Regressie-analyse

Cursus Bachelor Project 2 B&O

College 2

Harry B.G. Ganzeboom

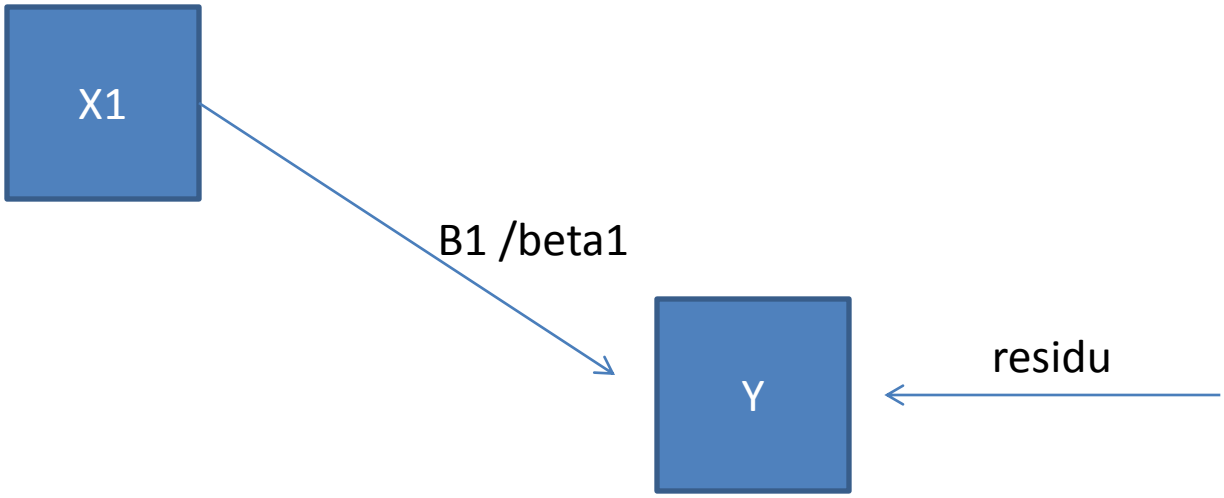
Agenda

- Lineaire regressie-model (herhaling)
  - Enkelvoudig (simple)
  - Meervoudig (multiple)
  - Regressie met dummy variabelen / lineariteit.
- Valkuilen (regression diagnostics)
- Relatie regressie- en factoranalyse
- Invloed van onbetrouwbaarheid

