Linear Models

Emilija Perković [1]

Dept. of Statistics
University of Washington

[1](Based on slides by Mathias Drton)

Linear model

- Data:

	Y	X_1	$\dots$	X_p
1	Y_1	x_{11}	$\dots$	x_{1p}
2	Y_2	x_{21}	$\dots$	x_{2p}
3	Y_3	x_{31}	$\dots$	x_{3p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	Y_n	x_{n1}	$\dots$	x_{np}

- $Y_1, \dots, Y_n$ are observations of a response and x_{ij} are features of the experimental units (including which treatment was applied).
- Linear model postulates

$$Y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \qquad i = 1, \dots, n,$$

where β is a vector of mean parameters and the ϵ_i are error terms with $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2$.
- When discussing normal population models, we will take $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Matrix setup

- Response and error vector

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- Design matrix

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

- Model in vector form (with error vector):

$$Y = X\beta + \epsilon$$

Covariance matrix

Definition

Let $Y = (Y_1, \dots, Y_p)$ be a random vector. The expectation of Y is the vector

$$\mathbb{E}[Y] = \begin{pmatrix} \mathbb{E}[Y_1] \\ \vdots \\ \mathbb{E}[Y_p] \end{pmatrix}.$$

The covariance matrix of Y in $\mathbb{R}^p$ is the symmetric matrix

$$\begin{aligned} \mathrm{Var}[Y] &= \mathbb{E}\big[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^{\mathsf{T}}\big] \\ &= \begin{pmatrix} \mathrm{Var}[Y_1] & Cov[Y_1, Y_2] & \cdots & Cov[Y_1, Y_p] \\ Cov[Y_1, Y_2] & \mathrm{Var}[Y_2] & \cdots & Cov[Y_2, Y_p] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[Y_1, Y_p] & Cov[Y_2, Y_p] & \cdots & \mathrm{Var}[Y_p] \end{pmatrix}. \end{aligned}$$

(The above expectation of a matrix is, as for a vector, taken componentwise.)

Covariance matrices

- If $A \in \mathbb{R}^{k\times p}$ and $b \in \mathbb{R}^k$ then

$$\mathbb{E}[AY + b] = A \cdot \mathbb{E}[Y] + b,$$
$$\mathrm{Var}[AY + b] = A \cdot \mathrm{Var}[Y] \cdot A^{\mathsf{T}}.$$

- A covariance matrix is positive semidefinite (all eigenvalues ≥ 0):

$$a^{\mathsf{T}}\mathrm{Var}[Y]a = \mathrm{Var}[a^{\mathsf{T}} Y] \geq 0 \qquad \forall a \in \mathbb{R}^p.$$

It is positive definite (all eigenvalues > 0) if $a^{\mathsf{T}}\mathrm{Var}[Y]a > 0$ for $a \neq 0$.

Least squares

Definition

A least squares estimator $\hat{\beta}$ is a choice of β that minimizes the sum of squared errors

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 = (Y - X\beta)^{\mathsf{T}}(Y - X\beta) = \|Y - X\beta\|^2.$$

- Gradient:

$$\frac{\partial}{\partial\beta}\|Y - X\beta\|^2 = -2X^{\mathsf{T}}(Y - X\beta) = 0 \qquad \Longleftrightarrow \qquad X^{\mathsf{T}}X\beta = X^{\mathsf{T}}Y$$

- If X has full column rank ($p \leq n$), the above normal equations have the unique solution

$$\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y.$$

Fitted values, residuals, hat matrix

Fitted values:

$$\hat{y} = X\hat{\beta} = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}y \in \mathbb{R}^n$$

Residuals:

$$e = y - \hat{y} = [I_n - X(X^\mathsf{T}X)^{-1}X^\mathsf{T}]y \in \mathbb{R}^n$$

Hat matrix ($\hat{y} = Hy$):

$$H = X(X^\mathsf{T}X)^{-1}X^\mathsf{T} \in \mathbb{R}^{n\times n}$$

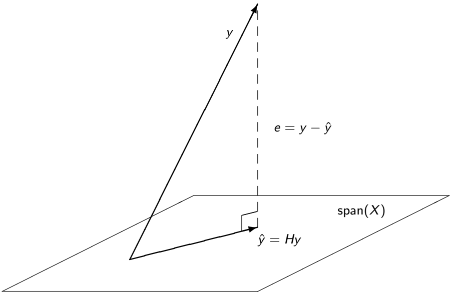
Proposition

The vector of residuals e is orthogonal to all vectors in $\mathcal{L}$, the column span of the design matrix X, $\mathcal{L} = \text{span}(X) = \{\, X\beta \;:\; \beta \in \mathbb{R}^p \,\}$. In particular, $e \perp \hat{y}$.

Proof.

