250A Linear Statistical Models A Review

Tomoki Okuno

Summer 2023

Contents

1 Span

- Given a vector space V over a field K, the span of $S \subseteq V$ can be defined as the set of all finite linear combinations of elements of S:

$$\operatorname{span}(S) = \left\{ \sum_{i=1}^{k} \lambda_i \mathbf{v}_i \;\middle|\; \mathbf{v}_i \in S, \lambda_i \in K \right\},$$

  which is a subspace of V. Clearly, $S \subset \operatorname{span}(S)$. We say S spans V.

2 Basis

- If $\mathbf{x}$ can be expressed as a linear combination: $\mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$, then $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is called a basis. A basis is not unique. For example, the followings both are a basis of $V = \mathbb{R}^3$.

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}, \quad \left\{ \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

- If $x_1, \ldots, x_k$ $(k < n)$ are linearly independent vectors, then they can be extended to form a basis for the n-dimensional vector space of V.

3 Subspace

- If S and T are subspaces of V, then $S \cap T$ (intersection) and $S + T = \{s + t \mid s \in S, t \in T\}$ are also subspaces of V. However, $S \cup T = \{s \text{ or } t \mid s \in S, t \in T\}$ (union) is not always a subspace of V.
- $S^\perp = \{\mathbf{v} \in V \mid (\mathbf{v}, \mathbf{s}) = 0, \ \forall \mathbf{s} \in S\}$ is a vector space (subspace). *Proof*: Let $\mathbf{v}_1, \mathbf{v}_2 \in S^\perp$ and $\alpha \in \mathbb{R}$,

$$\begin{aligned} (\alpha \mathbf{v}_1 + \mathbf{v}_2, \mathbf{s}) = \alpha(\mathbf{v}_1, \mathbf{s}) + (\mathbf{v}_2, \mathbf{s}) = 0 \;&\Rightarrow\; \alpha \mathbf{v}_1 + \mathbf{v}_2 \in S^\perp, \\ (\mathbf{0}, \mathbf{s}) = 0 \;&\Rightarrow\; \mathbf{0} \in S^\perp. \end{aligned}$$

- $N(\mathbf{A})$ is a vector space (subspace). *Proof*: Let $\mathbf{x}, \mathbf{y} \in N(\mathbf{A})$ and $\alpha \in \mathbb{R}$, then $\mathbf{A}(\alpha \mathbf{x} + \mathbf{y}) = 0 \Rightarrow \alpha \mathbf{x} + \mathbf{y} \in N(\mathbf{A})$ and $\mathbf{A0} = \mathbf{0} \Rightarrow \mathbf{0} \in N(\mathbf{A})$.

4 Inner product

- A vector space V is an *inner product space* if it is endowed with an inner product defined as $V \times V \to \mathbb{R}$, and has the following properties: For $x, y, z \in V$ and $\alpha, \beta \in \mathbb{R}$,
  (i) Symmetry: $(x, y) = (y, x)$
  (ii) Linearity: $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$
  (iii) Non-negative: $(x, x) \geq 0$ with equality if and only if $x = 0$.
- If $x_1, \ldots, x_n$ are orthogonal vectors in V with an inner product $(\cdot, \cdot)$, then they are linearly independent.

  Proof: Suppose $\sum_{i=1}^{n} \alpha_i x_i = 0$. Then

$$0 = (0, x_j) = \left(\sum_{i=1}^{n} \alpha_i x_i, x_j \right) = \alpha_j \|x_j\|^2 \quad \Rightarrow \quad \alpha_j = 0, \quad j = 1, \ldots, n.$$

- Even if x and y are orthogonal, they are not always linearly independent as either one can be zero.

- **Cauchy–Schwarz inequality**: $(x,y)^2 \le \|x\|^2\|y\|^2 \Leftrightarrow |(x,y)| \le \|x\|\|y\|$ with equality iff $x=0$ or $y=0$. *Proof*: Set $w_1 = x/\|x\|$ and $w_2 = y/\|y\|$. $0 \le (w_1 - w_2, w_1 - w_2) = 2(1-(w_1,w_2)) \Rightarrow (w_1,w_2) \le 1$.

  Example: Applying $x_i = \sqrt{a_i}$ and $y_i = 1/\sqrt{a_i}$ yields

$$\left(\sum_{i=1}^n 1\right)^2 \le \sum_{i=1}^n a_i \sum_{i=1}^n 1/a_i \quad \Rightarrow \quad \frac{n}{\sum_{i=1}^n 1/a_i} \le \frac{\sum_{i=1}^n a_i}{n},$$

  meaning that Harmonic mean $\le$ Arithmetic mean ($\le$ Geometric mean).

5 Some useful results for Matrices

- Let c_i be a vector with 1 for the i th element and 0 elsewhere.
- If $Ax = 0$ for $\forall x$, then $x = 0$: Setting $x = c_i$ leads to $Ax = a_i = 0$, where a_i is the ith column of A.
- If A is symmetric and $x'Ax = 0$ for $\forall x$, then $A = 0$: Setting $x = c_i$ leads to $x'Ax = a_{ii} = 0$. Further, setting $x = c_i + c_j$ $(i \ne j)$ leads to $x'Ax = a_{ii} + 2a_{ij} + a_{jj} = 0 \Rightarrow a_{ij} = 0$.
- If A is *not* symmetric, however, this is FALSE. Although $a_{ii} = 0$ still satisfies, we have $a_{ij} + a_{ji} = 0$ instead of $2a_{ij} = 0$. The counterexample is like this:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0 \quad \forall \mathbf{x}.$$

- If A is symmetric and nonsingular (usually variance-covariance matrix), then

$$\beta' A \beta - 2b'\beta = (\beta - A^{-1}b)'A(\beta - A^{-1}b) - b'A^{-1}b.$$

6 Trace and Eigenvalues

- Given a *square* matrix A, consider $Ax = \lambda x \Leftrightarrow (A - \lambda I)x = 0$, where $x \ne 0$. Then $A - \lambda I$ is always singular because otherwise (if nonsingular) $x = 0$, which contradicts the assumption. Thus, solving $|A - \lambda I| = 0$ obtains λ (eigenvalue) and the corresponding x (eigenvector).
- If A is an $n \times n$ symmetric with eigenvalues λ_i $(i = 1, \dots, n,$
  - $\mathrm{tr}(A) = \sum_{i=1}^n \lambda_i$ and $\det(A) = |A| = \prod_{i=1}^n \lambda_i$ by expanding $|\lambda I_n - A|$.
  - $\mathrm{tr}(A^k) = \mathrm{tr}[(T\Lambda T')^k] = \mathrm{tr}(T\Lambda^k T') = \mathrm{tr}(\Lambda^k) = \sum_{i=1}^n \lambda_i^k$ by the SD and the trace property.

7 Fundamental subspaces

- The space spanned by the *columns* of A, called the column space of A, is denoted by $\mathcal{C}(A)$.
- Let $A \in \mathbb{R}^{n \times p}$.

$$\begin{aligned} \text{Column space of } A &= \mathcal{C}(A) = \{Ax \mid x \in \mathbb{R}^p\}, \\ \text{Row space of } A &= \mathcal{R}(A) = \{A'x \mid x \in \mathbb{R}^n\} = \mathcal{C}(A'), \\ \text{Null space of } A &= \mathcal{N}(A) = \{x \in \mathbb{R}^p \mid Ax = 0\}, \\ \text{Left null space of } A &= \mathcal{N}(A') = \{x \in \mathbb{R}^n \mid A'x = 0\}. \end{aligned}$$

- $\mathcal{N}(A) = \mathcal{C}(A')^\perp$. *Proof*: If $x \in N(A)$, then $Ax = 0 \Rightarrow b'Ax = 0, \forall b \Rightarrow x'(A'b) = 0 \Rightarrow x \in C(A')^\perp$. Conversely, if $x \in \mathcal{C}(A')^\perp$, then $x'y = x'(A'b) = 0,\ \forall b \Rightarrow b'Ax = 0,\ \forall b \Rightarrow Ax = 0 \Rightarrow x \in \mathcal{N}(A)$.

- $(\Omega_1 \cap \Omega_2)^\perp = \Omega_1^\perp + \Omega_2^\perp$. Let $\Omega_i = \mathcal{N}(A_i)$ $(i = 1, 2)$. Then

$$\Omega_1 \cap \Omega_2 = \mathcal{N}(A_1) \cap \mathcal{N}(A_2) = \mathcal{N}\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$$

$$\Rightarrow (\Omega_1 \cap \Omega_2)^\perp = \mathcal{N}\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}^\perp = \mathcal{C}(A_1' \mid A_2') = \mathcal{C}(A_1') + \mathcal{C}(A_2') = \Omega_1^\perp + \Omega_2^\perp.$$

- (HW1) If $\mathcal{C}(A) \subseteq \mathcal{C}(B)$, show there exists C s.t. $A = BC$. What is $\text{rank}(C)$ if A has full column rank?

  Solution Let a_i be the ith column of A $(i = 1, \ldots, m)$, then since $a_i \in \mathcal{C}(A) \subseteq \mathcal{C}(B)$, there exists c_i such that $a_i = Bc_i$, so that $A = BC$. Then $m = \text{rank}(A) = \text{rank}(BC) \le \text{rank}(C) \le m$ follows $\text{rank}(C) = \text{rank}(A)$.

8 Rank

- $\text{rank}(A)$ is equivalent to the maximum number of linearly independent rows or columns.
- $\text{rank}(AB) \le \min(\text{rank}(A), \text{rank}(B))$ since the rows of AB are linear combinations of the rows of B and the columns of AB are linear combinations of the columns of A.
- If X is $n \times p$ of rank p and B is $p \times q$ of rank q, then $\text{rank}(XB) = q$.

  Proof 1: $XBa = X(Ba) = 0 \Rightarrow Ba = 0 \Rightarrow a = 0$. So, XB also has linearly independent columns.

  Proof 2: $q = \text{rank}(B) = \text{rank}[(X'X)^{-1}X'XB] \le \text{rank}(XB) \le \text{rank}(B) = q$.
- If A is any matrix and P and Q are any comfortable *nonsingular* matrices, then $\text{rank}(PAQ) = \text{rank}(A)$.

  Proof: $\text{rank}(A) \le \text{rank}(PAQ) \le \text{rank}(P^{-1}PAQQ^{-1}) = \text{rank}(A)$.
- (HW1) Suppose the columns of a comfortable matrix C are added to columns of A to form the augmented matrix $(A \mid C)$. Then $\text{rank}(A \mid C) \ge \text{rank}(A)$.

  Solution: Use the monotonicity of dimension. Let $\mathbf{A} = (\mathbf{a}_1 \cdots \mathbf{a}_p)$ and $\mathbf{C} = (\mathbf{c}_1 \cdots \mathbf{c}_q)$. Then

$$C(\mathbf{A}) = \text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_p\} \subseteq \text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_p, \mathbf{c}_1, \ldots, \mathbf{c}_q\} = C(\mathbf{A}|\mathbf{C}).$$

  Hence, $\dim(C(\mathbf{A})) \le \dim(C(\mathbf{A}|\mathbf{C}))$, or equivalently, $\text{rank}(\mathbf{A}) \le \text{rank}(\mathbf{A}|\mathbf{C})$.
- By the above and the SD, $\text{rank}(A) = \text{rank}(T'AT) = \text{rank}(\Lambda)$, i.e., $\text{rank}(A) =$ No. of nonzero eigenvalues.
- Any $n \times n$ symmetric matrix $\mathbf{A}$ has a set of n orthogonal eigenvectors and $C(\mathbf{A})$ is the space spanned by those eigenvectors corresponding to nonzero eigenvalues.

  Proof: Suppose $\lambda_{r+1} = \cdots = \lambda_n = 0$. Since $\mathbf{A} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}' = \sum_{i=1}^r \lambda_i \mathbf{t}_i \mathbf{t}_i'$,

$$\mathbf{Ax} = \sum_{i=1}^r \lambda_i \mathbf{t}_i \mathbf{t}_i' \mathbf{x} = \sum_{i=1}^r \lambda_i (\mathbf{t}_i' \mathbf{x}) \mathbf{t}_i, \quad \exists \mathbf{x},$$

  which means that $C(\mathbf{A})$ is spanned by $\mathbf{t}_1, \ldots, \mathbf{t}_r$.
- Let $\mathbf{X}$ be a $n \times p$ matrix of rank $r < p$ and $\mathbf{X}$ partitions $(\mathbf{X}_1 \mid \mathbf{X}_2)$, where $X_1 \in \mathbb{R}^{n \times r}$ and $X_2 \in \mathbb{R}^{n \times (p-r)}$, then show that $\mathbf{X} = \mathbf{X}_1\mathbf{L}$ where $\mathbf{L}$ is $r \times p$ of rank r. *Proof*: there exists $\mathbf{H}$ s.t. $\mathbf{X}_2 = \mathbf{X}_1\mathbf{H}$, so that

$$\mathbf{X} = (\mathbf{X}_1 \mid \mathbf{X}_1\mathbf{H}) = \mathbf{X}_1(\mathbf{I}_r \mid \mathbf{H}) := \mathbf{X}_1\mathbf{L} \;\Rightarrow\; r = \text{rank}(\mathbf{I}_r) \le \text{rank}(\mathbf{I}_r \mid \mathbf{H}) = \text{rank}(\mathbf{L}) \le \min(r, p) = r.$$

- If $\mathbf{A} \in \mathbb{R}^{n \times p}$ is of full column rank, $\mathbf{Ax} = 0 \Rightarrow \mathbf{x} = 0$ $(\mathbf{Ax} = 0 \Leftrightarrow \mathbf{x} = 0)$ since

$$(\mathbf{a}_1, \ldots, \mathbf{a}_p)\begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = \sum_{j=1}^p x_j \mathbf{a}_j = 0.$$

- Then $\mathbf{A}'\mathbf{A}$ is non-singular (invertible).

  Proof 1: Consider $\mathbf{A}'\mathbf{A}\mathbf{x} = 0$. If $\mathbf{A}'\mathbf{A}$ is not invertible, there must be a nonzero $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} = 0$, which contradicts the fact that $\mathbf{A}$ has full column rank.

  Proof 2: $\mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|^2 \geq 0$. Since $\mathbf{A}\mathbf{x} = \mathbf{0}$ holds iff $\mathbf{x} = 0$, $\mathbf{A}'\mathbf{A} \succ \mathbf{O} \Rightarrow \mathbf{A}'\mathbf{A}$ is nonsingular/invertible.

- Likewise, if $\mathbf{A} \in \mathbb{R}^{n\times p}$ is of full row rank, $\mathbf{A}'\mathbf{x} = 0 \Rightarrow \mathbf{x} = 0$.

- (HW1) Show the product of two full row rank matrices always full row rank.

  Solution: Let $\mathbf{A}$ and $\mathbf{B}$ be of full row rank. Then $(\mathbf{B}\mathbf{C})'\mathbf{x} = \mathbf{C}'(\mathbf{B}'\mathbf{x}) = \mathbf{0} \Rightarrow \mathbf{B}'\mathbf{x} = \mathbf{0} \Rightarrow \mathbf{x} = \mathbf{0}$.

- **Rank-nullity theorem**: If a matrix $A \in \mathbb{R}^{n\times p}$ with $\text{rank}(A) = r$

$$\dim\mathcal{C}(A) + \dim\mathcal{N}(A) = p \quad \text{or} \quad \text{rank}(A) + \text{nullity}(A) = p.$$

  Proof: Let $s = \dim\mathcal{N}(A)$ and $\alpha_1, \ldots, \alpha_s$ be a basis for $\mathcal{N}(A) \in \mathbb{R}^p$. Add $(p - s)$ linearly independent vectors $\beta_1, \ldots, \beta_{p-s}$ so that $\{\alpha_1, \ldots, \alpha_s, \beta_1, \ldots, \beta_{p-s}\}$ is a basis for $\mathbb{R}^p$. Then x can be written as:

$$x = \sum_{i=1}^{s} c_i\alpha_i + \sum_{j=1}^{p-s} d_j\beta_j \quad \Rightarrow \quad Ax = \sum_{j=1}^{p-s} d_j(A\beta_j) \quad \because A\alpha_i = 0,$$

  which means that any vector in $C(A)$ is spanned by $A\beta_1, \ldots, A\beta_{p-s}$. Next, we want to show that there are linearly independent vectors: Suppose

$$\sum_{j=1}^{p-s} \gamma_j(A\beta_j) = A\sum_{j=1}^{p-s} \gamma_j\beta_j = 0 \quad \Rightarrow \quad \sum_{j=1}^{p-s} \gamma_j\beta_j \in N(A),$$

  leading to

$$\sum_{j=1}^{p-s} \gamma_j\beta_j = \sum_{i=1}^{s} \delta_i\alpha_i \quad \Rightarrow \quad \sum_{j=1}^{p-s} \gamma_j\beta_j - \sum_{i=1}^{s} \delta_i\alpha_i = 0.$$

  Since $\{\alpha_1, \ldots, \alpha_s, \beta_1, \ldots, \beta_{p-s}\}$ is a basis for $\mathbb{R}^p$, $\gamma_j (= \delta_i) = 0$, $\forall i, j$. That is, $A\beta_1, \ldots, A\beta_{p-s}$ are linearly independent, or equivalently, $\{A\beta_1, \ldots, A\beta_{p-s}\}$ is a basis for $C(A)$ so that $p - s = \dim C(A) = \text{rank}(A) = r$.

- $\text{rank}(X'X) = \text{rank}(X) = \text{rank}(XX') = \text{rank}(X')$.

  Proof: Show $N(X'X) = N(X)$. $a \in N(X) \Rightarrow Xa = 0 \Rightarrow X'Xa = 0 \Rightarrow a \in N(X'X) \Rightarrow N(X) \subseteq N(X'X)$. Conversely, $a \in N(X'X) \Rightarrow X'Xa = 0 \Rightarrow \|Xa\|^2 = 0 \Rightarrow Xa = 0 \Rightarrow a \in N(X) \Rightarrow N(X'X) \subseteq N(X)$. Next

$$N(X'X) = N(X) \quad \Rightarrow \quad \dim N(X'X) = \dim N(X)$$

  Since both $X'X$ and X have the same p columns, by the rank-nullity theorem,

$$p - \text{rank}(X'X) = p - \text{rank}(X) \quad \Rightarrow \quad \text{rank}(X'X) = \text{rank}(X).$$

  In a similar way, $\text{rank}(XX') = \text{rank}(X')$. Since $\text{rank}(X'X) = \text{rank}(XX')$, we show the lemma.

- Also, $\text{rank}(X^+X) = \text{rank}(X) = \text{rank}(XX^+) = \text{rank}(X^+)$ holds, where X^+ is the Moore Penrose inverse (Midterm).

- $C(X'X) = C(X')$. *Proof*: $a \in C(X'X) \Rightarrow a = X'Xb = X'c, \exists c = Xb \Rightarrow a \in C(X')$, so $C(X'X) \subseteq C(X')$. However, $\dim(C(X'X)) = \dim(C(X'))$ by the above lemma, leading to $C(X'X) = C(X')$. This implies that we can always find one or more solutions to $X'X\beta = X'y$.

9 Symmetric and idempotent

- If A is symmetric, i.e., $A' = A$, then A^n is also symmetric.
- Symmetric matrices have only real eigenvalues:

  Proof 1: $\lambda\|x\|^2 = (\lambda x, x) = (Ax, x) = (x, A'x) = (x, Ax) = \lambda^*\|x\|^2 \Rightarrow \lambda^* = \lambda$.

  Proof 2: Let $\mathbf{Ax} = \lambda\mathbf{x} = (\alpha + i\beta)\mathbf{x}$ ($\mathbf{x} \neq 0$). Define $\mathbf{B} = (\mathbf{A} - (\alpha - i\beta)\mathbf{I})'(\mathbf{A} - (\alpha + i\beta)\mathbf{I})$. Then

  $$\mathbf{B} = \mathbf{A}^2 - 2\alpha\mathbf{A} + \alpha^2\mathbf{I} + \beta^2\mathbf{I} = (\mathbf{A} - \alpha\mathbf{I})^2 + \beta^2\mathbf{I} \quad \because \mathbf{A}' = \mathbf{A}.$$

  Since $\mathbf{Bx} = (\mathbf{A} - (\alpha - i\beta)\mathbf{I})'(\mathbf{Ax} - (\alpha + i\beta)\mathbf{x}) = \mathbf{0}$ by the assumption,

  $$0 = \mathbf{x}'\mathbf{Bx} = \mathbf{x}'(\mathbf{A} - \alpha\mathbf{I})^2\mathbf{x} + \beta^2\mathbf{x}'\mathbf{x} = \|(\mathbf{A} - \alpha\mathbf{I})\mathbf{x}\|^2 + \beta^2\|\mathbf{x}\|^2 \quad \because (\mathbf{A} - \alpha\mathbf{I})' = (\mathbf{A} - \alpha\mathbf{I}).$$

  The last two terms are both nonnegative, so $\beta = 0$.
- The eigenvectors corresponding to distinct eigenvalues are orthogonal to each other.

  Proof: Let $\mathbf{Ax} = \lambda_1\mathbf{x}$ and $\mathbf{Ay} = \lambda_2\mathbf{y}$ ($\lambda_1 \neq \lambda_2$). Then

  $$\lambda_1(\mathbf{x}, \mathbf{y}) = (\lambda_1\mathbf{x}, \mathbf{y}) = (\mathbf{Ax}, \mathbf{y}) = (\mathbf{x}, \mathbf{A}'\mathbf{y}) = (\mathbf{x}, \mathbf{Ay}) = (\mathbf{x}, \lambda_2\mathbf{y}) = \lambda_2(\mathbf{x}, \mathbf{y}) \;\Rightarrow\; (\mathbf{x}, \mathbf{y}) = 0.$$

- If $A^2 = A$, A is said to be idempotent. A symmetric and idempotent matrix is called a *projection matrix*, whose eigenvalues are 0 or 1 as $\lambda^2 x = \lambda(Ax) = A^2x = Ax = \lambda x \Rightarrow \lambda$.
  - X^+X is a projection matrix, where X^+ is the Moore-Penrose inverse (Midterm): $(X^+X)' = X^+X$ and $(X^+X)(X^+X) = X^+X$. So does XX^+.
- If A is *symmetric* and orthogonal, i.e., $A'A = AA' = I$, then row and columns of A are orthogonal each other. Also, $\det(A'A) = |A|^2 = 1 \;\Rightarrow\; |A| \pm 1$, which does *not* mean eigenvalues are ± 1 (e.g., $\pm 0.5, \pm 2$).
- If $A \in \mathbb{R}^{n \times n}$ with rank $r < n$ is (symmetric) and idempotent, by above and the spectral decomposition,

  $$T'AT = \Lambda = \begin{pmatrix} I_r & O \\ O & O \end{pmatrix} \quad \Rightarrow \quad A = T\Lambda T' = (\underbrace{T_1}_{n\times r} \mid T_2) \begin{pmatrix} I_r & O \\ O & O \end{pmatrix} \begin{pmatrix} T_1' \\ T_2' \end{pmatrix} = T_1T_1',$$

  where $AT_1 = T_1$ ($\lambda_1 = \cdots = \lambda_r = 1$) and $AT_2 = O$ ($\lambda_{r+1} = \cdots = \lambda_n = 0$). Note that
  - $t_1 \ldots, t_r \in C(A) = R(A)$, while $t_{r+1}, \ldots, t_n \in N(A)$, $i = r+1, \ldots, n$.
  - Note: Unlike T, T_1 and T_2 are *not* orthogonal as they are not square. Since T_1 has orthogonal columns, however, $T_1'T_1 = I_r$.
- Positive definite and semi-positive definite are defined only to symmetric matrices.
  - If $\mathbf{A}$ is p.d. $\Rightarrow |\mathbf{A}| > 0 \Rightarrow \mathbf{A}$ is non-singular.
  - If $\mathbf{A}$ is idempotent, then $\text{rank}(\mathbf{A}) = \text{tr}(\mathbf{A})$ = the number of eigenvalues 1.

