STA6246 Theory of Linear Models

Instructor Prof. James Hobert
Taken by Yu Zheng

Content List:

Review of Linear Algebra

Def: A set of vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(n)}$ is *linearly dependent* if there exist coefficients c_j, $j = 1, \cdots, n$, not all zero, such that

$$\sum_{j=1}^{n} c_j \mathbf{x}^{(j)} = \mathbf{0}.$$

This set of vectors is *linearly independent* if $\sum_{j=1}^{n} c_j \mathbf{x}^{(j)} = \mathbf{0}$ implies $c_j \equiv 0$.

Def: Two vectors are *orthogonal* to each other, written $\mathbf{x} \perp \mathbf{y}$, if

$$\mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x} = \sum_j x_j y_j = 0.$$

A set of vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(n)}$ are mutually orthogonal iff $\mathbf{x}^{(j)T} \mathbf{x}^{(j)} = 0$ for all $i \neq j$.

- A set of mutually orthogonal nonzero vectors are also linearly independent.

Def: A *vector space* $\mathcal{S}$ is a set of vectors that are closed under addition and scalar multiplication, that is, if $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are in $\mathcal{S}$, then $c_1 \mathbf{x}^{(1)} + c_2 \mathbf{x}^{(2)}$ is in $\mathcal{S}$.

Def: A vector space $\mathcal{S}$ is said to be *generated* by a set of vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(n)}$ if for every $\mathbf{x} \in \mathcal{S}$, there exist some coefficients c_j so that we can write

$$\mathbf{x} = \sum_j c_j \mathbf{x}^{(j)}.$$

Def: If a vector space $\mathcal{S}$ is generated by a set of linearly independent vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(n)}$, then this set of vectors form a *basis* for the space $\mathcal{S}$.

Def: The set of all linear combinations of a set of vectors is called the *span* of that set.

- $\mathbf{x} \in span\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(n)}\}$ if and only if there exists constants c_j such that $\mathbf{x} = \sum_j c_j \mathbf{x}^{(j)}$.

Def: The number of vectors in the basis for a vector space $\mathcal{S}$ is the *dimension* of the space $\mathcal{S}$, written $dim(\mathcal{S})$.

Def: The *rank* of a matrix $\mathbf{A}$ is the number of linear independent rows or columns, and denoted by $rank(\mathbf{A})$ or $r(\mathbf{A})$.

Def: The *column space* of a matrix, denoted by $\mathcal{C}(\mathbf{A})$, is the vector space spannned by the columns of the matrix, that is

$$\mathcal{C}(\mathbf{A}) = \{\mathbf{x} : \text{there exists a vectors } \mathbf{c} \text{ such that } \mathbf{x} = \mathbf{A}\mathbf{c}\}.$$

- $\dim(\mathcal{C}(\mathbf{A})) = rank(\mathbf{A})$
- $rank(\mathbf{AB}) \leq \min(rank(\mathbf{A}), rank(\mathbf{B}))$
- $\mathcal{C}(\mathbf{AB}) \subset \mathcal{C}(\mathbf{A})$
- If $\mathcal{C}(\mathbf{A}) \subset \mathcal{C}(\mathbf{B})$, then there exista a matrix $\mathbf{C}$ such that $\mathbf{A} = \mathbf{BC}$

Def: The *null space* of a matrix, denoted by $\mathcal{N}(\mathbf{A})$, is defined by $\mathcal{N}(\mathbf{A}) = \{\mathbf{y} : \mathbf{A}\mathbf{y} = \mathbf{0}\}$.

Notice that if the matrix $\mathbf{A}$ is $m \times n$, then vectors in $\mathcal{N}(\mathbf{A})$ have dimension n while vectors in $\mathcal{C}(\mathbf{A})$ have dimension m. Mathematically, this may be expressed as $\mathcal{C}(\mathbf{A}) \subset \mathbb{R}^m$ and $\mathcal{N}(\mathbf{A}) \subset \mathbb{R}^n$.

- If $\mathbf{A}$ has full-column rank, then $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$
- Suppose $\mathbf{A} \in \mathbb{R}^{m \times n}$, then $\dim(\mathcal{N}(\mathbf{A})) = n - r$ where $r = rank(\mathbf{A})$, or, more elegantly,

$$\dim(\mathcal{N}(\mathbf{A})) + \dim(\mathcal{C}(\mathbf{A})) = n.$$

Def: Two vector spaces $\mathcal{S}$ and $\mathcal{T}$ form *orthogonal complements* in $\mathbb{R}^m$ if and only if $\mathcal{S}, \mathcal{T} \subset \mathbb{R}^m$, $\mathcal{S} \cap \mathcal{T} = \{\mathbf{0}\}$, $\dim(\mathcal{S}) = r$, $\dim(\mathcal{T}) = n - r$, and every vector in $\mathcal{S}$ is orthogonal to every vector in $\mathcal{T}$.

- Let $\mathcal{S}$ and $\mathcal{T}$ be orthogonal complements in $\mathbb{R}^m$, then any vector $\mathbf{x} \in \mathbb{R}^m$ can be written as $\mathbf{x} = \mathbf{s} + \mathbf{t}$ where $\mathbf{s} \in \mathcal{S}$ and $\mathbf{t} \in \mathcal{T}$, and this decomposition is unique
- If $\mathbf{A}$ is an $m \times n$ matrix, then $\mathcal{C}(\mathbf{A})$ and $\mathcal{N}(\mathbf{A}^T)$ are orthogonal complements in $\mathbb{R}^m$
- Let $\mathcal{S}_1$ and $\mathcal{T}_1$ be orthogonal complements, as welle as $\mathcal{S}_2$ and $\mathcal{T}_2$; then if $\mathcal{S}_1 \subset \mathcal{S}_2$, then $\mathcal{T}_2 \subset \mathcal{T}_1$
- COnsider two vectors spaces $\mathcal{S}$ and $\mathcal{T}$. If $\mathcal{S} \subset \mathcal{T}$ and $\dim(\mathcal{S}) = \dim(\mathcal{T}) = k$, then $\mathcal{S} = \mathcal{T}$
- Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{b}$ a fixed vector. If $\mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}$ for all $\mathbf{x} \in \mathbb{R}^n$, then $\mathbf{A} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$
- If $\mathbf{B}\mathbf{x} = \mathbf{C}\mathbf{x}$ for all $\mathbf{x}$, then $\mathbf{B} = \mathbf{C}$
- Let $\mathbf{A}$ have full-column rank; then if $\mathbf{AB} = \mathbf{AC}$, then $\mathbf{B} = \mathbf{C}$
- If $\mathbf{C}^T\mathbf{C} = \mathbf{0}$, then $\mathbf{C} = \mathbf{0}$

Def: A system of equations $\mathbf{Ax} = \mathbf{c}$ is *consistent* iff there exists a solution $\mathbf{x}^*$ such that $\mathbf{Ax}^* = \mathbf{c}$

- A system of equations $\mathbf{Ax} = \mathbf{c}$ is consistent iff $\mathbf{c} \in \mathcal{C}(\mathbf{A})$

Def: A matrix $\mathbf{G}$ is a *generalized inverse* of the matrix $\mathbf{A}$ iff it satisfies $\mathbf{AGA} = \mathbf{A}$

- Let $\mathbf{A}$ be an $m \times n$ matrix with rank r. If $\mathbf{A}$ can be partitioned as below, with $rank(\mathbf{A}) = rank(\mathbf{C}) = r$,

  $$\mathbf{A} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix} \begin{matrix} r \\ m-r \end{matrix}$$

  so that $\mathbf{C}$ is nonsingular, then the matrix

  $$\mathbf{G} = \begin{bmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{matrix} r \\ n-r \end{matrix}$$

  is a generalized inverse of $\mathbf{A}$
- Let $\mathbf{A}$ be an $m \times n$ matrix with rank r. If $\mathbf{A}$ can be partitioned as below, with $rank(\mathbf{A}) = rank(\mathbf{F}) = r$,

  $$\mathbf{A} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix} \begin{matrix} m-r \\ r \end{matrix}$$

  so that $\mathbf{C}$ is nonsingular, then the matrix

  $$\mathbf{G} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}^{-1} \end{bmatrix} \begin{matrix} n-r \\ r \end{matrix}$$

  is a generalized inverse of $\mathbf{A}$
- Let $\mathbf{A}$ be an $m \times n$ matrix with rank r. Let $\mathbf{P}$ and $\mathbf{Q}$ be permutation matrices such that

  $$\mathbf{PAQ} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix}$$

  where $rank(\mathbf{A}) = rank(\mathbf{C}) = r$ and $\mathbf{C}$ is nonsingular. Then the matrix $\mathbf{G}$ below is a generalized inverse of $\mathbf{A}$:

  $$\mathbf{G} = \mathbf{Q} \begin{bmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{P}$$

- Let $\mathbf{Ax} = \mathbf{c}$ be a consistent system of equations and let $\mathbf{G}$ be a generalized inverse of $\mathbf{A}$, then $\mathbf{Gc}$ is a solution to the euqations $\mathbf{Ax} = \mathbf{c}$
- Let $\mathbf{Ax} = \mathbf{c}$ be a consistent system of equations and let $\mathbf{G}$ be a generalized inverse of $\mathbf{A}$; then $\tilde{\mathbf{x}}$ is a solution to the equations $\mathbf{Ax} = \mathbf{c}$ iff there exists a vector $\mathbf{z}$ such that $\tilde{\mathbf{x}} = \mathbf{Gc} + (\mathbf{I} - \mathbf{GA})\mathbf{z}$

Def: A square matrix $\mathbf{P}$ is *idempotent* iff $\mathbf{P}^2 = \mathbf{P}$

Def: A square matrix $\mathbf{P}$ is a *projection* onto the vector space $\mathcal{S}$ iff

1. $\mathbf{P}$ is idempotent,
2. for any $\mathbf{x}$, $\mathbf{Px} \in \mathcal{S}$. and
3. if $\mathbf{z} \in \mathcal{S}$, $\mathbf{Pz} = \mathbf{z}$

- $\mathbf{A}\mathbf{A}^g$ is a projection onto $\mathcal{C}(\mathbf{A})$
- $(\mathbf{I}-\mathbf{A}^g\mathbf{A})$ is a projection onto $\mathcal{N}(\mathbf{A})$
- A symmetric, idempotent matrix $\mathbf{P}$ that projects onto the vector space $\mathcal{S}$ is unique
- If a symmetric, idempotent matrix $\mathbf{P}$ projects onto $\mathcal{S}$, then $\mathbf{I}-\mathbf{P}$ projects onto its orthogonal complement
- $trace(\mathbf{AB}) = trace(\mathbf{BA})$
- $trace(\mathbf{A}^T\mathbf{A}) = \sum_i \sum_j A_{ij}^2$
- $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$
- $det\left(\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}\right) = |\mathbf{A}||\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}| = |\mathbf{D}||\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}|$
- The ***spectral decomposition*** of a symmetric matrix is $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T = \sum_j \lambda_j \mathbf{q}^{(j)}\mathbf{q}^{(j)T}$where $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues, ordered in the same way the eigenvectors are stacked as columns in $\mathbf{Q}$
- For a symmetric $n \times n$ matrix $\mathbf{A}$, $|\mathbf{A}| = \lambda_1 \times \cdots \times \lambda_n$ and $trace(\mathbf{A}) = \lambda_1 + \cdots + \lambda_n$, where λ_j are the eigenvalues of $\mathbf{A}$
- For a symmetric $n \times n$ matrix $\mathbf{A}$, the rank of it is equal to the number of nonzero eigenvalues

Def: A matrix $\mathbf{A}$ is *nonnegative definite* iff $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$

Def: A matrix $\mathbf{A}$ is *positive definite* iff $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x}$

- (***Cholesky factorization***) A square matrix $\mathbf{A}$ is positive definite iff there exists a nonsigunlar lower triangular matrix $\mathbf{L}$ such that $\mathbf{A} = \mathbf{L}\mathbf{L}^T$
- $\left(\mathbf{A}^{\frac{1}{2}}\right)^{-1} = \mathbf{A}^{-\frac{1}{2}} = \mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Q}^T$

Chapter 1: The GLM and Examples

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

$\mathbf{y}$: $N \times 1$ random vector of observable responses

$\mathbf{X}$: $N \times p$ matrix of known constants

$\mathbf{b}$: $p \times 1$ vector of unknown constants

$\mathbf{e}$: $N \times 1$ random vector of unobservable error such that $\mathbb{E}(\mathbf{e}) = \mathbf{0}$

It is called a linear model because the expected value of $\mathbf{y}$, $\mathbb{E}(\mathbf{y})$, is linear in $\mathbf{b}$.

Define $\mu(\mathbf{b}) = \mathbf{X}\mathbf{b} = \mathbb{E}(\mathbf{y})$. $\mu(\cdot)$ is a linear function: Fix $\mathbf{b}^{(1)}, \mathbf{b}^{(2)} \in \mathbb{R}^p$, and $a_1, a_2 \in \mathbb{R}$, then

$$\mu(a_1\mathbf{b}^{(1)} + a_2\mathbf{b}^{(2)}) = \mathbf{X}(a_1\mathbf{b}^{(1)} + a_2\mathbf{b}^{(2)}) = a_1\mathbf{X}\mathbf{b}^{(1)} + a_2\mathbf{X}\mathbf{b}^{(2)} = a_1\mu(\mathbf{b}^{(1)}) + a_2\mu(\mathbf{b}^{(2)})$$

Examples:

- One-sample problem (p=1): $y_1, \cdots, y_N$ i.i.d., $\mathbb{E}(y_1) = \mu$, $Var(y_1) = \sigma^2$. This is a special case of the GLM with $\mathbf{y} = (y_1, \cdots, y_N)^T$ and $\mathbf{X}\mathbf{b} = \mathbf{1}_N\mu$; $\mathbf{e}$ is just a vector of i.i.d. centered errors $e_1 \overset{d}{=} y_1 - \mu$.
- Simple linear regression (p=2): $y_i = \beta_0 + \beta_1 x_i + e_i$, $i = 1, 2, \cdots, N$, where $\{e_i\}_{i=1}^N$ uncorrelated r.v.s with $\mathbb{E}(e_i) = 0$ and $Var(e_i) = \sigma^2$. This is a special case of the GLM with $\mathbf{y} = (y_1, \cdots, y_N)^T$, $\mathbf{X}\mathbf{b} = \begin{pmatrix} 1 & x_0 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, $\mathbf{e} = (e_1, \cdots, e_N)^T$.
  - Actually we are making the assumption that the x_i's are measured without error.

    Eg. $y_i = \beta_0 + \beta_1 n_i + e_i$ (True Model)

    y_i is the crop yield, and n_i is the nitrogen in soil. We don't see n_i exactly; we see a distorted version $x_i = n_i + u_i$, where u_i is a random measurement error.

    $y_i = \beta_0 + \beta_1(n_i + u_i) + e_i$ (Observed Model) $\rightarrow$ Is this a special case of the GLM? No!
- Multiple regression (p=k+1): $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i$, $i = 1, 2, \cdots, N$, where $\{e_i\}_{i=1}^N$ uncorrelated r.v.s with $\mathbb{E}(e_i) = 0$ and $Var(e_i) = \sigma^2$. This is a special case of the GLM with $\mathbf{y} = (y_1, \cdots, y_N)^T$, $\mathbf{X}\mathbf{b} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{11} & \cdots & x_{1k} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_N \end{pmatrix}$, $\mathbf{e} = (e_1, \cdots, e_N)^T$.
  - $y_i = \beta_0 + \beta_1 i + \beta_2 i^2 + e_i$ (Yes)
  - $y_i = \beta_0 + \beta_1 \cos\left(\frac{2\pi i}{7}\right) + \beta_2 \log(i+1) + e_i$ (Yes)
  - $y_i = \beta_0 + \beta_1 e^{-\beta_2 x_i} + e_i$ (No)
  - $y_i = \frac{\beta_0}{\beta_1 + \beta_2 x_i} + e_i$ (No)
- One-way ANOVA: $y_{ij} = \mu + \alpha_i + e_{ij}$, $i = 1, \cdots, a$, $j = 1, \cdots, n_i$, where $\{e_{ij}\}_{i,j=1}^{a,n_i}$ uncorrelated r.v.s with $\mathbb{E}(e_{ij}) = 0$ and $Var(e_{ij}) = \sigma^2$. If we assume that $\{\alpha_i\}_{i=1}^a$ are unknown constants, then this is a special case of the GLM with $\mathbf{y} = (y_{11}, \cdots, y_{1n_1}, y_{21}, \cdots, y_{2n_2}, \cdots, y_{a1}, \cdots, y_{an_a})^T$, $\mathbf{X}\mathbf{b} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_a} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_a} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \end{pmatrix}$, $\mathbf{e} = (e_{11}, \cdots, e_{1n_1}, e_{21}, \cdots, e_{2n_2}, \cdots, e_{a1}, \cdots, e_{an_a})^T$.

  The comments:
  - $\mathbf{X}$ is not full rank

- If the a treatments were randomly selected from a population of treatments, then $\{\alpha_i\}_{i=1}^a$ must be considered random, and in this case, we have a so-called mixed model. However, if we redefine $\mathbf{Xb}$ and $\mathbf{e}$, this mixed model is still a special case of the GLM with $\mathbf{Xb} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \mu$, $\mathbf{e} = \begin{pmatrix} \alpha_1 + e_{11} \\ \alpha_1 + e_{12} \\ \vdots \\ \alpha_a + e_{a1} \\ \vdots \\ \alpha_a + e_{an_a} \end{pmatrix}$

  Here elements of $\mathbf{e}$ are no longer uncorrelated--but that's okay.

- Two-way nested model: $y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk}, i = 1, \cdots, a, j = 1, \cdots, b_i, k = 1, \cdots, n_{ij}$, where $\{e_{ijk}\}_{i,j,k=1}^{a,b_i,n_{ij}}$ uncorrelated r.v.s with $\mathbb{E}(e_{ijk}) = 0$ and $Var(e_{ijk}) = \sigma^2$. We'll assume the $\{\alpha_i\}_{i=1}^a$ and $\{\beta_{ij}\}_{i,j=1}^{a,b_i}$ are fixed treatment effects. This is a special case of the GLM.

  - Eg: Hospital costs from medical patients in two states (Florida & Pennsylvania)

    $b_1 = 2$ Florida (Dade county, Menrec county)

    $b_2 = 3$ Penn (A, B, E)

    $$\mathbf{y} = \begin{pmatrix} y_{F,D,1} \\ \vdots \\ y_{F,D,n_{11}} \\ y_{F,M,1} \\ \vdots \\ y_{F,M,n_{12}} \\ y_{P,A,1} \\ \vdots \\ y_{P,E,n_{23}} \end{pmatrix}, \mathbf{Xb} = \begin{pmatrix} \mathbf{1}_{n_{11}} & \mathbf{1}_{n_{11}} & \mathbf{0} & \mathbf{1}_{n_{11}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_{n_{12}} & \mathbf{1}_{n_{12}} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{n_{12}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_{n_{21}} & \mathbf{0} & \mathbf{1}_{n_{21}} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{n_{21}} & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_{n_{22}} & \mathbf{0} & \mathbf{1}_{n_{22}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{n_{22}} & \mathbf{0} \\ \mathbf{1}_{n_{23}} & \mathbf{0} & \mathbf{1}_{n_{23}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{n_{23}} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_F \\ \alpha_P \\ \beta_{F,D} \end{pmatrix}$$

    $p = a + 1 + \sum_{i=1}^a b_i$, $N = \sum_{i=1}^a \sum_{j=1}^{b_i} n_{ij}$

- Two-way crossed model: $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, i = 1, \cdots, a, j = 1, \cdots, b, k = 1, \cdots, n_{ij}$, where $\{e_{ijk}\}_{i,j,k=1}^{a,b,n_{ij}}$ uncorrelated r.v.s with $\mathbb{E}(e_{ijk}) = 0$ and $Var(e_{ijk}) = \sigma^2$. Assume all fixed effects. This is a special case of the GLM.

  Simple special case: $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, i = 1, \cdots, a, j = 1, \cdots, b$.

  $\rightarrow$Randomized block model: α_i-treatment effects, β_j-block effects. If block effects are considered fixed, this is a special case of the GLM.

$p = a + b + 1, N = ab, \mathbf{y} = (y_{11}, \cdots, y_{1b}, y_{21}, \cdots, y_{2b}, \cdots, y_{a1}, \cdots, y_{ab})^T$,

$$\mathbf{Xb} = \begin{pmatrix} \mathbf{1}_b & \mathbf{1}_b & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I}_b \\ \mathbf{1}_b & \mathbf{0} & \mathbf{1}_b & \cdots & \mathbf{0} & \mathbf{I}_b \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{1}_b & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_b & \mathbf{I}_b \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \\ \beta_1 \\ \vdots \\ \beta_b \end{pmatrix}$$

if block effects are considered random, then $\mathbf{Xb} = \begin{pmatrix} \mathbf{1}_b & \mathbf{1}_b & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_b & \mathbf{0} & \mathbf{1}_b & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_b & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_b \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \end{pmatrix}, \mathbf{e} = \begin{pmatrix} \beta_1 + e_{11} \\ \vdots \\ \beta_b + e_{1b} \\ \vdots \\ \beta_1 + e_{a1} \\ \vdots \\ \beta_b + e_{ab} \end{pmatrix}$.

- Analysis of covariate: $y_{ij} = \mu + \alpha_i + \beta x_{ij} + e_{ij}, i = 1, \cdots, a, j = 1, \cdots, n_i$, where $\{e_{ij}\}_{i,j=1}^{a,n_i}$ uncorrelated r.v.s with $\mathbb{E}(e_{ij}) = 0$ and $Var(e_{ij}) = \sigma^2$.
  - Eg: $y_{ij} =$ weight gain for individual j on diet $i, x_{ij} =$ initial weight

    $p = a + 2, N = \sum_{i=1}^{a} n_i, \mathbf{y} = (y_{11}, \cdots, y_{1n_1}, y_{21}, \cdots, y_{2n_2}, \cdots, y_{a1}, \cdots, y_{an_a})^T$,

    $$\mathbf{Xb} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} & \tilde{\mathbf{x}}_1 \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} & \tilde{\mathbf{x}}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{1}_{n_a} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_a} & \tilde{\mathbf{x}}_a \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \\ \beta \end{pmatrix}, \tilde{\mathbf{x}}_i = (x_{i1}, \cdots, x_{in_i})^T,$$

    $\mathbf{e} = (e_{11}, \cdots, e_{1n_1}, e_{21}, \cdots, e_{2n_2}, \cdots, e_{a1}, \cdots, e_{an_a})^T$.

We will gradually build up assumptions on $\mathbf{e}$:

- No assumptions. Just find $\mathbf{b}$ to minimize $\|\mathbf{y} - \mathbf{Xb}\|$
- Assume $\mathbb{E}(\mathbf{e}) = 0$. Find unbiased estimators of $\boldsymbol{\lambda}^T\mathbf{b}$, where $\boldsymbol{\lambda}^T = (\lambda_1, \cdots, \lambda_p)$ (Fixed vector)
- Assume $\{e_i\}_{i=1}^{n}$ are uncorrelated with $\mathbb{E}(\mathbf{e}) = \mathbf{0}$ and $Var(\mathbf{e}_{ij}) = \sigma^2$. Find the minimum variance unbiased estimator of $\boldsymbol{\lambda}^T\mathbf{b}$
- Assume $\mathbf{e}$ is MVN and develop tests and CIs concerning $\boldsymbol{\lambda}^T\mathbf{b}$

Chapter 2: The Linear Least Squares Problem

2.1 The Normal Equations

Assumption: $\mathbf{y} \in \mathbb{R}^N$ and $\mathbf{X} \in \mathbb{R}^{N \times p}$

Goal: Find the value of $\mathbf{b} \in \mathbb{R}^p$ that minimizes

$$Q(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$$

A value of $\mathbf{b}$ that minimizes $Q(\mathbf{b})$ will be called a ***least squares solution***

$$\mathcal{C}(\mathbf{X}) = \{\mathbf{v} \in \mathbb{R}^N : \mathbf{v} = \mathbf{X}\mathbf{w} \text{ for some } \mathbf{w} \in \mathbb{R}^p\}$$

Mathematically, trying to find the closest point in $\mathcal{C}(\mathbf{X})$ to the data $\mathbf{y}$

The gradient vector is given by

$$\frac{\partial Q}{\partial \mathbf{b}} = (\frac{\partial Q}{\partial b_1}, \frac{\partial Q}{\partial b_2}, \cdots, \frac{\partial Q}{\partial b_p})^T$$

The minimum of Q will occur where the gradient is zero, so we want to solve

$$\frac{\partial Q}{\partial \mathbf{b}} = \mathbf{0}$$

Result 2.1: Assume $\mathbf{a}$ is a $p \times 1$ vector, $\mathbf{b}$ is a $p \times 1$ vector, $\mathbf{A}$ is a $p \times p$ matrix. Then

- $\frac{\partial \mathbf{a}^T\mathbf{b}}{\partial \mathbf{b}} = \mathbf{a}$
- $\frac{\partial \mathbf{b}^T\mathbf{A}\mathbf{b}}{\partial \mathbf{b}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{b}$

proof: For the first result, note that $\mathbf{a}^T\mathbf{b} = a_1b_1 + a_2b_2 + \cdots + a_pb_p$, so $\left(\frac{\partial \mathbf{a}^T\mathbf{b}}{\partial \mathbf{b}}\right)_j = \frac{\partial \mathbf{a}^T\mathbf{b}}{\partial \mathbf{b}_j} = a_j$.

Similarly,

$$\mathbf{b}^T\mathbf{A}\mathbf{b} = \sum_{j=1}^{p}\sum_{k=1}^{p} A_{jk}b_jb_k,$$

and to form $\left(\frac{\partial \mathbf{b}^T\mathbf{A}\mathbf{b}}{\partial \mathbf{b}}\right)_j$, the terms that depend on b_j are

$$A_{jj}b_j^2 + \sum_{k \neq j}(A_{jk} + A_{kj})b_kb_j,$$

so the partial derivative w.r.t. b_j is

$$2A_{jj}b_j + \sum_{k \neq j} A_{jk}b_k + \sum_{k \neq j} A_{kj}b_k,$$

which is the j-th element of $(\mathbf{A} + \mathbf{A}^T)\mathbf{b}$.

Back to $Q(\mathbf{b})$, using Result 2.1,

$$\frac{\partial Q}{\partial \mathbf{b}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{b}.$$

Now, setting the gradient to 0 yields

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}.$$

They are the ***normal equations***.

Example:

- SLR: $y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, \cdots, N$

  $$\mathbf{X}^T\mathbf{X}\mathbf{b} = \begin{pmatrix} N & N\bar{\mathbf{x}} \\ N\bar{\mathbf{x}} & \sum_{i=1}^N x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} N\bar{\mathbf{y}} \\ \sum_{i=1}^N x_i y_i \end{pmatrix}$$

  Assume that $\sum_{i=1}^N (x_i - \bar{\mathbf{x}})^2 > 0$, i.e., the x_i's are NOT all the same. The solution to the N.E. is

  $$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - \bar{\mathbf{x}}) y_i}{\sum_{i=1}^N (x_i - \bar{\mathbf{x}})^2}$$
  $$\hat{\beta}_0 = \bar{\mathbf{y}} - \hat{\beta}_1 \bar{\mathbf{x}}$$

  If $\sum_{i=1}^N (x_i - \bar{\mathbf{x}})^2 = 0$, then there are infinitely many solutions to the N.E.s

  $$\hat{\beta}_1 = c$$
  $$\hat{\beta}_0 = \bar{\mathbf{y}} - c\bar{\mathbf{x}}$$

- Balanced one-way ANOVA:

  $y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, \cdots, a, j = 1, \cdots, n$

  $$\mathbf{X} = \begin{pmatrix} \mathbf{1}_n & \mathbf{1}_n & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_n & \mathbf{0} & \mathbf{1}_n & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_n & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_n \end{pmatrix}$$

  $$\mathbf{X}^T\mathbf{X}\mathbf{b} = \begin{pmatrix} n & n & n & \cdots & n \\ n & n & 0 & \cdots & 0 \\ n & 0 & n & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n & 0 & 0 & \cdots & n \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \end{pmatrix}, \quad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} y_{\cdot\cdot} \\ y_{1\cdot} \\ \vdots \\ y_{a\cdot} \end{pmatrix}$$

  $\mathbf{X}^T\mathbf{X}$ is singular, again, there are infinitely many solutions to the N.E.s

$$\hat{\mu} = c$$
$$\hat{\alpha}_i = \bar{y}_i - c$$

for any $c \in \mathbb{R}$.

Next, we will show that the N.E.s are consistent and we'll also show that

$$Q(\mathbf{b}) \text{ is minimized at } \hat{\mathbf{b}} \quad \Leftrightarrow \quad \hat{\mathbf{b}} \text{ is a solution to the N.E.s}$$

2.2 The Geometry of Least Squares

Recall that a vector space, $\mathcal{S}$, in $\mathbb{R}^m$ is a collection of vectors that is closed under addition and scalar multiplication. So if $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in \mathcal{S}$ and $c_1, c_2 \in \mathbb{R}$, then $c_1\mathbf{x}^{(1)} + c_2\mathbf{x}^{(2)} \in \mathcal{S}$.

Def A.4: Two vector spaces $\mathcal{S}$ and $\mathcal{T}$ form *orthogonal complements* in $\mathbb{R}^m$ if and only if $\mathcal{S}, \mathcal{T} \subset \mathbb{R}^m$, $\mathcal{S} \cap \mathcal{T} = \{\mathbf{0}\}$, $\dim(\mathcal{S}) + \dim(\mathcal{T}) = n$, and $\mathbf{s}^T\mathbf{t} = 0$ for every vector $\mathbf{s}$ in $\mathcal{S}$ and every vector $\mathbf{t}$ in $\mathcal{T}$.