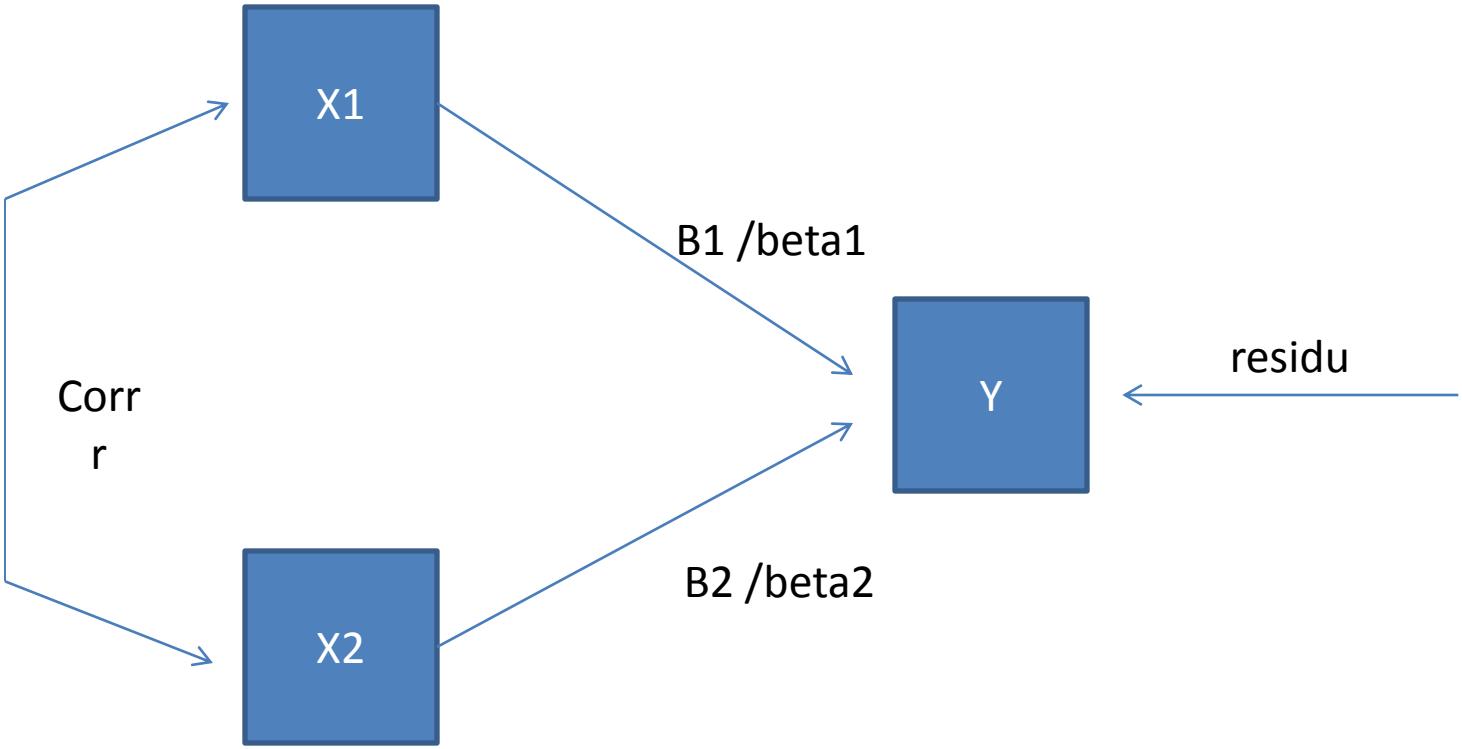
Pallant

- Ch. 13 is een mooie en korte inleiding in regressie-analyse. De collegestof geeft diepere achtergronden.
- In deze cursus doen we niet veel met Regression Diagnostics (164-166). Lees ze ***voorzichtig*** door. Je moet deze begrippen in het kopje van 164 wel kunnen uitleggen! Lees de rest ***zeer grondig***.
- Merk op verschil in terminologie:
  - Hierarchical multiple regression → stepwise (blockwise) forward.
  - Standard regression → Multiple regression.

Model: simple regression



Model: multiple regression



Korte demonstratie in SPSS

```
** MULTIPLE REGRESSIE STAPSGEWIJS **.

regr /dep=IDENT
  /enter=RESPECT
  /enter=PRIDE.

regr /stat=def change tol /dep=IDENT
  /enter=RESPECT /enter=PRIDE.
```

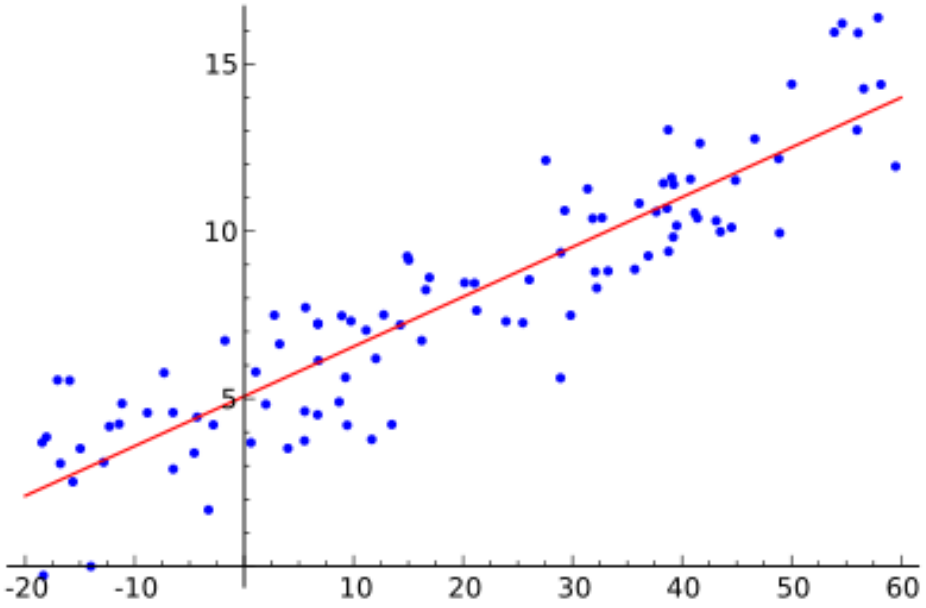
Het multipele regressiemodel

- Het meest gangbare model voor samenhangen tussen twee of meer variabelen is het meervoudige (multipele) lineaire regressiemodel:
  - $Y = B0 + B1*X1 + B2*X2 + .. + Bk*Xk + e$
- In woorden: hoe meer X1, X2 .. Xp, des te meer Y.
- Ook: als X1, X2, dan meer Y (discrete variant).

Het enkelvoudige regressiemodel

- In de meest eenvoudige variant wordt slechts één X met de Y in verband gebracht:
  - Y = B0 + B1*X1 + residu
    - Y Afhankelijke variabele
    - X1 Onafhankelijke variabele
    - B0 Intercept, constante
    - B1 Slope, hellingshoek, effect
    - e Error, residu.
- Het enkelvoudige (simple) regressiemodel heeft niet veel directe toepassing, maar als je dit eenmaal begrijpt, is meervoudige regressie een eitje.

