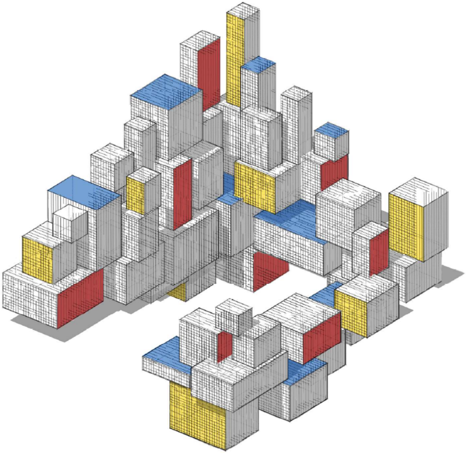
Normal Linear Models



Purpose

In this lecture we discuss:

- Properties of multivariate normal distributions
- Linear model + normal error terms = normal linear models.

Multivariate Normal Distribution

The multivariate normal (or Gaussian) distribution plays a central role in data science and machine learning.

Let $Z_1, \ldots, Z_n$ be independent and standard normal random variables. The joint pdf of $\boldsymbol{Z} = [Z_1, \ldots, Z_n]^\top$ is given by

$$f_{\mathbf{Z}}(\boldsymbol{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2} z_i^2} = (2\pi)^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2} \boldsymbol{z}^\top \boldsymbol{z}}, \quad \boldsymbol{z} \in \mathbb{R}^n.$$

We write $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix. Consider the affine transformation

$$\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{B}\,\boldsymbol{Z}$$

for some $m \times n$ matrix $\mathbf{B}$ and m-dimensional vector $\boldsymbol{\mu}$. Then $\boldsymbol{X}$ has a multivariate normal or *multivariate Gaussian* distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We write $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Affine Transforms are Again Normal

The following theorem states that any affine combination of independent multivariate normal random variables is again multivariate normal.

Theorem: Affine Transform of Normal Random Vectors

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_r$ be independent m_i-dimensional normal random vectors, with $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, \ldots, r$. Then, for any $n \times 1$ vector $\boldsymbol{a}$ and $n \times m_i$ matrices $\mathbf{B}_1, \ldots, \mathbf{B}_r$,

$$\boldsymbol{a} + \sum_{i=1}^{r} \mathbf{B}_i\, \boldsymbol{X}_i \sim \mathcal{N}\left(\boldsymbol{a} + \sum_{i=1}^{r} \mathbf{B}_i\, \boldsymbol{\mu}_i,\; \sum_{i=1}^{r} \mathbf{B}_i\, \boldsymbol{\Sigma}_i\, \mathbf{B}_i^\top\right).$$

Proof

Denote the n-dimensional random vector in the left-hand side of (**??**) by $\boldsymbol{Y}$. By definition, each $\boldsymbol{X}_i$ can be written as $\boldsymbol{\mu}_i + \mathbf{A}_i \boldsymbol{Z}_i$, where the $\{\boldsymbol{Z}_i\}$ are independent (because the $\{\boldsymbol{X}_i\}$ are independent), so that

$$\boldsymbol{Y} = \boldsymbol{a} + \sum_{i=1}^{r} \mathbf{B}_i \left(\boldsymbol{\mu}_i + \mathbf{A}_i \boldsymbol{Z}_i\right) = \boldsymbol{a} + \sum_{i=1}^{r} \mathbf{B}_i\, \boldsymbol{\mu}_i + \sum_{i=1}^{r} \mathbf{B}_i \mathbf{A}_i \boldsymbol{Z}_i,$$

which is an affine combination of independent standard normal random vectors. Hence, $\boldsymbol{Y}$ is multivariate normal.

By linearity of the expectation, we have $\mathbb{E}\boldsymbol{Y} = \boldsymbol{a} + \sum_{i=1}^{r} \mathbf{B}_i\, \boldsymbol{\mu}_i$. And the covariance matrix is

$$\mathbb{C}\text{ov}(\boldsymbol{Y}) = \sum_{i=1}^{r} \mathbb{C}\text{ov}(\mathbf{B}_i \mathbf{A}_i \boldsymbol{Z}_i) = \sum_{i=1}^{r} \mathbf{B}_i \mathbf{A}_i \mathbb{C}\text{ov}(\boldsymbol{Z}_i) \mathbf{A}_i^{\top} \mathbf{B}_i^{\top} = \sum_{i=1}^{r} \mathbf{B}_i \boldsymbol{\Sigma}_i \mathbf{B}_i^{\top}.$$

Marginals are Again Normal

The next theorem shows that the distribution of a subvector of a multivariate normal random vector is again normal.

