Steunpunt beleidsrelevant onderzoek 3de generatie

Fiscaliteit en Begroting

Spoor:

A3a2 IMPACT HERVORMING VERKEERSBELASTING
Wat drijft gezinnen in hun autokeuze?
Een continu-discreet model op gezinsniveau

Laurent Franckx
Hans Michiels

VITO Rapport 2014/TEM/R/15
VITO NV
Unit Transitie Energie & Milieu
2014

Algemeen secretariaat – Steunpunt beleidsrelevant Onderzoek
Fiscaliteit & Begroting
Henleykaai 84 – 9000 Gent – België
Tel: 0032 (0)9 243 29 06 – E-mail: vanessa.bombeeck@ugent.be
www.steunpuntfb.be

Steunpunt Fiscaliteit en Begroting
Spoor A3a2: Impact Hervorming Verkeersbelastingen

WAT DRIJFT GEZINNEN IN HUN AUTOKEUZE? EEN CONTINU-DISCREET MODEL OP GEZINSNIVEAU

Laurent Franckx
Hans Michiels

VITO Rapport 2014/TEM/R/15
VITO NV
Unit Transitie Energie & Milieu
2014

Laurent Franckx, Hans Michiels, Spoor A3a2: Impact Hervorming Verkeersbelasting. Wat drijft gezinnen in hun autokeuze? Een continu discreet model op gezinsniveau
2014 /TEM/R/15

Deel A3a2 van het Steunpunt Fiscaliteit en Begroting II heeft als doelstelling om de effecten van een hervorming van de verkeersbelastingen na te gaan, met het oog op de vergroening van de autofiscaliteit. Het rapport ontwikkelt een gecombineerd discreet/continu keuzemodel om na te gaan welke factoren een invloed uitoefenen op het aantal auto's en het type auto's in het bezit van een gezin en op het aantal kilometers dat het gezin jaarlijks aflegt. De factoren omvatten gezinskenmerken, technische autokenmerken en autokosten (incl. belastingen of subsidies). De analyse beschouwt enkel gezinnen met maximaal twee wagens in hun bezit.

Er worden drie stappen ondernomen. Eerst wordt aan de hand van discrete-keuze modellen de keuze gemodelleerd van de "klasse" auto die een gezin bezit, waarbij er rekening gehouden wordt met de kenmerken van de individuele modellen in elke klasse. De globale voorspellende waarde van het model is beperkt. Nochtans is de coëfficiënt van meerdere variabelen significant verschillend van nul, en heeft deze door de band genomen het verwachte teken. Zo neemt bij gezinnen met een laag inkomen, de kans dat een bepaalde autoklasse gekozen wordt, af indien de prijs ervan stijgt. Bij gezinnen met een hoog inkomen stelt men het omgekeerd effect vast, wat verklaard zou kunnen worden door een soort 'snob-effect'. Een toename van de brandstofkost leidt zoals verwacht tot een lagere kans dat een bepaalde klasse wordt gekozen. Voor gezinnen met één auto blijkt de coëfficiënt van de verkeersbelasting niet of slechts licht significant verschillend van nul. Voor gezinnen met twee auto's, daarentegen, oefent de verkeersbelasting wel degelijk een significante invloed uit.

Vervolgens wordt de keuze van het aantal auto's gemodelleerd voor gezinnen met twee of minder auto's. Voor dit model is de voorspellende waarde goed. De coëfficiënten van meerdere variabelen blijken statistisch significant, met het verwachte teken. Bijvoorbeeld, hoe hoger het gezinsinkomen en hoe groter de woon-werkafstand, hoe hoger het gemiddeld aantal auto's dat het gezin bezit. Gezinnen die in centrumsteden wonen of geregeld gebruik maken van trein of bus, bezitten gemiddeld minder wagens.

Tenslotte wordt het aantal kilometers geschat dat jaarlijks per auto wordt afgelegd, rekening houdend met het aantal auto's dat het gezin bezit. Bij dergelijke modellen stellen zich vaak problemen van endogeniteitsbias met betrekking tot de kenmerken van individuele modellen en zelf-selectie bias met betrekking tot het aantal auto's. Formele statistische testen hebben uitgewezen dat deze problemen zich hier niet stellen, zodat een OLS-schatting kan gebruikt worden. De globale voorspellende waarde van het afstandsmodel is beperkt. Nochtans oefenen er meerdere variabelen een significante invloed uit op de afgelegde afstanden. Zo blijkt de afgelegde afstanden per auto eerst te stijgen als functie van de woon-werkafstand, totdat deze een drempel overschrijdt. Deze drempel komt waarschijnlijk overeen met het punt waar de woon-werkafstand zo groot is geworden, dat het gemiddeld beter is om over te stappen op de trein. Een hogere brandstofkost gaat gepaard met kleinere jaarlijkse afstanden. Bij gezinnen met twee wagens blijkt dat het ouder en het kleiner automodel in het bezit van het gezin minder gebruikt worden dan de andere gezinswagen.

De beperkte voorspellende waarde van de modellen kan toegeschreven worden aan een samenspel van meerdere factoren. Ten eerste werden bepaalde belangrijke verklarende variabelen niet opgenomen, omdat de respons voor deze variabelen in het Onderzoek Verplaatsingsgedrag (OVG) te laag was. Ten tweede ontbraken vaak ook cruciale gegevens met betrekking tot de technische kenmerken en de prijs van de wagens. Ten derde ontbreken in het OVG meerdere

variabelen die rechtstreeks betrekking hebben op de activiteiten die verplaatsingen genereren. Ten vierde is de studie beperkt tot gezinnen die effectief eigenaar zijn van de gebruikte auto(s), waardoor de steekproef slechts betrekking heeft op een deel van de totale bevolking.

Verspreiding: Algemeen

Eindrapport

Steunpunt Fiscaliteit en Begroting II – Spoor A3a2: Impact Hervorming Verkeersbelasting

Wat drijft gezinnen in hun autokeuze? Een continu-discreet model op gezinsniveau

Laurent Franckx, Hans Michiels

Studie uitgevoerd in opdracht van: Steunpunt Fiscaliteit en Begroting II
2014/TEM/R/15

Februari 2014

VITO NV
Boeretang 200 - 2400 MOL - BELGIE
Tel. + 32 14 33 55 11 - Fax + 32 14 33 55 99
vito@vito.be - www.vito.be

BTW BE-0244.195.916 RPR (Turnhout)
Bank 375-1117354-90 ING
BE34 3751 1173 5490 - BBRUBEBB

Alle rechten, waaronder het auteursrecht, op de informatie vermeld in dit document berusten bij de Vlaamse Instelling voor Technologisch Onderzoek NV ("VITO"), Boeretang 200, BE-2400 Mol, RPR Turnhout BTW BE 0244.195.916. De informatie zoals verstrekt in dit document is vertrouwelijke informatie van VITO. Zonder de voorafgaande schriftelijke toestemming van VITO mag dit document niet worden gereproduceerd of verspreid worden noch geheel of gedeeltelijk gebruikt worden voor het instellen van claims, voor het voeren van gerechtelijke procedures, voor reclame of antireclame en ten behoeve van werving in meer algemene zin aangewend worden

INHOUD

LIJST VAN TABELLEN

LIJST VAN FIGUREN

LIJST VAN AFKORTINGEN

2SLS	Two stage least squares
ASC	Alternatief-specifieke constante
IV	Inclusive value
LL	Waarde van de loglikelihoodfunctie
MNL	Multinomial logit
MPV	Multi purpose vehicle (monovolume)
NL	Nested logit
OLS	Ordinary Least Squares
OVG	Onderzoek Verplaatsingsgedrag
SUV	Sports utility vehicle (auto met 4x4-look)
VARCOVAR	variantie-/covariantiematrix

HOOFDSTUK 1. ALGEMENE INLEIDING

1.1. DOELSTELLING

Deel A3a2 van het Steunpunt Fiscaliteit en Begroting II heeft als doelstelling om de effecten van een hervorming van de verkeersbelastingen na te gaan, met het oog op de vergroening van de autofiscaliteit. Op basis van een verkenning van de beschikbare data werd bij de start van het project beslist om twee benaderingen te gebruiken.

In de eerste benadering worden de beslissingen over het autobezit en autogebruik op het niveau van de individuele gezinnen empirisch geanalyseerd. Dit gebeurt met behulp van een gecombineerd discreet/continu keuzemodel. Hiermee wordt er nagegaan welke factoren een invloed uitoefenen op het aantal auto's en het type auto's in het bezit van een gezin en op het aantal kilometers dat het gezin met die auto's aflegt. De beïnvloedende factoren omvatten gezinskenmerken (aantal personen, leeftijdsklasse, inkomen,…), technische autokenmerken en autokosten (incl. belastingen of subsidies). Dit rapport bespreekt de verschillende stappen in de ontwikkeling van het model.

In de tweede benadering werd een discrete keuze model geschat voor de vraag naar nieuwe auto's per type op basis van gedetailleerde informatie op marktniveau van de auto's die aangekocht worden. Auto's zijn een typisch voorbeeld van gedifferentieerde producten. De methode is gebaseerd op de literatuur van industriële organisatie. In deze benadering wordt de keuze van een bepaald type van auto verklaard aan de hand van de autokenmerken. Op basis van de empirische analyse werd vervolgens een simulatiemodel ontwikkeld dat moet toelaten om de effecten van een hervorming van de verkeersbelastingen te berekenen. De methodologie en de resultaten worden beschreven in Mayeres & Vanhulsel (2014).

1.2. MODELLERINGSBENADERING

De doelstelling van het onderzoek bestaat erin om na te gaan welke factoren het autobezit en -gebruik van gezinnen beïnvloeden, en welke rol de verkeersbelastingen hierin spelen.

Meer bepaald is de doelstelling van dit onderzoek om volgende vragen te beantwoorden:

- Voor gezinnen die 1 auto bezitten: welke kenmerken van deze gezinnen, en van het beschikbare wagenpark bepalen de keuze van het model in het bezit van het gezin?
- Voor gezinnen die 2 auto's bezitten: welke kenmerken van deze gezinnen en van het beschikbare wagenpark bepalen de keuze van de modellen in het bezit van het gezin?
- Welke kenmerken van de gezinnen en van het beschikbare wagenpark bepalen of een gezin 0, 1 of 2 auto's bezit?
- Welke kenmerken van de auto's en van de gezinnen bepalen het aantal afgelegde kilometers?

Dit zal worden bestudeerd met behulp van data van het OVG (Onderzoek Verplaatsingsgedrag), die gecombineerd werden met gegevens van Febiac.

We beschouwen dus geen gezinnen met 3 of meer auto's. Aangezien de onderzoeksvraag betrekking heeft op auto*bezit*, en op het gebruik van auto's in het bezit van de gezinnen, beschouwen we geen gezinnen die 1 of meer bedrijfswagens tot hun beschikking hebben.

We wensen hier ook te benadrukken dat we hier kijken naar het auto*bezit*, en niet naar de beslissing om een auto aan te kopen. We modelleren hier dus in feite twee keuzen: (a) de initiële keuze van een specifieke auto samen met (b) de jaarlijks weerkerende keuze om de auto te behouden. In een ideale situatie zouden we deze keuze modelleren aan de hand van een volledig dynamisch model[1], maar omdat onze gegevens slechts op een jaar betrekking hebben, is dat niet mogelijk.

Uit de vraagstelling blijkt dat het beantwoorden van de hierboven opgesomde vragen twee verschillende econometrische tools vereist:

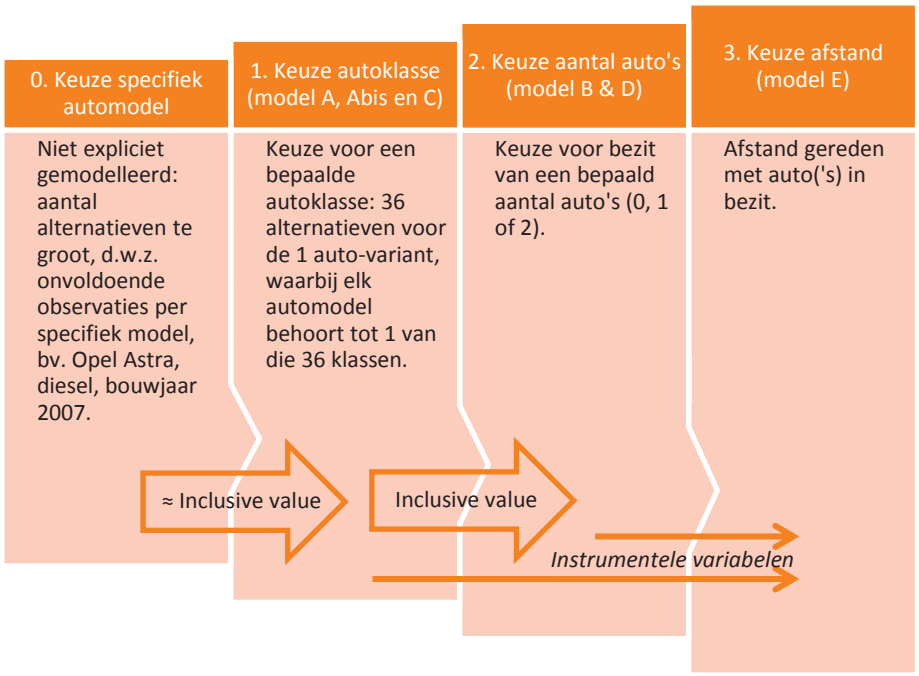
- De keuze van een specifiek automodel en van het aantal auto's zijn discrete variabelen: dit vereist het gebruik van discrete keuze modellen. We zullen hier gebruik maken van Multinomial Logit (MNL) en Nested Logit (NL) modellen – we verwijzen naar de technische bijlage achteraan dit rapport voor een beknopte samenvatting van deze benadering.
- De keuze van het aantal afgelegde kilometers is een continue variabele- daarvoor kunnen we lineaire regressie gebruiken.

Bovendien kunnen de antwoorden op deze vragen niet onafhankelijk van elkaar worden gegeven. De gezinnen zullen immers niet alleen kijken naar de kenmerken van het beschikbare wagenpark om te beslissen welke auto (of auto's) ze zullen kopen, maar ook om te beslissen hoeveel auto's ze zullen kopen: het is immers mogelijk dat zelfs het best mogelijk alternatief uit dat wagenpark minder nuttig is voor het gezin dan helemaal geen auto te bezitten.

Maar ook het aantal afgelegde kilometers kan niet los gezien worden van de autokeuze. Een gezin dat anticipeert dat het een groot aandeel kilometers op jaarbasis zal moeten afleggen, zal immers belang hechten aan andere parameters (brandstofverbruik, comfort) dan een gezin dat slechts een beperkt aantal kilometers moet afleggen. Zoals we hieronder zullen zien, heeft dat een impact op de te volgen modelleringstrategie.

Het model wordt dus opgebouwd aan de hand van een aantal gerelateerde blokken, weergegeven in Figuur 1. Submodel 1, 2 en 3 worden achtereenvolgens geschat om tot een geïntegreerd model te komen. Merk op dat submodel 1 en 2 elk nog zijn samengesteld uit een aantal afzonderlijke modellen, om rekening te kunnen houden met de specificiteiten van elke dataset. Zo valt bv. submodel 2 uiteen in twee modellen B en D: het eerste diende om de keuze tussen 0 en 1 auto's te modelleren, terwijl het tweede ook rekening houdt met het 2-auto-alternatief. De lettercodering in Figuur 1 is ook het systeem dat we aanhouden in de rest van de tekst.

[1] Een klassieke referentie is Hensher et al. (1992).



Figuur 1 Modelstructuur

1.2.1. SUBMODEL 0

Submodel 0 bevat de keuze voor een specifiek automodel (bv. Opel Astra diesel met bouwjaar 2007). Dit model wordt niet als dusdanig door ons gemodelleerd. Daarvoor bestaan twee redenen (zie Train, 1993, p. 142):

- Indien we de vraag modelleren op het niveau van de individuele merken en modellen, kunnen we ons alleen uitspreken over de merken en modellen die waargenomen zijn in de gebruikte steekproef (in dit geval, het OVG). Het aantal observaties binnen het OVG ligt echter te laag in vergelijking met het totaal aantal beschikbare modellen. Veel automodellen die beschikbaar waren in het jaar van de steekproef worden in het OVG zelfs helemaal niet waargenomen. Bovendien kunnen we dan niets zeggen over de keuze van nieuwe modellen in de toekomst.
- Het groot aantal verschillende merken en modellen (meer dan 30000) overschrijdt het maximaal aantal alternatieven dat onze econometrische software aankan.

Vandaar opteren we voor een alternatieve manier van werken.

1.2.2. SUBMODEL 1

In plaats van het model te schatten op het niveau van de individuele modellen, schatten we het model op het niveau van autoklassen (submodel 1). Een autoklasse is gedefinieerd door een carrosserietype, brandstoftype en bouwjaarklasse, bv. een kleine middenklasser op diesel van bouwjaar >2005 (zie ook Tabel 4).

Elk individueel beschikbaar model (dus niet alleen de modellen die effectief gekozen werden) wordt dan toegewezen aan een bepaalde autoklasse.

Qua structuur ziet men direct een analogie met de structuur van een Nested Logit model waarbij de “nests” overeenkomen met de autoklassen. In een Nested Logit model wordt het nut van een gegeven nest gemodelleerd als de som van twee componenten:

- Het nut dat voortvloeit uit elementen die gemeenschappelijk zijn aan alle elementen van de beschouwde nest.
- De verwachte waarde van het maximaal nut dat men kan halen uit de keuze van een van de elementen uit de nest. Immers, als een gezin kiest voor een auto uit een gegeven autoklasse, dan kiest het niet voor de “gemiddelde” auto uit deze klasse, maar voor de auto die het grootste nut oplevert voor het gezin. In een NL model wordt deze verwachte waarde voorgesteld aan hand van de zogenaamde “Inclusive value” (IV).

Onze benadering bouwt verder op een fundamenteel inzicht van McFadden (1978): indien het aantal alternatieven in een gegeven nest te groot wordt om het model te kunnen schatten op het niveau van de individuele alternatieven, dan kunnen we de IV benaderen aan de hand van volgende variabelen:

- het aantal onderliggende automodellen binnen elke autoklasse
- de gemiddeldes, varianties en covarianties van de kenmerken van de onderliggende automodellen binnen elke autoklasse

Dit model incorporeert dus informatie uit het voorgaande model door de “inclusive value” te benaderen. De intuïtie achter dit resultaat is dat, indien de onderliggende variabelen multivariaat normaal verdeeld zijn, de informatie over de gemiddeldes, de varianties en de covarianties volstaat om de verwachte waarde van het maximaal nut te benaderen.

Dus, in onze benadering hebben we voor elke autoklasse de gemiddelde waarde en de covariantiematrix berekend voor alle kenmerken van de onderliggende modellen. Deze waarden en het aantal onderliggende modellen zijn dan gebruikt voor de schatting van het model op het niveau van de autoklassen.

Deze benadering lost de twee bovenvermelde problemen op:

- Indien er nieuwe modellen op de markt komen, dan kunnen we ze classificeren binnen de bestaande autoklassen.
- Het aantal alternatieven wordt sterk beperkt, waardoor er zich ook geen informatica-technische problemen voordoen.

Een mogelijke beperking van deze benadering is dat de definitie van de autoklassen noodzakelijkerwijze gebaseerd is op beschikbare informatie in het OVG, die niet altijd alle relevante elementen weergeeft. Het is bijvoorbeeld niet mogelijk om te detecteren welke gezinnen een break gekozen hebben. Bijgevolg bevatten autoklassen zoals “stadswagen”, “kleine middenklasse”, “grote middenklasse” zowel de “break” versie als het basismodel. Het is dus mogelijk dat, voor bepaalde niet-geobserveerde kenmerken, de variantie binnen een gegeven autoklasse groter is dan tussen de autoklassen onderling. Dit zal dan een impact hebben op de globale voorspellende waarde van het model.

Merken we tenslotte op dat submodel 1 ook moet toegepast worden op de gezinnen met twee auto’s. Deze variante stelt een aantal specifieke problemen, die we verder (zie bespreking model C) zullen bespreken.

1.2.3. SUBMODEL 2

Vervolgens wordt de informatie uit submodel 1 gebruikt in de modellering van submodel 2, dat de kans op het bezitten van een bepaald aantal auto's weergeeft. De Inclusive Value[2] hoeft nu niet meer benaderd te worden. Ze kan immers exact berekend worden door de logsom te nemen van het resulterende nut uit het geschatte submodel A.

1.2.4. SUBMODEL 3

Als laatste onderdeel schatten we submodel 3: in dit submodel wordt de beslissing gemodelleerd voor het rijden van een bepaald aantal kilometers met elk van de wagens in het bezit van het gezin. In tegenstelling tot de vorige submodellen wordt hier dus een continue keuze gemodelleerd in plaats van een discrete keuze. De afgelegde afstand hangt af van zowel de gekozen autoklasse als het gekozen aantal auto's. Zoals hierboven reeds aangehaald, wordt de keuze van de afstand bovendien mee bepaald door een aantal variabelen (zoals het brandstofverbruik van de auto), die zowel een impact hebben op de afstandskeuze als (voorafgaand) op de keuze van de autoklasse en het aantal auto's. Deze verklarende variabelen zijn dus endogeen aan het model, en het gebruik van Ordinary Least Squares (OLS) zou dus leiden tot inconsistente schatters. Een instrumentele variabele-benadering is hier aangewezen. Dit wordt verder uitgewerkt in de bespreking van model E.

1.3. MODELSTAPPEN

We onderscheiden in dit rapport over het gezinsmodel een aantal logisch geordende blokken.

We starten met een beknopt overzicht van de **gebruikte data**. De nadruk ligt hier op een korte bespreking van de ruwe data en de belangrijkste bewerkingen die nadien gebeurden.

Daarna volgt een bespreking van de **geschatte modellen**.

In een eerste soort model zoeken we voor gezinnen met exact 1 auto in bezit een verklaring voor de keuze voor één specifieke autoklasse uit 36 mogelijke autoklassen (model A), gevolgd door de keuze tussen 0 en 1 auto (model B). Merk op dat voor het schatten van model B ook werd rekening gehouden met de gegevens van gezinnen zonder auto. Modellen A en B worden verderop besproken onder de sectie "Model voor gezinnen met maximum 1 auto".

Omdat we bij de keuze van de autoklassen voor gezinnen met 2 auto's in bezit (model C) gebonden zijn door een groot aantal mogelijke combinaties van 2 autoklassen, moesten we noodgedwongen het aantal mogelijke autoklassen beperken tot 30 (i.p.v. 36). Om consistentie te waarborgen doorheen de ganse modelketen werd daarom model A herschat als model A bis, dus met dit verschil dat er nu nog slechts keuze bestond uit 30 autoklassen. Voor gezinnen met niet 1 maar 2 auto's in bezit kon daarna model C geschat worden om de keuze voor een specifiek paar van autoklassen te verklaren.

[2] In dit geval: de verwachte waarde van het maximaal nut dat kan gehaald worden uit het bezit van 1 auto (in de nest met 1 auto) en van het maximaal nut dat kan gehaald worden uit het bezit van 2 auto's (in de nest met 2 auto's). Het "nut" van het bezitten van 0 auto's wordt genormaliseerd op 0.

Door de gezinnen zonder auto in bezit nog toe te voegen aan de dataset, en door gebruik te maken van de schattingsresultaten van model A bis en model C, kon model D worden geschat dat de keuze verklaart tussen het bezitten van 0 auto's, 1 auto of 2 auto's.

Tenslotte wordt de jaarlijks afgelegde afstand per auto geschat, zowel voor gezinnen met 1 auto in bezit als voor gezinnen met 2 auto's in bezit. Modellen A bis, C en D worden verderop besproken onder de sectie "Model voor gezinnen met maximum 2 auto's".

Tot slot vatten we de belangrijkste **conclusies** samen van het gezinsmodel dat in dit rapport wordt besproken.

HOOFDSTUK 2. GEBRUIKTE DATA

Verschillende databronnen werden gecombineerd om tot de tabellen te komen die gebruikt werden voor de schattingen. Onze 2 belangrijkste bronnen vormden de bevragingen binnen het Onderzoek Verplaatsingsgedrag (OVG) en de Febiac-autocatalogi.

- Uit het OVG werd versie 4.1 t.e.m. 4.3 gebruikt. In totaal namen in deze 3 versies van OVG 5046 gezinnen deel. Merk op dat niet alle gevraagde info door alle gezinnen werd ingevuld; zo merken we dat gevoelige info (bv. het inkomen) bij tamelijk wat gezinnen ontbreekt[3]. Een uitgebreid overzicht van beschrijvende statistieken over de verplaatsingen en de mobiliteitskenmerken van deze datasets kan men geordend per OVG-versie (4.1, 4.2 en 4.3) terugvinden op de website van MOW (Vlaamse Overheid – Departement MOW, 2014). Uit het OVG-onderzoek gebruiken wij info rond de belangrijkste gezinskenmerken (inkomen, gezinssamenstelling, beroepssituatie, etc.), de voertuigen in bezit (o.a. merk, model, brandstof en bouwjaar) en een steekproef van de afgelegde trips. Specifiek voor de socio-demografische kenmerken zien we dat er in OVG heel wat variabelen werden verzameld met een discreet karakter, bv. een inkomensniveau van niveau 1 tot 6 of een geslacht met niveau 1 of 2 (man of vrouw). Om hier terdege mee te kunnen omgaan in de schattingen werden zulke variabelen omgevormd tot dummy-variabelen[4].

  Omdat 'gewone' dummies het nadeel hebben dat het referentieniveau kan verward worden met de constante term van de nutsfunctie, opteren wij in deze studie voor het gebruik van effects-codering. Het enige verschil tussen effects-codering en 'gewone' dummy-codering zit in de waarde van de dummy voor het referentieniveau: waar die nog 0 was bij een gewone dummy neemt het referentieniveau altijd de waarde -1 aan bij effects-codering. Een voorbeeld werkt verduidelijkend. Stel dat we beschikken over een attribuut dat in OVG drie mogelijke waarden kan aannemen: oud, medium of jong. Om dit in 'gewone' dummy-vorm te coderen kiezen we een referentieniveau (stel: medium) en creëren we 1 dummy voor elk van de 2 overblijvende niveaus: dummy_oud en dummy_jong. De eerste neemt waarde 1 aan indien het oorspronkelijk attribuut 'oud' was, en 0 anders. De tweede neemt waarde 1 aan indien het oorspronkelijk attribuut 'jong' was, en 0 anders. Als we effects-codering toepassen, krijgen de dummyvariabelen de waarde -1 voor het gekozen referentieniveau (medium). Voor de rest blijft alles bij het oude. Zie Tabel 1.

Tabel 1 Illustratie dummy-codering versus effects-codering

Attribuut-niveau	Dummy_oud	Dummy_jong	Effects-code_oud	Effects-code_jong
Oud	1	0	1	0
Medium	0	0	-1	-1
Jong	0	1	0	1

[3] Indien een ontbrekende waarde voorkomt voor een variabele die wordt opgenomen in het geschat model, wordt dat gezin uit de schatting geweerd (zie verder).

[4] Een dummy neemt de waarde 1 of 0 aan, afhankelijk van de waarde van de oorspronkelijke variabele van dewelke hij is afgeleid.

- Uit de aangekochte databank van Febiac haalden we de technische kenmerken en prijzen van alle wagens die in België nieuw verkocht zijn sinds 1990. Het detailniveau voor de auto's gaat hier veel verder dan in OVG, met bijvoorbeeld info over prestatiekenmerken (bv. vermogen), afmetingen, uitrustingsniveaus, catalogusprijzen, etc. In totaal werden 33403 verschillende automodellen[5] overgehouden uit Febiac. Door het verschillend niveau van detail t.o.v. OVG werd elk van deze 33403 automodellen toegewezen aan een bepaalde autoklasse (zie modelstructuur hierboven), zodat de keuze van elk gezin voor een bepaalde autoklasse eenduidig was.
  Er werden aardig wat verbeteringen aangebracht aan de oorspronkelijke data van Febiac, omdat bijvoorbeeld het verbruik, afmetingen, etc. ontbraken of omdat de data soms zelf flagrant fout waren (bv. brandstoftype voor bepaald model dat nooit met die brandstof op de markt is gekomen). Op basis van de Febiac-data werden nog een aantal eigen berekeningen gedaan, bv. om de benaderende tweedehandsprijs te kennen (o.b.v. een aantal steekproeven op de tweedehandsrichtprijzen-databank van Autogids).

[5] Een automodel is 'verschillend' van alle andere automodellen indien het ervan verschilt op minstens 1 van volgende kenmerken: brandstoftype, bouwjaar, merk, model, aantal deuren, aantal zitplaatsen, cilinderinhoud, vermogen en verbruik.

HOOFDSTUK 3. KEUZEMODEL VOOR GEZINNEN MET MAXIMUM 1 AUTO

We geven in deze sectie een overzicht van de modellen die geschat werden voor de gezinnen met maximaal 1 auto. Enerzijds wordt de keuze voor een specifieke autoklasse geschat voor gezinnen met juist 1 auto in bezit (model A). We schatten zowel een MNL- als een NL-variant. Anderzijds schatten we een model dat de keuze tussen het bezitten van 0 of 1 auto verklaart (model B).

3.1. MODEL A: KEUZE AUTOKLASSE VOOR GEZINNEN MET 1 AUTO

3.1.1. MULTINOMIAL LOGIT (MNL) MODEL

In dit deel wordt het finale MNL-model besproken. Bij dit soort modellen gaan we ervan uit dat alle alternatieven even sterk op elkaar lijken, hetgeen niet noodzakelijk strookt met de realiteit, zie hoger bij de bespreking van de modelstructuur. Het grote voordeel van dit soort modellen is dat het berekenen van kansen dat bepaalde alternatieven gekozen worden, eenvoudig is.

Voor het modelleren van de keuze van de autoklasse werden alle beschikbare automodellen ingedeeld in 36 groepen[6]. Deze groepen werden ingedeeld op basis van volgende criteria: 3 leeftijdsklassen van de auto (bouwjaar <2001, 2001-2005, en >2005[7]), 6 carrosserietypes (stadswagens, kleine middenklassers, grote middenklassers & executives, cabrio & coupé, SUV, MPV[8]) en 2 brandstoftypes (benzine en diesel).

Door het model te schatten werd het belang van bepaalde variabelen op de keuze voor een welbepaald alternatief geëvalueerd. De geschatte coëfficiënten zeggen iets over de invloed van de variabelen op het nut van een bepaald alternatief voor de gezinnen, en daaruit voortvloeiend ook de kans dat een bepaald gezin de keuze maakt voor een bepaald alternatief.

De schattingsresultaten van het MNL-model worden hieronder weergegeven in Tabel 2. In Tabel 3 kan de lezer de definities terugvinden van de verschillende verklarende variabelen.

[6] Merk op: in model A bis en model C beperken we ons tot 30 groepen i.p.v. 36 groepen (autoklassen)
[7] Deze leeftijdsklassen komen overeen met de volgende groepen van Euronormen: ≤ Euro 2 (<2001), Euro 3 (2001-2005) en ≥ Euro 4 (>2005). Deze indeling is consistent met de veronderstellingen die men maakt bij de berekening van de Vlaamse BIV, indien de exacte Euronorm van een voertuig niet gekend is.
[8] Voorbeelden van elk van deze carrosserietypes zijn: Volkswagen Polo (stads), Volkswagen Golf (kleine middenklasser), Volkswagen Passat (grote middenklasser), Audi A6 of A8 (executive), BMW Z4 (cabrio), Audi A5 (coupé), Volkswagen Touareg (Sports utility vehicle of "SUV") en Volkswagen Touran (Multi purpose vehicle of MPV).

→ **Interpretatie schattingsresultaten**

Algemeen

Om te achterhalen in welke mate ons geschatte model de variatie in de geobserveerde keuzes verklaart, kunnen we een pseudo-R^2 (zie voetnoot[9]) berekenen. In de literatuur bestaat er aardig wat onenigheid over de zin van het berekenen van deze pseudo-R^2. Sommige auteurs vinden deze maatstaf enkel nuttig om de vergelijking te maken tussen gelijkaardige modellen onderling, die op dezelfde data en op dezelfde keuzes geschat zijn (IDRE – Institute for Digital Research and Education, 2014; Hensher et al., 2005). Hierbij bestaat het gevaar dat de veelal lage waardes van een pseudo-R^2, een publiek dat wel vertrouwd is met een R^2 voor lineaire schattingen, in verwarring kan brengen. Ander auteurs komen scherper uit de hoek en beschouwen de pseudo-R^2 als een slechte maatstaf voor 'goodness of fit'. Hoogstens kan het als een beschrijvende statistiek beschouwd worden (Econometric Software, Inc., 2014). In andere gevallen raadt men zelfs aan om helemaal te stoppen om die maatstaf te rapporteren, juist omdat die pseudo-R^2 voorkomt in zoveel vormen en mogelijke interpretaties dat het verwarrend wordt (Hoetker, 2007; Kaufman, 2014).

