

Kijkcijfers

het optimaliseren van reclame-
opbrengsten tijdens SBS-films

door:
Rob Konijn

begeleider:
dr. Sandjai Bhulai



Vrije Universiteit
Amsterdam

in samenwerking met:

TV Audience Solutions

Voor het optimaliseren van kijkcijfers



VERONICA





Voorwoord

Het BWI werkstuk is een van de laatste onderdelen van de studie Bedrijfswiskunde & Informatica. In dit werkstuk is het de bedoeling dat de drie deelgebieden van de BWI studie aan bod komen. In dit werkstuk wordt een (literatuur)onderzoek uitgevoerd over een onderwerp dat te maken heeft met de drie deelgebieden die in de BWI-studie aan bod komen: Bedrijfskunde, Wiskunde en Informatica.

In dit werkstuk zal worden beschreven hoe de omroep SBS haar reclameopbrengsten tijdens filmtuitzendingen kan maximaliseren. Achtereenvolgens zullen in dit werkstuk de meting van kijkcijfers, de samenhang tussen kijkcijfers en reclame-inkomsten, de betrouwbaarheid van kijkcijfers, het voorspellen van kijkcijfers voor films, en het inplannen van deze films aan bod komen.

Graag wil ik Robin Muurling van het bedrijf TVAS bedanken voor het wekken van mijn interesse voor het onderwerp, en voor het beschikbaar stellen van de data. Sandjai Bhulai wil ik graag bedanken voor het meedenken over de aanpak, en zijn enthousiasme tijdens de begeleiding van dit werkstuk.

Rob Konijn,

Januari 2008

Inhoudsopgave

Management Samenvatting.....	7
Probleemstelling	9
Wat zijn kijkcijfers en hoe worden ze gemeten?	11
Historie van de meting van kijkcijfers	11
Meting van de kijkcijfers nu	13
Maar ik ken helemaal niemand met een kastje.....	15
Definities.....	16
De steekproef.....	21
Weging.....	22
Relatie tussen kijkcijfers en reclame-inkomsten.....	25
Opbouw spottarief	25
De basisprijs	26
De doelgroep index.....	26
De product index.....	29
Maandindex	31
Marktindex	31
Spotlengte index	32
Betrouwbaarheid.....	33
Betrouwbaarheidsinterval van een enkele meting	33
Verschil tussen enkele metingen en gesommeerde resultaten	37
Het SCORE model.....	39
Een statistisch model voor het voorspellen van kijkcijfers	43
Het klassieke lineaire model	43
Algemene vorm van Generalized Linear Models (GLM).....	45
De link functie.....	48
Het schatten van de parameters	49
Model Diagnose	50
De Deviantie.....	52
Data analyse.....	57
Schaten van het logistische regressiemodel.....	59
Model diagnostiek en residu analyse logistische regressie model	61
Schaten van het lineaire regressiemodel	63
Residu analyse lineaire regressiemodel	65
Vergelijking MSE twee modellen.....	66
Conclusie en verbeteringen	66
Optimalisatie van reclameopbrengsten door middel van “scheduling”	69
Uitbreiding: Afhankelijkheid en het meerdere keren uitzenden van een film	71
Uitbreiding: verschillende reclamedoelgroepen	72
Reclame-inkomsten als responsvariabele.....	72
Verschillende statistische modellen voor verschillende groepen	72
Het aankoopbedrag van een film.....	75
Conclusie	79
Literatuurlijst	81
Appendix.....	83
Bijlage 1: Verkennende data analyse.....	83
Bijlage 2: Een deel van de R-code gebruikt tijdens het schatten van het statistische model	101
Bijlage 3: De wervingsmatrix.....	104
Bijlage 4: Filmtitels van de geanalyseerde films	107

Management Samenvatting

In deze scriptie wordt onderzocht welke films SBS het beste wanneer kan uitzenden. Het doel hierbij is om de reclameopbrengsten te maximaliseren.