Since $X^\mathsf{T}e = X^\mathsf{T}(y - X\hat{\beta}) = 0$, we have $e^\mathsf{T}X\alpha = 0$ for all $\alpha \in \mathbb{R}^p$. □

Geometry of least squares



Geometric view of linear models

- Model:
$$\mathbb{E}[Y] \in \mathcal{L}, \qquad \text{where } \mathcal{L} \subset \mathbb{R}^n \text{ is a linear space.}$$
- Fitted values $\hat{y}$ obtained by orthogonal projection onto $\mathcal{L}$, that is,
$$\hat{y} = \arg\min_{\mu \in \mathcal{L}} \|y - \mu\|^2.$$
- Fix a basis $\{x_1, \ldots, x_p\}$ of $\mathcal{L}$. Then the LSE $\hat{\beta}$ is the unique coefficient vector when writing $\hat{y}$ as a linear combination of $x_1, \ldots, x_p$:
$$\hat{y} = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p.$$
- Reference:

  Michael Wichura (2006). *The Coordinate-Free Approach to Linear Models.* Cambridge University Press.

Orthogonal projection

Theorem

Let y be any vector in $\mathbb{R}^n$, and suppose $\mathcal{L} \subset \mathbb{R}^n$ is a linear subspace.

(i) *There is a unique vector $\pi_{\mathcal{L}}(y) \in \mathcal{L}$ s.t. $y - \pi_{\mathcal{L}}(y) \perp v$ for all $v \in \mathcal{L}$.*

(ii) *A vector $v \in \mathcal{L}$ satisfies*
$$\|y - v\| = \min_{w \in \mathcal{L}} \|y - w\|$$
if and only if $v = \pi_{\mathcal{L}}(y)$.

Definition

The map $\pi_{\mathcal{L}} : \mathbb{R}^n \to \mathcal{L}$ is the orthogonal projection onto $\mathcal{L}$.

Orthogonal projection – proof

Proof.

(i) *Existence*: Pick an orthonormal basis $u_1, \dots, u_n$ of $\mathbb{R}^n$ such that $\mathcal{L} = \langle u_1, \dots, u_k \rangle$. Write $y = \sum_{i=1}^n \beta_i u_i$, and define $\pi_{\mathcal{L}}(y) = \sum_{i=1}^k \beta_i u_i$. Then the orthogonality claim follows because

$$y - \pi_{\mathcal{L}}(y) = \sum_{i=k+1}^{n} \beta_i u_i \perp u_1, \dots, u_k.$$

Uniqueness: If $v_1, v_2 \in \mathcal{L}$ satisfy that $y - v_1 \perp \mathcal{L}$ and $y - v_2 \perp \mathcal{L}$, then

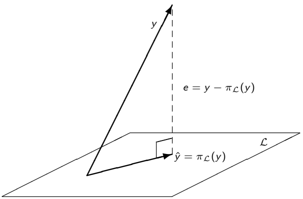
$$v_1 - v_2 = (y - v_1) - (y - v_2)$$

is a vector in $\mathcal{L}$ that is orthogonal to $\mathcal{L}$. It follows that $v_1 - v_2 = 0$.

(ii) Pythagoras:

$$\|y - v\|^2 = \|y - \pi_{\mathcal{L}}(y)\|^2 + \|\pi_{\mathcal{L}}(y) - v\|^2. \qquad \square$$

Geometry of least squares (again)



- Fitted values $\hat{y}$ and residuals e are always unique.
- Statistics that depend only on fitted values and residuals remain the same when changing design matrix X to $\tilde{X}$ with $\text{span}(X) = \text{span}(\tilde{X})$.

Properties of orthogonal projection

Lemma

(i) *The orthogonal projection* $\pi_{\mathcal{L}}$ *is a linear map.*

(ii) *Let* $\mathcal{L}^{\perp} = \{y \in \mathbb{R}^n : y \perp \mathcal{L}\}$ *be the orthogonal complement. Then*

$$\pi_{\mathcal{L}^{\perp}}(y) = y - \pi_{\mathcal{L}}(y).$$

(iii) *If* $\mathcal{L} = span(X)$ *for a matrix* $X \in \mathbb{R}^{n \times p}$ *of full column rank, then*

$$\pi_{\mathcal{L}}(y) \quad = \quad X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y.$$

(iv) *Let* $P \in \mathbb{R}^{n \times n}$*. The linear map* $y \mapsto Py$ *is an orthogonal projection if and only if*

$$P = P^2, \; P = P^{\mathsf{T}}.$$

In this case, $P = \pi_{span(P)}$*, and all eigenvalues of* P *are in* $\{0, 1\}$.

(v) *If* Q *is an orthogonal matrix then* $\pi_{Q\mathcal{L}}(Qy) = Q\pi_{\mathcal{L}}(y)$.

