KM 11(1983)
pag 31 - 39

Schatting van de parameters van het gegeneraliseerde logistische tijdreeksmodel.

G.H. Maassen*

Samenvatting:
De toepassingsmogelijkheden van het logistische tijdreeksmodel worden op welkome wijze verruimd, indien men het generaliseert tot $Y = C + M/(1 + K\exp(-Dt))$. Voor twee situaties worden beknopt methoden beschreven om de parameters te schatten: (1) de data volgen (bijna) exact een logistische curve en (2) de data worden gekenmerkt door onregelmatigheden. Voor de laatste (meest interessante) categorie heeft Levenbach een procedure voorgesteld die goed voldoet voor de bepaling van C,D en M. Aangetoond wordt dat zijn methode om K te bepalen sterk moet worden afgeraden.

De logistische groeicurve is zeer bekend geworden als model voor situaties, waarin sprake is van een fenomeen waarvan de omvang volgens een S-vormige curve groeit naar een zeker verzadigingsnivo. Talloos zijn de toepassingen, in allerlei gebieden van wetenschap.
De toepassingsmogelijkheden van de logistische groeicurve, die men algemeen met de volgende functie kan beschrijven

$$Y(t) = \frac{M}{1 + Ke^{-Dt}} \qquad (D,K \text{ en } M > 0),$$

worden echter nog aanmerkelijk verruimd, indien men de gedaante als volgt generaliseert:

$$(1) \qquad Y(t) = C + \frac{M}{1 + Ke^{-Dt}},$$

Voor een beter begrip van de eigenschappen van (1) en van de extra mogelijkheden ten opzichte van het traditionele logistische model, willen we met betrekking tot de parameters enkele opmerkingen maken.

- C is het asymptotisch startnivo voor de curve. Dat betekent dat nu ook verschijnselen gemodelleerd kunnen worden, waarvan de omvang het nulnivo niet dicht benadert, naarmate men verder in de tijd teruggaat.
- Het praktisch startnivo van de curve is gelijk aan: $C + M/(1 + K\exp(-Dt_0))$. Als men toestaat dat C negatief is, kunnen nu ook verschijnselen worden

---

* Instituut voor Pedagogische en Andragogische Wetenschappen
Heidelberglaan 1, 3584 CS Utrecht.

gemodelleerd die starten vanaf een nulpunt.

- De curve groeit volgens een S-vormig patroon asymptotisch naar een verzadigingsnivo C + M.
- De specifieke S-vorm van dit model wordt vastgelegd door de volgende relatie, waaraan curven van het type (1) voldoen:

$$(2) \qquad \frac{Y'(t)}{Y(t) - C} = D. \frac{M + C - Y(t)}{M} .$$

Dus op elk tijdstip is de relatieve groeisnelheid (gerelateerd aan de groei die op dat tijdstip is verwezenlijkt) evenredig aan de relatieve groeipotentie op dat moment. D is dan de evenredigheidsconstante. Uit de relatie zien we onder meer dat het gaat om curven die langzaam groeien als Y(t) de waarde C benadert of als Y(t) bijna het verzadigingsnivo bereikt heeft.

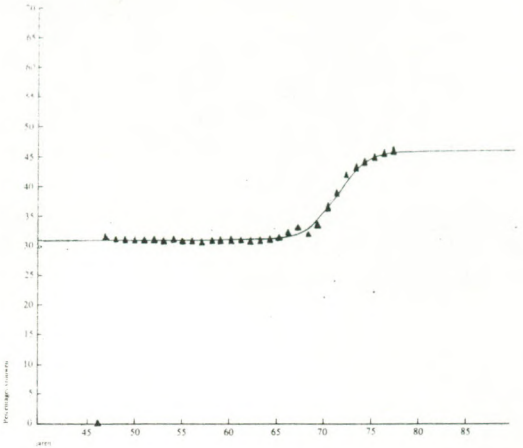
- De *vorm* van de curve wordt geheel bepaald door de waarden die C,D en M aannemen. K is daarvoor irrelevant. De waarde van K geeft informatie over de *plaats* van de curve ten opzichte van de tijdas (invullen van een andere waarde voor K betekent een verschuiving van de curve evenwijdig aan de tijdas). Belangrijk is vast te stellen, dat volgens (1) de functiewaarde van het punt met abcis ln K/D gelijk is aan C + M/2.
  Dus ln K/D geeft de plaats aan van het buigpunt van de curve, het punt waar de curve een maximale groeisnelheid heeft.

