Lineaire regressie

-Het toetsen van samenhang tussen twee variabelen-

Samenvatting

In de huidige onderwijsstructuur wordt van leerlingen op de middelbare scholen verwacht dat zij in een profielwerkstuk verslag doen van bijvoorbeeld een uitgewerkte scheikundige, biologische of natuurkundige proef. Vaak is de onderzoeksvraag op een dusdanige wijze geformuleerd, dat de leerling geïnteresseerd is in een systematisch verschil tussen groepen waarnemingen, bijvoorbeeld door het toepassen van twee behandelingen, of in een samenhang tussen twee grootheden. Voor toetsen van zulke verschillen of samenhangen heeft Wageningen University drie lesbrieven ontworpen. Dit is de derde van deze drie lesbrieven.
In deze lesbrief ga ik in op een situatie waarbij een leerling geïnteresseerd is in de samenhang tussen twee grootheden. Daartoe behandel ik Lineaire Regressie. Dit is een goede methode om de veronderstelde samenhang tussen twee grootheden in een lineair verband samen te vatten en statistisch te onderzoeken. Eerst probeer ik aan de hand van voorbeelden een leiddraad te geven om te bepalen in welke situaties Lineaire Regressie het best kan worden gebruikt. Daarna bespreek ik een strategie, waarin stap voor stap duidelijk wordt gemaakt hoe conclusies kunnen worden getrokken uit een aantal waarnemingen aan twee grootheden. Eén en ander illustreer ik aan de hand van voorbeelden.

Inleiding

Voor een project dat uit moet monden in een profielwerkstuk is het raadzaam om de volgende vijf fasen te doorlopen:

1. Probleemstelling
2. Planning
3. Verkenning
4. Uitvoering
5. Conclusie

Voor veel mensen komt de statistiek pas om de hoek kijken als een proef al is uitgevoerd, de waarnemingen klaar liggen om verwerkt te worden en om tot een conclusie te komen. Dit uitgangspunt is de meest voorkomende beginnersfout. Een conclusie kan slechts op een statistisch verantwoorde manier worden getrokken op basis van de waarnemingen als al vanaf het begin van het project duidelijk is met welke statistische methode men de uiteindelijke waarnemingen gaat verwerken. Tevens is het dan van groot belang om de waarnemingen op een dusdanige manier te verzamelen dat de beoogde methode ook kan worden gebruikt.

In verband met onzekerheid in de waarnemingen is het niet verstandig om te volstaan met één waarneming. Zo'n onzekerheid in de waarnemingen wordt ook wel *stochasticiteit*

genoemd. Wanneer een waarneming meerdere malen herhaald wordt, krijgt degene die de experimenten uitvoert een steeds beter beeld van de onzekerheid in de waarnemingen. Als men geïnteresseerd is in het effect van één bepaalde factor (bv kunstmest) op een zeker kenmerk (bv groei) van de onderzoekseenheden (bv. planten) dan zal men die factor variëren. Tegelijkertijd is het van belang om er op te letten dat alle overige factoren zoveel mogelijk gelijk worden gehouden. Dit is het zogenaamde '*ceteris paribus'* principe. Omdat het echter meestal onmogelijk is om alle overige factoren exact gelijk te houden, is het raadzaam om te loten wie in welke (behandelings)groep terechtkomt. Dit voorkomt een mogelijke verstrengeling van de effecten van niet constante factoren met het effect van de te onderzoeken factor.

Als men in één experiment gelijktijdig de effecten wil onderzoeken van twee factoren, b.v. herbicide en kunstmest, op de groei van planten, dan kunnen de waarnemingen worden gedaan aan vier groepen planten: (1) zonder kunstmest en met herbicide, (2) zonder kunstmest en zonder herbicide, (3) met kunstmest en met herbicide, (4) met kunstmest en zonder herbicide. Ook hier is het belangrijk om door loting te bepalen welke planten aan welke behandelingsgroep worden toegewezen.

In verslagen die op de middelbare school worden gemaakt over proeven is het vaak van belang om een samenhang tussen twee grootheden aan te tonen. Bijvoorbeeld levert het toevoegen van meer kunstmest aan sla na één week grotere slaplanten op? Het verband tussen deze twee grootheden, zeg de hoeveelheid toegevoegde kunstmest x en het gewicht van de sla y, kan redelijk worden beschreven door een rechte lijn: $y = a + bx$. Op grond van n waarnemingen aan x en y kunnen de meest aannemelijke schattingen van de as-afsnede a en de richtingscoëfficiënt b worden berekend.