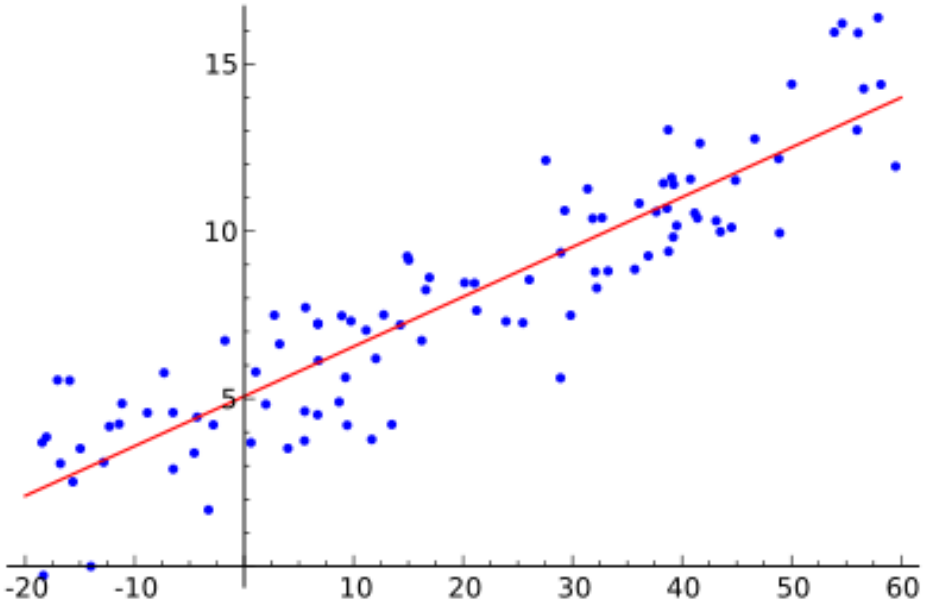
$$Y = B0 + B1.X$$



B0 = intercept: de waarde van Y als X 0 is

B1 = slope: de helling van de lijn. Dus de hoeveelheid Y die erbij komt als X één eenheid omhoog gaat

- Wat is het intercept? En wat is de slope?



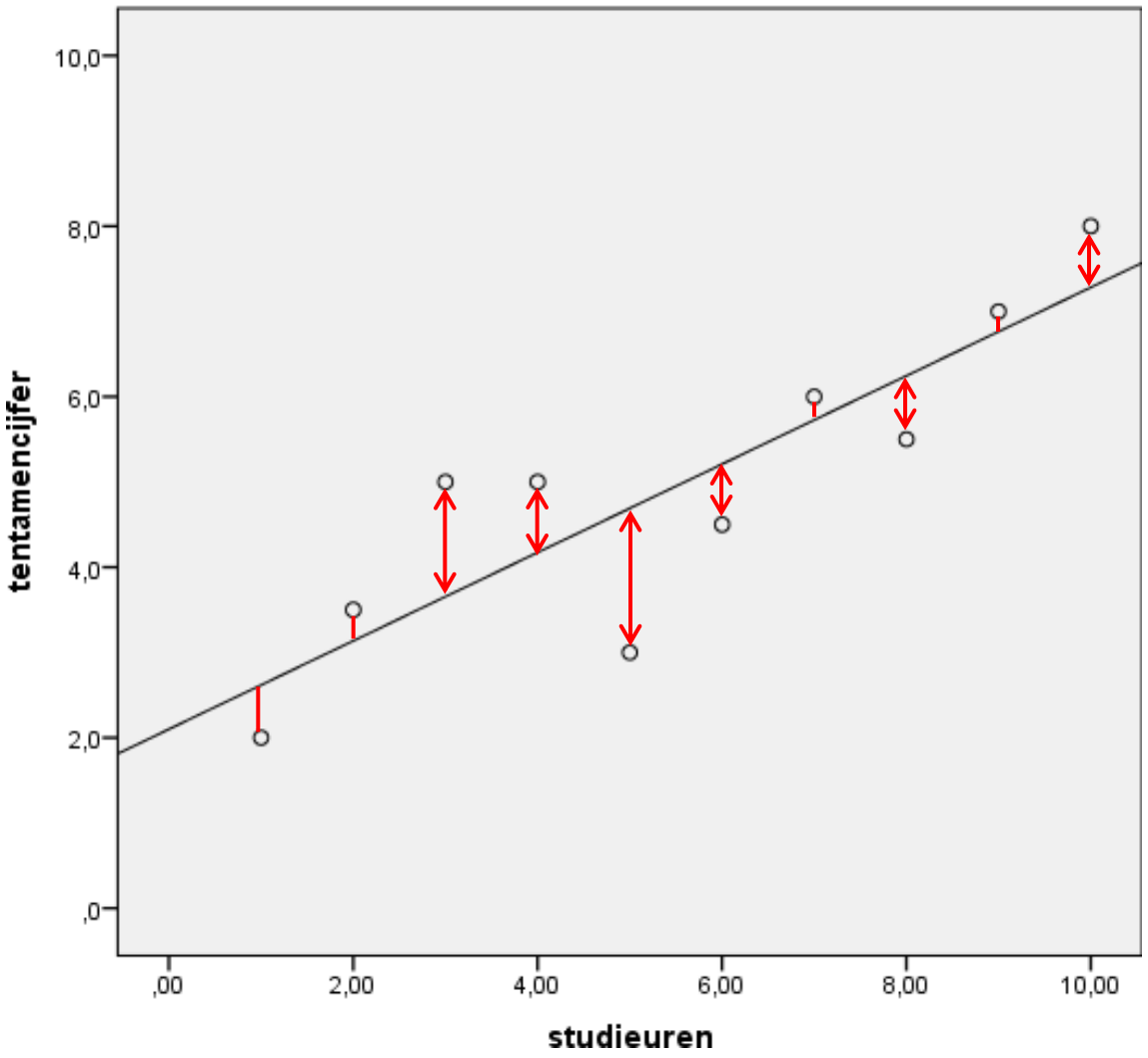
Intercept: bij X = 0, Y = 5. Het intercept is dus 5

Slope: bij X = 35 stijgt Y met 5 (van 5 naar 10). 5/35 is 0.14. De slope is dus 0.14

