Statistiek 2

12. Regressie en correlatie

Week 6

Even voorstellen

Dr. ir. Danny Blom

2014-2017: BSc Technische Wiskunde (TU/e)

2017-2019: MSc Industrial and Applied Mathematics (TU/e)

2019-2023: PhD operations research / toegepaste wiskunde (TU/e)

- Last-mile delivery PostNL
- Optimale zaalbezetting van theaters tijdens pandemie
- Optimaal toewijzen van donoren en nierpatienten

Vanaf 1 juni: universitair docent Statistiek en Operationele Analyse

Regressie

Regressie

Twee kansvariabelen $\underline{X}$ en $\underline{Y}$ waartussen een causaal verband bestaat

- Onafhankelijke variabele $\underline{X}$ en afhankelijke variabele $\underline{Y}$

Voorbeelden:

- Gemiddelde snelheid van een Scania Gryphus ($\underline{X}$) en het brandstofverbruik ($\underline{Y}$)
- Aantal operationele uren van een wapensysteem ($\underline{X}$) en de tijd tot de eerste onderhoudscyclus ($\underline{Y}$)

Focus op (enkelvoudig) lineaire regressie:

"als $\underline{X}$ stijgt, dan stijgt / daalt $\underline{Y}$"

Stappenplan lineaire regressie

Stap 1: Sanity check

- Is er daadwerkelijk een causaal verband tussen de variabelen?
- Teken een spreidingsdiagram. Is er een lineaire trend zichtbaar?

Stap 2: definieer het model $\underline{Y} = \beta_0 + \beta_1 X + \underline{\varepsilon}$

- $\beta_0 + \beta_1 X$: invloed van onafhankelijke variabele X op $\underline{Y}$ $\rightarrow \beta_0, \beta_1$ **<u>onbekend</u>**!
- $\underline{\varepsilon}$: invloed van andere factoren op $\underline{Y}$ (storingsterm)
- Aanname: $E(\underline{\varepsilon}) = 0,\ Var(\underline{\varepsilon}) = \sigma^2$ (**onbekend**!)

Stap 3: gegeven een steekproef, bereken schatters b_0 en b_1 voor β_0 en β_1:

- Regressielijn: $Y = b_0 + b_1 X$
- Wat kun je uit deze regressielijn concluderen?

Lineaire regressie

Defensie onderzoekt het benzineverbruik van de PNOD voertuigen. Hiertoe wordt op een speciaal circuit een tijdlang met constante snelheid gereden, waarna het benzineverbruik wordt vastgesteld in liters per 100 kilometer.

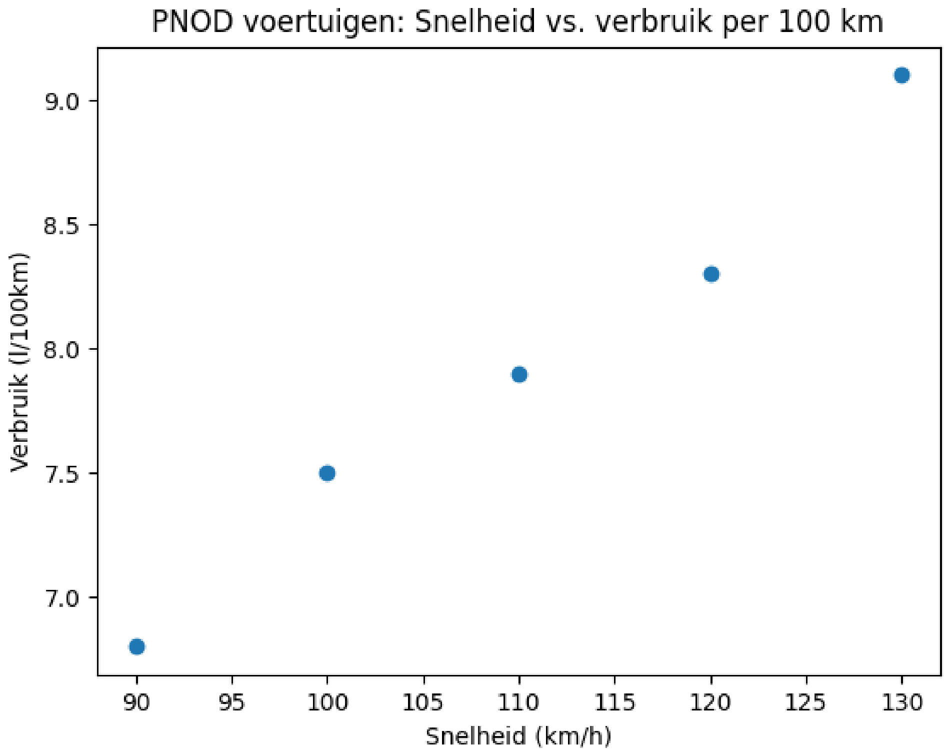
Snelheid (km/h)	90	100	110	120	130
Verbruik (l/100km)	6,8	7,5	7,9	8,3	9,1