Voor wat het waard is, rapporteren we hierna toch de pseudo-R^2 voor de geschatte modellen. We kijken er daarbij wel nauwgezet op toe dat we voor het basismodel exact dezelfde keuzes en steekproef gebruiken als bij het geschatte model.

Deze pseudo-R^2 is dus niet te vergelijken met de 'normale' R^2 die men gewoonlijk gebruikt om een Ordinary Least Squares (OLS)-schatting van een lineaire regressie te evalueren. Voor discrete-keuzemodellen berekent men een pseudo-R^2 door de loglikelihood-waarde (LL-waarde) van het geschat model te vergelijken met de LL-waarde van een restrictief model (ook 'basismodel'). Afhankelijk van de keuze van basismodel zal de pseudo-R^2 een andere waarde aannemen, dus het is belangrijk steeds te vermelden op welk basismodel de pseudo-R^2 berekend werd:

- Een eerste optie is om een <u>basismodel met gelijke aandelen</u> te schatten: in dat geval houdt men geen rekening met de informatie in de data en berekent men (het logaritme van) de waarschijnlijkheid dat men de waargenomen steekproef zal waarnemen, onder de hypothese dat alle alternatieven in de populatie een zelfde aandeel hebben. Concreet verloopt de schatting van dit basismodel door een generische constante op te nemen in de nutsfunctie van elk van de alternatieven, behalve 1.
- Een tweede optie is om een <u>basismodel met marktaandelen</u> te schatten: in dit geval gebruikt men de informatie aanwezig in de data en houdt men rekening met de frequentie van voorkomen van elk alternatief in de steekproef. Men berekent dus (het logaritme van) de waarschijnlijkheid dat men de waargenomen steekproef zal waarnemen, onder de hypothese dat de waargenomen frequenties gelijk zijn aan de frequentie in de populatie. Concreet wordt zo'n basismodel geschat door een alternatief-specifieke constante (ASC)[10] op te nemen in de nutsfunctie van elk van de alternatieven, behalve 1. In wat volgt

[9] Een pseudo-R^2 wordt berekend als $R^2 = 1 - \frac{LL_{geschat\ model}}{LL_{basismodel}}$, waarbij het 'basismodel' overeenkomt met een restrictie van het geschatte model (zie verderop in de tekst voor de 2 mogelijkheden). LL is waarde van de loglikelihood-functie van het geschatte model en wordt bij een modelschatting geminimaliseerd (want LL is negatief): op die manier wordt voor elke opgenomen variabele een coëfficiënt geschat die de kans maximaliseert dat een gezin uit de steekproef het alternatief zal kiezen dat het in werkelijkheid gekozen heeft.

[10] Een alternatief-specifieke constant of ASC wordt in de nutsfuncties opgenomen om de invloed van niet-geobserveerde factoren (i.e., factoren die niet als verklarende variabelen kunnen opgenomen konden worden) met een impact op de keuzebeslissing te vertegenwoordigen. De ASC geeft de gemiddelde invloed weer van de niet-geobserveerde factoren op de keuze voor een bepaald alternatief.

vermelden we de pseudo-R^2 volgens de tweede benadering (basismodel met marktaandelen) tenzij anders wordt vermeld.

Voor onderstaand MNL-model vonden we een pseudo-R^2 van 0.0163. Het geschatte MNL-model presteert duidelijk beter dan het basismodel volgens de LL ratio-test[11], waarbij we een teststatistiek (nl. het dubbel verschil tussen de LL-waarde van beide modellen) vergelijken met een kritische Chi^2-waarde die rekening houdt met het verschil in aantal geschatte parameters:

$$-2 \times \left(LL_{basismodel} - LL_{geschat\ model}\right) = -2 \times (-5535.614 + 5445.446) = 180.336$$
$$> \chi^2_{aantal\ nieuwe\ params\ geschat\ in\ geschatte\ model} = \chi^2_{42-35} = 14.067$$

Parameterschattingen

Tabel 2 MNL-variant van model A

Verklarende variabele		Geschatte coëfficiënt	t-statistiek	95% betrouwbaarheidsinterval	
1.	**PR11YLOW**	-0.22655E-04***	3.39	-0.35739E-04	-0.95705E-05
2.	**PR11YHIG**	0.26534E-04***	2.77	0.77305E-05	0.45337E-04
3.	**VOLMEDGZ**	-0.04105**	2.14	-0.07862	-0.00349
4.	**VOLGRGZ**	0.17806***	7.00	0.12821	0.22791
5.	**VOLYLOW**	-0.08156***	5.35	-0.11144	-0.05168
6.	**KWGHFDYG**	-0.00556***	2.86	-0.00937	-0.00174
7.	**KWGHFDOD**	0.00297*	1.79	-0.00028	0.00623
8.	**TRAFFTAX**	-0.00111	0.52	-0.00536	0.00314
9.	**FUELCOST**	-0.47033***	3.52	-0.73207	-0.20860
10.	**LOGRC**	1.94880***	3.85	0.95786	2.93973

*MNL-model geschat op 36 alternatieven (autoklassen). Aantal observaties gebruikt voor de schatting: 1740 van de 2342 gezinnen (602 gezinnen met een onbekende waarde voor 1 van de opgenomen variabelen werden geschrapt). De bekomen log likelihood-waarde bedraagt -5445. Een significantie van een parameter op het 1%, 5% of 10%-niveau wordt aangeduid met resp. ***, ** of *.*

Tabel 3 Verklaring variabelen opgenomen in het MNL-model

Verklarende variabele		Definitie
1.	**PR11YLOW**	Interactie tussen de **totale autoprijs** (aankoopprijs, BIV en CO2-premie, allen rekening houdend met de leeftijd van de auto) en een effects-coded **dummy voor lage** (netto ≤2000 EUR/mnd) **gezinsinkomens**
2.	**PR11YHIG**	Interactie tussen de **totale autoprijs** (aankoopprijs, BIV en CO2-premie, allen rekening houdend met de leeftijd van de auto) en een effects-coded **dummy voor hoge** (netto >4000 EUR/mnd) **gezinsinkomens**
3.	**VOLMEDGZ**	Interactie tussen **autovolume** (lengte x breedte x hoogte) en een effects-coded **dummy voor middelgrote gezinnen** (3 of 4 gezinsleden)
4.	**VOLGRGZ**	Interactie tussen **autovolume** (lengte x breedte x hoogte) en een effects-coded **dummy voor grote gezinnen** (>4 gezinsleden)
5.	**VOLYLOW**	Interactie tussen **autovolume** en een effects-coded **dummy voor lage** (netto ≤2000 EUR/mnd) **gezinsinkomens**
6.	**KWGHFDYG**	Interactie tussen **autovermogen** en een effects-coded **dummy voor jong gezinshoofd** (<40jr)
7.	**KWGHFDOD**	Interactie tussen **autovermogen** en een effects-coded **dummy voor oud gezinshoofd** (≥65jr)

[11] De nulhypothese (H_0) van deze test houdt in dat het geschatte model geen verbetering vormt van het basismodel.

Verklarende variabele	Definitie
8. **TRAFFTAX**	**Jaarlijkse verkeersbelasting** te betalen in 2011
9. **FUELCOST**	**Brandstofkost** (EUR/100km)
10. **LOGRC**	**LOGRC = LOG(N_i/N)**, d.i. de link met het onderliggend (niet-geschatte) model voor keuze van een specifiek automodel (i.t.t. de autoklasse)

We bespreken nu in detail de resultaten van het weerhouden MNL-model (zie Tabel 2).
Alle geschatte parameters zijn minstens significant op het 10%-niveau, behalve de parameter horende bij variabele 8. TRAFFTAX. Afgezien van die laatste en variabele 7 (KWGHFDOD) zijn ze alle minimaal significant op het 5%-niveau.

We nemen 2 autoprijsvariabelen op in interactie met het gezinsinkomen (variabele 1 & 2). We kunnen immers redelijkerwijze verwachten dat gezinnen met een lager inkomen sterker worden beïnvloed door de prijs dan de gezinnen met middelhoge en hoge inkomens. Uit de resultaten blijkt dat voor de gezinnen met een laag inkomen, de kans dat een bepaalde autoklasse gekozen wordt, negatief wordt beïnvloed indien de prijs van de autoklasse stijgt. Voor gezinnen met een hoog inkomen daarentegen is deze coëfficiënt positief, hetgeen impliceert dat hogere prijzen gepaard gaan met een hoger nut, hetgeen verklaard zou kunnen worden door een soort 'snob-effect'. Omdat voor de inkomensdummy effects-coding werd toegepast, kunnen we afleiden dat voor gezinnen met een gemiddeld inkomen deze interactievariabele een al dan niet significante coëfficiënt[12] heeft van ca. +0.22655-0.26534 = -0.03879E-04.

Verder nemen we het autovolume op in interactie met de gezinsgrootte (variabele 3 & 4) omdat we a priori verwachten dat grotere wagens van hoger nut zullen zijn voor grote gezinnen dan voor kleinere gezinnen. De geschatte coëfficiënten impliceren enerzijds dat grote gezinnen duidelijk een grotere auto weten te waarderen (positieve coëfficiënt), terwijl gezinnen met 1 of 2 gezinsleden dan weer liever een kleine auto blijken te hebben (niet-getoonde coëfficiënt van ca. -0.137, al dan niet-significant). Anderzijds stellen we vast dat gezinnen met 3 of 4 gezinsleden een kleinere kans hebben om een bepaalde autoklasse te kiezen wanneer het gemiddeld volume daarvan stijgt. Dit laatste kan een ietwat vreemde observatie lijken, maar we mogen niet vergeten dat de coëfficiënt van de 3de variabele de invloed omvat van het autovolume, buiten de invloeden van alle andere opgenomen variabelen. Zo zal bv. FUELCOST (zie verder) gecorreleerd zijn met het autovolume[13], en zo een deel van de invloed van een wijzigend autovolume incorporeren. Een andere mogelijke verklaring is dat het autovolume[14] geen perfecte weergave van de beschikbare binnenruimte van een auto.

Daarnaast werd ook nog de mogelijkheid voorzien dat de gekozen autogrootte afhangt van het gezinsinkomen. Bij opname in het model bleek enkel variabele 5 (VOLYLOW) een significant effect te hebben (VOLYHIGH werd daarom niet weerhouden in dit finale MNL-model). Uit de geschatte coëfficiënt leiden we af dat voor gezinnen met een laag inkomen een groter autovolume leidt tot een lager nut[15]. Aangezien VOLYHIGH niet-significant verschillend van 0 geschat werd, impliceert dit meteen ook dat de gemiddelde inkomens-gezinnen een positieve invloed ondervinden van een groter volume (niet-getoonde coëfficiënt van +0.08156).

[12] Het is echter niet mogelijk om een uitspraak te doen over het significantieniveau van deze variabele. Zelfs indien variabele 1 en 2 beide normaal verdeeld zouden zijn, kan men de betrouwbaarheidsintervallen niet zomaar bij elkaar optellen.
[13] Voor de 2342 gezinnen die deel uitmaken van de oorspronkelijke dataset vonden we een correlatie van 0.32 tussen FUELCOST en het autovolume.
[14] Autovolume werd berekend door de buitenmaten met elkaar te vermenigvuldigen (LxBxH). Hierdoor kreeg de breakversie van een bepaald automodel hetzelfde autovolume als de sedanversie van dat model.
[15] Dit is ook weer voor een stuk te wijten aan de hoge correlatie tussen volume en autoprijs (0.25).

Om in het model ook rekening te houden met het prestatiepotentieel van de gekozen wagen, bleek de opname van het vermogen de beste resultaten te geven (vergeleken met cilinderinhoud of topsnelheid). Bovendien bleek enkel het vermogen in interactie met de leeftijd van het gezinshoofd[16] (variabele 6 & 7) iets aan het model bij te dragen. Het is op zich niet zo vreemd dat het vermogen als hoofdeffect (d.w.z. niet-geïnterageerd met een andere variabele zoals leeftijd) geen significante bijdrage levert, aangezien we er vanuit gaan dat variabele 9. FUELCOST zeker in het model moet worden opgenomen. Deze brandstofkost is immers sterk gecorreleerd met het vermogen. Het opnemen van interacties tussen het vermogen en het inkomen lijken op het eerste zicht veel intuïtiever, maar werden hier niet-significant bevonden, vermoedelijk gedeeltelijk te wijten aan het feit dat de interactie prijs-inkomen reeds in het model was opgenomen. De geschatte coëfficiënten van de interactietermen vermogen-leeftijd gezinshoofd wijzen op een hoger nut voor oudere gezinshoofden (65+ jaar) en een lager nut voor jonge gezinshoofden (-40 jaar) bij een stijging van het autovermogen. Voor gezinshoofden uit de middelgroep (40-64 jaar) is de geschatte coëfficiënt dan +0.00259. Verder (zie Tabel 7) zullen we aantonen dat de coëfficiënten van variabele 6 en 7 niet langer statistisch significant zijn bij een nested logit model, wat er op kan wijzen dat dit resultaat eerder voortvloeit uit een verkeerde specificatie van het model.

Kijken we naar variabele 8., dan zien we dat de invloed van de verkeersbelasting insignificant blijkt te zijn. Deze bevinding verdient een woordje uitleg. Herinner u dat we alle auto's hebben ingedeeld in één van 36 mogelijke autoklassen. Die definitie van onze autoklassen bleek, gegeven de beschikbare data, een goed compromis tussen voldoende observaties per autoklasse en zoveel mogelijk variatie tussen de autoklassen. Nu is het wat betreft dit laatste punt natuurlijk irrealistisch om te verwachten dat er voldoende variatie zal bestaan tussen de autoklassen op *alle* mogelijke variabelen. Inderdaad, voor wat betreft de jaarlijkse verkeersbelasting blijkt de variantie tussen de autoklassen onderling (=inter-klasse variantie) slechts in 21 van de 36 gevallen groter te zijn dan de variantie binnen de autoklassen (=intra-klasse variantie). Vergelijken we dit met bv. de totale autoprijs-variabele, dan zien we dat in dat geval

$$\sigma^2_{inter-klasse} > \sigma^2_{intra-klasse}$$

voor 30 van de 36 autoklassen. Het feit dat de intra-klasse variantie voor TRAFFTAX zo hoog is voor 15 van de 36 gevallen vormt dus vermoedelijk de belangrijkste reden waarom deze variabele geen verklaringskracht heeft in het geschatte model.

Voor de 9de variabele (brandstofkost) werd een negatieve coëfficiënt gevonden. Dit wijst op een daling van het nut van een bepaalde autoklasse bij een stijging van de gemiddelde brandstofkost van die klasse.

Tot slot hebben we nog de coëfficiënt LOGRC (in de literatuur ook wel θ genoemd: dit is één van de links met het onderliggende submodel 0 voor de keuze van een specifiek automodel, zie de sectie "Modelleringsbenadering"). Dit is de parameter die hoort bij de variabele log(N_i/N), of het logaritme van de verhouding tussen het aantal automodellen horend bij een bepaalde autoklasse en het totaal aantal automodellen over alle autoklassen. Om consistent te zijn met de assumptie van nutsmaximalisatie moet deze coëfficiënt tussen 0 en 1 liggen (Hensher et al., 2005). De puntwaarde van de schatting zelf is weliswaar hoog, maar is niet significant verschillend van 1. We kunnen dus niet de hypothese verwerpen dat de coëfficiënt in het [0;1] interval ligt. Het feit dat deze parameter significant groter dan 0 geschat wordt, wijst er in ieder geval op dat het nut stijgt

[16] Het geboortejaar opnemen als continue variabele in plaats van als dummy, bleek geen significante schatters op te leveren.

wanneer het aantal automodellen in een bepaalde autoklasse groeit (en dus als de keuzemogelijkheden binnen deze klasse toenemen), ceteris paribus. De LOGRC variabele vormt dus de belangrijke link met het onderliggende (niet-geschatte) model voor de keuze voor een specifiek automodel. Het schatten van een model op specifiek automodelniveau zou enorm veel alternatieven opleveren, hetgeen praktische bezwaren zou opleveren tijdens de schatting. Vandaar dat dit model op specifiek automodelniveau impliciet wordt meegenomen in het model voor de keuze van de autoklasse.

Merk op dat er in deze MNL-variant van model A geen rekening werd gehouden met de varianties en covarianties op de kenmerken van de onderliggende automodellen binnen elke autoklasse (zie hoger). In de NL-variant van model A (hieronder) houden we wel rekening met die elementen.

Elasticiteiten

Bovenop deze parameterschattingen rapporteren we hierna ook nog de belangrijkste elasticiteiten. Aan de hand van deze resultaten willen we vragen kunnen beantwoorden zoals: hoe verandert de kans dat een bepaalde autoklasse gekozen wordt indien we attribuut x van die autoklasse wijzigen (directe elasticiteit)? Verder kan het ook interessant zijn om te bekijken hoe de kans op autoklasse 1 wijzigt indien een attribuut van autoklasse 2 wordt aangepast (kruiselingse elasticiteit). Hiertoe berekenen we puntelasticiteiten. Het zal blijken dat Tabel 4, die de definitie van de 36 autoklassen bevat, hierbij een handig hulpmiddel is. Merk op dat bij een keuzemodel de elasticiteiten gezins-specifiek zijn, maar om de resultaten overzichtelijk weer te geven berekenen we een gewogen gemiddelde[17] elasticiteit voor alle gezinnen in de steekproef.

De gevonden elasticiteiten voor een model als het onze zijn niet direct vergelijkbaar met de literatuur omdat het steeds een wijziging in kansen betreft, en elke elasticiteit dus een trade-off veronderstelt tussen de keuze voor een bepaalde klasse ten opzichte van alle andere mogelijke klassen. Resultaten tussen studies vergelijken onderstelt dan ook impliciet dat de definitie van de alternatieven over beide studies dezelfde is (de Jong et al., 2004). Voor studies die de impact bestuderen van een bepaald beleid op het totale autobezit (of –gebruik), is het rapporteren van elasticiteiten voor een gewijzigde vraag (i.p.v. gewijzigde kansen) wel mogelijk (zie bv. de Jong et al., 2001 of de Jong et al., 2009).

Tabel 4 Definitie 36 autoklassen

Autoklasse	Carrosserietype	Bouwjaar	Brandstof
1.	Stads	< 2001	benzine
2.	Stads	< 2001	diesel
3.	KleineMiddenklasse	< 2001	benzine
4.	KleineMiddenklasse	< 2001	diesel
5.	GroteMiddenklasse + executive	< 2001	benzine
6.	GroteMiddenklasse + executive	< 2001	diesel
7.	Cabrio + coupé	< 2001	benzine
8.	Cabrio + coupé	< 2001	diesel
9.	SUV	< 2001	benzine
10.	SUV	< 2001	diesel
11.	MPV	< 2001	benzine
12.	MPV	< 2001	diesel

[17] Om de elasticiteiten van de gezinnen te wegen gebruiken we de probability weighted sample enumeration-methode (PWSE), waarbij wordt rekening gehouden met de geschatte kansen voor elk gezin.

Autoklasse	Carrosserietype	Bouwjaar	Brandstof
13.	Stads	2001 - 2005	benzine
14.	Stads	2001 - 2005	diesel
15.	KleineMiddenklasse	2001 - 2005	benzine
16.	KleineMiddenklasse	2001 - 2005	diesel
17.	GroteMiddenklasse + executive	2001 - 2005	benzine
18.	GroteMiddenklasse + executive	2001 - 2005	diesel
19.	Cabrio + coupé	2001 - 2005	benzine
20.	Cabrio + coupé	2001 - 2005	diesel
21.	SUV	2001 - 2005	benzine
22.	SUV	2001 - 2005	diesel
23.	MPV	2001 - 2005	benzine
24.	MPV	2001 - 2005	diesel
25.	Stads	> 2005	benzine
26.	Stads	> 2005	diesel
27.	KleineMiddenklasse	> 2005	benzine
28.	KleineMiddenklasse	> 2005	diesel
29.	GroteMiddenklasse + executive	> 2005	benzine
30.	GroteMiddenklasse + executive	> 2005	diesel
31.	Cabrio + coupé	> 2005	benzine
32.	Cabrio + coupé	> 2005	diesel
33.	SUV	> 2005	benzine
34.	SUV	> 2005	diesel
35.	MPV	> 2005	benzine
36.	MPV	> 2005	diesel

FUELCOST

In onderstaande tabel vind je de directe elasticiteiten voor de variabele FUELCOST voor elk van de 36 autoklassen. Deze elasticiteit geeft dus de procentuele wijziging weer van de kans op autoklasse x, gegeven een prijsstijging van de brandstofkost voor die autoklasse x met 1%. De grote meerderheid van de geschatte elasticiteiten voor FUELCOST liggen tussen -3 en -7, hetgeen bv. voor autoklasse 1 impliceert dat bij een stijging van de gemiddelde brandstofkost van de autoklasse met 1%, de kans dat die autoklasse wordt gekozen daalt met ca. 5%. De vraag van de gezinnen is in deze dus zeer elastisch. We noteren ook enkele uitschieters met elasticiteiten groter (in absolute waarde) dan -7: het betreft hier duidelijk een deel van de autoklassen die van zichzelf al een hoog verbruik hebben (SUV, Coupé/cabrio, en oudere grote middenklasse & executives en MPV's) en waarbij de 1%-stijging van de sowieso al hoge brandstofkost in absolute termen dan ook sterker stijgt dan voor de zuinigere autoklassen (bv. stadswagens).

De kruiselingse elasticiteiten (niet getoond[18]) zijn allemaal positief, hetgeen logisch is want wanneer de brandstofkost van autoklasse x stijgt, zullen alle andere autoklassen interessanter worden en een hogere kans hebben om gekozen te worden. De kruiselingse elasticiteiten voor FUELCOST liggen allen in het interval [+0.0047 ; +0.4443]. Een zeer kleine kruiselingse elasticiteit impliceert een inelastische vraag. De laagste kruiselingse elasticiteiten werden waargenomen bij de autoklassen met een kleine kans om gekozen te worden / klein marktaandeel[19] (wegens een klein aantal onderliggende automodellen), bv. klasse 31. Bijvoorbeeld, wanneer de FUELCOST van

[18] De volledige tabellen met kruiselings elasticiteiten kunnen desgewenst op eenvoudig verzoek bekomen worden.

[19] Zie Hensher et al. (2005) en de vereenvoudiging van Louviere et al. (2000) voor de wiskundige formulering van de kruiselasticiteit. De kruiselasticiteit wordt dus beïnvloed door de a priori kans (marktaandeel) op het alternatief waarvoor een attribuut wordt gewijzigd.

autoklasse 31 met 1% stijgt zal de kans op autoklasse 1 slechts met 0.0037% stijgen, juist omdat autoklasse 31 zo'n klein aantal automodellen omvat (en dus een kleine impact heeft op de andere autoklassen). We vonden de grootste kruiselasticiteiten voor de autoklassen met een groot marktaandeel (wegens grote impact op de andere autoklassen). Bijvoorbeeld: wanneer de FUELCOST voor autoklasse 1 stijgt met 1%, zal de kans op autoklasse 3 met 0.4443% stijgen. Zie Tabel 4 voor de definitie van de klassen.

Tabel 5 Directe elasticiteiten voor FUELCOST, TRAFFTAX en LOGRC in het MNL-model

Autoklasse	Directe elasticiteit FUELCOST	Directe elasticiteit TRAFFTAX	Directe elasticiteit LOGRC
1.	-5.0450	-0.1780	-5.9675
2.	-3.9154	-0.2731	-9.1866
3.	-5.6467	-0.2567	-4.7170
4.	-3.9212	-0.3142	-6.8409
5.	-7.1437	-0.5551	-4.0436
6.	-4.4937	-0.4410	-6.0859
7.	-7.9885	-0.9500	-7.1244
8.	-4.7518	-0.3772	-13.5908
9.	-8.6353	-0.9209	-8.2978
10.	-6.1512	-0.7361	-8.8006
11.	-7.2352	-0.4670	-7.2986
12.	-4.7167	-0.3595	-7.4453
13.	-4.5585	-0.1954	-7.1254
14.	-3.1016	-0.2429	-9.0611
15.	-5.5521	-0.3042	-6.4472
16.	-3.4177	-0.3109	-6.9192
17.	-7.2138	-0.7326	-5.3922
18.	-4.2212	-0.4764	-6.0888
19.	-8.0254	-1.1422	-8.4608
20.	-4.7215	-0.5656	-13.9338
21.	-8.7637	-1.0327	-9.0165
22.	-6.0570	-0.7192	-9.2870
23.	-6.6892	-0.4504	-7.0497
24.	-4.2515	-0.3588	-6.4511
25.	-4.2863	-0.2035	-6.6732
26.	-2.9299	-0.2381	-7.5670
27.	-5.1398	-0.3110	-6.8785
28.	-3.3210	-0.3247	-6.1279
29.	-6.7039	-0.8207	-5.5484
30.	-3.9714	-0.4790	-5.1615
31.	-7.2957	-1.1154	-7.8398
32.	-4.1003	-0.4935	-10.6223
33.	-7.8413	-1.0881	-8.4037
34.	-5.0910	-0.5996	-7.1120
35.	-5.6738	-0.3681	-7.6816
36.	-4.1354	-0.3574	-5.4749

TRAFFTAX

Naast de resultaten voor FUELCOST toont Tabel 5 ook de directe elasticiteiten voor TRAFFTAX. Ondanks het feit dat de bijhorende parameter niet-significant werd bevonden in het model, kunnen we mogelijk interessante conclusies trekken uit de waardes van de berekende elasticiteiten. De waarde voor de directe elasticiteit is voor alle autoklassen <0, hetgeen impliceert

dat een hogere verkeersbelasting voor autoklasse x ook resulteert in een lagere kans dat autoklasse x gekozen wordt. In de meerderheid van de gevallen ligt de elasticiteit tussen -1 en 0. Voor slechts 4 autoklassen observeren we een relatief elastische respons: indien bv. voor autoklasse 19 een 1%-stijging van de verkeersbelasting wordt doorgevoerd, verwachten we o.b.v. het geschatte model een daling van de kans op autoklasse 19 met 1.14%. Ook hier weer merken we op dat de hoogste elasticiteiten worden waargenomen in de autoklassen met een hoge gemiddelde verkeersbelasting, hetgeen logisch is omdat een 1%-stijging van de verkeersbelasting een grotere absolute stijging met zich meebrengt dan voor de autoklassen met een lagere gemiddelde verkeersbelasting.

De kruiselasticiteiten (niet getoond) voor TRAFFTAX liggen allen in het interval [+0.0004 ; +0.0370]. Ook hier weer vonden we de laagste kruiselasticiteiten voor de autoklassen met een klein marktaandeel, bv. de kruiselasticiteit van een wijziging in TRAFFTAX voor autoklasse 20 op autoklasse 1 bedraagt slechts 0.0006. Voor autoklassen met een groter marktaandeel lag de kruiselasticiteit dan weer een stuk hoger (bv. indien de TRAFFTAX voor autoklasse 18 met 1% stijgt, zal de kans op autoklasse 22 met 0.0365% stijgen).

Andere variabelen

Elasticiteiten berekenen voor variabele LOGRC achten we weinig nuttig omdat de resultaten zeer moeilijk te interpreteren zijn: indien een bepaald model wordt toegevoegd aan de steekproef beïnvloedt dit immers zowel de teller als de noemer van het argument van de logaritme. Bovendien kan deze variabele niet door het beleid worden beïnvloed. Aangezien we elasticiteiten enkel kunnen berekenen voor continue variabelen, heeft het geen zin een gelijkaardig resultaat te tonen voor de interactievariabelen in het model. Met het oog op validatie kan het bestuderen van elasticiteiten van de (niet geïnterageerde) prijsvariabele echter zeer interessant zijn. Om een beter idee te krijgen van de grootteordes van de prijselasticiteit kunnen we daarom het model herschatten door variabele 1 en 2 te vervangen door een niet-geïnterageerde prijsvariabele (TOTPR11 = totale autoprijs[20]). De herschatte coëfficiënten worden hier niet getoond, omdat deze modelstructuur inferieur is aan de eerder getoonde resultaten. In Tabel 6 tonen we wel de directe elasticiteiten voor TOTPR11. Ook hier (net zoals bij bv. TRAFFTAX) valt op dat we hogere elasticiteiten noteren bij de autoklassen met een hoge gemiddelde waarde voor TOTPR11. Dit is niet zo vreemd omdat bij een MNL-model de directe elasticiteit direct afhangt van het niveau van het gewijzigde attribuut (Louviere et al., 2000).

Tabel 6 Directe elasticiteiten voor TOTPR11 in het gewijzigde MNL-model (variabele 1 en 2 vervangen door TOTPR11)

Autoklasse	Directe elasticiteit TOTPR11	Autoklasse	Directe elasticiteit TOTPR11
1.	-0.5247	**19.**	-14.4343
2.	-0.6135	**20.**	-9.9548
3.	-0.7469	**21.**	-7.9480
4.	-0.7681	**22.**	-7.1799
5.	-1.6177	**23.**	-4.7978
6.	-1.3362	**24.**	-4.7680
7.	-2.8176	**25.**	-6.4079
8.	-0.9745	**26.**	-7.4632
9.	-1.7467	**27.**	-8.9108

[20] Dit is exact de variabele die voorheen werd geïnterageerd met de 2 inkomensdummies om de variabelen 1 en 2 te bekomen. Zie ook Tabel 3 voor meer uitleg.

Autoklasse	Directe elasticiteit TOTPR11	Autoklasse	Directe elasticiteit TOTPR11
10.	-1.5712	**28.**	-8.8524
11.	-1.2415	**29.**	-19.7581
12.	-1.1538	**30.**	-15.2004
13.	-2.5195	**31.**	-29.3855
14.	-2.8589	**32.**	-16.0934
15.	-3.4876	**33.**	-19.1319
16.	-3.5741	**34.**	-17.0950
17.	-8.2253	**35.**	-9.3132
18.	-6.5806	**36.**	-10.0414

3.1.2. NESTED LOGIT (NL) MODEL

→ Interpretatie schattingsresultaten

Algemeen

Omdat we vermoeden dat bepaalde autoklassen betere substituten zijn voor elkaar dan andere, schatten we naast het MNL-model ook een gelijkaardige NL-structuur. Er werd een NL-model[21] geschat met 2 niveaus: op het hoogste niveau beschouwen we 3 nests volgens het bouwjaar van de autoklassen, met daaronder in elke nest 12 alternatieven (resp. autoklasse 1-12, 13-24, 25-36 voor nest 1, 2 en 3, zie Tabel 4). Bovendien werd in dit NL-model ook rekening gehouden met de variantie/covariantie van de opgenomen elementen (=schatter * variabele) in het model, gegeven de variantie-/covariantiematrix[22] (kortweg 'VARCOVAR') op de kenmerken van de onderliggende automodellen in elke autoklasse (zie McFadden, 1978; ook onder sectie "Modelleringsbenadering"). Om te achterhalen in welke mate ons geschatte NL-model de variatie in de geobserveerde keuzes verklaart, kunnen we juist zoals bij het MNL-model een pseudo-R^2 berekenen.

Voor onderstaand NL-model vonden we een pseudo-R^2 van 1-(-5431.81/-5535.61) = 0.0188: hiertoe werd het NL-model dus vergeleken met hetzelfde basismodel als hierboven (rekening houdend met geobserveerde marktaandelen in de steekproef). Volgens de LL ratio-test (5% significantie) presteert het geschatte NL-model significant beter dan het basismodel:

$$-2 \times \left(LL_{basismodel} - LL_{geschat\ model}\right) = -2 \times (-5535.614 + 5431.809) = 207.611$$
$$> \chi^2_{aantal\ nieuwe\ params\ geschat\ in\ geschatte\ model} = \chi^2_{46-35} = 19.675$$

[21] Naast het finale NL-model (hier getoond) werd geëxperimenteerd met verschillende andere NL-structuren met 1 niveau van nesting. Zo werd bv. een NL-model geschat met 2 bovenliggende nests SUV/geen SUV (d.w.z. een clustering met autoklassen 9, 10, 21, 22, 33, en 34 (SUV) versus al de rest (geen SUV); een NL-model met alle 6 verschillende carrosserietypes als nests; een NL-model met resp. carrosserietype GroteMiddenklasse+Executive, Stadswagens + KleineMiddenklasse of Cabrio+coupés versus alle overige types als bovenliggende nests; en een NL-model met het brandstoftype als bovenliggende nest. Verder werd ook een NL-model geschat met 2 niveaus van nesting: het brandstoftype als hoogste nests en daaronder een bijkomend nest-niveau met de 6 carrosserietypes. Alle geteste structuren bleken inferieur aan het finale (hier getoonde) model met de bouwjaarklasse als enige nestniveau.

[22] Dit is de matrix met de varianties van elk kenmerk voor alle automodellen opgenomen onder elke autoklasse, alsook de covarianties van elke mogelijke combinatie van 2 kenmerken van de automodellen onder elke autoklasse.

