Modelleren van Scanner Panel Data

Richard Paap

Het bestuderen van het individuele koopgedrag van huishoudens biedt meer inzicht in de effecten van promotie-activiteiten dan wanneer geaggregeerde data zoals totale verkopen worden gebruikt. In dit artikel zal een kort overzicht worden gegeven van simpele econometrische modellen, die gebruikt kunnen worden voor het beschrijven van aankopen van huishoudens. De modellen zullen worden geïllustreerd aan de hand van een bestaande data verzameling.

Inleiding

De introductie van streepjescodes op producten in combinatie met scanner kassa's maakt het mogelijk voor supermarkten om de totale verkopen van producten en merken per dag op eenvoudige wijze te bepalen. In combinatie met een vaste klantenkaart, zoals de bonuskaart van Albert Heijn, is het bovendien mogelijk om de bestedingen van individuele huishoudens te registreren. Per bezoek aan de supermarkt kan worden geregistreerd welke producten en merken worden gekocht, de prijs van de gekochte producten, en de prijzen van concurrerende producten, het tijdstip van de aankoop en de aanwezigheid van eventuele promotie-activiteiten op het tijdstip van aanschaf. Daarnaast zijn vaak ook een aantal kenmerken van de huishoudens bekend zoals gezinssamenstelling, leeftijd van het gezinshoofd en de postcode van het woonadres.

Een eenvoudige methode om deze data te bestuderen is het aggregeren van de aankopen van alle huishoudens tot totale verkopen van een merk per dag of per week. Het analyseren van deze aankopen over de tijd kan inzicht geven in de dynamische effecten van promotie-activiteiten. Het nadeel van het bestuderen van totale verkopen is dat veel informatie in de data wordt weggeaggregeerd. Een stijging van de verkopen van een merk door een prijspromotie kan immers door verschillende factoren worden verklaard. Het kan zijn dat huishoudens die het merk altijd kopen meer gaan consumeren. Een andere mogelijkheid is dat huishoudens die het merk normaliter niet kopen, tijdelijk of misschien permanent overstappen naar het betreffende merk door de gunstige prijs. Als laatste kan het ook voorkomen dat huishoudens meer kopen om een voorraad aan te leggen. Bij gebruik van geaggregeerde data kunnen deze factoren niet meer onderscheiden worden. Om een beter inzicht te krijgen in de precieze effecten van de prijsverlaging, is het daarom noodzakelijk om het aankoopgedrag van individuele huishoudens te bestuderen.

Het analyseren van aankopen van individuele huishoudens biedt ook andere voordelen. Zo kan bijvoorbeeld onderzoek worden gedaan naar merkloyaliteit, tussenaankooptijden (tijd tussen twee aankopen), en verschillen in preferenties van huishoudens, hetgeen met alleen totale verkopen onmogelijk is. Bovendien kunnen data op huishoudniveau aangewend worden voor marktsegmentatie analyses, die gebruikt kunnen worden voor doelgerichte reclame-campagnes. In dit artikel zal daarom een kort overzicht gegeven worden van simpele econometrische modellen die gebruikt worden om het aankoopgedrag van huishoudens te beschrijven. Alvorens we hiertoe overgaan, bekijken we in de volgende sectie een kleine scanner data verzameling uit de Verenigde Staten. Deze data verzameling zal tevens dienen als illustratie.

Data

Een scanner data verzameling verkregen met een vaste klantenkaart biedt geen garantie dat alle aankopen worden geregistreerd. Een consument kan zijn klantenkaart vergeten

variabele	Wisk	All	Eraplus	Solo	Tide	Surf
keuze in %	26.46	3.27	19.08	9.52	26.38	15.28
prijs / ounce in $cents	4.72	3.91	6.06	5.99	5.94	5.29
display in %	10.54	0.08	2.79	1.66	10.01	4.29
feature in %	12.46	0.08	3.69	2.18	10.76	5.34

Tabel 1: Keuze percentages, gemiddelde prijs, display en feature frequentie in de wasmiddelen data verzameling

of inkopen doen bij een concurrerende supermarkt. Om deze problemen uit te sluiten wordt vaak gebruik gemaakt van data verkregen van een huishoudpanel waarbij huishoudens alle gekochte producten thuis registreren met behulp van een eigen scanner.

