Linear Models in Statistics

Andrew Brown
(Notes by Shuai Wei)

September 17, 2023

Contents

Chapter 1

Matrix Algebra

1.1 Operation

Let $\vec{\mathbb{1}}_n$ be an $n \times 1$ column vector of 1's. Then $\vec{\mathbb{1}}_n^T \vec{\mathbb{1}}_n = n$ and $\vec{\mathbb{1}}_n \vec{\mathbb{1}}_n^T = J$, where J is an $n \times n$ square matrix of 1's.

Example 1.1. If A is $n \times p$, then sums of lines and columns are, respectively,

$$\vec{\mathbb{1}}_n^T A = \begin{bmatrix} \sum_i^n a_{i1} & \cdots & \sum_i^n a_{ip} \end{bmatrix} \text{ and } A\vec{\mathbb{1}}_p = \begin{bmatrix} \sum_{j=1}^p a_{1j} \\ \vdots \\ \sum_{j=1}^p a_{nj} \end{bmatrix}.$$

Theorem 1.2. *If A is $n \times p$ and B is $p \times m$, then $(AB)^T = B^T A^T$.*

Proof. Let $C = AB$, then $C = (c_{ij}) = (\sum_{k=1}^p a_{ik} b_{kj})$. Then $(AB)^T = C^T = (c_{ij})^T = (c_{ji}) = (\sum_{k=1}^p a_{jk} b_{ki}) = (\sum_{k=1}^p b_{ki} a_{jk}) = B^T A^T$. □

Proposition 1.3. Let A be $n \times m$ and B be $m \times p$ so that $A = \begin{bmatrix} \vec{\alpha}_1^T \\ \vdots \\ \vec{\alpha}_n^T \end{bmatrix}$, $B = (\vec{b}_1, \ldots, \vec{b}_p)$. Then

$$\begin{bmatrix} \vec{\alpha}_1^T B \\ \vdots \\ \vec{\alpha}_n^T B \end{bmatrix} = AB = \begin{bmatrix} A\vec{b}_1 & \ldots & A\vec{b}_p \end{bmatrix}.$$

Remark. $A\vec{b}$ is a linear combination of the columns of A, in which the coefficients are elements of $\vec{b}$. The columns of AB are linear combination of the columns of A. The coefficients for the jth column of AB are the elements of the jth column of B. We have the similar conclusion for the rows.

Theorem 1.4. *Let A be any $n \times p$ matrix. Then*

(a) $A^T A$ is $p \times p$ and its elements are products of the columns of A.

(b) AA^T *is* $n \times n$ *and its elements are products of the rows of* A*.*

(c) Both A^TA *and* AA^T *are symmetric.*

(d) If $A^TA = 0$*, then* $A = 0$*.*

1.2 Quadratic form

Definition 1.5. If A is a symmetric $n \times n$ matrix in $\mathbb{R}$ and $\vec{y}$ is a $n \times 1$ column vector, the product $\vec{x}^T A\vec{x} = \sum_{i=1}^n a_{ii}x_i^2 + \sum_{i \neq j} a_{ij}x_ix_j$ is called a *quadratic form.*

If A is a $n \times p$ matrix in $\mathbb{R}$, and $\vec{x}$ is $n \times 1$ and $\vec{y}$ is $p \times 1$, the product $\vec{x}^T A\vec{y} = \sum_{ij} a_{ij}x_iy_j$ is called a bilinear form.

1.3 Rank

Theorem 1.6. *Columns of* A *are linearly independent if and only if* $\vec{0}$ *is the unique solution of* $A\vec{c} = \vec{0}$*.*

Theorem 1.7. *If there is a non-zero solution for* $A\vec{c} = \vec{0}$*, then at least one of the column vectors* $\vec{a}_i$ *can be expressed as a linear combination of the other column vectors in the set.*

Theorem 1.8. *The maximum possible rank of an* $n \times p$ *matrix* A *is* $\min(n, p)$*. In a non-square matrix, the rows or columns are linearly dependent.*

Example 1.9. The rank of $A = \begin{bmatrix} 1 & -2 & 3 \\ 5 & 2 & 4 \end{bmatrix}$ is 2 since neither row is a multiple of the other. So the columns are linearly dependent. Then there exists not all zero constants c_1, c_2 and c_3 such that $c_1 \begin{bmatrix} 1 \\ 5 \end{bmatrix} + c_2 \begin{bmatrix} -2 \\ 2 \end{bmatrix} + c_3 \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Rewrite it as $\begin{bmatrix} 1 & -2 & 3 \\ 5 & 2 & 4 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ or $A\vec{c} = 0$. The solution to it is given by any multiple of $c = (14, -11, -12)^T$. In this case, the product $A\vec{c}$ is equal to $\vec{0}$ even though $A \neq 0$ and $\vec{c} \neq 0$. This is possible because of the linear dependence of the column vectors of A.

Remark. We can extend it to products of matrices. It is possible to find $A \neq 0$ and $B \neq 0$ such that $AB = 0$. Each linear combination of the columns of A is $\vec{0}$.

We can also exploit the linear dependence of rows or columns of a matrix to create expressions such as $AB = CB$, where $A \neq C$. Thus in a matrix equation, we cannot, in general, cancel a matrix from both sides of the equation.

There are two exceptions to this rule:

(a) If B is a full-rank **square** matrix, then $AB = CB$ implies $A = C$ (multiply by B^{-1} on the right);

(b) The other special case occurs when the expression holds for all possible values of the matrix common to both sides of the equation; for example, if $A\vec{x} = B\vec{x}$ for all possible values of x, then $A = B$. To see this, let $\vec{x} = (1, \ldots, 0)$. Then the first column of A equals the first column of B. Continuing in this fashion, we obtain $A = B$.

Theorem 1.10. *(a)* $\operatorname{rank}(AB) \leqslant \operatorname{rank}(A), \operatorname{rank}(B)$.

(b) If B and C are full-rank square matrices,

$$\operatorname{rank}(AB) = \operatorname{rank}(CA) = \operatorname{rank}(A) = \operatorname{rank}(CAB).$$

(c) $\operatorname{rank}(A) = \operatorname{rank}(A^T) = \operatorname{rank}(A^T A) = \operatorname{rank}(AA^T)$.

Proof. (a) $AB = A(\beta_1, \dots, \beta_k)$ and $AB = (\alpha_1^T, \dots, \alpha_m^T)^T B = (\alpha_1^T B, \dots, \alpha_m^T B)^T$.

(b) By (a), $\operatorname{rank}(A) = \operatorname{rank}(ABB^{-1}) \leqslant \operatorname{rank}(AB) \leqslant \operatorname{rank}(A)$. So $\operatorname{rank}(A) = \operatorname{rank}(AB)$. Also, $\operatorname{rank}(A) = \operatorname{rank}(C^{-1}CA) \leqslant \operatorname{rank}(CA) \leqslant \operatorname{rank}(A)$. So $\operatorname{rank}(A) = \operatorname{rank}(CA)$.

(c) Assume there exists x such that $A^T A\vec{x} = 0$, then $\vec{x}^T A^T A\vec{x} = 0$, i.e., $(A\vec{x})^T A\vec{x} = 0$. So $A\vec{x} = 0$. Thus, $\operatorname{rank}(A) \leqslant \operatorname{rank}(A^T A)$. □

Definition 1.11. The *column space* $C(A) = L(\vec{a}_1, \dots, \vec{a}_n) \subseteq \mathbb{R}^m$, where $A = (\vec{a}_1, \dots, \vec{a}_n) \in \mathbb{R}^{m\times n}$. The *null space* $N(A) = \{\vec{x} : A\vec{x} = 0\} \subseteq \mathbb{R}^n$.

Theorem 1.12. $C(A^T) \perp N(A)$ *and* $C(A) \perp N(A^T)$.

Theorem 1.13 (Fundamental Theorem of Linear Algebra, Part I)**.** *Let* $A \in \mathbb{R}^{m\times n}$ *with* $\operatorname{rank}(A) = r$*, then* $\dim(C(A)) = r$, $\dim(N(A)) = n - r$, $\dim(C(A^T)) = r$ *and* $\dim(N(A^T)) = m - r$.

1.4 Inverse

Theorem 1.14. *If A is nonsigular, then A^T is nonsingular and its inverse can be found as $(A^T)^{-1} = (A^{-1})^T$.*

Theorem 1.15. *If A and B are nonsingular matrices of the same size, then AB is nonsingular and $(AB)^{-1} = B^{-1}A^{-1}$.*

Theorem 1.16. *If A is symmetric and nonsingular and is partitioned as $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ and if $B = A_{22} - A_{21}A_{11}^{-1}A_{12}$, then provided A_{11}^{-1} and B^{-1} exists, the inverse of A is given by*

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}B^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}B^{-1} \\ -B^{-1}A_{21}A_{11}^{-1} & B^{-1} \end{bmatrix}$$

Theorem 1.17. *If a square matrix of the form $B + \vec{c}\vec{c}^T$ is nonsingular, where $\vec{c}$ is a vector and B is a nonsingular matrix, then $(B + \vec{c}\vec{c}^T)^{-1} = B^{-1} - \frac{B^{-1}\vec{c}\vec{c}^T B^{-1}}{1+\vec{c}^T B^{-1}\vec{c}}$.*

1.5 Positive Definite Matrices

Theorem 1.18. *In general, any quadratic form $\vec{y}^T A\vec{y}$ can be expressed as $\vec{y}^T \left(\frac{A+A^T}{2}\right)\vec{y}$, and thus the matrix of a quadratic form can always be choosen to be symmetric and thereby unique.*

Remark. We are usually only interested in symmetric p.d. or p.s.d..

Remark. Ths sums of squares in regression and analysis-of-variance can be expressed in the form $\vec{y}^T A\vec{y}$, where $\vec{y}$ is an observation vector. Such quadratic forms remains positive (or at least nonnegative) for all possible values of $\vec{y}$.

Definition 1.19. If the symmetric matrix A satisfies $\vec{y}^T A\vec{y} > 0$ for all possible $\vec{y}$ except $\vec{y} = 0$, then the quadratic form $\vec{y}^T A\vec{y}$ is said to be *positive definite*, and A is said to be a *positive definite matrix.*

If the symmetric matrix A satisfies $\vec{y}^T A\vec{y} \geqslant 0$ for any $\vec{y}$, then the quadratic form $\vec{y}^T A\vec{y}$ is said to be *positive definite*, and A is said to be a *positive semidefinite matrix.*

Theorem 1.20. *Let P be a nonsingular matrix.*

(a) If A is positive definite, then $P^T AP$ is positive definite.

(b) If A is positive semidefinite, then $P^T AP$ is positive semidefinite.

Proof. (a) $\vec{y}^T(P^T AP)\vec{y} = (P\vec{y})^T A(P\vec{y}) = 0$ if and only if $P\vec{y} = 0$ if and only if $\vec{y} = 0$.

(b) It is similar. □

Corollary 1.21. Let A be a $p \times p$ positive definite matrix and B be a $k \times p$ matrix of rank $k \leqslant p$. Then BAB^T is positive definite. In other cases, BAB^T is positive semidefinite.

Let A be a $p \times p$ positive definite matrix and B be a $p \times k$ matrix of rank $k \leqslant p$. Then $B^T AB$ is positive definite. In other cases, $B^T AB$ is positive semidefinite.

Proof. $\vec{y}^T BAB^T\vec{y} = (B^T\vec{y})^T A(B^T\vec{y}) > 0$ unless $B^T\vec{y} = 0$. Also, $B^T\vec{y} = 0$ if and only if $\vec{y} = 0$. □

Theorem 1.22. *A symmetric matrix A is positive definite if and only if there exists a nonsingular matrix P such that $A = P^T P$.*

Corollary 1.23. A positive definite matrix is nonsingular and all eigenvalues are positive.

Theorem 1.24. *Let B be an $n \times p$ matrix.*

(a) If $\operatorname{rank}(B) = p$, then $B^T B$ is positive definite.

(b) If $\operatorname{rank}(B) < p$, then $B^T B$ is positive semidefinite.

Proof. (a) $\vec{y}^T B^T By = (By)^T(By)$, which is a sum of squares and is thereby positive unless $B\vec{y} = 0$. Also, $B\vec{y} = y_1\vec{b_1} + \cdots + y_p\vec{b_p} = 0$ if and only if $\vec{y} = 0$.

(b) We can find $\vec{y} \neq 0$ such that $B\vec{y} = y_1\vec{b_1} + \cdots + y_p\vec{b_p} = 0$ since the columns of B are linearly dependent. Hence $y^T B^T By \geqslant 0$. □

Remark. We have a similar result for BB^T.

Theorem 1.25. *If A is positive definite, then A^{-1} is positive definite.*

Proof. $A = P^T P$, where P is nonsigular. Then P^{-1} is nonsingular and $A^{-1} = P^{-1}(P^T)^{-1} = P^{-1}(P^{-1})^T$. □

Theorem 1.26. *If A is positive definite, and is partitioned in the form $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, where A_{11} and A_{22} are square, then A_{11} and A_{22} are positive definite.*

Proof. $A_{11} = \begin{bmatrix} I & 0 \end{bmatrix} A \begin{bmatrix} I \\ 0 \end{bmatrix}$, where I is the same size as A_{11}. Since $\operatorname{rank}(I, 0) = \#$ of rows $< \#$ of columns, A_{11} is positive definite by Corollary 1.21. □

1.6 System of Equations

Definition 1.27. If the system of equations $A\vec{x} = \vec{c}$ has one or more solution vectors, it is said to be *consistent*. If the system has no solution, it is said to be *inconsistent*.

Remark. To illustrate the structure of a consistent system of equations $A\vec{x} = \vec{c}$, suppose that A is $p \times p$ of rank $r < p$. Then the rows of A are linearly dependent, and there exists some $\vec{b} \neq \vec{0}$ such that $\vec{b}^T A = b_1\vec{\alpha_1}^T + \cdots + b_p\vec{\alpha_p}^T = \vec{0}^T$. Since multiplication of $A\vec{x} = \vec{c}$ by $\vec{b}^T$ gives $\vec{b}^T A\vec{x} = \vec{b}^T\vec{c}$, i.e., $\vec{0}^T = \vec{b}^T\vec{c}$, i.e., $b_1c_1 + \cdots b_pc_p = \vec{0}$. Hence, in order for $A\vec{x} = \vec{c}$ to be consistent, the same linear relationships, if any, that exist among the rows of A must exist among the elements (rows) of $\vec{c}$. This is formalized by comparing the rank of A with the rank of the *augmented matrix* $(A, \vec{c})$.

Theorem 1.28. *The system of equations $A\vec{x} = \vec{c}$ has at least one solution vector $\vec{x}$ if and only if rank$(A) =$ rank$(A, \vec{c})$.*

Proof. Suppose $\text{rank}(A) = \text{rank}(A, \vec{c})$, so that appending $\vec{c}$ does not change the rank. Then $\vec{c}$ is a linear combination of the columns of A; that is; there exists $\vec{x}$ such that $x_1\vec{a_1} + \cdots + x_p\vec{a_p} = \vec{c}$, which can be written as $A\vec{x} = \vec{c}$. Thus, $\vec{x}$ is a solution.

Conversely, suppose there exists $\vec{x}$ such that $A\vec{x} = \vec{c}$. In general, $\text{rank}(A) \leqslant \text{rank}(A, \vec{c})$. Then $\text{rank}(A, \vec{c}) = \text{rank}(A, A\vec{x}) = \text{rank}[A(I, \vec{x})] \leqslant \text{rank}(A)$. Hence $\text{rank}(A) = \text{rank}(A, \vec{c})$. □

Theorem 1.29. *We care about system of equations because we want to solve $X^TX\vec{b} = X^T\vec{y}$. Since $C(X^T) = C(X^TX)$, we have it has at least one solution.*

1.7 Determinants

Theorem 1.30. *Assume A is a $n \times n$ matrix.*

(a) If A is singular, $|A| = 0$.

(b) If A is nonsingular, $|A| \neq 0$.

(c) If A is positive definite, $|A| > 0$.

(d) $\left|A^T\right| = |A|$.

(e) If A is nonsingular, $\left|A^{-1}\right| = \frac{1}{|A|}$.

(f) $|cA| = c^n|A|$.

Theorem 1.31. *If the square matrix A is partitioned as $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$. If A_{11} and A_{22} are square and nonsingular, then $|A| = |A_{11}|\left|A_{22} - A_{21}A_{11}^{-1}A_{12}\right| = |A_{22}|\left|A_{11} - A_{12}A_{22}^{-1}A_{21}\right|$.*

Corollary 1.32. Suppose $A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}$ or $A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$, where A_{11} and A_{22} are square. Then in either case $|A| = |A_{11}||A_{22}|$.

Theorem 1.33. *If A and B are square and the same size, then $|AB| = |A||B|$.*

Corollary 1.34. If A and B are square and the same size, thenn $|AB| = |BA|$ and $\left|A^2\right| = |A|^2$.

1.8 Inner products, Orthogonal vectors and matrices

Let $\vec{a}, \vec{b} \in \mathbb{R}^n$, then

$$\cos\theta = \frac{\vec{a}^T\vec{a} + \vec{b}^T\vec{b} - (\vec{b}-\vec{a})^T(\vec{b}-\vec{a})}{2\sqrt{(\vec{a}^T\vec{a})(\vec{b}^T\vec{b})}} = \frac{\vec{a}^T\vec{a} + \vec{b}^T\vec{b} - (\vec{b}^T\vec{b} + \vec{a}^T\vec{a} - 2\vec{a}^T\vec{b})}{2\sqrt{(\vec{a}^T\vec{a})(\vec{b}^T\vec{b})}} = \frac{\vec{a}^T\vec{b}}{\sqrt{(\vec{a}^T\vec{a})(\vec{b}^T\vec{b})}} = \frac{\langle a, b\rangle}{\|a\|\|b\|}.$$

When $\theta = 90^\circ$, $\vec{a}^T\vec{b} = 0$. So $\vec{a}$ and $\vec{b}$ are perpendicular.

A vector $\vec{b}$ can be normalized by $\vec{c} = \frac{\vec{b}}{\sqrt{\vec{b}^T\vec{b}}}$. Then $\vec{c}^T\vec{c} = 1$.

Theorem 1.35. *$\vec{x} \perp \vec{y}$ if and only if $\langle \vec{x}, \vec{y}\rangle = 0$.*

Theorem 1.36. *If $\vec{x}_1, \cdots, \vec{x}_k$ are all nonzero and mutually orthogonal, then $\vec{x}_1, \dots, \vec{x}_k$ are linearly independent.*

Theorem 1.37 (Pythogorean Theorem)**.** *Let $\vec{v}_1, \dots, \vec{v}_k$ be mutually orthogonal. Then*

$$\left\|\sum_{i=1}^{k} v_i\right\|^2 = \sum_{i=1}^{k}\|v_i\|^2.$$

Proof.

$$\left\|\sum_{i=1}^{k} \vec{v}_i\right\|^2 = \left\langle \sum_{i=1}^{k}\vec{v}_i, \sum_{i=1}^{k}\vec{v}_i \right\rangle = \sum_{i=1}^{k}\sum_{j=1}^{k}\langle \vec{v}_i, \vec{v}_j\rangle = \sum_{i=1}^{k}\langle \vec{v}_i, \vec{v}_i\rangle = \sum_{i=1}^{k}\|\vec{v}_i\|^2.$$

□

Definition 1.38. A set of $p \times 1$ vectors $\vec{c}_1, \dots, \vec{c}_p$ that are normalized and mutually orthogonal is said to be an *orthonormal set* of vectors. If the $p \times p$ matrix $C = (\vec{c_1}, \dots, \vec{c_p})$ has orthonormal columns, C is called an *orthogonal* matrix. Then $C^TC = I = CC^T$. Thus, an orthogonal matrix C has orthonormal rows as well as orthonormal columns. Moreover, if C is orthogonal, $C^T = C^{-1}$.

Remark. Multiplication of a vector by orthogonal matrix has the effect of rotating axes; that is, if $\vec{x}$ is transformed to $\vec{z} = C\vec{x}$, where C is orthogonal, then the distance from the origin to $\vec{z}$ is the same as the distance to x: $\vec{z}^T\vec{z} = (C\vec{x})^T(C\vec{x}) = \vec{x}^TC^TC\vec{x} = \vec{x}I\vec{x} = \vec{x}^T\vec{x}$. Hence, the transformation from $\vec{x}$ to $\vec{z}$ is a rotation.

Theorem 1.39. *If the $p \times p$ matrix C is orthogonal and if A is any $p \times p$ matrix, then*

(a) $|C| = 1$ or -1.

(b) $\left|C^TAC\right| = |A|$.

(c) $-1 \leqslant c_{ij} \leqslant 1$, where c_{ij} is any element of C.

Proof. It follows from that $CC^T = 1 = C^TC$. □

1.9 Projections

Definition 1.40. The *orthogonal projection* of $\vec{y}$ onto a vector $\vec{x}$ is the vector $\hat{\vec{y}}$ such that $\hat{\vec{y}} = b\vec{x}$, for some $b \in \mathbb{R}$ such that $(\vec{y} - \hat{\vec{y}}) \perp \vec{x}$. This implies $\vec{x}^T b\vec{x} = \hat{\vec{y}}^T \vec{x} = \langle \hat{\vec{y}}, \vec{x} \rangle = \langle \vec{y}, \vec{x} \rangle = \vec{y}^T \vec{x}$. We denote the projection of $\vec{y}$ onto $\vec{x}$ with $p(\vec{y}|\vec{x})$.

We get

$$b = (\vec{x}^T \vec{x})^{-1} \vec{x}^T \vec{y} = \begin{cases} \text{any constant}, & \vec{x} = \vec{0}, \\ \frac{\vec{y}^T \vec{x}}{|\vec{x}|^2}, & \vec{x} \neq 0. \end{cases}$$

1.9.1 Projection onto indicator vectors

Let $V \subseteq \mathbb{R}^n$, $A \subseteq \{1, \dots, n\}$ and define $\mathbb{1}_A =$ indicator vector of A.

Example 1.41. $X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} = (\mathbb{1}_{S_1}, \mathbb{1}_{S_2}, \mathbb{1}_{S_3})$, where $S_1 = \{1, 2\}, S_2 = \{3, 4\}, S_3 = \{5, 6\}$.

Let $S \in \{S_1, S_2, S_3\}$. Consider $p(\vec{y}|\mathbb{1}_S) = b\mathbb{1}_S$. Then $b = \frac{\vec{y}^T \mathbb{1}_S}{\|\mathbb{1}_S\|^2} = \frac{\sum_{i \in S} y_i}{|S|} = \bar{y}_S$. So $p(\vec{y}|\mathbb{1}_S) = \bar{y}_S \mathbb{1}_S$.

Theorem 1.42. *Let $V \leqslant W$ and $W \leqslant \mathbb{R}^n$, then $V^\perp = \{x \in \mathbb{R}^n : x \perp V\}$ is the orthogonal complement of V. $V^\perp \cap W$ is the orthogonal complement w.r.t. W and $\dim(V) + \dim(V^\perp \cap W) = \dim(W)$.*

Theorem 1.43 (Fundemental theorem of Linear Algebra, Part II)**.** *$N(A) = C^\perp(A^T)$, Or: $N(A^T) = C^\perp(A)$.*

Implication: This tells you exactly when $A\vec{x} = \vec{b}$ can be solved! $A\vec{x} = \vec{b}$ is solvable if and only if if $A^T \vec{y} = 0$, then $\vec{b}^T \vec{y} = \vec{0}$ for any $\vec{y}$.

Example 1.44. We are trying to solve $\vec{y} = X\vec{\beta}$, $X \in \mathbb{R}^{m \times n}$ with $m > n$. This is an overdetermined system when $\vec{y} \notin C(X)$.

We want to find $\hat{\vec{y}} \in C(X)$ that minimize $\left\|\vec{y} - \hat{\vec{y}}\right\|^2$, that is, $\min_{\vec{\beta}} \left\{ \left\|\vec{y} - X\vec{\beta}\right\| \right\}$. This is satisfied when $\vec{y} - \hat{\vec{y}} \perp\!\!\!\perp C(X)$. So $\vec{y} - \hat{\vec{y}} \in C^\perp(X)$ if and only if $\vec{y} - \hat{\vec{y}} \in N(X^T)$ if and only if $X^T(\vec{y} - X\hat{\vec{\beta}}) = 0$ if and only if $X^T \vec{y} = X^T X \hat{\vec{\beta}}$.

Definition 1.45. The projection of a vector $\vec{y}$ onto a subspace $V \subseteq \mathbb{R}^n$ is the vector $\hat{\vec{y}} \in V$ such that $(\vec{y} - \hat{\vec{y}}) \perp V$. This condition is equivalent to $(\vec{y} - \hat{\vec{y}})^T \vec{x} = 0$ for any $\vec{x} \in V$ if and only if $\vec{y}^T \vec{x} = \hat{\vec{y}}^T \vec{x}$ for any $\vec{x} \in V$. Also, such $\vec{z}$ is unique.

Theorem 1.46. *If $\vec{x}_1, \dots, \vec{x}_k$ are such that $L(\vec{x}_1, \dots, \vec{x}_k) = V$, then $\vec{z} = p(\vec{y}|V)$ if $\langle \vec{z}, \vec{x}_i \rangle = \langle \vec{y}, \vec{x}_i \rangle$ for $i = 1, \dots, k$.*

Theorem 1.47. *Let $\vec{v}_1, \dots, \vec{v}_k$ be an orthogonal basis for a subspace $V \leqslant R^n$. Then $p(\vec{y}|V) = \sum_{j=1}^k p(\vec{y}|\vec{v_j})$.*

Proof. Let $\vec{y} = \sum_{i=1}^k a_i\vec{v}_i$ for some $a_1,\ldots,a_k \in \mathbb{R}$. Then $\langle \vec{y}, \vec{v}_i\rangle = a_i\langle \vec{v}_i, \vec{v}_i\rangle$, so $a_i = \frac{\langle \vec{y},\vec{v}_i\rangle}{\|v_i\|^2}$ for $i = 1,\ldots,k$. Hence $\vec{y} = \sum_{i=1}^k a_i\vec{v}_i = \sum_{i=1}^k \frac{\langle \vec{y},\vec{v}_i\rangle}{\|v_i\|^2}\vec{v}_i = \sum_{i=1}^k p(\vec{y}|\vec{v}_j)$. □

Example 1.48. Consider Example 1.41 with $V = L(\mathbb{1}_{S_1}, \mathbb{1}_{S_2}, \mathbb{1}_{S_3}) = C(X) \subseteq \mathbb{R}^6$. Then

$$\begin{aligned} P(\vec{y}|V) &= p(\vec{y}|\mathbb{1}_{S_1}) + p(\vec{y}|\mathbb{1}_{S_2}) + p(\vec{y}|\mathbb{1}_{S_3}) = \bar{y}_{S_1}\mathbb{1}_{S_1} + \bar{y}_{S_2}\mathbb{1}_{S_2} + \bar{y}_{S_3}\mathbb{1}_{S_3} \\ &= \left(\frac{y_1+y_2}{2}, \frac{y_1+y_2}{2}, \frac{y_3+y_4}{2}, \frac{y_3+y_4}{2}, \frac{y_5+y_6}{2}, \frac{y_5+y_6}{2}\right). \end{aligned}$$

Theorem 1.49. *If $\vec{v}_1,\ldots,\vec{v}_k$ form an orthonormal basis for V, then $p(\vec{y}|V) = \sum_{i=1}^k \langle \vec{y}, \vec{v}_i\rangle\vec{v}_i$.*

Theorem 1.50. *Every subspace has an orthogonal basis. One method for finding such a basis is Gram-Schmidt process.*

Assume $\vec{x}_1,\ldots,\vec{x}_k$ form a basis for a subspace V. Take

$$\begin{aligned} \vec{v}_1 &= \vec{x}_1, \\ \vec{v}_2 &= \vec{x}_2 - p(\vec{x}_2|\vec{v}_1), \\ \vec{v}_3 &= \vec{x}_3 - p(\vec{x}_3|\vec{v}_1) - p(\vec{x}_3|\vec{v}_2), \\ &\vdots \\ \vec{v}_k &= x_k - \sum_{i=1}^{k-1} p(\vec{x}_k|\vec{v}_i). \end{aligned}$$

Proposition 1.51. Let $V = L(\vec{x}_1,\ldots,\vec{x}_k) = C(X)$, where $X = \begin{bmatrix}\vec{x}_1 & \cdots & \vec{x}_k\end{bmatrix}$. We want to project $\vec{y}$ onto $C(X)$, that is, $\hat{\vec{y}} = p\left(\vec{y}|C(X)\right)$. We need $\hat{\vec{y}} \in C(X)$ and $\langle \vec{y}, \vec{x}_i\rangle = \langle \hat{\vec{y}}, \vec{x}_i\rangle$ for $i = 1,\ldots,k$. But $\hat{\vec{y}} = \beta_1\vec{x}_1 + \cdots + \beta_k\vec{x}_k$ for some $\beta_1,\ldots,\beta_k \in \mathbb{R}$. So we need $\sum_{j=1}^k \beta_j\langle \vec{x}_j, \vec{x}_i\rangle = \langle \vec{y}, \vec{x}_i\rangle$ for $i = 1,\ldots,k$. Note that $X^TX = \begin{bmatrix}\vec{x}_1^T \\ \vdots \\ \vec{x}_k^T\end{bmatrix}\begin{bmatrix}\vec{x}_1 & \ldots & \vec{x}_k\end{bmatrix} = \begin{bmatrix}\vec{x}_1^T\vec{x}_1 & \cdots & \vec{x}_1^T\vec{x}_k \\ \vdots & \vdots & \vdots \\ \vec{x}_k^T\vec{x}_1 & \cdots & \vec{x}_k^T\vec{x}_k\end{bmatrix}$ and $X^T\vec{y} = \begin{bmatrix}\vec{x}_1^T \\ \vdots \\ \vec{x}_k^T\end{bmatrix}\vec{y} = \begin{bmatrix}\vec{x}_1^Ty \\ \vdots \\ \vec{x}_k^Ty\end{bmatrix}$. With $\vec{\beta} = (\beta_1,\ldots,\beta_k)^T$, the requirement to satisfy is $X^TX\vec{\beta} = X^T\vec{y}$.

