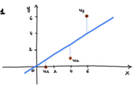
GEOMETRIC INTERPRETATION

Let's start with a simple example.

consider 3 statistical units (u_1, u_2, u_3); one covariate x_1 and the response y:

	x_1	y
u_1	1	0
u_2	4	1
u_3	6	6

$n=3$ $p=1$

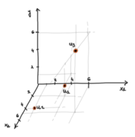


Our problem up to now was:

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

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 n points in a $(p+1)$-dimensional space : n points in $\mathbb{R}^{p+1}$

↓ # units ↓ # covariates + 1 (y)

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

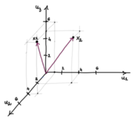
In the multiple linear model we have $Y = X\beta + \varepsilon$

where $X = [\underset{1}{x_1}\ \underset{2}{x_2}\ \dots\ \underset{p}{x_p}]$, 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 x_1 and x_2

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

→ any linear combination of x_1 and x_2 will lie on this plane.

If we call $X = [x_1\ x_2]$, $(n \times p) = (3 \times 2)$ matrix

$C(X) = \beta_1 x_1 + \beta_2 x_2$ the column space of X

$C(X)$ is a subspace of $\mathbb{R}^3$ of dimension 2 ⇒ any $\mu = \beta_1 x_1 + \beta_2 x_2$ will lie on $C(X)$.

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

In general, for an $(n \times p)$ matrix $X = [x_1\ x_2\ \dots\ x_p]$

$C(X) = \{\ \mu \in \mathbb{R}^n :\ \mu = X\beta = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p,\quad \beta \in \mathbb{R}^p\ \}$

$C(X)$ is the SUBSPACE of $\mathbb{R}^n$ generated by $(x_1, \dots, x_p)$.

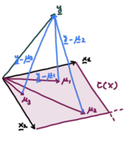
↳ p linearly indep. vectors ⇒ $C(X)$ has dimension p

The assumption of LINEARITY is equivalent to asking that the mean of Y belongs to $C(X)$.

When we introduce y_i, in general its will not lie on $C(X)$

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

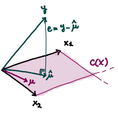
$y - X\beta$ is the difference between the response and that vector of $C(X)$.



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

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

⇒ we want $y - X\hat{\beta}$ to be orthogonal to $C(X)$ (hence $y - X\hat{\beta}$ is orthogonal to the columns $x_1, \dots, x_p$ of X)



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

⇒ $y - X\hat{\beta} \perp C(X)$

⇒ $y - X\hat{\beta} \perp x_j$ for all $j = 1, \dots, p$ *

* orthogonality: $\begin{cases} (y - X\hat{\beta})^T x_1 = 0 \\ \vdots \\ (y - X\hat{\beta})^T x_p = 0 \end{cases}$ ↓ normal equations

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

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

$P^T = P$ $\quad P^2 = P$

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

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

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

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

the vector e and X are orthogonal: $e^TX = 0 \iff X^Te = 0$

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

SUM OF SQUARES DECOMPOSITION

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

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

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

$y^Ty = y^T(P + I_n - P)y =$

$= y^TPy + y^T(I_n - P)y =$ (P and $(I_n - P)$ are symmetric and idempotent)

$= y^TP^TPy + y^T(I_n - P)^T(I_n - P)y =$ ⇒ $P = P^2 = P^TP$

$= \hat{y}^T\hat{y} + e^Te$

⇒ $y^Ty = \hat{y}^T\hat{y} + e^Te$

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

Consider a model that includes the intercept: $X = [\underset{1}{1_n}\ \underset{2}{x_2}\ \dots\ \underset{p}{x_p}]$, then $1_n \in C(X)$

and for the normal equations: $1_n^Te = 0$ ⇒ $\sum_{i=1}^n e_i = 0$

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

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

Let's consider

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

$= \underbrace{(y - 1\bar{y})^T(I_n - P)(y - 1\bar{y})}_{(a)} + \underbrace{(y - 1\bar{y})^TP(y - 1\bar{y})}_{(b)}$

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

$= e^Te - e^T1\bar{y} - \bar{y}1^Te + \bar{y}1^T1\bar{y} - \bar{y}1^TP1\bar{y}$

(b) $(y - 1\bar{y})^TP(y - 1\bar{y}) = y^TPy - y^TP1\bar{y} - \bar{y}1^TPy + \bar{y}1^TP1\bar{y}$

$= \hat{y}^T\hat{y} - \hat{y}^T1\bar{y} - \bar{y}1^T\hat{y} + \bar{y}1^TP1\bar{y}$

⇒ $(y - 1\bar{y})^T(y - 1\bar{y}) = e^Te + \bar{y}1^T1\bar{y} - \bar{y}1^TP1\bar{y} + \hat{y}^T\hat{y} - \hat{y}^T1\bar{y} - \bar{y}1^T\hat{y} + \bar{y}1^TP1\bar{y}$

$= e^Te + (\hat{y} - 1\bar{y})^T(\hat{y} - 1\bar{y})$

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

$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$ ⇒ DEVIANCE decomposition

SST $\quad$ SSR $\quad$ SSE

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

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 $1_n^Te = 0$ i.e. $\bar{e} = 0$, which holds only if $1_n \in C(X)$ ⇒ 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.