2 october 2025 - lec 2

SIMPLE LINEAR MODEL VIA ORDINARY LEAST SQUARES (OLS)

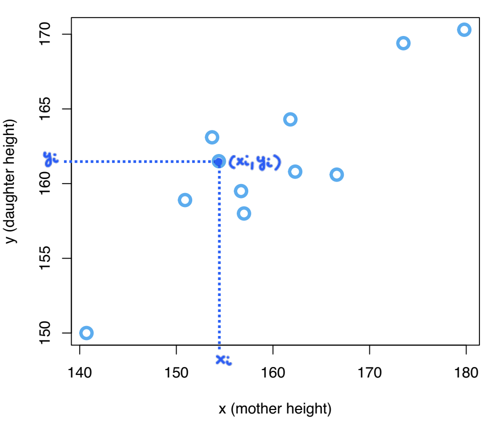
Assume that on n statistical units (individuals) we observe $(x_i; y_i)$, $i=1,\dots,n$.
Hence the data are $\underline{y}=(y_1,\dots,y_n)$ and $\underline{x}=(x_1,\dots,x_n)$
We consider that each y_i is realization of a r.v. Y_i, $i=1,\dots,n$ → sample space $\mathcal{Y}=\mathbb{R}^n$

We do not specify a distribution for $(Y_1,\dots,Y_n)$: we only make assumptions about the first two moments $\mathbb{E}[Y_i]$ and $var(Y_i)$.

We specify a simple linear model (only 1 covariate)
We estimate the model parameters only through "intuitive" considerations and a simple optimisation ("ordinary least squares" method)

We start with a simple example
relationship between the height of 11 mothers (x_i) and the height of their daughters (y_i).

	x	y
1	153.7	163.1
2	156.7	159.5
3	173.5	169.4
4	157.0	158.0
5	161.8	164.3
6	140.7	150.0
7	179.8	170.3
8	150.9	158.9
9	154.4	161.5
10	162.3	160.8
11	166.6	160.6



Intuition:
the simplest way to describe the relationship between two quantities is a straight line:

$$Y_i = \beta_1 + \beta_2 x_i \qquad i=1,\dots,n$$

However, such a relationship does not hold exactly: the points are not PERFECTLY ALIGNED.
hence we add an error term to take into account this discrepancy:

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad i=1,\dots,n$$

1st step: MODEL SPECIFICATION

Consider the model:

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad i=1,\dots,n$$

- Y_i: height of the i-th daughter
- $\beta_1 + \beta_2 x_i$: systematic component
- ε_i: ERROR TERM: the linear relationship is not exact

(β_1, β_2) are the REGRESSION COEFFICIENTS
We specified a straight line with the intercept (β_1)
We only observe 1 covariate, but we also introduce one additional "variable" taking value 1 for each individual.
The model matrix then is:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

⇒ β_1 is the INTERCEPT (coefficient of $\underline{1}$)
β_2 is the COEFFICIENT of x (slope)

ASSUMPTIONS on the independent variables
1. $x_1,\dots,x_n$ fixed and non-stochastic
2. the x_i can not be all equal (sample variance of $(x_1,\dots,x_n)$ must be $\neq 0$)

The systematic component is now fully specified, we need to define the stochastic component (ε).

ASSUMPTIONS on the STOCHASTIC COMPONENT
1. $\mathbb{E}[\varepsilon_i]=0$ for $i=1,\dots,n$
2. $Var(\varepsilon_i)=\sigma^2>0$ $i=1,\dots,n$ (common variance across subjects)
3. $Cov(\varepsilon_i,\varepsilon_k)=0$ if $i\neq k$, $i=1,\dots,n$ $k=1,\dots,n$

