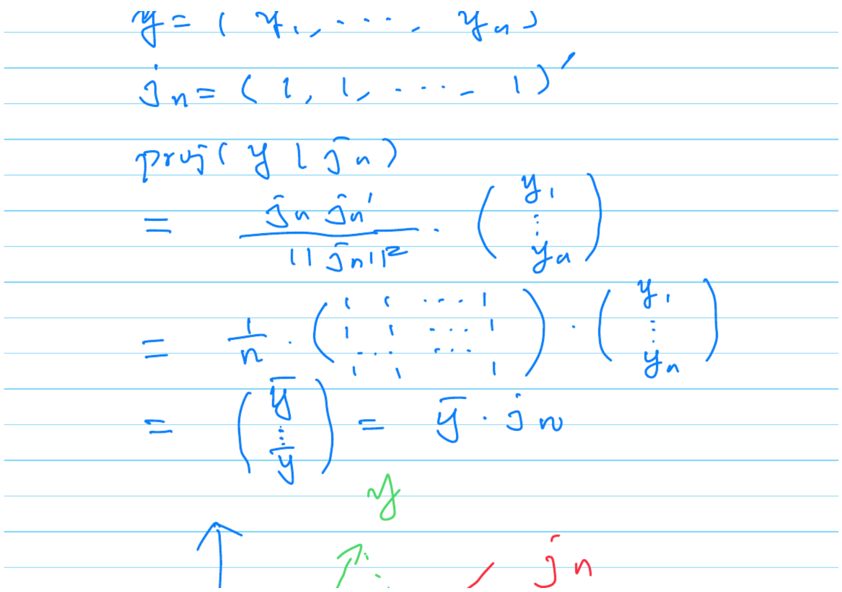

Figure 3.6: Projection onto Mean Vector

Pythagorean Theorem

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

Theorem 3.2 (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:*

$$||x + y||^2 = ||x||^2 + ||y||^2$$

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

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

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

$$||x + y||^2 = ||x||^2 + ||y||^2$$

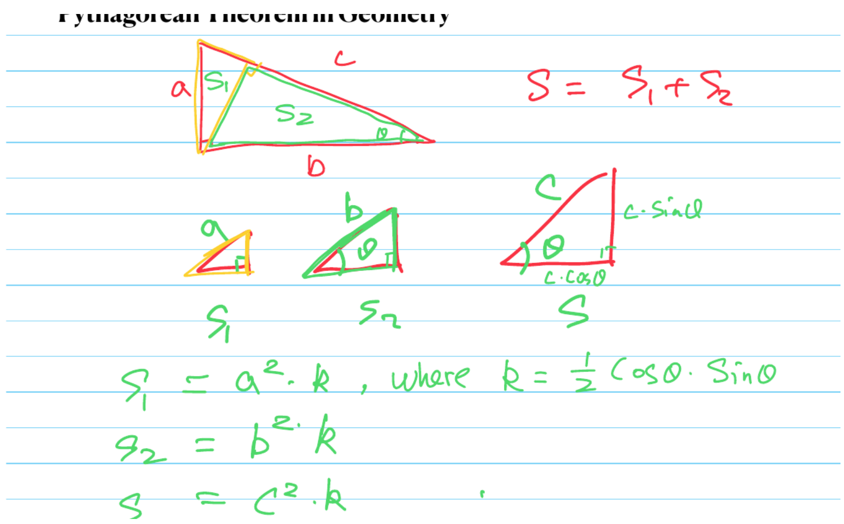

Figure 3.7: Pythagorean Theorem in Vector Space

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 3.3 (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 $\hat{y}$.*

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

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:

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

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

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

Applying the Pythagorean Theorem:

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

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

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

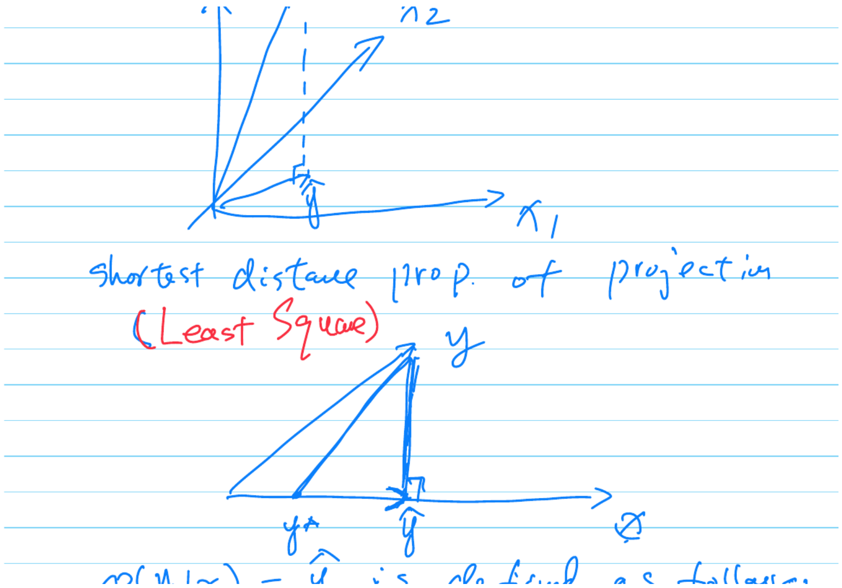

Figure 3.8: Least Square Property

3.2 General Vector Space

We now generalize our discussion from lines to broader spaces.

