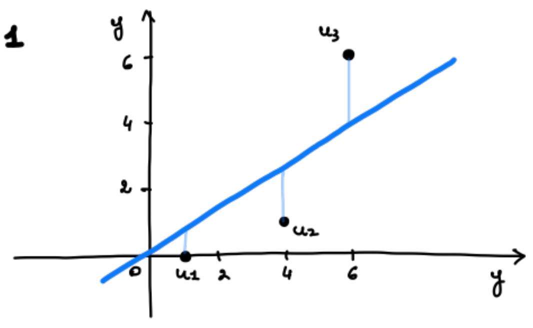
GEOMETRIC INTERPRETATION

Let's start with a simple example

consider 3 statistical units (u_1, u_2, u_3), one covariate x_i and the response y_i

	x_i	y_i
u_1	1	0
u_2	4	1
u_3	6	6

$n=3$ $p=1$

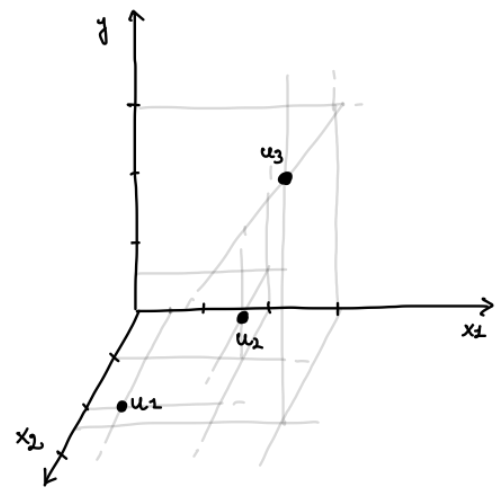


Our problem up to now was:

I look for the line that minimizes the "vertical distances"

$n = 3$ points in a $(p+1)$-dimensional space ($=$ # covariates $+ 1$)

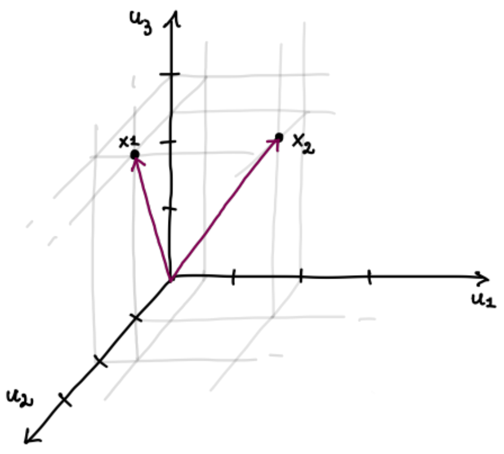
	x_{i1}	x_{i2}	y_i
u_1	1	4	0
u_2	4	2	1
u_3	6	5	6



In the multiple linear model we have $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$

where $X = [\underline{x}_1\ \underline{x}_2\ \dots\ \underline{x}_p]$, and the columns are p n-dimensional vectors

we can change perspective on the data: units are the axes, variables are vectors.



$p=2$ n-dimensional vectors
linearly independent
in an n-dimensional space

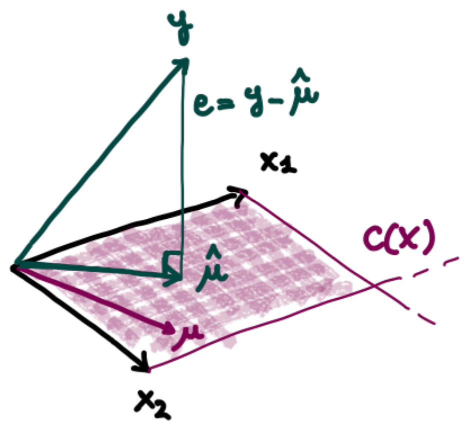
the 2 vectors identify a plane (2-dim space)

$\Rightarrow$ any linear combination of x_1 and x_2 will lie on this plane

If we call $X = [x_1\ x_2]$, $n \times p$ matrix,

$C(X) = a x_1 + b x_2$ the column space of X

subspace of $\mathbb{R}^n$ of dimension p



any $\mu = \beta_1 x_1 + \beta_2 x_2$ will lie on $C(X)$

For a given $(\beta_1, \beta_2) = \underline{\beta}$, $X\underline{\beta}$ is a vector in the subspace

When we introduce $\underline{y}$, in general it will not lie on $C(X)$

$\underline{y} - X\underline{\beta}$ is the difference between the response and that vector of $C(X)$

$(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta}) = S(\underline{\beta})$ is the squared length of the difference

$\Rightarrow$ minimizing $S(\underline{\beta})$ means finding, in $C(X)$, the vector $X\hat{\underline{\beta}}$ so that $\underline{y} - X\hat{\underline{\beta}}$ has minimum length.