1. $\mathbb{E}[\varepsilon_i]=0$ $i=1,\dots,n$ ABSENCE OF SYSTEMATIC ERROR

Implications for Y_i

$$\mathbb{E}[Y_i]=\mathbb{E}[\beta_1+\beta_2x_i+\varepsilon_i] \overset{\text{linearity of } \mathbb{E}}{=} \underbrace{\mathbb{E}[\beta_1+\beta_2x_i]}_{\text{non-stochastic}}+\underbrace{\mathbb{E}[\varepsilon_i]}_{0}=\beta_1+\beta_2x_i$$

What happens if there is a systematic error? i.e. $\mathbb{E}[\varepsilon_i]=c\neq 0$
$\mathbb{E}[Y_i]=\beta_1+\beta_2x_i+c=(\beta_1+c)+\beta_2x_i$
the systematic error c is inglobated into the intercept (not a problem)
it is equivalent to a model
$Y_i=\beta_1^*+\beta_2x_i+\varepsilon_i^*$ where $\beta_1^*=\beta_1+c$
$\varepsilon_i^*=\varepsilon_i-c \Rightarrow \mathbb{E}[\varepsilon_i^*]=0$

2. $Var(\varepsilon_i)=\sigma^2>0$ for all $i=1,\dots,n$ HOMOSCEDASTICITY OF THE ERRORS

Implications for Y_i:

$$var(Y_i)=var(\underbrace{\beta_1+\beta_2x_i}_{\text{non-stoch.}}+\varepsilon_i)=var(\varepsilon_i)=\sigma^2 \quad \forall i=1,\dots,n$$

⇒ homoscedasticity of the response

3. $cov(\varepsilon_i,\varepsilon_k)=0$ for $i\neq k$ THE ERRORS ARE UNCORRELATED

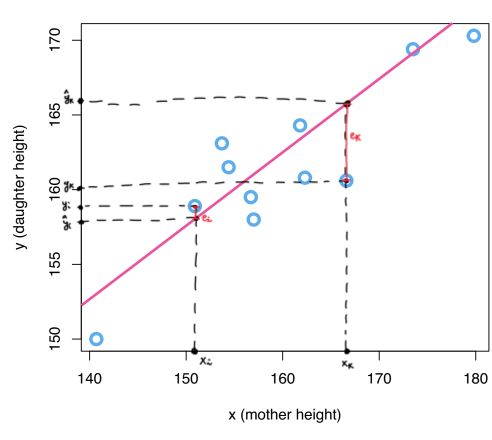
Implication for Y_i

$$cov(Y_i,Y_k)=cov(\underbrace{\beta_1+\beta_2x_i}_{\text{non-stochastic}}+\varepsilon_i\,;\,\underbrace{\beta_1+\beta_2x_k}+\varepsilon_k)=cov(\varepsilon_i,\varepsilon_k)=0$$

2nd step: ESTIMATE

what do we need to estimate? Unknown quantities are $(\beta_1,\beta_2,\sigma^2)$
Hence the PARAMETER SPACE is $\Theta=\mathbb{R}^2\times(0,+\infty)$

Every combination of (β_1,β_2) determines a specific line: how do we select the "best" line?
We need a criterion of what is a "good" line.
We want a line which is the closest to the observed points.



Consider this line: at each value of x_i corresponds one value of y_i that lies on the line
⇒ given x_i and fixed (β_1,β_2) we can compute the corresponding value of y_i (according to the line)

$$\hat{y}_i=\beta_1+\beta_2x_i \qquad \text{"predicted value"}$$

The discrepancy between the observed and the predicted value (at the observed locations x_i)
RESIDUAL: $e_i=y_i-\hat{y}_i$

A good line will have small residuals OVERALL.
- we could consider the sum of the residuals $\sum_{i=1}^n e_i$ and select the (β_1,β_2) that minimize it
  → not a good idea: positive and negative values cancel out.
- we could consider the sum of the absolute values $\sum_{i=1}^n |e_i|$ → mathematically not very practical
- we consider instead the sum of the SQUARED residuals

$$\sum_{i=1}^n e_i^2=\sum_{i=1}^n(y_i-\hat{y}_i)^2=\sum_{i=1}^n(y_i-\beta_1-\beta_2x_i)^2=S(\beta_1,\beta_2)$$

and take as an estimate of (β_1,β_2) the combination that minimizes it.