Let $X \in \mathbb{R}^{n\times k}$ with $n > k$. Since $\operatorname{rank}(X^TX) = \operatorname{rank}(X)$, $X^TX \in \mathbb{R}^{k\times k}$ is nonsingular if and only if $\operatorname{rank}(X) = k$. Then $\hat{\vec{\beta}} = (X^TX)^{-1}X^T\vec{y}$. So we can write $\hat{\vec{y}} = X\hat{\vec{\beta}} = X(X^TX)^{-1}X^T\vec{y} = P\vec{y}$, where P is called the *projection matrix* onto $C(X)$, P is also called the "*hat matrix*" denoted H. Also, $p\left(\vec{y}|C(X)\right) = \arg\min_{\vec{\beta}}\left\|\vec{y} - X\vec{\beta}\right\|^2$, which is the least square criterion.

Definition 1.52. A symmetrix matrix P is said to be an *(orthogonal) projection matrix* onto V if $P\vec{v} = \vec{v}$ for any $\vec{v} \in V$ and $P\vec{w} = 0$ for any $\vec{w} \in V^\perp$.

Remark. A $n \times n$ matrix P is said to be an *projection matrix* onto $\operatorname{Im}(P) =: V$ if $P^2 = P$. Then $\mathbb{R}^n = \operatorname{Im}(P) \oplus \operatorname{Ker}(P) = V \oplus V^\perp$. If P is orthogonal, then $PP^T = I$, so $P^T = IP^T = PPP^T = P$.

Theorem 1.53. *Let $P = X(X^TX)^{-1}X^T$, then*

(a) P is an orthogonal projection matrix onto X.

(b) $C(X) = C(P)$.

(c) P *is symmetric and idempotent.*

Theorem 1.54. *A matrix* P *is a projection onto* $C(P)$ *if and only if* P *is symmetric and idempotent.*

Theorem 1.55. *Projection matrices onto* V *are unique.*

We have seen that the projection onto $C(X)$ can be obtained with $X(X^TX)^{-1}X^T$. We can also find the projection starting from any orthogonal basis for $C(X)$.

Theorem 1.56. *Let* $q_1, \ldots, q_k$ *be an orthonormal basis for* $V \subseteq \mathbb{R}^n$. *Let* $Q = (q_1, \ldots, q_k)$. *Then* $QQ^T = \sum_{i=1}^k q_i q_i^T$ *is the projection matrix onto* V.

Example 1.57. Assume $V = L(\mathbb{1}_n)$. Let $Q = \frac{1}{\sqrt{n}}\mathbb{1}_n$. Then $P_V = QQ^T = \frac{1}{\sqrt{n}}\mathbb{1}_n\left(\frac{1}{\sqrt{n}}\mathbb{1}_n^T\right) = \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$. So $P_V\vec{x} = \begin{bmatrix} \frac{\sum_{i=1}^n x_i}{n} \\ \vdots \\ \frac{\sum_{i=1}^n x_i}{n} \end{bmatrix} = \bar{x}\mathbb{1}_n$. In this case verify: $P_V\vec{x} = \alpha\mathbb{1}_n$, where $\alpha = \frac{\langle x, \mathbb{1}_n \rangle}{\|\mathbb{1}_n\|^2} = \bar{x}$.

Theorem 1.58. *Let* $V \subseteq \mathbb{R}^n$ *be a vector space and let* V_0 *be a subspace of* V. *Let* P, P_0 *be the corresponding projection matrices. Then* $PP_0 = P_0$ *and* $P_0P = P_0$.

Theorem 1.59 (Orthogonal completement projection)**.** *Let* P *and* P_0 *be projection matrices with* $C(P_0) \subseteq C(P)$. *Then*

(a) $P - P_0$ *is the projection matrix (onto* $C(P - P_0)$*).*

(b) $C(P - P_0) = C^{\perp}(P_0) \cap C(P)$.

Example 1.60. Let $V = L(\mathbb{1}_n)$. Then $P_V = \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$. So $P_{V^\perp} = I - P_V = I - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$. Then $P_{V^\perp}\vec{x} = \left(I - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\right)\vec{x} = \vec{x} - \bar{x}\mathbb{1}_n$. Also, $\sum_{i=1}^n (x_i - \bar{x})^2 = (P_{V^\perp}\vec{x})^T P_{V^\perp}\vec{x}$.

1.10 Direct Sum

Definition 1.61. Subspaces $V_1, \ldots, V_k \subseteq \mathbb{R}^n$ are linearly independent ($\perp\!\!\!\perp$) if $\sum_{i=1}^k \vec{x}_i = \vec{0}$ with $x_i \in V_i$ for $i = 1, \ldots, k$, then $\vec{x}_i = \vec{0}$ for $i = 1, \ldots, k$.

Theorem 1.62. $V_1 \perp\!\!\!\perp V_2$ *if and only if* $V_1 \cap V_2 = \{\vec{0}\}$.

Definition 1.63. Let $V_1, \ldots, V_k \subseteq \mathbb{R}^n$ be subspaces. Let

$$V = \left\{ x : x = \sum_{i=1}^k x_i, x_i \in V_i, \forall i = 1, \ldots, k \right\},$$

then $V = V_1 + \cdots + V_k$, and if $V_1, \ldots, V_k$ are linearly independent, $V = V_1 \oplus \cdots \oplus V_k$.

Example 1.64. Let $V \subseteq \mathbb{R}^n$, since $V \cap V^\perp = \{\vec{0}\}$, we have $V \perp\!\!\!\perp V^\perp$ and then $\mathbb{R}^n = V \oplus V^\perp$.

Theorem 1.65. *Let $V_1, \dots, V_k \subseteq \mathbb{R}^n$ be subspaces and $x \in V = V_1 + \cdots + V_k$. The representation $x = \sum_{i=1}^k x_i$, $x_i \in V_i$ is unique if and only if $V_1, \dots, V_k$ are linearly independent.*

Theorem 1.66. *Let $\{v_{i1}, \dots, v_{in_i}\}$ be a basis for V_i for $i = 1, \dots, k$ and $V_1, \dots, V_k$ be linearly independent. Then $\{v_{11}, \dots, v_{1n_1}, \dots, v_{k1}, \dots, v_{kn_k}\}$ is a basis for $V = V_1 \oplus \cdots \oplus V_k$. So $\dim(V) = \sum_{i=1}^k \dim(V_i)$.*

Theorem 1.67 (Orthogonal decomposition)**.** *For any subspace $V \subseteq \mathbb{R}^n$ and any $\vec{x} \in \mathbb{R}^n$, there exist unique $\vec{x}_1, \vec{x}_2$ such that $\vec{x} = \vec{x}_1 + \vec{x}_2$, where $\vec{x}_1 \in V$ and $\vec{x}_2 \in V^\perp$. More specifically, $\vec{x}_1 = p(\vec{x}|V)$ and $\vec{x}_2 = p(\vec{x}|V^\perp)$.*

Theorem 1.68. *If $V = V_1 \oplus V_2$, then $V^\perp = V_1^\perp \cap V_2^\perp$.*

Theorem 1.69. *If $V = V_1 \oplus V_2$, then $P_V = P_{V_1} + P_{V_2}$ if and only if $V_1 \perp V_2$.*

Corollary 1.70. If $V = V_1 \oplus V_2$ and $V_1 \perp V_2$, then $P_{V_1} = P_V - P_{V_2}$.

Corollary 1.71. Let $V_0 \leqslant V \leqslant \mathbb{R}^n$, and $V = V_0 \oplus V_1$, where $V_1 = V \cap V_0^\perp$, since $V_0 \perp V_1$, $P_{V \cap V_0^\perp} = P_V - P_{V_0}$.

Theorem 1.72. *Let $\mathbb{R}^n \supseteq V = V_1 \oplus \cdots \oplus V_k$, where $V_1, \dots, V_k$ are mutually orthogonal. Then for $\vec{x} \in \mathbb{R}^n$, $p(\vec{x}|V) = \sum_{i=1}^k p(\vec{x}|V_i)$.*

Example 1.73.

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} + \vec{e}.$$

Note that $\mathbb{R}^6 = C(X) \oplus C^\perp(X)$ and X is not column full rank. Since $L(\vec{X}_1)$ and $L(\vec{X}_2, \vec{X}_3, \vec{X}_4)$ are not linearly independent,

$$C(X) = L(\vec{X}_1) + L(\vec{X}_2, \vec{X}_3, \vec{X}_4) = L(\vec{X}_1) \oplus \left[L(\vec{X}_2, \vec{X}_3, \vec{X}_4) \cap L^\perp(\vec{X}_1) \right],$$

where

$$L(\vec{X}_2, \vec{X}_3, \vec{X}_4) \cap L^\perp(\vec{X}_1) = \left\{ \begin{bmatrix} b \\ b \\ c \\ c \\ -b-c \\ -b-c \end{bmatrix}, b, c \in \mathbb{R} \right\}.$$

So

$$\begin{aligned} \hat{\vec{y}} = p\left(\vec{y}|C(X)\right) &= p\left(\vec{y}|L(\vec{X}_1)\right) + p\left(\vec{y}|L(\vec{X}_2, \vec{X}_3, \vec{X}_4) \cap L^\perp(\vec{X}_1)\right) \\ &= \overline{y}\mathbb{1}_n + \left(P_{L(\vec{X}_2, \vec{X}_3, \vec{X}_4)} - P_{L(\vec{X}_1)}\right)\vec{y} \\ &= \overline{y}\mathbb{1}_n + \left(P_{L(\vec{X}_2)} + P_{L(\vec{X}_3)} + P_{L(\vec{X}_4)}\right)\vec{y} - \overline{y}\vec{X}_1 \\ &= \overline{y}\mathbb{1}_n + \overline{y_1}\vec{X}_2 + \overline{y_2}\vec{X}_3 + \overline{y_3}\vec{X}_4 - \overline{y}(\vec{X}_2 + \vec{X}_3 + \vec{X}_4) \\ &= \overline{y}\mathbb{1}_n + (\overline{y_1} - \overline{y})\vec{X}_2 + (\overline{y_2} - \overline{y})\vec{X}_3 + (\overline{y_3} - \overline{y})\vec{X}_4. \end{aligned}$$

Also, $\mathbb{R}^6 = L(\vec{X}_1) \oplus \Big[L(\vec{X}_2, \vec{X}_3, \vec{X}_4) \cap L^{\perp}(\vec{X}_1)\Big] \oplus C^{\perp}(X)$, $I = P_{L(\vec{X}_1)} + P_{L(\vec{X}_2,\vec{X}_3,\vec{X}_4)\cap L^{\perp}(\vec{X}_1)} + P_{C^{\perp}(X)}$, and multiply by $\vec{y}$ on both sides, we have $\vec{y} = \hat{\vec{y_1}} + \hat{\vec{y_2}} + \vec{e}_1$.

1.11 Trace

Theorem 1.74. *(a) If A is $n \times p$ and B is $p \times n$, then* $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

(b) If A is $n \times p$, then $\mathrm{tr}(A^T A) = \sum_{i=1}^{p} \vec{a_i}^T \vec{a_i}$*, where $\vec{a_i}$ is the i-th column of A.*

(c) If A is $n \times p$, then $\mathrm{tr}(AA^T) = \sum_{i=1}^{p} \vec{a_i}^T \vec{a_i}$*, where $\vec{a_i}^T$ is the i-th row of A.*

(d) If $A = (a_{ij})$ is $n \times p$, then $\mathrm{tr}(A^T A) = \mathrm{tr}(AA^T) = \sum_{i=1}^{n} \sum_{j=1}^{p} a_{ij}^2$.

(e) If A is any $n \times n$ matrix and P is any $n \times n$ nonsingular matrix, then $\mathrm{tr}(P^{-1}AP) = \mathrm{tr}(A)$.

(f) If A is any $n \times n$ matrix and C is any $n \times n$ orthogonal matrix, then $\mathrm{tr}(C^{-1}AC) = \mathrm{tr}(A)$.

Proof. (e) $\mathrm{tr}(P^{-1}AP) = \mathrm{tr}(APP^{-1}) = \mathrm{tr}(A)$. □

1.12 Eigenvalues and Eigenvectors

Definition 1.75. For every square matrix A, a scalar λ and a nonzero vector $\vec{x}$ can be found such that $A\vec{x} = \lambda\vec{x}$. Note that, the vector $\vec{x}$ is transformed by A onto a multiple of itself, so that the point $A\vec{x}$ is on the line passing through $\vec{x}$ and the origin.

Definition 1.76. The *eigenspace* of A associated with eigenvalue λ is $N(A - \lambda I)$. Note the eigenspace is a vector space.

Theorem 1.77. *If* $\det(A) = 0$*, then $A\vec{v} = \vec{0}$ for some $\vec{v} \neq \vec{0}$, so $\lambda = 0$ is an eigenvalue of A.*

1.12.1 Functions of a Matrix

If λ is an eignenvalue of A with corresponding eigenvector x, then for certain functions $g(A)$, an eigenvalue is given by $g(\lambda)$ and $\vec{x}$ is the corresponding eigenvector of $g(A)$ as well as of A.

(a) If λ is an eigenvalue of A, then $c\lambda$ is an eigenvalue of cA, where c is an arbitrary constant such that $c \neq 0$. This is because $cA\vec{x} = c\lambda\vec{x}$. So $\vec{x}$ is also an eigenvector of cA corresponding to $c\lambda$.

(b) If λ is an eigenvalue of A and $\vec{x}$ is the corresponding eigenvector of A, then $c\lambda + k$ is an eigenvalue of the matrix $cA + kI$ and $\vec{x}$ is an eigenvector of $cA + kI$, where c and k are scalars.

To see this, we add $k\vec{x}$ to the above equation, $cA\vec{x} + k\vec{x} = c\lambda\vec{x} + k\vec{x}$. Then $(cA + kI)\vec{x} = (c\lambda + k)\vec{x}$.

(c) If λ is an eigenvalue of A, then λ^2 is an eigenvalue of A^2 since $A(A\vec{x}) = A(\lambda\vec{x})$, and then $A^2\vec{x} = \lambda A\vec{x} = \lambda(\lambda\vec{x}) = \lambda^2\vec{x}$. This can be extended to any power of A: $A^k\vec{x} = \lambda^k\vec{x}$.

(d) If λ is an eigenvalue of the **nonsingular** matrix A, then $1/\lambda$ is an eigenvalue of A^{-1} since $A^{-1}A\vec{x} = A^{-1}\lambda\vec{x}$, i.e., $\vec{x} = \lambda A^{-1}\vec{x}$, i.e., $A^{-1}\vec{x} = 1/\lambda\vec{x}$.

(e) If λ is an eigenvalue of A, then $(A^3 + 4A^2 - 3A + 5I)\vec{x} = A^3\vec{x} + 4A^2\vec{x} - 3A\vec{x} + 5\vec{x} = \lambda^3\vec{x} + 4\lambda^2\vec{x} - 3\lambda\vec{x} + 5\vec{x} = (\lambda^3 + 4\lambda^2 - 3\lambda + 5)\vec{x}$.

Theorem 1.78. *If λ is an eigenvalue of A, then $1 - \lambda$ is an eigenvalue of $I - A$. If $I - A$ is nonsingular, then $1/(1 - \lambda)$ is an eigenvalue of $(I - A)^{-1}$.*

Proof. It is sufficient to show when all the eigenvalues of A satisfy $-1 < \lambda_i < 1$, then $(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$. In the symmetric case, the spectral decomposition, $I - A = \mu^T\mu - \mu^T\Lambda\mu = \mu^T(I - \Lambda)\mu$. Then

$$\begin{aligned}(I - A)^{-1} &= \mu^T(I - \Lambda)^{-1}\mu = \mu^T\text{diag}\left(\frac{1}{1-\lambda_1}, \dots, \frac{1}{1-\lambda_n}\right)\mu \\ &= \mu^T\text{diag}\left(\sum_{k=0}^{\infty}\lambda_1^k, \dots, \sum_{k=0}^{\infty}\lambda_n^k\right)\mu = \mu^T\sum_{k=0}^{\infty}\Lambda^k\mu = \sum_{k=0}^{\infty}\mu^T\Lambda^k\mu = \sum_{k=0}^{\infty}A^k.\end{aligned}$$

In general, note that if $-1 < \lambda_i < 1$, then $\sum_{k=0}^{\infty} A^k < \infty$. So $(I - A)\sum_{k=0}^{\infty} A^k = \sum_{k=0}^{\infty} A^k - \sum_{k=0}^{\infty} A^{k+1} = A^0 = I$. Thus, $(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$. □

1.12.2 Symmetric matrices

Theorem 1.79. *Let A be an $n \times n$ symmetric matrix.*

(a) The eigenvalues $\lambda_1, \dots, \lambda_n$ of A are real.

(b) The eigenvectors $\vec{x}_1, \dots, \vec{x}_k$ of A corresponding to distinct eigenvalues $\lambda_1, \dots, \lambda_k$ are mutually orthogonal. The eigenvectors $\vec{x}_{k+1}, \dots, \vec{x}_n$ corresponding to the nondistinct eigenvalues can be choose to be mutually orthogonal to each other and to the other eigenvectors.

(c) If the eigenvectors are normalized and placed as columns of a matrix C, then C is an orthogonal matrix.

Definition 1.80. The number of times an eigenvalue λ in the characteristic poplynomial is the *algebraic multiplicity* of λ.

The number of linearly independent e-vectors associated with an e-val λ is the *geometric multiplicity* of λ.

Theorem 1.81. *In general, gemometric multiplicity is less than algeraic multiplicity.*

If geometric multiplicity is equal to algeraic multiplicity, then A is said to be diagonalizable.

Theorem 1.82 (Spectral Theorem)**.** *If A is an $n \times n$ symmetric matrix with eigenvalues $\lambda_1, \dots, \lambda_n$ and normalized eigenvectors $\vec{x}_1, \dots, \vec{x}_n$, then A can be expressed as $A = CDC^T = \sum_{i=1}^{n} \lambda_i\vec{x}_i\vec{x}_i^T$, which is called the* spectral decomposition *of A, where $D = \text{diag}(\lambda_1, \dots, \lambda_n)$ and C is the orthogonal matrix $C = (\vec{x}_1, \dots, \vec{x}_n)$.*

Proof. $A = AI = ACC^T = A(\vec{x}_1, \dots, \vec{x}_n)C^T = (A\vec{x}_1, \dots, A\vec{x}_n)C^T = (\lambda_1\vec{x}_1, \dots, \lambda_n\vec{x}_n)C^T = CDC^T$. □

Remark. The eigenvalues and eigenvectors basiccally summerize the "information" contained in a matrix A.

Since $\vec{x}_1, \dots, \vec{x}_n$ are orthonormal basis of A, $\vec{x}_i\vec{x}_i^T = P_i$ is the projection onto the 1-dimensional subspace $L(\vec{x}_i)$ for $i = 1, \dots, n$. Then $A = \sum_{i=1}^{n} \lambda_i P_i$.

Corollary 1.83. $C^TAC = D$, i.e., C diagonalizes A.

Theorem 1.84. *With the spectral decomposition, $\vec{x}^TA\vec{x} = \vec{x}^TCDC^T\vec{x} = \vec{y}^TD\vec{y}$, where $\vec{y} = C^T\vec{x}$. Let $y_i = (C^T\vec{x})_i$, then $\vec{y}^TD\vec{y} = \lambda_1\vec{y}_1^2 + \cdots + \lambda_n\vec{y}_n^2$. Note $\vec{y} = C^T\vec{x}$ rotates the axes of the space to align with the eigenvectors.*

Theorem 1.85. *For A symmetric and positive definite, we have $\vec{x}^TA\vec{x} = 1$ defines an ellipsoid in $\mathbb{R}^n$.*

1.12.3 Eigensystem of projection matrices

Theorem 1.86. *Consider the projection matrix P_V (onto V). Then*

(a) Every $\vec{0} \neq \vec{x} \in V$ is an eigenvector of P_V with eigenvalue 1.

(b) Every $\vec{0} \neq \vec{x} \in V^{\perp}$ is an eigenvector of P_V with eigenvalue 0.

(c) $\lambda = 1$ has multiplicity $\dim(V)$.

(d) $\lambda = 0$ has multiplicity $\dim(V^{\perp}) = n - \dim(V)$.

Thus, $\operatorname{tr}(P_V) = \sum_{i=1}^n \lambda_i = \dim(V)$ *and* $\operatorname{rank}(P_V) = \operatorname{tr}(P_V) = \sum_{i=1}^n \lambda_i$.

1.12.4 Positive Definite and Semidefinite Matrices

Theorem 1.87. *Let A be $n \times n$ with eigenvalues $\lambda_1, \ldots, \lambda_n$.*

(a) If A is positive definite, then $\lambda_i > 0$ for $i = 1, \ldots, n$.

(b) If A is positive semidefinite, then $\lambda_i \geqslant 0$ for $i = 1, \ldots, n$. The number of eigenvalues λ_i for which $\lambda_i > 0$ is the rank of A.

Proof. (a) For $i = 1, \ldots, n$, let $x_i \neq 0$ and $\vec{x_i}^TA\vec{x_i} = \lambda_i\vec{x_i}^T\vec{x_i}$, we have $\lambda_i = \frac{\vec{x_i}^TA\vec{x_i}}{\vec{x_i}^T\vec{x_i}} > 0$. □

Theorem 1.88. *If a matrix A is positive definite, we can find a* square root matrix $A^{1/2}$ *as follows.*

Since the eigenvalues of A are positive and $A = CDC^T = \left(CD^{1/2}C^T\right)\left(CD^{1/2}C^T\right)$, we have $A^{1/2} = CD^{1/2}C^T$, where $D^{1/2} = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$. The matrix $A^{1/2}$ is symmetric and has the property $A^{1/2}A^{1/2} = (A^{1/2})^2 = A$. Note $A^{1/2} = CD^{1/2}C^T$ is the unique symmetric square root matrix.

Remark. For positive semidefinite or positive definite matrix A, there exists a upper triangular B such that $A = B^TB$, such a factorization is called the Cholesky decomposition. $A = LU = LL^T = LDL^T = \left(LD^{1/2}\right)\left(LD^{1/2}\right)^T$. If A is positive definite, then the decomposition is unique.

1.12.5 Idempotent matrices

Definition 1.89. A square matrix A is said to be *idempotent* if $A^2 = A$. Many of the sums of squares in regression and analysis of variance can be expressed as quadratic forms $\vec{y}^TA\vec{y}$. The idempotence of A or of a product involving A will be used to establish that $\vec{y}^TA\vec{y}$ (or a multiple of $\vec{y}^TA\vec{y}$) has a chi-square distribution.

Theorem 1.90. *The only nonsingular idempotent matrix is the identity matrix I.*

Proof. If A is idenpotent and nonsingular, then $A^2 = A$, and the inverse A^{-1} exists. Then $A^{-1}A^2 = A^{-1}A$ and so $A = I$. □

Theorem 1.91. *If A is singular, symmetric, and idempotent, then A is positive semidefinite.*

Proof. Since $A = A^T$ and $A = A^2$, we have $A = A^2 = AA = A^TA$, which is positive semidefinite since $\text{rank}(A^T) < \#$ of columns. □

Theorem 1.92. *If A is an $n \times n$ symmetric idempotent matrix of rank r, then A has r eigenvalues equal to 1 and $n - r$ eigenvalues equal to 0.*

Proof. Since $A^2 = A$, $A^2\vec{x} = A\vec{x} = \lambda\vec{x}$. Also, $A^2\vec{x} = \lambda^2\vec{x}$. So $\lambda\vec{x} = \lambda^2\vec{x}$, i.e., $(\lambda - \lambda^2)\vec{x} = \vec{0}$. Since $\vec{x} \neq 0$, $\lambda - \lambda^2 = 0$. (Or: $A^k = (CDC^T) \times \cdots \times (CDC^T) = CD^kC^T = CDC^T = A$ if and only if λ_i is 0 or 1.) The number of nonzero eigenvalues is equal to $\text{rank}(A)$ according to the spectral decomposition. □

Corollary 1.93. If A is symmetric and idempotent of rank r, then $\text{rank}(A) = \text{tr}(A) = r$.

Theorem 1.94. *If A is an $n \times n$ idempotent matrix, P is an $n \times n$ nonsingular matrix, and C is an $n \times n$ orthogonal matrix, then*

(a) $I - A$ is idempotent.

(b) $A(I - A) = 0$ and $(I - A)A = 0$.

(c) $P^{-1}AP$ is idempotent.

(d) C^TAC is idempotent.

1.12.6 Vector and matrix calculus

Theorem 1.95. *Let $u = \vec{a}^T\vec{x} = \vec{x}^T\vec{a}$, where $\vec{a}^T = (a_1, \ldots, a_p)$ is a vector of constants. Then $\frac{\partial u}{\partial \vec{x}} = \vec{a}$.*

Proof. Since $\frac{\partial u}{\partial x_i} = \frac{\partial(a_1x_1 + \cdots + a_px_p)}{\partial x_i} = a_i$, and $\frac{\partial u}{\partial \vec{x}} = (\frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_p})^T = (a_1, \ldots, a_p)^T = \vec{a}$. □

Theorem 1.96. *Let $u = \vec{x}^TA\vec{x}$, where A is a symmetric matrix of constants. Then $\frac{\partial u}{\partial \vec{x}} = 2A\vec{x}$.*

1.13 Maximization or Minimization of a Function of a Vector

Consider a function $u = f(\vec{x})$ of the p variables in $\vec{x}$. Occasionally the situation requires the maximization or minimization of the function u, subject to q constraints on $\vec{x}$. We denote the constraints as $h_1(\vec{x}) = \vec{0}, \ldots, h_q(\vec{x}) = \vec{0}$, or $\vec{h}(\vec{x}) = \vec{0}$. We denote a vector of q unknown constants (the *Lagrange multipliers*) by $\vec{\lambda}$, and let $\vec{y}^T = (\vec{x}^T, \vec{\lambda}^T)$. We then let $v = u + \vec{\lambda}^T\vec{h}(\vec{x})$. The maximum of minimum of u subject to $\vec{h}(\vec{x}) = 0$ is obtained by solving the equations $\frac{\partial v}{\partial \vec{y}} = \vec{0}$.

Or, equivalently $\frac{\partial u}{\partial \vec{x}} + \frac{\partial \vec{h}}{\partial \vec{x}}\vec{\lambda} = \vec{0}$ and $\vec{h}(\vec{x}) = \vec{0}$, where $\frac{\partial \vec{h}}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \cdots \frac{\partial h_q}{\partial x_1} \\ \vdots & \vdots \\ \frac{\partial h_1}{\partial x_p} & \cdots \frac{\partial h_q}{\partial x_p} \end{bmatrix}$.

1.14 Generalized Inverses

Definition 1.97. A generalized inverse of an $n \times k$ matrix X is any $k \times n$ matrix X^- such that $XX^-X = X$.

Remark. (a) A generalized inverse always exists.

(b) A generalized inverse, in general, is not unique.

(c) If X is nonsingular, then $X^- = X^{-1}$.

Example 1.98. If $X = (1, 2, 3, 4)^T$, then the generalized inverses are $X_1^- = (1, 0, 0, 0)$, $X_2^- = (0, 1/2, 0, 0)$, $X_3^- = (0, 0, 1/3, 0)$ and $X_4^- = (0, 0, 0, 1/4)$.

