Statistics for Data Science: Week 9

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Linear Algebra Intermezzo- *continued*

Linear Subspaces, Orthogonal Projections, Gaussian Vectors

Definition (Multivariate Gaussian Distribution)

A random vector $\boldsymbol{Y}$ in $\mathbb{R}^d$ has the multivariate normal distribution if and only if $\boldsymbol{\beta}^\top \boldsymbol{Y}$ has the univariate normal distribution, $\forall \boldsymbol{\beta} \in \mathbb{R}^d$.

How can we use this definition to determine basic properties?

Recall that the *moment generating function* (MGF) of a random vector $\boldsymbol{W}$ in $\mathbb{R}^d$ is defined as

$$M_{\boldsymbol{W}}(\boldsymbol{\theta}) = \mathbb{E}[e^{\boldsymbol{\theta}^\top \boldsymbol{W}}], \qquad \boldsymbol{\theta} \in \mathbb{R}^d,$$

provided the expectation exists. When the MGF exists *it characterises the distribution of the random vector. Furthermore, two random vectors are independent if and only if their joint MGF is the product of their marginal MGF's.*

Most important facts about Gaussian vectors:

1. Moment generating function of $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$:

$$\boxed{M_{\boldsymbol{Y}}(\boldsymbol{u}) = \exp\left(\boldsymbol{u}^\top \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{u}^\top \boldsymbol{\Omega} \boldsymbol{u}\right)}.$$

2. $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$ and given $\boldsymbol{B}_{n\times p}$ and $\boldsymbol{\theta}_{n\times 1}$, then

$$\boldsymbol{\theta} + \boldsymbol{B}\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\theta} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top).$$

3. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ density, assuming $\boldsymbol{\Omega}$ nonsingular:

$$\boxed{f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Omega}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}}.$$

4. Constant density isosurfaces are ellipsoidal
5. Marginals of Gaussian are Gaussian (converse NOT true).
6. $\boldsymbol{\Omega}$ diagonal $\Leftrightarrow$ independent coordinates Y_j.
7. If $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$,

$$\boldsymbol{A}\boldsymbol{Y} \text{ independent of } \boldsymbol{B}\boldsymbol{Y} \iff \boldsymbol{A}\boldsymbol{\Omega}\boldsymbol{B}^\top = 0.$$

Proposition (Property 1: Moment Generating Function)

The moment generating function of $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ is

$$M_{\boldsymbol{Y}}(\boldsymbol{u}) = \exp\left(\boldsymbol{u}^\top \boldsymbol{\mu} + \tfrac{1}{2}\boldsymbol{u}^\top \boldsymbol{\Omega} \boldsymbol{u}\right)$$

Proof ($*$).

Let $\boldsymbol{u} \in \mathbb{R}^p$ be arbitrary. Then $\boldsymbol{u}^\top \boldsymbol{Y}$ is Gaussian with mean $\boldsymbol{u}^\top \boldsymbol{\mu}$ and variance $\boldsymbol{u}^\top \boldsymbol{\Omega} \boldsymbol{u}$. Hence it has moment generating function:

$$M_{\boldsymbol{u}^\top \boldsymbol{Y}}(t) = \mathbb{E}\left(e^{t\boldsymbol{u}^\top \boldsymbol{Y}}\right) = \exp\left\{t(\boldsymbol{u}^\top \boldsymbol{\mu}) + \frac{t^2}{2}(\boldsymbol{u}^\top \boldsymbol{\Omega} \boldsymbol{u})\right\}.$$

Now take $t = 1$ and observe that

$$M_{\boldsymbol{u}^\top \boldsymbol{Y}}(1) = \mathbb{E}\left(e^{\boldsymbol{u}^\top \boldsymbol{Y}}\right) = M_{\boldsymbol{Y}}(\boldsymbol{u}).$$

Combining the two, we conclude that

$$M_{\boldsymbol{Y}}(\boldsymbol{u}) = \exp\left(\boldsymbol{u}^\top \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{u}^\top \boldsymbol{\Omega} \boldsymbol{u}\right), \quad \boldsymbol{u} \in \mathbb{R}^p.$$

□

Proposition (Property 2: Affine Transformation)

For $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$ *and given* $\boldsymbol{B}_{n\times p}$ *and* $\boldsymbol{\theta}_{n\times 1}$*, we have*

$$\boldsymbol{\theta} + \boldsymbol{B}\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\theta} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)$$