DEF: the LEAST SQUARES estimate of (β_1,β_2) is the combination of values $(\hat\beta_1,\hat\beta_2)$ that minimizes $S(\beta_1,\beta_2)$

$$(\hat\beta_1,\hat\beta_2)=\underset{(\beta_1,\beta_2)\in\mathbb{R}^2}{\arg\min}\ S(\beta_1,\beta_2)=\underset{(\beta_1,\beta_2)\in\mathbb{R}^2}{\arg\min}\sum_{i=1}^n(y_i-\beta_1-\beta_2x_i)^2$$

We have hence turned a problem of estimation into an optimization.

THM: The least squares estimate of (β_1,β_2) is

$$\hat\beta_1=\bar y-\hat\beta_2\bar x$$

$$\hat\beta_2=\frac{\sum_{i=1}^n(x_i-\bar x)(y_i-\bar y)}{\sum_{i=1}^n(x_i-\bar x)^2}$$

where $\bar x=\frac{1}{n}\sum_{i=1}^n x_i$
$\bar y=\frac{1}{n}\sum_{i=1}^n y_i$ (sample mean).

Remark:
recall that the sample variance of $(x_1,\dots,x_n)$ is $s_X^2=\frac{1}{n-1}\sum_{i=1}^n(x_i-\bar x)^2$
(and similarly for s_Y^2)
the sample covariance is $s_{XY}=\frac{1}{n-1}\sum_{i=1}^n(x_i-\bar x)(y_i-\bar y)$
Hence $\hat\beta_2=\frac{s_{XY}}{s_X^2}$

Proof: we want to show that $\hat\beta_1,\hat\beta_2$ minimise $S(\beta_1,\beta_2)=\sum_{i=1}^n(y_i-\beta_1-\beta_2x_i)^2$.
We need to find the critical points (1st derivative $=0$)
and then check that $(\hat\beta_1,\hat\beta_2)$ is a minimum (2nd derivative >0)

$$\begin{cases}\dfrac{\partial S(\beta_1,\beta_2)}{\partial\beta_1}=0\\[2mm]\dfrac{\partial S(\beta_1,\beta_2)}{\partial\beta_2}=0\end{cases}\qquad\begin{cases}\sum_{i=1}^n 2(y_i-\beta_1-\beta_2x_i)(-1)=0\\ \sum_{i=1}^n 2(y_i-\beta_1-\beta_2x_i)(-x_i)=0\end{cases}$$

$$\begin{cases}\sum_{i=1}^n(y_i-\beta_1-\beta_2x_i)=0 & (A)\\ \sum_{i=1}^n x_i(y_i-\beta_1-\beta_2x_i)=0 & (B)\end{cases}$$

(A) $n\bar y-n\beta_1-n\beta_2\bar x=0$ (since $\sum_{i=1}^n y_i=n\bar y$)
$\beta_1=\bar y-\beta_2\bar x$

(B) $\sum_{i=1}^n x_iy_i-n\bar x\beta_1-\beta_2\sum_{i=1}^n x_i^2=0$ substituting $\beta_1=\bar y-\beta_2\bar x$
$\sum_{i=1}^n x_iy_i-n\bar x\bar y+n\beta_2\bar x^2-\beta_2\sum_{i=1}^n x_i^2=0$

$$\hat\beta_2=\frac{\sum_{i=1}^n x_iy_i-n\bar x\bar y}{\sum_{i=1}^n x_i^2-n\bar x^2}$$

but since
$(n-1)s_X^2=\sum_{i=1}^n(x_i-\bar x)^2=\sum_{i=1}^n x_i^2-n\bar x^2$ and
[$\sum(x_i^2+\bar x^2-2x_i\bar x)=\sum x_i^2+n\bar x^2-2\bar x\sum x_i=\sum x_i^2+n\bar x^2-2n\bar x^2$]
$(n-1)s_{XY}=\sum_{i=1}^n(x_i-\bar x)(y_i-\bar y)=\sum_{i=1}^n x_iy_i-n\bar x\bar y$
[$\sum(x_iy_i-x_i\bar y-y_i\bar x+\bar x\bar y)=\sum x_iy_i-n\bar x\bar y-n\bar x\bar y+n\bar x\bar y$]

we obtain $\hat\beta_2=\frac{s_{XY}}{s_X^2}$
and $\hat\beta_1=\bar y-\hat\beta_2\bar x$

