



UvA-DARE (Digital Academic Repository)

De zoekduur van EBB-schoolverlaters: definities en berekeningswijze

Berkhout, P.H.G.

Published in:
SEO-Rapport

[Link to publication](#)

Citation for published version (APA):

Berkhout, P. (2002). De zoekduur van EBB-schoolverlaters: definities en berekeningswijze. SEO-Rapport, 635a.

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <http://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

De zoekduur van EBB-schoolverlaters: definities en berekeningswijze

drs. P.H.G. Berkhout

Amsterdam, augustus 2002

“Het doel der Stichting is het verrichten van economische onderzoeken, zowel op het terrein der sociale economie als op dat der bedrijfseconomie, ten dienste van wetenschap en onderwijs, mede ten nutte van overheid en bedrijfsleven”

(art. 2 der stichtingsakte)

SEO-rapport 635A

ISSN 0926-2806

Copyright © 2002 SEO Amsterdam. Behoudens de in of krachtens de Auteurswet 1912 gestelde uitzonderingen mag niets uit deze uitgave worden verveelvoudigd en/of openbaar gemaakt op welke wijze dan ook zonder voorafgaande schriftelijke toestemming van de Stichting voor Economisch Onderzoek te Amsterdam.

Voorwoord

Naar aanleiding van het SEO-rapport “De bepaling van de arbeidsmarktrelevantie van opleidingen” (1990) werden in 1991 door de Stichting voor Economisch Onderzoek der Universiteit van Amsterdam (SEO) in samenwerking met het CBS de mogelijkheden onderzocht van jaarlijks terugkerende analyses over schoolverlaters in de Enquête Beroepsbevolking (EBB). Dat onderzoek resulteerde in een rapport over de arbeidsmarktrelevantie van opleidingen (Berkhout en Mot, 1992), dat in de daarop volgende jaren elk jaar een vervolg kreeg (Berkhout et al, 1993, 1994, 1995 en 1996). De uitkomsten vonden tot aan de dag van vandaag (augustus 2002) ook elk jaar hun weg naar de Schoolverlatersbrief van wat thans het Centrum voor Werk en Inkomen (CWI) heet en voorheen ‘Arbeidsvoorziening’ heette. De Schoolverlatersbrief heeft tot doel de belangrijkste uitkomsten van onderzoek over schoolverlaters bijeen te brengen ten behoeve van alle belanghebbenden. Ook werden in de meeste jaren de uitkomsten opgenomen in verschillende periodieken van het CBS.

In 2002 werd het onderzoek voor de tiende keer uitgevoerd. Ter gelegenheid van dit lustrum heeft de SEO dit rapport / research memorandum samengesteld. Het doel van deze notitie is een verantwoording te geven van de berekeningswijze van de zoekduren van EBB-schoolverlaters.

Dit schoolverlatersonderzoek kwam tot stand door de gezamenlijke inspanningen van een aantal mensen. Van zowel de SEO als het CBS werkten in de afgelopen 10 jaren verschillende onderzoekers mee aan het project. Van SEO-zijde waren dat: Esther Mot, Adriaan Kalwij, Tom Poot, Jacco Provoost, Jasper de Winter en Jaap Anne Korteweg. En van het CBS: Johan van der Valk, Jacques Thijssen, Mariëlle Reemers. Het project werd afwisselend door verschillende opdrachtgevers gefinancierd, te weten Arbeidsvoorziening Nederland, het Ministerie van Sociale Zaken en Werkgelegenheid en thans het CWI. Inspirerende begeleiding van de zijde van de opdrachtgevers werd verzorgd door achtereenvolgens: Maarten Wegerif, Ben Lok en Anske Bouman. Allen droegen bij aan dit werk.

1 Inleiding

Sinds het begin van de jaren 90 wordt op basis van gegevens uit de Enquête Beroepsbevolking (EBB) van het Centraal Bureau voor de Statistiek (CBS) de werkloosheidsduur van schoolverlaters geregistreerd. Deze gegevens worden jaarlijks onderzocht door de Stichting voor Economisch Onderzoek (SEO) der Universiteit van Amsterdam en worden onder meer gepubliceerd in de jaarlijkse ‘Schoolverlatersbrief’ van het Centrum voor Werk en Inkomen (CWI, voorheen: Arbeidsbureau).

Bij de berekening van de werkloosheidsduur – die in het vervolg wordt aangeduid met de term ‘zoekduur’ – komen in de praktijk twee problemen kijken, die om een oplossing vragen. Ten eerste wordt de zoekduur gemeten door tussen twee tijdstippen het aantal gehele maanden te berekenen. De waargenomen zoekduur T is dus een geheel getal en in feite slechts een grove benadering van de werkelijke zoekduur, die eerder in dagen zou moeten worden gemeten. Ter illustratie: het waarnemen van de waarde $T=2$, dat de werkelijke zoekduur ligt in het interval $[30, 90]$ dagen. Ten tweede is het voorkomen van onvoltooide zoekduren – dat zijn zoekduren van individuen die op het moment van enquêteren nog geen baan hebben gevonden – problematisch wanneer men groepsgemiddelden wil berekenen. Onvoltooide zoekduren zijn afgekapt bij het moment van enquêteren en vertegenwoordigen onvolledige informatie. Indien hiermee geen rekening wordt gehouden, worden berekende groepsgemiddelden neerwaarts vertekend. Beide problemen worden aangepakt door de zoekduren van schoolverlaters met behulp van een *mixed proportional hazard* model te analyseren. De per persoon verschillende *survival-functie* die met dit model wordt geschat, wordt vervolgens gebruikt om door middel van simulatie voor elke respondent de ‘werkelijke’ zoekduur te schatten en voor de onvoltooide zoekduren het restant te bepalen. De zoekduur die aldus ontstaat kan naar believen worden gegroepeerd en gemiddeld, zonder dat vertekening optreedt in de uitkomsten. Voorwaarde is natuurlijk wel dat de groepen voldoende omvang hebben, waarbij men als ondergrens 30 representatieve waarnemingen kan aanhouden.