Kleinste Kwadraten, Least Squares

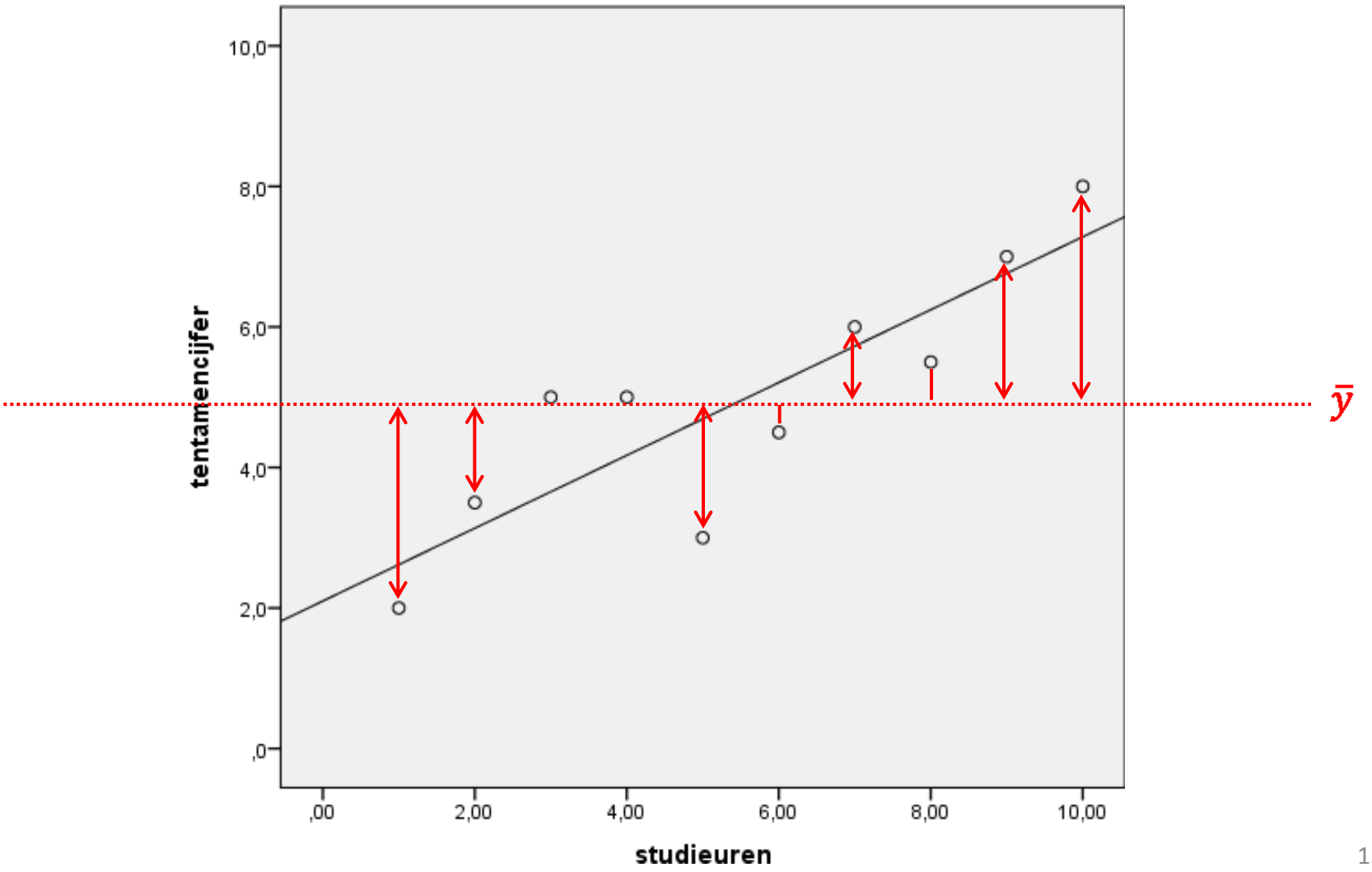
- De coefficienten van het model (B0 en B1) worden berekend door minimalisering van de kwadratensom:
- Y – Y^ = residu
- OLS: Sum(Y-Y^)**2 → minimaal.
- Voor deze minimalisering bestaat maar één oplossing en deze kan zonder iteraties worden berekend. Dat doet SPSS voor je.