Proof (∗).

$$\begin{aligned}
M_{\boldsymbol{\theta}+\boldsymbol{B}\boldsymbol{Y}}(u) &= \mathbb{E}\left[\exp\{\boldsymbol{u}^\top(\boldsymbol{\theta}+\boldsymbol{B}\boldsymbol{Y})\}\right] = \exp\left\{\boldsymbol{u}^\top\boldsymbol{\theta}\right\}\mathbb{E}\left[\exp\{(\boldsymbol{B}^\top\boldsymbol{u})^\top\boldsymbol{Y}\}\right] \\
&= \exp\left\{\boldsymbol{u}^\top\boldsymbol{\theta}\right\} M_{\boldsymbol{Y}}(\boldsymbol{B}^\top\boldsymbol{u}) \\
&= \exp\left\{\boldsymbol{u}^\top\boldsymbol{\theta}\right\}\exp\left\{(\boldsymbol{B}^\top\boldsymbol{u})^\top\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{u}^\top\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top\boldsymbol{u}\right\} \\
&= \exp\left\{\boldsymbol{u}^\top\boldsymbol{\theta} + \boldsymbol{u}^\top(\boldsymbol{B}\boldsymbol{\mu}) + \frac{1}{2}\boldsymbol{u}^\top\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top\boldsymbol{u}\right\} \\
&= \exp\left\{\boldsymbol{u}^\top(\boldsymbol{\theta} + \boldsymbol{B}\boldsymbol{\mu}) + \frac{1}{2}\boldsymbol{u}^\top\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top\boldsymbol{u}\right\}
\end{aligned}$$

And this last expression is the MGF of a $\mathcal{N}(\boldsymbol{\theta} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)$ distribution. □

Proposition (Property 3: Density Function)

Let $\boldsymbol{\Omega}_{p\times p}$ be nonsingular. The density of $\mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Omega}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}$$

Proof (∗).

Let $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^\top$ be a vector of iid $\mathcal{N}(0,1)$ random variables. Then, because of independence,

(a) the density of $\boldsymbol{Z}$ is

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \prod_{i=1}^{p} f_{Z_i}(z_i) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z_i^2\right) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z}\right).$$

(b) The MGF of $\boldsymbol{Z}$ is

$$M_{\boldsymbol{Z}}(\boldsymbol{u}) = \mathbb{E}\left\{\exp\left(\sum_{i=1}^{p} u_i Z_i\right)\right\} = \prod_{i=1}^{p} \mathbb{E}\{\exp(u_i Z_i)\} = \exp(\boldsymbol{u}^\top \boldsymbol{u}/2),$$

which is the MGF of a p-variate $\mathcal{N}(0, \boldsymbol{I})$ distribution.

$\overset{(a)+(b)}{\Longrightarrow}$ the $\mathcal{N}(0, \boldsymbol{I})$ density is $f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z}\right)$.

By the spectral theorem, $\boldsymbol{\Omega}$ admits a square root, $\boldsymbol{\Omega}^{1/2}$. Furthermore, since $\boldsymbol{\Omega}$ is non-singular, so is $\boldsymbol{\Omega}^{1/2}$.

Now observe that from our Property 2, we have $\boldsymbol{Y} \overset{d}{=} \boldsymbol{\Omega}^{1/2}\boldsymbol{Z} + \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$.

By the change of variables formula,

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= f_{\boldsymbol{\Omega}^{1/2}\boldsymbol{Z}+\boldsymbol{\mu}}(\boldsymbol{y}) \\ &= |\boldsymbol{\Omega}^{-1/2}| f_{\boldsymbol{Z}}\{\Omega^{-1/2}(\boldsymbol{y}-\boldsymbol{\mu})\} \\ &= \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Omega}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}. \end{aligned}$$

[Recall that to obtain the density of $\boldsymbol{W} = g(\boldsymbol{X})$ at $\boldsymbol{w}$, we need to evaluate $f_{\boldsymbol{X}}$ at $g^{-1}(\boldsymbol{w})$ but also multiply by the Jacobian determinant of g^{-1} at $\boldsymbol{w}$.]

$\square$

Proposition (Property 4: Isosurfaces)

The isosurfaces of a $\mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$ are $(p-1)$-dimensional ellipsoids centred at $\boldsymbol{\mu}$, with principal axes given by the eigenvectors of $\boldsymbol{\Omega}$ and with anisotropies given by the ratios of the square roots of the corresponding eigenvalues of $\boldsymbol{\Omega}$.