Een voorbeeld van een huishoudpanel scanner data verzameling is de aankopen van vloeibare wasmiddelen in Sioux Falls (South Dakota) verzameld door A.C. Nielsen. De data verzameling bevat 400 huishoudens met in totaal 2657 aankopen, die beginnen in juli 1986 en lopen tot juli 1988. Het betreft aankopen van de top zes nationale merken, Tide, Eraplus, Solo (Procter & Gamble) en Wisk, Surf en All (Lever Bros.). Bij elke aankoopmoment van een huishouden weten we de tijd sinds de laatste aankoop, merkkeuze, de prijs van de zes merken en eventuele promotie-activiteiten aangegeven met *display* en *feature*. Bovendien weten we de gezinsgrootte van ieder huishouden, de hoeveelheid eenheden gekochte wasmiddelen, en de uitgaven aan overige producten buiten de vloeibare wasmiddelen om.

Tabel 1 geeft een overzicht van de data. We zien dat Wisk en Tide het meeste gekozen worden. Dit zijn ook precies de merken die de meeste promotie voeren. Eraplus, Solo en Tide zijn de meest dure merken. All is relatief goedkoop, doet weinig aan promotie, en wordt weinig gekozen. De gemiddelde gezinsgrootte is ongeveer 2.8, de mediaan van het aantal weken tussen de aankopen is 40, terwijl de gemiddelde overige uitgaven per bezoek aan de supermarkt 35 dollar bedragen.

Econometrische Modellen

Voor het modelleren van het aankoopgedrag van huishoudens neemt men vaak de drie belangrijkste beslissingen voor een consument die regelmatig een product (bijvoorbeeld koffie of wasmiddelen) koopt als uitgangspunt:

1. Wanneer koop ik het betreffende product?
2. Hoeveel eenheden van het product koop ik?
3. Voor welk merk kies ik?

zie bijvoorbeeld Gupta (1988). Voor elke van deze beslissingen wordt een model geconstrueerd dat de uitkomst van de beslissing relateert aan verklarende variabelen zoals bijvoorbeeld, de prijs van het product, eventuele promotie-activiteiten, maar ook karakteristieken van het huishouden zoals gezinsgrootte en inkomen. In deze sectie zullen we in het kort drie simpele econometrische modellen bespreken die vaak worden toegepast.

Merkkeuze

In de econometrische literatuur heeft de merkkeuze vraag de meeste aandacht gekregen. Het merkkeuzegedrag van consumenten wordt vaak gemodelleerd door gebruik te maken van het concept van stochastische nutsmaximalisatie. Consumenten bepalen het nut van ieder merk dat afhangt van de eigenschappen van het merk, zoals bijvoorbeeld kwaliteit en smaak, maar ook van de prijs. Huishoudens worden vervolgens geacht te kiezen voor het merk met de hoogste nut. Deze theoretische modellen leiden vaak tot de bekende conditionele logit/probit modellen. In een conditioneel logit model bijvoorbeeld wordt de kans dat een huishouden i merk j kies uit de j beschikbare merken beschreven door

$$(1) \qquad \Pr[d_i = j] = \frac{\exp(\mu_j + x_{ij}\beta)}{\sum_{s=1}^{J} \exp(\mu_s + x_{is}\beta)}$$

voor $j=1,\ldots,J$ met $\mu_J = 0$ voor identificatie. De μ_j

variabele	parameter	standaardfout
Wisk	-0.065	0.077
All	-2.953	0.141
Eraplus	1.443	0.086
Solo	0.625	0.064
Tide	1.532	0.084
Surf	0	-
log prijs	-6.163	0.191
display	0.532	0.120
feature	0.867	0.113

Tabel 2: Parameter schattingen van een conditioneel logit model voor merkkeuze

parameters modelleren de basis preferentie voor merk j en de β parameters de effecten van de verklarende variabelen x_{ij} (zoals prijs van merk j ondervonden door huishouden i) op de keuze.