Definition 3.13 (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 $c = 0$).

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 3.14 (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:

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

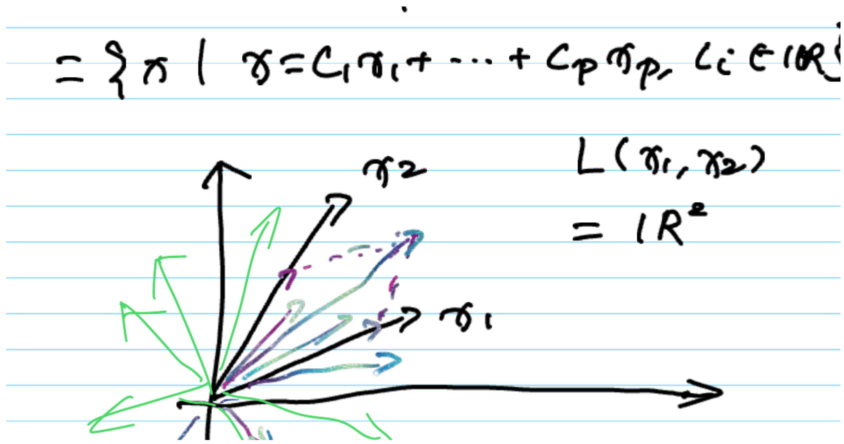

Figure 3.9: Spanned Vector Space

Column Space and Row Space

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

Definition 3.15 (Column Space)**.** For a matrix $X = (x_1, \dots, x_p)$, the **Column Space**, denoted $Col(X)$, is the vector space spanned by its columns:

$$Col(X) = L(x_1, \dots, x_p)$$

Definition 3.16 (Row Space)**.** The **Row Space**, denoted $Row(X)$, is the vector space spanned by the rows of the matrix X.

-

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 3.17 (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:

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

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

3.3 Rank of Matrices and Dim of Vector Space

Definition 3.18 (Rank)**.** The **Rank** of a matrix X, denoted Rank(X), is the maximum number of linearly independent columns in X. This is equivalent to the dimension of the column space:

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

3.3.0.1 Properties of Rank

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

Theorem 3.4 (Properties of Rank)**.**

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

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

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

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

Proof. Let X be an $n \times p$ matrix. Let r be the row rank of X. This means the dimension of the row space is r. Let $u_1, \dots, u_r$ be a basis for the row space of X (these are row vectors). Since every row of X is in the row space, each row $x_{i.}$ can be written as a linear combination of the basis vectors:

$$x_{i.} = c_{i1}u_1 + c_{i2}u_2 + \cdots + c_{ir}u_r \quad \text{for } i = 1, \dots, n$$

We can write this in matrix notation as:

$$X = CU$$

where C is an $n \times r$ matrix of coefficients c_{ij}, and U is an $r \times p$ matrix with rows $u_1, \dots, u_r$.

Now consider the columns of X. Since $X = CU$, the columns of X are linear combinations of the columns of C. Let $c^{(j)}$ be the j-th column of C. The columns of X lie in the space spanned by $\{c^{(1)}, \dots, c^{(r)}\}$. Thus, the column space of X, $Col(X)$, is a subspace of the column space of C.

$$\text{Dim}(Col(X)) \leq \text{Dim}(Col(C)) \leq r$$

The dimension of the column space of C is at most r (since C has only r columns). Therefore, Column Rank $\leq$ Row Rank.

Applying the same logic to X', we get Row Rank $\leq$ Column Rank. Combining these inequalities gives: **Row Rank = Column Rank**. □

Example: 2x3 Matrix

Consider the following 2×3 matrix:

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

Row Rank: The rows are $r_1 = (1, 0, 1)$ and $r_2 = (0, 1, 1)$. Neither is a multiple of the other, so they are linearly independent.

$$\text{Row Rank} = 2$$

Column Rank: The columns are $c_1 = \binom{1}{0}$, $c_2 = \binom{0}{1}$, and $c_3 = \binom{1}{1}$. Notice that $c_3 = c_1 + c_2$. The third column is dependent on the first two. However, c_1 and c_2 are independent (standard basis vectors).

$$\text{Column Rank} = 2$$

Thus, Rank(Row) = Rank(Col) = 2.

Orthogonality to a Subspace

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

Definition 3.19 (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.

$$y \perp V \iff y'x = 0 \quad \forall x \in V$$

Definition 3.20 (Orthogonal Complement)**.** The set of all vectors that are orthogonal to a subspace V is called the **Orthogonal Complement** of V, denoted $V^{\perp}$.

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

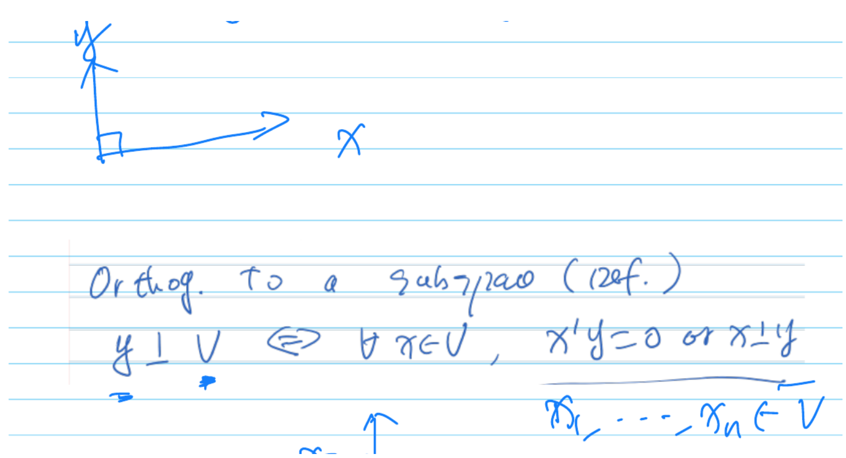

Figure 3.10: Orthogonal Complement

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 3.21 (Image and Kernel)**.**

