

Hoofdstuk 12 :

Regressie en correlatie

Marnix Van Daele

Marnix.VanDaele@UGent.be

Vakgroep Toegepaste Wiskunde en Informatica
Universiteit Gent

Regressielijn

Vraag : vind het verband $Y = F(X)$.

Werkwijze : steekproef waarbij X ingesteld en Y gemeten wordt.

Worden $(x_i, y_i), i = 1, \dots, n$ uitgezet in grafiek, dan zal dit een beeld geven over het functionele verband : lineair, kwadratisch, exponentieel, ...

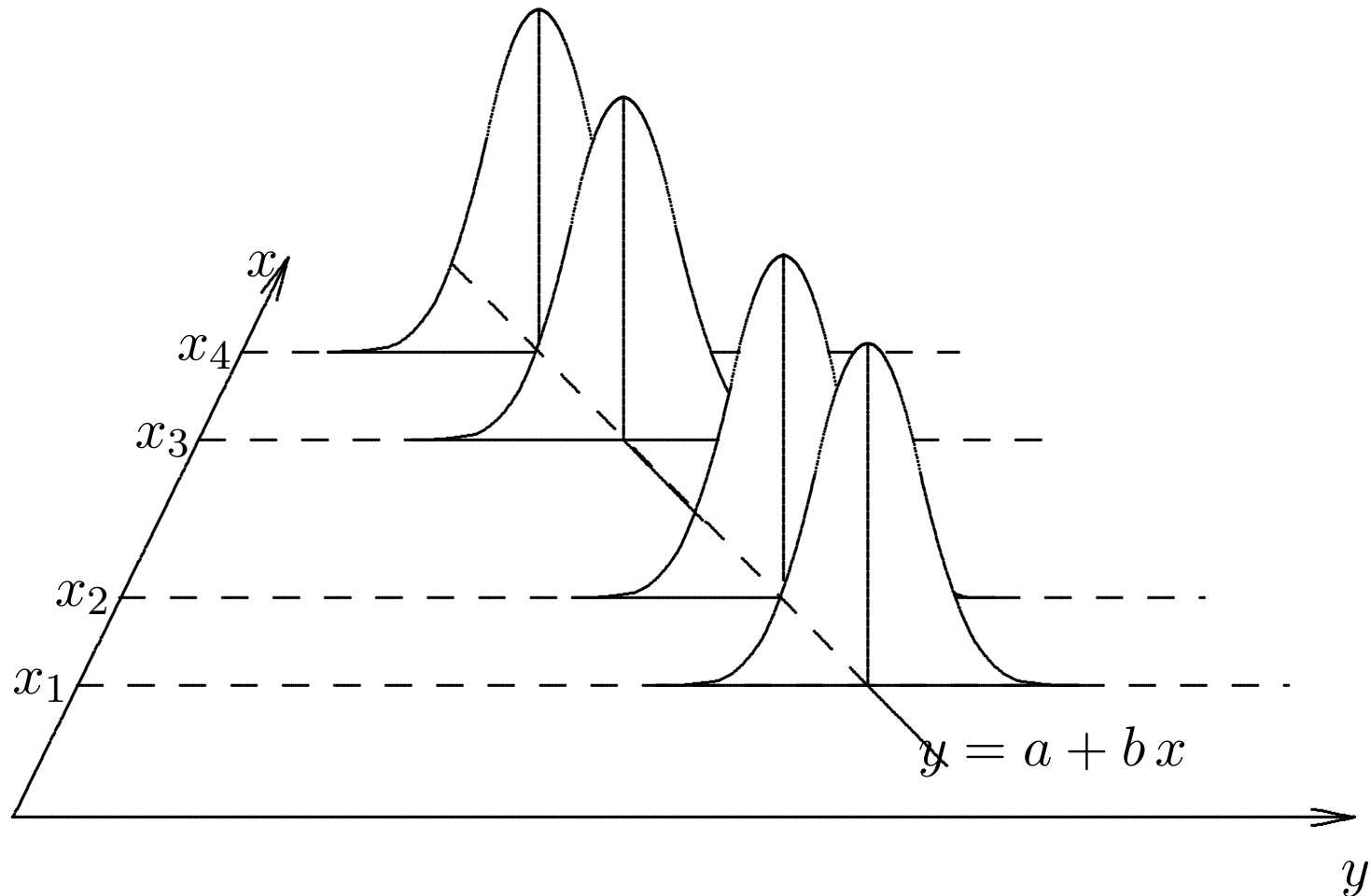
$y = F(x)$: regressielijn van Y op X

Maar : Y is onderhevig aan meetfouten

$$Y = F(X) + \epsilon \quad \epsilon : N(0, \sigma)$$

$$E[Y \mid X = x] = F(x)$$

Regressie : lineair model $Y = a + b X$



De best passende kromme

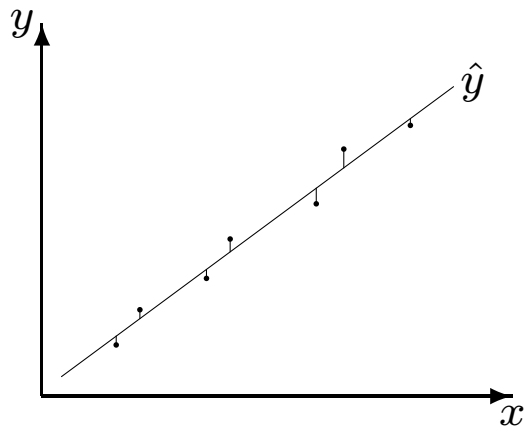
Eens de aard (lineair, kwadratisch, ...) van $F(x)$ bepaald is, moet binnen deze klasse van functie de beste oplossing nog gezocht worden.