A matrix $Q \in \mathbb{R}^{p \times p}$ *is* **orthogonal** *if* $QQ^{\mathsf{T}} = Q^{\mathsf{T}}Q = I$ *such that*

$$\langle Qx, Qy \rangle = x^{\mathsf{T}}Q^{\mathsf{T}}Qy = \langle x, y \rangle \qquad \forall x, y \in \mathbb{R}^p.$$

Proof.

(i) Follows from uniqueness of projections because

$$(\lambda y_1 + y_2) - [\lambda \pi_{\mathcal{L}}(y_1) + \pi_{\mathcal{L}}(y_2)] = \lambda[y_1 - \pi_{\mathcal{L}}(y_1)] + [y_2 - \pi_{\mathcal{L}}(y_2)] \perp \mathcal{L}.$$

(ii) Similar. (iii) See derivation of LSE.

(iv) $(\Rightarrow)$: First, $P^2 = P$ because $\pi_{\mathcal{L}} \circ \pi_{\mathcal{L}}(y) = \pi_{\mathcal{L}}(y)$ for all $y \in \mathbb{R}^n$. Second, $P^{\mathsf{T}} = P$ because, for all $y, z \in \mathbb{R}^n$,

$$\begin{aligned} y^{\mathsf{T}}Pz &= \left[y - \pi_{\mathcal{L}}(y) + \pi_{\mathcal{L}}(y)\right]^{\mathsf{T}}\pi_{\mathcal{L}}(z) = \pi_{\mathcal{L}}(y)^{\mathsf{T}}\pi_{\mathcal{L}}(z) \\ &= \pi_{\mathcal{L}}(y)^{\mathsf{T}}\left[\pi_{\mathcal{L}}(z) - z + z\right] = \pi_{\mathcal{L}}(y)^{\mathsf{T}}z = y^{\mathsf{T}}P^{\mathsf{T}}z. \end{aligned}$$

$(\Leftarrow)$: Follows from eigenvalue fact and property (v).

Eigenvalues: All vectors $v \in \text{span}(P)$ are eigenvectors for eigenvalue 1 because $Pv = P(P\beta) = P\beta = v$. The orthogonal complement $\text{span}(P)^{\perp} = \text{kernel}(P)$ contains only eigenvectors for eigenvalue 0.

(v) Follows because $Qy - Qz \perp Q\mathcal{L}$ iff $y - z \perp \mathcal{L}$. (Recall $\langle Qx, Qy \rangle = \langle x, y \rangle$ if Q orthogonal.) $\square$

Unbiased estimation

Define the residual sum of squares

$$\mathrm{SSE} = \|e\|^2 = \sum_{i=1}^{n} e_i^2.$$

Theorem

If $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}[\epsilon] = \sigma^2 I_n$, then the least squares estimator

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

is an unbiased estimator of β. Moreover,

$$\hat{\sigma}^2 = \frac{1}{n-p}\mathrm{SSE}$$

is an unbiased estimator of σ^2.

Lemma about quadratic forms

- $tr(A)$ - Trace of a matrix A, $tr(A) = \sum_i a_{ii}$
- properties: $tr(a \cdot A + b \cdot B) = a \cdot tr(A) + b \cdot tr(B)$, $tr(A \cdot B \cdot C) = tr(B \cdot C \cdot A) = tr(C \cdot B \cdot A)$

Lemma

Let Z be a random vector with $\mathbb{E}[Z] = \mu \in \mathbb{R}^p$ and $\mathrm{Var}[Z] = \Sigma \in \mathbb{R}^{p \times p}$. Let $A \in \mathbb{R}^{p \times p}$ be a matrix. Then,

$$\mathbb{E}[Z^\top A Z] = tr[A\Sigma] + \mu^\top A \mu$$

Proof.

Thinking of a real number as a 1×1 matrix, write

$$\mathbb{E}[Z^\top A Z] = \mathbb{E}[tr(Z^\top A Z)] = \mathbb{E}[tr(A Z Z^\top)].$$

Using the linearity of trace and expectation, we obtain

$$\begin{aligned}\mathbb{E}[Z^\top A Z] &= tr(A \cdot (\mathbb{E}[Z Z^\top])) = tr(A \cdot (\Sigma + \mu\mu^\top)) \\ &= tr(A\Sigma) + tr(A\mu\mu^\top) = tr(A\Sigma) + \mu^\top A \mu. \quad \square\end{aligned}$$

Proof of unbiasedness

Proof.

