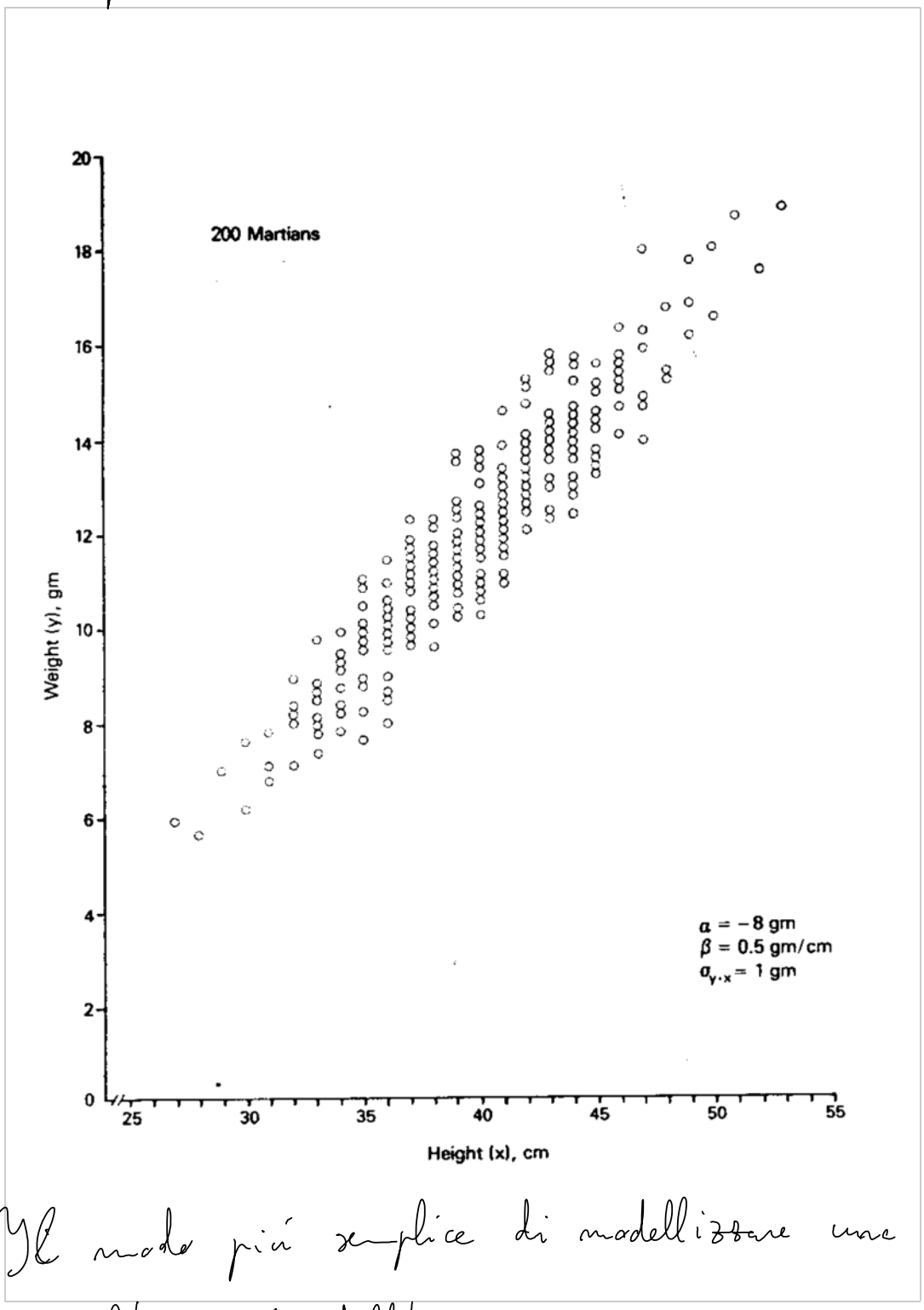
REGRESSIONE LINEARE

giovedì 5 giugno 2014 08.54

Supponiamo di registrare altezze e pesi dei 200 marziani, con questi risultati:



Il modo più semplice di modellizzare una dipendenza tra altezza X_i dell'i-esimo marziano e peso Y_i è

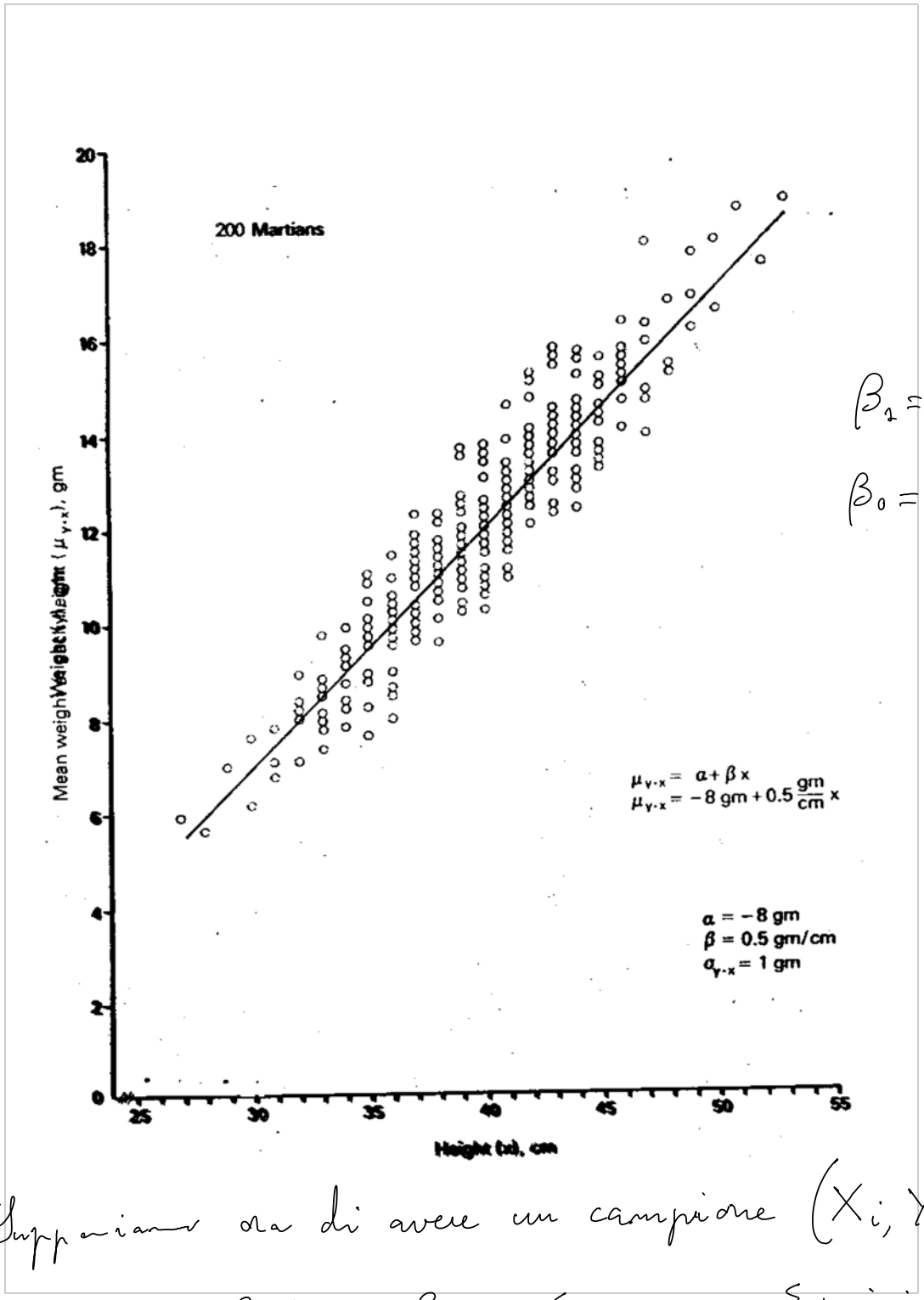
$$Y_i = \beta_1 X_i + \beta_0 + \varepsilon_i$$