principe der kleinste kwadraten : de best passende kromme is de kromme die de som van de kwadraten van de afwijkingen van de waargenomen waarden t.o.v. de hypothetische kromme minimaal maakt.

$$\text{minimaliseer } \sum_{i=1}^n (y_i - F(x_i))^2$$

Lineaire regressie : $Y = a + b X$

minimaliseer $\sum_{i=1}^n [y_i - F(x_i)]^2 = \sum_{i=1}^n [y_i - (a + b x_i)]^2 = S$



$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases} \implies \begin{cases} -2 \sum_{i=1}^n [y_i - (a + b x_i)] = 0 \\ -2 \sum_{i=1}^n [y_i - (a + b x_i)] x_i = 0 \end{cases}$$

$$\iff \begin{cases} n a + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases} \iff \begin{cases} a + b \bar{x} = \bar{y} \\ a \bar{x} + b \overline{x^2} = \overline{xy} \end{cases}$$

Lineaire regressie : $Y = a + b X$

$$\text{minimaliseer } \sum_{i=1}^n [y_i - F(x_i)]^2 = \sum_{i=1}^n [y_i - (a + b x_i)]^2$$

$$\Rightarrow \begin{cases} a \bar{x} + b \bar{x}^2 = \bar{y} \bar{x} \\ a \bar{x} + b \bar{x}^2 = \bar{x} \bar{y} \end{cases} \iff \begin{cases} b (\bar{x}^2 - \bar{x}^2) = \bar{x} \bar{y} - \bar{x} \bar{y} \\ a + b \bar{x} = \bar{y} \end{cases}$$

$$\Rightarrow \begin{cases} a = \bar{y} - b \bar{x} \\ b = \frac{\bar{x} \bar{y} - \bar{x} \bar{y}}{\bar{x}^2 - \bar{x}^2} = \frac{s_{xy}}{s_{xx}} \end{cases}$$

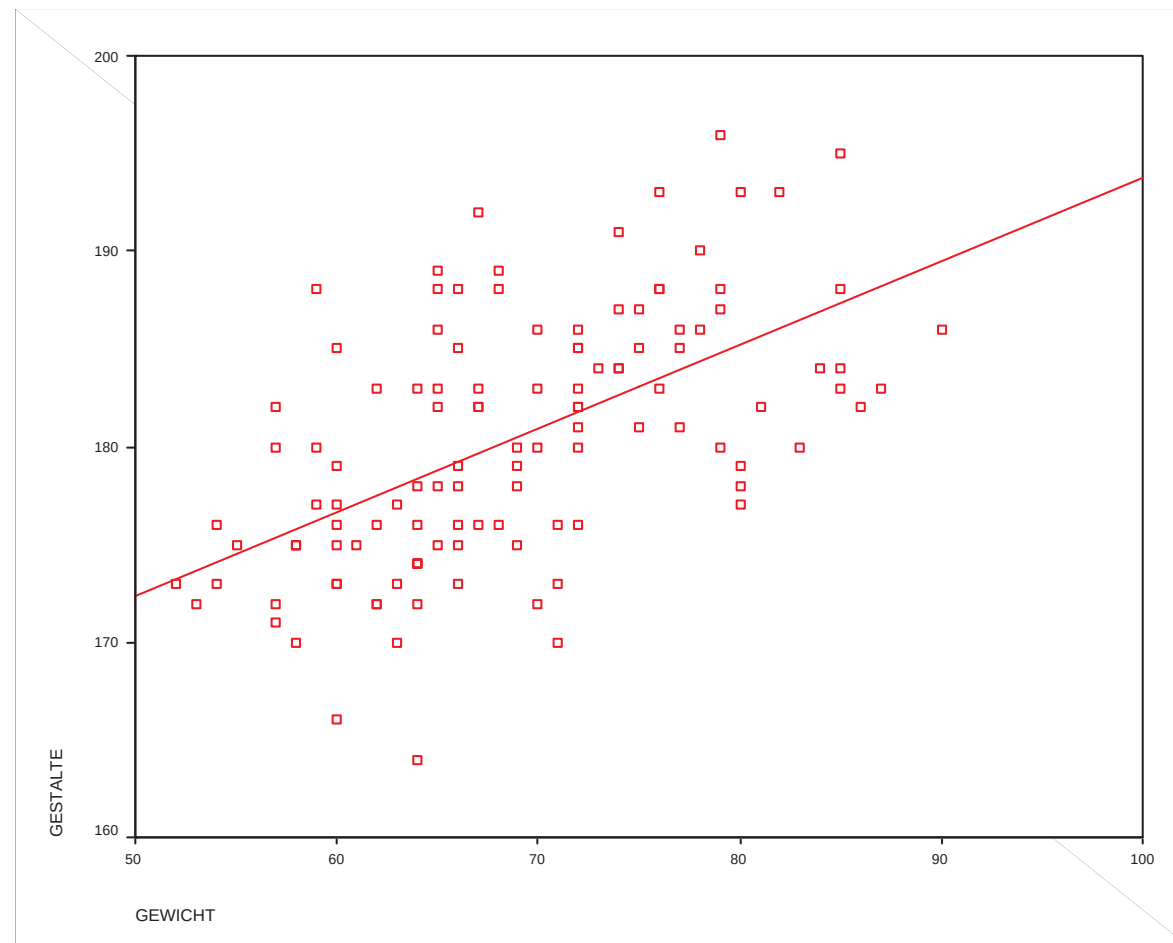
$y = a + b x$: lineaire regressielijn

b : regressiecoëfficiënt (van Y op X)

Voorbeeld : voorbeeldsteekproef

lineair verband tussen X =gewicht en Y =gestalte voor mannen :

regressielijn van Y op X : $y = 151.108 + 0.4261 x$.