Kijkcijfers worden gemeten aan de hand van een vaste steekproef onder de Nederlandse bevolking. Deze steekproef heeft een omvang van ongeveer 1200 huishoudens en omvat zo'n 2700 personen. De steekproef is samengesteld aan de hand van sociografische achtergrondvariabelen. De kijkdichtheid (kdh) van een programma is gedefinieerd als het gemiddelde percentage kijkers per seconde, gedurende dit programma. De absolute kijkcijfers kunnen worden verkregen door de kijkdichtheid (van een doelgroep) te vermenigvuldigen met het aantal inwoners in Nederland (van diezelfde doelgroep).

Omdat "het optimaliseren van kijkcijfers" en "het optimaliseren van reclameopbrengsten" niet helemaal hetzelfde is, wordt vervolgens wordt de opbouw van het spottarief uitgelegd. Dit is een soort "vertaling" van de kijkcijfers van een spot, naar de opbrengsten van die spot voor SBS.

Nadat alle begrippen en definities zijn uitgelegd, wordt de betrouwbaarheid van de meting van kijkcijfers behandeld. Er wordt uitgelegd hoe je een betrouwbaarheidsinterval kunt construeren voor enkele metingen, en voor meerdere metingen (bijvoorbeeld reclamecampagnes). Voor een "doorsnee" SBS film met een kijkdichtheid van 3%, (450.000 kijkers in de leeftijdscategorie 6 jaar en ouder), moet je denken aan een 95%-betrouwbaarheidsmarge van ongeveer 20%. Dit is uitgedrukt in een interval: [360.000, 540.000].

De reclameopbrengsten tijdens een bepaalde film, kunnen aan de hand van twee statistische modellen worden voorspeld: een statistisch model dat de kijkcijfers voorspelt, en een ander model dat deze kijkcijfers "vertaalt" naar reclameopbrengsten.

De kijkcijfers van een film kunnen goed worden voorspeld aan de hand van de eigenschappen van een film. Uit data analyse blijkt dat voor het voorspellen van kijkcijfers een logistisch regressiemodel beter werkt dan een lineair regressiemodel. Het verschil tussen deze modellen is, kort gezegd, dat het logistische regressiemodel een multiplicatief model is, dat wil zeggen dat de kijkcijfers met een percentage veranderen aan de hand van de eigenschappen van een film. In het geval van het lineaire regressiemodel veranderen de kijkcijfers met een absoluut aantal als een eigenschap wijzigt.

De belangrijkste eigenschappen om de kijkcijfers van een film te voorspellen zijn in volgorde van “belangrijkheid”: het uitzendtijdstip, of de film een actiefilm of avonturenfilm is, het aantal bioscoopbezoekers dat de film had, of de film in de winter (maanden oktober t/m februari) werd uitgezonden, het aantal kijkers naar het voorafgaande programma, en als laatste de Moviemeter rating. Verbeteringen van dit model kunnen worden gemaakt door een beter gevulde database, de toevoeging van concurrentiegegevens (programma's die door de concurrentie werden uitgezonden tijdens de film), en toevoeging van de kijkcijfers van de vorige keer dat een bepaalde film werd uitgezonden.

Het model dat de kijkcijfers “vertaalt” naar de reclameopbrengsten is voor een deel deterministisch. Dit komt doordat de kijkcijfers via een systeem van indexen (vermenigvuldigingsfactoren) worden omgerekend naar een reclameprijs. Hierbij moet je denken aan indices die per tijdstip en per maand verschillen. Het statistische gedeelte van dit model bestaat voornamelijk uit keuzes die de adverteerder maakt qua inkoopopties. Tijdens welke films besluit de adverteerder de “dure” inkoopopties te gebruiken? Helaas waren hiervan bij het maken van deze scriptie geen data aanwezig, en daarom kon dit niet worden onderzocht.