Theorem 1.99. *Let $X \in \mathbb{R}^{n\times k}$ with $\operatorname{rank}(X) = r$. Then*

(a) $\operatorname{rank}(X^-X) = \operatorname{rank}(XX^-) = \operatorname{rank}(X) = r$.

(b) $(X^-)^T$ is a generalized inverse of X^T.

(c) $X = X(X^TX)^-X^TX$.

Proof. (a) $\operatorname{rank}(X^-X) \leqslant \min(\operatorname{rank}(X^-), \operatorname{rank}) \leqslant \operatorname{rank}(X) = r$. Since $X = XX^{-1}X$, $\operatorname{rank}(X) \leqslant \min(\operatorname{rank}(X), \operatorname{rank}(X^-X)) \leqslant \operatorname{rank}(X^-X)$.

(b) $XX^-X = X$ implies $X^T(X^-)^TX^T = X^T$.

(c) Let $v \in \mathbb{R}^n$ and $v = v_1 + v_2$, where $v_1 \in C(X)$ and $v_2 \in C^\perp(X)$. Then $v_1 = Xb$ for some $b \in \mathbb{R}^n$, and $v^TX(X^TX)^-X^TX = v_1^TX(X^TX)^-X^TX = b^TX^TX(X^TX)^-X^TX = b^TX^TX = v^TX$. Since it is true for any v, we have $X = X(X^TX)^-X^TX$. □

Theorem 1.100. *$X(X^TX)^-X^T$ is the projection matrix onto $C(X)$.*

Proof. Let $v \in C\left(X(X^TX)^{-1}X^T\right) = C(X)$. Then there exists $b \in \mathbb{R}^n$ such that $v = Xb = X(X^TX)^-X^TXb = X(X^TX)^-X^Tb^*$. Then $v \in C\left(X(X^TX)^-X^T\right)$. So $C(X) \subseteq C\left(X(X^TX)^-X^T\right)$. Clearly, $C\left(X(X^TX)^-X^T\right) \subseteq C(X)$. Thus, $C(X) = C\left(X(X^TX)^-X^T\right)$. □

1.14.1 How we do find a generalized inverse?

Theorem 1.101. *Let $A \in \mathbb{R}^{n\times k}$ with* rank$=r$ *and A is partitioned as $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, where $A_{11} \in \mathbb{R}^{r\times r}$ with* rank r*. Then a generalized inverse is $A^- = \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}$.*

Corollary 1.102. Let $A \in \mathbb{R}^{n\times k}$ with *rank*$= r$ and A is partitioned as $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, where $A_{22} \in \mathbb{R}^{r\times r}$ with *rank* r. Then a generalized inverse is $A^- = \begin{bmatrix} 0 & 0 \\ 0 & A_{22}^{-1} \end{bmatrix}$.

Theorem 1.103. *In general, for $A \in \mathbb{R}^{n\times k}$ with* rank r*. To find a generalized inverse,*

(a) Find any non-singular $r \times r$ sub-matrix of A, say C (guranted to exist).

(b) Find $C^{-T} = \left(C^{-1}\right)^T = \left(C^T\right)^{-1}$.

(c) Replace elements of C w/ elements of C^{-T}.

(d) Relace the rest of A with zero.

(e) Transpose the resulting matrix.

Theorem 1.104. *Generalized inverses of symmetric matrices are not guranteed to be symmetric. But a symmetric generalized inverse always exists. We will always assume the generalized inverse is symmetric.*

Theorem 1.105. *If the system of equations $A\vec{x} = \vec{c}$ is consistent, then $x^* = A^- c$ is a solution.*

Proof. Since $A\vec{x} = \vec{c}$ and $AA^- A\vec{x} = A\vec{x}$, we have $AA^-\vec{c} = \vec{c}$. So $A\vec{x}^* = \vec{c}$. □

Theorem 1.106. *If $A\vec{x} = \vec{x}$ is consistent, then all possible solutions can be obtained via either of the following*

(a) Use a possible A^-, $\vec{x}^* = A^-\vec{c}$. $\vec{x}^{**} = A^-\vec{c} + (I - A^- A)\vec{h}$ for all possible h since $A(I - A^- A)\vec{h} = 0$.

(b) Use all possible A^- in $\vec{x}^* = A^-\vec{c}$.

Theorem 1.107. *The system of equations $A\vec{x} = \vec{c}$ has a solution if and only if for any generalized inverse A^-, it is true that $AA^-\vec{c} = \vec{c}$.*

1.15 Examples

Example 1.108. Prove $\text{rank}\left(P_{C(X)}\right) = \text{rank}(X)$.

Proof. By commutativity of the trace operator,

$$\text{rank}\left(P_{C(X)}\right) = \text{rank}\left(X\left(X^T X\right)^{-1} X^T\right) = \text{rank}\left(\left(X^T X\right)^{-1} X^T X\right) = \text{rank}\left(X^T X\right) = \text{rank}\left(X\right).$$

□

Chapter 2

Random Vectors and Matrices

Definition 2.1. A *random vector* or *random matrix* is a vector or matrix whose elements are random variables. Informally, a *random variable* is defined as a variable whose value depends on the outcome of a chance experiment.

Remark. In terms of experimental structure, we can distinguish two kinds of random vectors:

(a) A vector containing a measurement on each of n different individuals or experimental units. In this case, where the same variable is observed on each of n units selected at random, the n random variables $y_1, \ldots, y_n$ in the vector are typically uncorrelated and have the same variance. Consider the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i,\ i = 1, \ldots, n.$$

If we treat the x variables as constants, in which case, we have two random vectors:

$$\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \text{ and } \vec{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The y_i values are observable, but the ϵ_i's are not observable unless the $\beta_k's$ are known.

(b) A vector consisting of p different measurements on one individual or experimental unit. The p random variables thus obtained are typically correlated and have different variances. Consider regression of y on several random x variables. For the i^{th} individual in the sample, we observe the $k+1$ random variables $y_i, x_{i1}, \cdots, x_{ik}$ which constitute the random vector

$$(y_i, x_{i1}, \ldots, x_{ik})^T.$$

In some cases, the $k+1$ variables $y_i, x_{i1}, \ldots, x_{ik}$ are all measured using the same units or scale of measurement, but typically the scales differ.

2.1 Mean vector

The expected value of a $p \times 1$ random vector $\vec{y}$ is defined as the vector of expected values of the p random variables $y_1, \dots, y_p$ in $\vec{y}$:

$$E(\vec{y}) = E \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} E[y_1] \\ \vdots \\ E[y_p] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} = \vec{\mu}.$$

2.2 Covariance matrix for random vectors

Let Z be a $n \times p$ random matrix. Then expected value of Z

$$E[Z] = E \begin{bmatrix} z_{11} & \cdots & z_{1p} \\ \vdots & & \vdots \\ z_{n1} & \cdots & z_{np} \end{bmatrix} = \begin{bmatrix} E[z_{11}] & \cdots & E[z_{1p}] \\ \vdots & & \vdots \\ E[z_{n1}] & \cdots & E[z_{np}] \end{bmatrix}.$$

Note that

$$\Sigma = E[(\vec{y} - \vec{\mu})(\vec{y} - \vec{\mu})^T] = E[\vec{y}\vec{y}^T] - \vec{\mu}\vec{\mu}^T,$$

where

$$(\vec{y} - \vec{\mu})(\vec{y} - \vec{\mu})^T$$

is a random matrix, whose $(ij)^{\text{th}}$ element is $(y_i - \mu_i)(y_j - \mu_j)$.

Definition 2.2. For random vector $\vec{x} \in \mathbb{R}^k, \vec{y} \in \mathbb{R}^n$, let $\operatorname{Cov}(x_i, y_i) = \sigma_{ij}$. Then *(population) covariance matrix* of $\vec{x}$ and $\vec{y}$ is

$$\mathbb{R}^{k \times n} \ni \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \vdots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kn} \end{bmatrix} = \operatorname{Cov}(\vec{x}, \vec{y}) := \Sigma_{x,y}.$$

Theorem 2.3.

$$\operatorname{Cov}(\vec{x}, \vec{y}) = E\left[(\vec{x} - \vec{\mu}_x)(\vec{y} - \vec{\mu}_y)^T\right].$$

2.2.1 Correlation matrices

Theorem 2.4. *Define*

$$V_{\vec{x}} = \operatorname{diag}\left(\operatorname{Var}(\vec{x}_1), \dots, \operatorname{Var}(\vec{x}_k)\right).$$

$$\operatorname{Corr}(\vec{x}_i, \vec{x}_j) = \frac{\operatorname{Cov}(\vec{x}_i, \vec{x}_j)}{\sqrt{\operatorname{Var}(\vec{x}_i)\operatorname{Var}(\vec{x}_j)}}.$$

Then

$$\operatorname{Var}(\vec{x}) = V_{\vec{x}}^{1/2} \operatorname{Corr}(\vec{x}) V_{\vec{x}}^{1/2}.$$

$$\operatorname{Corr}(\vec{x}) = V_{\vec{x}}^{-1/2} \operatorname{Var}(\vec{x}) V_{\vec{x}}^{-1/2}.$$

$$\operatorname{Var}(\vec{x}, \vec{y}) = V_{\vec{x}}^{1/2} \operatorname{Corr}(\vec{x}, \vec{y}) V_{\vec{y}}^{1/2}.$$

$$\operatorname{Corr}(\vec{x}, \vec{y}) = V_{\vec{x}}^{-1/2} \operatorname{Var}(\vec{x}, \vec{y}) V_{\vec{y}}^{-1/2}.$$

2.2.2 Generalized Variance

A measure of overall variability in the population of $\vec{y}$ variables can be defined as the determinant of Σ:

$$\text{Generalized variance} = |\Sigma|.$$

If $|\Sigma|$ is small, the $\vec{y}$ variables are concentrated closer to μ than if $|\Sigma|$ is large. A smaller value of $|\Sigma|$ may also indicate that the variables $y_1, \dots, y_p$ are highly intercorrelated, in which case the $\vec{y}$ variables tend to occupy a subspace of the p dimensions.

2.3 Partitioned random vectors

Let

$$\vec{v} = \begin{bmatrix} \vec{y} \\ \vec{x} \end{bmatrix}$$

Then

$$\Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}$$

is a rectangular matrix.

2.4 Linear functions of random vectors

We often use lienar combinations of the variables $y_1, \dots, y_p$ from a random vector $\vec{y}$. Let

$$\vec{a} = (a_1, \dots, a_p)^T$$

be a vector of constants. Then the linear combination

$$z = a_1 y_1 + \cdots + a_p y_p = \vec{a}^T \vec{y}$$

is a random variable. Assume the mean of $\vec{y}$ is $\vec{\mu}$. Then the mean of z is

$$\mu_z = E[\vec{a}^T \vec{y}] = \vec{a}^T E(\vec{y}) = \vec{a}^T \vec{\mu}.$$

Suppose that we have several linear combinations of $\vec{y}$ with constant coefficients:

$$\begin{aligned} z_1 &= a_{11} y_1 + \cdots + a_{1p} y_p = \vec{a}_1^T \vec{y} \\ z_2 &= a_{21} y_1 + \cdots + a_{2p} y_p = \vec{a}_2^T \vec{y} \\ &\vdots \\ z_k &= a_{k1} y_1 + \cdots + a_{kp} y_p = \vec{a}_k \vec{y} \end{aligned}$$

These k linear functions can be written in the form

$$\vec{z} = A\vec{y}.$$

We often need A is full rank. Since $\vec{y}$ is a random vector, each z_i is a random variable and then $\vec{z}$ is a random vector.

Theorem 2.5. *Suppose $\vec{y}$ is a random vector, X is a random matrix, $\vec{a}$ and $\vec{b}$ are vectors of constants, and A and B are matrices of constants. Then*

(a) $E[A\vec{y}] = AE[\vec{y}]$.

(b) $E[\vec{a}^T X\vec{b}] = \vec{a}^T E[X]\vec{b}$.

(c) $E[AXB] = AE[X]B$.

Proof. (a) Assume A is $n \times p$ matrix.

$$\begin{aligned} E[A\vec{y}] &= E[(\vec{a}_1, \dots, \vec{a}_p)\vec{y}] \\ &= E[\vec{a}_1 y_1 + \cdots + \vec{a}_p y_p] \\ &= \vec{a}_1 E[y_1] + \cdots + \vec{a}_p E[y_p] \\ &= (\vec{a}_1, \dots, \vec{a}_p)(E[y_1], \dots, E[y_p])^T \\ &= AE[\vec{y}]. \end{aligned}$$

(b) Assume X is $n \times k$ matrix.

$$\begin{aligned} E[\vec{a}^T X\vec{b}] &= \vec{a}^T E[X\vec{b}] \\ &= \vec{a}^T E[(\vec{x}_1, \dots, \vec{x}_k)\vec{b}] \\ &= \vec{a}^T \left(E[\vec{x}_1]b_1 + \cdots + E[\vec{x}_k^T]b_k\right) \\ &= \vec{a}^T (E[\vec{x}_1], \dots, E[\vec{x}_k])(b_1, \dots, b_k)^T \\ &= \vec{a}^T E[X]\vec{b}. \end{aligned}$$

□

Theorem 2.6. *If $\vec{a}$ is a $p \times 1$ vector of constants and $\vec{y}$ is a $p \times 1$ random vector with covariance matrix Σ, then the variance*

$$\mathrm{Var}(\vec{a}^T \vec{y}) = \vec{a}^T \Sigma \vec{a}.$$

Proof.

$$\begin{aligned} \mathrm{Var}(\vec{a}^T \vec{y}) &= E\left[(\vec{a}^T\vec{y} - \vec{a}^T\vec{\mu})^2\right] = E\left[(\vec{a}^T)^2(\vec{y} - \vec{\mu})^2\right] \\ &= E\left[\vec{a}^T(\vec{y} - \vec{\mu})\vec{a}^T(\vec{y} - \vec{\mu})\right] = E\left[\vec{a}^T(\vec{y} - \vec{\mu})(\vec{y} - \vec{\mu})^T\vec{a}\right] \\ &= \vec{a}^T E\left[(\vec{y} - \vec{\mu})(\vec{y} - \vec{\mu})^T\right]\vec{a} = \vec{a}^T\Sigma\vec{a}. \end{aligned}$$

□

Corollary 2.7. If $\vec{a}$ and $\vec{b}$ are $p \times 1$ vectors of constants, then

$$\mathrm{Cov}(\vec{a}^T\vec{y}, \vec{b}^T\vec{y}) = \vec{a}^T\Sigma\vec{b}.$$

Theorem 2.8. *Let $\vec{z} = A\vec{y}$ and $\vec{w} = B\vec{y}$, where A is a $k \times p$ matrix of constants, B is an $m \times p$ matrix of constants, and $\vec{y}$ is a $p \times 1$ random vector with covariance matrix Σ. Assume Σ is positive definite.*

(a)

$$\mathrm{Cov}(\vec{z}) = \mathrm{Cov}(A\vec{y}) = A\Sigma A^T,$$

If rank$(A) = k \leqslant p$, $A\Sigma A^T$ is positive definite. In other cases, $A\Sigma A^T$ is positive semidefinite.

(b)

$$\operatorname{Cov}(\vec{z}, \vec{w}) = \operatorname{Cov}(A\vec{y}, B\vec{y}) = A\Sigma B^T.$$

Theorem 2.9. *Let $\vec{y}$ be a $p \times 1$ random vector and $\vec{x}$ be a $q \times 1$ random vector such that* $\operatorname{Cov}(\vec{y}, \vec{x}) = \Sigma_{yx}$. *Let A be a $k \times p$ matrix of constants and B be an $h \times q$ matrix of constants. Then*

$$\operatorname{Cov}(A\vec{y}, B\vec{x}) = A\Sigma_{yx}B^T.$$

Proof. Let $\vec{v} = \begin{bmatrix} \vec{y} \\ \vec{x} \end{bmatrix}$ and $C = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$. Then

$$\operatorname{Cov}\begin{bmatrix} A\vec{y} \\ B\vec{x} \end{bmatrix} = \operatorname{Cov}(C\vec{v}) = C\operatorname{Cov}(\vec{y}, \vec{x})C^T = C\begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{yx} & \Sigma_{xx} \end{bmatrix} C^T = \begin{bmatrix} A\Sigma_{yy}A^T & A\Sigma_{yx}B^T \\ B\Sigma_{xy}A^T & B\Sigma_{xx}B^T \end{bmatrix}.$$

So $\operatorname{Cov}(A\vec{y}, B\vec{x}) = A\Sigma_{yx}B^T$. One have that $\operatorname{Cov}(B\vec{x}, A\vec{y}) = B\Sigma_{xy}A^T$. □

Theorem 2.10 (Expectation of bilinear form)**.** *Let* $E[\vec{x}] = \vec{\mu}_x$, $E[\vec{y}] = \vec{\mu}_y$, $\operatorname{Cov}(\vec{x}, \vec{y}) = \Sigma_{\vec{x}\vec{y}}$. *Then*

$$E[\vec{x}^T A\vec{y}] = \operatorname{tr}(A\Sigma_{\vec{x}\vec{y}}^T) + \vec{\mu}_x A\vec{\mu}_y.$$

Chapter 3

Multivariate Normal Distribution

In order to make inferences, we often assume that the random vector of interest has a multivariate normal distribution.

3.1 Multivariate normal density function

We begin with independent standard normal random variable $z_1, \dots, z_p$, with $\mu_i = 0$ and $\sigma_i^2 = 1$ for any $i \in [p]$ and $\sigma_{ij} = 0$ for $i \neq j$. Let

$$\vec{z} = (z_1, \dots, z_p)^T,$$

where

$$E[\vec{z}] = \vec{0} \text{ and } \mathrm{Cov}(\vec{z}) = I,$$

and

$$z_i \sim N(0,1), \forall i \in [p].$$

We wish to transform $\vec{z}$ to a multivariate normal random vector

$$\vec{y} = (y_1, \dots, y_p)^T$$

with $E(\vec{y}) = \vec{\mu}$ and $\mathrm{Cov}(\vec{y}) = \Sigma$, where μ is any $p \times 1$ vector and Σ is any $p \times p$ positive definite matrix. The (joint) pdf of $\vec{z}$

$$\begin{aligned} f(\vec{z}) &= f(z_1, \dots, z_p) \\ &= g_1(z_1) \cdots g_p(z_p) \\ &= \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} \cdots \frac{1}{\sqrt{2\pi}} e^{-z_p^2/2} \\ &= \frac{1}{\left(\sqrt{2\pi}\right)^p} e^{-\sum_{i=1}^p z_i^2/2} \\ &= \frac{1}{\left(\sqrt{2\pi}\right)^p} e^{-z^T z/2} = \frac{1}{\left(\sqrt{2\pi}\right)^p} e^{-\|z\|_2^2/2}. \end{aligned}$$

We say

$$\vec{z} \sim N_p(\vec{0}, I),$$

where p is the dimension of the distribution and corresponds to the number of variables in $\vec{y}$. To transform $\vec{z}$ to $\vec{y}$ with arbitrary mean vector $E[\vec{y}] = \mu$ and arbitrary (positive definite) covariance matrix $\text{Cov}(\vec{y}) = \Sigma$, we define the transformation

$$\vec{y} = \Sigma^{1/2}\vec{z} + \vec{\mu},$$

where $\Sigma^{1/2}$ is the (symmetric) square root matrix. Note that

$$E[\vec{y}] = E[\Sigma^{1/2}\vec{z} + \vec{\mu}] = \Sigma^{1/2}E[\vec{z}] + \vec{\mu} = \vec{\mu}.$$

$$\text{Cov}(\vec{y}) = \text{Cov}(\Sigma^{1/2}\vec{z} + \vec{\mu}) = \Sigma^{1/2}\text{Cov}(\vec{z})\left(\Sigma^{1/2}\right)^T = \Sigma^{1/2}I\Sigma^{1/2} = \Sigma.$$

Note the analogy to

$$y = \sigma z + \mu.$$

The analogous expression for $\vec{y} = \Sigma^{1/2}\vec{z} + \vec{\mu}$ is

$$f(\vec{y}) = g(\vec{z})\underbrace{\left|\left|\Sigma^{-1/2}\right|\right|}_{\text{abs. det.}} = g(\vec{z})\left|\left|\Sigma^{1/2}\right|^{-1}\right|,$$

which parallels the absolute value expression $|dz/dy| = |1/\sigma|$ in the univariate case. The determinant

$$\left|\Sigma^{-1/2}\right|$$

is the Jacobian of the transformation. Since $\Sigma^{-1/2}$ is positive definite, we can dispense with the absolute value and then

$$f(\vec{y}) = g(\vec{z})\left|\Sigma^{-1/2}\right| = g(\vec{z})|\Sigma|^{-1/2}.$$

In order to express $\vec{z}$ in terms of $\vec{y}$, we need to obtain

$$\vec{z} = \Sigma^{-1/2}(\vec{y} - \vec{\mu}).$$

Then

$$\begin{aligned}
f(\vec{y}) &= g(\vec{z})|\Sigma|^{-1/2} \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^p |\Sigma|^{1/2}} e^{-\vec{z}^T\vec{z}/2} \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^p |\Sigma|^{1/2}} e^{-\left[\Sigma^{-1/2}\vec{(}\vec{y}-\vec{\mu})\right]^T\left[\Sigma^{-1/2}(\vec{y}-\vec{\mu})\right]/2} \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^p |\Sigma|^{1/2}} e^{-(\vec{y}-\vec{\mu})^T\Sigma^{-1}(\vec{y}-\vec{\mu})/2},
\end{aligned}$$

which is the multivariate normal density function with mean vector $\vec{\mu}$ and covariance matrix Σ. We say

$$\vec{y} \sim N_p(\vec{\mu}, \Sigma),$$

where the subscript p is the dimension of the p-variate normal distribution and indicates the number of variables, that is, $\vec{y}$ is $p \times 1$ and $\vec{\mu}$ is $p \times 1$ and Σ is $p \times p$. A comparison shows the standard distance

$$(\vec{y} - \vec{\mu})^T \Sigma^{-1} (\vec{y} - \vec{\mu})$$

in place of

$$\frac{(y - \mu)^2}{\sigma^2}$$

in the exponent and the square root of the generlized variance $|\Sigma|$ in place of σ^2 in the denominator. A small value of $|\Sigma|$ indicates that the $\vec{y}$'s are concentrated closer to $\vec{\mu}$ than is the case when Σ is large. A small value of $|\Sigma|$ may also indicate a high degree of multicollinearity among the variables. High *multicollinearity* indicates that the variables are highly intercorrelated, in which case the $\vec{y}$'s tend to occupy a subspace of the p dimensions.

Example 3.1. If $\vec{y}_n \sim N(\vec{\mu}_n, \Sigma_{n \times n})$ with $\Sigma_{n \times n}$ p.s.d., then

$$\vec{y}_n \overset{d}{=} A_{n \times p} \vec{z} + \vec{\mu}_n,$$

where $AA^T = \Sigma$ (e.g., Cholesky, then $n = p$), $\vec{z} = (z_1, \dots, z_p)$, and $z_i \overset{iid}{\sim} N(0, 1)$. If $\text{rank}(A) = n$, then Σ is positive definite. Then $|A| = \Sigma^{-1/2}$. If $\text{rank}(A) \neq n$, then $\vec{y}_n$ is still said to have a multivariate normal distribution, but its density DNE. It is an overdetermined distribution (degenerate case).

Theorem 3.2. *Let $\vec{\mu} \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ be symmetric p.s.d.. Then there exists a multivariate normal distribution with mean $\vec{\mu}$ and* Var-Cov *matrix Σ.*

Proof. For example, by Choleskys, there exists B such that $\Sigma = BB^T$. In this case, B is symmetric. Take $\vec{z} \sim N(\vec{0}, I)$, let

$$\vec{x} = B\vec{z} + \vec{\mu}.$$

Then $\vec{x} \sim N(\vec{\mu}, \Sigma)$. □

3.2 Moment generating functions

Definition 3.3. The MGF for a univariate variable y is defined as

$$M_y(t) = E[e^{ty}]$$

provided $E[e^{ty}]$ exists for every real number t in the neighorhood $-h < t < h$ for some positive number h. For the univariate normal $N(\mu, \sigma^2)$, the MGF of y is given by

$$M_y(t) = e^{\mu t + t^2 \sigma^2 / 2}.$$

Definition 3.4. For a random vector $\vec{y}$, the MGF is defined as

$$M_{\vec{y}}(\vec{t}) = E[e^{t_1 y_1 + \cdots + t_p y_p}] = E[e^{\vec{t}^T \vec{y}}].$$

By analogy, we have

$$\frac{\partial M_{\vec{y}}(\vec{0})}{\partial \vec{t}} = E[\vec{y}],$$

where $\frac{\partial M_{\vec{y}}(\vec{0})}{\partial \vec{t}}$ indicates that $\frac{\partial M_{\vec{y}}(\vec{t})}{\partial \vec{t}}$ is evaluated at $\vec{t} = \vec{0}$. Similarly,

$$\frac{\partial^2 M_{\vec{y}}(\vec{0})}{\partial t_r \partial t_s} = E[y_r y_s].$$

Theorem 3.5. *If $\vec{y}$ is distributed as $N_p(\vec{\mu}, \Sigma)$, its MGF is given by*

$$M_{\vec{y}}(\vec{t}) = e^{\vec{\mu}^T \vec{t} + \vec{t}^T \Sigma \vec{t}/2}.$$

Example 3.6. Let $\vec{z} \sim N(\vec{0}, I)$, then

$$M_{\vec{z}}(t) = E\left[e^{\vec{t}^T \vec{z}}\right] = E\left[e^{\sum_{i=1}^n t_i z_i}\right] = \prod_{i=1}^n E\left[e^{t_i z_i}\right] = \prod_{i=1}^n m_{z_i}(t_i) = \prod_{i=1}^n e^{\frac{t_i^2}{2}} = e^{\frac{1}{2}\vec{t}^T \vec{t}}.$$

Let $\vec{x} = A\vec{z} + \vec{\mu}$. Then

$$\begin{aligned} M_{\vec{x}}(\vec{t}) &= E\left[\exp\left(\vec{t}^T (A\vec{z} + \vec{\mu})\right)\right] = e^{\vec{t}^T \vec{\mu}} E\left[\exp\left((A^T \vec{t})^T \vec{z}\right)\right] \\ &= e^{\vec{t}^T \vec{\mu}} m_{\vec{z}}(A^T \vec{t}) = e^{\vec{t}^T \vec{\mu}} e^{\frac{1}{2} t^T A A^T t}. \end{aligned}$$

So $\vec{z} \sim N(\vec{\mu}, AA^T)$.

Corollary 3.7. If $\vec{y}$ is distributed as $N_p(\vec{\mu}, \Sigma)$, the MGF for $\vec{y} - \vec{\mu}$ is

$$M_{\vec{y}-\vec{\mu}}(\vec{t}) = e^{\vec{t}^T \Sigma \vec{t}/2}.$$

Theorem 3.8. *(a) If two random vectors have the same MGF, they have the same density.*

(b) Two random vectors are independent if and only if their MGFs factors into the product of their two seperate MGFs, that is, if $\vec{y}^T = (\vec{y_1}^T, \vec{y_1}^T)$ and $t^T = (\vec{t}_1^T, \vec{t}_2^T)$, then $\vec{y_1}$ and $\vec{y_2}$ are independnent if and only if

$$M_{\vec{y}}(\vec{t}) = M_{\vec{y}_1}(\vec{t}_1) M_{\vec{y}_2}(\vec{t}_2)$$

3.3 Properties of the multivariate normal distribution

Theorem 3.9. *Let $\vec{y} \sim N_p(\vec{\mu}, \Sigma)$. Let $\vec{a}$ be any $p \times 1$ vector of constants, and let A be any $k \times p$ matrix of constants with rank $k \leqslant p$. Then*

(a)

$$z = \vec{a}^T \vec{y} \sim N(\vec{a}^T \vec{\mu}, \vec{a}^T \Sigma \vec{a}).$$

Proof.

$$M_z(t) = E[e^{tz}] = E[e^{t\vec{a}^T \vec{y}}] = E[e^{(t\vec{a})^T \vec{y}}] = M_{\vec{y}}(t\vec{a}) = e^{(t\vec{a})^T \vec{\mu} + (t\vec{a})^T \Sigma (t\vec{a})/2} = e^{(\vec{a}^T \mu)t + (\vec{a}^T \Sigma \vec{a}) t^2/2}. \quad \square$$

(b)

$$A\vec{y} \sim N_k(A\vec{\mu}, A\Sigma A^T).$$

Proof.

$$M_{\vec{z}}(\vec{t}) = E[e^{\vec{t}^T \vec{z}}] = E[e^{\vec{t}^T A\vec{y}}] = e^{\vec{t}^T(A\vec{\mu})+\vec{t}^T(A\Sigma A^T)\vec{t}/2}. \qquad \square$$

Theorem 3.10. *If $\vec{y} \sim N_p(\vec{\mu}, \Sigma)$, then any $r \times 1$ subvector of $\vec{y}$ has an r-variate normal distribution with the same means, variance, and covariances as in the original p-variate normal distribution.*

Proof. Without loss of generality, let $\vec{y}^T = (\vec{y}_1^T, \vec{y}_2^T)$, where $\vec{y}_1$ is the $r \times 1$ subvector of interest. Let $\vec{\mu}$ and Σ be partitioned accordingly:

$$\mu = \begin{bmatrix} \vec{\mu_1} \\ \vec{\mu_2} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Define $A = (I_r, O)$. Then $A\vec{y} = \vec{y}_1$. Since $A\vec{\mu} = \vec{\mu}_1$ and $A\Sigma A^T = \Sigma_{11}$, $\vec{y}_1 \sim N_r(\vec{\mu}_1, \Sigma_{11})$. □

Theorem 3.11. *If $\vec{y}$ is $p \times 1$ and $\vec{x}$ is $q \times 1$ and*

$$\vec{v} = \begin{bmatrix} \vec{y} \\ \vec{x} \end{bmatrix} \sim N_{p+q}(\vec{\mu}, \Sigma),$$

then $\vec{y}$ and $\vec{x}$ are independent if $\Sigma_{\vec{y}\vec{x}} = 0$.