Voorbeeld : voorbeeldsteekproef

lineair verband tussen X =gewicht en Y =gestalte voor mannen :

regressielijn van Y op X : $y = 151.108 + 0.4261 x$.

Wat als de rollen van X en Y worden omgekeerd ?

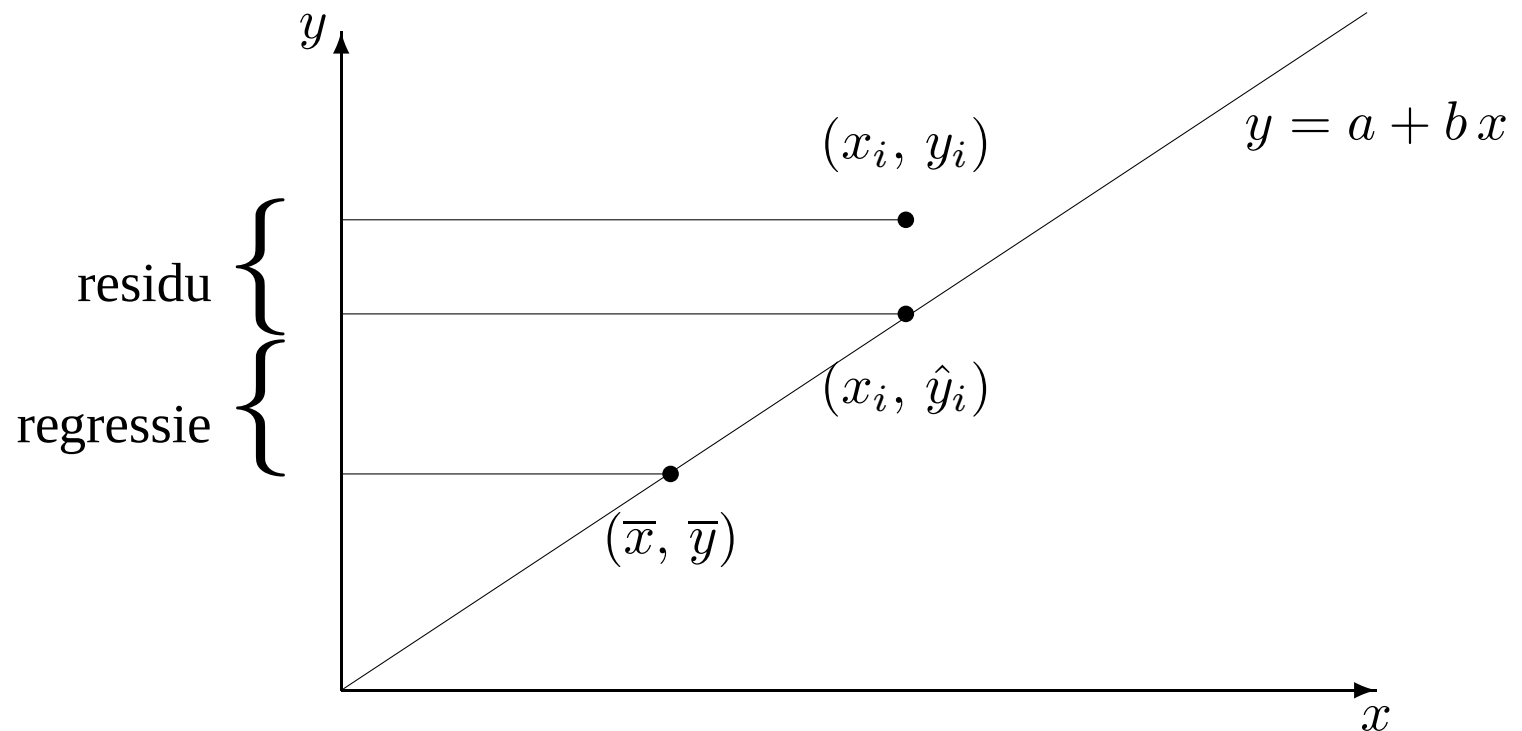
Y : ingesteld en X gemeten

regressielijn van Y op X : $x = -68.133 + 0.7592 y$

Dit is een andere rechte !

Lineaire regressie

$$y = a + b x \quad \text{met} \quad \begin{cases} a = \bar{y} - b \bar{x} \\ b = \frac{\overline{xy} - \bar{x} \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{s_{xy}}{s_{xx}} \end{cases} \quad \text{stel } \hat{y}_i = a + b x_i$$



$$\begin{aligned} y_i - \bar{y} &= (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \\ &= \text{residu} + \text{regressie} \end{aligned}$$

Lineaire regressie

$$y = a + b x \text{ met } \begin{cases} a = \bar{y} - b \bar{x} \\ b = \frac{\overline{xy} - \bar{x} \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{s_{xy}}{s_{xx}} \end{cases} \quad \text{stel } \hat{y}_i = a + b x_i$$

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

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

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

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

$$= s_{yy} - 2b s_{xy} + b^2 s_{xx} = s_{yy} - \frac{s_{xy}^2}{s_{xx}} \leq s_{yy}$$

