Hoofdstuk 10: Regressie

Inleiding
In dit deel zal uitgelegd worden hoe we statistische berekeningen kunnen maken als sprake is van één kwantitatieve responsvariabele en één kwantitatieve verklarende variabele. We gaan hierbij uit van de regressielijn $\hat{y}= b_0+ b_1x$, zoals in hoofdstuk 2 besproken is. In dit hoofdstuk proberen we echter uit te zoeken in hoeverre een berekende regressielijn een schatting is van de *ware* regressielijn die bij de populatie hoort. De regressielijn die bij de populatie hoort noteren we als $\beta_0+ \beta_1x$. In deze formule staat β_0 voor het intercept en β_1 voor de regressiecoëfficiënt. Deze waarden worden geschat aan de hand van b_0 en b_1.

10.1 Simpele lineaire regressie

Populaties
Simpele lineaire regressie wordt gebruikt om de relatie tussen een responsvariabele (y) en een verklarende variabele (x) te onderzoeken. We verwachten dat verschillende waarden van x samen zullen gaan met verschillende waarden van y. Stel: we willen de verandering in bloeddruk vastleggen voor twee experimentele groepen. De ene groep krijgt een echt medicijn en de andere groep krijgt een placebo. De behandeling (placebo of echt medicijn) kunnen we dan zien als een verklarende variabele en bloeddruk is dan de responsvariabele.

- De gemiddelde verandering in bloeddruk kan verschillend zijn in de twee populaties. Deze gemiddelden noemen we μ_1 en μ_2.
- Individuele veranderingen in bloeddruk variëren binnen elke populatie volgens de normaalverdeling. Dit betekent dat de meeste mensen binnen een groep ongeveer dezelfde bloeddruk hebben, terwijl een beperkt aantal mensen extreem afwijkt van de rest. Er wordt vanuit gegaan dat de standaarddeviaties van de populaties gelijk zijn.

Subpopulaties
Bij lineaire regressie kan de verklarende variabele (x) veel verschillende waarden aannemen. Je kunt bijvoorbeeld verschillende hoeveelheden van calcium geven aan verschillende groepen deelnemers. Deze waarden van x kunnen we zien als *subpopulaties:*

- Elke waarde van x gaat samen met één subpopulatie. Elke subpopulatie bestaat uit alle individuen in de populatie die dezelfde waarde van x hebben. Als we dus een experiment uitvoeren waarbij we de effecten van vijf verschillende hoeveelheden calcium op bloeddruk willen onderzoeken, dan bestuderen we vijf subpopulaties.

Het statistische model voor simpele lineaire regressie gaat er vanuit dat voor elke waarde van x de geobserveerde waarden van y normaal verdeeld zijn met een gemiddelde dat van x afhangt. We gebruiken het symbool μ_y om deze gemiddelden aan te geven. De gemiddelden μ_y kunnen veranderen als x volgens een vast patroon verandert. Bij simpele lineaire regressie gaan we er vanuit dat alle gemiddelden op een lijn liggen die gebaseerd is op x-waarden. Kort samengevat is er bij simpele lineaire regressie sprake van:

- Verandering van de gemiddelden van y wanneer x verandert. Alle gemiddelden liggen op een lijn. Daarom geldt: $\mu_y= \beta_0+ \beta_1x$. Dit is de *regressielijn van de populatie.*
- Individuele waarden van y (op basis van dezelfde x) variëren volgens de normaalverdeling. Deze normaalverdelingen hebben allemaal dezelfde standaarddeviatie.

Residuen
De regressielijn die we vinden is nooit perfect als het gaat om het voorspellen van y-waarden op basis van x-waarden. Daarom geldt:

- Data = fit+residu.
- Het fit-gedeelte bestaat uit de subpopulatie-gemiddelden die gevonden worden door middel van $\mu_y = \beta_0 + \beta_1 x$.
- Het residu-gedeelte staat voor de afwijkingen van de data vanaf de lijn die staat voor de populatiegemiddelden. We gaan ervan uit dat deze afwijkingen normaalverdeeld zijn en standaarddeviatie σ hebben. We gebruiken de Griekse letter ε als we het over het residu-gedeelte hebben. De ε-waarden kunnen gezien worden als 'ruis': het deel van de data dat niet verklaard kan worden met de regressielijn. Hierdoor zullen punten in een puntenwolk nooit helemaal op een rechte lijn liggen.

Het model voor simpele lineaire regressie
Het model voor simpele lineaire regressie gaat gepaard met de volgende feiten:

- Gegeven n aantal observaties van x en y, geldt: (x_1,y_1), (x_2,y_2),…, (x_n,y_n).
- De geobserveerde respons (y_i) gaat samen met verklaarde en onverklaarde elementen: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. In deze formule is $\beta_0 + \beta_1 x_i$ de gemiddelde respons wanneer $x=x_i$. De afwijkingen (ε_i) zijn onafhankelijk en normaalverdeeld. Ze hebben een gemiddelde van 0 en standaarddeviatie σ. De parameters van het model zijn dus: β_0, β_1 en σ.

