# Hoofdstuk 6

## 6.1

Hoewel lineaire regressies gewoonlijk gebruikt worden, zijn veel relaties nu eenmaal nier lineair. In deze situaties zou een lineaire regressie aangeven dat er geen relatie tussen de variabelen is, terwijl het simpelweg een non-lineaire relatie is. Hierdoor is het belangrijk om ook non-lineaire regressie te begrijpen.

Ten eerste kunnen we evalueren of een relatie lineair is of niet door naar een scatter-plot te kijken. Hierna moeten we het juiste model vinden. Non-lineaire regressie heeft natuurlijk een andere soort formule nodig. De normale regressie formule (van een rechte lijn; y = b1x1 + c) moet getransformeerd worden om gekromd te worden. Dit wordt gedaan door de verklarende variabele te kwadrateren (x  x2; alle waarden van de verklarende variabelen worden dan ook gekwadrateerd: 4 1 6, 12  144, etc.). De formule die we dan krijgen is als volgt: y = b1x1 + b2x12 + c.

Het kwadrateren van de verklarende variabele (x2) creëert een kwadratische functie, die kromlijnig is maar alleen 1 kromming heeft. Voor sommige modellen is het nodig dat de lijn meerdere krommingen heeft. Een functie waarbij de lijn 2 krommingen heeft is de derdegraads functie, en gaat als volgt: y = b1x1 + b2x12 + b3x13 + c.

Wees er bij deze formules van bewust dat er maar 1 verklarende variabele genoemd wordt. Deze variabele wordt simpelweg meerdere keren gebruikt in de formule, met verschillende b-coëfficiënten.

Om een lijn met een afnemend stijgende lijn te krijgen, moet je gebruik maken van de logaritmische functie. Een logaritme (log) wordt meestal gebruik met een base van 10. Dit betekent dat de log10 van een nummer het nummer is dat tot de macht van 10 het originele nummer wordt (dus log10(a) = b, omdat a = 10b). Omdat zo'n functie als gevolg heeft dat de verschillen tussen grotere waardes kleiner wordt dan de verschillen tussen kleine waardes, creëert het een kromme die geleidelijk af vlakt terwijl hij stijgt.

Ten slotte is er de inverse functie, die als volgt berekent wordt: y = 1/x. De inverse functie maakt grote waardes klein, en kleine waardes groot.

Dingen om in gedachten te houden tijdens het uitvoeren van een non-lineaire regressie:

- Je kan een constante toevoegen, zolang als deze toegevoegd wordt aan elke waarde van de variabele, zodat het de interpretatie van de resultaten niet beïnvloed. Hetzelfde geldt voor het vermenigvuldigen van de variabelen. Zo lang als zulke veranderingen consistent worden doorgevoerd, is er geen verandering in de gestandaardiseerde schattingen, de p-waardes, en dus de conclusies.
- Transformaties hebben effect op de vorm van de functie, maar ook op de distributie van de variabele. Dit betekent dat een transformatie ook het gevolg kan hebben dat een variabele niet meer een normaal distributie heeft. Het zou er zelfs voor kunnen zorgen dat er uitschieters ontstaan.
- De waardes van de verklarende variabele zelf hebben ook effect op de kromme van de functie.

- Niet alle transformaties zijn mogelijk. Een log van 0 heeft, bijvoorbeeld, geen waarde. Er moet dus soms een constante toegevoegd worden om ervoor te zorgen dat de serie waardes zich goed gedraagt.
- Als een lineair model en een non-lineair model beide de data even goed beschrijven, is het lineaire model altijd de beste optie omdat het simpeler is.
- Gebruik nooit stapsgewijze regressie als er non-lineaire variabelen in een model zitten, omdat zulke technieken een voorkeur hebben voor non-lineaire variabelen en zo het model onnodig complex kunnen maken.

## 6.2

Regressie analyse werkt best met een responsvariabele die een doorlopende serie aan waardes heeft. Maar er zijn ook genoeg situaties waarin een categorische responsvariabele (met data die bijvoorbeeld alleen bestaat uit “ja” en “nee”) geanalyseerd moet worden. Om dit te doem moeten de standaard regressie technieken wat aangepast worden. De meeste uitkomsten van een regressie analyse, zoals hellingen en distributies, zijn namelijk nonsens als er maar twee situaties mogelijk zijn volgens de data.