Ik behandel eerst voorbeelden van problemen waarbij lineaire regressie uitsluitsel kan geven. Vervolgens geef ik een stuk theorie om de achterliggende gedachte een beetje te begrijpen. Het stuk met formules, dat daarna volgt, is voor de liefhebbers. Daarna bespreek ik een strategie, waarin stap voor stap duidelijk wordt gemaakt hoe conclusies kunnen worden getrokken uit een bepaalde groep waarnemingen aan twee grootheden. Een samenvatting van de gemaakte stappen om de regressie met bijbehorende toetsing uit te voeren wordt het stappenplan genoemd en staat in een aparte paragraaf. Het stappenplan wordt geïllustreerd aan de hand van een voorbeeld.

Vraagstellingen die het gebruik van deze toets rechtvaardigen

De vraag "Heeft verhoging van het kooldioxide gehalte in de lucht een effect op de plantengroei?" kan leiden tot een proefopstelling waarbij planten gedurende een bepaalde periode worden opgekweekt bij normaal CO_2, en bijvoorbeeld bij een CO_2 gehalte dat 1%, 2%, 4% of 8% verhoogd CO_2 bevat. Stel dat er bij elke CO_2 concentratie een vijftal planten wordt opgekweekt gedurende 6 weken. Na de zes weken is het gewicht y van alle (in totaal) 25 planten bepaald. Na afloop zijn de coördinaten van 25 punten in het (x,y) vlak bekend waarvan er telkens 5 bij dezelfde x-waarde (=[CO_2 concentratie in proef]/[normale CO_2 concentratie]= 1; 1,01; 1,02; 1,04; 1,08) horen.

Lineaire regressie

-Het toetsen van samenhang tussen twee variabelen-

Of de hoeveelheid meststof die aan planten wordt gegeven mede de groei bepaalt kan experimenteel en statistisch worden onderzocht. Om het effect van een meststof te onderzoeken worden planten van een zelfde soort bij verschillende hoeveelheden mest geteeld. Na afloop van de onderzoekperiode wordt van alle planten het gewicht bepaald. De resultaten zijn samen te vatten als punten in het (x,y) vlak waarbij de x-waarde gelijk is aan hoeveelheid toegevoegde mest en de y-waarde stelt het gewicht van de plant voor.

Heeft een dieet dat gericht is op gewichtsverlies over een langere periode effect? Om zo'n effect te bepalen worden de personen die op dieet zijn wekelijks gewogen. Door op de x-as de tijd in weken uit te zetten tegen de gewichten (op de y-as) kan met regressie een lineair verband door de punten worden geschat. Met een toets kan vervolgens worden bepaald of er een daling optreedt in het gemeten gewicht.

Zijn ervaren bollenpellers sneller dan mensen die pas beginnen? Wellicht heb je wel eens bollen gepeld als vakantiebaan. Mensen die voor het eerst bollen pellen doen vaak veel langer over één mand dan mensen die dit al vaker hebben gedaan. Aangezien men vaak per mand betaald wordt en niet per uur, kan dit aardig wat frustratie opleveren. Om te onderzoeken of dit effect ook een rol speelt bij de kinderen in jouw klas kan je iedereen bijvoorbeeld 4 (kleine) manden met bollen geven. Noteer voor alle eerste manden de tijd die erover gedaan wordt. Meet tevens de tijd benodigd voor de tweede, derde en vierde mand. Door de punten waarbij mandnummer op de x-as staat en de werktijd voor die mand op de y-as kan een lineair verband worden geschat met behulp van lineaire regressie.

Theorie

In alle bovenstaande voorbeelden kan eerst de regressielijn worden geschat. Dan heb je het best passende verband bepaald. Echter het is op dat moment nog niet geheel duidelijk of het verband ook een verband van betekenis is. In de statistiek wordt een verband van betekenis significant genoemd. Om aan te tonen of de helling b of de as-afsnede a in een lineaire regressie significant is dient er een toets te worden uitgevoerd. Eerst geef ik een overzicht van de begrippen die noodzakelijk zijn om een statistische toets goed uit te kunnen voeren. Daarna ga ik over tot het berekenen van enkele grootheden in een voorbeeld. Het laatste deel van deze paragraaf is niet noodzakelijk voor het uitvoeren van de toets. Dit stuk heet dan ook achtergrondinformatie. Ik gebruik het in de inleiding beschreven voorbeeld over het gewicht y van slaplanten na één week te zijn gekweekt met toevoeging van een hoeveelheid kunstmest x als voorbeeld in deze paragraaf.