Dit rapport is als volgt gestructureerd. In Hoofdstuk 2 passeren de definities van de begrippen die in het onderzoek een rol spelen de revue. Hoofdstuk 3 beschrijft het model waarmee de zoekduren van EBB-schoolverlaters worden geanalyseerd. In hoofdstuk 4 wordt uiteengezet hoe de modeluitkomsten worden gebruikt om berekening van eenvoudige beschrijvende statistieken mogelijk te maken. Tevens wordt in dat hoofdstuk ingegaan op de betrouwbaarheid van de uitkomsten.

2 Definities

De Enquête Beroepsbevolking (EBB) is een onderzoek waarin het veldwerk gedurende het hele jaar plaatsvindt. In elke maand worden ongeveer 9000 adressen benaderd. Bij het huishouden op dat adres wordt een persoonlijk interview gehouden waarbij een vragenlijst met een interviewduur van circa een half uur wordt afgenomen. De vraagstelling omvat onder meer vragen over het hebben van werk, zoeken naar werk, kenmerken van de werkring, actueel en gevolgd onderwijs.

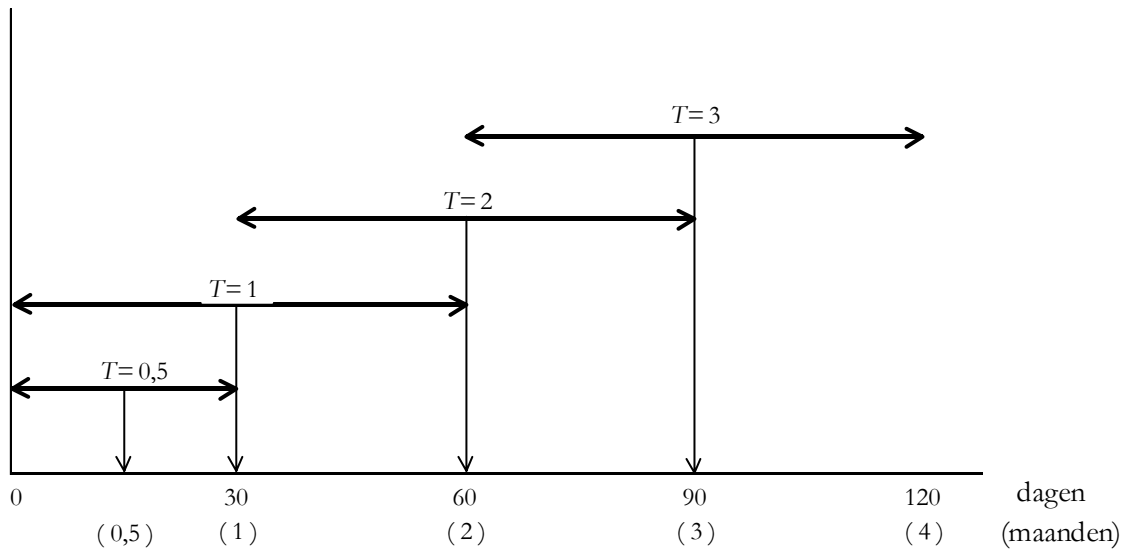
De definitie van het begrip schoolverlater maakt gebruik van het voortschrijdende karakter van de EBB. Een *schoolverlater* wordt gedefinieerd als iemand tussen 15 en 35 jaar, die op het moment van enquêteren niet langer dan één jaar geleden het voltijd-onderwijs verliet. Deze definitie impliceert dat iemand gedurende de periode van één jaar volgend op een periode van voltijd-onderwijs tot de schoolverlaters wordt gerekend. Een schoolverlater kan gedurende dat jaar deeltijd-onderwijs volgen, werken, werk zoeken of niet participeren op de arbeidsmarkt. Ook kan hij deze toestanden afwisselen en zelfs na dat jaar weer besluiten voltijd-onderwijs te gaan volgen. In dat laatste geval wordt hij op zeker moment in de toekomst weer een keer schoolverlater.