Theorem: Marginal Distributions of Normal Vectors

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be an n-dimensional normal random vector. Decompose $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_p \\ \boldsymbol{X}_q \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_p \\ \boldsymbol{\mu}_q \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_p & \boldsymbol{\Sigma}_r \\ \boldsymbol{\Sigma}_r^\top & \boldsymbol{\Sigma}_q \end{bmatrix},$$

where $\boldsymbol{\Sigma}_p$ is the upper left $p \times p$ corner of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_q$ is the lower right $q \times q$ corner of $\boldsymbol{\Sigma}$. Then, $\boldsymbol{X}_p \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$.

Proof

We give a proof assuming that $\mathbf{\Sigma}$ is positive semidefinite. Let $\mathbf{B}\mathbf{B}^\top$ be the (lower) Cholesky decomposition of $\mathbf{\Sigma}$. We can write

$$\begin{bmatrix} \boldsymbol{X}_p \\ \boldsymbol{X}_q \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_p \\ \boldsymbol{\mu}_q \end{bmatrix} + \underbrace{\begin{bmatrix} \mathbf{B}_p & \mathbf{O} \\ \mathbf{C}_r & \mathbf{C}_q \end{bmatrix}}_{\mathbf{B}} \begin{bmatrix} \boldsymbol{Z}_p \\ \boldsymbol{Z}_q \end{bmatrix},$$

where $\boldsymbol{Z}_p$ and $\boldsymbol{Z}_q$ are independent p- and q-dimensional standard normal random vectors. In particular, $\boldsymbol{X}_p = \boldsymbol{\mu}_p + \mathbf{B}_p \boldsymbol{Z}_p$, which means that $\boldsymbol{X}_p \sim \mathcal{N}(\boldsymbol{\mu}_p, \mathbf{\Sigma}_p)$, since $\mathbf{B}_p \mathbf{B}_p^\top = \mathbf{\Sigma}_p$.

By relabeling the elements of $\boldsymbol{X}$ we see that the Theorem implies that *any* subvector of $\boldsymbol{X}$ has a multivariate normal distribution. For example, $\boldsymbol{X}_q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{\Sigma}_q)$.

Conditionals are Again Normal

Theorem: Conditional Distributions of Normal Vectors

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be an n-dimensional normal random vector with $\det(\boldsymbol{\Sigma}) > 0$. Decompose $\boldsymbol{X}$ as:

$$\begin{bmatrix} \boldsymbol{X}_p \\ \boldsymbol{X}_q \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_p \\ \boldsymbol{\mu}_q \end{bmatrix} + \underbrace{\begin{bmatrix} \mathbf{B}_p & \mathbf{O} \\ \mathbf{C}_r & \mathbf{C}_q \end{bmatrix}}_{\mathbf{B}} \begin{bmatrix} \boldsymbol{Z}_p \\ \boldsymbol{Z}_q \end{bmatrix},$$

Then

$$\left(\boldsymbol{X}_q \mid \boldsymbol{X}_p = \boldsymbol{x}_p\right) \sim \mathcal{N}(\boldsymbol{\mu}_q + \boldsymbol{\Sigma}_r^\top \boldsymbol{\Sigma}_p^{-1}(\boldsymbol{x}_p - \boldsymbol{\mu}_p),\ \boldsymbol{\Sigma}_q - \boldsymbol{\Sigma}_r^\top \boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}_r).$$

As a consequence, $\boldsymbol{X}_p$ and $\boldsymbol{X}_q$ are *independent* if and only if they are *uncorrelated*; that is, if $\boldsymbol{\Sigma}_r = \mathbf{O}$ (zero matrix).

Proof

From the decomposition we see that $\boldsymbol{X}_p = \boldsymbol{\mu}_p + \mathbf{B}_p \boldsymbol{Z}_p$ and $\boldsymbol{X}_q = \boldsymbol{\mu}_q + \mathbf{C}_r \boldsymbol{Z}_p + \mathbf{C}_q \boldsymbol{Z}_q$. Consequently,

$$(\boldsymbol{X}_q \mid \boldsymbol{X}_p = \boldsymbol{x}_p) = \boldsymbol{\mu}_q + \mathbf{C}_r\, \mathbf{B}_p^{-1}(\boldsymbol{x}_p - \boldsymbol{\mu}_p) + \mathbf{C}_q \boldsymbol{Z}_q,$$

where $\boldsymbol{Z}_q \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$. It follows that $\boldsymbol{X}_q$, conditional on $\boldsymbol{X}_p = \boldsymbol{x}_p$, has a $\mathcal{N}(\boldsymbol{\mu}_q + \mathbf{C}_r\, \mathbf{B}_p^{-1}(\boldsymbol{x}_p - \boldsymbol{\mu}_p), \mathbf{C}_q \mathbf{C}_q^\top)$ distribution.