Bij het inroosteren van de films in een uitzendschema, dient rekening worden gehouden te worden met de verschillende reclamedoelgroepen. Dit kan worden gedaan door een inschatting te maken over welk deel van de totale reclamezendtijd voor welke doelgroep zal zijn. Vervolgens dient een statistisch model te worden gemaakt per reclamedoelgroep. Het inroosteren van films kan worden gezien als een “knapzakprobleem” (met uitbreidingen), en hiervoor is in het hoofdstuk “*optimalisatie van reclameopbrengsten door middel van scheduling*” een heuristiek (stappenplan) beschreven.

De laatste vraag die beantwoord is, is hoe je een inschatting kunt maken wat een “goede” aankoopprijs is van een film. Bij een periodeplanning hoort een totale reclameopbrengst in euro's. Het breakeven punt van de aankoopprijs voor een film is de som van het aantal euro's dat een film de periodeopbrengsten, nu en in de toekomst, doet stijgen. Uiteraard geldt: hoe lager de aankoopprijs, hoe beter. De verwachting is dat een film de eerste keer dat hij wordt uitgezonden meer kijkers trekt dan de tweede keer, en de tweede keer meer dan de derde keer etc. Om dit te onderzoeken is een verkennende analyse gedaan aan de hand van 15 films. Dit aantal bleek echter te weinig om conclusies aan te verbinden, dus verder onderzoek zal nodig zijn.

Probleemstelling

Tijdens mijn stage bij Nationale-Nederlanden heb ik (collega) Robin Muurling ontmoet. Hij vertelde me dat hij eigenaar is van het bedrijf TV Audience Solutions (TVAS). Dit bedrijf is gespecialiseerd in het voorspellen van kijkcijfers. Toen ik dit hoorde, leken kijkcijfers mij meteen een interessant onderwerp om ook eens te onderzoeken, en goed geschikt als onderwerp voor mijn werkstuk. Na mijn stage hebben Robin en ik afgesproken om de mogelijkheden te bespreken, en zijn tot het onderwerp “kijkcijfers voor films van SBS” terechtgekomen. Dit is een opdracht van TVAS die nog openstaat van SBS, en heeft dus direct een praktisch nut!

De probleemstelling kan als volgt worden beschreven: De opdracht is om voor de omroep SBS (die uitzendt op de kanalen SBS6, Net5 en Veronica), onderzoek te doen naar welke films het beste, wanneer uitgezonden kunnen worden. SBS heeft in haar programmering op de verschillende zenders een vaste uitzendtijd beschikbaar voor films. De vraag is hoe SBS de beschikbare films waarvan zij de uitzendrechten heeft, het beste kan “verdelen” over deze uitzendtijdstippen (ook wel “slots” genoemd). Hierbij wil SBS de reclame-inkomsten, die worden verdiend tijdens de reclameblokken behorend bij deze films, maximaliseren. De te onderzoeken tijden waarop SBS haar films uitzendt zijn: maandagavond 20.30 (SBS6), dinsdagavond 20.30 (Veronica), zondagavond 20.30 (Veronica), donderdagavond 20.30u (NET5) en vrijdagavond 22.30 (NET5). Een deelvraag is welke films, die recent zijn uitgekomen in 2006, het beste kunnen worden aangekocht door SBS, en wat een goede prijs is voor deze films.

Wat zijn kijkcijfers en hoe worden ze gemeten?

Om het probleem beschreven in de vorige paragraaf te kunnen oplossen, zullen we eerst moeten weten hoe kijkcijfers tot stand komen. In deze paragraaf zal worden uitgelegd hoe kijkcijfers worden gemeten. Maakt het bijvoorbeeld uit voor de kijkcijfers naar welk programma je zelf kijkt? Je leest het in dit hoofdstuk.