Result A.4: Let $\mathcal{S}$ and $\mathcal{T}$ be orthogonal complements in $\mathbb{R}^m$, then any vector $\mathbf{x} \in \mathbb{R}^m$ can be written as $\mathbf{x} = \mathbf{s} + \mathbf{t}$ where $\mathbf{s} \in \mathcal{S}$ and $\mathbf{t} \in \mathcal{T}$, and this decomposition is unique

Result A.5: If $\mathbf{A}$ is an $m \times n$ matrix, then $\mathcal{C}(\mathbf{A})$ and $\mathcal{N}(\mathbf{A}^T)$ are orthogonal complements in $\mathbb{R}^m$

Result A.6: Let $\mathcal{S}_1$ and $\mathcal{T}_1$ be orthogonal complements, as welle as $\mathcal{S}_2$ and $\mathcal{T}_2$; then if $\mathcal{S}_1 \subset \mathcal{S}_2$, then $\mathcal{T}_2 \subset \mathcal{T}_1$

Question: Are the N.E.s consistent? In other words, is it always true that $\mathbf{X}^T\mathbf{y} \in \mathcal{C}(\mathbf{X}^T\mathbf{X})$?

Lemma 2.1: $\mathcal{N}(\mathbf{X}^T\mathbf{X}) = \mathcal{N}(\mathbf{X})$

proof: $\mathbf{w} \in \mathcal{N}(\mathbf{X}) \Rightarrow \mathbf{X}\mathbf{w} = \mathbf{0} \Rightarrow \mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{0} \Rightarrow \mathbf{w} \in \mathcal{N}(\mathbf{X}^T\mathbf{X})$

$\mathbf{w} \in \mathcal{N}(\mathbf{X}^T\mathbf{X}) \Rightarrow \mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{0} \Rightarrow \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} = 0 \Rightarrow \|\mathbf{X}\mathbf{w}\| = 0 \Rightarrow \mathbf{X}\mathbf{w} = \mathbf{0} \Rightarrow \mathbf{w} \in \mathcal{N}(\mathbf{X})$

Result 2.2: $\mathcal{C}(\mathbf{X}^T\mathbf{X}) = \mathcal{C}(\mathbf{X}^T)$

proof: Result A.5 $\Rightarrow \mathcal{N}(\mathbf{X})$ and $\mathcal{C}(\mathbf{X}^T)$ are O.C.s, $\mathcal{N}(\mathbf{X}^T\mathbf{X})$ and $\mathcal{C}(\mathbf{X}^T\mathbf{X})$ are O.C.s

Result A.6 $\Rightarrow \mathcal{N}(\mathbf{X}) \subset \mathcal{N}(\mathbf{X}^T\mathbf{X}) \Rightarrow \mathcal{C}(\mathbf{X}^T\mathbf{X}) \subset \mathcal{C}(\mathbf{X}^T), \mathcal{N}(\mathbf{X}^T\mathbf{X}) \subset \mathcal{N}(\mathbf{X}) \Rightarrow \mathcal{C}(\mathbf{X}^T) \subset \mathcal{C}(\mathbf{X}^T\mathbf{X})$

Corollary 2.1: The N.E.s are consistent

proof: Result 2.2: $\mathbf{X}^T\mathbf{y} \in \mathcal{C}(\mathbf{X}^T) = \mathcal{C}(\mathbf{X}^T\mathbf{X})$

Result 2.3: $\hat{\mathbf{b}}$ is a solution to the N.E.s *iff* $\hat{\mathbf{b}}$ minimizes $Q(\mathbf{b}) = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$

*By Result A.12, $\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y}$ is a solution to the N.E.s

proof: By Corollary 2.1, we know that there exists a solution to the N.E.s, $\hat{\mathbf{b}}$.

$$
\begin{aligned}
Q(\mathbf{b}) &= (\mathbf{y}-\mathbf{X}\mathbf{b})^T(\mathbf{y}-\mathbf{X}\mathbf{b}) \\
&= (\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}+\mathbf{X}\hat{\mathbf{b}}-\mathbf{X}\mathbf{b})^T(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}+\mathbf{X}\hat{\mathbf{b}}-\mathbf{X}\mathbf{b}) \\
&= (\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})^T(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}) + 2(\hat{\mathbf{b}}-\mathbf{b})^T\mathbf{X}^T(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}) + (\mathbf{X}\hat{\mathbf{b}}-\mathbf{X}\mathbf{b})^T(\mathbf{X}\hat{\mathbf{b}}-\mathbf{X}\mathbf{b}) \\
&= Q(\hat{\mathbf{b}}) + \|\mathbf{X}(\hat{\mathbf{b}}-\mathbf{b})\|^2
\end{aligned}
$$

where the cross-product term vanishes since $\hat{\mathbf{b}}$ solves the N.E.s $\mathbf{X}^T(\mathbf{y}-\mathbf{X}^T\mathbf{b}) = \mathbf{0}$. Now since no other value of $\mathbf{b}$ can give a smaller value of $Q(\mathbf{b})$, clearly $\hat{\mathbf{b}}$ minimizes $Q(\mathbf{b})$.

For the other direction, we still have $\hat{\mathbf{b}}$ that solves N.E.s, and we now know that $\hat{\mathbf{b}}$ minimizes Q as well.

Suppose $\tilde{\mathbf{b}}$ minimizes Q. So $Q(\tilde{\mathbf{b}}) = Q(\hat{\mathbf{b}}) + \|\mathbf{X}(\hat{\mathbf{b}}-\tilde{\mathbf{b}})\|^2$. Since $Q(\tilde{\mathbf{b}}) = Q(\hat{\mathbf{b}})$, we know $\|\mathbf{X}(\hat{\mathbf{b}}-\tilde{\mathbf{b}})\|^2 = 0 \Rightarrow \mathbf{X}(\hat{\mathbf{b}}-\tilde{\mathbf{b}}) = \mathbf{0} \Rightarrow \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}\tilde{\mathbf{b}}$. Now $\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}^T\mathbf{X}\tilde{\mathbf{b}}$, so $\tilde{\mathbf{b}}$ satisfies the N.E.s.

Corollary 2.3: $\mathbf{X}\hat{\mathbf{b}}$ is invariant to the choice of a solution to the N.E.s.

proof: If $\hat{\mathbf{b}}$ and $\tilde{\mathbf{b}}$ are both solutions to N.E.s, then $Q(\tilde{\mathbf{b}}) = Q(\hat{\mathbf{b}}) + \|\mathbf{X}(\hat{\mathbf{b}}-\tilde{\mathbf{b}})\|^2$. So $\|\mathbf{X}(\hat{\mathbf{b}}-\tilde{\mathbf{b}})\| = 0$.

We have shown that the closest point in $\mathcal{C}(\mathbf{X})$ to $\mathcal{Y}$ is the vector of fitted values, $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$. The residual vector $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$ is in $\mathcal{N}(\mathbf{X}^T)$ since $\mathbf{X}^T\hat{\mathbf{e}} = \mathbf{X}^T(\mathbf{y}-\hat{\mathbf{y}}) = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\hat{\mathbf{b}} = 0$.

This unique orthogonal decomposition (from Result A.4) also yields a decomposition of sums of squares from the Pythagorean Theorem:

$$\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{e}}\|^2.$$

That is, the total sum of squares $\|\mathbf{y}\|^2$ is the sum of the regression sum of squares, or $SSR = \|\mathbf{X}\hat{\mathbf{b}}\|^2$, and error sum of squares, or $SSE = \|\hat{\mathbf{e}}\|^2$.

Now let's look at a different way of developing this decomposition using projections.

Result 2.4: $\mathbf{X}^T\mathbf{X}\mathbf{A} = \mathbf{X}^T\mathbf{X}\mathbf{B} \Leftrightarrow \mathbf{X}\mathbf{A} = \mathbf{X}\mathbf{B}$.

proof: ($\Leftarrow$) Obvious

($\Rightarrow$) We will show two different proofs:

- Algebraic: $\mathbf{X}^T\mathbf{X}(\mathbf{A}-\mathbf{B}) = \mathbf{0} \Rightarrow (\mathbf{A}-\mathbf{B})^T\mathbf{X}^T\mathbf{X}(\mathbf{A}-\mathbf{B}) = 0 \Rightarrow [\mathbf{X}(\mathbf{A}-\mathbf{B})]^T\mathbf{X}(\mathbf{A}-\mathbf{B}) = 0$
  $\Rightarrow (\text{Lemma A.1})\mathbf{X}(\mathbf{A}-\mathbf{B}) = \mathbf{0} \Rightarrow \mathbf{X}\mathbf{A} = \mathbf{X}\mathbf{B}$.
- Geometric: $\mathbf{X}^T\mathbf{X}(\mathbf{A}-\mathbf{B}) = 0$ implies that the columns of $\mathbf{X}(\mathbf{A}-\mathbf{B})$ are in $\mathcal{N}(\mathbf{X}^T)$. But the columns of $\mathbf{X}(\mathbf{A}-\mathbf{B})$ are also all in $\mathcal{C}(\mathbf{X})$. Since $\mathcal{C}(\mathbf{X})$ and $\mathcal{N}(\mathbf{X}^T)$ are O.C.s, the only vector that have in common is the zero vector. Thus, each conlumn of $\mathbf{X}(\mathbf{A}-\mathbf{B})$ must be $\mathbf{0}$, i.e., $\mathbf{X}(\mathbf{A}-\mathbf{B}) = \mathbf{0}$.

Some material from the appendix:

Def: A square matrix $\mathbf{P}$ is a *projection* onto the vector space $\mathcal{S}$ iff

1. $\mathbf{P}$ is idempotent,
2. for any $\mathbf{x}$, $\mathbf{P}\mathbf{x} \in \mathcal{S}$. and
3. if $\mathbf{z} \in \mathcal{S}$, $\mathbf{P}\mathbf{z} = \mathbf{z}$

Result A.14: $\mathbf{A}\mathbf{A}^g$ is a projection onto $\mathcal{C}(\mathbf{A})$.

proof: 1. $\mathbf{A}\mathbf{A}^g\mathbf{A}\mathbf{A}^g = \mathbf{A}\mathbf{A}^g$

2. $\mathbf{A}\mathbf{A}^g\mathbf{x} = \mathbf{A}(\mathbf{A}^g\mathbf{x}) \in \mathcal{C}(\mathbf{A})$
3. If $\mathbf{z} \in \mathcal{C}(\mathbf{A})$, then $\mathbf{z} = \mathbf{A}\mathbf{y}$ for some $\mathbf{y}$. $\mathbf{A}\mathbf{A}^g\mathbf{z} = \mathbf{A}\mathbf{A}^g\mathbf{A}\mathbf{y} = \mathbf{A}\mathbf{y} = z$

Result A.16: There is only one symmetric projection onto $\mathcal{S}$.

Result 2.5: $(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T$ is a generalized inverse of $\mathbf{X}$.

proof: $(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^g(\mathbf{X}^T\mathbf{X}) = \mathbf{X}^T\mathbf{X}$. By Result 2.4, we have $\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X} = \mathbf{X}$.

Theorem 2.1: Define $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T$. $\mathbf{P}_\mathbf{X}$ is **the** symmetric projection onto $\mathcal{C}(\mathbf{X})$.

proof: First, note that $\mathbf{P}_\mathbf{X} = \mathbf{A}\mathbf{A}^g$ where $\mathbf{X}$ is playing the role of $\mathbf{A}$. So by result A.14, $\mathbf{P}_\mathbf{X}$ is a projection onto $\mathcal{C}(\mathbf{X})$.

In order to establish symmetry, we first show that $\mathbf{P}_\mathbf{X}$ is invariant to the choice of G.I. of $\mathbf{X}^T\mathbf{X}$. Let $\mathbf{G}_1$ and $\mathbf{G}_2$ be two G.I.s of $\mathbf{X}^T\mathbf{X}$. Then $(\mathbf{X}^T\mathbf{X})\mathbf{G}_1(\mathbf{X}^T\mathbf{X}) = \mathbf{X}^T\mathbf{X} = (\mathbf{X}^T\mathbf{X})\mathbf{G}_2(\mathbf{X}^T\mathbf{X})$. By Result 2.4:

$$\mathbf{X}\mathbf{G}_1(\mathbf{X}^T\mathbf{X}) = \mathbf{X}\mathbf{G}_2(\mathbf{X}^T\mathbf{X}) \Rightarrow \mathbf{X}^T\mathbf{X}\mathbf{G}_1^T\mathbf{X}^T = \mathbf{X}^T\mathbf{X}\mathbf{G}_2^T\mathbf{X}^T.$$

Using Result 2.4 again:

$$\mathbf{X}\mathbf{G}_1^T\mathbf{X}^T = \mathbf{X}\mathbf{G}_2^T\mathbf{X}^T \Rightarrow \mathbf{X}\mathbf{G}_1\mathbf{X}^T = \mathbf{X}\mathbf{G}_2\mathbf{X}^T.$$

Hence, invariance.

Now, note that

$$(\mathbf{X}^T\mathbf{X})\mathbf{G}_1(\mathbf{X}^T\mathbf{X}) = \mathbf{X}^T\mathbf{X} \Rightarrow \mathbf{X}^T\mathbf{X}\mathbf{G}_1^T\mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{X}$$

which indicates that if $\mathbf{G}_1$ is a G.I. of $\mathbf{X}^T\mathbf{X}$, then so is G_1^T.

Finally,

$$\mathbf{P}_\mathbf{X}^T = \left[\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\right]^T = \mathbf{X}\left[(\mathbf{X}^T\mathbf{X})^g\right]^T\mathbf{X}^T = \mathbf{P}_\mathbf{X}.$$

Result 2.6: $\mathbf{I} - \mathbf{P}_\mathbf{X}$ is the unique symmetric projection onto $\mathcal{N}(\mathbf{X}^T)$.

proof: Follows immediately from corollary A.4.

Example: $\mathcal{S} = \{\mathbf{v} \in \mathbb{R}^N : \mathbf{v} = c\mathbf{1}_N \text{ for some } c \in \mathbb{R}\} = \mathcal{C}(\mathbf{1}_N)$

$\mathcal{T} = \{\mathbf{w} \in \mathbb{R}^N : \mathbf{1}_N^T\mathbf{w} = 0\} = \mathcal{N}(\mathbf{1}_N^T)$

- $\mathcal{S} \bigcap \mathcal{T} = \{0\}$
- $\dim(\mathcal{S}) + \dim(\mathcal{T}) = 1 + N - 1 = N$ (Theorem A.1)
- $\mathbf{v} \in \mathcal{S}, \mathbf{w} \in \mathcal{T}$, $\mathbf{v}^T\mathbf{w} = c\mathbf{1}_N^T\mathbf{w} = 0$

$\mathbf{P}_{\mathbf{1}_N} = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$, $\mathbf{P}_{\mathbf{1}_N}\mathbf{y} = \mathbf{1}_N\frac{1}{N}\mathbf{1}_N^T\mathbf{y} = \mathbf{1}_N\frac{1}{N}\sum_{i=1}^N y_i = \mathbf{1}_N\bar{y}$, $(\mathbf{I} - \mathbf{P}_{\mathbf{1}_N})\mathbf{y} = \mathbf{y} - \mathbf{P}_{\mathbf{1}_N}\mathbf{y} = \mathbf{y} - \mathbf{1}_N\bar{y}$.

Recall that we have the orthogonal decomposition: $\mathbf{y} = \mathbf{X}\hat{\mathbf{b}} + (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \hat{\mathbf{y}} + \hat{\mathbf{e}}$.

Now, $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y} = \mathbf{P}_\mathbf{X}\mathbf{y} \in \mathcal{C}(\mathbf{X})$ and $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} = (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{y} \in \mathcal{N}(\mathbf{X}^T)$.

Another way to think about minimizing $Q(\mathbf{b})$:

$$Q(\mathbf{b}) = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}} + \hat{\mathbf{y}} - \mathbf{X}\mathbf{b}\|^2 = \|\hat{\mathbf{y}} - \mathbf{X}\mathbf{b}\|^2 + \|\hat{\mathbf{e}}\|^2,$$

where the second piece cannot be minimized by varying $\mathbf{b}$. The first piece can be minimized to zero, because the equations

$$\mathbf{X}\mathbf{b} = \mathbf{P}_\mathbf{X}\mathbf{y} = \hat{\mathbf{y}}$$

are consistent because $\mathbf{P}_\mathbf{X}\mathbf{y} \in \mathcal{C}(\mathbf{X})$.

Result 2.7: The solutions to the N.E.s are the same as the solutions to the consistent equations $\mathbf{X}\mathbf{b} = \mathbf{P}_\mathbf{X}\mathbf{y}$

proof: Assume that $\hat{\mathbf{b}}$ solves $\mathbf{X}\mathbf{b} = \mathbf{P}_\mathbf{X}\mathbf{y}$. Then $\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}^T\mathbf{P}_\mathbf{X}\mathbf{y} = \mathbf{X}^T\mathbf{y}$, so $\hat{\mathbf{b}}$ solves N.E.s.

Now assume that $\hat{\mathbf{b}}$ solves N.E.s. Then $\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{P}_\mathbf{X}\mathbf{y} = \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y}$. Using Result 2.4, we have $\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y} = \mathbf{P}_\mathbf{X}\mathbf{y}$, so $\hat{\mathbf{b}}$ solves $\mathbf{X}\mathbf{b} = \mathbf{P}_\mathbf{X}\mathbf{y}$.

Theorem 2.2: If $\mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$, then $\mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{W}$ is the symmetric projection onto $\mathcal{C}\left((\mathbf{I} - \mathbf{P}_\mathbf{W})\mathbf{X}\right)$.

proof: We first show that $\mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{W}$ is idempotent.

$$(\mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{W})^2 = \mathbf{P}_\mathbf{X}^2 - \mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{W} - \mathbf{P}_\mathbf{W}\mathbf{P}_\mathbf{X} + \mathbf{P}_\mathbf{W}^2 = \mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{W} - \mathbf{P}_\mathbf{W}\mathbf{P}_\mathbf{X} + \mathbf{P}_\mathbf{W}.$$

Now, for any $\mathbf{z}$, $\mathbf{P}_\mathbf{W}\mathbf{z} \in \mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$. So $\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{W}\mathbf{z} = \mathbf{P}_\mathbf{W}\mathbf{z}$, $\forall \mathbf{z}$. Hence $\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{W} = \mathbf{P}_\mathbf{W}$ (Corollary A.1).

Then

$$(\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{W})^T = \mathbf{P}_W^T = \mathbf{P}_\mathbf{W}\mathbf{P}_\mathbf{X} = \mathbf{P}_\mathbf{W}$$
$$\Rightarrow (\mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{W})^2 = \mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{W} - \mathbf{P}_\mathbf{W} + \mathbf{P}_\mathbf{W} = \mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{W}.$$

Second, for any vector $\mathbf{u}$ can be decomposed as $\mathbf{u} = \mathbf{X}\mathbf{s} + \mathbf{t}$ where $\mathbf{t} \in \mathcal{N}(\mathbf{X}^T)$. So

$$
\begin{aligned}
(\mathbf{P_X} - \mathbf{P_W})\mathbf{U} &= (\mathbf{P_X} - \mathbf{P_W})(\mathbf{Xs} + \mathbf{t}) \\
&= \mathbf{P_X Xs} - \mathbf{P_W Xs} + \mathbf{P_X t} - \mathbf{P_W t} \\
&= \mathbf{Xs} - \mathbf{P_W Xs} + \mathbf{0} - \mathbf{P_W P_X t} \\
&= \mathbf{Xs} - \mathbf{P_W Xs} + \mathbf{0} - \mathbf{0} \\
&= (\mathbf{I} - \mathbf{P_W})\mathbf{Xs}.
\end{aligned}
$$

So for any $\mathbf{u}$, $(\mathbf{P_X} - \mathbf{P_W})\mathbf{u} \in \mathcal{C}((\mathbf{I} - \mathbf{P_W})\mathbf{X})$.

Third, if $\mathbf{y} \in \mathcal{C}((\mathbf{I} - \mathbf{P_W})\mathbf{X})$, then $\mathbf{y} = (\mathbf{I} - \mathbf{P_W})\mathbf{Xc}$ for some $\mathbf{c}$.

$$
\begin{aligned}
(\mathbf{P_X} - \mathbf{P_W})\mathbf{y} &= (\mathbf{P_X} - \mathbf{P_W})(\mathbf{I} - \mathbf{P_W})\mathbf{Xc} \\
&= (\mathbf{P_X} - \mathbf{P_W})\mathbf{Xc} - (\mathbf{P_X} - \mathbf{P_W})\mathbf{P_W Xc} \\
&= \mathbf{Xc} - \mathbf{P_W Xc} - \mathbf{P_X P_W Xc} + \mathbf{P_W Xc} \\
&= (\mathbf{I} - \mathbf{P_W})\mathbf{Xc} = \mathbf{y}
\end{aligned}
$$

2.3 Reparametrization

One-way ANOVA:

- $y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, 2, 3$
- $y_{ij} = d_i + e_{ij}$
- $y_{ij} = c_1 + c_2 \mathbb{1}_{\{1\}}(i) + c_3 \mathbb{1}_{\{2\}}(i) + e_{ij}$

These three versions of the ANOVA model should lead ti equivalent inferences.

Def: Two L.M.s, $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, where $\mathbf{X}$ is $N \times p$, and $\mathbf{y} = \mathbf{Wc} + \mathbf{e}$, where $\mathbf{W}$ is $N \times t$, are *equivalent* or *reparametrizations of each other*, if the two design matrices, $\mathbf{X}$ and $\mathbf{W}$, have the same column sapce, i.e., $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{W})$.

Back to ANOVA examples:

$$
\mathbf{Xb} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \mathbf{0} \\ \mathbf{1}_{n_3} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{n_3} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}
$$

$$
\mathbf{Wc} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} \\ \mathbf{1}_{n_3} & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}
$$

$\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{W})$: Note that in $\mathbf{X}$, column 4=column1-column2-column3.

Result 2.8: If two L.M.s are equivalent, then $\mathbf{P_X} = \mathbf{P_W}$.

Corollary 2.4: If two L.M.s are equivalent, then $\hat{\mathbf{y}}$ and $\hat{\mathbf{e}}$ are the same for the two models.

proof: $\hat{\mathbf{y}} = \mathbf{P_X y}$, $\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{P_X})\mathbf{y}$.

Assume $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{W})$. Result A.2 implies the existence of two matrices, $\mathbf{S}$ and $\mathbf{T}$, such that $\mathbf{W} = \mathbf{XT}$ and $\mathbf{X} = \mathbf{WS}$.

Result 2.9: Assume $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{W})$. If $\hat{\mathbf{c}}$ solves the N.E.s $\mathbf{W}^T\mathbf{W}\mathbf{c} = \mathbf{W}^T\mathbf{y}$, then $\hat{\mathbf{b}} = \mathbf{T}\hat{\mathbf{c}}$ solves the N.E.s $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$.

proof:

$$\mathbf{X}^T\mathbf{X}\mathbf{T}\hat{\mathbf{c}} = \mathbf{X}^T\mathbf{W}\hat{\mathbf{c}} = \mathbf{X}^T\mathbf{P}_\mathbf{W}\mathbf{y} = \mathbf{X}^T\mathbf{P}_\mathbf{X}\mathbf{y} = \mathbf{X}^T\mathbf{y}.$$

Example: $y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, 2, 3.$

$$\hat{\mathbf{b}} = \begin{pmatrix} \bar{y}_{\cdot\cdot} \\ \bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot} \\ \bar{y}_{2\cdot} - \bar{y}_{\cdot\cdot} \\ \bar{y}_{3\cdot} - \bar{y}_{\cdot\cdot} \end{pmatrix}$$

General solution:

$$\begin{pmatrix} 0 \\ \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{3\cdot} \end{pmatrix} + z \begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \end{pmatrix}, \quad \text{for any } z \in \mathbb{R}$$

$y_{ij} = c_1 + c_2\mathbf{I}_{\{1\}}(i) + c_3\mathbf{I}_{\{2\}}(i) + e_{ij}, i = 1, 2, 3$

$$\hat{c} = \begin{pmatrix} \bar{y}_{3\cdot} \\ \bar{y}_{1\cdot} - \bar{y}_{2\cdot} \\ \bar{y}_{2\cdot} - \bar{y}_{3\cdot} \end{pmatrix}$$

For $\mathbf{W}$ and $\mathbf{X}$ presented before, we have $\mathbf{W} = \mathbf{X}\mathbf{T}$, where $\mathbf{T}$ is give by

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\hat{\mathbf{b}} = \mathbf{T}\hat{\mathbf{c}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \bar{y}_{3\cdot} \\ \bar{y}_{1\cdot} - \bar{y}_{2\cdot} \\ \bar{y}_{2\cdot} - \bar{y}_{3\cdot} \end{pmatrix} = \begin{pmatrix} \bar{y}_{3\cdot} \\ \bar{y}_{1\cdot} - \bar{y}_{2\cdot} \\ \bar{y}_{2\cdot} - \bar{y}_{3\cdot} \\ 0 \end{pmatrix}$$

2.4 A Version of the QR Decomposition

Result: Let $\mathbf{W}_{N\times p}$ have full column rank. Then $\mathbf{W}$ can be decomposed as $\mathbf{W} = \mathbf{Q}\mathbf{R}$ where $\mathbf{Q}_{N\times p}$ is s.t. $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_p$ and $\mathbf{R}_{p\times p}$ is upper triangular, with positive elements on the diagonal.

Idea of the proof: Write the columns of $\mathbf{W}$ as $\mathbf{W}_{\cdot 1}, \mathbf{W}_{\cdot 2}, \cdots, \mathbf{W}_{\cdot p}$. We begin by constructing a set of orthogonal vectors $\mathbf{U}_{\cdot 1}, \mathbf{U}_{\cdot 2}, \cdots, \mathbf{U}_{\cdot p}$, each of which is a linear combination of the columns of $\mathbf{W}$, and such that for each $k = 1, 2, \cdots, p$, we have $span\{\mathbf{W}_{\cdot 1}, \mathbf{W}_{\cdot 2}, \cdots, \mathbf{W}_{\cdot k}\} = span\{\mathbf{U}_{\cdot 1}, \mathbf{U}_{\cdot 2}, \cdots, \mathbf{U}_{\cdot k}\}$.

Taking $k = p$ shows that $\mathcal{C}(\mathbf{W}) = \mathcal{C}(\mathbf{U})$, where $\mathbf{U}$ is the matrix whose columns are $\mathbf{U}_{\cdot 1}, \mathbf{U}_{\cdot 2}, \cdots, \mathbf{U}_{\cdot p}$.

Since the $\mathbf{U}_{\cdot i}$'s are orthogonal, we have $\mathbf{U}^T\mathbf{U} = \mathbf{D}$ where $\mathbf{D}$ is a diagonal matrix with k-th diagonal element equal to $\|\mathbf{U}_{\cdot k}\|^2$.

So, if we take $\mathbf{Q} = \mathbf{U}\mathbf{D}^{-\frac{1}{2}}$, then $\mathbf{Q}^T\mathbf{Q} = \mathbf{D}^{-1/2}\mathbf{U}^T\mathbf{U}\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{D}^{-1/2} = \mathbf{I}$.

proof: Because we construct $\mathbf{U}_{\cdot 1}, \mathbf{U}_{\cdot 2}, \cdots, \mathbf{U}_{\cdot p}$, Recall basic facts about L.S. theory:

$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$, $\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y}$, $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}$

$$\mathbf{X}^T\hat{\mathbf{e}} = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\hat{\mathbf{b}} = \mathbf{0}$$
$$\Rightarrow \hat{\mathbf{e}} \text{ is orthogonal to every column of } \mathbf{X}.$$

Step 1: $\mathbf{U}_{\cdot 1} = \mathbf{W}_{\cdot 1}$

Step 2: Regress $\mathbf{W}_{\cdot 2}$ on $\mathbf{U}_{\cdot 1}$; that is, fit the model $\mathbf{W}_{\cdot 2} = \mathbf{U}_{\cdot 1}b + \mathbf{e}$. Then

$$\hat{b}^{(2)} = \frac{\mathbf{U}_{\cdot 1}^T\mathbf{W}_{\cdot 2}}{\mathbf{U}_{\cdot 1}^T\mathbf{U}_{\cdot 1}}.$$

Now set $\mathbf{U}_{\cdot 2} = \mathbf{W}_{\cdot 2} - \mathbf{U}_{\cdot 1}\hat{b}^{(2)}$ (Essentially $\mathbf{U}_{\cdot 2}$ is $\hat{\mathbf{e}}$). So we have $\mathbf{U}_{\cdot 1} \perp \mathbf{U}_{\cdot 2}$.

Step 3: Regress $\mathbf{W}_{\cdot 3}$ on $\mathbf{U}_{\cdot 1}$ and $\mathbf{U}_{\cdot 2}$; that is, fit the model $\mathbf{W}_{\cdot 3} = (\mathbf{U}_{\cdot 1}\ \mathbf{U}_{\cdot 2})\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \mathbf{e}$. Then

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} \mathbf{U}_{\cdot 1}^T \\ \mathbf{U}_{\cdot 2}^T \end{pmatrix}(\mathbf{U}_{\cdot 1}\ \mathbf{U}_{\cdot 2}) = \begin{pmatrix} \|\mathbf{U}_{\cdot 1}\|^2 & 0 \\ 0 & \|\mathbf{U}_{\cdot 2}\|^2 \end{pmatrix}$$

and it is easy to solve N.E. $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$ directly. For $j = 1, 2$:

$$\left(\hat{\mathbf{b}}^{(3)}\right)_j = \frac{\mathbf{U}_{\cdot j}^T\mathbf{W}_{\cdot 3}}{\mathbf{U}_{\cdot j}^T\mathbf{U}_{\cdot j}}.$$

Now set $\mathbf{U}_{\cdot 3} = \mathbf{W}_{\cdot 3} - (\mathbf{U}_{\cdot 1}\ \mathbf{U}_{\cdot 2})\hat{\mathbf{b}}^{(3)} \Rightarrow \mathbf{U}_{\cdot 3} \perp \mathbf{U}_{\cdot j} (j = 1, 2)$.