Een toepassing van het gegeneraliseerde model vindt men in figuur 1, ontleend aan Hoksbergen en Maassen (1980)

In dit artikeltje besteden we aandacht aan de bepaling van de parameters C, D, M en K van het gegeneraliseerde logistische model.
Stel dat we beschikken over de data P_i van n opeenvolgende equidistante meetmomenten t_0 , t_1 , ..., t_{n-1}. Zonder verlies van algemeenheid kunnen de parameters D en K nu zo getransformeerd worden dat voor de corresponderende modelpunten Y_i geldt:

$$Y_i = C + \frac{M}{1 + Ke^{-Di}} \qquad (i = 0, 1, 2, \ldots, n-1)$$

Figuur 1.



Participatie (procentueel) van meisjes aan het VHMO/VWO sinds 1945.
Als het nulpunt van de tijdschaal op 1900 wordt gesteld heeft de curve de volgende functionele vorm:

$$0{,}309 + \frac{0{,}151}{1 + 1{,}157.10^{20}.e^{-0{,}651\ t}}$$

Laten we ons eerst richten op de ideale situatie waarin bekend is dat de datapunten exact aan een logistisch model voldoen (dus $P_i = Y_i$ voor alle i).
In deze situatie kunnen C,D,M en K worden berekend uit een willekeurig viertal equidistante datapunten. We zullen dit voor het gemak illustreren met behulp van P_0, P_1, P_2 en P_3.
Er geldt:

$$\frac{\dfrac{1}{P_2 - C} - \dfrac{1}{P_1 - C}}{\dfrac{1}{P_1 - C} - \dfrac{1}{P_0 - C}} = \frac{(1 + Ke^{-2D}) - (1 + Ke^{-D})}{(1 + Ke^{-D}) - (1 + K)} = e^{-D}.$$

ofwel:

$$(3a) \qquad \frac{(P_2 - P_1)(P_0 - C)}{(P_1 - P_0)(P_2 - C)} = e^{-D}.$$

Evenzo geldt:

$$(3b) \qquad \frac{(P_3 - P_2)\,(P_1 - C)}{(P_2 - P_1)\,(P_3 - C)} = e^{-D}.$$

Gelijkstellen van de linkerleden van (3a) en (3b) leidt tot een vierkantsvergelijking waaruit C kan worden opgelost:

$$(4) \qquad (R_1 - R_2)\,C^2 + \{R_2(P_1 + P_2) - R_1(P_0 + P_3)\}\,C + R_1 P_0 P_3 - R_2 P_1 P_2 = 0,$$

waarin $R_1 = \dfrac{P_2 - P_1}{P_1 - P_0}$ en $R_2 = \dfrac{P_3 - P_2}{P_2 - P_1}$.

Om de notaties in het vervolg te vereenvoudigen zullen we, nu we C kennen, de data als volgt transformeren: $P_i' = P_i - C$
De waarde voor parameter D kan nu uit een drietal punten worden berekend. Met behulp van bijv. (3a) vinden we:

$$(5) \qquad D = \ln \frac{(P_1' - P_0')\,P_2'}{(P_2' - P_1')\,P_0'} = \ln \frac{P_2'}{R_1 P_0'}.$$

Voor de bepaling van M kan worden uitgegaan van 2 punten, bijvoorbeeld:

$$\frac{M}{P_0'} - 1 = K \qquad \text{en} \qquad \frac{M}{P_1'} - 1 = Ke^{-D}.$$

Na eliminatie van K ontstaat:

$$(6) \qquad M = \frac{P_0'\,P_1'\,(1 - e^{D})}{P_1' - P_0'\,e^{D}} = \frac{R_1 P_0' - P_2'}{R_1 - P_2'/P_1'}.$$

