Les 6 Regressie en correlatie

Als we na twee kenmerken van elementen van een populatie kijken, is het een voor de hand liggende vraag of we aan de hand van de waarde van het eerste kenmerk een voorspelling kunnen doen voor de waarde van het tweede kenmerk. Bijvoorbeeld kunnen we ons afvragen of de prijs van een auto een indicatie geeft voor zijn levensduur, of de lengte van de wijsvinger iets zegt over de lengte van een persoon en of de resultaten van een toets in kansrekening iets te maken hebben met de resultaten in een toets over statistiek.

We zullen in deze les naar de samenhang tussen twee stochasten X en Y kijken, die gemeenschappelijke optreden. De belangrijkste vraag is hierbij of er een lineaire samenhang bestaat, dat wil zeggen of we Y kunnen schrijven in de vorm $Y = aX + b$.

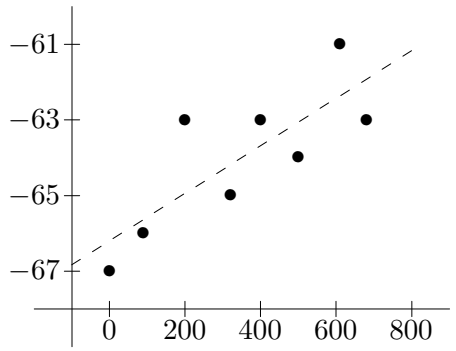
6.1 Regressie

Een vaak gebruikte aanpak om een idee van de samenhang van twee stochasten te krijgen, is een *scatterplot*, waarin de waarden voor de stochasten X en Y als coördinaten van punten in het 2-dimensionale $x - y$-vlak opgevat worden. Vaak kan men al aan de hand van een scatterplot een samenhang tussen de stochasten vast stellen, bijvoorbeeld als de punten enigszins op een lijn, op een parabool of op een exponentiële functie liggen.

Voorbeeld: Aan de zuidpool is het erg koud en men kan zich afvragen hou de temperatuur veranderd als men zich van de zuidpool verwijderd. Op een dag met $-67°C$ aan de zuidpool meet men in verschillende afstanden ook de temperatuur en krijgt de waarden in de volgende tabel:

i	1	2	3	4	5	6	7	8
afstand/km	0	90	200	320	400	500	610	680
temperatuur/$°C$	-67	-66	-63	-65	-63	-64	-61	-63

Als men de afstanden als x- en de temperaturen als y-coördinaten neemt, krijgt men het volgende plaatje. Daarbij is ook al een poging gedaan om een lijn door de waarden te leggen, die een samenhang tussen afstand en temperatuur zou kunnen beschrijven.



Figuur 18: Scatterplot met regressielijn.

Als we de grafiek van een functie van x door de punten leggen, nemen we een fundamentele beslissing: We maken x de *onafhankelijke* variabele en y de *afhankelijke* variabele. We spreken hierbij van *regressie* (van het Latijnse *regredere* = terugstappen), omdat we de afhankelijke variabele y op de onafhankelijke variabele x terug brengen. Merk op dat we alleen maar een samenhang tussen X en Y willen *beschrijven*, dit heeft niets met een eventuële causaliteit te maken. We maken helemaal geen uitspraak erover of X de *oorzaak* voor Y is of niet.

Als we nu beslissen dat we de grafiek van een functie $f(x)$ door de punten willen leggen, stelt zich de vraag hoe we zo'n grafiek kunnen bepalen. Omdat we y als afhankelijke variabele zien, is het een voor de hand liggende gedachte, de som van de kwadratische afstanden tussen de gevonden y-waarden en de grafiek te minimaliseren. Men spreekt dan van een *best fit* functie. Maar als we geen verdere aannamen over de functie $f(x)$ maken, is dit nog geen goed concept, want er laat zich aantonen dat er voor n punten altijd een veelterm van graad $n-1$ bestaat die precies door de n punten loopt. Hiervoor hebben we wel nodig dat de x-waarden van de punten verschillend zijn. We kunnen dus door n punten altijd een *perfecte fit* bereiken, maar die is vaak niet zo erg zinvol:

(i) De veelterm heeft in het algemeen gigantisch grote coëfficiënten en een grafiek met extreme stijgingen en geeft nauwelijks inzicht in de samenhang tussen de stochasten X en Y.

(ii) De punten die we in de scatterplot hebben getekend zijn vaak een steekproef van een veel grotere populatie (soms zelfs oneindig) en het voorspellingspotentiaal bij zo'n wilde functie is laag. We kunnen niet verwachten dat de y-waarde voor een x tussen of buiten de x_i goed door de functie beschreven wordt.

Voor een *fit* door een scatterplot neemt men daarom meestal geen willekeurige functie, maar een functie van bepaalde type die van een klein aantal parameters afhangt. Typische keuzes hiervoor zijn:

(1) Lineaire functies $f(x) = ax + b$.

(2) Kwadratische functies $f(x) = ax^2 + bx + c$.

(3) Exponentiële functies $f(x) = ae^{bx}$.

Het eenvoudigste maar ook meest belangrijke geval zijn de lineaire functies. Verschillende andere samenhangen laten zich door een transformatie van de variabelen op een lineaire samenhang terugbrengen. De volgende tabel geeft voor een aantal belangrijke relaties de hiervoor benodigde transformaties aan. Merk op dat de beste fit dan de kwadratische afstanden van de getransformeerde variabelen minimaliseert en niet de kwadratische afstanden bij de originele variabelen. Maar vaak is dit juist wenselijk, omdat bijvoorbeeld bij de exponentiële functie ae^{bx} de afwijkingen voor grotere waarden van x zeer groot worden en daarom de fit vooral door de grootste waarde van x bepaald zou worden.

veronderstelde samenhang	transformaties	lineaire samenhang
$y = \frac{a}{x} + b$	$x' = \frac{1}{x}, y' = y$	$y' = ax' + b$
$\frac{1}{y} = ax + b$	$x' = x, y' = \frac{1}{y}$	$y' = ax' + b$
$\frac{1}{y} = \frac{a}{x} + b$	$x' = \frac{1}{x}, y' = \frac{1}{y}$	$y' = ax' + b$
$y = be^{ax}$	$x' = x, y' = \log(y)$	$y' = ax' + \log(b)$
$y = bx^a$	$x' = \log(x), y' = \log(y)$	$y' = ax' + \log(b)$
$y = a\log(x) + b$	$x' = \log(x), y' = y$	$y' = ax' + b$

Om te herkennen dat er een samenhang van de vorm $y = be^{ax}$ bestaat, wordt vaak *logaritmisch* papier gebruikt, waarop de y-waarden al op een logaritmische schaal geplaatst worden. Op zo'n soort papier moeten de punten dan op een lijn liggen. Analoog wordt voor een functie van de vorm $y = bx^a$ *dubbel logaritmisch* papier gebruikt, waarbij zowel de x- en de y-waarden op logaritmische schalen ingetekend worden. Voor het menselijke oog is het veel makkelijker om te herkennen of punten enigszins dicht bij een lijn liggen, dan te zien of ze dicht bij de grafiek van een exponentiële of machtsfunctie liggen.