Now for $i \in \{3, 4, 5, \cdots, p-1\}$, at step $p+1$, we regress $\mathbf{W}_{i+1}$ on $\mathbf{U}_{\cdot 1}, \mathbf{U}_{\cdot 2}, \cdots, \mathbf{U}_{\cdot i}$. As above, for $j = 1, 2, c \ldots, i$,

$$\left(\hat{\mathbf{b}}^{(i+1)}\right)_j = \frac{\mathbf{U}_{\cdot j}^T\mathbf{W}_{\cdot i+1}}{\mathbf{U}_{\cdot j}^T\mathbf{U}_{\cdot j}}.$$

Now set $U_{\cdot i+1} = \mathbf{W}_{\cdot i+1} - (\mathbf{U}_{\cdot 1}\ \mathbf{U}_{\cdot 2} \cdots \mathbf{U}_{\cdot i})\hat{\mathbf{b}}^{(i+1)} \Rightarrow \mathbf{U}_{\cdot i+1} \perp \mathbf{U}_{\cdot j} (j = 1, 2, \cdots, i)$.

Now collect all of the $\hat{\mathbf{b}}$'s in a $p \times p$ matrix as follows:

$$\mathbf{S} = \begin{pmatrix} 1 & (\hat{\mathbf{b}^{(2)}})_1 & (\hat{\mathbf{b}}^{(3)})_1 & \cdots & (\hat{\mathbf{b}}^{(p)})_1 \\ 0 & 1 & (\hat{\mathbf{b}}^{(3)})_2 & \cdots & (\hat{\mathbf{b}}^{(p)})_2 \\ 0 & 0 & 1 & \cdots & (\hat{\mathbf{b}}^{(p)})_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & (\hat{\mathbf{b}}^{(p)})_{p-1} \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

More formally, $\mathbf{S}_{j,i+1} = \left(\hat{\mathbf{b}}^{(i+1)}\right)_j$ for $i = 1, 2, \cdots, p-1$ and $j = 1, 2, \cdots, i$. Diagonal elements are all 1, and all 0's below diagonal.'

Now for $i = 1, 2, \cdots, p-1$, we have

$$\mathbf{U}_{\cdot i+1} = \mathbf{W}_{\cdot i+1} - \sum_{j=1}^{i} \left(\hat{\mathbf{b}}^{(i+1)}\right) \mathbf{U}_{\cdot j} = \mathbf{W}_{\cdot i+1} - \sum_{j=1}^{i} \mathbf{S}_{j,i+1} \mathbf{U}_{\cdot j}$$

$$\Rightarrow \mathbf{W}_{\cdot i+1} = \mathbf{U}_{\cdot i+1} + \sum_{j=1}^{i} \mathbf{S}_{j,i+1} \mathbf{U}_{\cdot j} = \sum_{j=1}^{i+1} \mathbf{S}_{j,i+1} \mathbf{U}_{\cdot j}$$

$$\Rightarrow \mathbf{W}_{N \times p} = \mathbf{U}_{N \times p} \mathbf{S}_{p \times p}$$

Finally, we take $\mathbf{Q} = \mathbf{U}\mathbf{D}^{-\frac{1}{2}}$ and $\mathbf{R} = \mathbf{D}^{\frac{1}{2}}\mathbf{S}$. Then

$$\mathbf{Q}\mathbf{R} = \mathbf{U}\mathbf{D}^{-\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}\mathbf{S} = \mathbf{W}$$

Chapter 3: Estimability and Least Squares Estimators

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

We assume $\mathbb{E}\mathbf{e} = 0$, so $\mathbb{E}\mathbf{y} = \mathbf{X}\mathbf{b}$.

Fix $\boldsymbol{\lambda} \in \mathbb{R}^p$

Def 3.1: An estimator $t(\mathbf{y})$ is *unbiased* for $\boldsymbol{\lambda}^T\mathbf{b}$ if $\mathbf{E}t(\mathbf{y}) = \boldsymbol{\lambda}^T\mathbf{b}$, $\forall \mathbf{b} \in \mathbb{R}^p$.

Def 3.2: An estimator $t(\mathbf{y})$ is a *linear estimator* if $t(\mathbf{y}) = c + \mathbf{a}^T\mathbf{y}$ where $c \in \mathbb{R}$ and $\mathbf{a} \in \mathbb{R}^p$ are known constants.

Def 3.3: A function $\boldsymbol{\lambda}^T\mathbf{b}$ is *(linearly) estimable* if $\exists$ a linear unbiased estimator of it; otherwise $\boldsymbol{\lambda}^T\mathbf{b}$ is called *non-estimable*.

Result 3.1: $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable *iff* $\exists \mathbf{a} \in \mathbb{R}^N$ such that $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$, or, in other words, $\boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X}^T)$.

proof: ($\Leftarrow$) Assume there exists $\mathbf{a} \in \mathbb{R}^N$ such that $\mathbf{X}^T\mathbf{a} = \boldsymbol{\lambda}$. Then $\mathbb{E}(\mathbf{a}^T\mathbf{y}) = \mathbf{a}^T\mathbb{E}(\mathbf{y}) = \mathbf{a}^T\mathbf{X}\mathbf{b} = \boldsymbol{\lambda}^T\mathbf{b}$.

($\Rightarrow$) Assume that $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable. Then $\exists c, \mathbf{a}$ such that $\mathbb{E}(c + \mathbf{a}^T\mathbf{y}) = \boldsymbol{\lambda}^T\mathbf{b}$, $\forall \mathbf{b} \in \mathbb{R}^p$. Then, $c + \mathbf{a}^T\mathbf{X}\mathbf{b} - \boldsymbol{\lambda}^T\mathbf{b} = 0$, $\forall \mathbf{b} \in \mathbb{R}^p \Rightarrow c + (\mathbf{a}^T\mathbf{X} - \boldsymbol{\lambda}^T)\mathbf{b} = 0$, $\forall \mathbf{b} \in \mathbb{R}^p$. Using Result A.8, we have $c = 0$ and $\mathbf{a}^T\mathbf{X} - \boldsymbol{\lambda}^T = \mathbf{0}$ or $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$.

Example 3.1: One-way ANOVA with $a = n = 2$. $y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, 2, j = 1, 2$. $\mathbf{Xb} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix}$.

Q: Is α_1 estimable?

A: First, $\alpha_1 = \boldsymbol{\lambda}^T\mathbf{b} = (0, 1, 0)(\mu, \alpha_1, \alpha_2)^T$. Is $(0, 1, 0)^T \in \mathcal{C}(\mathbf{X}^T)$?

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\Rightarrow \sum_{i=1}^{4} a_i = 0, \quad a_1 + a_2 = 1, \quad a_3 + a_4 = 0$$

This is impossible. So $(0, 1, 0)^T \notin \mathcal{C}(\mathbf{X}^T) \Rightarrow \alpha_1$ is not estimable.

Result 3.1a: $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable if and only if $\boldsymbol{\lambda}^T\mathbf{b}$ is a linear combination of the expected values of the y_i's.

proof: ($\Rightarrow$) Assume that $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable. Then $\exists\ a \in \mathbb{R}^N$ such that $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$. Then $\boldsymbol{\lambda}^T\mathbf{b} = \mathbf{a}^T\mathbf{Xb} = \mathbf{a}^T\mathbb{E}(\mathbf{y})$.

($\Leftarrow$) Assume that $\boldsymbol{\lambda}^T\mathbf{b}$ is a linear combination of the elements of $\mathbf{Xb}$. Then, for all $\mathbf{b} \in \mathbb{R}^p$,
$\boldsymbol{\lambda}^T\mathbf{b} = \mathbf{a}^T\mathbf{Xb} \Rightarrow \boldsymbol{\lambda}^T = \mathbf{a}^T\mathbf{X}$ or $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$, i.e., $\boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X}^T)$. So $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable.

*If we take $\mathbf{a}^T = (\mathbf{e}^{(i)})^T$, then $\mathbf{a}^T\mathbb{E}\mathbf{y} = \mathbb{E}y_i$. So the expected value of any y_i is estimable.

Suppose $\boldsymbol{\lambda}^{(j)} \in \mathbb{R}^p, j = 1, 2, \cdots, k$. If $\boldsymbol{\lambda}^{(j)T}\mathbf{b}$ are all estimable, so is $\sum_{j=1}^{k} d_j\boldsymbol{\lambda}^{(j)T}\mathbf{b} = \left[\sum_{j=1}^{k} d_j\boldsymbol{\lambda}^{(j)}\right]^T\mathbf{b}$ for any $\{d_j\}_{j=1}^{k}$.

Def: A set of estimable functions $\{\boldsymbol{\lambda}^{(j)}\mathbf{b}\}_{j=1}^{k}$ is called *linearly independent* if the $\boldsymbol{\lambda}^{(j)}$'s are linearly independent.

Let $r = rank(\mathbf{X}^T) \leq p$ and $\dim(\mathcal{C}(\mathbf{X}^T)) = rank(X^T) = r$. Thus, any set of L.I. estimable functions cannot have more than r members.

Suppose that $\mathbf{X}$ has full column rank, $r = p$. Then $\mathcal{C}(\mathbf{X}^T) = \mathbb{R}^p$ and $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable for any $\boldsymbol{\lambda} \in \mathbb{R}^p$.

Q: How do we establish that $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable?

A: We have three methods.

Method 3.1: Show that $\boldsymbol{\lambda}^T\mathbf{b}$ can be written as a linear combination of the elements of $\mathbf{Xb} = \mathbb{E}(\mathbf{y})$.

Method 3.2: Construct a set of basis vectors for $\mathcal{C}(\mathbf{X}^T)$, call them $\mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(r)}$ and show that $\boldsymbol{\lambda} = \sum_{j=1}^{r} d_j\mathbf{v}^{(j)}$.

Method 3.3: Construct a set of basis vectors for $\mathcal{N}(\mathbf{X})$, call them $\mathbf{c}^{(1)}, \cdots, \mathbf{c}^{(p-r)}$. If $\boldsymbol{\lambda} \perp \mathbf{c}^{(j)}$ for $j = 1, 2, \cdots, p-r$, then $\boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X}^T)$.

Note that $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g$ is a projection onto $\mathcal{C}(\mathbf{X}^T\mathbf{X}) = \mathcal{C}(\mathbf{X}^T)$. So if $(\mathbf{X}^T\mathbf{X})^g$ has been calculated, and $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\lambda} = \boldsymbol{\lambda}$. Then $\boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X}^T) \Rightarrow \boldsymbol{\lambda}^T\mathbf{b}$ is estimable.

Examples:

- One-way ANOVA: $y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, \cdots, a, j = 1, \cdots, n_i$.

  $\mathbf{y} = (y_{11}, \cdots, y_{1n_1}, y_{21}, \cdots, y_{2n_2}, \cdots, y_{a1}, \cdots, y_{an_a})^T, \mathbf{Xb} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_a} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_a} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \end{pmatrix}, p = a + 1,$

  $N = \sum_{i=1}^a n_i, rank(\mathbf{X}) = r = a < p.$

  $\dim(\mathcal{C}(\mathbf{X}^T)) + \dim(\mathcal{N}(\mathbf{X})) = a + 1 \Rightarrow \dim(\mathcal{N}(\mathbf{X})) = 1$. To find a basis fo $\mathcal{N}(\mathbf{X})$, all we need is to find a single vector, $\mathbf{c}$, such that $\mathbf{Xc} = \mathbf{0}$. Note that $\mathbf{c} = \begin{pmatrix} 1 \\ -\mathbf{1}_a \end{pmatrix}$ is a basis for $\mathcal{N}(\mathbf{X})$. Now let's use this basis to figure out which functions $\boldsymbol{\lambda}^T\mathbf{b} = \lambda_0\mu + \sum_{i=1}^a \lambda_i\alpha_i$ are estimable. According to Method 3.3, $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable *iff* $\boldsymbol{\lambda}^T\mathbf{c} = 0$. Hence, $\mu + \alpha_i$ and $\alpha_i - \alpha_k (i \neq k)$ are estimable, but α_i is not estimable. Note that $\sum_{i=1}^a d_i\alpha_i$ is estimable iff $\sum_{i=1}^a d_i = 0$. Of course, when $\sum_{i=1}^a d_i = 0$, $\sum_{i=1}^a d_i\alpha_i$ is called a contrast.

Def 3.4: The *least sqaures (L.S.) estimator of the estimable function* $\boldsymbol{\lambda}^T\mathbf{b}$ is $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$ when $\hat{\mathbf{b}}$ is any solution to the N.E.s.

Result 3.2: $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$ is invariant to the choice of a solution to the N.E.s.

proof: Suppose that $\hat{\mathbf{b}}_1$ and $\hat{\mathbf{b}}_2$ are both solutions to the N.E.s. Then $\mathbf{X}^T\mathbf{X}(\hat{\mathbf{b}}_1 - \hat{\mathbf{b}}_2) = \mathbf{0}$. In other words, $\hat{\mathbf{b}}_1 - \hat{\mathbf{b}}_2 \in \mathcal{N}(\mathbf{X}^T\mathbf{X}) = \mathcal{N}(\mathbf{X})$. But $\boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X}^T) \perp \mathcal{N}(\mathbf{X})$. So $\boldsymbol{\lambda} \perp (\hat{\mathbf{b}}_1 - \hat{\mathbf{b}}_2)$, or $\boldsymbol{\lambda}^T(\hat{\mathbf{b}}_1 - \hat{\mathbf{b}}_2) = 0$, or $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_1 = \boldsymbol{\lambda}^T\hat{\mathbf{b}}_2$.

Result 3.3: The L.S. estimator $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$ of an estimable function, $\boldsymbol{\lambda}^T\mathbf{b}$, is a linear unbiased estimator of $\boldsymbol{\lambda}^T\mathbf{b}$.

proof: $\lambda^T\hat{\mathbf{b}} = \mathbf{a}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y} = \mathbf{a}^T\mathbf{P}_\mathbf{X}\mathbf{y}$.

$\mathbb{E}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) = \mathbb{E}(\mathbf{a}^T\mathbf{P}_\mathbf{X}\mathbf{y}) = \mathbf{a}^T\mathbf{P}_\mathbf{X}\mathbb{E}(\mathbf{y}) = \mathbf{a}^T\mathbf{P}_\mathbf{X}\mathbf{Xb} = \mathbf{a}^T\mathbf{Xb} = \boldsymbol{\lambda}^T\mathbf{b}$.

Example 1: Back to one-way ANOVA example: $y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, \cdots, a, j = 1, \cdots, n_i$.

Let's look at the least squares estimators of some estimable functions. Start with a solution to the N.E.s.

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} N & n_1 & n_2 & \cdots & n_a \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_a & 0 & 0 & \cdots & n_a \end{pmatrix}, \quad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} N\bar{y}_{\cdot\cdot} \\ n_1\bar{y}_{1\cdot} \\ n_2\bar{y}_{2\cdot} \\ \vdots \\ n_a\bar{y}_{a\cdot} \end{pmatrix}$$

A simple G.I. of $\mathbf{X}^T\mathbf{X}$ is:

$$\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1/n_1 & 0 & \cdots & 0 \\ 0 & 0 & 1/n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1/n_a \end{pmatrix}$$

This leads to the following general solution to the N.E.s:

$$\hat{\mathbf{b}} = \begin{pmatrix} 0 \\ \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \vdots \\ \bar{y}_{a\cdot} \end{pmatrix} + z \begin{pmatrix} -1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{for any } z \in \mathbb{R}.$$

L.S. estimator of $\boldsymbol{\lambda}^T\mathbf{b}$ is $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$:

$$\begin{aligned} \mu + \alpha_i: &\quad \boldsymbol{\lambda}^T\hat{\mathbf{b}} = (0 - z) + (\bar{y}_{i\cdot} + z) = \bar{y}_{i\cdot} \\ \alpha_i - \alpha_k: &\quad \boldsymbol{\lambda}^T\hat{\mathbf{b}} = (\bar{y}_{i\cdot} + z) - (\bar{y}_{k\cdot} + z) = \bar{y}_{i\cdot} - \bar{y}_{k\cdot} \\ \sum_{i=1}^{a} d_i\alpha_i: &\quad \boldsymbol{\lambda}^T\hat{\mathbf{b}} = \sum_{i=1}^{a} d_i(\bar{y}_{i\cdot} + z) = \sum_{i=1}^{a} d_i\bar{y}_{i\cdot} + z\sum_{i=1}^{a} d_i = \sum_{i=1}^{a} d_i\bar{y}_{i\cdot} \end{aligned}$$

Example 2: Two-way crossed model without interaction: $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$, $i = 1, \cdots, a$, $j = 1, \cdots, b$. $N = ab$ and $p = a + b + 1$.

$$\mathbf{y} = (y_{11}, \cdots, y_{1b}, y_{21}, \cdots, y_{2b}, \cdots, y_{a1}, \cdots, y_{ab})^T$$

$$\mathbf{Xb} = \begin{pmatrix} \mathbf{1}_b & \mathbf{1}_b & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I}_b \\ \mathbf{1}_b & \mathbf{0} & \mathbf{1}_b & \cdots & \mathbf{0} & \mathbf{I}_b \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{1}_b & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_b & \mathbf{I}_b \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \\ \beta_1 \\ \vdots \\ \beta_b \end{pmatrix}$$

$rank(\mathbf{X}) = a + b - 1$. Reason: The first column is the sum of columns 2 through $a + 1$ (and also the sum of columns $a + 2$ through $a + b + 1$). The second column is the sum of columns $a + 2$ through $a + b + 1$ minus the sum of columns 3 through $a + 1$. Remaining $a + b - 1$ columns are linearly independent.

Now, $dim(\mathcal{N}(\mathbf{X})) = 2$. Here is a basis for $\mathcal{N}(\mathbf{X})$:

$$\mathbf{c}^{(1)} = \begin{pmatrix} 1 \\ -\mathbf{1}_a \\ \mathbf{0}_b \end{pmatrix}, \quad \mathbf{c}^{(2)} = \begin{pmatrix} 1 \\ \mathbf{0}_a \\ -\mathbf{1}_b \end{pmatrix}$$

Now let's write $\boldsymbol{\lambda}^T\mathbf{b} = \lambda_0\mu + \sum_{i=1}^{a} \lambda_i\alpha_i + \sum_{j=1}^{b} \lambda_{a+j}\beta_j$.

Invoking Method 3.3: $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable *iff* $\lambda_0 - \sum_{i=1}^{a} \lambda_i = 0$ and $\lambda_0 - \sum_{j=1}^{b} \lambda_{a+j} = 0$.

Common estimable functions:

- $\mu + \alpha_i + \beta_j$
- $\alpha_i - \alpha_k$
- $\beta_j - \beta_k$
- $\sum_{i=1}^{a} d_i \alpha_i$ (fix $\sum_{i=1}^{a} d_i = 0$)
- $\sum_{j=1}^{b} f_j \beta_j$ (fixed $\sum_{j=1}^{d} f_j = 0$)

Suppose that $a = 3$ and $b = 4$. $rank(\mathbf{X}) = 6$. Here is a set of linearly independent estimable functions:

$$\{\mu + \alpha_1 + \beta_1, \alpha_2 - \alpha_1, \alpha_3 - \alpha_1, \beta_2 - \beta_1, \beta_3 - \beta_1, \beta_4 - \beta_1\}$$

Now, consider what happends when there is some missing data.

Eg 1:

	1	2	3	4
1	X	X	X	
2	X		X	
3			X	X

*column - B, row - A

So we do not observe $y_{14}, y_{22}, y_{24}, y_{31}, y_{32}$.

$$\mathbf{Xb} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

$rank(\mathbf{X}) = 6$. And $\mathcal{C}(\mathbf{X}^T)$ remains the same, so same set of 6 linearly independent estimable functions is srill valid.

Eg 2:

	1	2	3	4
1	X	X	X	
2	X	X	X	
3				X

*column - B, row - A

In this case, $rank(\mathbf{X}) = 5$. Basis for $\mathcal{N}(\mathbf{X})$:

$$\begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}, \text{ and } \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}$$

Here is a set of 5 linearly independent estimable functions:

$$\{\mu + \alpha_1 + \beta_1, \alpha_2 - \alpha_1, \beta_2 - \beta_1, \beta_3 - \beta_1, \mu + \alpha_3 + \beta_4\}.$$

Note that $\alpha_2 - \alpha_1$ and $\beta_4 - \beta_1$ are no longer estimable in this case. There is a simple reason for this:

$$\mathbb{E}y_{34} = \mu + \alpha_3 + \beta_4$$

This is the only component of $\mathbf{Xb}$ that involves either α_3 or β_4. We cannot differentiate between α_3 and β_4.

Back to Eg 1:

$\mathbb{E}(y_{33}) - \mathbb{E}(y_{13}) = \mu + \alpha_3 + \beta_3 - (\mu + \alpha_1 + \beta_3) = \alpha_3 - \alpha_1.$

$\mathbb{E}(y_{34}) - \mathbb{E}(y_{33}) + \mathbb{E}(y_{13}) - \mathbb{E}(y_{11}) = (\mu + \alpha_3 + \beta_4) - (\mu + \alpha_3 + \beta_3) + (\mu + \alpha_1 + \beta_3) - (\mu + \alpha_1 + \beta_1) = \beta_4 - \beta_1$
.

The pattern is connected in Eg1, but not in Eg2.

3.7 Reparameterization Revisited

- $\mathbf{y} = \mathbf{Xb} + \mathbf{e}\,(\mathbf{X}_{N\times p})$
- $\mathbf{y} = \mathbf{Wc} + \mathbf{e}\,(\mathbf{W}_{N\times t})$

Assume the models are equivalent, so $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{W})$. Then $\exists\ \mathbf{T}, \mathbf{S}$, such that $\mathbf{W} = \mathbf{XT}$ and $\mathbf{X} = \mathbf{WS}$.

Result 3.4: If $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable in the $\mathbf{X}$ model, and $\hat{\mathbf{c}}$ solves the N.E.s $\mathbf{W}^T\mathbf{Wc} = \mathbf{W}^T\mathbf{y}$, then $\boldsymbol{\lambda}^T\mathbf{T}\hat{\mathbf{c}}$ is the L.S. estimator of $\boldsymbol{\lambda}^T\mathbf{b}$.

proof: If $\mathbf{W}^T\mathbf{W}\hat{\mathbf{c}} = \mathbf{W}^T\mathbf{y}$, then by Result 2.9, we know that $\mathbf{X}^T\mathbf{XT}\hat{\mathbf{c}} = \mathbf{X}^T\mathbf{y}$, so the L.S. estimator of $\boldsymbol{\lambda}^T\mathbf{b}$ is $\boldsymbol{\lambda}^T\mathbf{T}\hat{\mathbf{c}}$.

Result 3.5: If $\mathbf{q}^T\mathbf{c}$ is estimable to the $\mathbf{W}$ model (so $\mathbf{q} \in \mathcal{C}(\mathbf{W}^T)$), then $\mathbf{q}^T\mathbf{Sb}$ is estimable in the $\mathbf{X}$ model, and the L.S. estimator is $\mathbf{q}^T\hat{\mathbf{c}}$ where $\mathbf{W}^T\mathbf{W}\hat{\mathbf{c}} = \mathbf{W}^T\mathbf{y}$.

proof: $\mathbf{q} \in \mathcal{C}(\mathbf{W}^T) \Rightarrow \mathbf{q} = \mathbf{W}^T\mathbf{a}$ for some $\mathbf{a}$. So $\mathbf{S}^T\mathbf{q} = \mathbf{S}^T\mathbf{W}^T\mathbf{a} = \mathbf{X}^T\mathbf{a} \in \mathcal{C}(\mathbf{X})$. So $\mathbf{q}^T\mathbf{Sb}$ is estimable in the $\mathbf{X}$ model. Now, the estimator is

$$\mathbf{q}^T\mathbf{S}\hat{\mathbf{b}} = \mathbf{q}^T\mathbf{ST}\hat{\mathbf{c}} = \mathbf{a}^T\mathbf{WST}\hat{\mathbf{c}} = \mathbf{a}^T\mathbf{XT}\hat{\mathbf{c}} = \mathbf{a}^T\mathbf{W}\hat{\mathbf{c}} = \mathbf{q}^T\hat{\mathbf{c}}.$$

*Note: The converse of Result 3.5 does not hold. That is, $\mathbf{S}^T\mathbf{q} \in \mathcal{C}(\mathbf{X}^T) \not\Rightarrow \mathbf{q} \in \mathcal{C}(\mathbf{W}^T)$. Indeed, $\mathbf{S}^T\mathbf{q} = \mathbf{X}^T\mathbf{a} = \mathbf{S}^T\mathbf{W}^T\mathbf{a}$. But this imples that $\mathbf{q} = \mathbf{W}^T\mathbf{a}$ only if $\mathbf{S}^T$ has full column rank (See corollary A.2).

The most common goal of reparameterization is to find a full-rank version of an overparameterized model. A full-rank version **always** exists, one simply has to delete columns of the design matrix that are linear combinations of the columns until there are no more such columns.

Recall that $\mathbf{X}_{N\times p} = \mathbf{W}_{N\times t}\mathbf{S}_{t\times p}$. If $\mathbf{W}$ has full column rank:

1. $\mathbf{W}^T\mathbf{W}$ is non-singular, and there is a unique solution to the $\mathbf{W}$ N.E.s
2. The matrix $\mathbf{S}$ is unique, so $\mathbf{Xb} = \mathbf{Wc} \Rightarrow \mathbf{WSb} = \mathbf{Wc} \Rightarrow \mathbf{Sb} = \mathbf{c}$. So, the parameter $\mathbf{c}$ has a unique representation in terms of the elements of $\mathbf{b}$.

Examples: Let's look at several different full-rank versions of the one-way ANOVA model.

- Standard model: $y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, \cdots, a, j = 1, \cdots, n_i$.

$$\mathbf{Xb} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_a} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_a} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \end{pmatrix}$$

- Cell means model: $y_{ij} = \mu_i + e_{ij}$

$$\mathbf{Z}\boldsymbol{\mu} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_a} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_a \end{pmatrix}$$

  $\mathbf{X}_{N\times(a+1)} = \mathbf{Z}_{N\times a}\mathbf{S}_{a\times(a+1)}$ where $\mathbf{S} = [\mathbf{1}_a\ \mathbf{I}_a]$.

  Let $\boldsymbol{\mu}^* = (\mu_1, \mu_2, \cdots, \mu_a)^T$. Then $\boldsymbol{\mu}^* = \mathbf{Sb} = [\mathbf{1}_a\ \mathbf{I}_a](\mu, \alpha_1, \cdots, \alpha_a)^T = (\mu + \alpha_1, \mu + \alpha_2, \cdots, \mu + \alpha_a)^T$.

- Cell reference model: $y_{ij} = c_1 + c_2\mathbf{I}_{\{1\}}(i) + c_3\mathbf{I}_{\{2\}}(i) + \cdots + c_a\mathbf{I}_{\{a-1\}}(i) + e_{ij}$

$$\mathbf{Wc} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_{a-1}} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_{a-1}} \\ \mathbf{1}_{n_a} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_a \end{pmatrix}$$

  $\mathbf{X}_{N\times(a+1)} = \mathbf{W}_{N\times a}\mathbf{S}_{a\times(a+1)}$ where $\mathbf{S} = [\mathbf{I}_a\ (1, -1, -1, \cdots, -1)^T]$.

  Let $\mathbf{c} = (c_1, c_2, \cdots, c_a)^T$. Then $\mathbf{c} = \mathbf{Sb} = [\mathbf{I}_a\ (1, -1, -1, \cdots, -1)^T] \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \end{pmatrix} = \begin{pmatrix} \mu + \alpha_a \\ \alpha_1 - \alpha_a \\ \vdots \\ \alpha_{a-1} - \alpha_a \end{pmatrix}$.

  Now, let's look at the relationship between $\mathbf{c}$ and $\boldsymbol{\mu}^*$:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{a-1} \\ \mu_a \end{pmatrix} = \begin{pmatrix} c_1 + c_2 \\ c_1 + c_3 \\ \vdots \\ c_1 + c_a \\ c_1 \end{pmatrix} \Rightarrow \begin{cases} c_1 = \mu_a \\ c_2 = \mu_1 - \mu_a \\ c_3 = \mu_3 - \mu_a \\ \vdots \\ c_a = \mu_{a-1} - \mu_a \end{cases}$$

3.8 Imposing Conditions on $\mathbf{b}$ to Get a Unique Solution to the Normal Equations

Example: One-way ANOVA: $y_{ij} = \mu + \alpha_i + e_{ij}$

N.E.s: $N\mu + \sum_{i=1}^{a} n_a \alpha_i = N\bar{y}_{\cdot\cdot},\ \mu + \alpha_i = \bar{y}_{i\cdot}, i = 1, 2, \cdots, a$

Here are those standard conditions on $\mathbf{b} = (\mu, \alpha_1, \cdots, \alpha_a)^T$ that are impossed to get a unqiue solution:

- $\alpha_1 = 0 \Rightarrow \hat{\mu} = \bar{y}_{a\cdot}, \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{a\cdot}, i = 1, 2, \cdots, a-1, \hat{\alpha}_a = 0$
- $\sum_{i=1}^{a} \alpha_i = 0 \Rightarrow \hat{\mu} = \frac{1}{a}\sum_{i=1}^{a} \bar{y}_{i\cdot}, \hat{\alpha}_i = \bar{y}_{i\cdot} - \hat{\mu}, i = 1, \cdots, a$
- $\sum_{i=1}^{a} n_i \alpha_i = 0 \Rightarrow \hat{\mu} = \bar{y}_{\cdot\cdot}, \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}, i = 1, \cdots, a$

Q: Can we always impose a set of conditions on $\mathbf{b}$ to obtain a unique solution to the N.E.s?

A:Yes.