Proof (∗).

Exercise: Use Property 3, and the spectral theorem. □

Proposition (Property 5: Coordinate Distributions)

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$. Then $Y_j \sim \mathcal{N}(\mu_j, \Omega_{jj})$.

Proof (∗).

Observe that $Y_j = (0\ ,\ 0\ , \ldots, \underbrace{1}_{j_{th}\text{ position}} , \ldots, \ 0\ ,\ 0)\boldsymbol{Y}$ and use Property 2. □

Proposition (Property 6: Diagonal $\boldsymbol{\Omega} \iff$ Independence)

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$. Then the Y_i are mutually independent if and only if $\boldsymbol{\Omega}$ is diagonal.

Proof ($*$).

Suppose that the Y_j are independent. Property 5 yields $Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ for some $\sigma_j > 0$. Thus the density of $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{j=1}^{p} f_{Y_j}(y_j) = \prod_{i=1}^{p} \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(y_j-\mu_j)^2}{\sigma_j^2}\right\}$$

$$= \frac{1}{(2\pi)^{p/2}\left|\mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)\right|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_p^{-2})(\boldsymbol{y}-\boldsymbol{\mu})\right\}.$$

Hence $\boldsymbol{Y} \sim \mathcal{N}\{\boldsymbol{\mu}, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)\}$, i.e. the covariance $\boldsymbol{\Omega}$ is diagonal.

Conversely, assume $\boldsymbol{\Omega}$ is diagonal, say $\boldsymbol{\Omega} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Then we can reverse the steps of the first part to see that the joint density $f_{\boldsymbol{Y}}(\boldsymbol{y})$ can be written as a product of the marginal densities $f_{Y_j}(y_j)$, thus proving independence.

$\square$

Proposition (Property 7: $\boldsymbol{AY}, \boldsymbol{BY}$ indep $\iff \boldsymbol{A\Omega B}^\top = 0$)

If $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$, and $\boldsymbol{A}_{m\times p}$, $\boldsymbol{B}_{d\times p}$ be real matrices. Then,

$$\boldsymbol{AY} \text{ independent of } \boldsymbol{BY} \iff \boldsymbol{A\Omega B}^\top = 0.$$

Proof (∗). [wlog assuming $\boldsymbol{\mu} = 0$ (simplifies the algebra)]

First assume $\boldsymbol{A\Omega B}^\top = 0$. Let $\boldsymbol{W}_{(m+d)\times 1} = \binom{\boldsymbol{AY}}{\boldsymbol{BY}}$ and $\boldsymbol{\theta}_{(m+d)\times 1} = \binom{\boldsymbol{u}_{m\times 1}}{\boldsymbol{v}_{d\times 1}}$.

$$\begin{aligned} M_{\boldsymbol{W}}(\boldsymbol{\theta}) &= \mathbb{E}[\exp\{\boldsymbol{W}^\top\boldsymbol{\theta}\}] = \mathbb{E}\left[\exp\left\{\boldsymbol{Y}^\top\boldsymbol{A}^\top\boldsymbol{u} + \boldsymbol{Y}^\top\boldsymbol{B}^\top\boldsymbol{v}\right\}\right] \\ &= \mathbb{E}\left[\exp\left\{\boldsymbol{Y}^\top(\boldsymbol{A}^\top\boldsymbol{u} + \boldsymbol{B}^\top\boldsymbol{v})\right\}\right] = M_{\boldsymbol{Y}}(\boldsymbol{A}^\top\boldsymbol{u} + \boldsymbol{B}^\top\boldsymbol{v}) \\ &= \exp\left\{\frac{1}{2}(\boldsymbol{A}^\top\boldsymbol{u} + \boldsymbol{B}^\top\boldsymbol{v})^\top\Omega(\boldsymbol{A}^\top\boldsymbol{u} + \boldsymbol{B}^\top\boldsymbol{v})\right\} \end{aligned}$$

$$= \exp\left\{\tfrac{1}{2}\left(\boldsymbol{u}^\top\boldsymbol{A\Omega A}^\top\boldsymbol{u} + \boldsymbol{v}^\top\boldsymbol{B\Omega B}^\top\boldsymbol{v} + \boldsymbol{u}^\top\underbrace{\boldsymbol{A\Omega B}^\top}_{=0}\boldsymbol{v} + \boldsymbol{v}^\top\underbrace{\boldsymbol{B\Omega A}^\top}_{=0}\boldsymbol{u}\right)\right\}$$