De *zoekduur* wordt gedefinieerd als het aantal gehele maanden tussen het moment van schoolverlaten en de aanvang van de eerste relevante baan. Het *moment van schoolverlaten* wordt verondersteld gelijk te zijn aan de laatste maand waarin voltijd-onderwijs werd genoten. Meer precies wordt als maand van schoolverlaten de maand genomen waarop de diploma wordt uitgereikt. De aanvang van de eerste *relevante* baan is de maand waarin de respondent een werkring begon, waarin hij niet als uitzend-, inval- of oproepkracht werkzaam was. Door een dergelijke restrictie op te leggen aan de gevonden baan, wordt niet elke willekeurige baan gezien als een 'echte' baan. De gedachte hier achter is dat naarmate men strengere eisen oplegt aan de gevonden baan, verschillen in de zoekduur meer zeggen over de arbeidsmarktrelevantie van opleidingen. (Teulings, 1992). De zoekduur van een individu wordt toegekend aan een bepaald kalenderjaar indien de eerste zoekmaand in dat kalenderjaar valt.

Zoekduur T wordt gemeten in hele maanden, omdat van zowel het begin- als het eindpunt alleen het jaar en de kalendermaand worden waargenomen. Indien men de 12 maanden nummert van 1 t/m 12 en de zoekduur gelijk stelt aan het verschil dan neemt de zoekduur de waarden 0, 1, 2, ... aan. Gesteld dat de werkelijke zoekduur Z in dagen wordt gemeten en dat het beginpunt willekeurig plaats vindt in de maand, dan is een correctie van de 0-waarneming noodzakelijk. Als $T=0$, dan vindt de schoolverlater nog in dezelfde kalender-

maand waarin hij zich op de arbeidsmarkt aanbiedt een relevante baan. De werkelijke zoekduur Z bedraagt in dat geval tenminste 1 dag en maximaal 31 dagen ofwel gemiddeld een halve maand. De zoekduur wordt in dat geval gecorrigeerd: $T=1/2$. Voor waarnemingen $T>0$, blijkt een correctie niet nodig. Als de passende baan volgens waarneming aanvangt in de opvolgende kalendermaand, dan wordt een zoekduur waargenomen van $T=1$. De werkelijke zoekduur Z is dan tenminste 1 dag en maximaal 62 dagen, ofwel gemiddeld één maand. In het algemeen geldt: als $T=k$ voor $k=1, 2, \dots$ dan is de werkelijke zoekduur Z geconcentreerd rond k met afwijking van maximaal ± 30 dagen. In Figuur 2.1 wordt dat grafisch weergegeven.

Figuur 2.1 Discrete waarneming van de zoekduur



Uit Figuur 2.1 blijkt dat geen één-op-één relatie bestaat tussen de werkelijke zoekduur Z in dagen en de gemeten zoekduur T uitgedrukt als verschil tussen de maandnummers. Een willekeurige zoekduur kan met even grote kans twee verschillende discrete waarden aannemen. Men kan dat in Figuur 2.1 nagaan door op een willekeurige positie op de horizontale as een verticale lijn te tekenen. Deze lijn snijdt altijd precies twee zoekduren T . Bovendien is het mogelijk dat twee waargenomen zoekduren T_x en T_y waarvoor geldt $T_x < T_y$ corresponderen met werkelijke zoekduren Z_x en Z_y , waarvoor geldt $Z_x > Z_y$. Bijvoorbeeld: persoon x heeft een werkelijke zoekduur van 34 dagen en persoon y een zoekduur van 50 dagen. Persoon y had dus een langere zoekduur dan x , maar toch kan het omgekeerde worden waargenomen. Dat wil zeggen: voor persoon y nemen we mogelijk $T=1$ waar en voor persoon x $T=2$. Dit impliceert dat zoekduur T met een *random* meetfout wordt waargenomen. In een duurmodel waarin een storingsterm in de hazardspecificatie is opgenomen, hoeft de meetfout in T niet te leiden tot structurele vertekening van de resultaten. In het onderhavige onderzoek wordt zo'n storingsterm opgenomen.

Negatieve zoekduren kunnen theoretisch ontstaan wanneer de baan eerder in de tijd aanvangt (of werd aanvaard) dan het moment van schoolverlaten. Kleine negatieve waarden van -1 , -2 en -3 worden gecorrigeerd, omdat het hier (waarschijnlijk) schoolverlaters betreft die feitelijk direct na schoolverlaten aan de slag gaan. De zoekduur wordt voor dergelijke waarnemingen gelijk gesteld aan een kleine positieve waarde: 0.01. Daarmee wordt verondersteld dat hun zoekduur ruim 7 uur bedraagt, zeg een goede nachtrust. Grote negatieve duren worden buiten de analyse gehouden met het argument dat ze met grote waarschijnlijkheid het gevolg zijn van fouten in de gegevens.

Van schoolverlaters is de laatst gevolgde opleiding in de vorm van een vijfcijferige code volgens de Standaard Onderwijs Indeling van 1987 (SOI 87) bekend. De opleiding betreft de laatst gevolgde opleiding. Deze hoeft niet noodzakelijkerwijs gelijk te zijn aan de hoogst behaalde opleiding. Het is namelijk mogelijk dat een schoolverlater de laatst gevolgde opleiding zonder diploma afsluit. Analyses vinden plaats op vier opleidingsniveaus:

- 1) vmbo (SOI 1^e digit 3)
- 2) bol (voorheen: mbo)/havo/vwo (SOI 1^e digit 4) (bol is nog niet ingeburgerd)
- 3) hbo (SOI 1^e digit 5)
- 4) wo (SOI 1^e digit 6)

Op elk niveau worden onderwijssectoren onderscheiden onder de namen:

- 1) algemeen
- 2) technisch
- 3) economisch
- 4) verzorgend
- 5) en eventueel: overig

De sectoren worden afgebakend aan de hand van de digits 2 t/m 5 van de SOI-code. In onderstaande tabel wordt dat weergegeven.