Het NL-model blijkt volgens de LL ratio-test ook significant beter te presteren dan het hiervoor geschatte MNL-model[23]:

$$-2 \times (LL_{MNL} - LL_{NL}) = -2 \times (-5445.446 + 5431.809) = 27.27466$$
$$> \chi^2_{aantal\ nieuwe\ params\ geschat\ in\ geschatte\ model} = \chi^2_{46-42} = 9.488$$

Parameterschattingen

Alle geschatte parameters zijn minstens significant op het 5%-niveau, behalve de parameter horende bij variabele 6 TRAFFTAX. We becommentariëren hierna enkel de opvallendste wijzigingen t.o.v. de parameterschattingen en interpretaties van het MNL-model. Zie Tabel 7 en Tabel 8.

Tabel 7 NL-variant van model A

Verklarende variabele	Geschatte coëfficiënt	t-statistiek	95% betrouwbaarheidsinterval	
1. PR11YLOW	-0.23201E-04***	4.14	-0.34195E-04	-0.12206E—04
2. PR11YHIG	0.24982E-04***	3.17	0.95378E-05	0.40426E-04
3. VOLMEDGZ	-0.02767**	2.21	-0.05216	-0.00318
4. VOLGRGZ	0.11653***	3.29	0.04716	0.18591
5. VOLYLOW	-0.03403**	2.36	-0.06230	-0.00577
6. TRAFFTAX	-0.00100	0.86	-0.00329	0.00129
7. FUELCOST	-0.24157**	2.22	-0.45474	-0.02840
8. LOGRC	0.98448**	2.29	0.14031	1.82865
9. OMTTXGG	-0.00044**	2.09	-0.00085	-0.00003
10. OMTTXGHY	-0.16931E-04**	2.38	-0.30885E-04	-0.29761E-05
11. OMTTXGHO	0.97515E-05**	2.20	0.10617E-05	0.18441E-04
12. IVOLD	0.52066***	2.99	0.17881	0.86252
13. IVMED	0.44059***	3.16	0.16763	0.71355
14. IVYNG	0.52955***	3.47	0.23086	0.82825

*NL-model geschat op 36 alternatieven (autoklassen). Aantal observaties gebruikt voor de schatting: 1740 van de 2342 (602 gezinnen met een onbekende waarde voor 1 van de opgenomen variabelen werden geschrapt). De bekomen log likelihood-waarde bedraagt -5432. Een significantie van een parameter op het 1%, 5% of 10%-niveau wordt aangeduid met resp. ***, ** of *.*

Tabel 8 Verklaring variabelen opgenomen in het NL-model

Verklarende variabele	Definitie
1. PR11YLOW	Interactie tussen de **totale autoprijs** (aankoopprijs, BIV en CO2-premie, allen rekening houdend met de leeftijd van de auto) en een effects-coded **dummy voor lage** (netto ≤2000 EUR/mnd) **gezinsinkomens**
2. PR11YHIG	Interactie tussen de **totale autoprijs** (aankoopprijs, BIV en CO2-premie, allen rekening houdend met de leeftijd van de auto) en een effects-coded **dummy voor hoge** (netto >4000 EUR/mnd) **gezinsinkomens**
3. VOLMEDGZ	Interactie tussen **autovolume** (lengte x breedte x hoogte) en een effects-coded **dummy voor middelgrote gezinnen** (3 of 4 gezinsleden)
4. VOLGRGZ	Interactie tussen **autovolume** (lengte x breedte x hoogte) en een effects-coded **dummy voor grote gezinnen** (>4 gezinsleden)

[23] Merk op dat het verschil in significantie niet uitsluitend te wijten is aan het verschil tussen MNL en NL, maar ook aan de opname van 3 covarianties op de kenmerken van de onderliggende automodellen. We kunnen uit deze Chi²-test dus niet noodzakelijk besluiten dat rekening houden met een neststructuur op zichzelf heeft volstaan om de performantie van het model te verbeteren.

5.	**VOLYLOW**	Interactie tussen **autovolume** en een effects-coded **dummy voor lage** (netto ≤2000 EUR/mnd) **gezinsinkomens**
6.	**TRAFFTAX**	**Jaarlijkse verkeersbelasting** te betalen in 2011
7.	**FUELCOST**	**Brandstofkost** (EUR/100km)
8.	**LOGRC**	**LOGRC = LOG(N_i/N)**, d.i. de link met het onderliggend (niet-geschatte) model voor keuze van een specifiek automodel (in tegenstelling tot de autoklasse)
9.	**OMTTXGG**	**Covariantie** van 1) **VOLGRGZ** (interactieterm) en 2) **TRAFFTAX**
10.	**OMTTXGHY**	**Covariantie** van 1) **KWGHFDYG** (interactieterm) en 2) **TRAFFTAX**
11.	**OMTTXGHO**	**Covariantie** van 1) **KWGHFDOD** (interactieterm) en 2) **TRAFFTAX**
12.	**IVOLD**	**Inclusive value** van de nest '**OLD**', d.i. auto's met bouwjaar <2001 (autoklasse 1-12)
13.	**IVMED**	**Inclusive value** van de nest '**MED**', d.i. auto's met bouwjaar 2001-2005 (autoklasse 13-24)
14.	**IVYNG**	**Inclusive value** van de nest '**YNG**', d.i. auto's met bouwjaar >2005 (autoklasse 25-36)

De variabelen die de interactie weergeven tussen prijs en inkomen behouden hetzelfde teken en blijven van dezelfde grootteorde.

De schattingen van de interactie autovolume-gezinsgrootte behouden ook het teken van de MNL-variant, maar de absolute waarde van de schatters daalt.

Voor variabele 5 (interactie volume–laag inkomen) vinden we een kleinere absolute waarde voor de geschatte coëfficiënt, alsook een gedaald significantieniveau (van 1% naar 5%).

De verkeersbelasting behoudt zijn weinig significante schatter met negatief teken.

De schatter voor variabele 7 (FUELCOST) moet t.o.v. het MNL-model wat aan significantie inboeten en verkleint in absolute waarde, maar behoudt wel het belangrijke negatieve teken.

De schatter voor LOGRC wordt qua magnitude ongeveer gehalveerd t.o.v. het MNL-model, waardoor de schatter nu tussen 0 en 1 ligt, consistent met de theorie van nutsmaximalisatie. Bovendien bevat het 95%-betrouwbaarheidsinterval het grootste deel van het interval [0 ; 1].

Twee variabelen uit het MNL-model zijn weggevallen wegens niet langer significant (interacties tussen autovermogen en leeftijd van het gezinshoofd).

Anderzijds bevat het NL-model ook 3 nieuwe variabelen in de nutsfuncties, en 3 geschatte parameters horende bij de zgn. inclusive values (IV) (zie Train, 1986 voor de definitie), die de link vormen tussen de 3 nests en de onderliggende alternatieven.

Startend van het MNL-model (zie hoger) en veralgemeend naar een NL-model op 2 niveaus, werden stapsgewijs de verschillende mogelijke varianties/covarianties van de aanwezige elementen in het MNL-model, toegevoegd aan de nutsfuncties van het NL-model, gebaseerd op de VARCOVAR-matrix van de kenmerken van de onderliggende automodellen in elke autoklasse Omdat we met interactievariabelen werken is het aantal mogelijke combinaties zeer hoog. Uiteindelijk bleken slechts 3 covarianties van de opgenomen elementen als significant in het model te worden weerhouden.

De geschatte parameter horend bij variabele 9 (OMTTXGG) heeft een negatief teken. Dit impliceert dat een hogere covariantie tussen VOLGRGZ en TRAFFTAX binnen een gegeven klasse gepaard gaat met een lager nut. Of met een voorbeeld: voor grote gezinnen geldt dat auto's met een groot volume en een hoge verkeersbelasting minder aantrekkelijk zijn dan auto's met een groot volume en een lagere verkeersbelasting. Dus, indien binnen een gegeven klasse, een hoge correlatie

bestaat tussen volume en verkeersbelasting, dan is deze klasse minder aantrekkelijk dan een klasse met een lagere correlatie tussen beide variabelen.

De geschatte parameter horende bij variabele 10 heeft ook een negatief teken. Merk vooreerst op dat deze variabele significant bevonden wordt, ook al werd één van de twee 'samenstellende' variabelen (nl. KWGHFDYG) niet in het NL-model weerhouden. Ook voor variabele 10. volgt uit een hogere covariantie tussen KWGHFDYG en TRAFFTAX, een lager nut. Jonge gezinshoofden (<40 jaar) ondervinden blijkbaar een significant lager nut indien de covariantie tussen autovermogen en verkeersbelasting hoog is.

De derde en laatste covariantie (OMTTXGHO) heeft een positieve parameter. Ook hier geldt weer dat één van de 2 samenstellende variabelen (KWGHFDOD) niet meer in het NL-model is opgenomen. De positieve waarde voor de schatter wijst op een stijgend nut bij hogere covarianties tussen autovermogen en verkeersbelasting voor oudere gezinshoofden (65+ jaar). Dus een autoklasse blijkt meer kans te hebben gekozen te worden door oude gezinshoofden indien het autovermogen én de verkeersbelasting beide eerder hoog liggen of beide eerder laag liggen. Dit kan verklaard worden door het feit dat het autovermogen sterk gecorreleerd[24] is met de cilinderinhoud (en die laatste is de beslissende parameter in de bepaling van de verkeersbelasting), dus de verkeersbelasting is sowieso hoog indien een auto een grote (en dus vaak krachtige) motor heeft. Een krachtige motor maakt een auto over het algemeen aantrekkelijk, ceteris paribus.
Het valt op dat alle 3 de covarianties gerelateerd zijn aan de verkeersbelasting. Al is het hoofdeffect van TRAFFTAX hiervoor insignificant bevonden, toch bleek het dus waardevol enkele covarianties met TRAFFTAX in de nutsfuncties mee te nemen.

Variabelen 12-14 vormen de ruggengraat van de NL-modelstructuur: deze inclusive values vormen de link tussen de nests en de onderliggende alternatieven behorend tot die nest. Het nut van een nest kan immers berekend worden als het nut van de variabelen die dezelfde waarde aannemen voor alle elementen van de nest, aangevuld met de inclusive value[25]. We zien dat elk van de 3 parameters een geschatte waarde tussen 0 en 1 heeft (incl. de 95%-betrouwbaarheidsintervallen), hetgeen consistent is met de theorie van nutsmaximalisatie. Immers, hoe hoger de IV-variabele (d.i. de verwachte waarde van het maximaal nut, onderliggend), hoe groter het nut voor de beslisser. Verder geldt: hoe hoger de geschatte IV-parameter, hoe lager de variatie (standaarddeviatie) tussen de niet-geobserveerde componenten van het nut van de onderliggende alternatieven in die nest.

Elasticiteiten

We berekenen zowel directe als kruiselingse puntelasticiteiten voor dezelfde 3 variabelen als voorheen (variabelen 6, 7 en 8).

FUELCOST

De directe elasticiteiten liggen ook in dit geval meestal tussen -3 en -8, wat terug wijst op een zeer elastische vraag. Voor de uitschieters gelden dezelfde conclusies als bij het MNL-model. Verder valt duidelijk op dat voor nest OLD en YNG de absolute waarde van de elasticiteiten kleiner is dan bij het MNL-model, terwijl de elasticiteit voor de MED-nest in absolute waarde steeds groter is vergeleken met het MNL-model.

[24] Correlatie tussen autovermogen en cilinderinhoud van 0.86 voor de geldige sample van 1740 gezinnen.
[25] Deze inclusive value kan best beschouwd worden als de verwachte waarde van het maximaal nut te verkrijgen door een keuze te maken uit alle alternatieven uit die nest.

Voor de kruiselings elasticiteiten is het grootste verschil met het MNL-model dat in het NL-model de kruiselasticiteiten van autoklassen binnen een bepaalde nest sterker op elkaar lijken dan kruiselasticiteiten van autoklassen uit 2 verschillende nests. De kruiselingse elasticiteiten (niet getoond[26]) zijn juist zoals bij het MNL-model allemaal positief. Dit is het verwachte resultaat: wanneer de brandstofkost van autoklasse x stijgt, zullen alle andere autoklassen interessanter worden en een hogere kans hebben om gekozen te worden.

De kruiselingse elasticiteiten[27] voor FUELCOST liggen allen in het interval [+0.0019 ; +0.9340], dus breder dan bij het MNL-model. We vonden de grootste kruiselasticiteiten voor de autoklassen die sterk op elkaar lijken (dit geldt algemeen). Bijvoorbeeld: wanneer de FUELCOST voor autoklasse 24 stijgt met 1%, zal de kans op autoklasse 22 (d.i. een autoklasse uit dezelfde nest) met 0.9340% stijgen. Zie Tabel 4 voor de definitie van de klassen. De laagste kruiselingse elasticiteiten werden waargenomen bij de autoklassen die niet sterk op elkaar lijken. Bijvoorbeeld, wanneer de FUELCOST van autoklasse 20 met 1% stijgt zal de kans op autoklasse 12 (d.i. een autoklasse uit een andere nest) slechts met 0.0026% stijgen, juist omdat autoklasse 20 en 12 zo'n slechte substituten zijn van elkaar.

Tabel 9 Directe elasticiteiten voor FUELCOST, TRAFFTAX en LOGRC in het NL-model

Autoklasse	Directe elasticiteit FUELCOST	Directe elasticiteit TRAFFTAX	Directe elasticiteit LOGRC
1.	-4.7270	-0.2941	-5.4996
2.	-3.8276	-0.4708	-8.8330
3.	-5.2067	-0.4174	-4.2779
4.	-3.6824	-0.5203	-6.3188
5.	-6.7196	-0.9209	-3.7410
6.	-4.1144	-0.7122	-5.4807
7.	-7.7964	-1.6352	-6.8388
8.	-4.6822	-0.6555	-13.1716
9.	-8.4950	-1.5977	-8.0288
10.	-6.0350	-1.2737	-8.4923
11.	-7.0519	-0.8027	-6.9967
12.	-4.4585	-0.5993	-6.9218
13.	-5.0107	-0.3788	-7.7034
14.	-3.5472	-0.4900	-10.1928
15.	-6.2614	-0.6051	-7.1512
16.	-3.6521	-0.5859	-7.2723
17.	-8.2698	-1.4812	-6.0798
18.	-4.5452	-0.9046	-6.4483
19.	-9.3112	-2.3371	-9.6550
20.	-5.4970	-1.1614	-15.9559
21.	-10.1401	-2.1072	-10.2610
22.	-6.9414	-1.4535	-10.4681

[26] De volledige tabellen met kruiselings elasticiteiten kunnen desgewenst op eenvoudig verzoek bekomen worden.

[27] De kruiselasticiteit van de kans op alternatief i voor gezin q m.b.t. een marginale wijziging van het k-de attribuut van alternatief j wordt berekend als: $E_{X_{jkq}}^{P_{iq}} = \frac{\partial P_{iq}}{\partial X_{jkq}} \times \frac{X_{jkq}}{P_{iq}}$, met E de elasticiteiten, P de kans en X het attribuut. Louviere et al. (2000) hebben aangetoond dat binnen eenzelfde nest (d.w.z. enkel geldig voor MNL of binnen een bepaalde nest van een NL) deze formule vereenvoudigt tot $E_{X_{jkq}}^{P_{iq}} = -\beta_{jk} \times X_{jkq} \times P_{jq}$. Ter volledigheid: de formule voor de directe prijselasticiteit luidt: $E_{X_{ikq}}^{P_{iq}} = \frac{\partial P_{iq}}{\partial X_{ikq}} \times \frac{X_{ikq}}{P_{iq}}$, en dit vereenvoudigt tot $E_{X_{ikq}}^{P_{iq}} = -\beta_{ik} \times X_{ikq} \times (1 - P_{iq})$ bij een MNL-structuur of binnen eenzelfde nest bij een NL-model.

23.	-7.5873	-0.9009	-7.8648
24.	-4.4636	-0.6643	-6.6618
25.	-3.9115	-0.3275	-5.9896
26.	-2.7504	-0.3942	-6.9868
27.	-4.8768	-0.5204	-6.4193
28.	-3.0352	-0.5233	-5.5087
29.	-6.4576	-1.3942	-5.2568
30.	-3.6469	-0.7758	-4.6619
31.	-7.0637	-1.9046	-7.4658
32.	-3.9696	-0.8425	-10.1146
33.	-7.5673	-1.8520	-7.9768
34.	-4.7961	-0.9961	-6.5899
35.	-5.3725	-0.6147	-7.1540
36.	-3.6600	-0.5579	-4.7660

TRAFFTAX

De waarde voor de directe elasticiteit is voor alle autoklassen <0, hetgeen impliceert dat een hogere verkeersbelasting voor autoklasse x ook resulteert in een lagere kans dat autoklasse x gekozen wordt. In de meerderheid van de gevallen ligt de elasticiteit tussen -1 en 0, maar toch zijn er nu meer autoklassen waarvoor we een relatief elastische respons observeren: voor autoklasse 19 en 21 ligt de elasticiteit zelfs lager dan -2. Indien bv. voor autoklasse 19 een 1%-stijging van de verkeersbelasting wordt doorgevoerd, verwachten we o.b.v. het geschatte model een daling van de kans op autoklasse 19 met 2.34%. Ook hier weer merken we op dat de hoogste elasticiteiten worden waargenomen in de autoklassen met een hoge gemiddelde verkeersbelasting, hetgeen logisch is omdat een 1%-stijging van de verkeersbelasting een grotere absolute stijging met zich meebrengt dan voor de autoklassen met een lagere gemiddelde verkeersbelasting. Voor alle autoklassen observeren we in absolute waarde een hogere elasticiteit dan voor het MNL-model.

De kruiselasticiteiten (niet getoond) voor TRAFFTAX liggen allen in het interval [+0.0004 ; +0.1390]. De bovengrens is nu een stuk hoger dan bij het MNL-model, vanwege de clustering in 3 nests waardoor autoklassen in één bepaalde nest sterker op elkaar lijken dan autoklassen uit verschillende nests, en dus bepaalde autoklassen sterker op elkaar lijken (betere substituten) dan in het MNL-model. Ook hier weer vonden we de laagste kruiselasticiteiten voor de autoklassen die slechte substituten van elkaar zijn, bv. de kruiselasticiteit van een wijziging in TRAFFTAX voor autoklasse 31 op autoklasse 24 bedraagt slechts 0.0006. Voor sterker op elkaar gelijkende autoklassen lag de kruiselasticiteit dan weer een stuk hoger, bv. indien de TRAFFTAX voor autoklasse 6 met 1% stijgt, zal de kans op autoklasse 12 met 0.1149% stijgen. Deze conclusies zijn een logisch gevolg van de toepassing van een neststructuur.

Andere variabelen

Variabelen 1-5 zijn interactietermen met 1 dummy, waarvoor elasticiteiten weinig betekenis hebben. Net zoals bij het MNL-model tonen we hier ook de directe elasticiteiten voor de TOTPR11-variabele, na een gelijkaardige herschatting van het model, nl. het vervangen van variabelen 1 en 2 door de TOTPR11-variabele. Consistent met eerdere bevindingen, zien we dat er hogere elasticiteiten gelden voor de autoklassen die sowieso al een hoge gemiddelde prijs hebben. Dit is te wijten aan het feit dat een 1%-stijging van deze variabele een grotere impact heeft indien het gemiddelde startbedrag al hoger is.

De waarde van de geschatte elasticiteiten ligt in het algemeen dicht bij de waarde die werd geschat voor het MNL model.

Tabel 10 Directe elasticiteiten voor TOTPR11 in het gewijzigde NL-model (variabele 1 en 2 vervangen door TOTPR11)

Autoklasse	Directe elasticiteit TOTPR11	Autoklasse	Directe elasticiteit TOTPR11
1.	-0.5355	19.	-17.2391
2.	-0.6257	20.	-11.9056
3.	-0.7621	21.	-9.4863
4.	-0.7837	22.	-8.5539
5.	-1.6472	23.	-5.7022
6.	-1.3623	24.	-5.5288
7.	-2.8673	25.	-6.8540
8.	-0.9935	26.	-8.0057
9.	-1.7805	27.	-9.5692
10.	-1.6020	28.	-9.4688
11.	-1.2661	29.	-21.2446
12.	-1.1745	30.	-16.2679
13.	-2.9684	31.	-31.6221
14.	-3.4033	32.	-17.3186
15.	-4.1376	33.	-20.5769
16.	-4.1840	34.	-18.3797
17.	-9.7914	35.	-9.9989
18.	-7.7160	36.	-10.6601

Ook de 3 covariantie-variabelen (variabelen 9-11) zijn berekend door in elk van de 3 gevallen telkens 1 dummy mee te nemen. Voor variabele 8 worden ook voor het NL-model geen elasticiteiten berekend omdat de interpretatie ervan moeilijk is en die bovendien niet door het beleid kan worden beïnvloed. Variabelen 12-14 tot slot zijn intern berekend door het model te optimaliseren waardoor de bijhorende elasticiteiten geen interpretatieve waarde kunnen gegeven worden.

3.2. MODEL B: KEUZE AANTAL AUTO'S VOOR GEZINNEN MET 0 OF 1 AUTO

Met dit model B trachten we de keuze tussen het bezitten van geen (0) of 1 auto te verklaren. In model A hebben we de kans geschat dat een bepaald gezin kiest voor een bepaalde autoklasse. Hiertoe werd het model geschat op alle gezinnen in de steekproef met 1 auto in bezit (2342 gezinnen). Aan deze set van 2342 gezinnen werden nu de gezinnen zonder auto toegevoegd, zodat beide mogelijke keuzes van model B (0 of 1 auto) effectief door bepaalde gezinnen gekozen worden in het model.

3.2.1. INTERPRETATIE SCHATTINGSRESULTATEN

→ Algemeen

Omdat we hier een model schatten met slechts 2 alternatieven (nl. het bezitten van één dan wel geen auto[28]), heeft de toepassing van een geneste structuur weinig zin. Daarom schatten we het model volgens de eenvoudiger MNL-structuur[29]. Hierbij bevat de nutsfunctie voor het 1 auto-alternatief alle variabelen[30] (zoals in Tabel 11 weergegeven), terwijl de nutsfunctie voor het 0 auto-alternatief enkel een alternatief-specifieke constante (ASC) bevat.

Het finale model werd samengesteld door stapsgewijs variabelen toe te voegen aan één van de 2 nutsfuncties. Omdat de keuze nu draait rond het aantal gekozen wagens, hoeven de kenmerken van de wagens nu niet meer opgenomen worden. Dit wordt immers reeds vervat in de inclusive value[31] die model A met model B linkt. We verwachten a priori dat het nut dat een gezin haalt uit het bezit van een bepaald aantal auto's (0 of 1) dus niet enkel wordt bepaald door een aantal socio-demografische kenmerken (bv. diploma of gezinsgrootte) van het gezin, maar ook door de inclusive value, of m.a.w. de verwachte waarde van het maximale nut dat kan bekomen worden door 1 auto te kiezen (gegeven de onderliggende autoklassen).
Merk op dat veel van de socio-demografische factoren die we ter beschikking hadden, sterk met elkaar gecorreleerd zijn. Denken we bijvoorbeeld maar aan het verband tussen het gezinsinkomen en het hoogst behaalde diploma. In zulke gevallen werd er telkens voor gekozen om een zo hoog mogelijke modelfit te bekomen, en coëfficiënten met een zo hoog mogelijke significantie.

Om te achterhalen in welke mate ons geschatte model de variatie in de geobserveerde keuzes verklaart, kunnen we weer een pseudo-R^2 berekenen. Als we opnieuw vergelijken met een basismodel dat enkel rekening houdt met waargenomen marktaandelen, bekomen we een pseudo-R^2 van 0.2164. De LL ratio-test bevestigt dat dit MNL-model beter presteert dan het basismodel:

$$-2 \times \left(LL_{basismodel} - LL_{geschat\ model}\right) = -2 \times (-871.669 + 683.056) = 377.225$$
$$> \chi^2_{aantal\ nieuwe\ params\ geschat\ in\ geschatte\ model} = \chi^2_{8-1} = 14.067$$

[28] Merk op dat de steekproef voor deze schattingen beperkt werd tot gezinnen met 1 auto of geen auto. Alle schattingsresultaten moeten dan ook in die zin geïnterpreteerd worden. Een schatter geeft hier dus weer hoe de kansen op 0 dan wel 1 auto wijzigt, gegeven dat er uitsluitend gezinnen met 0 of 1 auto bestaan. Die laatste assumptie wordt in een later stadium verder versoepeld.
[29] Strikt genomen passen we wel een NL-structuur toe, maar dan in 2 stadia: full information maximum likelihood (FIML) voor de keuze van de autoklasse (model A), en op basis van de resultaten daarvan de keuze van het aantal voertuigen (model B).
[30] Voor het berekenen van de inclusive value tussen model A en model B (hier de LOGSUM genoemd), gebruikten we het NL-model dat hiervoor werd besproken.
[31] Verder zullen we zien dat die inclusive value door de variabele LOGSUM wordt vertegenwoordigd.

→ **Parameterschattingen**

Tabel 11 Schattingen voor model B

Verklarende variabele		Geschatte coëfficiënt	t-statistiek	95% betrouwbaarheidsinterval	
1.	**GHFD_VR**	-0.31186***	4.02	-0.46386	-0.15987
2.	**SINGLE**	-0.66583***	8.24	-0.82418	-0.50748
3.	**BEDIENDE**	0.34034***	3.09	0.12464	0.55604
4.	**Y_LOW**	-0.49158***	5.14	-0.67889	-0.30428
5.	**DIPL_L**	-0.54464***	5.38	-0.74313	-0.34615
6.	**DIPL_H**	0.38033***	2.80	0.11431	0.64634
7.	**GEMTH_GR**	-0.34818***	4.44	-0.50172	-0.19465

*MNL-model geschat op 2 alternatieven (0 auto / 1 auto). Aantal observaties gebruikt voor de schatting: 1757 van de 2899 gezinnen (1142 gezinnen met een onbekende waarde voor 1 van de opgenomen variabelen werden geschrapt). De bekomen log likelihood-waarde bedraagt -683. Een significantie van een parameter op het 1%, 5% of 10%-niveau wordt aangeduid met resp. ***, ** of *.*

Tabel 12 Verklaring variabelen opgenomen in model B

Verklarende variabele		Definitie
0.	**LOGSUM**	De logsom of **inclusive value** van het **1-auto alternatief**: dit komt overeen met de verwachte waarde van het maximale nut dat kan verkregen worden door een specifieke onderliggende autoklasse te kiezen. Deze variabele vormt dus de link met het onderliggende model A, in ons geval in de NL-variant.
1.	**GHFD_VR**	Een effects-coded **dummy** voor een **vrouwelijk gezinshoofd**
2.	**SINGLE**	Een effects-coded **dummy** voor gezinnen met slechts **1 gezinslid**
3.	**BEDIENDE**	Een effects-coded **dummy** voor **respondenten** die als **hoofdberoep** onder de categorie '**bediende**' vallen (i.t.t. arbeider of zelfstandige)
4.	**Y_LOW**	Een effects-coded dummy voor **lage** (netto ≤2000 EUR/mnd) **gezinsinkomens**
5.	**DIPL_L**	Een effects-coded **dummy** voor **gezinshoofd** met als hoogst behaalde **diploma 'lager onderwijs' of minder**
6.	**DIPL_H**	Een effects-coded **dummy** voor **gezinshoofd** met als hoogst behaalde **diploma 'hoger onderwijs'**
7.	**GEMTH_GR**	Een effects-coded **dummy** voor gezinnen met een **grootstad of Vlaamse centrumstad als thuisadres.** Zou beschouwd kunnen worden als een proxy voor beschikbaarheid van goed openbaar vervoer.

Alle geschatte parameters zijn significant op het 1%-niveau. De specifieke schattingen en de verklaringen van de opgenomen variabelen kan men terugvinden in Tabel 11 en Tabel 12. Merk op dat alle variabelen werden opgenomen in de nutsfunctie van het 1-auto alternatief (zie hoger); dus positieve coëfficiënten wijzen op variabelen met een positieve invloed op het 1-auto-alternatief en negatieve coëfficiënten wijzen op variabelen met een positieve invloed op het 0-auto alternatief.

We vinden een niet-significante[32] parameter voor variabele 0 "LOGSUM" (om die reden ook niet getoond in Tabel 11). Deze variabele is een weergave van de verwachte waarde van het maximaal nut van het kiezen van 1 auto, en vormt als dusdanig de link met het onderliggende model A voor de keuze van de autoklasse. Uit deze niet-significante coëfficiënt blijkt dus dat de verwachte waarde van het maximaal nut van het bezitten van 1 auto geen invloed heeft op de keuze van het aantal auto's.

[32] Zoals men verderop kan zien, worden in model D (keuze tussen 0, 1 of 2 auto's) de coëfficiënten van de LOGSUM-variabelen wel significant verschillend van 0 bevonden.

Bij de interpretatie van dit resultaat moeten we zeer voorzichtig zijn. De LOGSUM geeft immers weer wat de verwachte waarde is van het maximaal nut van het kiezen van 1 auto, gegeven de kenmerken van het gezin, én gegeven de technische kenmerken van alle beschikbare autoklassen. Vermits alle gezinnen uit dezelfde autoklassen kunnen kiezen, brengt dit met zich mee dat:

(a) Gezinnen met andere socio-economische kenmerken de voorkeur zullen geven aan andere autoklassen. Dit verklaart waarom het verwacht maximaal nut varieert over de gezinnen. Het gebrek aan significantie van deze coëfficiënt betekent dus dat verschillen in kenmerken tussen de gezinnen alleen *rechtsreeks* een impact hebben op de keuze van het aantal auto's, en niet via hun impact op de keuze van een specifieke klasse.

(b) Sommige technische kenmerken leiden tot de toename van het verwacht nut van een bepaalde autoklasse voor alle gezinnen. Dit heeft als gevolg dat de waarschijnlijkheid stijgt dat deze klasse wordt gekozen (gegeven dat een gezin 1 wagen bezit) voor alle gezinnen, maar ook dat het maximaal verwachte nut van het bezitten van een auto toeneemt. Dit soort effecten blijkt hier dus geen significante invloed uit te oefenen op het aantal auto's.

Uit de schatting horend bij variabele 1. leren we dat een vrouwelijk gezinshoofd (t.o.v. een mannelijk gezinshoofd) de kans verhoogt dat een gezin 0 auto's bezit. Merk op dat de dummy voor een vrouwelijk gezinshoofd positief gecorreleerd is met de dummy voor lage inkomens (+0.274): het is dus mogelijk dat de causaliteit vooral via het inkomensniveau werkt.

Kijken we naar de schatter voor variabele 2, dan zien we dat gezinnen bestaande uit 1 persoon een hogere kans hebben op het bezitten van 0 auto's dan de gezinnen met meer dan 1 gezinslid. Er werd beslist om deze variabele SINGLE op te nemen in het model, eerder dan de dummies voor groot gezin of middelgroot gezin, omdat de modelfit beter bleek te zijn voor een model met SINGLE.

Voor variabele 3 (BEDIENDE) vonden we dan weer een positieve coëfficiënt. Voor de dummies voor beroepsklassen ('zelfstandige' en 'arbeider'[33]) werd in beide gevallen een niet-significante parameter geschat, en deze dummies werden daarom uit de geschatte nutsfunctie verwijderd. De positieve parameter voor BEDIENDE duidt erop dat bediendes een hogere kans hebben op het bezit van 1 auto dan niet-beroepsactieve respondenten.. Daarnaast werd ook een alternatieve specificatie overwogen met een dummy die aangeeft of de respondent al dan niet beroepsactief was, maar die leverde een (iets) slechtere modelfit op dan degene die hier wordt gepresenteerd.

De schatter voor variabele 4. (Y_LOW) duidt op een lagere kans op het bezit van 1 voertuig voor gezinnen met een laag inkomen, vergeleken met gezinnen met een gemiddeld inkomen. Initieel werd ook een dummy voor hoge inkomens (>4000 EUR/mnd) meegenomen, maar de geschatte coëfficiënt daarvoor werd niet-significant bevonden. M.a.w., voor gezinnen met hoge inkomens kunnen we niet stellen dat ze een hogere kans hebben op het bezitten van een bepaald aantal auto's in vergelijking met gezinnen met gemiddelde inkomens.

Voor variabele 5 en 6 (diploma) werd ietwat verrassend[34] in beide gevallen een significante coëfficiënt geschat. Gezinshoofden met een lage scholingsgraad hebben een grotere kans dan gezinshoofden met een gemiddelde scholingsgraad (hoogste diploma 'middelbaar onderwijs') om 0

[33] De 3 dummies voor de beroepscategorie (BEDIENDE, ARBEIDER, ZELFSTND) werd effects-gecodeerd met als referentieniveau NTACTIEF, d.w.z. de niet-beroepsactieve respondenten.