$$residuals\ SS = \sum (residual)^2 = \sum (y - \hat{y})^2$$

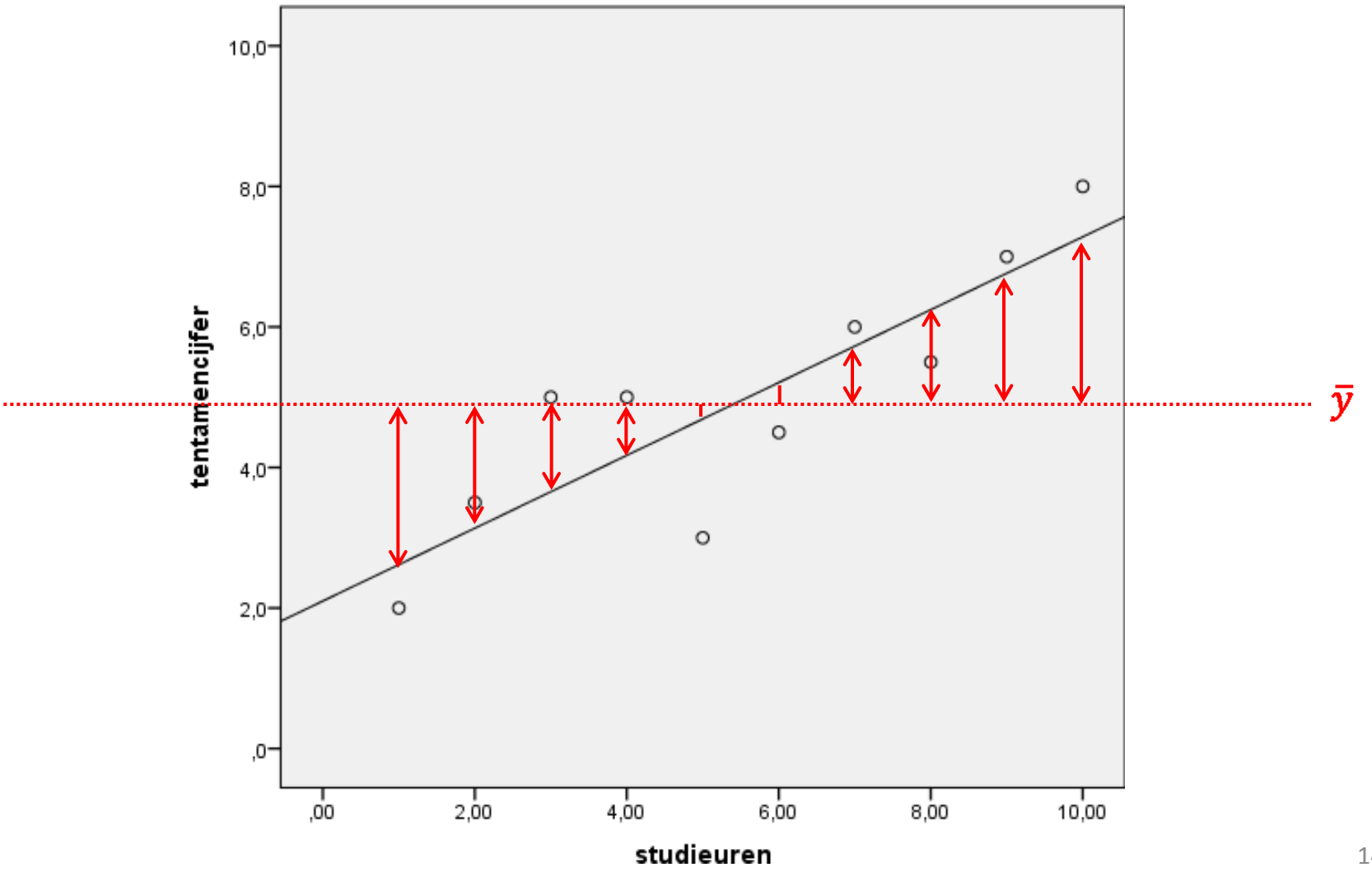


$$\hat{y} = 2.1 + 0.518x$$

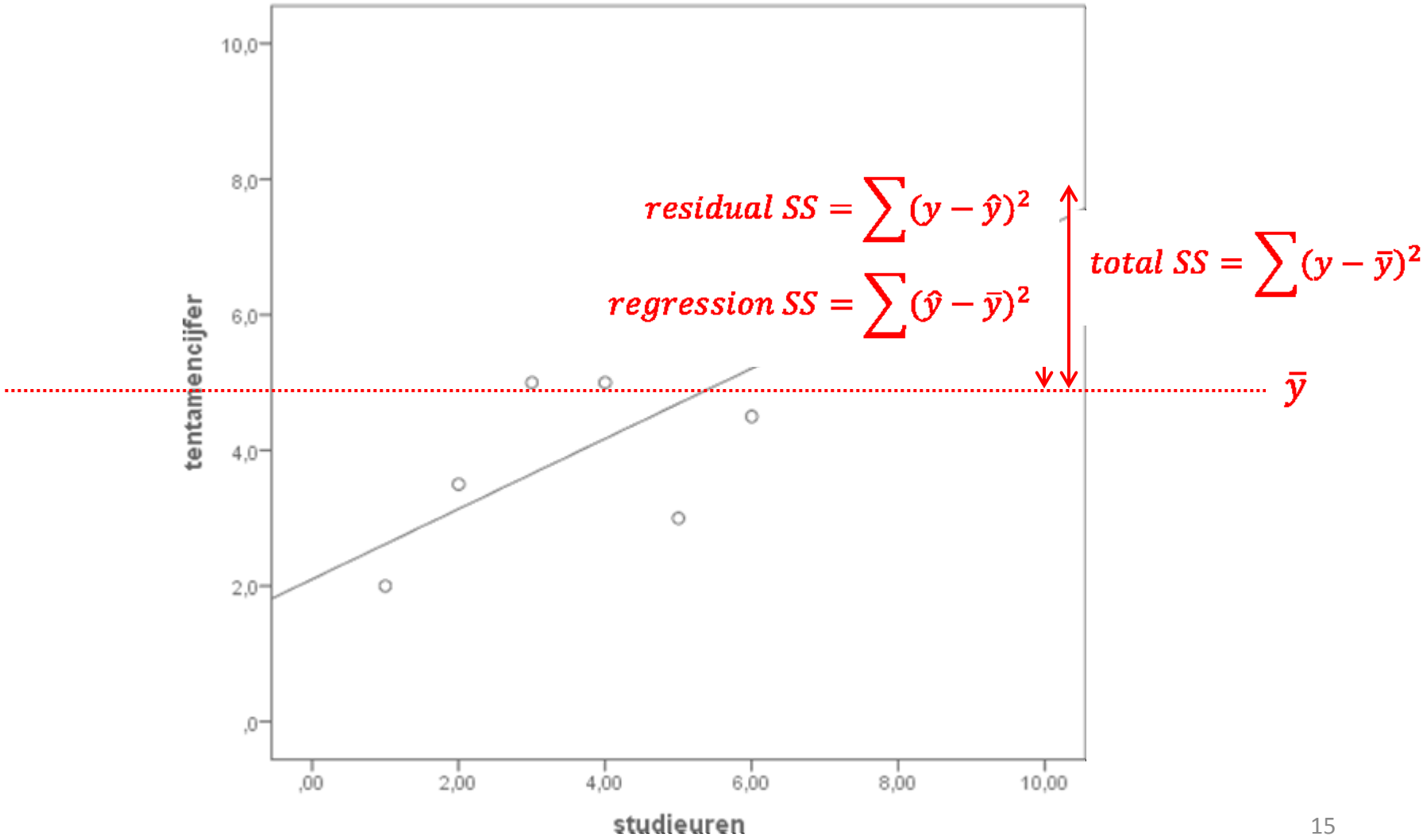
$$total\ SS = \sum (deviatie)^2 = \sum (y - \bar{y})^2$$



