STA437 Notes

Ian Zhang

July 29, 2024

Contents

1 Matrix Algebra and Random Vectors **3**
1.1 Vectors . . . 3
1.1.1 Vector Spaces . . . 3
1.1.2 Inner Product Spaces . . . 3
1.1.3 Length . . . 3
1.1.4 Angle . . . 3
1.1.5 Projections . . . 4
1.1.6 Cauchy-Schwarz Inequality . . . 4
1.2 Matrices . . . 4
1.2.1 Properties of Matrices . . . 4
1.2.2 Orthogonal Matrices . . . 5
1.2.3 Eigenvalues/Eigenvectors . . . 5
1.2.4 Symmetric Matrices . . . 6
1.2.5 Definite Matrices . . . 6
1.2.6 Square Root Matrix . . . 7
1.2.7 Extended Cauchy-Schwarz Inequality . . . 7
1.2.8 Idempotence . . . 7
1.2.9 Orthogonal Projection Matrices . . . 8
1.3 Random Vectors . . . 8
1.3.1 Expected Value . . . 8
1.3.2 Variance-Covariance Matrix . . . 8
1.3.3 Correlation Matrix . . . 9
1.3.4 Standard Deviation Matrix . . . 9
1.3.5 Mean/Covariance of Linear Combinations of Random Variables . . . 10
1.3.6 Partitioning Σ and μ . . . 10

2 Sampling Geometry and Random Sampling **11**
2.1 Random Samples . . . 11
2.2 Descriptive Statistics . . . 11
2.2.1 Sample Mean as a Projection . . . 12
2.2.2 Deviation Vectors . . . 13
2.3 Sample Mean, Covariance, and Correlation as Matrix Operations . . . 13
2.4 Sample Values of Linear Combinations of Random Variables . . . 14
2.4.1 Sampling Distributions of Estimators . . . 15
2.5 Generalized Variance . . . 15
2.5.1 Other Measures of Generalized Variance . . . 16

3 Multivariate Normal Distribution **16**
3.1 Multivariate Normal Density . . . 16
3.1.1 Bivariate Normal Distribution . . . 17
3.2 Key Properties of Multivariate Normal Distribution . . . 17
3.2.1 Property 1 . . . 17
3.2.2 Property 2 . . . 17
3.2.3 Property 3 . . . 18
3.2.4 Property 4 . . . 18
3.3 Constant Probability Density Contours . . . 19
3.4 Distribution of $(X-\mu)^T\Sigma^{-1}(X-\mu)$ . . . 19
3.5 Sampling from the Multivariate Normal Distribution and MLE . . . 19
3.5.1 Multivariate Normal Likelihood . . . 19

4 Linear Regression **20**
4.1 Least Squares Estimation . . . 20
4.1.1 Univariate . . . 20
4.1.2 Multivariate . . . 21
4.1.3 Sampling Properties of the LSE . . . 23
4.2 Sum of Squares Decomposition . . . 23

5 Principal Component Analysis **24**
5.1 PCA Formulation . . . 24
5.2 Total Population Variance . . . 25
5.3 Sample Variance by PC . . . 25

1 Matrix Algebra and Random Vectors

1.1 Vectors

A vector x with p elements is a $p \times 1$ matrix

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix}^T$$

1.1.1 Vector Spaces

A vector space V over a field $\mathbb{F}$ is a set with 2 operators that satisfy certain axioms.

1.1.2 Inner Product Spaces

An inner product of vectors in a vector space V is a mapping

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$$

The **dot product** is defined as

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

1.1.3 Length

The length of a vector x is L_x, where

$$L_x = \sqrt{x^T x} = \sqrt{\sum_{i=1}^{p} x_i^2} \qquad L_{cx} = |c| \sqrt{L_x}$$

where c is a constant. If we choose $c = L_x^{-1}$, then cx is the unit vector with length 1 in the direction of x since $c > 0$.

1.1.4 Angle

Suppose the angle between $x, y \in \mathbb{R}^2$ is θ. Then

$$\cos(\theta) = \frac{\langle x, y \rangle}{L_x L_y}$$

Thus if x, y are perpendicular,

$$\cos(\theta) = 0 \implies x^T y = 0$$

since L_x and L_y are nonzero.

1.1.5 Projections

The projection of x onto y has length

$$L_x|\cos(\theta)| = \left|\frac{\langle x, y\rangle}{L_y}\right|$$

This vector has direction y, so the projection of x onto y is

$$\frac{x^T y}{y^T y} y$$

1.1.6 Cauchy-Schwarz Inequality

Let $x, y \in \mathbb{R}^p$. Then

$$(x^T y)^2 \leq (x^T x)(y^T y) = L_x^2 L_y^2$$

with equality holding iff $x = cy$ for some $c \in \mathbb{R}$.

Proof. If either x, y are zero, then $LHS = RHS = 0$. Suppose $x, y \neq 0$ and consider $x - cy$. If $x - cy \leq 0$, then

$$\begin{aligned} 0 &< (x - cy)^T (x - cy) \\ &= x^T x - 2cx^T y + c^2 y^T y \end{aligned}$$

Since $y^T y > 0$, then the quadratic in c $x^T x - 2cx^T y + c^2 y^T y$ has no real roots, thus $4(x^T y)^2 - 4(x^T x)(y^T y) \leq 0$, so

$$(x^T y)^2 \leq (x^T x)(y^T y)$$

■

1.2 Matrices

Definition. The rank of a matrix A is the number of independent rows/columns of A.

