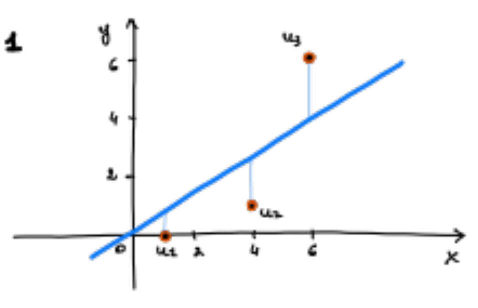
GEOMETRIC INTERPRETATION

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

$n=3 \quad p=1$

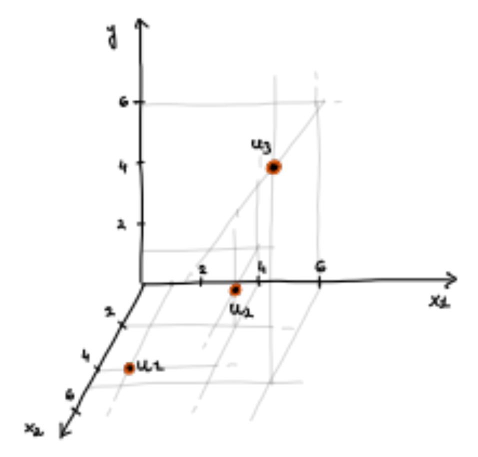


Our problem up to now was:

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

If we consider now the same units, but 2 covariates x_{i1} and x_{i2}

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



We represent $\underset{\#\text{ units}}{n}$ points in a $\underset{\#\text{ covariates } + 1\,(y)}{(p+1)}$-dimensional space : n points in $\mathbb{R}^{p+1}$

where the coordinates of each point are the values assumed by the p covariates and the response.

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

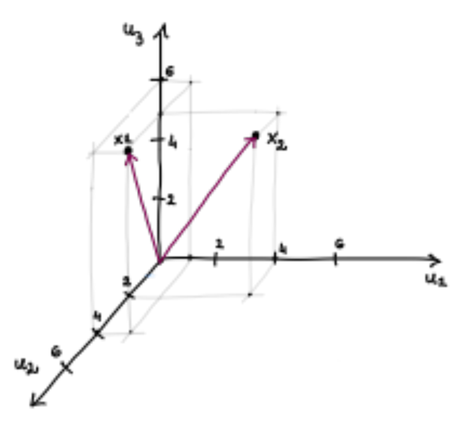
where $X = [\underline{x}_1 \ \underline{x}_2 \ \dots \ \underline{x}_p]$, and the columns are p n-dimensional vectors

→ we can change perspective on the data: now UNITS ARE THE AXES

VARIABLES ARE VECTORS

We represent p vectors in a n-dimensional space : p n-dimensional vectors in $\mathbb{R}^n$

The coordinates of each vector are the observations of that variable on the n units



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

On this space, we can define the set of all possible LINEAR COMBINATIONS of $\underline{x}_1, \dots, \underline{x}_p$

$\mathcal{C}(X) = \{ \underline{\mu} \in \mathbb{R}^n : \ \underline{\mu} = X\underline{\beta} = \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2 + \dots + \beta_p \underline{x}_p, \quad \underline{\beta} \in \mathbb{R}^p \}$

calligraphic "C"

In particular, $\mathcal{C}(X)$ is the SUBSPACE of $\mathbb{R}^n$ generated by $(\underline{x}_1, \dots, \underline{x}_p)$.

↳ p linearly indep. vectors

$\Rightarrow \mathcal{C}(X)$ has dimension p

In our example, the 2 vectors identify a plane (2-dim space)

→ any linear combination of $\underline{x}_1$ and $\underline{x}_2$ will lie on this plane

If we call $X = [\underline{x}_1 \ \underline{x}_2]$, $(n \times p) = (3 \times 2)$ matrix

$\mathcal{C}(X) = \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2$ the column space of X

$\mathcal{C}(X)$ is a subspace of $\mathbb{R}^3$ of dimension 2 $\Rightarrow$ any $\underline{\mu} = \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2$ will lie on $\mathcal{C}(X)$.

