3.3 Effectief gebruikte gewichten verplaatsingen

Tabel 7. Gewichten die aan de invuldagen en maanden zijn toegekend om de steekproef representatiever te maken

Label Maand van Invuldag	Sample freq.	Pop.freq.	Pop.prop.	Expected freq.	Chi-2	Weight
1	593	31	0.0846995	587.052	0.0602668	0.989969
2	704	29	0.079235	549.178	43.647	0.780082
3	620	31	0.0846995	587.052	1.8492	0.946858
4	516	30	0.0819672	568.115	4.78063	1.101
5	531	31	0.0846995	587.052	5.35186	1.10556
6	562	30	0.0819672	568.115	0.0658146	1.01088
7	545	31	0.0846995	587.052	3.01228	1.07716
8	532	31	0.0846995	587.052	5.1626	1.10348
9	556	30	0.0819672	568.115	0.258341	1.02179
10	589	31	0.0846995	587.052	0.00646458	0.996693
11	548	30	0.0819672	568.115	0.712186	1.03671
12	635	31	0.0846995	587.052	3.91621	0.924491

De gewichten waarmee we verplaatsingen willen vermenigvuldigen, zijn berekend op personenniveau. Dit wil zeggen dat we nagaan of er in elke maand een voldoende aantal personen ondervraagd is die zich hadden kunnen verplaatsen. Deze personen krijgen elk hun gewicht mee, zoals bepaald bij 'gewichten personen'. Op deze wijze bekomen we voldoende gegevens per maand, maar worden maanden waarin mensen zich minder verplaatsen niet kunstmatig opgehoogd. Het gewicht voor de maand februari is 0.78 (Tabel 7), niet omdat er te veel verplaatsingen waren in februari, maar omdat in die maand we meer personen ondervraagd hebben. De verschillen tussen 1e en 2e invuldag zijn niet zo groot dat er andere daggewichten voor nodig zijn.

4 Technische aspecten i.v.m. de statistische verwerking

De bekomen data werden uitgezuiverd aan de hand van strikte regels (Nuyts & Zwerts 2000), en verwerkt met behulp van het statistische pakket SAS. De meeste resultaten zijn voorgesteld als frequentietabellen. Voor sommige vragen waarbij we willen weten op welke wijze een variabele afhangt van meerdere andere variabelen is lineaire of logistische regressie gebruikt.

4.1 Frequentietabellen en regressie

Voor de meeste analyses gebruiken we enkel frequentietabellen. Hiermee kunnen we het verband tonen tussen twee variabelen of, indien handig geschikt, eventueel tussen drie variabelen. In bepaalde gevallen willen we echter de invloed kennen die meerdere variabelen hebben op één andere variabele. Bijvoorbeeld: hoe hangt het aantal personenwagens af van het geslacht van het gezinshoofd, de leeftijd van het gezinshoofd, het gezinsinkomen en het aantal gezinsleden. Dit doen we via regressie.

In de OVG's passen we *multivariate* regressie toe: we proberen steeds een verband te leggen tussen 1 afhankelijke variabele en verscheidene onafhankelijke variabelen[3]. Afhankelijk van de mogelijke waarden van de afhankelijke variabele gebruiken we een ander 'type' regressie: *lineaire* regressie of *logistische* regressie.

[3] Multi-variate regressie in tegenstelling tot *univariate* regressie waar men 1 afhankelijke variabele probeert te begrijpen met behulp van 1 onafhankelijke variabele.

Lineaire regressie is de meest 'klassieke' regressie. Deze regressie is van de vorm:
$Y = aX1 + bX2 + ...cXn + d$, met Y de afhankelijke variabele en X1 tot Xn n onafhankelijke variabelen. De regressie heet lineair omdat alle variabelen, zowel de afhankelijke als de onafhankelijke, lineair gebruikt worden (=zonder er kwadraten of andere functies op toe te passen).

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:

$$\ln\left(\frac{P}{1-P}\right) = aX1 + bX2 + ...cXn + d,$$

P is dan de kans dat iemand een rijbewijs heeft.

In principe is regressie uitgevonden om continue variabelen te vergelijken met continue variabelen, b.v. lengte van armen i.f.v. totale lengte. Bij dit onderzoek zijn er echter verscheidene variabelen die opgedeeld zijn in klassen. We kunnen hier op verschillende manieren mee omgaan in de regressie.

- Indien de geklasseerde waarden oorspronkelijk continu waren, dan kunnen we de klassen vervangen door hun midden. Dit levert een (oplosbaar) praktisch probleem voor de laatste klasse, daar die in principe geen midden heeft. Een ander nadeel is dat men ervan uitgaat dat elke onafhankelijke variabele een lineaire invloed heeft op de afhankelijke variabele. Dit is in praktijk niet waar. Een stijging van het inkomen van 40.000 BEF per maand, indien men 20.000 BEF per maand verdient, heeft een totaal andere invloed op de mobiliteit van deze persoon dan een stijging van het inkomen van 40.000 BEF per maand, indien men reeds 200.000 BEF per maand verdient. Dit is dan wel op te vangen door de variabelen te transformeren, maar dan vermindert het inzicht in het uiteindelijke model.