The first claim is easily verified,

$$\mathbb{E}[\hat{\beta}] = (X^\top X)^{-1} X^\top \mathbb{E}[Y] = (X^\top X)^{-1} X^\top X\beta = \beta.$$

For the second claim, note that

$$\mathrm{SSE} = e^\top e = Y^\top (I_n - H)^\top (I_n - H) Y = Y^\top (I_n - H) Y.$$

By the lemma, the expectation of this quadratic form in Y is

$$\begin{aligned}\mathbb{E}[\mathrm{SSE}] &= tr\big[\,(I_n - H)\mathrm{Var}[Y]\,\big] + (X\beta)^\top \cancel{(I_n - H)X\beta} \\ &= \sigma^2 \cdot tr(I_n - H) = \sigma^2 \cdot \Big[tr(I_n) - tr(X(X^\top X)^{-1} X^\top)\Big] = \sigma^2 (n - p).\end{aligned}$$

because $tr(H) = tr((X^\top X)^{-1} X^\top X) = tr(I_p) = p$, or simply, because H has eigenvalues 0 and 1 of multiplicities $n - p$ and p. □

Examples of linear hypotheses

- Linear model:

$$Y \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

- Some treatment effects are zero:

$$H_0 : \beta_1 = \beta_2 = 0$$

- Some treatment effects are equal:

$$H_0 : \beta_1 = \beta_2 = \beta_3$$

- But other hypotheses could be of interest:

$$H_0 : \beta_2 = 2\beta_1$$

General F-test

- Linear model:

$$Y = X\beta + \epsilon, \qquad \mathbb{E}[\epsilon] = 0, \quad \mathrm{Var}[\epsilon] = \sigma^2 I_n.$$

- Mean vector $\mu \in \mathbb{E}[Y]$ is contained in the linear space

$$\mathcal{L} = \mathrm{span}(X) = \{\, X\beta \;:\; \beta \in \mathbb{R}^p \,\}.$$

- Assume $n > p$, in which case $\mathrm{SSE} = \|\pi_{\mathcal{L}^\perp}(Y)\|^2 \neq 0$ with probability one.

- The examples on the previous slide involved null hypotheses that correspond to linear subspaces $\mathcal{H} \subsetneq \mathcal{L}$ of the form

$$\mathcal{H} = \{\, X\beta \;:\; \beta \in \mathbb{R}^p \text{ and } A\beta = 0 \,\}$$

  for different choices of a matrix A.
- In what follows, assume that $q = \dim(\mathcal{H}) < p = \dim(\mathcal{L})$.

Linear models Covariance matrices Least squares F-statistic Multivariate normal distribution Inference

General F-test

Definition

The statistic

$$F = \frac{\frac{1}{p-q}\,[\mathrm{SSE}(\text{reduced model}) - \mathrm{SSE}(\text{full model})]}{\frac{1}{n-p}\,\mathrm{SSE}(\text{full model})} = \frac{\frac{1}{p-q}\left(\|\pi_{\mathcal{H}^\perp}(Y)\|^2 - \|\pi_{\mathcal{L}^\perp}(Y)\|^2\right)}{\frac{1}{n-p}\|\pi_{\mathcal{L}^\perp}(Y)\|^2}$$

is the F-statistic for the testing problem

$$H_0 : \mu \in \mathcal{H} \quad \text{vs.} \quad H_1 : \mu \in \mathcal{L}\backslash\mathcal{H}.$$

- The F statistic can be used for randomization tests or for an F-test in inference based on normal population models.
- In normal population-based inference, the F-test rejects H_0 if $F > f_{p-q,n-p,\alpha}$, where $f_{p-q,n-p,\alpha}$ is the $1-\alpha$ quantile of the $F_{p-q,n-p}$ distribution. (Proof later)

Remarks on F-test

- F-test of $H_0 : \beta_2 = \beta_3 = 0$ in model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ can be computed using the R command

$$\texttt{anova}\,(\texttt{lm}(\texttt{y} \sim \texttt{x1}), \texttt{lm}(\texttt{y} \sim \texttt{x1} + \texttt{x2} + \texttt{x3}))$$

- How about the following R commands?

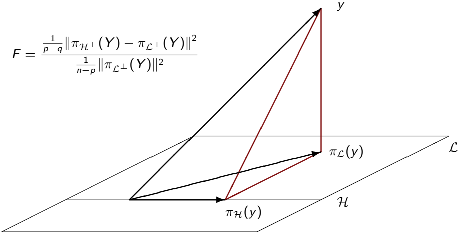
$$\texttt{z} = \texttt{x1} + \texttt{x2}$$
$$\texttt{anova}\,(\texttt{lm}(\texttt{y} \sim \texttt{z}), \texttt{lm}(\texttt{y} \sim \texttt{x1} + \texttt{x2} + \texttt{x3}))$$

- With this and any other test:

$$\text{Statistical significance} \not\Rightarrow \text{practical importance}$$

Geometry of the F-Test

$$F = \frac{\frac{1}{p-q}\|\pi_{\mathcal{H}^\perp}(Y) - \pi_{\mathcal{L}^\perp}(Y)\|^2}{\frac{1}{n-p}\|\pi_{\mathcal{L}^\perp}(Y)\|^2}$$