Om zulke data te analyseren moet de data zelf verandert worden. Dit is dus anders dan eerdere transformaties, waar veranderingen altijd op de verklarende variabele werden toegepast. De transformatie die hier nodig is heet de ‘logit’ tranformatie, wat betekent dat de hierop gebaseerde regressie analyse een logistische regressie heet.

Één manier om de data te transformeren is door gebruik te maken van kans (‘probabilities’) i.p.v. de categorische variabele. Bijvoorbeeld, laten we zeggen dat we meerdere groepen aan het analyseren zijn waarin de responsvariabele het hebben van een huisdier is. De categorische variabele zou een 1 geven voor het hebben van een huisdier, en een 0 voor het niet hebben van een huisdier, wat leidt tot een totaal nummer per groep van hoeveel mensen in die groep een huisdier hebben. Het gebruik van kans zou er voor zorgen dat iedere groep een nummer tussen 0 en 1 krijgt dat aangeeft wat de kans is dat iemand in die groep een huisdier heeft (door het delen van de hoeveelheid mensen met huisdieren door de hoeveelheid mensen in de groep). Op deze manier wordt er een continuüm gecreëerd.

Dit is echter maar een gelimiteerd continuüm, omdat de nummers niet hoger dan 1 of lager dan 0 kunnen zijn. Een regressie analyse zou leiden tot nummers buiten dit gebied, nummers die dan dus betekenisloos zijn. Dus, kans is hier niet geheel de oplossing. Laten we daarom kijken naar ‘odds’. Hierbij maken we gebruik van de ‘odds ratio’, die de waarschijnlijkheid van een gebeurtenis aangeeft.

Hoewel het gebruik van ‘odds’ betekent dat geen waarde voorspeld door een regressie buiten bereik is, zijn waardes onder 0 hiermee nog steeds onmogelijk. We moeten dus nog één laatste stap maken: het berekenen van de ‘logit’. De ‘logit’ is simpelweg het natuurlijk logaritme (log) van de ‘odds ratio’. Als de ‘logit’ gebruikt wordt als waardes voor de variabele is het mogelijk om waardes onder de 0 te hebben.

Deze drie vormen van waardes zijn dus aan elkaar verbonden. Met één hiervan kunnen de anderen altijd berekend worden:

Probability ↔ Odds ↔ Logit

Lineaire regressie maakt gebruik van ‘ordinary least squares’ (OLS) regressie. Dit is een regressie techniek die de lijn vindt die het best bij de data past door de restwaarden (het verschil tussen de data punten en de regressie lijn) te kwadrateren (om zo geen negatieve waardes te krijgen) en deze bij elkaar op te tellen. De lijn waarbij het hieruit verkregen nummer het laagst is, is de lijn die best bij de data past.

Logistische regressie kan echter geen gebruik maken van OLS, vanwege de soort vergelijkingen die in deze vorm van regressie aanwezig zijn. In plaats hiervan wordt er gebruik gemaakt van ‘maximum likelihood’ (ML). Door dit gebruik van ML krijgt logistische regressie een iets andere output dan dat van OLS regressie, maar gelukkig niet te anders.

ML schatting maakt gebruik van de ‘log likelihood (LL) function’. Dit is een functie die aangeeft hoe goed de parameter schattingen bij de daadwerkelijke data passen. Aan de hand van deze uitkomst worden de parameter schattingen aangepast en wordt de LL opnieuw berekend. Deze aanpassingen gaan door tot dat er geen verbeteringen meer gemaakt kunnen worden.

Het kan over het algemeen gezegd worden dat hoe groter de LL is, hoe beter het model bij de data zal passen.

Zoals eerder besproken is het mogelijk om hiërarchisch de variabelen in een regressie te evalueren. Eerder werd genoemd hoe je dit in een OLS regressie kan doen, maar dit si natuurlijk anders in een logistische regressie. Gelukkig voert SPSS automatisch de berekeningen uit om verbeteringen te evalueren wanneer een variabele aan een model wordt toegevoegd.