con $\varepsilon_i \sim N(0, \sigma^2_{Y|X})$

La retta $y = \beta_1 x + \beta_0$ si dice retta delle medie

In questo modello ci sono 3 parametri: $\beta_1, \beta_0, \sigma^2_{Y|X}$

β_1 = coeff. angolare, β_0 = ordinata all'origine



$\beta_1 = 0{,}5\ g/cm$

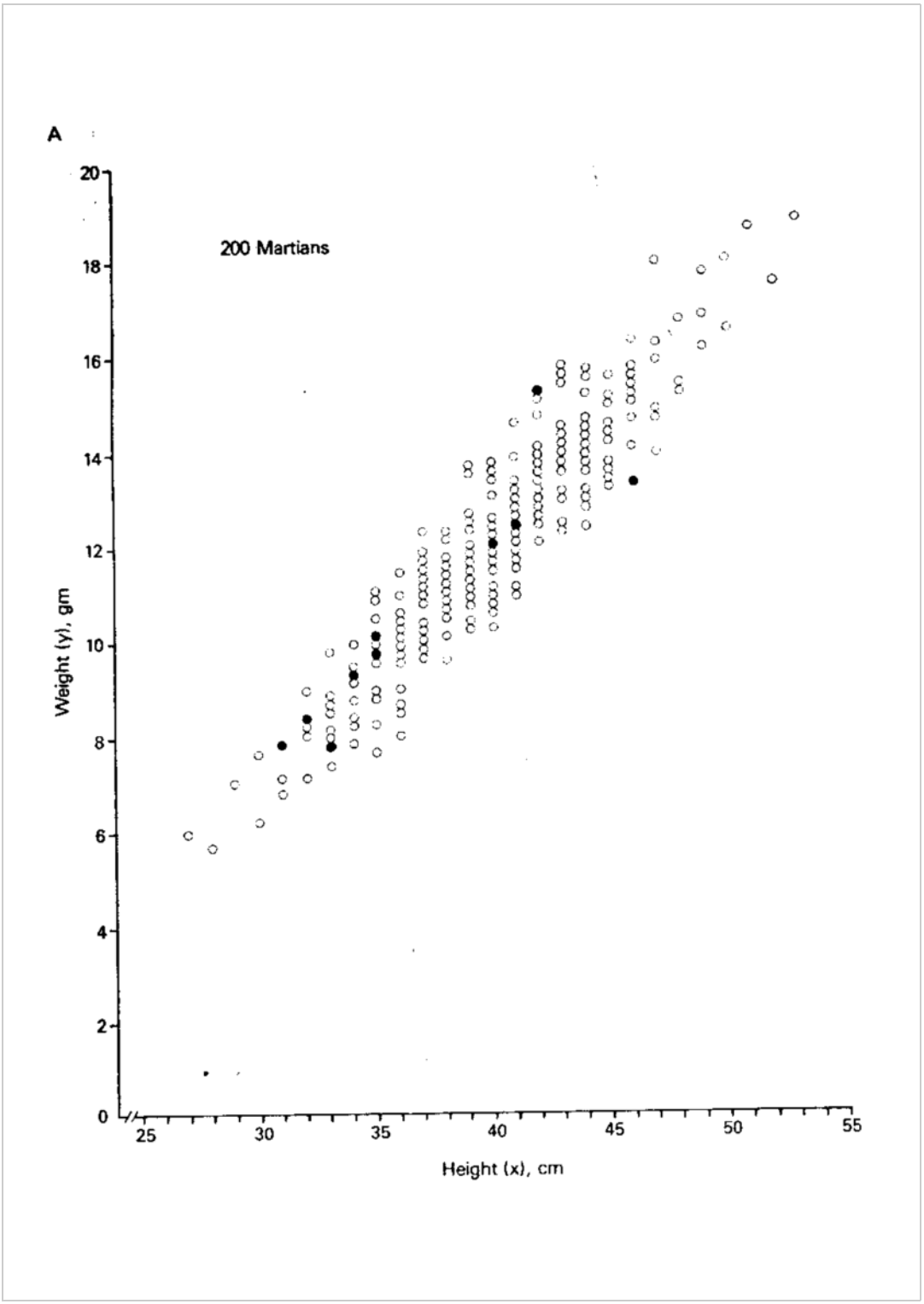
$\beta_0 = -8\ g$

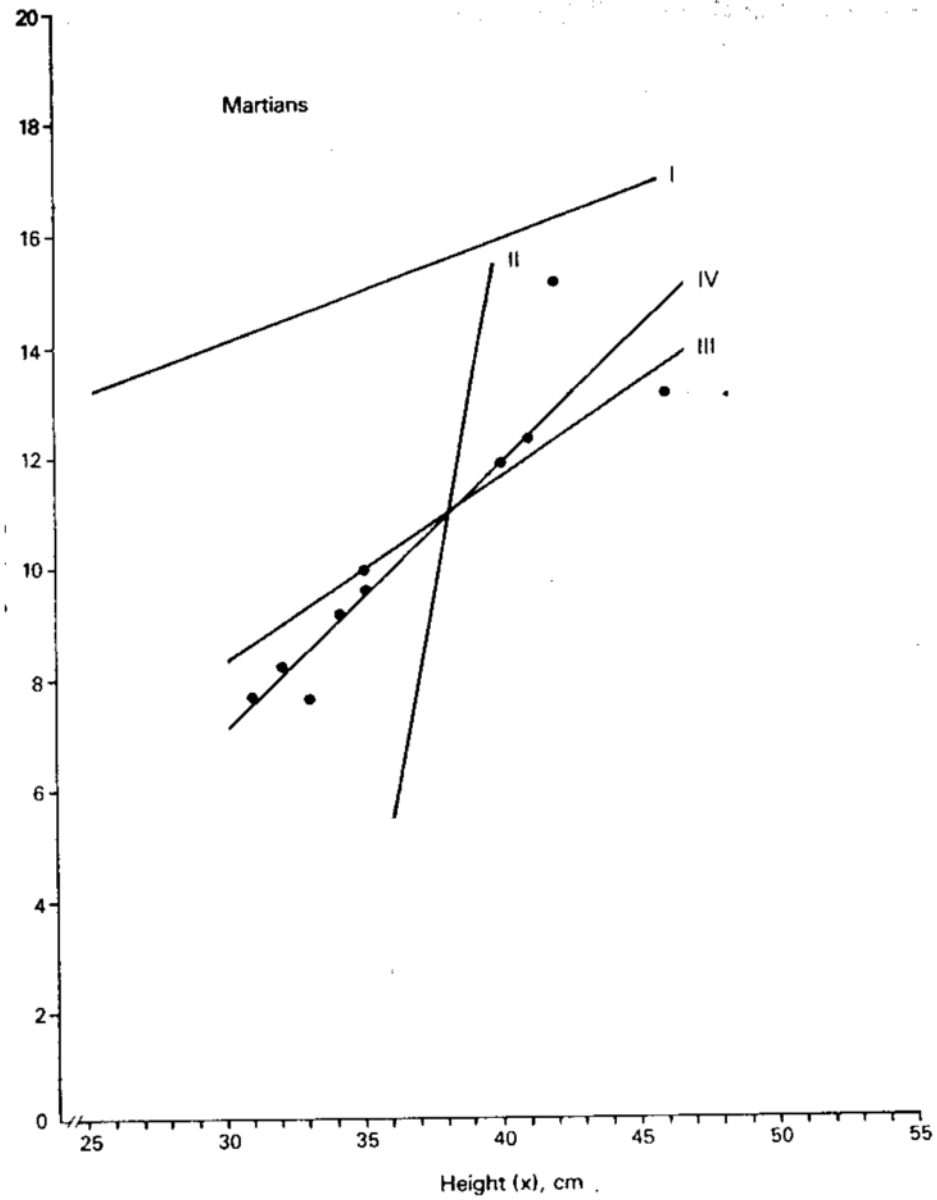
Supponiamo di avere un campione (X_i, Y_i), $i = 1, \dots, n$

con $Y_i = \beta_1 X_i + \beta_0 + \varepsilon_i$, con ε_i i.i.d. $\sim N(0, \sigma^2_{Y|X})$