Om een statistische toets te kunnen uitvoeren is het noodzakelijk een veronderstelling te formuleren. Met het formuleren van een zogenaamde *nulhypothese* wordt de collectie kansverdelingen voor het juiste onderliggende statistische model ingeperkt. De nulhypothese sluit aan bij de tot nu toe aangenomen veronderstellingen (de traditie). Een nulhypothese heeft altijd een tegenhanger, die de *alternatieve hypothese* wordt genoemd. Deze is zodanig geformuleerd dat hij zegt dat het onderliggende statistische model niet

beperkt is tot de collectie modellen onder de nulhypothese. Elke statistische toets geeft de mogelijkheid om op grond van de waarnemingen te besluiten of de nulhypothese al of niet verworpen dient te worden. Voordat men een experiment uitvoert heeft men op grond van kennis of van een redenering soms al een idee of waarnemingen y groter of kleiner zullen zijn als de ingestelde waarde x groter is. In de proef waarbij slaplanten worden opgekweekt met verschillende hoeveelheden mest verwacht je dat planten die met toevoeging van een grotere hoeveelheid mest groeien ook groter zullen worden.

Als in een (alternatieve) hypothese het woord 'groter' of 'kleiner' voorkomt, dan heb je te maken met een *eenzijdig* te toetsen hypothese: afwijkingen van de waarde van een toetsingsgrootheid naar een bepaalde kant wijzen namelijk op ondersteuning van de nulhypothese en afwijkingen de andere kant op leveren aanwijzingen voor het alternatief. Wanneer van tevoren niet duidelijk is naar welke kant een afwijking uit zou kunnen vallen, dan leveren afwijkingen naar beide kanten aanwijzingen voor het alternatief. In het geval van het effect van de meststoffen A en B is er vooraf geen enkele aanwijzing welke beter zou zijn dan de ander. Op deze manier wordt er *tweezijdig* getoetst.

Het kan gebeuren dat de nulhypothese ten onrechte wordt verworpen. De nulhypothese is dan waar, maar gedurende het uitvoeren van de toets is toch het besluit gevallen om hem te verwerpen. Naarmate de kans dat de nulhypothese onterecht wordt verworpen kleiner wordt is de uitkomst van een statistische toets betrouwbaarder. De *onbetrouwbaarheid* van een statistische toets is gelijk aan (het maximum van) de kans dat de nulhypothese onterecht wordt verworpen. De onbetrouwbaarheidsdrempel wordt meestal aangeduid met de Griekse letter α. Veel voorkomende ingestelde waarden van α zijn 0,05 en 0,10. Als een toets wordt uitgevoerd met een onbetrouwbaarheidsdrempel α van 5%, dan is de kans dat de nulhypothese onterecht wordt verworpen dus maximaal 0,05. Bij een tweezijdige toets duiden afwijkingen naar boven of naar beneden op de alternatieve hypothese en de onbetrouwbaarheid waarbij in een tabel moet worden afgelezen is dan $\alpha/2$. Ik hoop dit duidelijk te maken in de uitgewerkte voorbeelden.

Voor het uitvoeren van een statistische toets wordt altijd een uit de waarnemingen afgeleide grootheid gebruikt. Deze wordt de *toetsingsgrootheid* genoemd. Voor regressie kan zowel worden getoetst of de asafsnede a als de richtingscoëfficiënt b van een bepaalde van tevoren veronderstelde waarde afwijkt. Over de aanpak daarvan volgt later meer.

Als ik in het voorbeeld over de slaplanten even uitga van de fictieve gegevens uit tabel 1, dan kan ik hier proberen uit te leggen met formules hoe lineaire regressie in zijn werk gaat. Eerst even herhalen hoe de *gemiddelden* $\bar{x}$ en $\bar{y}$ van respectievelijk x en y worden berekend. Alle ingestelde waarden van x worden opgeteld en door het aantal opgetelde getallen (=*n*)