Define $rank(X) = r \leq p$.

The conditions on $\mathbf{b}$ will be specified as follows:

$$\mathbf{C}\mathbf{b} = \mathbf{0}$$

where $\mathbf{C} \in \mathbb{R}^{(p-r)\times p}$, $\mathbf{b} \in \mathbb{R}^{p\times 1}$, $\mathbf{0} \in \mathbb{R}^{(p-r)\times 1}$, and $rank(\mathbf{C}) = p - r$.

Adding these conditions to the N.E.s yields

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} \\ \mathbf{C} \end{pmatrix}\mathbf{b} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{0} \end{pmatrix} \qquad (1)$$

or equivalently,

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{C} \end{pmatrix}\mathbf{b} = \begin{pmatrix} \mathbf{P}_{\mathbf{X}}\mathbf{y} \\ \mathbf{0} \end{pmatrix} \qquad (2)$$

($\hat{\mathbf{b}}$ solves N.E.s $\Leftrightarrow$ $\hat{\mathbf{b}}$ solves $\mathbf{X}\mathbf{b} = \mathbf{P}_{\mathbf{X}}\mathbf{y}$)

equations (2) will have a unique solution if $rank(\begin{pmatrix} \mathbf{X} \\ \mathbf{C} \end{pmatrix}) = p$. That is, if $\mathcal{C}((\mathbf{X}^T \quad \mathbf{C}^T)) = \mathbb{R}^p$. Thus, choosing $\mathbf{C}$ is equivalent to adding L.I. rows to $\mathbf{X}$ until its rank is p. Equivalently, we can add L.I. columns to $\mathbf{X}^T$ until it has rank p. The columns of $\mathbf{C}^T$must be L.I. of one another, and of the collumns of $\mathbf{X}^T$. Thus, the columns of $\mathbf{C}^T$ **CANNOT** be in $\mathcal{C}(\mathbf{X}^T)$. In other words, the columns of $\mathbf{C}^T$ must correspond to non-estimable functions.

Suppose $\mathbf{C}$ satisfies

$$rank(\mathbf{C}) = p - r \qquad (3)$$
$$\mathcal{C}(\mathbf{X}^T) \cap \mathcal{C}(\mathbf{C}^T) = \{0\} \qquad (4)$$

Together, (3) and (4) implies

$$rank(\begin{pmatrix} \mathbf{X} \\ \mathbf{C} \end{pmatrix}) = p$$

Example: $y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, 2, 3$

$p = 4$, $r = 3$. A basis for $\mathcal{C}(\mathbf{X}^T)$ is given by

$$\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Consider the restriction $\sum_{i=1}^{3} \alpha_i = 0$. In the form $\mathbf{Cb} = \mathbf{0}$, we have

$$\begin{pmatrix} 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \mathbf{0}$$

So $\mathbf{C}^T = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}$.

Q: Is $(0, 1, 1, 1)^T$ linearly independent of the basis vectors for $\mathcal{C}(\mathbf{X}^T)$?

A: Yes. It is because $(0, 1, 1, 1)^T = a_1(1, 1, 0, 0)^T + a_2(1, 0, 1, 0)^T + a_3(1, 0, 0, 1)^T$ is impossible.

Lemma 3.1: If $\mathbf{C}$ satisfies (3) and (4), then (1) is equivalent to

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} \\ \mathbf{C}^T\mathbf{C} \end{pmatrix} \mathbf{b} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{0} \end{pmatrix} \qquad (6)$$

And, (6) is equivalent to

$$(\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})\mathbf{b} = \mathbf{X}^T\mathbf{y} \qquad (7)$$

proof: Recall from Lemma 2.1 that $\mathcal{N}(\mathbf{C}) = \mathcal{N}(\mathbf{C}^T\mathbf{C})$. Thus $\mathbf{Cb} = \mathbf{0} \Leftrightarrow \mathbf{C}^T\mathbf{Cb} = \mathbf{0}$.

Now, if (6) holds, then

$$\mathbf{X}^T\mathbf{Xb} = \mathbf{X}^T\mathbf{y} \quad \text{and} \quad \mathbf{C}^T\mathbf{Cb} = \mathbf{0}.$$

(7) follows by just adding these equations. Now assume that (7) holds, and rewrite it as follows:

$$\mathbf{C}^T\mathbf{C}\mathbf{b} = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{b}$$

The RHS is in $\mathcal{C}(\mathbf{X}^T)$, but the only vectors that $\mathcal{C}(\mathbf{X}^T)$ and $\mathcal{C}(\mathbf{C}^T)$ have in common is $\mathbf{0}$. So $\mathbf{C}^T\mathbf{C}\mathbf{b} = \mathbf{0}$ and $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$.

Result 3.6: Assume $\mathbf{C}$ satisfies (3) and (4). Then

1. $\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C}$ is non-singular
2. $(\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})^{-1}\mathbf{X}^T\mathbf{y}$ is the unique solution to (1)
3. $(\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})^{-1}$ is a G.I. of $\mathbf{X}^T\mathbf{X}$
4. $\mathbf{C}(\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})^{-1}\mathbf{X}^T = \mathbf{0}$
5. $\mathbf{C}(\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T = \mathbf{I}$

proof:

1. $\mathbb{R}^p = \mathcal{C}((\mathbf{X}^T \quad \mathbf{C}^T)) = \mathcal{C}\left((\mathbf{X}^T \quad \mathbf{C}^T)\begin{pmatrix}\mathbf{X}\\ \mathbf{C}\end{pmatrix}\right) = \mathcal{C}(\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})$

   So, $\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C}$ is a $p \times p$ matrix with rank p. So it is non-singular.
2. The unique solution of $(\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})\mathbf{b} = \mathbf{X}^T\mathbf{y}$ is $(\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})^{-1}\mathbf{X}^T\mathbf{y}$. Now, the result follows from Lemma 3.1
3. From 2, we have $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{y}$ for $\forall \mathbf{y} \in \mathbb{R}^N$ because of the first equations in (1). So $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})^{-1}\mathbf{X}^T = \mathbf{X}^T$. Postmultiply by $\mathbf{X}$:

   $$\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{X} \Rightarrow (\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})^{-1} \text{ is a G.I. of } \mathbf{X}^T\mathbf{X}$$

4. From 2, we have $\mathbf{C}(\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{0}$ for $\forall \mathbf{y} \in \mathbb{R}^N$.

3.9 Constrained Parameter Space

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

Suppose we wish to restrict $\mathbf{b}$ to a subspcae of $\mathbb{R}^p$ given by

$$\mathcal{T} = \{\mathbf{b} \in \mathbb{R}^p : \mathbf{P}^T\mathbf{b} = \delta\},$$

where we will insist that the matrix $\mathbf{P} \in \mathbb{R}^{p\times q}$ is of full-column rank.

Note: This may seem similar to Section 3.8, but here we do NOT insist that the columns of $\mathbf{P}$ correspond to non-estimable functions.

Def: $\boldsymbol{\lambda}^T\mathbf{b}$ is *estimable in the constrained model* if $\exists\, c \in \mathbb{R}$ and $\mathbf{a} \in \mathbb{R}^N$ s.t. $\mathbb{E}(c + \mathbf{a}^T\mathbf{y}) = \boldsymbol{\lambda}^T\mathbf{b}, \forall \mathbf{b} \in \mathcal{T}$.

*Note: If $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable in the unconstrained model, then it is estimable in the constrained model. So the constraint can only increase the number of estimable functions.

Result 3.7: In the constrained model, $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable *iff* $\exists\, \mathbf{a} \in \mathbb{R}^N$ and $\mathbf{d} \in \mathbb{R}^q$ s.t. $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a} + \mathbf{P}\mathbf{d}$.

proof: ($\Leftarrow$):
$\mathbb{E}(\mathbf{d}^T\delta + \mathbf{a}^T\mathbf{y}) = \mathbf{d}^T\delta + \mathbf{a}^T\mathbf{X}\mathbf{b} = \mathbf{d}^T\delta + (\boldsymbol{\lambda} - \mathbf{P}\mathbf{d})^T\mathbf{b} = \mathbf{d}^T\delta + \boldsymbol{\lambda}^T\mathbf{b} - \mathbf{d}^T\mathbf{P}^T\mathbf{b} = \mathbf{d}^T\delta + \boldsymbol{\lambda}^T\mathbf{b} - \mathbf{d}^T\delta = \boldsymbol{\lambda}^T\mathbf{b}$.

($\Rightarrow$): Let $\mathbf{W}$ be a matrix such that $\mathcal{C}(\mathbf{W}) = \mathcal{N}(\mathbf{P}^T)$. Fix $\mathbf{b}_* \in \mathcal{T}$. Then for any $\mathbf{z}$ we have
$\mathbf{P}^T(\mathbf{b}_* + \mathbf{W}\mathbf{z}) = \mathbf{P}^T\mathbf{b}_* + \mathbf{P}^T\mathbf{W}\mathbf{z} = \delta$. Now, since $\mathbb{E}(c + \mathbf{a}^T\mathbf{y}) = \boldsymbol{\lambda}^T\mathbf{b}, \forall \mathbf{b} \in \mathcal{T}$, we have

$$\begin{aligned} & c + \mathbf{a}^T\mathbf{X}\mathbf{b} = \boldsymbol{\lambda}^T\mathbf{b}, \quad \forall \mathbf{b} \in \mathcal{T} \\ \Rightarrow & c + \mathbf{a}^T\mathbf{X}(\mathbf{b}_* + \mathbf{W}\mathbf{z}) = \boldsymbol{\lambda}^T(\mathbf{b}_* + \mathbf{W}\mathbf{z}), \quad \forall \mathbf{z} \\ \Rightarrow & (\boldsymbol{\lambda} - \mathbf{X}^T\mathbf{a})^T\mathbf{W}\mathbf{z} - c - \mathbf{a}^T\mathbf{X}\mathbf{b}_* + \boldsymbol{\lambda}^T\mathbf{b}_* = 0, \quad \forall \mathbf{z} \end{aligned}$$

Using Result A.8, we have $(\boldsymbol{\lambda} - \mathbf{X}^T\mathbf{a})^T\mathbf{W} = \mathbf{0}$. Thus $\mathbf{W}^T(\boldsymbol{\lambda} - \mathbf{X}^T\mathbf{a}) = \mathbf{0} \Rightarrow (\boldsymbol{\lambda} - \mathbf{X}^T\mathbf{a}) \in \mathcal{N}(\mathbf{W}^T) = \mathcal{C}(\mathbf{P}) \Rightarrow \exists$ $\mathbf{d} \in \mathbb{R}^q$, s.t. $\boldsymbol{\lambda} - \mathbf{X}^T\mathbf{a} = \mathbf{P}\mathbf{d}$.

Note that the existence of $\mathbf{a} \in \mathbb{R}^N$ and $\mathbf{d} \in \mathbb{R}^q$ s..t. $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a} + \mathbf{P}\mathbf{d}$ means that

$$\begin{aligned} \boldsymbol{\lambda} &= \begin{pmatrix} \mathbf{X}^T & \mathbf{P} \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{d} \end{pmatrix} \\ \Rightarrow \boldsymbol{\lambda} &\in \mathcal{C}\left(\begin{pmatrix} \mathbf{X}^T & \mathbf{P} \end{pmatrix}\right) \supset \mathcal{C}(\mathbf{X}^T) \end{aligned}$$

Now consider minimizing $Q(\mathbf{b}) = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$ subject to the constraint that $\mathbf{P}^T\mathbf{b} = \delta$.

Method of Lagrange Multipliers

$$L(\mathbf{b}, \boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) + 2\boldsymbol{\theta}^T(\mathbf{P}^T\mathbf{b} - \delta)$$

$$\begin{aligned} \frac{\partial L(\mathbf{b}, \boldsymbol{\theta})}{\partial \mathbf{b}} &= -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{b}) + 2\mathbf{P}\boldsymbol{\theta} \\ \frac{\partial L(\mathbf{b}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= 2(\mathbf{P}^T\mathbf{b} - \delta) \end{aligned}$$

Setting these to zeo yields the R.N.E.s:

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{b} \\ \boldsymbol{\theta} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \delta \end{pmatrix}$$

Are the restricted normal equations (R.N.E.s) consistent?

Let's recall how we establish that the N.E.s ($\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$) are consistent. We showed that $\mathcal{N}(\mathbf{X}^T\mathbf{X}) = \mathcal{N}(\mathbf{X})$, which implies that $\mathcal{C}(\mathbf{X}^T\mathbf{X}^T) = \mathcal{C}(\mathbf{X}^T)$. Note that it would have been sufficient to show that $\mathcal{N}(\mathbf{X}^T\mathbf{X}) \subseteq \mathcal{N}(\mathbf{X})$ because by result A.6, it follows that $\mathcal{C}(\mathbf{X}^T\mathbf{X}) \supseteq \mathcal{C}(\mathbf{X}^T)$. Then $\mathbf{X}^T\mathbf{y} \in \mathcal{C}(\mathbf{X}^T) \Rightarrow \mathbf{X}^T\mathbf{y} \in \mathcal{C}(\mathbf{X}^T\mathbf{X})$.

Result 3.8: The R.N.Es are consistent.

proof: Claim

$$\begin{pmatrix}\mathbf{X}^T\mathbf{y}\\ \delta\end{pmatrix} \in \mathcal{C}\left(\begin{pmatrix}\mathbf{X}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{P}^T\end{pmatrix}\right)$$

Indeed, there esits a vector

$$\mathbf{u} = \begin{pmatrix}\mathbf{u}_1 \in \mathbb{R}^{N\times 1}\\ \mathbf{u}_2 \in \mathbb{R}^{p\times 1}\end{pmatrix} \quad \text{s.t.} \quad \begin{pmatrix}\mathbf{X}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{P}^T\end{pmatrix}\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix} = \begin{pmatrix}\mathbf{X}^T\mathbf{y}\\ \delta\end{pmatrix}$$

where $\mathbf{u}_1 = \mathbf{y}$ and $\mathbf{u}_2$ exists because we assume that $\delta \in \mathcal{C}(\mathbf{P}^T)$.

Thus, it is enough to show that

$$\mathcal{C}\left(\begin{pmatrix}\mathbf{X}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{P}^T\end{pmatrix}\right) \subseteq \mathcal{C}\left(\begin{pmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{P}\\ \mathbf{P}^T & \mathbf{0}\end{pmatrix}\right).$$

To start, let $\mathbf{v} = \begin{pmatrix}\mathbf{v}_1 \in \mathbb{R}^{p\times 1}\\ \mathbf{v}_2 \in \mathbb{R}^{q\times 1}\end{pmatrix}$ be s.t.

$$\begin{pmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{P}\\ \mathbf{P}^T & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}.$$

Then, we have (i) $\mathbf{X}^T\mathbf{X}\mathbf{v}_1 + \mathbf{P}\mathbf{v}_2 = \mathbf{0}$ and (ii) $\mathbf{P}^T\mathbf{v}_1 = \mathbf{0}$.

$$\text{(i)} \Rightarrow \mathbf{v}_1^T\mathbf{X}^T\mathbf{X}\mathbf{v}_1 + \mathbf{v}_1^T\mathbf{P}\mathbf{v}_2 = \mathbf{0} \overset{\text{(ii)}}{\Rightarrow} \mathbf{v}_1^T\mathbf{X}^T\mathbf{X}\mathbf{v}_1 = \mathbf{0} \Rightarrow \mathbf{X}\mathbf{v}_1 = \mathbf{0}.$$

But by (i), $\mathbf{P}\mathbf{v}_2 = \mathbf{0}$

$$\begin{aligned}
&\Rightarrow \begin{pmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{P}\end{pmatrix}\begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}\\
&\Rightarrow \mathcal{N}\left(\begin{pmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{P}\\ \mathbf{P}^T & \mathbf{0}\end{pmatrix}\right) \subseteq \mathcal{N}\left(\begin{pmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{P}\end{pmatrix}\right)\\
&\overset{\text{Result A.6}}{\Rightarrow} \mathcal{C}\left(\begin{pmatrix}\mathbf{X}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{P}^T\end{pmatrix}\right) \subseteq \mathcal{C}\left(\begin{pmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{P}\\ \mathbf{P}^T & \mathbf{0}\end{pmatrix}\right).
\end{aligned}$$

Let $\begin{pmatrix}\hat{\mathbf{b}}_H\\ \hat{\boldsymbol{\theta}}_H\end{pmatrix}$ denote a generic solution to the R.N.E.s

Result 3.9: $\hat{\mathbf{b}}_H$ minimizes $Q(\mathbf{b})$ subject to $\mathbf{b} \in \mathcal{T}$.

proof: Suppose $\mathbf{P}^T\tilde{\mathbf{b}} = \delta$.

$$\begin{aligned}
Q(\tilde{\mathbf{b}}) &= (\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}})^T(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}}) = \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_H + \mathbf{X}(\hat{\mathbf{b}}_H - \tilde{\mathbf{b}})\right)^T\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_H + \mathbf{X}(\hat{\mathbf{b}}_H - \tilde{\mathbf{b}})\right)\\
&= Q(\hat{\mathbf{b}}_H) + \|\mathbf{X}(\hat{\mathbf{b}}_H - \tilde{\mathbf{b}})\|^2
\end{aligned}$$

since $(\hat{\mathbf{b}}_H - \tilde{\mathbf{b}})^T\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_H) = (\hat{\mathbf{b}}_H - \tilde{\mathbf{b}})^T\mathbf{P}\hat{\boldsymbol{\theta}}_H = \left[\mathbf{P}^T(\hat{\mathbf{b}}_H - \tilde{\mathbf{b}})\right]^T\hat{\boldsymbol{\theta}}_H = 0$ (by constraint). So, we have $Q(\hat{\mathbf{b}}_H) \leq Q(\tilde{\mathbf{b}})$ for all $\tilde{\mathbf{b}} \in \mathcal{T}$, with equality *iff* $\|\mathbf{X}(\hat{\mathbf{b}}_H - \tilde{\mathbf{b}})\|^2 = 0$, or $\mathbf{X}\hat{\mathbf{b}}_H = \mathbf{X}\tilde{\mathbf{b}}$.

Result 3.10: Let $\tilde{\mathbf{b}}$ satisfiy $\mathbf{P}^T\tilde{\mathbf{b}} = \delta$. $Q(\hat{\mathbf{b}}_H) = Q(\tilde{\mathbf{b}})$ *iff* $\exists\, \tilde{\boldsymbol{\theta}}$ s.t. $\begin{pmatrix}\tilde{\mathbf{b}} & \tilde{\boldsymbol{\theta}}\end{pmatrix}$ solves the R.N.E.s.

proof: ($\Leftarrow$) If $\begin{pmatrix}\hat{\mathbf{b}}_H & \hat{\boldsymbol{\theta}}_H\end{pmatrix}$ and $\begin{pmatrix}\tilde{\mathbf{b}} & \tilde{\boldsymbol{\theta}}\end{pmatrix}$ both solve the R.N.E.s, then $Q(\hat{\mathbf{b}}_H) = Q(\tilde{\mathbf{b}}) = \min$.

($\Rightarrow$) From the proof of Result 3.9, we have $Q(\hat{\mathbf{b}}_H) = Q(\tilde{\mathbf{b}}) \Leftrightarrow \mathbf{X}\hat{\mathbf{b}}_H = \mathbf{X}\tilde{\mathbf{b}} \Rightarrow \mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_H = \mathbf{X}^T\mathbf{X}\tilde{\mathbf{b}}$. Now, since $\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_H + \mathbf{P}\hat{\boldsymbol{\theta}}_H = \mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\tilde{\mathbf{b}} + \mathbf{P}\hat{\boldsymbol{\theta}}_H$ we have that $\begin{pmatrix}\tilde{\mathbf{b}} & \hat{\boldsymbol{\theta}}_H\end{pmatrix}$ solves the R.N.E.s,

Chapter 4: Gauss–Markov Model

Suppose Σ is an $n \times n$ covariance matrix associated with $Z = (Z_1, \cdots, Z_n)^T$. Then for $v \in \mathbb{R}^n$, we have

$$v^T\Sigma v = Var(v^T Z) = Var(\sum_{i=1}^n v_i Z_i) \geq 0 \quad \Rightarrow \quad \Sigma \text{ is non-negative definite.}$$

When is Σ positive definite? The only way that $v^T\Sigma v = 0$ when $v \neq 0$ is if $Var(\sum_{i=1}^n v_i Z_i) = 0$. This means that $\sum_{i=1}^n v_i Z_i$ must be constant. If no such relationship exists, then Σ is positive definite.

Gauss-Markov Model

$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, $\mathbb{E}\mathbf{e} = 0$, $Cov(\mathbf{e}) = \sigma^2\mathbf{I}$.

So, each element of $\mathbf{e}$ has mean 0, variance σ^2 and $Cov(e_i, e_j) = 0 (i \neq j)$.

Let's derive the variance of the L.S. estimator of the estimable function $\boldsymbol{\lambda}^T\mathbf{b}$.

Of course, L.S. estimator of $\boldsymbol{\lambda}^T\mathbf{b}$ is $\boldsymbol{\lambda}^T\hat{\mathbf{b}} = \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y}$. We know that $\mathbb{E}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) = \boldsymbol{\lambda}^T\mathbf{b}$.

$$\begin{aligned} Var(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) &= Var(\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y}) \\ &= \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T Cov(\mathbf{y})[\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T]^T \\ &= \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda} \end{aligned}$$

Note that $\mathbf{X}^T\mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T$ projects onto $\mathcal{C}(\mathbf{X}^T\mathbf{X}) = \mathcal{C}(\mathbf{X}^T)$ and $\boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X}^T)$, so $\mathbf{X}^T\mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda} = \boldsymbol{\lambda}$. So

$$Var(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) = \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\lambda}.$$

Example:

SLR: $y_i = \beta_0 + \beta_1 X_i + e_i$. Find $Var(\hat{\beta}_1)$.

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} N & N\bar{X} \\ N\bar{X} & \sum_{i=1}^N X_i^2 \end{pmatrix}$$

$$\beta_1 = \boldsymbol{\lambda}^T\mathbf{b} = \begin{pmatrix}0 & 1\end{pmatrix}\begin{pmatrix}b_0 \\ b_1\end{pmatrix}$$

$$
\begin{aligned}
Var(\hat{\beta}_1) &= \sigma^2 \boldsymbol{\lambda}^T (\mathbf{X}^T\mathbf{X})^g \boldsymbol{\lambda} \\
&= \sigma^2 \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} N & N\bar{X} \\ N\bar{X} & \sum_{i=1}^N X_i^2 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\
&= \frac{\sigma^2}{NS_{XX}} \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^N X_i^2 & -N\bar{X} \\ -N\bar{X} & N \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\
&= \frac{\sigma^2}{S_{XX}}
\end{aligned}
$$

Theorem 4.1 (Gauss-Markov Theorem): Assume that $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable. Under the G-M model, the L.S. estimator $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$ is **the** *best (minimum variance) linear unbiased estimator (BLUE).*

proof: Suppose that $c + \mathbf{d}^T\mathbf{y}$ is an unbiased estimator of $\boldsymbol{\lambda}^T\mathbf{b}$. Then $c + \mathbf{d}^T\mathbf{X}\mathbf{b} = \boldsymbol{\lambda}^T\mathbf{b}$, $\forall \mathbf{b} \in \mathbb{R}^p$. This implies $c = 0$ and $\mathbf{d}^T\mathbf{X} = \boldsymbol{\lambda}^T$. Now,

$$
\begin{aligned}
Var(c + \mathbf{d}^T\mathbf{y}) = Var(\mathbf{d}^T\mathbf{y}) &= Var(\boldsymbol{\lambda}^T\hat{\mathbf{b}} + \mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) \\
&= Var(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) + Var(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) + 2Cov(\boldsymbol{\lambda}^T\hat{\mathbf{b}}, \mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) \\
&= Var(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) + Var(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) + 2Cov(\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y}, [\mathbf{d} - \mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda}]^T\mathbf{y}) \\
&= Var(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) + Var(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) + 2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T Cov(\mathbf{y})[\mathbf{d} - \mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda}] \\
&= Var(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) + Var(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) + 2\sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g[\mathbf{X}^T\mathbf{d} - \mathbf{X}^T\mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda}] \\
&= Var(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) + Var(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) + 2\sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g(\boldsymbol{\lambda} - \boldsymbol{\lambda}) \\
&= Var(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) + Var(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) \\
&\geq Var(\boldsymbol{\lambda}^T\hat{\mathbf{b}})
\end{aligned}
$$

with equality *iff*

$$
\begin{aligned}
0 &= Var(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) \\
&= Var(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y}) \\
&= Var\left([\mathbf{d} - \mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda}]^T\mathbf{y}\right) \\
&= [\mathbf{d} - \mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda}]^T Cov(\mathbf{y})[\mathbf{d} - \mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda}] \\
&= \sigma^2\|\mathbf{d} - \mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda}\|^2
\end{aligned}
$$

but this norm square is 0 *iff* $\mathbf{d}^T = \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T \Rightarrow \mathbf{d}^T\mathbf{y} = \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y} = \boldsymbol{\lambda}^T\hat{\mathbf{b}} \Rightarrow$ BLUE is unique.

Note that the crucial step in the proof of G-M theorem is showing that $Cov(\boldsymbol{\lambda}^T\hat{\mathbf{b}}, \mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) = 0$. Of course, $\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}$ is an unbiased estimator of 0.

Result 4.1: The BLUE $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$ is uncorrelated with all linear unbiased estimator of 0.

proof: Suppose $\mathbb{E}(c + \mathbf{a}^T\mathbf{y}) = c + \mathbf{a}^T\mathbf{X}\mathbf{b} = 0, \forall \mathbf{b} \in \mathbb{R}^p$. It follows that $c = 0$ and $\mathbf{a}^T\mathbf{X} = \mathbf{0} \Rightarrow \mathbf{a} \in \mathcal{N}(\mathbf{X}^T)$. So,

$$\begin{aligned} Cov(\boldsymbol{\lambda}^T\hat{\mathbf{b}}, \mathbf{a}^T\mathbf{y}) &= Cov(\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y}, \mathbf{a}^T\mathbf{y}) \\ &= \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T Cov(\mathbf{y})\mathbf{a} \\ &= \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{a} \\ &= 0 \end{aligned}$$

*The G-M theorem can be extended to the case where we have r linearly independent estimable functions $\{\boldsymbol{\lambda}^{(j)T}\mathbf{b}\}_{j=1}^r$. (Assume $rank(\mathbf{X}) = r$)

Let $\boldsymbol{\Lambda}$ be a $p \times r$ matrix whose jth column is $\boldsymbol{\lambda}^{(j)}$. Then $\boldsymbol{\Lambda}^T\hat{\mathbf{b}}$ is the $r \times 1$ vector of L.S. estimators, and $\mathbb{E}(\boldsymbol{\Lambda}^T\hat{\mathbf{b}}) = \boldsymbol{\Lambda}^T\mathbf{b}$,

$$\begin{aligned} Cov(\boldsymbol{\Lambda}^T\hat{\mathbf{b}}) &= Cov(\boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y}) \\ &= \boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T Cov(\mathbf{y})[\boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T]^T \\ &= \sigma^2\boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\Lambda} \\ &= \sigma^2\boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\Lambda}. \end{aligned}$$

Q: Is $\sigma^2\boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\Lambda}$ positive definite?

A: Yes. Note that $\boldsymbol{\Lambda} = \mathbf{X}^T\mathbf{S}$, so $\boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\Lambda} = \mathbf{S}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{S} = \mathbf{S}^T\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X}\mathbf{S} = (\mathbf{P}_\mathbf{X}\mathbf{S})^T\mathbf{P}_\mathbf{X}\mathbf{S}$.

Because $\mathbf{X}^T\mathbf{P}_\mathbf{X}\mathbf{S} = \mathbf{X}^T\mathbf{S} = \boldsymbol{\Lambda}$, we have

$$r = rank(\boldsymbol{\Lambda}) \leq \min\{rank(\mathbf{X}^T), rank(\mathbf{P}_\mathbf{X}\mathbf{S})\} \leq rank(\mathbf{P}_\mathbf{X}\mathbf{S}) \Rightarrow rank(\mathbf{P}_\mathbf{X}\mathbf{S}) = r.$$

So, $\sigma^2\boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\Lambda}$ is positive definite.