$$regression\ SS = \sum (\hat{y} - \bar{y})^2$$



$$r^2 = \frac{total\ SS - residual\ SS}{total\ SS} = \frac{regression\ SS}{total\ SS}$$



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,863[a]	,745	,713	,97293

a. Predictors: (Constant), studieuren

$\sum(y-\hat{y})^2$ $\sum(\hat{y}-\bar{y})^2$

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22,152	1	22,152	23,402	,001[b]
	Residual	7,573	8	,947		
	Total	29,725	9			

a. Dependent Variable: tentamencijfer

b. Predictors: (Constant), studieuren

$\sum(y-\bar{y})^2$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,100	,665		3,160	,013
	studieuren	,518	,107	,863	4,838	,001

a. Dependent Variable: tentamencijfer

Gestandaardiseerde regressie

- In gestandaardiseerde regressie worden Y en X allebei gestandaardiseerd tot Z-variabelen

  $ZX = (X - M(X)) / SD(X)$

  $ZY = (Y - M(Y)) / SD(Y)$
- $ZY = 0 + beta*ZX + residu$
- $-1 < beta < 1$
- Het voordeel van beta boven B is dat je onmiddellijk de sterkte van een invloed ziet.
- Bij enkelvoudige regressie is beta gelijk aan de correlatie.
- Het blijft een regressiecoefficient: beta is een slope die aangeeft hoeveel SD(Y) je krijgt voor 1 SD(X).
- De intercept van een gestandaardiseerde regressie is altijd 0.

Demonstratie in SPSS

```
** gestandaardiseerd **.

desc respect pride ident /save.

desc Zrespect Zpride Zident.

regr /dep=Zident /enter=Zpride
Zrespect
```

Hoe te lezen? Van onder naar boven..

- Het model:
  - Ongestandaardiseerde regressie-coefficienten B (als je de eenheden van X en Y inhoudelijk kunt interpreteren)
  - Gestandaardiseerde regressie-coefficienten BETA (als de eenheden feitelijk betekenisloos zijn)
- Significantie:
  - Standard Error (SE): geschatte steekproeffluctuatie
  - T-waarde = B/SE: hoeveel SE ligt de B van 0 af! 0 is de bij de nulhypothese veranderstelde waarde.
  - Sig.: Kans dat je deze B zou verkrijgen als H0 waar is.
  - Confidence interval: 95% kans dat de populatiecoefficient in dit interval valt.

ANOVA en R2

- De ANOVA-tabel geeft de opdeling van de kwadraatsommen SS:
  - SS-Total
  - SS-Regression
  - SS-Residual
- SS-total = SS-regression + SS-residual
- Verklaarde variantie:
  - R2 = SS-regression / SS-Total
  - R = √R2
  - Adj. R2: R2 rekening houdend met aantal X-variabelen.
- Vrijheidsgraden DF:
  - DF-total = N-1
  - DF-regression = aantal B-coeffienten
  - DF-residu = DF-total – DF-regression
- Mean squares MS = SS / ndf.
- F-test: MS-regression / MS-residual. (DF1, DF2). SPSS zoekt de significantie voor je op.
- De F-test toetst de significantie van de verandering in R2. Dit is met name interessant bij stapsgewijze regressie, waarin je meerdere modellen vergelijkt

Multipele regressie

- Bij multipele regressie komen er gewoon meer X-variabelen en evenveel B-coefficienten bij.
- De coëfficiënten B staan nu voor ***partiële*** effecten.
- Dit betekent dat je B1 nu het effect van X1 op Y is terwijl X2 constant wordt gehouden. Evenzo is B2 het effect van X2 op Y als X1 constant wordt gehouden.
- Aan deze partiële interpretatie dankt het regressie-model zijn kracht en bruikbaarheid.

Hoe houdt multiple regressie constant? (1)

- De interpretatie van multipele regressiecoefficienten is ***partieel***: B1 geeft de invloed van X1 op Y, terwijl X2 constant wordt gehouden; B2 geeft de invloed van X2 op Y terwijl X1 constant wordt gehouden.
- Regressie-analyse is daarmee geschikt om de invloed van met elkaar correleren X-vars op Y te bestuderen.
- Maar hoe doet regressie dat eigenlijk?
- Constant houden van X1 betekent eigenlijk: terwijl we binnen vaste waarden van X2 kijken.