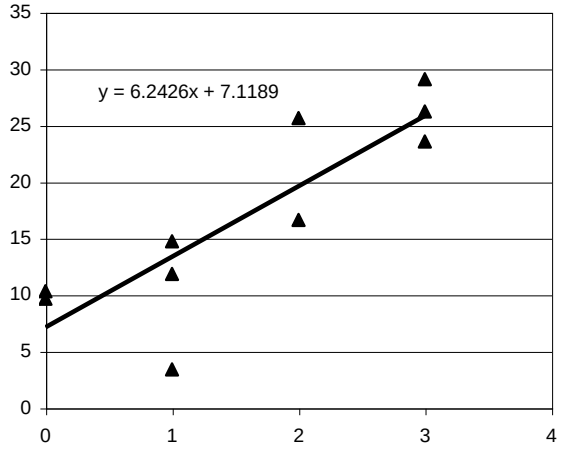
Relatieve kunst-mestgift (x)	**Gewicht slaplantje (y)**
0	10,32
0	9,64
1	11,84
1	14,74
1	3,40
2	16,62
2	25,63
3	23,57
3	29,08
3	26,23

Tabel 1: Fictieve gegevens voor gewicht van slaplanten na 1 week opgegroeid te zijn bij verschillende hoeveelheden kunstmest

gedeeld (analoog voor y). In formulevorm is dat

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

waarin voor het huidige voorbeeld geldt dat n=10 (steekproefgrootte, want er zijn 10 getalparen).
Voor degenen die nog niet bekend zijn met het sommatieteken "Σ" volgt hier een voorbeeld om uit te leggen hoe je verkort kunt opschrijven dat je de getallen 1 tot en met 100 optelt. De formule daarvoor is: $\sum_{i=1}^{100} i$



Figuur 1: Plaatje van de punten uit tabel 1 en de berekende regressielijn.

De interesse gaat echter niet alleen uit naar de gemiddelden van de x- en y-waarden, maar ook naar de spreiding van de gegevens rond dat gemiddelde (de afwijkingen van het gemiddelde). Daartoe worden de *steekproefvarianties* van x en y berekend door van elke waargenomen waarde zijn gemiddelde waarde af te trekken, het resultaat daarvan te kwadrateren, alle zo verkregen getallen bij elkaar op te tellen en uiteindelijk te delen door het aantal waarnemingen – 1. Door het kwadrateren van de afwijkingen kan deze grootheid alleen maar positieve waarden aannemen. In formulevorm ziet dat er als volgt uit.

Achtergrondinformatie

Formules voor de steekproefvarianties van x en y (resp. $(s_x)^2$ en $(s_x)^2$):

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

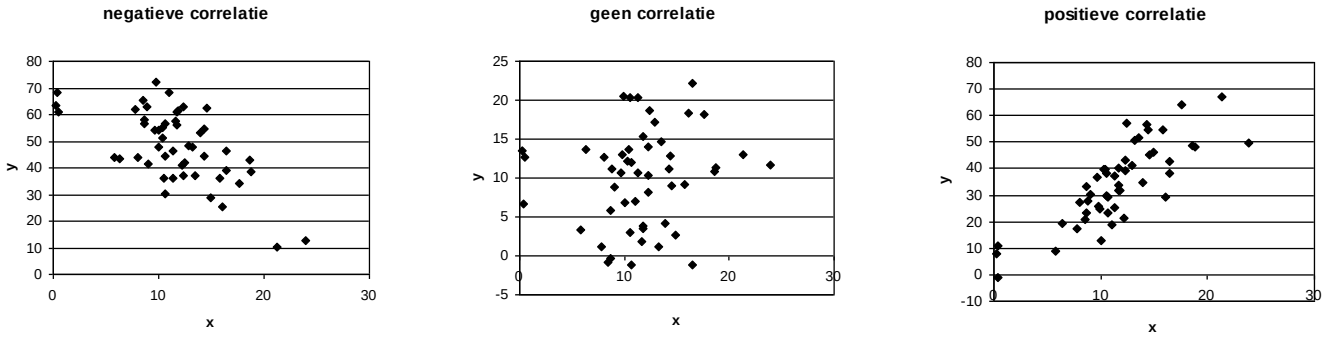
Tot nu toe zijn de gebruikte termen herhalingen uit wat je bij statistiek al hebt gehad. Ik neem echter aan dat het woord *steekproefcovariantie* bij een ieder onbekend is. De manier, waarop je deze berekent, is analoog aan de berekening van de steekproefvarianties, maar nu vermenigvuldig je de afwijking in de x-richting met die in de y-richting. Daar nu geen kwadraten worden genomen kan deze berekende grootheid zowel positieve als negatieve waarden aannemen.