Lineaire regressie

$$y = a + b x \quad \text{met} \quad \begin{cases} a = \bar{y} - b \bar{x} \\ b = \frac{\overline{xy} - \bar{x} \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{s_{xy}}{s_{xx}} \end{cases} \quad \text{stel } \hat{y}_i = a + b x_i$$

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \leq \sum_{i=1}^n (y_i - \bar{y})^2$$

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

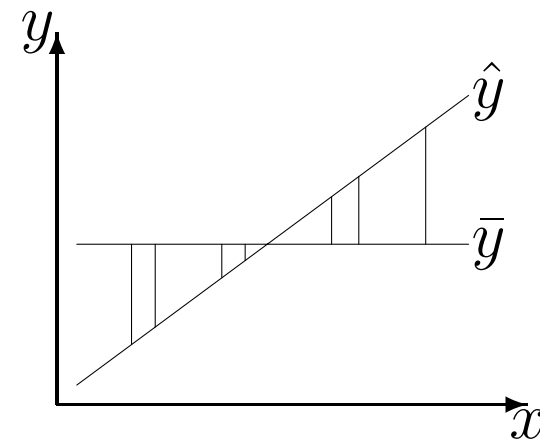
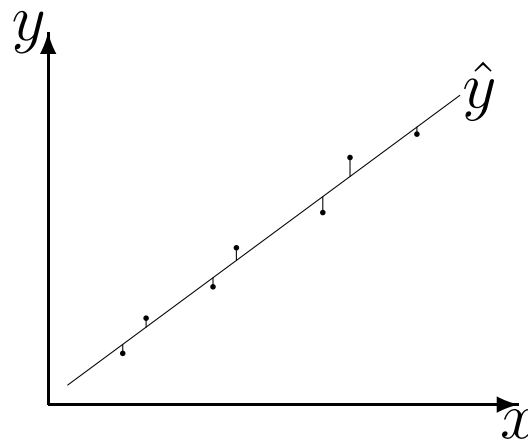
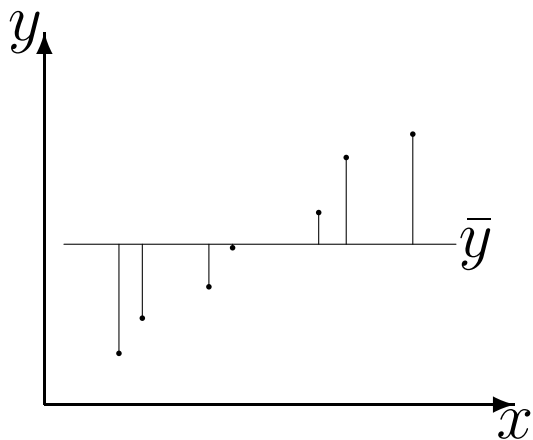
Componenten van variabiliteit

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

Som van kwadraten
rond het gemiddelde

Som van kwadraten
t.o.v. de regressielijn

Som van kwadraten
te wijten aan regressie



Componenten van variabiliteit

$$\sum_{i=1}^n (y_i - \bar{y})^2$$

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

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

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

als

$$\sum_{i=1}^n (y_i - \hat{y}_i) (\hat{y}_i - \bar{y}) = 0$$

Componenten van variabiliteit

$$\sum_{i=1}^n (y_i - \hat{y}_i) (\hat{y}_i - \bar{y})$$

$$= \sum_{i=1}^n (y_i - \hat{y}_i) (a - \bar{y} + b x_i)$$

$$= (a - \bar{y}) \sum_{i=1}^n (y_i - \hat{y}_i) + b \sum_{i=1}^n x_i (y_i - \hat{y}_i)$$

$$= (a - \bar{y}) \sum_{i=1}^n (y_i - a - b x_i) + b \sum_{i=1}^n x_i (y_i - a - b x_i)$$

$$= (a - \bar{y}) n (\bar{y} - a - b \bar{x}) + b n (\overline{xy} - a \bar{x} - b \overline{x^2}) = 0$$

$$\text{want } \begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases} \implies \begin{cases} a + b \bar{x} = \bar{y} \\ a \bar{x} + b \overline{x^2} = \overline{xy} \end{cases}$$

a en b

De waarden a en b zijn steekproefafhankelijk, dus het zijn op hun beurt steekproefwaarden van veranderlijken A en B .

We noemen de gevonden waarden $\hat{a}$ en $\hat{b}$ en ze stellen schattingen voor van de ware a en de ware b .

$$\sigma_A = \frac{\sigma}{s_x} \sqrt{\frac{x^2}{n}} \quad \sigma_B = \frac{\sigma}{s_x \sqrt{n}} \quad Y = F(X) + \epsilon \quad \epsilon : N(0, \sigma)$$

Betekenis σ : ongeveer 95 % van alle meetpunten liggen hoogstens 2σ van de regressielijn af.

$$\sigma^2 \approx \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{n}{n-2} \left(s_y^2 - \frac{s_{xy}^2}{s_x^2} \right)$$

$$\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 : \chi^2(n-2)$$

Toevalsveranderlijken

Op A en B kunnen bvb. testen uitgevoerd worden.

$$H_0 : b = b_0$$

$$H_1 : b \neq b_0$$

Als H_0 waar is, is $\frac{B - b_0}{\sigma_B} : T(n - 2)$

Voorbeeldsteekproef : gestalte op gewicht voor mannen

met $b_0 = 0$: voor $\alpha = 5\%$ verwerpen we H_0 als $|t| > 1.98$.

$$\sigma_B = \frac{\sigma}{s_x \sqrt{n}} \quad \sigma^2 \approx \frac{1}{n - 2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\sigma = \sqrt{\frac{1544.006}{115}} \quad \text{en} \quad s_x = 8.56$$

$$t = \frac{0.426}{0.039} = 10.76.$$

Dit is een zeer significant resultaat.