$= M_{\boldsymbol{AY}}(\boldsymbol{u})M_{\boldsymbol{BY}}(\boldsymbol{v})$ (joint MGF = product of marginal MGFs, thus independence)

For the converse, assume that $\boldsymbol{AY}$ and $\boldsymbol{BY}$ are independent. Then, $\forall \boldsymbol{u}, \boldsymbol{v}$,

$$M_{\boldsymbol{W}}(\boldsymbol{\theta}) = M_{\boldsymbol{AY}}(\boldsymbol{u}) M_{\boldsymbol{BY}}(\boldsymbol{v}), \quad \forall \boldsymbol{u}, \boldsymbol{v},$$

$$\implies \exp\left\{\frac{1}{2}\left(\boldsymbol{u}^\top \boldsymbol{A}\Omega\boldsymbol{A}^\top \boldsymbol{u} + \boldsymbol{v}^\top \boldsymbol{B}\Omega\boldsymbol{B}^\top \boldsymbol{v} + \boldsymbol{u}^\top \boldsymbol{A}\Omega\boldsymbol{B}^\top \boldsymbol{v} + \boldsymbol{v}^\top \boldsymbol{B}\Omega\boldsymbol{A}^\top \boldsymbol{u}\right)\right\}$$

$$= \exp\left\{\frac{1}{2}\boldsymbol{u}^\top \boldsymbol{A}\Omega\boldsymbol{A}^\top \boldsymbol{u}\right\} \exp\left\{\frac{1}{2}\boldsymbol{v}^\top \boldsymbol{B}\Omega\boldsymbol{B}^\top \boldsymbol{v}\right\}$$

$$\implies \exp\left\{\frac{1}{2} \times 2\boldsymbol{v}^\top \boldsymbol{A}\Omega\boldsymbol{B}^\top \boldsymbol{u}\right\} = 1$$

$$\implies \boldsymbol{v}^\top \boldsymbol{A}\Omega\boldsymbol{B}^\top \boldsymbol{u} = 0, \qquad \forall\ \boldsymbol{u}, \boldsymbol{v},$$

$$\implies \boldsymbol{A}\Omega\boldsymbol{B}^\top = 0.$$

□

Reminder:

Definition (χ^2 distribution)

Let $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I}_{p\times p})$. Then $\|\boldsymbol{Z}\|^2 = \sum_{j=1}^{p} Z_j^2$ is said to have the chi-square (χ^2) distribution with p degrees of freedom; we write $\|\boldsymbol{Z}\|^2 \sim \chi_p^2$.

[Thus, χ_p^2 is the distribution of the sum of squares of p real independent standard Gaussian random variates.]

Definition (F distribution)

Let $V \sim \chi_p^2$ and $W \sim \chi_q^2$ be independent random variables. Then $(V/p)/(W/q)$ is said to have the F distribution with p and q degrees of freedom; we write $(V/p)/(W/q) \sim F_{p,q}$.

Proposition (Gaussian Quadratic Forms)

1. *If* $\boldsymbol{Z} \sim \mathcal{N}(0_{p\times 1}, \boldsymbol{I}_{p\times p})$ *and* $\boldsymbol{H}$ *is a projection of rank* $r \leq p$,

$$\boldsymbol{Z}^\top \boldsymbol{H}\boldsymbol{Z} \sim \chi^2_r.$$

2. $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$ *with* $\boldsymbol{\Omega}$ *nonsingular* $\Longrightarrow$

$$(\boldsymbol{Y} - \boldsymbol{\mu})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \sim \chi^2_p.$$

Gaussian Linear Regression: Likelihood and Geometry

General formulation:

$$Y_i | x_i \overset{ind}{\sim} \text{Distribution}\{g(x_i)\}, \quad i = 1, \ldots, n.$$

Simple Normal Linear Regression:

$$\begin{cases} \text{Distribution} = \mathcal{N}\{g(x), \sigma^2\} \\ g(x) = \beta_0 + \beta_1 x \end{cases}$$

Resulting Model:

$$Y_i \overset{ind}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$\Updownarrow$$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} \mathcal{N}(0, \sigma^2)$$