Is $(\hat\beta_1,\hat\beta_2)$ a minimum? We compute the Hessian

$$H=\begin{bmatrix}\dfrac{\partial^2S(\beta_1,\beta_2)}{\partial\beta_1^2} & \dfrac{\partial^2S(\beta_1,\beta_2)}{\partial\beta_1\partial\beta_2}\\[2mm] \dfrac{\partial^2S(\beta_1,\beta_2)}{\partial\beta_2\partial\beta_1} & \dfrac{\partial^2S(\beta_1,\beta_2)}{\partial\beta_2^2}\end{bmatrix}=\begin{bmatrix}2n & 2n\bar x\\ 2n\bar x & 2\sum_{i=1}^n x_i^2\end{bmatrix}$$

$$\det(H)=4n\sum_{i=1}^n x_i^2-4n^2\bar x^2=4n\left(\sum_{i=1}^n x_i^2-n\bar x^2\right)=4n\sum_{i=1}^n(x_i-\bar x)^2>0$$

since $\det(H)>0$ and $H_{11}=2n>0$, $(\hat\beta_1,\hat\beta_2)$ is a minimum of $S(\beta_1,\beta_2)$
Moreover, it is the global minimum. □

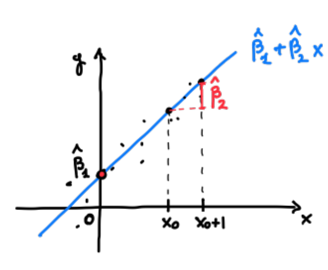
Remarks:
- we did not use the assumptions on ε_i
- we used the assumption on the x_i: what happens if $x_i=x_0$ for all $i=1,\dots,n$?
  $(x_i-\bar x)=0\ \forall i \Rightarrow s_X^2=0$ and $s_{XY}=0 \Rightarrow \hat\beta_2=\frac{0}{0}$ not defined

- once we estimate $(\hat\beta_1,\hat\beta_2)$, we automatically obtain $\hat y=\hat\beta_1+\hat\beta_2x$, i.e. the estimated regression line.
- $\hat y$ allows us to make predictions: given a generic value x, we predict the corresponding value of the response.
  As usual, careful with extrapolation, i.e., estimating the response for a value of x outside of the observed range of $(x_1,\dots,x_n)$.

- **INTERPRETATION of $(\hat\beta_1,\hat\beta_2)$**

we have estimated a line $\hat y=\hat\beta_1+\hat\beta_2x$

$\hat\beta_1$ is the intercept, i.e., the predicted value of y when $x=0$.
Not always interpretable! E.g. with the heights example: height $=0$ is meaningless



Now consider two individuals observed at $x_1=x_0$ and $x_2=x_0+1$
The predicted values are
$\hat y_1=\hat\beta_1+\hat\beta_2x_0$
$\hat y_2=\hat\beta_1+\hat\beta_2(x_0+1)$

let's study the difference in their predicted values

$$\hat y_2-\hat y_1=\hat\beta_1+\hat\beta_2(x_0+1)-\hat\beta_1-\hat\beta_2x_0=\hat\beta_2x_0+\hat\beta_2-\hat\beta_2x_0=\hat\beta_2$$

Hence $\hat\beta_2$ is the expected change in y when I increase x of 1 unit
i.e., in general, the parameter β_2:

$$\beta_2=\mathbb{E}[Y\mid x=x_0+1]-\mathbb{E}[Y\mid x=x_0]$$