Schatten en voorspellen

De regressielijn kan gebruikt worden

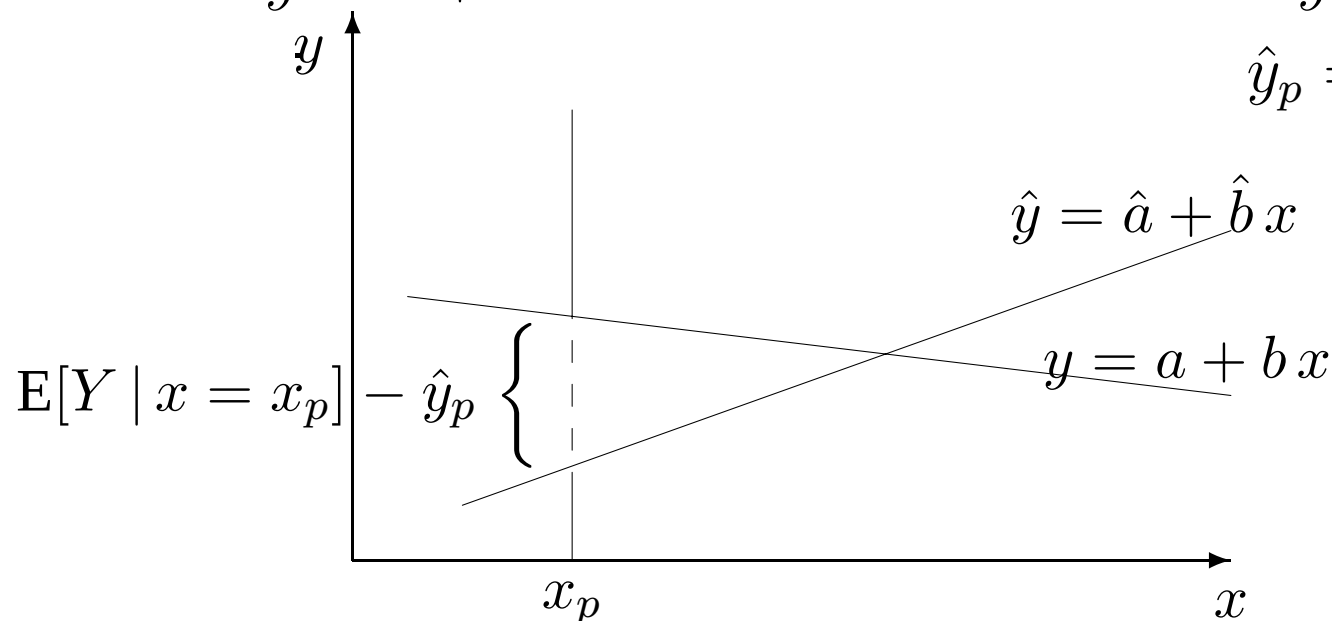
- om een gemiddelde waarde $E[Y \mid X = x_p]$ te Schatten
- om een toekomstige meetwaarde y_p te voorspellen

Schatten van $E[Y | X = x_p]$

Ware verband $y = a + b x$

Geschatte verband $\hat{y} = \hat{a} + \hat{b} x$

$$\hat{y}_p = \hat{a} + \hat{b} x_p$$



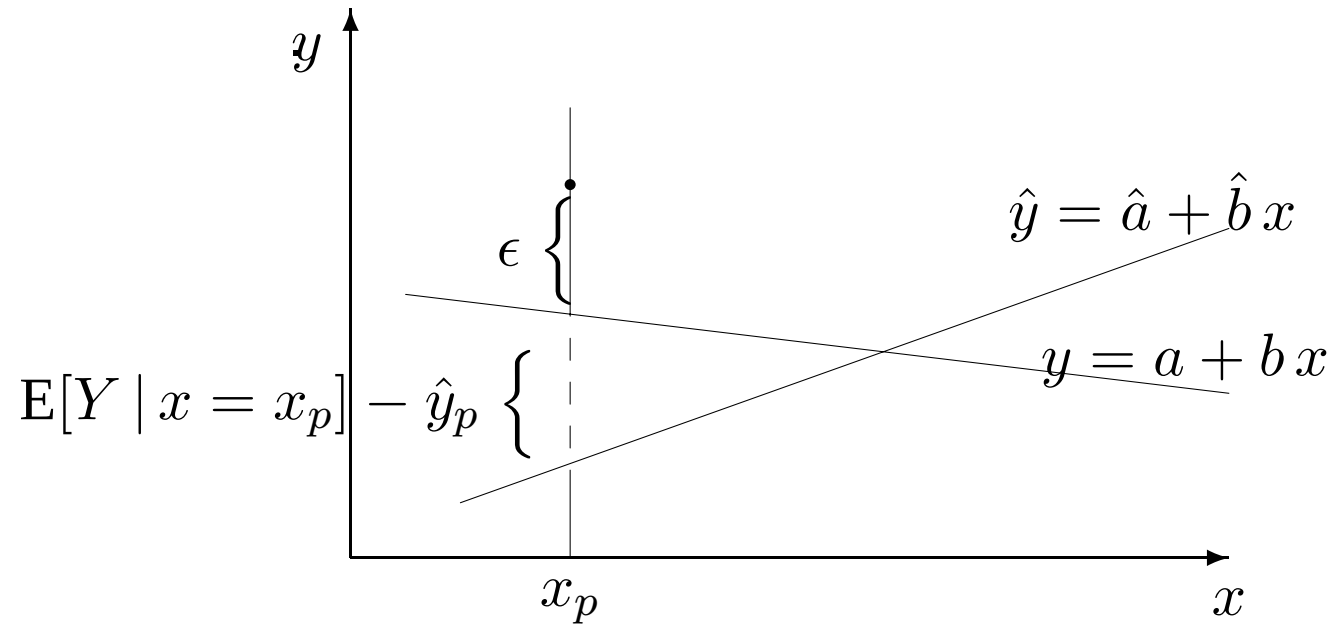
$\hat{y}_p$: onvertekende schatter voor $E[Y | X = x_p]$