6.2 De regressielijn

We gaan nu voor een verzameling $(x_1, y_1), \ldots, (x_n, y_n)$ van punten de lijn $y = f(x) = ax + b$ bepalen zo dat de kwadratische afstanden van de y_i van deze lijn minimaal worden. Deze lijn heet de *regressielijn* door de punten (x_i, y_i). De voorwaarde aan de regressielijn is, dat de functie

$$V(a, b) := \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

minimaal wordt. Bij functies van één veranderlijke vindt men een minimum door de nulpunten van de afgeleide te zoeken. Dit lukt bij functies van meer veranderlijken net zo, waarbij men apart de *partiële afgeleiden* naar de verschillende variabelen bekijkt. Bij een partiële afgeleide worden de andere variabelen als constanten behandeld.

In ons geval hebben we

$$\frac{\partial}{\partial a} V(a, b) = \sum_i 2(ax_i + b - y_i)x_i \quad \text{en} \quad \frac{\partial}{\partial b} V(a, b) = \sum_i 2(ax_i + b - y_i).$$

Als we de twee partiële afgeleiden gelijk aan 0 zetten, volgt uit $\frac{\partial}{\partial a} V(a, b) = 0$

$$a \sum_i x_i^2 + b \sum_i x_i = \sum_i x_i y_i$$

en uit $\frac{\partial}{\partial b} V(a, b) = 0$ krijgen we

$$a \sum_i x_i + nb = \sum_i y_i.$$

Dit zijn twee *lineaire vergelijkingen* in de twee onbekenden a en b, die we in principe rechtstreeks zouden kunnen oplossen, maar die enigszins onoverzichtelijke coëfficiënten hebben. We gaan de coëfficiënten daarom met behulp van gemiddelden iets handiger noteren en definiëren hiervoor:

$$\overline{x} = \frac{1}{n}\sum_i x_i, \quad \overline{x^2} = \frac{1}{n}\sum_i x_i^2, \quad \overline{y} = \frac{1}{n}\sum_i y_i, \quad \overline{xy} = \frac{1}{n}\sum_i x_i y_i$$

dan worden de vergelijkingen

$$a\overline{x^2} + b\overline{x} = \overline{xy} \text{ en } a\overline{x} + b = \overline{y}.$$

Uit de laatste vergelijking volgt rechtstreeks dat

$$b = \overline{y} - a\overline{x}$$

en als we dit in de eerste vergelijking invullen, hebben we $a\overline{x^2} + \overline{x}\,\overline{y} - a\overline{x}^2 = \overline{xy}$ of te wel $a(\overline{x^2} - \overline{x}^2) = \overline{xy} - \overline{x}\,\overline{y}$, dus

$$a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}$$

en

$$b = \frac{\overline{y}\overline{x^2} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2}$$

Merk op: We hebben gezien dat $\overline{y} = a\overline{x} + b$ geldt, dus gaat de regressielijn in het bijzonder door het zwaartepunt $(\overline{x}, \overline{y})$ van de punten (x_i, y_i).

Voorbeeld: In het voorbeeld met de afstanden en temperaturen hebben we $n = 8$, $\overline{x} = 350$, $\overline{x^2} = 174375$, $\overline{y} = -64$ en $\overline{xy} = -22073.75$. Hieruit volgt $a \approx 0.006289$ en $b = \overline{y} - a\overline{x} \approx -66.201$. Dit zijn natuurlijk juist de coëfficiënten van de lijn die in Figuur 18 ingetekend is.

Ook nu lijken de uitdrukkingen voor de coëfficiënten a en b van de regressielijn nog te ingewikkeld om ze te kunnen onthouden, maar als we ons aan de kansrekening herinneren, kunnen we een heel mooi ezelsbruggetje bouwen.

Als x_i en y_i de waarden van twee uniform verdeelde discrete stochasten X en Y met n mogelijke waarden zijn, dan is $\overline{x}$ juist de verwachtingswaarde van X en $\overline{y}$ is de verwachtingswaarde van Y. Verder geldt dat $\overline{x^2} - \overline{x}^2 = E[X^2] - E[X]^2 = Var(X)$, dus de noemer van a is juist de variantie van X. Voor het product van de stochasten geldt $\overline{xy} - \overline{x}\,\overline{y} = E[X \cdot Y] - E[X] \cdot E[Y] = Cov(X, Y)$, dus is de teller van a de *covariantie* van X en Y. Met deze interpretatie hebben we dus

$$a = \frac{Cov(X, Y)}{Var(X)}$$

en dit ziet er een heel stuk beter uit.

Variantie-analyse en de correlatiecoëfficiënt

Oorspronkelijk hadden we de variantie van een uniform verdeelde stochast X met n waarden gedefinieerd door $Var(X) = \frac{1}{n}\sum_i (x_i - \overline{x})^2$ en door dit met $Var(X) = \overline{x^2} - \overline{x}^2$ te vergelijken, krijgen we de volgende identiteit, die ons vaak handig te pas zal komen:

$$\sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - n\overline{x}^2 = n\overline{x^2} - n\overline{x}^2.$$

(Natuurlijk kunnen we deze identiteit ook rechtstreeks narekenen, maar dit hebben we in feite al eerder gedaan, en er is geen reden om vervelend werk twee keer te doen.)

Ook voor de covariantie hadden we een andere schrijfwijze gezien, namelijk $Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$. Als we de verwachtingswaarden weer als gemiddelden schrijven, krijgen we hiermee $Cov(X,Y) = \frac{1}{n}\sum_i (x - \overline{x})(y - \overline{y})$ en dit geeft de identiteit

$$\sum_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_i x_i y_i - \frac{1}{n}(\sum_i x_i)(\sum_i y_i) = n\overline{xy} - n\overline{x}\,\overline{y}.$$

Bij elkaar genomen kunnen we de stijgingscoëfficiënt a van de regressielijn dus ook schrijven als

$$a = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}.$$

We voeren nu nog eens nieuwe notaties in, om deze uitdrukking voor de stijgingscoëfficiënt van de regressielijn handiger te kunnen schrijven, namelijk

$$v_{xx} := \sum_i (x_i - \overline{x})^2, \quad v_{yy} := \sum_i (y_i - \overline{y})^2, \quad v_{xy} := \sum_i (x_i - \overline{x})(y_i - \overline{y})$$

dan geldt (heel kort)

$$a = \frac{v_{xy}}{v_{xx}}.$$

We hebben tot nu toe altijd gezegd dat y de afhankelijke en x de onafhankelijke variabele is. Maar we kunnen natuurlijk de rollen van x en y omdraaien en bij een regressielijn de kwadratische afstanden tot de x-waarden minimaliseren.

De lijn $x = a'y + b'$ die we als regressielijn voor x afhankelijk van y krijgen, zal in het algemeen afwijken van de lijn $y = ax + b$ die y met betrekking tot x uitdrukt. Er laat zich zelfs aantonen dat de twee lijnen alleen maar overeenkomen als de punten precies op een lijn liggen.