The proof is completed by observing that

$$\boldsymbol{\Sigma}_r^\top \boldsymbol{\Sigma}_p^{-1} = \mathbf{C}_r \mathbf{B}_p^\top (\mathbf{B}_p^\top)^{-1} \mathbf{B}_p^{-1} = \mathbf{C}_r\, \mathbf{B}_p^{-1},$$

and

$$\boldsymbol{\Sigma}_q - \boldsymbol{\Sigma}_r^\top \boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}_r = \mathbf{C}_r \mathbf{C}_r^\top + \mathbf{C}_q \mathbf{C}_q^\top - \mathbf{C}_r \mathbf{B}_p^{-1} \underbrace{\boldsymbol{\Sigma}_r}_{\mathbf{B}_p \mathbf{C}_r^\top} = \mathbf{C}_q \mathbf{C}_q^\top.$$

The next few results are about the relationships between the normal, chi-squared, Student, and F distributions, defined in Table **??**. Recall that the chi-squared family of distributions, denoted by χ^2_n, are simply $\mathsf{Gamma}(n/2, 1/2)$ distributions, where the parameter $n \in \{1, 2, 3, \ldots\}$ is called the degrees of freedom.

Theorem: Normal and χ^2 Distributions

If $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is an n-dimensional normal random vector with $\det(\boldsymbol{\Sigma}) > 0$, then

$$(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2_n.$$

Proof

Let $\mathbf{B}\mathbf{B}^\top$ be the Cholesky decomposition of $\boldsymbol{\Sigma}$, where $\mathbf{B}$ is invertible. Since $\boldsymbol{X}$ can be written as $\boldsymbol{\mu} + \mathbf{B}\boldsymbol{Z}$, where $\boldsymbol{Z} = [Z_1, \ldots, Z_n]^\top$ is a vector of independent standard normal random variables, we have

$$(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) = (\boldsymbol{X} - \boldsymbol{\mu})^\top (\mathbf{B}\mathbf{B}^\top)^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{Z}^\top \boldsymbol{Z} = \sum_{i=1}^{n} Z_i^2.$$

Using the independence of $Z_1, \ldots, Z_n$, the moment generating function of $Y = \sum_{i=1}^n Z_i^2$ is given by

$$\mathbb{E}\,\mathrm{e}^{sY} = \mathbb{E}\,\mathrm{e}^{s(Z_1^2 + \cdots + Z_n^2)} = \mathbb{E}\,[\mathrm{e}^{sZ_1^2} \cdots \mathrm{e}^{sZ_n^2}] = \left(\mathbb{E}\,\mathrm{e}^{sZ^2}\right)^n,$$

where $Z \sim \mathcal{N}(0, 1)$. The moment generating function of Z^2 is

$$\mathbb{E}\,\mathrm{e}^{sZ^2} = \int_{-\infty}^{\infty} \mathrm{e}^{sz^2} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}(1-2s)z^2} \mathrm{d}z = \frac{1}{\sqrt{1-2s}},$$

so that $\mathbb{E}\, e^{sY} = \left(\frac{1}{2}/(\frac{1}{2} - s)\right)^{\frac{n}{2}}$, $s < \frac{1}{2}$, which is the moment generating function of the $\mathsf{Gamma}(n/2, 1/2)$ distribution; that is, the χ^2_n distribution. The proof is completed using the uniqueness of the moment generating function.

Consequently, if $\boldsymbol{X} = [X_1, \ldots, X_n]^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, then the squared length $\|\boldsymbol{X}\|^2 = X_1^2 + \cdots + X_n^2$ has a χ^2_n distribution.

If instead $X_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \ldots$, independently, then $\|\boldsymbol{X}\|^2$ is said to have a noncentral χ^2_n distribution.

This distribution depends on the $\{\mu_i\}$ only through the norm $\|\boldsymbol{\mu}\|$. We write $\|\boldsymbol{X}\|^2 \sim \chi^2_n(\theta)$, where $\theta = \|\boldsymbol{\mu}\|$ is the noncentrality parameter.