Suppose that $\mathbb{E}(\mathbf{C}^T\mathbf{y}) = \boldsymbol{\Lambda}^T\mathbf{b}, \forall \mathbf{b} \in \mathbb{R}^p$. ($\mathbf{C} \in \mathbb{R}^{N\times r}$)

So, $\mathbf{C}^T\mathbf{X}\mathbf{b} = \boldsymbol{\Lambda}^T\mathbf{b}, \forall \mathbf{b} \in \mathbb{R}^p \Rightarrow \mathbf{C}^T\mathbf{X} = \boldsymbol{\Lambda}^T$. Now, $Cov(\mathbf{C}^T\mathbf{y}) = \mathbf{C}^T Cov(\mathbf{y})\mathbf{C} = \sigma^2\mathbf{C}^T\mathbf{C}$ and

$$\begin{aligned} Cov(\mathbf{C}^T\mathbf{y}) - Cov(\boldsymbol{\Lambda}^T\hat{\mathbf{b}}) &= \sigma^2\left[\mathbf{C}^T\mathbf{C} - \boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\Lambda}\right] \\ &= \sigma^2\left[\mathbf{C}^T\mathbf{C} - \mathbf{C}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{C}\right] \\ &= \sigma^2\mathbf{C}^T(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{C} \end{aligned}$$

Note that, for $\mathbf{v} \in \mathbb{R}^r$, we have

$$\begin{gathered} \mathbf{v}^T\mathbf{C}^T[\mathbf{I} - \mathbf{P}_\mathbf{X}]\mathbf{C}\mathbf{v} = \mathbf{v}^T\mathbf{C}(\mathbf{I} - \mathbf{P}_\mathbf{X})(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{C}\mathbf{v} = ((\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{C}\mathbf{v})^T(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{C}\mathbf{v} \geq 0 \\ \Rightarrow \mathbf{C}^T[\mathbf{I} - \mathbf{P}_\mathbf{X}]\mathbf{C} \text{ is non-negative definite} \\ \Rightarrow Cov(\mathbf{C}^T\mathbf{y}) - Cov(\Lambda^T\hat{\mathbf{b}}) \text{ is non-negative definite.} \end{gathered}$$

Fix $\mathbf{v} \in \mathbb{R}^r$, $\mathbf{v} \neq \mathbf{0}$. Let's compare $\mathbf{v}^T\mathbf{C}^T\mathbf{y}$ with $\mathbf{v}^T\boldsymbol{\Lambda}^T\hat{\mathbf{b}}$.

$$Var(\mathbf{v}^T\mathbf{C}^T\mathbf{y}) - Var(\mathbf{v}^T\boldsymbol{\Lambda}^T\hat{\mathbf{b}}) = \mathbf{v}^T Cov(\mathbf{C}^T\mathbf{y})\mathbf{v} - \mathbf{v}^T Cov(\boldsymbol{\Lambda}^T\hat{\mathbf{b}})\mathbf{v} = \mathbf{v}^T(Cov(\mathbf{C}^T\mathbf{y}) - Cov(\Lambda^T\hat{\mathbf{b}}))\mathbf{v} \geq 0.$$

Suppose $r = rank(\mathbf{X}) = p$. Then any functions $\boldsymbol{\lambda}^T\mathbf{b}$ are estimable. So we can take $\boldsymbol{\Lambda} = \mathbf{I}_p$, which yields

$$Cov(\hat{\mathbf{b}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

We also have

$$Cov(\tilde{\mathbf{b}}) - Cov(\hat{\mathbf{b}}) \text{ is non-negative definite,}$$

where $\tilde{\mathbf{b}}$ is any generic unbasied estimator of $\mathbf{b}$.

4.3 Variance estimation

Let $\mathbf{Z}$ be random variable such that $\mathbb{E}\mathbf{Z} = \mu$ and $Cov(\mathbf{Z}) = \mathbf{\Sigma}$. Then if $\mathbf{A}$ is a fixed matrix, we have

$$\mathbb{E}(\mathbf{Z}^T\mathbf{A}\mathbf{Z}) = \mu^T\mathbf{A}\mu + tr(\mathbf{A}\mathbf{\Sigma}).$$

proof: See Page 76 in the textbook.

Fact: $tr(\mathbf{I} - \mathbf{P_X}) = N - r$, where $r = rank(\mathbf{X})$.

Reason: $tr(\mathbf{I} - \mathbf{P_X}) = tr(\mathbf{I}) - tr(\mathbf{P_X})$. Since $\mathbf{P_X}$ is idempotent, $rank(\mathbf{P_X}) = tr(\mathbf{P_X})$, but $\mathcal{C}(\mathbf{P_X}) = \mathcal{C}(\mathbf{X})$, so $rank(\mathbf{P_X}) = rank(\mathbf{X}) = r$.

Recall that $SSE = \hat{\mathbf{e}}^T\hat{\mathbf{e}} = \mathbf{y}^T(\mathbf{I} - \mathbf{P_X})\mathbf{y}$. Define $\hat{\sigma}^2 = SSE/(N - r)$.

Result 4.2: Under the G-M model, $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

proof:

$$\begin{aligned}\mathbb{E}(\mathbf{y}^T(\mathbf{I} - \mathbf{P_X})\mathbf{y}) &= \mathbf{b}^T\mathbf{X}^T(\mathbf{I} - \mathbf{P_X})\mathbf{X}\mathbf{b} + tr((\mathbf{I} - \mathbf{P_X})Cov(\mathbf{y})) \\ &= 0 + \sigma^2 tr(\mathbf{I} - \mathbf{P_X}) \\ &= \sigma^2(N - r).\end{aligned}$$

4.5 The Aitken Model and Generalized Least Squares

Aitken model: $\mathbf{y} = \mathbf{A}\mathbf{b} + \mathbf{e}$, $\mathbb{E}(\mathbf{e}) = \mathbf{0}$, $Cov(\mathbf{e}) = \sigma^2\mathbf{V}$, $\mathbf{V}$ is a fixed positive definite matrix.

The key to dealing with this model is the decomposition $\mathbf{R}\mathbf{V}\mathbf{R}^T = \mathbf{I}$.

Two ways to find $\mathbf{R}$.

1. Cholesky factorization: $\mathbf{V} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is an invertible lower-triangular matrix. Then $\mathbf{R} = \mathbf{L}^{-1}$ and $\mathbf{R}\mathbf{V}\mathbf{R}^T = \mathbf{L}^{-1}\mathbf{V}(\mathbf{L}^{-1})^T = \mathbf{L}^{-1}\mathbf{L}\mathbf{L}^T(\mathbf{L}^{-1})^T = \mathbf{I}$.
2. Spectral decomposition: $\mathbf{V} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$. Here $\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}^{-1/2}\mathbf{Q}^T$ (symmetric) and

$$\mathbf{R}\mathbf{V}\mathbf{R}^T = \mathbf{Q}\mathbf{\Lambda}^{-1/2}\mathbf{Q}^T\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\mathbf{Q}\mathbf{\Lambda}^{-1/2}\mathbf{Q}^T = \mathbf{Q}\mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}\mathbf{\Lambda}^{-1/2}\mathbf{Q}^T = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}.$$

Now, we use $\mathbf{R}$ to transform the Aitken model:

$$\mathbf{Ry} = \mathbf{RXb} + \mathbf{Re}$$
$$\Leftrightarrow \mathbf{z} = \mathbf{Ub} + \mathbf{f}$$

$\mathbb{E}(\mathbf{f}) = \mathbb{E}(\mathbf{Re}) = \mathbf{R}\mathbb{E}(\mathbf{e}) = \mathbf{0}$, $Cov(\mathbf{f}) = Cov(\mathbf{Re}) = \mathbf{R}Cov(\mathbf{e})\mathbf{R}^T = \sigma^2\mathbf{RVR}^T = \sigma^2\mathbf{I}$.

Thus, the transformed model is nothing but a G-M model. We will solve problems using G-M theory, and then transform back.

Estimability: $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable $\Leftrightarrow \boldsymbol{\lambda} \in \mathcal{C}(\mathbf{U}^T)$. But $\mathcal{C}(\mathbf{U}^T) = \mathcal{C}(\mathbf{X}^T\mathbf{R}^T) = \mathcal{C}(\mathbf{X}^T)$ since $\mathbf{R}^T$ is invertible.

Generalized least squares

$\mathbf{U}^T\mathbf{Ub} = \mathbf{U}^T\mathbf{z} \Rightarrow (\mathbf{RX})^T\mathbf{RXb} = (\mathbf{RX})^T\mathbf{z} \Rightarrow \mathbf{X}^T\mathbf{R}^T\mathbf{RXb} = \mathbf{X}^T\mathbf{R}^T\mathbf{Ry} \Rightarrow \mathbf{X}^T\mathbf{V}^{-1}\mathbf{Xb} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$. Here we have used $\mathbf{V} = \mathbf{R}^{-1}(\mathbf{R}^T)^{-1} = (\mathbf{R}^T\mathbf{R})^{-1} \Rightarrow \mathbf{V}^{-1} = \mathbf{R}^T\mathbf{R}$.

We call $\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Xb} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ *Aitken equations*, and call the solution $\hat{\mathbf{b}}_{GLS}$.

Theorem (Aitken's Theorem): Let $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ with $\mathbb{E}\mathbf{e} = 0$ and $Cov(\mathbf{e}) = \sigma^2\mathbf{V}$ where $\mathbf{V}$ is known positive definite. If $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable, then $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{GLS}$ is the BLUE for $\boldsymbol{\lambda}^T\mathbf{b}$.

From the results in Chapter 2, we know that $\hat{\mathbf{b}}_{GLS}$ minimizes

$$\|\mathbf{z} - \mathbf{Ub}\|^2 = \|\mathbf{R}(\mathbf{y} - \mathbf{Xb})\|^2 = (\mathbf{y} - \mathbf{Xb})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb})$$

Examples:

- SLR: $y_i = \beta_0 + \beta_1 x_i + e_i$, $\mathbb{E}(e_i) = 0$, $Var(\mathbf{e}) = \sigma^2\mathbf{V}$ where $\mathbf{V}$ is diagonal.

$$(\mathbf{y} - \mathbf{Xb})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb}) = \sum_{i=1}^{N} \frac{1}{V_{ii}}(y_i - \beta_0 - \beta_1 x_i)^2 \quad \text{"weighted L.S."}$$

- Regression through the origin with heteroskedastic errors. $y_i = \beta_1 x_i + e_i$, $\mathbb{E}e_i = 0$, $Var(e_i) = \sigma^2 x_i^2$, $Cov(e_i, e_j) = 0\ (i \neq j)$.

$$\mathbf{V} = \begin{pmatrix} x_1^2 & & & \\ & x_2^2 & & \\ & & \ddots & \\ & & & x_N^2 \end{pmatrix} \quad \mathbf{R} = \begin{pmatrix} x_1^{-1} & & & \\ & x_2^{-1} & & \\ & & \ddots & \\ & & & x_N^{-1} \end{pmatrix}$$

  $\mathbf{Ry} = (y_1/x_1, y_2/x_2, \cdots, y_N/x_N)^T$, $\mathbf{RX} = \mathbf{1}_N$.

  Transformed model: $\frac{y_i}{x_i} = \beta + \frac{e_i}{x_i}$.

  $\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} = N$, $\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = \sum_{i=1}^N \frac{y_i}{x_i}$, $\hat{\mathbf{b}}_{GLS} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = \frac{1}{N}\sum_{i=1}^N \frac{y_i}{x_i}$,
  $Var(\hat{\mathbf{b}}_{GLS}) = \frac{1}{N^2}\sum_{i=1}^N Var(y_i/x_i) = \frac{1}{N^2}\sum_{i=1}^N \sigma^2 = \sigma^2/N$

Q: Is $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{OLS}$ ever BLUE under the Aitken model?

Result 4.3 (Generalization of Result 4.1): Let $\mathbf{y}$ be a random variable s.t. $Var(y_i) < \infty$. The linear estimator $\mathbf{t}^T\mathbf{y}$ is BLUE for $\mathbb{E}(\mathbf{t}^T\mathbf{y})$ *iff* it is uncorrelated with all linear unbiased estimator of 0.

proof: ($\Leftarrow$) Let $\mathbf{a}^T\mathbf{y}$ be such that $\mathbb{E}(\mathbf{a}^T\mathbf{y}) = \mathbb{E}(\mathbf{t}^T\mathbf{y})$.

$$\begin{aligned} Var(\mathbf{a}^T\mathbf{y}) &= Var(\mathbf{t}^T\mathbf{y} + \mathbf{a}^T\mathbf{y} - \mathbf{t}^T\mathbf{y}) \\ &= Var(\mathbf{t}^T\mathbf{y}) + Var(\mathbf{a}^T\mathbf{y} - \mathbf{t}^T\mathbf{y}) + 2Cov(\mathbf{t}^T\mathbf{y}, \mathbf{a}^T\mathbf{y} - \mathbf{t}^T\mathbf{y}) \\ &= Var(\mathbf{t}^T\mathbf{y}) + Var(\mathbf{a}^T\mathbf{y} - \mathbf{t}^T\mathbf{y}) \\ &\geq Var(\mathbf{t}^T\mathbf{y}) \end{aligned}$$

($\Rightarrow$) Suppose $\mathbb{E}(\mathbf{h}^T\mathbf{y}) = 0$. Need to show that $Cov(\mathbf{t}^T\mathbf{y}, \mathbf{h}^T\mathbf{y}) = 0$. If $\mathbf{h}^T\mathbf{y}$ is constant then the result clearly holds. So assume that $\mathbf{h}^T\mathbf{y}$ is not constant. Let $Cov(\mathbf{t}^T\mathbf{y}, \mathbf{h}^T\mathbf{y}) = c$ and $Var(\mathbf{h}^T\mathbf{y}) = d > 0$. Now consider an alternative linear unbiased estimator of $\mathbb{E}(\mathbf{t}^T\mathbf{y})$: $\mathbf{a}^T\mathbf{y} = \mathbf{t}^T\mathbf{y} - (\frac{c}{d})\mathbf{h}^T\mathbf{y}$, and we have

$$\begin{aligned} Var(\mathbf{a}^T\mathbf{y}) &= Var(\mathbf{t}^T\mathbf{y}) + (\frac{c}{d})^2 Var(\mathbf{h}^T\mathbf{y}) - 2\frac{c}{d}Cov(\mathbf{t}^T\mathbf{y}, \mathbf{h}^T\mathbf{y}) \\ &= Var(\mathbf{t}^T\mathbf{y}) + \frac{c^2}{d} - 2\frac{c^2}{d} \\ &= Var(\mathbf{t}^T\mathbf{y}) - \frac{c^2}{d} \end{aligned}$$

But $\mathbf{t}^T\mathbf{y}$ is BLUE, so c must be 0.

Corollary 4.1: Under the Aitken model, $\mathbf{t}^T\mathbf{y}$ is the BLUE for $\mathbb{E}(\mathbf{t}^T\mathbf{y})$ *iff* $\mathbf{Vt} \in \mathcal{C}(\mathbf{X})$.

proof: It is enough to show that $\mathbf{t}^T\mathbf{y}$ is uncorrelated with all linear unbiased estimators of 0 *iff* $\mathbf{Vt} \in \mathcal{C}(\mathbf{X})$.

First, $\mathbb{E}(\mathbf{h}^T\mathbf{y}) = 0 \Leftrightarrow \mathbf{h}^T\mathbf{X}\mathbf{b} = 0, \forall \mathbf{b} \in \mathbb{R}^p \Leftrightarrow \mathbf{h}^T\mathbf{X} = \mathbf{0} \Leftrightarrow \mathbf{h} \in \mathcal{N}(\mathbf{X}^T)$.

($\Leftarrow$) Suppose $\mathbb{E}(\mathbf{h}^T\mathbf{y}) = 0$. Then $\mathbf{h} \in \mathcal{N}(\mathbf{X}^T)$ and $Cov(\mathbf{t}^T\mathbf{y}, \mathbf{h}^T\mathbf{y}) = \mathbf{t}^T Cov(\mathbf{y})\mathbf{h} = \sigma^2\mathbf{t}^T\mathbf{V}\mathbf{h} = 0$.

($\Rightarrow$) If $Cov(\mathbf{t}^T\mathbf{y}, \mathbf{h}^T\mathbf{y}) = \sigma^2\mathbf{t}^T\mathbf{V}\mathbf{h} = 0$ for $\forall \mathbf{h} \in \mathcal{N}(\mathbf{X}^T)$, then $\mathbf{V}^T\mathbf{t} \perp \mathcal{N}(\mathbf{X}^T) \Rightarrow \mathbf{Vt} \in \mathcal{C}(\mathbf{X})$.

Result 4.4: Under the Aitkin model, all OLS estimators are BLUE (that is, each $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{OLS}$ is BLUE for the corresponding estimable $\boldsymbol{\lambda}^T\mathbf{b}$) *iff* $\exists$ a matrix $\mathbf{Q}$ s.t. $\mathbf{VX} = \mathbf{XQ}$.

proof: ($\Leftarrow$) A generic OLS estuimator takes the form

$$\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{OLS} = \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y} = \mathbf{t}^T\mathbf{y}$$

where $\mathbf{t} = \mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda}$.

Now, if $\mathbf{VX} = \mathbf{XQ}$, we have

$$\mathbf{Vt} = \mathbf{VX}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda} = \mathbf{XQ}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X})$$

so $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{OLS}$ is BLUE by corollary 4.1.

($\Rightarrow$) Let $\mathbf{X}^T\mathbf{X} = (\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}, \cdots, \boldsymbol{\lambda}^{(p)})$; that is, define $\boldsymbol{\lambda}^{(j)}$ to be the j-th column of $\mathbf{X}^T\mathbf{X}$, $j = 1, 2, \cdots, p$.

Now, $\boldsymbol{\lambda}^{(j)T}\hat{\mathbf{b}}_{OLS} = \mathbf{t}^{(j)T}\mathbf{y}$ where $\mathbf{t}^{(j)} = \mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda}^{(j)}$. Since we know that $\boldsymbol{\lambda}^{(j)T}\hat{\mathbf{b}}_{OLS}$ is BLUE, $\mathbf{V}\mathbf{t}^{(j)} \in \mathcal{C}(\mathbf{X})$, or, in other words, $\exists\, \mathbf{q}^{(j)}$ s.t. $\mathbf{V}\mathbf{t}^{(j)} = \mathbf{X}\mathbf{q}^{(j)} \Rightarrow \mathbf{V}\underbrace{(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}, \cdots, \mathbf{t}^{(p)})}_{\mathbf{T}} = \mathbf{X}\underbrace{(\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \cdots, \mathbf{q}^{(p)})}_{\mathbf{Q}}$. But

$$\begin{aligned}
\mathbf{V}\mathbf{T} &= \mathbf{V}\left[\mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda}^{(1)} \quad \mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda}^{(2)} \quad \cdots \quad \mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\boldsymbol{\lambda}^{(p)}\right] \\
&= \mathbf{V}\mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T[\boldsymbol{\lambda}^{(1)} \quad \boldsymbol{\lambda}^{(2)} \quad \cdots \quad \boldsymbol{\lambda}^{(p)}] \\
&= \mathbf{V}\mathbf{X}\underbrace{[(\mathbf{X}^T\mathbf{X})^g]^T\mathbf{X}^T}_{\mathbf{P}_\mathbf{X}}\mathbf{X} = \mathbf{V}\mathbf{X}
\end{aligned}$$

Example: Back to heteroskedastic regression through the origin.

$y_i = \beta x_i + e_i, \mathbb{E}e_i = 0, Var(e_i) = \sigma^2 x_i^2, Cov(e_i, e_j) = 0\,(i \neq j)$. Assume $x_i \neq 0, \forall i$.

$$\mathbf{V} = \begin{pmatrix} x_1^2 & & & \\ & x_2^2 & & \\ & & \ddots & \\ & & & x_N^2 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

We showed that $\hat{\mathbf{b}}_{GLS} = \frac{1}{N}\sum_{i=1}^N \frac{y_i}{x_i}$ and $Var(\hat{\mathbf{b}}_{GLS}) = \frac{\sigma^2}{N}$. Of course,

$$\hat{\mathbf{b}}_{OLS} = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2} \quad Var(\hat{\mathbf{b}}_{OLS}) = \frac{\sum_{i=1}^N x_i^2 Var(y_i)}{\left(\sum_{i=1}^N x_i^2\right)^2} = \sigma^2 \frac{\sum_{i=1}^N x_i^4}{\left(\sum_{i=1}^N x_i^2\right)^2}.$$

Here is a slick way of comparing these variances:

Let U be a random variable taking values $x_1^2, x_2^2, \cdots, x_N^2$ with equal probabilities $1/N$. Then $\mathbb{E}U = \frac{1}{N}\sum x_i^2$ and $\mathbb{E}U^2 = \frac{1}{N}\sum x_i^4$, and

$$\begin{aligned}
Var(U) = \frac{1}{N}\sum x_i^4 - \left(\frac{1}{N}\sum x_i^2\right)^2 &\geq 0 \\
\Rightarrow \frac{1}{N}\sum x_i^4 &\geq \frac{1}{N^2}\left(\sum x_i^2\right)^2 \\
\Rightarrow N\sum x_i^4 \geq \left(\sum x_i^2\right)^2, \quad \text{or} \quad \frac{\sum x_i^4}{(\sum x_i^2)^2} &\geq \frac{1}{N}
\end{aligned}$$

with equality *iff* $Var(U) = 0$; that is, *iff* $|x_i| = c > 0$ for $\forall i = 1, 2, \cdots, N$.

Of course, if $|x_i| = c > 0, \forall i = 1, 2, \cdots, N$,

$$\begin{aligned}
\hat{\mathbf{b}}_{GLS} &= \frac{1}{N}\sum \frac{y_i}{x_i} = \frac{1}{N}\sum \frac{x_i y_i}{x_i^2} = \frac{1}{N}\sum \frac{x_i y_i}{c^2} \\
&= \frac{1}{Nc^2}\sum x_i y_i = \frac{\sum x_i y_i}{\sum x_i^2} = \hat{\mathbf{b}}_{OLS}.
\end{aligned}$$

In other words, the OLS estimator $\hat{\mathbf{b}}_{OLS}$ is BLUE *iff* $|x_i| = c > 0, \forall i$.

Now, let's consider Result 4.4 in this context. When does there exist a $\mathbf{Q}$ such that $\mathbf{VX} = \mathbf{XQ}$?

$$\mathbf{VX} = \begin{pmatrix} x_1^3 \\ x_2^3 \\ \vdots \\ x_N^3 \end{pmatrix} \quad \mathbf{XQ} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} q$$

The only way that $\mathbf{VX} = \mathbf{XQ}$ is if $x_i^3 = x_i q, \forall i \Rightarrow x_i^2 = q = c^2 > 0, \forall i \Rightarrow |x_i| = c > 0, \forall i$.

Example: $\mathbf{y} = \mathbf{Xb} + \mathbf{e}, \mathbb{E}\mathbf{e} = \mathbf{0}, Cov(\mathbf{e}) = \sigma^2\mathbf{V} = \sigma^2\mathbf{I}_N + \tau^2\mathbf{1}_N\mathbf{1}_N^T = \sigma^2(\mathbf{I}_N + \frac{\tau^2}{\sigma^2}\mathbf{1}_N\mathbf{1}_N^T)$. Assume that $\mathbf{Xb} = (\mathbf{1}_N \quad \mathbf{X}^*)\begin{pmatrix} b_1 \\ \mathbf{b}_2 \end{pmatrix}$

Can we find $\mathbf{Q}$ s.t. $\mathbf{VX} = \mathbf{XQ}$?

$$\begin{aligned} \sigma^2\mathbf{VX} &= (\sigma^2\mathbf{I}_N + \tau^2\mathbf{1}_N\mathbf{1}_N^T)\,(\mathbf{1}_N \quad \mathbf{X}^*) \\ &= \left(\sigma^2\mathbf{1}_N + N\tau^2\mathbf{1}_N \quad \sigma^2\mathbf{X}^* + \tau^2\mathbf{1}_N\mathbf{1}_N^T\mathbf{X}^*\right) \end{aligned}$$

$$\mathbf{XQ} = (\mathbf{1}_N \quad \mathbf{X}^*)\begin{pmatrix} Q_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix} = (\mathbf{1}_N Q_{11} + \mathbf{X}^*\mathbf{Q}_{21} \quad \mathbf{1}_N\mathbf{Q}_{12} + \mathbf{X}^*\mathbf{Q}_{22})$$

Matching up terms, we have

$$\begin{aligned} &Q_{11} = \sigma^2 + N\tau^2 \quad \mathbf{Q}_{21} = \mathbf{0} \quad \mathbf{Q}_{12} = \tau^2\mathbf{1}_N^T\mathbf{X}^* \quad \mathbf{Q}_{22} = \sigma^2\mathbf{I}_{p-1} \\ &\Rightarrow \sigma^2\mathbf{VX} = \mathbf{XQ} \quad \text{or} \quad \mathbf{VX} = \mathbf{XQ}', \text{ where } \mathbf{Q}' = \mathbf{Q}/\sigma^2 \\ &\Rightarrow \text{OLS estimators are BLUE.} \end{aligned}$$

4.7 Best Estimation in a Constrained Parameter Space

Recall the "constrained" problem from Section 3.9

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}, \quad \text{where } \mathbf{b} \in \mathcal{T} = \left\{\mathbf{b} \in \mathbb{R}^p : \mathbf{P}^T\mathbf{b} = \delta\right\}$$

$\mathbf{P}$ is a fixed $p \times q$ matrix of rank q and $\delta \in \mathcal{C}(\mathbf{P}^T)$ is fixed.

Recall the restricted normal equations:

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{b} \\ \boldsymbol{\theta} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \delta \end{pmatrix}$$

The R.N.E.s are consistent (for all $\mathbf{y} \in \mathbb{R}^N$ and $\delta \in \mathcal{C}(\mathbf{P}^T)$). As before, let $\begin{pmatrix} \hat{\mathbf{b}}_H \\ \hat{\boldsymbol{\theta}}_H \end{pmatrix}$ denote a generic solution to the R.N.E.s.

We now show that $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_H$ is BLUE for estimable $\boldsymbol{\lambda}^T\mathbf{b}$ in the constrained model.

Lemma 4.2: Assume that $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable in the constrained model. Then the following equations are consistent:

$$\begin{pmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{P}\\ \mathbf{P}^T & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{pmatrix}=\begin{pmatrix}\boldsymbol{\lambda}\\ \mathbf{0}\end{pmatrix}\quad(\star)$$

proof: Result 3.7a implies that $\boldsymbol{\lambda}=\mathbf{X}^T\mathbf{a}+\mathbf{P}\mathbf{d}$ for some $\mathbf{a}\in\mathbb{R}^N$ and $\mathbf{d}\in\mathbb{R}^q$.

Now consider the R.N.E.s with $\mathbf{y}=\mathbf{a}$ and $\delta=\mathbf{0}$:

$$\begin{pmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{P}\\ \mathbf{P}^T & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{w}_1\\ \mathbf{w}_2\end{pmatrix}=\begin{pmatrix}\mathbf{X}^T\mathbf{a}\\ \mathbf{0}\end{pmatrix}$$

Since $\mathbf{0}\in\mathcal{C}(\mathbf{P}^T)$, these equations are consistent. Let $\mathbf{w}^*=\begin{pmatrix}\mathbf{w}_1^*\\ \mathbf{w}_2^*\end{pmatrix}$ be a solution. Then

$$\begin{pmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{P}\\ \mathbf{P}^T & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{w}_1^*\\ \mathbf{w}_2^*+\mathbf{d}\end{pmatrix}=\begin{pmatrix}\mathbf{X}^T\mathbf{X}\mathbf{w}_1^*+\mathbf{P}(\mathbf{w}_2^*+\mathbf{d})\\ \mathbf{P}^T\mathbf{w}_1^*\end{pmatrix}=\begin{pmatrix}\mathbf{X}^T\mathbf{a}+\mathbf{P}\mathbf{d}\\ \mathbf{0}\end{pmatrix}=\begin{pmatrix}\boldsymbol{\lambda}\\ \mathbf{0}\end{pmatrix}.$$

Lemma 4.3: $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_H$ is a linear unbiased estimator of $\boldsymbol{\lambda}^T\mathbf{b}$ (in the constrained model).

proof: Let $\mathbf{v}^*=\begin{pmatrix}\mathbf{v}_1^*\\ \mathbf{v}_2^*\end{pmatrix}$ be a solution to $(\star)$. Then

$$\begin{aligned}\boldsymbol{\lambda}^T\hat{\mathbf{b}}_H=\begin{pmatrix}\boldsymbol{\lambda}^T & \mathbf{0}\end{pmatrix}\begin{pmatrix}\hat{\mathbf{b}}_H\\ \hat{\boldsymbol{\theta}}_H\end{pmatrix}&\overset{(\star)}{=}\begin{pmatrix}\mathbf{v}_1^{*T} & \mathbf{v}_2^{*T}\end{pmatrix}\begin{pmatrix}\mathbf{X}\mathbf{X}^T & \mathbf{P}\\ \mathbf{P}^T & \mathbf{0}\end{pmatrix}\begin{pmatrix}\hat{\mathbf{b}}_H\\ \hat{\boldsymbol{\theta}}_H\end{pmatrix}\\ &=\begin{pmatrix}\mathbf{v}_1^{*T} & \mathbf{v}_2^{*T}\end{pmatrix}\begin{pmatrix}\mathbf{X}^T\mathbf{y}\\ \delta\end{pmatrix}=\mathbf{v}_1^{*T}\mathbf{X}^T\mathbf{y}+\mathbf{v}_2^{*T}\delta\rightarrow\text{linear estimator}\end{aligned}$$

$$\begin{aligned}\mathbb{E}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}_H)&=\mathbb{E}(\mathbf{v}_1^{*T}\mathbf{X}^T\mathbf{y}+\mathbf{v}_2^{*T}\delta)\\ &=\mathbf{v}_1^{*T}\mathbf{X}^T\mathbf{X}\mathbf{b}+\mathbf{v}_2^{*T}\delta\\ &\overset{(\star)}{=}(\boldsymbol{\lambda}-\mathbf{P}\mathbf{v}_2^*)^T\mathbf{b}+\mathbf{v}_2^{*T}\delta\\ &=\boldsymbol{\lambda}^T\mathbf{b}-\mathbf{v}_2^{*T}(\mathbf{P}^T\mathbf{b}-\delta)\\ &=\boldsymbol{\lambda}^T\mathbf{b}.\end{aligned}$$

Result 4.5: Under the G-M assumptions, $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_H$ is the BLUE of the estimable function $\boldsymbol{\lambda}^T\mathbf{b}$ in the constrained model.

proof: Let $\mathbf{a}^T\mathbf{y}+c$ be an unbiased estimator of $\boldsymbol{\lambda}^T\mathbf{b}$ under the constraint. Again, $\boldsymbol{\lambda}=\mathbf{X}^T\mathbf{a}+\mathbf{P}\mathbf{d}$ for some $\mathbf{a}\in\mathbb{R}^N$ and $\mathbf{d}\in\mathbb{R}^q$. Now,

$$Var(\mathbf{a}^T\mathbf{y}+c)=Var(\boldsymbol{\lambda}^T\hat{\mathbf{b}}_H)+Var(\mathbf{a}^T\mathbf{y}+c-\boldsymbol{\lambda}^T\hat{\mathbf{b}}_H)+Cov(\boldsymbol{\lambda}^T\hat{\mathbf{b}}_H,\mathbf{a}^T\mathbf{y}+c-\boldsymbol{\lambda}^T\hat{\mathbf{b}}_H)$$

From the previous proof, we know that $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_H=\mathbf{v}_1^{*T}\mathbf{X}^T\mathbf{y}+\mathbf{v}_2^{*T}\delta$. Thus,

$$\begin{aligned} Cov(\boldsymbol{\lambda}^T\hat{\mathbf{b}}_H, \mathbf{a}^T\mathbf{y} + c - \boldsymbol{\lambda}^T\hat{\mathbf{b}}_H) &= Cov(\mathbf{v}_1^{*T}\mathbf{X}^T\mathbf{y}, \mathbf{a}^T\mathbf{y} - \mathbf{v}_1^{*T}\mathbf{X}^T\mathbf{y}) \\ &= Cov(\mathbf{v}_1^{*T}\mathbf{X}^T\mathbf{y}, (\mathbf{a} - \mathbf{X}\mathbf{v}_1^*)^T\mathbf{y}) \\ &= \mathbf{v}_1^{*T}\mathbf{X}^T Cov(\mathbf{y})(\mathbf{a} - \mathbf{X}\mathbf{v}_1^*) \\ &= \sigma^2\mathbf{v}_1^{*T}(\mathbf{X}^T\mathbf{a} - \mathbf{X}^T\mathbf{X}\mathbf{v}_1^*) \\ &\overset{(\star)}{=} \sigma^2\mathbf{v}_1^{*T}(\boldsymbol{\lambda} - \mathbf{P}\mathbf{d} - (\boldsymbol{\lambda} - \mathbf{P}\mathbf{v}_2^*)) \\ &= \sigma^2\mathbf{v}_1^{*T}\mathbf{P}(\mathbf{v}_2^* - \mathbf{d}) \\ &\overset{(\star)}{=} 0. \end{aligned}$$

Chapter 5: Distributional Theory

5.2 Multivariate Normal Distribution

Let $\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2)$ and $\mathbf{Y} = a\mathbf{X} + b$ for $a, b \in \mathbb{R}$. Then $\mathbf{Y} \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

What about the multivariate version of this?

