Lecture 1 - Oct 17

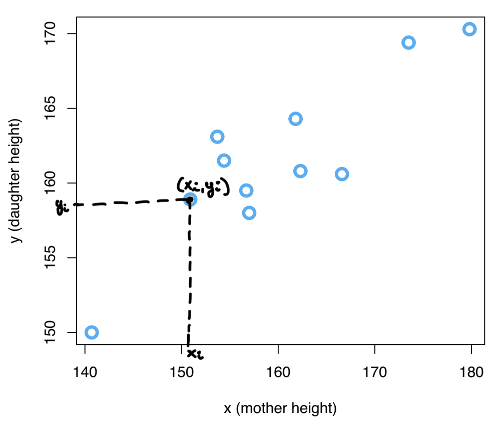
SIMPLE LINEAR MODEL VIA ORDINARY LEAST SQUARES (OLS)

Consider a linear regression, without the normality assumption for $Y_1, \dots, Y_n$.
We only make assumptions about the first two moments.

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, $i = 1, \dots, n$ $\rightarrow$ sample space $S = \mathcal{Y}^n = \mathbb{R}^n$
simple example: relationship between the height of 11 mothers (x_i) and the height of their daughters.

	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 may not hold exactly, in the sense that 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: the linear relationship is not "exact"

(β_1, β_2) regression coefficients
we only observe 1 covariate, but we also introduce one additional "variable" taking value 1 for each individual.

the model matrix hence is
$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$\Rightarrow$ β_1 is the intercept
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) $Var(\varepsilon_i) = \sigma^2 > 0 \quad i = 1, \dots, n$ (common variance across subjects)
(3) $cov(\varepsilon_i, \varepsilon_k) = 0$ if $i \neq k$, $i = 1, \dots, n \quad k = 1, \dots, n$

(1) $\mathbb{E}[\varepsilon_i] = 0 \quad i = 1, \dots, n$
"Absence of systematic error"
Implications for Y_i
$$\mathbb{E}[Y_i] = \mathbb{E}[\beta_1 + \beta_2 x_i + \varepsilon_i] \overset{\text{linearity of } \mathbb{E}}{=} \underbrace{\mathbb{E}[\beta_1 + \beta_2 x_i]}_{\text{non-stochastic}} + \underbrace{\mathbb{E}[\varepsilon_i]}_{0} = \beta_1 + \beta_2 x_i$$

What happens if there is a systematic error? i.e. $\mathbb{E}[\varepsilon_i] = c \neq 0$
$\mathbb{E}[Y_i] = \beta_1 + \beta_2 x_i + c = (\beta_1 + c) + \beta_2 x_i$
the systematic error c is inglobated in the intercept (no big deal!)
it is equivalent to a model
$Y_i = \beta_1^* + \beta_2 x_i + \varepsilon_i^*$ where $\beta_1^* = \beta_1 + c$
$\varepsilon_i^* = \varepsilon_i - c \Rightarrow \mathbb{E}[\varepsilon_i^*] = 0$

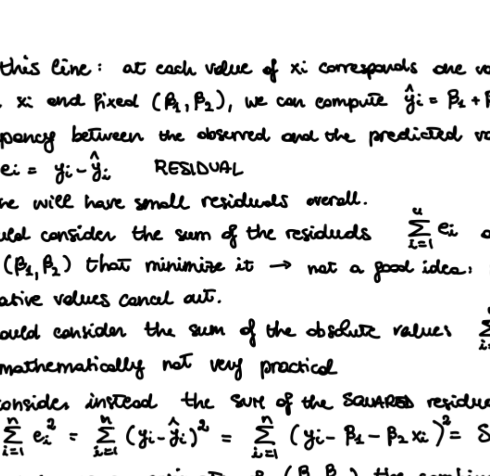
(2) $var(\varepsilon_i) = \sigma^2 > 0$ for all $i = 1, \dots, n$
"Homoscedasticity of the errors"
Implications for Y_i:
$$var(Y_i) = var(\underbrace{\beta_1 + \beta_2 x_i}_{\text{non-stoch.}} + \varepsilon_i) = var(\varepsilon_i) = \sigma^2 \quad \forall i = 1, \dots, n$$
$\Rightarrow$ homoscedasticity of the response