[34] "Verrassend" omdat we a priori zouden verwachten dat de scholingsgraad sterk gecorreleerd is met andere variabelen die al in het model waren opgenomen (bv. LOGSUM, Y_LOW, BEDIENDE, etc.). De coëfficiënten van variabele 5 en 6 moeten geïnterpreteerd worden als het effect van de scholingsgraad, bovenop de effecten die al door de andere opgenomen variabelen worden geïncorporeerd.

dan 1 auto te kiezen. Gezinshoofden met een universitair of hogeschool-diploma hebben dan weer een grotere kans om 1 auto te kiezen dan gezinnen met een gemiddelde scholingsgraad.

De 7de variabele is een dummy die weergeeft of de woonplaats van het betreffende gezin in een grootstad of Vlaamse centrumstad gelegen is. Deze dummy kan dus eventueel beschouwd worden als een proxy voor goede voorzieningen van openbaar vervoer, of van de nabijheid van mogelijke bestemmingen (zoals winkels en vrijetijdactiviteiten)[35]. Het negatief teken voor deze variabele is hetgeen we hadden verwacht: gezinnen die in een grotere stad wonen hebben een hogere kans dan mensen uit rurale gemeenten (het referentieniveau) om 0 auto's te bezitten. Vermits de coëfficiënt voor de dummy voor middelgrote steden en gemeenten niet significant verschilt van 0, wordt hij niet getoond in Tabel 11.

→ **Elasticiteiten**

Omdat alle opgenomen variabelen in de nutsfunctie dummyvariabelen zijn, heeft het weinig betekenis om voor deze een puntelasticiteit te berekenen. Het berekenen van boogelasticiteiten zou een oplossing kunnen bieden bij traditionele dummies, maar doordat wij alle dummies hebben gecodeerd via effects-coding (zie hoger) is het interpreteren van boogelasticiteiten daarvoor zeer moeilijk en zou het weinig bijdragen aan de resultaten van het model.

[35] We moeten er echter mee rekening houden dat deze variabele mogelijk ook een proxy is voor het gezinsinkomen.

HOOFDSTUK 4. KEUZEMODEL VOOR GEZINNEN MET MAXIMUM 2 AUTO'S

In deze sectie schatten we een aantal modellen voor gezinnen met maximaal 2 auto's in bezit. Om een consistent verhaal op te bouwen, herschatten we eerst het vroegere model A met een gewijzigd aantal autoklassen (model A bis). Hierna schatten we het model dat de autoklassekeuze voor gezinnen met juist 2 auto's tracht te verklaren (model C). Verder wordt, naar analogie met model B, een model voorgesteld dat de keuze tussen het bezitten van 0, 1 of 2 auto's weergeeft (model D). Finaal schatten we een model dat de afgelegde afstand met elk van de auto's in bezit tracht te verklaren (model E).

4.1. MODEL A BIS: KEUZE AUTOKLASSE VOOR GEZINNEN MET 1 AUTO

Tot hier toe hebben we ons beperkt tot de analyse van gezinnen met maximaal 1 auto. We breiden hier het model nu uit met de gezinnen die 2 auto's bezitten (vanaf model C). Om hier toe te kunnen overgaan moeten we echter ook een aantal aanpassingen doorvoeren aan het model voor gezinnen met 1 auto (model A wordt herschat tot model A bis).

Het verschil met model A is dat we hier nu nog maar rekening houden met 30 autoklassen i.p.v. 36 (de stadswagens en kleine middenklassers uit Tabel 4 werden samengevoegd). De dataset voor de 1 auto-gezinnen is nu dus een pak kleiner geworden. De reden voor deze wijziging van 36 naar 30 autoklassen ligt in de grenzen die ons door de NLOGIT-software worden opgelegd. Voor het schatten van de keuze van autoklasse voor gezinnen met 2 auto's is het immers belangrijk dat we de grens van maximaal 500 alternatieven niet overschrijden, en dit is enkel haalbaar indien we het aantal mogelijke autoklassen beperken.

In dit uitgebreid model kunnen gezinnen dus kiezen tussen:

- 30 autoklassen als ze 1 auto kiezen
- 30*31/2 = 465 combinaties van autoklassen als ze 2 auto's kiezen

Dit resulteert in 495 mogelijke keuzen.

4.1.1. MNL-VARIANT

Het loont de moeite om te testen of het bekomen MNL-resultaat (en evt. ook NL-resultaat) van model A bis niet te fel afwijkt van hetgeen we hoger vonden onder "Model A".

De schattingsresultaten voor het MNL-model worden in Tabel 13 weergegeven en vergeleken met de resultaten van model A. De pseudo-R^2 voor dit model A bis bedraagt 0.0165, quasi identiek aan de waarde die we eerder vonden voor model A (0.0163).

De geschatte coëfficiënten voor model A bis wijken over het algemeen niet sterk af die van Model A. Wat opvalt is de (zij het licht) significante schatter voor variabele 8 (TRAFFTAX). Anderzijds bevat

het 95% betrouwbaarheidsinterval voor variabele 10 (LOGRC) nu niet langer een deel van het interval [0;1].

Tabel 13 MNL-variant van model A versus model A bis

Verklarende variabele	Geschatte coëfficiënt Model A	Geschatte coëfficiënt Model A bis	95% betrouwbaar-heidsinterval Model A		95% betrouwbaar-heidsinterval Model A bis	
1. PR11YLOW	-0.22655E-04***	-0.22061E-04***	-0.35739E-04	-0.95705E-05	-0.35149E-04	-0.89718E-05
2. PR11YHIG	0.26534E-04***	0.26326E-04***	0.77305E-05	0.45337E-04	0.74624E-05	0.45190E-04
3. VOLMEDGZ	-0.04105**	-0.03688*	-0.07862	-0.00349	-0.07517	0.00142
4. VOLGRGZ	0.17806***	0.17916***	0.12821	0.22791	0.12927	0.22904
5. VOLYLOW	-0.08156***	-0.07488***	-0.11144	-0.05168	-0.10611	-0.04365
6. KWGHFDYG	-0.00556***	-0.00649***	-0.00937	-0.00174	-0.01062	-0.00236
7. KWGHFDOD	0.00297*	0.00356**	-0.00028	0.00623	0.00008	0.00703
8. TRAFFTAX	-0.00111	-0.00170*	-0.00536	0.00314	-0.00369	0.00029
9. FUELCOST	-0.47033***	-0.39258***	-0.73207	-0.20860	-0.55728	-0.22787
10. LOGRC	1.94880***	1.61810***	0.95786	2.93973	1.02633	2.20986

4.1.2. NL-VARIANT

Ook hier vergelijken we model A met een nieuw model A bis, met dit verschil dat we nu telkens de NL-variant van de 2 modellen met elkaar vergelijken i.p.v. de MNL-variant.

Merk op dat de pseudo-R^2 van het model A bis ook hier weer fractie hoger ligt dan voorheen: 0.0192 t.o.v. 0.0188 voor de initiële NL-variant van model A.

De parameterschattingen zelf verschillen niet erg sterk tussen model A en model A bis, behalve dan voor variabele 8 (LOGRC). Hiervoor werd nu niet alleen een parameter geschat in het interval [0;1]; bovendien is het betrouwbaarheidsinterval kleiner dan in model A waardoor het voor een groter deel in het [0;1]-interval ligt.

Om qua aantal autoklassen consistent te zijn doorheen de ganse modelstructuur, verkiezen we om met de resultaten van model A bis (variant NL) te blijven werken voor het berekenen van de inclusive values die nodig zijn voor het schatten van model D.

Tabel 14 NL-variant van model A versus model A bis

Verklarende variabele		Geschatte coëfficiënt Model A	Geschatte coëfficiënt Model A bis	95% betrouwbaarheidsinterval Model A		95% betrouwbaarheidsinterval Model A bis	
1.	**PR11YLOW**	-0.23201E-04***	-0.23423E-04***	-0.34195E-04	-0.12206E-04	-0.34493E-04	-0.12353E-04
2.	**PR11YHIG**	0.24982E-04***	0.25876E-04***	0.95378E-05	0.40426E-04	0.10062E-04	0.41690E-04
3.	**VOLMEDGZ**	-0.02767**	-0.02404*	-0.05216	-0.00318	-0.04824	0.00015
4.	**VOLGRGZ**	0.11653***	0.11426***	0.04716	0.18591	0.04421	0.18431
5.	**VOLYLOW**	-0.03403**	-0.03086**	-0.06230	-0.00577	-0.05845	-0.00327
6.	**TRAFFTAX**	-0.00100	-0.00118*	-0.00329	0.00129	-0.00239	0.00003
7.	**FUELCOST**	-0.24157**	-0.17428**	-0.45474	-0.02840	-0.31230	-0.03625
8.	**LOGRC**	0.98448**	0.69615**	0.14031	1.82865	0.15656	1.23575
9.	**OMTTXGG**	-0.00044**	-0.00045**	-0.00085	-0.00003	-0.00089	-0.00002
10.	**OMTTXGHY**	-0.16931E-04**	-0.17303E-04**	-0.30885E-04	-0.29761E-05	-0.31733E-04	-0.28721E-05
11.	**OMTTXGHO**	0.97515E-05**	0.99640E-05**	0.10617E-05	0.18441E-04	0.10889E-05	0.18839E-04
12.	**IVOLD**	0.52066***	0.45917***	0.17881	0.86252	0.15296	0.76538
13.	**IVMED**	0.44059***	0.46530***	0.16763	0.71355	0.16832	0.76227
14.	**IVYNG**	0.52955***	0.55531***	0.23086	0.82825	0.23055	0.88006

4.2. MODEL C: KEUZE AUTOKLASSE VOOR GEZINNEN MET 2 AUTO'S

We testen hier enkel een MNL-variant omdat het niet duidelijk is hoe we de nests zouden moeten definiëren in het geval van een NL-model: het gaat immers altijd om de keuze voor een PAAR van auto's en die paren vallen niet eenvoudig in een bepaald vakje te duwen. Merk op dat we net zoals bij het ontwerp van model A bis uitgaan van 30 (en niet langer 36) autoklassen.

Bij de opbouw van de nutsfuncties[36] werd gestart van de variabelen die we eerder al gebruikten in het model voor de schatting van de autoklasse-keuze voor gezinnen met 1 auto, met dit verschil echter dat we hier werken met de som van de technische kenmerken van beide wagens in plaats van het technisch kenmerk van 1 auto[37]. Merk op dat we m.b.t. het autovolume naast de som ook een variabele opnemen die de verwachte waarde weergeeft van de absolute waarde van het verschil in autovolume tussen de 2 gekozen autoklassen. De redenering hierachter is dat we verwachten dat gezinnen graag voertuigen bezitten van verschillende groottes (bv. een familiewagen en een kleinere wagen voor plezierritten), zie Train (1986).

[36] Merk op dat we bij de schatting van model C o.w.v. programmatorische beperkingen geen alternatief-specifieke constantes (ASC's) konden opnemen, in lijn met Train (1986).

[37] Een bijkomend (subtiel) verschil t.o.v. model A is dat de inkomensklassen geherdefinieerd werden: de groep met hoge inkomens begint nu al vanaf een netto-maandinkomen hoger dan 3000 EUR, ten opzichte van 4000 EUR vroeger. De reden voor deze aanpassing was een eindeloos itereren van de optimalisatiesoftware indien we bleven bij de inkomensdefinitie uit model A.

4.2.1. INTERPRETATIE SCHATTINGSRESULTATEN

Tabel 15 Schattingsresultaten voor model C

Verklarende variabele		Geschatte coëfficiënt	t-statistiek	95% betrouwbaarheidsinterval	
1.	**PR11TYLOW**	-0.29852E-04***	4.91	-0.41779E-04	-0.17924E-04
2.	**PR11TYHIG**	0.27980E-04***	7.50	0.20671E-04	0.35288E-04
3.	**VOLTGRGZ**	0.05347***	2.62	0.01354	0.09340
4.	**VOLTYLOW**	0.08116***	3.75	0.03877	0.12355
5.	**KWTGHFDYG**	-0.00675***	5.44	-0.00918	-0.00432
6.	**TRAFFTAXT**	-0.00372***	14.80	-0.00421	-0.00323
7.	**FUELCOSTT**	-0.03434***	2.80	-0.05834	-0.01034
8.	**LOGRCT**	0.09933***	4.03	0.05107	0.14759
9.	**EXP_VOLD**	0.71508***	4.70	0.41670	1.01346

*MNL-model geschat op 465 alternatieven (paren van autoklassen). Aantal observaties gebruikt voor de schatting: 866 van de 1145 gezinnen (279 gezinnen met een onbekende waarde voor 1 van de opgenomen variabelen werden geschrapt). De bekomen log likelihood-waarde bedraagt -8422. Een significantie van een parameter op het 1%, 5% of 10%-niveau wordt aangeduid met resp. ***, ** of *.*

Tabel 16 Verklaring variabelen opgenomen in model C

Verklarende variabele		Definitie
1.	**PR11TYLOW**	Interactie tussen de **som van de totale autoprijs** (aankoopprijs, BIV en CO2-premie, allen rekening houdend met de leeftijd van de auto) **van de 2 auto's** en een effects-coded **dummy voor lage** (netto ≤2000 EUR/mnd) **gezinsinkomens**
2.	**PR11TYHIG**	Interactie tussen de **som van de totale autoprijs** (aankoopprijs, BIV en CO2-premie, allen rekening houdend met de leeftijd van de auto) **van de 2 auto's** en een effects-coded **dummy voor hoge** (netto >3000 EUR/mnd) **gezinsinkomens**
3.	**VOLTGRGZ**	Interactie tussen **de som van de autovolumes** (lengte x breedte x hoogte) **van de 2 auto's** en een effects-coded **dummy voor grote gezinnen** (>4 gezinsleden)
4.	**VOLTYLOW**	Interactie tussen **de som van de autovolumes van de 2 auto's** en een effects-coded **dummy voor lage** (netto ≤2000 EUR/mnd) **gezinsinkomens**
5.	**KWTGHFDYG**	Interactie tussen **de som van het autovermogen van de 2 auto's** en een effects-coded **dummy voor jong gezinshoofd** (<40jr)
6.	**TRAFFTAXT**	**Som van de jaarlijkse verkeersbelasting van de 2 auto's**, te betalen in 2011
7.	**FUELCOSTT**	**Som van de brandstofkost van de 2 auto's** (EUR/100km)
8.	**LOGRCT**	**LOGRCT = LOG(aantal mogelijke combinaties van automodellen binnen het paar van autoklassen/totaal aantal mogelijke paren van automodellen over alle autoklassen)**, d.i. de link met het onderliggend (niet-geschatte) model voor keuze van een specifiek paar automodellen (i.t.t. de autoklassen)
9.	**EXP_VOLD**	Verwachte absolute waarde van het **verschil in autovolume** (lengte x breedte x hoogte) tussen de 2 auto's

In Tabel 15 en Tabel 16 tonen we respectievelijk de schattingsresultaten en de definitie van elke variabele. De niet-significant bevonden variabelen worden hier niet getoond: het gaat over VOLTMEDGZ, KWTGHFDOD en VOLTYHIG[38]. Merk op dat we voor model C geen pseudo-R^2 kunnen berekenen zoals voorheen, omdat de gebruikte software niet toelaat om meer dan 150 variabelen

[38] VOLTMEDGZ = "interactie tussen de som van de autovolumes van de 2 auto's en een effects-coded dummy voor middelgrote gezinnen (3 of 4 gezinsleden)"; KWTGHFDOD = "interactie tussen de som van het autovermogen van de 2 auto's en een effects-coded dummy voor een oud gezinshoofd (≥ 65jr)" ; VOLTYHIG = "Interactie tussen de som van de autovolumes van de 2 auto's en een effects-coded dummy voor hoge (netto >3000 EUR/mnd) gezinsinkomens". De tegenhanger van die laatste werd eerder ook al niet-significant bevonden in model A, terwijl de tegenhangers van de eerste twee voorheen wel significant bevonden werden in model A (zie Tabel 2).

te schatten. De nutsfuncties van het te schatten basismodel zouden immers enkel bestaan uit 465 (aantal alternatieven) -1 = 464 alternatief-specifieke constantes (ASC's), hetgeen de toegestane limiet overschrijdt. Om die reden beperken we ons tot een bespreking van de geschatte coëfficiënten.

Alle geschatte parameters zijn significant op het 1%-niveau en hebben het verwachte teken.

Voor variabele 1 en 2 zien we de kans op een bepaald paar van autoklassen afneemt voor gezinnen met lagere inkomens indien de som van de gemiddelde prijzen van die 2 autoklassen stijgt. Voor gezinnen met hoge inkomens merken zien we opnieuw dat er een soort 'snobeffect' bestaat.

Uit de schatter voor variabele 3 leren we dat grote gezinnen een hogere kans hebben op een bepaald paar van autoklassen indien de som van de autovolumes uit beide autoklassen stijgt.

Ook hier zien we dat voor gezinnen met een laag inkomen, een grotere som van de autovolumes over beide autoklassen een positieve invloed heeft op de kans om gekozen te worden (variabele 4.).

Net zoals in model A voelen gezinnen met een jong gezinshoofd zich over het algemeen minder aangetrokken tot paren van autoklassen met een grotere som van de vermogens (variabele 5.).

Anders dan in model A komt de jaarlijkse verkeersbelasting nu echter wel naar voren als significant verschillend van 0. Uit de schatter voor variabele 6 blijkt dat een hogere som van de verkeersbelasting over de 2 autoklassen resulteert in een significant lagere kans om gekozen te worden, ceteris paribus.

Ook logisch is het negatief teken voor variabele 7: hoe hoger de som van de brandstofkost per kilometer, hoe lager de kans dat een gezin zal kiezen voor een bepaald paar van autoklassen.

De schatter voor variabele 8 toont aan dat een hoger aantal mogelijke automodelcombinaties in het paar van autoklassen een positieve invloed heeft op de kans dat een bepaald paar van autoklassen gekozen wordt, hetgeen ook weer in de lijn van de verwachtingen ligt. Bovendien ligt de geschatte coëfficiënt in het interval [0 ; 1] en is het 95% betrouwbaarheidsinterval ook volledig een deelverzameling van dit interval. Dit impliceert dat deze schatter consistent is met de assumpties voor nutsmaximalisatie (zie hoger).

Variabele 9 geeft de verwachte waarde weer van de absolute waarde van het verschil in autovolume tussen beide autoklassen[39]. Dit is een goede indicator voor de mate van complementariteit tussen beide gekozen autoklassen. Zoals verwacht ondervinden de gezinnen een positieve invloed van een groter verschil in autovolume tussen beide gekozen autoklassen, ceteris paribus.

4.3. MODEL D: KEUZE AANTAL AUTO'S VOOR GEZINNEN MET 0, 1 OF 2 AUTO'S

Bij het ontwerp van model D hebben we de keuze van een gezin voor een bepaald aantal auto's gemodelleerd, waarbij de gezinnen de keuze hadden uit het bezitten van geen auto, 1 auto dan wel 2 auto's. In die zin is dit model een uitbreiding op model B.

[39] De wiskundige uitwerking van deze variabele wordt uitvoerig beschreven in Train, K. (1986) pp. 160-162.

Door de opname van een extra alternatief ten opzichte van model B (nl. de keuze voor het bezitten van 2 auto's) waren we echter wel genoodzaakt een belangrijke wijziging door te voeren aan het model. Zoals reeds aangegeven in de beschrijving van model A bis en model C, gaan we nu niet langer uit van 36 mogelijke alternatieven in de onderliggende autoklasse-keuze, maar slechts 30 autoklassen. De reden hiervoor is te vinden in de beperking op het aantal alternatieven in de gebruikte software.

Als inputtabel voor de schattingen nemen we in dit geval zowel de gezinnen met 1 auto, geen auto als de gezinnen met 2 auto's mee. Zoals voorheen werden enkel gezinnen met privé-wagens op benzine of diesel in bezit weerhouden in de analyse. Bij de schatting van het model werd verondersteld dat elk gezin 3 alternatieven heeft: 0 auto, 1 auto of 2 auto's in bezit. Bij het bezitten van 1 auto of 2 auto's houden we bovendien rekening met de verwachte waarde van het maximaal nut dat voortvloeit uit het kiezen van 1 of 2 auto's. Op die manier voorzien we een link tussen dit model D en de onderliggende modellen A bis en C.

4.3.1. INTERPRETATIE SCHATTINGSRESULTATEN

→ Algemeen

We schatten een MNL-model[40], omdat een geneste structuur bij 3 alternatieven weinig zin heeft. Bij het ontwerp van de nutsfuncties nemen we enkel een ASC op voor het 0-auto-alternatief, terwijl de nutsfunctie voor het 1-auto- en 2-auto-alternatief alle verklarende variabelen bevatten (zie Tabel 17). We schatten bovendien voor alle verklarende variabelen een alternatief-specifieke parameter, zodat we een beter zicht krijgen op de verschillende effecten van de variabelen op de keuze van beide alternatieven (1 auto of 2 auto's).

We zijn gestart van dezelfde set variabelen die significant bevonden werden in model B. Het finale model werd samengesteld door stapsgewijs enkele nieuwe variabelen toe te voegen aan de nutsfuncties voor het 1-auto- en 2-auto-alternatief. Zo werd de LOGSUM-variabele opnieuw toegevoegd, alsook enkele nieuwe variabelen die enkele paradoxen uit het afstandsmodel (model E, zie verder) zouden kunnen verklaren: variabelen die duiden op het gebruik van andere modi dan de auto (BUSWEK, FIETSWEK, TREINWEK), de leeftijd van het gezinshoofd (lineair en kwadratisch) en de woon-werkafstand (al dan niet kwadratisch).

Om te achterhalen in welke mate ons geschatte model de variatie in de geobserveerde keuzes verklaart, kunnen we weer een pseudo-R^2 berekenen. Als we opnieuw vergelijken met een basismodel dat enkel rekening houdt met waargenomen marktaandelen, bekomen we een pseudo-R^2 van 0.2627. De LL ratio-test bevestigt dat dit MNL-model beter presteert dan het basismodel:

$$-2 \times \left(LL_{basismodel} - LL_{geschat\ model}\right) = -2 \times (-2267.109 + 1671.502) = 1191.213$$
$$> \chi^2_{aantal\ nieuwe\ params\ geschat\ in\ geschatte\ model} = \chi^2_{30-2} = 41.337$$

[40] Merk op dat we ook hier weer strikt genomen een NL-structuur toepassen, maar dan in 2 stadia: FIML voor de keuze van de autoklasse(n) (model A bis en model C), en op basis daarvan de keuze voor het aantal voertuigen (model D).

→ **Parameterschattingen**

Tabel 17 Schattingen voor model D

Verklarende variabele	Geschatte coëfficiënt	t-statistiek	95% betrouwbaarheidsinterval	
1. LOGSUM2	0.21965***	5.22	0.13718	0.32011
2. LOGSUM1	0.35784**	2.41	0.06685	0.64884
3. GHFD_VR2	-0.18368*	1.68	-0.39797	0.03061
4. GHFD_VR1	-0.23966***	2.88	-0.40255	-0.07678
5. SINGLE2	-2.20145***	5.97	-2.92470	-1.47819
6. SINGLE1	-0.79457***	7.75	-0.99544	-0.59371
7. LEDENA1	-0.19888**	2.16	-0.37902	-0.01873
8. Y_LOW2	-1.86867***	12.33	-2.16566	-1.57169
9. Y_LOW1	-0.39457***	3.16	-0.63910	-0.15005
10. Y_HIGH2	1.02555***	7.90	0.77103	1.28006
11. DIPL_L2	-0.76937***	5.70	-1.03382	-0.50491
12. DIPL_L1	-0.60847***	5.50	-0.82530	-0.39165
13. DIPL_H2	0.57568***	3.48	0.25169	0.89968
14. DIPL_H1	0.44842***	3.02	0.15779	0.73904
15. GEMTH_GR2	-0.39526***	3.78	-0.60002	-0.19050
16. GEMTH_GR1	-0.22657***	2.60	-0.39722	-0.05593
17. BUSWEK2	-1.08041***	9.11	-1.31287	-0.84795
18. BUSWEK1	-0.68349***	7.76	-0.85608	-0.51089
19. FIETSWEK1	0.14135***	3.05	0.05062	0.23208
20. TREINWEK2	-0.57566***	2.92	-0.96167	-0.18965
21. TREINWEK1	-0.49402***	2.89	-0.82939	-0.15865
22. LFTGHFD2	0.20841***	5.09	0.12815	0.28868
23. LFTGHFD1	0.16489***	5.92	0.11027	0.21952
24. LFTGHFDE2_2	-0.00189***	5.16	-0.00261	-0.00117
25. LFTGHFDE2_1	-0.00145***	6.08	-0.00191	-0.00098
26. WNWRK2	0.06027***	5.01	0.03668	0.08387
27. WNWRK1	0.03116***	3.07	0.01126	0.05106
28. WNWRKE2_2	-0.00023***	2.86	-0.00039	-0.00007

*MNL-model geschat op 3 alternatieven (0 auto / 1 auto / 2 auto's). Aantal geldige observaties gebruikt voor de schatting: 2407. De bekomen log likelihood-waarde bedraagt -1672. Een significantie van een parameter op het 1%, 5% of 10%-niveau wordt aangeduid met resp. ***, ** of *.*

Tabel 18 Verklaring variabelen opgenomen in model D

Verklarende variabele	Definitie
1. LOGSUM2	De logsom of **inclusive value** van het **2-auto-alternatief**: dit komt overeen met de verwachte waarde van het maximale nut dat kan verkregen worden door een paar van specifieke onderliggende autoklassen te kiezen. Deze variabele vormt dus de link met het onderliggende model C.
2. LOGSUM1	De logsom of **inclusive value** van het **1-auto-alternatief**: dit komt overeen met de verwachte waarde van het maximale nut dat kan verkregen worden door een specifieke onderliggende autoklasse te kiezen. Deze variabele vormt dus de link met het onderliggende model A bis.
3. GHFD_VR2	Een effects-coded **dummy** voor een **vrouwelijk gezinshoofd** voor het **2-auto-alternatief**
4. GHFD_VR1	Een effects-coded **dummy** voor een **vrouwelijk gezinshoofd** voor het **1-auto-alternatief**
5. SINGLE2	Een effects-coded **dummy** voor gezinnen met slechts **1 gezinslid** voor het **2-auto-alternatief**
6. SINGLE1	Een effects-coded **dummy** voor gezinnen met slechts **1 gezinslid** voor het **1-auto-alternatief**
7. LEDENA1	Het (discreet) **aantal leden in het gezin**, voor het **1-auto-alternatief**

Verklarende variabele	Definitie
8. **Y_LOW2**	Een effects-coded **dummy** voor **lage** (netto ≤2000 EUR/mnd) **gezinsinkomens** voor het **2-auto-alternatief**
9. **Y_LOW1**	Een effects-coded **dummy** voor **lage** (netto ≤2000 EUR/mnd) **gezinsinkomens** voor het **1-auto-alternatief**
10. **Y_HIGH2**	Een effects-coded **dummy** voor **hoge** (netto >4000 EUR/mnd) **gezinsinkomens** voor het **2-auto-alternatief**
11. **DIPL_L2**	Een effects-coded **dummy** voor **gezinshoofd** met als hoogst behaalde **diploma 'lager onderwijs' of minder**, voor het **2-auto-alternatief**
12. **DIPL_L1**	Een effects-coded **dummy** voor **gezinshoofd** met als hoogst behaalde **diploma 'lager onderwijs' of minder**, voor het **1-auto-alternatief**
13. **DIPL_H2**	Een effects-coded **dummy** voor **gezinshoofd** met als hoogst behaalde **diploma 'hoger onderwijs'**, voor het **2-auto-alternatief**
14. **DIPL_H1**	Een effects-coded **dummy** voor **gezinshoofd** met als hoogst behaalde **diploma 'hoger onderwijs'**, voor het **1-auto-alternatief**
15. **GEMTH_GR2**	Een effects-coded **dummy** voor gezinnen met een **grootstad of Vlaamse centrumstad als thuisadres**, voor het **2-auto-alternatief.** Zou beschouwd kunnen worden als een proxy voor beschikbaarheid van goed openbaar vervoer.
16. **GEMTH_GR1**	Een effects-coded **dummy** voor gezinnen met een **grootstad of Vlaamse centrumstad als thuisadres**, voor het **1-auto-alternatief.** Zou beschouwd kunnen worden als een proxy voor beschikbaarheid van goed openbaar vervoer.
17. **BUSWEK2**	Een effects-coded **dummy** voor respondenten die **minstens wekelijks gebruikmaken van de modus 'bus'**, voor het **2-auto-alternatief**
18. **BUSWEK1**	Een effects-coded **dummy** voor respondenten die **minstens wekelijks gebruikmaken van de modus 'bus'**, voor het **1-auto-alternatief**
19. **FIETSWEK1**	Een effects-coded **dummy** voor respondenten die **minstens wekelijks gebruikmaken van de modus 'fiets'**, voor het **1-auto-alternatief**
20. **TREINWEK2**	Een effects-coded **dummy** voor respondenten die **minstens wekelijks gebruikmaken van de modus 'trein'**, voor het **2-auto-alternatief**
21. **TREINWEK1**	Een effects-coded **dummy** voor respondenten die **minstens wekelijks gebruikmaken van de modus 'trein'**, voor het **1-auto-alternatief**
22. **LFTGHFD2**	De **leeftijd** van het **gezinshoofd**, voor het **2-auto-alternatief**
23. **LFTGHFD1**	De **leeftijd** van het **gezinshoofd**, voor het **1-auto-alternatief**
24. **LFTGHFDE2_2**	Het **kwadraat** van de **leeftijd** van het **gezinshoofd,** voor het **2-auto-alternatief** = (LFTGHFD2)^2
25. **LFTGHFDE2_1**	Het **kwadraat** van de **leeftijd** van het **gezinshoofd,** voor het **1-auto-alternatief** = (LFTGHFD1)^2
26. **WNWRK2**	**Afstand** (km) van de **woonplaats** tot het vaste **werk- of schooladres**, voor het **2-auto-alternatief**
27. **WNWRK1**	**Afstand** (km) van de **woonplaats** tot het vaste **werk- of schooladres**, voor het **1-auto-alternatief**
28. **WNWRKE2_2**	**Kwadraat** van de **afstand** (km) van de **woonplaats** tot het vaste **werk- of schooladres**, voor het **2-auto-alternatief** = (WNWRK2)^2

Alle geschatte parameters zijn significant op het 1%-niveau, behalve LOGSUM1, LEDENA1 (beide significant op 5%-niveau) en GHFD_VR1 (op 10%-niveau). De specifieke schattingen en de verklaringen van de opgenomen variabelen kan men terugvinden in Tabel 17 en Tabel 18. Merk op dat alle variabelen ofwel werden opgenomen in de nutsfuncties van het 1-auto alternatief ofwel het 2-auto-alternatief (zie hoger). Bijvoorbeeld: positieve coëfficiënten voor variabelen met suffix '2' wijzen op variabelen met een positieve invloed op het 2-auto-alternatief en negatieve coëfficiënten voor variabelen met suffix '2' wijzen op variabelen met een negatieve invloed op het 2-auto alternatief.

De coëfficiënten voor variabele 1 en 2 wijzen op een duidelijk positieve bijdrage van het maximale nut dat men kan halen uit het bezit van een of twee wagens. Zoals hierboven bij de bespreking van model B reeds uitgelegd, betekent dat dat de waarschijnlijkheid dat het gezin een bepaald aantal auto's kiest, mee wordt bepaald door (a) de technische kenmerken van het beschikbare wagenpark

(b) de manier waarop de socio-economische kenmerken van het gezin de gekozen autoklassen beïnvloedt.

De schattingen voor variabele 3 en 4 tonen aan dat gezinnen met een vrouwelijk gezinshoofd een lagere kans hebben op het bezitten van 1 of 2 auto's dan gezinnen met een mannelijk gezinshoofd, ceteris paribus. Bovendien is de kans op het bezitten van 1 auto nog kleiner dan het bezitten van 2 auto's, indien een vrouw aan het hoofd van het gezin staat.

Gezinnen bestaande uit slechts 1 gezinslid hebben een grotere kans op het bezitten van 0 auto's dan op het bezitten van 1 auto en zeker dan op het bezitten van 2 auto's. Dit blijkt uit de parameterschattingen van variabele 5 en 6, en gegeven dat deze dummies op 0 werden genormaliseerd voor het 0-auto-alternatief.

Kijken we naar het totaal aantal gezinsleden (variabele 7), dan zien we dat die een negatieve coëfficiënt krijgt in de nutsfunctie van het 1-auto-alternatief (voor het 2-auto-alternatief werd die niet-significant bevonden). Hieruit leiden we af dat bij een groeiend aantal gezinsleden, de kans op het 1-auto-alternatief zal dalen, terwijl de kans op het 0-auto- of 2-auto-alternatief niet significant wijzigt, ceteris paribus. Het feit dat deze variabele slechts significant is op het 5%-niveau heeft er vermoedelijk mee te maken dat er een zekere overlap bestaat tussen de dummies 5 en 6, en de discrete variabele 7.