Hoe houdt regressie-analyse constant? (2)

- Multipele regressie werkt eigenlijk als volgt:
- X2 = BB0 + BB1*X1 → ResX1 = X2 - BB0 + BB1*X1
- X1 = BB0 + BB2*X2 → ResX2 = X1 - BB0 + BB1*X2
- ResX1 en ResX2 zijn de residuen van X1 (resp. X2) uit de voorspelling van X2 (resp. X1). ***Het zijn de stukjes van X1 (resp. X2) die NIET beinvloed worden door X2 (resp. X1).***
- Regr /dep=Y /enter=ResX1.
- Regr /dep=Y /enter=ResX2.
- Dit geeft dezelfde coefficienten als: Regr /dep=Y /enter=X1 X2.
- Illustratie in SPSS.

Demonstratie in SPSS

```
** Hoe houdt regressie analyse constant? ** .

regr /dep=pride /enter=respect
/save=residual(ResPride).
regr /dep=respect /enter=pride
/save=residual(ResRespect).

regr /dep=Ident /enter=ResPride.
regr /dep=Ident /enter=ResRespect.

regr /dep=ident /enter=pride Respect.
```

Valkuil 1: Intercept

- Interpretatie van het intercept B0. Dit is de ***verwachte waarde van Y als alle X-variabelen 0 zijn***.
- Dit is alleen maar een zinvolle grootheid als 0 inderdaad een geldige waarde van X is.
- Het is goede praktijk om de X-variabelen zo te coderen dat je weet wat 0 is.
- Bv: codeer de variabele ‘Sekse’ als (0) man en (1) vrouw. Leeftijd: trek gemiddelde leeftijd af (centreren).

Valkuil 2: (multi)collineariteit

- Collineariteit is de mate waarin de X-variabelen met elkaar samenhangen:
  - Collineariteit: correlatie tussen twee X-variabelen
  - ***Multi***-collineariteit: voorspelbaarheid van één X uit alle andere X-variabelen.
- Multi-collineariteit kun je niet (direct) zien aan de correlaties tussen de X-vars!!
- Als de X-variabelen sterk met elkaar samenhangen worden de schattingen van B instabiel:
  - SE worden heel groot
  - Verklaarde variantie R2 neemt juist niet toe

Valkuil 2: multi-collineariteit

- SPSS geeft diagnostische toetsen of de collineariteit niet te sterk is:
  - TOL (tolerance): onverklaarde variantie van een X uit andere X-en.
  - VIF: 1/Tolerance.
- Advies: VIF < 10. TOL > .10.
- Problemen van multi-collineariteit zijn echter afhankelijk van N: bij grote N krijg je wel stabiele schattingen, ook al is er sterke samenhang tussen X-vars.
- Merk op dat multipele regressie-analyse juist bedoeld is om collineaire situaties te analyseren: als de X-vars niet zouden correleren, kun je net zo goed enkelvoudige regressie doen!

Valkuil 3: perfecte collineariteit (singulariteit)

- Als X-variabelen perfect uit elkaar voorspelbaar zijn, kun je de B-coefficienten niet schatten.
- Ook niet als je heel veel data hebt!
- Voorbeelden:
  - Leeftijd en geboortejaar
  - Mannen versus vrouwen (dummy-variabelen): je vergelijkt dan met een referentie (=weggelaten) categorie!

Valkuil 4: Lineariteitsaanname

- Regressie-modellen zouden beter “lineaire” modellen kunnen heten: relaties worden samengevat via een rechte lijn: $y = a + b.x$.
- Maar lang niet alle relaties zijn lineair. Denk aan:
  - Historisch trends met breuken en golven
  - Verschillen tussen leeftijdsgroepen

Lineariteitstest

- **Means** levert een test of linearity voor het geval je X categorisch (gemaakt) is (AgeCat).
- Je kunt deze test of linearity ook zelf maken via dummy variabelen.
- De test bij means is een nuttig en snel begin, maar als je het via dummy variabelen zelf doet, heb je meer mogelijkheden.

Dummy-variabelen

- Een **set** dummyvariabelen wordt afgeleid van een categorische variabelen:
  - **Recode Educat (1=1)(else=0) into EduCat1.**
  - **Recode Educat (2=1)(else=0) into EduCat2.**
  - **Recode Educat (3=1)(else=0) into EduCat3.**
  - **Recode Educat (4=1)(else=0) into EduCat4.**
  - **Recode Educat (5=1)(else=0) into EduCat5.**
- Dit levert vijf 0/1 variabelen op die we in een multiple regressie als voorspeller gebruiken ipv de oorspronkelijke EduCat**.**
- **Y = B0 + B1*Educat1 + B2*Educat2 + B3*Educat3 + B4*Educat4 + B5*Educat5**

Demonstratie in SPSS

```
regr /dep=pinc /enter=educat1 educat2
educat3 educat4 educat5.

regr /dep=pinc /enter=educat2 educat3
educat4 educat5.
```

Regressie met dummy-variabelen

- Dummy-variabelen zijn perfect collineair!
- Ze moeten dat ook zijn!!
- Je kunt de regressievergelijking goed interpreteren als je één dummy-var weglaat.
- Het is heel belangrijk te weten welke deze ***referentie*** is.
- SPSS kiest meestal voor de dummy-variabele die hoort bij de categorie met de grootste N.
- Soms is het handiger om een andere te kiezen.
- De effecten van de dummy-variabelen moet je interpreteren als een verschil met de intercept – die de verwachte waarde voor de referentie voorstelt.