$$\hat{y}_p : N(E[Y | X = x_p], \sigma_{\hat{y}_p})$$

$$\sigma_{\hat{y}_p} = \sqrt{\frac{\sigma^2}{n} \left[1 + \frac{(x_p - \bar{x})^2}{s_x^2} \right]}$$

$$\frac{E[Y | X = x_p] - \hat{y}_p}{\sigma_{\hat{y}_p}} \sim T(n - 2)$$

Voorspellen van toekomstige waarde



$$y - \hat{y} = (y - E[Y | X = x]) + (E[Y | X = x] - \hat{y})$$

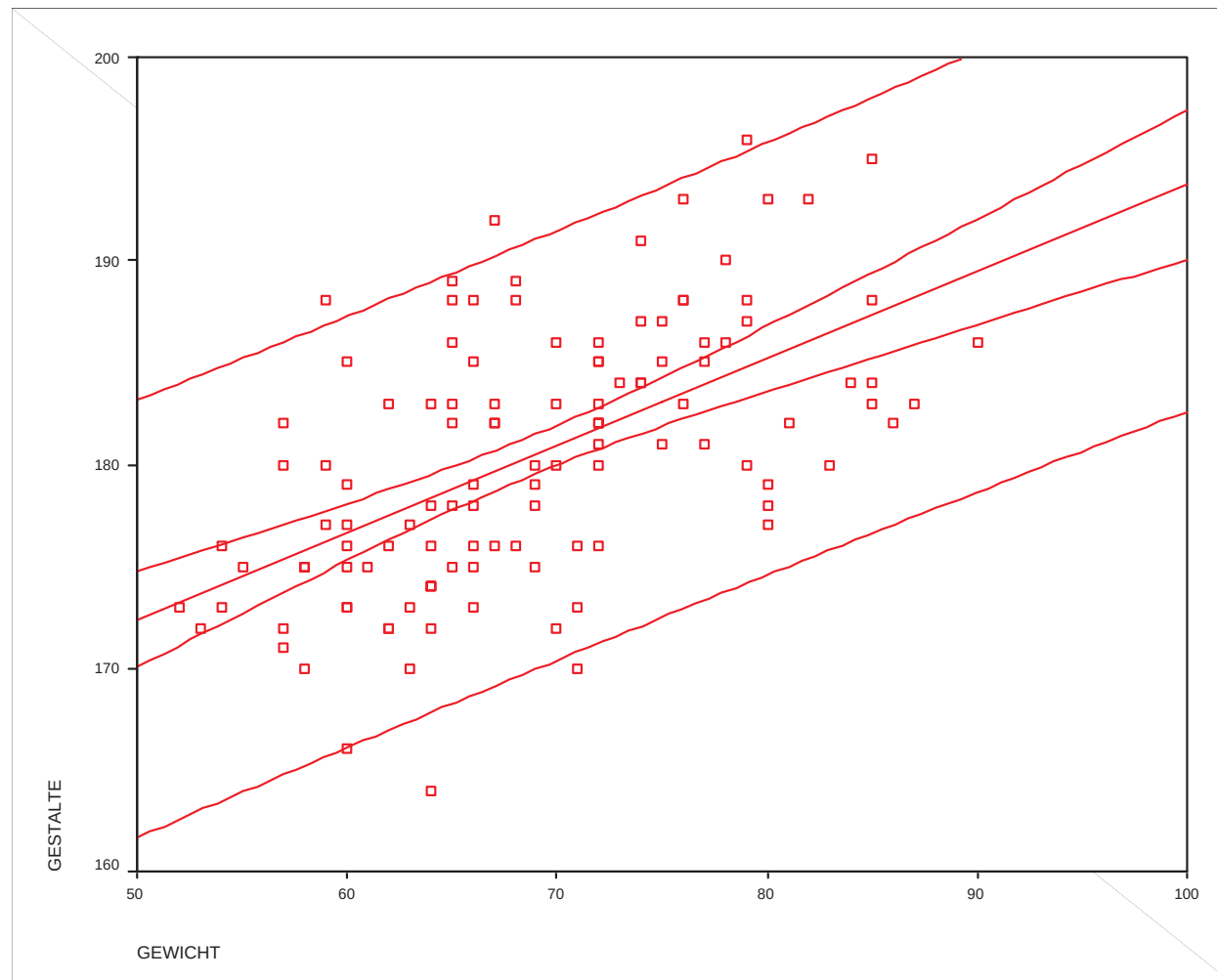
$E[Y | X = x] - \hat{y}$: fout in de regressie

$y - E[Y | X = x]$: random fout ϵ .

$$\begin{aligned} \sigma_{y-\hat{y}}^2 &= \sigma_y^2 + \sigma_{\hat{y}}^2 = \sigma^2 + \frac{\sigma^2}{n} \left[1 + \frac{(x_p - \bar{x})^2}{s_x^2} \right] \\ &= \sigma^2 \left(1 + \frac{1}{n} \left[1 + \frac{(x_p - \bar{x})^2}{s_x^2} \right] \right) \end{aligned}$$

Schatten en voorspellen

Zowel voor het schatten als voor het voorspellen kunnen betrouwbaarheidsintervallen opgesteld worden.