- $\text{rank}(A_{n\times p}) \leq \min(n, p)$

1.2.1 Properties of Matrices

- $(AB)^T = B^T A^T$
- $(A^{-1})^T = (A^T)^{-1}$
- $(AB)^{-1} = B^{-1} A^{-1}$

- $\text{rank}(A) = \text{rank}(A^T) = \text{rank}(AA^T) = \text{rank}(A^TA)$
- $A_{n\times n}$ is invertible iff $\text{rank}(A) = n$
- $\det(AB) = \det(A)\det(B)$
- $\det(A) = \det(A^T)$
- $\det(A^{-1}) = \frac{1}{\det(A)}$
- $\det(cA_{k\times k}) = c^k \det(A_{k\times k})$
- $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$
- $\text{tr}(AB) = \text{tr}(BA)$
- $\text{tr}(P^{-1}AP) = \text{tr}(A)$ where P is invertible and of same dimension as A

1.2.2 Orthogonal Matrices

Definition. An orthogonal matrix $Q \in \mathbb{R}^{k\times k}$ is characterized by

$$QQ^T = Q^TQ = I$$

or equivalently, $Q^T = Q^{-1}$.

A matrix Q is orthogonal if all its rows have length 1 and are mutually orthogonal.

Proposition. If Q is orthogonal, then $\det(Q) = \pm 1$.

Proof. $\det(Q) = \det(Q^T) = \det(Q^{-1}) = \frac{1}{\det(Q)} \implies |\det(Q)| = 1.$ ■

1.2.3 Eigenvalues/Eigenvectors

An eigenvector x of a matrix A with eigenvalue λ satisfies

$$Ax = \lambda x$$

where $x \neq 0$. To solve for λ, we solve $\det(A - \lambda I) = 0$.
Properties: For arbitrary $A \in \mathbb{R}^{k\times k}$,

1. $\text{tr}(A) = \sum_{i=1}^{n} \lambda_i$
2. $\det(A) = \prod_{i=1}^{n} \lambda_i$
3. $P^{-1}AP$ and P have the same eigenvalues

1.2.4 Symmetric Matrices

Definition. A square matrix A is symmetric if $A = A^T$.

Properties: For a symmetric $A_{k\times k}$, let

$$(\lambda_1, e_1), (\lambda_2, e_2), \cdots, (\lambda_k, e_k)$$

be its pairs of eigenvalues and eigenvectors, which are assumed to be normalized (i.e., $L_{e_i} = 1$ for all i). Then

1. $\lambda_k \in \mathbb{R}$
2. All the eigenvectors are mutually orthogonal, or equivalently, $\langle e_i, e_j \rangle = 0$

The **spectral decomposition for symmetric matrices** is the matrix

$$A = P\Lambda P^T = \sum_{i=1}^{k} \lambda_i e_i e_i^T$$

where

$$P = \begin{bmatrix} e_1 & e_2 & \cdots & e_k \end{bmatrix}$$
$$\Lambda = \operatorname{diag}\begin{bmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_k \end{bmatrix}$$

- $A^n = P\Lambda^n P$ where
$$\Lambda^n = \operatorname{diag}\begin{bmatrix} \lambda_1^n & \cdots & \lambda_k^n \end{bmatrix}$$

1.2.5 Definite Matrices

Definition. A symmetric matrix $A_{k\times k}$ is nonnegative definite if for all $x \in \mathbb{R}^k$,

$$x^T A x \geq 0$$

If so, we say $A \succeq 0$.

Definition. A symmetric matrix $A_{k\times k}$ is positive definite if for all $x \neq 0$,

$$x^T A x > 0$$

If so, we say $A \succ 0$.

Proposition. A symmetric matrix A is positive definite iff all the eigenvalues of A are *strictly* positive.

- Nonnegative definite if all eigenvalues are nonnegative

Let A be a $k \times k$ positive definite matrix, thus $A = P\Lambda P^T$. A's inverse is the matrix

$$A^{-1} = P\Lambda^{-1}P^T$$

where

$$\Lambda^{-1} = \text{diag}\begin{bmatrix} \frac{1}{\lambda_1} & \cdots & \frac{1}{\lambda_k} \end{bmatrix}$$

1.2.6 Square Root Matrix

Definition. The square root matrix of a positive definite matrix A is

$$A^{\frac{1}{2}} = P\Lambda^{\frac{1}{2}}P^T$$

Properties:

- $A^{\frac{1}{2}}$ is symmetric
- $A^{\frac{1}{2}}A^{\frac{1}{2}} = A$
- $A^{\frac{1}{2}}A^{-\frac{1}{2}} = A^{-\frac{1}{2}}A^{\frac{1}{2}} = I$

1.2.7 Extended Cauchy-Schwarz Inequality

Let $x, y \in \mathbb{R}^p$ and B be a $p \times p$ positive definite matrix. Then

$$(x^T y)^2 \leq (x^T Bx)(y^T B^{-1} y)$$

with equality holding iff $x = cB^{-1}y$, $c \in \mathbb{R}$.