- Een andere mogelijkheid is de klassen vervangen door een zelfbepaalde waarde. De waarde wordt zo bepaald, dat een univariate regressie van deze variabele zo performant mogelijk is. Indien de waarden handig gekozen worden verhoogt dit de performantie van het uiteindelijke model. Het nadeel is dat de bepaling van de waarden steeds iets arbitrairs heeft, en dat het uiteindelijke model moeilijker te interpreteren is.

- Een derde mogelijkheid is de k klassen vervangen door k-1 dummy variabelen (=ja/nee of 0/1 variabelen). Dit wil zeggen dat we één referentieklasse kiezen, en voor elke andere klasse een variabele die zegt of de waarneming ertoe behoort of niet. Dit heeft twee nadelen. Het aantal variabelen kan snel oplopen. Indien men echter genoeg waarnemingen heeft, en het regressiemodel eerder manueel opbouwt dan SAS de variabelen te laten kiezen, dan kan dit probleem omzeild worden. Een ander nadeel is dat soms alle onafhankelijke variabelen dummy variabelen zijn. Lineaire regressie veronderstelt dat alle variabelen samen een multivariate normaalverdeling hebben. Dat is erg onwaarschijnlijk indien alle variabelen dummy variabelen zijn. Het gevolg hiervan is dat de schatting van de coëfficiënten iets minder correct is dan verhoopt kon worden. Het grote voordeel van het gebruik van dummy variabelen is dat elke klasse zijn eigen coëfficiënt krijgt, en dat het uiteindelijke regressiemodel erg inzichtelijk is.

We willen de regressiemodellen zo overzichtelijk mogelijk houden, en hebben daarom gekozen voor het gebruik van dummy variabelen.

Bij regressie kan men rekening houden met verschillende variabelen, bijvoorbeeld vrouw zijn, of jonger dan 25 jaar, maar ook met combinaties van dergelijke variabelen: vrouwen jonger dan 25 jaar. Hoe meer men combineert, hoe kleiner het aantal waarnemingen in de doelgroep. In principe houdt SAS bij de berekening van de relevantie (= significantie) van een bepaalde combinatie rekening met

het aantal waarnemingen in de betrokken groep. Kleine groepen hebben minder kans om significant te zijn. Om mogelijke overfitting[4] te voorkomen, hebben we op voorhand reeds groepen met minder dan 25 waarnemingen uitgesloten. We hebben eveneens getracht zoveel mogelijk beïnvloedende factoren te betrekken alhoewel dit niet steeds mogelijk is (b.v. de afstand tot een bepaalde bushalte is opgenomen in de regressie, de ritfrequentie van de bus(sen) evenwel niet). In die zin moeten we de regressieresultaten enigszins relativeren.

4.2 Betrouwbaarheid van de resultaten.

De fout bij de berekening van de parameters bij regressie kan men berekenen door een veelvoud te nemen van de standaarderror op deze parameter. De standaarderror σ wordt door SAS mee berekend en ook getoond in de output. De schatting van de parameter heeft ongeveer een normaal verdeling. Om bijvoorbeeld een betrouwbaarheidsinterval te krijgen van 95%, neemt men de schatting van de parameter ± 1.96*σ.

Bij proporties geldt een vergelijkbaar resultaat. Men kan dan de standaarderror berekenen aan de hand van de bekomen proportie en het gebruikte aantal in de steekproef. Voor een proportie p en een steekproefaantal n vindt men met een betrouwbaarheid van 95% :

$$p \pm 1.96 * \sqrt{\frac{p*(1-p)}{n}}\,.$$

Bij een steekproefaantal van 2000 waarnemingen, en een proportie van 0.10 vinden we dan

$$0.10 \pm 1.96 * \sqrt{\frac{0.10*(0.90)}{2000}} = 0.10 \pm 0.013\,.$$

We hebben bij de steekproef dus een proportie gevonden van 0.10, en we zijn 95% zeker dat de proportie voor de populatie ligt tussen 0.087 en 0.113.
De onzekerheid stijgt hoe dichter de waarde van de proportie bij 0.5 ligt. Bij p=0.5 vinden we 0.5 ± 0.022. De onzekerheid is dubbel zo groot als bij een proportie van 0.10. Vanzelfsprekend stijgt de fout ook bij een kleiner aantal waarnemingen.

Als globale regel kunnen we stellen dat we, zolang we 2000 waarnemingen hebben, voor alle proporties de onzekerheid van de proportie in de buurt ligt van 2%. Stijgt het aantal waarnemingen, of is de proportie verder verwijderd van 0.5, dan daalt de fout.

5 Vergelijking telefonisch/postaal bevraagden en enkel postaal bevraagden

Er zijn 3 groepen van mensen die enkel postaal bevraagd zijn:
(1) Huishoudens waarmee geen telefonisch contact mogelijk is omdat ze geen vaste telefoonaansluiting hebben.
(2) Huishoudens waarmee geen telefonisch contact mogelijk is omdat een fout nummer aan het betrokken gezin gekoppeld was.
(3) Huishoudens waarmee geen telefonisch contact mogelijk is omdat ze weliswaar een vaste telefoonaansluiting hebben maar enkel een geheim nummer.

[4] 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.