Bereken de regressielijn waarbij het verbruik wordt verklaard uit de gekozen snelheid.

Lineaire regressie

Stap 1: sanity check

- Als een auto sneller rijdt, dan stijgt ook het benzineverbruik (check)
- Spreidingsdiagram: er is een duidelijke lineaire trend zichtbaar (check)



Lineaire regressie

Stap 2: We definieren een lineair regressiemodel $\underline{Y} = \beta_0 + \beta_1 X + \underline{\varepsilon}$:

- Onafhankelijke variabele X: snelheid in km/h
- Afhankelijke variabele Y: benzineverbruik in liters per 100 km

De storingsterm $\underline{\varepsilon}$ bevat in dit geval mogelijke andere factoren:

- Weersomstandigheden
- Staat van het wegdek
- Etc.

Lineaire regressie

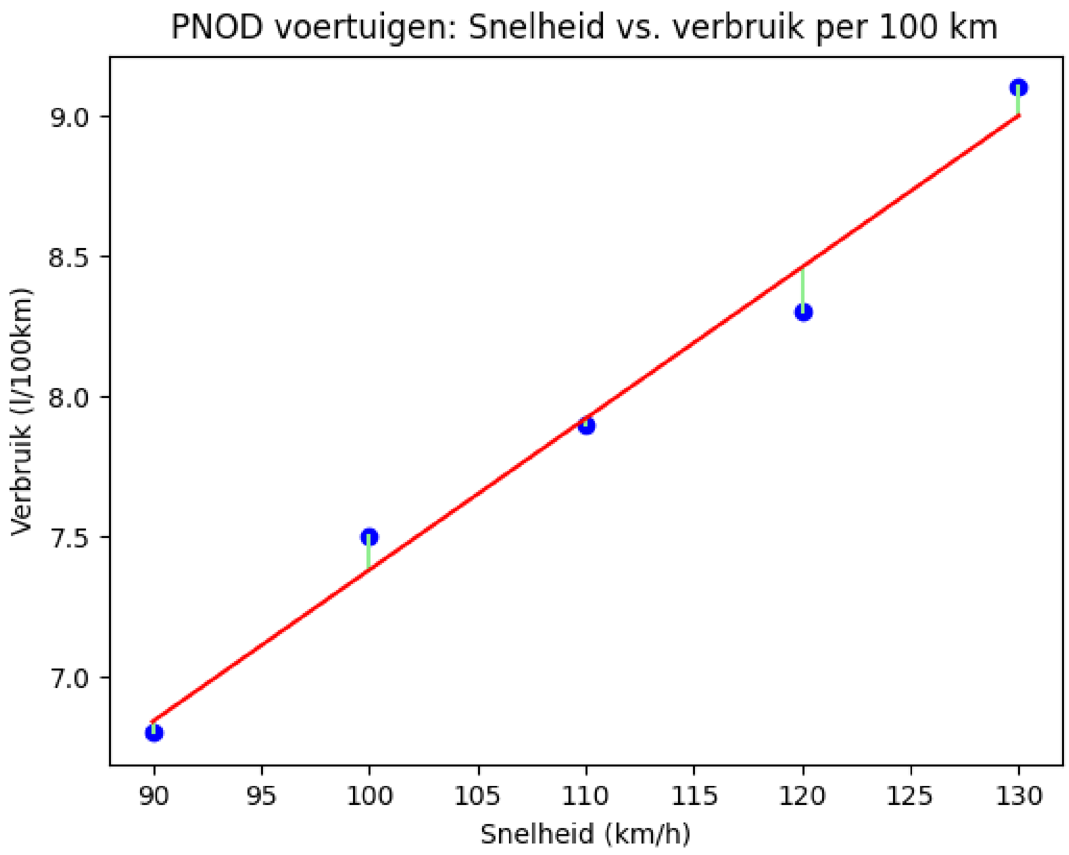
Stap 3: Gegeven een steekproef, bereken schatters b_0 en b_1 voor onbekende parameters β_0 en β_1:

Idee: vind een lijn $Y = b_0 + b_1 X$ die het "best" past bij de steekproef

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

De vraag is nu: hoe definieren we "best passend"?

Kleinste kwadratenmethode



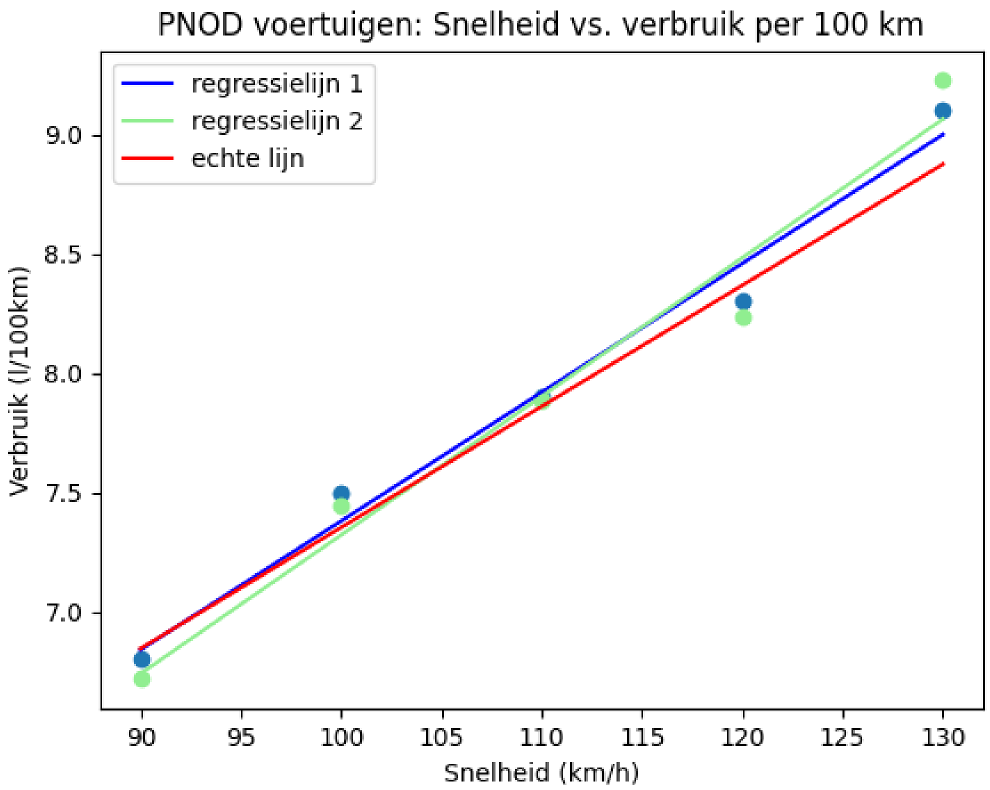
"Bereken de lijn $Y = b_0 + b_1 X$ waarvoor de som van de kwadraten van de lengte van de groene segmenten zo klein mogelijk is"

Algemene formules:

$$b_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2}, \quad b_0 = \overline{Y} - b_1 \overline{X}$$

Twee intervallen: lineaire regressie

Echter: andere steekproef → (net iets) andere regressielijn



Hoe kun je nu de Y-waarde voorspellen bij een nieuw datapunt X_0?