Proof. If either x or y is zero, then $LHS = RHS = 0$. Suppose $x, y \neq 0$. Then

$$\begin{aligned} x^T y &= x^T I y \\ &= x^T B^{\frac{1}{2}} B^{-\frac{1}{2}} y && B \text{ is positive definite} \\ &= (B^{\frac{1}{2}}x)^T B^{-\frac{1}{2}} y && \text{property of } B^{\frac{1}{2}} \end{aligned}$$

Applying the $C - S$ inequality to $B^{\frac{1}{2}}x$ and $B^{-\frac{1}{2}}y$ proves the claim. ■

1.2.8 Idempotence

Definition. A matrix A is idempotent if $A^2 = A$.

- If A is idempotent, then so is $I - A$

1.2.9 Orthogonal Projection Matrices

Definition. A square matrix A is an orthogonal projection matrix if

$$A^2 = A = A^T$$

Note that in general, a projection matrix A is idempotent and symmetric.
Let P be an orthogonal projection matrix. Then

$$\langle x, Py\rangle = \langle Px, y\rangle = \langle Px, Py\rangle$$

To show the first equality,

$$\langle x, Py\rangle = x^T Py = (P^T x)^T y = (Px)^T y = \langle Px, y\rangle$$

To show the second, we use the first to show

$$\langle Px, Py\rangle = \langle x, P^2 y\rangle = \langle x, Py\rangle$$

1.3 Random Vectors

Definition (Random Vectors)**.** A random vector is a vector with random variables as elements.

1.3.1 Expected Value

The expected value of a random vector X is vector

$$\mu = E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

If A, B are deterministic matrices (constant matrices), then

$$E(AX + B) = AE(X) + B$$

1.3.2 Variance-Covariance Matrix

The variance-covariance matrix

$$\Sigma = \mathrm{Cov}(X) = E[(X-\mu)(X-\mu)^T] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix}$$

where

$$\sigma_{ij} = \text{Cov}(X_i, X_j)$$

Since covariance is symmetric, then

$$\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$$

Σ is symmetric. By definition,

$$\sigma_{ij} = \sum_{x_i} \sum_{x_j} (x_i - \mu_i)(x_j - \mu_j) p_{ij}(x_i, x_j) = \sigma_{ji}$$

1.3.3 Correlation Matrix

The correlation matrix is a matrix

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}$$

where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$$

is the correlation between X_i and X_j. Since correlation is proportionate to covariance, which is symmetric, then $\rho_{ij} = \rho_{ji}$ for all i, j, thus ρ is symmetric.

1.3.4 Standard Deviation Matrix

The standard deviation matrix is a matrix

$$V^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

It is worth noting that

$$\Sigma = V^{\frac{1}{2}} \rho V^{\frac{1}{2}} \implies \rho = V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}}$$

In the univariate case, we have

$$\text{Var}(X) = E(X^2) - \mu^2$$

In the multivariate case,

$$\Sigma = E[(X - \mu)(X - \mu)^T] = E[XX^T] - \mu\mu^T$$

1.3.5 Mean/Covariance of Linear Combinations of Random Variables

Let $c^T = \begin{bmatrix} c_1 & \cdots & c_p \end{bmatrix}$ be a vector of constants and consider the linear combination $c^T X$. Then

$$E(c^T X) = c^T E(X) \qquad \text{Var}(c^T X) = c^T \Sigma c$$

If C is a $q \times p$ deterministic matrix, then if we define the vector

$$Z_{q\times 1} = C_{q\times p} X_{p\times 1}$$

as the collection of q linear combinations of the X_i, then

$$\mu_Z = C\mu_X \qquad \Sigma_Z = C\Sigma_X C^T$$

Proof. The case of μ_Z is trivial by definition of expectation. To show Σ_Z,

$$\begin{aligned}
\Sigma_Z &= E[(Z-\mu_Z)(Z-\mu_Z)^T] \\
&= E[(CX - C\mu_X)(CX - C\mu_X)^T] \\
&= CE[(X-\mu_X)(X-\mu_X)^T]C^T \\
&= C\Sigma_X C^T
\end{aligned}$$

as required. ■

1.3.6 Partitioning Σ and μ

Partition X into 2 groups

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_q \\ \cdots \\ X_{q+1} \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} X^{(1)}_{q\times 1} \\ \cdots \\ X^{(2)}_{(p-q)\times 1} \end{bmatrix}$$

Then

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_q \\ \cdots \\ \mu_{q+1} \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \mu^{(1)}_{q\times 1} \\ \cdots \\ \mu^{(2)}_{(p-q)\times 1} \end{bmatrix}$$

and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \vdots & \Sigma_{12} \\ \dots & \vdots & \dots \\ \Sigma_{21} & \vdots & \Sigma_{22} \end{bmatrix}$$

where Σ_{11} is $q \times q$, Σ_{22} is $(p-q) \times (p-q)$, Σ_{12} is $q \times (p-q)$ and $\Sigma_{21} = \Sigma_{12}^T$.

2 Sampling Geometry and Random Sampling

2.1 Random Samples

Suppose

$$X = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$$

is a collection of n samples of size p from a joint distribution $f_X(x) = f_X(x_1, \dots, x_p)$.

