FoDA L11

Linear Regression

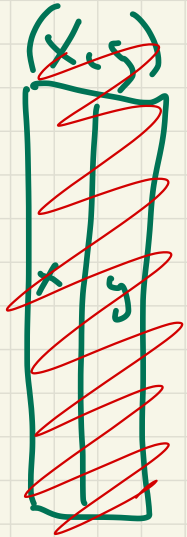
Input $(X, y) = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$

$X \in \mathbb{R}^{n \times 1}$

$y \in \mathbb{R}^n$

$x_i \in \mathbb{R}$ explanatory variable

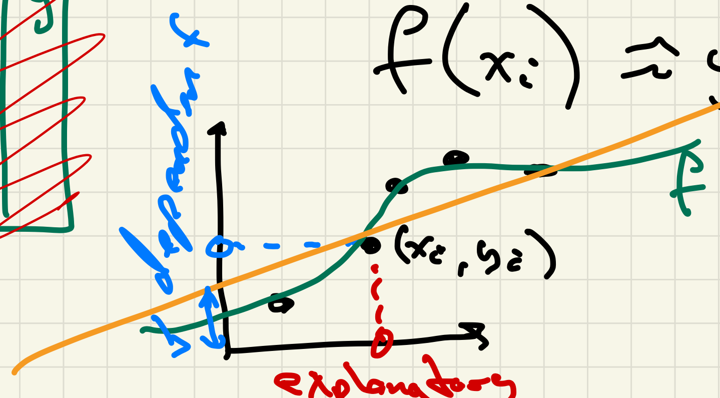
$y_i \in \mathbb{R}$ dependent variable



$f : \mathbb{R}^1 \to \mathbb{R}^1$ (d)

$x_i \qquad y_i$

$f(x_i) \approx y_i$



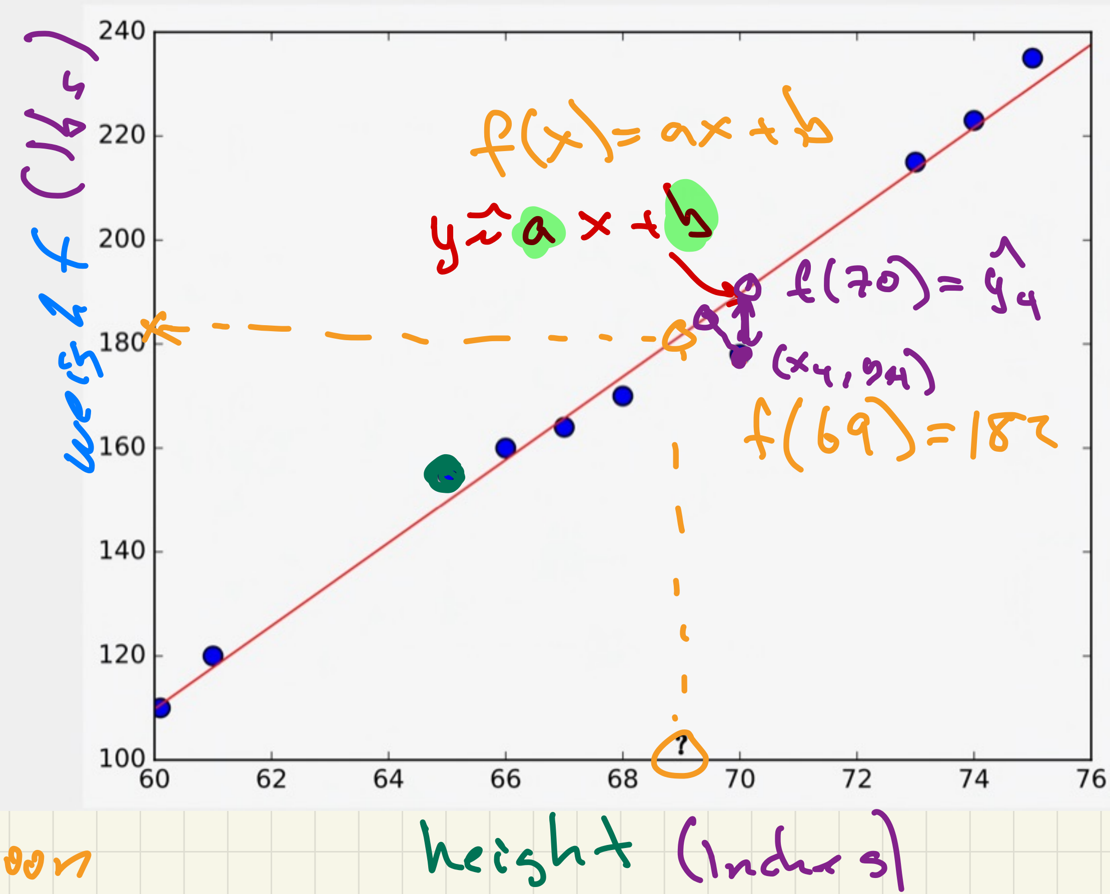
linear function

$y_i \approx a \, x_i + b$

slope (a), offset (b)

height (in)	weight (lbs)
66	160
68	170
60	110
70	178
65	155
61	120
74	223
73	215
75	235
67	164
69	?