Antwoord 1: regressielijn $Y = b_0 + b_1 X$ → puntschatting $Y_0 = b_0 + b_1 X_0$

Antwoord 2: betrouwbaarheidsintervallen!

Twee intervallen: lineaire regressie

Gegeven is een regressielijn $Y = b_0 + b_1 X$

Vragen:

- Wat is een 95% betrouwbaarheidsinterval voor het gemiddelde verbruik bij een snelheid van 105 km/h? (<u>schattingsprobleem</u>)

- Wat is een 95%-voorspellingsinterval van het individuele verbruik van een auto die 105 km/h rijdt? (<u>voorspellingsprobleem</u>)

Welk van deze intervallen is het grootst?

Schattingsprobleem

Stap 1: Bereken een schatting van $Var(\underline{\varepsilon}) = \sigma^2$:

$$s_{\underline{\varepsilon}}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum Y_i^2 - b_0 \sum Y_i - b_1 \sum X_i Y_i}{n-2}$$

Stap 2: Bereken een schatting van de variantie van $\mu = E(\underline{Y}|X_0)$:

$$s_\mu = s_{\underline{\varepsilon}} \sqrt{\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum \left(X_i - \overline{X}\right)^2}}$$

Stap 3: Bereken het $100(1-\alpha)\%$ $-$betrouwbaarheidsinterval:

$$\left[Y_0 - t_{\frac{\alpha}{2}}[n-2] * s_\mu, Y_0 + t_{\frac{\alpha}{2}}[n-2] * s_\mu\right]$$

Voorspellingsprobleem

Stap 1: Bereken een schatting van $Var(\underline{\varepsilon}) = \sigma^2$:

$$s_{\underline{\varepsilon}}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum Y_i^2 - b_0 \sum Y_i - b_1 \sum X_i Y_i}{n-2}$$

Stap 2: Bereken een schatting van de variantie van $Y \mid X_0$ (forecast):

$$s_f = s_{\underline{\varepsilon}} \sqrt{1 + \frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum\left(X_i - \overline{X}\right)^2}}$$

Stap 3: Bereken het $100(1-\alpha)\%$ −betrouwbaarheidsinterval :

$$\left[Y_0 - t_{\frac{\alpha}{2}}[n-2] * s_f, Y_0 + t_{\frac{\alpha}{2}}[n-2] * s_f\right]$$

Twee intervallen: lineaire regressie

Gegeven is een regressielijn $Y = b_0 + b_1 X$

Vragen:

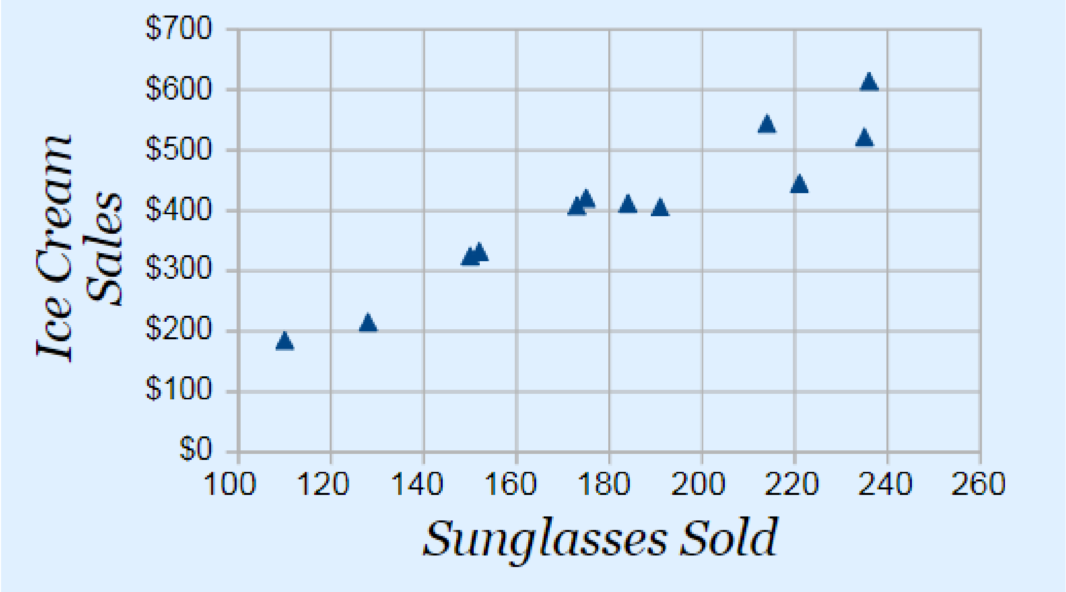
- Wat is een 95% betrouwbaarheidsinterval voor het gemiddelde verbruik bij een snelheid van 105 km/h? (schattingsprobleem)
- Wat is een 95%-voorspellingsinterval van het individuele verbruik van een auto die 105 km/h rijdt? (voorspellingsprobleem)

Welk van deze intervallen is het grootst?

Correlatie

Correlatie

Verbanden tussen twee kansvariabelen zijn niet altijd causaal!!!



Derde variabele probleem:

- Er bestaat een andere variabele (bv. buitentemperatuur / aantal zonuren per dag) dat invloed heeft op beide kansvariabelen

Pearson's correlatiecoëfficiënt

Gegeven datapunten $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, is Pearson's correlatiecoëfficiënt gedefinieerd als:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum\left(X_i - \overline{X}\right)^2\left(Y_i - \overline{Y}\right)^2}} = \frac{\overline{XY} - \overline{X} * \overline{Y}}{\sqrt{\left(\overline{X^2} - \bar{X}^2\right)\left(\overline{Y^2} - \bar{Y}^2\right)}}$$

Dit is een getal tussen -1 en 1!

We moeten dus eerst de grootheden $\overline{X}, \overline{Y}, \overline{X^2}, \overline{Y^2}$ en $\overline{XY}$ berekenen!

Stappenplan correlatie

Stap 1: sanity check

- Teken een spreidingsdiagram. Is er een bepaalde trend zichtbaar?

Stap 2: bereken Pearson's correlatiecoëfficiënt r

Stap 3: conclusie over correlatie met behulp van spreidingsdiagram en r

Correlatie

Defensie onderzoekt het benzineverbruik van de PNOD voertuigen. Hiertoe wordt op een speciaal circuit een tijdlang met constante snelheid gereden, waarna het benzineverbruik wordt vastgesteld in liters per 100 kilometer.