2.2 Descriptive Statistics

Let $x_{11}, \dots, x_{n1}$ be n observations on the first variable. Then

$$\bar{x}_1 = \frac{1}{n} \sum_{i=1}^{n} x_{i1}$$

is the sample mean of the first variable, thus

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik}$$

is the sample mean for the kth for all $k \in \{1, \dots, p\}$.

$$s_k^2 = s_{kk} = \frac{1}{n} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2$$

is the sample variance, so $\sqrt{s_{kk}}$ is the sample standard deviation.

$$s_{ik} = \frac{1}{n} \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

is the sample covariance for $i, k \in \{1, \dots, p\}$. Finally, the correlation coefficient is for $i, k \in \{1, \dots, p\}$ is

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n} (x_j - \bar{x}_i)^2} \sqrt{\sum_{j=1}^{n} (x_j - \bar{x}_k)^2}}$$

With these definitions, we can define sample mean and covariance as

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} \qquad S_n = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{1p} & \cdots & s_{pp} \end{bmatrix}$$

If we instead use

$$s_{ik} = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

as the sample covariance between X_i and X_p, then we can define an unbiased estimator of Σ as

$$E(S) = \Sigma \implies E(S_n) = \frac{n-1}{n}\Sigma$$

Finally, define sample correlation as

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix}$$

2.2.1 Sample Mean as a Projection

Consider the data

$$X_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & \cdots & y_p \end{bmatrix}$$

Define the vector

$$\mathbb{1}_n^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}$$

thus $\frac{1}{\sqrt{n}}\mathbb{1}_n$ has length 1.

Claim. The projection of y_i onto $\frac{1}{\sqrt{n}}\mathbb{1}_n$ is the vector $\bar{x}_i\mathbb{1}_n = \begin{bmatrix} \bar{x}_i & \bar{x}_i & \cdots & \bar{x}_i \end{bmatrix}^T$.

Proof.

$$\frac{\langle y_i, \frac{1}{\sqrt{n}}\mathbb{1}_n\rangle}{\langle \frac{1}{\sqrt{n}}\mathbb{1}_n, \frac{1}{\sqrt{n}}\mathbb{1}_n\rangle}\frac{1}{\sqrt{n}}\mathbb{1}_n = \left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n} x_{ji}\right)\frac{1}{\sqrt{n}}\mathbb{1}_n = \frac{1}{n}\sum_{j=1}^{n} x_{ji}\mathbb{1}_n = \bar{x}_i\mathbb{1}_n$$

■

2.2.2 Deviation Vectors

For each y_i defined above, define the deviation vector

$$d_i = y_i - \bar{x}_i \mathbb{1}_n = \begin{bmatrix} x_{1i} - \bar{x}_i & \cdots & x_{ni} - \bar{x}_i \end{bmatrix}^T$$

Since $\bar{x}_i \mathbb{1}_n$ is the projection of y_i onto $\frac{1}{\sqrt{n}}\mathbb{1}_n$, then $d_i \perp \bar{x}_i \mathbb{1}_n$. Notice how by definition,

$$\langle d_i, d_i \rangle = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2 = (n-1)s_{ii}$$

for all $i \in \{1, \ldots, p\}$. This shows that the length of the deviation vectors are proportionate to the standard deviations, so if d_i is long, then the variance in that direction is *large*, and if short, then the variance is *small*.

$$L_{d_i} = \sqrt{(n-1)s_{ii}}$$

For $i \neq j$, we have

$$\langle d_i, d_j \rangle = \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) = (n-1)s_{ij}$$

Since $\langle x, y \rangle = L_x L_y \cos(\theta)$ where θ is the angle between x and y, then this shows

$$L_{d_i} L_{d_j} \cos(\theta_{ij}) = (n-1)\sqrt{s_{ii}}\sqrt{s_{jj}} \cos(\theta_{ij})$$

By definition of sample correlation, this shows that $r_{ij} = \cos(\theta_{ij})$, or equivalently, the sample correlation between X_i and X_j is the cosine of the angle between their respective deviation vectors of their observations.
This implies that

$$S = [s_{ij}]_{ij} = \left[\frac{1}{n-1} \langle d_i, d_j \rangle \right]_{ij}$$

Additionally, the **sum of the elements of a deviation vector must equal** 0**.**

2.3 Sample Mean, Covariance, and Correlation as Matrix Operations

Paired with the definitions above, the sample mean can be rewritten as

$$\bar{x} = \frac{1}{n} X^T \mathbb{1}_n = \frac{1}{n} \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Define the $n \times p$ matrix of deviations

$$X - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T X = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix} = \begin{bmatrix} d_1 & d_2 & \cdots & d_p \end{bmatrix}$$

Claim. $I - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$ is an orthogonal projector.

Proof. Since I and $\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$ are symmetric, then

$$\begin{aligned} \left\langle \left(I - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\right)x, y \right\rangle &= \langle x, y\rangle - \left\langle \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T x, y \right\rangle \\ &= \langle x, y\rangle - \left\langle x, \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T y \right\rangle \\ &= \left\langle x, \left(I - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\right) y \right\rangle \end{aligned}$$

which shows symmetry. To show idempotence,

$$\left(I - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\right)^2 = I - \frac{2}{n}\mathbb{1}_n\mathbb{1}_n^T + \left(\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\right)^2 = I - \frac{2}{n}\mathbb{1}_n\mathbb{1}_n^T + \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T = 1 - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$$

as required. ■

Using this matrix, we define the $p \times p$ sample covariance matrix using

$$(n-1)S = \left(X - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T X\right)^T \left(X - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T X\right) = X^T\left(I - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\right)X$$