10 Projections on Subspaces

- Let P_Ω and P_ω be the projection matrix onto $\Omega = C(X)$ and $\omega \subseteq \Omega$.
- $P_\Omega(I - P_\Omega) = O \Rightarrow I - P_\Omega = P_{\Omega^\perp}$, that is, $I - P_\Omega$ projects onto $\Omega^\perp$.
- Since $P_\omega P_\Omega = P_\Omega P_\omega = P_\omega$, we have $P_\omega(P_\Omega - P_\omega) = O$, meaning that $P_\Omega - P_\omega$ projects onto $\omega^\perp \cap \Omega$.
- $(I - P_\Omega)(P_\Omega - P_\omega) = (P_\Omega - P_\omega) - P_\Omega(P_\Omega - P_\omega) = O \;\Rightarrow\; I - P_\Omega \perp\!\!\!\perp P_\Omega - P_\omega$.

- If A_1 is any matrix such that $\omega = \mathcal{N}(A_1) \cap \Omega$, then $\omega^\perp \cap \Omega = \mathcal{C}(P_\Omega A_1')$.

  Proof: Since $\omega^\perp = (\mathcal{N}(A_1) \cap \Omega)^\perp = \mathcal{C}(A_1') + \Omega^\perp$, if $x \in \omega^\perp \cap \Omega$, then $x = P_\Omega x = P_\Omega[A_1'\alpha + (I - P_\Omega)\beta] = P_\Omega A_1'\alpha \in \mathcal{C}(P_\Omega A_1') \Rightarrow \omega^\perp \cap \Omega \subseteq C(P_\Omega A_1')$. Conversely, if $x \in C(P_\Omega A_1')$, then $x \in C(P_\Omega) = \Omega$. Also, if $z \in \omega = \mathcal{N}(A_1) \cap \Omega$, then $x'z = \alpha' A_1 P_\Omega z = \alpha' A_1 z = 0$, so that $x \in \omega^\perp \cap \Omega \Rightarrow C(P_\Omega A_1') \subseteq \omega^\perp \cap \Omega$.

- If A_1 is a $q \times n$ matrix of rank q, then $\text{rank}(P_\Omega A_1') = q$ if and only if $C(A_1') \cap \Omega^\perp = 0$.

  Proof: We have $\text{rank}(\mathbf{P}_\Omega \mathbf{A}_1') \leq \text{rank}(\mathbf{A}_1) = q$. Let $\mathbf{A}_1' = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_q) \in \mathbb{R}^{n \times q}$ and suppose $\text{rank}(\mathbf{P}_\mathbf{\Omega} \mathbf{A}_1') < q$. Then there exists nonzero $\sum_i c_i a_i \in \mathcal{C}(\mathbf{A}_1')$ such that $\mathbf{P}_\Omega \mathbf{A}_1' \mathbf{c} = \sum_i c_i \mathbf{P}_\Omega \mathbf{a}_i = 0$ that is perpendicular to Ω. Hence, $C(\mathbf{A}_1') \cap \Omega^\perp \neq 0$, which is a contradiction.

11 Positive (semi-) definite

- $\mathbf{A}$ is positive definite iff $\mathbf{x}'\mathbf{A}\mathbf{x} > 0,\ \forall \mathbf{x} \neq \mathbf{0}$ or iff all leading minors have positive determinant. If $\mathbf{A}$ is positive definite, $\mathbf{A}$ is clearly non-singular.
- $\mathbf{A}$ is positive semi-definite if $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0,\ \forall \mathbf{x} \neq 0$.
- The diagonal elements of a p.d. matrix are all positive: Setting $x = e_i$ leads to $x'Ax = a_{ii} > 0$, $\forall i$.
- If $\mathbf{A}$ is p.d., there exists the non-singular and symmetric matrix $\mathbf{A}^{\frac{1}{2}}$ such that $\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}$

  Proof: Since $\mathbf{A}$ is symmetric and has only positive eigenvalues, by spectral decomposition,

$$\mathbf{A} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}' = \mathbf{T}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{T}' = (\mathbf{T}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{T}')(\mathbf{T}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{T}') = \mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} \quad \text{since } \mathbf{T}'\mathbf{T} = \mathbf{I}.$$

- If $\mathbf{A}$ is p.s.d., we also have $\mathbf{A}$ s.t. $\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}$, but $\mathbf{A}$ is singular ($|\mathbf{A}| = 0$).
- If $\mathbf{A}$ is p.s.d., then $\mathbf{X}'\mathbf{A}\mathbf{X} = \mathbf{O} \Rightarrow \mathbf{A}\mathbf{X} = \mathbf{O}$. Note $\mathbf{A}$ is *singular* so that $\mathbf{A}^{-1}$ and $\mathbf{A}^{-1/2}$ does not exist.

  Proof: For $\forall \mathbf{a}$, $\mathbf{a}'\mathbf{X}'\mathbf{A}\mathbf{X}\mathbf{a} = \|\mathbf{A}^{1/2}\mathbf{X}\mathbf{a}\|^2 = 0 \Rightarrow \mathbf{A}^{1/2}\mathbf{X}\mathbf{a} = \mathbf{0} \Rightarrow \mathbf{A}\mathbf{X}\mathbf{a} = \mathbf{0}$ (not $\mathbf{X}\mathbf{a} = \mathbf{0}$), so $\mathbf{A}\mathbf{X} = \mathbf{O}$.
- **Simultaneous diagonalization**: If $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{B} \succeq \mathbf{O}$, then there exists $\mathbf{U}$ ($|\mathbf{U}| \neq 0$) s.t.

$$\mathbf{U}'\mathbf{A}\mathbf{U} = \mathbf{I}, \quad \mathbf{U}'\mathbf{B}\mathbf{U} = \mathbf{D} = \text{diag}(d_1, \ldots, d_n).$$

  Proof: By definition of positive definite, we can assume $\mathbf{A}$ and $\mathbf{B}$ are symmetric. Also, $\mathbf{A} \succ \mathbf{O}$ implies that $\mathbf{A}^{1/2}$ exists, so that $\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}$ is symmetric. By the spectral decomposition,

$$\mathbf{T}'\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}\mathbf{T} = (\mathbf{A}^{-\frac{1}{2}}\mathbf{T})'\mathbf{B}(\mathbf{A}^{-\frac{1}{2}}\mathbf{T}) = \mathbf{U}'\mathbf{B}\mathbf{U} = \mathbf{D} \succeq \mathbf{O},$$

  where $\mathbf{U} = \mathbf{A}^{-\frac{1}{2}}\mathbf{T}$. Then $\mathbf{U}'\mathbf{A}\mathbf{U} = \mathbf{T}'\mathbf{A}^{-\frac{1}{2}}\mathbf{A}\mathbf{A}^{-\frac{1}{2}}\mathbf{T} = \mathbf{T}'\mathbf{T} = \mathbf{I}$ as $\mathbf{T}$ is orthogonal.

  - If $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{B} \succ \mathbf{O}$ and $\mathbf{A} \succ \mathbf{B}$, then 1) $|\mathbf{A}| > |\mathbf{B}|$ and 2) $\mathbf{B}^{-1} - \mathbf{A}^{-1} \succ \mathbf{O}$.

    Proof of (1): Since $\mathbf{U}$ is nonsingular, $\mathbf{I} - \mathbf{D} = \mathbf{U}'(\mathbf{A} - \mathbf{B})\mathbf{U} \succ \mathbf{O}$. Hence, $d_i < 1$ for $\forall i$. Hence,

$$0 < |\mathbf{I}| - |\mathbf{D}| = |\mathbf{U}'|(|\mathbf{A}| - |\mathbf{B}|)|\mathbf{U}| = (|\mathbf{A}| - |\mathbf{B}|)|\mathbf{U}'\mathbf{U}| = (|\mathbf{A}| - |\mathbf{B}|)|\mathbf{A}|^{-1} \ \Rightarrow\ |\mathbf{A}| - |\mathbf{B}| > 0.$$

    Proof of (2): We have $\mathbf{A}^{-1}\mathbf{U}\mathbf{U}'$ and $\mathbf{B}^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}'$, so that

$$\mathbf{B}^{-1} - \mathbf{A}^{-1} = \mathbf{U}(\mathbf{D}^{-1} - \mathbf{I})\mathbf{U}' \succ \mathbf{O} \quad \because \mathbf{D}^{-1} - \mathbf{I} \succ \mathbf{O}.$$

- If $\mathbf{A}$ is an $n \times n$ p.d. and $\mathbf{B}$ is an $n \times n$ symmetric matrix, then $\mathbf{A} - t\mathbf{B}$ is p.d. for $|t|$ sufficiently small. *Brief proof*: The ith leading minor determinant of $\mathbf{A} - t\mathbf{B}$ is a function of t, which is positive when $t = 0$. Since the function is continuous, it will be positive for $|t| < \delta_i$ for δ_i sufficiently small. Let $\delta = \min(\delta_1, \ldots, \delta_n)$, then all the leading minor determinants will be positive for $|t| < \delta$.

- If L is positive definite then for any b,

$$\max_{h"h\neq 0}\left[\frac{(h'b)^2}{h'Lh}\right]=b'L^{-1}b.$$

Proof: Use Cauchy–Schwarz inequality: $(u'v)^2\leq\|u\|^2\|v\|^2$. Suppose $u\neq 0$, then we have

$$\frac{(u'v)^2}{\|u\|^2}\leq\|v\|^2$$

Further let $u=L^{1/2}h$ $(h\neq 0)$ and $v=L^{-1/2}b$ as $L\succ O$, then

$$\frac{(h'b)^2}{h'Lh}\leq b'L^{-1}b$$

with the equality holds when $L^{1/2}h=cL^{-1/2}b\Rightarrow cb=Lh$, where c is a scalar.

12 Eigenvalue Application

- Let A be an $n\times n$ symmetric matrix, then

$$\max_{x:x\neq 0}\left(\frac{x'Ax}{x'x}\right)=\lambda_{\text{MAX}},\quad \min_{x:x\neq 0}\left(\frac{x'Ax}{x'x}\right)=\lambda_{\text{MIN}}$$

and these values occur when x is the eigenvector corresponding to the λ_{MAX} and λ_{MIN}, respectively.

Proof: Suppose $\lambda_1\geq\cdots\geq\lambda_n$. By the spectral decomposition, $T'AT=\Lambda$. Setting $x=Ty$ leads to

$$\frac{x'Ax}{x'x}=\frac{y'T'ATy}{y'T'Ty}=\frac{y'\Lambda y}{y'y}=\frac{\sum_{i=1}^n\lambda_i y_i^2}{\sum_{i=1}^n y_i^2}\leq\lambda_1$$

with equality when $y=e_1\Rightarrow x=Te_1=t_1$. Also,

$$\frac{x'Ax}{x'x}=\frac{y'T'ATy}{y'T'Ty}=\frac{y'\Lambda y}{y'y}=\frac{\sum_{i=1}^n\lambda_i y_i^2}{\sum_{i=1}^n y_i^2}\geq\lambda_n$$

with equality when $y=e_n\Rightarrow x=Te_n=t_n$.

- (HW1) Show the minimum and maximum eigenvalues of

$$B=\frac{2b}{2b-1}I_n-\frac{1_n1_n'}{2b-1},\quad b>\frac{1}{2}.$$

Solution: For $x\neq 0$,

$$\frac{x'Bx}{x'x}=\frac{2b}{2b-1}-\frac{1}{2b-1}\frac{(1_n'x)^2}{x'x}.$$

By Cauchy–Schwarz inequality,

$$\frac{x'Bx}{x'x}\geq\frac{2b}{2b-1}-\frac{1}{2b-1}\frac{\|1_n\|^2\|x\|^2}{x'x}=\frac{2b}{2b-1}-\frac{n}{2b-1}=\frac{2b-n}{2b-1}=\lambda_{\text{MIN}}$$

with equality iff $x=c1_n$. Also,

$$\frac{x'Bx}{x'x}\leq\frac{2b}{2b-1}=\lambda_{\text{MAX}}$$

with equality iff $1_n'x=0$, i.e., $1_n\perp x$.

13 Partitioned Matrix

- Basic determinant properties

$$\left|\begin{pmatrix} I & B \\ O & I \end{pmatrix}\right| = \left|\begin{pmatrix} I & O \\ C & I \end{pmatrix}\right| = |I| = 1, \quad \left|\begin{pmatrix} A_{11} & O \\ O & I \end{pmatrix}\right| = |A_{11}|, \quad \left|\begin{pmatrix} I & O \\ O & A_{22} \end{pmatrix}\right| = |A_{22}|$$

  follow

$$\begin{aligned} \left|\begin{pmatrix} I & O \\ A_{21} & A_{22} \end{pmatrix}\right| &= \left|\begin{pmatrix} I & O \\ O & A_{22} \end{pmatrix}\right| \left|\begin{pmatrix} I & O \\ A_{22}^{-1}A_{21} & I \end{pmatrix}\right| = |A_{22}|, \\ \left|\begin{pmatrix} A_{11} & O \\ A_{21} & A_{22} \end{pmatrix}\right| &= \left|\begin{pmatrix} A_{11} & O \\ O & I \end{pmatrix}\right| \left|\begin{pmatrix} I & O \\ A_{21} & A_{22} \end{pmatrix}\right| = |A_{11}||A_{22}|. \end{aligned}$$

- Let $A_{11.2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$ (Schur complement), then $|A| = |A_{22}||A_{11.2}| = |A_{11}||A_{22.1}|$ since

$$|A| = \left|1 \cdot \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \cdot 1\right| = \left|\begin{pmatrix} I & -A_{12}A_{22}^{-1} \\ O & I \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} I & O \\ -A_{22}^{-1}A_{21} & I \end{pmatrix}\right| = \left|\begin{pmatrix} A_{11.2} & O \\ O & A_{22} \end{pmatrix}\right|.$$

- (HW2) Let $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times n}$ then $|I_n + AB| = |I_m + BA|$. *Proof*:

$$|\mathbf{I}_m + \mathbf{AB}| = \begin{vmatrix} \mathbf{I}_m + \mathbf{AB} & \mathbf{O} \\ \mathbf{B} & \mathbf{I}_n \end{vmatrix} = \begin{vmatrix} \mathbf{I}_m & \mathbf{A} \\ \mathbf{O} & \mathbf{I}_n \end{vmatrix} \begin{vmatrix} \mathbf{I}_m & -\mathbf{A} \\ \mathbf{B} & \mathbf{I}_n \end{vmatrix} = \begin{vmatrix} \mathbf{I}_m & -\mathbf{A} \\ \mathbf{B} & \mathbf{I}_n \end{vmatrix} \begin{vmatrix} \mathbf{I}_m & \mathbf{A} \\ \mathbf{O} & \mathbf{I}_n \end{vmatrix} = \begin{vmatrix} \mathbf{I}_m & \mathbf{O} \\ \mathbf{B} & \mathbf{I}_n + \mathbf{BA} \end{vmatrix} = |\mathbf{I}_n + \mathbf{BA}|.$$

- (HW3) If a partition matrix

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \succeq \mathbf{O},$$

  then $N(\mathbf{A}_{22}) \subset N(\mathbf{A}_{12})$ and $C(\mathbf{A}_{21}) \subset C(\mathbf{A}_{22})$.

  Proof: Let $\mathbf{x}' = (\mathbf{x}_1' \ \alpha\mathbf{x}_2')$, where $\mathbf{x}_2 \in N(\mathbf{A}_{22})$ and $\alpha \in \mathbb{R}$.

$$\begin{aligned} 0 \le \mathbf{x}'\mathbf{A}\mathbf{x} &= \mathbf{x}_1'\mathbf{A}_{11}\mathbf{x}_1 + \alpha\mathbf{x}_2'\mathbf{A}_{21}\mathbf{x}_1 + \alpha\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2 + \alpha^2\mathbf{x}_2'\mathbf{A}_{22}\mathbf{x}_2 \\ &= \mathbf{x}_1'\mathbf{A}_{11}\mathbf{x}_1 + 2\alpha\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2 \quad \text{since } \mathbf{A}_{21}' = \mathbf{A}_{12}, \ \mathbf{A}_{22}\mathbf{x}_2 = \mathbf{0}. \end{aligned}$$

  To satisfy that RHS ≥ 0 for $\forall\alpha$, $\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2$ has to be zero for $\forall\mathbf{x}_1$. Then, $\mathbf{A}_{12}\mathbf{x}_2 = 0$. Hence, $N(\mathbf{A}_{22}) \subset N(\mathbf{A}_{12})$. It follows from this relationship that

$$\begin{aligned} N(\mathbf{A}_{22}) \subset N(\mathbf{A}_{12}) &\Leftrightarrow C(\mathbf{A}_{22}')^{\perp} \subset C(\mathbf{A}_{12}')^{\perp} \\ &\Leftrightarrow C(\mathbf{A}_{22})^{\perp} \subset C(\mathbf{A}_{21})^{\perp} \quad \text{since } \mathbf{A}_{22}' = \mathbf{A}_{22}, \mathbf{A}_{12}' = \mathbf{A}_{21} \\ &\Leftrightarrow C(\mathbf{A}_{21}) \subset C(\mathbf{A}_{22}). \end{aligned}$$

14 Inverse Matrix

- **Sherman-Morrison-Woodbury formula**: Let A and B be nonsingular $m\times m$ and $n\times n$ matrices, respectively, and let U be $m\times n$ and V be $n\times m$. Then

$$\begin{aligned} (A + UBV)^{-1} &= A^{-1} - A^{-1}UB(B + BVA^{-1}UB)^{-1}BVA^{-1} \\ &= A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1}. \end{aligned}$$

  Proof: Pre- or post- multiply by $A + UBV$ to get I_m.

- Setting $B = 1$, $U = \pm u \in \mathbb{R}^m$, and $V = v' \in \mathbb{R}^m$, we have

$$(A \pm uv')^{-1} = A^{-1} \mp \frac{A^{-1}uv'A^{-1}}{1 \pm v'A^{-1}u}.$$

15 Generalized inverse

- Let $\mathbf{A} \in \mathbb{R}^{n\times m}$ with rank of $r < \min(n,m)$ (not full rank), then there exists $\mathbf{A}^-$, s.t. (i) $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$.
- Such a matrix always exists and is called *a* generalized inverse or g-inverse (HW1).

  Proof: If $\mathbf{A}$ is non-singular, then $\mathbf{B} = \mathbf{A}^{-1}$ is unique.

  If $\mathbf{A}$ is singular, suppose $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $\text{rank}(\mathbf{A}) = r$. By the rank factorization, we obtain $\mathbf{A} = \mathbf{C}\mathbf{R}$, where $\mathbf{C} \in \mathbb{R}^{m\times r}$ is full column rank and $\mathbf{R} \in \mathbb{R}^{r\times n}$ is full row rank. Since $\mathbf{A}\mathbf{B}\mathbf{A} = (\mathbf{C}\mathbf{R})\mathbf{B}(\mathbf{C}\mathbf{R}) = \mathbf{C}(\mathbf{R}\mathbf{B}\mathbf{C})\mathbf{R}$, we want to find $\mathbf{B}$ s.t. $\mathbf{R}\mathbf{B}\mathbf{C} = \mathbf{I}$ so that $\mathbf{A}\mathbf{B}\mathbf{A} = \mathbf{A}$. As mentioned before, $\mathbf{C}'\mathbf{C}$ and $\mathbf{R}\mathbf{R}'$ are non-singular even though $\mathbf{A}$ is singular. Hence, there always exists

  $$\mathbf{B} = \mathbf{R}'(\mathbf{R}\mathbf{R}')^{-1}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'$$

  such that $\mathbf{R}\mathbf{B}\mathbf{C} = \mathbf{I}$.

- $\mathbf{A}^-$ is not unique. There are several ways of getting it: If $\mathbf{A}^-$ is a g-inverse, then
  - $\mathbf{G} = \mathbf{A}^- + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{W}$ ($\mathbf{W} \neq \mathbf{O}$) is also a g-inverse since $\mathbf{A}\mathbf{G}\mathbf{A} = \mathbf{A}(\mathbf{A}^- + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{W})\mathbf{A} = \mathbf{A}\mathbf{A}^-\mathbf{A} + (\mathbf{A} - \mathbf{A}\mathbf{A}^-\mathbf{A})\mathbf{W}\mathbf{A} = \mathbf{A}$, or
  - $\mathbf{G} = \mathbf{A}^- + \mathbf{u}\mathbf{v}'$ ($\mathbf{u}\mathbf{v}' \neq \mathbf{O}$) is also a g-inverse, where $\mathbf{u} \in N(\mathbf{A})$ s.t. $\mathbf{u} \neq \mathbf{0}$ or $\mathbf{v} \in N(\mathbf{A}')$ s.t. $\mathbf{v} \neq \mathbf{0}$, since $\mathbf{A}\mathbf{G}\mathbf{A} = \mathbf{A}\mathbf{A}^-\mathbf{A}^- + (\mathbf{A}\mathbf{u})\mathbf{v}'\mathbf{A} = \mathbf{A}$.
- Taking transpose of the above property yields $\mathbf{A}'(\mathbf{A}^-)'\mathbf{A}' = \mathbf{A}'$, leading to $(\mathbf{A}')^- = (\mathbf{A}^-)'$.
- A solution(s) to $\mathbf{A}\mathbf{x} = \mathbf{b}$ is $\mathbf{x} = \mathbf{A}^-\mathbf{b}$, which is not unique, as $\mathbf{A}(\mathbf{A}^-\mathbf{b}) = \mathbf{A}\mathbf{A}^-\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x} = \mathbf{b}$.
- If $\mathbf{A}^-$ also satisfies three more conditions: (ii) $\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-$, (iii) $(\mathbf{A}\mathbf{A}^-)' = \mathbf{A}\mathbf{A}^-$, and (iv) $(\mathbf{A}^-\mathbf{A})' = \mathbf{A}^-\mathbf{A}$, then $\mathbf{A}^-$ is denoted by $\mathbf{A}^+$, which is called the **Moore-Penrose inverse**.
- Moore-Penrose inverse $\mathbf{A}^+$ is unique. If $\mathbf{B}^+$ is another Moore-Penrose inverse, then

  $$\begin{aligned}\mathbf{B}^+ &= \mathbf{B}^+\mathbf{A}\mathbf{B}^+ = \mathbf{B}^+\mathbf{A}\mathbf{A}^+\mathbf{A}\mathbf{B}^+ = \mathbf{A}'(\mathbf{B}^+)'\mathbf{A}^+(\mathbf{B}^+)'\mathbf{A}' = \mathbf{A}'(\mathbf{A}^+)'\mathbf{A}'(\mathbf{B}^+)'\mathbf{A}^+(\mathbf{B}^+)'\mathbf{A}'(\mathbf{A}^+)'\mathbf{A}' \\ &= \mathbf{A}^+\mathbf{A}\mathbf{B}^+\mathbf{A}\mathbf{A}^+\mathbf{A}\mathbf{B}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+\mathbf{A}\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+.\end{aligned}$$

- (HW5) Show $\mathcal{C}(\mathbf{A}^+) = \mathcal{C}(\mathbf{A}')$.