De waarde voor parameter K kan uit elk willekeurig punt worden berekend:

$$(7) \qquad K = e^{Di}\,\frac{M - P_i'}{P_i'}.$$

Gewoonlijk bevatten data, waaraan men een logistische curve wil aanpassen, onregelmatigheden. De resultaten van bovenstaande methode worden dan afhankelijk van het viertal punten dat men kiest. Bovendien kunnen kleine onregelmatigheden grote ontsporingen in de oplossing tot gevolg hebben. Voor data uit de praktijk is de methode dan ook in het algemeen niet geschikt. Levenbach (1973) geeft een methode die wel geschikt is voor situaties, waarin men een curve van de vorm (1) aan zijn data wil aanpassen. Zijn methode maakt op een elegante wijze gebruik van het algoritme ter bepaling van de coëfficiënten in een lineaire multipele regressievergelijking, zoals uit onderstaande korte beschrijving mag blijken. Uitgangspunt is de vergelijking (2), die men ook in de volgende vorm kan gieten:

$$Y'(t) = \alpha\, Y(t) + \beta\, Y^2(t) + \gamma \ ,$$

ofwel:

$$R(t) = \alpha + \beta\, Y(t) + \gamma/Y(t) \ . \tag{8}$$

Hierin is dan: $R(t) = \dfrac{Y'(t)}{Y(t)}$ de relatieve groeisnelheid,

$$C = (\sqrt{\alpha^2 - 4\beta\gamma} - \alpha)/2\beta \ , \tag{9a}$$

$$D = \sqrt{\alpha^2 - 4\beta\gamma} \ , \tag{9b}$$

$$M = -\sqrt{\alpha^2 - 4\beta\gamma}\ /\beta \ . \tag{9c}$$

Voor onze datapunten P_i $(i = 0,1,2,\ldots,n-1)$ kunnen we nu schrijven:

$$\hat{R}_i = \alpha + \beta P_i + \gamma/P_i + \varepsilon_i \ , \tag{10}$$

waarin $\hat{R}_i$ een schatting is voor $R(t_i)$, berekend uit de data. Bruikbaar is:

$$\hat{R}_i = q_i s_i \frac{\ln (q_i/s_i)}{q_i - s_i} \ ,$$

omdat men kan bewijzen, dat $\hat{R}_i$ ligt tussen

$$q_i = \frac{P_{i+1} - P_i}{(t_{i+1} - t_i)P_{i+1}} \qquad \text{en} \qquad s_i = \frac{P_i - P_{i-1}}{(t_i - t_{i-1})P_{i-1}} \ ,$$

die op zich al als mogelijke schatters kunnen worden beschouwd. De residuele componenten ε_i in (10) zijn klein als de data P_i ongeveer een gegeneraliseerd logistisch patroon te zien geven en als α, β en γ een logistische curve constitueren die qua vorm overeenkomt met dat patroon.
Met behulp van het algoritme ter berekening van de coëfficiënten in een lineaire multipele regressievergelijking kunnen, zonder daar verdere assumpties met betrekking tot kansverdelingen aan te verbinden, die waarden van α, β en γ berekend worden waarvoor $\Sigma\ \varepsilon_i^2$ wordt geminimaliseerd.
Uit (9a), (9b) en (9c) kunnen dan de parameterwaarden voor een logistische curve worden bepaald.

In vele gevallen dat de auteur van dit artikeltje ervan gebruik heeft gemaakt, is de methode van Levenbach in staat gebleken logistische curven op te sporen, die qua vorm goed overeenkomen met het patroon in data die ongeveer een logistische curve volgen.

Ons rest nog de bepaling van K, dus de bepaling van de plaats van de best passende curve. Ook in aansluiting op de zoëven beschreven methode zouden we K kunnen berekenen volgens formule (7). Maar bij data met onregelmatigheden is het natuurlijk onverstandig zich op één punt te verlaten.
Levenbach hanteert een formule die als uitbreiding van (7) beschouwd kan worden. Wordt voor elk van de n datapunten formule (7) opgesteld en vervolgens het product genomen, dan ontstaat:

$$K^n = \exp(D\Sigma i)\ \Pi\ \left(\frac{M - P_i'}{P_i'}\right) \quad ,$$

ofwel:

$$(11) \qquad K = \exp\left\{\frac{1}{n}\ \Sigma\ \left(\ln\frac{M - P_i'}{P_i'} + Di\right)\right\} \quad .$$