Regressieparameters schatten
Zoals eerder gezegd willen we de regressielijn die we op basis van een steekproef gevonden hebben, gebruiken om een regressielijn te maken voor de populatie. De regressielijn voor een steekproef is: $\hat{y} = b_0 + b_1 x$. In deel B was al uitgelegd hoe b_0 en b_1 gevonden kunnen worden:

- $b_1 = r(s_y / s_x)$. In deze formule staat r voor de correlatie tussen y en x. De rest van de formule maakt gebruik van de standaarddeviaties van y en x.
- b_0= : y-streepje - $b_1\bar{x}$.
- Het residu is: e_i =(geobserveerde y-waarde) – (voorspelde y-waarde). Dit is hetzelfde als: $y_i - \hat{y}_i$. Dit is weer hetzelfde als: $y_i - b_0 - b_1 x_i$. De residuen (e_i) corresponderen met de residuen ε_i. De e_i tellen op tot 0 en de ε_i komen uit een populatie met een gemiddelde van 0.

Dan moet nog de laatste parameter (σ) berekend worden. Deze parameter meet in hoeverre y-waarden van de populatie-regressielijn *afwijken.* Om deze parameter te berekenen, maken we daarom gebruik van residuen.

- Eerst berekenen we de variantie van de regressielijn die bij de populatie hoort (σ^2). Dit doen we door de variantie van de steekproef te gebruiken: $s^2 = (\Sigma e^2_i)/ n-2$. Dit is hetzelfde als: $\Sigma(y_i - \hat{y}_i)^2/n-2$.
- Vervolgens trekken we de wortel uit de variantie (s^2) om σ te vinden.

Betrouwbaarheidsintervallen
Betrouwbaarheidsintervallen kunnen in het algemeen gevonden worden middels de formule: schatting± $t^*SE_{schatting}$. Voor β_0 en β_1 kunnen afzonderlijk betrouwbaarheidsintervallen berekend worden:

- Het betrouwbaarheidsinterval voor het intercept β_0 is: $b_0 \pm t^* SEb_0$.
- Het betrouwbaarheidsinterval voor de regressiecoëfficiënt β_1 is: $b_1 \pm t^* SEb_1$.
- In deze formules is t* de waarde voor t(n-2) met gebied C tussen –t* en t*.

Significantietoetsen
De nulhypothese stelt dat de regressiecoëfficiënt in de populatie 0 is ($\beta_1 = 0$). Om deze hypothese te toetsen maken we gebruik van een *toetsstatistiek:*

- $t = b_1/SEb_1$. De vrijheidsgraden zijn n-2. De nulhypothese kan zowel eenzijdig als tweezijdig getoetst worden.
- Als er tweezijdig getoetst wordt, moet de p-waarde uit de t-tabel eerst vermenigvuldigd worden om een conclusie te trekken over de nulhypothese. Als blijkt dat de alternatieve hypothese aangenomen moet worden, dan betekent dit dat er een relatie bestaat tussen x en y in de populatie. Let op: een hele kleine p-waarde zegt bij deze significantietoets niet dat we een sterke relatie hebben gevonden tussen x en y. Er mag dan alleen geconcludeerd worden dat er sprake is van een relatie, maar de grootte van de relatie is niet duidelijk.

Betrouwbaarheidsintervallen voor de gemiddelde respons
Voor elke waarde van x (ook wel x* genoemd) is de gemiddelde y-waarde in de subpopulatie:

- $\mu_y = b_0 + b_1 x^*$.
- Het bijbehorende betrouwbaarheidsinterval voor de gemiddelde respons is: $\mu_y \pm t^* SE_u$. In deze formule is t* de waarde voor t(n-2) met gebied C tussen –t* en t*.

Voorspellingsintervallen
Soms willen we een waarde van een y voorspellen die ver buiten de y-waarden in de data ligt. In dat geval maken we gebruik van een voorspellingsinterval. Eerst moet een steekproef van n aantal observaties getrokken worden. Vervolgens moet het 95% betrouwbaarheidsinterval berekend worden voor een specifieke x-waarde (x*).

- Het *voorspellingsinterval voor een toekomstige observatie* van y uit de subpopulatie van x* is: $\hat{y} \pm t^* SE_{y\text{-dakje}}$. In deze formule staat t* voor de waarde van t(n-2) met gebied C tussen –t* en t*.