Formule voor de steekproefcovariantie tussen x en y is: $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Lineaire regressie

-Het toetsen van samenhang tussen twee variabelen-

Met behulp van de steekproefcovariantie wordt een idee verkregen over de samenhang van de twee grootheden x en y. Een positieve covariantie wijst erop dat y groter wordt als x groter wordt, terwijl een negatieve covariantie erop wijst dat y juist kleiner wordt naarmate x groter waarden aanneemt. Omdat de berekende covariantie afhangt van de variantie in de x en y richting hebben statistici een correlatiecoëfficiënt bedacht, waarbij door de steekproefstandaard afwijking in de x- en y-richting wordt gedeeld.



Figuur 2: Uit plaatjes van de punten waarmee de regressie-analyse wordt gedaan kan al een indruk worden verkregen van het teken van de correlatiecoëfficiënt.

Formule voor de schatter van de correlatiecoëfficiënt: $r_{xy} = \frac{S_{xy}}{S_x \cdot S_y}$

De correlatiecoëfficiënt heeft dus hetzelfde teken (positief (+) of negatief (–)) als de covariantie (zie figuur 2 voor bijbehorende plaatjes). Als de correlatiecoëfficiënt klein is, dan is het waarschijnlijk dat x en y niet op een systematische manier tegelijk variëren.

De regressieanalyse en de toetsing

In een regressieanalyse bepaal je een lineair verband tussen y en x van de vorm $\bar{y} = a + bx$. Als de moederverdeling van de y variabele normaal verdeeld is, dan zijn de schatters voor de as-afsnede a en de helling b ook normaal verdeeld. De formules voor het bepalen van a en b zijn

$b = \frac{s_{xy}}{s_x^2}$ en $a = \bar{y} - b\bar{x}$.

Het doen van een regressie-analyse in deze lesbrief wordt behandeld als een recept in het computerprogramma Excel (zie blz. 10).

Nadat in Excel het recept is uitgevoerd, kan je toetsen of de richtingscoëfficiënt al of niet significant afwijkt van een veronderstelde waarde b_0. De te toetsen nulhypothese is dan $b=b_0$.(de richtingscoefficient wijkt niet af van b_0). Deze kan worden getoetst tegen een alternatieve hypothese $b \neq b_0$ (tweezijdig), $b < b_0$ of $b > b_0$ (beiden eenzijdig). Dit gebeurt met de toetsingsgrootheid $T_b = \frac{b - b_0}{s_b}$

Eveneens kan worden getoetst of de as-afsnede a significant afwijkt van een veronderstelde waarde a_0. De nulhypothese $a = a_0$ (de as-afsnede wijkt niet af van a_0)

wordt dan getoetst tegen een alternatieve hypothese $a \neq a_0$ (tweezijdig), $a < a_0$ of $a > a_0$ (beiden eenzijdig). Dit gebeurt met de toetsingsgrootheid $T_a = \frac{a - a_0}{s_a}$.

Voor verdere uitleg zie het stappenplan verderop.

Uitvoering

Gebruik van Excel

In veel spreadsheet programma's kunnen de berekeningen die voor een lineaire regressie noodzakelijk zijn eenvoudig worden uitgevoerd. Omdat ik er van uit ga dat op veel middelbare scholen het Microsoft Office pakket de standaard is op de computers, vertel ik hier hoe de berekeningen in Microsoft Excel kunnen worden uitgevoerd.

In dit spreadsheetprogramma zijn de kolommen (verticaal) gekenmerkt door letters,- in kolom A kunnen bijvoorbeeld een heleboel getallen ingevuld worden (zie onderstaande tabel). De rijen (horizontaal) worden aangeduid met cijfers. Zodoende is elke cel (zo'n rechthoekje) uniek vastgelegd door een letter-cijfercombinatie. In de onderstaande tabel staat het getal 94,52 in een notatie met decimale punt in plaats van decimale komma in de cel B9. De Engelstalige versie van Excel werkt namelijk met decimale punten en niet met komma's, terwijl de Nederlandstalige versie met decimale komma's werkt.

Ik ga er van uit dat in de eerste twee kolommen per kolom 10 getallen zijn ingevuld. In de eerste kolom staan dan 10 x-waarden en in de tweede kolom 10 y-waarden (zie onderstaande tabel voor een voorbeeld).