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

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

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

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

Theorem 3.5 (Relationship between Kernel and Row Space)**.** *The kernel of X is the orthogonal complement of the row space of X:*

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

Nullity Theorem

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

Theorem 3.6 (Rank-Nullity Theorem)**.** *For an $n \times p$ matrix X:*

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

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

-

Rank Inequalities

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

Theorem 3.7 (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:*

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

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:

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

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:

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

Combining these gives the result.

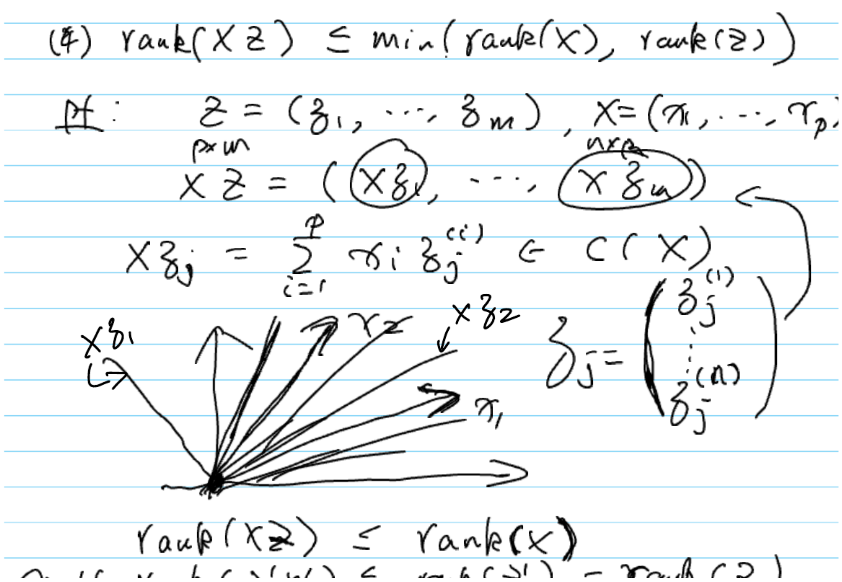

Figure 3.11: Rank of Matrix Product

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 3.8 (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:*

$$\textit{Rank}(AX) = \textit{Rank}(X)$$

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

$$\textit{Rank}(XB) = \textit{Rank}(X)$$

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

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

Thus, Rank$(AX) =$ Rank(X).

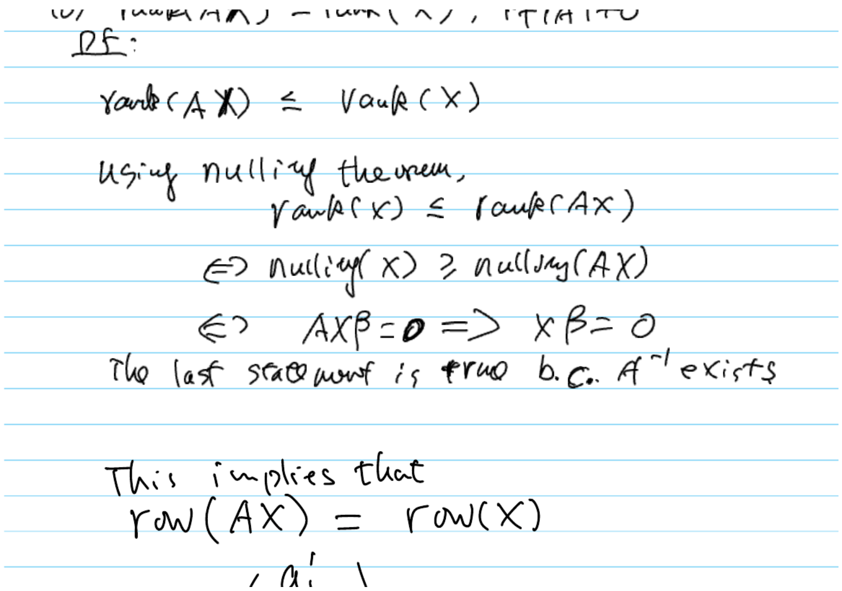

Figure 3.12: Rank Preservation with Invertible Matrices

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 3.9 (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:*

$$Rank(X'X) = Rank(X)$$
$$Rank(XX') = Rank(X)$$

Proof Strategy: 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':

$$v'X'Xv = 0 \implies (Xv)'(Xv) = 0 \implies ||Xv||^2 = 0 \implies Xv = 0$$

So Ker(X) = 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.