Om de nieuwe regressielijn $x = a'y + b'$ te bepalen, zouden we de hele procedure kunnen herhalen, maar dit is helemaal niet nodig. Als we de coördinaten (x_i, y_i) verruilen en dus de punten (y_i, x_i) bekijken, hoeven we in onze formules alleen maar alle letters x door y te vervangen en andersom. Omdat dit voor v_{xy} geen verschil maakt, krijgen we voor de stijgingscoëfficiënt a' de uitdrukking

$$a' := \frac{v_{xy}}{v_{yy}}.$$

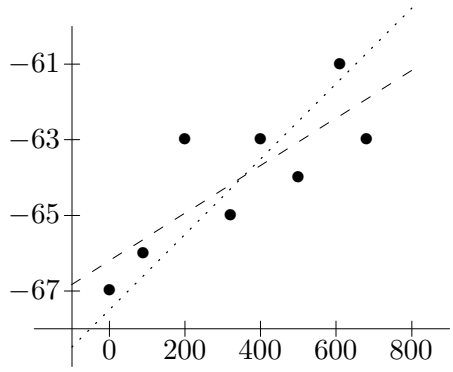
Omdat ook deze regressielijn door het zwaartepunt van de punten moet lopen, geldt verder dat

$$b' = \overline{x} - a'\overline{y}.$$

Om nu de twee regressielijnen met elkaar te kunnen vergelijken, moeten we de vergelijking $x = a'y + b'$ van de nieuwe regressielijn naar y oplossen. Dit geeft

$$y = \frac{1}{a'}x - \frac{b'}{a'} = \frac{v_{yy}}{v_{xy}}x - \frac{b'}{a'}.$$

Voorbeeld: In het voorbeeld hadden we voor x als onafhankelijke variabele de regressielijn $y = 0.006289x - 66.201$ gevonden die de kwadratische afstanden in y-richting minimaliseert. Als we in plaats hiervan naar de regressielijn van x in afhankelijkheid van y kijken, krijgen we de nieuwe regressielijn $y = 0.009962x - 67.487$, die de kwadratische afstanden in x-richting minimaliseert. In Figuur 19 zijn deze twee lijnen te zien, de eerste als stippellijn en de tweede als puntjeslijn. Het is duidelijk dat voor de tweede lijn de afstanden in x-richting kleiner zijn dan voor de eerste lijn.



Figuur 19: Regressielijnen voor x en y als onafhankelijke variabele.

Als we nu de stijgingscoëfficiënten $a = \frac{v_{xy}}{v_{xx}}$ en $\frac{1}{a'} = \frac{v_{yy}}{v_{xy}}$ van de twee regressielijnen vergelijken, krijgen we

$$\frac{a}{\frac{1}{a'}} = a \cdot a' = r^2 \text{ of } \frac{a}{r} = \frac{r}{a'} \quad \text{met} \quad r^2 = \frac{v_{xy}^2}{v_{xx}v_{yy}}.$$

Het getal r^2 is dus een maat voor de afwijking van de stijgingscoëfficiënten van elkaar en hoe dichter de punten bij een lijn liggen, hoe dichter zal r^2 bij 1 liggen.

We definiëren nu

$$r := \frac{v_{xy}}{\sqrt{v_{xx}v_{yy}}}$$

en noemen dit de *correlatiecoëfficiënt* van de punten $(x_1, y_1), \ldots, (x_n, y_n)$.

Een interpretatie van de correlatiecoëfficiënt krijgen we door een redenering die erg op de variantie-analyse lijkt. We kijken hiervoor naar de kwadratische afwijkingen van de y-waarden van het gemiddelde $\overline{y}$ en analyseren hoeveel van deze afwijking door de regressielijn verklaard wordt. Hoe groter het aandeel

van de verklaarde afwijking, hoe beter klopt de aanname van een lineaire samenhang.

Voor iedere waarde y_i noteren we met $\hat{y}_i = ax_i + b$ de volgens de regressielijn verwachte waarde bij x_i, dan heet het verschil $y_i - \hat{y}_i$ tussen de werkelijke waarde en de door de regressielijn voorspelde waarde het *residu* bij x_i. We noemen $(\hat{y}_i - \overline{y})^2$ de (door de regressielijn) *verklaarde afwijking* en $(y_i - \hat{y}_i)^2$ de *onverklaarde afwijking* van y_i van het gemiddelde $\overline{y}$.

We tonen nu aan dat de totale kwadratische afwijking juist de som van de verklaarde en van de onverklaarde afwijking is, d.w.z. we laten zien dat

$$v_{yy} = \sum_i (y_i - \overline{y})^2 = \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{\text{onverklaard}} + \underbrace{\sum_i (\hat{y}_i - \overline{y})^2}_{\text{verklaard}}.$$

Er geldt

$$(y_i - \overline{y})^2 = \big((y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})\big)^2 = (y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}),$$

dus moeten we laten zien dat

$$\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum_i (y_i - ax_i - b)(ax_i + b - \overline{y}) = 0.$$

Door uitwerken vinden we dat

$$(y_i - ax_i - b)(ax_i + b - \overline{y}) = ax_i(y_i - ax_i - b) + b(y_i - ax_i - b) - \overline{y}(y_i - ax_i - b)$$

en de lineaire vergelijkingen waaruit we de coëfficiënten a en b hebben gevonden waren juist

$$\sum_i x_i y_i = a \sum_i x_i^2 + b \sum_i x_i \quad \text{en} \quad \sum_i y_i = a \sum_i x_i + bn.$$

Daarom hebben we

$$\sum_i x_i(y_i - ax_i - b) = 0, \quad \text{en} \sum_i (y_i - ax_i - b) = 0$$

en hieruit volgt dat inderdaad

$$\sum_i (y_i - ax_i - b)(ax_i + b - \overline{y}) = 0.$$

We weten nu dat

$$\frac{\sum_i (y_i - \hat{y})^2}{v_{yy}} + \frac{\sum_i (\hat{y}_i - \overline{y})^2}{v_{yy}} = 1,$$

en $\frac{\sum_i (\hat{y}_i - \overline{y})^2}{v_{yy}}$ is juist het aandeel van de totale kwadratische afwijking dat door de regressielijn verklaard wordt. We gaan nu aantonen dat dit aandeel precies het kwadraat r^2 van de correlatiecoëfficiënt is, dus we laten zien dat

$$\frac{\sum_i (\hat{y}_i - \overline{y})^2}{v_{yy}} = \frac{v_{xy}^2}{v_{xx} v_{yy}} = r^2.$$

Om het rekenwerk iets eenvoudiger te houden, gaan we ervan uit dat we de paren (x_i, y_i) zo hebben verschoven dat $\overline{x} = 0$ en $\overline{y} = 0$ is. In dit geval geldt

$$a = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{v_{xy}}{v_{xx}} \quad \text{en} \quad b = 0.$$

Hieruit volgt

$$\frac{\sum_i (\hat{y}_i - \overline{y})^2}{v_{yy}} = \frac{\sum_i (a x_i)^2}{v_{yy}} = \frac{a^2 \sum_i x_i^2}{v_{yy}} = a^2 \frac{v_{xx}}{v_{yy}} = \frac{v_{xy}^2}{v_{xx}^2} \frac{v_{xx}}{v_{yy}} = \frac{v_{xy}^2}{v_{xx} v_{yy}} = r^2.$$

In het bijzonder volgt hiermee ook, dat $0 \leq r^2 \leq 1$ en dus $-1 \leq r \leq 1$. Omdat r^2 het aandeel van de kwadratische afwijking aangeeft dat door de regressielijn verklaard wordt, noemt men r^2 ook de *coëfficiënt van beslistheid* (coëfficiënt of determination).