Jargon: Y is *response variable* and x is *explanatory variable* (or *covariate*)
Linearity: Linearity is in the *parameters*, not the *explanatory variable*.
Example: Flexibility in what we define as explanatory:

$$Y_j = \beta_0 + \beta_1 \underbrace{\sin(x_j)}_{x_j^*} + \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \text{Normal}(0, \sigma^2).$$

Example: Sometimes a transformation may be required:

$$Y_j = \beta_0 e^{\beta_1 x_j} \eta_j, \quad \eta_j \overset{iid}{\sim} \text{Lognormal}$$

$$\log(\cdot) \downarrow \qquad\qquad \uparrow \exp(\cdot)$$

$$\log Y_j = \log \beta_0 + \beta_1 x_j + \log \eta_j, \quad \log \eta_j \overset{iid}{\sim} \text{Normal}$$

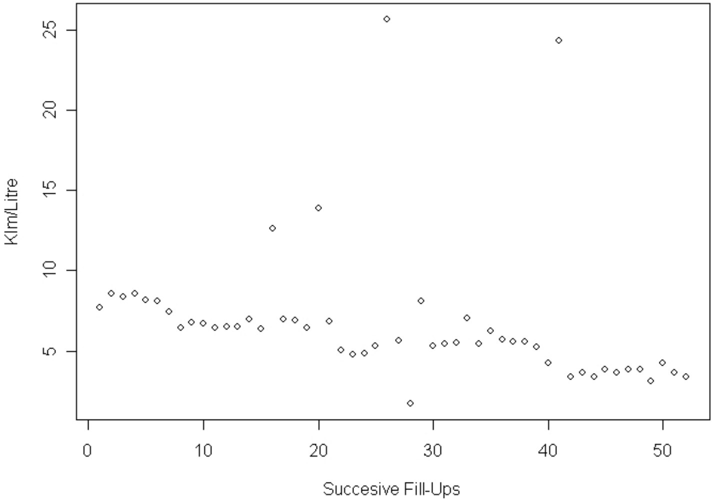
Data Structure:
For $i = 1, \ldots, n$, pairs

$$(x_i, y_i) \longrightarrow \begin{cases} x_i \text{ fixed values of } x \\ Y_i \text{ random output } Y_i \text{ when input is } x_i \end{cases}$$

Example: Professor's Van

Fillup	Km/L
1	7.72
2	8.54
3	8.35
4	8.55
5	8.16
6	8.12
7	7.46
8	6.43
9	6.74
10	6.72

Example: Professor's Van



Instead of $x_i \in \mathbb{R}$ could have $\boldsymbol{x}_i^\top \in \mathbb{R}^q$:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_q x_{iq} + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} \mathcal{N}(0, \sigma^2).$$

Letting $p = q + 1$, this can be summarised via matrix notation:

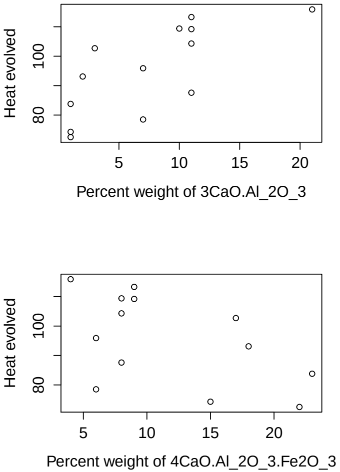
$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \ldots & x_{1q} \\ 1 & x_{21} & & x_{2q} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nq} \end{pmatrix}}_{\boldsymbol{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

$$\Longrightarrow \underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times p}{\boldsymbol{X}} \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_{n\times n})$$

$\boldsymbol{X}$ is called the design matrix.

Example: Cement Heat Evolution

Case	$3CaO.Al_2O_3$	$3CaO.SiO_2$	$4Cao.Al_2O_3.Fe_2O_3$	$2CaO.SiO_2$	Heat
1	7.00	26.00	6.00	60.00	78.50
2	1.00	29.00	15.00	52.00	74.30
3	11.00	56.00	8.00	20.00	104.30
4	11.00	31.00	8.00	47.00	87.60
5	7.00	52.00	6.00	33.00	95.90
6	11.00	55.00	9.00	22.00	109.20
7	3.00	71.00	17.00	6.00	102.70
8	1.00	31.00	22.00	44.00	72.50
9	2.00	54.00	18.00	22.00	93.10
10	21.00	47.00	4.00	26.00	115.90
11	1.00	40.00	23.00	34.00	83.80
12	11.00	66.00	9.00	12.00	113.30
13	10.00	68.00	8.00	12.00	109.40