Historie van de meting van kijkcijfers

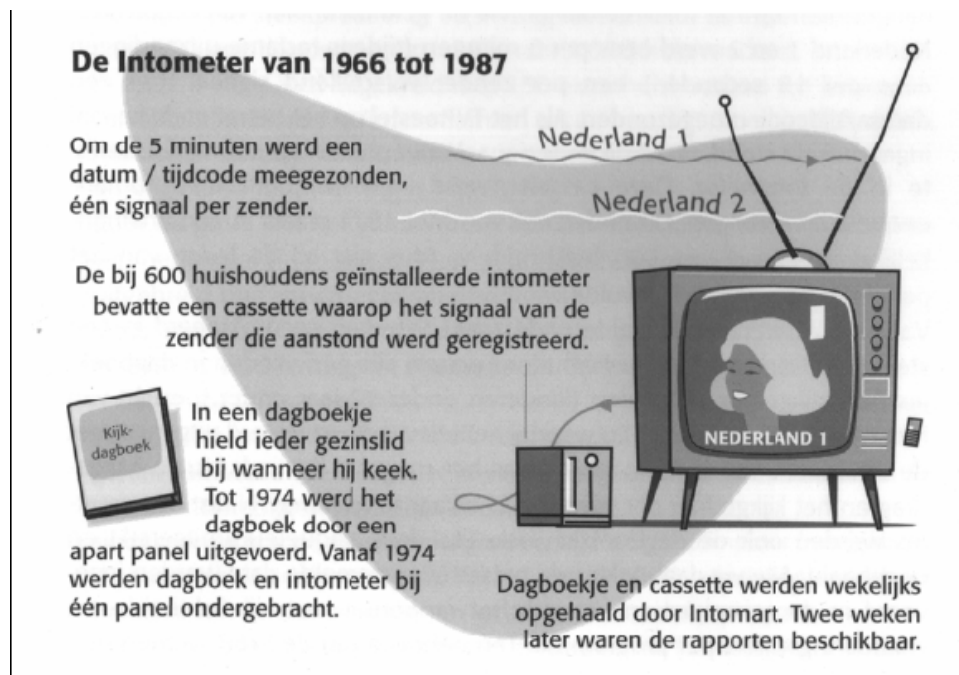
Het kijkonderzoek is in Nederland gestart op 2 januari 1965 door Intomart GfK in opdracht van de Publieke Omroep (voorheen NOS en omroepen). In 1967 is de Ster als opdrachtgever erbij gekomen. Van 1965 tot medio 1987 werd het kijkgedrag gemeten in een panel van 1.500 personen van 12 jaar en ouder dat een dagboek bijhield waarin de televisieprogramma's van een week van de zenders Nederland 1 en 2 stonden vermeld. De panelleden kruisten de programma's aan, die zij voor tenminste de helft hadden gezien en gaven deze programma's een waardering in de vorm van een rapportcijfer. Tot 1974 werden de dagboeken per post heen en teruggezonden (Intomart GfK e.a., 2007).

The image shows a sample of a television viewing diary (kijkcijferdagboek) from Intomart GfK. The diary is open to three pages, showing the layout for recording viewing habits. The pages are dated: **woensdag 20 juni**, **donderdag 19 maart**, and **donderdag 21 maart**. Each page lists television programs for **nederland 1** and **nederland 2**. The programs are listed in a grid format, with columns for **GEZIEN** (seen) and **OOR-DEEL** (share). The programs listed include: **DE BESTE A...**, **VRIENDE...**, **KENNEM...**, **STER-REK...**, **JOU...**, **LAST FRONTIER**, **PABLO**, **DAKTARI**, **KINDERKRANT**, **STER REKLAME**, **JOURNAAL**, **NIEUWS VOOR DOVEN EN SLECHTHORENDE**, **STAATSLOTERIJ**, **AMSTRAAT**, **PASPOORT**, **SESAMSTRAAT**, **JEUGDJOURNAAL**, and **STER REKLAME**. The diary is a physical document with a grid-like structure for recording viewing habits.