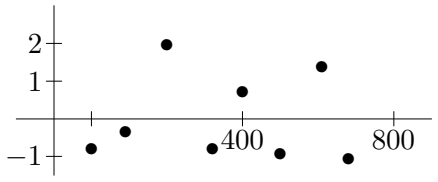
Voorbeeld: In het voorbeeld hadden we voor de twee regressielijnen de stijgingen $a = 0.006289$ en $\frac{1}{a'} = 0.009962$ gevonden, hieruit volgt dat $r^2 = 0.631$ en $r = 0.795$. Dit kunnen we ook uit de kwadratische afwijkingen afleiden:

Er geldt dat $\sum_i (y_i - \overline{y})^2 = 26$ en voor de residuën krijgen we de waarden

i	1	2	3	4	5	6	7	8
afstand	0	90	200	320	400	500	610	680
residu $y_i - \hat{y}_i$	−0.80	−0.36	1.94	−0.81	0.69	−0.94	1.36	−1.08

Voor de door de regressielijn verklaarde kwadratische afwijking krijgen we in dit voorbeeld $\sum_i (\hat{y}_i - \overline{y})^2 = 16.415$ en de onverklaarde afwijking is $\sum_i (y_i - \hat{y}_i)^2 = 9.585$ en we hebben $r^2 = \frac{16.415}{26} = 0.631$.

Om een idee te krijgen of een regressielijn wel een redelijke benadering geeft, is het vaak handig om alleen maar naar de residuën $y_i - \hat{y}_i$ te kijken. Als de residuën een soort van patroon laten zien, is er waarschijnlijk iets mis met de lineaire samenhang tussen de x- en de y-waarden, terwijl een willekeurige verdeling rond de x-as een goed teken is. In Figuur 20 zijn de residuën voor het voorbeeld met de temperaturen in verschillende afstanden van de zuidpool te zien. Er valt in dit geval geen duidelijk patroon op.



Figuur 20: Residuën $y_i - \hat{y}_i$ voor een regressielijn.

Steekproeven

Voor twee stochasten X en Y die twee kenmerken van een populatie beschrijven, heeft de stochast (X, Y) voor de paren van waarden een 2-dimensionale kansverdeling. Net als bij de gewone kansverdelingen kan deze 2-dimensionale kansverdeling discreet of continu zijn.

In het discrete geval is de kansverdeling door de kansen $P(X = x, Y = y)$ bepaald, in het continue geval door een dichtheidsfunctie $f(x, y)$ met verdelingsfunctie

$$F(x, y) := \int_{u \le x, v \le y} f(u, v)\, dv\, du.$$

De integratie over het gebied $-\infty < u \le x, -\infty < v \le y$ wordt hierbij door de twee in elkaar geschakelde gewone integraties berekend, dus

$$F(x, y) := \int_{-\infty}^{x} \big(\int_{-\infty}^{y} f(u, v)\, dv \big) du$$

waarbij (net als bij de partiële afgeleiden) de variabele van de buitenste integratie in de binnenste integraal als constante beschouwd wordt.

De paren $(x_1, y_1), \ldots, (x_n, y_n)$ vatten we nu op als steekproef voor de 2-dimensionale kansverdeling van de stochast (X, Y). Hierbij zijn dan natuurlijk de x_i steekproefwaarden voor X en de y_i steekproefwaarden voor Y, en we definiëren de steekproefvarianties (zo als altijd) door

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2, \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

Analoog definiëren we ook een *steekproefcovariantie* door

$$s_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}),$$

dan kunnen we de stijgingscoëfficiënt a van de regressielijn door de punten (x_i, y_i) en de correlatiecoëfficiënt r schrijven als

$$a = \frac{s_{xy}}{s_x^2} \quad \text{en} \quad r = \frac{s_{xy}}{s_x s_y}.$$

In het bijzonder hebben we

$$(y - \overline{y}) = a(x - \overline{x}) = \frac{s_{xy}}{s_x^2} (x - \overline{x}) = \frac{s_y}{s_x} r (x - \overline{x}).$$

Als we nu (op de inmiddels welbekende manier) de twee variabelen x en y op z-coördinaten transformeren, zien we dat

$$\frac{y - \overline{y}}{s_y} = \frac{s_{xy}}{s_x^2 s_y} (x - \overline{x}) = \frac{s_{xy}}{s_x s_y} \frac{x - \overline{x}}{s_x} = r \frac{x - \overline{x}}{s_x}$$

dus vinden we ook hier weer een nieuwe interpretatie van de correlatiecoëfficiënt.

6.3 Het lineaire regressie model

Zelfs als we veronderstellen, dat er een lineaire samenhang tussen de stochasten X en Y bestaat, kan men niet verwachten dat de punten (x_i, y_i) van de

steekproef precies op een lijn liggen. De aanname dat de afwijkingen van de lijn toevallige fouten zijn leidt tot het *lineaire regressie model*, waarbij men een lineaire samenhang $Y = \alpha X + \beta$ tussen de stochasten verondersteld, die door een fout term verruisd is. Dit betekent dat de y-waarden y_i in de steekproef voor een gegeven waarde x_i van de vorm

$$y_i = \alpha x_i + \beta + \varepsilon_i$$

zijn, waarbij ε_i een foutterm is. Men neemt verder aan dat voor alle waarden van x_i de fouttermen normaal verdeeld met verwachtingswaarde 0 zijn. In principe zou de variantie van de fouttermen van de x-waarde x_i kunnen afhangen, maar om het model hanteerbaar te houden, gaat men ook hier ervan uit, dat de varianties van alle fouttermen gelijk zijn. De stochast E_i die de foutterm bij x_i aangeeft, heeft dus in het lineaire regressie model een normale verdeling met verwachtingswaarde 0 en variantie σ^2.

Er laat zich aantonen dat het lineaire regressie model juist is als de kansverdeling van (X, Y) een 2-dimensionale normale verdeling is. Voor twee *onafhankelijke* normaal verdeelde stochasten X en Y met $E[X] = \mu_X$, $E[Y] = \mu_Y$, $Var(X) = \sigma_X^2$, $Var(Y) = \sigma_Y^2$ is de gemeenschappelijke 2-dimensionale normale verdeling van (X, Y) gegeven door de dichtheidsfunctie

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} e^{-\frac{1}{2}((\frac{x-\mu_X}{\sigma_X})^2+(\frac{y-\mu_Y}{\sigma_Y})^2)}$$

die juist het product van de aparte dichtheidsfuncties voor X en Y is.