  Proof: If $\mathbf{x} \in \mathcal{C}(\mathbf{A}^+)$, then $\mathbf{x} = \mathbf{A}^+\mathbf{u} = \mathbf{A}^+\mathbf{A}\mathbf{A}^+\mathbf{u} = \mathbf{A}'(\mathbf{A}^+)'\mathbf{A}^+\mathbf{u} \in \mathcal{C}(\mathbf{A}')$ for some $\mathbf{u}$. If $\mathbf{x} \in C(\mathbf{A}')$, then $\mathbf{x} = \mathbf{A}'\mathbf{w} = (\mathbf{A}^+\mathbf{A})'\mathbf{A}'\mathbf{w} = \mathbf{A}^+\mathbf{A}\mathbf{A}'\mathbf{w} \in \mathcal{C}(\mathbf{A}^+)$ for some $\mathbf{w}$.

16 Decomposition

- Rank factorization: $\underbrace{\mathbf{A}}_{n\times p} = \underbrace{\mathbf{C}}_{n\times r}\underbrace{\mathbf{R}}_{r\times p}$, where $\mathbf{C}$ has full column rank and $\mathbf{R}$ has full row rank. Then $(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\mathbf{C} = \mathbf{R}\mathbf{R}'(\mathbf{R}\mathbf{R}')^{-1} = \mathbf{I}_r$.
- (HW1) If $PA'A = QA'A$, then $PA' = QA'$ for any comfortable matrices P and Q.

  Solution: If $\mathbf{A}$ is non-singular, or $\mathbf{A}^{-1}$ exists, $\mathbf{P}\mathbf{A}'\mathbf{A} = \mathbf{Q}\mathbf{A}'\mathbf{A} \;\Rightarrow\; \mathbf{P}\mathbf{A}' = \mathbf{Q}\mathbf{A}'$.

  If $\mathbf{A}$ is singular and $\text{rank}(\mathbf{A}) = r$, we have $\mathbf{A} = \mathbf{C}\mathbf{R}$ by the rank factorization, where $\mathbf{C} \in \mathbb{R}^{n\times r}$ is full column rank and $\mathbf{R} \in \mathbb{R}^{r\times n}$ is full row rank. Then

  $$\mathbf{P}\mathbf{A}'\mathbf{A} = \mathbf{Q}\mathbf{A}'\mathbf{A} \;\Rightarrow\; (\mathbf{P} - \mathbf{Q})\mathbf{A}'\mathbf{A} = \mathbf{O} \;\Rightarrow\; (\mathbf{P} - \mathbf{Q})\mathbf{R}'\mathbf{C}'\mathbf{C}\mathbf{R} = \mathbf{O}.$$

  Note that the $r \times r$ matrices $\mathbf{C}'\mathbf{C}$ and $\mathbf{R}\mathbf{R}'$ are non-singular or invertible because we have

  $$\text{rank}(\mathbf{C}'\mathbf{C}) = \text{rank}(\mathbf{C}) = \text{rank}(\mathbf{R}\mathbf{R}') = \text{rank}(\mathbf{R}) = r \text{ (full rank)}$$

  Thus, multiplying by $\mathbf{R}'(\mathbf{R}\mathbf{R}')^{-1}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'$ (g-inverse of $\mathbf{A}$), we obtain

  $$(\mathbf{P} - \mathbf{Q})\mathbf{R}'\mathbf{C}'\mathbf{C}\mathbf{R}[\mathbf{R}'(\mathbf{R}\mathbf{R}')^{-1}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'] = \mathbf{O} \;\Rightarrow\; (\mathbf{P} - \mathbf{Q})\mathbf{R}'\mathbf{C}' = \mathbf{O} \;\Rightarrow\; (\mathbf{P} - \mathbf{Q})\mathbf{A}' = \mathbf{O}.$$

- QR factorization (Gram-Schmidt algorithm): Suppose $\mathbf{A} \in \mathbb{R}^{n\times k}$ and $\mathbf{Q} = (\mathbf{q}_1 \cdots \mathbf{q}_k)$, where

$$\mathbf{q}_i = \frac{\mathbf{a}_i - \sum_{j=1}^{i-1}(\mathbf{a}_i, \mathbf{q}_j)\mathbf{q}_j}{\|\mathbf{a}_i - \sum_{j=1}^{i-1}(\mathbf{a}_i, \mathbf{q}_j)\mathbf{q}_j\|}, \ 1 \le i \le k \quad \text{(orthonormal columns)}.$$

  Then $\mathbf{A} = \mathbf{QR}$, where $\mathbf{R}$ is an upper triangle. QR decomposition is often used to solve the linear least squares problem.

  Application: Consider normal equations: $X'X\beta = X'y$. Solving $\widehat{\beta} = (X'X)^{-1}X'y$ is computationally costly. If we obtain $X = QR$, then the normal equations become

$$R'Q'QR\beta = R'Q'y \quad \Rightarrow \quad R'R\beta = R'Q'y \quad \Rightarrow \quad (R')^{-1}R'R\beta = (R')^{-1}R'Q'y \quad \Rightarrow \quad R\beta = Q'y.$$

  Since R is an upper triangular, it is easier to compute β by solving this from the last element of β.

- Spectral decomposition: If $\mathbf{A}$ is a $n \times n$ symmetric matrix, then $\mathbf{A} = \mathbf{T\Lambda T}' = \sum_i \lambda_i \mathbf{t}_i\mathbf{t}_i'$, or $\mathbf{T}'\mathbf{A}\mathbf{T} = \mathbf{\Lambda}$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$ and $\mathbf{T}$ is an orthogonal matrix (not symmetric in general) with eigenvectors. The columns of $\mathbf{T}$ are eigenvectors, which form an orthogonal basis for $\mathbb{R}^n$.
  - $C(\mathbf{A})$ is spanned by its eigenvector: $Ax = \sum_i \lambda_i \mathbf{t}_i\mathbf{t}_i'\mathbf{x} = \sum_i \lambda_i(\mathbf{t}_i'\mathbf{x})\mathbf{t}_i \in C(\mathbf{A})$.

- Singular value decomposition: Let $\mathbf{A} \in \mathbb{R}^{n\times p}$ with rank of r,

$$\begin{aligned}
\underbrace{\mathbf{A}}_{n\times p} &= \underbrace{(\mathbf{S}_r \mid \mathbf{S}_{p-r})}_{n\times p}\begin{pmatrix}\mathbf{D}_r & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{pmatrix}\underbrace{\begin{pmatrix}\mathbf{T}_r'\\ \mathbf{T}_{p-r}'\end{pmatrix}}_{p\times p} \quad \text{(normal form)}\\
&= \underbrace{\mathbf{S}_r}_{n\times r}\mathbf{D}_r\underbrace{\mathbf{T}_r'}_{r\times p} \quad \text{(reduced form)}\\
&= \sum_{i=1}^{r}\sigma_i\mathbf{s}_i\mathbf{t}_i' \quad \text{(outer product form)},
\end{aligned}$$

  where $\mathbf{D}_r = \text{diag}(\sigma_1, \dots, \sigma_r)$ for $\sigma_1 \ge \cdots \ge \sigma_r > 0$. $\mathbf{S}_r'\mathbf{S}_r = \mathbf{T}_r'\mathbf{T}_r = \mathbf{I}_r$ (Converse is not identity!).
  - Solution 1: Find σ_i^2 (eigenvalues) and $\mathbf{t}_i$ (eigenvectors) by solving $\mathbf{A}'\mathbf{A}\mathbf{t}_i = \sigma_i^2\mathbf{t}_i$. Then

$$\mathbf{s}_i = \frac{\mathbf{A}\mathbf{t}_i}{\sigma_i}, \quad i = 1, \dots, r,$$

    where $\mathbf{s}_i'\mathbf{s}_j = \mathbf{t}_i\mathbf{A}'\mathbf{A}\mathbf{t}_j/(\sigma_i\sigma_j) = (\sigma_j/\sigma_i)\mathbf{t}_i\mathbf{t}_j = \delta_{ij}$, i.e., $\mathbf{S}_r$ is orthogonal as well as $\mathbf{T}_r$.
  - Solution 2: Find σ_i^2 and $\mathbf{s}_i$ by solving $\mathbf{A}\mathbf{A}'\mathbf{s}_i = \sigma_i^{-1}\mathbf{A}\mathbf{A}'\mathbf{A}\mathbf{t}_i = \sigma_i\mathbf{A}\mathbf{t}_i = \sigma_i^2\mathbf{s}_i$. Then

$$\mathbf{t}_i = \frac{\mathbf{A}'\mathbf{s}_i}{\sigma_i}, \quad i = 1, \dots, r.$$

  - The Moore–Penrose inverse: $\mathbf{A}^+ = \mathbf{T}\mathbf{D}_r^{-1}\mathbf{S}'$ that satisfies the following four properties: (i) $\mathbf{A}\mathbf{A}^+\mathbf{A} = (\mathbf{S}\mathbf{D}_r\mathbf{T}')(\mathbf{T}\mathbf{D}_r^{-1}\mathbf{S}')(\mathbf{S}\mathbf{D}_r\mathbf{T}') = \mathbf{S}\mathbf{D}_r\mathbf{T}' = \mathbf{A}$, (ii) $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{T}\mathbf{D}_r^{-1}\mathbf{S}' = \mathbf{A}^+$, (iv) $(\mathbf{A}^+\mathbf{A})' = [(\mathbf{T}\mathbf{D}_r^{-1}\mathbf{S}')(\mathbf{S}\mathbf{D}_r\mathbf{T}')]' = (\mathbf{T}\mathbf{T}')' = \mathbf{T}\mathbf{T}' = \mathbf{A}^+\mathbf{A}$, and (iii) $(\mathbf{A}\mathbf{A}^+)' = \mathbf{S}\mathbf{S}' = \mathbf{A}\mathbf{A}^+$.

- (HW2) Find the SVD of $\mathbf{X}$ whose first row is $(1, 0, 0, 0)$ and the second row is $(-1, 0, 0, 0)$.

  Solution: $r = \text{rank}(\mathbf{X}) = 1$ and

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix}2 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}$$

implies that the eigenvector greater than 0 is $\lambda = 2$. Thus, $\sigma_1 = \sqrt{2}$. The corresponding eigenvector is

$$(2\mathbf{I} - \mathbf{X}'\mathbf{X})\mathbf{t}_1 = \mathbf{0} \;\Rightarrow\; \mathbf{t}_1 = (1, 0, 0, 0)'.$$

Then,

$$\mathbf{s}_1 = \frac{\mathbf{X}\mathbf{t}_1}{\sigma_1} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Hence,

$$\begin{aligned}\mathbf{X} = \mathbf{S}_r\mathbf{D}_r\mathbf{T}_r' = \mathbf{s}_1\sigma_1\mathbf{t}_1' &= \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}(\sqrt{2})\begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix} \quad \text{(reduced form)} \\ = \mathbf{S}\mathbf{D}\mathbf{T}' = (\mathbf{s}_1\ \mathbf{s}_2)\begin{pmatrix} \sigma_1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{t}_1' \\ \mathbf{t}_2' \end{pmatrix} &= \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad \text{(normal form)}\end{aligned}$$

In the normal form of SVD, $\mathbf{s}_2$ and $\mathbf{t}_2$ orthogonal to $\mathbf{s}_1$ and $\mathbf{t}_1$ were chosen, respectively.

- **Cholesky's decomposition**: If $\mathbf{A}$ is p.d., there exists a *unique* upper triangular matrix $\mathbf{R}$ with positive diagonal elements such that $\mathbf{A} = \mathbf{R}'\mathbf{R}$. This is useful for efficient numerical solutions, e.g., Monte Carlo simulations. The Cholesky decomposition is roughly twice as efficient as the LU decomposition for solving systems of linear equations.

17 Expectation and Variance-covariance

- For a random matrix $\mathbf{Z}$ and comfortable matrices, $E(\mathbf{AZB} + \mathbf{C}) = \mathbf{A}E(\mathbf{Z})\mathbf{B} + \mathbf{C}$.
- $\operatorname{Cov}(\mathbf{X}, \mathbf{Y}) = E[(\mathbf{X} - E(\mathbf{X}))(\mathbf{Y} - E(\mathbf{Y})')]$ and $\operatorname{Cov}(\mathbf{X}, \mathbf{X}) = \operatorname{Var}(\mathbf{X})$.
- $\operatorname{Cov}(\mathbf{AX}, \mathbf{BY}) = \mathbf{A}\operatorname{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}'$.
- $\operatorname{Var}(a\mathbf{X} + b\mathbf{Y}) = a^2\operatorname{Var}(\mathbf{X}) + ab[\operatorname{Cov}(\mathbf{X}, \mathbf{Y}) + \operatorname{Cov}(\mathbf{Y}, \mathbf{X})] + b^2\operatorname{Var}(\mathbf{Y})$. Note $\operatorname{Cov}(\mathbf{X}, \mathbf{Y}) \neq \operatorname{Cov}(\mathbf{Y}, \mathbf{X})$.
- $E(\mathbf{x}'\mathbf{A}\mathbf{x}) = \operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$, where $\boldsymbol{\Sigma} = \operatorname{Var}(\mathbf{X})$.
  - $E[(\mathbf{x} - \mathbf{b})'\mathbf{A}(\mathbf{x} - \mathbf{b})] = \operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \mathbf{b})'\mathbf{A}(\boldsymbol{\mu} - \mathbf{b})$ as $\operatorname{Var}(\mathbf{X} - \mathbf{b}) = \operatorname{Var}(\mathbf{X})$.
  - If $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$, $E(\mathbf{x}'\mathbf{A}\mathbf{x}) = \sigma^2\operatorname{tr}(\mathbf{A}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} = \sigma^2(\text{sum of the coefficient of } X_i^2) + (\mathbf{x}'\mathbf{A}\mathbf{x})_{\mathbf{x}=\boldsymbol{\mu}}$

18 Multivariate normal distribution

- If $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the density is $f(\mathbf{Y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = Ce^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}$, where $C = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}$:

  Proof: By SD, $\boldsymbol{\Sigma} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'$, where $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_p)$ and let $\mathbf{Z} = \mathbf{T}'(\mathbf{y} - \boldsymbol{\mu}) \Rightarrow \mathbf{y} = \mathbf{T}\mathbf{z} + \mathbf{u}$, then

  $$1 = \int_{\mathbf{y}\in\mathbb{R}^p} Ce^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}d\mathbf{y} = \int_{\mathbf{z}\in\mathbb{R}^p} Ce^{-\frac{1}{2}\mathbf{z}'\boldsymbol{\Lambda}^{-1}\mathbf{z}}|J|d\mathbf{z} = \int_{\mathbf{z}\in\mathbb{R}^p} Ce^{-\frac{1}{2}\sum_{i=1}^p z_i^2/\lambda_i}d\mathbf{z}$$

  since $|J| = |\det(d\mathbf{y}/d\mathbf{x})| = |\det(\mathbf{T})| = |\pm 1| = 1$. Further

  $$\int_{\mathbf{z}\in\mathbb{R}^p} Ce^{-\frac{1}{2}\sum_{i=1}^p z_i^2/\lambda_i}d\mathbf{z} = C\prod_{i=1}^p\int_{-\infty}^{\infty} e^{-\frac{1}{2}z_i^2/\lambda_i}dz_i = C\prod_{i=1}^p(\sqrt{2\pi\lambda_i}) = C(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}.$$

- For the above, $\mathbb{E}(\mathbf{z}) = \mathbf{0} \Rightarrow \mathbb{E}(\mathbf{y}) = \mathbf{u}$ and $\operatorname{Cov}(\mathbf{Y}) = \operatorname{Cov}(\mathbf{TZ} + \mathbf{u}) = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}' = \boldsymbol{\Sigma}$.
- Mgf of $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is $\psi_{\mathbf{Y}}(\mathbf{t}) = \exp(\boldsymbol{\mu}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})$.

  Proof: If $\boldsymbol{\mu} = 0$, the mgf of $\mathbf{y}_0 \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ is

  $$\mathbb{E}(e^{\mathbf{t}'\mathbf{y}_0}) = C\int e^{\mathbf{t}'\mathbf{y}_0}e^{-\frac{1}{2}\mathbf{y}_0'\boldsymbol{\Sigma}^{-1}\mathbf{y}_0}d\mathbf{y}_0 = C\int e^{-\frac{1}{2}[(\mathbf{y}_0-\boldsymbol{\Sigma}\mathbf{t})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_0-\boldsymbol{\Sigma}\mathbf{t})-\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}]}d\mathbf{y}_0 = e^{\frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}},$$

  so that $\mathbb{E}(e^{\mathbf{t}'\mathbf{y}}) = \mathbb{E}(e^{\mathbf{t}'(\mathbf{y}_0+\boldsymbol{\mu})}) = e^{\mathbf{t}'\boldsymbol{\mu}}\mathbb{E}(e^{\mathbf{t}'\mathbf{y}_0}) = e^{\boldsymbol{\mu}'\mathbf{t}+\frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}$.

- Let $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$ with rank m (full row rank), then $\mathbf{x} \sim N_m(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. Note: $\mathbf{A}$ must have full row rank to ensure $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' \succ \mathbf{O}$; otherwise $\mathbf{x}\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'\mathbf{x}$ can be zero for nonzero $\mathbf{x}$.
- All subsets of $\mathbf{y}$ are multivariate normal: Take $\mathbf{A} = (\mathbf{I}_k \mid \mathbf{O}) \in \mathbb{R}^{k\times n}$, $\mathbf{A}\mathbf{y} = (y_1, \dots, y_k) \sim N_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.
- For $\mathbf{a} \in \mathbf{R}^n \backslash \{\mathbf{0}\}$, $\mathbf{a}'\mathbf{y} \sim N_1(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$, i.e., a linear combination of y_i's is univariate normal.
- Suppose $\mathbb{E}(\mathbf{Y}) = \boldsymbol{\mu}$ and $\text{Var}(\mathbf{Y}) = \boldsymbol{\Sigma}$. $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \;\Leftrightarrow\; \mathbf{a}'\mathbf{Y}$ has a univariate normal for all $\mathbf{a}$.

  Proof: ($\Rightarrow$) See above. ($\Leftarrow$) If $\mathbf{t}'\mathbf{Y}$ has a univariate normal for all $\mathbf{t}$. By assumption, $\mathbf{t}'\mathbf{Y} \sim N(\mathbf{t}'\boldsymbol{\mu}, \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})$ and hence the mgf of $\mathbf{t}'\mathbf{Y}$ is $M_{\mathbf{t}'\mathbf{Y}}(s) = \mathbb{E}[e^{s(\mathbf{t}'\mathbf{Y})}] = \exp[(\mathbf{t}'\boldsymbol{\mu})s + \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}s^2/2]$. Putting $s = 1$ yields $M_{\mathbf{t}'\mathbf{Y}}(1) = \mathbb{E}(e^{\mathbf{t}'\mathbf{Y}}) = \exp[\mathbf{t}'\boldsymbol{\mu} + \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2] = M_{\mathbf{Y}}(\mathbf{t})$, which means that $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Yet, even though all marginals of $\mathbf{X}$ are normal, $\mathbf{X}$ may *not* be normally distributed (See 250A HW).
- Consider the joint density of $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \quad \sim \quad N_{p+q}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),$$

  then $X \perp\!\!\!\perp Y \Leftrightarrow \Sigma_{12} = \Sigma_{21}' = O_{p\times q}$ so that $f_{X,Y}(x, y \mid \mu_1, \mu_2, \Sigma) = f_X(x \mid \mu_1, \Sigma_{11}) f_Y(y \mid \mu_2, \Sigma_{22})$.

  Proof: Use MGF. $\psi_{X,Y}(t) = e^{t'\mu + t'\Sigma t} = e^{t_1'\mu_1 + t_1'\Sigma_{11}t_1 + t_2'\mu_2 + t_2'\Sigma_{22}t_2} = \psi_X(t_1)\psi_Y(t_2)$.
- In the above setting, the conditional density of X given $Y = y$ is

$$X \mid Y = y \quad \sim \quad N_p(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y - \mu_2), \Sigma_{11.2})$$

  The proof is below.
- Theorem 2.5: Let $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{U} = \mathbf{A}\mathbf{Y}$ and $\mathbf{V} = \mathbf{B}\mathbf{Y}$. Then $\mathbf{U} \perp \mathbf{V} \Leftrightarrow \text{Cov}[\mathbf{U}, \mathbf{V}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' = 0$.

19 Conditional multivariate normal distribution

- If $\mathbf{A}_{22}$ is invertible and given that

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \in \mathbb{R}^{p+q}, \quad \mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \in \mathbb{R}^{(p+q)\times(p+q)}.$$

  Let $\mathbb{E}(\mathbf{X}_1) = \boldsymbol{\mu}_1$ and $\mathbb{E}(\mathbf{X}_2) = \boldsymbol{\mu}_2$.

  Consider the transformation

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{O} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}.$$

  Since this is a linear transformation, the joint distribution is also multivariate normal with $\mathbb{E}(\mathbf{Y}) = \boldsymbol{\mu}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\boldsymbol{\mu}_2$, $\mathbb{E}(\mathbf{X}_2) = \boldsymbol{\mu}_2$. and covariance matrix

$$\text{Var}\begin{pmatrix} \mathbf{Y} \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{O} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{O} & \mathbf{I}_q \end{pmatrix}' = \begin{pmatrix} \mathbf{A}_{11.2} & \mathbf{O}' \\ \mathbf{O} & \mathbf{A}_{22} \end{pmatrix},$$

  which implies that $\mathbf{Y}$ and $\mathbf{X}_2$ are uncorrelated and then independent. Thus, the conditional distribution of $\mathbf{Y} \mid \mathbf{X}_2 = \mathbf{x}_2$ is the same as the marginal distribution of $\mathbf{Y}$:

$$\mathbf{Y} \mid \mathbf{X}_2 = \mathbf{x}_2 \sim N_p(\boldsymbol{\mu}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\boldsymbol{\mu}_2,\ \mathbf{A}_{11.2}).$$

  Further, because of this independence, $\mathbf{X}_1 = \mathbf{Y} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{X}_2$ given $\mathbf{X}_2 = \mathbf{x}_2$ is distributed as

$$\begin{aligned} \mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2 &\sim N_p(\boldsymbol{\mu}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\boldsymbol{\mu}_2 + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{x}_2,\ \mathbf{A}_{11.2}) \\ &\sim N_p(\boldsymbol{\mu}_1 + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2),\ \mathbf{A}_{11.2}) \end{aligned}$$

- If $\mathbf{A}_{22}$ is not invertible, consider the transformation with g-inverse of $\mathbf{A}_{22}$

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p & -\mathbf{A}_{12}\mathbf{A}_{22}^- \\ \mathbf{O} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}.$$

Then, covariance matrix

$$\begin{aligned} \operatorname{Var}\begin{pmatrix} \mathbf{Y} \\ \mathbf{X}_2 \end{pmatrix} &= \begin{pmatrix} \mathbf{I}_p & -\mathbf{A}_{12}\mathbf{A}_{22}^- \\ \mathbf{O} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & \mathbf{O}' \\ -\mathbf{A}_{22}^-\mathbf{A}_{21} & \mathbf{I}_q \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^-\mathbf{A}_{21} & \mathbf{A}_{12} - \mathbf{A}_{12}\mathbf{A}_{22}^-\mathbf{A}_{22} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & \mathbf{O}' \\ -\mathbf{A}_{22}^-\mathbf{A}_{21} & \mathbf{I}_q \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{A}_{11.2} - (\mathbf{A}_{12} - \mathbf{A}_{12}\mathbf{A}_{22}^-\mathbf{A}_{22})\mathbf{A}_{22}^-\mathbf{A}_{21} & \mathbf{A}_{12} - \mathbf{A}_{12}\mathbf{A}_{22}^-\mathbf{A}_{22} \\ \mathbf{A}_{21} - \mathbf{A}_{22}\mathbf{A}_{22}^-\mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}. \end{aligned}$$

For the top left,

$$\mathbf{A}_{11.2} - (\mathbf{A}_{12} - \mathbf{A}_{12}\mathbf{A}_{22}^-\mathbf{A}_{22})\mathbf{A}_{22}^-\mathbf{A}_{21} = \mathbf{A}_{11.2} - \mathbf{A}_{12}\mathbf{A}_{22}^-\mathbf{A}_{21} + \mathbf{A}_{12}\mathbf{A}_{22}^-\mathbf{A}_{22}\mathbf{A}_{22}^-\mathbf{A}_{21} = \mathbf{A}_{11.2}.$$

For the top right and bottom left,

$$\begin{aligned} \mathbf{A}_{12} - \mathbf{A}_{12}\mathbf{A}_{22}^-\mathbf{A}_{22} &= \mathbf{H}'\mathbf{A}_{22} - \mathbf{H}'\mathbf{A}_{22}\mathbf{A}_{22}^-\mathbf{A}_{22} = \mathbf{O} \\ \mathbf{A}_{21} - \mathbf{A}_{22}\mathbf{A}_{22}^-\mathbf{A}_{21} &= \mathbf{A}_{22}\mathbf{H} - \mathbf{A}_{22}\mathbf{A}_{22}^-\mathbf{A}_{22}\mathbf{H} = \mathbf{O}. \end{aligned}$$

since $C(\mathbf{A}_{21}) \subseteq C(\mathbf{A}_{22})$ implies that there exists $\mathbf{H}$ such that $\mathbf{A}_{21} = \mathbf{A}_{22}\mathbf{H}$ and $\mathbf{A}_{12} = \mathbf{H}'\mathbf{A}_{22}$.