Kolom A	Kolom B
1	48.19
2	35.55
3	62.9
4	82.42
5	84.17
6	95.73
7	36.06
8	70.25
9	94.52
10	62.61

Selecteer een leeg stukje van het Excel werkblad met een breedte van 2 kolommen en een hoogte van 5 rijen. Type in de lege cel in de linksboven hoek van het geselecteerde stuk in de Engels-talige versie van Excel "= LINEST(B1:B10, A1:A10, 1, 1)" en in de Nederlands-talige versie van Excel "=LIJNSCHAT(B1:B10; A1:A10; 1; 1)". Om Excel de juiste berekeningen te laten uitvoeren moeten nu de toetsen Ctrl, Shift en Enter tegelijkertijd ingedrukt worden. Het resultaat van deze handeling is dat de 10 cellen worden gevuld met getallen (zie rechter tabel hieronder voor de uitvoer van de Engelstalige versie van Excel).

b	a
s_b	s_a
$(r_{xy})^2$	s
F	d.f.
ss_{reg}	ss_{res}

2.738182	52.18
2.413053	14.9726
0.138639	21.91764
1.287629	8
618.5553	3843.064

In de linkertabel hier juist boven staan op corresponderende plaatsen de symbolen weergegeven die aangeven wat het getal in de rechtertabel betekent; ik bespreek achtereenvolgens enkele (niet alle) symbolen, die voor onderzoek naar de samenhang tussen de grootheden x en y kunnen worden gebruikt. De in het onderhavige probleem berekende waarden staan er, tussen haakjes, achter gegeven:

b: Dit is de richtingscoëfficiënt van de geschatte lijn (2,74).
a: Dit is de as-afsnede van de geschatte lijn (52,18).

Dit houdt in dat de geschatte lijn gegeven wordt als y=52,18 + 2,74 x.

s_b: Dit is de standaardfout (de wortel uit de variantie) van b (2,41).
s_a: Dit is de standaardfout van a.(14,97).
d.f.: Dit is het aantal vrijheidsgraden (8); dit aantal is gelijk aan het aantal waarnemingen min het aantal geschatte parameters. De geschatte parameters bij een lineaire regressie zijn de twee getallen a en b. In het bovenstaande voorbeeld zijn tien punten met hun x en y waarden gegeven. Het aantal vrijheidsgraden is dan het aantal punten min het aantal geschatte parameters = 10 - 2 = 8.

De som van de kwadraten van het verschil tussen de gemeten y waarden en de gemiddelde y waarde is hier $ss_{tot} = \sum_{i=1}^{10}(y_i - \bar{y})^2 = 4461{,}619 = ss_{reg} + ss_{res}$. Dit wordt ook wel de kwadraatsom van de fouten genoemd (engels: sum of squared errors).

ss_{reg}: dit is het stukje van de totale kwadraatsom dat door de regressie wordt verklaard.
ss_{res}: dit is het stukje van de totale kwadraatsom dat niet door de regressie wordt verklaard, de onverklaarde of residuele kwadraatsom.

Naarmate er een groter deel van de variantie wordt verklaard (dat houdt in dat ss_{reg} groot is ten opzichte van $ss_{reg} + ss_{res}$) past de geschatte lijn beter bij de waarnemingen.

In EXCEL is het ook mogelijk om de as-afsnede op nul te zetten. In dat geval is de opdracht "= LINEST(B1:B10; A1:A10; 0; 1)" en wordt y = b x geschat. Vergeet niet Ctrl, Shift en Enter tegelijk in te drukken om de tabel van vijf rijen en twee kolommen gevuld te krijgen. Het zij opgemerkt dat dit consequenties heeft voor de toetsing (zie stappenplan).

Stappenplan lineaire regressie

Voor het systematisch uitwerken van een toetsingsprocedure voor de lineaire regressie is het volgende schema van toepassing:

1. Formuleer het probleem in woorden, bv.
   (a) Ik verwacht dat y groter is bij grotere x.
   (b) Ik verwacht dat er een niet gespecificeerde samenhang is tussen x en y.
   (c) Ik verwacht dat er een positieve asafsnede is bij de samenhang tussen x en y