Note: $\pi_{\mathcal{H}^\perp}(y) - \pi_{\mathcal{L}^\perp}(y) = \pi_{\mathcal{L}}(y) - \pi_{\mathcal{H}}(y)$ and
$\|\pi_{\mathcal{H}^\perp}(y) - \pi_{\mathcal{L}^\perp}(y)\|^2 = \|\pi_{\mathcal{H}^\perp}(y)\|^2 - \|\pi_{\mathcal{L}^\perp}(y)\|^2 = \|\pi_{\mathcal{L}}(y)\|^2 - \|\pi_{\mathcal{H}}(y)\|^2$

Univariate normal distribution

- A random variable X has the **standard normal distribution**, in symbols, $X \sim N(0,1)$, if it has the density
$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \ \ x \in \mathbb{R}.$$
- For $a, b \in \mathbb{R}$, define $Y = aX + b$.
  - If $a = 0$, then $P(Y = b) = 1$, so Y is constant with probability 1.
  - If $a \neq 0$, then Y has density
$$f_Y(y) = \frac{1}{\sqrt{2\pi a^2}} e^{-\frac{(y-b)^2}{2a^2}}, \ \ y \in \mathbb{R}.$$
  - Note: distribution of Y depends only on a^2 and b.
- A r.v. Y has the **normal distribution** with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \geq 0$, denoted $N(\mu, \sigma^2)$, if Y has the same distribution as $\sigma X + \mu$.
  - Y has the familiar density
$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}, \ \ y \in \mathbb{R}.$$
  - As the names of the parameters suggest, $\mathbb{E}[Y] = \mu$ and $\mathrm{Var}[Y] = \sigma^2$.

Standard normal distribution in $\mathbb{R}^p$

- If $X_1, \dots, X_p \overset{iid}{\sim} N(0,1)$, then the random vector $X = (X_1, \dots, X_p)^\mathsf{T}$ is said to have the **(p-variate) standard normal distribution**.
- The joint density of X is
$$\phi_p(x) = \prod_{i=1}^{p} \left(\frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} \right) = \frac{1}{\sqrt{(2\pi)^p}} e^{-\sum_i x_i^2/2} = \frac{1}{\sqrt{(2\pi)^p}} e^{-\|x\|^2/2}.$$
- A matrix $Q \in \mathbb{R}^{p\times p}$ is **orthogonal** if $QQ^\mathsf{T} = Q^\mathsf{T}Q = I$ such that
$$\langle Qx, Qy \rangle = x^\mathsf{T} Q^\mathsf{T} Q y = \langle x, y \rangle \qquad \forall x, y \in \mathbb{R}^p.$$

Lemma (Orthogonal invariance)

If $Q \in \mathbb{R}^{p\times p}$ is orthogonal and X is standard normal in $\mathbb{R}^p$, then QX is also p-variate standard normal.

Proof. Since $\det(Q) = \det(Q^\mathsf{T}) = \pm 1$, the random vector $Y = QX$ has density
$$f_Y(y) = f_X(Q^\mathsf{T} y) \cdot \left|\det(Q^\mathsf{T})\right| \propto e^{-\|y\|^2/2}.$$

Multivariate normal distribution

As in univariate case, define general normal distribution via affine transformations.

Definition

A random vector Y in $\mathbb{R}^p$ follows a **multivariate normal distribution** if there exists a k-variate standard normal random vector X such that

$$Y = AX + b,$$

for some matrix $A \in \mathbb{R}^{p\times k}$ and vector $b \in \mathbb{R}^p$.

Theorem

If X and Y are multivariate normal random vectors in $\mathbb{R}^p$ with $\mathbb{E}[X] = \mathbb{E}[Y]$ and $\mathrm{Var}[X] = \mathrm{Var}[Y]$ then X and Y have the same distribution.

Notation

We write $N_p(\mu, \Sigma)$ to denote the p-variate normal distribution with mean vector μ and covariance matrix Σ. If X is standard normal in $\mathbb{R}^p$, then $\mu = 0$ and $\Sigma = I_p$ is the identity matrix.

Mean vector and cov. matrix determine normal distribution

Proof. Suppose $X = AU + b$ and $Y = CV + d$ for two standard normal random vectors U and V.

(i) First, observe that $b = \mathbb{E}[X] = \mathbb{E}[Y] = d$. WLOG, let $b = d = 0$.

(ii) Adding zero-columns if necessary, assume that $A, C \in \mathbb{R}^{p\times k}$.

(iii) One can show that since $AA^{\mathsf{T}} = \mathrm{Var}[X] = \mathrm{Var}[Y] = CC^{\mathsf{T}}$, there exists a $k \times k$ orthogonal matrix Q s.t. $C = AQ$. The theorem is then proven because QV is standard normal.