Define the $p \times p$ standard deviation matrix

$$D^{\frac{1}{2}}_{p\times p} = \begin{bmatrix} \sqrt{s_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{s_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{s_{pp}} \end{bmatrix}$$

Then the sample correlation matrix is

$$R = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$$

where

$$D^{-\frac{1}{2}} = \left(D^{\frac{1}{2}}\right)^{-1} = \operatorname{diag}\begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & \cdots & \frac{1}{\sqrt{s_{pp}}} \end{bmatrix}$$

2.4 Sample Values of Linear Combinations of Random Variables

Consider again

$$c^T X = c_1 X_1 + \cdots + c_p X_p$$

Then

$$\text{Sample mean} = c^T \bar{x}$$

$$\text{Sample variance of } c^T X = c^T S c$$

For a deterministic $q \times p$ matrix C, the q linear combinations $C_{q\times p} X_p$ have

$$\text{Sample mean vector} = C\bar{x}$$

$$\text{Sample covariance matrix} = CSC^T$$

2.4.1 Sampling Distributions of Estimators

Let $X_1, \ldots, X_n$ be random samples from a joint distribution with mean vector μ and covariance matrix Σ.

- $\bar{X}$ is an unbiased estimator of μ with $E(\bar{X}) = \mu$
- The covariance matrix of $\bar{X}$ is $\text{Cov}(\bar{X}) = \frac{1}{n}\Sigma$
- $E(S_n) = \frac{n-1}{n}\Sigma$ so S is an unbiased estimator of Σ

2.5 Generalized Variance

The sample covariance matrix is given by

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix}$$

which contains p variances and $\frac{p(p-1)}{2}$ covariances. To express a single value for the variation expressed by S, use

$$\text{Generalized sample variance} = \det(S)$$

The generalized sample variance is proportionate to the square of the volume generated by the p deviation vectors $d_1, \ldots, d_p$.

$$\det(S) = (n-1)^{-p} \det \begin{bmatrix} d_1 & \cdots & d_p \end{bmatrix}$$

since $S = \left[\frac{1}{n-1}\langle d_i, d_j \rangle\right]_{ij}$. Thus, the generalized variance is 0 iff the deviation vectors are linearly dependent.

- If $n \leq p$, then $\det(S) = 0$

2.5.1 Other Measures of Generalized Variance

$$\text{Generalized variance of standardized variables} = \det(R)$$

Since $R = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$ and $\det(D^{-\frac{1}{2}}) = \prod_{i=1}^{p} \frac{1}{\sqrt{s_{ii}}}$, then

$$\det(R) = \det(D^{-\frac{1}{2}}) \det(S) \det(D^{-\frac{1}{2}}) = (s_{11} s_{22} \cdots s_{pp})^{-1} \det(S)$$

$$\text{Total sample variance} = \operatorname{tr}(S) = \sum_{i=1}^{p} s_{ii}$$

3 Multivariate Normal Distribution

3.1 Multivariate Normal Density

Definition. A p-dimensional random vector X has multivariate normal distribution if every linear combination of its components, $c^T X$ has a univariate normal distribution.

If $X \sim \mathcal{N}(\mu, \sigma^2)$, the pdf is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

for $x \in \mathbb{R}$. Rewrite the exponent as

$$-\frac{(x-\mu)^2}{2\sigma^2} = -\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)$$

If $X \sim \mathcal{N}_p(\mu, \Sigma)$, the exponent is generalized to

$$-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)$$

thus the pdf is

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}} \det(\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

If $\det(\Sigma) = 0$, we say X is degenerate.
If $Z_1, \ldots, Z_p$ are independent with $Z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, the density of

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix}$$

is $\mathcal{N}_p(\mu, \Sigma)$ with $\mu = \begin{bmatrix} \mu_1 & \cdots & \mu_p \end{bmatrix}^T$ and $\Sigma = \operatorname{diag}\begin{bmatrix} \sigma_1^2 & \cdots & \sigma_p^2 \end{bmatrix}$

3.1.1 Bivariate Normal Distribution

If $p = 2$, then

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix}$$

and

$$\Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$$

Thus

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}} \exp\left[-\frac{1}{2\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_{11}} + \frac{(x_2-\mu_2)^2}{\sigma_{22}} - 2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2)\right)\right]$$

3.2 Key Properties of Multivariate Normal Distribution

3.2.1 Property 1

If $X \sim \mathcal{N}_p(\mu, \Sigma)$, then

$$a^T X \sim \mathcal{N}(a^T\mu, a^T\Sigma a)$$

If we take $a = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T$, then $a^T X = X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$. In general, the marginal disrtibution over the X_i is $\mathcal{N}(\mu_i, \sigma_i)$. Generalizing, if we consider the q linear combinations $A_{q\times p}X_{p\times 1}$, then

$$AX \sim \mathcal{N}_q(A\mu, A\Sigma A^T)$$

3.2.2 Property 2

All subsets of X are normally distributed. If we partition $X_{p\times 1}$ into $X^{(1)}_{q\times 1}$ and $X^{(2)}_{(p-q)\times 1}$, then $X^{(1)}_{q\times 1} \sim \mathcal{N}_q(\mu^{(1)}, \Sigma_{11})$ where Σ_{11} is $q \times q$. More precisely, if we consider

$$X = \begin{bmatrix} X_1 & \cdots & X_q & \vdots & X_{q+1} & \cdots & X_p \end{bmatrix}^T = \begin{bmatrix} X^{(1)} \\ \cdots \\ X^{(2)} \end{bmatrix}$$