Figuur 1: voorbeeld van een kijkcijferdagboek (Intomart GfK e.a., 2007)

Naast de dagboekmeting op persoonsniveau vond vanaf 1966 ook een aanvullende meting plaats op toestelniveau. Een door Philips dochter (Electrologica) speciaal hiertoe voor Intomart GfK ontwikkelde meter (de kijkmeter) werd aangesloten op het televisietoestel. Als de tv aanstond, werd op een cassettebandje een signaal opgeslagen dat aangaf naar welke zender werd gekeken. Vanaf 1974 werden de beide onderzoekssystemen gecombineerd in één steekproef van circa 600 huishoudens waarin alle gezinsleden in dagboekjes hun kijkgedrag bijhielden (kinderen onder 12 jaar onder toezicht van hun ouder of verzorger) en waarbij het televisietoestel werd uitgerust met een kijkmeter (Intomart GfK e.a., 2007).

In figuur 2 wordt beschreven hoe dit in zijn werk ging.



Figuur 2: Hoe kijkcijfers tot 1987 werden bepaald (Intomart GfK e.a., 2007)

Vanaf 1987 werden de kijkcijfers op een andere manier gemeten. Voor het nieuwe onderzoek werd gekozen voor de 4900-Persoonsmeter. Met deze 4900-meter (zie figuur 3) konden alle te ontvangen zenders, het videokijkgedrag en waarderingen van programma's worden gemeten, de medewerking was voor de panelleden veel minder belastend dan het invullen van een dagboekje en de resultaten waren de ochtend na uitzending direct online beschikbaar (Intomart GfK e.a., 2007).



Figuur 3: de 4900-Persoonsmeter (Intomart GfK e.a., 2007)

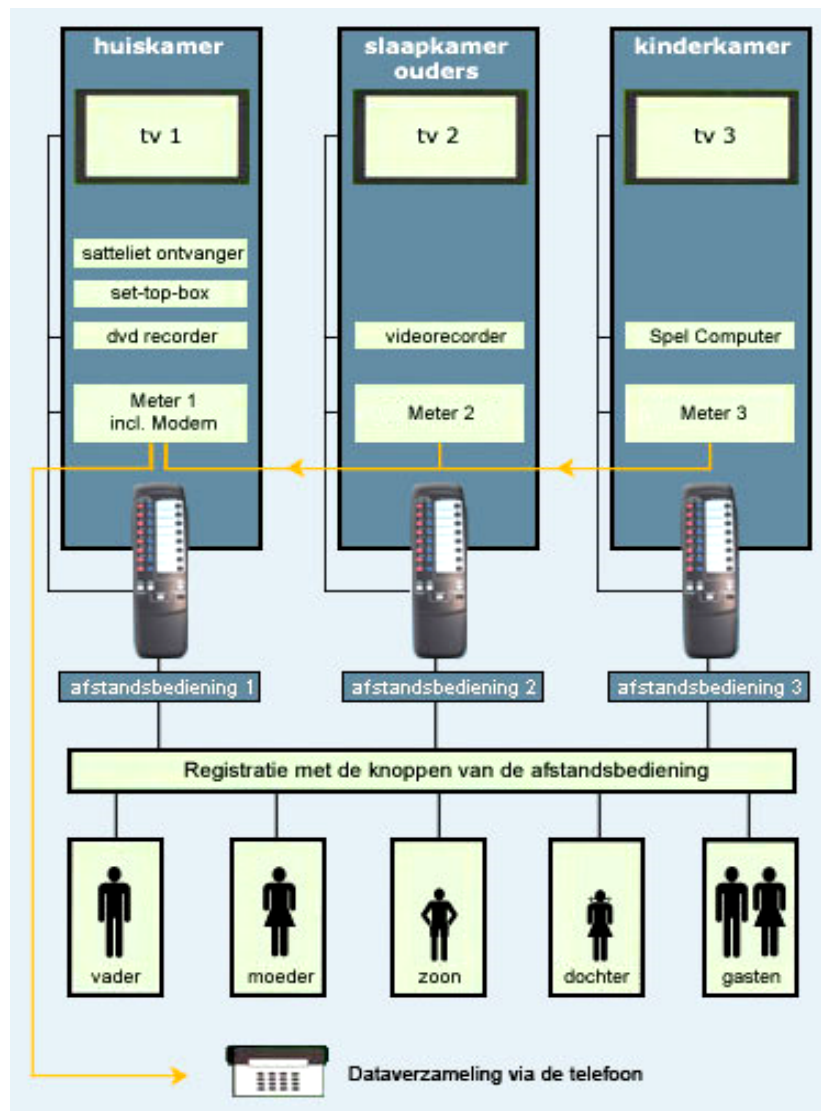
Sinds 1987 is er veel veranderd in de televisiewereld. Er zijn veel (commerciële) televisiezenders bij gekomen. Ook zijn de kijkcijfers “belangrijker” geworden. Dit komt doordat reclame-inkomsten grotendeels de enige inkomsten zijn van de commerciële zenders, en deze hangen af van de kijkcijfers. Hoe de reclame-inkomsten precies afhangen van de kijkcijfers wordt beschreven in de paragraaf “Relatie tussen kijkcijfers en reclame-inkomsten”. Omdat het belang van kijkcijfers steeds groter is geworden, is ook de steekproef uitgebreid om de betrouwbaarheid van de kijkcijfers te vergroten. De steekproef is opgehoogd van 700 huishoudens naar 850 huishoudens in 1991, vervolgens naar 1000 huishoudens in 1995, en naar 1250 huishoudens in 1999 (Intomart GfK e.a., 2007). Over de betrouwbaarheid en validiteit van de steekproef is meer te lezen in het hoofdstuk “betrouwbaarheid”.