Deze methode maakt dus gebruik van alle datapunten en lijkt een aanvaardbaar alternatief.
Toch moet deze methode in het algemeen sterk worden afgeraden. In de praktijk is de kans zelfs groot dat (11) tot ernstiger onnauwkeurigheden leidt dan (7), mits men in het laatste geval z'n punt met enige zorg kiest. We kunnen de bezwaren van (11) duidelijker naar voren brengen, indien we de discussie toespitsen op het buigpunt van de gezochte curve. Daartoe schrijven we (11) iets anders:

$$\frac{\ln K}{D} = \frac{1}{nD} \sum \ln e^{Di} \frac{M - P_i'}{P_i'} = \frac{1}{n} \sum \frac{\ln Ki}{D} .$$

Dat betekent dat de tijdscoördinaat van het buigpunt van de gezochte curve gelijk is aan het gemiddelde van de coördinaten van de buigpunten van alle curven van gelijke vorm (dus met dezelfde waarden voor C,D en M) die getrokken kunnen worden door de punten P_0 , P_1 ,....,P_{n-1} .
Deze methode wordt problematisch als de data punten bevatten die schommelen rond het startnivo, respectievelijk het verzadigingsnivo.
Data uit de praktijk hebben meestal de eigenschap dat ze niet die meetnauwkeurigheid bezitten, dat ze deze nivo's monotoon met de tijd willekeurig dicht zullen naderen. Ze zullen rond deze nivo's gaan schommelen met een spreiding die onafhankelijk is van het tijdstip. (Kunnen we ervan opaan, dat de data die meetnauwkeurigheid wel bezitten dan maken we ons druk om niets: de methode uit de eerste helft van dit artikeltje zal waarschijnlijk goed voldoen).
Stel nu dat de gezochte curve de parameterwaarden C,D,K_o en M heeft. En dat een zeker datapunt P hoort bij het tijdstip $t_i = T_i + (\ln K_o)/D$ (T_i is het tijdsverloop voorbij het buigpunt van de gezochte curve). Zij $K(T_i)$ de K-waarde van de gelijkvormige curve door P. Er gelden dan:

$$P = C + \frac{M}{1 + K(T_i)\, e^{-Dt_i}} = C + \frac{M}{1 + \frac{K(T_i)}{K_o} e^{-DT_i}} ,$$

$$Y(T_i) = C + \frac{M}{1 + K_o e^{-Dt_i}} = C + \frac{M}{1 + e^{-DT_i}} .$$

De 'onregelmatigheid' is dan gelijk aan:

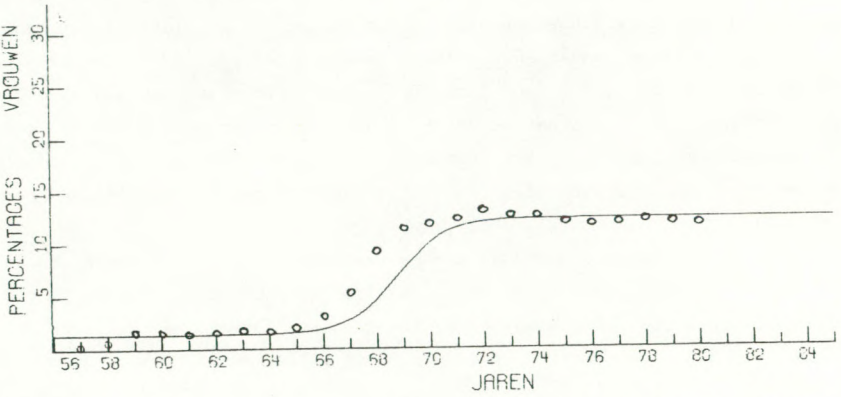
$$Y(T_i) - P = \frac{M \left(\frac{K(T_i)}{K_o} - 1\right) e^{-DT_i}}{\left(1 + \frac{K(T_i)}{K_o} e^{-DT_i}\right)\left(1 + e^{-DT_i}\right)} = \frac{(P - C)\left(\frac{K(T_i)}{K_o} - 1\right)}{e^{DT_i} + 1} .$$

Voor $T_i \to \infty$ geldt $Y(T_i) \to C + M$ en dus $K(T_i) \to \infty$ (bij vaste P), asymptotisch gelijk aan een exponentiële functie van T_i.