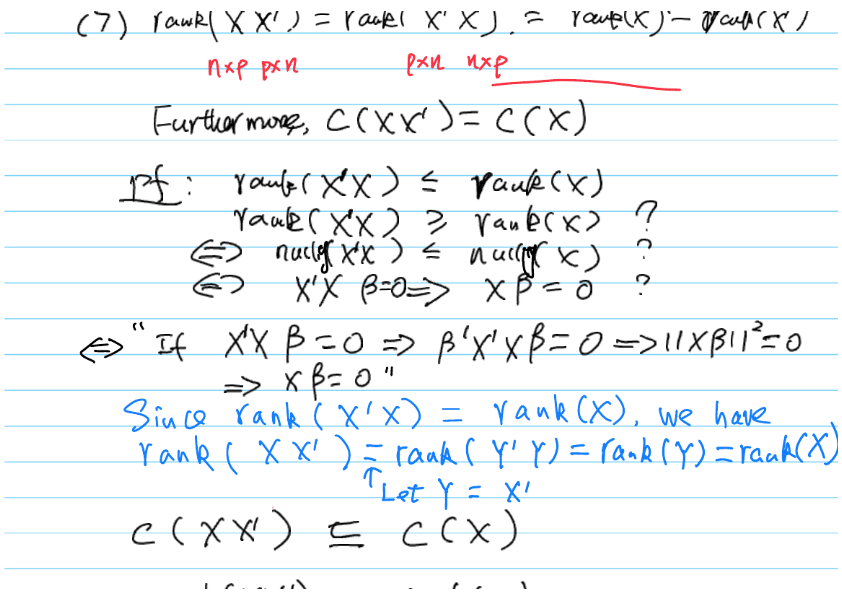

Figure 3.13: Rank of Gram Matrix

Column Space of XX'

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

Theorem 3.10 (Column Space Equivalence)**.** *The column space of XX' is identical to the column space of X:*

$$Col(XX') = Col(X)$$

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

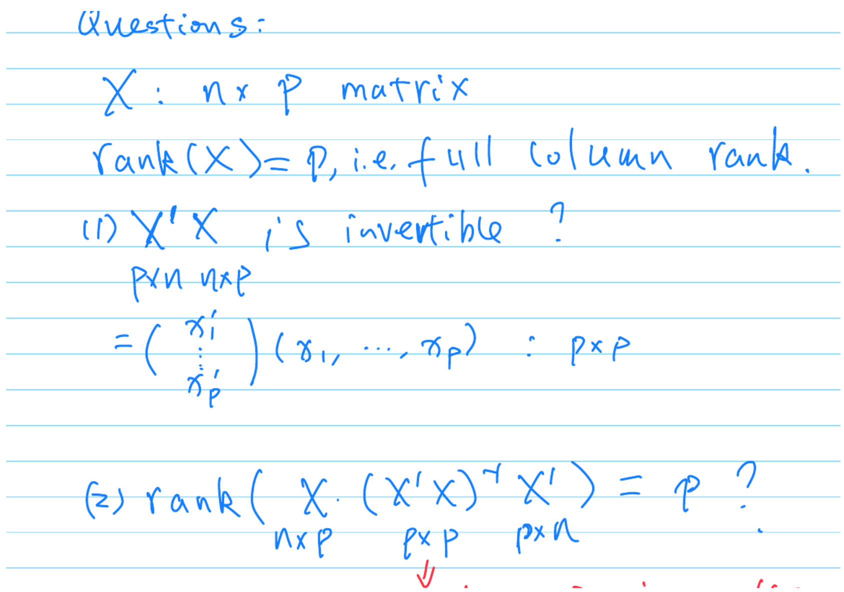

Figure 3.14: Column Space Equivalence

3.4 Projection via Orthonormal Basis

Basis and Dimension

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

Definition 3.22 (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 3.23 (Orthonormal Basis)**.** A basis $\{q_1, \dots, q_k\}$ is called an **Orthonormal Basis** if:

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

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

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

$$||q_i||^2 = q_i' q_i = 1$$

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

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

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

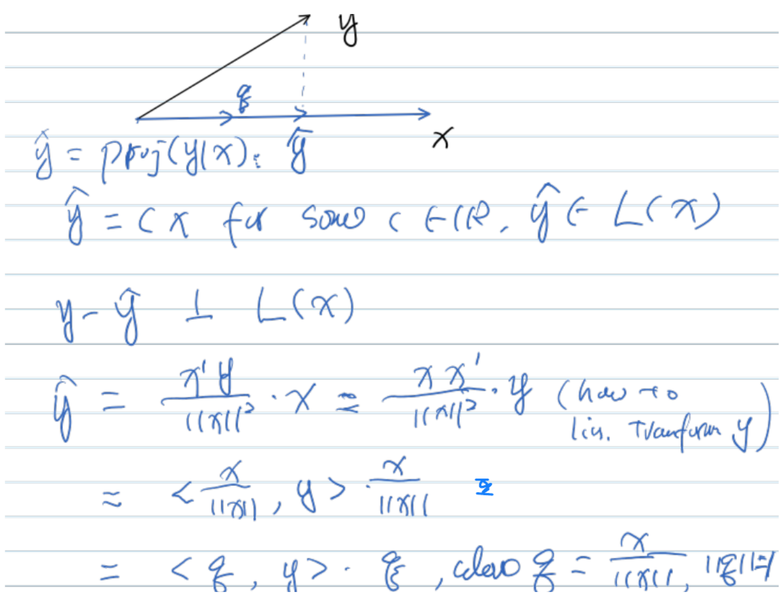

Figure 3.15: Projection onto a Subspace

The Projection Theorem

This theorem establishes the existence and uniqueness of the projection vector for any subspace, regardless of the basis used.

Theorem 3.11 (Projection Theorem)**.** *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:*

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

Equivalently:

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

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 $q_1, \dots, q_k$.