Als X en Y niet onafhankelijk zijn, is $\rho := \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$ de correlatiecoëfficiënt van X en Y. In dit geval heeft de 2-dimensionale normale verdeling van (X, Y) de dichtheidsfunctie

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}((\frac{x-\mu_X}{\sigma_X})^2-2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y}+(\frac{y-\mu_Y}{\sigma_Y})^2)}.$$

Ook als het paar (X, Y) geen 2-dimensionale normale verdeling heeft, biedt het lineaire regressie model in veel gevallen een redelijke aanpak, omdat (net als bij de gewone kansverdelingen) de verdeling van (X, Y) vaak goed door een normale verdeling benaderd wordt en dus ook het lineaire regressie model bij benadering klopt.

Schatters

De coëfficiënten a en b van de regressielijn door de punten (x_i, y_i) zien we nu als schatting voor de parameters α en β van het lineaire regressie model. Om dit verder te analyseren, noemen we de schatters, die op een concrete steekproef de waarden a en b geven A en B en de stochast die de verdeling van de y-waarden voor een vaste x_i beschrijft, noemen we Y_i. Volgens onze veronderstelling heeft dan Y_i een normale verdeling met verwachtingswaarde $\mu_i := \alpha x_i + \beta$ en variantie σ^2. Verder krijgen we ook een schatting e_i voor de foutterm ε_i, door $e_i := y_i - (ax_i + b)$ te definiëren.

We gaan nu de verwachtingswaarden en varianties van de schatters A en B bepalen. Hiervoor merken we eerst op dat uit de definitie van het gemiddelde $\overline{x}$ volgt, dat $\sum_i (x_i - \overline{x}) = 0$. Hieruit krijgen we

$$v_{xx} = \sum_i (x_i - \overline{x})(x_i - \overline{x}) = \sum_i (x_i - \overline{x})x_i - \overline{x} \underbrace{\sum_i (x_i - \overline{x})}_{=0} = \sum_i (x_i - \overline{x})x_i$$

en

$$v_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_i (x_i - \overline{x})y_i - \overline{y} \underbrace{\sum_i (x_i - \overline{x})}_{=0} = \sum_i (x_i - \overline{x})y_i.$$

Uit het laatste volgt voor de stijgingscoëfficiënt a van de regressielijn, dat

$$a = \frac{v_{xy}}{v_{xx}} = \sum_i \frac{x_i - \overline{x}}{v_{xx}} y_i$$

en dus kunnen we de schatter A schrijven als

$$A = \sum_i \frac{x_i - \overline{x}}{v_{xx}} Y_i.$$

Voor de verwachtingswaarde $E[A]$ krijgen we hiermee

$$\begin{aligned} E[A] &= \sum_i \frac{x_i - \overline{x}}{v_{xx}} E[Y_i] = \sum_i \frac{x_i - \overline{x}}{v_{xx}} (\alpha x_i + \beta) \\ &= \frac{\alpha}{v_{xx}} \sum_i (x_i - \overline{x})x_i + \frac{\beta}{v_{xx}} \underbrace{\sum_i (x_i - \overline{x})}_{=0} = \frac{\alpha}{v_{xx}} v_{xx} = \alpha, \end{aligned}$$

dus is A een zuivere schatter voor α.

Omdat de Y_i onafhankelijk zijn en variantie σ^2 hebben, krijgen we voor de variantie $Var(A)$ dat

$$Var(A) = \sum_i (\frac{x_i - \overline{x}}{v_{xx}})^2 Var(Y_i) = \frac{v_{xx}}{v_{xx}^2} \sigma^2 = \frac{\sigma^2}{v_{xx}}.$$

Merk op dat de schatting a van α beter wordt naarmate de spreiding v_{xx} van de x-waarden in de steekproef groter wordt. Dit zou men ook verwachten, want de stijging van een regressielijn door punten met sterk verspreide x-waarden is minder gevoelig tegen schommelingen in de y-waarden van de punten dan een lijn door punten met x-waarden die dicht bij elkaar liggen.

Voor de schatter B gebruiken we nu dat a en b samenhangen door de relatie $\overline{y} = a\overline{x} + b$. Omdat de x-waarden x_i vast gekozen zijn, is $\overline{x}$ bij alle steekproeven hetzelfde en de verdeling van $\overline{y}$ wordt beschreven door de stochast $\frac{1}{n}\sum_i Y_i$. Voor de schatters A en B geldt dus de relatie

$$\frac{1}{n} \sum_i Y_i = A\overline{x} + B.$$

en hieruit volgt voor $B = \frac{1}{n}\sum_i Y_i - A\overline{x}$:

$$E[B] = \frac{1}{n}\sum_i E[Y_i] - E[A]\overline{x} = \frac{1}{n}\sum_i(\alpha x_i + \beta) - \alpha\overline{x} = \alpha\overline{x} + \frac{1}{n}\,n\beta - \alpha\overline{x} = \beta.$$

Ook B is dus een zuivere schatter voor de coëfficiënt β van het lineaire regressie model. Voor de variantie $Var(B)$ geldt:

$$Var(B) = \frac{1}{n^2}\sum_i Var(Y_i) + \overline{x}^2 Var(A) = \frac{1}{n^2}n\sigma^2 + \frac{\overline{x}^2}{v_{xx}}\sigma^2 = (\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}})\sigma^2.$$

Omdat $v_{xx} = n\overline{x^2} - n\overline{x}^2$, geldt $\overline{x}^2 = \overline{x^2} - \frac{1}{n}v_{xx}$ en dus kunnen we $Var(B)$ ook schrijven als

$$Var(B) = (\frac{\overline{x^2}}{v_{xx}})\sigma^2.$$

Ten slotte gaan we de schatter $C = \sum_i E_i^2$ analyseren, die als schatting de som $\sum_i e_i^2$ van de kwadraten van de residuën $e_i = y_i - \hat{y}_i$ geeft. We hadden al gezien dat $v_{yy} = \sum_i(y_i - \overline{y})^2 = \sum_i(y_i - \hat{y}_i)^2 + \sum_i(\hat{y}_i - \overline{y})^2$ is. Hieruit volgt dat

$$\sum_i e_i^2 = \sum_i(y_i - \hat{y}_i)^2 = v_{yy} - \sum_i(\hat{y}_i - \overline{y})^2 = v_{yy} - \frac{v_{xy}^2}{v_{xx}} = v_{yy} - a^2 v_{xx}.$$

Als we $\overline{Y} := \frac{1}{n}\sum_i Y_i$ schrijven, volgt hieruit voor de schatter $C = \sum_i E_i^2$ dat

$$C = \sum_i(Y_i - \overline{Y})^2 - v_{xx}A^2 = \sum_i Y_i^2 - n\overline{Y}^2 - v_{xx}A^2.$$