We know this gives partitions such that

$$\mu = \begin{bmatrix} \mu^{(1)} \\ \cdots \\ \mu^{(2)} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \vdots & \Sigma_{12} \\ \cdots & \vdots & \cdots \\ \Sigma_{21} & \vdots & \Sigma_{22} \end{bmatrix}$$

Let $A = \begin{bmatrix} I_{q\times q} & \vdots & 0_{q\times(p-q)} \end{bmatrix}$. Then by property 1,

$$X^{(1)} \sim \mathcal{N}_q(A\mu, A\Sigma A^T) = \mathcal{N}_q(\mu^{(1)}, \Sigma_{11})$$

3.2.3 Property 3

Let $X = \begin{bmatrix} X^{(1)}_{q_1\times 1} & \vdots & X^{(2)}_{q_2\times 1} \end{bmatrix}^T$. Then $X^{(1)} \perp\!\!\!\perp X^{(2)}$ iff $\Sigma_{12} = 0_{q_1\times q_2}$.

3.2.4 Property 4

If $X \sim \mathcal{N}_p(\mu, \Sigma)$ and $X = \begin{bmatrix} X_1 & \vdots & X_2 \end{bmatrix}^T$ with $\det(\Sigma) > 0$, then the random vector $X_1|X_2 = x_2$, is Normal with mean

$$\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

and covariance

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Proof. Define

$$A = \begin{bmatrix} I_{q\times q} & \vdots & -\Sigma_{12}\Sigma_{22}^{-1} \\ \cdots & \vdots & \cdots \\ 0_{(p-q)\times q} & \vdots & I_{(p-q)\times(p-q)} \end{bmatrix}$$

Then

$$A(X-\mu) \sim \mathcal{N}_p\left(0, \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & \Sigma_{22} \end{bmatrix}\right)$$

Since

$$A(X-\mu) = \begin{bmatrix} X_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \\ \cdots \\ X_2 - \mu_2 \end{bmatrix}$$

and $X_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \sim \mathcal{N}_q(0, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$, then if we condition on $X_2 = x_2$, we have $X_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \sim \mathcal{N}_q(0, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ as well. By property 1, we have

$$X_1|X_2 = x_2 \sim \mathcal{N}_q\left(\mu_1 + \sigma_{12}\sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

as required. ■

3.3 Constant Probability Density Contours

Let $X \sim \mathcal{N}_p(\mu, \Sigma)$ so

$$f(x) = \frac{1}{\det(2\pi\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

The constant probability density contour is the set

$$\{x : (x-\mu)^T\Sigma^{-1}(x-\mu) = c^2\}$$

which is the surface of an ellipsoid centered at μ. This means the axes of the ellipsoid are $\pm c\sqrt{\lambda_i}e_i$ where (λ_i, e_i) are eigenvalue/eigenvector pairs of Σ for all $i \in \{1, \ldots, p\}$.

3.4 Distribution of $(X-\mu)^T\Sigma^{-1}(X-\mu)$

Let $X \sim \mathcal{N}_p(\mu, \Sigma)$ with $\det(\Sigma) > 0$. Then

$$(X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi^2_p$$

Thus, the probability $1-\alpha$ is assigned to

$$\{x : (x-\mu)^T\Sigma^{-1}(x-\mu) \leq \chi^2_p(\alpha)\}$$

where $\chi^2_p(\alpha)$ is the 100αth percentile of χ^2_p.

3.5 Sampling from the Multivariate Normal Distribution and MLE

Theorem (Multivariate CLT)**.** Let $X_1, \ldots, X_n$ be independent samples from a population with mean vector μ and covariance Σ. Then

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}_p(0, \Sigma)$$

3.5.1 Multivariate Normal Likelihood

Let $X_1, \ldots, X_n$ be random samples from $\mathcal{N}_p(\mu, \Sigma)$. Then their joint density function is

$$\prod_{j=1}^{n} \frac{1}{|\det(2\pi\Sigma)|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)\right]$$
$$= \frac{1}{|\det(2\pi\Sigma)|^{\frac{n}{2}}} \exp\left[-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)\right]$$

Fact: Let $A \in \mathbb{R}^{k\times k}$ be symmetric. Then

$$x^TAx = \text{tr}(x^TAx) = \text{tr}(Axx^T)$$

Claim. $\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) = \text{tr}\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\mu)(x_j-\mu)^T\right)\right]$

Proof. By the fact above,

$$\begin{aligned}\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) &= \sum_{j=1}^{n}\text{tr}\left(\Sigma^{-1}(x_j-\mu)(x_j-\mu)^T\right)\\ &= \text{tr}\left(\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\mu)(x_j-\mu)^T\right)\right) \quad \text{by distributivity}\end{aligned}$$

■

Furthermore,

$$\sum_{j=1}^{n}(x_j-\mu)(x_j-\mu)^T = \sum_{j=1}^{n}(x_j-\bar{x})(x_j-\mu)^T + n(\bar{x}-\mu)(\bar{x}-\mu)^T$$

Thus, the likelihood function of (μ,Σ) is

$$L(\mu,\Sigma) = \frac{1}{|\det(2\pi\Sigma)|^{\frac{n}{2}}}\exp\left[-\frac{1}{2}\text{tr}\left(\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^T\right)\right) - \frac{n}{2}(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)\right]$$