De inkomensdummies voor gezinnen met een laag inkomen (variabele 8 en 9) wijzen op een dalende kans op het bezitten van 1 auto over geen auto's, en van 2 auto's over 1 auto, als het gezin een laag inkomen heeft. Voor hoge inkomens vonden we enkel een positieve impact op de kans op het bezitten van 2 auto's (variabele 10).

Naast het inkomensniveau heeft ook het hoogst behaalde diploma een significante invloed in de nutsfuncties. Zo hebben gezinshoofden met als hoogst behaalde diploma 'lager onderwijs' of minder (variabele 11 en 12), een dalende kans op het bezit van 1 auto over geen auto, en van 2 auto's over 1 auto. Voor gezinshoofden met een diploma 'hoger onderwijs' ligt de kans op het bezitten van 2 auto's hoger dan de kans op het bezitten van 1 auto, en die van 1 auto hoger dan die van geen auto (variabele 13 en 14).

Gezinnen met een grootstad of Vlaamse centrumstad als thuisbasis hebben een grotere kans op het bezitten van 0 auto's dan op het bezitten van 1 auto, en zeker dan het bezitten van 2 auto's. Dit blijkt uit de geschatte coëfficiënten van variabele 15 en 16.

Gezinnen die minstens op wekelijkse basis gebruikmaken van openbaar vervoersmodi 'bus' en 'trein', hebben een kleinere kans op het bezitten van 2 auto's dan het bezitten van 1 auto, en ook een grotere kans op het bezitten van 1 auto dan geen auto. Dit blijkt uit de schattingen van variabele 17, 18, 20 en 21. Ook de effecten van metro- en tramgebruik werden initieel in de schatting opgenomen, maar werden niet significant bevonden en dus niet weerhouden in Tabel 17. Voor het 1-auto-alternatief werd verder nog een significant effect gevonden voor variabele 19, een dummy die aangeeft of men minstens wekelijks gebruikmaakt van de fiets. Uit de schatting blijkt dat respondenten die minstens wekelijks gebruikmaken van de fiets, een hogere kans hebben op het bezitten van 1 auto dan op het bezitten van geen auto of 2 auto's[41].

[41] Het is moeilijk de causaliteit van deze effecten van het gebruik van alternatieve modi in te schatten. Het is immers best mogelijk dat een gezin slechts 1 auto bezit omdat veel wordt gebruikgemaakt van openbaar vervoer. Maar ook het omgekeerde is denkbaar: een gezin maakt juist veel gebruik van openbaar vervoer omdat ze slechts 1 auto bezitten.

Uit de schattingen horende bij variabele 22 en 24 leren we dat voor het 2-auto-alternatief, een gezinshoofd met stijgende leeftijd eerst een nut zal ondervinden van het bezit van 2 auto's, maar eenmaal een bepaalde leeftijd bereikt zal het nut van het bezit van 2 auto's weer dalen. Voor het 2-auto-alternatief wordt de hoogste kans op het bezit van 2 auto's bereikt op de leeftijd van ca. 55 jaar[42], ceteris paribus. Dit impliceert dat het bezit van meerdere auto's slechts interessant blijkt indien men zich in een zeer specifieke levensfase bevindt, bv. iets oudere kinderen, budgettaire ruimte, etc.
Vergelijken we die resultaten met de schattingen voor variabele 23 en 25, dan nemen we daar een gelijkaardig parabolisch patroon waar. Het enige verschil is dat de gezinshoofdleeftijd waarop de kans op het bezit van 1 auto maximaal wordt, iets hoger ligt (57 jaar) dan voor het 2-auto-alternatief. Dit is in lijn met hetgeen we a priori zouden verwachten.

Resten ons nog de variabelen die betrekking hebben op de afstand tussen de woonplaats en het vaste werk- of schooladres van de respondent. Ook hier leiden we af uit de schattingen horende bij variabele 26 en 28 dat het nut (en dus de kans) van het 2-auto-alternatief een parabolisch verloop kent dat initieel oploopt naarmate de woon-werkafstand stijgt, en daarna weer daalt. Dit ligt in lijn met hetgeen we a priori zouden verwachten, omdat voor zeer korte afstanden (te voet, fiets) en zeer lange afstanden (trein) vaak vermoedelijk andere modi (dan de auto) geprefereerd worden. Voor het 1-auto-alternatief vonden we geen indicaties voor een dergelijk parabolisch verloop, en gaan we dus uit van een lineaire schatter: als de woon-werkafstand stijgt zal de kans op het bezit van 1 auto ook stijgen, ceteris paribus.

[42] Het punt (leeftijd gezinshoofd) waarop de probabiliteit op het bezitten van 2 auto's maximaal is, kon worden berekend door de kans op het bezit van 2 auto's af te leiden naar de variabele LFTGHFD, die vergelijking gelijk aan 0 te stellen, en op te lossen naar LFTGHFD.

HOOFDSTUK 5. MODEL E: AFSTANDSMODEL

5.1. INLEIDING

De vragen die we tot nu toe hebben bekeken hadden altijd betrekking op discrete keuzen, zoals:

- Hoeveel auto's bezit een gezin?
- Gegeven het aantal auto's dat een gezin bezit, wat is het type auto dat een gezin kiest?

De vraag die we in dit hoofdstuk beschouwen is echter continu van aard:

- Indien een huishouden één voertuig bezit, wat is dan de verwachte waarde van het aantal kilometers die het huishouden per jaar met dat voertuig zal afleggen, rekening houdende met de kenmerken van het gezin én van het gekozen voertuig?
- Indien een huishouden twee voertuigen bezit, wat is dan de verwachte waarde van het aantal kilometers die het huishouden per jaar met elk voertuig zal afleggen? Bij deze vraag zullen we niet alleen kijken naar de hierboven vermelde elementen, maar ook rekening houden met de kenmerken van het andere voertuig in het bezit van het huishouden.

De aard van de vragen die we hier stellen is dus fundamenteel verschillend van de eerdere vragen. Voor het schatten van dergelijke continue modellen kunnen we normaal gezien terugvallen op vertrouwde regressiemodellen.

Er is echter een belangrijke complicatie: het continu model kan niet onafhankelijk worden beschouwd van de resultaten van de discrete keuze modellen die we hiervoor hebben geschat.

In de *structuur van het model* kijken we inderdaad *achtereenvolgens* naar drie beslissingen: (a) het aantal auto's dat een gezin bezit, (b) het type auto (c) en het aantal afgelegde kilometers. In *werkelijkheid* worden deze beslissingen door een gezin *niet sequentieel* genomen: een gezin neemt beslissingen met betrekking tot het aantal en type auto's, rekening houdende met de afstanden die het verwacht af te leggen. Bijgevolg zullen een aantal factoren die de jaarlijks afgelegde afstanden beïnvloeden, ook een impact hebben op de keuze van het aantal auto's en van het type auto. Het gebruik van de standaard methode der kleinste kwadraten (Ordinary Least Squares – OLS) zal dan leiden tot inconsistente schattingen van de vraagfunctie.

De oplossing voor het probleem zal afhangen van de achterliggende oorzaak van de inconsistentie: voor bepaalde problemen zullen we moeten overgaan tot het gebruik van Instrumentele Variabelen (IV), voor andere zullen we moeten overgaan tot een correctie van zelfselectie bias.

We illustreren beide problemen aan de hand van een voorbeeldje. We volgen hierbij Train (1993, Hoofdstuk 5) – voor een meer exhaustieve benadering verwijzen we naar dezelfde bron.

5.1.1. GEBRUIK VAN INSTRUMENTELE VARIABELEN

Veronderstellen we eenvoudigheidshalve dat het aantal afgelegde kilometers (*D*) op jaarbasis van twee parameters afhangt: het inkomen van het gezin, *Y*, en de kostprijs per kilometer, *P*. De vraagfunctie kan dan als volgt worden geformuleerd (waarbij *e* de foutenterm is):

$$D = b_1 + b_2 P + b_3 Y + e$$

Deze foutenterm *e* stelt de som voor van de invloeden van de niet-waargenomen parameters. Een cruciale hypothese in het gebruik van OLS is dat deze foutenterm onafhankelijk is van de onafhankelijke variabelen (hier *Y* en *P*).

Dit zal echter meestal niet het geval zijn. Stel bijvoorbeeld dat het huishouden anticipeert dat het op jaarbasis veel kilometers zal moeten afleggen. Er zijn hiervoor meerdere redenen denkbaar: bijvoorbeeld omdat het gezinshoofd veel kilometers moet afleggen om professionele redenen, of omdat de kinderen hobby's hebben waarvoor lange verplaatsingen nodig zijn, of omdat een van de gezinsleden zorg moet dragen voor een hulpbehoevende ouder die ver weg woont. Deze motieven worden echter niet altijd geobserveerd in de steekproef, en hun impact zal dus een invloed uitoefenen op de grootte van de foutenterm. Het gaat hier telkens over elementen die op korte termijn niet kunnen gewijzigd worden. Daarom kunnen we er ook van uitgaan dat dergelijke huishoudens een auto zullen kiezen die relatief zuinig is in het gebruik. Het statistisch gevolg van deze keuze is dan dat de prijs en de foutenterm gecorreleerd zijn: een gezin dat omwille van niet-geobserveerde motieven meer kilometers aflegt dan men zou verwachten op basis van de geobserveerde variabelen[43], zal ook een auto kiezen met een lagere kost per kilometer.

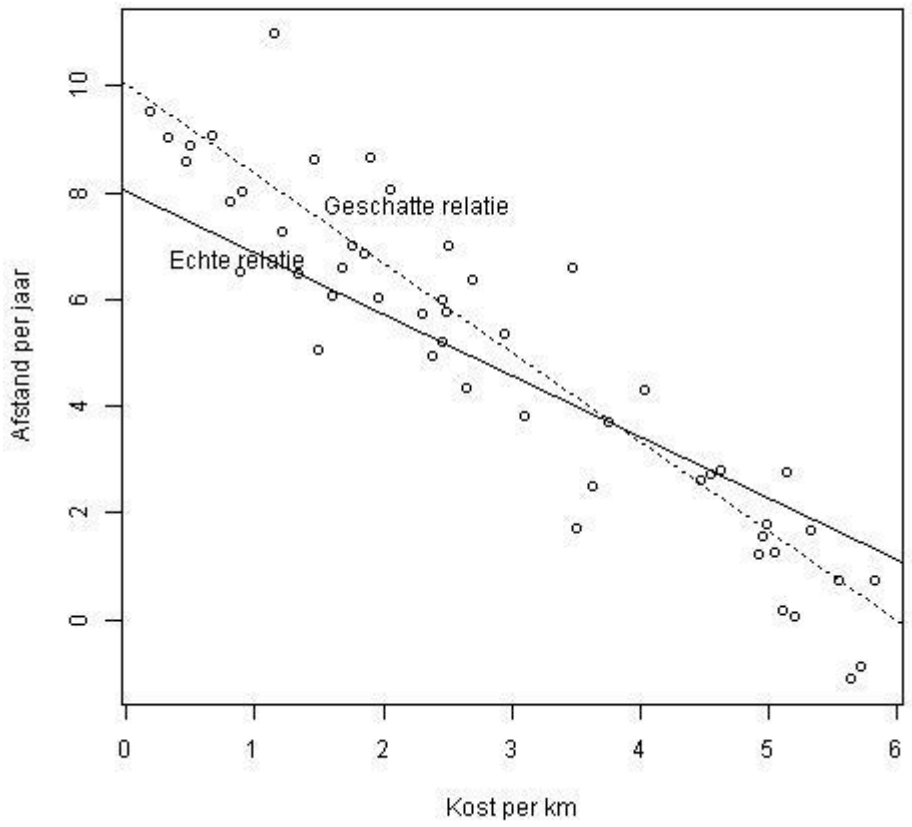
Het probleem is dus dat de beslissing om een bepaald type auto te kopen (de discrete keuze die we al hebben gemodelleerd) ook een beslissing inhoudt met betrekking tot de kostprijs per kilometer van het automodel, en dus ook mee beïnvloed wordt door de afstanden die een huishouden verwacht af te leggen.

Dit kan geïllustreerd worden aan de hand van een voorbeeld.

In Figuur 2 stelt de doorlopende rechte lijn de "echte" relatie voor tussen het aantal afgelegde kilometers en de kostprijs per kilometer. Uit de redenering hierboven weten we dat, indien een auto een lage kost per kilometer heeft, het aantal afgelegde kilometers hoger zal liggen dan men zou verwachten op basis van andere geobserveerde gegevens (zoals, in dit voorbeeld, het gezinsinkomen). Daardoor zullen de meeste observaties bij lage eenheidskosten boven de "echte" lijn liggen. Het omgekeerde gaat op voor hoge eenheidskosten. Het resultaat is dat men bij gebruik van OLS de "stippellijn" zal schatten als de relatie tussen afstand en eenheidskost: de absolute waarde van de richtingscoëfficiënt van deze rechte is hoger, en dus zal men de gevoeligheid van de afgelegde afstanden als functie van de kost overschatten.

Het gebruik van OLS leidt dus tot een systematische fout in de schatting van de relatie tussen de afgelegde afstand en de kostprijs per kilometer. Een uitbreiding van de steekproef zal dit probleem niet verhelpen: de geschatte parameters zullen dus inconsistent zijn.

[43] Zodat er een grotere fout zal ontstaan tussen de reële waarde van de afgelegde afstand en de voorspelling op basis van OLS.



Figuur 2: Endogeniteitsbias

De oplossing voor dit probleem bestaat er in dat men het model zal schatten in twee fasen. In een eerste fase zal de endogene variabele (in dit geval, de eenheidskost) geschat worden als een functie van variabelen waarvan men redelijkerwijze kan uitgaan dat ze echt exogeen zijn. In een tweede fase zal men dan het aantal afgelegde kilometers schatten als een functie van de geschatte waarde van de eenheidskost.

Deze manier van werken (Two Stage Least Squares of 2SLS) is equivalent met de methode van de Instrumentele Variabelen, waarbij de exogene variabelen dienst doen als instrumenten – we verwijzen naar elke standaard handboek econometrie (zie bijvoorbeeld Greene, 2012) voor een gedetailleerde bespreking van deze methode.

De keuze van de instrumenten is echter verre van evident. Enerzijds moeten deze instrumenten gecorreleerd zijn met de endogene variabele[44]: anders kan men op basis van de instrumenten geen nauwkeurige voorspelling maken van deze endogene variabele. Anderzijds mogen deze instrumenten niet gecorreleerd zijn met de foutenterm, want dan verplaatst men gewoon het probleem.

[44] In het voorbeeld dat we hier beschouwen, de kostprijs per kilometer.

De eenheidskost van een wagen wordt bijvoorbeeld mee bepaald door variabelen zoals het vermogen van de motor en het gewicht van de auto. Dus, vanuit het standpunt van het eerste instrument zijn zij potentieel interessante instrumenten.

Maar kunnen we er van uit gaan dat ze niet gecorreleerd zijn met de foutenterm? Sommigen zouden kunnen betogen dat mensen die relatief grote afstanden afleggen meer belang zullen hechten aan het rijcomfort, en dat wordt mede bepaald door de accelaratie van een auto, dat op zijn beurt wordt bepaald door factoren zoals het gewicht van de auto en het vermogen.

We komen hieronder terug op dit specifiek punt. We zullen verder ook bespreken welke formele testen dienen uitgevoerd te worden.

5.1.2. CORRECTIE VOOR ZELF-SELECTIE

Het tweede probleem vloeit voort uit de bedenking dat het aantal voertuigen in een gezin ook niet onafhankelijk is van het aantal kilometers dat het gezin verwacht af te leggen met de wagen. Een gezin kan immers op basis van een aantal parameters (woon-werkafstand, aantal familieleden, hobby's van de familieleden, beschikbaarheid van openbaar vervoer, afstand tot familieleden en vrienden) redelijk anticiperen of het met 1 wagen zal toekomen. Dat betekent dat de parameters die een invloed zullen uitoefenen op het aantal afgelegde kilometers ook een invloed zullen uitoefenen op het aantal wagens in het gezin.

Het lijkt ook redelijk om ervan uit te gaan dat een bepaalde parameter op een andere manier het aantal kilometers zal afleggen wanneer een gezin 1 auto bezit dan wanneer het er 2 bezit. Daarom zullen we twee aparte modellen schatten, in functie van het aantal auto's dat een gezin bezit.

Maar dit creëert een nieuwe bron van inconsistentie.

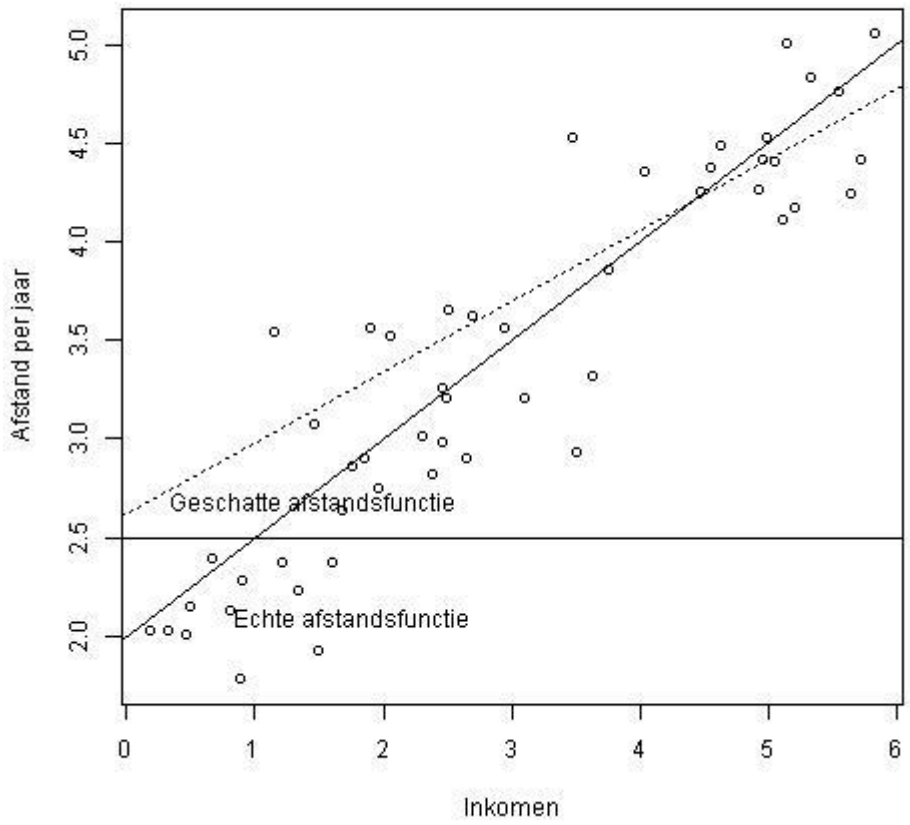
Beschouwen we bijvoorbeeld een huishouden met een laag inkomen. De kans dat een huishouden met een laag inkomen twee auto's kiest is lager dan voor een huishouden met een hoog inkomen, ceteris paribus. Indien dat arm huishouden toch twee auto's kiest, dan zal dat te wijten zijn aan (niet-geobserveerde) factoren die meestal ook tot gevolg hebben dat het huishouden meer kilometers zal afleggen dan men zou verwachten op basis van geobserveerde factoren. Dus, voor gezinnen met lage inkomens en 2 auto's, zal de foutenterm op de afstandsfunctie groot zijn.

Ook hier kunnen we de oorsprong van de bias grafisch illustreren[45]. In Figuur 2 stellen we de "echte" relatie tussen het inkomen en het aantal afgelegde kilometers (in 10^4 km) voor aan de hand van een lineaire functie:

$$D = b_1 + b_3 Y + e$$

Aangezien we niet alle relevante factoren kunnen opnemen in het model, zullen we waarnemingen aan beide kanten van de rechte verspreid liggen. Stel nu dat gezinnen die meer dan 20.000 km afleggen een tweede auto kopen. Indien we de relatie tussen de jaarlijks afgelegde afstanden en het inkomen enkel schatten op basis van de waarnemingen voor gezinnen met twee auto's, dan zal de "afstandsfunctie" geschat worden als de rechte stippellijn.

[45] Merk op dat deze illustratie niet rechtstreeks van toepassing is op het hier beschouwde model, vermits wij de jaarlijks afgelegde afstanden per wagen beschouwen, niet per gezin.



Figuur 3: Zelfselectiebias

Dit betekent dat, voor gezinnen met twee auto's, de gevoeligheid van de afgelegde afstanden als functie van het inkomen zal worden onderschat.

Deze waarneming gaat op voor alle variabelen die de probabiliteit beïnvloeden van het aantal wagens dat wordt gekozen: ze zijn allemaal gecorreleerd met de foutenterm indien men OLS toepast.

De foutenterm voor gezinnen met twee wagens kan dan gesplitst worden in twee termen:

$$e_2 = E(e_2) + \vartheta$$

Waar $E(e_2)$ een functie is van de waarschijnlijkheid dat een gezin twee wagens kiest, en ϑ een foutenterm is die niet gecorreleerd is met variabelen die het aantal auto's beïnvloeden. $E(e_2)$ is de correctieterm voor de selectiviteit.

Indien men $E(e_2)$ kent, dan kan men de volgende functie schatten aan de hand van OLS:

$$D = b_1 + b_3 Y + E(e_2) + \vartheta$$

De overblijvende moeilijkheid is dan om de correctieterm te bepalen.

Laten we eerst een paar bijkomende variabelen definiëren:

- σ is de standaardafwijking op het OLS model
- *J=1,…,M* zijn de verschillende discrete keuzemogelijkheden
- P_j is de waarschijnlijkheid dat alternatief J wordt gekozen
- r_j is de correlatiecoëfficiënt tussen de foutenterm in het lineair model en de niet-observeerbare variabelen voor elk alternatief in het discrete keuzemodel

In het geval dat wij beschouwen zijn de M alternatieven: (1) het gezin kiest twee auto's, (2) het gezin kiest 1 auto (3) en het gezin kiest geen auto.

Laten we nu de correctieterm beschouwen voor de afstandsfunctie voor alternatief 1. Dubin en McFadden (1984) hebben aangetoond dat, als de waarschijnlijkheid dat een bepaald aantal auto's wordt gekozen logit verdeeld is, en de OLS foutentermen normaal verdeeld zijn, de correctieterm dan gelijk is aan:

$$\sigma \frac{\sqrt{6}}{\pi} \sum_{j=2,\cdots,M} r_j \left(\frac{P_j ln(P_j)}{1-P_j} + ln(P_1) \right)$$

Aangezien de keuzewaarschijnlijkheden reeds gekend zijn (zie Model D), kan voor elke j de term σr_j met OLS geschat worden samen met de afstandsfunctie.

5.1.3. BESCHRIJVENDE STATISTIEKEN: ALGEMEEN

Om het afstandsmodel te schatten beschikken we over twee belangrijke bronnen van informatie. Enerzijds hebben we op basis van de Febiac databank zicht op de belangrijkste technische en economische kenmerken van de auto's. Anderzijds biedt het Onderzoek Verplaatsingsgedrag (OVG) ons een berg aan informatie, niet alleen met betrekking tot de socio-economische kenmerken van de huishoudens, maar ook met betrekking tot hun verplaatsingsgedrag.

Laten we eerst even de gebruikte gegevens uit de **Febiac databank** nader bekijken. Hierbij wensen we te benadrukken dat we, in alles wat volgt, verwijzen naar de specifieke kenmerken van een merk en model[46], en niet naar klasse-gemiddeldes zoals bij het discrete keuzemodel: in een continu model is er immers geen reden om gegevens te groeperen.

De belangrijkste auto-specifieke economische determinant van het wagengebruik is de kostprijs per kilometer. We beschikken echter niet over gegevens met betrekking tot de onderhouds- en slijtagekost per kilometer, en beperken ons bijgevolg tot de **brandstofkost** (in EUR) per km. Van de andere economische parameters (de kost van de aankoop en de jaarlijkse verkeersbelasting) verwachten we niet dat ze het *gebruik* van de auto beïnvloeden, aangezien het hier gaat over *vaste* kosten. Een verkennende statistische analyse heeft inderdaad uitgewezen dat het effect van deze variabelen niet statistisch significant was.

Daarnaast gaan we er ook van uit dat mensen met een minder comfortabele auto minder geneigd zullen zijn om lange afstanden af te leggen. Daarom nemen we het volume van de auto mee als

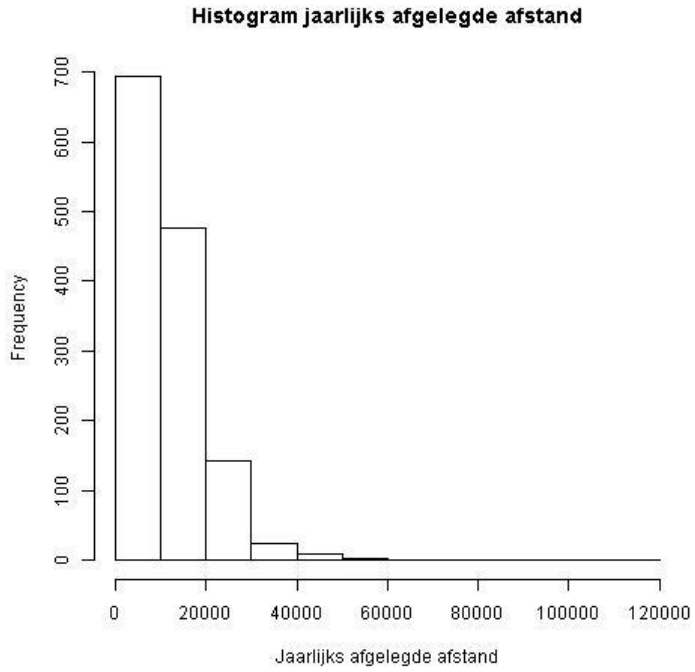
[46] De enige beperking is dat een bepaalde merk-model combinatie soms kan verwijzen naar meerdere varianten. We hebben dan altijd verondersteld dat het gezin de goedkoopste variant koos.

indicator van het comfortniveau – andere indicatoren zoals de maximale snelheid en het vermogen per eenheid gewicht bleken bij een verkennende analyse niet statistisch significant te zijn. Naast het brandstofverbruik per kilometer, zal ook het volume van de auto waarschijnlijk geplaagd worden door endogeniteitsbias.

Daarnaast hebben we ook gegevens gebruikt uit het **OVG**.

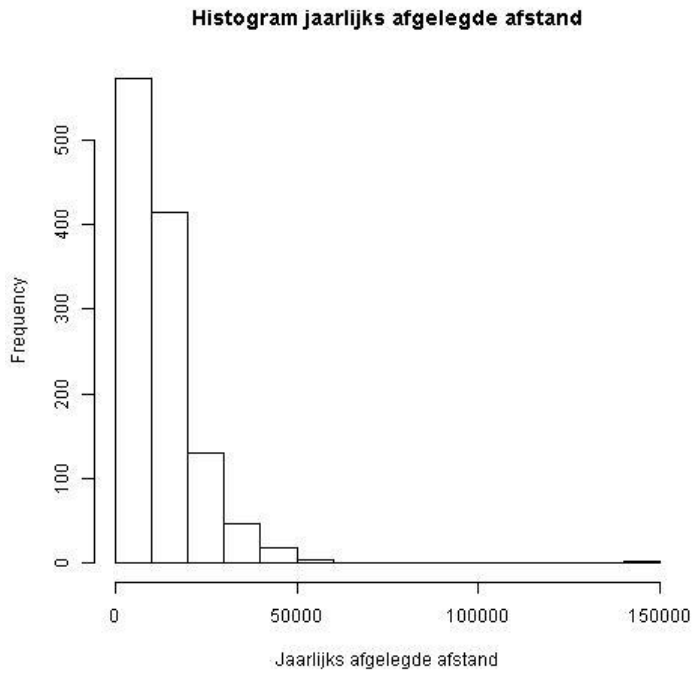
Uit het voertuigkeuze model weten we al dat het aantal voertuigen per gezin mee wordt bepaald door een aantal socio-economische variabelen. Hier zullen we alleen kijken naar de variabelen waar er inderdaad belangrijke verschillen bestaan tussen gezinnen met 1 auto en gezinnen met 2 auto's.

Laten we eerst kijken naar de afhankelijke variabele: de **jaarlijks afgelegde afstanden** per auto. We zien onmiddellijk dat alle percentielen en het gemiddelde van deze variabele lager liggen bij gezinnen met een auto dan bij gezinnen met twee auto's. We moeten hier ook bij opmerken dat het gaat over de afgelegde afstanden pér auto, niet per gezin. Dit wijst er op dat het verplaatsingsgedrag bij gezinnen met twee auto's van een heel andere aard is dan gezinnen met 1 auto[47], en dat het dus inderdaad aangewezen is om twee aparte modellen te schatten.



Figuur 4: verdeling van de jaarlijks afgelegde afstand (gezinnen met 1 wagens)

[47] We moeten ook voor ogen blijven houden dat we hier enkel gezinnen beschouwen die de gebruikte auto's ook effectief in hun bezit hebben. We beschouwen hier dus niet het gedrag van gezinnen met een of meerdere bedrijfswagens.



Figuur 5: verdeling van de jaarlijks afgelegde afstand (gezinnen met 2 wagens)

Tabel 19: kwartielen van de jaarlijks afgelegde afstand per wagen

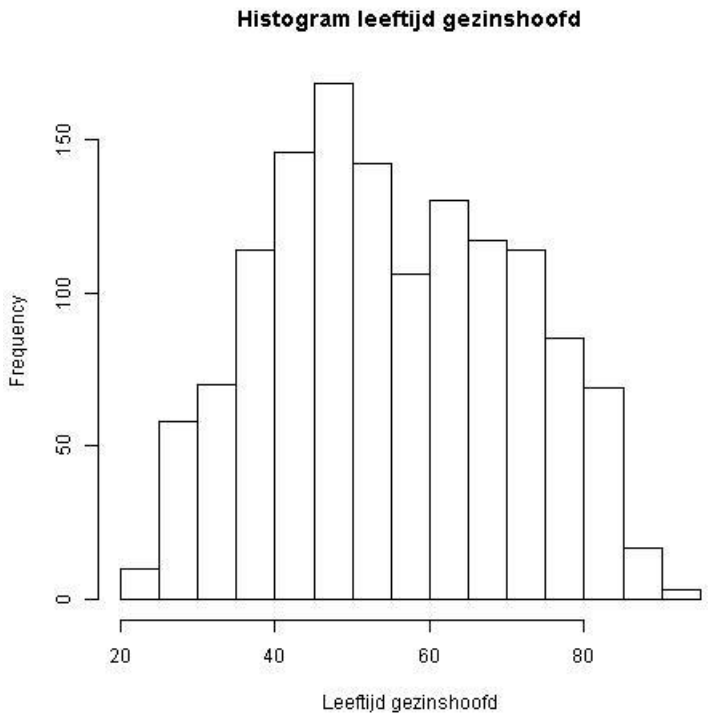
Aantal auto's per gezin	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5	6000	10000	12540	17000	120000
2	10	7000	11000	14050	20000	150000

Vervolgens beschouwen we een aantal onafhankelijke variabelen die een bron kunnen zijn van selectiebias.

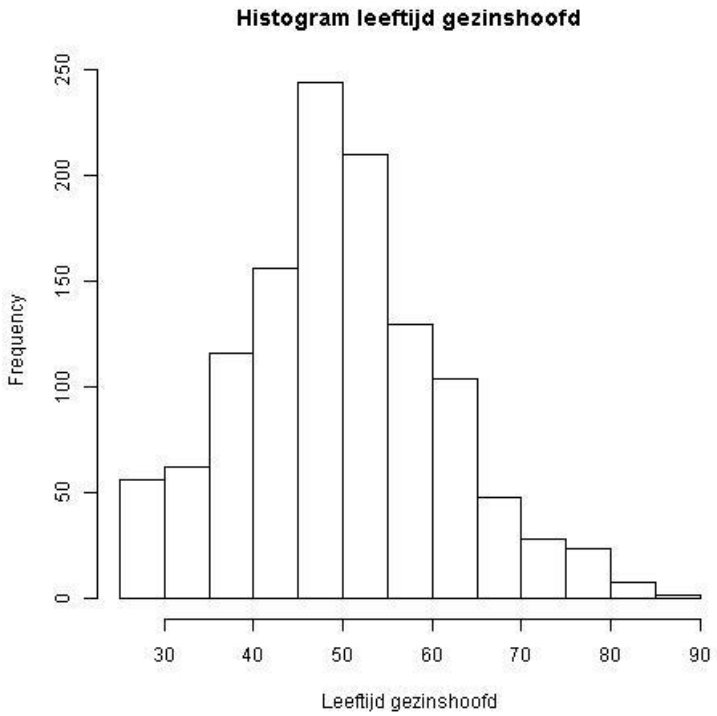
Een eerste punt heeft betrekking op de verdeling van de **leeftijden van de gezinshoofden**.