Heat evolved
80
100
5
10
15
20
Percent weight of 3CaO.Al_2O_3
Heat evolved
80
100
30
40
50
60
70
Percent weight of 3CaO.SiO_2
Heat evolved
80
100
5
10
15
20
Percent weight of 4CaO.Al_2O_3.Fe2O_3
Heat evolved
80
100
10
20
30
40
50
60
Percent weight of 2CaO.SiO_2

Model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\Updownarrow$$

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_{n\times n})$$

Observe: $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ for given fixed design matrix $\boldsymbol{X}$, i.e.:

$$(Y_1, x_{11}, \ldots, x_{1q}), \ldots, (Y_i, x_{i1}, \ldots, x_{iq}), \ldots, (Y_n, x_{n1}, \ldots, x_{nq})$$

Likelihood and Loglikelihood

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \right\}$$

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2} \left\{ n \log 2\pi + n \log \sigma^2 + \frac{1}{\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \right\}$$

Whatever the value of σ, the log-likelihood is maximised when $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ is minimised. Hence, the MLE of $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right\} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(-\boldsymbol{X}\boldsymbol{\beta})$$

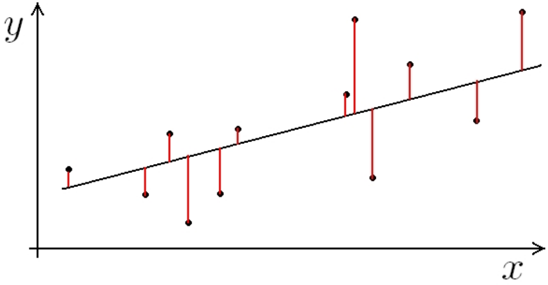
Obtain minimum by solving:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\
0 &= \frac{\partial(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \, \frac{\partial(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}{\partial(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})} \quad \text{(chain rule)} \\
0 &= \boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \quad \text{(normal equations)} \\
\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} &= \boldsymbol{X}^\top\boldsymbol{Y} \\
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \quad (\text{if } \boldsymbol{X} \text{ has rank } p)
\end{aligned}$$

The MLE $\hat{\boldsymbol{\beta}}$ is called the least squares estimator because it is a result of minimising

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \underbrace{\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_q x_{iq})^2}_{\text{sum of squares}}.$$

Thus we are trying to find the $\boldsymbol{\beta}$ that gives the hyperplane with minimum sum of squared vertical distances from our observations.



Residuals: $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, so that $\boldsymbol{e} = (e_1, \ldots, e_n)^\top$, with

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_q x_{iq}$$

"Regression Line" is such that $\sum e_i^2$ is minimised over all $\boldsymbol{\beta}$.

Fitted Values: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, so that $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^\top$, with

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_q x_{iq}$$

Since the MLE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$ for all values of σ^2, we have

$$
\begin{aligned}
\hat{\sigma}^2 &= \arg\max_{\sigma^2} \left\{ \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}, \sigma^2) \right\} \\
&= \arg\max_{\sigma^2} \ell(\hat{\boldsymbol{\beta}}, \sigma^2) \\
&= \arg\max_{\sigma^2} \; -\frac{1}{2} \left\{ n \log \sigma^2 + \frac{1}{\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) \right\}.
\end{aligned}
$$

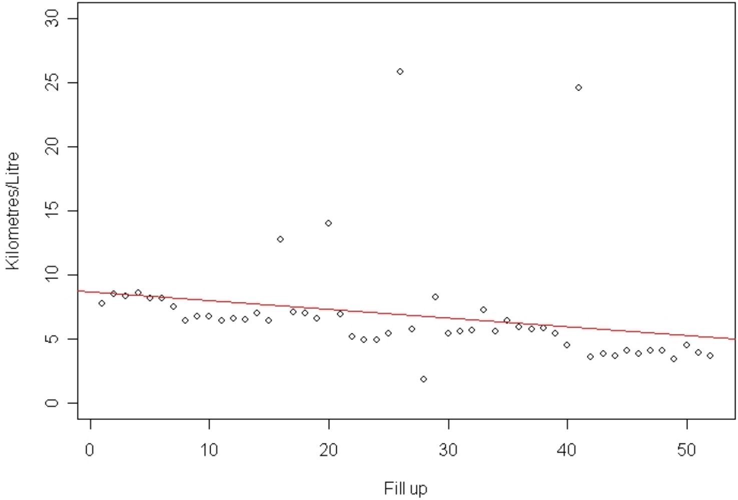
Differentiating and setting equal to zero yields

$$
\hat{\sigma}^2 = \frac{1}{n} (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}).
$$