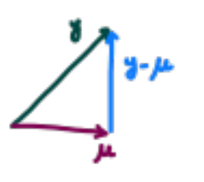
The assumption of LINEARITY is equivalent to asking that the mean of $\underline{Y}$ belongs to $\mathcal{C}(X)$.

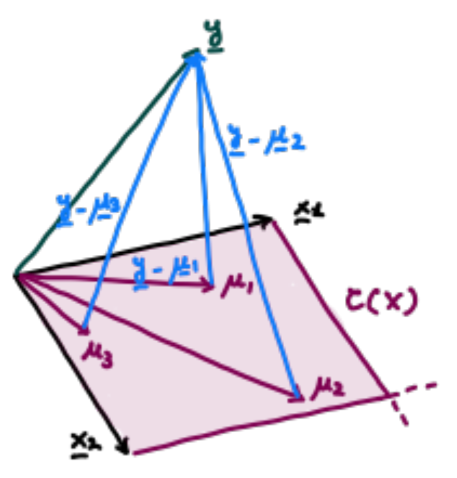
For a given $(\beta_1, \beta_2) = \underline{\beta}$, $\underline{\mu} = X\underline{\beta}$ is a vector in the subspace

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

Now, consider $\underline{y}$ and a generic vector of $\mathcal{C}(X)$ $\underline{\mu} = X\underline{\beta}$.

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

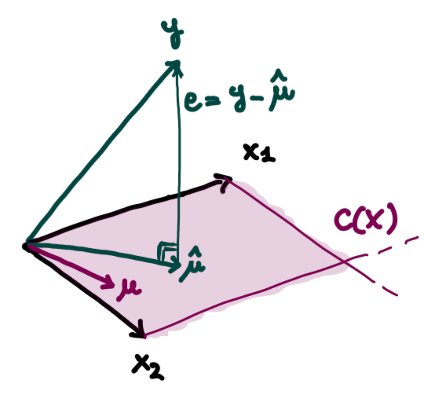




$(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta}) = S(\underline{\beta})$ is the squared length of the difference

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

$\Rightarrow$ we want $\underline{y} - X\underline{\beta}$ to be orthogonal to $\mathcal{C}(X)$ (hence $\underline{y} - X\underline{\beta}$ is orthogonal to the columns $\underline{x}_1, \dots, \underline{x}_p$ of X)



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

$\Rightarrow \underline{y} - X\hat{\underline{\beta}} \perp \mathcal{C}(X)$

$\Rightarrow \underline{y} - X\hat{\underline{\beta}} \perp \underline{x}_j$ for all $j = 1, \dots, p$ *

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

↓

normal equations

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

$(n \times n)$, symmetric, idempotent, with rank $= p$

$P^T = P \qquad P^2 = P$

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

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

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

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

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

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

SUM OF SQUARES DECOMPOSITION

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

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

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

$$\begin{aligned} \underline{y}^T\underline{y} &= \underline{y}^T (P + I_n - P)\underline{y} = \\ &= \underline{y}^T P \underline{y} + \underline{y}^T (I_n - P)\underline{y} = \\ &= \underline{y}^T P^T P \underline{y} + \underline{y}^T (I_n - P)^T (I_n - P)\underline{y} = \\ &= \hat{\underline{y}}^T \hat{\underline{y}} + \underline{e}^T\underline{e} \end{aligned}$$

(P and $(I_n - P)$ are symmetric and idempotent)

$\Rightarrow P = P^2 = P \cdot P = P^T P$

$$\Rightarrow \underline{y}^T\underline{y} = \hat{\underline{y}}^T\hat{\underline{y}} + \underline{e}^T\underline{e}$$

or, equivalently $\|\underline{y}\|^2 = \|\hat{\underline{y}}\|^2 + \|\underline{e}\|^2$