Correlatie

D.m.v. regressie : de best passende rechte

Vraag : in hoeverre liggen die punten effectief rond een rechte ?

Dit wordt uitgedrukt door de **correlatiecoëfficiënt** r .

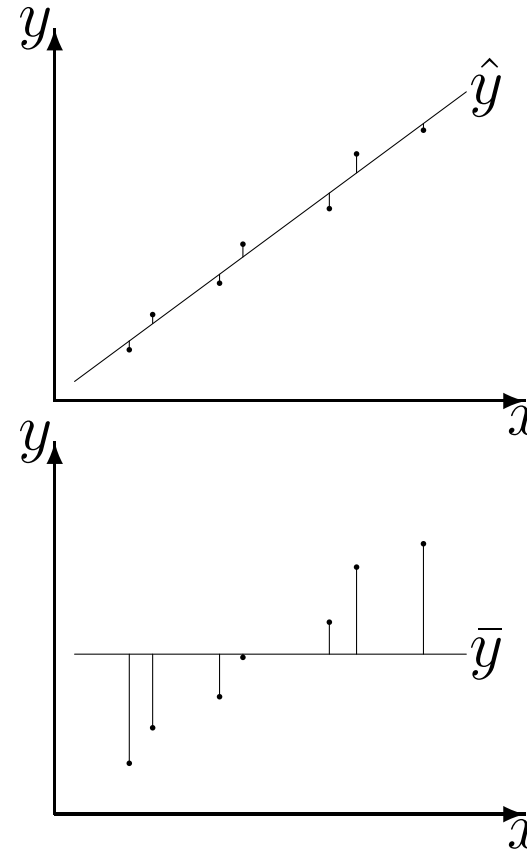
$$r = \frac{\overline{xy} - \bar{x} \bar{y}}{s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = s_{yy} - \frac{s_{xy}^2}{s_{xx}} = s_{yy} \left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2} \right) = s_{yy} (1 - r^2)$$

$$1 - r^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Correlatiecoëfficiënt

$$1 - r^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$



$1 - r^2$ bepaalt het percentage (van de som van de kwadraten van de afwijkingen t.o.v. het gemiddelde) dat niet kan verklaard worden door regressie.

Correlatie

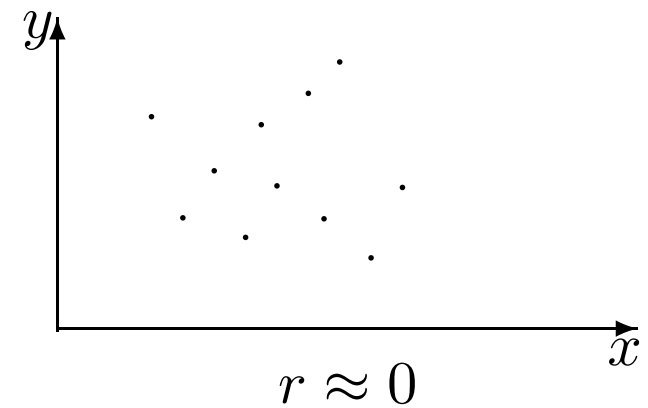
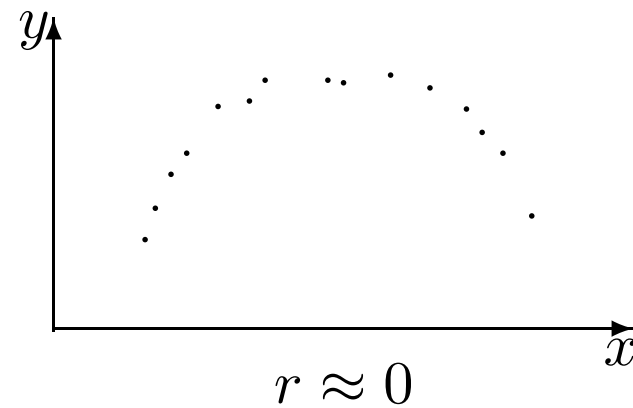
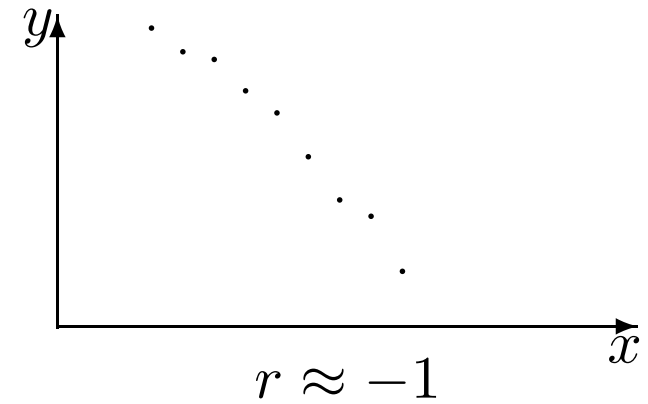
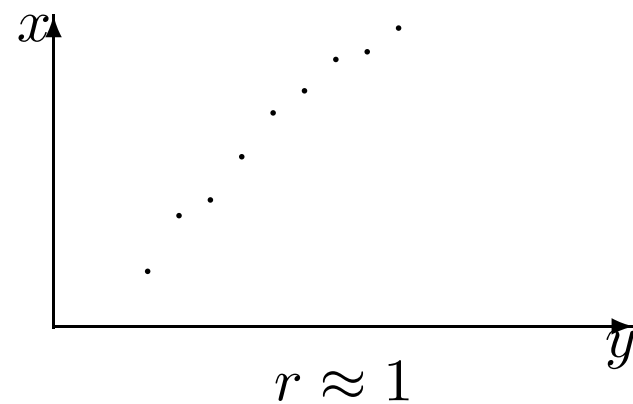
$$r = \frac{\overline{xy} - \bar{x} \bar{y}}{s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