Meting van de kijkcijfers nu

Sinds 2002 tot nu worden de kijkcijfers gemeten door de Stichting Kijk Onderzoek (SKO). Op dit moment is de panelomvang 1220 huishoudens. In deze huishoudens wonen in totaal 2700 personen (Stichting Kijk Onderzoek, 2007).

De kijkcijfers worden sinds 1987 gemeten met behulp van een kastje (zie figuur 3). Elk huishouden dat meedoet aan het kijkonderzoek krijgt op alle televisietoestellen en randapparatuur een 'kastje', een kijkmeter. De kijkmeter bestaat uit een display waarop vragen en instructies te lezen zijn, zoals 'Wie kijkt?'. Daarnaast bevat de meter een afstandsbediening met onder meer aparte persoonsknoppen voor elk lid van het huishouden. De kijkmeter kan en slaat deze informatie op. De kijkmeter legt vast wat er in het televisietoestel gebeurt, maar kan niet 'zien' wat zich in de kamer afspeelt en wie er naar de televisie kijken. Daarom moeten alle personen van 3 jaar en ouder die kijken zich aanmelden. Dat aanmelden gebeurt door de eigen persoonsknop op de afstandsbediening in te drukken. De display meldt dan wie er kijkt. Zodra

iemand is aangemeld, slaat de kijkmeter het televisiekijkgedrag van deze persoon op. Stopt diegene met kijken, dan moet hij of zij zich afmelden. Elke nacht worden alle geregistreerde gegevens van de voorgaande dag doorgestuurd aan de centrale computer bij Intomart GfK te Hilversum. Dit gebeurt automatisch via een modem en de telefoon, zonder dat de panelleden dit merken (Stichting Kijk Onderzoek, 2007). Zie ook figuur 4 voor een beschrijving van hoe dit in zijn werk gaat.



Figuur 4: Het meten van kijkcijfers in een huishouden (Stichting Kijk Onderzoek, 2007a)

Zoals in figuur 4 is weergegeven, wordt ook het gebruik van videorecorders en spelcomputers geregistreerd. Ook tellen alle tv toestellen in het huishouden mee voor de kijkcijfermeting.