Uit onderstaande histogrammen en tabellen merken we volgende punten op:

- De minimumleeftijd van het gezinshoofd in gezinnen met twee wagens ligt hoger dan bij gezinnen met 1 wagen.
- De maximumleeftijd van het gezinshoofd in gezinnen met twee wagens ligt lager dan bij gezinnen met 1 wagen.
- Zowel de mediane als de gemiddelde leeftijd van het gezinshoofd in gezinnen met twee wagens ligt lager dan bij gezinnen met 1 wagen.
- Het derde kwartiel van de leeftijd van het gezinshoofd in gezinnen met twee wagens ligt lager dan bij gezinnen met 1 wagen.



Figuur 6: leeftijdsverdeling van het gezinshoofd (gezinnen met 1 wagen)



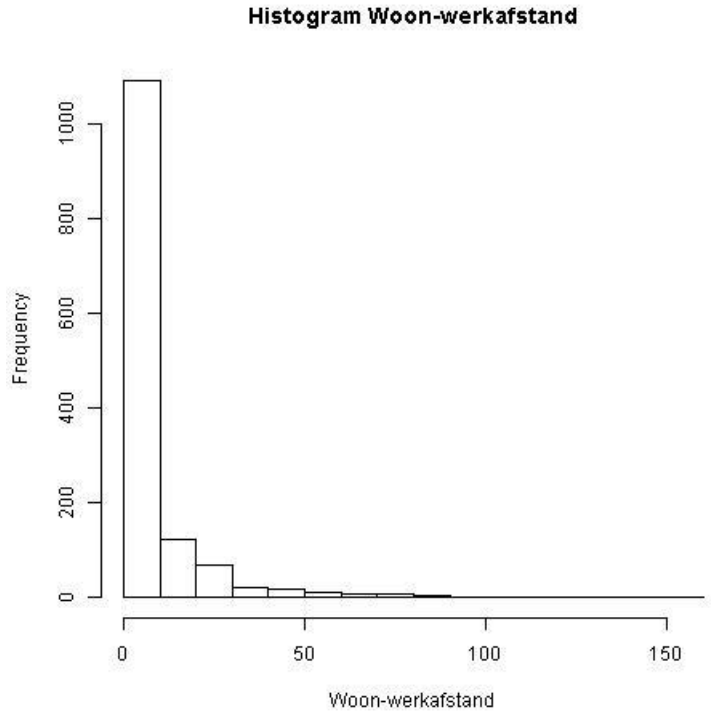
Figuur 7: leeftijdsverdeling van het gezinshoofd (gezinnen met 2 wagens)

Tabel 20: kwartielen leeftijd van het gezinshoofd (gezinnen met 1 wagen)

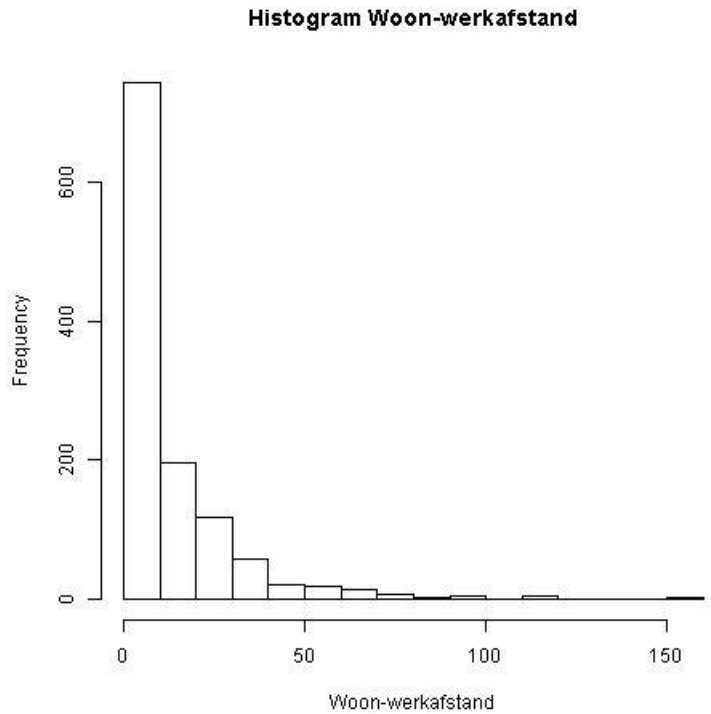
Aantal auto's per gezin	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	22	43	54	55.65	68	93
2	25	43	49	50.16	57	86

Deze vaststellingen suggereren dat gezinnen met twee wagens een ander leeftijdsprofiel hebben dan gezinnen met 1 wagen: het "typisch" leeftijdsprofiel van deze gezinnen is dat van mensen die kinderen in huis hebben die nog niet onafhankelijk zijn op mobiliteitsvlak. Dit werd bevestigd in het keuzemodel.

Een tweede punt heeft betrekking op de verdeling van **de woon-werkafstanden**.



Figuur 8: verdeling van de woon-werkafstand (gezinnen met 1 wagen)



Figuur 9: verdeling van de woon-werkafstand (gezinnen met 2 wagens)

Uit Tabel 21 kunnen we direct opmerken dat de mediaan, het gemiddelde en het derde kwartiel van de woon-werkafstanden in gezinnen met twee auto's beduidend hoger liggen dan bij gezinnen met 1 auto. Deze waarden zijn wel voor beide types gezinnen veel lager dan de maximale woon-werkafstand.

Uit de histogrammen zien we ook dat de frequentie van de geobserveerde woon-werkafstanden in gezinnen met twee auto's weliswaar heel sterk daalt wanneer de woon-werkafstand toeneemt, maar dat deze daling ook voor grote waarden duidelijk minder snel verloopt dan bij gezinnen met 1 auto. De significantie van deze parameter werd bevestigd in het keuzemodel.

We merken ook op dat zowel de woon-werkafstand als het totaal aantal afgelegde kilometers rechts-scheef verdeeld zijn.

Tabel 21: kwartielen woon-werkafstand (gezinnen met 1 wagen)

Aantal auto's per gezin	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	0	0	0	6.708	7.5	160
2	0	0	6.5	12.59	16	160

5.2. MODEL VOOR GEZINNEN MET 1 WAGEN

Hieronder bespreken we het vraagmodel voor gezinnen die 1 wagen bezitten. Ten eerste overlopen we de data die we gebruiken voor de schatting van het model. Vervolgens bespreken we de resultaten van de OLS schattingen die we hebben uitgevoerd. Zoals hierboven besproken, zijn er echter twee mogelijke bronnen van inconsistentie in de schattingen: de endogeniteit van het type auto dat worden gekozen enerzijds, en de endogeniteit van het aantal auto's per gezin anderzijds. We bespreken achtereenvolgens beide problemen, en tonen aan dat ze in dit geval minder belangrijk zijn dan men zou kunnen verwachten.

5.2.1. BESCHRIJVENDE STATISTIEKEN VOOR HET 1 AUTO-MODEL

Hieronder verkennen we grafisch de relatie tussen de jaarlijks afgelegde afstanden en de belangrijkste kenmerken van de gezinnen die 1 auto bezitten.

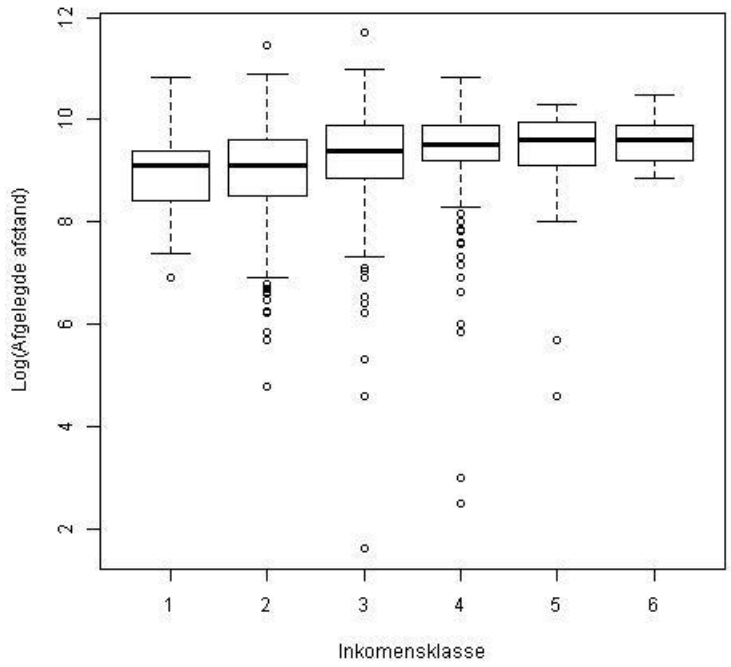
In de eerste plaats verwachten we ons er aan dat het **inkomen** van het gezin een impact zal hebben op het aantal afgelegde kilometers. Het inkomen van het gezin bepaalt immers niet alleen de financiële middelen die beschikbaar zijn om zich te verplaatsen, maar ook de mate waarin het gezin activiteiten kan betalen waar het zich naartoe zal verplaatsen.

In het OVG wordt het totaal netto-gezinsinkomen gerapporteerd aan de hand van zes inkomensklassen (EUR per maand):

Tabel 22: effects coding voor de inkomensklassen

Inkomensklasse	Grenzen
1	0-1000
2	1001-2000
3	2001-3000
4	3001-4000
5	4001-5000
6	5000+

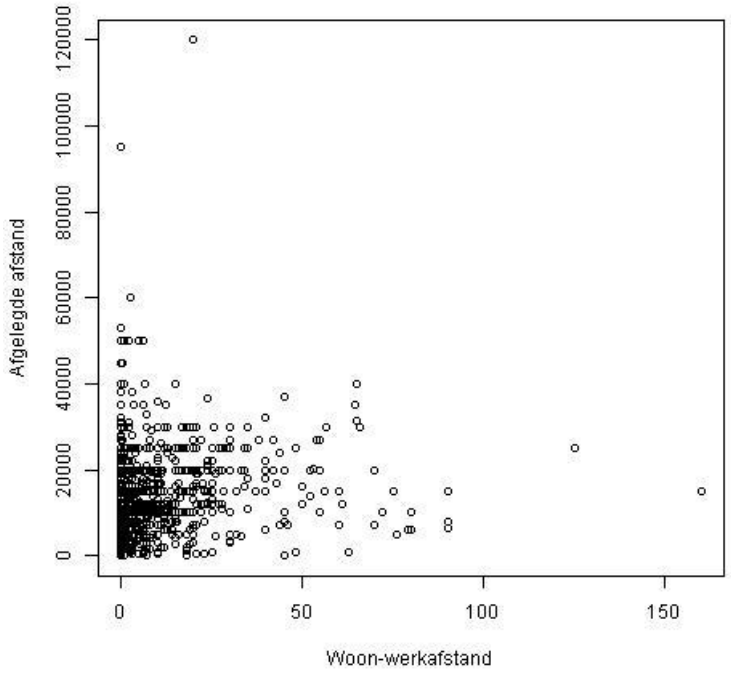
Figuur 10 stelt een doosdiagram voor van de logaritme van de afgelegde afstand op jaarbasis als functie van dit gezinsinkomen. Voor elke inkomensklasse stelt de "doos" het interval voor tussen het eerste en het derde kwartiel van het logaritme van de afgelegde afstand, terwijl de dikke horizontale lijn in de doos de mediaan van de waarnemingen voorstelt. De horizontale lijnen onder en boven de "doos" komen overeen met (respectievelijk) de kleinste en de grootste waarde die binnen een afstand van 1,5 maal de grootte van de doos verwijderd zijn van (respectievelijk) het eerste en het derde kwartiel.



Figuur 10: doosdiagram inkomen-afgelegde afstand

Op basis van het doosdiagram zien we dat een hogere inkomensklasse meestal gepaard gaat met een hogere mediane jaarlijkse afstand. Tegelijkertijd zien we ook dat de spreiding rond deze mediaan sterk varieert in functie van de inkomensklasse, en vooral hoog is voor gezinsinkomens tussen de 1000 en de 4000 EURO. Het is ook voor deze klassen dat we de meeste outliers waarnemen. De statistische analyse zal bevestigen dat het inkomen uiteindelijk slechts een zeer beperkte voorspellende waarde heeft.

Een tweede mogelijke determinant van het totaal aantal kilometer dat een gezin per jaar aflegt, is het **minimum aantal kilometer dat voortvloeit uit "verplichte" activiteiten**. We hebben er hierboven reeds op gewezen dat een aantal verplaatsingen rechtstreeks voortvloeien uit "verplichtingen" van het gezin, zoals verplaatsingen naar het werk of naar naaste familieleden. We beschikken echter alleen over een indicator met betrekking tot de afstand tussen de woonplaats en de werkplaats, "vastkm".



Figuur 11: afgelegde km versus woon-werklafstand

Figuur 11 stelt het aantal afgelegde kilometers op jaarbasis en de woon-werkafstand voor. Uit deze grafiek valt geen duidelijke tendens af te leiden. Dit is echter voor een groot deel te wijten aan de enorm grote concentratie van *vastkm* rond relatief lage waarden.

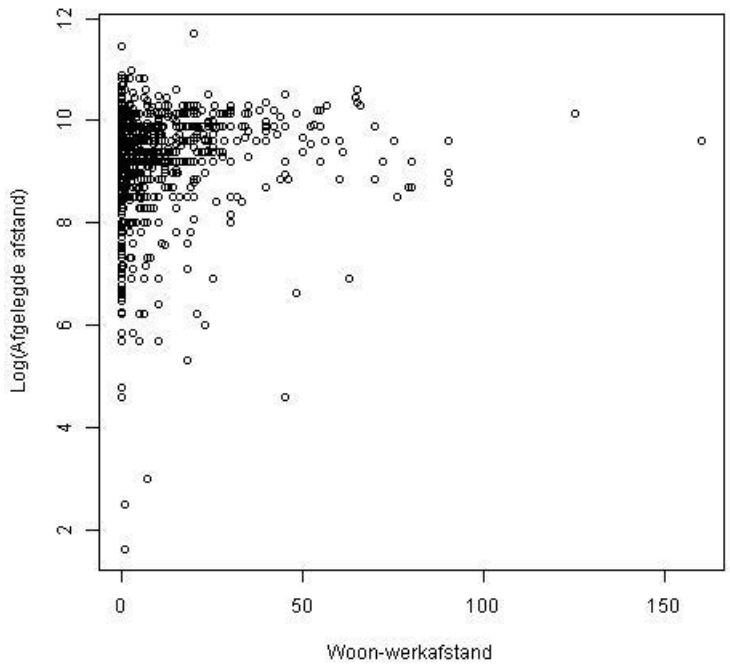
Dit hoeft ons niet te verbazen als we kijken naar de kerngetallen van de spreiding van *vastkm* (zie Tabel 23): 75 % van de gezinnen bevinden zich op minder dan 7,5 km van hun plaats van tewerkstelling. Daarnaast stellen we echter vast dat een minderheid van de gezinnen[48] zich ver tot zeer ver moet verplaatsen naar de werkplaats.

[48] Hier moeten we wel opmerken dat het hier enkel gaat over de gezinnen met 1 auto. In het keuzemodel is aangetoond dat de beslissing om geen tweede auto te kopen mee beïnvloed wordt door de woon-werkafstand.

Tabel 23: kerngetallen van de verdeling van vastkm

Min.	1st Qu.	Median	Mean	3rd Qu.	Max
0.000	0.000	0.000	6.708	7.500	160.000

Als we het aantal afgelegde kilometers op een logaritmische schaal voorstellen, kunnen we slechts een licht stijgende tendens ontwaren (zie Figuur 12). Bij een univariate lineaire regressie bleek de coëfficiënt van *vastkm* echter wel degelijk significant, en kon bovendien de nulhypothese van homoscedasticiteit niet verworpen worden[49] – we hebben daarom deze variabele weerhouden voor de multivariate regressie.



Figuur 12: log afgelegde km versus woon-werklafstand

Een derde mogelijke variabele is de **leeftijd van het gezinshoofd** - hier uitgedrukt als het verschil tussen het observatiejaar van de studie en het geboortejaar. We kunnen ons er immers aan verwachten dat zowel het aantal verplaatsingen als de lengte van de verplaatsingen samenhangt met de periode in het leven waarin het gezinshoofd zich bevindt.

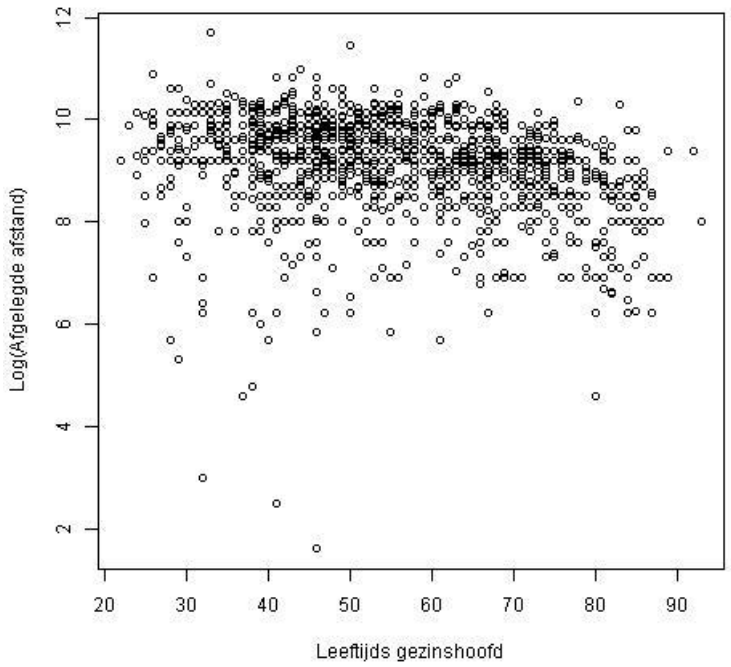
Op Figuur 13 zien we eerst een lichte tendens tot toename van het aantal afgelegde kilometer, tot wanneer het gezinshoofd tussen de 30 en de 50 jaar is. Vervolgens daalt de afgelegde afstand weer.

Bij een lineaire regressie van het logaritme van de afstand als een kwadratische functie van het logaritme van de leeftijd van het gezinshoofd bleek de coëfficiënt van zowel de lineaire term als de

[49] Homoscedasticiteit betekent dat de variantie van de foutenterm niet afhangt van de waarde van de onafhankelijke variabele, in dit geval de woon-werkafstand. Gedetailleerde resultaten beschikbaar op een eenvoudig verzoek.

kwadratische term significant. De nulhypothese van homoscedasticiteit werd echter wel verworpen op een significantieniveau van 2,5%.[50] We hebben deze variabele weerhouden voor de multivariate regressie.

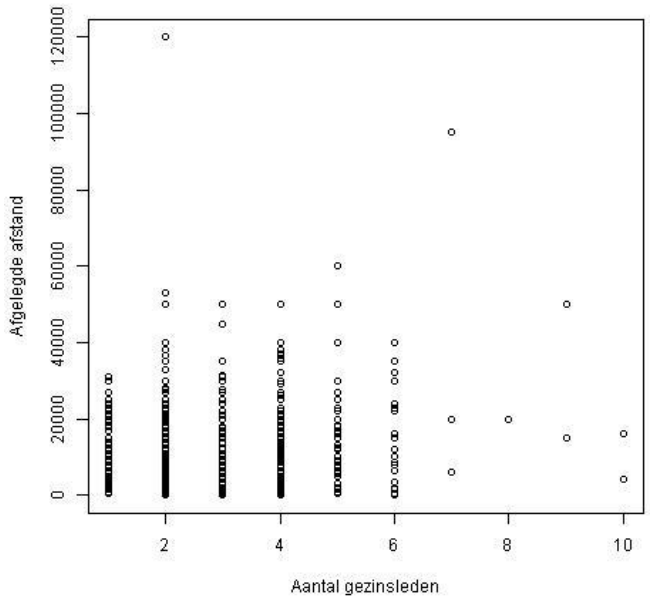
Aangezien deze variabele ook een impact heeft op het aantal auto's dat een gezin kiest, zal hij waarschijnlijk een bron zijn van zelf-selectiebias.



Figuur 13: log afgelegde km versus log leeftijd gezinshoofd

Ook van het **aantal leden van het gezin** verwachten we dat deze een impact zullen hebben op het aantal afgelegde kilometers. Uit het autokeuze-model is al gebleken dat deze een impact heeft op het aantal auto's dat een gezin bezit. Uit Figuur 14 kunnen we echter geen duidelijke invloed afleiden. Uit een verkennende multivariate analyse is gebleken dat deze variabele inderdaad geen statistisch significante impact had, en deze is niet weerhouden voor verdere analyse.

[50] Gedetailleerde resultaten beschikbaar op een eenvoudig verzoek.



Figuur 14: afgelegde km versus aantal gezinsleden

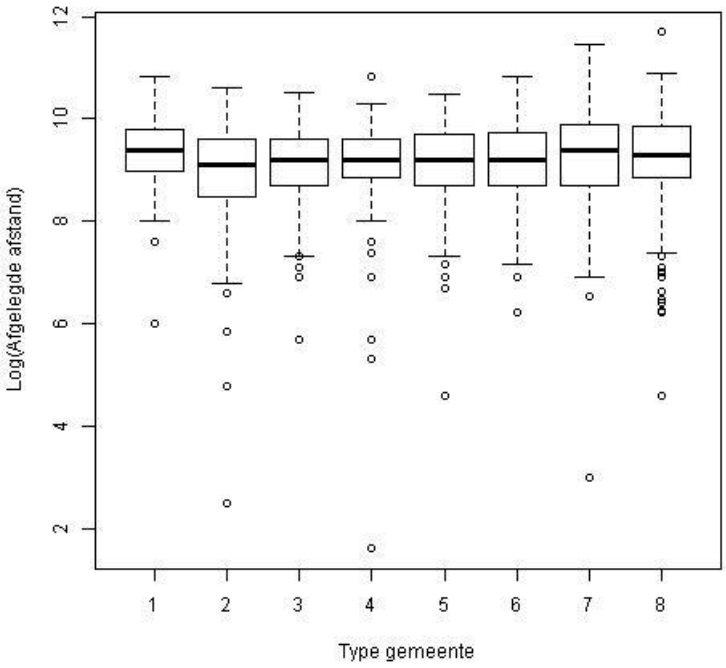
Het OVG bevat ook informatie met betrekking tot het **type gemeente** waarin het gezin woont. Er wordt het onderscheid gemaakt tussen volgende categorieën:

Tabel 24: effects coding voor het type woonplaats

Type gemeente	Grenzen
1	Vlaams stedelijk gebied rond Brussel
2	grootstedelijk gebied centrumgemeenten
3	regionaalstedelijk gebied centrumgemeenten
4	grootstedelijk gebied randgemeenten
5	regionaalstedelijk gebied randgemeenten
6	structuurondersteunend kleinstedelijk gebied
7	kleinstedelijk gebied op provinciaal niveau
8	buitengebied

Er bestaat een uitgebreide literatuur waarin wordt betoogd dat mensen in dichtbebouwde stedelijke gebieden zich minder verplaatsen, en zich vooral minder met de auto verplaatsen, dan mensen in gebieden met een lagere bevolkingsdichtheid (voor een recente samenvatting van de literatuur, zie Boussauw 2011). Het keuzemodel heeft aangetoond dat deze variabele inderdaad een invloed kan uitoefenen op het aantal auto's dat een gezin bezit.

In Figuur 15 valt geen duidelijk patroon te herkennen, hoewel er indicaties zijn dat de jaarlijkse verplaatsingen in grootstedelijke gebieden inderdaad kleiner zijn. Deze variabele wordt dus meegenomen in de multivariate regressie.



Figuur 15: afgelegde km versus type woonplaats

Een laatst aandachtspunt heeft betrekking op het gebruik van **mogelijke alternatieve vervoersmiddelen**. Het keuzemodel heeft al aangetoond dat het frequent gebruik van trein, bus of fiets een impact kan hebben op het aantal auto's dat een gezin bezit.

Voor alle modi wordt het onderscheid gemaakt tussen 5 categorieën van gebruiksfrequentie

Tabel 25: effects coding voor de gebruiksfrekwentie van alternatieve modi

Gebruiksklasse	Grenzen
1	nooit of <1x per jaar
2	1 tot enkele keren per jaar
3	1 tot enkele keren per maand
4	1 tot enkele keren per week
5	dagelijks

Uit Figuur 19 tot en met Figuur 23 (in bijlage) blijkt dat er geen duidelijk patroon te ontwaren valt. Het gebruik van alternatieve modi heeft dus blijkbaar vooral een impact op het aantal auto's per gezin, en niet op het effectief gebruik van de auto indien het gezin slechts 1 auto bezit. Na een verkennende statistische analyse is alleen het gebruik van de motor weerhouden voor de multivariate regressie.

We hebben ook een variante uitgeprobeerd met een "gecombineerde" indicator voor het gebruik van het openbaar vervoer, maar ook deze bleek niet statistisch significant.

Daarnaast zijn er een aantal **variabelen die we niet hebben opgenomen** omdat er te veel observaties ontbraken voor deze variabelen. Hierdoor kromp de omvang van de steekproef in die mate, dat de voorspellende kracht van het model afnam. Het gaat hier over de variabelen die betrekking hebben op: deelname aan een carpoolsysteem, het werktijdenregime van de respondent, het gebruik van de auto voor professionele verplaatsingen, en het geslacht van het gezinshoofd.

Tenslotte geven we de **correlatiematrix** weer voor de continue variabelen (enkel indien de correlatie groter is dan 0,5). Op basis van Tabel 26 kunnen we volgende besluiten trekken:

- Er is slechts een zeer beperkte collineariteit tussen de belangrijkste continue verklarende variabelen.
- De cilinderinhoud, het vermogen en het gewicht zouden redelijke instrumenten kunnen zijn voor het brandstofverbruik en het volume, in de mate dat zij zelf onafhankelijk zijn van de jaarlijkse afstanden die een gezin verwacht af te leggen.

Tabel 26: correlatiematrix voor het 1-auto afstandsmodel

	totpr11	trafftax	vastkm	ghfdgb	fuel_con	ledena	vol	cyl	kw	gewleeg
totpr11	1.00	.	.	.	.	.	.	.	0.58	0.51
trafftax	.	1.00	.	.	.	.	.	0.91	0.72	0.64
vastkm	.	.	1.00	.	.	.	.	.	.	.
ghfdgb	.	.	.	1.00	.	.	.	.	.	.
fuel_con	.	.	.	.	1.00	.	.	.	0.56	.
ledena	.	.	.	.	.	1.00	.	.	.	.
vol	.	.	.	.	.	.	1.00	0.59	.	0.86
cyl	.	0.91	.	.	.	.	0.59	1.00	0.81	0.78
kw	0.58	0.72	.	.	0.56	.	.	0.81	1.00	0.65
gewleeg	0.51	0.64	.	.	.	.	0.86	0.78	0.65	1.00

5.2.2. RESULTATEN VAN DE OLS SCHATTINGEN

Als eerste model hebben we een OLS model geschat met de jaarlijks afgelegde afstand als de afhankelijke variabele.

Tabel 27: verklaring van de gebruikte variabelen

Naam van de variabele	Betekenis
(Intercept)	Constante term van de regressie
vastkm	Woon-werkafstand
I(2011- ghfdgb)	De leeftijd van het gezinshoofd in 2011
totink*i*	De effects code van inkomensklasse i (zoals gedefinieerd in Tabel 22)
gemthuistypei	De effects code van het type woonplaats i (zoals gedefinieerd in Tabel 24))
vol	Het volume van de auto
gmotori	De effects code voor de frequentie van het gebruik van de motor i (zoals gedefinieerd in Tabel 25))
fuel_con	Het brandstofverbruik in EUR per 100 km

We hebben daarbij meerdere modelspecificaties uitgetest. We geven in Tabel 28 het weerhouden OLS model weer. In dit model wordt de afgelegde afstand geschat als een lineaire functie van de onafhankelijke variabelen. Daarnaast hebben we ook een log-lineair model geschat. De resultaten van het log-lineair model zijn grotendeels vergelijkbaar met deze van het lineair model[51], maar de interpretatie van de coëfficiënten van de "effects codes" is minder intuïtief duidelijk.

Dit model is geschat aan de hand van 1345 waarnemingen.

Tabel 28: OLS schatting voor het 1-auto afstandsmodel

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2081.9916	3420.7781	0.609	0.54287	
vastkm	109.6693	34.3054	3.197	0.00142	**
I((vastkm)^2)	-0.8082	0.4113	-1.965	0.0496	*
I(2011- ghfdgb)	189.0042	106.1952	1.78	0.07534	.
I((2011-ghfdgb)^2)	-2.8828	0.9388	-3.071	0.00218	**
totink1	-810.7889	1060.4201	-0.765	0.44465	
totink2	-1423.5554	579.0789	-2.458	0.01409	*
totink3	215.8619	584.285	0.369	0.71185	
totink4	-510.1942	682.0434	-0.748	0.45457	
totink5	105.478	1208.7951	0.087	0.93048	
gemthuistype1	1801.972	1129.9514	1.595	0.11101	
gemthuistype2	-2152.3556	736.0103	-2.924	0.00351	**
gemthuistype3	-1002.4367	615.4596	-1.629	0.1036	
gemthuistype4	-1224.3277	957.923	-1.278	0.20144	

[51] En beschikbaar op eenvoudige aanvraag.

gemthuistype5	-466.9044	887.3905	-0.526	0.59887	
gemthuistype6	-7.4463	727.1263	-0.01	0.99183	
gemthuistype7	1784.5079	627.1088	2.846	0.0045	**
vol	882.1992	141.2945	6.244	5.75E-10	***
gmotor1	2601.7336	1181.1399	2.203	0.02779	*
gmotor2	-3417.6329	2389.566	-1.43	0.15289	
gmotor3	0.6413	1876.5755	0	0.99973	
gmotor4	-5195.043	2759.4302	-1.883	0.05997	.
fuel_con	-250.6279	94.0797	-2.664	0.00782	**
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	
Residual standard error:	8486	on 1322 degrees of freedom			
Multiple R-squared:	0.1788				
Adjusted R-squared:	0.16				
F-statistic:	13.08	on 22 and 1322 DF,			
p-value:	< 2.2e-16				

Bij de schattingsresultaten valt op dat de variantie van de verklarende variabelen slechts 17,88% van de variantie van de afhankelijke variabele verklaart. Aangezien onze databank een uitgebreid gamma van technische en socio-economische gegevens bevat, kunnen we concluderen dat slechts een klein deel van de jaarlijks afgelegde afstanden per auto kan verklaard worden aan de hand van de beschikbare gegevens. We zullen hier later op terugkomen.

Op basis van de Breusch-Pagan test kunnen we de nulhypothese van homoscedasticiteit niet verwerpen, zodat er geen reden bestaat om over te gaan tot een Weighted Least Squares schatting.

Tabel 29: Breusch-Pagan test voor de OLS schatting voor het 1-auto afstandsmodel

BP = 20.9048
df = 22
p-value = 0.5266

Laten we nu de coëfficiënten van de verschillende weerhouden variabelen dichterbij bekijken.

Ten eerste stellen we vast dat de **woon-werkafstand** (*vastkm*) inderdaad de jaarlijks afgelegde afstanden beïnvloedt. Zowel de lineaire term als de kwadratische term zijn statistisch significant. De kleinere significantie van de kwadratische term kan gemakkelijker begrepen worden als we teruggrijpen naar Tabel 23: de grote meerderheid van de waarnemingen voor *vastkm* liggen onder de 10 km, zodat elke schatting voor grotere afstanden onvermijdelijk onderworpen is aan een zeer grote mate van onzekerheid.

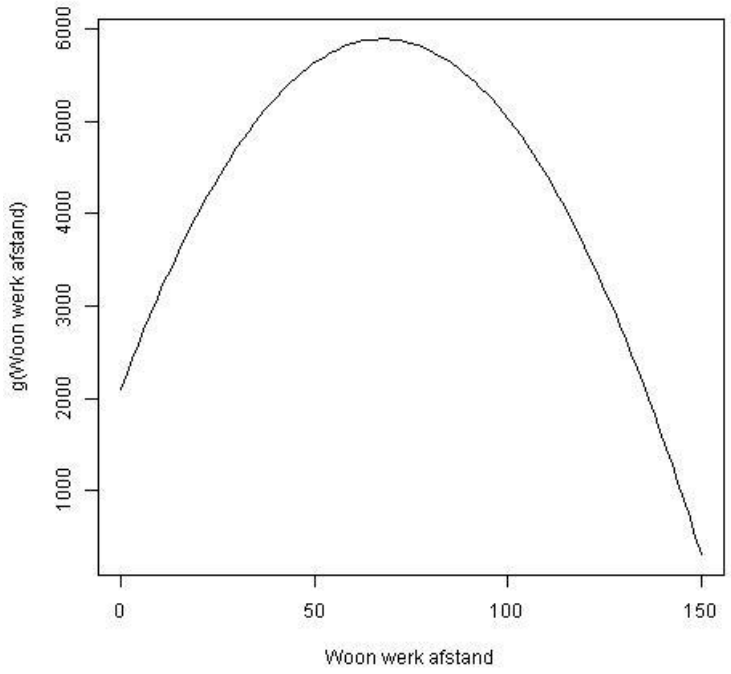
Nochtans denken we dat het interessant kan zijn om deze kwadratische term te weerhouden in de modelspecificatie. Laten we bijvoorbeeld kijken naar de evolutie van de jaarlijks afgelegde

afstanden, enkel in functie van de woonwerkafstand. In Figuur 16 stellen we de geschatte functie voor:

$$y = g(x) = -0.81 * x^2 + 109.67 * \mathrm{x} + 2082$$

Deze functie wordt gemaximaliseerd als x = 68 km.