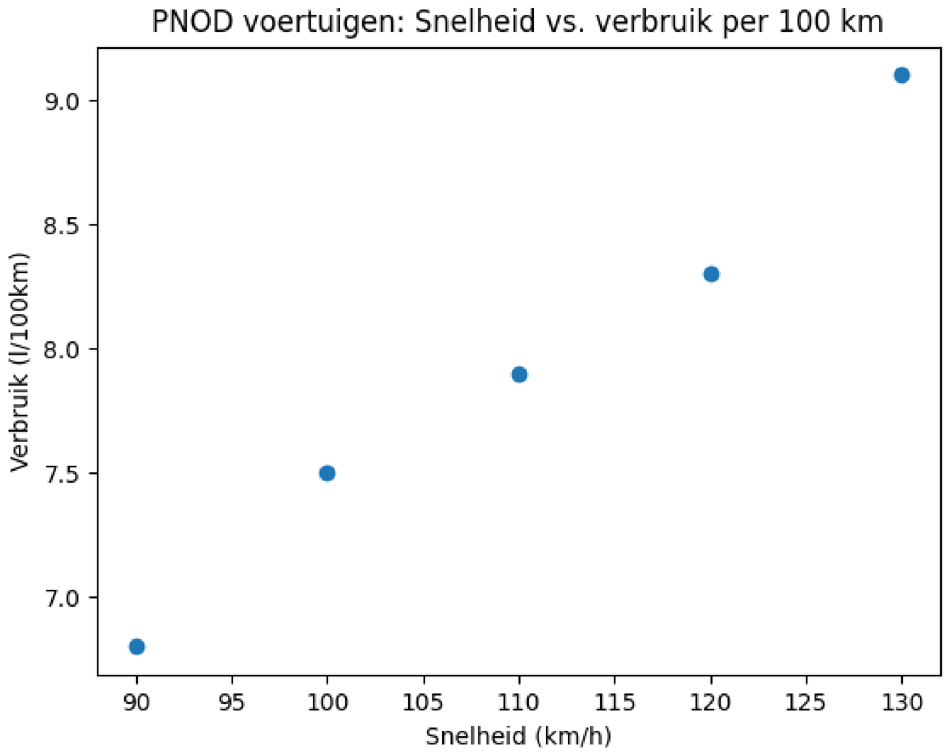
Snelheid (km/h)	90	100	110	120	130
Verbruik (l/100km)	6,8	7,5	7,9	8,3	9,1

Wat kunnen we op statistische verantwoorde wijze zeggen over de correlatie tussen snelheid en verbruik?

Correlatie

Stap 1: sanity check

- Spreidingsdiagram: er is een duidelijke lineaire trend zichtbaar (check)



Stappenplan: correlatie

Stap 2: We berekenen eerst $\overline{X}, \overline{Y}, \overline{X^2}, \overline{Y^2}$ en $\overline{XY}$

Verbruik (Y)	Snelheid (X)	XY	X^2	Y^2
6,8	90	612	8100	46,24
7,5	100	750	10000	56,25
7,9	110	869	12100	62,41
8,3	120	996	14400	68,89
9,1	130	1183	16900	82,81
$\boldsymbol{\overline{Y} = 7,92}$	$\boldsymbol{\overline{X} = 110}$	$\boldsymbol{\overline{XY} = 882}$	$\boldsymbol{\overline{X^2} = 12300}$	$\boldsymbol{\overline{Y^2} = 63,32}$

Vul in in de formule:

$$r = \frac{\overline{XY} - \overline{X} * \overline{Y}}{\sqrt{(\overline{X^2} - \overline{X}^2)(\overline{Y^2} - \overline{Y}^2)}} = \frac{882 - 110 * 7{,}92}{\sqrt{(12300 - 110^2)(63{,}32 - 7{,}92^2)}} \approx 0{,}9912$$