De fysieke “registratie” van de zender waarnaar wordt gekeken gebeurt op meerdere manieren: De eerste manier heet Picture Matching. Picture Matching is een systeem dat continu samples neemt van het beeld dat op de televisie verschijnt en de kenmerken van die samples digitaal vastlegt. De zodanig verkregen informatie wordt ’s nacht verzonden naar de Intomart GfK-computers. Daar worden de gegevens vergeleken met dezelfde informatie, opgenomen van de belangrijkste zenders op zogenaamde reference sites. Vergelijking van deze samples door matching bepaalt dan naar welke zender is gekeken. Op deze wijze kan het gehele kijktraject van de voorgaande dag worden gereconstrueerd. Het gehele proces wordt dagelijks uitgevoerd en de gegevens zijn, zoals gebruikelijk, de volgende morgen beschikbaar. (Intomart GfK e.a., 2007)

De Picture Matching techniek werkt niet altijd goed, vooral als beelden erg op elkaar lijken zoals bij internationale voetbalwedstrijden, of als een uitzending op meerdere zenders tegelijk wordt uitgezonden zoals Koninginnedag of belangrijk nieuws. In het geval het niet duidelijk is, wordt ook de programmacode met elkaar vergeleken. Ook is er recent een derde zenderherkenningstechniek operationeel in de kijkmeter van het kijkonderzoek: Enhanced Audio Matching. Het betreft een techniek die gebruik maakt van het matchen van audio-samples voor het herkennen van zenders. Deze techniek lijkt op Picture Matching en is gebaseerd op een (voor het menselijke oor onhoorbare) geluidscode. (Intomart GfK e.a., 2007)

Bij het afspelen van een zelf opgenomen videoband wordt de datum, het tijdstip en de zender van de opname herkend. Alle uitzendingen die binnen 7 dagen na het moment van uitzending op normale snelheid worden bekeken (dus exclusief snelspoelen) worden in een aparte rapportage als uitgesteld kijkgedrag gerapporteerd. Dit uitgestelde kijkgedrag naar zelf opgenomen videobanden is als een aanvullend bestand bij de ruwe beschikbaar. Al het kijkgedrag naar videobanden (al dan niet zelf opgenomen) is in de dagelijkse rapportering samengevoegd onder de noemer “video”. (Intomart GfK e.a., 2007)

Maar ik ken helemaal niemand met een kastje

Tijdens deze scriptie is met meerdere mensen gesproken die vrijwel allemaal dezelfde opmerking hadden over het meten van kijkcijfers met de kastjes: maar ik ken helemaal niemand die zo’n kastje heeft! In tabel 1 worden de kansen berekend dat je niemand kent met een kastje, aan de hand van het aantal huishoudens dat je kent.

aantal huishoudens dat je kent	kans dat 1 persoon geen huishouden kent met een kastje	kans dat 10 personen geen huishouden kennen met een kastje
25	0,998	0,980
50	0,996	0,961
100	0,992	0,923

Tabel 1 : kans dat je niemand kent die een kijkcijfer kastje heeft

In tabel 1 is de linker kolom het aantal huishoudens dat je goed genoeg kent om hun “tv-situatie” thuis te weten. Dit ligt misschien eerder rond de 25 dan rond de 100. In ieder geval is de kans groot dat je niemand kent met een kastje.

Definities

Nadat de data gemeten zijn, worden per uitgezonden programma of reclamespot, standaard enkele berekeningen uitgevoerd. In deze paragraaf worden enkele veel voorkomende begrippen uitgelegd.

De kijkdichtheid

De kijkdichtheid (kdh) van een programma is gedefinieerd als het gemiddelde percentage kijkers per seconde, gedurende dit programma. De respondenten in de steekproef worden daarbij gewogen met een gewicht bepaald door de mate waarin ze voorkomen in de totale populatie van kijkers in Nederland. Meer over deze weging staat in de paragraaf: “de steekproef”.

$$kdh_i = \frac{\sum_{r=1}^R w_r p_{r,i}}{\sum_{r=1}^R w_r} \cdot 100\%$$

1)

(Intomart GfK e.a., 2007)

Hierin is:

Kdh_i : de kijkdichtheid voor programma i

R: aantal respondenten in de steekproef

W_r : gewicht van respondent r

$P_{r,i}$: kijkkans van respondent r voor programma i

Met de kijkkans wordt bedoeld: het aantal secondes dat een respondent heeft gekeken naar het programma, gedeeld door het totale aantal secondes dat het programma duurt.