Naarmate de data meer punten rond het verzadigingsnivo bevatten, wordt de kans groter, dat een onregelmatigheid van een zekere grootte plaatsvindt op een tijdstip ver van het buigpunt van de gezochte curve. De corresponderende $(\ln K_i)/D$- bijdrage gaat in de berekening van (11) als uitbijter fungeren. Analoog voor punten rond het startnivo. In de praktijk zal men geconfronteerd worden met de paradoxale situatie dat de bepaling van K volgens (11) instabieler wordt, naarmate men gebruik maakt van meer datapunten, in het bijzonder wanneer er sprake is van een overwicht aan punten rond hetzij startnivo, hetzij verzadigingsnivo.

Tot wat voor wonderlijke curven (11) kan leiden wordt geïllustreerd in figuur 2.

Figuur 2



Participatie van meisjes (procentueel) aan het Hoger Technisch Beroepsonderwijs. (Waarde voor K van de 'bestpassende curve' bepaald volgens (11)).

Dat Levenbach niet op dit probleem stuitte is veroorzaakt doordat punten in de buurt van het startnivo, respectievelijk verzadigingsnivo, niet in zijn data voorkwamen. Maar het probleem is serieus, want juist data die zulke punten bevatten zullen de onderzoeker op het idee brengen een logistische curve aan te passen.

In het onderzoek van Hoksbergen & Maassen is een oplossing voor het probleem gezocht, gebaseerd op het idee dat voor de bepaling van C,D en M het gebruik van alle punten niet noodzakelijk is en voor de bepaling van K zelfs ongewenst. Er werd een voor de hand liggende maat voor de fit van de logistische curve gekozen. Bijvoorbeeld de gemiddelde absolute waarde of de standaarddeviatie van de verticaal gemeten afstanden tussen datapunten en curve. In de gevallen dat er een overwicht aan punten rond het startnivo bestond werden voor de meest recente m jaren C,D,M en K berekend volgens de methode van Levenbach.

De procedure werd herhaald bij toenemende waarde van m en uit de zo ontstane familie van krommen werden de curve met de beste fit gekozen.
In de minder vaak voorkomende gevallen dat er een overwicht bestaat aan punten rond het verzadigingsnivo kan een analoge procedure worden gevolgd, waarbij dan wordt gebruik gemaakt van de vroegste m punten (m weer variabel). Deze methode leverde steeds bevredigende resultaten. Hinderlijke punten verdwijnen aldus of brengen storingen in die elkaar opheffen. Maar wellicht mogen we zo niet meer verwachten dan dat althans in het oog lopende onnauwkeurigheden worden weggewerkt. De methode houdt een zekere instabiliteit die haar eigenlijk minder geschikt maakt als middel voor een objectieve vaststelling van het punt waar zich de grootste groei in het onderzochte fenomeen voltrekt.
Eigenlijk is het verwonderlijk dat er in de literatuur zoveel voorstellen zijn gedaan voor de berekening van parameter K, en dat niet algemeen het idee aanvaard wordt dat het verstandiger is alle informatie in de data te gebruiken om de fit van een curve te optimaliseren dan dat men deze informatie alleen maar gebruikt voor de bepaling van K.
Dit idee kan op een wijze worden uitgewerkt die gemakkelijk voor een computer is te programmeren:

(1) Bereken parameterwaarden voor C,D en M op basis van alle bruikbare datapunten (dat zijn alle punten waarvoor $\hat{R}_i$ bestaat) volgens de methode van Levenbach.

(2) Maak een beginschatting voor (lnK)/D door (7), respectievelijk (11), toe te passen op één of liever meer punten in de buurt van het buigpunt van de gezochte kromme. Bepaal de fit van het zo gevonden model.

(3) Breng een kleine verschuiving aan in (lnK)/D, naar 'links' en naar 'rechts'. Stel vast in welke richting de fit verbetert. Met aansluitende iteratieslagen, corresponderend met verdergaande verschuivingen in (lnK)/D in die richting, kan men de fit nog verbeteren tot aan zekere stopcriteria is voldaan.
Deze methode kan niet elegant genoemd worden, maar ze is wel doeltreffend gebleken.

Literatuur

Hoksbergen, R.A.C. en Maassen, G.H.: Hoever gaat de onderwijsemancipatie van de vrouw?; Mens en Maatschappij, jrg. 55 (1980), nr. 1, pg. 5-38.

Levenbach, Hans: A generalization of the logistic growth function and its estimation.; Statistica Neerlandica, vol. 27 (1973), nr. 1, pg.47-54.