Tabel 2.1 Onderwijssectoren en SOI-code

	algemeen	economisch	technisch	verzorgend
1 ^e SOI-digit 1–4	2 ^e /3 ^e digit 00–04	2 ^e /3 ^e digit 60–69	2 ^e /3 ^e digit 20–49	2 ^e /3 ^e digit 05–19, 50–59, 70–79
1 ^e SOI-digit 5–6	nvt	2 ^e /3 ^e digit 60–69	2 ^e /3 ^e digit 20–49	2 ^e /3 ^e digit 05–19, 50–59, 70–79

Bron: CBS

Deze indeling is overigens anno 2002 aan revisie toe, omdat met de invoering van de Wet Educatie en Beroepsonderwijs (WEB) in schooljaar 1997/1998 het lagere en middelbare beroepsonderwijs is geherstructureerd. Dit heeft geleid tot een nieuw onderwijsclassificatie, de SOI 98. Deze classificatie is tot op heden niet gebruikt voor de analyses, maar dat zal binnenkort wel gebeuren.

3 Model

Econometrische modelbouw begint vaak met de keuze van een verdelingsfunctie, die de waargenomen stochastiek van het te modelleren fenomeen moet ondervangen. In het duurmodel is het uitgangspunt de *hazardrate* θ ofwel de *uittredevoet*. De uittredevoet is gedefinieerd als de voorwaardelijke kans per oneindig klein tijdinterval dt om vanuit een zekere toestand op tijdstip t over te gaan naar een andere. In het geval van schoolverlaters beschrijft θ de overgang van werkloosheid naar betaald werk. Verondersteld wordt daarmee dat de schoolverlater ten minste voor een korte tijd na schoolverlaten (bijvoorbeeld een goede nachtrust lang) werkloos is. Tijdvariabele t meet daarbij de (individuele) verstreken tijd in de toestand van beschouwing, in dit geval dus werkloosheid. In formele notatie:

$$\theta(t) \equiv \lim_{dt \rightarrow 0} \frac{P(t \leq T < t + dt \mid T \geq t)}{dt} = \lim_{dt \rightarrow 0} \frac{P(t \leq T < t + dt)}{P(T \geq t)} = \frac{f(t)}{1 - F(t)} \quad (3.1)$$

De uittredevoet is in het geval van continue tijd, zoals in bovenstaande vergelijking, een intensiteit en geen kans. Essentieel in de definitie is de voorwaarde $T \geq t$ die bepaalt dat alleen de overlevenden tot het tijdstip t in beschouwing worden genomen; zij die eerder uitstroomden doen niet meer mee. Instructief voor begrip van de uittredevoet en het onderscheid dat de voorwaarde aanbrengt, is het te plaatsen in de context van de levensduur van mensen. De onvoorwaardelijke kans bij de geboorte om 65 te worden geeft het percentage weer van alle mensen - dus ook zij die dat niet halen - die in het 66-ste levensjaar sterven. De uittredevoet geeft weer welk deel van de 65-jarigen voor hun 66-ste verjaardag sterft. Uit definitie (3.1) volgt dat een één-op-één relatie bestaat tussen de uittredevoet en de verdelingsfunctie $F(t)$ met bijbehorende dichtheidsfunctie $f(t) = \partial F(t) / \partial t$. Met andere woorden: met de keuze van een bepaalde functionele vorm voor $\theta(t)$ wordt impliciet voor een zekere verdelingsfunctie gekozen en andersom. De functie $1 - F(t)$ wordt in de theorie van duurmodellen ook wel de *survival function* genoemd, vaak aangeduid met $S(t)$.