($\sigma^2_{Y|X}$ = varianza del rumore)

Vogliamo stimare i parametri β_1, β_0 e $\sigma^2_{Y|X}$


A
200 Martians
Weight (y), gm
Height (x), cm
0
2
4
6
8
10
12
14
16
18
20
25
30
35
40
45
50
55



Per decidere quali sono gli stimatori "migliori", prendiamo quelli che minimizzano il quadrato delle distanze delle y:

$$f(b_0, b_1) = \sum_1^n{}_i \left(Y_i - \underbrace{b_1 X_i - b_0}_{\text{della retta delle medie}} \right)^2$$

Y_i: dei dati

Gli stimatori migliori sono i b_0, b_1 che minimizzano f:

$$\sum X_i Y_i - n \bar{X} \bar{Y}$$

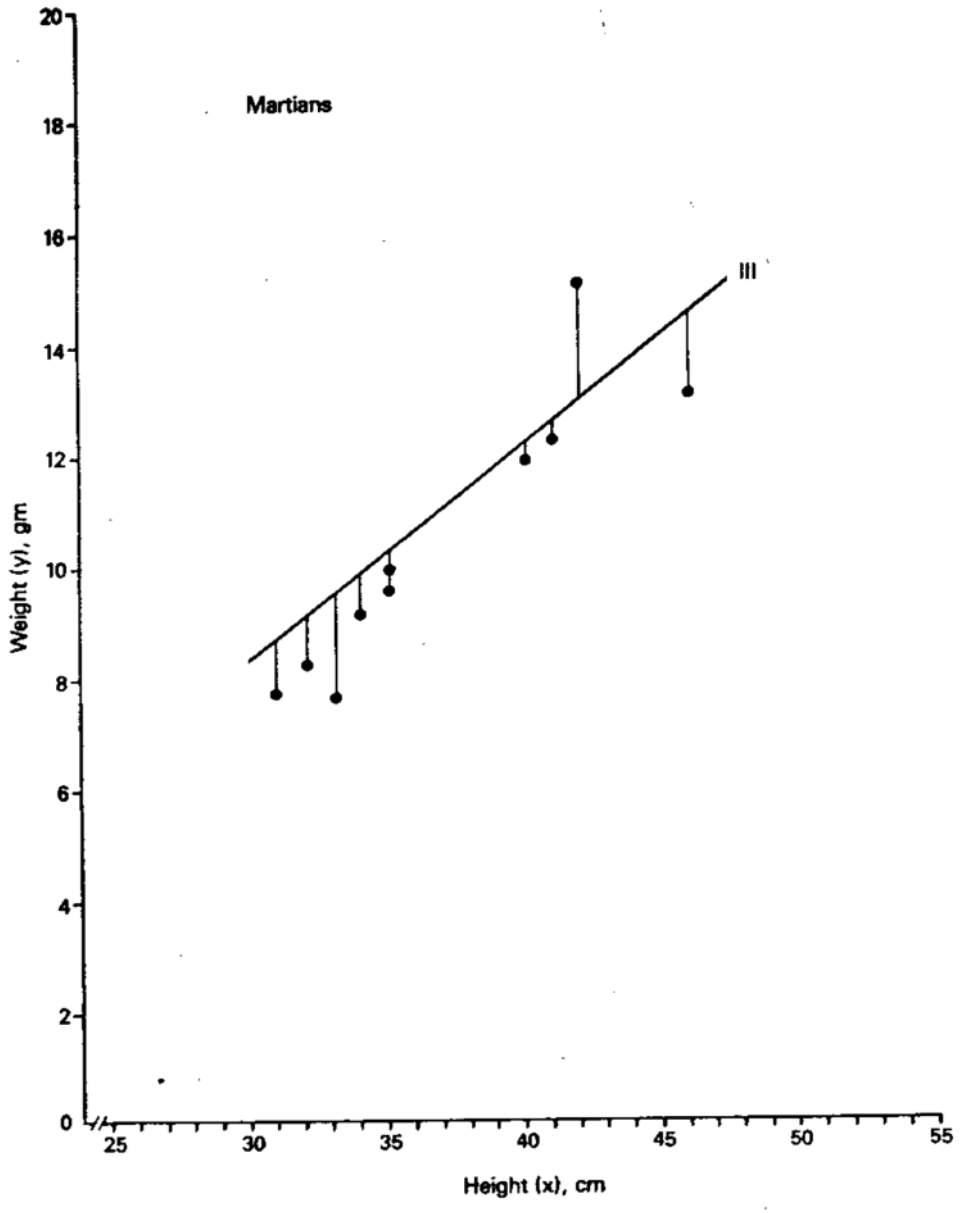
$$b_1 = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}, \qquad b_0 = \bar{Y} - b_1 \bar{X}$$