Projections of Normal Vectors

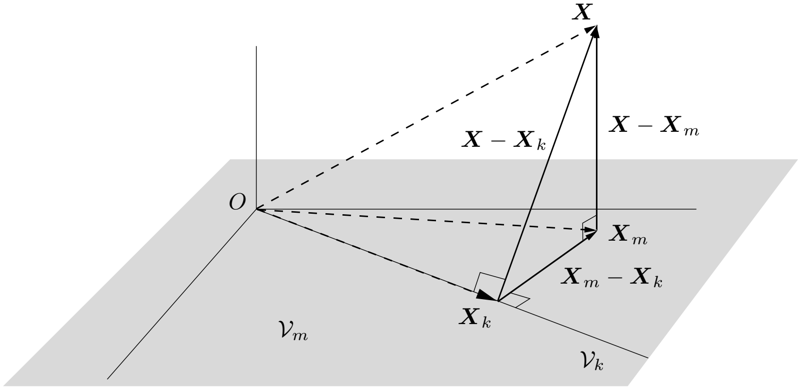
Non-central χ^2 distributions frequently occur when considering *projections* of multivariate normal random variables.

Theorem: Normal and Noncentral χ^2 Distributions

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_n)$ be an n-dimensional normal random vector and let $\mathcal{V}_k \subset \mathcal{V}_m$ be linear subspaces of dimensions k and m, respectively, with $k < m \leqslant n$. Let $\boldsymbol{X}_k$ and $\boldsymbol{X}_m$ be orthogonal projections of $\boldsymbol{X}$ onto $\mathcal{V}_k$ and $\mathcal{V}_m$, and let $\boldsymbol{\mu}_k$ and $\boldsymbol{\mu}_m$ be the corresponding projections of $\boldsymbol{\mu}$. Then, the following holds.

1. The random vectors $\boldsymbol{X}_k$, $\boldsymbol{X}_m - \boldsymbol{X}_k$, and $\boldsymbol{X} - \boldsymbol{X}_m$ are independent.
2. $\|\boldsymbol{X}_k\|^2 \sim \chi^2_k(\|\boldsymbol{\mu}_k\|)$, $\|\boldsymbol{X}_m - \boldsymbol{X}_k\|^2 \sim \chi^2_{m-k}(\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_k\|)$, and $\|\boldsymbol{X} - \boldsymbol{X}_m\|^2 \sim \chi^2_{n-m}(\|\boldsymbol{\mu} - \boldsymbol{\mu}_m\|)$.

Projections and Pythagoras



Proof

Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be an orthonormal basis of $\mathbb{R}^n$ such that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ spans $\mathcal{V}_k$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ spans $\mathcal{V}_m$.

We can write the orthogonal projection matrices onto $\mathcal{V}_j$, as $\mathbf{P}_j = \sum_{i=1}^{j} \boldsymbol{v}_i \boldsymbol{v}_i^\top$, $j = k, m, n$, where $\mathbf{P}_n$ is simply the identity matrix.

Let $\mathbf{V} := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n]$ and define $\boldsymbol{Z} := [Z_1, \ldots, Z_n]^\top = \mathbf{V}^\top \boldsymbol{X}$. Recall that any orthogonal transformation such as $\boldsymbol{z} = \mathbf{V}^\top \boldsymbol{x}$ is *length preserving*; that is, $\|\boldsymbol{z}\| = \|\boldsymbol{x}\|$.

To prove the first statement of the theorem, note that $\mathbf{V}^\top \boldsymbol{X}_j = \mathbf{V}^\top \mathbf{P}_j \boldsymbol{X} = [Z_1, \ldots, Z_j, 0, \ldots, 0]^\top$, $j = k, m$.

It follows that $\mathbf{V}^\top (\boldsymbol{X}_m - \boldsymbol{X}_k) = [0, \ldots, 0, Z_{k+1}, \ldots, Z_m, 0, \ldots, 0]^\top$ and $\mathbf{V}^\top (\boldsymbol{X} - \boldsymbol{X}_m) = [0, \ldots, 0, Z_{m+1}, \ldots, Z_n]^\top$.

Moreover, being a linear transformation of a normal random vector, $\boldsymbol{Z}$ is also normal, with covariance matrix $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_n$.

Proof

In particular, the $\{Z_i\}$ are *independent*. This shows that $\boldsymbol{X}_k$, $\boldsymbol{X}_m - \boldsymbol{X}_k$ and $\boldsymbol{X} - \boldsymbol{X}_m$ are independent as well.