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

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

3.5 Projection via Projection Matrices

Theorem 3.12 (Projection Matrix (Orthonormal))**.** *The projection $\hat{y}$ can be written as:*

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

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

$$P = QQ'$$

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 3.13 (Properties of Projection Matrices)**.** *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 of Idempotence: If $\hat{y} = Py$ is already in the subspace $Col(X)$, then projecting it again should not change it.

$$P(Py) = Py \implies P^2y = Py \quad \forall y$$

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 3.4 (ANOVA as Projection)**.** Consider a one-way ANOVA model:

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

where i represents the group and j represents the observation within the group.

We can define dummy variables (indicators) for each group. Let x_1 be the indicator for group 1, x_2 for group 2, etc. These vectors are mutually orthogonal because an observation cannot belong to two groups simultaneously.

The projection of the data vector y onto the space spanned by these indicators is the sum of the projections onto each group vector:

$$\hat{y} = \text{proj}(y|x_1) + \text{proj}(y|x_2) + \ldots$$

Since the indicators are orthogonal, this simplifies to calculating the mean for each group. The fitted value for any observation y_{ij} is simply the group mean $\bar{y}_{i.}$.

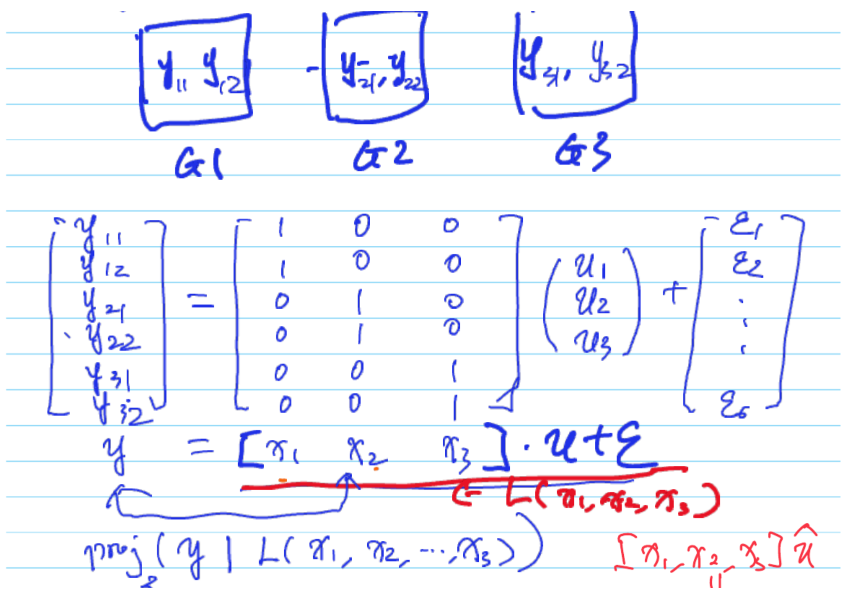

Figure 3.16: ANOVA Projection Geometry

Projection onto Orthogonal Complement

If P is the projection matrix onto a subspace V, we can easily define the projection onto the orthogonal complement $V^\perp$ (the "error" space).

Definition 3.25 (Projection onto Complement)**.** The matrix $M = I - P$ is the projection matrix onto the orthogonal complement $Col(X)^\perp$.

Properties of M: 1. **Idempotent:** $M^2 = (I-P)(I-P) = I-2P+P^2 = I-2P+P = I-P = M$. 2. **Symmetric:** $M' = (I-P)' = I-P' = I-P = M$. 3. **Orthogonal to P:** $PM = P(I-P) = P-P^2 = 0$.

This matrix M produces the residuals: $e = My = (I-P)y = y-\hat{y}$.

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.

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

1. **Step 1:** Normalize the first vector.
$$q_1 = \frac{x_1}{||x_1||}$$

2. **Step 2:** Project x_2 onto q_1 and subtract it to find the orthogonal component.
$$v_2 = x_2 - (x_2' q_1) q_1$$
Then normalize:
$$q_2 = \frac{v_2}{||v_2||}$$

3. **Step k:** Subtract the projections onto all previous q vectors.
$$v_k = x_k - \sum_{j=1}^{k-1} (x_k' q_j) q_j$$
$$q_k = \frac{v_k}{||v_k||}$$

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

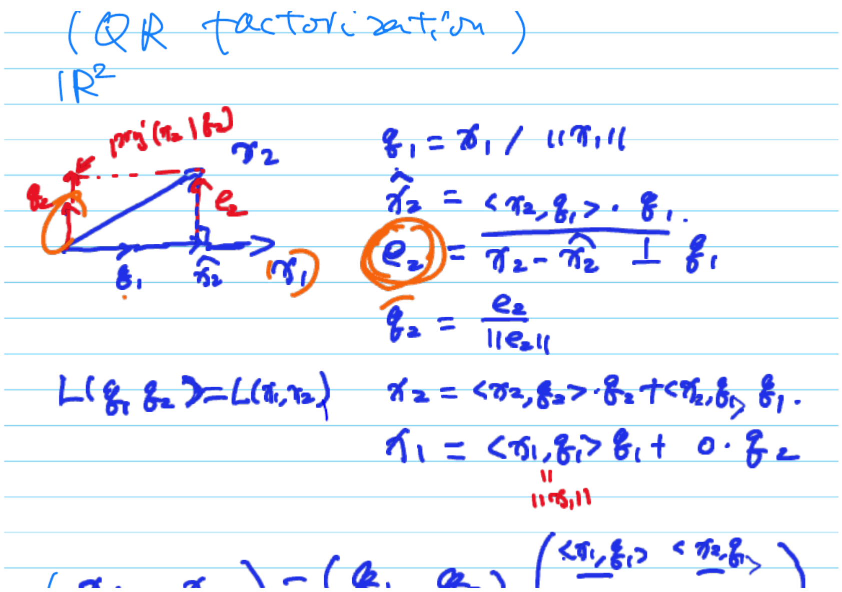

Figure 3.17: Gram-Schmidt Process

Projection Matrix Definition

We now formally derive the projection matrix P for the general case where we project y onto the column space of a matrix X.