Corollary 3.12. If $\vec{y} \sim N_p(\vec{\mu}, \Sigma)$ and $\text{Cov}(A\vec{y}, B\vec{y}) = A\Sigma B^T = \mathbf{0}$, then $A\vec{y}$ and $B\vec{y}$ are independent.

Lemma 3.13. Let $\vec{y} \sim N(\vec{\mu}, \Sigma)$,

$$\vec{y} = \begin{bmatrix} \vec{y_1} \\ \vec{y_2} \end{bmatrix}, \quad \vec{\mu} = \begin{bmatrix} \vec{\mu_1} \\ \vec{\mu_2} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where $\Sigma_{21} = \Sigma_{12}^T$. Let $\vec{y}_{2|1} = \vec{y}_2 - \Sigma_{21}\Sigma_{11}^{-1}\vec{y}_1$. Then $\vec{y}_{2|1} \perp\!\!\!\perp \vec{y}_1$ with $\vec{y}_1 \sim N_p(\vec{\mu}, \Sigma_{11})$ and $\vec{y}_{2|1} \sim N_{n-p}(\vec{\mu}_{2|1}, \Sigma_{22|1})$, where

$$\vec{\mu}_{2|1} = \vec{\mu}_2 - \Sigma_{21}\Sigma_{11}^{-1}\vec{\mu}_1,$$
$$\Sigma_{22|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

Proof. Let

$$\vec{y_1} = C_1\vec{y}, \quad C_1 = \begin{bmatrix} I & 0 \end{bmatrix};$$
$$\vec{y_2} = C_2\vec{y}, \quad C_2 = \begin{bmatrix} -\Sigma_{21}\Sigma_{11}^{-1} & I \end{bmatrix}.$$

$$\text{Cov}\left(\vec{y}_1, \vec{y}_{2|1}\right) = \text{Cov}\left(C_1\vec{y}, C_2\vec{y}\right) = C_1\Sigma C_2^T = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \end{bmatrix} \begin{bmatrix} -\Sigma_{11}^{-1}\Sigma_{12} \\ I \end{bmatrix} = \mathbf{0}. \qquad \square$$

Theorem 3.14. *Let $\vec{y} \sim N(\vec{\mu}, \Sigma)$,*

$$\vec{y} = \begin{bmatrix} \vec{y_1} \\ \vec{y_2} \end{bmatrix} \quad \vec{\mu} = \begin{bmatrix} \vec{\mu_1} \\ \vec{\mu_2} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where $\Sigma_{21} = \Sigma_{12}^T$. Then

$$\vec{y}_2|\vec{y}_1 \sim N_{n-p}\left(\vec{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\vec{y}_1 - \vec{\mu}_1), \Sigma_{22|1}\right).$$

Proof. Since $\vec{y}_{2|1} \perp\!\!\!\perp \vec{y}_1$, we have

$$\vec{y}_{2|1}|\vec{y}_1 \overset{d}{=} \vec{y}_{2|1} \sim N_{n-p}\left(\vec{\mu}_{2|1}, \Sigma_{22|1}\right).$$

But $\vec{y}_2 = \vec{y}_{2|1} + \Sigma_{21}\Sigma_{11}^{-1}\vec{y}_1$, and $\Sigma_{21}\Sigma_{11}^{-1}\vec{y}_1|\vec{y}_1 = \Sigma_{21}\Sigma_{11}^{-1}\vec{y}_1$. Then $\vec{y}_2|\vec{y}_1 \overset{d}{=} \vec{y}_{2|1} + \Sigma_{21}\Sigma_{11}^{-1}\vec{y}_1$. So,

$$\vec{y}_2|\vec{y}_1 \sim N_{n-p}(\vec{\mu}^*, \Sigma_{22|1}),$$

where

$$\vec{\mu}^* = \vec{\mu}_{2|1} + \Sigma_{21}\Sigma_{11}^{-1}\vec{y}_1 = \vec{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\vec{y}_1 - \vec{\mu}_1).$$

Thus,

$$\vec{y}_2|\vec{y}_1 \sim N_{n-p}\left(\vec{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\vec{y}_1 - \vec{\mu}_1)), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right). \qquad \square$$

3.4 Partial/Multiple Correlation

Example 3.15. Educational psychologist studies the relationship between height y_1 and reading ability y_2 of children based on scores of a standard test. For 200 children in grades 3,4 and 5, he measured y_1 and y_2, finding a sample correlation 0.56. Is there a linear association between height and reading ability? Yes, but only because we are ignoring one (or more) "lurking variable". Under children with more years of schooling tend to taller. The partial correlation coefficient is a measure of linear relationship between two variables with the linear affect of one or more variables removed.

Theorem 3.16. *Let $\vec{v} \sim N_{p+q}(\vec{\mu}, \Sigma)$,*

$$\vec{v} = \begin{bmatrix}\vec{x}\\ \vec{y}\end{bmatrix}, \quad \Sigma = \begin{bmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix},$$

where $\Sigma_{yx} = \Sigma_{xy}^T$, $\vec{x} = (v_1, \dots, v_p)^T$ and $\vec{y} = (v_{p+1}, \dots, v_{p+q})^T$. Then

$$\text{Var}(\vec{y}|\vec{x}) = \Sigma_{\vec{y}\vec{y}} - \Sigma_{\vec{y}\vec{x}}\Sigma_{\vec{x}\vec{x}}^{-1}\Sigma_{\vec{x}\vec{y}} := \Sigma_{\vec{y}|\vec{x}}.$$

Definition 3.17. Let $\sigma_{ij|1,\dots,p} = (\Sigma_{\vec{y}|\vec{x}})_{ij}$. The *partial correlation coefficient* of y_i and y_j given $\vec{x}$ is given by

$$\rho_{ij|1,\dots,p} = \frac{\sigma_{ij|1,\dots,p}}{(\sigma_{ii|1,\dots,p})^{1/2}(\sigma_{jj|1,\dots,p})^{1/2}}.$$

Like ordinary correlation, it is still true that

$$-1 \leqslant \rho_{ij|1,\dots,p} \leqslant 1.$$

Example 3.18. If v_1 = age, v_2 = height, v_3 = reading ability and set $\vec{x} = v_1$, $\vec{y} = (v_2, v_3)^T$. Then $\rho_{23|1}$ is the partial correlation of height and reading ability after removing the linear effect of age. We expect $\rho_{23|1} \approx 0$.

The multiple correlation coefficient measures the linear association between one variable and a group of others. Let $\vec{v} \sim N_{p+q}(\vec{\mu}, \Sigma)$,

$$\vec{v} = \begin{bmatrix}\vec{x}\\ y\end{bmatrix}, \quad \Sigma = \begin{bmatrix}\Sigma_{xx} & \sigma_{\vec{x}y}\\ \sigma_{y\vec{x}} & \sigma_{yy}\end{bmatrix}.$$

Then

$$E[\vec{y}|\vec{x}] = \mu_y + \sigma_{y\vec{x}}\Sigma_{\vec{x}\vec{x}}^{-1}(\vec{x} - \vec{\mu}_x) = \mu_{y|\vec{x}}.$$

Definition 3.19. The *square multiple correlation coefficient* between y and $\vec{x}$ is defined as

$$\vec{\rho}^{2}_{y\vec{x}} = \frac{\operatorname{Cov}(\vec{\mu}_{y|\vec{x}}, y)}{\operatorname{Var}(\vec{\mu}_{y|\vec{x}})^{1/2}\operatorname{Var}(y)^{1/2}} = \left(\frac{\sigma_{y\vec{x}}\Sigma^{-1}_{\vec{x}\vec{x}}\sigma_{\vec{x}y}}{\sigma_{yy}}\right)^{1/2}.$$

It is still true that $0 \leqslant \rho_{\vec{x},y} \leqslant 1$.

Remark. Note that $\rho^2_{y,\vec{x}}$ gives the strength of linear association, but not direction. (It is the correlation between y and $E[y|\vec{x}]$.

Remark. The sample squared multiple correlation coefficient is called the *coefficient of determination*, denoted as R^2.

Chapter 4

Distribution of Quadratic Forms in $\vec{y}$

In chapter 3 and 4, we discussed some properties of linear functions of the random vector $\vec{y}$. We now consider quadratic forms in $\vec{y}$. We will find it useful in later chapters to express a sum of squares as a quadratic form $\vec{y}^T A\vec{y}$. In this format, we will be able to show that certain sums of squares have chi-square distributions and are independent, thereby leading to F tests.

4.1 Sum of Squares

Example 4.1. Note

$$\sum_{i=1}^{n} y_i^2 = \vec{y}^T\vec{y} = \vec{y}^T I\vec{y}.$$

Since

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\vec{j}^T\vec{y} = \frac{1}{n}\vec{y}^T\vec{j},$$

we have

$$n\bar{y}^2 = n\left(\frac{1}{n}\vec{y}^T\vec{j}\right)\left(\frac{1}{n}\vec{j}^T\vec{y}\right) = \frac{1}{n}\vec{y}^T\vec{j}\vec{j}^T\vec{y} = \vec{y}^T\left(\frac{1}{n}J\right)\vec{y}.$$

So

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \vec{y}^T\left(I-\frac{1}{n}J\right)\vec{y}.$$

Hence

$$\sum_{i=1}^{n} y_i^2 = \left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) + n\bar{y}^2$$

can be written in terms of quadratic forms as

$$\vec{y}^T I\vec{y} = \vec{y}^T\left(I-\frac{1}{n}J\right)y + \vec{y}^T\left(\frac{1}{n}J\right)\vec{y}.$$

Remark. The matrices of the three quadratic forms have the following properties:

(a) $I = \left(I - \frac{1}{n}J\right) + \frac{1}{n}J$.

(b) $I, I - \frac{1}{n}J$ and $\frac{1}{n}J$ are idempotent.

(c) $\left(I - \frac{1}{n}J\right)\left(\frac{1}{n}J\right) = 0$.

Using theorems given later in this chapter (and assuming normality of the y_i's), these three properties lead to the conclusion that

$$\sum_{i=1}^{n}(y_i - \bar{y})^2/\sigma^2 \text{ and } n\bar{y}^2/\sigma^2$$

have chi-square distributions and are independent.

4.2 Mean and Variance of Quadratic Forms

Theorem 4.2. *If $E[\vec{y}] = \vec{\mu}$ and $\operatorname{Var}(\vec{y}) = \Sigma$, and if A is symmetric matrix of constants, then*

$$E[\vec{y}^T A\vec{y}] = \operatorname{tr}(A\Sigma) + \vec{\mu}^T A\vec{\mu}.$$

Proof. Since $\Sigma = E[\vec{y}\vec{y}^T] - \vec{\mu}\vec{\mu}^T$, we have

$$\begin{aligned} E[\vec{y}^T A\vec{y}] &= E\left[\operatorname{tr}(\vec{y}^T A\vec{y})\right] = E\left[\operatorname{tr}(A\vec{y}\vec{y}^T)\right] = \operatorname{tr}\left(E[A\vec{y}\vec{y}^T]\right) \\ &= \operatorname{tr}\left(AE[\vec{y}\vec{y}^T]\right) = \operatorname{tr}\left(A(\Sigma + \vec{\mu}\vec{\mu}^T)\right) = \operatorname{tr}\left(A\Sigma + A\vec{\mu}\vec{\mu}^T\right) \\ &= \operatorname{tr}(A\Sigma) + \vec{\mu}^T A\vec{\mu}. \end{aligned}$$

□

Example 4.3. Assume y_i's are iid with $E(\vec{y}) = \mu\vec{j}$ and variance $\operatorname{Cov}(\vec{y}) = \sigma^2 I$, then

$$\begin{aligned} E[s^2] &= \frac{1}{n-1}E\left[\vec{y}^T\left(I - \frac{1}{n}J\right)\vec{y}\right] \\ &= \frac{1}{n-1}\left(\operatorname{tr}\left(\left(I - \frac{1}{n}J\right)(\sigma^2 I)\right) + \mu\vec{j}^T\left(I - \frac{1}{n}J\right)\mu\vec{j}\right) \\ &= \frac{1}{n-1}\left(\sigma^2(n-1) + \mu^2\left(\vec{j}^T\vec{j} - \frac{1}{n}\vec{j}^T\vec{j}\vec{j}^T\vec{j}\right)\right) \\ &= \sigma^2. \end{aligned}$$

Example 4.4. Assume x_i's are iid with $E(\vec{x}) = \mu\vec{\mathbb{1}}_n$ and variance $\operatorname{Cov}(\vec{x}) = \sigma^2 I$. Consider

$$Q(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \|P_{V^\perp}\vec{x}\| = \vec{x}^T(I - P_V)\vec{x},$$

where $V = L(\mathbb{1}_n)$. Then

$$\begin{aligned} E[Q(x)] &= E\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\ &= \operatorname{tr}\left((I - P_V)(\sigma^2 I)\right) + \mu\vec{\mathbb{1}}_n^T(I - P_V)\vec{\mu}\mathbb{1}_n \\ &= \sigma^2\operatorname{tr}(I - P_V) + \mu\left(\vec{\mathbb{1}}_n^T - \vec{\mathbb{1}}_n\right)\vec{\mathbb{1}}_n = \sigma^2(n-1). \end{aligned}$$

Alternatively, define $\vec{y} = \vec{x} - \bar{x}\mathbb{1}_n = (x_1 - \bar{x}, \dots, x_n - \bar{x})^T$. Then $\vec{y} = P_{V^\perp}\vec{x}$ and

$$Q(\vec{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \vec{y}^T I \vec{y}.$$

$$\operatorname{Var}(\vec{y}) = \operatorname{Var}(P_{V^\perp}\vec{x}) = P_{V^\perp}\operatorname{Var}(\vec{x})(P_{V^\perp})^T = \sigma^2 P_{V^\perp}.$$

So

$$E[Q(x)] = \operatorname{tr}\left(I(\sigma^2 P_{V^\perp})\right) + 0 = \sigma^2(n-1).$$

Theorem 4.5. *If $\vec{y} \sim N_p(\vec{\mu}, \Sigma)$, then the MGF of $\vec{y}^T A\vec{y}$ is*

$$M_{\vec{y}^T A\vec{y}}(t) = |I - 2tA\Sigma|^{-1/2} e^{-\vec{\mu}^T[I-(I-2tA\Sigma)^{-1}]\Sigma^{-1}\vec{\mu}/2}.$$

Theorem 4.6. *If $\vec{y} \sim N_p(\vec{\mu}, \Sigma)$, then*

$$\operatorname{Cov}(\vec{y}, \vec{y}A\vec{y}) = 2\Sigma A\vec{\mu}.$$

Proof.

$$\begin{aligned}
\operatorname{Cov}(\vec{y}, \vec{y}^T A\vec{y}) &= E\left[(\vec{y} - E[\vec{y}])(\vec{y}^T A\vec{y} - E[\vec{y}A\vec{y}])\right] \\
&= E\left[(\vec{y} - \vec{\mu})(\vec{y}^T A\vec{y} - \vec{\mu}^T A\vec{\mu} - \operatorname{tr}(A\Sigma)\right] \\
&= E\left[(\vec{y} - \vec{\mu})\left((\vec{y} - \vec{\mu})^T A(\vec{y} - \vec{\mu}) + 2(\vec{y} - \vec{\mu})^T A\vec{\mu} - \operatorname{tr}(A\Sigma)\right)\right] \\
&= 0 + 2\Sigma A\vec{\mu} - 0,
\end{aligned}$$

since all third central moments of the multivariate normal are zero. □

Corollary 4.7. If $\vec{y} \sim N_p(\vec{\mu}, \Sigma)$, and B is $k \times p$ matrix of constants. Then

$$\operatorname{Cov}(B\vec{y}, \vec{y}^T A\vec{y}) = 2B\Sigma A\vec{\mu}.$$

Theorem 4.8. *Let $\vec{v} = \begin{bmatrix}\vec{y} \\ \vec{x}\end{bmatrix}$ be a partitioned random vector with $E = \begin{bmatrix}\vec{y} \\ \vec{x}\end{bmatrix} = \begin{bmatrix}\vec{\mu}_y \\ \vec{\mu}_x\end{bmatrix}$ and*

$$\operatorname{Cov}\begin{bmatrix}\vec{y} \\ \vec{x}\end{bmatrix} = \begin{bmatrix}\vec{\Sigma}_{yy} & \Sigma_{yx} \\ \vec{\Sigma}_{xy} & \Sigma_{xx}\end{bmatrix}.$$

Then

$$E(\vec{x}^T A\vec{y}) = \operatorname{tr}(A\Sigma_{yx}) + \vec{\mu}_x^T A\vec{\mu}_y.$$

4.3 Noncentral Chi-square Distribution

Recall if $z_1, \dots, z_n$ are iid $N(0,1)$, then $u = \sum_{i=1}^n z_i^2 = \vec{z}^T\vec{z} \sim \chi^2(n)$ and

$$f(u) = \frac{u^{n/2-1}e^{-y/2}}{\Gamma(n/2)2^{n/2}},\ u > 0,$$

where

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx,$$
$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1),$$
$$\Gamma(1/2) = \sqrt{\pi}.$$
$$E(u) = n,$$
$$\mathrm{Var}(u) = 2n,$$
$$M_u(t) = \frac{1}{(1-2t)^{n/2}}.$$

Theorem 4.9. *Let $X \sim N(\mu, 1)$ and $Y = X^2$, then*

$$f_Y(y) = \sum_{k=0}^\infty f_{\chi^2_{2k+1}} P(K = k),$$

where $K \sim \mathrm{Poi}(\mu^2/2)$. Hierarchically,

$$Y|K \sim \chi^2_{2K+1},$$
$$K \sim \mathrm{Poi}(\lambda).$$

Theorem 4.10. *To generate Y:*

(a) Generate $K \sim \mathrm{Poi}(\lambda)$;

(b) Generate $2K+1$ χ^2_1 random variables and add them up, that is Y. (i.e., a χ^2 r.v. with random df.)

Definition 4.11. Assume $y_i \sim N(\mu_i, 1), i \in [n]$ are independent, then

$$v = \sum_{i=1}^n y_i^2 = \vec{y}^T \vec{y}$$

is called the *noncentral chi-square distribution* and is denoted by

$$\chi^2(n, \lambda).$$

The *noncentrality parameter* λ is defined as

$$\lambda = \frac{1}{2}\sum_{i=1}^n \mu_i^2 = \frac{1}{2}\vec{\mu}^T\vec{\mu}.$$

Then

$$E\left[\sum_{i=1}^n (y_i - \mu_i)^2\right] = n,$$
$$E\left[\sum_{i=1}^n y_i^2\right] = \sum_{i=1}^n E[y_i^2] = \sum_{i=1}^n (\mu_i^2 + 1) = n + 2\lambda.$$

Theorem 4.12 (additive)**.** *If $v_1, \dots, v_k$ are independently distributed as*

$$\chi^2(n_i, \lambda_i),$$

then

$$\sum_{i=1}^k v_i \sim \chi^2\left(\sum_{i=1}^k n_i, \sum_{i=1}^k \lambda_i\right).$$

Proof. By Theorem 4.9. □

Theorem 4.13. *If $v \sim \chi^2(n, \lambda)$, then*

$$E[v] = n + 2\lambda,$$

$$\text{Var}(v) = 2n + 8\lambda,$$

$$M_v(t) = \frac{1}{(1-2t)^{n/2}} e^{-\lambda[1-1/(1-2t)]},\ t < \frac{1}{2}.$$

Proof. (?) Let $K \sim \text{Poi}(\lambda)$, then

$$\begin{aligned} E[v] &= E\left[E[v|K]\right] = E[2K+n] = 2\lambda + n, \\ \text{Var}(v) &= \text{Var}\left(E[v|K]\right) + E\left[\text{Var}(v|K)\right] \\ &= \text{Var}(2K+n) + E\left[2(2K+n)\right] \\ &= \text{Var}(2K) + E[4K+2n] \\ &= 4\lambda + 4\lambda + 2n \\ &= 2n + 8\lambda. \end{aligned}$$

Lastly, since the MGF of χ^2_ν is

$$\frac{1}{(1-2t)^{\nu/2}},\ t < \frac{1}{2}.$$

$$M_v(t) = E\left[e^{tv}\right] = E\left[E\left[\left. e^{tv}\right| K\right]\right] = E\left[\left(\frac{1}{1-2t}\right)^{\frac{2K+n}{2}}\right].$$

□

Remark. $\chi^2(n,0) = \chi^2(n) = \chi^2_n$.

Theorem 4.14. *If $\vec{y} \sim N_n(\vec{\mu}, I)$, then*

$$\|\vec{y} - \vec{\mu}\|^2 = (\vec{y} - \vec{\mu})^T(\vec{y} - \vec{\mu}) \sim \chi^2_n.$$

4.4 Distribution of Quadratic Forms

Theorem 4.15. *If $\vec{y} \sim N_n(\vec{\mu}, \Sigma)$, then*

$$(\vec{y} - \vec{\mu})^T \Sigma^{-1} (\vec{y} - \vec{\mu}) = \left(\Sigma^{-1/2}(\vec{y} - \vec{\mu})\right)^T \left(\Sigma^{-1/2}(\vec{y} - \vec{\mu})\right) = \vec{z}^T \vec{z} \sim \chi^2_n,$$

where $\vec{z} = \Sigma^{-1/2}(\vec{y} - \vec{\mu}) \sim N(\vec{0}, I)$.

Theorem 4.16. *If $\vec{y} \sim N_n(\vec{\mu}, \Sigma)$, then the* MGF *of $C = \vec{y}^T A \vec{y}$ is*

$$M_c(t) = |I - 2tA\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\vec{\mu}^T \left[I - (I - 2tA\Sigma)^{-1}\right] \Sigma^{-1}\vec{\mu}\right\}.$$

We saw previously that a projection P_V onto a space $V \subseteq \mathbb{R}^n$ with $\dim(V) = k$, has k-evals 1 and $n - k$-evals 0. The idea can be extended to any idempotent matrix.

Theorem 4.17. *If $A \in \mathbb{R}^{n \times n}$ is idempotent of rank r, then A has r e-vals = 1 and $n - r$ e-vals = 0.*

Proof. Since $\operatorname{rank}(A) = r$, then $\dim(C(A)) = r$. So take $\{\vec{x}_1, \dots, \vec{x}_r\}$ as a basis for $C(A)$, then for $i \in [k]$, $\vec{x}_i = A\vec{z}$, for some $\vec{z}$. Then

$$A\vec{x}_i = A(A\vec{z}) = A\vec{z} = \vec{x}_i.$$

So each $\vec{x}_i$ is an e-vect corresponding to $\lambda = 1$. But there are r LIN such vectors, so the gem. mult. of $\lambda = 1$ is r, which is less than algebraic multiplicity. But $\operatorname{rank}(N(A)) = n - r$, then $\lambda = 0$ has alg. mult. of at least $n - r$. So A has r-vals = 1 and $n - r$ e-vals = 0. □

Corollary 4.18. If A is idempotent, by Theorem 4.17, $\operatorname{rank}(A) = \sum_i \lambda_i$, so

$$\operatorname{rank}(A) = \operatorname{tr}(A).$$

Theorem 4.19. *Let $\vec{y} \sim N_p(\vec{\mu}, \Sigma)$, let A be a symmetric matrix of constants of rank r, and let $\lambda = \frac{1}{2}\vec{\mu}^T A \vec{\mu}$. Then*

$$\vec{y}^T A \vec{y} \sim \chi^2(r, \lambda) \iff A\Sigma \textit{ idempotent.}$$

Corollary 4.20. If $\vec{y} \sim N_p(\vec{0}, \sigma^2 I)$, then

$$\vec{y}^T A \vec{y}/\sigma^2 \sim \chi^2(r) \iff A \text{ is idempotent of rank } r.$$

Corollary 4.21. If $\vec{y} \sim N_p(\vec{\mu}, \sigma^2 I)$, then

$$\vec{y}^T A \vec{y}/\sigma^2 \sim \chi^2(r, \lambda) \iff A \text{ is idempotent of rank } r,$$

where $\lambda = \frac{\vec{\mu}^T A \vec{\mu}}{2\sigma^2}$.

Corollary 4.22. If $\vec{y} \sim N_p(\vec{\mu}, \sigma^2 I)$, and let P_V be the projection matrix onto a subspace $V \subseteq \mathbb{R}^n$ of dimension $r \leqslant n$. Then

$$\vec{y}^T P_V \vec{y}/\sigma^2 = \frac{1}{\sigma^2}\|p(\vec{y}|V)\|^2 \sim \chi^2(r, \lambda).$$

where $\lambda = \frac{\vec{\mu}^T P_V \vec{\mu}}{2\sigma^2} = \frac{1}{2\sigma^2}\|p(\vec{\mu}|V)\|^2$.

Corollary 4.23. Suppose $\vec{y} \sim N_p(\vec{\mu}, \Sigma)$ and let $\vec{c} \in \mathbb{R}^p$ be arbitrary, then

$$(\vec{y} - \vec{c})^T \Sigma^{-1} (\vec{y} - \vec{c}) \sim \chi^2(n, \lambda),$$

where $\lambda = \frac{1}{2}(\vec{\mu} - \vec{c})^T \Sigma^{-1} (\vec{\mu} - \vec{c})$.

Example 4.24. Assume $\vec{y} \sim N_n(\mu\vec{j}, \sigma^2 I)$, then

$$(n-1)s^2/\sigma^2 \sim \chi^2(n-1).$$

Proof.

$$(n-1)s^2/\sigma^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \vec{y}^T\left[\frac{I-(1/n)J}{\sigma^2}\right]\vec{y}.$$

Since $\frac{I-(1/n)J}{\sigma^2}\sigma^2 I = I - \frac{1}{n}J$ is idempotent,

$$\text{rank}\left(\frac{I-(1/n)J}{\sigma^2}\right) = \text{rank}\,(I-(1/n)J) = \text{tr}\,(I-(1/n)J) = n-1,$$

and

$$\lambda = \frac{\mu\vec{j}^T\left(I - \frac{1}{n}J\right)\mu\vec{j}}{2\sigma^2} = \frac{\mu^2(n-n)}{2\sigma^2} = 0,$$

we have $(n-1)s^2/\sigma^2 \sim \chi^2(n-1)$. □

The classical linear model has the form $\vec{y} \sim N_n(\vec{\mu}, \sigma^2 I)$, where

$$\vec{\mu} = X\vec{\beta} \in V = L(\vec{x}_1, \ldots, \vec{x}_k) = C(X).$$

We are interested in the statistical properties of $\hat{\vec{y}} = p(\vec{y}|V)$ and function of the residual vector $\vec{y} - \hat{\vec{y}} = p(\vec{y}|V^{\perp})$.