We will soon see that a better (unbiased) estimator is

$$
S^2 = \frac{1}{n-p} (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}).
$$

Example: Professor's Van



$$\hat{\beta}_0 = 8.6 \quad \hat{\beta}_1 = -0.068 \quad S^2 = 17.4$$

There are two <u>dual</u> geometrical viewpoints that one may adopt:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & & x_{2q} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{(n-1)1} & x_{(n-1)2} & \cdots & x_{(n-1)q} \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- Row geometry: focus on the n OBSERVATIONS
- Column geometry: focus on the p covariates

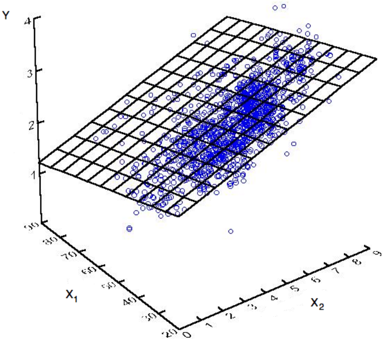
Both are useful, usually for different things:

- Row geometry useful for exploratory analysis.
- Column geometry useful for theoretical analysis.

Both geometries give useful, but different, intuitive interpretations of the least squares estimators.

Corresponds to the "scatterplot geometry" – (data space)

- n points in $\mathbb{R}^p$
- each corresponds to an observation
- least squares parameters give parametric equation for a hyperplane
- hyperplane has property that it minimizes the sum of squared vertical distances of observations from the plane itself over all possible hyperplanes



- Fitted values are vertical projections (NOT orthogonal projections!) of observations onto plane, residuals are signed vertical distances of observations from plane.

Adopt the dual perspective:

- Consider the entire vector $\boldsymbol{Y}$ as a single point living in $\mathbb{R}^n$
- Then consider each variable (column of $\boldsymbol{X}$) as a point also in $\mathbb{R}^n$

What is the interpretation of the p-dimensional vector $\hat{\boldsymbol{\beta}}$, and the n-dimensional vectors $\hat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ in this dual space?

Turns out there is another important plane here: the plane spanned by the variable vectors (the column vectors of $\boldsymbol{X}$).

Recall that this is the *column space* of $\boldsymbol{X}$, denoted by $\mathcal{M}(\boldsymbol{X})$.

$$\text{Recall:} \quad \underbrace{\mathcal{M}(\boldsymbol{X})}_{\text{Column Space}} := \{\boldsymbol{X}\gamma : \gamma \in \mathbb{R}^p\}$$

Q: What does $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ mean?

A: $\boldsymbol{Y}$ is [some element of $\mathcal{M}(\boldsymbol{X})$] + [Gaussian disturbance].

Any realisation of $\boldsymbol{Y}$ will lie outside $\mathcal{M}(\boldsymbol{X})$ (almost surely). MLE estimates $\boldsymbol{\beta}$ by minimising

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

Thus we search for a $\boldsymbol{\beta}$ giving the element of $\mathcal{M}(\boldsymbol{X})$ with the minimum distance from $\boldsymbol{Y}$.

Hence $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the projection of $\boldsymbol{Y}$ onto $\mathcal{M}(\boldsymbol{X})$:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} := \underbrace{\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top}_{\boldsymbol{H}}\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}.$$

$\boldsymbol{H}$ is the hat matrix (because it puts a hat on $\boldsymbol{Y}$!)