Next, observe that $\|\boldsymbol{X}_k\| = \|\mathbf{V}^\top \boldsymbol{X}_k\| = \|\boldsymbol{Z}_k\|$, where $\boldsymbol{Z}_k := [Z_1, \ldots, Z_k]^\top$. The latter vector has independent components with variances 1, and its squared norm has therefore (by definition) a $\chi^2_k(\theta)$ distribution. The noncentrality parameter is

$$\theta = \|\mathbb{E}\boldsymbol{Z}_k\| = \|\mathbb{E}\boldsymbol{X}_k\| = \|\boldsymbol{\mu}_k\|,$$

again by the length-preserving property of orthogonal transformations. This shows that $\|\boldsymbol{X}_k\|^2 \sim \chi^2_k(\|\boldsymbol{\mu}_k\|)$.

The distributions of $\|\boldsymbol{X}_m - \boldsymbol{X}_k\|^2$ and $\|\boldsymbol{X} - \boldsymbol{X}_m\|^2$ follow by analogy.

Quotients of Squared Norms

The above projection theorem is frequently used in the statistical analysis of *normal linear models*. In typical situations $\boldsymbol{\mu}$ lies in the subspace $\mathcal{V}_m$ or even $\mathcal{V}_k$ — in which case $\|\boldsymbol{X}_m - \boldsymbol{X}_k\|^2 \sim \chi^2_{m-k}$ and $\|\boldsymbol{X} - \boldsymbol{X}_m\|^2 \sim \chi^2_{n-m}$, independently. The (scaled) quotient then turns out to have an F distribution — a consequence of the following theorem.

Theorem: Relationship Between χ^2 and F Distributions

Let $U \sim \chi^2_m$ and $V \sim \chi^2_n$ be independent. Then,

$$\frac{U/m}{V/n} \sim \mathsf{F}(m, n).$$

Proof

For notational simplicity, let $c = m/2$ and $d = n/2$. The pdf of $W = U/V$ is given by $f_W(w) = \int_0^\infty f_U(wv)\, v\, f_V(v)\, \mathrm{d}v$. Substituting the pdfs of the corresponding Gamma distributions, we have

$$\begin{aligned} f_W(w) &= \int_0^\infty \frac{(wv)^{c-1}\, \mathrm{e}^{-wv/2}}{\Gamma(c)\, 2^c}\, v\, \frac{v^{d-1} \mathrm{e}^{-v/2}}{\Gamma(d)\, 2^d}\, \mathrm{d}v \\ &= \frac{\Gamma(c+d)}{\Gamma(c)\,\Gamma(d)}\, \frac{w^{c-1}}{(1+w)^{c+d}}. \end{aligned}$$

The density of $Z = \frac{n}{m}\frac{U}{V}$ is given by

$$f_Z(z) = f_W(z\, m/n)\, m/n.$$

The proof is completed by comparing the resulting expression with the pdf of the F distribution.

Normal Linear Models

Normal linear models combine the simplicity of the linear model with the tractability of the Gaussian distribution. They are the principal model for traditional statistics, and include the classic linear regression and analysis of variance models.

Definition 1: Normal Linear Model

In a normal linear model the response Y depends on a p-dimensional explanatory variable $\boldsymbol{x} = [x_1, \ldots, x_p]^\top$, via the linear relationship

$$Y = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Multivariate Normal Response Vector

Thus, a normal linear model is a linear model with normal error terms.

In our usual notation, the corresponding normal linear model for the whole training set $\{(\boldsymbol{x}_i, Y_i)\}$ has the form

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is the model matrix comprised of rows $\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

Consequently, $\boldsymbol{Y}$ can be written as $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \sigma \boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, so that $\boldsymbol{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$.

Estimating $\boldsymbol{\beta}$

It follows that the joint density of $\boldsymbol{Y}$, conditional on the explanatory variables, is given by

$$g(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}) = (2\pi\sigma^2)^{-\frac{n}{2}} \, \mathrm{e}^{-\frac{1}{2\sigma^2} ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||^2}.$$

Estimation of the parameter $\boldsymbol{\beta}$ can be performed via

- the least-squares method, or
- the maximum likelihood method.

It is clear that for every value of σ^2 the likelihood is maximal when $\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ is minimal.

As a consequence, the maximum likelihood estimate for $\boldsymbol{\beta}$ is the same as the least-squares estimate $\widehat{\boldsymbol{\beta}} = \mathbf{X}^{+}\boldsymbol{y}$.

Estimating σ^2

The maximum likelihood estimate of σ^2 is equal to

$$\widehat{\sigma^2} = \frac{\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{n},$$

where $\widehat{\boldsymbol{\beta}}$ is the maximum likelihood estimate (least squares estimate in this case) of $\boldsymbol{\beta}$.