Therefore, as for the previous case,

$$\begin{aligned} \mathbf{Y} \mid \mathbf{X}_2 = \mathbf{x}_2 &\sim N_p(\boldsymbol{\mu}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^-\boldsymbol{\mu}_2,\ \mathbf{A}_{11.2}) \\ \Rightarrow \mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2 &\sim N_p(\boldsymbol{\mu}_1 + \mathbf{A}_{12}\mathbf{A}_{22}^-(\mathbf{x}_2 - \boldsymbol{\mu}_2),\ \mathbf{A}_{11.2}) \end{aligned}$$

20 Multivariate t distribution

Let $Y = (Y_1, Y_2, \dots, Y_p)'$ is said to have a multivariate t distribution if its PDF is given by

$$f(y) = \frac{\Gamma(\frac{1}{2}(\nu+n))}{(\pi\nu)^{n/2}\Gamma(\frac{1}{2}\nu)} |\Sigma|^{-1/2} \left[1 + \frac{(y-\mu)'\Sigma^{-1}(y-\mu)}{\nu}\right]^{-(p+\nu)/2},$$

where $\Sigma \succ O$. We say $Y \sim t_p(\nu, \mu, \Sigma)$. This distribution has the following properties:

- If $\Sigma = (\sigma_{ij})$, then $(Y_i - \mu_i)/\sqrt{\sigma_{ii}} \sim t_\nu$.
- $(Y-\mu)'\Sigma^{-1}(Y-\mu) \sim F_{n,\nu}$.

21 Quadratic form

- Let X be a p-dimensional random variable with mean μ and covariance Σ (not assumed normal yet). Consider the quadratic form $Q = x'Ax$ for some comfortable A. Then $\mathbb{E}(Q) = \operatorname{tr}(A\Sigma) + \mu' A \mu$. *Proof*:

$$\begin{aligned} \mathbb{E}(x'Ax) = \mathbb{E}[\operatorname{tr}(x'Ax)] = \mathbb{E}[\operatorname{tr}(Axx')] &= \mathbb{E}[\operatorname{tr}(A(x-\mu)(x-\mu)' + A\mu\mu')] \\ &= \operatorname{tr} A\mathbb{E}[((x-\mu)(x-\mu)'] + \operatorname{tr}(A\mu\mu'). \end{aligned}$$

- Example: Consider the mean of a sample variance $S^2 = \sum_{i=1}^n (x_i - \overline{x})^2/(n-1)$, where $x_i \sim N(\mu, \sigma^2)$:

$$(n-1)S^2 = \sum_{i=1}^n (x_i - \overline{x})^2 = x'\left(I_n - \frac{1_n 1_n'}{n}\right)x := x'Ax,$$

so that $\mathbb{E}(n-1)S^2 = \operatorname{tr}(A\Sigma) + x'Ax|_{x=\mu} = \sigma^2 \operatorname{tr}(A) + 0 = \sigma^2(n-1) \ \Rightarrow\ \mathbb{E}S^2 = \sigma^2$.

- Let $y \sim N_p(0, \mathbf{I}_p)$ and let A be symmetric. Then $Q = y'Ay \sim \chi^2_r(0) \Leftrightarrow A$ is idempotent of rank r:

  Proof: ($\Leftarrow$) Using the spectral decomposition of A,

$$\begin{aligned} Q &= y'T\Lambda T'y = z'\Lambda z = \sum_{i=1}^{r} z_i^2 \sim \chi^2_r(0) \quad \because z = T'y \sim N_n(0,\ T'T = I_n) \\ &= y'T_1T_1'y = z_r'z_r = \sum_{i=1}^{r} z_i^2 \sim \chi^2_r(0) \quad \because z_r = T_1'y \sim N_r(0,\ T_1'T_1 = I_r) \end{aligned}$$

  ($\Rightarrow$) Express the MGF of $Q = y'Ay \sim \chi^2_r$ with A, which is known. For $t < 1/2$,

$$\frac{1}{(1-2t)^{r/2}} = E(e^{Qt}) = \int (2\pi)^{-p/2} \exp\left[-\frac{y'(I-2tA)y}{2}\right] dy = \frac{1}{|I-2tA|^{1/2}} = \prod_{i=1}^{p} \frac{1}{(1-2t\lambda_i)^{1/2}}$$

  by SD. It follows that r of p eigenvalues have to be 1 and the others 0 so that A is idempotent.

- If $y \sim N_p(0, \mathbf{\Sigma})$, then $Q = y'Ay \sim \chi^2_r(0) \Leftrightarrow A\Sigma$ is idempotent of rank r, or equivalently, $A\Sigma A = A$.

  Proof: Let $x = \Sigma^{-1/2}y \sim N(0, I_p) \Rightarrow y = \Sigma^{1/2}x$, then $Q = x'\Sigma^{1/2}A\Sigma^{1/2}x$. By the above theorem,

$$\begin{aligned} Q = x'\Sigma^{1/2}A\Sigma^{1/2}x \sim \chi^2_r \quad &\Leftrightarrow \quad x'\Sigma^{1/2}A\Sigma^{1/2}x \text{ is idempotent of rank } r \\ &\Leftrightarrow \quad (\Sigma^{1/2}A\Sigma^{1/2})(\Sigma^{1/2}A\Sigma^{1/2}) = \Sigma^{1/2}A\Sigma^{1/2} \\ &\Leftrightarrow \quad A\Sigma A = A \quad \Leftrightarrow \quad A\Sigma A\Sigma = A\Sigma \end{aligned}$$

  with $r = \text{rank}(\Sigma^{1/2}A\Sigma^{1/2}) = \text{tr}(\Sigma^{1/2}A\Sigma^{1/2}) = \text{tr}(A\Sigma) = \text{rank}(A\Sigma)$.

  Another solution: $r = \text{rank}(A) = \text{rank}(A\Sigma A) \le \text{rank}(A\Sigma) \le \text{rank}(A)$ and Mgf of $Q = \mathbf{y'Ay}$ is

$$E\left(e^{Qt}\right) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \int e^{-\frac{1}{2}\mathbf{y}'(\mathbf{\Sigma}^{-1}-2t\mathbf{A})\mathbf{y}'} d\mathbf{y} = \frac{1}{|\mathbf{\Sigma}^{-1}-2t\mathbf{A}|^{\frac{1}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} = \frac{1}{|\mathbf{I}-2t\mathbf{A\Sigma}|^{\frac{1}{2}}}.$$

  provided that $|t|$ is small enough. Note that if $\Sigma \succ O$ and $A' = A$, then $\Sigma + tA$ is also p.d. for small $|t|$.

- Let $y \sim N_p(0, \mathbf{I}_p)$, A_i is symmetric and $Q_i = y'A_iy \sim \chi^2_{r_i}$ for $i = 1, 2$. Then $Q_1 \perp Q_2 \Leftrightarrow A_1A_2 = O$.

  Proof: ($\Rightarrow$) $Q_1 \perp Q_2 \Rightarrow Q_1 + Q_2 = y'(A_1 + A_2)y \sim \chi^2_{r_1+r_2} \Rightarrow A_1 + A_2$ is idempotent by above, that is,

$$(A_1 + A_2)^2 = A_1 + A_2 \quad \Rightarrow \quad A_1A_2 + A_2A_1 = O.$$

  Left and right multiplications by A_1yield $A_1A_2 + A_1A_2A_1 = A_1A_2A_1 + A_2A_1 \Rightarrow A_1A_2 = A_2A_1 = O$.

  ($\Leftarrow$) Suppose $A_1A_2 = 0$. Find the Mgf of Q_1 and Q_2:

$$\begin{aligned} \psi_{Q_1,Q_2}(t_1, t_2) = E(e^{t_1Q_1+t_2Q_2}) &= \int (2\pi)^{-p/2} \exp\left[-\frac{1}{2}y'(I - 2t_1A_1 - 2t_2A_2)y\right] dy \\ &= \frac{1}{|I - 2t_1A_1 - 2t_2A_2|^{1/2}} \\ &= \frac{1}{|I-2t_1A_1|^{\frac{1}{2}}} \frac{1}{|I-2t_2A_2|^{\frac{1}{2}}} \quad \because A_1A_2 = O \\ &= \psi_{Q_1}(t_1)\psi_{Q_2}(t_2), \end{aligned}$$

  meaning that $Q_1 \perp Q_2$.

- If $y \sim N_p(0, \Sigma)$, A_i is symmetric and $Q_i = y'A_iy \sim \chi^2_{r_i}$ for $i = 1, 2$. Then $Q_1 \perp Q_2 \Leftrightarrow A_1\Sigma A_2 = O$.

  Proof 1: Same process as the above: ($\Rightarrow$) $Q_1 \perp Q_2 \Rightarrow (A_1 + A_2)\Sigma$ is idempotent by above, that is,

$$(A_1 + A_2)\Sigma(A_1 + A_2) = A_1 + A_2 \quad \Rightarrow \quad A_1\Sigma A_2 + A_2\Sigma A_1 = O \quad \Rightarrow \quad A_1\Sigma A_2 = A_2\Sigma A_1 = O.$$

($\Leftarrow$) Suppose $A_1\Sigma A_2 = 0$, $E(e^{t_1Q_1+t_2Q_2}) = |\Sigma|^{-\frac{1}{2}}|\Sigma^{-1}-2tA_1-2tA_2|^{-\frac{1}{2}} = |I-2tA_1\Sigma|^{-\frac{1}{2}}|I-2tA_2\Sigma|^{-\frac{1}{2}}$.

Proof 2: Let $x = \Sigma^{-\frac{1}{2}}y \sim N_p(0, \mathbf{I}_p)$, then $Q_i = x'\Sigma^{\frac{1}{2}}A_1\Sigma^{\frac{1}{2}}x$. Hence, by the above theorem,

$$Q_1 \perp Q_2 \quad \Leftrightarrow \quad \Sigma^{\frac{1}{2}}A_1\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}A_2\Sigma^{\frac{1}{2}} = O \quad \Leftrightarrow \quad A_1\Sigma A_2 = O.$$

- Let $y \sim N_p(0, I_p)$. If $Q_1 - Q_2 \geq 0$ and $Q_i = y'A_iy \sim \chi^2_{r_i}$ for $i = 1, 2$ then

$$Q_1 - Q_2 \perp\!\!\!\perp Q_2, \quad Q_1 - Q_2 \sim \chi^2_{r_1-r_2}.$$

Proof: Since $Q_1 - Q_2 = y'(A_1 - A_2)y \geq 0$, $\forall y \in \mathbb{R}^p$, take $z \in N(A_1)$ to obtain

$$0 \leq z'(A_1 - A_2)z = -z'A_2z \leq 0 \quad \because A_2 \succeq O$$

so that $z'A_2z = z'A_2^2z = \|A_2z\|^2 = 0 \Rightarrow A_2z = 0 \Rightarrow z \in N(A_2)$. So we have $N(A_1) \subseteq N(A_2)$. Specifically, $(I_p - A_1)y \in N(A_1)$ since $A_1(I_p - A_1)y = (A_1 - A_1^2)y = 0$. It follows that

$$A_2(I_p - A_1)y = 0,\ \forall y \quad \Rightarrow \quad A_2 - A_2A_1 = O \quad \text{and} \quad A_2 - A_1A_2 = O \quad \because A_1' = A_1, A_2' = A_2.$$

Using the equation to get $(A_1 - A_2)A_2 = A_1A_2 - A_2^2 = A_1A_2 - A_2 = O \Rightarrow Q_1 - Q_2 \perp\!\!\!\perp Q_2$ and

$$\begin{aligned}(A_1 - A_2)^2 &= A_1 - A_1A_2 - A_2A_1 + A_2 = A_1 - A_2,\\ \text{rank}(A_1 - A_2) &= \text{tr}(A_1 - A_2) = \text{tr}(A_1) - \text{tr}(A_2) = r_1 - r_2,\end{aligned}$$

which shows $Q_1 - Q_2 \sim \chi^2_{r_1-r_2}(0)$.

- If $\mathbf{y} \sim N_p(\mathbf{m}, \mathbf{I}_p)$ and $\mathbf{A}$ is **idempotent** of rank k. Then $(\mathbf{y}-\mathbf{a})'\mathbf{A}(\mathbf{y}-\mathbf{a}) \sim \chi^2_k((\mathbf{m}-\mathbf{a})'\mathbf{A}(\mathbf{m}-\mathbf{a}))$.

Proof: Let $\mathbf{z} = \mathbf{y} - \mathbf{a}$, then $\mathbf{z} \sim N_p(\mathbf{m} - \mathbf{a}, \mathbf{I}_p)$. By the spectral decomposition, we obtain

$$\mathbf{A} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}' = \mathbf{T}_1\mathbf{T}_1',$$

where $\mathbf{T}$ is orthogonal, $\mathbf{T}_1$ has k column eigenvectors corresponding to eigenvalues 1 ($\mathbf{A}\mathbf{T}_1 = \mathbf{T}_1$), and $\mathbf{T}_2$ consists of $p-k$ column eigenvectors corresponding to eigenvalues 0 ($\mathbf{A}\mathbf{T}_2 = \mathbf{O}$). Then

$$\mathbf{z}'\mathbf{A}\mathbf{z} = (\mathbf{T}_1'\mathbf{z})'\mathbf{T}_1'\mathbf{z} \sim \chi^2_k(\|\mathbf{T}_1'(\mathbf{m}-\mathbf{a})\|^2) \sim \chi^2_k((\mathbf{m}-\mathbf{a})'\mathbf{A}(\mathbf{m}-\mathbf{a}))$$

since $\mathbf{T}_1'\mathbf{z} \sim N_k(\mathbf{T}_1'(\mathbf{m}-\mathbf{a}), \mathbf{I}_k)$ and $\|\mathbf{T}_1'(\mathbf{m}-\mathbf{a})\|^2 = (\mathbf{m}-\mathbf{a})'\mathbf{A}(\mathbf{m}-\mathbf{a})$.

- **Important!** In general, what if $y \sim N_p(\mu, \Sigma)$? We can write

$$Q = y'Ay = y'\Sigma^{-1/2}T(T'\Sigma^{1/2}A\Sigma^{1/2}T)T'\Sigma^{-1/2}y,$$

where T is orthogonal such that $T'(\Sigma^{1/2}A\Sigma^{1/2})T = D = (\lambda_1, \ldots, \lambda_p)$ by spectral decomposition of $\Sigma^{1/2}A\Sigma^{1/2}$. Note that $\text{rank}(D) = \text{rank}(\Sigma^{1/2}A\Sigma^{1/2}) = \text{rank}(A)$. Further let $z = T'\Sigma^{-1/2}y \sim N_p(T'\Sigma^{-1/2}\mu, I_p)$, so that

$$Q = z'Dz = \sum_{i=1}^p \lambda_i z_i^2, \quad \text{where} \quad z_i \sim N(t_i'\Sigma^{-1/2}\mu, 1) \Rightarrow z_i^2 \sim \chi^2_1\left((t_i'\Sigma^{-1/2}\mu)^2 = \mu'\right).$$

Hence, $Q = y'Ay$ is a weighted linear combination of independent noncentral χ^2 r.v.s with one degree of freedom and noncentrality parameters $\theta_i = (t_i'\Sigma^{-1/2}\mu)^2$.

The weights are non-zero eigenvalues of $\Sigma^{1/2}A\Sigma^{1/2}$, or equivalently, eigenvalues of $A\Sigma$ or ΣA because

$$|\Sigma^{1/2}A\Sigma^{1/2} - \lambda I_p| = |\Sigma^{1/2}||A\Sigma - \lambda I_p||\Sigma^{-1/2}| = |A\Sigma - \lambda I_p| = |\Sigma A - \lambda I_p|.$$

- Ex.1: Special case: When $A = \Sigma^{-1}$, then $\Sigma^{1/2}A\Sigma^{1/2} = I_p$, so that $D = I_p$ and $Q = z'z \sim \chi^2_p(\theta)$, where

$$\theta = \mu'\Sigma^{-1/2}\left(\sum_{i=1}^p t_it_i'\right)\Sigma^{-1/2}\mu = \mu'\Sigma^{-1}\mu.$$

- Ex.2: Common case: When $\Sigma = I_p$ and A is idempotent with $\text{rank}(A) = r \le p$, then

$$Q = \sum_{i=1}^{r} z_i^2 \quad \sim \quad \chi_r^2(\theta), \quad \text{where} \quad z_i^2 \sim \chi_1^2(\theta_i = \mu' t_i t_i' \mu)$$

with the noncentral parameter

$$\theta = \sum_{i=1}^{p} \theta_i = \sum_{i=1}^{r} \mu' t_i t_i' \mu = \mu' \left(\sum_{i=1}^{r} t_i t_i' \right) \mu = \mu' A \mu.$$

- Ex.3: When $\Sigma^{1/2} A \Sigma^{1/2}$ is idempotent, or equivalently $A\Sigma$ is idempotent, in other word, $A\Sigma A = A$ with $\text{rank}(A\Sigma) = r \le p$, then $D = I_r$ so that

$$Q = \sum_{i=1}^{r} z_i^2 \quad \sim \quad \chi_r^2(\theta), \quad \text{where} \quad z_i^2 \sim \chi_1^2(\theta_i = \mu' \Sigma^{-1/2} t_i t_i' \Sigma^{-1/2} \mu)$$

with the noncentral parameter

$$\theta = \sum_{i=1}^{p} \theta_i = \sum_{i=1}^{r} \mu' \Sigma^{-1/2} t_i t_i' \Sigma^{-1/2} \mu = \mu' \Sigma^{-1/2} \left(\sum_{i=1}^{r} t_i t_i' \right) \Sigma^{-1/2} \mu = \mu' A \mu.$$

22 Non-central chi-square distribution

- Define: Let $X_i \overset{ind}{\sim} N(\mu_i, 1)$, $i = 1, \ldots, n$ and $\mu = (\mu_1, \ldots, \mu_n)'$, then $Y = \sum_{i=1}^n$ is said to have a noncentral χ^2 distribution with n degrees of freedom and non-centrality parameter $\delta^2 = \sum_{i=1}^n \mu_i^2 = \|\mu\|^2$, or $Y \sim \chi_n^2(\|\mu\|^2)$. Why does the distribution of Y depend only on n and $\|\mu\|^2$.
- This Y can be expressed as the sum of a noncentral χ^2 with 1 df and a central χ^2 with $n-1$ dfs.

  Proof: Let $a_1 = \mu/\|\mu\|$ so that $a_1' a_1 = (\mu'\mu)/\|\mu\|^2 = 1$. Construct A with linearly independent rows:

$$A = \begin{pmatrix} a_1' \\ a_2' \\ \vdots \\ a_n' \end{pmatrix} \quad \text{s.t.} \quad A'A = AA' = I_n \ (a_i' a_j = \delta_{ij}).$$

  Then we have

$$W = AX \quad \sim \quad N_n\left(\begin{pmatrix} \|\mu\| \\ \mathbf{0} \end{pmatrix}, \ I_n \right)$$

  as $a_1'\mu = \|\mu\|$, $a_i'\mu = a_i'(a_1\|\mu\|) = 0$, $i = 2, \ldots, n$, and $\text{Cov}(W) = A\,\text{Cov}(X)A' = AA' = I_n$. Hence,

$$Y = \sum_{i=1}^{n} X_i^2 = X'X = X'A'AX = W'W = W_1^2 + \sum_{i=2}^{n} W_i^2,$$

  where $W_1 \sim N(\|\mu\|, 1) \Rightarrow W_1^2 \sim \chi_1^2(\|\mu\|^2)$ and $W_i \sim N(0,1)$, $i > 1 \Rightarrow \sum_{i=2}^n W_i^2 \sim \chi_{n-1}^2(0)$.
- Let $\delta = \|\mu\|^2$, then the mean of Y is

$$E(Y) = E(W_1^2) + E\left(\sum_{i=2}^{n} W_i^2 \right) = E((W_1 - \delta)^2 + \delta^2) + n - 1 = \text{Var}(W_1) + \delta^2 + n - 1 = n + \delta,$$

  while the variance of Y, $2n + 4\delta$, would be hard to obtain from this property.

- The pdf of Y is given by

$$f_Y(y;n,\delta) = \sum_{i=0}^{\infty} \frac{e^{-\delta/2}(\delta/2)^i}{i!} f_{X_{n+2i}}(y), \quad X_{n+2i} \sim \chi^2_{n+2i}(0),$$

which is a Poisson-weighted mixture of *central* chi-square distributions (See 250A HW).

- (HW8 in 250B). Equivalently, we say that, if $V \mid K \sim \chi^2_{p+2k}(0)$ and $K \sim \text{Pois}(\alpha'\alpha/2)$, then $V \sim \chi^2_p(\alpha'\alpha)$. In addition, if $U \sim N_p(\alpha, I)_p$, then since $U'U \sim \chi^2_p(\alpha'\alpha)$

$$E\left(\frac{1}{U'U}\right) = E\left[E\left(\frac{1}{U'U}\Big|K\right)\right] = E\left(\frac{1}{p+2K-2}\right).$$

- The MGF is given by

$$M_Y(t) = \exp\left(\frac{\delta t}{1-2t}\right)\frac{1}{(1-2t)^{n/2}}, \quad t < \frac{1}{2}.$$

Using $\psi(t) = \log M_Y(t) = \delta t/(1-2t) - (n/2)\log(1-2t)$, the mean and variance are

$$E(Y) = \psi'(0) = \frac{\delta}{(1-2t)^2} + \frac{n}{1-2t}\Big|_{t=0} = n+\delta,$$
$$\text{Var}(Y) = \psi''(0) = \frac{4\delta}{(1-2t)^3} + \frac{2n}{(1-2t)^2}\Big|_{t=0} = 2n+4\delta.$$

23 Non-central t distribution

- Suppose $X \sim N(\theta,1)$, $V \sim \chi^2_\nu(0)$, and $X \perp\!\!\!\perp V$, then $T = X/\sqrt{V/\nu}$ has a noncentral t distribution with noncentrality parameter θ, $T \sim t_\nu(\theta)$. The pdf of T is complicated. Mgf pf T does not exist.
- Using the above expression, the mean of T is

$$E(T) = E(X)\sqrt{\nu}E(V^{-1/2}) = \theta\sqrt{\frac{\nu}{2}}\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}, \quad \nu > 1;$$

otherwise, it does not exist.