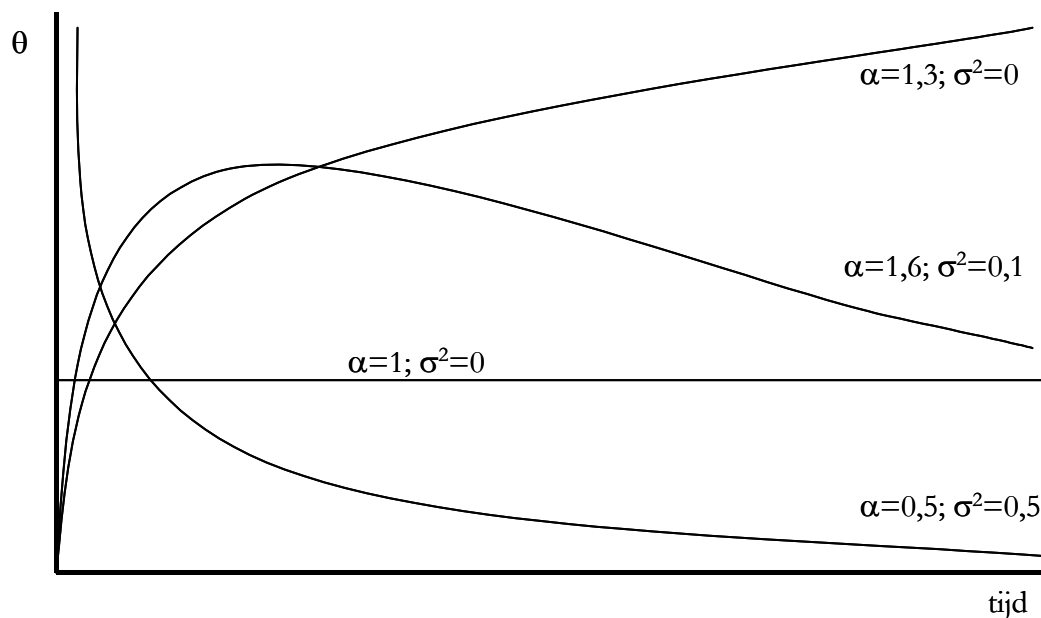
Een speciale klasse van duurmodellen ontstaat door de hazardfunctie te vermenigvuldigen met een positieve stochast ν met verwachting gelijk aan één en een vrij variërende variantie σ^2 . Deze stochast is de analogie van de storingsterm in kleinste kwadraten regressie en wordt opgevoerd om niet-waargenomen heterogeniteit van respondenten te ondervangen. Deze modellen worden ook wel *frailty models* genoemd (Hougaard, 1986). Wordt voor stochast ν de gamma verdeling gekozen in combinatie met de (proportionele) Weibull hazard functie, dan resulteert de Burr verdeling. Dit model behoort tot de klasse van *mixed proportional hazard*

(MPH) modellen. In het vervolg refereren we aan dit model met naam “Burr model”. De uittredevoet van het Burr model heeft de volgende functionele vorm:.

$$\theta_m(t) = \frac{\lambda \alpha t^{\alpha-1}}{1 + \sigma^2 \lambda t^\alpha} \quad \text{waarbij } \lambda = \exp(x'\beta) \quad (3.2)$$

Het verloop van de uittredevoet in de tijd kan in het Burr model allerlei vormen aannemen: constant, monotoon dalend, monotoon stijgend en eerst stijgend dan dalend. In Figuur 3.2 worden voorbeelden daarvan weergegeven. Het Burr model behoort zowel tot de klasse van *frailty models* als tot die van MPH-modellen. Het model heeft het Exponentiële model ($\alpha = 1, \sigma^2 = 0$), het Weibull model ($\sigma^2 = 0$) en het Log-logistische model ($\sigma^2 = 1$) als speciale gevallen. Verder zij opgemerkt dat behoudens in de eerste twee speciale gevallen, het proportionele effect van een willekeurige variabele x op de uittredevoet in het Burr model niet op ieder tijdstip hetzelfde is. Met andere woorden: de eigenschap dat het verloop van de uittredevoet in de tijd en het effect van variabelen gescheiden zijn, is niet van toepassing indien sprake is van niet-waargenomen heterogeniteit ($\sigma^2 > 0$). Technische details van het Burr model worden in de bijlage gegeven. Voor goede introducties in de theorie der duurmodellen in het algemeen wordt verwezen naar Kalbfleish and Prentice (1980), Kiefer (1988) en Lancaster (1990).

Figuur 3.2 Mogelijke vormen van de uittredevoet in het Burr model



De survivalfunctie heeft de volgende vorm:

$$S_m(t) = (1 + \sigma^2 \lambda t^\alpha)^{-1/\sigma^2} \quad (3.3)$$

Vergelijking (3.3) wordt gebruikt om, nadat de parameters α , σ^2 en λ geschat zijn, onvoltooide zoekduren te voorspellen. De zoekduren van EBB-schoolverlaters worden met behulp van het Burr-model geanalyseerd. Dit model is als standaardprocedure in STATA (versie 7) beschikbaar.

De verklarende variabelen van de zoekduur van schoolverlaters verschijnen in de vector x in de term λ van vergelijkingen (3.2) en (3.3). Daartoe worden gerekend: geslacht, leeftijd, etniciteit, landsdeel (noord, oost, zuid en west) en een dummy voor het behalen van het diploma voor de laatst gevolgde opleiding. Daarnaast wordt de toestand van de arbeidsmarkt per maand gekarakteriseerd door middel van enkele macro-gegevens: het aantal openstaande vacatures (3-maands interpolaties), het aantal geregistreerde werklozen (voortschrijdend 3-maandsgemiddelde), de kapitaalmarktrente en de CBS-producentenvertrouwensindex. Deze grootheden variëren per maand. Aan alle waargenomen zoekduren worden deze grootheden per maand gekoppeld. Ter illustratie hiervan een voorbeeld. Stel dat de zoekduur van een schoolverlater begint in mei 1999 en eindigt in oktober 1999. De waargenomen zoekduur bedraagt dan $T=5$. Deze waarneming valt dan uiteen in vijf records: mei 99, juni 99, juli 99, augustus 99 en september 99. De standen van vacatures, werklozen, rente en vertrouwen in deze maanden worden aan het bestand gekoppeld. De vier grootheden worden als *‘in de tijd variërende’* variabelen opgenomen. Het opnemen van deze variabelen betekent dat een bestand van 15 duizend schoolverlaters wordt opgeblazen tot meer dan 60 duizend ‘waarnemingen’.