prediction



Measure Error

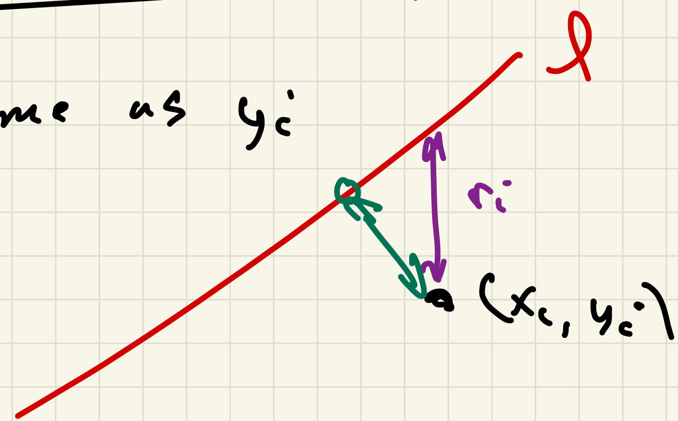
residual for model $\ell : \mathbb{R} \to \mathbb{R}$

on data point (x_i, y_i)

$$r_i = (y_i - \hat{y}_i) = (y_i - \ell(x_i))$$

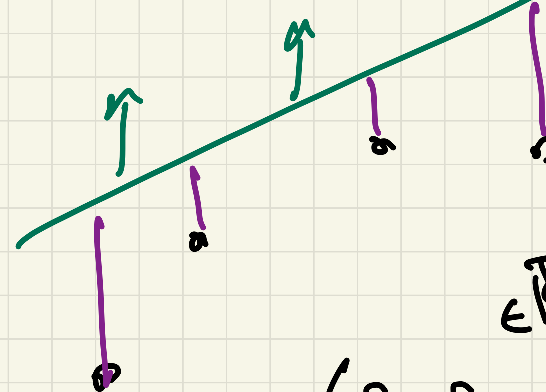
---

units of r_i same as y_i



Overall error

Sum $\sum_{i=1}^{n} r_i < 0$



$\in \mathbb{R}^n$

$r = (r_1, r_2, \dots, r_n)$

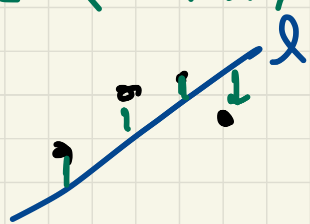
$r_i = y_i - \hat{y}_i$

Sum of Squared Errors

$\|r\|_1 = \sum_{i=1}^{n} |r_i|$ or $\sum_{i=1}^{n} r_i^2 = \|r\|_2^2$

↳ closed form algorithm

$$SSE((x,y), \ell) = \sum_{i=1}^{n} (r_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \ell(x_i))^2$$



$y_i = \ell(x_i) + N(0, \sigma^2)$ $\ell(x) = ax + b$

Input $(x,y) = \{(x_1,y_1), \ldots, (x_n,y_n)\} \in \mathbb{R} \times \mathbb{R}$

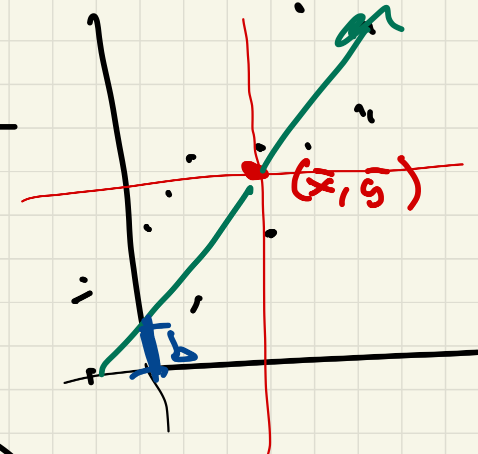
Goal Find $l(x) = ax + b$

↳ value a, b

to minimize

$$g(a,b) = SSE((x,y), l) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Solving for a, b



- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\qquad$ $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

- Normalize data (center)

$$\bar{X} = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$$

$$\bar{Y} = (y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y})$$

1. $a = \dfrac{\langle \bar{X}, \bar{Y} \rangle}{\|\bar{X}\|^2}$ $\qquad$ 2. $b = \bar{y} - a\bar{x}$

$\theta = \text{angle}(\bar{X}, \bar{Y})$

$$= \frac{\|\bar{X}\| \cdot \|\bar{Y}\| \cdot \cos(\theta)}{\|X\|^2} = \frac{\|\bar{Y}\|}{\|\bar{X}\|} \cos(\theta)$$

weight ($\|\bar{Y}\|$)

height ($\|\bar{X}\|$)