- Example 1: Let $X_i \sim N(\mu,\sigma^2)$ and $S^2 = \sum_{i=1}^n (x_i - \overline{x})^2/(n-1)$. What is the dist. of $\sqrt{n}(\overline{X}-a)/S$?

$$\frac{\sqrt{n}(\overline{X}-a)}{\sigma} \sim N\left(\frac{\sqrt{n}(\mu-a)}{\sigma}, 1\right), \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}(0).$$

It follows that

$$T = \frac{\sqrt{n}(\overline{X}-a)/\sigma}{\sqrt{(n-1)S^2/(\sigma^2(n-1))}} = \frac{\sqrt{n}(\overline{X}-a)}{S} \quad \sim \quad t_{n-1}\left(\theta = \frac{\sqrt{n}(\mu-a)}{\sigma}\right).$$

- Example 2: Let $Y_i \overset{iid}{\sim} N(\mu,\sigma^2)$, $i = 1,2,3,4$. Find k such that

$$T = k\frac{(\overline{Y}-\mu_0)}{\sqrt{(y_1-y_2)^2 + (y_1+y_2-2y_3)^2/3 + (y_1+y_2+y_3-3y_4)^2/6}}$$

has a noncentral density. Note that $n=4$. Since $\overline{Y} \sim N(\mu,\sigma^2/4)$

$$X = \frac{2(\overline{Y}-\mu_0)}{\sigma} \sim N\left(\theta = \frac{2(\mu-\mu_0)}{\sigma}, 1\right).$$

Want to have the quadratic form for the denominator. Suppose

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1/\sqrt{3} & 1/\sqrt{3} & -2/\sqrt{3} & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} & -3/\sqrt{6} \end{pmatrix} \in \mathbb{R}^{3\times 4}$$

Then

$$W = Ay = \begin{bmatrix} y_1 - y_2 \\ (y_1 + y_2 - 2y_3)/\sqrt{3} \\ (y_1 + y_2 + y_3 - 3y_4)/\sqrt{6} \end{bmatrix} \quad \sim \quad N_3(A\mu = 0, \sigma^2 AA' = 2\sigma^2 I_3),$$

so that we have $W'W/(2\sigma^2) \sim \chi_3^2(0)$. Therefore,

$$T = \frac{X}{\sqrt{W'W/(2\sigma^2 \times 3)}} = \frac{2\sqrt{6}(\overline{Y} - \mu_0)}{\sqrt{W'W}} \quad \sim \quad t_3(\theta) \quad \Rightarrow \quad k = 2\sqrt{6}.$$

- Let (X_i, Y_i), $i = 1, \ldots, n$ be a random sample from the bivariate normal distribution with parameters m_1, m_2, v_1^2, v_2^2, and correlation r. If d is a fixed constant, find a constant k so that

$$T = \frac{k(\overline{X} - \overline{Y} - d)}{\sqrt{\sum_{i=1}^n (X_i - Y_i - \overline{X} + \overline{Y})^2}}.$$

Proof: Let $Z_i = X_i - Y_i$, $\overline{Z} = \overline{X} - \overline{Y}$ and $\mathbf{A} = (1 \;\; -1)$. Then

$$Z_i = \mathbf{A}\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left(\mathbf{Am} = m_1 - m_2, \mathbf{A\Sigma A'} = \nu^2\right),$$

where $\nu^2 = \nu_1^2 - 2r\nu_1\nu_2 + \nu_2^2$. It follows that

$$W := \frac{\sqrt{n}(\overline{Z} - d)}{\nu} \sim N\left(\frac{\sqrt{n}(m_1 - m_2 - d)}{\nu}, 1\right), \quad V := \frac{\sum_{i=1}^n (Z_i - \overline{Z})^2}{\nu^2} \sim \chi_{n-1}^2(0)$$

Since $\overline{Z} \perp \sum_{i=1}^n (Z_i - \overline{Z})^2$ and hence $W \perp V$ even though $r \neq 0$. Thus,

$$\frac{W}{\sqrt{\frac{V}{n-1}}} = \frac{\sqrt{n}(\overline{Z} - d)/\nu}{\sqrt{\frac{\sum_{i=1}^n (Z_i - \overline{Z})^2/\nu^2}{n-1}}} = \frac{\sqrt{n(n-1)}(\overline{Z} - d)}{\{\sum_{i=1}^n (Z_i - \overline{Z})^2\}^{1/2}} \sim t_{n-1}\left(\frac{\sqrt{n}(m_1 - m_2 - d)}{\nu}\right),$$

which is equivalent to T if $k = \sqrt{n(n-1)}$.

24 Non-central F distribution

- Suppose $X \sim \chi_n^2(\theta)$, $Y \sim \chi_m^2(0)$, and $V \perp\!\!\!\perp W$, then $F = (X/n)/(Y/m)$ has a noncentral F distribution with noncentrality parameter θ, $F \sim F_{n,m}(\theta)$. The pdf of F is complicated. Mgf of F does not exist.
- Using the above expression, the mean of T is

$$E(F) = \frac{m}{n}E(X)E(V^{-1}) = \frac{m}{n}\frac{n+\theta}{m-2}, \quad m > 2;$$

otherwise, it does not exist. Also, the variance of F is

$$\begin{aligned} \mathrm{Var}(F) = \left(\frac{m}{n}\right)^2 \mathrm{Var}\left(\frac{X}{Y}\right) &= \left(\frac{m}{n}\right)^2 \left[E(X^2)E(Y^{-2}) - E(X)^2 E(Y^{-1})^2\right] \\ &= \left(\frac{m}{n}\right)^2 \left[\frac{2n + 4\theta + (n+\theta)^2}{(m-2)(m-4)} - \frac{(n+\theta)^2}{(m-2)^2}\right] \\ &= 2\frac{(n+2\theta)(m-2) + (n+\theta)^2}{(m-2)^2(m-4)}\left(\frac{m}{n}\right)^2, \quad m > 4. \end{aligned}$$

- (Final) Let $x_1 = (1,1,1,1,1)'$ and $x_2 = (1,1,0,0,0)'$, $\theta = (6,6,2,2,2)'$, and $Y\ N_5(\theta, I_5)$. Let $V = \mathcal{L}(x_1, x_2)$ and let $\widehat{Y}$ be the orthogonal projection of Y onto V. Find a constant K so that

$$F = \frac{K\|\widehat{Y}\|^2}{\|Y - \widehat{Y}\|^2}.$$

Solution: We have $\|\widehat{Y}\| = Y'PY \sim \chi_2^2(\theta' P\theta)$ and $\|Y - \widehat{Y}\|^2 = Y'QY \sim \chi_3^2(\theta' Q\theta)$. Since $P_V\theta = \theta$ (need to calculate), $\theta' P\theta = \|\theta\|^2 = 84$ and $\theta' Q\theta = 0$. Hence, $F \sim F_{2,3}(84)$ with $K = 1.5$.

25 Independence theorem and lemma

- Let $y_p \sim N_p(0, I_p)$, $u = Ay$ and $v = By$. If $\mathrm{Cov}(u, v) = AB' = O$, then $u \perp\!\!\!\perp v$ and $u'u \perp\!\!\!\perp v'v$.
- Further let

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \in \mathbb{R}^{k\times p}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \in \mathbb{R}^{l\times p}$$

and $A_1 \in \mathbb{R}^{k_1\times p}$ and $B_1 \in \mathbb{R}^{l_1\times p}$ have linearly independent rows. Then

$$C = \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} \in \mathbb{R}^{(k_1+l_1)\times p}$$

is of full row rank since

$$C'x = (A_1' \mid B_1')\begin{pmatrix} x_{k_1} \\ x_{l_1} \end{pmatrix} = 0 \quad \Rightarrow \quad A_1' x_{k_1} = B_1' x_{l_1} = 0 \quad \Rightarrow \quad x = 0.$$

Let $u_1 = A_1 y$ and $v_1 = B_1 y$, then $u \perp\!\!\!\perp v \Rightarrow u_1 \perp\!\!\!\perp v_1$ since

$$u = Ay = \begin{pmatrix} A_1 y \\ A_2 y \end{pmatrix} = \begin{pmatrix} A_1 y \\ HA_1 y \end{pmatrix} = \begin{pmatrix} I_{k_1} \\ H_1 \end{pmatrix} u_1, \quad v = \begin{pmatrix} I_{l_1} \\ H_2 \end{pmatrix} v_1.$$

- **Craig's theorem**: If $y \sim N_p(0, \Sigma)$ and $Q_i = y' A_i y$. Then $Q_i \perp Q_j \Leftrightarrow A_i \Sigma A_j = O$.

  Proof: $\Leftarrow$ is derived from joint mgf of Q_i and Q_j, i.e., $E(e^{t_1 Q_1 + t_2 Q_2})$. The other direction is difficult.

  - This also holds in the general case: $y \sim N_p(m, \Sigma)$ (see HW5).
  - Especially, if $\Sigma = I$, then $Q_i \perp Q_j \Leftrightarrow A_i A_j = O$.

- **Loynes' lemma**. If $\mathbf{M}^2 = \mathbf{M} = \mathbf{M}'$, $\mathbf{P} = \mathbf{P}' \succeq \mathbf{O}$, and $\mathbf{I} - \mathbf{M} - \mathbf{P} \succeq \mathbf{O}$, then $\mathbf{MP} = \mathbf{PM} = \mathbf{O}$.

  Proof: Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} = \mathbf{Mx}$ then $\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{Mx} = \mathbf{y}'\mathbf{MMx} = \mathbf{y}'\mathbf{My}$. By assumption,

$$0 \le \mathbf{y}'(\mathbf{I} - \mathbf{M} - \mathbf{P})\mathbf{y} = -\mathbf{y}'\mathbf{Py} \le 0 \quad \because \mathbf{P} \succeq O.$$

  Hence, $\mathbf{y}'\mathbf{Py} = 0 \Rightarrow \|\mathbf{Py}\| = 0 \Rightarrow \mathbf{Py} = \mathbf{PMx} = 0$ for $\forall \mathbf{x} \Rightarrow \mathbf{PM} = \mathbf{O}$ and $(\mathbf{PM})' = \mathbf{MP} = \mathbf{O}$.

- **Marsaglia-Garaybill's Lemma**. If $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_q$ are symmetric $n \times n$ matrices, then **any of two** of the following statements imply the third:

  (i) $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_q$ are idempotent;

  (ii) $\mathbf{D}_i \mathbf{D}_j = \mathbf{O}$, $\forall i \neq j$.

  (iii) $\mathbf{D} = \mathbf{D}_1 + \mathbf{D}_2 + \cdots + \mathbf{D}_q$ is idempotent;

  Proof:

  - (i) + (ii) $\to$ (iii): $\mathbf{D}^2 = \left(\sum_{i=1}^q \mathbf{D}_i\right)^2 = \sum_{i=1}^q \mathbf{D}_i^2 + \sum_{i\neq j} \mathbf{D}_i \mathbf{D}_j = \sum_{i=1}^q \mathbf{D}_i = \mathbf{D}$.

- (i) + (iii) → (ii): Consider

$$\mathbf{I} - \mathbf{D}_i - \mathbf{D}_j = (\mathbf{I} - \mathbf{D}) + (\mathbf{D} - \mathbf{D}_i - \mathbf{D}_j).$$

  $\mathbf{I}-\mathbf{D} \succeq \mathbf{O}$ by (iii) and $\mathbf{D}-\mathbf{D}_i-\mathbf{D}_j = \sum_{k\neq i,j} \mathbf{D}_k \succeq \mathbf{O}$ by (i), so that $\mathbf{I}-\mathbf{D}_i-\mathbf{D}_j \succeq \mathbf{O} \Rightarrow \mathbf{D}_i\mathbf{D}_j = \mathbf{O}$ by Loynes' lemma.

- (ii) + (iii) → (i): Let $\mathbf{D}_i\mathbf{x} = \lambda_i\mathbf{x}$ for $\mathbf{x} \neq \mathbf{0}$. If $\lambda \neq 0$, then, by (ii),

$$\mathbf{D}\mathbf{x} = \frac{\mathbf{D}\mathbf{D}_i\mathbf{x}}{\lambda_i} = \frac{\mathbf{D}_i^2\mathbf{x}}{\lambda_i} = \mathbf{D}_i\mathbf{x},$$

  which implies that $\mathbf{D}_i$ has the same nonzero eigenvalues of $\mathbf{D}$. By (iii), $\lambda_i = 1$.

- **Cochran's theorem.** Let $\mathbf{y} \sim N_p(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{y}'\mathbf{y} = \sum_{i=1}^k Q_i = \sum_{i=1}^k \mathbf{y}'\mathbf{A}_i\mathbf{y}$, where $\text{rank}(\mathbf{A}_i) = r_i$, $i = 1, \ldots, k$. Then the following statements are equivalent:

  (i) $Q_i \perp Q_j$ for $1 \leq i \neq j \leq k$;

  (ii) $Q_i \sim \chi^2_{r_i}(0)$, $i = 1, \ldots, k$;

  (iii) $\sum_{i=1}^k r_i = p$.

  Proof:

  - (i) → (ii): $Q_i \perp Q_j \Rightarrow \mathbf{A}_i\mathbf{A}_j = \mathbf{O}$ by Craig, and $\mathbf{I} = \sum_{i=1}^k \mathbf{A}_i$ is idempotent. By **MG lemma**, $\mathbf{A}_i$ is idempotent, so that $Q_i \sim \chi^2_{r_i}(0)$.

    (Another solution) $Q_i \perp Q_j \Rightarrow Q_1 \perp Q_2 + \cdots + Q_k \Rightarrow \mathbf{A}_1(\mathbf{A}_2 + \cdots + \mathbf{A}_k) = \mathbf{O}$ by Craig $\Rightarrow \mathbf{A}_1(\mathbf{I} - \mathbf{A}_1) = \mathbf{O} \Rightarrow \mathbf{A}_1$ is idempotent.

  - (ii) → (iii): Since $\mathbf{I} = \sum_{i=1}^k \mathbf{A}_k$, and $\mathbf{A}_i$ is idempotent,

$$\sum_{i=1}^k r_i = \sum_{i=1}^k \text{tr}(\mathbf{A}_i) = \text{tr}\left(\sum_{i=1}^k \mathbf{A}_i\right) = \text{tr}(\mathbf{I}_p) = p.$$

  - (iii) → (i): Let $\alpha_1, \ldots, \alpha_{r_i}$ be eigenvalues of $\mathbf{A}_i$ and $\mathbf{T}$ be an orthogonal matrix such that $\mathbf{T}'\mathbf{A}_i\mathbf{T} = \text{diag}(\alpha_1, \ldots, \alpha_{r_i}, 0, \ldots 0)$ by the spectral decomposition. Then we can write

$$\mathbf{I} = \mathbf{T}'\mathbf{T} = \mathbf{T}'\mathbf{A}_i\mathbf{T} + \mathbf{T}'\mathbf{A}_{(-i)}\mathbf{T},$$

    where $\mathbf{A}_{(-i)} = \sum_{j\neq i}^k \mathbf{A}_j$, meaning that $\mathbf{T}'\mathbf{A}_{(-i)}\mathbf{T}$ also has to be orthogonal. Suppose $\mathbf{T}'\mathbf{A}_{(-i)}\mathbf{T} = \text{diag}(\beta_1, \ldots, \beta_p)$. Then

$$\mathbf{I} = \text{diag}(\alpha_1 + \beta_1, \ldots, \alpha_{r_i} + \beta_{r_1}, \beta_{r_i+1}, \ldots, \beta_p)$$

    yields $\beta_{r_i+1} = \cdots = \beta_p = 1$. Hence, $\text{rank}(\mathbf{T}'\mathbf{A}_{(-i)}\mathbf{T}) \geq p - r_i$. **However**,

$$\text{rank}(\mathbf{T}'\mathbf{A}_{(-i)}\mathbf{T}) = \text{rank}(\mathbf{A}_{(-i)}) = \text{rank}\left(\sum_{j\neq i}^k \mathbf{A}_j\right) \leq \sum_{j\neq i}^k \text{rank}(A_j) = p - r_i.$$

    leading to $\text{rank}(\mathbf{T}'\mathbf{A}_{(-i)}\mathbf{T}) = p - r_i \Rightarrow \beta_1 = \cdots = \beta_{r_i} = 0 \Rightarrow \alpha_1 = \cdots = \alpha_{r_i} = 1 \Rightarrow \mathbf{A}_i$ is (symmetric and) idempotent for $i = 1, \ldots, k$. This result and the fact that $\mathbf{I} = \sum_{i=1}^k \mathbf{A}_i$ is idempotent follow $\mathbf{A}_i\mathbf{A}_j = \mathbf{O}$ by **MG lemma** and hence $Q_i \perp\!\!\!\perp Q_j$ by Craig's theorem if $\mathbf{\Sigma} = \mathbf{I}$.

26 Orthogonal Projection

- Let $\Omega \subseteq V = \mathbb{R}^n$. Any $y \in V$ can be written *uniquely* as $y = u + v$, where $u \in \Omega$, $v \in \Omega^\perp$.

  Proof: Suppose $\dim(\Omega) = r$. Let $\{x_1, \ldots, x_r\}$ be an **orthogonal basis** for Ω. Expand this to an orthogonal basis for V by adding $\{x_{r+1}, \ldots, x_k\}$. Then $y \in V$ is expressed as

  $$y = \sum_{i=1}^{r} \alpha_i x_i + \sum_{i=r+1}^{k} \alpha_i x_i = u + v.$$

  If there are u_1, u_2, v_1, v_2 such that $u_i \in \Omega$ and $v_i \in \Omega^\perp$, $i = 1, 2$. Then we have $u_1 + v_1 = u_2 + v_2 \Rightarrow u_1 - u_2 = v_2 - v_1 \in \Omega \cap \Omega^\perp = \{0\} \Rightarrow u_1 = u_2$ and $v_1 = v_2$, which shows the uniqueness.

- Orthogonal projection of y on Ω is u, and then $y - u \in \Omega^\perp$ (residual). P_Ω such that $P_\Omega y = u \in \Omega$ is called the orthogonal projection matrix of y on Ω.
  - Let $\mathbf{P}_\Omega \mathbf{y} = \mathbf{u} \in \Omega$, then $\mathbf{y} - \mathbf{u} = (\mathbf{I} - \mathbf{P}_\Omega)\mathbf{y} = \mathbf{v} \in \Omega^\perp$
  - **Claim**: P_Ω is unique. *Proof*: If there are two such matrices P_Ω and $\widetilde{P_\Omega}$, then $P_\Omega y = u = \widetilde{P_\Omega} y \Rightarrow (P_\Omega - \widetilde{P_\Omega}) y = 0$ for $\forall y \in \mathbb{R}^n \Rightarrow \widetilde{P_\Omega} = P_\Omega$.

- How to find P_Ω: Again, let $\dim(\Omega) = r$ and $\{x_1, \ldots, x_r, x_{r+1}, \ldots, x_k\}$ be an orthogonal basis for $V = \mathbb{R}^n$. WLOG, assume that $\{x_1, \ldots, x_r\}$ is an orthogonal basis for Ω. Then, if $y \in V$, we can write

  $$y = \sum_{i=1}^{r} \alpha_i x_i + \sum_{i=r+1}^{k} \alpha_i x_i = u + v, \quad u \in \Omega,\ v \in \Omega^\perp.$$

  If $\ell = 1, \ldots, r$, then $(x_\ell, y) = x_\ell' y = \alpha_\ell$. Hence, the orthogonal projection u is given by

  $$u = \sum_{i=1}^{r} \alpha_i x_i = \sum_{i=1}^{r} (x_i' y) x_i = (x_1, \ldots, x_r) \begin{pmatrix} x_1' y \\ \vdots \\ x_r' y \end{pmatrix} = TT'y = P_\Omega y.$$

  Note that T has orthogonal columns but is not an orthogonal matrix as T is not symmetric.

- **Very importantly** (again), we can write $P = TT'$, where T has orthogonal columns (not symmetric), i.e., $T'T' = I_r$. T is *not* unique as there are an infinite number of orthogonal basis; but, P is unique.

- From above, P is the orthogonal projection matrix if and only if P is symmetric and idempotent.

  Proof:($\Rightarrow$) If $P = TT'$, which is obviously symmetric, and $(TT')(TT') = T(T'T)T' = TT'$ (idempotent). ($\Leftarrow$) If P is symmetric and idempotent, $P = U\Lambda U' = U_1 U_1'$, where $U = (U_1 \mid U_2)$ and U_1 has r orthogonal columns.

- Hence, we can write $y = P_\Omega y + (I_p - P_\Omega) y = u + v$, where $u \in \Omega$ and $v \in \Omega^\perp$.

- $\mathbf{P}_\Omega = \mathbf{X}(\mathbf{X}'\mathbf{X})^- \mathbf{X}'$ is the orthogonal projection matrix onto $\Omega = C(\mathbf{X})$:
  - Symmetric: $(\mathbf{X}(\mathbf{X}'\mathbf{X})^- \mathbf{X}')' = \mathbf{X}[(\mathbf{X}'\mathbf{X})^-]'\mathbf{X}' = \mathbf{X}[(\mathbf{X}'\mathbf{X})']^-\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$ since $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A} \Rightarrow \mathbf{A}'(\mathbf{A}^-)'\mathbf{A}' = \mathbf{A}' = \mathbf{A}'(\mathbf{A}')^-\mathbf{A}' \Rightarrow (\mathbf{A}^-)' = (\mathbf{A}')^-$ if $\mathbf{A}$ is symmetric.
  - Idempotent: We want to show $\mathbf{P}_\Omega^2 = \mathbf{P}_\Omega$ or $(\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}')(\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}') = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$, but we cannot use the second property of the Moore-Penrose inverse.

Proof: By the property of a g-inverse: $\mathbf{AA^-A} = \mathbf{A}$,

$$\begin{aligned}
\mathbf{X'X(X'X)^-X'X} = \mathbf{X'X} &\Rightarrow (\mathbf{X}^+)'\mathbf{X'X(X'X)^-X'X} = (\mathbf{X}^+)'\mathbf{X'X} \\
&\Rightarrow (\mathbf{XX}^+)'\mathbf{X(X'X)^-X'X} = (\mathbf{XX}^+)'\mathbf{X} \\
&\Rightarrow \mathbf{XX}^+\mathbf{X(X'X)^-X'X} = \mathbf{XX}^+\mathbf{X} \\
&\Rightarrow \mathbf{X(X'X)^-X'X} = \mathbf{X} \\
&\Rightarrow \mathbf{X(X'X)^-X'X(X'X)^-X'} = \mathbf{X(X'X)^-X'} \\
&\Rightarrow \mathbf{P}_\Omega^2 = \mathbf{P}_\Omega.
\end{aligned}$$

- If ω is a subspace of Ω (i.e., $\omega \subseteq \Omega$), $P_\omega P_\Omega = P_\Omega P_\omega = P_\omega$. *Proof*: Let $y \in V$, then $P_\omega y \in \omega \subseteq \Omega$. Then $P_\Omega(P_\omega y) = P_\omega y \Rightarrow (P_\Omega P_\omega - P_\omega)y = 0, \forall y$, so $P_\Omega P_\omega = P_\omega$. Take transpose to get $P_\omega P_\Omega = P_\omega$.
- Consider $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \in \mathbb{R}^n$, where $\mathbf{X}$ is not full rank and $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$. Then solving normal equations yields fitted vector $\widehat{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X(X'X)^-X'y} = \mathbf{P}_\Omega\mathbf{y}$, which is always UNIQUE even though $\mathbf{X}$ is *not* full column rank, in other words, $\mathbf{(X'X)^-}$ and $\widehat{\boldsymbol{\beta}} = \mathbf{(X'X)^-X'y}$ are *not* unique.