If $k \le p$ and $\mathrm{rank}(A) = k$, we may take $Q = A^{\mathsf{T}}C(C^{\mathsf{T}}C)^{-1}$. Indeed,

$$AA^{\mathsf{T}} = CC^{\mathsf{T}} \implies AA^{\mathsf{T}}\times C(C^{\mathsf{T}}C)^{-1} = CC^{\mathsf{T}}\times C(C^{\mathsf{T}}C)^{-1} \implies AQ = C,$$

and Q is orthogonal because

$$AQ = C \implies (C^{\mathsf{T}}C)^{-1}C^{\mathsf{T}} \times AQ = (C^{\mathsf{T}}C)^{-1}C^{\mathsf{T}} \times C \implies Q^{\mathsf{T}}Q = I.$$

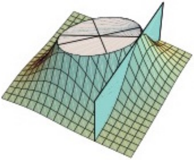
General case/geometry (for students with mathematical background): Proposition 12.13 in the book ‘Brownian Motion’ by Mörters & Peres. □

Density

Theorem

If Σ is positive definite, then $N_p(\mu, \Sigma)$ has joint density

$$f_{\mu,\Sigma}(x) = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} \times \exp\left\{ -\frac{1}{2}(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)\right\}.$$



(Galton, 1886, $p = 2$)

Proof. Represent $Y \sim N_p(\mu, \Sigma)$ as $Y = AX + \mu$, where $X \sim N_p(0, I_p)$ is standard normal and $A \in \mathbb{R}^{p\times p}$ invertible. Then

$$\begin{aligned} f_Y(y) &= f_X(A^{-1}(y-\mu)) \cdot |\det(A^{-1})| \\ &= \frac{1}{\sqrt{(2\pi)^p}} \exp\left\{ -\frac{1}{2}(y-\mu)^\mathsf{T} A^{-\mathsf{T}} A^{-1}(y-\mu)\right\} \cdot \frac{1}{\det(AA^\mathsf{T})^{1/2}}. \end{aligned}$$

Now note that $\Sigma = AA^\mathsf{T}$ and $\Sigma^{-1} = A^{-\mathsf{T}}A^{-1}$. □

Linear models Covariance matrices Least squares F-statistic Multivariate normal distribution Inference

Linear transformations and marginal distribution

We have defined multivariate normal distribution by means of linear transformations of standard normal random vectors.

Lemma

If $X \sim N_p(\mu, \Sigma)$, then

$$AX + b \sim N_p(A\mu + b, A\Sigma A^\mathsf{T}).$$

Consider partitioned random vector

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

As a consequence of the above lemma with $A = (I, 0)$, it holds that:

Theorem

The marginal distribution of X_1 is normal, namely,

$$X_1 \sim N(\mu_1, \Sigma_{11}).$$

Independence

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Theorem

The subvectors X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$.

Proof.

($\Rightarrow$): Independence implies zero covariance.

($\Leftarrow$): We can choose A_1 and A_2 such that X has same distribution as

$$\begin{pmatrix} A_1 Z_1 \\ A_2 Z_2 \end{pmatrix} = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \quad \text{with} \quad \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N_p(0, I). \qquad \square$$

If X is not jointly normal, then $\Sigma_{12} = \Sigma_{21}^{\top} = 0$ does not imply independence of X_1 and X_2 (marginals can still be normal).

Distribution of Estimators

Theorem

In a linear model with normal distribution assumption (N):

$$Y \sim N_n\left(\mu, \sigma^2 I_n\right) \quad \textit{with} \quad \mu = X\beta,\ \beta \in \mathbb{R}^p,$$

the LS estimators $\hat{\beta}$ and $\hat{\sigma}^2$ are independent and distributed as

(i) $\hat{\beta} \sim N_p\left(\beta, \sigma^2 (X^\top X)^{-1}\right)$,

(ii) $\dfrac{\hat{\sigma}^2}{\sigma^2} \cdot (n-p) \sim \chi^2_{n-p}$.

Geometric intuition:

- $(\hat{Y}, e)$ linear transformation of Y and thus jointly multivariate normal;
- $\hat{\beta}$ is a linear function of $\hat{Y}$, and $\hat{\sigma}^2$ is a function of e;
- $\hat{Y} \perp e$ implies independence.

For the proof of (i), note that $\hat{\beta}$ is normal, we have shown $\mathbb{E}[\hat{\beta}] = \beta$ before, and

$$\mathrm{Var}[\hat{\beta}] = (X^\top X)^{-1} X^\top \mathrm{Var}[Y] X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}.$$

Proof.

Proof.

