2 Interpretatie van de regressies

Bij een lineire regressie drukken we een bepaalde variabele die veel getalwaarden kan aannemen, bijvoorbeeld aantal dagen carpoolen per jaar, uit als een lineaire functie van ander variabelen bv. 'vrouw zijn' en leeftijd tussen 16 en 24 jaar.

Dan is de regressie van de vorm:
$Y = aX1 + bX2 + c$, met Y het aantal dagen dat men met iemand meerijdt, X1 en X2 onafhankelijke variabelen, hier 'vrouw zijn' en leeftijd tussen 16 en 24 jaar, en a, b en c door SAS (software-programma) berekende constanten.

Indien we een regressie willen berekenen voor een variabele die enkel 'ja' of 'nee' kan zijn, zoals het bezit van een rijbewijs, dan kunnen we geen gewone lineaire regressie toepassen, maar wel een logistische regressie. De logistische regressie lijkt op een gewone regressie, maar op de afhankelijke variabele wordt eerst een logistische transformatie toegepast.

De regressie is van de vorm:
$\ln\left(\frac{P}{1-P}\right) = aX1 + bX2 + c$, met P de kans dat iemand een rijbewijs heeft, en net zoals bij lineaire regressie, X1 en X2 onafhankelijke variabelen, hier 'vrouw zijn' en leeftijd tussen 16 en 24 jaar, en a, b en c door SAS berekende constanten.

We kunnen deze vergelijking ook schrijven als:

$$De\,kans\,op\,een\,rijbewijs = \frac{1}{1+e^{-(aX1+bX2+c.)}}.$$

Dit maakt het (iets) eenvoudiger om de getallen te interpreteren.

Tabel 1. Fictief voorbeeld van een logistische regressie om de begrippen uit te leggen. Afhankelijke variabele is rijbewijsbezit

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	4.5595	0.2098	472.1404	0.0001	.	.
VROUW	1	-1.2984	0.1815	51.2019	0.0001	-0.351515	0.273
LFT1624	1	-0.1966	0.3123	0.3962	0.5291	-0.022571	0.822

Voor wie niet echt geïnteresseerd is in de exacte getalwaarde, maar enkel in het feit of iemand meer of minder kans heeft op een rijbewijs volstaat volgende vuistregel. Als de '*Parameter Estimate*' positief is dan stijgt de kans op een rijbewijs; indien de Parameter Estimate negatief is dan daalt de kans op een rijbewijs.
Voorbeeld uit Tabel 1: De 'parameter estimate' van vrouw is "-1.2984". Deze parameter estimate is negatief, dus daalt de kans op een rijbewijs indien de persoon in kwestie een vrouw is.

De volledige betekenis van deze logistische regressie is:

$$De\,kans\,op\,een\,rijbewijs = \frac{1}{1+e^{-(4.5595-1.2984\,'indien\,vrouw'-1.966\,'indien\,tussen\,16\,en\,24\,jaar')}}$$

Als variabelen niet in de regressie voorkomen, wil dit zeggen dat ze geen toegevoegde waarde meer hebben *bij alle variabelen die reeds in het model zitten*. Zo blijkt dat een aanzienlijk aantal respondenten geen treinstation in de buurt van hun huis hebben, maar ook niet in de buurt van hun werk. Vaak is dit een reden om een ander vervoermiddel te nemen. Het is van belang dat er op één plaats geen station is. Maar het extra probleem dat er op de andere plaats ook geen station is, is erg beperkt. Daardoor verschijnt dit vaak niet meer in de regressies. Welke afstand (halte thuis of halte op het werk) er in het model opgenomen wordt, is zuiver bepaald door de statistische berekeningen. De meest significante variabelen blijven over.
We hebben dus meestal veel meer variabelen uitgeprobeerd, dan dat er uiteindelijk in de regressie overblijven.

Ook het aantal variabelen dat in de regressies is opgenomen is zuiver statistisch bepaald. De variabelen met een significantie kleiner dan 5% (P < 5%) zijn opgenomen in het model. Dit heeft tot gevolg dat er soms variabelen verschijnen die we niet verwacht hadden, en die we zelfs bij nader inzien niet kunnen verklaren[2]. Het is ook mogelijk om meer 'gericht' modellen te maken. We kunnen bijvoorbeeld enkel variabelen met P<5% weerhouden als we kunnen begrijpen waarom ze relevant zijn, en variabelen waarvan we niet begrijpen waarvan hun impact komt, weerhouden we enkel bij P<1%, of P<0.1%. Dit is duidelijk minder wetenschappelijk, maar het levert een model op waarvan we alle aspecten (denken te) begrijpen, en een model dat ook eenvoudiger uit te leggen is aan de buitenwereld, bv. voor het sturen van beleidsbeslissingen. Anderzijds kunnen we ook voor op voorhand bepaalde variabelen in het model dwingen, om hun P-waarde te kennen. Het kan voor een overheidsbeslissing van belang zijn of de afstand tot een bushalte met 7% kans irrelevant is voor het nemen van het openbaar vervoer, of met 60% kans geen impact heeft op een stijging van het gebruik van het openbaar vervoer. In het geval van P =7% is het de moeite om andere analyses te doen, of zelfs ander onderzoek te verrichten om meer zekerheid te krijgen of er nu wel of niet een invloed is, in het geval dat P=60% is er gewoonweg geen verband.
We hebben er heel bewust niet voor gekozen om op het eerste zicht vreemde variabelen weg te laten of andere variabelen in de regressie te dwingen.. Enerzijds omdat dit een eerste poging was om met regressie beter zicht te krijgen op het gebruik van de vervoermiddelen. En dan is het voorzichtig om gewoon het terrein af te tasten zonder zelf te veel in te grijpen. Anderzijds omdat dergelijke ingrepen enkel zin hebben indien men heel gericht antwoorden zoekt op bepaalde vragen. En ook daarvoor is het nu nog te vroeg.