Met de relatie $Var(X) = E[X^2] - E[X]^2$, dus $E[X^2] = Var(X) + E[X]^2$, volgt hieruit

$$\begin{aligned}
E[C] &= \sum_i E[Y_i^2] - nE[\overline{Y}^2] - v_{xx}E[A^2] \\
&= \sum_i(Var(Y_i) + E[Y_i]^2) - n(Var(\overline{Y}) + E[\overline{Y}]^2) - v_{xx}(Var(A) + E[A]^2) \\
&= \sum_i(\sigma^2 + (\alpha x_i + \beta)^2) - n(\frac{1}{n}\sigma^2 + (\alpha\overline{x} + \beta)^2) - v_{xx}(\frac{1}{v_{xx}}\sigma^2 + \alpha^2) \\
&= n\sigma^2 + \sum_i(\alpha x_i + \beta)^2 - \sigma^2 - n(\alpha\overline{x} + \beta)^2 - \sigma^2 - v_{xx}\alpha^2 \\
&= (n-2)\sigma^2 + \sum_i(\alpha^2 x_i^2 + 2\alpha\beta x_i + \beta^2) - n\alpha^2\overline{x}^2 - 2\alpha\beta n\overline{x} - n\beta^2 - v_{xx}\alpha^2 \\
&= (n-2)\sigma^2 + \alpha^2(n\overline{x^2} - n\overline{x}^2 - v_{xx}) \\
&= (n-2)\sigma^2.
\end{aligned}$$

In de laatste stap hebben we hierbij gebruik ervan gemaakt dat $v_{xx} = \sum_i(x_i - \overline{x})^2 = n\overline{x^2} - n\overline{x}^2$. Uit $E[C] = (n-2)\sigma^2$ volgt in het bijzonder:

$$\frac{1}{n-2}\sum_i E_i^2 \text{ een zuivere schatter voor } \sigma^2.$$

De reden dat we in dit geval 2 vrijheidsgraden verliezen is, dat we de twee coëfficiënten a en b uit de steekproefwaarden hebben geschat.

Betrouwbaarheidsintervallen voor de parameters van het model

Volgens onze succesvolle strategie analyseren we nu weer onze schatters A en B om betrouwbaarheidsintervallen rond de schattingen a en b voor de coëfficiënten α en β van het lineaire regressie model te vinden.

Zo als altijd verschuiven we de schatter zo dat zijn verwachtingswaarde 0 wordt en delen vervolgens door zijn standaardafwijking, dus we kijken naar $Z := \frac{A-E[A]}{\sqrt{Var(A)}}$. Er laat zich aantonen dat de zo verkregen stochast

$$Z := \frac{A-\alpha}{\frac{\sigma}{\sqrt{v_{xx}}}} = \frac{A-\alpha}{\sigma}\sqrt{v_{xx}}$$

een standaard-normale verdeling heeft, dus kunnen we met de z-waarden betrouwbaarheidsintervallen definiëren. Er geldt dat

$$\begin{aligned} P(|Z| \leq z_{\frac{1-\gamma}{2}}) &= P(\alpha - z_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{v_{xx}}} \leq A \leq \alpha + z_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{v_{xx}}}) \\ &= P(A - z_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{v_{xx}}} \leq \alpha \leq A + z_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{v_{xx}}}) = \gamma. \end{aligned}$$

De stijgingscoëfficiënt a die we uit de steekproef hebben berekend, levert dus op betrouwbaarheidslevel γ (let op: α is nu de parameter van het lineaire regressie model) voor α het betrouwbaarheidsinterval

$$\left[a - z_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{v_{xx}}}, a + z_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{v_{xx}}}\right].$$

In de meeste gevallen zullen we de variantie σ^2 van de fouttermen niet kennen, in dit geval moeten we σ door de schatting

$$s := \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

vervangen. Hiermee krijgen we de stochast

$$T_\alpha := \frac{A-\alpha}{\frac{s}{\sqrt{v_{xx}}}} = \frac{A-\alpha}{s}\sqrt{v_{xx}}$$

en er laat zich aantonen dat T_α een Student-t verdeling met $n-2$ vrijheidsgraden heeft.

We hoeven dus in het boven gevonden betrouwbaarheidsinterval voor α alleen maar de kritieke z-waarde $z_{\frac{1-\gamma}{2}}$ door de kritieke t-waarde $t_{\frac{1-\gamma}{2}}$ van een Student-t verdeling met $n-2$ vrijheidsgraden te vervangen en de standaardafwijking σ door de schatting s en krijgen zo (bij onbekende variantie σ^2) op betrouwbaarheidslevel γ het volgende betrouwbaarheidsinterval voor α:

$$\left[a - t_{\frac{1-\gamma}{2}}\frac{s}{\sqrt{v_{xx}}}, a + t_{\frac{1-\gamma}{2}}\frac{s}{\sqrt{v_{xx}}}\right].$$

De stochast T_α laat ook weer het verband tussen de regressielijn en de variantie-analyse zien. Men kan aantonen dat T_α^2 een F-verdeling met 1 en $n-2$ vrijheidsgraden heeft.

Met behulp van de betrouwbaarheidsintervallen kunnen we nu ook toetsen voor de coëfficiënt α van het lineaire regressie model definiëren. De meest belangrijke vraag is hierbij meestal, of de steekproef (x_i, y_i) evidentie geeft tegen de nulhypothese

$$H_0 : \alpha = 0.$$

Bij een tweezijdige toets zullen we de nulhypothese $\alpha = 0$ op een significantie level van $1-\gamma$ verwerpen, als

$$|a| > z_{\frac{1-\gamma}{2}} \frac{\sigma}{\sqrt{v_{xx}}} \text{ bij bekende } \sigma^2 \quad \text{en} \quad |a| > t_{\frac{1-\gamma}{2}} \frac{s}{\sqrt{v_{xx}}} \text{ bij onbekende } \sigma^2.$$

Voorbeeld: In ons voorbeeld hadden we $a = 0.00629$ gevonden. We berekenen verder dat $v_{xx} = 415000$ en $\sum_i e_i^2 = 9.585$. Hieruit krijgen we de schatting $s^2 = 1.598$ voor σ^2 en dus voor σ de schatting $s = 1.264$. Om de nulhypothese $\alpha = 0$ te toetsen, moeten we $\frac{a}{s}\sqrt{v_{xx}}$ met de t-waarden van een Student-t verdeling met $8-2 = 6$ vrijheidsgraden vergelijken. Voor $\gamma = 95\%$ hebben we $t_{\frac{1-\gamma}{2}} = t_{6,0.025} = 2.45$ en voor onze waarden geldt dat $\frac{a}{s}\sqrt{v_{xx}} = 3.21$, dus kunnen we op een significantielevel van 5% de nulhypothese $\alpha = 0$ verwerpen. Het betrouwbaarheidsinterval dat we op level 5% voor α vinden is

$$\left[a - 2.45 \cdot \frac{s}{\sqrt{v_{xx}}}, a + 2.45 \cdot \frac{s}{\sqrt{v_{xx}}}\right] = [0.00148, 0.01110].$$

De reden dat we hier zo een relatief groot interval voor α krijgen ligt in het feit dat er bij een regressielijn door slechts 8 punten geen grote zekerheid over de stijging kan bestaan.