Valkuil 5: Onbetrouwbaarheid van de X-variabelen (niet behandeld)

- Het regressiemodel veronderstelt dat je X-variabelen volledig betrouwbaarheid gemeten zijn.
- Als er toch onbetrouwbaarheid in de onafhankelijke variabelen is overgebleven, verzwakt dat de B- en Beta-coefficienten.
- De gevolgen kunnen met name dramatisch zijn, wanneer je onderzoeksvraag gaat over de relatieve sterkte van B- of Beta-coefficienten.

Onbetrouwbaarheid van Y

- Als je je Y met onbetrouwbaarheid gemeten is:
  - Vermindert dat de verklaarde variantie
  - Verzwakt dat de Beta-coefficienten
  - Vergroot dat de SE,
- ***Maar blijven de B-coefficienten (ongeveer) hetzelfde!***

Demonstratie in SPSS

```
** EFFECT VAN ONBETROUWBAARHEID IN **.

set seed 180453.
compute loting1=uniform(1).
compute loting2=uniform(1).

sort cases by loting1.
compute xPride=Pride.
if (loting2 > .80) xPride=lag(Pride).

regr /dep=ident /enter=xPride Respect.
```

Nog wat meer valkuilen (Pallant) (niet behandeld)

- Pallant bediscussieert:
  - N of cases (??)
  - Outliers
  - Normaliteit (van residuen)
  - Homoskedasticiteit
  - Onafhankelijkheid van waarnemingen
- Het zijn allemaal belangrijke problemen, maar er is een belangrijke kanttekening bij: vertekeningen treden met name op in de schatting van de SE's, niet zozeer bij de B's. Het doet er allemaal veel toe, bij de significantietesten.

Nogmaals: gestandaardiseerd of ongestandaardiseerd?

- Gestandaardiseerde modellen zijn een herberekening van het ongestandaardiseerde model.
- Ongestandaardiseerde coefficienten van twee X-vars kun je alleen vergelijken als de meeteenheid dezelfde is, bv. tussen stapsgewijze modellen, of tussen groepen (mannen en vrouwen).
- Gestandaardiseerde coëfficiënten kun (ook) vergelijken tussen variabelen: je kunt er de relatieve sterkte van effecten mee bepalen.
- Gestandaardiseerde effecten zijn niet gemakkelijk te interpreteren bij 0/1 X-variabelen – zoals dummy variabelen.

Regressie- en Factor-analyse

- Factoranalyse is weinig anders dan een verzameling van meerdere multipele regressiemodellen, waarbij de onafhankelijke variabelen ongemeten (latent) zijn.
- In de praktijk gebruiken we hier bijna altijd gestandaardiseerde coefficienten; de intercepten zijn dan altijd 0 en blijven onvermeld.

Factor- en regressieanalyse

- Y1 = B11*F1 + B12*F2 + residu
- Y2 = B21*F1 + B22*F2 + residu
- Y3 = B31*F1 + B32*F2 + residu
- Y4 = B41*F1 + B42*F2 + residu

Met simple structure:

- Y1 = B11*F1 + 0*F2 + residu
- Y2 = B21*F1 + B22*F2 + residu
- Y3 = 0*F1 + B32*F2 + residu
- Y4 = 0*F1 + B42*F2 + residu

Regressie en causaliteit

- Regressie-coefficienten worden vaak binnen een causaal (oorzaak-gevolg) model geinterpreteerd.
- Causale interpretatie is alleen geldig als aan twee voorwaarden voldaan is:
  - *No reversed causation* = causale volgorde: X kan oorzaak van Y zijn = X gaat vooraf aan Y (en niet andersom)
  - *No confounding:* X en Y worden niet beinvloed door een achterliggende variabele Z.
- Causale volgorde volgt uit je design of theoretische overwegingen; confounders moet je meten en via regressie constant houden.

Waarom heet ‘regressie’ zo?

- Regressiemodel is een niet zo duidelijke naam. Beter zou zijn “lineaire modellen” of “least-squares” modellen.
- De term “regressie” is afkomstig uit de biologie, waarin het model oorspronkelijke ontwikkeld is om de samenhang in lengte tussen ouders en kinderen te modelleren.
- Kinderen worden niet even lang als hun ouders: lange ouders hebben kinderen die gemiddeld korter zijn dan zijzelf, korte ouders hebben kinderen die gemiddeld langer zijn dan zijzelf.
- De ‘regressie’-coeffecient meet dan de mate waarin dit verband niet perfect (1.0) is, i.e. de lengte van kinderen ‘regresseert naar het gemiddelde’.

Stappenplan regressie-analyse

- Stap 1: Check variabelen: meetniveau, missing values? Codeer categorische variabelen als 0/1 (dummy) variabelen. Zorg voor 0-punt bij andere variabelen duidelijk is (centreren).
- Stap 2: Causale volgorde: wat is X en wat is Y?
- Stap 3: Bereken de regressiemodellen, bij voorkeur stapsgewijs (regress /enter=).
- Stap 4: Interpreteer eerst het ongestandaardiseerde en gestandaardiseerde regressie-model, daarna de significantie.
- Stap 5: Rapporteer B en SE, en/of beta/t en R2.

Mediatie en moderatie

- Volgende week mediatie-analyse, derde week moderatie analyse:
- Mediatie: verloopt de invloed van X op Y via M?
- Moderatie: hangt de invloed van X op Y af Z?
- Baron & Kenny beginnen met Moderatie, wij doen het andersom.
- Over geen van beide geeft Pallant directe informatie. Beide maken gebruik van het multipele regressie-model zoals besproken in Pallant, CH 13.