  Proof (again): Set $\mathbf{P}_\Omega$ and $\widetilde{\mathbf{P}}_\Omega$, where $\mathbf{P}_\Omega\mathbf{y} = \mathbf{u} = \widetilde{\mathbf{P}}_\Omega\mathbf{y} \Rightarrow (\mathbf{P}_\Omega - \widetilde{\mathbf{P}}_\Omega)\mathbf{y} = \mathbf{0}$ for $\forall\mathbf{y} \Rightarrow \mathbf{P}_\Omega = \widetilde{\mathbf{P}}_\Omega$.

27 Gauss Markov's theorem

- Consider $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\theta} \in C(\mathbf{X})$, where $\mathbf{X}$ has full column rank and $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. If $\widehat{\boldsymbol{\beta}}$ is an ordinary least square (OLS) estimate of $\boldsymbol{\beta}$, i.e., $\widehat{\boldsymbol{\beta}} = \mathbf{(X'X)^{-1}X'y}$, then $\widehat{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ has the property that $\mathbf{c'X}\widehat{\boldsymbol{\beta}} = \mathbf{c}'\widehat{\boldsymbol{\theta}}$ is the best linear unbiased estimator (BLUE) of $\mathbf{c'X}\boldsymbol{\beta} = \mathbf{c}'\boldsymbol{\theta}$, $\forall\mathbf{c}$.

  Proof: Suppose $a'y$ (linear combination of $y = (y_1, \ldots, y_n)'$ is a linear unbiased estimator of $c'\theta$, i.e., $\mathbb{E}(a'y) = c'\theta$. Since $\mathbb{E}(y) = X\beta = \theta$, $a'X\beta = c'X\beta$, $\forall\beta \Rightarrow a'X = c'X$. Also,

  $$\mathrm{Var}(a'y) = \sigma^2 a'a, \quad \mathrm{Var}(c'\widehat{\theta}) = \sigma^2 c' \,\mathrm{Var}(X\widehat{\beta})c = \sigma^2 c'(X(X'X)^{-1}X')c = \sigma^2 a' P_{C(X)} a,$$

  so that $\mathrm{Var}(a'y) - \mathrm{Var}(c'\widehat{\theta}) = \sigma^2 a'(I_n - P_X)a \succeq O \Rightarrow \mathrm{Var}(a'y) \succeq \mathrm{Var}(c'\widehat{\theta})$, which is minimum variance.
- Similarly, $c'\widehat{\beta}$ is BLUE for $c'\beta$: Suppose $a'y$ is a linear unbiased estimator of $c'\beta$, then $a'X\beta = c'\beta$, $\forall\beta \Rightarrow a'X = c'$. Then $\mathrm{Var}(a'y) - \mathrm{Var}(c'\widehat{\beta}) = \sigma^2 a'a - \sigma^2 c'(X'X)^{-1}c = \sigma^2 a'(I_n - P_X)a \succeq O$.

28 Estimability

- Consider $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \in \mathbb{R}^n$, where $\mathrm{rank}(\mathbf{X}) = r < p$ (not full rank) and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$ and $\boldsymbol{\beta} : p \times 1$.
- In a less-than-full-rank model, $\widehat{\beta}$ is not unique so that β is not estimable, that is, there is *no* linear unbiased estimate for β. Proof can be done by contradiction.

  Proof: If there *is* a linear unbiased estimator for β, we have $\mathbb{E}(a_i'y) = \beta_i$, $i = 1, \ldots, p$. Setting $A' = (a_1, \ldots, a_p)$ leads to $E(Ay) = \beta \Rightarrow AX\beta = \beta$, $\forall\beta \Rightarrow AX = I_p$. However, $p = \mathrm{rank}(I_p) = \mathrm{rank}(AX) \le \mathrm{rank}(X) = r$, which contradicts with $r < p$.
- However $\widehat{\theta} = \widehat{y} = X\widehat{\beta}$ *is* unique, so each element θ_i of $\theta = X\beta$ can be estimated as $\widehat{\theta}_i = x_i'\widehat{\beta}$.
- Definition: The parametric function $a'\beta$ is said to be estimable if it has a linear unbiased estimate, $b'Y$.
- By the discussion in Gauss Markov theorem, a $a'\beta$ is estimable if there exists a vector b such that $\mathbb{E}(b'Y) = a'\beta \Rightarrow b'X\beta = a'\beta, \forall\beta \Rightarrow X'b = a \in C(X')$ or $a' = b'X$.
- **Theorem**: $a'\beta$ is estimable if and only if $a' = a'(X'X)^-X'X$.

  Proof: ($\Leftarrow$) $a'(X'X)^-X'X = a' \Rightarrow a = X'X(X'X)^-a \in C(X')$. ($\Rightarrow$) If $a'\beta$ is estimable, then $a' = b'X \Rightarrow a'(X'X)^-X'X = b'X(X'X)^-X'X = b'P_XX = b'X = a'$.

- $a'\beta$ is estimable $\Leftrightarrow a \in C(X')$: $a'\beta = E(b'Y) = b'X\beta, \forall \beta$, so that $a' = b'X$ or $a = X'b$.
- By the above, $\text{Var}[a'\beta] = a' \text{Var}[(X'X)^{-}X'y]a = \sigma^2 a'[(X'X)^{-}X'X(X'X)^{-}]a = \sigma^2 a'(X'X)^{-}a$.
- If $a'\beta$ is estimable, $a'\widehat{\beta}$ is unique. *Proof* $a' = b'X \Rightarrow a'\beta = b'X\beta = b'\theta$. Similarly, $a'\widehat{\beta} = b'X\widehat{\beta} = b'\widehat{\theta}$, which is unique. By theorem for BLUE, $b'\widehat{\theta}$ is the BLUE of $b'\theta$, so that $a'\widehat{\theta}$ is the BLUE of $a'\theta$.
  - Since the GLS estimate is simply the OLS for a transformed model, $a'\widehat{\beta}_W$ is the BLUE of $a'\beta$. This implies that the OLS estimate $a'\widehat{\beta}$ is not BLUE in a less-than-full-rank model, although this still be unbiased. That is $E(a'\widehat{\beta}) = E(a'\widehat{\beta}_w) = a'\beta$, but $\text{var}[b'Y] \geq \text{var}[a'\widehat{\beta}] \geq \text{var}[a'\widehat{\beta}_W]$.
- $a'\mathbb{E}(\widehat{\beta})$ is an estimable function of β. *Proof*: $a'\mathbb{E}(\widehat{\beta}) = a'\mathbb{E}[(X'X)^{-}X'Y] = a'(X'X)^{-}X'X\beta = c'\beta$, where $c = X'X(X'X)^{-}a \in C(X')$.
- If $a'\widehat{\beta}$ is **invariant** with respect to $\widehat{\beta}$, $a'\beta$ is estimable. See HW.
- Suppose that $E(Y) = X\beta$ and $\text{Var}(Y) = \sigma^2 I_n$. $a'Y$ is the linear unbiased estimate of $E(a'Y)$ with minimum variance iff $\text{cov}(a'Y, b'Y) = 0$ for all b such that $E(b'Y) = 0$ (i.e., $b'X = 0'$).

  Proof: Suppose $c'Y = (a+b)'Y$. Then $E(c'Y) = c'X\beta = a'X\beta = E(a'Y)$ for $\forall b$ s.t. $E(b'Y) = 0$. Further $\text{var}(c'Y) = \text{var}(a'Y) + \text{var}(b'Y) + \text{cov}(a'Y, b'Y) \geq \text{var}(a'Y)$ with equality iff $\text{cov}(a'Y, b'Y) = 0$.
- **Example**: Consider a one-way ANOVA, $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $i = 1, \ldots, a$ (No. of group), $j = 1, \ldots, n_i$ (No. of obs in the ith group). Let $n = \sum_{i=1}^{a} n_i$. Then the model can be written as

$$\mathbb{E}(y) = X\beta \in \mathbb{R}^n$$
$$\mathbb{E}\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \vdots \\ y_{a1} \\ \vdots \\ y_{an_a} \end{pmatrix} = \underbrace{\begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 & \cdots & 0 \\ 1_{n_2} & 0 & 1_{n_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 1_{n_1} & 0 & 0 & \cdots & 1_{n_a} \end{pmatrix}}_{n\times(a+1)} \underbrace{\begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_a \end{pmatrix}}_{(a+1)\times 1},$$

where X has less than full rank as $\text{rank}(X) = a < (a+1)$. Calculate $(X'X)^{-}X'X$:

$$X'X = \underbrace{\begin{pmatrix} n & n_1 & \cdots & n_a \\ n_1 & n_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ n_a & 0 & \cdots & n_a \end{pmatrix}}_{(a+1)\times(a+1)} \quad \Rightarrow \quad (X'X)^{-} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & n_1^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_a^{-1} \end{pmatrix}.$$

Hence, the condition for $c'\beta$ to be estimable is $c' = c'(X'X)^{-}X'X$:

$$\begin{aligned}(c_0, c_1, \ldots, c_a) &= (c_0, c_1, \ldots, c_a)\begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & n_1^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_a^{-1} \end{pmatrix}\begin{pmatrix} n & n_1 & \cdots & n_a \\ n_1 & n_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ n_a & 0 & \cdots & n_a \end{pmatrix} \\ &= \left(\sum_{i=1}^{a} c_i, c_1, \ldots, c_a\right) \quad \Rightarrow \quad c_0 = \sum_{i=1}^{a} c_i.\end{aligned}$$

E.g., $(0, 1, -1, 0\ldots, 0)\beta = \tau_1 - \tau_2$ is estimable, but $(1, -1, 0, 0\ldots, 0)\beta = \mu - \tau_1$ is *not* estimable.

29 Distribution theory

- Consider $y = X\beta + \epsilon \sim N_n(X\beta, \sigma^2 I_n)$, where $X : n \times p$, $\text{rank}(X) = p$, and $\beta : p \times 1$. $Q_X = I_n - P_X$.
- $\widehat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$. So, the pdf is

$$f(\widehat{\beta} \mid \beta, \sigma^2) = (2\pi\sigma^2)^{-p/2}|(X'X)|^{-1/2} \exp\left[-\frac{(\widehat{\beta}-\beta)'X'X(\widehat{\beta}-\beta)}{2\sigma^2}\right].$$

- $(\widehat{\beta}-\beta)'X'X(\widehat{\beta}-\beta)/\sigma^2 \sim \chi_p^2(0)$ by above.
- $\widehat{\beta} \perp\!\!\!\perp y - \widehat{y}$ since $\text{Cov}(\widehat{\beta}, y - \widehat{y}) = \text{Cov}((X'X)^{-1}X'y, Q_X y) = (X'X)^{-1}X'(\sigma^2 I_n)Q_X' = 0$.
- $\widehat{\beta} \perp\!\!\!\perp S^2 = (y-\widehat{y})'(y-\widehat{y})/(n-p) = y'Q_X y/(n-p)$ by above.
- $\text{SSE}/\sigma^2 = (n-p)S^2/\sigma^2 = y'Q_X y/\sigma^2 \sim \chi_r^2(0)$, where $r = \text{rank}(Q_X) = n-p$.

  Proof: $y'Q_X'y = (y-X\beta)'Q_X(y-X\beta) = \epsilon' Q_X \epsilon$ and Q_X is symmetric and idempotent of rank $n-p$.
  - $E(\text{SSE}/\sigma^2) = n - p \Rightarrow E(\text{SSE}/(n-p)) = \sigma^2$.
  - Another solution: $E(y'Qy) = \text{tr}(Q\sigma^2 I_n) + \mu Q \mu = \sigma^2(n-p)$.
- MLE of β coincides with the least square estimate for β: $\widehat{\beta} = (X'X)^{-1}X'y$.
- MLE of σ^2 is $\widehat{\sigma}^2_{\text{MLE}} = \text{SSE}/n = \|y - X\widehat{\beta}\|^2/n$, which is biased, while $\widehat{\sigma}^2 = \text{SSE}/(n-p)$ is unbiased.
- The information matrix is given by

$$I = -E\left(\frac{\partial \ell^2}{\partial\theta\partial\theta'}\right) = \text{Var}\left[\frac{\partial \ell}{\partial \theta}\right] = \begin{bmatrix} -E\left(\frac{\partial \ell^2}{\partial\beta\partial\beta'}\right) & -E\left(\frac{\partial \ell^2}{\partial\beta\partial\sigma^2}\right) \\ -E\left(\frac{\partial \ell^2}{\partial\sigma^2\partial\beta'}\right) & -E\left(\frac{\partial \ell^2}{(\partial\sigma^2)^2}\right) \end{bmatrix} = \begin{pmatrix} \frac{X'X}{\sigma^2} & 0 \\ 0' & \frac{n}{2\sigma^4} \end{pmatrix},$$

  which gives us the multivariate Cramer-Rao lower bound for unbiased estimates of (β, σ^2), namely,

$$I^{-1} = \begin{pmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}.$$

  Since $\text{Var}(\widehat{\beta}) = \sigma^2(X'X)^{-1}$, which attains the lower bound, $\widehat{\beta}$ is said to be the minimum variance unbiased estimate (**MVUE**) of β.
- If columns in X are orthogonal each other, i.e., $X = (x_1, \ldots, x_p)$ and $x_i \perp\!\!\!\perp x_j$, $i \neq j$. Then since $X'X = \text{diag}(x_1'x_1, \ldots, x_p'x_p)$, the OLS estimate is given by

$$\widehat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} (x_1'x_1)^{-1} & & O \\ & \ddots & \\ O & & (x_p'x_p)^{-1} \end{pmatrix} \begin{pmatrix} x_1'Y \\ \vdots \\ x_p'Y \end{pmatrix} \quad \Rightarrow \quad \widehat{\beta}_j = (x_j'x_j)^{-1}x_j'Y,$$

  meaning that the OLS estimate of β_j, $\widehat{\beta}_j$, is unchanged if any of the other β_k $(k \neq j)$ equals zero. Also,

$$\text{SSE} = Y'Y - Y'P_XY = Y'Y - \widehat{\beta}'X'Y = Y'Y - \sum_{j=1}^p \widehat{\beta}_j x_j'Y = Y'Y - \sum_{j=1}^p \widehat{\beta}_j^2 \|x_j\|^2,$$

  which implies that if $\beta_j = 0$, the only change in the SSE is the *addition* of the term $\widehat{\beta}_j x_j'Y$ or $\widehat{\beta}_j\|x_j\|^2$.
- Example: Suppose x_{ij} are standardized so that for $j = 1, \ldots, p$, the sample mean is $\sum_i x_{ij} = 0$ and the sample variance $\sum_i x_{ij}^2 = c$. We now show that $(1/p)\sum_{j=1}^p \text{var}(\widehat{\beta}_j)$ is minimized when the column of X are mutually orthogonal.

Proof: Since the first column of X is unity, we have

$$X'X = \begin{pmatrix} n & 0' \\ 0 & C \end{pmatrix} \quad \Rightarrow \quad (X'X)^{-1} = \begin{pmatrix} n^{-1} & 0' \\ 0 & C^{-1} \end{pmatrix}.$$

$$\Rightarrow \sum_{j=1}^{p} \mathrm{var}(\widehat{\beta}_j) = \mathrm{tr}[\mathrm{Var}(\widehat{\beta})] = \sigma^2\,\mathrm{tr}[(X'X)^{-1}] = \sigma^2[\mathrm{tr}(C^{-1}) + n^{-1}] = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1},$$

where $\lambda_1 = n$ and λ_j $(j \geq 2)$ are eigenvalues of C. $\mathrm{tr}(C) = c(p-1) = \sum_j \lambda_j$ gives $\lambda_j = c$. So, there exists an orthogonal matrix T s.t. $C = T\Lambda T' = cI_p$, so that the column of X must be mutually orthogonal.

30 MLE for multivariate normal without using vector/matrix derivative

- Suppose $y_1, \ldots, y_n$ be a random sample from $N_p(\mu, V)$.
- Let $A = \sum_{i=1}^{n}(y_i - \overline{y})(y_i - \overline{y})' \succ O$, then log-likelihood is

$$\begin{aligned}
\ell(\mu, V) &= C - \frac{n}{2}\log|V| - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)'V^{-1}(y_i - \mu) \\
&= C - \frac{n}{2}\log|V| - \frac{1}{2}\sum_{i=1}^{n}(y_i - \overline{y})'V^{-1}(y_i - \overline{y}) - \frac{1}{2}\sum_{i=1}^{n}(\overline{y} - \mu)'V^{-1}(\overline{y} - \mu) \\
&\leq C - \frac{n}{2}\log|V| - \frac{1}{2}\mathrm{tr}(V^{-1}A)
\end{aligned}$$

with equality (maximum) when $\mu = \widehat{\mu} = \overline{y}$.

- Further let $\lambda_1, \ldots, \lambda_n$ be eigenvalues of $A^{1/2}V^{-1}A^{1/2}$,

$$\begin{aligned}
\ell(\widehat{\mu}, V) &= C - \frac{n}{2}\log|V| - \frac{1}{2}\mathrm{tr}(V^{-1}A) \\
&= C + \frac{n}{2}\log|V^{-1}| - \frac{1}{2}\mathrm{tr}(V^{-1}A) + \frac{n}{2}\log|A| - \frac{n}{2}\log|A| \\
&= \widetilde{C} - \frac{n}{2}\log|V^{-1}A| - \frac{1}{2}\mathrm{tr}(V^{-1}A) \\
&= \widetilde{C} - \frac{n}{2}\log|A^{1/2}V^{-1}A^{1/2}| - \frac{1}{2}\mathrm{tr}(A^{1/2}V^{-1}A^{1/2}) \\
&= \widetilde{C} - \frac{n}{2}\log\prod_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i=1}^{n}\lambda_i \\
&= \widetilde{C} - \frac{n}{2}\sum_{i=1}^{n}(n\log\lambda_i - \lambda_i).
\end{aligned}$$

Hence $\partial\ell/\partial\lambda_i = 0 \Rightarrow \widehat{\lambda}_i = n \Rightarrow A^{1/2}\widehat{V}^{-1}A^{1/2} = nI_n \Rightarrow \widehat{V}^{-1} = nA^{-1} \Rightarrow \widehat{V} = A/n$. Note $\partial^2\ell/\partial\lambda_i^2 < 0$.

31 Generalized Least Square Estimate

- Consider $y = X\beta + \epsilon$, where $\mathrm{Cov}(\epsilon) = \sigma^2 V$. Let $\widetilde{*} = V^{-1/2}*$, then $\widetilde{y} = \widetilde{X}\beta + \widetilde{\epsilon}$ and $\mathrm{Cov}(\widetilde{\epsilon}) = \sigma^2 I_n$, so

$$\widehat{\beta}_W = (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{y} = (X'V^{-1}X)^{-1}X'V^{-1}y,$$

which is said to be the *generalized* least square (GLS) estimate. If V is diagonal (not identity), then this can be called *weighted* least square (WLS) estimate.

- SSE (or RSS, residual sum of squares) is

$$\text{SSE} = (\widetilde{Y} - \widetilde{X}\widehat{\beta}_W)'(\widetilde{Y} - \widetilde{X}\widehat{\beta}_W) = (Y - X\widehat{\beta}_W)'V^{-1}(Y - X\widehat{\beta}_W)$$

- Let $P_{\widetilde{X}}$ be the orthogonal projection such that $P_{\widetilde{X}}y = \widetilde{X}\widehat{\beta}_W$, then

$$\text{SSE} = (\widetilde{Y} - \widetilde{X}\widehat{\beta}_W)'(\widetilde{Y} - \widetilde{X}\widehat{\beta}_W) = \widetilde{Y}'(I - P_{\widetilde{X}})\widetilde{Y} = (\widetilde{Y} - \widetilde{X}\beta)'(I - P_{\widetilde{X}})(\widetilde{Y} - \widetilde{X}\beta) = \widetilde{\epsilon}(I - P_{\widetilde{X}})\widetilde{\epsilon}$$
$$\Rightarrow \frac{\text{SSE}}{\sigma^2} = \frac{(Y - X\widehat{\beta}_W)'V^{-1}(Y - X\widehat{\beta}_W)}{\sigma^2} = \frac{\widetilde{\epsilon}(I - P_{\widetilde{X}})\widetilde{\epsilon}}{\sigma^2} \sim \chi^2_{n-p}(0)$$

as $\widetilde{\epsilon} \sim N(0, \sigma^2 I_n)$ and $\text{rank}(I_n - P_{\widetilde{X}}) = \text{tr}(I_n - P_{\widetilde{X}}) = n - p$.

- If $\epsilon \sim N_n(0, \sigma^2 V)$, then $\widehat{\beta}_W \sim N_p(\beta,\ \sigma^2(X'V^{-1}X)^{-1})$
- Suppose $V = \text{diag}(\omega_1, \ldots, \omega_n)$, where $\omega_i = \text{Var}(y_i)$. If ω_i depends only on the values of X, then errors are heteroscedastic.
- (Important) Let $\widehat{\beta}$ be OLS estimate. $\widehat{\beta}_W = \widehat{\beta} \ \Leftrightarrow\ C(V^{-1}X) = C(X) \ \Leftrightarrow\ C(VX) = C(X)$.

  Solution: We can write $Y = Y_1 + Y_2$, where $Y_1 \in \mathcal{C}(X)$ and $Y_2 \in \mathcal{C}(X)^\perp$. First, since $Y_1 \in \mathcal{C}(X)$, we write $Y_1 = Xa$, $\exists a$. Hence,

$$(X'V^{-1}X)^{-1}X'V^{-1}Y_1 = (X'V^{-1}X)^{-1}X'V^{-1}Xa = a = (X'X)^{-1}X'Y_1.$$

  Hence, need to show $(X'V^{-1}X)^{-1}X'V^{-1}Y_2 = (X'X)^{-1}X'Y_2$. Since $Y_2 \in \mathcal{C}(X)^\perp = \mathcal{N}(X')$. we have

$$(X'V^{-1}X)^{-1}X'V^{-1}Y_2 = 0 \quad \Leftrightarrow \quad X'V^{-1}Y_2 = 0.$$

  This holds iff $Y_2 \in \mathcal{N}(X'V^{-1}) = \mathcal{C}(V^{-1}X)^\perp$. Thus, $C(V^{-1}X)^\perp \subseteq C(X)^\perp \Leftrightarrow C(V^{-1}X) \supseteq C(X)$. However, by $\text{rank}(V^{-1}X) = \text{rank}(X)$ and the rank-nullity theorem, $C(V^{-1}X) = C(X)$. Finally,

$$\mathcal{C}(V^{-1}X) = \mathcal{C}(X) \quad \Leftrightarrow \quad V^{-1}X = XW \quad \Leftrightarrow \quad X = VXW \quad \Leftrightarrow \quad \mathcal{C}(X) = \mathcal{C}(VX).$$

  where W is a nonsingular matrix.