Minder belangrijk dan A is de schatter B voor de verschuiving β op de y-as. Ook hier krijgt men door de transformatie $Z := \frac{B-E[B]}{\sqrt{Var(B)}}$ en standaard-normaal verdeelde stochast, namelijk

$$Z := \frac{B-\beta}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}}}}.$$

Dit geeft op onbetrouwbaarheidslevel $1-\gamma$ voor β het betrouwbaarheidsinterval

$$\left[b - z_{\frac{1-\gamma}{2}}\,\sigma\,\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}}}, b + z_{\frac{1-\gamma}{2}}\,\sigma\,\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}}}\right].$$

Ook in dit geval geeft vervangen van σ door de schatting $s = \sqrt{\frac{\sum_i e_i^2}{n-2}}$ een stochast T_β met een Student-t verdeling met $n-2$ vrijheidsgraden, namelijk

$$T_\beta := \frac{B-\beta}{s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}}}}.$$

Hieruit volgt bij onbekende variantie σ^2 het betrouwbaarheidsinterval

$$\left[b - t_{\frac{1-\gamma}{2}}\, s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}}}, b + t_{\frac{1-\gamma}{2}}\, s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}}}\right]$$

voor β.

Betrouwbaarheidsintervallen voor de waarden

Uit de schatters A en B kunnen we voor een willekeurige waarde x door

$$M_x := Ax + B$$

een schatter voor het gemiddelde μ_x van de y-waarden voor de x-waarde x maken. Merk op dat we volgens het lineaire regressie model nog steeds ervan uit gaan dat de y-waarden voor x een normale verdeling rond $\mu_x = \alpha x + \beta$ hebben.

De verwachtingswaarde van M_x is $E[M_x] = \alpha x + \beta = \mu_x$ en dus is M_x een zuivere schatter voor μ_x.

Om de variantie van M_x te bepalen, merken we eerst op dat we hadden gezien dat $\overline{y} = a\overline{x} + b$, waaruit volgt dat voor een punt (x, y) op de regressielijn geldt dat $y - \overline{y} = a(x - \overline{x})$. Voor de schatter M_x volgt hieruit, dat

$$M_x - \overline{Y} = A(x - \overline{x}) \text{ en dus } M_x = \overline{Y} + A(x - \overline{x})$$

waarbij $\overline{Y}$ en A onafhankelijke stochasten zijn. Voor de variantie van M_x volgt hieruit

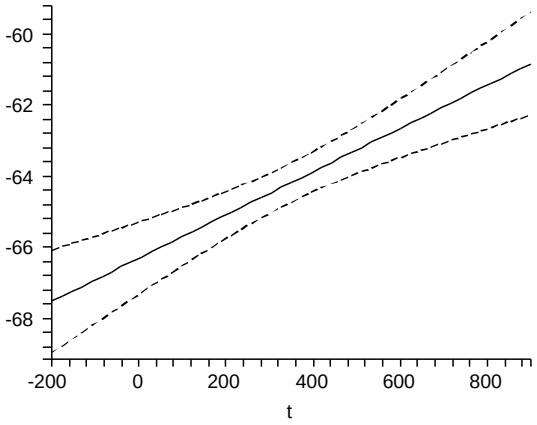
$$Var(M_x) = Var(\overline{Y}) + (x-\overline{x})^2 Var(A) = \frac{\sigma^2}{n} + (x-\overline{x})^2 \frac{\sigma^2}{v_{xx}} = (\frac{1}{n} + \frac{(x-\overline{x})^2}{v_{xx}})\sigma^2.$$

Omdat B juist de schatter M_x voor $x = 0$ is, moeten we hier voor $x = 0$ de variantie van B terugvinden, en dit is inderdaad het geval, want we hadden gevonden dat $Var(B) = (\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}})\sigma^2$.

Merk op dat een betrouwbaarheidsinterval voor μ_x afhangt van de afstand tussen x en $\overline{x}$. Voor x dicht bij het gemiddelde is het betrouwbaarheidsinterval minder groot dan voor verder verwijderde. In Figuur 21 is dit voor ons voorbeeld duidelijk te zien, de twee krommen rond de regressielijn geven de grenzen van het betrouwbaarheidsinterval voor μ_x op een betrouwbaarheidslevel van 90% aan.

Als we nu een interval voor de y-waarden voor een zekere x-waarde willen schatten, weten we dat we volgens het lineaire regressie model de schatting $y = ax + b + e$ hebben, waarbij de foutterm e normaal verdeeld met variantie σ^2 is. Als we de schatter die de verdeling van de y-waarden voor de x-waarde x beschrijft met Y_x noteren, volgt hieruit dat

$$Y_x = Ax + B + E_x = M_x + E_x$$



Figuur 21: Betrouwbaarheidsintervallen voor de gemiddelden μ_x.

waarbij E_x normaal verdeeld met verwachtingswaarde 0 en variantie σ^2 is. Hieruit krijgen we

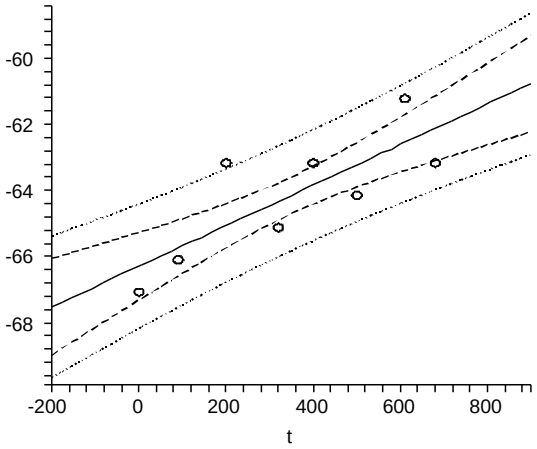
$$E[Y_x] = E[M_x] + E[E_x] = \mu_x,$$

dus is ook Y_x een zuivere schatter voor $\mu_x = \alpha x + \beta$.

Voor de variantie geldt

$$\begin{aligned} Var(Y_x) &= Var(M_x) + Var(E_x) = (\frac{1}{n} + \frac{(x-\overline{x})^2}{v_{xx}})\sigma^2 + \sigma^2 \\ &= (1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{v_{xx}})\sigma^2. \end{aligned}$$

Het zal geen verrassing zijn, dat de schatter Y_x voor de y-waarden een groter interval rond μ_x geeft dan de schatter M_x voor het gemiddelde van de y-waarden.



Figuur 22: Betrouwbaarheidsintervallen voor de y-waarden.

In Figuur 22 zijn voor ons voorbeeld de betrouwbaarheidsintervallen op betrouwbaarheidslevel 90% voor de y-waarden samen met de (kleinere) betrouw-

baarheidsintervallen voor de μ_x te zien. Omdat ook de punten voor de paren (x_i, y_i) aangegeven zijn, kunnen we in het bijzonder herkennen dat 7 van de 8 waarden binnen het betrouwbaarheidsinterval voor de y-waarden liggen, zo als we dat bij een onbetrouwbaarheid van 10% zouden kunnen verwachten.