$\Rightarrow$ we want $\underline{y} - X\hat{\underline{\beta}}$ to be orthogonal to $C(X)$

(hence $\underline{y} - X\hat{\underline{\beta}}$ is orthogonal to the columns $\underline{x}_1, \dots, \underline{x}_p$ of X)

orthogonality: $\begin{cases} (\underline{y} - X\hat{\underline{\beta}})^T \underline{x}_1 = 0 \\ \vdots \\ (\underline{y} - X\hat{\underline{\beta}})^T \underline{x}_p = 0 \end{cases}$

Indeed, $\hat{\mu} = X\hat{\underline{\beta}}$ is the ORTHOGONAL PROJECTION of $\underline{y}$ onto $C(X)$

$\hat{\mu} = X\hat{\underline{\beta}} = X\underbrace{(X^TX)^{-1}X^T\underline{y}}_{\hat{\underline{\beta}}} = P\underline{y}$ and $P = X(X^TX)^{-1}X^T$ is the projection matrix (check: it is symmetric and idempotent)

The vector of residuals $\underline{e} = \underline{y} - \hat{\underline{\mu}} = \underline{y} - P\underline{y} = (I_n - P)\underline{y}$ is also a projection of $\underline{y}$:

$\underline{e}$ is the projection of $\underline{y}$ on the subspace of $\mathbb{R}^n$ perpendicular to $C(X)$: $\underline{e} \perp C(X)$.

$(I_n - P)$ is also a projection matrix (check) of rank $n-p$ (it projects on the space $\perp C(X)$)

$\Rightarrow$ the vector of fitted values $\underline{\mu}$ and the vector of residuals $\underline{e}$ are perpendicular: $\underline{e}^T\underline{\mu} = 0$

the vector $\underline{e}$ and X are orthogonal: $\underline{e}^T X = 0 \iff X^T\underline{e} = 0$

$X^T(\underline{y} - X\hat{\underline{\beta}}) = 0$ $\rightarrow$ the normal equation

the least squares estimate decomposes the response vector into two orthogonal components

$\underline{y} = \hat{\underline{\mu}} + \underline{e} = \hat{\underline{y}} + \underline{e} = \hat{\underline{y}} + (\underline{y} - \hat{\underline{y}})$

thanks to the orthogonality between $\underline{e}$ and $\hat{\underline{\mu}} = \hat{\underline{y}}$ we can write

$\|\underline{y}\|^2 = \|\underline{e}\|^2 + \|\hat{\underline{y}}\| = \underline{y}^T\underline{y} = \underline{e}^T\underline{e} + \hat{\underline{y}}^T\hat{\underline{y}}$ (T. Pitagora)

Consider a model which includes the intercept: $X = [\underline{1}_n\ \underline{x}^{(2)} \dots \underline{x}^{(p)}]$, then $\underline{1}_n \in C(X)$

and for the normal equations: $\underline{1}_n^T\underline{e} = 0 \Rightarrow \sum_{i=1}^n e_i = 0$

moreover, $\underline{1}_n^T\underline{e} = \underline{1}^T(\underline{y} - \hat{\underline{y}}) = \underline{1}^T\underline{y} - \underline{1}^T\hat{\underline{y}} = 0$

$= n\bar{y} - n\bar{\hat{y}} \Rightarrow \bar{y} = \bar{\hat{y}}$

$\underline{y} = \hat{\underline{y}} + \underline{e} \Rightarrow \underline{y} - \underline{1}_n \cdot \bar{y} = \hat{\underline{y}} - \underline{1}_n\bar{y} + \underline{e}$

$\Rightarrow \|\underline{y} - \underline{1}_n\bar{y}\|^2 = \|\hat{\underline{y}} - \underline{1}_n\bar{y}\|^2 + \|\underline{e}\|^2$

$\Rightarrow \underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SSE}$ $\Rightarrow$ DEVIANCE decomposition