Theorem 4.25. *Let $V \leqslant \mathbb{R}^n$ have dimension k, and let $\vec{y} \in \mathbb{R}^n$ a random vector with mean $E[\vec{y}] = \vec{\mu}$. Then*

(a)

$$E[p(\vec{y}|V)] = p(\vec{\mu}|V).$$

(b) If $\text{Var}(\vec{y}) = \sigma^2 I$, then

$$\text{Var}(p(\vec{y}|V)) = \sigma^2 P_V,$$

$$E\left[\|p(\vec{y}|V)\|^2\right] = \sigma^2 k + \|p(\vec{\mu}|V)\|^2.$$

(c) If further, $\vec{y} \sim N_n(\vec{\mu}, \sigma^2 I)$, then

$$p(\vec{y}|V) \sim N_n\left(p(\vec{\mu}|V), \sigma^2 P_V\right),$$

$$\frac{1}{\sigma^2}\|p(\vec{y}|V)\|^2 = \frac{1}{\sigma^2}\vec{y}^T P_V \vec{y} \sim \chi_k^2\left(\frac{1}{2\sigma^2}\|p(\vec{\mu}|V)\|^2\right).$$

Proof.

$$E[p(\vec{y}|V)] = E[P_V\vec{y}] = P_V E[\vec{y}] = P_V\vec{\mu} = p(\vec{\mu}|V).$$

$$\text{Var}\,(p(\vec{y}|V)) = \text{Var}\,(P_V\vec{y}) = P_V(\sigma^2 I)P_V^T = \sigma^2 P_V.$$

$$E\left[\|p(\vec{y}|V)\|^2\right] = E\left[\vec{y}^T P_V \vec{y}\right] = \text{tr}(\sigma^2 P_V) + \vec{\mu}^T P_V \vec{\mu} = \sigma^2 k + \|p(\vec{\mu}|V)\|^2.$$

□

$$\text{Sample space } \mathbb{R}^n \begin{cases} \text{Mutually orthogonal model space: } V_1, \ldots, V_k. \\ \text{Error space.} \end{cases}$$

Theorem 4.26. *Let $V_1, \ldots, V_k$ be mutually orthogonal subspaces of $\mathbb{R}^n$ with dimensions $d_1, \ldots, d_k$, respectively. Let $\vec{y} \in \mathbb{R}^n$ with mean $E[\vec{y}] = \vec{\mu}$. Let*

$$\hat{\vec{y}}_i = p(\vec{y}|V_i) = P_{V_i}\vec{y}, \forall i \in [k],$$

$$\vec{\mu}_i = p(\vec{\mu}|V_i) = P_{V_i}\vec{\mu}, \forall i \in [k].$$

(a) If $\mathrm{Var}(\vec{y}) = \sigma^2 I$, then

$$\mathrm{Cov}(\hat{y}_i, \hat{y}_j) = 0, \forall i \neq j.$$

(b) If $\vec{y} \sim N_n(\vec{\mu}, \sigma^2 I)$, then $\hat{y}_1, \ldots, \hat{y}_k$ are independent with $\hat{y}_i \sim N(\vec{\mu}_i, \sigma^2 P_{V_i})$.

(c) If $\vec{y} \sim N_n(\vec{\mu}, \sigma^2 I)$, then $\|\hat{y}_1\|^2, \ldots, \|\hat{y}_k\|^2$ are independent with

$$\frac{1}{\sigma^2}\|\hat{y}_i\|^2 \sim \chi^2_{d_i}\left(\frac{1}{2\sigma^2}\|\mu_i\|^2\right).$$

Proof. (a) For $i \neq j$,

$$\mathrm{Cov}(\hat{y}_i, \hat{y}_j) = \mathrm{Cov}(P_{V_i}\vec{y}, P_{V_j}\vec{y}) = \sigma^2 P_{V_i} P_{V_j} = 0,$$

since $P_{V_i} P_{V_j} \vec{z} = \vec{0}$ for any $\vec{z}$.

(b)

$$E[\hat{y}_i] = E[P_{V_i}\vec{y}] = P_{V_i}\vec{\mu} = \vec{\mu}_i.$$

$$\mathrm{Var}(\hat{y}_i) = \mathrm{Var}(P_{V_i}\vec{y}) = P_{V_i}\sigma^2 I P_{V_i}^T = \sigma^2 P_{V_i}.$$ □

Example 4.27. Let $\vec{y} = X\vec{\beta} + \vec{e}$, where $X \in \mathbb{R}^{n \times p}$, $\mathrm{rank}(X) = p$ and $\vec{e} \sim N(0, \sigma^2 I)$. Take

$$P_V = I - X(X^T X)^{-1} X^T,$$

then $\mathrm{SSE} = \vec{y}^T P_V \vec{y}$. Then

$$\frac{\mathrm{SSE}}{\sigma^2} = \vec{y}^T \left(\frac{I - X(X^T X)^{-1} X}{\sigma^2}\right) \vec{y}.$$

But $p(E[\vec{y}]|V) = P_V E[\vec{y}] = P_V X\vec{\beta} = 0$. By Theorem 4.26(c), we have

$$\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2_{n-p}.$$

4.5 Independence of Linear Forms and Quadratic Forms

Lemma 4.28. If $\vec{y} \sim N_n(\vec{\mu}, \Sigma)$, then

$$\mathrm{Cov}(\vec{y}, \vec{y}^T A\vec{y}) = 2\Sigma A\vec{\mu}.$$

Corollary 4.29. If $B \in \mathbb{R}^{k\times n}$ be constant and $\vec{y} \sim N_n(\vec{\mu}, \Sigma)$, then

$$\mathrm{Cov}(B\vec{y}, \vec{y}A\vec{y}) = 2B\Sigma A\vec{\mu}.$$

Theorem 4.30. *Suppose that B is a $k\times p$ matrix of constants, A is a $p\times p$ symmetric of constants, and $\vec{y} \sim N_p(\vec{\mu}, \Sigma)$. Then*

$$B\vec{y} \perp\!\!\!\perp \vec{y}^T A\vec{y} \Longleftrightarrow B\Sigma A = \mathbf{0}.$$

Corollary 4.31. If A is $p \times p$ symmetric and $\vec{y} \sim N_p(\vec{\mu}, \sigma^2 I)$, then

$$B\vec{y} \perp\!\!\!\perp \vec{y}^T A\vec{y} \Longleftrightarrow BA = \mathbf{0}.$$

Example 4.32. Assume $\vec{y} \sim N_n(\mu\vec{j}, \sigma^2 I)$. Since

$$s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1) = \vec{y}^T\left(I - \frac{1}{n}J\right)\vec{y},$$

$\bar{y} = \frac{1}{n}\vec{j}^T y$ and $\frac{1}{n}\vec{j}^T\left(I - \frac{1}{n}J\right) = \frac{1}{n}\vec{j}^T - \frac{1}{n^2}n\vec{j}^T = 0$, (or: $\mathbb{1}_n^T P_{L^\perp(\mathbb{1}_n)} = \left(P_{L^\perp(\mathbb{1}_n)}\mathbb{1}_n\right)^T = 0$,) we have $\bar{y} \perp\!\!\!\perp s^2$.

Theorem 4.33. *Let A and B be symmetric matrices of constants. If $\vec{y} \sim N_p(\vec{\mu}, \Sigma)$, then*

$$\vec{y}^T A\vec{y} \perp\!\!\!\perp \vec{y}^T B\vec{y} \Longleftrightarrow A\Sigma B = \mathbf{0}.$$

Corollary 4.34. Let A and B be symmetric matrices of constants and $\vec{y} \sim N_p(\vec{\mu}, \sigma^2 I)$, then

$$\vec{y}^T A\vec{y} \perp\!\!\!\perp \vec{y}^T B\vec{y} \Longleftrightarrow AB = 0. \text{ (or equivalently } BA = 0.)$$

Example 4.35. Express

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

as

$$\vec{y}^T\vec{y} = \vec{y}^T\left(I - \frac{1}{n}J\right)\vec{y} + \vec{y}^T\left(\frac{1}{n}J\right)\vec{y}.$$

Assume $\vec{y} \sim N_n(\mu\vec{j}, \sigma^2 I)$. Since $JJ = nJ$, we have $\left(I - \frac{1}{n}J\right)\left(\frac{1}{n}J\right) = 0$. So

$$\vec{y}^T\left(I - \frac{1}{n}J\right)\vec{y} \perp\!\!\!\perp \vec{y}^T\left(\frac{1}{n}J\right)\vec{y},$$

which is obvious since we have shown $s^2 \perp\!\!\!\perp \bar{y}$.

4.6 Noncentral F and t Distribution

4.6.1 Noncentral F distribution

Recall if $u \sim \chi^2(p)$, $v \sim \chi^2(q)$ and u and v are independent, then

$$w = \frac{u/p}{v/q} \sim F(p,q).$$

Moreover

$$E(w) = \frac{q}{q-2},$$

$$\operatorname{Var}(w) = \frac{2q^2(p+q-2)}{p(q-1)^2(q-4)}.$$

Definition 4.36. Suppose that $u \sim \chi^2(p,\lambda)$, while $v \sim \chi^2(q)$ with u and v independent. Then

$$z = \frac{u/p}{v/q} \sim F(p,q,\lambda),$$

the *noncentral F distribution* with noncentrality parameter λ, where λ is the same noncentrality parameter as in the distribution of noncentral chi-square distribution.

$$E[z] = \frac{q}{q-2}\left(1 + \frac{2\lambda}{p}\right),$$

which of course, greater than $E[w]$.

Remark. When an F statistic is used to test a hypothesis H_0, the distribution will typically be central if the (null) hypothesis is true and noncentral if the hypothesis is false. Thus the noncentral F distribution can often be used to evaluate the power of an F test. The *power* of a test is the probability of rejecting H_0 for a given value of λ. If F_α is the upper α percentage point of the central F distribution, then the power, $P(p,q,\alpha,\lambda)$ can be defined as

$$P(p,q,\alpha,\lambda) = P(z \geqslant F_\alpha),$$

where z is the noncentral F random variable.

Example 4.37. $\vec{y} = X\vec{\beta} + \vec{e}$, where $\vec{e} \sim N_n(\vec{0}, \sigma^2 I)$, then $\vec{y} \sim N_n(\vec{X}\beta, \sigma^2 I)$,

$$\mathrm{SSE}_{\text{full}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \vec{y}^T(I - X(X^TX)^{-1}X^T)\vec{y} = \vec{y}^T A\vec{y},$$

where $A = I - P_{C(X)}$. Then $AX = 0$. Let $X = (X_1, X_2)$,

$$\vec{\beta} = \begin{bmatrix} \vec{\beta}_1 \\ \vec{\beta}_2 \end{bmatrix}.$$

Consider testing $H_0 : \vec{\beta}_2 = \vec{0}$, then

$$H_0 : \text{Reduced model: } \vec{y} = X_1\vec{\beta}_1 + \vec{e}.$$

So

$$\mathrm{SSE}_{\mathrm{red}} = \vec{y}^T \left(I - X_1(X_1^T X_1)^{-1} X_1^T \right) \vec{y}.$$

The extra SS is given by

$$\mathrm{SS}_{H_0} = \mathrm{SSE}_{\mathrm{red}} - \mathrm{SSE}_{\mathrm{full}} = \vec{y}^T \left(X(X^T X)^{-1} X^T - X_1(X_1^T X_1)^{-1} X_1^T \right) \vec{y} = \vec{y}^T B \vec{y},$$

where

$$B = P_{C(X)} - P_{C(X_1)} = X(X^T X)^{-1} X^T - X_1(X_1^T X_1)^{-1} X_1^T.$$

Then $AB = AP_{C(X)} - AP_{C(X_1)}$. But $AP_{C(X)} = (I - P_{C(X)})P_{C(X)} = 0$. Also,

$$AX_1 = AX \begin{bmatrix} I \\ 0 \end{bmatrix} = 0.$$

So $AP_{C(X_2)} = 0$. Thus, $AB = 0$. So $\mathrm{SS}_{H_0} \perp\!\!\!\perp \mathrm{SSE}_{\mathrm{full}}$. Hence we can use these independent SS. to build approximate F-statistics to test $H_0 : \vec{\beta}_2 = 0$.

Theorem 4.38. *Let $\vec{y} = X\vec{\beta} + \vec{e}$, where $X \in \mathbb{R}^{n \times k}$ with* rank *$r \leqslant k$ and $\vec{e} \sim N(\vec{0}, \sigma^2 I)$. Let $V_1 \subseteq C(X)$ be a subspace with* $\dim(V_1) = r_1 \leqslant r$. *Let $\hat{y} = p(\vec{y}|C(X))$. Then*

$$F = \frac{\|p(\vec{y}|V_1)\|^2 / r_1}{\left\|\vec{y} - \hat{\vec{y}}\right\|^2 / (n-r)} \sim F_{r_1, n-r}\left(\frac{1}{2\sigma^2} \left\| p(X\vec{\beta}|V_1) \right\|^2 \right).$$

Proof. $\vec{y} - \hat{\vec{y}} = P(\vec{y}|C^\perp(X))$, so we know from previous example, $Q_1 = \|p(\vec{y}|V_1)\|^2$ and $Q_2 = \left\|p(\vec{y}|C^\perp(X))\right\|^2$ are independent r.v.'s and that since $\mathrm{rank}(P_{V_1}) = r_1$,

$$Q_1/\sigma^2 \sim \chi^2_{r_1}\left(\frac{1}{2\sigma^2} \left\| p(X\vec{\beta}|V_1) \right\|^2 \right),$$

$$Q_2/\sigma^2 \sim \chi^2_{n-r}.$$

□

4.6.2 Noncentral t Distribution

Recall if $z \sim N(0,1)$ and $u \sim \chi^2(p)$ and z and u are independent, then

$$t = \frac{z}{\sqrt{u/p}} \sim t(p).$$

Definition 4.39. Suppose that $y \sim N(\mu, 1)$ and $u \sim \chi^2(p)$ and y and u are independent. Then

$$t = \frac{y}{\sqrt{u/p}} \sim t(p, \mu),$$

the *noncentral t distribution* with p degrees of freedom and noncentrality parameter μ.

Example 4.40. If $y \sim N(\mu, \sigma^2)$, then

$$t = \frac{y/\sigma}{\sqrt{u/p}} \sim t(p, \mu/\sigma)$$

since $y/\sigma \sim N(\mu/\sigma, 1)$.

Theorem 4.41. *Let $y_1, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then for any constant μ_0,*

$$T = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}(\lambda),$$

where

$$\lambda = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}.$$

Proof. Let $u = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}$, then $u \sim N(\frac{\mu-\mu_0}{\sigma/\sqrt{n}}, 1)$. Note that

$$T = \frac{\frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{s^2/\sigma^2}} = \frac{u}{\sqrt{\chi^2_{n-1}/n-1}}. \qquad \square$$

Chapter 5

Simple Linear Regresssion

5.1 The Model

Definition 5.1. The *simple linear regression* model for n observations can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n.$$

The designation *simple* indicates that there is only one x to predict the response y, and *linear* means that the above model is linear in β_0 and β_1. In this chapter, we assume that y_i and ϵ_i are random variables and the values of x_i are known constants, which means that the same values of $x_1, \ldots, x_n$ would be used in repeated sampling. To complete the above model, we make the following additional assumptions:

(a) $E(\epsilon_i) = 0$ for $i = 1, \ldots, n$, or equivalently, $E[y_i] = \beta_0 + \beta_1 x_i$.

(b) $\mathrm{Var}(\epsilon) = \sigma^2$ or $i = 1, \ldots, n$, or equivalently, $\mathrm{Var}(y_i) = \sigma^2$.

(c) $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$, or equivalently, $\mathrm{Cov}(y_i, y_j) = 0$.

Remark. (a) Assumption 1 states y_i depends only on x_i and that all other variation in y_i is random.

(b) Assumption 2 asserts the variance of ϵ or y does not depend on the value of x_i, known as the assumption of homoscedasticity, homogeneous variance or constant variance.

(c) Assumption 3 says ϵ or y variables are uncorrelated with each other. (If we add a normality assumption, the ϵ or y variables will be independent.)

5.2 Estimation of β_0, β_1 and σ^2

Using a random sample of n observations $y_1, \ldots, y_n$ and the accompanying fixed values $x_1, \ldots, x_n$, we can estimate the parameters β_0, β_1 and σ^2. To obtain the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we use the method of least squares, which does not require any distributional assumptions.

In the least-square approach, we seek estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squares of the deviations $y_i - \hat{y}_i$ of the n observed y_i's from their predicted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

$$\hat{\vec{\epsilon}}^T\hat{\vec{\epsilon}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Note that the predicted value $\hat{y}_i$ estimate $E[y_i]$, not y_i; that is, $\hat{\beta}_0 + \hat{\beta}_1 x_i$ estimates $\beta_0 + \beta_1 x_i$, not $\beta_0 + \beta_1 x_i + \epsilon_i$. A better notation would be $\hat{E}[y_i]$, but $\hat{y}_i$ is commonly used. By differentiating the above formula, we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\overline{x}\cdot\overline{y}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

where

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}.$$

Note the three assumptions in section 1 were not used in deriving the least-square estimators. It is not necessary that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be based on $E[y_i] = \beta_0 + \beta_1 x_i$, that is, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ can be fit to a set of data for which

$$E[y_i] \neq \beta_0 + \beta_1 x_i.$$

However, if the three assumptions holds, then the least-square estimator $\hat{\beta}_0$ and $\hat{\beta}_1$ are **unbiased** and have **minimum** variance among all linear unbiased estimators. Use the three assumptions, we obtain the following means and variance of $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\begin{aligned} E[\hat{\beta}_1] &= \beta_1, \\ E[\hat{\beta}_0] &= \beta_0, \\ \mathrm{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}, \\ \mathrm{Var}(\hat{\beta}_0) &= \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]. \end{aligned}$$

Note in discussing $E[\hat{\beta}_1]$ and $\mathrm{Var}(\hat{\beta}_1)$, for example, we are considering random variation of $\hat{\beta}_1$ from sample for sample. It is assumed that n values $x_1, \ldots, x_n$ would remain the same in future samples so that $\mathrm{Var}(\hat{\beta}_1)$ and $\mathrm{Var}(\hat{\beta}_0)$ are **constant**. Furthermore, we see that when $\sum_{i=1}^n (x_i - \overline{x})^2$ is maximized, $\mathrm{Var}(\hat{\beta}_1)$ is minimized. By assumption 2 in section 1, σ^2 is the same for each y_i. Then $\sigma^2 = E\left[(y_i - E[y_i])^2\right]$ for each i. Using $\hat{y}_i$ as an estimator of $E[y_i]$, we estimate σ^2 by an average from the sample, that is

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2} = \frac{\mathrm{SSE}}{n-2}.$$

The deviation $\hat{\epsilon}_i = y_i - \hat{y}_i$ is often called the *residual* of y_i and SSE is called the *residual sum of squares*. With $n-2$ in the denominator, s^2 is an unbiased estimator of σ^2. Intuitively, we divide by $n-2$ instead of $n-1$ because $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ has two estimated parameters and should thereby be a better estimator of $E[y_i]$ than $\overline{y}$. Thus we expect $\mathrm{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 < \sum_{i=1}^n (y_i - \overline{y})^2$. Using some algebra, we indeed have

$$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 - \frac{\left[\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})\right]^2}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

5.3 Hypothesis Test and Confidence Interval for β_1

Typically, hypothesis about β_1 are of more interest than hypothesis about β_0 since our first priority is to determine whether there is a linear relationship between y and x. In order to obtain a test for $H_0 : \beta_1 = 0$, we assume that $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Then

(a)

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right).$$

(b)

$$(n-2)s^2/\sigma^2 \sim \chi^2(n-2).$$

(c)

$$\hat{\beta}_1 \perp\!\!\!\perp s^2.$$

(We can not use the distribution of $\hat{\beta}_1$ directly since σ is unknown.) From these three properties it follows that

$$t = \frac{\hat{\beta}_1}{s/\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}} = \frac{\hat{\beta}_1/\sigma/\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}{\sqrt{(n-2)s^2/\sigma^2/n-2}} \sim t(n-2, \delta),$$

where

$$\delta = \frac{E[\hat{\beta}_1]}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} = \frac{\beta_1}{\sigma/\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}.$$

If $\beta_1 = 0$, then $t \sim t(n-2)$. For a two sided alternative hypothesis $H_1 : \beta_1 \neq 0$, we reject $H_0 : \beta_1 = 0$ if $|t| \geqslant t_{\alpha/2,n-2}$, where $t_{\alpha/2,n-2}$ is the upper $\alpha/2$ percentage point of the **central** $t(n-2)$ distribution and α is the desired significant level of the test (probability of rejecting). Alternatively, we reject H_0 if the p value $p < \alpha$. For a two sided test, the p value is defined as twice the probability that $t(n-2)$ exceeds the absolute value of the observe t. A $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}.$$

5.4 Coefficient of Determination

Definition 5.2. The *coefficient of determination* r^2 is defined as

$$r^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\sum_{i=1}^n (\hat{y}_i - \overline{y})}{\sum_{i=1}^n (y_i - \overline{y})^2},$$

where SSR is the regression sum of squares and SST is the total sum of squares.

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$$

Thus r^2 gives the proportion of variation in y that is explained by the model or, equivalently, accounted for by the regression on x.

We have labeled as r^2 because it is the same as the square of the *sample correlation coefficient* r between y and x

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \overline{x})^2 \sum_{i=1}^n (y_i - \overline{y})^2}}.$$

Chapter 6

Multiple Regression: Estimation

6.1 Introduction

In multiple regression, we attempt to predict a dependent or response variable y on the basis of an assumed linear relationship with several independent or predictor variables $x_1, x_2, \dots, x_k$. In addition to constructing a model for prediction, we may wish to assess the extent of the relationship between y and the x variables. For this purpose, we use the multiple correlation coefficient R. In this chapter, y is a continuous random variable and the x variables are fixed con- stants (either discrete or continuous) that are controlled by the experimenter.

6.2 The model

$$\vec{y} = X\vec{\beta} + \vec{\epsilon}.$$

The assumptions on ϵ_i and y_i are

$$E[\vec{\epsilon}] = 0 \text{ or } E[\vec{y}] = X\vec{\beta}.$$

$$\text{Cov}(\vec{\epsilon}) = \sigma^2 I \text{ or } \text{Cov}(\vec{y}) = \sigma^2 I.$$

In summary, if $\vec{\epsilon} \sim N_n(0, \sigma^2 I)$, then

$$\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I).$$

In multiple linear regression, the explanatory variables are usually LIN. So we assume $\text{rank}(X) = k + 1$. Also, we assume $n > k + 1$, so the model involves data reduction (i.e., summerization). If $r = k + 1$, there is no reduction, just transformation.

Remark. The $\vec{\beta}$ arguments are called (partial) regression coefficients. Mathematically, the partial derivative of $E[\vec{y}]$ w.r.t x_1 is β_1. Thus, β_1 indicates the change in $E[\vec{y}]$ with a unit increase in x_1 when the other predictors are held constant. Statistically, β_1 shows the effect x_1 on $E[\vec{y}]$ in the presence of the other x's. This effect would typically be different from the effect of x_1 on $E[\vec{y}]$ if the other x's were not present in the model. For example

$$\vec{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

will usually be different from β_0^* and β_1^* in $\vec{y} = \beta_0^* + \beta_1^* x + \epsilon^*$. If $\vec{x}_1$ and $\vec{x}_2$ are orthogonal, that is, if $\vec{x}_1^T \vec{x}_2 = 0$ or if $(\vec{x}_1 - \bar{x}_1 I_n)^T(\vec{x}_2 - \bar{x}_2 I_n) = 0$, then $\beta_0 = \beta_0^*$ and $\beta_1 = \beta_1^*$.

Remark. If we are only interested in prediction, we usually only need to find $\hat{\vec{\mu}} = X\hat{\vec{\beta}}$. Often, we are interested in estimating $\hat{\vec{\beta}}$.

6.3 Estimation of $\vec{\beta}$ and σ^2

6.3.1 Least Squares Estimator for $\vec{\beta}$

6.3.2 Properties of the Least-Square Estimator $\hat{\vec{\beta}}$

Theorem 6.1. *If $E[\vec{y}] = X\vec{\beta}$, then $\hat{\beta}$ is an unbiased estiamtor for $\vec{\beta}$.*

Theorem 6.2. *If $\mathrm{Cov}(\vec{y}) = \sigma^2 I$, the covariance matrix for $\hat{\vec{\beta}}$ is given by $\sigma^2(X^TX)^{-1}$.*

Theorem 6.3. *If $\vec{\epsilon} \sim \mathrm{MVN}(\vec{0}, \sigma^2 I)$, then*

$$\hat{\vec{\beta}} \sim N_{k+1}\left(\vec{\beta}, \sigma^2(X^TX)^{-1}\right).$$

Remark. We will see that under certain conditions, $\hat{\vec{\beta}}$ is still asymptotically norm even when $\vec{\epsilon}$ is not normally distributed.

Theorem 6.4 (Gauss-Markov Theorem)**.** *If $E[\vec{y}] = X\vec{\beta}$ and $\mathrm{Cov}(\vec{y}) = \sigma^2 I$, the least-squares estimators $\hat{\beta}_j$ for $j = 0, \ldots, k$ have minimum variance among all linear unbiased estimator (*BLUE*).*

Theorem 6.5. *Gauss-Markov does not depend on normality. $\hat{\vec{\beta}}$ is BLUE regardless of the error distribution. If we add the normal errors assumptions, the LSE has minimum variance among all unbiased estimators (*UMVUE*).*

Proof. Consider the CLM

$$\vec{y} = X\vec{\beta} + \vec{e},\ \vec{e} \sim N(0, \sigma^2 I).$$

Then

$$\begin{aligned}
f(\vec{y}|\vec{\beta}, \sigma^2) &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})\right) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(\vec{y}^T\vec{y} - 2\vec{\beta}^T\vec{x}^T\vec{y} + \vec{\beta}^T X^T X\vec{\beta}\right)\right) \\
&= C(\vec{\beta}, \sigma^2) \exp\left(-\frac{1}{2\sigma^2}\vec{y}^T\vec{y} + \frac{\vec{\beta}^T}{\sigma^2}X^T\vec{y}\right) \\
&= C(\vec{\beta}, \sigma^2) \exp\left(\theta_1 T_1(\vec{y}) + \sum_{j=2}^{k+2} \theta_j T_j(\vec{y})\right),
\end{aligned}$$

where $\theta_1 = -\frac{1}{2\sigma^2}$, $T_1(\vec{y}) = \vec{y}^T\vec{y}$.

$$\begin{bmatrix} \theta_2 \\ \vdots \\ \theta_{k+2} \end{bmatrix} = \frac{1}{\sigma^2}\vec{\beta}, \quad \begin{bmatrix} T_2(\vec{y}) \\ \vdots \\ T_{k+2}(\vec{y}) \end{bmatrix} = X^T\vec{y}.$$

So the density is an exponential family with complete sufficient statistic

$$\begin{bmatrix} \vec{y}^T\vec{y} \\ X^T\vec{y} \end{bmatrix}.$$

Since $\vec{a}^T\hat{\vec{\beta}} = \vec{a}^T(X^TX)^{-1}X^T\vec{y}$ is a function of $X^T\vec{y}$ and is unbiased for $a^T\vec{\beta}$, $\vec{a}^T\hat{\vec{\beta}}$ is UMVUE for $\vec{a}^T\vec{\beta}$. Further, $s^2 = \frac{1}{n-k-1}(\vec{y} - X\hat{\vec{\beta}})^T(\vec{y} - X\hat{\vec{\beta}})$ is unbiased for σ^2 and

$$s^2 = \frac{\vec{y}^T\vec{y} - \vec{y}^TX(X^TX)^{-1}(X^T\vec{y})}{n-k-1}.$$

Thus, s^2 is UMVUE for σ^2. □

Corollary 6.6. If $E[\vec{y}] = X\vec{\beta}$ and $\text{Cov}(\vec{y}) = \sigma^2 I$, the BLUE of $\vec{a}^T\vec{\beta}$ is $\vec{a}^T\hat{\vec{\beta}}$, where $\hat{\vec{\beta}}$ is the least squares estimator $\hat{\vec{\beta}} = (X^TX)^{-1}X^T\vec{y}$.

Remark. Take $\vec{a} = (0, 0, \dots, 1, \dots, 0)^T$ to see that $\hat{\beta}_j$ is the BLUE for β_j.

Theorem 6.7.

$$\text{Var}(a^T\hat{\vec{\beta}}) = a^T \, \text{Var}(\hat{\vec{\beta}}) = \sigma^2[a^T(X^TX)^{-1}a].$$

So if the columns of X are mutually orthogonal, the elements of $\hat{\vec{\beta}}$ are uncorrelated. For a given set of explanatory variables, the values at which they are observed affect the variance (precison) of the resulting estimators.