Correlatie

Als we stochasten X en Y met zekere kansverdelingen en met een gemeenschappelijke kansverdeling voor (X, Y) veronderstellen, dan is de correlatiecoëfficiënt gedefinieerd door

$$\rho := \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}.$$

Als we de paren (x_i, y_i) als steekproef zien, waarbij de y-waarden voor de x-waarde x_i door de stochast Y_i worden beschreven, krijgen we de schatter

$$R := \frac{\sum_i (x_i - \overline{x})(Y_i - \overline{Y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (Y_i - \overline{Y})^2}}$$

die voor een concrete steekproef de correlatiecoëfficiënt

$$r := \frac{v_{xy}}{\sqrt{v_{xx} v_{yy}}}$$

als schatting voor ρ geeft.

We gaan er nu van uit dat het lineaire regressie model inderdaad van toepassing is, dus dat (X, Y) een 2-dimensionale normale verdeling heeft. Dan kunnen we uit r iets over ρ concluderen, als we op R de *Fisher transformatie* toepassen:

$$V := \frac{1}{2} \log(\frac{1+R}{1-R})$$

heeft (bij benadering) een normale verdeling met

$$E[V] = \frac{1}{2} \log(\frac{1+\rho}{1-\rho}) \quad \text{en} \quad Var(V) = \frac{1}{n-3}.$$

Hiermee kunnen we betrouwbaarheidsintervallen voor ρ definiëren en kunnen de nulhypothese $\rho = 0$ toetsen. Er geldt

$$P\left(\frac{1}{2} \log(\frac{1+\rho}{1-\rho}) - \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}} \leq \frac{1}{2} \log(\frac{1+r}{1-r}) \leq \frac{1}{2} \log(\frac{1+\rho}{1-\rho}) + \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}}\right) = \gamma,$$

dus is

$$\left[\frac{1}{2} \log(\frac{1+r}{1-r}) - \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}}, \frac{1}{2} \log(\frac{1+r}{1-r}) + \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}}\right]$$

een betrouwbaarheidsinterval op significantie level $1-\gamma$ voor $\frac{1}{2} \log(\frac{1+\rho}{1-\rho})$.

Om een betrouwbaarheidsinterval voor ρ te krijgen, bepalen we de inverse functie van $f(x) = \frac{1}{2} \log(\frac{1+x}{1-x})$, d.w.z. we lossen $y = \frac{1}{2} \log(\frac{1+x}{1-x})$ naar x op. Er

geldt $e^{2y} = \frac{1+x}{1-x}$, dus is $e^{2y}-1 = \frac{1+x}{1-x} - \frac{1-x}{1-x} = \frac{2x}{1-x}$ en $e^{2y}+1 = \frac{1+x}{1-x} + \frac{1-x}{1-x} = \frac{2}{1-x}$. Hieruit volgt dat

$$x = \frac{e^{2y}-1}{e^{2y}+1} \Leftrightarrow y = \frac{1}{2}\log(\frac{1+x}{1-x}).$$

Met

$$v_1 := \frac{1}{2}\log(\frac{1+r}{1-r}) - \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}} \quad \text{en} \quad v_2 := \frac{1}{2}\log(\frac{1+r}{1-r}) + \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}}$$

vindt men zo dat

$$P\left(\frac{e^{2v_1}-1}{e^{2v_1}+1} \leq \rho \leq \frac{e^{2v_2}-1}{e^{2v_2}+1}\right) = \gamma$$

en dit geeft

$$\left[\frac{e^{2v_1}-1}{e^{2v_1}+1}, \frac{e^{2v_2}-1}{e^{2v_2}+1}\right]$$

als betrouwbaarheidsinterval op significantielevel $1-\gamma$ voor ρ.

Omdat voor $\rho = 0$ geldt dat $\frac{1}{2}\log(\frac{1+\rho}{1-\rho}) = 0$ kunnen we de nulhypothese

$$H_0 : \rho = 0$$

op onbetrouwbaarheidslevel $1-\gamma$ heel makkelijk toetsen: De nulhypothese wordt verworpen als

$$|\frac{1}{2}\log(\frac{1+r}{1-r})| > \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}}.$$

Ten slotte merken we op dat onder de veronderstelling dat (X, Y) een 2-dimensionale normale verdeling heeft, de toets op de nulhypothese $\rho = 0$ equivalent is met de toets dat de stijgingscoëfficiënt $\alpha = 0$ is. Er geldt namelijk dat de stochast T_α geschreven kan worden als

$$T_\alpha = \frac{R}{\sqrt{\frac{1-R^2}{n-2}}} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}.$$

Dit zien we als volgt in: We hebben gezien dat voor een gegeven steekproef (x_i, y_i) de schatter T_α de schatting

$$t = \frac{a-\alpha}{\sqrt{\frac{\sum_i e_i^2}{n-2}}}\sqrt{v_{xx}}$$

geeft. Onder de veronderstelling van de nulhypothese $\alpha = 0$ en door invullen van $a = \frac{v_{xy}}{v_{xx}}$ laat zich dit herschrijven tot

$$t = \frac{a}{\sqrt{\sum_i e_i^2}}\sqrt{v_{xx}(n-2)} = \frac{v_{xy}}{v_{xx}}\sqrt{\frac{v_{xx}(n-2)}{\sum_i e_i^2}} = \frac{v_{xy}}{\sqrt{v_{xx}}}\sqrt{\frac{(n-2)}{\sum_i e_i^2}}.$$

Aan de andere kant geldt dat r^2 het aandeel van de door de regressielijn verklaarde kwadratische afwijking v_{yy} is, daarom is $1-r^2$ juist het onverklaarde aandeel, dus

$$1 - r^2 = \frac{\sum_i e_i^2}{v_{yy}}.$$

Met $r = \frac{v_{xy}}{\sqrt{v_{xx}v_{yy}}}$ volgt hieruit, dat

$$\frac{r}{\sqrt{1-r^2}} = \frac{v_{xy}}{\sqrt{v_{xx}v_{yy}}}\sqrt{\frac{v_{yy}}{\sum_i e_i^2}} = \frac{v_{xy}}{\sqrt{v_{xx}}}\sqrt{\frac{1}{\sum_i e_i^2}},$$

dus geldt inderdaad dat

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

We kunnen dus de nulhypothese

$$H_0 : \alpha = 0$$

ook met behulp van de correlatiecoëfficiënt r van de steekproef (x_i, y_i) toetsen en we verwerpen de nulhypothese op een significantielevel $1-\gamma$ als

$$|\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}| > t_{\frac{1-\gamma}{2}}.$$

Belangrijke begrippen in deze les

- scatterplot
- regressie
- beste fit
- regressielijn
- correlatiecoëfficiënt
- lineair regressie model
- Fisher transformatie