4 Zoekduren van EBB-schoolverlaters

4.1 Berekeningswijze

Het in Hoofdstuk 3 beschreven model wordt gebruikt voor twee doeleinden. Ten eerste worden *onvoltooide zoekduren* waargenomen. Dat zijn de zoekduren van schoolverlaters die op het moment van enquêteren nog geen baan hebben gevonden. De enquêteur kwam toevallig eerder langs dan de baan, met als gevolg dat de zoekduur (voor de onderzoeker) niet kon worden voltooid. Een onvoltooide zoekduur bevat echter wel waardevolle informatie, maar die informatie is alleen niet helemaal volledig. We weten feitelijk alleen hoe lang de werkzoekende *ten minste* werkloos was, niet hoe lang die werkloosheid *precies* duurde. Dergelijke zoekduren kunnen wel worden benut in modelschattingen, en dat is nodig ook, omdat het weglaten van dergelijke waarnemingen de kans met zich meebrengt dat een selectieve groep schoolverlaters (die met langere zoekduren) buiten beschouwing blijft. Ook in de berekening van eenvoudige statistieken, zoals het gemiddelde, geven onvoltooide zoekduren problemen. Weglaten levert vertekening op, maar ze zonder meer gebruiken ook. In beide gevallen wordt het werkelijke gemiddelde structureel onderschat. In de EBB is een aanzienlijk deel van de waargenomen zoekduren onvoltooid. In Tabel 4.1 wordt het percentage voltooid en onvoltooid weergegeven naar kalenderjaar van schoolverlaten voor de periode 1995 t/m 2001.

Tabel 4.1 Onvoltooide en voltooide zoekduren (percentages)

	1995	1996	1997	1998	1999	2000	2001
voltooid	35	44	58	62	64	63	60
onvoltooid	65	56	42	38	36	37	40

Bron: CBS

Uit Tabel 4.1 blijkt dat een aanzienlijk deel – variërend van 65 tot 36% – van de zoekduren van schoolverlaters in de EBB onvoltooid is. Het aandeel varieert met de conjunctuur. In de gespannen arbeidsmarkt van 1998-2000 is het aantal onvoltooide zoekduren relatief laag, in de ruime arbeidsmarkt van 1995 is het aantal relatief hoog. Het geschatte model, waarin overigens ook voor de conjunctuurbeweging wordt gecorrigeerd, wordt gebruikt om het restant van de onvoltooide duren te simuleren, zodat berekening van eenvoudige statistieken mogelijk wordt. Dat gaat als volgt. De modelschattingen geven voor iedere respondent de parameters van zijn individuele survivalfunctie $S_i(t)$. Laat T_{onv} een onvoltooide zoekduur zijn, dan wordt een willekeurige trekking gedaan in het interval $[S_i(T_{\text{onv}}), S_i(30)]$. Het restant van de zoekduur wordt vervolgens gelijk gesteld aan $S_i^{-1}(\text{trekking}) - T_{\text{onv}}$. De gesimuleerde

zoekduur is dus tenminste even lang als de waargenomen onvoltooide duur en nooit langer dan 30 maanden.

Het tweede doel waarvoor het model wordt aangewend is om de zoekduren ook ‘achter de komma’ van waarden te voorzien. Zoals in Figuur 2.1 grafisch wordt uitgebeeld, wordt alleen het aantal hele maanden tussen het moment van schoolverlaten en de aanvang van de eerste baan waargenomen. De werkelijke zoekduur wordt als het ware geclusterd op de tijdas waargenomen. Zo betekent bijvoorbeeld het waarnemen van de waarde $T=2$, dat de werkelijke zoekduur ligt in het interval $[30, 90]$ dagen. Voor berekening van groeps-gemiddelden is dat geen groot probleem. De waarde van T ligt in het midden van het interval en dus zullen afwijkingen ‘wegmiddelen’ wanneer men het gemiddelde bepaalt¹. Een goede reden om de voltooide zoekduren² toch met behulp van simulatie cijfers achter de komma te geven, is omdat dan ook de mediaan van groepen schoolverlaters kan worden bepaald. Zou men de mediaan bepalen met de waargenomen T , dan resulteert een geheel getal. Dat getal geeft niet alleen de mediaan niet goed weer, ook wordt het vergelijken met andere groepen lastig omdat verschillen in medianen in hele maanden worden uitgedrukt. De simulatieprocedure is analoog aan die voor de onvoltooide duren. De modelschattingen geven voor iedere respondent de parameters van zijn individuele survivalfunctie $S_i(t)$. Laat $T_{\text{vol}} \geq 1$ een voltooide zoekduur zijn, dan wordt een willekeurige trekking gedaan in het interval $[S_i(T_{\text{vol}}-1), S_i(T_{\text{vol}}+1)]$. De zoekduur is gelijk aan S_i^{-1} (trekking). Voor zoekduur $T_{\text{vol}}=1/2$ geldt een aangepast trekkingsinterval: $[S_i(0), S_i(1)]$.