$$1 - r^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \geq 0 \implies -1 \leq r \leq 1$$

Liggen alle punten exact op een rechte, dan is $y_i = \hat{y}_i$,
zodat $1 - r^2 = 0$

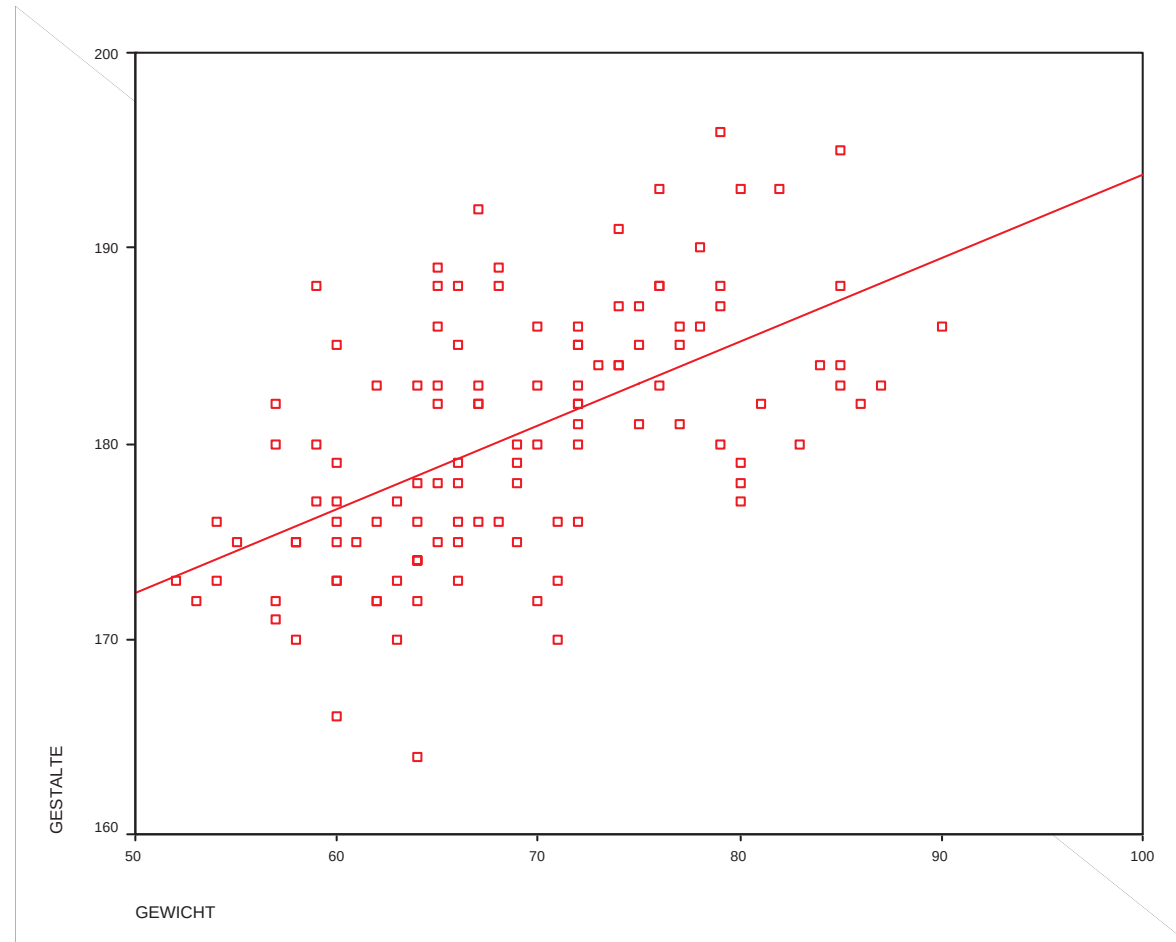
Correlatie

r drukt uit in hoeverre er een **lineair** verband is tussen X en Y



Voorbeeld : voorbeeldsteekproef

X = gewicht en Y = gestalte voor mannen :
regressielijn van Y op X : $y = 151.108 + 0.4261 x$.



$r^2 = 0.323$: slechts 32 % van de deviaties verklaard worden
door regressie

Correlatie

$$r = \frac{s_{xy}}{s_x s_y} : \text{correlatiecoëfficiënt (voor steekproeven)}$$

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} : \text{correlatiecoëfficiënt (voor populaties)}$$

r is de waarde van een veranderlijke R

$$\frac{R \sqrt{n-2}}{\sqrt{1-R^2}} : T(n-2)$$

De test $\rho = 0$ is statistisch volledig equivalent met de test $b = 0$.
niet-parametrische test : op basis van Spearman rang coëfficiënt