(ii) & Independence: Consider first the canonical case with design matrix

$$X = \begin{pmatrix} Z \\ 0 \end{pmatrix}, \quad \text{for some full rank matrix } Z \in \mathbb{R}^{p \times p}.$$

Then the orthogonal projection of a vector y onto

$$\mathrm{span}(X) = \{y \in \mathbb{R}^n \ : \ y_{p+1} = \cdots = y_n = 0\}$$

is $\pi_{\mathrm{span}(X)}(y) = (y_1, \ldots, y_p, 0, \ldots, 0)^\mathsf{T}$. Thus,

$$\hat{Y} = (Y_1, \ldots, Y_p, 0, \ldots, 0)^\mathsf{T} \quad \text{and} \quad e = (0, \ldots, 0, Y_{p+1}, \ldots, Y_n)^\mathsf{T}$$

Given the special form of the design matrix, $Y_i \sim N(0, \sigma^2)$ for all $i > p$. Therefore,

$$\mathrm{SSE} \ = \ \|Y - \hat{Y}\|^2 \ = \ \sigma^2 \sum_{i=p+1}^{n} \left(\frac{Y_i}{\sigma}\right)^2 \ \sim \ \sigma^2 \chi^2_{n-p}$$

Since $\hat{\beta}$ is a function of $(Y_1, \ldots, Y_p)$ and $\hat{\sigma}^2$ is a function of $(Y_{p+1}, \ldots, Y_n)$, the two estimators are independent. □

Proof

Proof.

In the general case with arbitrary full rank design matrix, there is an orthogonal matrix Q such that

$$QX = \begin{pmatrix} Z \\ 0 \end{pmatrix}, \quad \text{for some full rank matrix } Z \in \mathbb{R}^{p \times p}.$$

Define the rotated response $\tilde{Y} = QY \sim N(QX\beta, \sigma^2 I_n)$. Then $\tilde{Y}_i \sim N(0, \sigma^2)$ for all $i > p$. Consequently,

$$\begin{aligned} \mathrm{SSE} \ &= \ \|Y - \hat{Y}\|^2 \ = \ \|Q(Y - \pi_{\mathrm{span}(X)}(Y))\|^2 \\ &= \ \|QY - \pi_{Q\mathrm{span}(X)}(QY)\|^2 \ = \ \sum_{i=p+1}^{n} \tilde{Y}_i^2 \ \sim \ \sigma^2 \cdot \chi^2_{n-p}. \end{aligned}$$

Since $\hat{\beta}$ is a function of $(\tilde{Y}_1, \ldots, \tilde{Y}_p)$ and $\hat{\sigma}^2$ is a function of $(\tilde{Y}_{p+1}, \ldots, \tilde{Y}_n)$, the two estimators are independent. □

Distribution of standardized estimator

The variance of $\hat{\beta}_j$, the j-th diagonal entry of the cov. matrix $\mathrm{Var}[\hat{\beta}]$, is

$$\mathrm{Var}[\hat{\beta}_j] = \sigma^2 (X^\top X)^{-1}_{jj}.$$

Estimating σ^2 by $\hat{\sigma}^2$ we obtain the standard error

$$SE[\hat{\beta}_j] = \hat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}.$$

Theorem

Under the normal distribution assumption (N), the ratio

$$\frac{\hat{\beta}_j - \beta_j}{SE[\hat{\beta}_j]}$$

follows a t-distribution with $n-p$ degrees of freedom (denoted t_{n-p}).

Proof of theorem.

Recall that the t_m distribution is the distribution of a ratio

$$\frac{Z}{\sqrt{\frac{1}{m}W}}$$

where (i) $Z \sim N(0,1)$, (ii) $W \sim \chi^2_m$, and (iii) Z and W are independent.

In the present context, define the two independent random variables

$$Z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 (X^\top X)^{-1}_{jj}}} \sim N(0,1) \qquad \text{and} \qquad W = \frac{\hat{\sigma}^2}{\sigma^2}\cdot(n-p) \sim \chi^2_{n-p}.$$

Then the claimed t_{n-p} distribution holds because we can write

$$\frac{\hat{\beta}_j - \beta_j}{SE[\hat{\beta}_j]} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 (X^\top X)^{-1}_{jj}}}\cdot\sqrt{\frac{1}{\hat{\sigma}^2/\sigma^2}} = \frac{Z}{\sqrt{\frac{1}{n-p}W}}. \qquad \square$$

T-tests and confidence intervals

T-test for $H_0 : \beta_j = \beta_j^*$ vs. $H_1 : \beta_j \neq \beta_j^*$ uses statistic

$$T_j = \frac{\hat{\beta}_j - \beta_j^*}{SE[\hat{\beta}_j]},$$

which under H_0 follows a t_{n-p} distribution. The p-value for the test is

$$2P(t_{n-p} > |T_j|).$$

An exact $(1-\alpha)$-confidence interval for β_j is given by

$$\left(\hat{\beta}_j - t_{n-p,1-\alpha/2} \cdot SE[\hat{\beta}_j],\ \ \hat{\beta}_j + t_{n-p,1-\alpha/2} \cdot SE[\hat{\beta}_j]\right).$$

Here, $t_{n-p,1-\alpha/2}$ is the critical value defined by the equation

$$P(T < t_{n-p,1-\alpha/2}) = 1 - \alpha/2, \text{ and } P(T > t_{n-p,1-\alpha/2}) = \alpha/2,$$

for a random variable $T \sim t_{n-p}$.