Na de bovenbeschreven simulatiestappen beschikken we over een zoekduurvariabele, zeg Z^* , waarvan zonder problemen het gemiddelde of de mediaan voor groepen van schoolverlaters kan worden berekend. De random trekkingen die gebruikt worden voor de simulaties brengen met zich mee dat Z^* voor elke waarneming verandert als opnieuw wordt gesimuleerd. Groepsgemiddelden kunnen hierdoor licht variëren. Om dit te voorkomen, kunnen de trekkingen worden gebaseerd op *seeds*, die afhankelijk zijn van persoons-kenmerken.

¹ Dat is overigens niet helemaal waar. Meestal zal door de scheefheid van de onderliggende verdelingsfunctie meer massa onder T te vinden zijn dan erboven. Het effect hiervan is echter pas in de tweede decimaal van het gemiddelde merkbaar en dus praktisch verwaarloosbaar.

² De onvoltooide duren hebben als gevolg van de simulatie van het restant al cijfers achter de komma

4.2 Betrouwbaarheid

Hoe betrouwbaar zijn de gemiddelden, medianen en eventueel andere beschrijvende statistieken, wanneer gebruik wordt gemaakt van Z^* ? De methode gebruikt alle beschikbare informatie zo goed mogelijk om een schatting te maken van de werkelijke zoekduur Z . De methode herbergt echter wel enige impliciete veronderstellingen, die op zich niet stringent of onrealistisch zijn, maar wel het vermelden waard zijn. Zo wordt verondersteld dat de zoekduur van schoolverlaters kan worden gemodelleerd met een zij het flexibel, maar desondanks niet minder abstract en theoretisch model. De weerbarstige werkelijkheid wordt altijd gesimplificeerd wanneer modellen hun intrede doen, en dat geldt ook hier. Met het gekozen model wordt verondersteld dat het verloop van de uittredevoet van schoolverlaters in de tijd in één van de opgelegde functionele vormen past, terwijl dat niet zo hoeft te zijn. Daarnaast wordt verondersteld dat we de zoekduur kunnen verklaren met een simpele lineaire combinatie van waargenomen karakteristieken, en dat we onvoltooide zoekduren kunnen voorspellen aan de hand van die karakteristieken en de geschatte relatie. In de praktijk moeten we ermee rekening houden dat we slechts een beperkt aantal verklarende variabelen waarnemen en dat geschatte relatie niet meer dan eerste-orde benadering is.

Het moge duidelijk zijn dat de berekende Z^* meestal zal afwijken van de werkelijke niet waargenomen Z . Dat is op zich geen probleem, omdat deze afwijkingen in groeps-gemiddelden van grote groepen tegen elkaar zullen wegvallen. Er zijn twee bronnen van onzuiverheid, die de betrouwbaarheid van beschrijvende groepsstatistieken in gevaar brengen. De eerste is de groepsomvang. De centrale limietstelling uit de statistiek leert dat bij kleine groepen van zeg 30 waarnemingen de variantie van een steekproefgemiddelde veel groter is dan bij zeg 100 waarnemingen. Het is dus van belang niet al te kleine groepen in beschouwing te nemen. De tweede bron van onnauwkeurigheid is het aantal onvoltooide zoekduren in een groep. Naarmate dit aantal groter is, wordt het aantal op basis van het model voorspelde zoekduren groter. Dat is een reden om gemiddelden van groepen met relatief veel oorspronkelijk onvoltooide zoekduren minder te vertrouwen dan wanneer slechts enkele onvoltooide zoekduren in de een groep worden waargenomen. Wanneer sprake is van ‘relatief veel’ onvoltooide zoekduren blijft een kwestie van smaak, maar in elk geval is meer dan 50% veel, omdat dan de voorspelde waarnemingen in de meerderheid zijn.

Merk op dat de groepsmediaan pas onzuiver wordt, indien meer dan 50% onvoltooid is, omdat de mediaan in dat geval een simulatieresultaat is en niet een werkelijke waarneming. De groepsmediaan blijft in het hele bereik van 0–49% onvoltooide zoekduren hetzelfde. Met het oog hierop verdient het presenteren van groepsmedianen eigenlijk de voorkeur boven groepsgemiddelden.

5 Literatuur

Berkhout, P.H.G. en E.S. Mot (1992), De arbeidsmarktrelevantie van opleidingen, OAV-rapport 92-08, Centraal bureau voor de arbeidsvoorziening, Rijswijk

Berkhout, P.H.G. en E.S. Mot (1993), De arbeidsmarktrelevantie van opleidingen: vervolgstudie, OAV-rapport 93-07, Centraal bureau voor de arbeidsvoorziening, Rijswijk

Berkhout, P.H.G. en J.C. Provoost (1994), Schoolverlaters op zoek naar een baan: een duuranalyse, SEO-rapport nr 339, Stichting voor Economisch Onderzoek der Universiteit van Amsterdam

Berkhout, P.H.G. en J.C. Provoost (1995), Schoolverlaters op zoek naar een baan: 1992-1994, SEO-rapport nr 359, Stichting voor Economisch Onderzoek der Universiteit van Amsterdam