Als illustratie schatten we een conditioneel logit model voor de merkkeuze voor de hierboven beschreven data verzameling. Tabel 2 toont de parameter schattingen. We zien dat prijs, zoals verwacht, een negatieve invloed heeft op merkkeuze. De variabelen display en feature (0/1 dummy's) hebben beide een significante positieve invloed op keuze. Het voorgestelde logit model is echter te simpel om de merkkeuzes van huishoudens nauwkeurig te beschrijven. Sinds de introductie van logit modellen voor merkkeuzes door Guadagni en Little (1983) zijn er veel aanpassingen aan het model voorgesteld. Een beperking van het model is dat verondersteld wordt dat ieder huishouden dezelfde basis preferenties heeft. Door het toelaten van verschillende parameters per huishouden kan flink in verklaringskracht worden gewonnen. Een tweede verbetering betreft het toelaten van dynamiek in het model. Hiermee kan bijvoorbeeld gekeken worden naar de effecten van promoties over tijd. Een simpele manier om dit te doen is door het opnemen van een vertraagde merkkeuze dummy. Met meer geavanceerde modellen kunnen zelfs de korte termijn effecten onderscheiden worden van de lange termijn effecten, zie Paap en Franses (2000). Een laatste verbetering betreft het toelaten van *consideration sets*, waarbij consumenten eerst het aantal merken beperken alvorens over te gaan tot hun definitieve keuze, zoals beschreven in Van Nierop (2000) en Vroomen (2001).

Tussenaankooptijd

Voor het modelleren van de tijd tussen twee aankopen wordt vaak gebruikt gemaakt van zogenaamde duurmodellen, zie bijvoorbeeld Jain and Vilcassim (1991). Vaak wordt niet expliciet gekeken naar de aankooptijd zelf, maar naar de kans dat een product gekocht wordt gegeven dat het niet gekocht is sinds de laatste aankoop. In continue tijd vertaalt dit zich tot de zogenaamde hazard functie die gedefinieerd is als de ratio van een dichtheidsfunctie ($f(t)$) en een overlevingsfunctie ($1-F(t)$)

$$\lambda_0(t) = \frac{f(t)}{1-F(t)} \qquad (2)$$

waarbij t de tijd in weken of dagen is. De dichtheidsfunctie $f(t)$ beschrijft de verdeling van tussenaankooptijden waarvoor verschillende mogelijkheden zijn, zoals een exponentiële verdeling, een Weibull verdeling, of een log-logistische verdeling. Verklarende variabelen worden op een simpele wijze aan de standaard hazard functie $\lambda_0(t)$ toegevoegd

$$\lambda(t_i \mid x_i) = \exp(-x_i\gamma)\lambda_0(t_i) \qquad (3)$$

waarbij t_i de tussenaankooptijd van huishouden i is. Dit model wordt het proportionele hazard model genoemd. De parameter γ beschrijft het effect van de verklarende variabelen x_i op de conditionele kans dat het product gekocht wordt.

Het eerste deel van Tabel 3 toont de parameterschattingen van een proportionele hazard model voor de tussenaankooptijden van vloeibare wasmiddelen. De hazard functie komt voor uit een log-logistische verdeling voor de tusssenaankooptijden. We zien dat grote gezinnen een kortere tussenaankooptijd hebben zoals verwacht. De hoeveelheid gekochte wasmiddelen bij de vorige aankoop heeft een positief effect. Deze variabele wordt vaak opgenomen als een simpele manier om te corrigeren voor

	Duurmodel		*Poisson model*	
variabele	parameter	standaardfout	parameter	standaardfout
constante	-0.284	0.132	-0.165	0.085
Δ log prijs	0.104	0.054	-0.045	0.054
display	-0.034	0.079	0.109	0.064
feature	0.133	0.072	0.131	0.058
log (tijd sinds vorige aankoop)	-	-	0.079	0.018
aantal gekocht vorige aankoop	0.091	0.021	-	-
gezinsgrootte	-0.127	0.014	0.052	0.013
overige uitgaven	-0.007	0.042	0.093	0.026

Tabel 3: Parameter schattingen van een duurmodel voor tussenaankooptijd en een Poisson model voor gekochte hoeveelheid

voorraadvorming. Een hogere prijs impliceert een langere tussenaankooptijd. Display en feature zijn beiden niet significant en lijken geen effect te hebben.

Het geschatte duurmodel is wederom vrij simpel. Omdat de parameter γ gelijk is over de huishoudens, wordt bijvoorbeeld verondersteld dat huishoudens hetzelfde reageren op prijsveranderingen. Een ander punt van kritiek is dat prijzen en promoties tussen twee aankopen kunnen veranderen. Het feit dat een huishouden nu vloeibare wasmiddel koopt, zal daarom ook afhangen van het door hem geobserveerde prijspatroon tussen de aankopen, zie Gupta (1991). Het hierboven geschatte model houdt hier geen rekening mee.