Normal Equations

We want to find $\hat{y} = X\beta$ such that the residual $(y - \hat{y})$ is orthogonal to the column space $Col(X)$. This means the residual must be orthogonal to every column of X.

$$X'(y - X\beta) = 0$$

Expanding this gives the famous **Normal Equations**:

$$X'y - X'X\beta = 0 \implies X'X\beta = X'y$$

Theorem 3.14 (Least Squares Estimator)**.** *If $X'X$ is invertible (i.e., X has full column rank), the unique solution for β is:*

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

Let $X = (x_1, \cdots, x_p)$: $n \times p$ matrix

We want to project y to $C(X)$

That is, we want to find $\beta \in \mathbb{R}^p$ s.t.

$$y - X\beta \perp C(X)$$

$$\iff y - X\beta \perp x_i, \text{ for } i = 1, \cdots, p$$

$$\iff x_i'(y - X\beta) = 0, \text{ for each } i$$

$$\iff X'(y - X\beta) = 0$$

$$\iff \underset{p \times n}{X'}\ \underset{n \times 1}{y} = X'X\underset{n \times 1}{\beta} \leftarrow \text{normal equation}$$

Figure 3.18: Normal Equations Derivation

The Matrix P

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

Definition 3.26 (General Projection Matrix)**.** The projection of y onto $Col(X)$ is given by:

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

Thus, the projection matrix P is defined as:

$$P = X(X'X)^{-1}X'$$

Relationship with QR Decomposition

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

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

$$\begin{aligned}
P &= QR((QR)'(QR))^{-1}(QR)' \\
&= QR(R'Q'QR)^{-1}R'Q' \\
&= QR(R'R)^{-1}R'Q' \\
&= QRR^{-1}(R')^{-1}R'Q' \\
&= QQ'
\end{aligned}$$

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

Properties of P

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

Theorem 3.15 (Properties of P)**.** *The matrix $P = X(X'X)^{-1}X'$ satisfies:*

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

$$\mathit{tr}(P) = \mathit{tr}(X(X'X)^{-1}X') = \mathit{tr}((X'X)^{-1}X'X) = \mathit{tr}(I_p) = p$$

Projection onto Complement

As before, the projection onto the orthogonal complement (the residual maker matrix) is $M = I - P$.

Definition 3.27 (Residual Maker Matrix M)**.**

$$M = I - X(X'X)^{-1}X'$$

This matrix projects y onto the null space of X' (the orthogonal complement of the column space of X).

3.6 Projections onto Nested Subspaces

In hypothesis testing (like comparing a null model to an alternative model), we often deal with nested subspaces.

Definition 3.28 (Nested Models)**.** Consider two models: 1. **Reduced Model (M_0):** $y \in Col(X_0)$ 2. **Full Model (M_1):** $y \in Col(X_1)$

We say the models are nested if the column space of the reduced model is contained entirely within the column space of the full model:

$$Col(X_0) \subseteq Col(X_1)$$

Usually, X_1 is constructed by adding columns to X_0: $X_1 = [X_0, X_{new}]$.

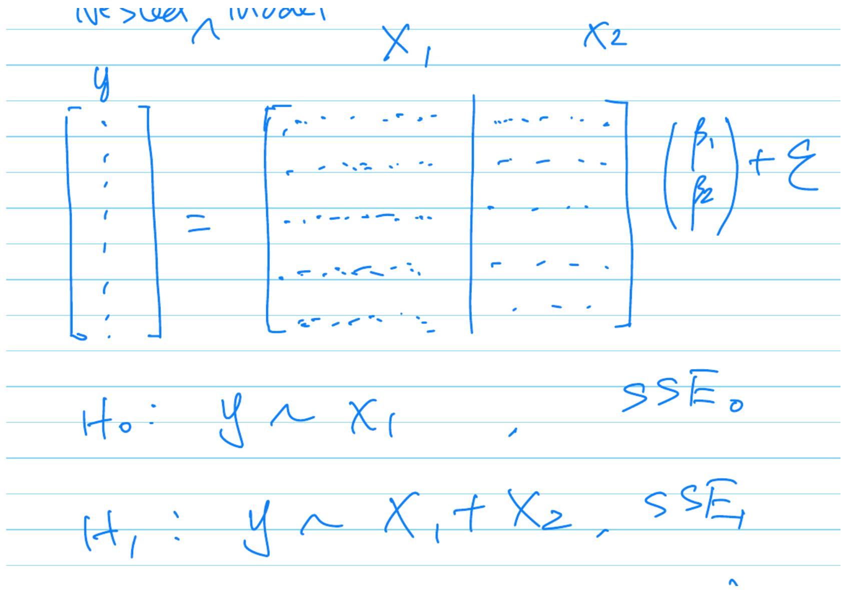

Figure 3.19: Nested Models Concept

Projection Composition

Let P_0 be the projection matrix onto $Col(X_0)$ and P_1 be the projection matrix onto $Col(X_1)$. Since $Col(X_0) \subseteq Col(X_1)$, we have important relationships between these matrices.