Theorem 6.8. *The predicted values $\hat{y}$ is invariant to a full-rank linear transformation on the x's.*

Proof. Assume a full rank linear transformation of X is given by $Z = XH$, where H is square and of full rank. In the orthogonal model,

$$\hat{\mu} = X(X^TX)^{-1}X^T\vec{y} = P_{C(X)}\vec{y}.$$

In the transformation model,

$$\hat{\mu} = Z(Z^TZ)^{-1}Z^T\vec{y} = P_{C(Z)}\vec{y} = P_{C(XH)}\vec{y}.$$

It is sufficient to show $C(X) = C(XH)$. □

Theorem 6.9. *If $\vec{x} = (1, x_1, \dots, x_k)^T$ and consider rescaling the predictors $\vec{z} = (1, c_1x_1, \dots, c_kx_k)^T$, then $\hat{y} = \hat{\vec{\beta}}^T\vec{x} = \hat{\vec{\beta}}_z^T\vec{z}$, where $\hat{\vec{\beta}}_z$ is the least square estimator from the regression of y on $\vec{z}$.*

Proof. Take $H = \text{diag}(1, c_1, \dots, c_k)$. □

6.3.3 An estimation for σ^2

Theorem 6.10. *If* $E[\vec{y}] = X\vec{\beta}$ *and* $\mathrm{Cov}(\vec{y}) = \sigma^2 I$, *then*

$$E[s^2] = \sigma^2.$$

Corollary 6.11. An unbiased estimator of $\mathrm{Cov}(\hat{\vec{\beta}})$ is given by

$$\widehat{\mathrm{Cov}}(\hat{\vec{\beta}}) = s^2(X^T X)^{-1}.$$

6.4 Geometry of least squares

6.4.1 Parameter Space, data Space, and prediction space

$\vec{\beta}$ can be viewed as a single point in $\mathbb{R}^{k+1}$, which is called the parameter space. $\vec{y}$ can be viewed as a single point in $\mathbb{R}^n$, which is called the data space. The columns of $X = (\mathbb{1}_n, \vec{x}_1, \ldots, \vec{x}_k)$ including $\mathbb{1}_n$ are points in the data space. Note that because we assumed that X is of rank $k+1$, these vectors are linearly independent. The set of all possible linear combinations of the columns of X constitutes a subspace of the data space, which is called the prediction space ($\mathbb{R}^{k+1}$), which is a proper complete subspace of the data space $\mathbb{R}^n$. Note that $\vec{y}$ is not in the prediction space, is known and $E[\vec{y}]$ is in the prediction space. Multiple linear regression can be understood geometrically as the process of finding a sensible estimate of $E[\vec{y}]$ in the prediction space and then determining the vector in the parameter space that is associated with this estimate. The estimate of $E[\vec{y}]$ is denoted as $\hat{\vec{y}}$, and the associated vector in the parameter space is denoted as $\hat{\vec{\beta}}$. A reasonable geometric idea is to estimate $E[\vec{y}]$ using the point in the prediction space that is closest to $\vec{y}$. It turns out that $\hat{\vec{y}}$, the closest point in the prediction space to $\vec{y}$, can be found by noting that the difference vector $\hat{\vec{\epsilon}} = \vec{y} - \hat{\vec{y}}$ must be orthogonal (perpendicular) to the prediction space (linear analysis). Furthermore, because the prediction space is spanned by the columns of X, the point $\hat{\vec{y}}$ must be such that $\hat{\vec{\epsilon}}$ is orthogonal to the columns of X. We therefore seek $\hat{\vec{y}}$ such that

$$X^T \hat{\vec{\epsilon}} = \vec{0},$$

or

$$X^T(\vec{y} - \hat{\vec{y}}) = X^T(\vec{y} - X\hat{\vec{\beta}}) = X^T\vec{y} - X^T X\hat{\vec{\beta}} = \vec{0},$$

which implies that $X^T X\hat{\vec{\beta}} = X^T\vec{y}$.

6.5 Normal model

6.5.1 Maximum Likelihood Estimation

LS tells us how to estimate parameter $\vec{\beta}$, but doesn't help much with other parameters (or other distributional properties). ML provides a general criterion for finding estimators of unknown parameters. It provides access to a much broader class of estimators than LS at the expense of stronger distribution assumptions.

Definition 6.12. Assume a random vector $\vec{y}$ has pdf/pmf $f(\vec{y}|\vec{r})$, which depends on an unknown $\vec{r} \in \Gamma \subseteq \mathbb{R}^p$. If Γ is an open set, then $\hat{r}$ satisfies

$$\frac{\partial l(\hat{\vec{r}})}{\partial \vec{r}} = 0.$$

In CLM, the parameters are $\vec{\beta}$ and σ^2, i.e.,

$$r = \begin{bmatrix} \vec{\beta} \\ \sigma^2 \end{bmatrix}$$

Theorem 6.13. *If $\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I)$, where $X \in \mathbb{R}^{n\times(k+1)}$ of rank $k+1 < n$, the* MLE *of $\vec{\beta}$ and σ^2 are*

$$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y},$$

$$\hat{\sigma}^2 = \frac{1}{n}\|\vec{y} - p\left(\vec{y}|C(X)\right)\|^2 = \frac{1}{n}\left\|p\left(\vec{y}|C^{\perp}(X)\right)\right\|^2 = \frac{\text{SSE}}{n}.$$

Remark. A "better" estimator is

$$s^2 = \frac{\left\|\vec{y} - X\hat{\vec{\beta}}\right\|^2}{n - \text{rank}(X)},$$

even if X is not full rank. It can be shown that s^2 is the best (quadratic) unbiased estimator of σ^2 under the spherical error CLM, i.e., $\text{Var}(\vec{e}|X) = \sigma^2 I$, where normality is not required.

Example 6.14. For the intercept only model, $y_i = \beta_0 + e_i$, $X\vec{\beta} = \mathbb{1}_n \beta_0$. Then $\hat{\beta}_0 = \bar{y}$. Since $\text{rank}(X) = 1$,

$$s^2 = \frac{1}{n-1}\left\|\vec{y} - \hat{\vec{y}}\right\|^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

6.5.2 Properties of $\hat{\vec{\beta}}$ and $\hat{\sigma}^2$

Theorem 6.15. *Assume $\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I)$, then*

(a)

$$\hat{\vec{\beta}} \sim N_{k+1}\left(\vec{\beta}, \sigma^2 (X^T X)^{-1}\right).$$

(b)

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

(c)

$$\hat{\vec{\beta}} \perp\!\!\!\perp s^2.$$

6.6 Generalized Least-Squares: $\mathbf{Cov}(\vec{y}) = \sigma^2 V$

In simple linear regression, larger values of x_i may lead to larger values of $\mathrm{Var}(y_i)$. In either simple or multiple regression, if $y_1, \ldots, y_n$ accur at sequential points in time, they are typically correlated. Then we use the model

$$\vec{y} = X\vec{\beta} + \vec{\epsilon}, \quad E[\vec{y}] = X\vec{\beta}, \quad \mathrm{Var}(\vec{y}) = \Sigma = \sigma^2 V,$$

where V is known positive definite matrix.

Remark. The model above is not a Gaussian-Markov model.

Example 6.16. Uncorrelated, hetoskedastic errors

$$y_i = \beta_0 + \beta_1 x_i + e_i, \ i = 1, \ldots, n,$$

where e_i's are independent with $E[e_i] = 0$ and $\mathrm{Var}(e_i) = \sigma^2 x_i$. Then

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

$$V = \mathrm{diag}(x_1, \ldots, x_n).$$

In this case, we might estimate $\vec{\beta}$ with

$$\arg\min_\beta \sum_{i=1}^n \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{x_i} = \arg\min_\beta \sum_{i=1}^n \frac{1}{x_i}(y_i - \beta_0 - \beta_1 x_i)^2.$$

This is weighted least square (lower variance implies greater weight in the sum). Note that

$$\sum_{i=1}^n \frac{1}{x_i}(y_i - \beta_0 - \beta_1 x_i)^2 = (\vec{y} - X\vec{\beta})^T C^{-1}(\vec{y} - X\vec{\beta}).$$

A simple transformation can make the model be a Gaussian-Markov model. Since V is known p.d., then there exists a nonsingular Q such that

$$V = QQ^T.$$

Then

$$Q^{-1}\vec{y} = (Q^{-1}X)\vec{\beta} + Q^{-1}\vec{e}.$$

Then $E[Q^{-1}\vec{e}] = 0$ and $\mathrm{Var}(Q^{-1}\vec{e}) = \sigma^2 I$. Then the GLS criterion is

$$\begin{aligned}(Q^{-1}\vec{e})^T(Q^{-1}\vec{e}) &= (Q^{-1}\vec{y} - Q^{-1}X\vec{\beta})^T(Q^{-1}\vec{y} - Q^{-1}X\vec{\beta}) \\ &= (\vec{y} - X\vec{\beta})^T(QQ^T)^{-1}(\vec{y} - X\vec{\beta}) \\ &= (\vec{y} - X\vec{\beta})^T V^{-1}(\vec{y} - X\vec{\beta}).\end{aligned}$$

Remark. The GLS estimates $\vec{\beta}$ by minimizing squared statiscal distance instead of Euclidean distance, i.e., it takes into account the covariances among y_i's. Most of time, V arbitrary = "generalized least squares", V diagonal = "weighted least square".

Theorem 6.17.

$$\mathrm{SSE} = \left(Q^{-1}\vec{y} - Q^{-1}X\hat{\vec{\beta}}\right)^T \left(Q^{-1}\vec{y} - Q^{-1}X\hat{\vec{\beta}}\right) = (Q^{-1}\vec{y})^T P_{(Q^{-1}X)^\perp} Q^{-1}\vec{y}.$$

Proof. It follows from

$$\mathrm{rank}\left(P_{(Q^{-1}X)^\perp}\right) = n - \mathrm{rank}\left(Q^{-1}X\right) = n - \mathrm{rank}(X) = n - k - 1. \qquad \square$$

Theorem 6.18.

$$\hat{\vec{\beta}} \perp\!\!\!\perp \mathrm{SSE}$$

Proof. Let $\tilde{X} = Q^{-1}X$. Since

$$\hat{\vec{\beta}} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}\left(Q^{-1}\vec{y}\right) = B\left(Q^{-1}\vec{y}\right),$$

we have

$$\mathrm{SSE} = \left(Q^{-1}\vec{y}\right)^T \left(I - \tilde{X}\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}^T\right)\left(Q^{-1}\vec{y}\right) = \left(Q^{-1}\vec{y}\right)^T A \left(Q^{-1}\vec{y}\right).$$

By the distribution of quadratic form, since $B(\sigma^2 I)A = 0$, we have

$$\hat{\vec{\beta}} \perp\!\!\!\perp \mathrm{SSE}. \qquad \square$$

Theorem 6.19. *(a) The* BLUE *of $\vec{\beta}$ is*

$$\hat{\vec{\beta}} = (X^TV^{-1}X)^{-1}X^TV^{-1}\vec{y}.$$

(b)

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

(c)

$$s^2 = \frac{(\vec{y} - X\hat{\vec{\beta}})^TV^{-1}(\vec{y} - X\hat{\vec{\beta}})}{n-k-1} = \frac{\vec{y}^T\left[V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}\right]\vec{y}}{n-k-1}.$$

(d) If $\vec{e} \sim N(\vec{0}, \sigma^2 V)$, the MLE*'s of $\vec{\beta}$ and σ^2 are*

$$\hat{\vec{\beta}} = (X^TV^{-1}X)^{-1}X^TV^{-1}\vec{y}.$$

$$\hat{\sigma}^2 = \frac{(\vec{y} - X\hat{\vec{\beta}})^TV^{-1}(\vec{y} - X\hat{\vec{\beta}})}{n-k-1} = \frac{\vec{y}^T\left[V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}\right]\vec{y}}{n-k-1}.$$

6.7 Model Misspecification

What happens if we use OLS when GLS is appropriate? Suppose the true model is GLS, then the BLUE

$$\hat{\vec{\beta}} = (X^TV^{-1}X)^{-1}X^TV^{-1}\vec{y}.$$

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

But we estimate $\vec{\beta}$ with

$$\hat{\vec{\beta}}^* = (X^TX)^{-1}X^T\vec{y}.$$

Then

$$E[\vec{a}^T\hat{\vec{\beta}}^*] = \vec{a}^TE[(X^TX)^{-1}X^T\vec{y}] = \vec{a}^T\beta.$$

So $\hat{\vec{\beta}}$ and $\hat{\vec{\beta}}^*$ are unbiased. However,

$$\mathrm{Var}(\hat{\vec{\beta}}^*) = (X^TX)^{-1}X^T\mathrm{Var}(\vec{y})X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}X^TVX(X^TX)^{-1}.$$

$$\mathrm{Var}(\vec{a}^T\hat{\vec{\beta}}^*) = \sigma^2\vec{a}^T(X^TX)^{-1}X^TVX(X^TX)^{-1}\vec{a}.$$

$$\mathrm{Var}(\vec{a}^T\hat{\vec{\beta}}) = \sigma^2\vec{a}^T(X^TV^{-1}X)^{-1}\vec{a}.$$

Notice that

$$\begin{aligned}
&\mathrm{Var}(\vec{a}^T\hat{\vec{\beta}}^*) - \mathrm{Var}(\vec{a}^T\hat{\vec{\beta}})\\
&= \sigma^2\vec{a}^T(X^TX)^{-1}\left[X^TVX - (X^TX)(X^TV^{-1}X)^{-1}(X^TX)\right](X^TX)^{-1}\vec{a}\\
&= \sigma^2\vec{a}^T(X^TX)^{-1}X^T\left[V - X(X^TV^{-1}X)^{-1}X^T\right]X(X^TX)^{-1}\vec{a}\\
&= \sigma^2\vec{a}^T(X^TX)^{-1}X^TV^{\frac{1}{2}}\left[I - V^{-\frac{1}{2}}X(X^TV^{-1}X)^{-1}X^TV^{-\frac{1}{2}}\right]V^{\frac{1}{2}}X(X^TX)^{-1}\vec{a}\\
&= \sigma^2\vec{a}^T\vec{z}^TP_{C(\tilde{X})^\perp}\vec{z} \geqslant 0
\end{aligned}$$

where $z = (X^TX)^{-1}X^TV^{\frac{1}{2}}$ and $\tilde{X} = V^{-\frac{1}{2}}X$. What happens if we misspecify $E[\vec{y}]$? Consider the special case, and partition the model

$$\vec{y} = X\vec{\beta} + \vec{e} = (X_1, X_2)\begin{bmatrix}\vec{\beta}_1\\ \vec{\beta}_2\end{bmatrix} + \vec{e} = X_1\vec{\beta}_1 + X_2\vec{\beta}_2 + \vec{e}.$$

If we leave out $X_2\vec{\beta}_2$ when it should be included, we are underfitting. If we include $X_2\vec{\beta}_2$ when it doesn't belong to the true model, we are overfitting.

6.7.1 Underfitting

We write the reduced model as

$$\vec{y} = X_1\vec{\beta}_1^* + \vec{\epsilon}^*,$$

$$\mathrm{Var}(\vec{e}^*) = \sigma^2I.$$

Then the mean and Var-Cov of $\beta_1^* = (X_1^TX_1)^{-1}X_1^T\vec{y}$ are

(a)

$$E[\hat{\vec{\beta}}_1] = \vec{\beta}_1 + A\vec{\beta},$$

where $A = (X_1^T X_1)^{-1} X_1^T X_2$.

(b)

$$\mathrm{Var}(\hat{\vec{\beta}}_1^*) = \sigma^2 (X_1^T X_1)^{-1}.$$

This means that underfitting makes $\hat{\vec{\beta}}_1$ biased an amount that depends on both the excluded and included explanatory variables.

Corollary 6.20. If $X_1^T X_2 = \mathbf{0}$, that is, if columns of X_1 are orthogonal to the columns of X_2, then

$$E[\hat{\vec{\beta}}_1^*] = \beta_1.$$

Remark. $\sigma^2 (X_1^T X_1)^{-1}$ is not equal to the corresponding block of Var-Cov matrix for $\hat{\vec{\beta}}$ obtained from the full model.

Theorem 6.21. *Let* $\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y}$ *from the full model be partitioned as* $\begin{bmatrix} \hat{\vec{\beta}}_1 \\ \hat{\vec{\beta}}_2 \end{bmatrix}$ *and let* $\hat{\vec{\beta}}_1^* = (X_1^T X_1)^{-1} X_1^T \vec{y}$ *be the estimator from the reduced model. Then*

$$\mathrm{Var}(\hat{\vec{\beta}}_1) - \mathrm{Var}(\hat{\vec{\beta}}_1^*) = AB^{-1}A^T,$$

which is a p.s.d. matrix, where

$$A = (X_1^T X_1)^{-1} X_1^T X_2,$$
$$B = X_2^T X_2 - X_2^T X_1 A.$$

Proof.

$$\begin{aligned}
\mathrm{Var}(\hat{\vec{\beta}}) = \mathrm{Var} \begin{bmatrix} \vec{\beta}_1 \\ \vec{\beta}_2 \end{bmatrix} &= \sigma^2 (X^T X)^{-1} = \sigma^2 \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix}^{-1} \\
&= \sigma^2 \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}^{-1} = \sigma^2 \begin{bmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{bmatrix}
\end{aligned}$$

We know

$$H^{11} = H_{11}^{-1} + H_{11}^{-1} H_{12} B^{-1} H_{21} H_{11}^{-1},$$

where $B = H_{22} - H_{21} H_{11}^{-1} H_{12}$. Also, we know

$$\mathrm{Var}(\hat{\vec{\beta}}_1^*) = \sigma^2 (X_1^T X_1)^{-1} = \sigma^2 H_{11}^{-1}.$$

Thus,

$$\begin{aligned}
\mathrm{Var}(\hat{\vec{\beta}}_1) - \mathrm{Var}(\hat{\vec{\beta}}_1^*) &= \sigma^2 (H^{11} - H_{11}^{-1}) \\
&= \sigma^2 \left(H_{11}^{-1} H_{12} B^{-1} H_{21} H_{11}^{-1} \right) \\
&= \sigma^2 \left[(X_1^T X_1)^{-1} X_1^T X_2 B^{-1} X_2^T X_1 (X_1^T X_1)^{-1} \right] \\
&= \sigma^2 AB^{-1}A^T.
\end{aligned}$$

We can show $AB^{-1}A$ is p.s.d. □

Remark. Underfitting reduces te variance of regression parameter estimators, but introduces bias. Overfitting proces unbiased estimators, but increases the variance of these estimators.

6.8 The model in centered form

It's sometimes useful to "center" the explanatory variables when fitting the multple regression model, for example, in expressing certain hypothesis tests.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \\ &= \alpha + \beta_(x_{i1} - \overline{x}_1) + \cdots + \beta_k(x_{ik} - \overline{x}_k) + \epsilon_i,\ i \in [n], \end{aligned}$$

where

$$\alpha = \beta_0 + \beta_1\overline{x}_1 + \cdots + \beta_k\overline{x}_k.$$

In matrix form,

$$\vec{y} = \begin{bmatrix} \mathbb{1}_n & X_c \end{bmatrix} \begin{bmatrix} \alpha \\ \vec{\beta}_1 \end{bmatrix} + \vec{e},$$

where

$$\vec{\beta}_1 = (\beta_1, \ldots, \beta_k)^T,$$

$$X_c = (I - \frac{1}{n}J)X_1,$$

where $I - \frac{1}{n}J$ is sometimes called the centered matrix. The least square estimation:

$$\begin{aligned} \begin{bmatrix} \hat{\alpha} \\ \hat{\vec{\beta}} \end{bmatrix} &= \left[\begin{bmatrix} \mathbb{1}_n \\ X_c \end{bmatrix} \begin{bmatrix} \mathbb{1}_n & X_c \end{bmatrix} \right]^{-1} \begin{bmatrix} \mathbb{1}_n^T \\ X_c^T \end{bmatrix} \vec{y} \\ &= \begin{bmatrix} n & \vec{0}^T \\ \vec{0} & X_c^T X_c \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{1}_n^T \\ X_c^T \end{bmatrix} \vec{y} \\ &= \begin{bmatrix} n^{-1} & \vec{0}^T \\ \vec{0} & (X_c^T X_c)^{-1} \end{bmatrix} \begin{bmatrix} n\overline{y} \\ X_c^T \vec{y} \end{bmatrix} \\ &= \begin{bmatrix} \overline{y} \\ (X_c^T X_c)^{-1} X_c^T \vec{y} \end{bmatrix}. \end{aligned}$$

Thus, $\hat{\alpha} = \overline{y}$ and $\hat{\beta}_1 = (X_c^T X_c)^{-1} X_c^T \vec{y}$.

Remark. These estimators are the same as the usual least-squares estimators with the adjustment

$$\hat{\beta}_0 = \hat{\alpha} - \hat{\beta}_1\overline{x}_1 - \hat{\beta}_k\overline{x}_k = \overline{y} - \hat{\vec{\beta}}_1\overline{\vec{x}}.$$

Then

$$\widehat{E[y_i]} = \hat{\alpha} + \hat{\beta}_1(x_{i1} - \overline{x}_1) + \cdots + \hat{\beta}_k(x_{ik} - \overline{x}_k),$$

so the regression plane passes through the point $(\overline{y}, \overline{x}_1, \ldots, \overline{x}_k.)$

6.9 SST,SSE,SSR,R^2

Theorem 6.22. *In general,*

$$\begin{aligned}\vec{y}^T(I-P_{C(X)})\vec{y} &= \text{SSE}\\ &= (\vec{y}-X\hat{\beta})^T(\vec{y}-X\hat{\vec{\beta}})\\ &= (\vec{y}-X\hat{\vec{\beta}})^T\vec{y}-(\vec{y}-X\hat{\beta})^TX\hat{\vec{\beta}}\\ &= \vec{y}^T\vec{y}-\hat{\vec{\beta}}6TX^T\vec{y}-(X^T\vec{y}-X^TX\hat{\vec{\beta}})^T\hat{\vec{\beta}}\\ &= \vec{y}^T\vec{y}-\hat{\vec{\beta}}^TX^T\vec{y}.\end{aligned}$$

Since

$$\text{SST} = \vec{y}^T\vec{y}-n\bar{y}^2 = \vec{y}^T\vec{y}-\hat{\vec{\beta}}^TX^T\vec{y}+\hat{\vec{\beta}}^TX^T\vec{y}-n\bar{y}^2,$$

we have

$$\vec{y}^T\left(P_{C(X)}-P_{C(\mathbb{1}_n)}\right)\vec{y} = \text{SSR} = \hat{\vec{\beta}}^TX^T\vec{y}-n\bar{y}^2.$$

In the centered case,

$$\begin{aligned}\text{SSE} &= \vec{y}^T\vec{y}-(\hat{\alpha}\ \hat{\vec{\beta}}_1^T)(\mathbb{1}_n\ X_c)^T\vec{y}\\ &= \vec{y}^T\vec{y}-n\bar{y}^2-\hat{\beta}_1^TX_c^T\vec{y}\\ &= \text{SST}-\text{SSR}\,.\end{aligned}$$

So

$$\text{SSR} = \hat{\beta}_1^TX_c^T\vec{y} = \hat{\beta}_1X_c^TX_c(X_c^TX_c)^{-1}X_c^T\vec{y} = \hat{\beta}_1^TX_c^TX_c\hat{\beta}_1 = (X_c\hat{\vec{\beta}}_1)^T(X_c\hat{\vec{\beta}}_1).$$

Definition 6.23.

SST: corrected total sum of squares,

which quantifies total variability in the data.

SSR: regression sum of squares,

which quantifies the variability in the data that can be explained by the regression terms.

Definition 6.24. The proportion of the total SS due to the regression,

$$R^2 := \frac{\text{SSR}}{\text{SST}},$$

which is called the *coefficient of determination.* It is the sample estimation of the squared multiple correlation coefficient.

Theorem 6.25 (Facts)**.** *(a)*

$$0 \leqslant R^2 \leqslant 1.$$

(b) If all the $\hat{\beta}_j$'s were 0, except for $\hat{\beta}_0$, R^2 would be 0. (This event has probability 0 for continuous data). If all the y value fell on the fitted surface, that is, if $y_i = \hat{y}_i$ for $i = 1, \ldots, n$, then $R^2 = 1$.

(c)

$$R = \sqrt{R^2} = r_{\vec{y}\hat{\vec{y}}},$$

which is the sample correlation between y_i's and the $\hat{y}_i$'s.

(d) Adding a variable x to the model cannot decrease the value of R^2.

(e) If $\beta_1 = \cdots = \beta_k = 0$, then

$$E[R^2] = \frac{k}{n-1}.$$

Note that the $\hat{\beta}_j$'s will not be 0 when the β_j's are 0.

Proof. Note that $\vec{y} \sim N_n(\beta_0 \mathbb{1}_n, \sigma^2 I)$. We know

$$\frac{\text{SSE}}{\sigma^2} = \vec{y}^T \left(\frac{I - P_{C(X)}}{\sigma^2} \right) \vec{y} \sim \chi^2_{n-k-1},$$

$$\frac{\text{SSR}}{\sigma^2} = \vec{y}^T \left(\frac{P_{C(X)} - P_{L(\mathbb{1}_n)}}{\sigma^2} \right) \vec{y} \sim \chi^2_{k},$$

and

$$\frac{\text{SSE}}{\sigma^2} \perp\!\!\!\perp \frac{\text{SSR}}{\sigma^2}.$$

Now reall that if $W \sim \chi^2_m$, $V \sim \chi^2_n$ and $W \perp\!\!\!\perp V$, then

$$W + V \perp\!\!\!\perp \frac{W}{W+V}.$$

Then

$$E[W] = E\left[\frac{W}{W+V}(W+V) \right] = E\left[\frac{W}{W+V} \right] E[W+V].$$

So

$$E\left[\frac{W}{W+V} \right] = \frac{E[W]}{E[W+V]}.$$

Thus,

$$E[R^2] = \frac{E[\text{SSR}]}{E[\text{SSR} + \text{SSE}]} = \frac{k}{n-1}. \qquad \square$$

(f) R^2 cannot be partitioned into k components, each of which is uniquely attributable to an x_j, unless the x's are x's are mutually orthogonal, that is,

$$\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{im} - \bar{x}_m) = 0, \textit{ for } \; j \neq m.$$

(g) R^2 is invariant to a full rank linear transformation of x's and to a scale change on y, but not invariant to a joint linear transformation or $[y, x]$.

(h)

$$R = \cos(\theta),$$

where θ is the angle between mean-corrected $\vec{y}$ and mean-corrected $\hat{y}$.

$$R^2 = \frac{\left\|\hat{\vec{y}} - \bar{y}\mathbb{1}_n\right\|^2}{\left\|\vec{y} - \bar{y}\mathbb{1}_n\right\|^2} = \frac{(X\hat{\beta})^T\vec{y} - n\bar{y}^2}{\vec{y}^T\vec{y} - n\bar{y}^2} = \cos^2\theta.$$

Also,

$$\text{SST} = \left\|\vec{y} - \bar{y}\mathbb{1}_n\right\|^2 = \left\|\vec{y} - \hat{\vec{y}}\right\|^2 + \left\|\hat{\vec{y}} - \bar{y}\mathbb{1}_n\right\|^2.$$

We can see that if k is a relatively large fraction of n, it is possible to have a large value of R^2 that is not meaningful. In this case, x's that do not contribute to predicting y may appear to do so in a particular example, and the estimated regression equation may not be a useful estimator of the population model. To correct for this tendency, Ezekeil proposed a bias-corrected of R^2.

$$R_a^2 = \frac{\left(R^2 - \frac{k}{n-1}\right)(n-1)}{n-k-1} = \frac{(n-1)R^2 - k}{n-k-1}.$$

6.10 Examples

Example 6.26. In the linear model $\vec{y} = X\vec{\beta} + \vec{e}, e \sim N(0, \Sigma)$, show that the BLUE of $\vec{\beta}$ is equal to the OLS estimator if and only if there exists a nonsingular matrix F such that

$$\Sigma X = XF.$$

In other words, show that $(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\vec{y} = (X^TX)^{-1}X^T\vec{y}$ if and only if $\Sigma X = XF$.

Proof. "⇒".

$$\begin{aligned}\left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1} = (X^TX)^{-1}X^T &\Longrightarrow X^TX\left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1} = X^T\\ &\Longrightarrow \Sigma^{-1}X\left(X^T\Sigma^{-1}X\right)^{-1}X^TX = X\\ &\Longrightarrow X\left(X^T\Sigma^{-1}X\right)^{-1}X^TX = \Sigma X.\end{aligned}$$

"⇐".

$$\begin{aligned}\Sigma^{-1}XF = X &\Longrightarrow F^TX^T\Sigma^{-1} = X^T\\ &\Longrightarrow F^TX^T\Sigma^{-1}X = X^TX\\ &\Longrightarrow \left(X^T\Sigma^{-1}X\right)^{-1}F^{-T} = \left(X^TX\right)^{-1}\\ &\Longrightarrow \left(X^T\Sigma^{-1}X\right)^{-1}F^{-T}X^T = \left(X^TX\right)^{-1}X^T.\end{aligned}$$

Since $XF^{-1} = \Sigma^{-1}X$, we have $F^{-T}X^T = X^T\Sigma^{-1}$. □

Chapter 7

Multiple Regression: Tests of Hypotheses and Confidence Intervals

We will assume throughout the chapter that

$$\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I),$$

where X is $n \times (k+1)$ of rank $k+1 < n$.