De 'parameter estimate' bij het intercept geeft de waarde van de regressie in de referentiesituatie. Dit impliceert dan ook dat er een referentiesituatie bepaald wordt. Ook dit hebben we aan de statistiek overgelaten: we hebben de statistiek de significant afwijkende variabelen laten zoeken. Wat niet afwijkt is dan de referentiesituatie. Hiervoor geldt dezelfde opmerking als hierboven. We hadden zelf

[2] Strikt gezien kan dit een gevolg zijn van het gebruikte statistische criterium. De variabelen met een significantie kleiner dan 5% zijn behouden in het model. Dit wil zeggen dat, als er maar 5% kans is dat een variabele bij wijze van pech door de steekproef relevant lijkt, maar in werkelijkheid toch niet relevant is, dat we dan de variabele behouden. De redenering daarachter is: '5% kans is zo klein, dat kan geen toeval meer zijn'.
We kunnen dit ook minder positief formuleren: indien we 100 variabelen proberen om een model te maken, dan kunnen er door zuiver pech 5 geselecteerd worden die significant lijken, maar het eigenlijk niet zijn. Welnu, voor deze modellen hebben we ongeveer 60 variabelen uitgeprobeerd. Normaal gezien worden alle relevante variabelen geselecteerd, maar we moeten er rekening mee houden dat er ook variabelen geselecteerd zijn die toch niet relevant zijn. Het is verleidelijk om te stellen dat dit de variabelen zijn waarvan we de impact niet begrijpen.
Met behulp van meer geavanceerde statistische, maar helaas ook meer arbeidsintensieve technieken, is het mogelijk om overfitting met grotere zekerheid uit te sluiten. Overfitting heeft plaats als men een variabele toevoegt die belangrijk lijkt, maar het eigenlijk niet is. Het is een variabele die voor deze steekproef significant is, maar dat bij een andere steekproef niet meer zou zijn. Door deze variabele toe te voegen lijkt het dus alsof de regressie verbetert, maar in werkelijkheid weten we daardoor niets meer over de populatie.

kunnen ingrijpen, maar zolang we niet zeker weten hoe en waarom, is het voorzichtig om dit niet te doen. Dat blijkt bijvoorbeeld uit het feit dat de referentiesituatie voor de bus-, tram en treinhaltes verschilt van regressie tot regressie.

3 Niet-becommentarieerde tabellen

3.1 Gebruik van vervoermiddelen en ligging van de woning

Aan de respondenten werd gevraagd de ligging van hun woonplaats aan te geven. Men kon kiezen tussen 3 mogelijkheden: in het centrum of dichtbebouwd gebied, afgelegen of tussen in.

Bijkomend onderzoek (OVG-perceptie afstand en bebouwingsindex; Nuyts, Princen, Zwerts, 2000) heeft uitgewezen dat de interpretatie door de respondenten niet altijd even consequent is en overeenkomt met een "professionele" definitie. Bij de interpretatie van de onderstaande tabellen dient hiermee rekening te worden gehouden.

Tabel 2. Verdeling van de personen volgens frequentie van wagengebruik en ligging van de woning

GAUTO(Gebruik van de auto) LIGGING(Ligging van de woonplaats)

Frequency Percent Row Pct Col Pct	in het centrum/dicht bebouwd gebied	niet in het centrum, niet afgelegen	afgelegen	Total
nooit	226.58	199.56	20.41	446.54
	3.73	3.29	0.34	7.36
	50.74	44.69	4.57	
	8.55	6.49	5.90	
dagelijks	1397.2	1806.6	225.62	3429.4
	23.02	29.77	3.72	56.51
	40.74	52.68	6.58	
	52.73	58.80	65.18	
1 tot enkele keren per week	889.62	924.1	88.044	1901.8
	14.66	15.23	1.45	31.34
	46.78	48.59	4.63	
	33.57	30.08	25.44	
1 tot enkele keren per maand	109.42	114.52	8.8258	232.76
	1.80	1.89	0.15	3.84
	47.01	49.20	3.79	
	4.13	3.73	2.55	
1 tot enkele keren per jaar	27.079	27.875	3.2314	58.185
	0.45	0.46	0.05	0.96
	46.54	47.91	5.55	
	1.02	0.91	0.93	
Total	2649.93	3072.64	346.129	6068.7
	43.67	50.63	5.70	100.00