In math statistics class, we defined the MVN distribution through the density. We said $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\mathbf{V}_{p\times p}$ is positive definite, if

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{V}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

Now, what about a linear function of $\mathbf{X}$? Suppose $\mathbf{A}_{q\times p}$ and $\mathbf{b} \in \mathbb{R}^q$. Define $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$. Is $\mathbf{Y}$ MVN?

$$Cov(\mathbf{Y}) = \mathbf{A}Cov(\mathbf{X})\mathbf{A}^T = \mathbf{A}\mathbf{V}\mathbf{A}^T$$

If $q > p$, then $\mathbf{A}\mathbf{V}\mathbf{A}^T$ cannot be positive definite since $rank(\mathbf{A}\mathbf{V}\mathbf{A}^T) < q$.

So $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ need not be MVN in our old "density" sense.

However, $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ is MVN in a more general sense.

Def: The *mgf* of a p-dimensional r.v. is defined as

$$M_{\mathbf{X}}(t) = \mathbb{E}e^{\mathbf{t}^T\mathbf{X}}$$

provided there exists $h > 0$ such that this expectation exists (is finite) for all $\mathbf{t} \in S$, where
$S := \{\mathbf{t} \in \mathbb{R}^p : t_i \in (-h, h), i = 1, 2, \cdots, p\}$.

Def: Let $\mathbf{Z} = (Z_1, \cdots, Z_p)^T$ where $\{Z_i\}_{i=1}^p$ are i.i.d. $\mathcal{N}(0, 1)$. Then $\mathbf{Z}$ has the *standard multivariate normal distribution*, denoted by $\mathcal{N}_p(\mathbf{0}, \mathbf{I})$. Its pdf is given by

$$p_{\mathbf{Z}}(\mathbf{z}) = (2\pi)^{-p/2} e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}}$$

Q: What is the mgf of $\mathbf{Z}$ (assuming existence)?

Fix $\mathbf{t} \in \mathbb{R}^p$,

$$\begin{aligned} M_{\mathbf{Z}}(\mathbf{t}) = \mathbb{E}e^{\mathbf{t}^T\mathbf{Z}} &= \int_{\mathbb{R}^p} (2\pi)^{-p/2} e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z} + \mathbf{t}^T\mathbf{z}}\, d\mathbf{z} \\ &= \int_{\mathbb{R}^p} (2\pi)^{-p/2} e^{-\frac{1}{2}\sum_{i=1}^n (z_i^2 - 2t_i z_i)}\, d\mathbf{z} \\ &= \prod_{i=1}^p \int_{\mathbb{R}} (2\pi)^{-1/2} e^{-\frac{1}{2}(z_i^2 - 2t_i z_i)}\, dz_i \\ &= \prod_{i=1}^p \int_{\mathbb{R}} (2\pi)^{-1/2} e^{-\frac{1}{2}(z_i - t_i)^2} e^{\frac{1}{2}t_i^2}\, dz_i \\ &= \prod_{i=1}^p e^{\frac{1}{2}t_i^2} = e^{\frac{1}{2}\mathbf{t}^T\mathbf{t}} \end{aligned}$$

Note that the expectation exists (is finite) for all $\mathbf{t} \in \mathbb{R}^p$.

Result 5.1: Let X_1, X_2 be two r.v.s. If the mgf's exist, and $M_{X_1}(t) = M_{X_2}(t)$ for all t in an open square around the origin, then $X_1 \overset{d}{=} X_2$.

Result 5.2: Suppose $X_1, X_2, \cdots, X_p$ have mgfs $M_{X_i}(t_i), i = 1, 2, \cdots, p$. Let $X = (X_1^T, X_2^T, \cdots, X_p^T)^T$ have mgf $M_X(t)$, where $t = (t_1^T, t_2^T, \cdots, t_p^T)^T$. Then $X_1, X_2, \cdots, X_p$ are mutually independent *iff*

$$M_X(t) = \prod_{i=1}^p M_{X_i}(t_i)$$

for all t in an open square around the origin.

Now suppose $\mathbf{Z} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ and let $\mathbf{X} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$. Then

$$M_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[e^{\mathbf{t}^T(\boldsymbol{\mu} + \mathbf{A}\mathbf{Z})}\right] = e^{\mathbf{t}^T\boldsymbol{\mu}}\mathbb{E}[e^{\mathbf{t}^T\mathbf{A}\mathbf{Z}}] = e^{\mathbf{t}^T\boldsymbol{\mu}}\mathbb{E}[e^{(\mathbf{A}^T\mathbf{t})^T\mathbf{Z}}] = e^{\mathbf{t}^T\boldsymbol{\mu}} M_{\mathbf{Z}}(\mathbf{A}^T\mathbf{t}) = e^{\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\mathbf{A}\mathbf{A}^T\mathbf{t}}.$$

Thus, the distribution of $\mathbf{X} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$ depends on $\mathbf{A}$ only through $\mathbf{A}\mathbf{A}^T$. Indeed, if $\mathbf{B}\mathbf{B}^T = \mathbf{A}\mathbf{A}^T$, then $\boldsymbol{\mu} + \mathbf{A}\mathbf{Z} \overset{d}{=} \boldsymbol{\mu} + \mathbf{B}\mathbf{Z}$. Of course, $\mathbf{A}\mathbf{A}^T = Cov(\mathbf{X})$.

Def: Let $\boldsymbol{\mu} \in \mathbb{R}^p$ and let $\mathbf{V}_{p\times p}$ be non-negative definite. We say that $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$ if

$$M_{\mathbf{X}}(\mathbf{t}) = e^{\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\mathbf{V}\mathbf{t}}.$$

If $\mathbf{V}$ is not positive definite, then $\mathbf{X}$ does not have a density.

Example: $\mathbf{Z} \sim \mathcal{N}_2(\mathbf{0}, \mathbf{I})$. $\mathbf{X} = \mathbf{A}\mathbf{Z}$ where $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Then

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} Z_1 \\ 0 \end{pmatrix}$$

$$\mathbf{A}\mathbf{A}^T = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\mathbf{X} \sim \mathcal{N}_2 \left(\mathbf{0}, \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \right)$$

$\mathbf{X}$ lies on this set: $L = \{(u, v) \in \mathbb{R}^2 : v = 0\} = \mathbb{R} \times \{0\}$.

Note that, if $p(x, y)$ is a density on $\mathbb{R}^2$, that is, $p(x, y) \geq 0$ and $\int_{\mathbb{R}^2} p(x, y) dx dy = 1$, then $\int_L p(x, y) dx dy = 0$.

Result 5.3: If $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$ where $\mathbf{a}_{q \times 1}$ and $\mathbf{B}_{q \times p}$, then

$$\mathbf{Y} \sim \mathcal{N}_q(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\mathbf{V}\mathbf{B}^T).$$

proof:
$M_{\mathbf{Y}}(\mathbf{t}) = \mathbb{E}e^{\mathbf{t}^T\mathbf{Y}} = \mathbb{E}e^{\mathbf{t}^T(\mathbf{a}+\mathbf{B}\mathbf{X})} = e^{\mathbf{t}^T\mathbf{a}}\mathbb{E}e^{\mathbf{t}^T\mathbf{B}\mathbf{X}} = e^{\mathbf{t}^T\mathbf{a}}M_{\mathbf{X}}(\mathbf{B}^T\mathbf{t}) = e^{\mathbf{t}^T\mathbf{a}}e^{\mathbf{t}^T\mathbf{B}\boldsymbol{\mu}+\frac{1}{2}\mathbf{t}^T\mathbf{B}\mathbf{V}\mathbf{B}^T\mathbf{t}} = e^{\mathbf{t}^T(\mathbf{a}+\mathbf{B}\boldsymbol{\mu})+\frac{1}{2}\mathbf{t}^T\mathbf{B}\mathbf{V}\mathbf{B}^T\mathbf{t}}$.

Corollary 5.1: If $\mathbf{X}$ is MVN, then the joint distribution of any subset of the components of $\mathbf{X}$ is also MVN.

proof: Partition $\mathbf{X}$, $\boldsymbol{\mu}$ and $\mathbf{V}$ as follows:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}, \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}$$

Now, let $\mathbf{W} = \mathbf{B}\mathbf{X}$ where $\mathbf{B} = (\mathbf{I}_{p_1} \quad \mathbf{0})$, so $\mathbf{W} = \mathbf{X}_1$. Then from Result 5.3, we have

$$\mathbf{X}_1 \sim \mathcal{N}_{p_1} \left((\mathbf{I}_{p_1} \quad \mathbf{0}) \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, (\mathbf{I}_{p_1} \quad \mathbf{0}) \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{p_1} \\ \mathbf{0} \end{pmatrix} \right) \Rightarrow \mathbf{X}_1 \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \mathbf{V}_{11})$$

Corollary 5.2: Suppose $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{V}$ is positive definite. Then

- $\exists$ nonsingular $\mathbf{A}$ s.t. $\mathbf{V} = \mathbf{A}\mathbf{A}^T$
- $\mathbf{A}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$
- the pdf of $\mathbf{X}$ is $(2\pi)^{-p/2}|\mathbf{V}|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$.

proof is HW.

Result 5.4: Assume $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$. Partition as follows

$$\mathbf{X}=\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \vdots\\ \mathbf{X}_m\end{pmatrix}\begin{matrix}p_1\\ p_2\\ \vdots\\ p_m\end{matrix},\quad \boldsymbol{\mu}=\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\\ \vdots\\ \boldsymbol{\mu}_m\end{pmatrix}\begin{matrix}p_1\\ p_2\\ \vdots\\ p_m\end{matrix},\quad \mathbf{V}=\begin{pmatrix}\mathbf{V}_{11} & \mathbf{V}_{12} & \cdots & \mathbf{V}_{1m}\\ \mathbf{V}_{21} & \mathbf{V}_{22} & \cdots & \mathbf{V}_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{V}_{m1} & \mathbf{V}_{m2} & \cdots & \mathbf{V}_{mm}\end{pmatrix}\begin{matrix}p_1\\ p_2\\ \vdots\\ p_m\end{matrix}$$

Then $\mathbf{X}_1,\mathbf{X}_2,\cdots,\mathbf{X}_m$ are mutually independent $\Leftrightarrow \mathbf{V}_{ij}=\mathbf{0}, \forall i\neq j$.

proof: ($\Rightarrow$) $\mathbf{V}_{ij}=\mathbb{E}\left[(\mathbf{X}_i-\boldsymbol{\mu}_i)(\mathbf{X}_i-\boldsymbol{\mu}_i)^T\right]\overset{ind}{=}[\mathbf{E}(\mathbf{X}_i-\boldsymbol{\mu}_i)][\mathbf{E}(\mathbf{X}_i-\boldsymbol{\mu}_i)]^T=\mathbf{0}$.

($\Leftarrow$) $M_{\mathbf{X}}(\mathbf{t})=e^{\mathbf{t}^T\boldsymbol{\mu}+\frac{1}{2}\mathbf{t}^T\mathbf{V}\mathbf{t}}=e^{\sum_{i=1}^m\mathbf{t}_i^T\boldsymbol{\mu}_i+\sum_{i=1}^m\mathbf{t}_i^T\mathbf{V}_{ii}\mathbf{t}_i}=\prod_{i=1}^m e^{\mathbf{t}_i^T\boldsymbol{\mu}+\frac{1}{2}\mathbf{t}_i^T\mathbf{V}_{ii}\mathbf{t}_i}=\prod_{i=1}^m M_{\mathbf{X}_i}(\mathbf{t}_i)$. Result now follows from Result 5.2.

Corollary 5.3: $\mathbf{X}\sim\mathcal{N}_p(\boldsymbol{\mu},\mathbf{V})$. Let $\mathbf{Y}_1=\mathbf{a}_1+\mathbf{B}_1\mathbf{X}$ and $\mathbf{Y}_2=\mathbf{a}_2+\mathbf{B}_2\mathbf{X}$. Then $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent *iff* $\mathbf{B}_1\mathbf{V}\mathbf{B}_2^T=\mathbf{0}$.

proof: $\mathbf{Y}=\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{pmatrix}=\begin{pmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{pmatrix}+\begin{pmatrix}\mathbf{B}_1\\ \mathbf{B}_2\end{pmatrix}\mathbf{X}$. According to Result 5.3, we have

$$\mathbf{Y}\sim\mathcal{N}\left(\begin{pmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{pmatrix}+\begin{pmatrix}\mathbf{B}_1\\ \mathbf{B}_2\end{pmatrix}\boldsymbol{\mu},\begin{pmatrix}\mathbf{B}_1\\ \mathbf{B}_2\end{pmatrix}\mathbf{V}\begin{pmatrix}\mathbf{B}_1 & \mathbf{B}_2\end{pmatrix}^T\right)=\mathcal{N}\left(\begin{pmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{pmatrix}+\begin{pmatrix}\mathbf{B}_1\\ \mathbf{B}_2\end{pmatrix}\boldsymbol{\mu},\begin{pmatrix}\mathbf{B}_1\mathbf{V}\mathbf{B}_1^T & \mathbf{B}_1\mathbf{V}\mathbf{B}_2^T\\ \mathbf{B}_2\mathbf{V}\mathbf{B}_1^T & \mathbf{B}_2\mathbf{V}\mathbf{B}_2^T\end{pmatrix}\right)$$

Result follows immediately from Result 5.4.

5.3 Chi-Square and Related Distributions

Suppose that $\mathbf{Z}\sim\mathcal{N}_p(\mathbf{0},\mathbf{I})$. Define $U=\mathbf{Z}^T\mathbf{Z}=\sum_{i=1}^p Z_i^2$.

$$\begin{aligned}M_U(t)&=\mathbb{E}e^{t\sum_{i=1}^p Z_i^2}=\mathbb{E}\prod_{i=1}^p e^{tZ_i^2}=\prod_{i=1}^p\mathbb{E}e^{tZ_i^2}\\&=\prod_{i=1}^p\int_{\mathbb{R}}(2\pi)^{-1/2}e^{-\frac{1}{2}(z_i^2-2tz_i^2)}dz_i\\&=\prod_{i=1}^p\int_{\mathbb{R}}(2\pi)^{-1/2}e^{-\frac{1}{2}(1-2t)z_i^2}dz_i\\&=\prod_{i=1}^p(1-2t)^{-1/2}\\&=(1-2t)^{-\frac{p}{2}}\quad(t<\frac{1}{2})\end{aligned}$$

This implies that

$$U\sim\text{Gamma}(\frac{p}{2},2)=\chi_p^2$$

$$p_U(u)=\frac{1}{\Gamma(p/2)2^{p/2}}u^{\frac{p}{2}-1}e^{-\frac{p}{2}u}I_{\mathbb{R}^+}(u).$$

Def 5.6: Let $J \sim \text{Poisson}(\phi)$ ($\phi > 0$). Let $U|J = j \sim \chi^2_{p+2j}$. Then $U \sim \chi^2_p(\phi)$, that is, noncentral χ^2 with p degrees of freedom and noncentrality parameter ϕ. The pdf of $U \sim \chi^2_p(\phi)$ is given by

$$p_U(u) = \sum_{j=0}^{\infty} \left[\frac{e^{-\phi}\phi^j}{j!} \right] \left[\frac{u^{\frac{p+2j-2}{2}} e^{-\frac{u}{2}}}{\Gamma(\frac{p+2j}{2}) 2^{\frac{p+2j}{2}}} \right] I_{\mathbb{R}^+}(u).$$

Result 5.5: If $U \sim \chi^2_p(\phi)$, then

$$M_U(t) = (1-2t)^{-\frac{p}{2}} e^{\frac{2\phi t}{1-2t}}$$

proof:

$$\begin{aligned} M_U(t) = \mathbb{E}e^{tU} = \mathbb{E}[\mathbb{E}[e^{tU}|J]] = \mathbb{E}[(1-2t)^{-(p+2J)/2}] \quad (t < \frac{1}{2}) \\ &= \sum_{j=0}^{\infty} (1-2t)^{-(p+2j)/2} \left[\frac{e^{-\phi}\phi^j}{j!} \right] \\ &= (1-2t)^{-p/2} e^{-\phi} \sum_{j=0}^{\infty} \frac{\left(\frac{\phi}{1-2t}\right)^j}{j!} \\ &= (1-2t)^{-p/2} e^{-\phi} e^{\frac{\phi}{1-2t}} \\ &= (1-2t)^{-p/2} e^{\frac{2\phi t}{1-2t}}. \end{aligned}$$

Result 5.6: If $U \sim \chi^2_p(\phi)$, then $\mathbb{E}U = p + 2\phi$, $Var(U) = 2p + 8\phi$.

proof: Easy; use $\mathbb{E}U = \mathbb{E}[\mathbb{E}[U|J]]$ and $Var(U) = Var[\mathbb{E}[U|J]] + \mathbb{E}[Var[U|J]]$.

Result 5.7: If $\{U_i\}_{i=1}^m$ are independent $\chi^2_{p_i}(\phi_i)$, then $U = \sum_{i=1}^m U_i \sim \chi^2_{\sum p_i}(\sum \phi_i)$.

proof: Easy; $M_U(t) = \prod_{i=1}^m M_{U_i}(t)$.

Result 5.8: If $X \sim \mathcal{N}(\mu, 1)$, then $U = X^2 \sim \chi^2_1(\mu^2/2)$.

proof:

$$\begin{aligned} M_U(t) = \mathbb{E}e^{tX^2} &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2} e^{tx^2}\, dx \\ &= e^{-\mu^2/2} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2 - 2x\mu - 2tx^2)}\, dx \\ &= e^{-\mu^2/2 + \mu^2/(2(1-2t))} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)\left(x - \frac{\mu}{1-2t}\right)^2}\, dx \\ &= (1-2t)^{-\frac{1}{2}} e^{\frac{\mu^2}{2}\frac{2t}{1-2t}}. \end{aligned}$$

Result 5.9: If $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{I})$, then $W = \mathbf{X}^T\mathbf{X} \sim \chi^2_p(\boldsymbol{\mu}^T\boldsymbol{\mu}/2)$.

proof: $W = \sum_{i=1}^p U_i$ where $U_i \sim \chi^2_1(\mu_i^2/2) \Rightarrow W \sim \chi^2_p(\sum \mu_i^2/2)$ (By Result 5.7).

Result 5.10: If $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$ where $\mathbf{V}$ is nonsingular, then $W = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} \sim \chi^2_p(\frac{1}{2}\boldsymbol{\mu}^T\mathbf{V}^{-1}\boldsymbol{\mu})$.

proof: $\exists$ a nonsingular matrix $\mathbf{A}$ s.t. $\mathbf{A}^T\mathbf{A} = \mathbf{V}$. Define $\mathbf{Z} = \mathbf{A}^{-T}\mathbf{X}$. Then
$\mathbf{Z} \sim \mathcal{N}_p(\mathbf{A}^{-T}\boldsymbol{\mu}, \mathbf{A}^{-T}\mathbf{V}\mathbf{A}^{-1}) = \mathcal{N}_p(\mathbf{A}^{-T}\boldsymbol{\mu}, \mathbf{I})$. Now, by Result 5.9, we have

$$W = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} = \mathbf{X}^T\mathbf{A}^{-1}\mathbf{A}^{-T}\mathbf{X} = \mathbf{Z}^T\mathbf{Z} \sim \chi^2_p\left(\frac{\boldsymbol{\mu}^T\mathbf{A}^{-1}\mathbf{A}^{-T}\boldsymbol{\mu}}{2}\right)$$
$$\Rightarrow W \sim \chi^2_p\left(\frac{1}{2}\boldsymbol{\mu}^T\mathbf{V}^{-1}\boldsymbol{\mu}\right).$$

Result 5.11: Let $U \sim \chi^2_p(\phi)$. Then $P(U > c)$ is strictly increasing in ϕ for fixed c and p.

*More explanation: Suppose $\phi' > \phi$, $U' \sim \chi^2_p(\phi')$, $U \sim \chi^2_p(\phi)$. U' is stochastically larger than U.
$P(U' > c) > P(U > c)$, or $F_{U'}(t) < F_U(t)$.

proof: Define

$$v_k = P(\chi^2_k > c) = \int_c^\infty \frac{v^{k/2-1}e^{-v/2}}{\Gamma(k/2)2^{k/2}}\,dv$$

Claim $v_{k+2} > v_k$:

$$\begin{aligned}
v_{k+2} &= \int_c^\infty \frac{v^{(k+2)/2-1}e^{-v/2}}{\Gamma((k+2)/2)2^{(k+2)/2}}\,dv \\
&= \frac{1}{k\Gamma(k/2)2^{k/2}}\int_c^\infty v^{k/2}e^{-v/2}\,dv \\
&= \frac{1}{k\Gamma(k/2)2^{k/2}}\left[-2v^{k/2}e^{-v/2}\Big|_c^\infty + k\int_c^\infty v^{k/2-1}e^{-v/2}\,dv\right] \\
&= \frac{2c^{k/2}e^{-c/2}}{k\Gamma(k/2)2^{k/2}} + \frac{1}{\Gamma(k/2)2^{k/2}}\int_c^\infty v^{k/2-1}e^{-v/2}\,dv \\
&= \text{something positive} + v_k
\end{aligned}$$

So claim is established.

Now $P(U > c) = \sum_{j=0}^\infty \left[\frac{e^{-\phi}\phi^j}{j!}\right] v_{p+2j}$ and

$$\begin{aligned}
\frac{d}{d\phi}P(U>c) &= \frac{d}{d\phi}\left[e^{-\phi}v_p + \sum_{j=1}^{\infty}\left[\frac{e^{-\phi}\phi^j}{j!}\right]v_{p+2j}\right] \\
&= -e^{-\phi}v_p + \sum_{j=1}^{\infty}\left[\frac{e^{-\phi}\phi^{j-1}}{(j-1)!} - \frac{e^{-\phi}\phi^j}{j!}\right]v_{p+2j} \\
&= \sum_{j=1}^{\infty}\frac{e^{-\phi}\phi^{j-1}}{(j-1)!}v_{p+2j} - \sum_{j=0}^{\infty}\frac{e^{-\phi}\phi^j}{j!}v_{p+2j} \\
&= \sum_{k=0}^{\infty}\frac{e^{-\phi}\phi^k}{k!}v_{p+2(k+1)} - \sum_{j=0}^{\infty}\frac{e^{-\phi}\phi^j}{j!}v_{p+2k} \\
&= \sum_{k=0}^{\infty}\frac{e^{-\phi}\phi^k}{k!}[v_{p+2(k+1)} - v_{p+2k}] \\
&> 0 \quad \text{(By claim)}.
\end{aligned}$$

Def 5.7: $U_1 \perp U_2, U_1 \sim \chi^2_{p_1}, U_2 \sim \chi^2_{p_2}$. Then $F = \frac{U_1/p_1}{U_2/p_2} \sim F_{p_1,p_2}$.

Def 5.8: $U_1 \perp U_2, U_1 \sim \chi^2_{p_1}(\phi), U_2 \sim \chi^2_{p_2}$. Then $F = \frac{U_1/p_1}{U_2/p_2} \sim F_{p_1,p_2}(\phi)$ (*Noncentral F*).

Result 5.13: $W \sim F_{p_1,p_2}(\phi)$. For fixed p_1, p_2 and $c > 0$, $P(W > c)$ is strictly increasing in ϕ.

*In other words, if $\phi' > \phi$ and $W' \sim F_{p_1,p_2}(\phi')$ and $W \sim F_{p_1,p_2}(\phi)$, then W' is sochastically larger than W.

proof:

$$\begin{aligned}
P(W>c) &= P\left(\frac{U_1/p_1}{U_2/p_2} > c\right) = P\left(U_1 > \frac{p_1}{p_2}cU_2\right) \\
&= \int_0^{\infty} P\left(U_1 > \frac{p_1}{p_2}cU_2 \middle| U_2 = u_2\right) f_{U_2}(u_2)\,du_2 \\
&= \int_0^{\infty} P\left(U_1 > \frac{p_1}{p_2}cu_2\right) f_{U_2}(u_2)\,du_2
\end{aligned}$$

$W' = \frac{U_1'/p_1}{U_2/p_2}, W = \frac{U_1/p_1}{U_3/p_2}, U_1 \sim \chi^2_{p_1}(\phi), U_1' \sim \chi^2_{p_1}(\phi'), U_2 \sim \chi^2_{p_2}, U_3 \sim \chi^2_{p_2}$.

$$P(W'>c) - P(W>c) = \int_0^{\infty} \underbrace{\left[P\left(U_1' > \frac{p_1}{p_2}cu_2\right) - P\left(U_1 > \frac{p_1}{p_2}cu_2\right)\right]}_{>0 \text{ for all } u_2>0 \text{ due to Result 5.11}} f_{U_2}(u_2)\,du_2 > 0.$$

Def 5.9: $U \sim \mathcal{N}(\mu, 1)$ and $V \sim \chi^2_k, U \perp V$. Then $T = \frac{U}{\sqrt{V/k}} \sim t_k(\mu)$ *noncentral student's t distribution* with k d.f. and noncentrality parameter μ. If $\mu = 0$, this is the usual student's distribution.

*Note that $T \sim t_k(\mu), T = \frac{U}{\sqrt{V/k}}, T^2 = \frac{U^2}{V/k} \sim \frac{\chi^2_1(\mu^2/2)}{\chi^2_k} = F_{1,k}(\mu^2/2)$.

5.4 Distribution of Quadratic Forms

Lemma 5.1: Let $\mathbf{A}$ be a $p \times p$ symmetric matrix. Then $\mathbf{A}$ is idempotent with rank s *iff* $\exists$ a $p \times s$ matrix $\mathbf{G}$ s.t. $\mathbf{G}^T\mathbf{G} = \mathbf{I}_s$ and $\mathbf{G}\mathbf{G}^T = \mathbf{A}$.

proof: ($\Leftarrow$) $\mathbf{A}^2 = \mathbf{G}\mathbf{G}^T\mathbf{G}\mathbf{G}^T = \mathbf{G}\mathbf{G}^T = \mathbf{A}$ so $\mathbf{A}$ is idempotent. $s = rank(\mathbf{G}^T\mathbf{G}) = rank(\mathbf{G}\mathbf{G}^T) = rank(\mathbf{A})$.

($\Rightarrow$) Since $\mathbf{A}^2 = \mathbf{A}$, all eigenvalues of $\mathbf{A}$ are either 0 or 1, and $rank(\mathbf{A})$ is equal to the number of nonzero eigenvalues. Since $\mathbf{A}$ is symmetric, we have a spectral decomposition

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T = \begin{pmatrix}\mathbf{Q}_1 & \mathbf{Q}_2\end{pmatrix}\begin{pmatrix}\mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{Q}_1^T \\ \mathbf{Q}_2^T\end{pmatrix} = \mathbf{Q}_1\mathbf{Q}_1^T$$

$\mathbf{Q}_1^T\mathbf{Q}_1 = \mathbf{I}_s$ because columns of $\mathbf{Q}_1$ are orthogonal. So we have $\mathbf{G} = \mathbf{Q}_1$.

Result 5.14: Let $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{I})$. If $\mathbf{A}$ is symmetric and idempotent with rank s, then

$$\mathbf{X}^T\mathbf{A}\mathbf{X} \sim \chi^2_s\left(\frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}\right).$$

proof: By Lemma 5.1, $\exists$ $\mathbf{G}$ s.t. $\mathbf{G}^T\mathbf{G} = \mathbf{I}_s$ and $\mathbf{G}\mathbf{G}^T = \mathbf{A}$.

Now, $\mathbf{G}^T\mathbf{X} \sim \mathcal{N}_s(\mathbf{G}^T\boldsymbol{\mu}, \mathbf{I}_s)$. Result 5.9 $\Rightarrow \mathbf{X}^T\mathbf{A}\mathbf{X} = (\mathbf{G}^T\mathbf{X})^T\mathbf{G}\mathbf{X} \sim \chi^2_s\left(\frac{1}{2}\boldsymbol{\mu}^T\mathbf{G}\mathbf{G}^T\boldsymbol{\mu}\right) = \chi^2_s\left(\frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}\right)$.