7.1 Test on a subset of the β's

Testing linear hypothesis amounts to putting constraints on the model space and comparing the constrained model vs the unconstrained model. Assume model

$$\vec{y} = \vec{\mu} + \vec{e},$$

where

$$\vec{\mu} = X\vec{\beta} \in C(X), \quad \vec{e} \sim N(0, \sigma^2 I).$$

Q: Is $\vec{\mu} \in V_0 \subseteq C(X)$? For example, is $\mu \subseteq C(X_0)$, where X_0 consists of a subset of the columns of X. Without loss of generality, arrange the linear model so the terms that we want to test appear last in the linear predictor:

$$\vec{y} = X\vec{\beta} + \vec{e} = X_1\vec{\beta}_1 + X_2\vec{\beta}_2 + \vec{e},$$

where $X_1 \in \mathbb{R}^{n\times(k+1-h)}$ and $X_2 \in \mathbb{R}^{n\times h}$. Thus, $\vec{\beta_1} = (\beta_0, \beta_1, \ldots, \beta_{k-h})^T$ and $\vec{\beta_2} = (\beta_{k-h+1}, \ldots, \beta_k)^T$. Under $H_0 : \vec{\beta}_2 = \vec{0}$,

$$\vec{y} = X_1\vec{\beta}_1^*, \quad \vec{e}^* \sim N(0, \sigma^2 I).$$

The problem is to test

$$H_0 : \vec{\mu} \in C(X_1) \text{ vs } H_1 : \vec{\mu} \notin C(X_1),$$

under the maintained hypothesis that

$$\vec{C}(X) = C\left([X_1, X_2]\right).$$

So we need a test statistic whose magnitude measures the strength of the evidence agaist H_0. If the test statistic is "large enough", we reject H_0. How large? It depends on the predictors. An α-level test rejects if the p-value $< \alpha$ for some pre-specify α. Note: for hypothesis testing and confidence intervals, we will need normality as well. Under RM, $\mu \subseteq C(X_1) \subseteq C(X)$, so

$$\text{RM true} \Longrightarrow \text{FM true.}$$

The least square estimators of $\vec{\mu}$, $P_{C(X_1)}\vec{y}$ and $P_{C(X)}\vec{y}$ estimate the same thing. Then

$$P_{C(X)}\vec{y} - P_{C(X_1)}\vec{y} = \left(P_{C(X)} - P_{C(X_1)}\right)\vec{y}$$

should be small under $H_0 : \vec{\mu} \in C(X_1)$. Measure the size of $\left(P_{C(X)} - P_{C(X_1)}\right)\vec{y}$ with

$$\left\|\left(P_{C(X)} - P_{C(X_1)}\right)\vec{y}\right\|^2 = \vec{y}^T\left(P_{C(X)} - P_{C(X_1)}\right)\vec{y}.$$

Note

$$\begin{aligned} E\left[\vec{y}^T\left(P_{C(X)} - P_{C(X_1)}\right)\vec{y}\right] &= \sigma^2 \dim\left(C^{\perp}(X_1) \cap C(X)\right) + \vec{\mu}^T\left(P_{C(X)} - P_{C(X_1)}\right)\vec{\mu} \\ &= \sigma^2 h + \left(P_{C(X)}\vec{\mu} - P_{C(X_1)}\vec{\mu}\right)^T\left(P_{C(X)}\vec{\mu} - P_{C(X_1)}\vec{\mu}\right). \end{aligned}$$

Under H_0, $\vec{\mu} \in C(X_1) \subseteq C(X)$, then

$$P_{C(X)}\vec{\mu} - P_{C(X_1)}\vec{\mu} = \vec{\mu} - \vec{\mu} = \vec{0}.$$

But under H_1, $\vec{\mu} \in C(X)$, but $\vec{\mu} \notin C(X_1)$, so

$$P_{C(X)}\vec{\mu} - P_{C(X_1)}\vec{\mu} = \vec{\mu} - \vec{\mu}_0 \neq 0.$$

In other words,

$$E\left[\vec{y}^T\left(P_{C(X)} - P_{C(X_1)}\right)\vec{y}\right] = \left\{\begin{array}{cc} \sigma^2 h, & \text{under } H_0 \\ \sigma^2 h + \|\vec{\mu} - \vec{\mu}_0\|^2, & \text{under } H_1 \end{array}\right.$$

This suggests the test statistic

$$\frac{\left\|\hat{\vec{y}} - \hat{\vec{y}}_0\right\|^2/h}{\sigma^2}\left\{\begin{array}{cc} \approx 1, & \text{under } H_0 \\ > 1, & \text{under } H_1 \end{array}\right. .$$

Typically, σ^2 is unnknown and must be estimated under FM, the appropriate estimator is

$$s^2 = \frac{\left\|\vec{y} - \hat{\vec{y}}\right\|^2}{n-k-1}.$$

(s^2 is valid under both H_0 and H_1.) So the test statistic is

$$\frac{\left\|\hat{\vec{y}} - \hat{\vec{y}}_0\right\|^2/h}{\left\|\vec{y} - \hat{\vec{y}}^2\right\|/(n-k-1)}\left\{\begin{array}{cc} \approx 1, & \text{under } H_0 \\ > 1, & \text{under } H_1 \end{array}\right. .$$

Theorem 7.1. *Suppose* $\vec{y} \sim N_n(X\hat{\beta}, \sigma^2 I)$*, where* $X\vec{\beta} = X_1\beta_1 + X_2\beta_2$*. Let*

$$\begin{aligned} \hat{\vec{y}} &= p(\vec{y}|C(X)) = P_{C(X)}\vec{y}, \\ \hat{\vec{y}}_0 &= p(\vec{y}|C(X_1)) = P_{C(X_1)}\vec{y}, \\ \vec{\mu}_0 &= p(\vec{\mu}|C(X_1)) = P_{C(X_1)}\vec{\mu} = P_{C(X_1)}(X\vec{\beta}). \end{aligned}$$

Then

(a)

$$\frac{1}{\sigma^2}\left\|\vec{y} - \hat{\vec{y}}\right\|^2 = \frac{1}{\sigma^2}\vec{y}^T\left(I - P_{C(X)}\right)\vec{y} \sim \chi^2_{n-k-1}.$$

(b)

$$\frac{1}{\sigma^2}\left\|\hat{\vec{y}} - \hat{\vec{y}}_0\right\|^2 = \frac{1}{\sigma^2}\vec{y}^T\left(P_{C(X)} - P_{C(X_1)}\right)\vec{y} \sim \chi^2_h(\lambda_1),$$

where

$$\begin{aligned} \lambda_1 &= \frac{1}{2\sigma^2}\left\|\left(P_{C(X)} - P_{C(X_1)}\right)\vec{\mu}\right\|^2 = \frac{1}{2\sigma^2}\left\|\vec{\mu} - \vec{\mu}_0\right\|^2 \\ &= \frac{1}{2\sigma^2}\left\|P_{C(X)}(X_1\vec{\beta}_1 + X_2\vec{\beta}_2) - P_{C(X_1)}(X_1\vec{\beta}_1 + X_2\vec{\beta}_2)\right\|^2 \\ &= \frac{1}{2\sigma^2}\left\|X_1\vec{\beta}_1 + X_2\vec{\beta}_2 - X_1\vec{\beta}_1 - P_{C(X_1)}X_2\vec{\beta}_2\right\|^2 \\ &= \frac{1}{2\sigma^2}\left\|P_{C^\perp(X_1)}X_2\vec{\beta}_2\right\|^2 \\ &= \frac{1}{2\sigma^2}\vec{\beta}_2^T X_2^T P_{C^\perp(X_1)}X_2\vec{\beta}_2. \end{aligned}$$

(c)

$$\frac{1}{\sigma^2}\left\|\vec{y} - \hat{\vec{y}}\right\|^2 \perp\!\!\!\perp \frac{1}{\sigma^2}\left\|\hat{\vec{y}} - \hat{\vec{y}}_0\right\|^2.$$

Proof. (c)

$$\begin{aligned} \left\|\vec{y} - \hat{\vec{y}}\right\|^2 &= \left\|p(\vec{y}|C^\perp(X))\right\|^2, \\ \left\|\vec{y} - \hat{\vec{y}}_0\right\|^2 &= \left\|p(\vec{y}|C(X) \cap C^\perp(X_1))\right\|^2, \end{aligned}$$

and

$$C(X) \cap X^\perp(X_1) \subseteq C(X),$$

so it is mutually orthogonal projections. □

Theorem 7.2. *Let* $\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I)$ *and define an* F *statistic as follows:*

$$\begin{aligned} F &= \frac{\left\|\hat{\vec{y}} - \hat{\vec{y}}_0\right\|^2/h}{s^2} = \frac{\vec{y}^T\left(P_{C(X)} - P_{C(X_1)}\right)\vec{y}/h}{\vec{y}^T\left(I - P_{C(X)}\right)\vec{y}/(n-k-1)} \\ &= \begin{cases} F_{n,n-k-1}, & \text{under } H_0 \\ F_{n,n-k-1}(\lambda_1), & \text{under } H_1 \end{cases}. \end{aligned}$$

Proof. Under H_0, $\lambda_1 = 0$. □

Remark. The α-level F-test for $H_0 : \vec{\beta}_2 = \vec{0}$ vs $H_1 : \vec{\beta}_2 \neq \vec{0}$ is to reject H_0 if $F > F_{n,n-k-1,1-\alpha}$, where $F_{n,n-k-1,1-\alpha}$ is the upper α^{th}-quantile of $F_{n,n-k-1}$. Equivalently, reject H_0 if $P(X > F) < \alpha$, where $X \sim F_{n,n-k-1}$.

Theorem 7.3. *The F test is always a right tail test.*

Definition 7.4.

$$\begin{aligned}
\left\|\hat{\vec{y}} - \hat{\vec{y}}_0\right\|^2 &= \left\|\vec{y} - \hat{\vec{y}}_0\right\|^2 - \left\|\vec{y} - \hat{\vec{y}}\right\|^2 \\
&= \mathrm{SSE}_{\mathrm{RM}} - \mathrm{SSE}_{\mathrm{FM}} \\
&= \mathrm{SSR}_{\mathrm{FM}} - \mathrm{SSR}_{\mathrm{RM}} \\
&= \mathrm{SS}(\vec{\beta}_2|\vec{\beta}_1) \\
&= \text{“extra” regression sum of squares due to } \beta_2 \text{ after counting for } \beta_1 \\
&= \text{type I SS.}
\end{aligned}$$

The difference in df_E's (dimension of error space) is

$$(n - (k + 1 - h)) - (n - (k + 1)) = h.$$

Then the test statistic can be written as

$$F = \frac{\frac{\mathrm{SSE}_{\mathrm{RM}} - \mathrm{SSE}_{\mathrm{FM}}}{\mathrm{df}_E(\mathrm{RM}) - \mathrm{df}_E(\mathrm{FM})}}{\mathrm{SSE}_{\mathrm{FM}}/\mathrm{df}_E(\mathrm{FM})}.$$

The test is summerized in an AVOVA table.

ANOVA Table

Source of Variation	df	Sum of Squares	Mean Squares	F
Due to $\vec{\beta}_2$ adjusted for $\vec{\beta}_1$	h	$\mathrm{SS}(\vec{\beta}_2\vert\vec{\beta}_1) = \vec{y}^T P_{C(X) \cap C^{\perp}(X_1)} \vec{y}$	$\mathrm{MS}(\vec{\beta}_2\vert\vec{\beta}_1) = \mathrm{SS}(\vec{\beta}_2\vert\vec{\beta}_1)/h$	$\frac{\mathrm{MS}(\vec{\beta}_2\vert\beta_1)}{\mathrm{MSE}}$
Error	$n - k - 1$	$\mathrm{SSE} = \vec{y}^T \left(I - P_{C(X)}\right) \vec{y}$	$\mathrm{MSE} = \mathrm{SSE}/(n - k - 1)$	
Corrected Total	$n - 1$	$\mathrm{SST} = \sum_i (y_i - \overline{y})^2$		

An additional column is sometimes added to the AVOVA table for expected mean squares $E(\mathrm{MS})$.

$$\begin{aligned}
E\left[\mathrm{MS}(\vec{\beta}_2|\vec{\beta}_1)\right] &= \frac{1}{h} E\left[\mathrm{SS}(\vec{\beta}_1|\vec{\beta}_1)\right] \\
&= \frac{\sigma^2}{h} E\left[\frac{\mathrm{SS}(\vec{\beta}_2|\vec{\beta}_1)}{\sigma^2}\right] \\
&= \frac{\sigma^2}{h}(h + 2\lambda_1) \\
&= \sigma^2 + \frac{1}{h}\|\mu - \mu_0\|^2.
\end{aligned}$$

Also, $E[\mathrm{MSE}] = \sigma^2$. Thus,

$$\frac{E[\mathrm{MS}(\vec{\beta}_2|\vec{\beta}_1)]}{E[\mathrm{MSE}]} = 1, \text{under } H_0.$$

Note any mean square can be regarded as an estimate of its expected value. So MSE estimates σ^2 (always), and $MS(\vec{\beta}_2|\vec{\beta}-1)$ estimates σ^2 under H_0 and estimates $\sigma^2+c,\ c>0$ under H_1. F behaves as

$$F\left\{\begin{array}{ll}\approx 1, & \text{under } H_0\\ = 1, & \text{under } H_1\end{array}\right. .$$

7.2 Test of Overall Regression

We note in last chapter that the probblem associated with both overfitting and underfitting motivates us to seek an optimal model. Hypothesis testing is a formal tool for, among other things, choosing between a residual model and an associated full model. The hypothesis H_0, expresses the reduced model in terms of values of a subset of the β_j's in $\vec{\beta}$. The alternative hypothesis, H_1, is associated with the full model. Partition $\vec{\beta}$ so that

$$\vec{\beta}_1=\beta_0,$$

$$\vec{\beta}_2=(\beta_1,\ldots,\beta_k)^T.$$

Then

$$\vec{y}=X_1\vec{\beta}_1+X_2\vec{\beta}_2+\vec{e}.$$

In this case, $H_0:\vec{\beta}_2=0$ is equivalent to

$$H_0:\beta_1=\cdots=\beta_k=0,$$

which says that explanatory variables $x_1,\ldots,x_k$ have no linear effect (don't predict) the response. This is called the overall test. Under H_0,

$$\hat{y}_0=p\left(\vec{y}|C(X_1)\right)=p\left(\vec{y}|L(\mathbb{1}_n)\right)=\bar{y}\mathbb{1}_n,$$

and $h=k$. So the numerator of the F-statistic is

$$\begin{aligned}\frac{1}{k}\vec{y}^T\left(P_{C(X)}-P_{L(\mathbb{1}_n)}\right)\vec{y}&=\frac{1}{k}\left(\vec{y}^TP_{L(\mathbb{1}_n)}\vec{y}-\vec{y}^TP_{L(\mathbb{1}_n)}\vec{y}\right)\\&=\frac{1}{k}\left(P_{C(X)}\vec{y}\right)^T\vec{y}-\vec{y}^TP^T_{L(\mathbb{1}_n)}P_{L(\mathbb{1}_n)}\vec{y}\\&=\frac{1}{k}\left(\hat{\beta}^TX^T\vec{y}-n\bar{y}^2\right)\\&=\frac{\text{SSR}}{k}=\text{MSR}\,.\end{aligned}$$

So the test statistic for the overall regression is

$$F=\frac{\text{SSR}/k}{\text{SSE}/(n-k-1)}=\frac{\text{MSR}}{\text{MSE}}\sim\left\{\begin{array}{cc}F_{k,n-k-1}, & \text{under } H_0:\beta_1=\cdots=\beta_k=0\\ F_{k,n-k-1}(\lambda_1), & \text{under } H_1.\end{array}\right. ,$$

where

$$\begin{aligned}
\lambda_1 &= \frac{1}{2\sigma^2}\left\|\left(P_{C(X)} - P_{C(\mathbb{1}_n)}\right)\left(X_1\vec{\mathbb{1}}_n + X_2\beta_2\right)\right\|^2 \\
&= \frac{1}{2\sigma^2}\left\|\left(P_{C(X)} - P_{C(\mathbb{1}_n)}\right)X_2\beta_2\right\|^2 \\
&= \frac{1}{2\sigma^2}\left\|X_2\vec{\beta}_2 - P_{C(\mathbb{1}_n)}X_2\vec{\beta}_2\right\|^2 \\
&= \frac{1}{2\sigma^2}\vec{\beta}_2^T X_2^T P_{L^\perp(\mathbb{1}_n)} X_2\vec{\beta}_2 \\
&\left(= \frac{1}{2\sigma^2}\vec{\beta}_2^T X_c^T X_c\vec{\beta}_2.\right)
\end{aligned}$$

ANOVA Table

Source of Variation	df	Sum of Squares	Mean Squares	F
Due to $\vec{\beta}_2$	k	$\mathrm{SSR} = \hat{\vec{\beta}}_1^T\vec{y} = \hat{\vec{\beta}}^T X^T\vec{y} - n\bar{y}^2$	SSR/k	$\frac{\mathrm{MSR}}{\mathrm{MSE}}$
Error	$n-k-1$	$\mathrm{SSE} = \vec{y}^T\left(I - P_{C(X)}\right)\vec{y}$	$\mathrm{SSE}/(n-k-1)$	
Total	$n-1$	$\mathrm{SST} = \sum_i (y_i - \bar{y})^2$		

7.3 F test in terms of R^2

Theorem 7.5. *The F statistic for testing $H_0 : \vec{\beta}_2 = 0$ in the full rank linear model $\vec{y} = X_1\vec{\beta}_1 + X_2\vec{\beta}_2 + \vec{e}$ can be written in terms of R^2 as*

$$F = \frac{\left(R^2_{FM} - R^2_{\mathrm{RM}}\right)/h}{(1 - R^2_{FM})/(n-k-1)}.$$

Proof. Exercise. □

Corollary 7.6. The F statistic for overall regression can be written in terms of R^2 as

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}.$$

Proof. For this hypothesis, $h = \dim(\vec{\beta}_2) = k$. Then it is sufficient to show that $R_{\mathrm{RM}} = 0$. The reduced model is $\vec{y} = \beta_0\mathbb{1}_n + \vec{e}$. Then $(X\hat{\vec{\beta}})_{\mathrm{RM}} = \bar{y}\mathbb{1}_n$. So

$$R^2_{\mathrm{RM}} = \frac{\left\|\hat{\vec{y}} - \bar{y}\mathbb{1}_n\right\|^2}{\left\|\vec{y} - \bar{y}\mathbb{1}_n\right\|^2} = 0.$$ □

7.4 The General Linear Hypothesis Tests for $H_0 : C\vec{\beta} = \vec{0}$ and $C\vec{\beta} = \vec{t}$

7.4.1 The test for $H_0 : C\beta = \vec{0}$

The hypothesis $H_0 : C\vec{\beta} = \vec{0}$, where C is a known $q \times (k+1)$ coefficient matrix of rank $q \leqslant k+1$, is known as the general linear hypothesis. The alternative hypothesis is $H_1 : C\vec{\beta} \neq \vec{0}$.

Example 7.7. The hypothesis $H_0 : \vec{\beta_2} = 0$ in the overall regression can be expressed in the form $H_0 : C\vec{\beta} = \vec{0}$ as follows

$$H_0 : C\vec{\beta} = (\vec{0}, I_k) \begin{bmatrix} \vec{\beta_1} \\ \vec{\beta_2} \end{bmatrix} = \vec{\beta_2} = \vec{0}.$$

Similarly, the hypothesis $H_0 : \vec{\beta_2} = \vec{0}$ in the test on a subset of $\vec{\beta}$ can be expressed in the form $H_0 : C\vec{\beta} = \vec{0}$:

$$H_0 : C\vec{\beta} = (0, I_h) \begin{bmatrix} \vec{\beta_1} \\ \vec{\beta_2} \end{bmatrix} = \vec{\beta_2} = \vec{0}.$$

Example 7.8. $H_0 : \beta_1 = \beta_2$ can be expressed as

$$H_0 : \vec{\beta_1} - \vec{\beta_2} = 0 \Longleftrightarrow C^T\vec{\beta} = 0,$$

where $C = (0, 1, -1, 0 \cdots, 0)^T$.

Example 7.9. $H_0 : \beta_1 = \cdots = \beta_k$ when $k = 4$ can be written as $C\vec{\beta} = 0$, where

$$C = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix},$$

or

$$C = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{bmatrix}.$$

Example 7.10. The formulation $H_0 : C\vec{\beta} = \vec{0}$ also allows for more general hypotheses such as

$$H_0 : 2\beta_1 - \beta_2 = \beta_2 - 2\beta_3 + 3\beta_4 = \beta_1 - \beta_4 = 0,$$

which can be expressed as follows:

$$H_0 : \begin{bmatrix} 0 & 2 & -1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 3 \\ 0 & 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Remark. $\text{rank}(C) = q$ ensures that we don't have any redundant hypothesis in $H_0 : C\vec{\beta} = t$.

Remark. The test statistic for $H_0 : C\vec{\beta} = \vec{0}$ is based on comparing $C\hat{\vec{\beta}}$ to its null value 0, using squared statistical distance of the form

$$\begin{aligned} Q &= \left(C\hat{\vec{\beta}} - E_0\left[C\hat{\vec{\beta}}\right]\right)^T \left(\widehat{\text{Var}_0}\left(C\hat{\vec{\beta}}\right)\right)^{-1} \left(C\hat{\vec{\beta}} - E_0\left[C\hat{\vec{\beta}}\right]\right) \\ &= \left(C\hat{\vec{\beta}}\right)^T \left(\widehat{\text{Var}_0}\left(C\hat{\vec{\beta}}\right)\right)^{-1} C\hat{\vec{\beta}}, \end{aligned}$$

where 0 indicates that expectation is taken w.r.t. the null model. Since

$$\hat{\vec{\beta}} \sim N_{k+1}\left(\vec{\beta}, \sigma^2 (X^T X)^{-1}\right),$$

we have

$$C\hat{\vec{\beta}} \sim N_q(C\vec{\beta}, \sigma^2 C(X^T X)^{-1} C^T).$$

Estimating σ^2 with $s^2 = \frac{\text{SSE}}{n-k-1}$, we obtain

$$\widehat{\text{Var}}\left(C\hat{\vec{\beta}}\right) = s^2 C(X^T X)^{-1} C^T.$$

So

$$\begin{aligned} Q &= \left(C\hat{\vec{\beta}}\right)^T \left[C(X^T X)^{-1} C^T\right]^{-1} C\vec{\beta} \\ &= \frac{\left(C\hat{\vec{\beta}}\right)^T \left[C(X^T X)^{-1} C^T\right]^{-1} C\hat{\vec{\beta}}}{\text{SSE}/(n-k-1)}. \end{aligned}$$

To use Q as a test statistic, we need its distribution. We denote the sum of squares due to $C\vec{\beta}$ (due to hypothesis) as SSH, i.e.,

$$\text{SSH} = \left(C\hat{\vec{\beta}}\right)^T \left[C(X^T X)^{-1} C^T\right]^{-1} C\hat{\vec{\beta}} = \text{SS due to } H_0.$$

Theorem 7.11. *If $\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I)$ and C is $q \times (k+1)$ of rank $q \leqslant k+1$, then*

(a)

$$C\hat{\beta} \sim N_q[C\vec{\beta}, \sigma^2 C(X^T X)^{-1} C^T].$$

(b)

$$\frac{\text{SSH}}{\sigma^2} = \frac{\left(C\hat{\vec{\beta}}\right)^T \left[C(X^T X)^{-1} C^T\right]^{-1} C\hat{\vec{\beta}}}{\sigma^2} \sim \chi^2(q, \lambda),$$

where

$$\lambda = \frac{(C\beta)^T \left[C(X^T X)^{-1} C^T\right]^{-1} C\vec{\beta}}{2\sigma^2}.$$

(c)

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-k-1).$$

(d)

$$\text{SSH} \perp\!\!\!\perp \text{SSE}.$$

Proof. (b) Since

$$C\hat{\vec{\beta}} \sim N_{k+1}[\vec{\beta}, \sigma^2 C(X^T X)^{-1} C^T],$$

the result follows by the distribution of Quadratic form.

(d) Since $\hat{\vec{\beta}} \perp\!\!\!\perp$ SSE, we have SSH $= f(\vec{\beta}) \perp\!\!\!\perp$ SSE. □

Theorem 7.12. *Let $\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I)$ and define the statistic*

$$F = \frac{Q}{q} = \frac{\mathrm{SSH}\,/q}{\mathrm{SSE}\,/(n-k-1)} = \frac{(C\vec{\beta})^T[C(X^TX)^{-1}C^T]^{-1}C\hat{\vec{\beta}}/q}{\mathrm{SSE}\,/(n-k-1)},$$

where C is $q \times (k+1)$ of rank $q \leqslant k+1$ and $\hat{\vec{\beta}} = (X^TX)^{-1}X^T\vec{y}$. Then

$$F \sim \left\{ \begin{array}{c} F_{q,n-k-1}, \text{ under } H_0 : C\vec{\beta} = \vec{0} \\ F_{q,n-k-1}(\lambda), \text{ under } H_1 : C\vec{\beta} \neq \vec{0} \end{array} \right. ,$$

where

$$\lambda = \frac{(C\vec{\beta})^T[C(X^TX)^{-1}C^T]^{-1}C\vec{\beta}}{2\sigma^2}.$$

The F test for $H_0 : C\beta = \vec{0}$ is usually called the general linear hypothesis test. The degrees of freedom q is the number of linear combinations in $C\vec{\beta}$.

Theorem 7.13. *The F test in Theorem 7.12 for the general linear hypothesis $H_0 : C\vec{\beta} = \vec{0}$ is a full-reduced-model test.*

Proof. Under H_0, $C\vec{\beta} = 0$,

$$C(X^TX)^{-1}X^TX\vec{\beta} = \vec{0},$$

$$C(X^TX)^{-1}X^T\mu = 0,$$

$$T^T\vec{\mu} = \vec{0},$$

where $T = X(X^TX)^{-1}C^T$. Then under H_0, $\mu = X\vec{\beta} \in C(X)$ and $\mu \in C^{\perp}(T)$. Then

$$\mu \in C(X) \cap C^{\perp}(T) =: V_0 \subseteq C(X).$$

Since under $H_1 : C\vec{\beta} \neq \vec{0}$, $\mu \in C(X)$, but $\mu \notin V_0$. So the two hypothesis are nested. Test for nested models is of the form

$$F = \frac{\vec{y}^T\left(P_{C(X)} - P_{C(X_1)}\right)\vec{y}/h}{\mathrm{SSE}/(n-k-1)}.$$

Replace $P_{C(X_1)}$ with $P_{V_0} = P_{C(X)} - P_{C(T)}$, and replace h with $\dim(C(X)) - \dim(V_0)$, which is the reduction of dimension of the model space in moving from FM to RM. Note

$$\operatorname{rank}(T) = \operatorname{rank}(T^T) \geqslant \operatorname{rank}(T^TX) = \operatorname{rank}\left(C(X^TX)^{-1}X^TX\right) = \operatorname{rank}(C) = q.$$

Also

$$\operatorname{rank}(T) = \operatorname{rank}(T^TT) = \operatorname{rank}\left(C(X^TX)^{-1}X^TX(X^TX)^{-1}C^T\right) = \operatorname{rank}\left(C(X^TX)^{-1}C^T\right) \leqslant q.$$

So

$$\operatorname{rank}(T) = q = \dim\left(C(X)\right) - \dim(V_0).$$

Thus, the FM vs RM F statistic is

$$F = \frac{\vec{y}\left(P_{C(X)} - P_{V_0}\right)\vec{y}/q}{\mathrm{SSE}/(n-k-1)} = \frac{\vec{y}\left(P_{C(X)} - \left(P_{C(X)} - P_{C(T)}\right)\right)\vec{y}/q}{\mathrm{SSE}/(n-k-1)} = \frac{\vec{y}^T P_{C(T)}\vec{y}/q}{\mathrm{SSE}/(n-k-1)},$$

where

$$\begin{aligned}\vec{y}P_{C(T)}\vec{y} &= \vec{y}^T T(T^T T)^{-1} T^T \vec{y} \\ &= \vec{y}^T X(X^T X)^{-1} C^T \left[C(X^T X)^{-1} X^T X (X^T X)^{-1} C^T\right]^{-1} C(X^T X)^{-1} X^T \vec{y} \\ &= \hat{\vec{\beta}} C^T \left[C(X^T X)^{-1} C^T\right]^{-1} C\hat{\vec{\beta}}.\end{aligned}$$

Thus, this is the test for the general linear hypothesis. □

7.4.2 The test for $H_0 : C\vec{\beta} = \vec{t}$

We assume that the system of equations $C\vec{\beta} = \vec{t}$ is consistent, that is, rank(C) =rank$(C, \vec{t})$.