2. Formuleer de nulhypothese en de alternatieve hypothese in woorden. Op grond van de nulhypothese en de alternatieve hypothese bepaal je of je eenzijdig of tweezijdig gaat toetsen.
   (a) nulhypothese: er is geen verband tussen y en x: b = 0; alternatieve hypothese: er is een positief verband tussen y en x: b > 0 (nu ga je eenzijdig toetsen)
   (b) nulhypothese: er is geen verband tussen y en x: b = 0; alternatieve hypothese: er is een verband tussen y en x: b ≠ 0 (nu ga je tweezijdig toetsen)
   (c) nulhypothese: de as-afsnede is nul: a = 0; alternatieve hypothese: er is een positieve as-afsnede: a > 0 (nu ga je eenzijdig toetsen)
3. Bepaal de toetsingsgrootheid T. Geef aan of je verwacht dat T grote of kleine waarden aanneemt als de alternatieve hypothese waar is. Bij een tweezijdige toets verwacht je dat T ofwel grotere ofwel kleinere waarden aanneemt onder de alternatieve hypothese. In dat laatste geval betekenen middelmatige waarden van T een ondersteuning van de nulhypothese.
4. Kies voor de onbetrouwbaarheidsdrempel α een waarde waarmee je de toets gaat uitvoeren (veelal 0,05 of 0,10).
5. Lees in de tabel aan het einde van deze les brief de kritieke waarde(n) af en bepaal het kritieke gebied. Hierbij is het aantal vrijheidsgraden (afgekort d.f. van "degrees of freedom") het aantal x-waarden –2 als je a en b schat en het aantal x-waarden –1 als je a op nul zet en b schat.
6. Voer de regressie-analyse met Excel uit.
7. Bepaal de waarde van de toetsingsgrootheid T, met behulp van de output van Excel.
8. Trek op een statistische verantwoorde manier een conclusie en vertel het resultaat vervolgens in je eigen woorden.

Als bovenstaande procedure stap voor stap wordt gevolgd kan voor elk probleem waarbij samenhang tussen twee grootheden wordt bekeken een verantwoorde conclusie worden getrokken.

Uitgewerkte voorbeelden

De invloed van bidtijd op de leeftijd waarop mensen overlijden

Recent Amerikaans onderzoek claimt dat de levensverwachting van mensen wordt verhoogd door regelmatig te bidden. In een studie in een Amerikaans dorp werd de gemiddelde leeftijd waarop mensen overleden via lineaire regressie gerelateerd aan de gemiddelde tijd die per dag aan bidden werd besteed. De resultaten zijn samengevat in tabel 2.

Bidtijd (x in min)	**Gem. leeftijd (y in jaren)**
1	65
5	69
10	75
30	81
45	83
60	81

Tabel 2: gegevens uit een Amerikaans dorp over dagelijkse tijd besteed aan bidden.

1. Naarmate mensen gemiddeld genomen meer bidden kunnen ze een hogere leeftijd bereiken.
2. nulhypothese: er is geen verband tussen y en x: b = 0; alternatieve hypothese: er is een positief verband tussen y en x: b > 0 (nu ga je eenzijdig toetsen)

3. $T_b = \frac{b-0}{s_b}$ T_b neemt grote waarden aan als de alternatieve hypothese waar is.
4. De onbetrouwbaarheidsdrempel α is 0,10.
5. Er zijn 6 punten en dus lees je af bij d.f. = 6 –2 =4 en α = 0,10: de afgelezen kritieke waarde tcrit = 1,533 en het kritieke gebied is $1{,}533 \leq T_b \leq \infty$
6. Voer de regressie-analyse met Excel uit. (zie output in tabel 3)
7. De waarde van de toetsingsgrootheid T_b = 0,266/ 0,077=3,455
8. T_b is groter dan de kritieke waarde en ligt dus in het kritieke gebied. De nulhypothese wordt dus verworpen. Elke minuut dat een persoon gemiddeld meer bidt levert een systematisch hogere leeftijd op waarop hij/zij sterft.

0,266	68,963
0,077	2,554
0,751	4,095
12,060	4
202,25	67,081

Tabel 3: output van Excel voor gegevens uit tabel 2

N.B. Het is niet zo dat deze conclusie de causaliteit die in het proces verborgen zit adequaat beschrijft. Het kan zo zijn dat de mensen die langer bidden ook gezondere leefgewoonten hebben die ervoor zorgen dat ze daadwerkelijk langer leven.