32 Add Regressions to a Model

- Assume that $\mathbb{E}(y) = X\beta$, where $\text{Var}(\epsilon) = \sigma^2 I$ and then $\widehat{\beta} = (X'X)^{-1}X'y$.
- Consider another model $G : \mathbb{E}(y) = X\beta + Z\gamma$, where the columns of X and Z are linearly independent.
- We can write the model G as

$$\mathbb{E}(y) = (X\ Z)\begin{pmatrix}\beta \\ \gamma\end{pmatrix} = W\delta.$$

- In a special case, if $X'Z = O$, i.e., they have columns that are orthogonal to one each other, then

$$\widehat{\delta}_G = \begin{pmatrix}\widehat{\beta}_G \\ \widehat{\gamma}_G\end{pmatrix} = (W'W)^{-1}W'y = \begin{pmatrix}X'X & O \\ O & Z'Z\end{pmatrix}^{-1}\begin{pmatrix}X' \\ Z'\end{pmatrix}y = \begin{pmatrix}(X'X)^{-1}X'y \\ (Z'Z)^{-1}Z'y\end{pmatrix}, \quad \Rightarrow \quad \widehat{\beta}_G = \widehat{\beta}.$$

- In the general case, let $P_X = X(X'X)^{-1}X'$ and $Q_X = I - P_X$. Then we can write G model as

$$\begin{aligned}\mathbb{E}(y) &= X\beta + P_X Z\gamma + Q_X Z\gamma \\ &= X[\beta + (X'X)^{-1}X'Z\gamma] + Q_X Z\gamma \\ &= X\alpha + Q_X Z\gamma.\end{aligned}$$

Since $XQ_X' = XQ_X = O$, as for the specific case,

$$\begin{aligned}
\widehat{\alpha} &= (X'X)^{-1}X'y = \widehat{\beta}_G + (X'X)^{-1}X'Z\widehat{\gamma}_G = \widehat{\beta}_G + L\widehat{\gamma}_G, \\
\widehat{\gamma}_G &= (Z'Q_X'Q_XZ)^{-1}Z'Q_Xy = (Z'Q_XZ)^{-1}Z'Q_Xy = MZ'Q_Xy, \\
\widehat{\beta}_G &= (X'X)^{-1}X'(y - Z\widehat{\gamma}_G) = \widehat{\beta} - L\widehat{\gamma}_G,
\end{aligned}$$

where $L = (X'X)^{-1}X'Z$ and $M = (Z'Q_XZ)^{-1}$.

Check that $Z'Q_XZ$ is nonsingular: Suppose $Z'Q_XZa = 0$. Then

$$a'Z'Q_XZa = \|Q_XZa\| = 0 \Rightarrow Q_XZa = 0 \Rightarrow Za = P_XZa = X(X'X)^{-1}X'Za \in C(X).$$

However, we have $X \perp\!\!\!\perp Z \Rightarrow C(Z) \cap C(X) = \{0\}$, so that $a = 0$, meaning that $Z'Q_XZ$ is invertible.

- Variance-covariance matrix is a bit complicated.

$$\begin{aligned}
\mathrm{Var}(\widehat{\gamma}_G) &= \sigma^2 MZ'Q_XZM = \sigma^2 M \\
\mathrm{Cov}(\widehat{\beta}, \widehat{\gamma}_G) &= \mathrm{Cov}((X'X)^{-1}X'y, MZ'Q_Xy) = \sigma^2(X'X)^{-1}X'Q_XZM = O \\
\mathrm{Var}(\widehat{\beta}_G) &= \mathrm{Var}(\widehat{\beta} - L\widehat{\gamma}_G) = \mathrm{Var}(\widehat{\beta}) + L\,\mathrm{Var}(\widehat{\gamma}_G)L' = \sigma^2[(X'X)^{-1} + LML'], \\
\mathrm{Cov}(\widehat{\beta}_G, \widehat{\gamma}_G) &= \mathrm{Cov}(\widehat{\beta} - L\widehat{\gamma}_G,\ \widehat{\gamma}_G) = O - L\,\mathrm{Var}(\widehat{\gamma}_G) = -\sigma^2 LM.
\end{aligned}$$

To summarize,

$$\mathrm{Cov}\begin{pmatrix}\widehat{\beta}_G \\ \widehat{\gamma}_G\end{pmatrix} = \sigma^2 \begin{pmatrix}(X'X)^{-1} + LML' & -LM \\ -ML' & M\end{pmatrix}.$$

We also see that $\mathrm{Var}(\widehat{\beta}_G) = \sigma^2[(X'X)^{-1} + LML'] \succeq \sigma^2(X'X)^{-1} = \mathrm{Var}(\widehat{\beta})$ because

$$a'LML'a = \|M^{1/2}L'a\|^2 \geq 0 \quad \Rightarrow \quad LML' \succeq O,$$

which means that adding regressors does not decrease the variance-covariance of β estimate.

- Let P_W be the orthogonal projection matrix on $C(W)$.

$$\begin{aligned}
\widehat{y}_{C(W)} = P_Wy = X\widehat{\beta}_G + Z\widehat{\gamma}_G \\
&= X(\widehat{\beta} - L\widehat{\gamma}_G) + Z\widehat{\gamma}_G \\
&= P_Xy + Q_XZ\widehat{\gamma}_G \quad \because XL = P_XZ \\
&= (P_X + Q_XZMZ'Q_X)y,\ \forall y.
\end{aligned}$$

It follows that $P_W = P_X + Q_XZ(Z'Q_XZ)^{-1}Z'Q_X$.

- Using this, SSE in G model is given by

$$\begin{aligned}
\mathrm{SSE}_G = y'(I - P_W)y &= y'(I - P_X - Q_XZ(Z'Q_XZ)^{-1}Z'Q_X)y \\
&= \mathrm{SSE} - y'Q_XZMZ'Q_Xy \preceq \mathrm{SSE}
\end{aligned}$$

since $y'Q_XZMZ'Q_Xy = \|M^{1/2}Z'Q_Xy\|^2 \geq 0 \Rightarrow Q_XZMZ'Q_X \succ O$, meaning that adding regressors does not increase SSE.

33 Estimate under linear constraints

- Consider $y = X\beta + \epsilon$, where $\mathbb{E}(\epsilon) = 0$ and $\mathrm{Cov}(\epsilon) = \sigma^2 I_n$. Assume $X : n \times p$ and $\beta : p \times 1$.
- First, suppose $\mathrm{rank}(X) = p$ (full column rank).
- Want to estimate β such that $A\beta = c$, where $A : q \times p$ and $c : q \times 1$.

- The first method uses **Lagrange multiplier**: $f(\beta) = \|y - X\beta\|^2 + \lambda'(A\beta - c)$, where $\lambda \in \mathbb{R}^q$.

$$\frac{\partial f(\beta)}{\partial \beta} = -2X'(y - X\beta) + A'\lambda, \quad \frac{\partial f(\beta)}{\partial \lambda} = A\beta - c.$$

  Both derivatives equal to zero gives

$$\widehat{\beta}_H = \widehat{\beta} - (X'X)^{-1}A'\widehat{\lambda}_H/2 \quad \Rightarrow \quad A\widehat{\beta}_H = A\widehat{\beta} - A(X'X)^{-1}A'\widehat{\lambda}_H/2 = c$$

  so that $\lambda_H/2 = [A(X'X)^{-1}A']^{-1}(A\widehat{\beta} - c)$ and hence

$$\widehat{\beta}_H = \widehat{\beta} - (X'X)^{-1}A\widehat{\lambda}_H/2 = \widehat{\beta} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\widehat{\beta} - c).$$

- Second approach assumes there exists β_0 s.t. $A\beta_0 = c$. Then

$$\widetilde{y} = y - X\beta_0 = X(\beta - \beta_0) + \epsilon := X\gamma + \epsilon.$$

  Let $\theta = X\gamma \in C(X)$ and $A_1 = A(X'X)^{-1}X'$. Then

$$A_1\theta = A(X'X)^{-1}X'X\gamma = A(\beta - \beta_0) = 0 \quad \Rightarrow \quad \theta \in N(A_1).$$

  Thus, $\theta = C(X) \cap N(A_1) = \Omega \cap N(A_1) \equiv \omega \subseteq \Omega$. It follows that $\widehat{\theta} = P_\omega \widetilde{y}$.

- **Lemma 1**: If $\omega \subseteq \Omega$, then $P_\omega = P_\Omega - P_{\omega^\perp \cap \Omega}$.

  Proof: $\omega \subseteq \Omega \Rightarrow P_\omega P_\Omega = P_\omega$. So, $(P_\omega y)'(P_\Omega - P_\omega)y = y'P_\omega(P_\Omega - P_\omega)y = 0$, leading to

$$P_\omega \perp\!\!\!\perp P_\Omega - P_\omega \quad \Rightarrow \quad P_\Omega - P_\omega = P_{\omega^\perp \cap \Omega}.$$

- **Lemma 2**: $\omega^\perp \cap \Omega = C(P_\Omega A_1')$.

  Proof: Show that $\omega^\perp \cap \Omega \subseteq C(P_\Omega A_1')$ and $C(P_\Omega A_1') \subseteq \omega^\perp \cap \Omega$. We use

$$\omega^\perp \cap \Omega = (\Omega \cap N(A_1))^\perp \cap \Omega = (\Omega^\perp + C(A_1')) \cap \Omega.$$

  First, if $x \in \omega^\perp \cap \Omega = (\Omega^\perp + C(A_1')) \cap \Omega$, then

$$x = P_\Omega[A_1'\alpha + (I - P_\Omega)\beta] = P_\Omega A_1'\alpha \in C(P_\Omega A_1') \quad \Rightarrow \quad \omega^\perp \cap \Omega \subseteq C(P_\Omega A_1').$$

  Conversely, if $z \in C(P_\Omega A_1')$ and set $x \in \omega = \Omega \cap N(A_1)$, then $\exists b$

$$z'x = (P_\Omega A_1'b)'x = b'A_1P_\Omega x = b'A_1x = 0 \quad \Rightarrow \quad z \in \omega^\perp = \omega^\perp \cap \Omega \quad \Rightarrow \quad C(P_\Omega A_1') \subseteq \omega^\perp \cap \Omega.$$

- Therefore, the estimate of $\theta = X(\beta - \beta_0)$ is

$$\begin{aligned}
\widehat{\theta}_H = P_\omega\widetilde{y} &= (P_\Omega - P_{\omega^\perp \cap \Omega})\widetilde{y} \quad \text{by lemma 1} \\
&= (P_\Omega - P_{C(P_\Omega A_1')})\widetilde{y} \quad \text{by lemma 2} \\
&= (P_\Omega - P_\Omega A_1'(A_1P_\Omega A_1')^{-1}A_1P_\Omega)(y - X\beta_0) \\
&= P_\Omega(y - X\beta_0) - P_\Omega A_1'(A_1P_\Omega A_1')^{-1}A_1P_\Omega(y - X\beta_0) \\
&= X(\widehat{\beta} - \beta_0) - P_\Omega A_1'(A_1P_\Omega A_1')^{-1}A_1X(\widehat{\beta} - \beta_0) \\
&= X(\widehat{\beta} - \beta_0) - X(X'X)^{-1}A'(A(X'X)^{-1}A)^{-1}A(\widehat{\beta} - \beta_0) \quad \because P_\Omega A_1' = X(X'X)^{-1}A' \\
&= X(\widehat{\beta} - \beta_0) - X(X'X)^{-1}A'(A(X'X)^{-1}A)^{-1}(A\widehat{\beta} - c).
\end{aligned}$$

  Since we can write $\widehat{\theta}_H = X(\widehat{\beta}_H - \beta_0)$,

$$\begin{aligned}
& X\widehat{\beta}_H = X\widehat{\beta} - X(X'X)^{-1}A'(A(X'X)^{-1}A)^{-1}(A\widehat{\beta} - c) \\
\Rightarrow \quad & \widehat{\beta}_H = \widehat{\beta} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\widehat{\beta} - c) \quad \because \text{premultiply by } (X'X)^{-1}X'
\end{aligned}$$

  Indeed, $A\widehat{\beta}_H = A\widehat{\beta} - (A\widehat{\beta} - c) = c$. We can use $(X'X)^-$, which is more complicated as shown in the next section.

- $\text{var}(\widehat{\beta}_{Hj}) \le \text{var}(\widehat{\beta}_j)$ as

$$\begin{aligned}
\text{Var}(\widehat{\beta}_H) &= \text{Var}[(I - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A)\widehat{\beta}] \\
&= \sigma^2[(X'X)^{-1} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A(X'X)^{-1}] \\
&\preceq \sigma^2(X'X)^{-1} = \text{Var}(\widehat{\beta}).
\end{aligned}$$

- Show $\|Y - \widehat{Y}_H\|^2 = \|Y - \widehat{Y}\|^2 + \|\widehat{Y} - \widehat{Y}_H\|^2$ *wisely.*

  Proof: Need to show $(Y - \widehat{Y})'(\widehat{Y} - \widehat{Y}_H) = 0$. Let P_ω be the projection matrix onto $\omega = N(A_1) \cap \Omega$. Since $P_\Omega P_\omega = P_\omega P_\Omega = P_\omega$,

$$(Y - \widehat{Y})'(\widehat{Y} - \widehat{Y}_H) = Y'(I - P_\Omega)(P_\Omega - P_\omega)Y = Y'(P_\Omega - P_\omega - P_\Omega + P_\Omega P_\omega)Y = 0.$$

34 Design matrix of less than full rank

- Consider the randomized block design with two treatments and two blocks: $Y_{ij} = \mu + \alpha_i + \gamma_j + \epsilon_{ij}$, $i, j = 1, 2$. Then the model is

$$E(Y) = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \gamma_1 \\ \gamma_2 \end{pmatrix} = X\beta,$$

  where the columns X are linearly *dependent* ($\text{rank}(X) = 3$).

- We have two options for X to be of full rank. First, set $\alpha_2 = 0$ and $\gamma_2 = 0$, i,e, regard them as reference:

$$E(Y) = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \gamma_1 \end{pmatrix}$$

  and the second is that we use two identifiability constraints, $H\beta = 0$ or $\sum_i \alpha_i = 0$ and $\sum_j \gamma_j = 0$:

$$\begin{pmatrix} \theta \\ 0 \end{pmatrix} = \begin{pmatrix} X \\ H \end{pmatrix} \beta = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \gamma_1 \\ \gamma_2 \end{pmatrix},$$

  where the augmented matrix now has linearly *independent* columns. Thus given θ, β becomes unique,

- Suppose $\text{rank}(X) = r < p$ and still $A\beta = c \in \mathbb{R}^q$, where A (full row rank) with a_i' in rows and c are known. If each of $a_i'\beta$ is estimable for $i = 1, \ldots, q$, then $\forall m_i$, such that $\mathbb{E}(m_i'y) = a_i'\beta \Rightarrow m_i'X = a_i'$. Hence, we have $A = MX$, where $M = (m_1, \ldots, m_q)' \in \mathbb{R}^{q \times n}$ with rank q as $q = \text{rank}(A) = \text{rank}(MX) \le \text{rank}(M) \le q$.

- Recall that we consider $\mathbb{E}(\widetilde{y}) = X\gamma = \theta \in \Omega = C(X)$. Then $M\theta = MX\gamma = A(\beta - \beta_0) = 0$, so that $\theta \in N(M) \cap \Omega := \omega \subseteq \Omega$. Using this, we form $X\widehat{\gamma}_H = \widehat{\theta}_H = P_\omega \widetilde{y} = (P_\Omega - P_{C(P_\Omega M')})\widetilde{y}$. Since

$P_\Omega M' = X(X'X)^- X'M' = X(X'X)^- A'$ and $MP_\Omega M' = A(X'X)^- A'$, we also get the same formula:

$$\begin{aligned}
X(\widehat{\beta}_H - \beta_0) &= (P_\Omega - P_{C(P_\Omega M')})\widetilde{y} \\
&= (P_\Omega - P_\Omega M'(MP_\Omega M')^- MP_\Omega)(y - X\beta_0) \\
&= (I_n - P_\Omega M'(MP_\Omega M')^- M)(P_\Omega y - P_\Omega X\beta_0) \\
&= (I_n - P_\Omega M'(MP_\Omega M')^- M)X(\widehat{\beta} - \beta_0) \\
&= X(\widehat{\beta} - \beta_0) - P_\Omega M'(MP_\Omega M')^- MX(\widehat{\beta} - \beta_0) \\
&= X(\widehat{\beta} - \beta_0) - X(X'X)^- A'(A(X'X)^- A')^- A(\widehat{\beta} - \beta_0)
\end{aligned}$$

and similarly to full rank X,

$$\begin{aligned}
X\widehat{\beta}_H &= X\widehat{\beta} - X(X'X)^- A'(A(X'X)^- A')^- A(\widehat{\beta} - \beta_0) \\
\Rightarrow X'X\widehat{\beta}_H &= X'X\widehat{\beta} - X'X(X'X)^- A'(A(X'X)^- A')^- A(\widehat{\beta} - \beta_0) \\
\Rightarrow X'X\widehat{\beta}_H &= X'X\widehat{\beta} - A'(A(X'X)^- A')^- A(\widehat{\beta} - \beta_0)
\end{aligned}$$

since $X'X(X'X)^- A' = X'X(X'X)^- X'M' = X'P_\Omega M' = X'M' = A'$.

Moreover, importantly, if $A\beta = c$, where A is of less than full rank, then $\beta = A^- c$ is a solution (but not unique) as $A\beta = A(A^- c) = AA^- A\beta = A\beta = c$. Using this fact, we have

$$\begin{aligned}
\widehat{\beta}_H &= (X'X)^- X'X\widehat{\beta} - (X'X)^- A'(A(X'X)^- A')^- A(\widehat{\beta} - \beta_0) \\
&= (X'X)^- X'y - (X'X)^- A'(A(X'X)^- A')^- A(\widehat{\beta} - \beta_0) \\
&= \widehat{\beta} - (X'X)^- A'(A(X'X)^- A')^- (A\widehat{\beta} - c).
\end{aligned}$$

- Claim that $A(X'X)^- A'$ is invertible (nonsingular), i.e., $[A(X'X)^- A']^- = [A(X'X)^- A']^{-1}$. Since $A(X'X)^- A' = MP_\Omega M' = MP_\Omega P_\Omega M'$, enough to show $P_\Omega M'$ has full column rank ($\text{rank}(P_\Omega M') = q$).

  Lemma 3: $P_\Omega M'$ has full column rank $\Leftrightarrow \Omega^\perp \cap C(M') = \{0\}$. *Proof*:

  - ($\Rightarrow$) Suppose $\text{rank}(P_\Omega M') = q$ and set $z \in \Omega^\perp \cap C(M')$. First $z \in C(M')$ leads to $z = M'a$, $\exists a$. Further, since $z \in \Omega^\perp = C(X)^\perp = N(X')$, $0 = X'z = X'M'a = A'a \Rightarrow a = 0$ since A has full *row* rank. Hence $z = 0 \Rightarrow \Omega^\perp \cap C(M') = \{0\}$.
  - ($\Leftarrow$) Show the contraposition: $\text{rank}(P_\Omega M') < q \Rightarrow \Omega^\perp \cap C(M') \neq \{0\}$. Suppose $\text{rank}(P_\Omega M') < q$. For $\exists \alpha \in \mathbb{R}^q \backslash \{\mathbf{0}\}$ s.t. $\sum_{i=1}^q \alpha_i (P_\Omega m_i) = P_\Omega \sum_{i=1}^q \alpha_i m_i = 0$, so that $\sum_{i=1}^q \alpha_i m_i \in \Omega^\perp \cap C(M') \backslash \{0\}$.

  In conclusion, if $\Omega^\perp \cap C(M') = \{0\}$ or equivalently, $\text{rank}(P_\Omega M') = q$ (full column rank), $A(X'X)^- A'$ is invertible, so that $A\widehat{\beta}_H = A\widehat{\beta} - A(\widehat{\beta} - \beta_0) = c$ even if X has less than full column rank.

35 Hypothesis testing under linear constraints

- Go back to the condition where X has full column rank and then the constrained estimate of β is

$$\widehat{\beta}_H = \widehat{\beta} - (X'X)^{-1} A'[A(X'X)^{-1} A']^{-1} (A\widehat{\beta} - c). \tag{*}$$

- Want to test $H : A\beta = c$ vs $H_A : A\beta \neq c$, where A (full row rank) and c are known.
- Under H, the sum of square errors is given by

$$\text{SSE}_H = \|y - X\widehat{\beta}_H\|^2 = \|y - X\widehat{\beta} + X\widehat{\beta} - X\widehat{\beta}_H\|^2 = \text{SSE} + (\widehat{\beta}_H - \widehat{\beta})' X'X(\widehat{\beta}_H - \widehat{\beta})$$

  Substituting (*) into $\widehat{\beta}_H$ provides

$$\text{SSE}_H - \text{SSE} = \|\widehat{Y} - \widehat{Y}_H\|^2 = (A\widehat{\beta} - c)'[A(X'X)^{-1} A']^{-1} (A\widehat{\beta} - c).$$

Here $\widehat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1}) \Rightarrow A\widehat{\beta} - c \sim N_q(0, \sigma^2 A(X'X)^{-1}A')$ under H_0, leading to

$$\frac{\text{SSE}_H - \text{SSE}}{\sigma^2} = (A\widehat{\beta} - c)'[\sigma^2 A(X'X)^{-1}A']^{-1}(A\widehat{\beta} - c) \quad \sim \quad \chi_q^2(0).$$

- Note that *without* restrictions, what is the expectation of the difference in SSE? We have

$$A\widehat{\beta} - c \quad \sim \quad N_p(A\beta - c,\ \sigma^2 A(X'X)^{-1}A).$$

Hence, let $Z = A\widehat{\beta} - c$ and $B = A(X'X)^{-1}A$,

$$\begin{aligned}\mathbb{E}[\text{SSE}_H - \text{SSE}] = \mathbb{E}[Z'B^{-1}Z)] &= \text{tr}(B^{-1}(\sigma^2 B)) + (A\beta - c)'B^{-1}(A\beta - c) \\ &= \sigma^2 q + (A\beta - c)'[A(X'X)^{-1}A]^{-1}(A\beta - c) \\ &= \sigma^2 q + (\text{SSE}_H - \text{SSE})_{\widehat{\beta}=\beta}.\end{aligned}$$

implying that without restriction $(\text{SSE}_H - \text{SSE})/\sigma^2 \sim \chi_q^2(\lambda)$, where $\lambda = (A\beta - c)'B^{-1}(A\beta - c)$.

- We also have $\text{SSE}/\sigma^2 = y'Q_x y/\sigma^2 \sim \chi_{n-p}^2(0)$. Therefore, the F statistic for testing H_0 is

$$F = \frac{(\text{SSE}_H - \text{SSE})/q}{\text{SSE}/(n-p)} = \frac{n-p}{q}\frac{\text{SSE}_H - \text{SSE}}{\text{SSE}} \quad \sim \quad F_{q,n-p}(0) \quad \text{under } H_0.$$

This is not enough! Need to show $\text{SSE}_H - \text{SSE} \perp\!\!\!\perp \text{SSE}$.

Proof: Since $X'Q_X = X'(I_n - P_X) = O$, $X'y \perp\!\!\!\perp Q_X y$ by Craig's theorem, leading to

$$X'y \perp\!\!\!\perp Q_X y \ \Rightarrow \ (X'X)^{-1}X'y \perp\!\!\!\perp y'Q_X y \ \Rightarrow \ \widehat{\beta} \perp\!\!\!\perp \text{SSE} \ \Rightarrow \ \text{SSE}_H - \text{SSE} \perp\!\!\!\perp \text{SSE}$$

as $\text{SSE}_H - \text{SSE}$ is a function of $\widehat{\beta}$.

- Let $S_H^2 = (\text{SSE}_H - \text{SSE})/q$ and $S^2 = \text{SSE}/(n-p)$. From above, $\mathbb{E}[S_H^2] = \sigma^2 + \delta$, where $\delta \geq 0$ as $A(X'X)^{-1}A \succ O$ and $E(S^2) = \sigma^2$. When $H : A\beta = c$ is true, $\delta = 0$ so that $E(S_H^2)$ is also unbiased for σ^2, that is, $F = S_H^2/S^2 \approx 1$. When H is false, $\delta > 0$ and by $E(S_H^2) > E(S^2)$ and $S_H^2 \perp\!\!\!\perp S$,

$$F = E\left[\frac{S_H^2}{S^2}\right] = E[S_H^2]E\left[\frac{1}{S^2}\right] > \frac{E[S_H^2]}{E[S^2]} > 1.$$

Thus, we reject H if F is significantly large (Λ is small).