Linear models Covariance matrices Least squares F-statistic Multivariate normal distribution Inference

Validity of confidence interval

Proof.

The random interval

$$\left(\hat{\beta}_j \pm t_{n-p,1-\alpha/2} \cdot SE[\hat{\beta}_j]\right)$$

contains the true parameter β_j if and only if

$$-t_{n-p,1-\alpha/2} < \frac{\hat{\beta}_j - \beta_j}{SE[\hat{\beta}_j]} < t_{n-p,1-\alpha/2}.$$

By the symmetry of the t distribution, the probability of the latter event is

$$1 - 2 \cdot P\left(\frac{\hat{\beta}_j - \beta_j}{SE[\hat{\beta}_j]} > t_{n-p,1-\alpha/2}\right) = 1 - 2 \cdot \alpha/2 = 1 - \alpha. \qquad \square$$

Distribution theory for the F-test

- Let $\mathcal{L} = \text{span}(X) = \{ X\beta \ : \ \beta \in \mathbb{R}^p \}$.
- Test $H_0 : \mu \in \mathcal{H}$ vs. $H_1 : \mu \in \mathcal{L} \backslash \mathcal{H}$ for a linear subspace $\mathcal{H} \subsetneq \mathcal{L}$.
- Suppose $p = \dim(\mathcal{L})$ and $q = \dim(\mathcal{H}) < p$.

Theorem

Under the normal distribution assumption (N), if $H_0 : \mu \in \mathcal{H}$ is true then the F-statistic for the testing problem has an $F_{p-q,n-p}$ distribution.

Under the alternative H_1, the F-statistic has a non-central F-distribution.

Proof.

Proof.

(a) Canonical case:

$$\mathcal{H} = \langle e_1, \dots, e_q \rangle = \{ y \in \mathbb{R}^n \ : \ y_i = 0 \ \forall i > q \}$$
$$\mathcal{L} = \langle e_1, \dots, e_p \rangle = \{ y \in \mathbb{R}^n \ : \ y_i = 0 \ \forall i > p \}$$

$$\pi_{\mathcal{H}}(Y) = \Big(Y_1, \dots, Y_q, 0, \dots, 0 \Big)^{\mathsf{T}},$$
$$\pi_{\mathcal{L}}(Y) = \Big(Y_1, \dots, Y_q, Y_{q+1}, \dots, Y_p, 0, \dots, 0 \Big)^{\mathsf{T}}.$$

Therefore,

$$F = \frac{\frac{1}{p-q} \left(\|\pi_{\mathcal{H}^\perp}(Y)\|^2 - \|\pi_{\mathcal{L}^\perp}(Y)\|^2 \right)}{\frac{1}{n-p} \|\pi_{\mathcal{L}^\perp}(Y)\|^2} = \frac{\frac{1}{p-q} \sum_{i=q+1}^{p} Y_i^2 / \sigma^2}{\frac{1}{n-p} \sum_{i=p+1}^{n} Y_i^2 / \sigma^2}$$

follows an $F_{p-q,n-p}$ distribution. □

Proof.

.

(b) General case:

Since $\mathcal{H} \subseteq \mathcal{L}$, there exists an orthogonal matrix $Q \in \mathbb{R}^{n \times n}$ such that

$$Q\mathcal{H} = \langle e_1, \ldots, e_q \rangle = \{y \in \mathbb{R}^n \ : \ y_i = 0 \ \forall i > q\}$$

$$Q\mathcal{L} = \langle e_1, \ldots, e_p \rangle = \{y \in \mathbb{R}^n \ : \ y_i = 0 \ \forall i > p\}$$

Orthogonal transformations preserve lengths and angles, e.g., $(Q\mathcal{H})^\perp = Q(\mathcal{H})^\perp$. Thus,

$$\begin{aligned} F &= \frac{\frac{1}{p-q}\left(\|\pi_{\mathcal{H}^\perp}(Y)\|^2 - \|\pi_{\mathcal{L}^\perp}(Y)\|^2\right)}{\frac{1}{n-p}\|\pi_{\mathcal{L}^\perp}(Y)\|^2} \\ &= \frac{\frac{1}{p-q}\left(\|\pi_{(Q\mathcal{H})^\perp}(QY)\|^2 - \|\pi_{(Q\mathcal{L})^\perp}(QY)\|^2\right)}{\frac{1}{n-p}\|\pi_{(Q\mathcal{L})^\perp}(QY)\|^2} \end{aligned}$$

is the F-statistic for the testing problem $H_0 : \tilde{\mu} \in Q\mathcal{H}$ vs. $\tilde{\mu} \in Q\mathcal{L}$ based on observation of $\tilde{Y} = QY \sim N(\tilde{\mu}, \sigma^2 \cdot I_n)$. This is the canonical case. $\square$