Optimizing this function tells us that the MLEs of μ and Σ respectively are

$$\hat{\mu} = \bar{X} \qquad \hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}(X_j-\bar{X})(X_j-\bar{X})^T = \frac{n-1}{n}S$$

Lemma. Given a $p\times p$ symmetric positive definite matrix B and a scalar $b>0$, it follows that

$$\frac{1}{|\Sigma|^b}e^{-\text{tr}\left(\Sigma^{-1}B\right)/2} \le \frac{1}{|B|^b}(2b)^{pb}e^{-bp}$$

for all positive definite $\Sigma_{p\times p}$, with equality when

$$\Sigma = \frac{1}{2b}B$$

4 Linear Regression

4.1 Least Squares Estimation

4.1.1 Univariate

Consider the model

$$Y = \beta_0 + \beta_1 Z + \varepsilon$$

By Least Squares, we want to minimize

$$S(b_0, b_1) = \sum_{j=1}^{n} (y_j - b_o - b_1 z_j)^2$$

assuming data points $(z_1, y_1), \ldots, (z_n, y_n)$ and where the b_0, b_1 are trial values of β_0, β_1. Setting

$$\frac{\partial S}{\partial b_0} = 0 = \frac{\partial S}{\partial b_1}$$

we have

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{z} \qquad \hat{\beta}_1 = \frac{\sum_{j=1}^{n} (y_j - \bar{y})(z_j - \bar{z})}{\sum_{j=1}^{n} (z_j - \bar{z})^2}$$

where $\hat{\beta}_0, \hat{\beta}_1$ are the Least Squares Estimates of β_0, β_1.

4.1.2 Multivariate

Consider the model

$$Y = \beta_0 + \beta_1 Z_1 + \cdots + \beta_r Z_r + \varepsilon$$

From n observations on Y, the model becomes

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 z_{11} + \cdots + \beta_r z_{1r} + \varepsilon_1 \\ &\vdots \\ Y_n &= \beta_n + \beta_1 z_{n1} + \cdots + \beta_r z_{nr} + \varepsilon_n \end{aligned}$$

Make the following assumptions:

1. $E[\varepsilon_i] = 0$
2. $\text{Var}[\varepsilon_i] = \sigma^2$
3. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$

In matrix notation,

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

or equivalently,

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & \cdots & z_{1r} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

By the assumptions above,

- $E[\boldsymbol{\varepsilon}] = 0$
- $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I$

Let $\boldsymbol{b}$ be a trial value of $\boldsymbol{\beta}$. We want to minimize

$$S(\boldsymbol{b}) = \sum_{j=1}^{n}(y_j - b_0 - b_1 z_{j1} - \cdots - b_r z_{jr})^2 = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b})$$

which will produce the LSE $\hat{\boldsymbol{\beta}}$.

Definition. The deviations

$$\hat{\varepsilon}_j = y_j - \hat{\beta}_0 - \hat{\beta}_1 z_{j1} - \cdots - \hat{\beta}_r z_{jr}$$

are the residuals.

Note that

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}$$

Suppose $\boldsymbol{Z}$ has full rank with $r + 1 \leq n$.

Claim. The LSE $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}$$

Proof. By definition,

$$\begin{aligned} S(\boldsymbol{b}) &= (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b}) \\ &= (\boldsymbol{y}^T - (\boldsymbol{Z}\boldsymbol{b})^T)(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{b} - (\boldsymbol{Z}\boldsymbol{b})^T\boldsymbol{y} - (\boldsymbol{Z}\boldsymbol{b})^T(\boldsymbol{Z}\boldsymbol{b}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{b} - \boldsymbol{b}^T\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{b} \end{aligned}$$

Then taking the partial derivatives,

$$\frac{\partial S}{\partial \boldsymbol{b}} = -2\boldsymbol{Z}^T\boldsymbol{y} + 2(\boldsymbol{Z}^T\boldsymbol{Z})\boldsymbol{b} = 0 \implies \hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}\boldsymbol{y}$$

■

Definition. The hat matrix is the matrix

$$\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$$

Note that H is both idempotent and symmetric, thus H is an orthogonal projector.

It follows that

$$\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y}$$

where $\hat{\boldsymbol{y}}$ is the fitted value of $\boldsymbol{y}$ and satisfies $\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}$. This implies that $\boldsymbol{H}$ projects onto the space spanned by the columns of $\boldsymbol{Z}$, which is the set of all linear combinations

of the predictors. Additionally,

$$\boldsymbol{Z}^T(I - \boldsymbol{H}) = 0$$

Finally, the residuals $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ satisfy $\boldsymbol{Z}^T\hat{\boldsymbol{\varepsilon}} = 0$, $\boldsymbol{y}^T\hat{\boldsymbol{\varepsilon}} = 0$.