Met betrekking tot het sterfproces heeft men natuurlijk altijd in het achterhoofd dat ook al bidt iemand nooit hij/zij toch zo'n 70 jaar oud kan worden. Om deze veronderstelling te toetsen kan je met dezelfde gegevens ook met lineaire regressie aan de slag, ook al heb je geen waarnemingen bij bidtijden van gemiddeld nul minuten. Hierbij dient wel de opmerking gemaakt te worden dat dit niet helemaal zuiver is, omdat je buiten het interval van waargenomen bidtijden kijkt.

Mensen die (bijna) nooit bidden kunnen gemiddeld een leeftijd van 70 jaar bereiken.

1. nulhypothese: mensen die nooit bidden worden gemiddeld 70 jaar (dit is de waarde van de as-afsnede): a = 70; alternatieve hypothese: mensen die nooit bidden worden niet gemiddeld 70 jaar: $a \neq 0$ (nu ga je tweezijdig toetsen)
2. $T_a = \frac{a-70}{s_a}$ T_a neemt relatief grotere of relatief kleinere waarden aan als de alternatieve hypothese waar is.
3. De onbetrouwbaarheidsdrempel α is 0,10.
4. Er zijn 6 punten en dus lees je af bij d.f. = 6 –2 =4 en α = 0,05, omdat het om de eenzijdige onbetrouwbaarheid draait in de tabel: $t_{crit,rechts}$ = 2,132 en het kritieke gebied bestaande uit twee delen is $-\infty \leq T_a \leq -2{,}132$ en $2{,}132 \leq T_a \leq \infty$
5. Voer de regressie-analyse met Excel uit. (zie output in tabel 3)
6. De waarde van de toetsingsgrootheid T_a = (68,963 – 70)/2,554=-0,406;
7. T_a ligt niet in het kritieke gebied en de nulhypothese wordt dus niet verworpen. De gemiddelde leeftijd die mensen die niet bidden kunnen bereiken wijkt niet aantoonbaar af van 70 jaar.

Opdrachten

De invoering van de euro in Nederland

Een consumentenorganisatie wil onderzoeken of een bepaalde supermarktketen het "omprijzen" naar euro's aangrijpt om de prijzen van voedingsmiddelen te verhogen. Alle prijzen in de tabel zijn gegeven in guldens. In tabel 4 staan de prijzen vóór en na de invoering van de euro. Wanneer de prijzen niet veranderd zijn dan wijkt de regressielijn niet af van de 45 graden lijn y=x. Onderzoek met behulp van Excel en het stappenplan of dit het geval is (neem $\alpha = 0{,}05$).

Prijs vóór (x)	Prijs na (y)
1,19	1,29
2,25	2,18
3,69	3,72
4,89	4,96
0,89	0,99
2,39	2,53
1,49	1,52
3,59	3,72
1,98	1,96

Tabel 4: prijzen in guldens voor en na invoering van de euro.

Samenhang tussen buitentemperatuur en hoeveelheid verkocht ijs

Temperatuur (x in C)	Verkocht ijs (y in l)
21,3	57,8
28,8	74,7
24,6	64,8
22,4	59,9
29,1	75,1
25,8	68,3
27,9	71,2
23,0	59,9

Tabel 5: temperatuur met het verkochte ijs in liters.

Een ijsventer wenst de dagelijkse hoeveelheid ijs die hij op een zomerse weekdag verkoopt te relateren aan de buitentemperatuur die hij afleest bij vertrek uit zijn huis. Hij verwacht dat elke graad hoger die hij afleest leidt tot de verkoop van 2 extra liter ijs. De verkochte hoeveelheid ijs in liters en de afgelezen buitentemperatuur op 8 dagen zijn gegeven in tabel 5. Onderzoek met behulp van Excel en het stappenplan of de verwachting van de ijsboer ook ondersteund wordt door de gegevens (neem $\alpha = 0{,}10$).

Suggesties voor verder onderzoek

Documentatie

Bolle, E.A.W, J.H.M. Lenoir en J.N.M. van Loon (1974) Statistiek: wiskundige statistiek. Kluwer, Deventer, pp. 248

Groen, W.E., A.J. Hakkert en W.H.H. van der Maaten (1988) Keuze onderwerp wiskunde A: Correlatie en regressie, Wolters- Noordhoff, Groningen, pp. 47

Oriëntatie op vervolgonderwijs

Het onderwerp van deze lesmodule kom je ook tegen bij de meeste opleidingen van Wageningen University. Kijk voor meer informatie op www.wageningenuniversity.nl.

Auteur: Lia Hemerik, Biometris, leerstoelgroep Wiskundige en Statistische Methoden