We zien dus dat de afgelegde afstanden per auto eerst stijgen als functie van de woonwerkafstand, tot op het punt dat deze een drempel overschrijden. Deze drempel komt waarschijnlijk overeen met het punt waar de woonwerkafstand zo groot is geworden, dat het gemiddeld beter is om over te stappen op de trein of om een tweede wagen aan te schaffen[52]. De endogeniteit van het aantal wagens zullen we verder testen, maar de endogeniteit van de modale keuze overschrijdt de scope van deze studie[53].



Figuur 16: afgelegde afstand versus woon-werkafstand voor gezinnen met 1 auto

Ten tweede zien we dat de jaarlijks afgelegde afstand inderdaad een **kwadratische functie is van de leeftijd van het gezinshoofd**.

[52] Vergeten we niet dat we in het keuzemodel hebben aangetoond dat de woonwerkafstand inderdaad een impact heeft op het aantal auto's in het bezit van het gezin.

[53] Een mogelijke manier om dit testen zou er in bestaat om de waarden voor de variabele *gtrein* voor lage waarden van *vastkm* te vergelijken met deze voor hoger waarden van *vastkm.* Het fundamenteel probleem is echter dat, in werkelijkheid, de keuze van het aantal auto's en de frequentie van het gebruik van het openbaar vervoer samen worden bepaald in functie van de totale afstanden die een gezin op jaarbasis verwacht af te leggen. Om dit volledig correct te schatten zouden we deze problemen simultaan moeten schatten in een multimodaal keuzemodel. Dit is een mogelijke piste voor verder onderzoek.

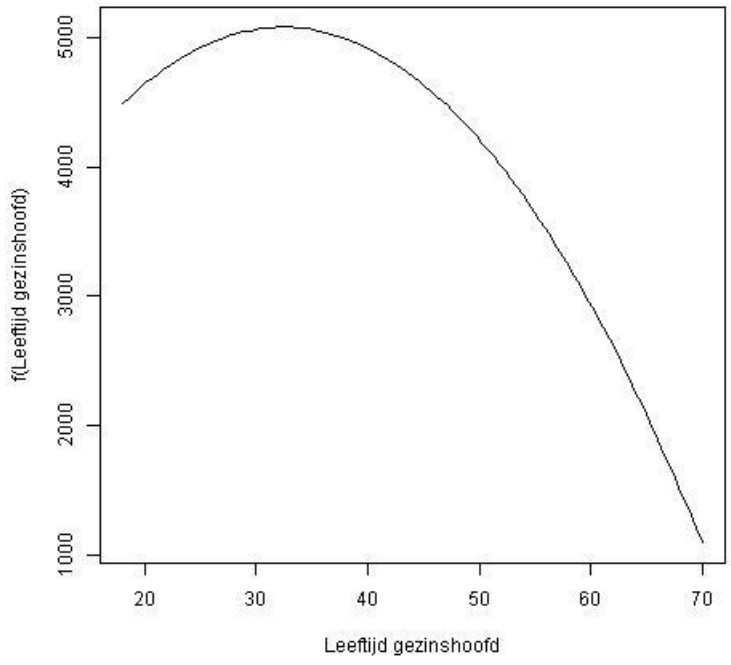
Ook in dit geval is het interessant om te kijken naar de evolutie van de jaarlijks afgelegde afstanden, enkel in functie van de leeftijd van het gezinshoofd. In Figuur 17 stellen we volgende functie voor:

$$y = f(x) = -2{,}88 * x^2 + 189 * \mathrm{x} + 2082$$

Deze functie wordt gemaximaliseerd als x = 33.

Er kan geen onmiddellijke interpretatie gegeven worden aan deze waarde. Het lijkt bijvoorbeeld zeer aannemelijk dat het aantal afgelegde kilometers daalt op het moment dat mensen op pensioen gaan – maar dan zou men verwachten dat de dalende tendens zich pas zou inzetten op een latere leeftijd. Een andere determinant van de afgelegde afstanden zijn waarschijnlijk kinderen die nog niet de leeftijd hebben bereikt waar ze over hun eigen auto beschikken – maar ook voor deze verklaring zou men de piek later verwachten.

Een meer overtuigende verklaring zou zijn dat, als de afgelegde afstand op het niveau van het gezin een bepaalde drempel overschrijdt, het gezin overgaat tot de aankoop van een tweede wagen. De daling van het aantal afgelegde kilometers vanaf 32 jaar zou dan vooral de weergave zijn van een afname van het aantal gezinnen die vanaf die leeftijd over slechts een auto beschikken. De gezinnen die dan overblijven in de steekproef zijn dan gezinnen die sowieso minder kilometers afleggen. In het keuzemodel is inderdaad aangetoond dat de leeftijd van het gezinshoofd een impact heeft op het aantal auto's. De endogeniteit zal verder expliciet worden getest.



Figuur 17: afgelegde afstand versus leeftijd gezinshoofd

De derde grote groep van variabelen hebben betrekking op de **inkomensklasse** van het gezin. Aangezien het OVG alleen inkomenscategorieën rapporteert, hebben we het inkomen van het gezin gemodelleerd aan de hand van “effects coding”.

We hebben de inkomensklasse > 5000 EUR als referentieniveau gedefinieerd waarvoor elke “effects code” de waarde -1 aanneemt.

We stellen vast dat enkel voor “totink2” de coëfficiënt statistisch significant is. Dit betekent dat een gezin met een inkomen tussen de 1000 en de 2000 EUR per maand gemiddeld 1424 km minder aflegt dan het “gemiddeld” gezin. De verwachte waarde van de coëfficiënt van het referentieniveau wordt dan berekend als: 812,8762 + 1441,9994 – 189,2737 + 526,1802 + 132,4785 = 2423,199.

De standaardafwijking van deze coëfficiënt kan dan berekend worden als de vierkantswortel van de som van de kwadraten van de standaardfouten op deze coëfficiënten:

$$\sqrt{579^2 + 584^2 + 682^2 + 1209^2 + 1130^2} = 1930$$

Het 99% betrouwbaarheidsinterval van deze coëfficiënt is dan:

(2423,199 – 3 * 1930; 2423,199 + 3 * 1930) = (--3369; 8215).

De coëfficiënt van het referentieniveau is dus niet significant verschillend van nul.

De vierde categorie variabelen hebben betrekking op de kenmerken van de **woonplaats van het gezin**. Aangezien deze kenmerken worden samengevat aan de hand van een categorische variabele, hebben we ook hier gebruik gemaakt van een “effects coding”, met het buitengebied als referentieniveau.

Enkel voor type 2 en type 7 bleken de coëfficiënten statistisch significant: zoals verwacht is de invloed op de afgelegde afstanden van het wonen in centrumgemeenten van grootstedelijke gebieden negatief, en de invloed van het wonen in kleinstedelijke gebieden positief.

We kunnen nu dezelfde redenering toepassen als bij de berekening van de parameters van het referentieniveau voor het inkomen. We zien dan dat de verwachte waarde van de coëfficiënt van het referentieniveau gelijk is aan: 1267, en de standaardfout 2196. Dus ook hier zien we dat de invloed van het referentieniveau niet statistisch significant is.

Ten vijfde is de coëfficiënt van het **volume van wagen** sterk significant en positief: gezinnen met een grotere auto rijden meer. Zoals hierboven reeds opgemerkt is deze variabele echter waarschijnlijk endogeen.

Ten zesde is ook het **gebruik van de motor** een categorische variabele, waar we “dagelijks” gebruik als referentieniveau genomen hebben.

Hier blijken alleen de twee meest extreme vormen van motorgebruik een invloed te hebben op het aantal afgelegde kilometers met de auto: mensen die nooit de motor gebruiken, rijden significant meer met de auto, en mensen die de motor wekelijks gebruiken, rijden (licht) significant minder met de auto.

De coëfficiënt van het referentieniveau is 6010.301, en de standaardfout 4273. De invloed van het referentieniveau is dus niet statistisch significant.

Ten zevende heeft de **brandstofkost per kilometer** zoals verwacht een significant negatieve invloed op het aantal afgelegde kilometers met de auto. Maar zoals hierboven uitgelegd, is dit een endogene variabele.

We gaan daarom nu over tot de analyse van het model met instrumentele variabelen.

5.2.3. RESULTATEN VAN DE SCHATTING MET INSTRUMENTELE VARIABELEN

Zoals hierboven reeds aangegeven, zijn er minstens twee variabelen (de brandstofkost per kilometer en het volume van de auto) waarvan we kunnen verwachten dat die endogeen zijn in het model: indien men verwacht dat men veel kilometers af zal leggen op jaarbasis, dan zal men (ceteris paribus) een zuinigere en grotere auto kiezen.

We hebben daarom het model opnieuw geschat met instrumentele variabelen. We hebben daarbij volgende instrumenten gebruikt: de prijs van de auto, de jaarlijkse verkeersbelasting, het aantal leden in het gezin, het aantal fietsen in het gezin, het gebruik van alternatieve modi, en het werktijdregime van het gezinshoofd.

Tabel 30: IV schatting voor het 1-auto afstandsmodel

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	180.8785	3858.54	0.047	0.962618	
vastkm	123.7998	35.3458	3.503	0.000476	***
I((vastkm)^2)	-0.985	0.4245	-2.32	0.020481	*
I(2011- ghfdgb)	152.5368	109.3581	1.395	0.163299	
I((2011-ghfdgb)^2)	-2.4466	0.976	-2.507	0.012302	*
totink1	-510.03	1082.454	-0.471	0.63759	
totink2	-1283.04	591.0404	-2.171	0.030123	*
totink3	191.6528	590.6154	0.324	0.745613	
totink4	-718.002	696.9121	-1.03	0.303075	
totink5	-185.227	1229.838	-0.151	0.880306	
gemthuistype1	1911.566	1143.43	1.672	0.094804	.
gemthuistype2	-2057.51	745.6241	-2.759	0.00587	**
gemthuistype3	-857.799	626.204	-1.37	0.17097	
gemthuistype4	-1362.58	970.3092	-1.404	0.160471	
gemthuistype5	-574.638	898.5192	-0.64	0.522583	
gemthuistype6	-39.9051	734.9135	-0.054	0.956705	
gemthuistype7	1671.663	636.1194	2.628	0.008691	**
vol	1483.259	326.9079	4.537	6.22E-06	***
gmotor1	2649.057	1196.652	2.214	0.027018	*
gmotor2	-3396.63	2415.13	-1.406	0.159841	

gmotor3	125.7557	1897.211	0.066	0.947161	
gmotor4	-5891.47	2811.377	-2.096	0.036309	*
fuel_con	-622.787	241.6134	-2.578	0.010056	*
Diagnostic	tests:				
	df1	df2	statistic	p-value	
Weak instruments	29	1295	18.309	<2e-16	***
Wu-Hausman	2	1320	2.148	0.117	
Sargan	27	NA	24.433		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	
Residual standard error:	8574	on 1322 degrees of freedom			
Multiple R-squared:	0.1615				
Adjusted R-squared:	0.1475				
Wald test	11.95	on 22 and 1322 DF,			
p-value:	< 2.2e-16				

Bekijken we eerst even de teststatistieken met betrekking tot de keuze van de schattingsmethode.

De eerste test-statistiek "weak instruments" heeft betrekking op de correlatie tussen de instrumenten en de onafhankelijke variabelen: indien deze te klein is, dan zal een schatting met Instrumentele Variabelen zeer onnauwkeurig zijn. Om dit te testen wordt een statistiek berekend (zie Green (2012) p 250 voor de details) die een F verdeling volgt. Indien de waarde van deze statistiek lager dan 10 is, dan worden de instrumenten als "zwak" beschouwd. Dat is hier dus niet het geval.

De tweede test-statistiek "Wu-Hausman" test of de IV schatting een verbetering inhoudt ten opzichte van de OLS schatting. De beschouwde nul hypothese is dat zowel de OLS als de IV schatters consistent zijn. Als dat het geval is, dan kunnen we beter de OLS schattingen gebruiken, aangezien de covariantiematrix van de OLS schatters asymptotisch kleiner is.

Onder de nul-hypothese volgt de Wu-Hausman statistiek een χ^2 verdeling met het aantal endogene variabelen als aantal vrijheidsgraden (zie Greene (2012) p 237 voor meer details). In dit geval kunnen we de nulhypothese niet verwerpen.

De derde test-statistiek "Sargan" test of de instrumenten gecorreleerd zijn met de residuen. De waarde van de Sargan test moet vergeleken worden met een chi-kwadraat met significantieniveau 95% waarbij het aantal vrijheidsgraden gelijk is aan het verschil tussen het aantal instrumenten en het aantal endogene variabelen. Als de waarde kleiner is dan de kritische waarde, dan kunnen we nulhypothese niet verwerpen dat de instrumenten niet gecorreleerd met de residuen. Als de waarde groter is, dan verwerpen we de nulhypothese wel: sommige van de instrumenten (maar men kan niet zeggen welke) zijn dan gecorreleerd met de residuen en zijn dus niet geldig. In dit geval wordt de nulhypothese niet verworpen.

We kunnen dus besluiten dat onze instrumenten "goed" gekozen zijn: ze zijn niet gecorreleerd met de foutenterm, en ze zijn wel gecorreleerd met de gebruikte regressoren**. Aangezien we echter de**

nulhypothese niet kunnen verwerpen dat ook de OLS schatters consistent zijn, verkiezen we het OLS model, aangezien dit model nauwkeuriger is.

5.2.4. RESULTATEN VAN DE SCHATTING MET CORRECTIE VOOR ZELF-SELECTIE

Tenslotte hebben we het 1 auto model ook nog geschat met correctie voor zelf-selectie. De verklarende variabelen in dit model zijn dezelfde als in het OLS model, maar we hebben twee termen toegevoegd: *corr_cars_nu* is de DubinMcFadden correctieterm voor het "twee auto" alternatief, en *corr_cars_0* is de DubinMcFadden correctieterm voor het "0 auto" alternatief.

Tabel 31: schatting voor het 1-auto afstandsmodel met Dubin-McFadden correctie

	Estimate	Std.Error	t	value	Pr(>|t|)
(Intercept)	4595.6871	3768.8445	1.219	0.22291	
vastkm	102.1987	35.9988	2.839	0.0046	**
I((vastkm)^2)	-0.7724	0.4218	-1.831	0.06727	.
I(2011- ghfdgb)	106.4316	116.0607	0.917	0.35929	
I((2011- ghfdgb)^2)	-2.1262	1.0336	-2.057	0.03988	*
totink1	-16.5944	1233.825	-0.013	0.98927	
totink2	-941.172	845.5738	-1.113	0.26589	
totink3	-37.0803	656.4225	-0.056	0.95496	
totink4	-870.329	716.2755	-1.215	0.22455	
totink5	-196.6466	1357.098	-0.145	0.88481	
gemthuistype1	1987.8361	1135.0275	1.751	0.08012	.
gemthuistype2	-1958.189	744.4526	-2.630	0.00863	**
gemthuistype3	-913.4155	618.3855	-1.477	0.13989	
gemthuistype4	-1223.7164	957.9702	-1.277	0.20168	
gemthuistype5	-564.4079	889.3543	-0.635	0.52578	
gemthuistype6	-55.7532	727.595	-0.077	0.93893	
gemthuistype7	1628.9089	636.2034	2.560	0.01057	*
vol	852.1694	142.3608	5.986	2.77E-09	***
gmotor1	2607.1482	1181.1999	2.207	0.02747	*
gmotor2	-3228.9473	2390.7846	-1.351	0.17706	
gmotor3	-129.8802	1880.1954	-0.069	0.94494	
gmotor4	-5182.8357	2758.307	-1.879	0.06047	.
fuel_con	-249.9021	94.0391	-2.657	0.00797	**
corr_cars_nu	-1409.2041	1244.5724	-1.132	0.25772	
corr_cars_0	1901.9068	1060.5563	1.793	0.07315	.
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	
Residual standard error:	8482	on 1320 degrees of freedom			

Multiple R-squared:	0.1808	
Adjusted R-squared:	0.1659	
F-statistic:	12.14	on 24 and 1320 DF
p-value:	< 2.2e-16	

We zien dat *corr_cars_nu* niet significant verschillend van nul is, en dat *corr_cars_0* het slechts is met een 10% drempel.

We hebben daarom ook de totale significantie van de correctietermen geschat. We zien hieronder dat de nulhypothese dat *corr_cars_nu = corr_cars_0 = 0* niet kan verworpen worden[54].

Tabel 32: significantietest voor de DubinMcFadden correctietermen

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1322	9.52E+10				
2	1320	9.50E+10	2	233840164	1.6253	0.1972

Hoewel de correctietermen niet significant verschillend van nul zijn, stellen we toch vast dat de toevoeging van deze termen de significantie van meerdere variabelen (leeftijd, inkomen, de kwadratische term voor de woon-werkafstand) in belangrijke mate verlaagd.

In elk geval kunnen we de **nulhypothese niet verwerpen dat de correlatie tussen de foutentermen van het afstandsmodel en de foutenterm van het keuzemodel nul is**. Bijgevolg weerhouden we het OLS model zonder correctieterm.

5.3. MODEL VOOR GEZINNEN MET 2 WAGENS

5.3.1. BESCHRIJVENDE STATISTIEKEN

Hoewel de kenmerken van de gezinnen met twee wagens verschillen van die van gezinnen met 1 wagen, is de analyse van de beschrijvende statistieken grotendeels gelijklopend. De belangrijkste verschillen hebben betrekking op de verdeling van de afhankelijke variabelen, en van een beperkt aantal onafhankelijke variabelen zoals de woon-werkafstand en de leeftijd van het gezinshoofd. Deze verschillen zijn hierboven reeds besproken. We zullen daarom de andere resultaten hier niet reproduceren - ze zijn op eenvoudige aanvraag beschikbaar.

5.3.2. RESULTATEN VAN DE OLS SCHATTINGEN

In wat volgt, schatten we de afgelegde afstanden voor *elke* wagen van het gezin. We hebben daarbij dezelfde onafhankelijke variabelen gebruikt als bij het 1 auto-gezin. Daarnaast hebben we er ook rekening mee gehouden dat gezinnen soms twee auto's bezitten met twee verschillende gebruiksprofielen. Zo kan de grootste auto gebruikt worden voor het gezinslid die de grootste

[54] Onder de nulhypothese volgt de F waarde een F verdeling waarvan het aantal vrijheidsgraden gelijk is aan het aantal geteste hypothesen (in dit geval, 2) en het verschil tussen de omvang van de steekproef en het aantal onafhankelijke variabelen (in dit geval, 1345 – 25 = 1320).

afstanden aflegt om professionele redenen, terwijl de kleinere auto vooral wordt gebruikt door het gezinslid dat dichtbij huis werkt. De grotere auto zal normaal gezien ook gebruikt worden voor langere verplaatsingen om privé-redenen. We creëren daarom drie bijkomende dummy variabelen, die aangeven welke auto de oudste is (*older_car*) , welke de goedkoopste (*cheap_car*) en welke de kleinste (*small_car*).

Tabel 33 geeft het weerhouden OLS model weer. Deze is gebaseerd op een steekproef met 1188 waarnemingen (594 gezinnen met twee wagens).

Tabel 33: OLS schatting voor het 2-auto afstandsmodel

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2290.3014	3327.8772	0.688	0.491454	
Vastkm	137.2902	37.826	3.63	0.000296	***
I((vastkm)^2)	-0.7404	0.4079	-1.815	0.069777	.
totink1	-3499.859	2970.3241	-1.178	0.238929	
totink2	-2564.2338	1009.6132	-2.54	0.011221	*
totink3	595.7108	840.5237	0.709	0.47863	
totink4	57.3232	816.9244	0.07	0.944071	
totink5	1973.8203	1040.824	1.896	0.058156	.
gemthuistype1	-978.9672	1564.4178	-0.626	0.531589	
gemthuistype2	-2777.3173	1552.387	-1.789	0.073866	.
gemthuistype3	155.4574	900.9183	0.173	0.863032	
gemthuistype4	2088.6777	1600.6602	1.305	0.192192	
gemthuistype5	859.8023	1188.2824	0.724	0.469478	
gemthuistype6	-372.0287	1249.585	-0.298	0.765969	
gemthuistype7	1273.8538	880.1921	1.447	0.148099	
vol	1082.7289	209.4062	5.17	2.75E-07	***
gmotor1	658.2269	1497.3419	0.44	0.660311	
gmotor2	5984.3335	2487.1906	2.406	0.016282	*
gmotor3	-2683.2124	2234.0017	-1.201	0.229967	
gmotor4	-1650.8558	3077.7554	-0.536	0.591797	
fuel_con	23.0891	122.4619	0.189	0.850486	
older_car	-1472.7303	653.0934	-2.255	0.024319	*
cheap_car	-399.2922	715.928	-0.558	0.577139	
small_car	-2379.8382	831.8411	-2.861	0.0043	**
gtram1	-2094.4403	1186.9325	-1.765	0.077898	.
gtram2	-2557.3713	1279.6279	-1.999	0.045893	*
gtram3	-1643.4318	1551.0377	-1.06	0.289562	
gtram4	-2239.9742	1922.1249	-1.165	0.244112	
gtrein1	2041.7605	756.9718	2.697	0.007093	**
gtrein2	2037.5634	790.3954	2.578	0.010063	*

gtrein3	1843.6011	1360.1308	1.355	0.175536
gtrein4	-1842.0633	1305.4438	-1.411	0.158495
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
Residual standard error:	10880	on 1156 degrees of freedom		
Multiple R-squared:	0.1329			
Adjusted R-squared:	0.1097			
F-statistic:	5.718	on 31 and 1156 DF		
p-value:	< 2.2e-16			

Op basis van de Breusch-Pagan test kunnen we ook hier de nulhypothese van homoscedasticiteit niet verwerpen.

Tabel 34: Breusch-Pagan test voor de OLS schatting voor het 2-auto afstandsmodel

BP = 24.481
df = 22
p-value = 0.3225

De globale fit van dit model ligt lager dan bij in het model voor gezinnen met 1 wagen.

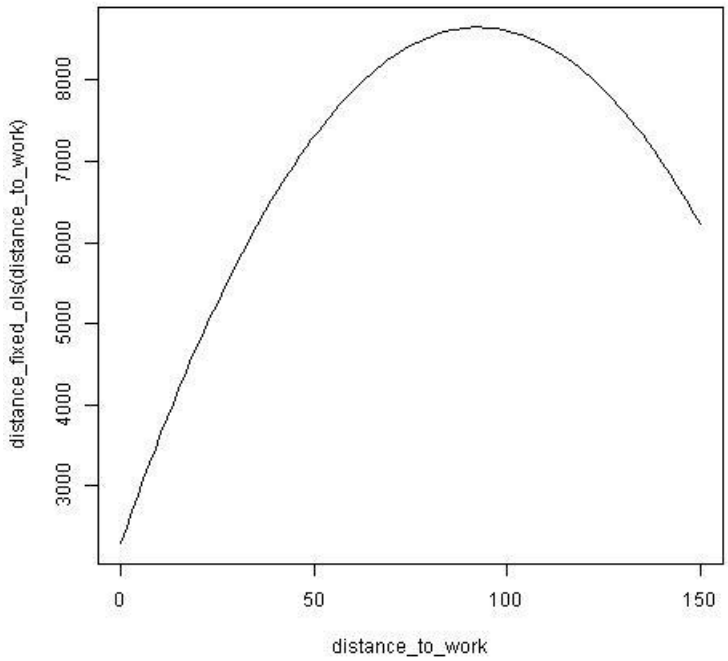
Ten opzichte van het model voor gezinnen met 1 wagen stellen we een aantal belangrijke veranderingen vast.

Ten eerste verdwijnt elke invloed van de **leeftijd van het gezinshoofd**. Zoals hierboven reeds geopperd, is dat misschien omdat gezinnen pas overgaan tot het aankopen van een tweede auto als meerdere gezinsleden de auto nodig hebben om zich naar het werk, naar school of naar vrijetijdactiviteiten te verplaatsen. Bij dergelijke gezinnen zou er weinig marge zijn om het aantal afgelegde kilometers op jaarbasis aan te passen.

Dergelijke motieven zijn echter voor een deel leeftijdsgebonden – onze bespreking van de beschrijvende statistieken duidde aan dat het leeftijdsprofiel van het gezinshoofd in gezinnen met 2 auto's verschilt van het leeftijdsprofiel in gezinnen met 1 auto, en overeenkomt met de leeftijd van mensen die (a) werken (b) een gezin met minderjarige kinderen hebben. Het keuzemodel heeft inderdaad bevestigd dat de leeftijd van het gezinshoofd een impact heeft op de keuze van het aantal auto's.

Ten tweede is de significantie van de kwadratische term voor de **woonwerkafstand** sterk afgenomen, waardoor het maximum nu wordt bereikt voor een woonwerkafstand van 93 km. Een mogelijke verklaring hiervoor is dat de gezinnen die zich beperken tot 1 auto, dat doen omdat ze door hun woonplaats gemakkelijker toegang hebben tot vervoer per trein, en daarvoor reeds vanaf een kortere woonwerkafstand de overstap maken naar de trein (eerder dan een tweede gezinswagen aan te schaffen). Om deze hypothese te kunnen toetsen zouden we echter moeten kunnen beschikken over bijkomende data.

Een andere mogelijke verklaring is zuiver statistisch-technisch van aard: onze analyse hierboven heeft bevestigd dat, in gezinnen met 2 auto's, de woon-werkafstand duidelijk hoger ligt dan bij gezinnen met 1 auto. Voor deze gezinnen is er dus een bredere steekproef aan woon-werkafstanden beschikbaar, waardoor de geschatte coëfficiënt voor de kwadratische term nauwkeuriger is.



Figuur 18: afgelegde afstand versus woon-werkafstand voor gezinnen met 2 auto's

Ten derde is de invloed van het **gezinsinkomen** gelijkaardig gebleven. Al bij al blijft deze invloed relatief zwak. Een mogelijke verklaring hiervoor is dat, zoals aangetoond in het keuzemodel, het gezinsinkomen ook een rol speelt in de keuze van het aantal auto's en in de kenmerken van de gekozen auto's. Indien een gezin verwacht dat ze veel kilometers moet afleggen op jaarbasis, zal ze misschien eerder besparen op bepaalde attributen van de auto's dan op het aantal afgelegde kilometers.

Ten vierde is de invloed van het **type gemeente** nog verder afgenomen. Waarschijnlijk is ook dit omdat, zoals aangetoond in het keuzemodel, het type gemeente ook een invloed heeft op het aantal auto's, waardoor de invloed van het type gemeente zich eerder zal laten voelen op het niveau van het totaal aantal kilometers dat door het gezin wordt afgelegd, eerder dan op het aantal kilometers per auto. Om deze hypothese te kunnen toetsen is bijkomend werk nodig.

Ten vijfde is de invloed van het **volume** zeer significant gebleven, terwijl het **brandstofverbruik** geen significante invloed meer heeft.

Ten zesde blijft de invloed van het **gebruik van een motor** zeer beperkt.

Ten zevende zien we dat er twee significante termen zijn bijgekomen die betrekking hebben op de relatieve positie van de beschouwde auto ten opzichte van de andere auto in bezit van het gezin.

De **oudste** van de twee gezinswagens wordt significant minder gebruikt dan de nieuwe. Hetzelfde gaat op met betrekking tot het **kleiner** model. Zoals hierboven reeds besproken zijn beide resultaten in overeenstemming met wat we zouden verwachten.

We merken ook op dat er twee alternatieve modi zijn die geen significante invloed hadden op de afgelegde afstanden in het 1 auto-model, maar wel in het 2- auto-model: de tram en de trein. Ook hier gaat het over de categorische variabelen *gtram* en *gtrein*, die op dezelfde manier zijn gedefinieerd als *gmotor*.

Zo blijkt dat bij gezinnen met twee wagens het **tramgebruik** een impact heeft op het aantal afgelegde kilometers. We kunnen dezelfde procedure toepassen als bij het 1 auto model om aan te tonen dat ook de coëfficiënt voor het referentieniveau niet significatief verschillend van nul is.

Het teken van de geschatte coëfficiënten valt echter moeilijk te verklaren, want het impliceert dat dagelijkse tramgebruikers meer kilometers afleggen dan andere tramgebruikers (met inbegrip van mensen die zelden of nooit de tram nemen). We moeten hierbij wel de kanttekening plaatsen dat dit contra-intuïtief teken van de coëfficiënt alleen significant is voor mensen die minder dan eens per maand de tram nemen.

Tenslotte blijkt dat bij gezinnen met twee wagens het **treingebruik** ook een impact heeft op het aantal afgelegde kilometers. Ook hier is het teken van het referentieniveau niet significatief. Het teken van de geschatte coëfficiënten ligt wel in de lijn van de verwachtingen, want het impliceert dat frequente treingebruikers minder kilometers afleggen dan andere treingebruikers. Ook hier moeten we opmerken dat het resultaat alleen significatief is voor mensen die minder dan eens per maand de trein nemen.

5.3.3. RESULTATEN VAN DE SCHATTING MET INSTRUMENTELE VARIABELEN

Ook in het twee-auto model zijn er minstens twee variabelen (de brandstofkost per kilometer en het volume van de auto) waarvan we kunnen verwachten dat die endogeen zijn in het model: indien men verwacht dat men veel kilometers af zal leggen op jaarbasis, dan zal men een zuinigere en grotere auto kiezen.

We hebben daarom het model opnieuw geschat met instrumentele variabelen. We hebben daarbij dezelfde instrumenten gebruikt als bij het een-auto model. Daarnaast hebben we ook de leeftijd van het gezinshoofd (dat niet significant is gebleken als regressor) als instrument gebruikt.

Tabel 35: IV schatting voor het 2-auto afstandsmodel

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	87.7391	5510.63	0.016	0.9873	
vastkm	141.7094	38.4628	3.684	0.00024	***
I((vastkm)^2)	-0.7836	0.4158	-1.885	0.05974	.
totink1	-3664.76	3096.076	-1.184	0.23678	
totink2	-2512.02	1018.149	-2.467	0.01376	*
totink3	717.0636	887.9653	0.808	0.41952	

totink4	152.8517	828.7089	0.184	0.8537	
totink5	1962.748	1046.222	1.876	0.0609	.
gemthuistype1	-921.424	1577.319	-0.584	0.55922	
gemthuistype2	-2728.21	1556.048	-1.753	0.07982	.
gemthuistype3	111.5158	903.0765	0.123	0.90175	
gemthuistype4	2044.433	1602.47	1.276	0.20228	
gemthuistype5	882.5155	1189.344	0.742	0.45823	
gemthuistype6	-420.623	1251.483	-0.336	0.73686	
gemthuistype7	1312.605	884.8438	1.483	0.13823	
vol	1121.908	611.9404	1.833	0.06701	.
gmotor1	659.0196	1498.308	0.44	0.66013	
gmotor2	5933.616	2493.284	2.38	0.01748	*
gmotor3	-2702.71	2248.327	-1.202	0.22957	
gmotor4	-1663.69	3166.91	-0.525	0.59945	
fuel_con	158.9388	323.7844	0.491	0.62361	
older_car	-1507.76	666.2377	-2.263	0.02381	*
cheap_car	-291.399	725.1148	-0.402	0.68786	
small_car	-2159.67	1317.866	-1.639	0.10153	
gtram1	-2112.22	1188.452	-1.777	0.07578	.
gtram2	-2566.19	1280.742	-2.004	0.04534	*
gtram3	-1622.49	1555.39	-1.043	0.2971	
gtram4	-2202.59	1927.117	-1.143	0.2533	
gtrein1	2058.722	765.0445	2.691	0.00723	**
gtrein2	2062.681	794.8657	2.595	0.00958	**
gtrein3	1816.456	1378.275	1.318	0.18779	
gtrein4	-1898.33	1345.163	-1.411	0.15845	
Diagnostic	tests:				
	df1	df2	statistic	p-value	
Weak instruments	18	1140	28.944	<2e-16	***
Wu-Hausman	2	1154	0.465	0.628	
Sargan	16	NA	14.182	0.585	
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	
Residual standard error:	10890	on 1156 degrees of freedom			
Multiple R-squared:	0.1319				
Adjusted R-squared:	0.1086				
F-statistic:	5.168	on 31 and 1156 DF			
p-value:	< 2.2e-16				

De teststatistieken met betrekking tot de gebruikte instrumenten geven resultaten die bijna perfect analoog zijn aan de resultaten voor het 1 auto-model.