4.1.3 Sampling Properties of the LSE

For $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}$,

- $E(\hat{\boldsymbol{\beta}}) = \beta$
- $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$

The residuals have properties

- $E(\hat{\boldsymbol{\varepsilon}}) = 0$
- $\text{Cov}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(I - \boldsymbol{H})$

4.2 Sum of Squares Decomposition

Definition. The Residual Sum of Squares (RSS) is as in

$$RSS = \sum_{j=1}^{n}(y_j - \beta_0 - \beta_1 z_{j1} - \cdots - \beta_r z_{jr})^2 = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{Z}\hat{\boldsymbol{\beta}}$$

Definition. The Total Sum of Squares is defined as

$$\boldsymbol{y}^T\boldsymbol{y} = (\hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}})^T(\hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}) = \hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}$$

Additionally, since

$$\bar{y} = \bar{\hat{y}}$$

then

$$\boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2 = \hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} - n(\bar{\hat{y}})^2 + \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}$$

which gives

$$\sum_{j=1}^{n}(y_j - \bar{y})^2 = \sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2 + \sum_{j=1}^{n}\hat{\varepsilon}_j^2$$

Definition. The coefficient of determination is

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{j=1}^{n}\hat{\varepsilon}_j^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2} \\ &= \frac{\sum_{j=1}^{n}(y_j - \bar{y})^2}{\sum_{j=1}^{n}(y_j - \bar{y}_j)^2} \end{aligned}$$

5 Principal Component Analysis

Lemma. Let B be a $p \times p$ symmetric positive definite matrix with eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \leq \lambda_p > 0$$

with associated eigenevectors $e_1, \ldots, e_p$. Then

$$\begin{aligned}
\max_{x \neq 0} \frac{x^T B x}{x^T x} &= \lambda_1 \text{ at } x = e_1 \\
\min_{x \neq 0} \frac{x^T B x}{x^T x} &= \lambda_p \text{ at } x = e_p \\
\max_{x \perp e_1, \ldots, e_k} \frac{x^T B x}{x^T x} &= \lambda_{k+1} \text{ at } x = e_{k+1}
\end{aligned}$$

5.1 PCA Formulation

Let random vector $X^T = \begin{bmatrix} X_1 & \cdots & X_p \end{bmatrix}$ have covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and corresponding normalized eigenvectors $e_1, \ldots, e_p$. Consider linear combinations

$$\begin{aligned}
Y_1 &= a_1^T X = a_{11} X_1 + \cdots + a_{1p} X_p \\
&\vdots \\
Y_p &= a_p^T X = a_{p1} X_1 + \cdots + a_{pp} X_p
\end{aligned}$$

thus

- $\text{Var}(Y_i) = a_i^T \Sigma a_i$
- $\text{Cov}(Y_i, Y_j) = a_i^T \Sigma a_j$

Definition. The principal components are those uncorrelated linear combinations $Y_1, Y_2, \ldots$ whose variances are as large as possible subject to $a_i^T a_i = 1$.

In other words, the ith component is the linear combination $a_i^T X$ that maximizes $\text{Var}(a_i^T X)$ subject to $a_i^T a_i = 1$ and $\text{Cov}(a_i^T X, a_j^T X) = 0$ for all $j < i$. From the lemma, this implies that the ith component is given by setting $a_i = e_i$ since $e_i^T e_i = 0$, $e_i \perp e_j$ for all $j < i$ and $\text{Cov}(a_i^T X, a_j^T X) = a_i^T \Sigma a_j$ where Σ is symmetric positive definite. Additionally, it is implied that $\text{Var}(e_i^T X) = \lambda_i$.

The ith principal component is

$$Y_i = e_i^T X$$

5.2 Total Population Variance

Consider the principal components $Y_i = e_i^T X$. By definition

$$\text{tr}\,(\Sigma) = \sum_{i=1}^{p} \text{Var}(X_i)$$

Since the e_i are normalized, then $e_i e_i^T = I$, thus $\Sigma = \Sigma e_i e^T$. Since Σ is symmetric, then $e_i^T \Sigma e_i = \text{tr}\left(e_i^T \Sigma e_i\right) = \text{tr}\left(\Sigma e_i e_i^T\right)$, thus

$$\text{tr}\,(\Sigma) = \text{tr}\left(\Sigma e_i e_i^T\right) = \text{tr}\left(e_i^T \Sigma e_i\right) = \sum_{i=1}^{p} \text{Var}(Y_i) = \sum_{i=1}^{p} \lambda_i$$

Define the total population variance as

$$\begin{aligned}\text{tr}\,(\Sigma) &= \sum_{i=1}^{p} \sigma_{ii} \\ &= \sum_{i=1}^{p} \lambda_i\end{aligned}$$

The **proportion of the total population variance due to the *k*th principal component** is

$$\frac{\lambda_k}{\lambda_1 + \cdots + \lambda_p}$$

for $k = 1, \ldots, p$. To choose how many components to retain, we consider the magnitude of an eigenvalue vs its number (i.e.: $|\lambda_i|$ vs i). This is so we can keep the predictors that have a significant explanation for population variance and can remove the ones with a weaker explanation (prevents over fitting).

5.3 Sample Variance by PC

Let the data $x_1, \ldots, x_n$ be n independent drawings from a p-dimensional population with sample covariance matrix S with eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ and associated normalized eigenvectors $\hat{e}_1, \ldots, \hat{e}_p$. The first PC maximizes

$$\frac{a_1^T S a_1}{a_1^T a_1}$$

which is equal to $\hat{\lambda}_i$ and attained when $a_1 = \hat{e}_1$. It follows that the ith PC is

$$\hat{y}_i = \hat{e}_i^T x$$

where sample covariance of $(\hat{y}_i, \hat{y}_k)$ is 0 (i.e.: maintain uncorrelated errors).