Leads to geometric derivation of the MLE of $\boldsymbol{\beta}$:

- Choose $\hat{\boldsymbol{\beta}}$ to minimise $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, so

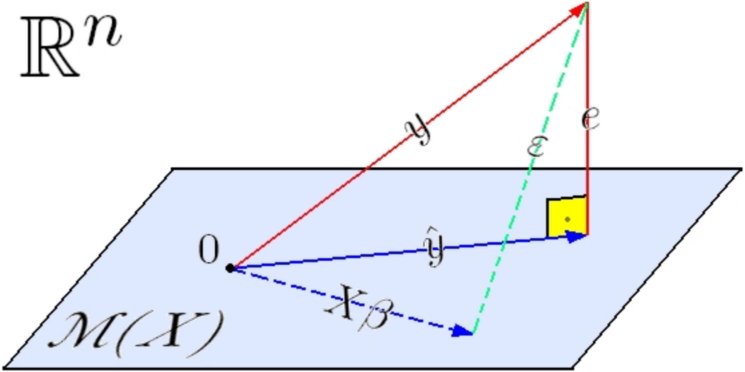
$$\hat{\beta} = \arg\min \|\boldsymbol{Y} - \boldsymbol{X}\beta\|^2.$$

- $\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \min_{\boldsymbol{\gamma}\in\mathcal{M}(\boldsymbol{X})} \|\boldsymbol{Y} - \gamma\|^2$
- But the unique γ that yields $\min_{\boldsymbol{\gamma}\in\mathcal{M}(\boldsymbol{X})} \|\boldsymbol{Y} - \gamma\|^2$ is $\gamma = \boldsymbol{P}\boldsymbol{Y}$.
- Here $\boldsymbol{P}$ is the projection onto the column space of $\boldsymbol{X}$, $\mathcal{M}(\boldsymbol{X})$.
- Since $\boldsymbol{X}$ is of full rank, $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$.
- So $\gamma = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$
- $\hat{\boldsymbol{\beta}}$ will now be the unique (since $\boldsymbol{X}$ non-singular) vector of coordinates of γ with respect to the basis of columns of $\boldsymbol{X}$.
- So

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

which implies that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$

The (Column) Geometry of Least Squares



Important facts that will repeatedly be made use of:

1. $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\varepsilon$.
2. $\hat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ are orthogonal, i.e. $\hat{\boldsymbol{Y}}^\top \boldsymbol{e} = 0$
3. Pythagoras: $\boldsymbol{Y}^\top \boldsymbol{Y} = \hat{\boldsymbol{Y}}^\top \hat{\boldsymbol{Y}} + \boldsymbol{e}^\top \boldsymbol{e} = \boldsymbol{Y}^\top \boldsymbol{H}\boldsymbol{Y} + \varepsilon^\top (\boldsymbol{I} - \boldsymbol{H})\varepsilon$

Derivation:

1. $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{X}\beta + \varepsilon) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\beta + (\boldsymbol{I} - \boldsymbol{H})\varepsilon = (\boldsymbol{I} - \boldsymbol{H})\varepsilon$
2. $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \implies \hat{\boldsymbol{Y}}^\top \boldsymbol{e} = \boldsymbol{Y}^\top \boldsymbol{H}^\top (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = 0$
3. $\boldsymbol{Y}^\top \boldsymbol{Y} = (\boldsymbol{H}\boldsymbol{Y} + (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y})^\top (\boldsymbol{H}\boldsymbol{Y} + (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}) = \hat{\boldsymbol{Y}}^\top \hat{\boldsymbol{Y}} + \boldsymbol{e}^\top \boldsymbol{e} + \underbrace{2\boldsymbol{Y}\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}}_{=0}$.

Could also assume slightly different model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \frac{\varepsilon_i}{\sqrt{w_i}}, \quad \varepsilon_i \overset{ind}{\sim} \mathcal{N}(0, \sigma^2), \quad w_i > 0$$

$$\Updownarrow$$

$$Y_i \overset{ind}{\sim} N\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq}, \frac{\sigma^2}{w_i}\right).$$

With the w_j *known* weights (example: each Y_j is an average of w_j measurements).

Arises often in practice (e.g., in sample surveys), but also arises in theory (will see in GLM).

Transformation:

$$\boldsymbol{Y}^* = \boldsymbol{W}^{1/2}\boldsymbol{Y}, \quad \boldsymbol{X}^* = \boldsymbol{W}^{1/2}\boldsymbol{X}$$

with

$$\boldsymbol{W}_{n\times n} = \mathrm{diag}(w_1, \ldots, w_n)$$

Leads to usual scenario. In this notation we obtain:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= [(\boldsymbol{X}^*)^\top \boldsymbol{X}^*]^{-1}(\boldsymbol{X}^*)^\top \boldsymbol{Y}^* \\ &= (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{Y}\end{aligned}$$

Similarly:

$$S^2 = \frac{1}{n-p}\boldsymbol{Y}^\top \left[\boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}\right] \boldsymbol{Y}$$