Ten eerste blijkt uit de test-statistiek “weak instruments” dat de correlatie tussen de instrumenten en de regressoren voldoende hoog is.

Ten tweede, op basis van de test-statistiek “Wu-Hausman” kunnen we de hypothese niet verwerpen dat de OLS schatters ook consistent zijn als de IV schatters het zijn.

Ten derde kunnen we op basis van de “Sargan” test de nulhypothese niet verwerpen dat de instrumenten niet gecorreleerd zijn met de residuen.

We kunnen dus ook hier besluiten dat onze **instrumenten “goed” gekozen zijn**: ze zijn niet gecorreleerd met de foutenterm, en ze zijn wel gecorreleerd met de gebruikte regressoren. **Aangezien we echter niet de nulhypothese kunnen verwerpen dat ook de OLS schatters consistent zijn, verkiezen we het OLS model**, aangezien dit model nauwkeuriger is.

5.3.4. RESULTATEN VAN DE SCHATTING MET CORRECTIE VOOR ZELF-SELECTIE

Tenslotte hebben we het 2 auto model ook nog geschat met correctie voor zelf-selectie. De verklarende variabelen in dit model zijn dezelfde als in het OLS model, maar we hebben twee termen toegevoegd: *corr_cars_nu* is de DubinMcFadden correctieterm voor het “een auto” alternatief, en *corr_cars_0* is de DubinMcFadden correctieterm voor het “0 auto” alternatief.

Tabel 36: schatting voor het 2-auto afstandsmodel met Dubin-McFadden correctie

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6975.6456	4495.717	1.552	0.12103	
vastkm	98.435	45.5981	2.159	0.03107	*
I((vastkm)^2)	-0.4098	0.4601	-0.891	0.37332	
totink1	-2354.4666	3092.961	-0.761	0.44667	
totink2	-1272.8479	1381.298	-0.921	0.35699	
totink3	1149.6901	947.6676	1.213	0.22531	
totink4	-331.2264	896.3826	-0.37	0.71181	
totink5	688.3815	1337.457	0.515	0.60687	
gemthuistype1	-806.4346	1569.087	-0.514	0.60738	
gemthuistype2	-2667.8379	1562.575	-1.707	0.08803	.
gemthuistype3	298.0022	906.0584	0.329	0.74229	
gemthuistype4	2071.3801	1605.3	1.29	0.19719	
gemthuistype5	766.1516	1192.108	0.643	0.52056	
gemthuistype6	-417.4241	1253.242	-0.333	0.73914	
gemthuistype7	1109.9191	892.7719	1.243	0.21404	
vol	1051.8994	210.3915	5	6.63E-07	***
gmotor1	641.479	1497.205	0.428	0.6684	
gmotor2	5917.4667	2489.79	2.377	0.01763	*
gmotor3	-2555.2844	2238.401	-1.142	0.25387	

gmotor4	-1727.7983	3083.262	-0.56	0.57533	
fuel_con	36.0056	122.9756	0.293	0.76974	
older_car	-1479.0506	652.9879	-2.265	0.02369	*
cheap_car	-399.7472	715.8013	-0.558	0.57664	
small_car	-2434.6618	832.446	-2.925	0.00352	**
gtram1	-1960.7652	1196.528	-1.639	0.10155	
gtram2	-2381.7721	1290.386	-1.846	0.06518	.
gtram3	-1679.8462	1563.016	-1.075	0.28271	
gtram4	-2044.5486	1976.709	-1.034	0.3012	
gtrein1	1892.0577	768.9715	2.461	0.01402	*
gtrein2	1962.1992	798.373	2.458	0.01413	*
gtrein3	1921.1299	1362.155	1.41	0.1587	
gtrein4	-1936.7262	1307.818	-1.481	0.13891	
corr_cars_nu	4847.5088	3910.012	1.24	0.21531	
corr_cars_0	-3701.5331	3781.472	-0.979	0.32785	
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	
Residual standard error:	10880	on 1154 degrees of freedom			
Multiple R-squared:	0.1348				
Adjusted R-squared:	0.11				
F-statistic:	5.447	on 33 and 1154 DF			
p-value:	< 2.2e-16				

Ook hier stellen we vast dat de twee correctietermen niet significant verschillend van nul zijn. Wel zien we een verdere verlaging van het significantieniveau voor de woon-werkafstand, het inkomen en het gemeentetype, drie variabelen waarvan we weten dat ze ook het aantal auto's beïnvloeden.

5.3.5. TOTALE AFSTAND PER GEZIN ALS AFHANKELIJKE VARIABELE

Onze doelstelling is om de jaarlijks afgelegde afstand per auto te voorspellen. Men zou echter kunnen opperen dat, voor een gezin, de totaal afgelegde afstand per gezin relevanter is dan de afstand per auto.

We hebben daarom ook een model geschat voor de totaal afgelegde afstand per gezin. Dit model presteert echter slechter dan het model per individuele auto – gedetailleerde resultaten zijn beschikbaar op aanvraag.

HOOFDSTUK 6. CONCLUSIES

In dit rapport hebben we achtereenvolgens volgende vragen bekeken:

- Gegeven het aantal auto's dat een gezin bezit, wat zijn de kenmerken van deze auto's?
- Hoeveel auto's kiest een gezin dat twee of minder auto's bezit?
- Gegeven het aantal auto's dat een gezin bezit, hoeveel kilometers legt het gezin per jaar met elke auto af?

We bespreken hieronder nog even de belangrijkste vaststellingen per vraag, en trekken daarna besluiten van methodologische aard.

6.1.1. BELANGRIJKSTE VASTSTELLINGEN

Ten eerste hebben we de keuze van de auto-kenmerken gemodelleerd op het niveau van "klassen" van auto's, waarbij we rekening houden met de belangrijkste kenmerken (gemiddeldes, maar ook varianties en covarianties) van de individuele modellen die elke klasse uitmaken. De belangrijkste conclusies zijn:

- Voor gezinnen met 1 auto, hebben we vastgesteld dat de keuze van de gezinnen het best worden gemodelleerd aan de hand van een Nested Logit model, waarbij de "nests" worden bepaald door het bouwjaar van de auto's. Het geprefereerd model wordt samengevat in Tabel 14. Hoewel de globale voorspellende waarde van het model beperkt blijft, stellen we vast dat de coëfficiënt van meerdere variabelen (de prijs van de wagen in interactie met het inkomen, het volume van de wagen in interactie met de gezinsgrootte, het volume in interactie met het inkomen, de brandstofkost, het onderliggend aantal modellen, een aantal covariantietermen) significant verschillend van nul is, en door de hand genomen het verwachte teken heeft. De coëfficiënt van de verkeersbelasting blijkt echter niet of (naar gelang het aantal beschouwde klassen) slechts licht significant verschillend van nul.

- Voor gezinnen met 2 auto's, werd de waarschijnlijkheid geschat van bepaalde combinaties van auto's. Het geprefereerd model wordt samengevat in Tabel 15. De resultaten voor dit model zijn grotendeels gelijklopend met de resultaten voor het 1 auto model: de globale voorspellende kracht is beperkt, hoewel meerdere coëfficiënten een hoog significantieniveau hebben. Een belangrijk verschil is wel dat, voor gezinnen met twee auto's, de totale verkeersbelasting wel degelijk een significante invloed uitoefent.

Ten tweede hebben we de keuze van het aantal auto's gemodelleerd voor gezinnen met twee of minder auto's[55]. Het geprefereerd model wordt samengevat in Tabel 17. Voor dit model bekomen we wel een redelijke voorspellende waarde[56]. We stellen vast dat meerdere variabelen een significante invloed uitoefenen op de keuzewaarschijnlijkheid: de verwachte waarde van het maximaal nu dat men kan halen uit het bezit van 1 of 2 wagens, het geslacht van het gezinshoofd,

[55] We hebben ook apart de keuze gemodelleerd van het aantal auto's voor gezinnen met 0 of 1 auto's.
[56] Pseudo-R^2 van 0.26.

het aantal gezinsleden, het gezinsinkomen, het diploma van het gezinshoofd, het type gemeente waar het gezin woont, de frequentie waarmee het gezin gebruik maakt van andere modi, de leeftijd van het gezin en de woon-werk afstand.

Tenslotte hebben we het aantal kilometers geschat dat jaarlijks per auto wordt afgelegd, rekening houdende met het aantal auto's dat het gezin bezit.

We kunnen redelijkerwijze verwachten dat zowel het aantal auto's als bepaalde kenmerken van de auto beïnvloed worden door het aantal kilometers dat een gezin verwacht af te leggen op jaarbasis. Indien we een verschillend model schatten voor gezinnen met 1 auto als voor gezinnen met 2 auto's, dan kan dit soort modellen dus vertekend worden door twee belangrijke bronnen van bias: endogeniteitsbias met betrekking tot de kenmerken van individuele modellen en zelf-selectie bias met betrekking tot het aantal auto's. Formele statistische testen hebben echter uitgewezen dat deze elementen hier geen significante invloed uitoefenen. We kunnen ons daarom beperken tot het rapporteren van de resultaten van een OLS schatting.

Voor gezinnen met 1 wagen wordt het geprefereerd model samengevat in Tabel 28. We stellen we vast dat de globale voorspellende waarde van het model beperkt is (aangepaste R^2 van 0.16). Nochtans zijn er meerdere variabelen die een significante invloed uitoefenen op de afgelegde afstanden: de woon-werkafstand, de leeftijd van het gezinshoofd, het gezinsinkomen, het type woonplaats, het volume van de wagen, de frequentie waarmee de motor wordt gebruikt als alternatieve vervoersmodus en tenslotte de brandstofkost per kilometer. Vaste kosten zoals de aankoopprijs of de jaarlijkse verkeersbelasting speken geen rol.

De resultaten voor gezinnen met twee wagens zijn grotendeels gelijklopend – zie Tabel 33 voor een samenvatting van het geprefereerd model. Wel stellen we vast dat de leeftijd van het gezinshoofd en het brandstofverbruik geen significante invloed meer uitoefenen. Ook de significantie van andere termen neemt over het algemeen af. De frequentie waarmee tram en trein gebruikt worden als alternatieve modi blijkt ook een significante invloed uit te oefenen. Het ouder en het kleiner model in het bezit van het gezin worden, zoals verwacht, minder gebruikt dan de andere gezinswagen.

6.1.2. METHODOLOGISCHE NABESCHOUWINGEN

Uiteindelijk blijkt dat alleen het aantal auto's dat een gezin bezit nauwkeurig kan worden geschat. Het lijkt ons belangrijk om mogelijke oorzaken te begrijpen en aanbevelingen te formuleren voor toekomstig onderzoek.

Een eerste belangrijke vaststelling is dat veel **antwoorden in het OVG onvolledig of onnauwkeurig** worden ingevuld.

We hebben hierboven reeds aangehaald dat bepaalde belangrijke verklarende variabelen niet werden opgenomen in het model, omdat de respons voor deze variabelen te laag was. Hierdoor kromp de omvang van steekproef in die mate dat de voorspellende waarde van het model nog verder afnam. Het resultaat is dus wel dat een aantal mogelijk cruciale variabelen (deelname aan een carpoolsysteem, het werktijdenregime van de respondent, het gebruik van de auto voor professionele verplaatsingen) niet zijn opgenomen in het model.

Nog veel problematischer zijn echter de onnauwkeurige antwoorden. Het is immers mogelijk om flagrant foute antwoorden te elimineren uit de steekproef. Er bestaat echter geen waterdichte methode om onnauwkeurige, maar geloofwaardige, antwoorden te identificeren.

Een tweede punt heeft betrekking op de **kenmerken van wagens**.

Gezinnen rapporteren het merk en het model van elke wagen, maar er bestaan soms meerdere varianten per model. Het is echter niet mogelijk om de specifieke variant te identificeren die het gezin heeft gekozen. We hebben dan telkens de technische kenmerken genomen van de goedkoopste variant.

Daarnaast ontbraken soms technische kenmerken: we hebben deze, in de mate van het mogelijke, vervangen door benaderende waarden. Ook de tweedehandsprijs is gebaseerd op eigen berekeningen, eerder dan op gerapporteerde cijfers.

Ten derde stellen we vast dat een aantal **mogelijke verklarende variabelen van het verplaatsingsgedrag niet worden opgenomen** in de huidige versie van het OVG.

Zoals hierboven aangehaald, is de transportvraag een afgeleide vraag, die voortvloeit uit de activiteiten waar mensen aan wensen deel te nemen. De woon-werkafstand is een variabele die een duidelijke indicatie geeft van de verplaatsingen die voortvloeien uit de activiteit “werk”. Voor andere activiteiten (zoals verplaatsingen naar vrijetijdsactiviteiten en familie) beschikken we echter slechts over proxy variabelen zoals het aantal gezinsleden – uit onze resultaten blijkt dat deze niet volstaan om het verplaatsingsgedrag nauwkeurig te voorspellen. Wij zouden dus willen aanbevelen dat toekomstige versies van het OVG meer de nadruk zouden leggen op variabelen die meer rechtstreeks betrekking hebben op de activiteiten die verplaatsingen genereren.

Een ander punt heeft betrekking op het gebruik van andere modi. Het OVG bevat zeer veel informatie met betrekking tot het gebruik van andere vervoersmodi dan de auto. Het probleem is echter dat het *gebruik* van deze andere modi even goed een endogene variabele is als het gebruik van de wagen. Men zou dus strikt genomen een model moeten ontwikkelen waarbij het gebruik van alle vervoersmodi simultaan wordt voorspeld, *gegeven* het aanbod van alle individuele modi. Het type gemeente waarin het gezin woont is weliswaar een proxy voor de beschikbaarheid van openbaar vervoer, maar dit volstaat niet. Het vinden van aanvaardbare indicatoren voor het aanbod van openbaar vervoer is echter niet evident: de nabijheid van een bushalte zegt bijvoorbeeld niets over de frequentie van het aanbod, of over de kwaliteit van de aansluitingen. We laten dit open als een onderwerp voor verder onderzoek.

Een vierde belangrijk punt is dat we ons hebben beperkt tot gezinnen **die effectief eigenaar zijn van hun auto(s),** waardoor onze steekproef slechts betrekking heeft op een deel van de totale bevolking.

Aangezien gebruikers van een bedrijfswagen niet geconfronteerd worden met de volledige kosten van hun autogebruik, zullen zowel het type auto dat ze gebruiken, als het aantal kilometers die ze afleggen, in sterke mate afwijken van deze van de gezinnen die we hier beschouwen. Het lijkt ons daarom nog altijd beter om een apart model te schatten voor deze gezinnen.

Het probleem is echter dat we ons aan kunnen verwachten dat gezinnen die beschikken over een of meerdere bedrijfswagen niet representatief zijn: zij zullen waarschijnlijk meer kilometers afleggen om professionele redenen, en zullen waarschijnlijk ook over een hoger gezinsinkomen beschikken. Het resultaat is dan dat, in onze steekproef, het verplaatsingsgedrag van de gezinnen

met hogere inkomens niet representatief zal zijn voor alle gezinnen in de inkomensklassen. Dit is duidelijk een prioritair onderwerp voor verder onderzoek. Een cruciaal element daarbij zal dan bestaan in het schatten van de reële kosten van bedrijfswagens voor de gezinnen die ze gebruiken.

Een vijfde punt heeft betrekking op de **definitie van de autoklassen**. We hebben in het autokeuzemodel de individuele merken en modellen gegroepeerd in klassen op basis van het carrosserietype. De resultaten van het keuzemodel doen vermoeden dat de criteria die we gebruikt hebben voor deze groepering niet de criteria zijn die gezinnen gebruiken voor de keuze van een specifieke autoklasse. Het zoeken van nieuwe classificatiecriteria, misschien op basis van een formele clusteranalyse, is ook een mogelijk onderwerp voor verder onderzoek.

LITERATUURLIJST

Boussauw K. (2011). Ruimte, regio en mobiliteit. Aspecten van ruimtelijke nabijheid en duurzaam verplaatsingsgedrag in Vlaanderen. Garant Uitgevers

de Jong G., Fox J., Pieters M., Daly A. & Smith R. (2004). *A comparison of car ownership models*, **Transport Reviews 24**(4): 379-408, uitgegeven door White Rose University Consortium.

de Jong G. & Gunn H. (2001). *Recent Evidence on Car Cost and Time Elasticities of Travel Demand in Europe*, **Journal of Transport Economics and Policy 35**(2): 137-160, uitgegeven door White Rose University Consortium.

de Jong G., Kouwenhoven M., Geurs K., Bucci P. & Tuinenga J. (2009). *The Impact of Fixed and Variable Costs on Household Car Ownership*, **Journal of Choice Modelling 2**(2): 179-199, uitgegeven door JOCM.

Dubin, J. A & McFadden, D. L, 1984. "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption," Econometrica, Econometric Society, vol. 52(2), pages 345-62, March.

Econometric Software, Inc., *Frequently Asked Questions - Questions about R-squareds*, geraadpleegd op 3/2/2014 via
http://www.limdep.com/support/faq/

Hensher D., Smith N.C., Milthrope F.W. and Barnard P.O. (1992), *Dimensions of Automobile Demand. A Longitudinal Study of Household Automobile Onwership and Use*. Elsevier Science Publishers. Amsterdam

Hensher D., Rose J. & Greene W. (2005). *Applied Choice Analysis - A Primer*, ISBN 978-0-521-60577-9, 717 pp, uitgegeven door Cambridge University Press, Cambridge.

Hoetker G. (2007). *The Use of Logit and Probit Models in Strategic Management Research: Critical Issues*, **Strategic Management Journal 28**: 331-343, uitgegeven door John Wiley & Sons, Ltd.

IDRE - Institute for Digital Research and Education, *FAQ: What are pseudo R-squareds?*, geraadpleegd op 3/2/2014 via
http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm

Kaufman J., *Can pseudo-R-squareds from logistic regressions be compared and used as a measure of fit?*, geraadpleegd op 3/2/2014 via
http://andrewgelman.com/2009/11/03/can_pseudo-r-sq/

Louviere, J., Hensher, D., & Swait, J. 2000. *Stated Choice Methods - Analysis and Applications* Cambridge, Cambridge University Press.

Mayeres, I. & M. Vanhulsel (2014), Simulatiemodel voor de hervorming van de verkeersbelastingen, Rapport Steunpunt Fiscaliteit en Begroting II.

McFadden D. (1978). *Spatial Interaction Theory and Planning Models - [25] Modelling the Choice of Residential Location*, ISBN 0886-0416, Karlqvist A., Lundqvist L., Snickars F. & Weibull J. (eds.), 388 pp, uitgegeven door North Holland Publishing Company, Amsterdam.

Train K. (1986). *Qualitative Choice Analysis - Theory Econometrics, and an Application to Automobile Demand*, ISBN 0-262-20055-4, 247 pp, uitgegeven door The MIT Press, Cambridge, Massachusetts.

Vlaamse Overheid - Departement Mobiliteit en Openbare Werken (MOW). Onderzoek Verplaatsingsgedrag Vlaanderen - Rapport OVG Vlaanderen 4 - Tabellen / Overzicht van enkele belangrijke mobiliteitskenmerken. Mobiel Vlaanderen, geraadpleegd op 3/2/2014 via http://www.mobielvlaanderen.be/ovg/ovg04.php?a=19&nav=11

BIJLAGE A

Technische bijlage: de gebruikte keuzemodellen

Het schatten van discrete keuzemodellen komt neer op het schatten van een aantal nutsfuncties *U*.

De schattingsprocedure gaat ervan uit dat elk individu *n* nutsmaximalisatie nastreeft, dus dat alternatief *i* gekozen wordt indien het bijhorend nut groter is dan het nut van alle andere alternatieven: $U_{in} > U_{jn}$ voor alle *j*. De 'echte' nutsfunctie *U* kan echter nooit geschat worden omdat we niet alle verklarende variabelen kennen of omdat we niet over de juiste data beschikken. We zullen het nut *U* dus moeten benaderen d.m.v. het 'representatieve nut' *V*, die geschat wordt op basis van waarneembare gegevens. De overblijvende foutenterm *e* stelt dan de gemeenschappelijke impact voor van de gegevens die wij niet kunnen waarnemen:

$$U_{in} = V_{in} + e_{in}$$

Afhankelijk van de assumpties die we maken rond de verdeling van de foutenterm *e*, kunnen we andere types modellen schatten (bv. multinomial logit versus nested logit). De kans dat een individu *n* een bepaald alternatief *i* kiest, wordt bijgevolg gemodelleerd als volgt:

$$P_{in} = Prob(U_{in} > U_{jn}, voor\ alle\ j \neq i)$$

De meest eenvoudige kansberekeningen zijn mogelijk indien we ervan uitgaan dat alle alternatieven even sterk op elkaar lijken. We veronderstellen dan dat de foutentermen e_{in} onafhankelijk van elkaar maar identiek verdeeld zijn volgens een Weibull-verdeling - we gebruiken een 'multinomial logit' (MNL)-model. Het gevolg is dat de kans op alternatief *i* berekend kan worden als:

$$P_{in} = \frac{e^{V_{in}}}{\sum_{all\ j} e^{V_{jn}}}$$

In dit geval is de relatieve kans dat alternatief *i* en *j* worden gekozen ($P_{in}/P_{jn} = e^{Vin}/e^{Vjn}$) onafhankelijk van de beschikbaarheid van andere alternatieven – dit resultaat bij MNL modellen wordt ook wel eens de "Independence of irrelevant alternatives" (IIA) genoemd in de literatuur.

Een nested logit (NL)-structuur daarentegen veronderstelt niet langer dat alle alternatieven even sterk op elkaar lijken met betrekking tot de niet-waarneembare kenmerken. Gelijkaardige[57] alternatieven bevinden zich dan in een "nest".

Dit heeft een invloed op het berekenen van de kans op alternatief *i* behorende tot nest *k*:

$$P_{in} = P_{in|nest\ k} \times P_{nest\ k}$$

De eerste factor is de voorwaardelijke kans dat alternatief *i* wordt gekozen, gegeven dat het individu een alternatief uit nest *k* kiest. De tweede factor stelt de "onvoorwaardelijke" kans voor dat het individu een alternatief uit nest *k* kiest.

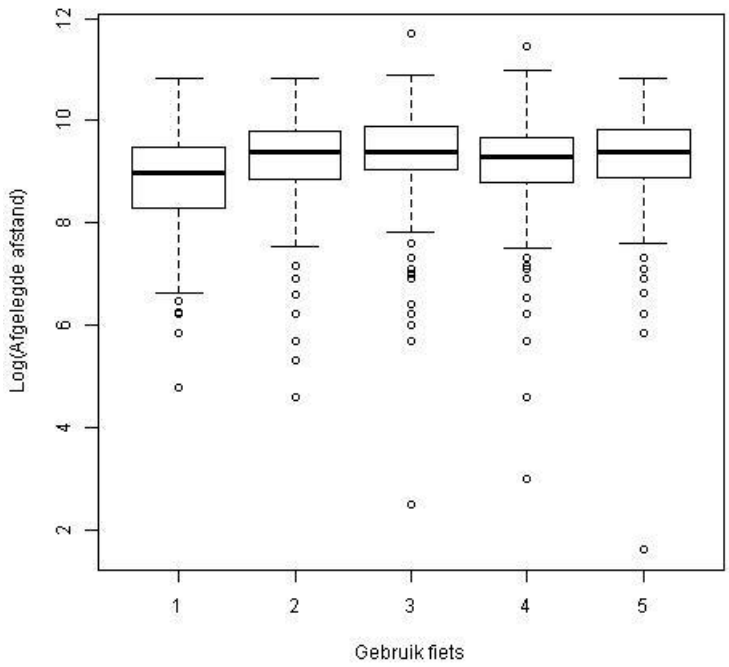
[57] Nogmaals, "gelijkaardig" met betrekking tot de niet-observeerbare kenmerken.

De eerste factor wordt berekend zoals in het MNL-geval van hierboven. Voor het berekenen van de tweede factor ($P_{nest\ k}$) moeten we er echter rekening mee houden dat het nut van het individu wordt bepaalde door twee elementen:

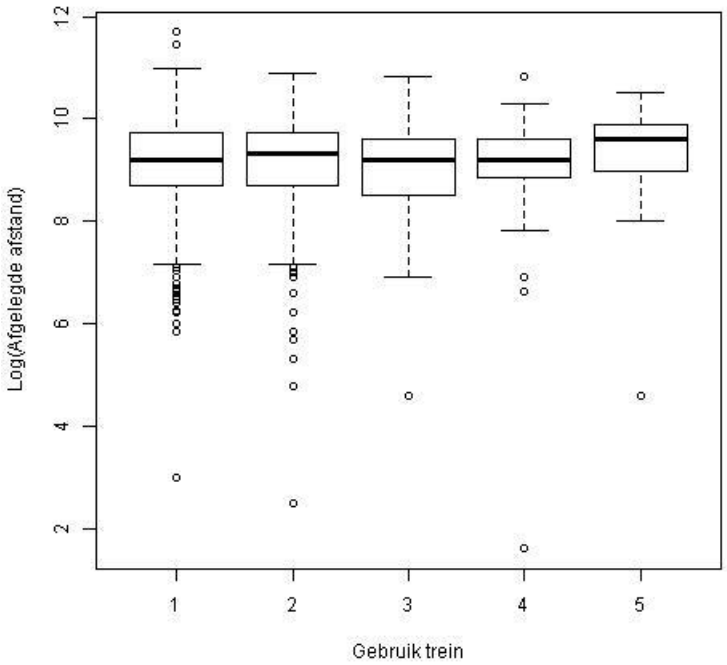
- Het nut dat voortvloeit uit elementen die gemeenschappelijk zijn aan alle elementen van de beschouwde nest.
- De verwachte waarde van het maximaal nut dat men kan halen uit de keuze van een van de elementen uit de nest. In een NL model wordt deze verwachte waarde voorgesteld aan hand van de zogenaamde “Inclusive value” (IV).

Voor meer achtergrond en wiskundige onderbouwing van de intuïties in deze technische bijlage verwijzen we naar Train (1986), hoofdstuk 1, 2 en 4.

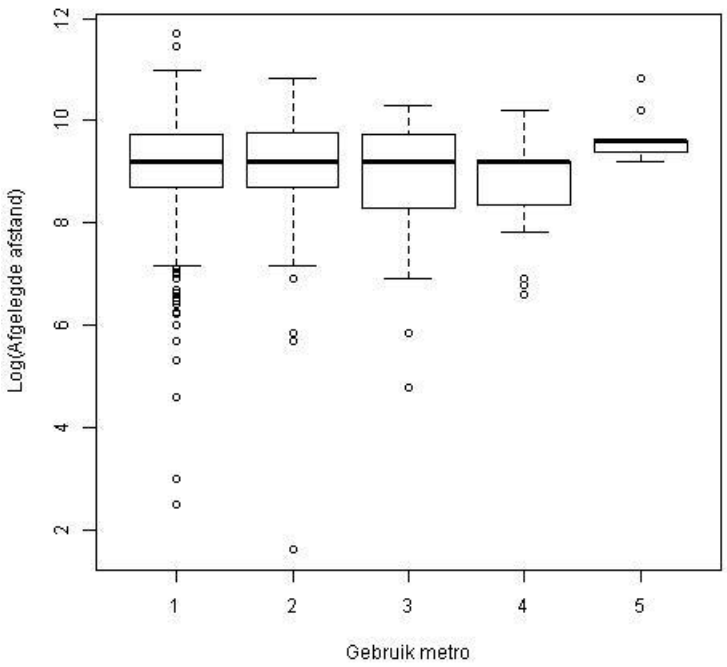
Doosdiagrammen voor alternatieve modi



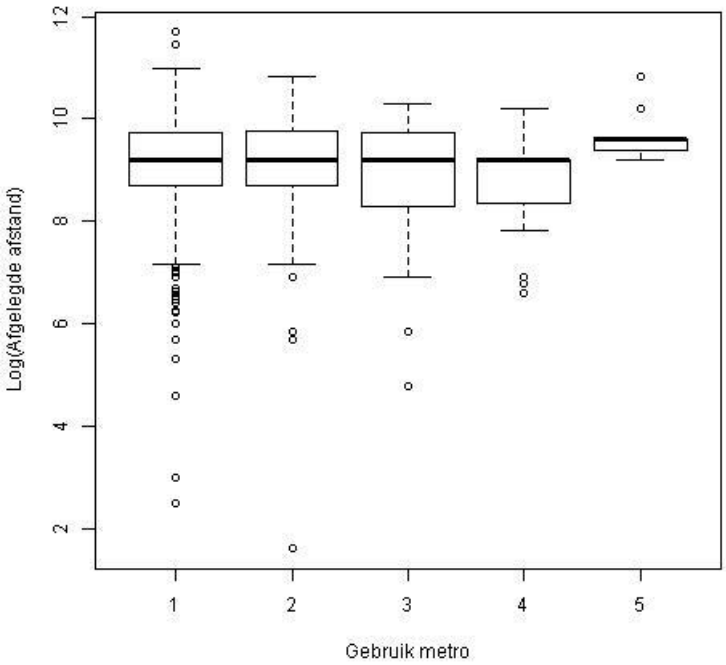
Figuur 19: afgelegde km versus gebruik fiets



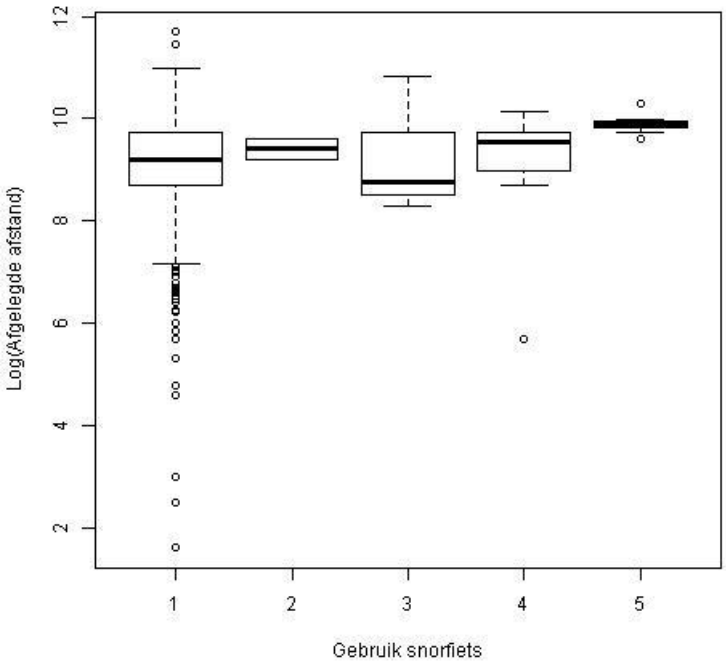
Figuur 20: afgelegde km versus gebruik trein



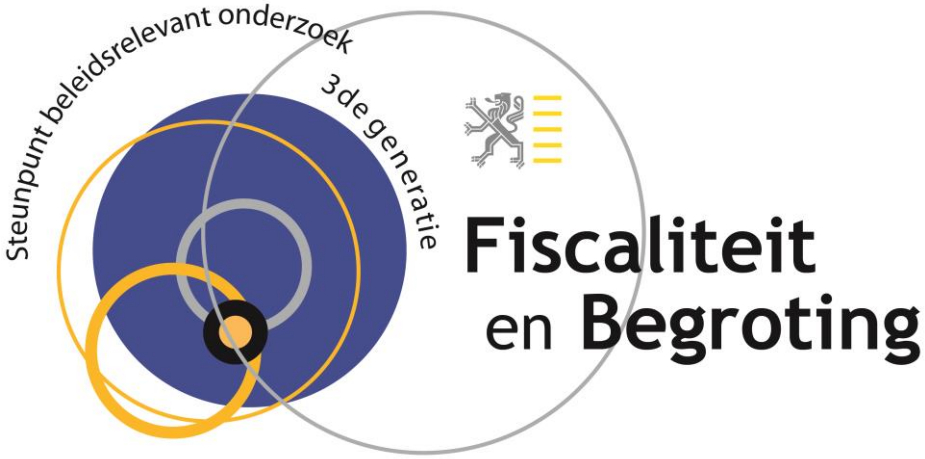
Figuur 21: afgelegde km versus gebruik metro



Figuur 22: afgelegde km versus gebruik moto



Figuur 23: afgelegde km versus gebruik snor- en motorfiets



UNIVERSITEIT GENT

vito
vision on technology

KATHOLIEKE UNIVERSITEIT LEUVEN