Consider a model that includes the intercept : $X = [\underline{1}_n \ \underline{x}_2 \ \dots \ \underline{x}_p]$, then $\underline{1}_n \in \mathcal{C}(X)$

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

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

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

Let's consider

$$\begin{aligned} (\underline{y} - \underline{1}\bar{y})^T(\underline{y} - \underline{1}\bar{y}) &= (\underline{y} - \underline{1}\bar{y})^T(I_n - P + P)(\underline{y} - \underline{1}\bar{y}) \\ &= \underbrace{(\underline{y} - \underline{1}\bar{y})^T(I_n - P)(\underline{y} - \underline{1}\bar{y})}_{(a)} + \underbrace{(\underline{y} - \underline{1}\bar{y})^T P (\underline{y} - \underline{1}\bar{y})}_{(b)} \end{aligned}$$

(a) $(\underline{y} - \underline{1}\bar{y})^T(I_n - P)(\underline{y} - \underline{1}\bar{y}) = \underline{y}^T(I_n - P)\underline{y} - \underline{y}^T(I_n - P)\underline{1}\bar{y} - \bar{y}\underline{1}^T(I_n - P)\underline{y} + \bar{y}\underline{1}^T(I_n - P)\underline{1}\bar{y}$

$= \underline{e}^T\underline{e} - \cancel{\underline{e}^T\underline{1}}\bar{y} - \bar{y}\cancel{\underline{1}^T\underline{e}} + \bar{y}\underline{1}^T\underline{1}\bar{y} - \bar{y}\underline{1}^T P \underline{1}\bar{y}$

(b) $(\underline{y} - \underline{1}\bar{y})^T P (\underline{y} - \underline{1}\bar{y}) = \underline{y}^T P \underline{y} - \underline{y}^T P \underline{1}\bar{y} - \bar{y}\underline{1}^T P \underline{y} + \bar{y}\underline{1}^T P \underline{1}\bar{y}$

$= \hat{\underline{y}}^T\hat{\underline{y}} - \hat{\underline{y}}\underline{1}\bar{y} - \bar{y}\underline{1}^T\hat{\underline{y}} + \bar{y}\underline{1}^T P \underline{1}\bar{y}$

$$\begin{aligned} \Rightarrow (\underline{y} - \underline{1}\bar{y})^T(\underline{y} - \underline{1}\bar{y}) &= \underline{e}^T\underline{e} + \bar{y}\underline{1}^T\underline{1}\bar{y} - \cancel{\bar{y}\underline{1}^T P \underline{1}\bar{y}} + \hat{\underline{y}}^T\hat{\underline{y}} - \hat{\underline{y}}\underline{1}\bar{y} - \bar{y}\underline{1}^T\hat{\underline{y}} + \cancel{\bar{y}\underline{1}^T P \underline{1}\bar{y}} \\ &= \underline{e}^T\underline{e} + (\hat{\underline{y}} - \underline{1}\bar{y})^T(\hat{\underline{y}} - \underline{1}\bar{y}) \end{aligned}$$

$$\Rightarrow (\underline{y} - \underline{1}\bar{y})^T(\underline{y} - \underline{1}\bar{y}) = (\hat{\underline{y}} - \underline{1}\bar{y})^T(\hat{\underline{y}} - \underline{1}\bar{y}) + (\underline{y} - \hat{\underline{y}})^T(\underline{y} - \hat{\underline{y}})$$

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

$\Rightarrow$ DEVIANCE decomposition

This is the same decomposition that we found in the simple LM.

Also in this case, we can define the coefficient of determination $R^2 = \frac{SSR}{SST}$.

Its interpretation does not change.

Notice that we derived the decomposition using the fact that $\underline{1}^T\underline{e} = 0$ i.e. $\bar{e} = 0$, which holds only if $\underline{1} \in \mathcal{C}(X) \Rightarrow$ if the model includes the intercept.

If we don't include the intercept, in general $SST \neq SSR + SSE$, and R^2 loses its interpretation.