Result 5.15: Let $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$ with $\mathbf{V}$ nonsingular. Let $\mathbf{A}$ be a symmetric matrix. If $\mathbf{A}\mathbf{V}$ is idempotent with rank s, then

$$\mathbf{X}^T\mathbf{A}\mathbf{X} \sim \chi^2_s\left(\frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}\right)$$

proof: $\mathbf{V} = \mathbf{\Gamma}\mathbf{\Gamma}^T$. Define $\mathbf{B} = \mathbf{\Gamma}^T\mathbf{A}\mathbf{\Gamma}$ so that $\mathbf{A} = (\mathbf{\Gamma}^T)^{-1}\mathbf{B}\mathbf{\Gamma}^{-1}$. Now $\mathbf{A}\mathbf{V} = \mathbf{A}\mathbf{V}\mathbf{A}\mathbf{V} \Rightarrow \mathbf{A} = \mathbf{A}\mathbf{V}\mathbf{A}$ because $\mathbf{V}$ is nonsingular. Thus $\mathbf{B} = \mathbf{\Gamma}^T\mathbf{A}\mathbf{\Gamma} = \mathbf{\Gamma}^T\mathbf{A}\mathbf{V}\mathbf{A}\mathbf{\Gamma} = \mathbf{\Gamma}^T\mathbf{A}\mathbf{\Gamma}\mathbf{\Gamma}^T\mathbf{A}\mathbf{\Gamma} = \mathbf{B}^2$. So $\mathbf{B}$ is idempotent. $\mathbf{B}$ is also symmetric. Also, $rank(\mathbf{B}) = rank(\mathbf{A}) = rank(\mathbf{A}\mathbf{V}) = s$.

Now, $\mathbf{Y} = \mathbf{\Gamma}^{-1}\mathbf{X} \sim \mathcal{N}_p(\mathbf{\Gamma}^{-1}\boldsymbol{\mu}, \mathbf{I}_p)$. Result 5.14 $\Rightarrow$
$\mathbf{X}^T\mathbf{A}\mathbf{X} = \mathbf{X}^T(\mathbf{\Gamma}^T)^{-1}\mathbf{\Gamma}^T\mathbf{A}\mathbf{\Gamma}\mathbf{\Gamma}^{-1}\mathbf{X} = \mathbf{Y}^T\mathbf{B}\mathbf{Y} \sim \chi^2_s\left(\frac{1}{2}\boldsymbol{\mu}^T(\mathbf{\Gamma}^{-1})^T\mathbf{B}\mathbf{\Gamma}^{-1}\boldsymbol{\mu}\right) = \chi^2_s\left(\frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}\right)$.

Back to G-M model $\mathbf{y} \sim \mathcal{N}_N(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I})$, $r = rank(\mathbf{X})$. We will apply Result 5.15 with $\boldsymbol{\mu} = \mathbf{X}\mathbf{b}$ and $\mathbf{V} = \sigma^2\mathbf{I}$.

First, take $\mathbf{A} = \frac{1}{\sigma^2}(\mathbf{I} - \mathbf{P}_\mathbf{X})$. Then $\mathbf{A}\mathbf{V} = \mathbf{I} - \mathbf{P}_\mathbf{X}$ which is idempotent and has rank $N - r$. By Result 5.15,
$\mathbf{y}^T\mathbf{A}\mathbf{y} = \frac{\|\hat{\mathbf{e}}\|^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2_{N-r}$; noncentrality parameter is 0 since $\frac{1}{2\sigma^2}(\mathbf{X}\mathbf{b})^T(\mathbf{I} - \mathbf{P}_\mathbf{X})(\mathbf{X}\mathbf{b}) = 0$.

Second, take $\mathbf{A} = \frac{1}{\sigma^2}\mathbf{P}_\mathbf{X}$. Then $\mathbf{A}\mathbf{V} = \mathbf{P}_\mathbf{X}$, which is idempotent and has rank r. By Result 5.15,
$\mathbf{y}^T\mathbf{A}\mathbf{y} = \frac{\|\hat{\mathbf{y}}\|^2}{\sigma^2} = \frac{SSR}{\sigma^2} \sim \chi^2_r\left(\frac{1}{2\sigma^2}\|\mathbf{X}\mathbf{b}\|^2\right)$; noncentrality parameter is $\frac{1}{2\sigma^2}(\mathbf{X}\mathbf{b})^T\mathbf{P}_\mathbf{X}(\mathbf{X}\mathbf{b}) = \frac{1}{2\sigma^2}\|\mathbf{X}\mathbf{b}\|^2$.

Now note that $\begin{pmatrix}\hat{\mathbf{y}}\\ \hat{\mathbf{e}}\end{pmatrix} = \begin{pmatrix}\mathbf{P_X y}\\ (\mathbf{I}-\mathbf{P_X})\mathbf{y}\end{pmatrix} = \begin{pmatrix}\mathbf{P_X}\\ \mathbf{I}-\mathbf{P_X}\end{pmatrix}\mathbf{y}$. By Result 5.3, we have

$$\begin{aligned}\begin{pmatrix}\hat{\mathbf{y}}\\ \hat{\mathbf{e}}\end{pmatrix} &\sim \mathcal{N}_{2N}\left(\begin{pmatrix}\mathbf{P_X}\\ \mathbf{I}-\mathbf{P_X}\end{pmatrix}\mathbf{Xb}, \begin{pmatrix}\mathbf{P_X}\\ \mathbf{I}-\mathbf{P_X}\end{pmatrix}\sigma^2\mathbf{I}\begin{pmatrix}\mathbf{P_X} & \mathbf{I}-\mathbf{P_X}\end{pmatrix}\right)\\ &\sim \mathcal{N}_{2N}\left(\begin{pmatrix}\mathbf{Xb}\\ \mathbf{0}\end{pmatrix}, \sigma^2\begin{pmatrix}\mathbf{P_X} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}-\mathbf{P_X}\end{pmatrix}\right)\end{aligned}$$

Corollary 5.3 $\Rightarrow \hat{\mathbf{y}}$ and $\hat{\mathbf{e}}$ are independent. Hence, SSE and SSR are also independent. Thus

$$F = \frac{\|\hat{\mathbf{y}}\|^2/r}{\|\hat{\mathbf{e}}\|^2/(N-r)} \sim F_{r,N-r}\left(\frac{1}{2\sigma^2}\|\mathbf{Xb}\|^2\right)$$

Result 5.16: Let $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{A}$ is a symmetric matrix, and $\mathbf{B}$ is a $q \times p$ matrix. If $\mathbf{BVA} = \mathbf{0}$, then $\mathbf{BX}$ and $\mathbf{X}^T\mathbf{AX}$ are independent.

proof: $\mathbf{A} = \mathbf{Q\Lambda Q}^T = \begin{pmatrix}\mathbf{Q}_1 & \mathbf{Q}_2\end{pmatrix}\begin{pmatrix}\mathbf{\Lambda}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{Q}_1^T\\ \mathbf{Q}_2^T\end{pmatrix}$ where $\mathbf{Q}^T\mathbf{Q} = \mathbf{QQ}^T = \mathbf{I}$ and $\mathbf{\Lambda}_1$ is diagonal with rank $= rank(A) = s$. So $\mathbf{A} = \mathbf{Q}_1\mathbf{\Lambda}_1\mathbf{Q}_1^T$.

Result 5.3 $\Rightarrow$

$$\begin{pmatrix}\mathbf{B}\\ \mathbf{Q}_1^T\end{pmatrix}\mathbf{X} \sim \mathcal{N}_{q+s}\left(\begin{pmatrix}\mathbf{B\mu}\\ \mathbf{Q}_1^T\boldsymbol{\mu}\end{pmatrix}, \begin{pmatrix}\mathbf{BVB}^T & \mathbf{BVQ}_1\\ \mathbf{Q}_1^T\mathbf{VB}^T & \mathbf{Q}_1^T\mathbf{VQ}_1\end{pmatrix}\right)$$

So, if $\mathbf{BVQ}_1 = \mathbf{0}$, then $\mathbf{BX}$ and $\mathbf{Q}_1^T\mathbf{X}$ are independent. But note that $\mathbf{BX}$ and $\mathbf{Q}_1^T\mathbf{X}$ are independent, then so are $\mathbf{BX}$ and $(\mathbf{Q}_1^T\mathbf{X})^T\mathbf{\Lambda}_1(\mathbf{Q}_1^T\mathbf{X}) = \mathbf{X}^T\mathbf{AX}$.

Finally, if $\mathbf{BVA} = \mathbf{0}$, then $\mathbf{BVQ}_1\mathbf{\Lambda}_1\mathbf{Q}_1^T = \mathbf{0} \Rightarrow \mathbf{BVQ}_1\mathbf{\Lambda}_1\mathbf{Q}_1^T\mathbf{Q}_1 = \mathbf{0} \Rightarrow \mathbf{BVQ}_1\mathbf{\Lambda}_1 = \mathbf{0} \Rightarrow \mathbf{BVQ}_1 = \mathbf{0}$.

Corollary 5.4: Let $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$. Suppose $\mathbf{A}$ and $\mathbf{B}$ are symmetric. If $\mathbf{BVA} = \mathbf{0}$, then $\mathbf{X}^T\mathbf{AX}$ and $\mathbf{X}^T\mathbf{BX}$ are independent.

proof: HW.

5.5 Cochran's Theorem

Theorem 5.1: Let $\mathbf{y} \sim \mathcal{N}_N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. Suppose that $\{\mathbf{A}_i\}_{i=1}^k$ are symmetric, idempotent matrices s.t. $rank(\mathbf{A}_i) = s_i$ and $\sum_{i=1}^k \mathbf{A}_i = \mathbf{I}_N$. Then $\sum_{i=1}^k s_i = N$, $\{\mathbf{y}^T\mathbf{A}_i\mathbf{y}\}_{i=1}^k$ are independent, and $\frac{1}{\sigma^2}\mathbf{y}^T\mathbf{A}_i\mathbf{y} \sim \chi^2_{s_i}\left(\frac{1}{2\sigma^2}\boldsymbol{\mu}^T\mathbf{A}_i\boldsymbol{\mu}\right)$.

proof: $N = \text{trace}(\mathbf{I}_N) = \text{trace}(\sum_{i=1}^k \mathbf{A}_i) = \sum_{i=1}^k \text{trace}(\mathbf{A}_i) = \sum_{i=1}^k s_i$.

By Lemma 5.1, for each $i = 1, 2, \cdots, k$, $\exists$ a $N \times s_i$ matrix $\mathbf{Q}_i$ s.t. $\mathbf{Q}_i\mathbf{Q}_i^T = \mathbf{A}_i$ and $\mathbf{Q}_i^T\mathbf{Q}_i = \mathbf{I}_{s_i}$ (Note that $rank(\mathbf{Q}_i) = s_i$).

Define $\mathbf{Q} = (\mathbf{Q}_1 \quad \mathbf{Q}_2 \quad \cdots \quad \mathbf{Q}_k)_{N\times N}$. Now $\mathbf{Q}\mathbf{Q}^T = \sum_{i=1}^k \mathbf{Q}_i\mathbf{Q}_i^T = \sum_{i=1}^k \mathbf{A}_i = \mathbf{I}_N$. Since $\mathbf{Q}$ is square, it must be orthogonal, so $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_N$.

$$\mathbf{Q}^T\mathbf{Q} = \begin{pmatrix}\mathbf{Q}_1^T\\ \mathbf{Q}_2^T\\ \vdots\\ \mathbf{Q}_k^T\end{pmatrix}(\mathbf{Q}_1 \quad \mathbf{Q}_2 \quad \cdots \quad \mathbf{Q}_k) = \begin{pmatrix}\mathbf{I}_{s_1} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{s_2} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_{s_k}\end{pmatrix} \Rightarrow \mathbf{Q}_i^T\mathbf{Q}_j = \mathbf{0}, \forall i \neq j.$$

Now $\mathbf{Q}^T\mathbf{y} = \begin{pmatrix}\mathbf{Q}_1^T\mathbf{y}\\ \mathbf{Q}_2^T\mathbf{y}\\ \vdots\\ \mathbf{Q}_k^T\mathbf{y}\end{pmatrix} \sim \mathcal{N}(\mathbf{Q}^T\boldsymbol{\mu}, \sigma^2\mathbf{I}_N) \Rightarrow \{\mathbf{Q}_i^T\mathbf{y}\}_{i=1}^k$ are independent and $\mathbf{Q}_i^T\mathbf{y} \sim \mathcal{N}(\mathbf{Q}_i^T\boldsymbol{\mu}, \sigma^2\mathbf{I}_{s_i})$, i.e., $\frac{1}{\sigma}\mathbf{Q}_i^T\mathbf{y} \sim \mathcal{N}(\mathbf{Q}_i^T\boldsymbol{\mu}/\sigma, \mathbf{I}_{s_i})$.

Now, since $\mathbf{y}^T\mathbf{A}_i\mathbf{y} = \|\mathbf{Q}_i^T\mathbf{y}\|^2$, $\{\mathbf{y}^T\mathbf{A}_i\mathbf{y}\}_{i=1}^k$ are also independent.

And, finally, we have

$$\frac{1}{\sigma^2}\mathbf{y}^T\mathbf{A}_i\mathbf{y} = \frac{1}{\sigma^2}\|\mathbf{Q}_i^T\mathbf{y}\|^2 \sim \chi^2_{s_i}\left(\frac{1}{2\sigma^2}\boldsymbol{\mu}^T\mathbf{A}_i\boldsymbol{\mu}\right).$$

Example: One-way ANOVA

$y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, 2, \cdots, a, j = 1, 2, \cdots, n_i, e_{ij} \sim \mathcal{N}(0, \sigma^2)$ independent.

$$\mathbf{X}\mathbf{b} = \begin{pmatrix}\mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \mathbf{1}_{n_a} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_a}\end{pmatrix}\begin{pmatrix}\mu\\ \alpha_1\\ \alpha_2\\ \vdots\\ \alpha_a\end{pmatrix}$$

$$\mathbf{A}_1 = \mathbf{P}_{\mathbf{1}_N}, \quad \mathbf{A}_2 = \mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{1}_N}, \quad \mathbf{A}_3 = \mathbf{I} - \mathbf{P}_{\mathbf{X}}$$

Clearly, $\mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3 = \mathbf{I}$.

$rank(\mathbf{P}_{\mathbf{1}_N}) = 1, rank(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = N - a \Rightarrow rank(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{1}_N}) = a - 1.$

$\mathcal{C}(\mathbf{1}_N) \subset \mathcal{C}(\mathbf{X}) \overset{\text{Thm 2.2}}{\Rightarrow} \mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{1}_N}$ is symmetric projection onto $\mathcal{C}((\mathbf{I} - \mathbf{P}_{\mathbf{1}_N})\mathbf{X})$.

$$\mathbf{y}^T\mathbf{A}_1\mathbf{y} = SSM = N\bar{y}^2 \qquad \frac{SSM}{\sigma^2} \sim \chi^2_1\left(\frac{1}{2\sigma^2}(\mathbf{X}\mathbf{b})^T\mathbf{P}_{\mathbf{1}_N}\mathbf{X}\mathbf{b}\right)$$

$$\mathbf{y}^T\mathbf{A}_2\mathbf{y} = SSA_{cfm} = \sum_{i=1}^a n_i\bar{y}_{i\cdot} - N\bar{y}^2 \qquad \frac{SSA_{cfm}}{\sigma^2} \sim \chi^2_{a-1}\left(\frac{1}{2\sigma^2}(\mathbf{X}\mathbf{b})^T(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{1}_N})\mathbf{X}\mathbf{b}\right)$$

$$\mathbf{y}^T\mathbf{A}_3\mathbf{y} = SSE = \sum_{i=1}^a\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 \qquad \frac{SSE}{\sigma^2} \sim \chi^2_{N-a}\left(\frac{1}{2\sigma^2}(\mathbf{X}\mathbf{b})^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{X}\mathbf{b}\right) = \chi^2_{N-a}$$

ANOVA Table:

Source	DF	Projection	SS	Noncentrality
Mean	1	$\mathbf{P}_{\mathbf{1}_N}$	SSM=$N\bar{y}^2$	$\frac{1}{2\sigma^2}N(\mu+\bar{\alpha})^2$
Group	a-1	$\mathbf{P}_{\mathbf{X}}-\mathbf{P}_{\mathbf{1}_N}$	SSA$_{cfm}=\sum_i n_i\bar{y}_{i\cdot}^2-N\bar{y}^2$	$\frac{1}{2\sigma^2}\sum_i n_i(\alpha_i-\bar{\alpha})^2$
Error	N-a	$\mathbf{I}-\mathbf{P}_{\mathbf{X}}$	SSE=$\sum_{i,j}(y_{ij}-\bar{y}_{i\cdot})^2$	0

Chapter 6: Statistical Inference

6.2 Results from Statistical Theory

Assume $\mathbf{y}\sim\mathcal{N}_N(\mathbf{Xb},\sigma^2\mathbf{I})$.

$$\begin{aligned}f(\mathbf{y}|\mathbf{b},\sigma^2)&=(2\pi\sigma^2)^{-N/2}\exp\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{Xb})^T(\mathbf{y}-\mathbf{Xb})\}\\&=(2\pi\sigma^2)^{-N/2}\exp\{-\frac{1}{2\sigma^2}\mathbf{b}^T\mathbf{X}^T\mathbf{Xb}\}\exp\{-\frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{y}+\frac{1}{\sigma^2}\mathbf{b}^T\mathbf{X}^T\mathbf{y}\}\end{aligned}$$

By the factorization theorem, we see that $(\mathbf{y}^T\mathbf{y},\mathbf{X}^T\mathbf{y})$ is a sufficient statistic for $(\mathbf{b},\sigma^2)$.

Result 6.1: Assume $\mathbf{y}\sim\mathcal{N}_N(\mathbf{Xb},\sigma^2\mathbf{I})$. $(\mathbf{y}^T\mathbf{y},\mathbf{X}^T\mathbf{y})$ is a minimal sufficient statistic for $(\mathbf{b},\sigma^2)$.

proof: According to Theorem 6.2.B in C&B, it is enough to show that for two different responses, $\mathbf{y}_1,\mathbf{y}_2\in\mathbb{R}^N,\mathbf{y}_1\neq\mathbf{y}_2$,

$$\frac{f(\mathbf{y}_1|\mathbf{b},\sigma^2)}{f(\mathbf{y}_2|\mathbf{b},\sigma^2)}\text{ is constant in }(\mathbf{b},\sigma^2)\quad\Leftrightarrow\quad(\mathbf{y}_1^T\mathbf{y}_1,\mathbf{X}^T\mathbf{y}_1)=(\mathbf{y}_2^T\mathbf{y}_2,\mathbf{X}^T\mathbf{y}_2)$$

First

$$\frac{f(\mathbf{y}_1|\mathbf{b},\sigma^2)}{f(\mathbf{y}_2|\mathbf{b},\sigma^2)}=\exp\left\{\frac{1}{2\sigma^2}(\mathbf{y}_2^T\mathbf{y}_2-\mathbf{y}_1^T\mathbf{y}_1)+\frac{1}{\sigma^2}\mathbf{b}^T(\mathbf{X}^T\mathbf{y}_1-\mathbf{X}^T\mathbf{y}_2)\right\}$$

($\Leftarrow$) Obvious

($\Rightarrow$) $\forall(\mathbf{b},\sigma^2)\in\mathbb{R}^p\times\mathbb{R}^+$, $\frac{1}{2\sigma^2}(\mathbf{y}_2^T\mathbf{y}_2-\mathbf{y}_1^T\mathbf{y}_1)+\frac{1}{\sigma^2}\mathbf{b}^T(\mathbf{X}^T\mathbf{y}_1-\mathbf{X}^T\mathbf{y}_2)=c$ where c does not depend on $(\mathbf{b},\sigma^2)$. For fixed $\sigma^2>0$, $\frac{1}{2}(\mathbf{y}_2^T\mathbf{y}_2-\mathbf{y}_1^T\mathbf{y}_1)-c\sigma^2+\mathbf{b}^T(\mathbf{X}^T\mathbf{y}_1-\mathbf{X}^T\mathbf{y}_2)=0,\forall\mathbf{b}\in\mathbb{R}^p$. By Result A.8, $\mathbf{X}^T\mathbf{y}_1-\mathbf{X}^T\mathbf{y}_2=\mathbf{0}\Rightarrow$ $\frac{1}{2\sigma^2}(\mathbf{y}_2^T\mathbf{y}_2-\mathbf{y}_1^T\mathbf{y}_1)=c,\forall\sigma^2>0\Rightarrow\mathbf{y}_2^T\mathbf{y}_2-\mathbf{y}_1^T\mathbf{y}_1=0$.

Corollary 6.1: $\mathbf{y}\sim\mathcal{N}_N(\mathbf{Xb},\sigma^2\mathbf{I})$. $(SSE,\mathbf{X}^T\mathbf{y})$ is also minimal sufficient for $(\mathbf{b},\sigma^2)$.

Q: How does least squares related to M.L. (maximum likelihood)?

A: Almost the same.

Result 6.3: Assume $\mathbf{y} \sim \mathcal{N}_N(\mathbf{Xb}, \sigma^2\mathbf{I})$. Let $\hat{\mathbf{b}}$ be a solution to the N.E.s. $(\hat{\mathbf{b}}, SSE/N)$ is a ML estimator of $(\mathbf{b}, \sigma^2)$.

proof: Recall that $Q(\mathbf{b}) = (\mathbf{y} - \mathbf{Xb})^T(\mathbf{y} - \mathbf{Xb})$.

$$L(\mathbf{b}, \sigma^2|\mathbf{y}) = (2\pi)^{-N/2}(\sigma^2)^{-N/2}e^{-\frac{1}{2\sigma^2}Q(\mathbf{b})}$$

For any $\sigma^2 > 0$, $e^{-\frac{1}{2\sigma^2}Q(\mathbf{b})}$ is maximized by minimizing $Q(\mathbf{b})$. Of course, $Q(\mathbf{b})$ is minimized at $\hat{\mathbf{b}}$. So, we now can say for all $(\mathbf{b}, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}^+$,

$$L(\mathbf{b}, \sigma^2|\mathbf{y}) \leq L(\hat{\mathbf{b}}, \sigma^2|\mathbf{y})$$

Now

$$\log L(\hat{\mathbf{b}}, \sigma^2|\mathbf{y}) = -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}Q(\hat{\mathbf{b}}) + \text{constant}$$

$$\frac{d}{d\sigma^2}\log L(\hat{\mathbf{b}}, \sigma^2|\mathbf{y}) = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}Q(\hat{\mathbf{b}}) \overset{!}{=} 0$$

$$\Rightarrow \hat{\sigma^2} = \frac{SSE}{N}$$

$$\frac{d^2}{d(\sigma^2)^2}\log L(\hat{\mathbf{b}}, \sigma^2|\mathbf{y})\Big|_{\hat{\sigma^2}} < 0 \Rightarrow \text{ Maximum}$$

Finally, we can say: for all $(\mathbf{b}, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}^+$, $L(\mathbf{b}, \sigma^2|\mathbf{y}) \leq L(\hat{\mathbf{b}}, \sigma^2|\mathbf{y}) \leq L(\hat{\mathbf{b}}, \hat{\sigma^2}|\mathbf{y})$.

Corollary 6.3: Assume $\mathbf{y} \sim \mathcal{N}_N(\mathbf{Xb}, \sigma^2\mathbf{I})$. The ML estimator of an estimable function, $\boldsymbol{\lambda}^T\mathbf{b}$, is $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$ when $\hat{\mathbf{b}}$ solve the N.E.s.

proof: Invariance of MLE.

6.3 Testing the General Linear Hypothesis

$$H : \mathbf{K}^T\mathbf{b} = \mathbf{m} \qquad A : \mathbf{K}^T\mathbf{b} \neq \mathbf{m}$$

Assume:

- $\mathbf{K}_{p\times s}$ has rank s. If $\mathbf{K}$ is not full column rank, there will be redundant constrains.
- Components of $\mathbf{K}^T\mathbf{b}$ are estimable. In other words, each column of $\mathbf{K}$ is in $\mathcal{C}(\mathbf{X}^T)$. Thus $\mathbf{K} = \mathbf{X}^T\mathbf{A}$ for some $\mathbf{A}_{N\times s}$.

Example: Two-way crossed model with interaction

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \quad i = 1, 2, \cdots, a, j = 1, 2, \cdots, b, k = 1, 2, \cdots, n_{ij}$$

Q: How do we test for no interaction?

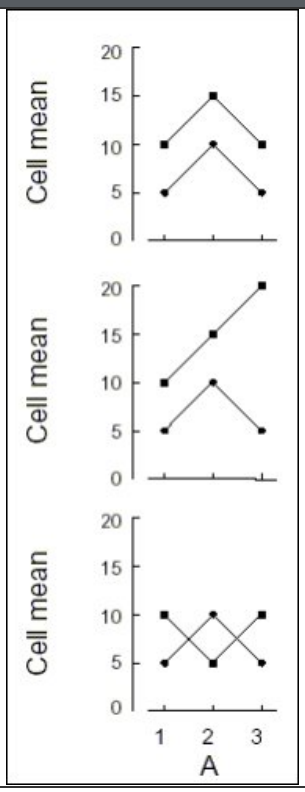
Maybe

$$\mathbf{K}^T\mathbf{b} = \begin{pmatrix}\mathbf{0}_{ab\times(a+b+1)} & \mathbf{I}_{ab}\end{pmatrix}\begin{pmatrix}\mu\\ \alpha_1\\ \vdots\\ \alpha_a\\ \beta_1\\ \vdots\\ \beta_b\\ \gamma_{11}\\ \vdots\\ \gamma_{ab}\end{pmatrix}, \quad \mathbf{m} = \mathbf{0}$$

Doesn't work since γ_{ij} is not estimable.

Let's think about a special case with $K = 1$, $a = 3$ and $b = 2$.

Q: What does no interaction really mean?



The difference in response between levels i and i' of factor A does not depend on factor B.

$$\mathbb{E}(y_{ij} - y_{i'j}) = \mathbb{E}(y_{ij'} - y_{i'j'}), \quad i \neq i', j \neq j'$$
$$\Rightarrow \mu + \alpha_i + \beta_j + \gamma_{ij} - (\mu + \alpha_{i'} + \beta_j + \gamma_{i'j}) = \mu + \alpha_i + \beta_{j'} + \gamma_{ij'} - (\mu + \alpha_{i'} + \beta_{j'} + \gamma_{i'j'})$$
$$\text{or } \gamma_{ij} - \gamma_{i'j} - \gamma_{ij'} + \gamma_{i'j'} = 0$$

$(i, i') = (1, 2), (i, i') = (1, 3), (i, i') = (2, 3)$

When $a = 3$ and $b = 2$, we only need 2 such equations to test the hypothesis on no interaction:

- $\gamma_{11} - \gamma_{21} - \gamma_{12} + \gamma_{22} = 0$
- $\gamma_{11} - \gamma_{31} - \gamma_{12} + \gamma_{32} = 0$

Subtract first from second, you get $\gamma_{21} - \gamma_{31} - \gamma_{22} + \gamma_{32} = 0$.

In general, we need $(a-1)(b-1)$ equations to test for no interaction. This makes sense as the rank of $\mathbf{X}$ in the full model is ab, and in the model without interaction, the rank of $\mathbf{X}$ is $a + b - 1$, and $ab - (a + b - 1) = (a-1)(b-1)$.

Back to the general case.

Let's use "first principles" (or "do what makes sense") to construct a statistical test for $H : \mathbf{K}^T\mathbf{b} = \mathbf{m}$.

The BLUE of $\mathbf{K}^T\mathbf{b}$ is $\mathbf{K}^T\hat{\mathbf{b}} = \mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y}$.

$$\mathbf{K}^T\hat{\mathbf{b}} \sim \mathcal{N}_s\left(\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}\mathbf{b}, \sigma^2\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\mathbf{K}\right)$$

But $\left[\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}\right]^T = \underbrace{(\mathbf{X}^T\mathbf{X})\left[(\mathbf{X}^T\mathbf{X})^g\right]^T}_{\text{projection onto } \mathcal{C}(\mathbf{X}^T\mathbf{X}) = \mathcal{C}(\mathbf{X}^T)}\ \mathbf{K} = \mathbf{K}$. So

$$\mathbf{K}^T\hat{\mathbf{b}} \sim \mathcal{N}_s\left(\mathbf{K}^T\mathbf{b}, \sigma^2\underbrace{\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}}_{\mathbf{H}}\right)$$

Result 6.4: The $s \times s$ matrix $\mathbf{H} = \mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}$ is nonsingular, and hence positive definite.

proof: $\mathbf{K}_{p\times s} = \mathbf{X}^T_{p\times N}\mathbf{A}_{N\times s}$, $rank(\mathbf{X}) = r$. Now,
$\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K} = \mathbf{A}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{A} = \mathbf{A}^T\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X}\mathbf{A} = (\mathbf{P}_\mathbf{X}\mathbf{A})^T\mathbf{P}_\mathbf{X}\mathbf{A}$. If $\mathbf{P}_\mathbf{X}\mathbf{A}_{N\times s}$ has full rank, then
$(\mathbf{P}_\mathbf{X}\mathbf{A})^T\mathbf{P}_\mathbf{X}\mathbf{A}$ is invertible, which is what we want.