(3) $cov(\varepsilon_i, \varepsilon_k) = 0$ for $i \neq k$
"the errors are uncorrelated"
Implication for Y_i
$$cov(Y_i, Y_k) = cov(\underbrace{\beta_1 + \beta_2 x_i}_{\text{non-stochastic}} + \varepsilon_i, \underbrace{\beta_1 + \beta_2 x_k}_{\text{non-stochastic}} + \varepsilon_k) = cov(\varepsilon_i, \varepsilon_k) = 0$$

2nd step: ESTIMATE

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

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.
$\Rightarrow$ given x_i and fixed (β_1, β_2), we can compute $\hat{y}_i = \beta_1 + \beta_2 x_i$
The discrepancy between the observed and the predicted value (at the observed locations) is
$$e_i = y_i - \hat{y}_i \qquad \text{RESIDUAL}$$
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 $\rightarrow$ not a good idea: positive and negative values cancel out.
- we could consider the sum of the absolute values $\sum_{i=1}^n |e_i|$
$\rightarrow$ mathematically not very practical
- we consider instead the sum of the SQUARED residuals
$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2 = S(\beta_1, \beta_2)$$
and take as an estimate of (β_1, β_2) the combination that minimize it.

DEF: the LEAST SQUARES estimate of (β_1, β_2) is the combination of values $(\hat{\beta}_1, \hat{\beta}_2)$ that minimizes $S(\beta_1, \beta_2)$, i.e.
$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{(\beta_1, \beta_2) \in \mathbb{R}^2}{\arg\min} S(\beta_1, \beta_2) = \underset{(\beta_1, \beta_2) \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2$$

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 $\hat{\beta}_1, \hat{\beta}_2$ minimize $S(\beta_1, \beta_2) = \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2$.
We need to find the critical points (1st derivative $= 0$)
and then check that they are minimum (2nd derivative > 0)

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

$$\begin{cases} \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i) = 0 & (A) \\ \sum_{i=1}^n x_i (y_i - \beta_1 - \beta_2 x_i) = 0 & (B) \end{cases}$$

(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) $\sum_{i=1}^n x_i y_i - n\bar{x}\beta_1 - \beta_2 \sum_{i=1}^n x_i^2 = 0$ substituting $\beta_1 = \bar{y} - \beta_2 \bar{x}$
$\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} + n\beta_2 \bar{x}^2 - \beta_2 \sum_{i=1}^n x_i^2 = 0$
$$\hat{\beta}_2 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$
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
$\sum (x_i^2 + \bar{x}^2 - 2x_i\bar{x}) = \sum x_i^2 + n\bar{x}^2 - 2\bar{x}\sum x_i = \sum x_i^2 + n\bar{x} - 2n\bar{x}$
$(n-1)s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$
$\sum (x_i y_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y}) = \sum x_i y_i - n\bar{x}\bar{y} - \cancel{n\bar{x}\bar{y}} + \cancel{n\bar{x}\bar{y}}$

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
$$H = \begin{bmatrix} \frac{\partial^2 S(\beta_1, \beta_2)}{\partial \beta_1^2} & \frac{\partial^2 S(\beta_1, \beta_2)}{\partial \beta_1 \partial \beta_2} \\ \frac{\partial^2 S(\beta_1, \beta_2)}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 S(\beta_1, \beta_2)}{\partial \beta_2^2} \end{bmatrix} = \begin{bmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum_{i=1}^n x_i^2 \end{bmatrix}$$
$$\det(H) = 4n \sum_{i=1}^n x_i^2 - 4n^2\bar{x}^2 = 4n\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right) = 4n \sum_{i=1}^n (x_i - \bar{x})^2 > 0$$
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 $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$, 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 $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$
$\hat{\beta}_1$ is the intercept, i.e., the predicted value of y when $x = 0$. (Not always interpretable!) eg heights

Now consider two individuals observed at $x_1 = x_0$ and $x_2 = x_0 + 1$
The predicted values are $\hat{y}_1 = \hat{\beta}_1 + \hat{\beta}_2 x_0$
$\hat{y}_2 = \hat{\beta}_1 + \hat{\beta}_2 (x_0 + 1)$
$$\hat{y}_2 - \hat{y}_1 = \cancel{\hat{\beta}_1} + \hat{\beta}_2(x_0 + 1) - \cancel{\hat{\beta}_1} - \hat{\beta}_2 x_0 = \cancel{\hat{\beta}_2 x_0} + \hat{\beta}_2 - \cancel{\hat{\beta}_2 x_0} = \hat{\beta}_2$$
Hence $\hat{\beta}_2$ is the expected change in y when I increase x of 1 unit