Theorem 3.16 (Composition of Projections)**.** *If $Col(P_0) \subseteq Col(P_1)$, then:*

1. *$P_1 P_0 = P_0$ (Projecting onto the small space, then the large space, keeps you in the small space).*
2. *$P_0 P_1 = P_0$ (Projecting onto the large space, then the small space, is the same as just projecting onto the small space).*

Difference of Projections

The difference between the two projection matrices, $P_1 - P_0$, is itself a projection matrix.

Theorem 3.17 (Difference Projection)**.** *The matrix $P_\Delta = P_1 - P_0$ is an orthogonal projection matrix onto the subspace $Col(X_1) \cap Col(X_0)^\perp$. This subspace represents the "extra" information in the full model that is orthogonal to the reduced model.*

Properties:

1. ***Symmetric:*** $(P_1 - P_0)' = P_1 - P_0$.
2. ***Idempotent:*** $(P_1 - P_0)(P_1 - P_0) = P_1 - P_0 P_1 - P_1 P_0 + P_0 = P_1 - P_0 - P_0 + P_0 = P_1 - P_0$.
3. ***Orthogonality:*** $(P_1 - P_0)P_0 = P_1 P_0 - P_0 = P_0 - P_0 = 0$.

Decomposition of Sum of Squares

This geometry allows us to decompose the total vector y into three orthogonal components:

1. $\hat{y}_0$ (The fit of the reduced model)
2. $\hat{y}_1 - \hat{y}_0$ (The improvement from the reduced to the full model)
3. $y - \hat{y}_1$ (The residual of the full model)

$$y = \hat{y}_0 + (\hat{y}_1 - \hat{y}_0) + (y - \hat{y}_1)$$

Squaring the norms (applying Pythagoras):

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

This equation is the foundation for the F-test in ANOVA and regression.

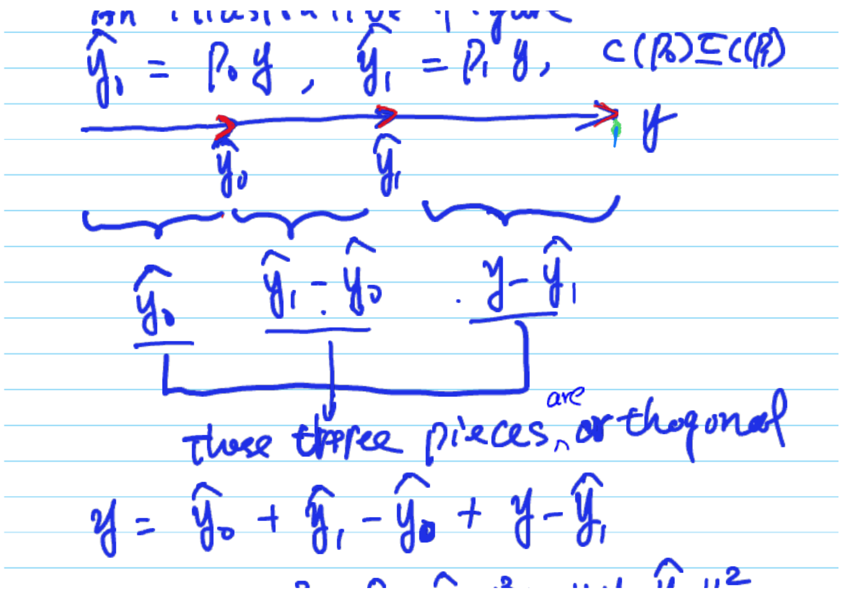

Figure 3.20: Decomposition of Sum of Squares

ANOVA Sum of Squares

We apply the decomposition of sum of squares to the specific case of Analysis of Variance.

Definition 3.29 (ANOVA Decomposition)**.** The Total Sum of Squares (TSS), or the squared norm of y (often centered), can be split into components:

1. **Residual Sum of Squares (Full Model):**

$$RSS_1 = ||y - \hat{y}_1||^2 = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

This represents the "Within Group" sum of squares.

2. **Difference Sum of Squares:**

$$RSS_0 - RSS_1 = ||\hat{y}_1 - \hat{y}_0||^2 = \sum_i n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

This represents the "Between Group" sum of squares ($SS_{between}$).

This relationship confirms that:

$$TSS = SS_{within} + SS_{between}$$

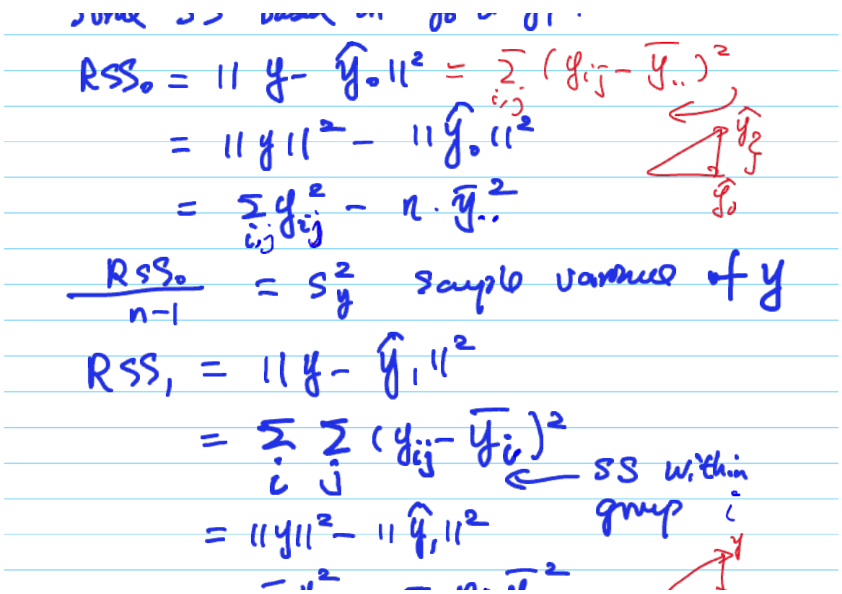

Figure 3.21: ANOVA Sum of Squares Derivation

Projections in Orthogonal Spaces

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

Theorem 3.18 (Orthogonal Decomposition)**.** *If $\mathbb{R}^n$ is the direct sum of orthogonal subspaces $V_1, V_2, \dots, V_k$:*

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

where $V_i \perp V_j$ for all $i \neq j$.

Then any vector y can be uniquely written as:

$$y = x_1 + x_2 + \cdots + x_k$$

where $x_i \in V_i$.

Furthermore, each component x_i is simply the projection of y onto the subspace V_i:

$$x_i = P_i y$$

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

$$I_n = P_1 + P_2 + \cdots + P_k$$

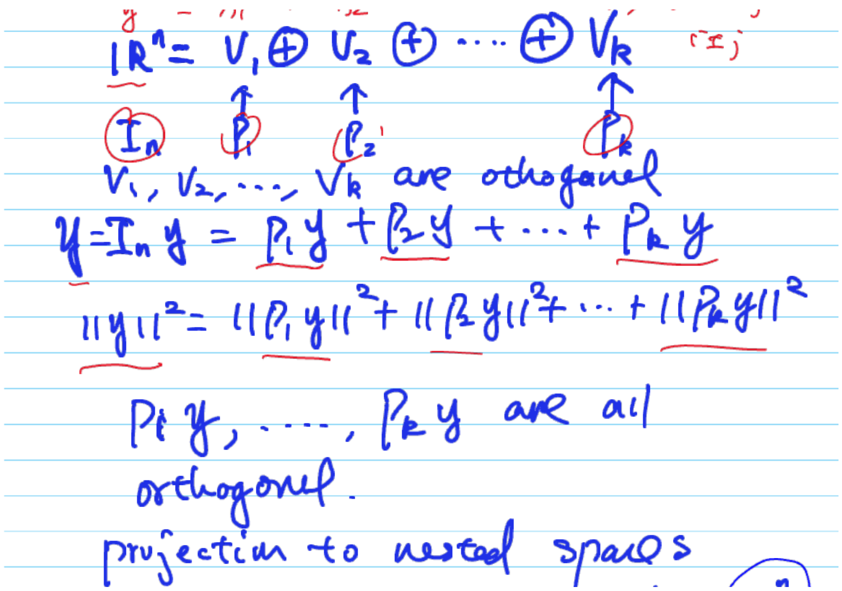

Figure 3.22: Orthogonal Space Decomposition

The following diagram summarizes the relationships between the vector spaces and projections discussed in this lecture.

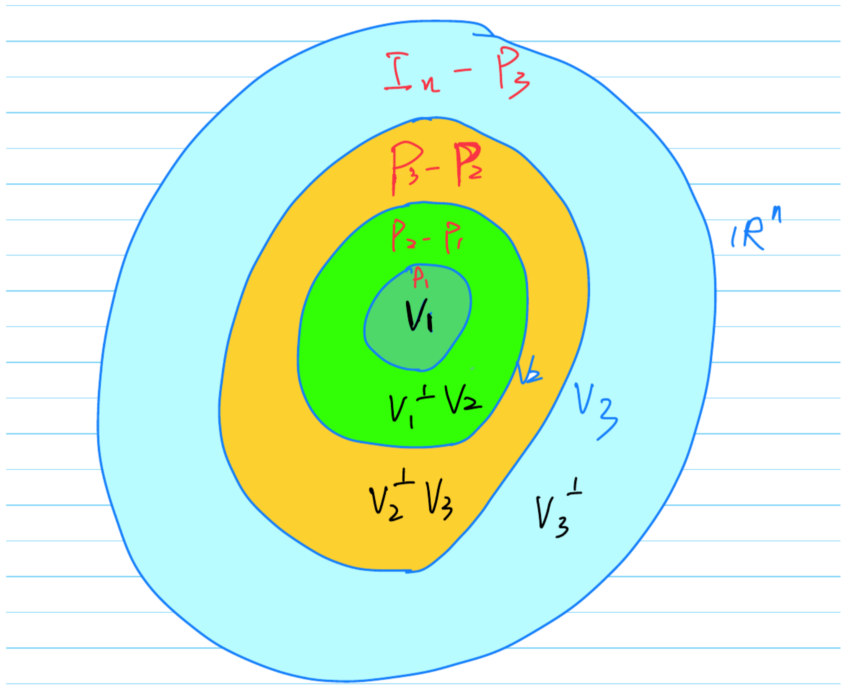

Figure 3.23: Summary Diagram