$y = b_1 x + b_0$ <u>retta di regressione</u>

<u>nota</u>: $\sum_1^n X_i^2 - n\bar{X}^2 = (n-1) s_x^2$. Allo stesso modo,

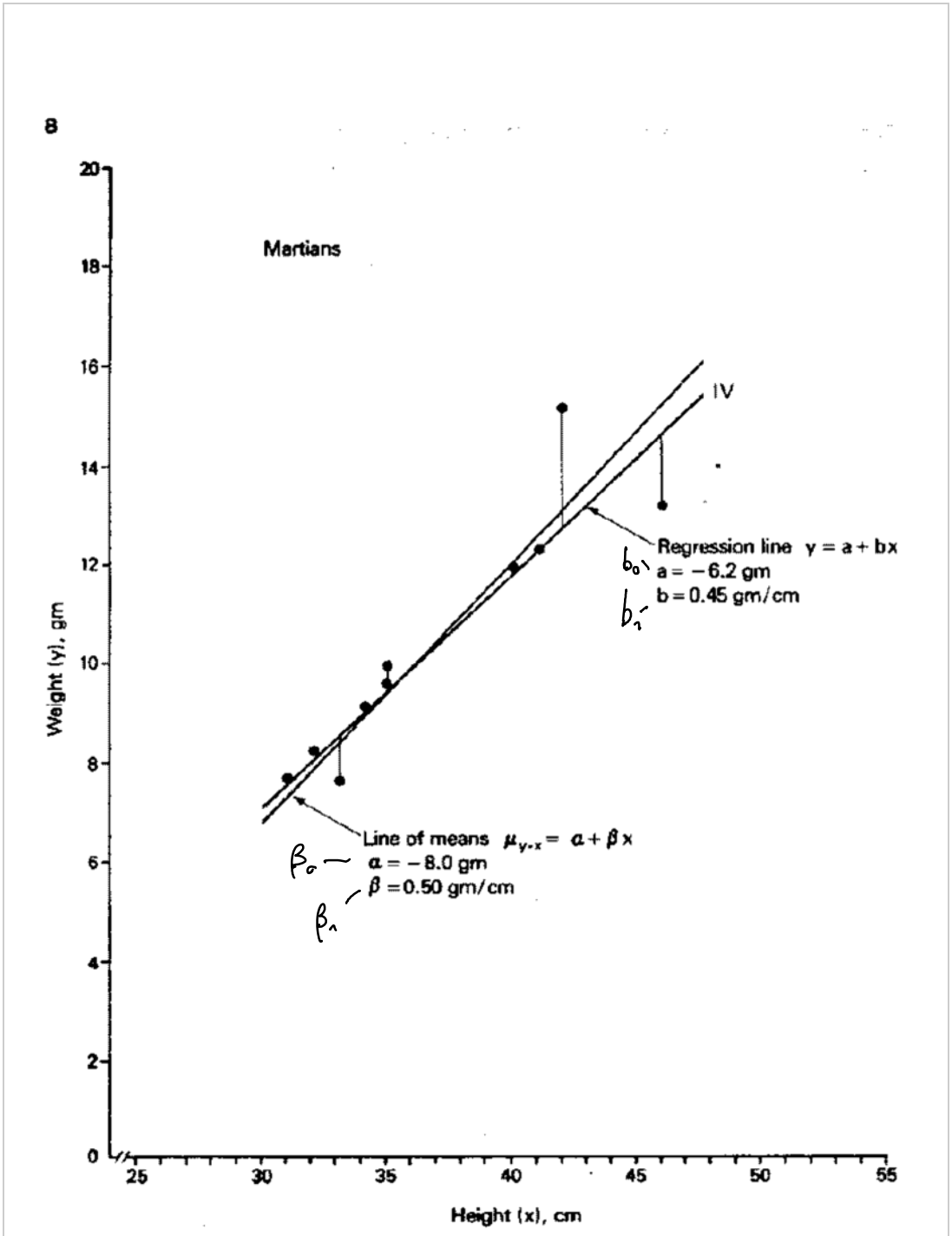
definiamo $s_{XY} = \frac{1}{n-1}\left(\sum_1^n X_i Y_i - n\bar{X}\bar{Y}\right)$