Theorem 7.14. *If $\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I)$ and C is $q \times (k+1)$ of rank $q \leqslant k+1$, then*

(a)

$$C\hat{\vec{\beta}} - \vec{t} \sim N_q(C\vec{\beta} - \vec{t}, \sigma^2 C(X^T X)^{-1} C^T],$$

(b)

$$\frac{\mathrm{SSH}}{\sigma^2} = \frac{(C\hat{\vec{\beta}} - \vec{t})^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\vec{\beta}} - \vec{t})}{\sigma^2} \sim \chi^2(q, \lambda),$$

where

$$\lambda = \frac{(C\beta - \vec{t})^T [C(X^T X)^{-1} C^T](C\vec{\beta} - \vec{t})}{2\sigma^2}.$$

(c)

$$\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-k-1),$$

(d)

$$\mathrm{SSH} \perp\!\!\!\perp \mathrm{SSE}.$$

Theorem 7.15. *Let $\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I)$ and define the statistic*

$$F = \frac{\mathrm{SSH}/q}{\mathrm{SSE}/(n-k-1)} = \frac{(C\vec{\beta} - \vec{t})^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\vec{\beta}} - \vec{t})/q}{\mathrm{SSE}/(n-k-1)},$$

where C is $q \times (k+1)$ of rank $q \leqslant k+1$ and $\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y}$. Then

$$F \sim \begin{cases} F_{q,n-k-1}, \text{ under } H_0 : C\vec{\beta} = \vec{t} \\ F_{q,n-k-1}(\lambda), \text{ under } H_1 : C\vec{\beta} \neq \vec{t} \end{cases},$$

where

$$\lambda = \frac{(C\vec{\beta} - \vec{t})^T [C(X^T X)^{-1} C^T]^{-1} (C\vec{\beta} - \vec{t})}{2\sigma^2}.$$

7.5 Tesing on β_j and $\vec{a}^T\vec{\beta}$

Important special cases of the general linear test are $H_0 : \beta_j = 0$ or $H_0 : \vec{a}^T\vec{\beta} = 0$, then

$$C = \vec{a}^T \text{ and } q = 1,$$

and

$$F = \frac{(\vec{a}^T\hat{\vec{\beta}})^T[\vec{a}^T(X^TX)^{-1}\vec{a}]^{-1}\vec{a}^T\hat{\vec{\beta}}}{\text{SSE}/n-k-1} = \frac{(\vec{a}^T\hat{\vec{\beta}})^2}{s^2\vec{a}^T(X^TX)^{-1}\vec{a}} \sim F_{1,n-k-1}, \text{ under } H_0 : \vec{a}^T\vec{\beta} = 0.$$

Since the F statistic has 1 and $n-k-1$ degrees of freedom, we can equivalently use the t statistic

$$t = \frac{\vec{a}^T\hat{\vec{\beta}}}{s\sqrt{\vec{a}^T(X^TX)^{-1}\vec{a}}} \sim t_{n-k-1} \text{ under } H_0.$$

A special case: $\vec{a} = (0, 0, \ldots, 1, 0, \ldots, 0)^T$, where 1 is in the $(j+1)$th position. Then

$$\vec{a}^T\hat{\vec{\beta}} = \hat{\beta}_j,$$

$$\vec{a}^T(X^TX)^{-1}\vec{a} = (X^TX)^{-1}_{j+1,j+1}.$$

Thus,

$$F = \frac{\hat{\beta_j}^2}{s^2\{(X^TX)^{-1}\}_{j+1,j+1}},$$

and

$$t = \frac{\hat{\beta}_j}{s\sqrt{\{(X^TX)^{-1}\}_{j+1,j+1}}} = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)},$$

since $\text{Cov}(\hat{\vec{\beta}}) = \sigma^2(X^TX)^{-1}$.

7.6 Confidence Interval and Prediction Intervals

Hypothesis tests and confidence intervals are essentially two ways of approaching the same problem. Recall

(a) For an α-level test of the form $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, a $100(1-\alpha)\%$ confidence region is "set of all values θ_0 s.t. H_0 would not be rejected at the α level". In other words, it is the acceptance region of the α-level test.

(b) θ_0 is outside of a $100(1-\alpha)\%$ confidence region for θ if and only if an α-level test of $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ is rejected.

(c) In other words, we invert the statistical tests we have derived to obtain confidence region.

7.6.1 Confidence Region for $\vec{\beta}$

Under H_0,

$$F = \frac{(C\vec{\beta})^T[C(X^TX)^{-1}C^T]^{-1}C\hat{\vec{\beta}}/\sigma^2 q}{\mathrm{SSE}\,/\sigma^2(n-k-1)} = \frac{\chi^2_q/q}{\chi^2_{n-k-1}/n-k-1} \sim F_{q,n-k-1}.$$

The distribution of F is the same for all values of $\vec{\beta}$ and σ^2, and thus a pivotal quantity, which we can use to derive confidence regions

$$P\left(\frac{\left(C\hat{\vec{\beta}} - C\vec{\beta}\right)^T \left(C(X^TX)^{-1}C^T\right)^{-1}\left(C\hat{\vec{\beta}} - C\beta\right)}{qs^2} \leqslant F_{q,n-k-1,1-\alpha}\right) = 1-\alpha.$$

As a function of $\vec{\beta}$,

$$\frac{\left(C\hat{\vec{\beta}} - C\vec{\beta}\right)^T \left(C(X^TX)^{-1}C^T\right)^{-1}\left(C\hat{\vec{\beta}} - C\beta\right)}{qs^2} = F_{q,n-k-1,1-\alpha},$$

is the equation of an ellipsoid centered at $C\hat{\vec{\beta}}$, with orientation determined by $C(X^TX)^{-1}C^T$.

Special cases:

(a) If C is equal to I, then $\vec{t}$ equal to $\vec{\beta}$ and q becomes to $k+1$, we obtain

$$\frac{(\hat{\vec{\beta}} - \vec{\beta})^T X^T X(\hat{\vec{\beta}} - \vec{\beta})}{(k+1)s^2} \sim F_{k+1,n-k-1}.$$

$$P\left((\hat{\vec{\beta}} - \vec{\beta})^T X^T X(\hat{\vec{\beta}} - \vec{\beta})/(k+1)s^2 \leqslant F_{\alpha,k+1,n-k-1}\right) = 1-\alpha.$$

Then $100(1-\alpha)\%$ joint confidence region for $\vec{\beta}$ is

$$S = \{\vec{\beta} : (\hat{\vec{\beta}} - \vec{\beta})^T X^T X(\hat{\vec{\beta}} - \vec{\beta}) \leqslant (k+1)s^2 F_{k+1,n-k-1,1-\alpha}\}.$$

For $k=1$, this region can be plotted as an ellipse in two dimensions. For $k>1$, the ellipsoidal region is unwieldy to interpret and report, and we therefore consider intervals for the individual β_j's or for $\vec{a}^T\vec{\beta}$.

(b) Let $C = \vec{a}^T$, then $t = \vec{a}^T\vec{\beta}$.

$$\frac{\left(\vec{a}^T\hat{\vec{\beta}} - \vec{a}^T\vec{\beta}\right)^2}{s^2\vec{a}^T(X^TX)^{-1}\vec{a}} \sim F_{1,n-k-1}.$$

Then

$$\frac{\vec{a}^T\hat{\vec{\beta}} - \vec{a}^T\vec{\beta}}{s\sqrt{\vec{a}^T(X^TX)^{-1}\vec{a}}} \sim t_{n-k-1}.$$

Then

$$P\left(t_{n-k-1,\frac{\alpha}{2}} \leqslant \frac{\vec{a}^T\hat{\vec{\beta}} - \vec{a}^T\vec{\beta}}{s\sqrt{\vec{a}^T(X^TX)^{-1}\vec{a}}} \leqslant t_{n-k-1,1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

Rearranging, we find a $100(1-\alpha)\%$ C.I. about $\vec{\alpha}^T\vec{\beta}$ is

$$\vec{\alpha}^T\hat{\vec{\beta}} \pm t_{n-k-1,1-\frac{\alpha}{2}}s\sqrt{\vec{a}^T(X^TX)^{-1}\vec{a}}.$$

(c) Take $\vec{a} = (0,\dots,1,\dots,0)$, where 1 is in the j^{th} position. Similarly, $100(1-\alpha)\%$ C.I. about $\vec{\alpha}^T\vec{\beta}$ is

$$\hat{\beta}_j \pm t_{n-k-1,1-\frac{\alpha}{2}}s\sqrt{\{(X^TX)^{-1}\}_{j+1,j+1}}.$$

7.6.2 Confidence Interval for $E(y_0)$

Let $\vec{x}_0 = (1, x_{01},\dots,x_{0k})^T$ denote a particular choice of $\vec{x} = (1, x_1,\dots,x_k)^T$. Note that $\vec{x}_0$ need not be one of the $\vec{x}^T$s in the sample; that is, $\vec{x}_0^T$ need not be a row of X. If $\vec{x_0}$ is very far outside the area covered by the sample however, the prediction may be poor. Let y_0 be an observation corresponding to $\vec{x}_0$. Then

$$y_0 = \vec{x}_0^T\vec{\beta} + \vec{e},$$

where $\vec{e} \sim N(0,\sigma^2 I)$, and $\vec{\beta}$ and σ^2 are the same. Then $E[y_0] = \vec{x}_0^T\vec{\beta}$. We wish to find a confidence interval for $E[y_0]$, that is, for the mean of the distribution of y-values corresponding to $\vec{x}_0$. The minimum variance unbiased estimator of $E[y_0]$ is given by

$$\widehat{E[y_0]} = \vec{x}_0^T\hat{\beta}.$$

Since that are of the form $\vec{a}^T\vec{\beta}$ and $\vec{a}^T\hat{\vec{\beta}}$, respectively, we obtain a $100(1-\alpha)\%$ confidence interval for $E[y_0] = \vec{x}_0^T\vec{\beta}$

$$\vec{x}_0^T\hat{\vec{\beta}} \pm t_{\alpha/2,n-k-1}s\sqrt{\vec{x}_0^T(X^TX)^{-1}\vec{x}_0}.$$

Remark. We are sometimes interested in simultaneous intervals about several values of the explanatory variables (or for the entire regression line.) Let

$$A_i = \text{“event that } i^{\text{th}} \text{ interval captured the true mean response”}.$$

For example, $P(A_i) = 0.95$ for $i = 1,\dots,4$, then

$$P\left(\bigcap_{i=1}^{4} A_i\right) < 0.95.$$

So each interval needs to made wider to attain an overall C.I. level of 0.95. Let

$$B_i = \text{“Type I error on } i^{\text{th}} \text{ test”}.$$

Then

$$P(\text{“at least one type I error”}) = P\left(\bigcup_{i}^{n} B_i\right) \leqslant \sum_{i=1}^{n} P(B_i).$$

Instead of α on an individual test, use

$$\alpha^* = \frac{\alpha}{n} = \frac{\alpha}{\text{\# of tests}}.$$

Then

$$P\left(\bigcup_{i=1}^{n} B_i\right) \leqslant n\alpha^* = \alpha.$$

See chapter 8.67 in the textbook.

7.6.3 Prediction interval for a future observation y_0

A "confidence interval" for a future observation y_0 corresponding to $\vec{x}_0$ is called a *prediction interval.* We speak of a prediction interval rather than a condidence interval because y_0 is an individual observation and is thereby a random variable rather than a parameter. To be $100(1-\alpha)\%$ confident that the interval contains y_0, the prediction interval will clearly have to be wider than a condidence interval for the parameter $E[y_0]$. Since

$$y_0 = \vec{x}_0^T\vec{\beta} + \vec{\epsilon}_0,$$

we predict y_0 by $\hat{y}_0 = \vec{x}_0^T\hat{\vec{\beta}}$, which is also the estimator of $E[y_0] = \vec{x}_0^T\vec{\beta}$. The random variables y_0 and $\hat{y}_0$ are independent because y_0 is a future observation to be obtained independently of the n observation used to compute $\hat{y}_0 = \vec{x}_0^T\hat{\vec{\beta}}$. Hence

$$\mathrm{Var}(y_0 - \hat{y}_0) = \mathrm{Var}(y_0 - \vec{x}_0^T\hat{\vec{\beta}}) = \mathrm{Var}(\vec{x_0}^T\vec{\beta} + \epsilon_0 - \vec{x}_0^T\hat{\vec{\beta}}).$$

Since $\vec{x_0}^T\vec{\beta}$ is a constant, this becomes

$$\begin{aligned}\mathrm{Var}(y_0 - \hat{y}_0) &= \mathrm{Var}(\epsilon_0) + \mathrm{Var}(\vec{x}_0^T\hat{\vec{\beta}})\\ &= \sigma^2 + \sigma^2\vec{x}_0^T(X^TX)^{-1}\vec{x}_0\\ &= \sigma^2[1 + \vec{x}_0^T(X^TX)^{-1}\vec{x}_0],\end{aligned}$$

which is estimated by

$$s^2[1 + \vec{x}_0^T(X^TX)^{-1}\vec{x}_0].$$

It can be shown that $E[y_0 - \hat{\vec{y}}_0] = 0$ and that s^2 is independent of both y_0 and $\hat{y}_0 = \vec{x}_0^T\hat{\vec{\beta}}$. Therefore, the t statistic

$$t = \frac{y_0 - \hat{y}_0}{s\sqrt{1 + \vec{x}_0^T(X^TX)^{-1}\vec{x}_0}} \sim t(n-k-1).$$

A $100(1-\alpha)\%$ prediction interval is

$$\hat{y}_0 \pm t_{1-\alpha/2,n-k-1}\, s\sqrt{1 + \vec{x}_0^T(X^TX)^{-1}\vec{x}_0},$$

which is wider than the C.I. of $E[y_0]$.

7.6.4 Confidence Interval for σ^2

Since

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

we have

$$P\left[\chi^2_{1-\frac{\alpha}{2},n-k-1} \leqslant \frac{(n-k-1)s^2}{\sigma^2} \leqslant \chi^2_{\alpha/2,n-k-1}\right] = 1-\alpha,$$

where $\chi^2_{\alpha/2,n-k-1}$ is the upper $\alpha/2$ percentage point of the chi-square distribution.

7.7 Likelihood Ratio Tests

The tests in the previous sections were derived using informal methods based on finding sums of squares that have chi-square distributions and are independent. These same tests can be obtained more formlly by the likelihood ratio approch. We describe the likelihood ratio method in the simple context of testing $H_0 : \vec{\beta} = 0$ vs $H_1 : \vec{\beta} \neq 0$. For a random sample

$$\vec{y} = (y_1, \ldots, y_n)^T \sim N_n(X\vec{\beta}, \sigma^2 I),$$

the likelihood function is given as

$$L(\vec{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(\vec{y}-X\vec{\beta})^T(\vec{y}-X\vec{\beta})/2\sigma^2}.$$

The likelihood ratio method compares the maximum value of $L(\vec{\beta}, \sigma^2)$ restriced by $H_0 : \vec{\beta} = \vec{0}$ to the maximum value of $L(\vec{\beta}, \sigma^2)$ under $H_1 : \vec{\beta} \neq \vec{0}$, which is essentially unrestriced. We denote the maximum value of $L(\vec{\beta}, \sigma^2)$ restriced by $\vec{\beta} = \vec{0}$ as

$$\max_{\theta \in \Theta_0} L(\vec{\beta}, \sigma^2)$$

and the unrestriced maximum as

$$\max_{\theta \in \Theta} L(\vec{\beta}, \sigma^2).$$

If $\vec{\beta}$ is equal (or close) to $\vec{0}$, then $\max_{\theta \in \Theta_0} L(\vec{\beta}, \sigma^2)$ should be close to $\max_{\theta \in \Theta} L(\vec{\beta}, \sigma^2)$. If not, we would conclude that $\vec{y} = (y_1, \ldots, y_n)^T$ apparently did not come from $N_n(X\vec{\beta}, \sigma^2 I)$ with $\vec{\beta} = \vec{0}$.

Definition 7.16. The likelihood ratio test (LRT) statistic for testing $H_0 : \vec{\theta} \in \Theta_0$ vs $H_1 : \vec{\theta} \notin \Theta_0$ is

$$\lambda(\vec{y}) = \frac{\sup_{\vec{\theta} \in \Theta_0} L(\vec{\theta}|\vec{y})}{\sup_{\vec{\theta} \in \Theta} L(\vec{\theta}|\vec{y})}$$

It is clear that $0 \leqslant \lambda(\vec{y}) \leqslant 1$. Smaller values of $\lambda(\vec{y})$ would favor H_1 and larger values would favor H_0. We thus reject H_0 if $\lambda(\vec{y}) \leqslant c$, where c is chosen to that $P(\lambda(\vec{y}) \leqslant c) = \alpha$ if H_0 is true.

Theorem 7.17. *If*

$$\vec{y} = X_1\vec{\beta}_1 + X_2\vec{\beta}_2 + e \sim N_n(X\vec{\beta}, \sigma^2 I),$$

the F-test for $H_0 : \vec{\beta}_2 = \vec{0}$ is equivalent to the LRT.

Proof. We use the maximum likelihood estimators

$$\hat{\vec{\beta}} = (X^TX)^{-1}X^T\vec{y}$$

$$\hat{\sigma}^2_{\text{FM}} = \frac{(\vec{y} - X\hat{\vec{\beta}})^T(\vec{y} - X\hat{\vec{\beta}})}{n} = \frac{\text{SSE}_{\text{FM}}}{n}.$$

Then

$$\begin{aligned}\sup_{\sigma^2>0, \vec{\beta}\in\mathbb{R}^{k+1}} L(\vec{\beta}, \sigma^2|\vec{y}) &= \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}} e^{-(\vec{y}-X\hat{\vec{\beta}})^T(\vec{y}-X\hat{\vec{\beta}})/2\hat{\sigma}^2} \\ &= (2\pi)^{-\frac{n}{2}}(\hat{\sigma}^2_{\text{FM}})^{-\frac{n}{2}}e^{-\frac{n}{2}}.\end{aligned}$$

For RM,

$$L_{\text{RM}}(\vec{\beta}_1^*, \sigma^2) = (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{(\vec{y} - X_1\vec{\beta}_1)^T(\vec{y} - X_1\beta_1)}{2\sigma^2}\right).$$

Similarly, we have

$$\hat{\vec{\beta}}_1^* = (X_1^TX_1)^{-1}X_1^T\vec{y},$$

$$\sigma^2_{\text{RM}} = \frac{(\vec{y} - X\hat{\vec{\beta}}^*)^T(\vec{y} - X\hat{\vec{\beta}}^*)}{n} = \frac{\text{SSE}_{\text{RM}}}{n}.$$

Then

$$\sup_{\sigma^2>0, \vec{\beta}_1^*\in\mathbb{R}^{k+1-h}} L(\vec{\beta}_1^*, \sigma^2|\vec{y}) = (2\pi)^{-\frac{n}{2}}(\hat{\sigma}^2_{\text{RM}})^{-\frac{n}{2}}e^{-\frac{n}{2}}.$$

Thus, the LRT statistic is

$$\lambda(\vec{y}) = \frac{\sup_{\vec{\theta}\in\Theta_0} L(\vec{\beta}_1^*, \sigma^2|\vec{y})}{\sup_{\vec{\beta},\sigma62\in\Theta} L(\vec{\theta}|\vec{y})} = \frac{(2\pi)^{-\frac{n}{2}}(\hat{\sigma}^2_{\text{RM}})^{-\frac{n}{2}}e^{-\frac{n}{2}}}{(2\pi)^{-\frac{n}{2}}(\hat{\sigma}^2_{\text{FM}})^{-\frac{n}{2}}e^{-\frac{n}{2}}} = \left(\frac{\text{SSE}_{\text{FM}}}{\text{SSE}_{\text{RM}}}\right)^{\frac{n}{2}}.$$

Rejection region:

$$\{\vec{y} : \lambda(\vec{y}) \leqslant c\},\ c \in (0, 1).$$

$$\begin{aligned}\lambda(y) \leqslant c &\iff \frac{\text{SSE}_{\text{FM}}}{\text{SSE}_{\text{RM}}} \geqslant\leqslant c \\ &\iff \frac{\text{SSE}_{\text{RM}}}{\text{SSE}_{\text{FM}}} \geqslant c^{-\frac{2}{n}} \\ &\iff \frac{\text{SSE}_{\text{RM}} - \text{SSE}_{\text{FM}}}{\text{SSE}_{\text{FM}}} \geqslant c^{-\frac{2}{n}} - 1 \\ &\iff \frac{h}{n-k-1}\frac{\vec{y}\left(P_{C(X)} - P_{C(X_1)}\right)\vec{y}}{\vec{y}\left(I - P_{C(X)}\right)\vec{y}/(n-k-1)} \geqslant c^{-\frac{2}{n}} - 1 \\ &\iff F \geqslant \frac{h}{n-k-1}\left(c^{-\frac{2}{n}} - 1\right),\end{aligned}$$

where

$$F = \frac{\vec{y}\left(P_{C(X)} - P_{C(X_1)}\right)\vec{y}}{\vec{y}\left(I - P_{C(X)}\right)\vec{y}/(n-k-1)} \sim F_{h,n-k-1}.$$

Thus, the LRT and that F test are equivalent. □

Theorem 7.18. *If $\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I)$, then the F test for $C\vec{b}eta = \vec{t}$ is equivalent to the* LRT.

Proof. To derive the LRT for the gen. lin. hypothesis $H_0 : C\vec{\beta} = \vec{t}$, we need to find

$$\sup_{\sigma^2>0, \vec{\beta}: C\vec{\beta}=t} L(\beta, \sigma^2).$$

We can do this with Lagrange multiplier,

$$\begin{aligned}\rho(\vec{\beta}, \sigma^2) &= l(\vec{\beta}, \sigma^2) + \vec{\lambda}^T(C\vec{\beta} - \vec{t}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})}{2\sigma^2} + \lambda^T(C\vec{\beta} - \vec{t}).\end{aligned}$$

Let

$$\begin{aligned}\hat{\sigma}^2 &= \frac{(\vec{y} - X\hat{\vec{\beta}})^T(\vec{y} - X\hat{\vec{\beta}})}{n}, \\ \hat{\vec{\beta}} &= (X^TX)^{-1}X^T\vec{y}.\end{aligned}$$

Differentiating with respect to $\vec{\beta}$, $\vec{\lambda}$ and σ^2, we obtain

$$\frac{\partial\rho}{\partial\vec{\beta}} = \frac{1}{2\sigma^2}(2X^T\vec{y} - 2X^TX\vec{\beta}) + C^T\vec{\lambda} = \vec{0},$$

$$\frac{\partial\rho}{\partial\vec{\lambda}} = C\vec{\beta} - \vec{t} = \vec{0},$$

$$\frac{\partial\rho}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) = 0.$$

From the first equation, we have

$$\hat{\vec{\beta}}_0 = (X^TX)^{-1}X^T\vec{y} + \hat{\sigma}_0^2(X^TX)^{-1}C^T\vec{\lambda}.$$

Then

$$C\hat{\vec{\beta}}_0 = C\hat{\vec{\beta}} + \sigma^2 C(X^TX)^{-1}C^T\vec{\lambda} = \vec{t}.$$

Then

$$\lambda = -\left[C(X^TX)^{-1}C^T\right]^{-1}\frac{C\hat{\vec{\beta}} - \vec{t}}{\hat{\sigma}^2}.$$

Solving for $\vec{\lambda}$ and plugging in

$$\hat{\vec{\beta}}_0 = \hat{\vec{\beta}} - (X^TX)^{-1}C^T[C(X^TX)^{-1}C^T]^{-1}(C\hat{\vec{\beta}} - \vec{t}),$$

and

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}(\vec{y} - X\hat{\vec{\beta}}_0)^T(\vec{y} - X\hat{\vec{\beta}}_0) \\ &= \frac{1}{n}\left(\vec{y} - X\hat{\vec{\beta}} + X(X^TX)^{-1}C^T\left(C(X^TX)^{-1}C^T\right)^{-1}\left(C\hat{\vec{\beta}} - \vec{t}\right)\right)^T \\ &\quad\cdot\frac{1}{n}\left(\vec{y} - X\hat{\vec{\beta}} + X(X^TX)^{-1}C^T\left(C(X^TX)^{-1}C^T\right)^{-1}\left(C\hat{\vec{\beta}} - \vec{t}\right)\right).\end{aligned}$$

Let

$$\vec{d}_1 = \vec{y} - X\hat{\vec{\beta}}_0,$$
$$\vec{d}_2 = X(X^TX)^{-1}C^T\left(C(X^TX)^{-1}C^T\right)\left(C\hat{\vec{\beta}} - \vec{t}\right).$$

Then

$$\begin{aligned}\vec{d}_1^T\vec{d}_2 &= (\vec{y} - X\hat{\vec{\beta}}_0)^T X(X^TX)^{-1}C^T\left(C(X^TX)^{-1}C^T\right)^{-1}\left(C\hat{\vec{\beta}} - \vec{t}\right)\\ &= (\vec{y} - X\hat{\beta})(X\vec{z}) = 0.\end{aligned}$$

$$\begin{aligned}\hat{\sigma}_0^2 &= \frac{1}{n}(\vec{d}_1 + \vec{d}_2)^T(\vec{d}_1 + \vec{d}_2) = \frac{1}{n}\left[\vec{d}_1^T\vec{d}_1 + \vec{d}_2^T\vec{d}_2 + 2\vec{d}_1^T\vec{d}_2\right]\\ &= \frac{1}{n}\vec{d}_1^T\vec{d}_1 + \frac{1}{n}\vec{d}_2^T\vec{d}_2 = \mathrm{SSE}_{\mathrm{FM}} + \vec{d}_2^T\vec{d}_2.\end{aligned}$$

But

$$\begin{aligned}d_2^Td_2 = &\left(C\hat{\vec{\beta}} - \vec{t}\right)^T\left(C(X^TX)^{-1}C^T\right)^{-1}C(X^TX)^{-1}X^T\cdot\\ &\cdot X(X^TX)^{-1}C^T\left(C(X^TX)^{-1}C^T\right)\left(C\hat{\vec{\beta}} - \vec{t}\right)\\ =& \left(C\hat{\vec{\beta}} - \vec{t}\right)^T\left(C(X^TX)^{-1}C^T\right)^{-1}C(X^TX)^{-1}C^T\left(C(X^TX)^{-1}C^T\right)^{-1}\left(C\hat{\vec{\beta}} - \vec{t}\right)\\ =& \left(C\hat{\vec{\beta}} - \vec{t}\right)^T\left(C(X^TX)^{-1}C^T\right)^{-1}\left(C\hat{\vec{\beta}} - \vec{t}\right)\\ =& \mathrm{SSH}_0\,.\end{aligned}$$

Thus,

$$\sigma_0^2 = \frac{\mathrm{SSE}_{\mathrm{FM}}}{n} + \frac{\mathrm{SSH}_0}{n}.$$

Then the LRT statistic is and

$$\begin{aligned}\lambda(\vec{y}) &= \frac{\sup_{\sigma^2>0, C\vec{\beta}=\vec{0}} L(\vec{\beta}, \sigma^2|\vec{y})}{\sup_{\sigma^2>0, \vec{\beta}\in\mathbb{R}^{k+1}} L(\vec{\beta}, \sigma^2|\vec{y})}\\ &= \left(\frac{\mathrm{SSE}_{\mathrm{FM}}}{\mathrm{SSE}_{\mathrm{RM}}}\right)^{n/2} = \left(\frac{\mathrm{SSE}_{\mathrm{FM}}}{\mathrm{SSE}_{\mathrm{FM}} + \mathrm{SSH}_0}\right)^{n/2}\\ &= \left(\frac{1}{1 + \mathrm{SSH}_0 / \mathrm{SSE}_{\mathrm{FM}}}\right)^{n/2}\\ &= \left(\frac{1}{1 + qF/(n-k-1)}\right)^{n/2},\end{aligned}$$

where

$$F = \frac{\mathrm{SSH}/q}{\mathrm{SSE}/n-k-1}.$$

Then $\lambda(\vec{y}) \leqslant c,\ c \in (0,1)$ is equivalent to $F \geqslant F_{q,n-k-1,1-\alpha}$. □