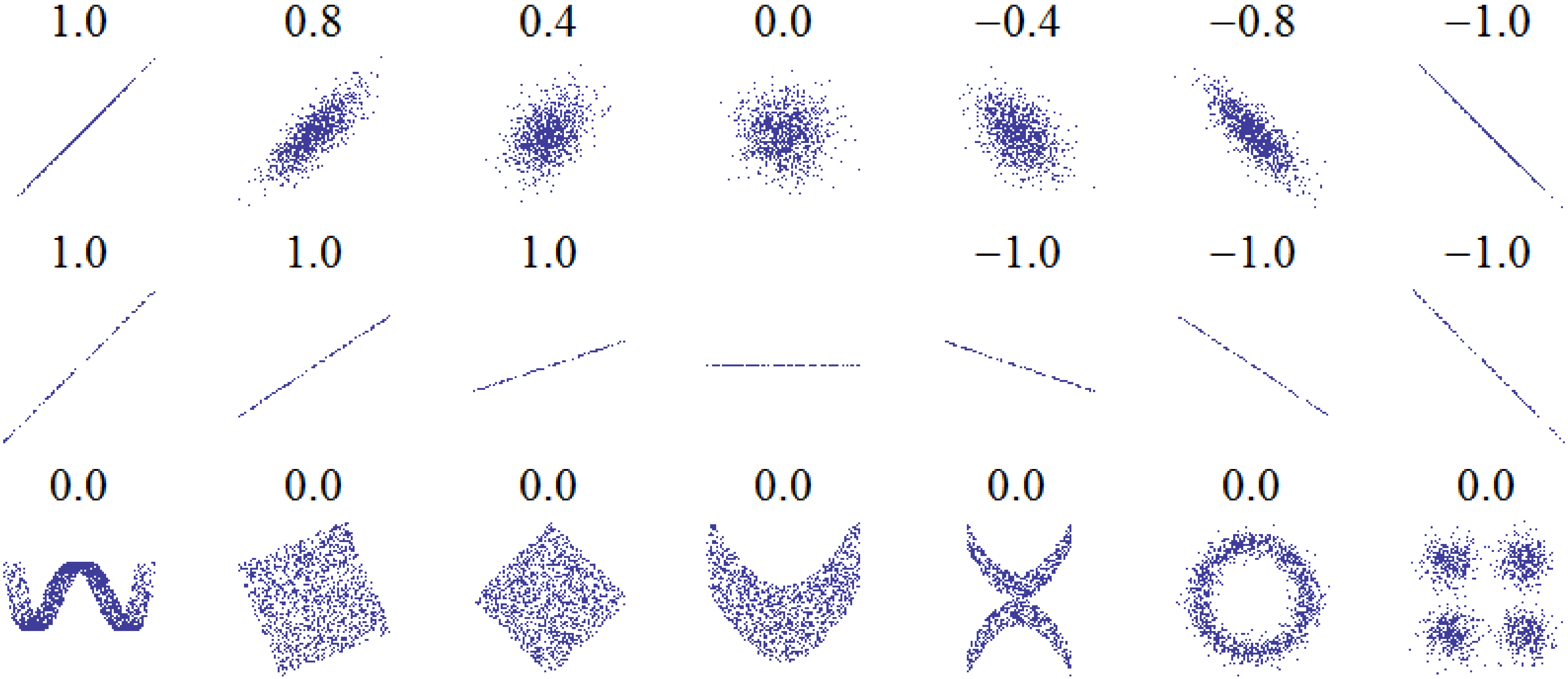
Stappenplan: correlatie

Stap 3: Waardes van r kunnen ruwweg in deze klassen worden opgedeeld.

$\|r\|$	r^2 (afgerond)	Interpretatie verband
< 0,3	< 0,1	zeer zwak
0,3 - 0,5	0,1 - 0,25	zwak
0,5 - 0,7	0,25 - 0,5	matig
0,7 - 0,85	0,5 - 0,75	sterk
0,85 - 0,95	0,75 - 0,9	zeer sterk
> 0,95	> 0,9	uitzonderlijk sterk (suspect!)

De gevonden correlatie is uitzonderlijk sterk! $(r \approx 0{,}9912)$

Correlatie en samenhang



1. Van sterk positieve naar sterk negatieve correlatie.
2. Sterke correlatie heeft niets met de richtingscoëfficiënt te maken.
3. $r = 0$ impliceert niet onafhankelijkheid.

Samenvatting

Regressie	Correlatie
• onafhankelijke variabele X • afhankelijke variabele Y • oorzaak → gevolg • Model: $\underline{Y} = \beta_0 + \beta_1 X + \underline{\varepsilon}$ • Regressielijn $Y = a + bX$ • Betrouwbaarheidsintervallen	• Niet noodzakelijk causaal verband • Lineaire samenhang • Pearson's r: lineaire samenhang

Vragen?

Veel succes bij het tentamen!