<u>covarianza campionaria</u>: è uno stimatore di $Cov(X, Y)$



Passiamo ricevere

$$b_1 = \frac{S_{XY}}{S_X^2}$$



b_0

b_1

β_0

β_1

stima della varianza del rumore:

$$\sigma^2_{Y|X} = Var[\varepsilon_i] = Var[Y_i - \beta_1 X_i - \beta_0]$$

$$\sigma^2_{Y|X} = \mathrm{Var}[\varepsilon_i] = \mathrm{Var}[Y_i - \beta_1 X_i - \beta_0]$$

può essere stimata da

$$s^2_{Y|X} := \frac{1}{n-2} \sum_1^n (Y_i - b_1 X_i - b_0)^2$$

che è una stimatore corretta, cioè $E[s^2_{Y|X}] = \sigma^2_{Y|X}$

y_i può dimostrare che $s^2_{Y|X} = \frac{n-1}{n-2}\left(s^2_Y - b_1^2 s^2_X\right)$

esempio (es. 9.1) retta di regressione per

X_i	Y_i	X_i^2	Y_i^2	$X_i Y_i$
3	2,7	9	7,29	8,1
4	3,6	16	12,96	14,4
1	2,1	1	4,41	2,1
2	2,6	4	6,76	5,2
TOT. 10	11,0	30	31,42	29,8

$n = 4$

$$\bar{X} = \frac{\sum X_i}{n} = \frac{10}{4} = 2,5 \qquad \bar{Y} = \frac{\sum Y_i}{n} = \frac{11}{4} = 2,75$$

$$s^2_X = \frac{1}{n-1}\left(\sum X_i^2 - n\bar{X}^2\right) = \frac{1}{3}\left(30 - 4 \cdot 2,5^2\right) = 1,67$$

$$s^2_Y = \frac{1}{n-1}\left(\sum_1^n Y_i^2 - n\bar{Y}^2\right) = \frac{1}{3}\left(31,42 - 4 \cdot 2,75^2\right) = 0,39$$

$$s_{XY} = \frac{1}{n-1}\left(\sum X_i Y_i - n\bar{X}\bar{Y}\right) = \frac{1}{3}\left(29,8 - 4 \cdot 2,5 \cdot 2,75\right) = 0,27$$

$$b_1 = \frac{s_{XY}}{s^2_X} = \frac{0,27}{1,67} = 0,46$$

$$b_0 = \bar{Y} - b_1 \bar{X} = 2,75 - 0,46 \cdot 2,5 = 1,60$$

$$s^2_{Y|X} = \frac{n-1}{n-2}\left(s^2_Y - b_1^2 s^2_X\right) = \frac{3}{2}\left(0,39 - 0,46^2 \cdot 1,67\right) = 0,056$$

$$s_{Y|X} = \sqrt{s^2_{Y|X}} = \sqrt{0,056} = 0,236$$