Berkhout, P.H.G. (1996), Schoolverlaters op zoek naar een baan: 1992-1995, SEO-rapport nr 381, Stichting voor Economisch Onderzoek der Universiteit van Amsterdam

Hougaard, P. (1986), Survival models for heterogeneous populations derived from stable distributions, *Biometrika* 73, pp. 387-396

Kalbfleish J.D. and R.L. Prentice (1980), *The statistical analysis of failure time data*, John Wiley and Sons, New York

Kiefer, N.M. (1988), Economic duration data and hazard functions, *Journal of Economic Literature* 26, p. 646-679

Lancaster, T. (1990), *The econometric analysis of transition data*, Cambridge University press

Mood, A.M., F.A. Graybill and D.C. Boes, (1974) *Introduction to the theory of statistics*, 3rd edition, McGraw-Hill

Teulings C.N., m.m.v. D. Webbink (1990a), De bepaling van de arbeidsmarktrelevantie van opleidingen, SEO-rapport nr XXX

6 Technische bijlage

Het uitgangspunt van *frailty models* is dat in de specificatie van de uittredevoet multiplicatief een storingsterm wordt opgenomen. Met andere woorden, conditioneel op een onbekende ν , wordt de uittredevoet verondersteld gelijk te zijn aan:

$$\theta(t | \nu) = \nu\theta(t) \quad (\text{A.1})$$

waarbij ν een positieve stochast is met verwachting 1 en variantie σ^2 . De conditionele survival functie kan nu worden geschreven als een functie van de survival functie $S(t)$ corresponderend met $\theta(t)$:

$$S(t | \nu) = \exp\left\{-\int_0^t \nu\theta(u)du\right\} = \exp\left\{-\nu \int_0^t \frac{f(u)}{S(u)} du\right\} = \{S(t)\}^\nu \quad (\text{A.2})$$

De conditionele dichtheidsfunctie is dan gelijk aan:

$$f(t | \nu) = -S'(t | \nu) = \nu f(t) \{S(t)\}^{\nu-1} \quad (\text{A.3})$$

De onconditionele dichtheidsfunctie f_m - met de subscript m van mixture om hem te onderscheiden van de dichtheid corresponderend met $S(t)$ en $\theta(t)$ - kan worden afgeleid door ν met behulp van dichtheid $g(\nu)$ door middel van integratie te elimineren:

$$f_m(t) = \int_0^\infty f(t | \nu)g(\nu)d\nu = \int_0^\infty \nu f(t) \{S(t)\}^{\nu-1} g(\nu)d\nu \quad (\text{A.4})$$

De onconditionele survival functie is per definitie gelijk aan:

$$S_m(t) = 1 - \int_0^t f_m(u)du \quad (\text{A.5})$$

Onder veronderstelling dat storingsterm ν verdeeld is volgens de gamma-verdeling met verwachting 1 en variantie σ^2 , geparametriseerd als $\text{Gamma}(1/\sigma^2, \sigma^2)$, dan is de onconditionele survival functie $S_m(t)$ gelijk aan:

$$S_m(t) = \{1 - \sigma^2 \ln S(t)\}^{-1/\sigma^2} \quad (\text{A.6})$$

Het Burr model (zie ook Lancaster 1990, p. 65-70) ontstaat nu door voor de basis hazardfunctie die van het Weibull model te kiezen. Dat betekent dat $\theta(t) = \alpha \lambda t^{\alpha-1}$, waaruit volgt dat $S(t) = \exp(-\lambda t^\alpha)$ en dat de onconditionele survival functie gelijk is aan:

$$S_m(t) = (1 + \sigma^2 \lambda t^\alpha)^{-1/\sigma^2} \quad \text{waarbij} \quad \lambda = \exp(x'\beta) \quad (\text{A.7})$$

De uittredevoet is

$$\theta_m(t) = \frac{\lambda \alpha t^{\alpha-1}}{1 + \sigma^2 \lambda t^\alpha} \quad (\text{A.8})$$

De verwachting en de mediaan zijn gelijk aan:

$$E_m(T) = \frac{1}{\alpha} \left(\frac{1}{\sigma^2 \lambda} \right)^{\frac{1}{\alpha}} \cdot \frac{\Gamma(1/\alpha) \Gamma(\sigma^{-2} - 1/\alpha)}{\Gamma(\sigma^{-2})} \quad (\text{A.9})$$

$$\text{Mediaan}_m(T) = \left(\frac{2^{\sigma^2} - 1}{\sigma^2 \lambda} \right)^{1/\alpha} \quad (\text{A.10})$$

waarbij $\Gamma(\cdot)$ de gammafunctie³ voorstelt. Schatting van het model wil zeggen dat waarden voor de parameters α , β en σ worden geschat. Dat gebeurt aan de hand van de maximum-likelihood methode. Zelf programmeren van de likelihoodfunctie is mogelijk, maar wel bewerkelijk. Thans zijn veel duurmodellen, zoals ook het hier beschreven model, in sommige statistische pakketten als standaard procedure beschikbaar.

³ Zie bijvoorbeeld; Mood, A.M. et al. (1974), p. 534