Now, $\mathbf{X}^T\mathbf{P}_\mathbf{X}\mathbf{A} = \mathbf{X}^T\mathbf{A} = \mathbf{K}$. $s = rank(\mathbf{K}) \leq \min\{rank(\mathbf{X}^T), rank(\mathbf{P}_\mathbf{X}\mathbf{A}\} \leq rank(\mathbf{P}_\mathbf{X}\mathbf{A}) \Rightarrow rank(\mathbf{P}_\mathbf{X}\mathbf{A}) = s$.

$$\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m} \sim \mathcal{N}_N(\mathbf{K}^T\mathbf{b} - \mathbf{m}, \sigma^2\mathbf{H})$$

Result 5.10 $\Rightarrow$

$$(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m})^T(\sigma^2\mathbf{H})^{-1}(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m}) \sim \chi^2_s(\phi)$$

where $\phi = \frac{1}{2}(\mathbf{K}^T\mathbf{b} - \mathbf{m})(\sigma^2\mathbf{H})^{-1}(\mathbf{K}^T\mathbf{b} - \mathbf{m})$.

Recall that $\mathbf{K} = \mathbf{X}^T\mathbf{A}$, so $\mathbf{K}^T\hat{\mathbf{b}} = \mathbf{A}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y} = \mathbf{A}^T\mathbf{P}_\mathbf{X}\mathbf{y}$. But we know that $\mathbf{P}_\mathbf{X}\mathbf{y}$ and $(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{y}$ are independent, so $\mathbf{K}^T\hat{\mathbf{b}}$ and SSE are also independent.

Finally,

$$F = \frac{(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m})^T\mathbf{H}^{-1}(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m})/s}{SSE/(N-r)} \sim F_{s,N-r}(\phi).$$

Under $H : \mathbf{K}^T\mathbf{b} = \mathbf{m}$, $F \sim F_{s,N-r}$.

Also, we know from Result 5.13 that F is stochastically increasing in ϕ. It therefore makes sense to reject when F is large.

Let's reject when $F > F_{s,N-r,\alpha}$. This is level α test, and it is an unbiased because $P(rejecting)$ is the smallest when $\phi = 0$.

Example 6.5: One-way ANOVA

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, 2, \cdots, a, j = 1, 2, \cdots, n_i$$

$H : \alpha_1 - \alpha_2 = 0, \alpha_1 - \alpha_3 = 0, \cdots, \alpha_1 - \alpha_a = 0.$

$$\mathbf{K}^T = \begin{pmatrix} 0 & 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & 0 & \cdots & -1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix}, \quad \mathbf{m} = \mathbf{0}.$$

If H is true, then $\alpha_1 = \alpha_2 = \cdots = \alpha_a$.

Note that we could have specified H in a different way, e.g., $H_* = \alpha_1 - \alpha_2 = 0, \alpha_2 - \alpha_3 = 0, \cdots, \alpha_{a-1} - \alpha_a = 0$.

Let's construct the F statistic for H. Here $N = \sum_{i=1}^a n_i$, $p = a + 1$, $s = a - 1$.

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_a} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_a} \end{pmatrix}, \quad (\mathbf{X}^T\mathbf{X})^g = \begin{pmatrix} 0 & & & & \\ & \frac{1}{n_1} & & & \\ & & \frac{1}{n_2} & & \\ & & & \ddots & \\ & & & & \frac{1}{n_a} \end{pmatrix} = \begin{pmatrix} 0 & & \\ & \frac{1}{n_1} & \\ & & \mathbf{D}_* \end{pmatrix}$$

This yields

$$\hat{\mathbf{b}} = \begin{pmatrix} 0 \\ \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \vdots \\ \bar{y}_{a\cdot} \end{pmatrix}$$

$$\mathbf{K}\hat{\mathbf{b}} - \mathbf{m} = \begin{pmatrix} \mathbf{0} & \mathbf{1}_{a-1} & -\mathbf{I}_{a-1} \end{pmatrix} \begin{pmatrix} 0 \\ \bar{y}_{1\cdot} \\ \vdots \\ y_{a\cdot} \end{pmatrix} = \begin{pmatrix} \bar{y}_{1\cdot} - \bar{y}_{2\cdot} \\ \bar{y}_{1\cdot} - \bar{y}_{3\cdot} \\ \vdots \\ \bar{y}_{1\cdot} - \bar{y}_{a\cdot} \end{pmatrix}$$

$$\begin{aligned} \mathbf{H} = \mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K} &= \begin{pmatrix} \mathbf{0} & \mathbf{1}_{a-1} & -\mathbf{I}_{a-1} \end{pmatrix} \begin{pmatrix} 0 & & \\ & \frac{1}{n_1} & \\ & & \mathbf{D}_* \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{1}_{a-1}^T \\ -\mathbf{I}_{a-1} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{0} & \frac{1}{n_1}\mathbf{1}_{a-1} & -\mathbf{D}_* \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{1}_{a-1}^T \\ -\mathbf{I}_{a-1} \end{pmatrix} \\ &= \mathbf{D}_* + \frac{1}{n_1}\mathbf{1}_{a-1}\mathbf{1}_{a-1}^T \end{aligned}$$

Using the result from Problem A.75:

$$\left[\mathbf{D}_* + \frac{1}{n_1}\mathbf{1}_{a-1}\mathbf{1}_{a-1}^T\right]^{-1} = \mathbf{D}_*^{-1} - \frac{1}{N}\mathbf{D}_*^{-1}\mathbf{1}_{a-1}\mathbf{1}_{a-1}^T\mathbf{D}_*^{-1}$$

$$\begin{aligned} &(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m})^T[\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}]^{-1}(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m}) \\ =&(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m})^T\mathbf{D}_*^{-1}(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m}) - \frac{1}{N}(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m})^T\mathbf{D}_*^{-1}\mathbf{1}_{a-1}\mathbf{1}_{a-1}^T\mathbf{D}_*^{-1}(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m}) \\ =&\sum_{i=2}^{a} n_i(\bar{y}_{1\cdot} - \bar{y}_{i\cdot})^2 - \frac{1}{N}\left[\sum_{i=2}^{a} n_i(\bar{y}_{1\cdot} - \bar{y}_{i\cdot})\right]^2 \\ =&\sum_{i=2}^{a} n_i(\bar{y}_{1\cdot} - \bar{y}_{i\cdot})^2 - \frac{1}{N}[N(\bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot})]^2 \\ =&\cdots \\ =&\sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \qquad \left(\text{Using } \sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y}_{1\cdot} + \bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot})^2 = \cdots\right) \end{aligned}$$

Clearly, $SSE = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$. So our F-statistic is

$$F = \frac{\left(\sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2\right)/(a-1)}{\left(\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2\right)/(N-a)} \overset{\text{under } H}{\sim} F_{a-1,N-a}$$

Recall that we could have used H_* instead of H.

Q: Will we get the same test?

A: Yes.

In general, suppose that we have two equivalent hypothesis: $H : \mathbf{K}^T\mathbf{b} = \mathbf{m}, \quad H_* : \mathbf{K}_*^T\mathbf{b} = \mathbf{m}_*$.

Equivalence $\Rightarrow S = \left\{\mathbf{b} \in \mathbb{R}^p : \mathbf{K}^T\mathbf{b} = \mathbf{m}\right\} = \left\{\mathbf{b} \in \mathbb{R}^p : \mathbf{K}_*^T\mathbf{b} = \mathbf{m}_*\right\} = S_*$

Result A.13 implies that points in S can be expressed as

$$\mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{m} + (\mathbf{I} - \mathbf{P}_\mathbf{K})\mathbf{z}, \quad \mathbf{z} \in \mathbb{R}^p$$

Since $S = S_*$, we have

$$\begin{gathered}\mathbf{K}_*^T\left[\mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{m} + (\mathbf{I} - \mathbf{P}_\mathbf{K})\mathbf{z}\right] = \mathbf{m}_*, \quad \forall \mathbf{z} \in \mathbb{R}^p \\ \text{or } \mathbf{K}_*^T\mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{m} - \mathbf{m}_* + \mathbf{K}_*^T(\mathbf{I} - \mathbf{P}_\mathbf{K})\mathbf{z} = \mathbf{0}, \quad \forall \mathbf{z} \in \mathbb{R}^p\end{gathered}$$

Result A.8 implies that

- $\mathbf{K}_*^T(\mathbf{I} - \mathbf{P}_\mathbf{K}) = \mathbf{0}$, or $\mathbf{P}_\mathbf{K}\mathbf{K}_* = \mathbf{K}_*$
- $\mathbf{K}_*^T\mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{m} = \mathbf{m}_*$

$\mathbf{P}_\mathbf{K}\mathbf{K}_* = \mathbf{K}_* \Rightarrow \mathcal{C}(\mathbf{K}_*) \subseteq \mathcal{C}(\mathbf{K})$. If we reverse the roles of $\mathbf{K}$ and $\mathbf{K}_*$ in the previous argument, we will get $\mathcal{C}(\mathbf{K}) \subseteq \mathcal{C}(\mathbf{K}_*)$.

So $\mathcal{C}(\mathbf{K}) = \mathcal{C}(\mathbf{K}_*) \Rightarrow \exists$ a nonsingular matrix $\mathbf{Q}$ such that $\mathbf{K}_{p\times s}\mathbf{Q}_{s\times s}^T = \mathbf{K}_{*p\times s}$.

It follows from $\mathbf{K}_*^T\mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{m} = \mathbf{m}_*$ that

$$\mathbf{Q}\mathbf{K}^T\mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{m} = \mathbf{m}_* \Rightarrow \mathbf{Q}\mathbf{m} = \mathbf{m}_*$$

Also,

$$\mathbf{K}\mathbf{Q}^T = \mathbf{K}_* \Rightarrow \mathbf{K}^T\mathbf{K}\mathbf{Q}^T = \mathbf{K}^T\mathbf{K}_* \Rightarrow \mathbf{Q}^T = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{K}_* \Rightarrow \mathbf{Q} = \mathbf{K}_*^T\mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}.$$

Now, let's look at the numerator of the F-statistic for testing $H_* : \mathbf{K}_*^T\mathbf{b} = \mathbf{m}_*$:

$$\begin{aligned}&(\mathbf{K}_*^T\hat{\mathbf{b}} - \mathbf{m}_*)^T\left[\mathbf{K}_*^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}_*\right]^{-1}(\mathbf{K}_*^T\hat{\mathbf{b}} - \mathbf{m}_*) \\ =&(\mathbf{Q}\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{Q}\mathbf{m})^T\left[\mathbf{Q}\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}\mathbf{Q}^T\right]^{-1}(\mathbf{Q}\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{Q}\mathbf{m}) \\ =&(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m})\left[\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^T\mathbf{K}\right]^{-1}(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m})\end{aligned}$$

What happens when $s = 1$ so that $\mathbf{K}^T$ is a $1 \times p$ vector?

$H : \mathbf{K}^T\mathbf{b} = \mathbf{m}$

Now we can do both one-sided and two-sided test: $A_1 : \mathbf{K}^T\mathbf{b} > \mathbf{m} \quad A_2 : \mathbf{K}^T\mathbf{b} \neq \mathbf{m}$.

$$\mathbf{K}^T\hat{\mathbf{b}} \sim \mathcal{N}\left(\mathbf{K}^T\mathbf{b}, \sigma^2\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}\right)$$

$$\Rightarrow \frac{\frac{\mathbf{K}^T\hat{\mathbf{b}}-\mathbf{m}}{\sqrt{\sigma^2\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}}}}{\sqrt{\frac{SSE}{\sigma^2(N-r)}}} = \frac{\mathcal{N}(\mu,1)}{\sqrt{\frac{\chi^2_{N-r}}{N-r}}}$$

$$\Rightarrow t = \frac{\mathbf{K}^T\hat{\mathbf{b}}-\mathbf{m}}{\sqrt{\hat{\sigma^2}\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}}} \sim t_{N-r}(\mu), \quad \mu = \frac{\mathbf{K}^T\mathbf{b}-\mathbf{m}}{\sqrt{\sigma^2\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}}}$$

For A_1, we reject H if $t > t_{N-r,\alpha}$

For A_2, we reject H if $|t| > t_{N-r,\alpha/2}$

Example (Full and reduced models):

Let's return to the test for no interaction in the two-way model. $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$, $i = 1, 2, \cdots, a$, $j = 1, 2, \cdots, b$, $k = 1, 2, \cdots, n$.

$$\mathbf{Xb} = \begin{pmatrix} \mathbf{1}_n & \mathbf{1}_n & \mathbf{0} & \mathbf{0} & \mathbf{1}_n & \mathbf{0} & \mathbf{1}_n & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_n & \mathbf{1}_n & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_n & \mathbf{0} & \mathbf{1}_n & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}_n & \mathbf{0} & \mathbf{0} & \mathbf{1}_n & \mathbf{0} & \mathbf{1}_n & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_n \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \vdots \\ \gamma_{32} \end{pmatrix}$$

$$\mathbb{E}\mathbf{y} = \mathbf{Xb} = \begin{pmatrix} \mathbf{X}_0 & \mathbf{X}_1 \end{pmatrix} \begin{pmatrix} \mathbf{b}_0 \\ \mathbf{b}_1 \end{pmatrix} = \mathbf{X}_0\mathbf{b}_0 + \mathbf{X}_1\mathbf{b}_1$$

where $\mathbf{X}_0$ consists of the first 6 columns of $\mathbf{X}$.

We would like to test $H_1 : \mathbf{X}_1\mathbf{b}_1 = \mathbf{0}$, but components of $\mathbf{X}_1\mathbf{b}_1$ are not estimable.

Recall the alternative parametrization:

$$\mathbf{y} = \mathbf{X}_0\mathbf{c}_0 + (\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1\mathbf{c}_1 + \mathbf{e}$$

Note that

$$\mathbb{E}[(\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}] = (\mathbf{I} - \mathbf{P}_{\mathbf{X}_0}) \begin{pmatrix} \mathbf{X}_0 & (\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1 \end{pmatrix} \begin{pmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \end{pmatrix} = \begin{pmatrix} \mathbf{0} & (\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1 \end{pmatrix} \begin{pmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \end{pmatrix} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1\mathbf{c}_1$$

Thus the components of $(\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1\mathbf{c}_1$ are estimable.

$H : (\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1\mathbf{c}_1 = \mathbf{0}$.

Problem: $(\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1$ may not be full rank.

Let $rank(\mathbf{X}) = r$, $rank(\mathbf{X}_0) = r_0$, and $rank((\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1) = s = r - r_0$. Just pick out s L.I. estimable functions from $(\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1\mathbf{c}_1$ and perform the corresponding F-test.

6.4 The Likelihood Ratio Test (LRT)

$\mathbf{y} \sim \mathcal{N}_N(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I})$

Want to test $H : \mathbf{K}^T\mathbf{b} = \mathbf{m}$, alternative $A : \mathbf{K}^T\mathbf{b} \neq \mathbf{m}$

Full parameter space: $\Omega = \{(\mathbf{b}, \sigma^2) : \mathbf{b} \in \mathbb{R}^p, \sigma^2 > 0\}$

Reduced parameter space: $\Omega_0 = \{(\mathbf{b}, \sigma^2) : \mathbf{b} \in \mathbb{R}^p, \mathbf{K}^T\mathbf{b} = \mathbf{m}, \sigma^2 > 0\}$

Likelihood function: $L(\mathbf{b}, \sigma^2|\mathbf{y}) = (2\pi\sigma^2)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2}Q(\mathbf{b})}$, where $Q(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})$

LRT statistic:

$$\phi(\mathbf{y}) = \frac{\sup_{(\mathbf{b},\sigma^2)\in\Omega_0} L(\mathbf{b}, \sigma^2|\mathbf{y})}{\sup_{(\mathbf{b},\sigma^2)\in\Omega} L(\mathbf{b}, \sigma^2|\mathbf{y})}$$

Reject when $\phi(\mathbf{y})$ is "small".

Recall that we can maximize $L(\mathbf{b}, \sigma^2|\mathbf{y})$ sequentially. First, w.r.t. $\mathbf{b}$, and then w.r.t. σ^2. Suppose we have the minimizers of $Q(\mathbf{b})$: $Q(\hat{\mathbf{b}})$ and $Q(\hat{\mathbf{b}}_H)$.

Now, we know for fixed $Q(\mathbf{b})$, the maximizer of $L(\mathbf{b}, \sigma^2|\mathbf{y})$ in σ^2 is $\hat{\sigma^2} = \frac{Q(\mathbf{b})}{N}$. Thus,

$$\phi(\mathbf{y}) = \frac{\left(2\pi\frac{Q(\hat{\mathbf{b}}_H)}{N}\right)^{-\frac{N}{2}} e^{-\frac{1}{2Q(\hat{\mathbf{b}}_H)/N}Q(\hat{\mathbf{b}}_H)}}{\left(2\pi\frac{Q(\hat{\mathbf{b}})}{N}\right)^{-\frac{N}{2}} e^{-\frac{1}{2Q(\hat{\mathbf{b}}_H)/N}Q(\hat{\mathbf{b}})}} = \left(\frac{Q(\hat{\mathbf{b}}_H)}{Q(\hat{\mathbf{b}})}\right)^{-\frac{N}{2}}$$

$$\begin{aligned}\left(\frac{Q(\hat{\mathbf{b}}_H)}{Q(\hat{\mathbf{b}})}\right)^{-\frac{N}{2}} < c &\Leftrightarrow \frac{Q(\hat{\mathbf{b}}_H)}{Q(\hat{\mathbf{b}})} > c^{-\frac{N}{2}} \\ &\Leftrightarrow \frac{Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}})}{Q(\hat{\mathbf{b}})} > c^{-\frac{N}{2}} - 1 \\ &\Leftrightarrow \frac{(Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}}))/s}{Q(\hat{\mathbf{b}})/(N-r)} > \frac{N-r}{s}(c^{-\frac{N}{2}} - 1)\end{aligned}$$

Example 6.5 (One-way ANOVA):

$Q(\hat{\mathbf{b}}) = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2 = \mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$.

How do we minimize $Q(\mathbf{b})$ subject to $\alpha_1 = \alpha_2 = \cdots = \alpha_a$?

$$\begin{aligned}\inf_{(\mathbf{b},\sigma^2)\in\Omega_0}\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - (\mu + \alpha_{ij}))^2 &= \inf_{\mu,\alpha\in\mathbb{R}}\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - (\mu+\alpha))^2 \\ &= \inf_{\delta\in\mathbb{R}}\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \delta)^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2\end{aligned}$$

Then $Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}}) = \sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$.

$$\frac{(Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}}))/s}{Q(\hat{\mathbf{b}})/(N-r)} = \frac{\sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2/(a-1)}{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2/(N-a)} = F \overset{H}{\sim} F_{a-1,N-a}$$

Full v.s. Reduced Models:

$$\mathbb{E}\mathbf{y} = \mathbf{X}\mathbf{b} = \begin{pmatrix}\mathbf{X}_0 & \mathbf{X}_1\end{pmatrix}\begin{pmatrix}\mathbf{b}_0 \\ \mathbf{b}_1\end{pmatrix} = \mathbf{X}_0\mathbf{b}_0 + \mathbf{X}_1\mathbf{b}_1$$

Want to test: $H : \mathbf{X}_1\mathbf{b}_1 = \mathbf{0}$, but components may not be estimable.

$$\mathbf{y} = \mathbf{X}_0\mathbf{c}_0 + (\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1\mathbf{c}_1 + \mathbf{e}$$

$rank(\mathbf{X}) = r$, $rank(\mathbf{X}_0) = r_0$, and $rank((\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1) = s = r - r_0$

How do we use this test with the LRT?

$$\begin{aligned}Q(\hat{\mathbf{b}}_H) &= \mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} = SSE(\text{reduced}) \\ Q(\hat{\mathbf{b}}) &= \mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y} = SSE(\text{full}) \\ \Rightarrow Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}}) &= \mathbf{y}^T(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}\end{aligned}$$

Now,

$$\underbrace{\mathbf{P}_{\mathbf{X}_0}}_{\text{rank } r} + \underbrace{\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_0}}_{\text{rank } s=r-r_0} + \underbrace{\mathbf{I} - \mathbf{P}_{\mathbf{X}}}_{\text{rank } N-r} = \mathbf{I}$$

All these are projections.

By Cochran's Theorem, $\mathbf{y}^T\mathbf{P}_{\mathbf{X}_0}\mathbf{y}/\sigma^2$, $\mathbf{y}^T(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}/\sigma^2$, and $\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}/\sigma^2$ are independent noncentral χ^2's with noncentrality parameters $(\mathbf{X}\mathbf{b})^T\mathbf{P}_{\mathbf{X}_0}(\mathbf{X}\mathbf{b})/2\sigma^2$, $(\mathbf{X}\mathbf{b})^T(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_0})(\mathbf{X}\mathbf{b})/2\sigma^2$ and 0.

$$\frac{[Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}})]/s}{Q(\hat{\mathbf{b}})/(N-r)} = \frac{\mathbf{y}^T(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}/s}{\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}/(N-r)} = F \sim F_{s,N-r}(\phi), \quad \phi = \frac{(\mathbf{X}\mathbf{b})^T(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_0})(\mathbf{X}\mathbf{b})}{2\sigma^2}$$

It is easy to see that if $H : \mathbf{X}_1\mathbf{b}_1 = \mathbf{0}$, then $\phi = 0$.

6.5 First Principles Test and LRT

Theorem 6.1: Let $\mathbf{K}^T\mathbf{b}$ be a set of linearly independent estimable functions. Also, let $\hat{\mathbf{b}}_H$ be part of a solution to the R.N.s with constraint $\mathbf{K}^T\mathbf{b} = \mathbf{m}$ (instead of $\mathbf{P}^T\mathbf{b} = \boldsymbol{\delta}$). Then

$$Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}}) = (\hat{\mathbf{b}}_H - \hat{\mathbf{b}})^T\mathbf{X}^T\mathbf{X}(\hat{\mathbf{b}}_H - \hat{\mathbf{b}}) = (\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m})^T\left[\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}\right]^{-1}(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m})$$

(That is, the two tests are the same)

proof:

$$\begin{aligned} Q(\hat{\mathbf{b}}_H) &= (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_H)^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_H) \\ &= (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} + \mathbf{X}\hat{\mathbf{b}} - \mathbf{X}\hat{\mathbf{b}}_H)^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} + \mathbf{X}\hat{\mathbf{b}} - \mathbf{X}\hat{\mathbf{b}}_H) \\ &= Q(\hat{\mathbf{b}}) + (\hat{\mathbf{b}} - \hat{\mathbf{b}}_H)^T\mathbf{X}^T\mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}_H) + 2(\hat{\mathbf{b}} - \hat{\mathbf{b}}_H)^T\underbrace{\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})}_{=\mathbf{0}} \end{aligned}$$

R.N.E.s:

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{K} \\ \mathbf{K}^T & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{b} \\ \boldsymbol{\theta} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{m} \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}_H) = \mathbf{X}^T\mathbf{y} - (\mathbf{X}^T\mathbf{y} - \mathbf{K}\hat{\boldsymbol{\theta}}_H) = \mathbf{K}\hat{\boldsymbol{\theta}}_H \quad (\star)$$

Premultiplying $(\star)$ by $\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g$ yields

$$\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}_H) = \mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}\hat{\boldsymbol{\theta}}_H$$

But $\left[\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}\right]^T = \mathbf{X}^T\mathbf{X}[(\mathbf{X}^T\mathbf{X})^g]^T\mathbf{K} = \mathbf{K}$ since columns of $\mathbf{K}$ are in $\mathcal{C}(\mathbf{X}^T)$. Thus,

$$\begin{aligned} \mathbf{K}^T(\hat{\mathbf{b}} - \hat{\mathbf{b}}_H) &= \mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}\hat{\boldsymbol{\theta}}_H \\ \Rightarrow \left[\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}\right]^{-1}\mathbf{K}^T(\hat{\mathbf{b}} - \hat{\mathbf{b}}_H) &= \hat{\boldsymbol{\theta}}_H. \end{aligned}$$

Now, premultiplying $(\star)$ by $(\hat{\mathbf{b}} - \hat{\mathbf{b}}_H)^T$ yields

$$\begin{aligned} Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}}) &= (\hat{\mathbf{b}} - \hat{\mathbf{b}}_H)^T\mathbf{X}^T\mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}_H) \\ &= (\hat{\mathbf{b}} - \hat{\mathbf{b}}_H)^T\mathbf{K}\hat{\boldsymbol{\theta}}_H \\ &= (\hat{\mathbf{b}} - \hat{\mathbf{b}}_H)^T\mathbf{K}\left[\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}\right]^{-1}\mathbf{K}^T(\hat{\mathbf{b}} - \hat{\mathbf{b}}_H) \\ &= (\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{K}^T\hat{\mathbf{b}}_H)^T\left[\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}\right]^{-1}(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{K}^T\hat{\mathbf{b}}_H) \\ &= (\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m})^T\left[\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}\right]^{-1}(\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m}) \end{aligned}$$

Corollary 6.4: Let $\mathbf{K}^T\mathbf{b}$ be a set of L.I. estimable functions and $\hat{\mathbf{b}}$ a solution to the N.E.s. We can find $\hat{\mathbf{b}}_H$ by solving for $\mathbf{b}$ in the following equation:

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y} - \mathbf{K}\left[\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}\right]^{-1}[\mathbf{K}^T\hat{\mathbf{b}} - \mathbf{m}]$$

proof: From R.N.E.s, we have

$$\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_H + \mathbf{K}\hat{\boldsymbol{\theta}}_H = \mathbf{X}^T\mathbf{y}$$

But we know from the proof of Theorem 6.1 that

$$\hat{\boldsymbol{\theta}}_H = \left[\mathbf{K}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{K}\right]^{-1}\mathbf{K}^T(\hat{\mathbf{b}} - \hat{\mathbf{b}}_H)$$

Now plug in, and rearrange.

Scheffe method for dealing with multiple comparisons

Suppose $\mathbf{A}$ is a positive definite matrix and let $\mathbf{w}$ be a vector. We want

$$\max_{\mathbf{x}\neq 0} \frac{(\mathbf{x}^T\mathbf{w})^2}{\mathbf{x}^T\mathbf{A}\mathbf{x}}$$

$$\frac{(\mathbf{x}^T\mathbf{w})^2}{\mathbf{x}^T\mathbf{A}\mathbf{x}} = \frac{\langle \mathbf{A}^{1/2}\mathbf{x}, \mathbf{A}^{-1/2}\mathbf{w}\rangle^2}{\|\mathbf{A}^{1/2}\mathbf{x}\|^2} \overset{\text{C-S}}{\leq} \|\mathbf{A}^{-1/2}\mathbf{w}\|^2$$

but we attain equality when $\mathbf{A}^{1/2}\mathbf{x} = \mathbf{A}^{-1/2}\mathbf{w}$.

$$\Rightarrow \max_{\mathbf{x}\neq 0} \frac{(\mathbf{x}^T\mathbf{w})^2}{\mathbf{x}^T\mathbf{A}\mathbf{x}} = \mathbf{w}^T\mathbf{A}^{-1}\mathbf{w} \quad (\star\star)$$

Now suppose that $\tau = \boldsymbol{\Lambda}^T\mathbf{b}$ where $\boldsymbol{\Lambda}_{p\times s}$ has L.I. columns in $\mathcal{C}(\mathbf{X}^T)$. Define

$$R(\mathbf{u}, c) = \left(\mathbf{u}^T\hat{\tau} - c\hat{\sigma}\sqrt{\mathbf{u}^T\mathbf{H}\mathbf{u}}, \mathbf{u}^T\hat{\tau} + c\hat{\sigma}\sqrt{\mathbf{u}^T\mathbf{H}\mathbf{u}}\right)$$

where $\mathbf{H} = \boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\Lambda}$, $\hat{\sigma}^2 = SSE/(N-r)$, $\mathbf{u} \in \mathbb{R}^s$, $c > 0$.

Goal: Find $c > 0$, s.t.

$$P\left(\mathbf{u}^T\tau \in R(\mathbf{u}, c), \forall \mathbf{u} \in \mathbb{R}^s\right) = 1 - \alpha.$$

If the above holds, we can form individual C.I.s for as many linear combinations of the τ_j's as we like, and the whole collection will hold simultaneously at level α, or with probability $1-\alpha$.

First

$$P(\mathbf{u}^T\tau \in R(\mathbf{u}, c)) = P\left(\left|\frac{\mathbf{u}^T\tau - \mathbf{u}^T\hat{\tau}}{\hat{\sigma}\sqrt{\mathbf{u}^T\mathbf{H}\mathbf{u}}}\right| \leq c\right).$$

Thus,

$$
\begin{aligned}
P(\mathbf{u}^T\tau \in R(\mathbf{u}, c), \forall \mathbf{u} \in \mathbb{R}^s) &= P\left(\max_{\mathbf{u}\in\mathbb{R}^s, \mathbf{u}\neq\mathbf{0}} \left|\frac{\mathbf{u}^T\tau - \mathbf{u}^T\hat{\tau}}{\hat{\sigma}\sqrt{\mathbf{u}^T\mathbf{H}\mathbf{u}}}\right| \leq c\right) \\
&= P\left(\max_{\mathbf{u}\in\mathbb{R}^s, \mathbf{u}\neq\mathbf{0}} \frac{[\mathbf{u}^T(\tau - \hat{\tau})]^2}{\hat{\sigma}^2\mathbf{u}^T\mathbf{H}\mathbf{u}} \leq c^2\right) \\
&\overset{(\star\star)}{=} P\left((\hat{\tau} - \tau)^T\mathbf{H}^{-1}(\hat{\tau} - \tau)/\hat{\sigma}^2 \leq c^2\right) \\
&= P\left((\hat{\tau} - \tau)^T\mathbf{H}^{-1}(\hat{\tau} - \tau)/s\hat{\sigma}^2 \leq \frac{c^2}{s}\right)
\end{aligned}
$$

But

$$
\frac{(\hat{\tau} - \tau)^T\mathbf{H}^{-1}(\hat{\tau} - \tau)/s}{\hat{\sigma}^2} \sim F_{s,N-r}
$$

So, all we need to do is setting $\frac{c^2}{s} = F_{s,N-r,\alpha}$ so that $P(\mathbf{u}^T\tau \in R(\mathbf{u}, c), \forall \mathbf{u} \in \mathbb{R}^s) = 1 - \alpha$. This yields

$$
c = \sqrt{sF_{s,N-r,\alpha}}
$$