10.2 Meer over simpele lineaire regressie

Analyse van variantie (ANOVA) voor regressie
Door middel van *analyse van variantie* (ANOVA) kunnen we uitzoeken in hoeverre data verklaard kan worden door het deel dat past bij de regressielijn (fit) en het deel dat daarvan afwijkt (residuen). De totale variatie in y wordt uitgedrukt door de afwijkingen y_i – y-streepje. Als deze afwijkingen allemaal 0 zouden zijn, dan zouden alle observaties gelijk zijn en zou er geen variatie in y zijn. Er zijn twee redenen waarom y_i niet gelijk is aan het gemiddelde van y:

- De waarden van y_i gaan samen met verschillende waarden van x en zijn daarom verschillend.
- Individuele observaties zullen van het gemiddelde verschillen, omdat er variatie is *binnen* de subpopulatie die bij een specifieke x-waarde hoort.

Het model
Zoals eerder gezegd maken we bij lineaire regressie gebruik van het model data= fit + residuen. Als we hier in termen van variantie naar gaan kijken, dan wordt dit:

- SST = SSM + SSE. Hierbij staat SST voor de totale variantie, SSM voor de variantie die door het model wordt verklaard en SSE voor de variantie die niet door het model wordt verklaard (error). SS staat voor *´sum of squares´.*
- SST wordt berekend met de formule: $\Sigma(y_i\text{-y-streepje})^2$.
- SSM wordt berekend met de formule: $\Sigma(\hat{y}_i\text{-y-streepje})^2$.
- SSE wordt berekend met de formule: $\Sigma(y_i - \hat{y}_i)^2$.

Vrijheidsgraden en MS *(mean square)*
Daarnaast is het ook mogelijk om voor elke bron van variantie de bijbehorende vrijheidsgraden uit te rekenen. Er wordt uitgegaan van een soortgelijke formule: DFT = DFM + DFE. In deze formule staat DF voor *vrijheidsgraden (degrees of freedom)*. De vrijheidsgraden die bij het totaal, het model en de error horen, worden als volgt berekend:

- DFT = n-1.
- DFM = 1
- DFE = n-2.

We vinden de MS voor elke bron van variantie door de SS te delen door de bijbehorende vrijheidsgraden (DF). Als de MS voor het totaal gevonden moet worden, dan wordt dat dus gedaan door SST/ DFT te berekenen. De proportie verklaarde variantie (r^2) kan als volgt gevonden worden:

- SSM/SST. Het resultaat laat ons zien hoeveel van de variantie in y wordt verklaard door het model.

De ANOVA-toets
De nulhypothese dat de regressiecoëfficiënt (β_1) van de populatie 0 is, kunnen we aan de hand van de *F-toets* toetsen. De nulhypothese zegt dus eigenlijk dat x en y in de populatie geen lineaire samenhang vertonen. De F-toets vinden we als volgt:

- F= MSM/MSE.

Als de nulhypothese waar is, dan heeft deze F-toets een distributie van 1 vrijheidsgraad in de noemer en n-2 vrijheidsgraden in de teller: F(1,n-2). Deze vrijheidsgraden horen bij MSM en MSE. Net zoals er veel t-toetsen bestaan, zijn er ook veel F-toetsen. Als de regressiecoëfficiënt niet 0 is ($\beta_1 \neq 0$), dan is MSM relatief groot ten opzichte van MSE. Dit betekent dat grote waarden van F bewijs tegen de nulhypothese geven. We toetsen in dit verband altijd tweezijdig.

Schema
De informatie die tot nu toe gegeven is, wordt kort in de onderstaande *ANOVA-tabel* samengevat:

BRON (source)	Vrijheidsgraden (DF)	SS (sum of squares)	MS (mean square)	F
Model	1	$\Sigma(\hat{y}_i$-y-streepje$)^2$	SSM/DFM	MSM/MSE
Error	n-2	$\Sigma(y_i - \hat{y}_i)^2$	SSE/DFE	
Totaal	n-1	$\Sigma(y_i$-y-streepje$)^2$	SST/DFT	

Correlatie binnen de populatie toetsen

We kunnen ook toetsen of er een correlatie tussen twee variabelen in de populatie bestaat. We gebruiken de Griekse letter ρ om de *populatiecorrelatie* weer te geven. Als x en y beide normaalverdeeld zijn, dan is ρ=0 hetzelfde als zeggen dat x en y in de populatie onafhankelijk zijn. Dit betekent dat er geen enkele relatie tussen x en y in de populatie bestaat. De alternatieve hypothese kan zowel eenzijdig als tweezijdig geformuleerd worden. Om de hypothese ρ=0 te toetsen, maken we gebruik van de volgende stappen om de t-toets te berekenen:

- Eerst vermenigvuldigen we de correlatie (r) met de wortel uit n-2. In deze formule staat n voor de grootte van de steekproef.
- Vervolgens delen we dit getal door de wortel uit 1- r^2.

De gevonden t-toets is hetzelfde als de t-toets die we zouden vinden als we de hypothese $\beta_1=0$ hadden getoetst. Dit betekent dat als er geen correlatie in de populatie bestaat, dat de regressiecoëfficiënt 0 is.