Aangekochte Hoeveelheid

De aangekochte hoeveelheid zou bijvoorbeeld beschreven kunnen worden met een standaard regressiemodel. Vaak is het echter alleen mogelijk om een product in discrete aantallen te kopen. In dat geval wordt vaak gebruikt gemaakt van een Poisson verdeling, waarbij de locatie parameter een functie is van verklarende variabelen zoals $\exp(x_i\delta)$. Gegeven aankoop wordt de kans dat een huishouden i k eenheden koopt gelijkgesteld aan

$$(4) \qquad \Pr[y_i = k] = \frac{\exp(-\exp(x_i\delta))(\exp(x_i\delta))^{y_i - 1}}{(y_i - 1)!}$$

voor $k>0$. De parameter δ beschrijft het effect van de verklarende variabelen x_i op de aangekocht hoeveelheid.

Het tweede deel van Tabel 3 toont de parameter schattingen van een Poisson model voor de aangekochte hoeveelheden vloeibare wasmiddelen. We zien dat prijs en display geen significante invloed hebben. Feature heeft een significante positieve invloed op de hoeveelheid. Wanneer het langer geleden is dat vloeibare wasmiddelen zijn gekocht dan wordt er significant meer gekocht. Grote gezinnen kopen meer wasmiddelen en er worden ook meer eenheden gekocht als er grote hoeveelheden boodschappen worden gedaan.

Het geschatte model is wederom vrij simpel en voor praktische toepassingen is het raadzaam om te kijken of het model verbeterd kan worden. Men kan bijvoorbeeld heterogeniteit in de beslissingen van consumenten toestaan of door het opnemen van vertraagde aankoophoeveelheden dynamiek toelaten.

Waarschijnlijk kan bovendien de tussenaankooptijd en de hoeveelheidsbeslissing niet los van elkaar bekeken worden, zodat een gemeenschappelijk model voor beide grootheden noodzakelijk is.

Conclusie

In dit artikel hebben we een kort overzicht gegeven van econometrische modellen voor het beschrijven van het aankoopgedrag van individuele huishoudens. De modellen die besproken zijn, zijn echter simpel in vergelijking met toepassingen in de empirische marketing literatuur. Vooral de onderwerpen (niet-waargenomen) heterogeniteit en het modelleren van dynamische structuren hebben daar recent veel aandacht gekregen. Het goed modelleren van heterogeniteit is namelijk zeer belangrijk om een helder inzicht te krijgen in de veranderingen in het (keuze)gedrag van huishoudens voor het analyseren van de effecten van promoties over de tijd.

Personalia

Correspondentie-adres:	Richard Paap
	Erasmus Universiteit Rotterdam
	RIBES, kamer H16-17
	Postbus 1738
	3000 DR Rotterdam
E-mail:	paap@few.eur.nl

Referenties

[1] Jain, D.C. en N.J. Vilcassim, 1991, Investigating Household Purchase Timing Decisions: A Conditioneel Hazard Function Approach, *Marketing Science*, 10, 1-23.

[2] Guadagni, P.E. and J.D.C. Little, 1983, A Logit Model of Brand Choice Calibrated on Scanner Data, *Marketing Science*, 2, 203-238.

[3] Gupta, S., 1988, Impact of Sales Promotions on When, What, and How Much to Buy, *Journal of Marketing Research*, 25, 342-355.

[4] Gupta, S., 1991, Stochastic Models of Interpurchase Time with Time-Dependent Covariates, *Journal of Marketing Research*, 28, 1-15.

[5] Paap, R. en P.H. Franses, 2000, A Dynamic Multinomial Probit Model for Brand Choice with Different Long-run and Short-run Effects of Marketing-mix Variables, *Journal of Applied Econometrics*, 15, 717-744.

[6] Van Nierop, E., 2000, Modelleren van Consideration Sets in Merkkeuzemodellen, *Medium Econometrische Toepassingen*, 4, 4-6.

[7] Vroomen, B., 2001, Using Artificial Neural Networks to Model Consideration Sets and Brand Choice, *Medium Econometrische Toepassingen*, 2, 4-7.