- Exercise: If $H : A\beta = c$ is true,

$$F = \frac{n-p}{q}\frac{\text{SSE}_H - \text{SSE}}{\text{SSE}} = \frac{n-p}{q}\frac{\epsilon'(P - P_H)\epsilon}{\epsilon'(I_n - P)\epsilon},$$

where $P_H = P - X(X'X)^{-1}A'B^{-1}A(X'X)^{-1}X'$ is symmetric and idempotent.

Proof: The denominator is obvious. Show that $\text{SSE}_H - \text{SSE} = \|\widehat{Y} - \widehat{Y}_H\|^2 = \epsilon'(P - P_H)\epsilon$. We have

$$\begin{aligned}\text{SSE}_H - \text{SSE} &= (A\widehat{\beta} - c)'[A(X'X)^{-1}A']^{-1}(A\widehat{\beta} - c) \\ &= (\widehat{\beta} - \beta)'A'[A(X'X)^{-1}A']^{-1}A(\widehat{\beta} - \beta) \\ &= (Y - X\beta)'X(X'X)^{-1}A'B^{-1}A(X'X)^{-1}X'(Y - X\beta) \\ &= \epsilon'(P - P_H)\epsilon.\end{aligned}$$

- Example (The Straight Line). Let $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$ or $E(Y) = X\beta$, where $X = (1, x)$ and $\beta = (\beta_0, \beta_1)$. Then we have

$$\begin{aligned}X'X &= \begin{pmatrix} n & 1'x \\ 1'x & x'x \end{pmatrix} = \begin{pmatrix} n & n\overline{x} \\ n\overline{x} & \sum_i x_i^2 \end{pmatrix} \\ \Rightarrow \quad (X'X)^{-1} &= \frac{1}{n\sum_i(x_i - \overline{x})^2}\begin{pmatrix} \sum_i x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{pmatrix} = \frac{1}{\sum_i(x_i - \overline{x})^2}\begin{pmatrix} \frac{1}{n}\sum_i x_i^2 & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}\end{aligned}$$

and so $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{x}$ and $\widehat{\beta}_1 = \sum_i (Y_i - \overline{Y})(x_i - \overline{x}) / \sum_i (x_i - \overline{x})^2$. Note that since $\mathrm{Var}(\widehat{\beta}) = \sigma^2 (X'X)^{-1}$, the correlation coefficient of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, ρ, is

$$\rho = \frac{\mathrm{cov}(\widehat{\beta}_0, \widehat{\beta}_1)}{\sqrt{\mathrm{var}(\widehat{\beta}_0)\,\mathrm{var}(\widehat{\beta}_1)}} = \frac{-n\overline{x}}{\sqrt{n\sum_i x_i^2}}.$$

F statistic for testing

– $H : \beta_1 = c$ is

$$\begin{aligned} F &= \frac{(A\widehat{\beta} - c)'[A(X'X)^{-1}A']^{-1}(A\widehat{\beta} - c)/q}{\mathrm{SSE}/(n-2)} \\ &= \frac{(\widehat{\beta}_1 - c)\left[1/\sum_i (x_i - \overline{x})^2\right]^{-1}(\widehat{\beta}_1 - c)/1}{S^2} \\ &= \frac{(\widehat{\beta}_1 - c)^2}{S^2/\sum_i (x_i - \overline{x})^2} \quad \sim \quad F_{1,n-2}. \end{aligned}$$

– $H : \beta_0 = c$ is

$$F = \frac{(\widehat{\beta}_0 - c)^2}{S^2 \sum_i x_i^2 / \left[n \sum_i (x_i - \overline{x})^2\right]} \quad \sim \quad F_{1,n-2}.$$

Also, the fitted value is given by $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i = \overline{Y} + \widehat{\beta}_1 (x_i - \overline{x})$ and hence

$$\begin{aligned} \mathrm{SSE} = \sum_i (Y_i - \widehat{Y}_i)^2 &= \sum_i \left[Y_i - \overline{Y} - \widehat{\beta}_1 (x_i - \overline{x})\right] \\ &= \sum_i (Y_i - \overline{Y})^2 - 2\widehat{\beta}_1 \sum_i (Y_i - \overline{Y})(x_i - \overline{x}) + \widehat{\beta}_1^2 \sum_i (x_i - \overline{x})^2 \\ &= \sum_i (Y_i - \overline{Y})^2 - \widehat{\beta}_1^2 \sum_i (x_i - \overline{x})^2 \\ &= \sum_i (Y_i - \overline{Y})^2 - \sum_i (\widehat{Y}_i - \overline{Y})^2. \end{aligned}$$

so that $\sum_i (Y_i - \overline{Y})^2 = \sum_i (Y_i - \widehat{Y}_i)^2 + \sum_i (\widehat{Y}_i - \overline{Y})^2 = \sum_i (Y_i - \widehat{Y}_i)^2 + r^2 \sum_i (Y_i - \overline{Y})^2$, where

$$r^2 = \frac{\sum_i (\widehat{Y}_i - \overline{Y})^2}{\sum_i (Y_i - \overline{Y})^2} = \frac{\widehat{\beta}_1^2 \sum_i (x_i - \overline{x})^2}{\sum_i (Y_i - \overline{Y})^2} = \frac{\left[\sum_i (Y_i - \overline{Y}))(x_i - \overline{x})\right]^2}{\sum_i (Y_i - \overline{Y})^2 \sum_i (x_i - \overline{x})^2},$$

which is the square of the *sample correlation* between Y and x. We can write

$$(1 - r^2) \sum_i (Y_i - \overline{Y})^2 = \sum_i (Y_i - \widehat{Y}_i)^2 = \mathrm{SSE}.$$

Using SSE and r, the F statistic for testing

– $H : \beta_1 = 0$ is

$$F = \frac{\widehat{\beta}_1^2 \sum_i (x_i - \overline{x})^2}{\mathrm{SSE}/(n-2)} = \frac{\widehat{\beta}_1^2 \sum_i (x_i - \overline{x})^2 (n-2)}{(1 - r^2) \sum_i (Y_i - \overline{Y})^2} = \frac{r^2 (n-2)}{1 - r^2}.$$

– $H : \beta_0 = 0$ is

$$F = \frac{n\widehat{\beta}_0^2 \sum_i (x_i - \overline{x})^2}{\sum_i x_i^2 \mathrm{SSE}/(n-2)} = \frac{n\overline{Y}^2 \sum_i (x_i - \overline{x})^2}{(1 - r^2) \sum_i (Y_i - \overline{Y})^2 \sum_i x_i^2}.$$

36 Exercise under linear constraints (1)

- Consider the standard linear model: $y = X\beta + \epsilon \in \mathbb{R}^n$, where $X : n \times p$, $\beta p \times 1$ and $\epsilon \sim N_n(0, \sigma^2 I_n)$.
- Here, set $n = 10$ and $p = 3$ and suppose X has *orthonormal* columns.
- We also have $X'y = (1, 2, 3)'$ and $y'y = 20$.
- Set $H : A\beta = c \in \mathbb{R}^q$, in particular, $\beta_1 + \beta_2 + \beta_3 = 2$ $(q = 1) \Rightarrow A = (1, 1, 1)$ and $c = 2$.
- (a) Find $\widehat{\beta}_H$. First $\widehat{\beta} = (X'X)^{-1}X'y = I_3(1, 2, 3) = (1, 2, 3)$. Hence,

$$\begin{aligned}\widehat{\beta}_H &= \widehat{\beta} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\widehat{\beta} - c) \\ &= \begin{pmatrix}1\\2\\3\end{pmatrix} - \frac{4}{3}\begin{pmatrix}1\\1\\1\end{pmatrix} = \frac{1}{3}\begin{pmatrix}-1\\2\\5\end{pmatrix}.\end{aligned}$$

- (b) Determine $\mathrm{Cov}(\widehat{\beta}_H)$.

$$\begin{aligned}\mathrm{Cov}(\widehat{\beta}_H) &= \mathrm{Cov}[(I_3 - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A)\widehat{\beta}] \\ &= \mathrm{Cov}[(I_3 - 3^{-1}1_3'1_3)\widehat{\beta}] \\ &= \sigma^2(I_3 - 3^{-1}1_3'1_3)^2 \preceq \sigma^2 I_3 = \mathrm{Cov}(\widehat{\beta}),\end{aligned}$$

  which implies that $\widehat{\beta}_H$ is biased but has lower variance than $\widehat{\beta}$, which is unbiased without this constraint.
- (c) Calculate F test statistics. First, we have $P_X = X(X'X)^{-1}X' = XX' \neq I_n$ and then

$$\mathrm{SSE} = yQy = y'y - y'Py = 20 - \|X'y\|^2 = 20 - 14 = 6.$$

  For the numerator,

$$\mathrm{SSE}_H - \mathrm{SSE} = (A\widehat{\beta} - c)'[A(X'X)^{-1}A']^{-1}(A\widehat{\beta} - c) = \frac{16}{3}.$$

  Therefore,

$$F = \frac{(\mathrm{SSE}_H - \mathrm{SSE})/q}{\mathrm{SSE}/(n-p)} = \frac{(16/3)/1}{6/(10-3)} = \frac{56}{9} = 6.22 \quad \sim \quad F_{1,7}.$$

  Since $F_{1,7,0.95} = 5.59$, we reject $H : \beta_1 + \beta_2 + \beta_3 = 2$.

37 Exercise under linear constraints (2)

- Given $Y = \theta + \epsilon$, where $\epsilon \sim N_4(0, \sigma^2 I_4)$ and $1_4'\theta = 0$, show the F-statistic for testing $H : \theta_1 = \theta_3$.
- First, using the Lagrange multiplier, that is, from $f(\theta) = \|Y - \theta\|^2 - \lambda(1'\theta)$, we have

$$\widehat{Y}_i = \widehat{\theta}_i = Y_i - \overline{Y} \quad \Leftrightarrow \quad \widehat{Y} = \widehat{\theta} = (I_4 - 11'/4)Y$$

  so that the denominator of the F-statistic is

$$S^2 = \frac{\mathrm{SSE}}{n-p} = \frac{\|Y - \widehat{Y}\|^2}{4-3} = 4\overline{Y}^2 = \frac{(1'Y)^2}{4}.$$

- For the numerator, we have two solutions, but $X'X$ is 3×3, so the calculation of $A(X'X)^{-1}A'$ would not be wise. Hence, use the Lagrange multiplier again. Set $\theta_1 = \theta_3$ then

$$f(\theta) = (Y_1 - \theta_1)^2 + (Y_2 - \theta_2)^2 + (Y_3 - \theta_1)^2 + (Y_4 - \theta_4)^2 - \lambda(2\theta_1 + \theta_2 + \theta_4).$$

Solving the above, we have

$$\widehat{Y}_{1H} = \widehat{Y}_{3H} = \frac{Y_1 + Y_3}{2} - \overline{Y}, \quad \widehat{Y}_{2H} = \widehat{Y}_2, \quad \widehat{Y}_{4H} = \widehat{Y}_4.$$

Then the numerator of the F is

$$\frac{\text{SSE}_H - \text{SSE}}{q} = \frac{\|\widehat{Y}_H - \widehat{Y}\|^2}{1} = (\widehat{Y}_{1H} - \widehat{Y}_1)^2 + (\widehat{Y}_{3H} - \widehat{Y}_3)^2 = 2\left(\frac{Y_1 - Y_3}{2}\right)^2 = \frac{(Y_1 - Y_3)^2}{2}$$

- Therefore, the F statistic is

$$F = \frac{(Y_1 - Y_3)^2/2}{(1'Y)^2/4} = \frac{2(Y_1 - Y_3)^2}{(Y_1 + Y_2 + Y_3 + Y_4)^2} \quad \sim \quad F_{1,1}.$$

38 Likelihood Ratio Test

- Let Θ_0 and Θ be the null space and the whole space, respectively, and $\widehat{\theta}_0$ and $\widehat{\theta}$ be MLEs of θ for each space. Then LRT statistic and its asymptotic distribution are

$$-2\ln\Lambda = -2\ln\frac{\max_{\theta\in\Theta_0} f(x|\theta)}{\max_{\theta\in\Theta} f(x|\theta)} = -2\log\frac{L_{H_0}(\widehat{\theta}_0)}{L_{H_A}(\widehat{\theta})} \quad \xrightarrow{D} \quad \chi_r^2(0),$$

where $r = \dim(\Theta) - \dim(\Theta_0)$, i.e., difference in the number of parameters.

- If $y \sim N_n(\beta, \sigma^2)$. Then

$$\widehat{\sigma}_H^2 = \frac{(y - X\widehat{\beta}_0)'(y - X\widehat{\beta}_0)}{n}, \quad \widehat{\sigma}^2 = \frac{(y - X\widehat{\beta})'(y - X\widehat{\beta})}{n},$$

so that

$$L_{H_0}(\widehat{\beta}_H, \widehat{\sigma}_H^2) = (2\pi\widehat{\sigma}_0^2)^{-n/2}e^{-n/2}, \quad L_{H_A}(\widehat{\beta}, \widehat{\sigma}^2) = (2\pi\widehat{\sigma}^2)^{-n/2}e^{-n/2},$$

leading to

$$\Lambda = \frac{L_{H_0}(\widehat{\beta}_0, \widehat{\theta}_0)}{L_{H_A}(\widehat{\beta}, \widehat{\theta})} = \left(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}^2}\right)^{-n/2}, \quad -2\log\Lambda = n(\log\widehat{\sigma}_0^2 - \log\widehat{\sigma}^2).$$

We reject H if $\Lambda < c$. Λ is not a convenient test statistic.

- Instead, using this notation, we have

$$F = \frac{(\text{SSE}_H - \text{SSE})/q}{\text{SSE}/(n-p)} = \frac{n-p}{q}\left(\frac{\text{SSE}_H}{\text{SSE}} - 1\right) = \frac{n-p}{q}\left(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}^2} - 1\right) = \frac{n-p}{q}\left(\Lambda^{-2/n} - 1\right).$$

We reject H when $F > F_{q,n-p,1-\alpha}$.

39 Jensen's inequality

- The direction of the inequality depends on the sign of $f''(X)$. How to remember?
- We know that $\text{Var}(X) = E(X^2) - E(X)^2 \geq 0 \Rightarrow E(X^2) \geq E(X)^2$. So if $f(x) = x^2$, then $E(f(X)) \geq f(E(X))$. This implies that

$$\begin{aligned} f(x) \text{ is a convex function } (f''(x) > 0) \quad &\Rightarrow \quad E(f(X)) \geq f(E(X)) \\ f(x) \text{ is a concave function } (f''(x) < 0) \quad &\Rightarrow \quad E(f(X)) \leq f(E(X)). \end{aligned}$$

If $f(x) = x^{-1}, x > 0$, since $f''(x) = 2x^{-3} > 0$ $(x > 0)$, $E(f(X)) \geq f(E(X)) \Rightarrow E(X^{-1}) \geq (E(X))^{-1}$.

40 Iterative Algorithms

- Consider a model with log-likelihood $\ell(\gamma)$. Want to find $\widehat{\gamma}$, the MLE of γ, by the iterative process.
- Fisher's method of scoring:

$$\begin{aligned}
\gamma^{(m+1)} &= \gamma^{(m)} - \left\{\mathbb{E}\left(\frac{\partial^2 \ell}{\partial\gamma\partial\gamma'}\right)\right\}^{-1}_{\gamma^{(m)}} \left(\frac{\partial \ell}{\partial \gamma}\right)_{\gamma^{(m)}} \\
&= \gamma^{(m)} + \left\{\mathbb{E}\left(\frac{\partial \ell}{\partial\gamma}\frac{\partial \ell}{\partial\gamma'}\right)\right\}^{-1}_{\gamma^{(m)}} \left(\frac{\partial \ell}{\partial \gamma}\right)_{\gamma^{(m)}} \\
&= \gamma^{(m)} + I(\gamma^{(m)})^{-1}\ell'(\gamma^{(m)}),
\end{aligned}$$

  i.e., γ is updated by adding the product of the *inverse* **Fisher** information and the **score** function.
- Newton method:

$$\begin{aligned}
\gamma^{(m+1)} &= \gamma^{(m)} - \left(\frac{\partial^2 \ell}{\partial\gamma\partial\gamma'}\right)^{-1}_{\gamma^{(m)}} \left(\frac{\partial \ell}{\partial \gamma}\right)_{\gamma^{(m)}} \\
&= \gamma^{(m)} - H(\gamma^{(m)})^{-1}\ell'(\gamma^{(m)}),
\end{aligned}$$

  i.e., γ is updated by subtracting the product of the inverse Hessian and the score function.
- Derivation (from 250B HW4): First, find the MLE of $\boldsymbol{\theta}$, say $\widehat{\boldsymbol{\theta}}$, such that

$$\mathbf{u}(\widehat{\boldsymbol{\theta}}) = \frac{\partial \ell(\widehat{\boldsymbol{\theta}}, \mathbf{y})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

  Taylor expansion of $\mathbf{u}(\widehat{\boldsymbol{\theta}})$ around an initial value $\boldsymbol{\theta}_0$ up to the first order gives

$$\mathbf{u}(\widehat{\boldsymbol{\theta}}) \approx \mathbf{u}(\boldsymbol{\theta}_0) + \frac{\partial \mathbf{u}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{u}(\boldsymbol{\theta}_0) + \mathbf{H}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0} \;\Rightarrow\; \widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - \mathbf{H}^{-1}(\boldsymbol{\theta}_0)\mathbf{u}(\boldsymbol{\theta}_0).$$

41 Miscellaneous Exercises

- (Midterm) True or False: For any linear models, it is always true that the sum of residuals equals 0.

  Solution. False; the sum of residuals is $1'e = 1'(I-P)Y = 0$ only if $1'P = 1'$ that is $1_n \in C(P)$.
- (Midterm) True or False. Let S be a $n \times p$ matrix and T be a $n \times q$ matrix and both have full column rank. Let P_S be the orthogonal projection matrix onto $C(S)$ and assume further that columns in S are linearly independent of those in T. Then $T'(I - P_S)T$ is nonsingular.

  Solution: Let $Q_S = I - P_S$. For $a \in \mathbb{R}^q$, suppose $a'T'Q_STa = 0$. Since $a'T'Q_STa = \|Q_S^{1/2}Ta\|^2$,

$$a'T'Q_STa = 0 \;\Leftrightarrow\; \|Q_S^{1/2}Ta\|^2 = 0 \;\Leftrightarrow\; Q_STa = \mathbf{0} \;\Leftrightarrow\; Ta = S(S'S)^{-1}S'Ta \;\Leftrightarrow\; a = \mathbf{0}$$

  as $S \perp\!\!\!\perp T$ implies $C(S) \cap C(T) = \{\mathbf{0}\}$.
- Given the predictor $\widehat{Y} = x\widehat{\beta} \in \mathbb{R}$, where $X = (1, x_1, \ldots, x_{p-1})$. Show that $\widehat{Y}$ has a minimum variance of σ^2/n at the x point $x_j = \overline{x}_{\cdot j}$ $(j = 1, 2, \ldots, p-1)$.

  Solution: We know that $Y_i = x_i'\beta$, where $x_i = (1, x_{i1}, \ldots, x_{i,p-1})$, and $Y_i = \widetilde{x}_i'\beta$, where $\widetilde{x}_i = (1, x_{i1} - \overline{x}_{\cdot 1}, \ldots, x_{i,p-1} - \overline{x}_{\cdot,p-1})$ (after scaling) have the same $\widehat{\beta}_j$, $j = 1, \ldots, p-1$ with a different $\widehat{\beta}_0$. Then

$$\begin{aligned}
&\operatorname{Var}(\widehat{\beta}) = \sigma^2 \begin{pmatrix} n & 0' \\ 0 & C \end{pmatrix}^{-1} = \sigma^2 \begin{pmatrix} 1/n & 0' \\ 0 & C^{-1} \end{pmatrix} \\
\Rightarrow \quad &\operatorname{var}(\widehat{Y}) = \sigma^2 x' \begin{pmatrix} 1/n & 0' \\ 0 & C^{-1} \end{pmatrix} x = \sigma^2\left(\frac{1}{n} + v'C^{-1}v\right) \geq \frac{\sigma^2}{n} \quad \because C \succ O
\end{aligned}$$

  with equality iff $v = 0 \Leftrightarrow x_j = \overline{x}_{\cdot j}$ $(j = 1, 2, \ldots, p-1)$.

- (HW2) Show that $\|x\| = \|y\|$ iff there exists an orthogonal matrix T such that $Tx = y$ using the householder transformation matrix H, which is symmetric and orthogonal.

  Proof: If $Tx = y$, then $y'y = x'T'Tx = x'x \Rightarrow \|x\| = \|y\|$ since L2 norm is always positive.

  If $\|x\| = \|y\|$, then $\|x\|e_1 = \|y\|e_1 \Rightarrow H_1 x = H_2 y \Rightarrow H_2 H_1 x = y$, where $T = H_2 H_1$ is orthogonal.

- (HW4) Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be independent random samples from $N(\mu_1, v_1^2)$ and $N(\mu_2, v_2^2)$, and let S_1^2 and S_2^2 denote the sample variances. Then what is the distribution of

$$\frac{k(X_1 + X_2)}{|Y_1 - Y_2|} \quad \text{and} \quad \frac{k[(X_1 - c)^2 + (X_2 - c)^2]}{S_2^2}.$$

  Solution: First, $X_1 \perp X_2$ and $Y_1 \perp Y_2$ follow

$$\frac{X_1 + X_2}{\sqrt{2}\nu_1} \sim N\left(\frac{\sqrt{2}\mu_1}{\nu_1}, 1\right), \quad \frac{Y_1 - Y_2}{\sqrt{2}\nu_2} \sim N(0,1) \;\Rightarrow\; \frac{(Y_1 - Y_2)^2}{2\nu_2^2} \sim \chi_1^2(0),$$

  respectively. Further, since $(X_1 + X_2) \perp (Y_1 - Y_2)$, we have

$$\frac{(X_1 + X_2)/(\sqrt{2}\nu_1)}{\sqrt{(Y_1 - Y_2)^2/(2\nu_2^2)}} = \frac{\nu_2}{\nu_1} \frac{X_1 + X_2}{|Y_1 - Y_2|} \sim t_1\left(\frac{\sqrt{2}\mu_1}{\nu_1}\right),$$

  which is the given statistic if $k = \nu_2/\nu_1$. Secondly, $X_1 \perp X_2$ leads to

$$\frac{(X_i - c)^2}{\nu_1^2} \sim \chi_1^2\left(\left(\frac{\mu_1 - c}{\nu_1}\right)^2\right) \;\Rightarrow\; \frac{(X_1 - c)^2 + (X_2 - c)^2}{\nu_1^2} \sim \chi_2^2\left(2\left(\frac{\mu_1 - c}{\nu_1}\right)^2\right)$$

  and we know $(n_2 - 1)S_2^2/\nu_2^2 \sim \chi_{n_2-1}^2(0)$. Since $X_i \perp\!\!\!\perp S_2^2$ that is a function of Y_i,

$$\frac{\frac{[(X_1-c)^2+(X_2-c)^2]/\nu_1^2}{2}}{\frac{(n_2-1)S_2^2/\nu_2^2}{n_2-1}} = \frac{\nu_2^2}{2\nu_1^2} \frac{(X_1 - c)^2 + (X_2 - c)^2}{S_2^2} \sim F_{2,n_2-1}\left(2\left(\frac{\mu_1 - c}{\nu_1}\right)^2\right),$$

  which is the given statistic if $k = \nu_2^2/(2\nu_1^2)$.