De minuut waarin een programma begint en eindigt, wordt afgekapt op hele minuten. Wanneer een programma X bijvoorbeeld begint om 18.00.40 uur en eindigt om 18.15.50 uur, dan wordt de kijkdichtheid van programma X berekend door het gemiddelde percentage kijkers per seconde te berekenen van 18.00.00 uur tot en met 18.14.59 uur.

Bovenstaande maatregel is belangrijk bij het bepalen van de kijkdichtheid van een spot, omdat een spot vaak een korte tijd duurt. De kijkdichtheid van een spot is gedefinieerd als het gemiddelde percentage kijkers per seconde, gedurende de lengte van de spot. Deze wordt ook berekend met formule 1), maar omdat de berekeningen “per minuut” gedaan worden zal de berekening van een spot meestal 1 minuut betreffen. Wanneer een spot van 30 seconden bijvoorbeeld begint om 20.25.40 uur en eindigt om 20.26.10 uur, dan wordt de kijkdichtheid van de spot berekend door het gemiddelde percentage kijkers per seconde te berekenen voor de minuut van 20.25.00 uur tot en met 20.25.59 uur.

Het marktaandeel

Het marktaandeel is gedefinieerd als het percentage kijkers naar een uitzending of tijdvak, gebaseerd op het totale kijkerspubliek. Het marktaandeel wordt berekend door de kijkdichtheid van een programma of tijdvak door de kijkdichtheid totaalzenders gedurende het betreffende programma c.q. tijdvak te delen.

Een Gross Rating Point (GRP)

Een Gross Rating Point is 1% kijkdichtheid binnen een bepaalde doelgroep. De term wordt vooral gebruikt als men spreekt over de kijkdichtheid van spots en commercials. Dit begrip zal vaak terugkomen in het hoofdstuk “relatie tussen kijkcijfers en reclame-inkomsten” (SBS, 2007).

Het bereik

Bij het meten van het bereik wordt onderscheid gemaakt tussen het programmabereik en het spotbereik.

Het programmabereik

Het programmabereik is gedefinieerd als het percentage kijkers dat minimaal één keer voor een bepaalde tijd naar één programma van een reeks programma's heeft gekeken. Een eenmalig programma wordt daarbij als een speciaal geval van een reeks programma's beschouwd. Als ondergrens voor het programmabereik geldt (als conventie in de markt) dat men tenminste 60 seconden aaneengesloten naar een programma moet hebben gekeken.

In de formule 2) kan bepaald worden of iemand naar één of meerdere programma's moet hebben gekeken om in het programmabereik mee te tellen, hoe lang hij ernaar moet hebben gekeken, en of dit aaneengesloten was of niet. Hoe lang iemand naar één of meerdere programma's moet hebben gekeken kan óf in minuten óf als percentage worden opgegeven.

$$bereik_I = \frac{\sum_{r=1}^R w_r d_{r,I}}{\sum_{r=1}^R w_r} \cdot 100\%$$

$$d_{r,I} = \begin{cases} 1 & j_{r,I} \geq J \\ 0 & j_{r,I} < J \end{cases}$$

2)

- I : reeks van programma's
- $d_{r,I}$: 'dummy' die aangeeft of respondent volgens specifieke definitie in het programmabereik valt
- J : criterium voor het specifieke programmabereik
- R : aantal respondenten in de steekproef
- w_r : gewicht van respondent r
- $j_{r,I}$: volgens de specifieke voorwaarden gecalculeerd persoonlijk criterium van respondent r voor programmareeks I

(Intomart GfK e.a., 2007)

Het spotbereik

Zoals eerder vermeld, is er veel geld gemoeid met reclamecontracten. Daarom is er ook een standaard berekening voor het spotbereik, die samenhangt met de berekening van de kijkdichtheid. (Intomart GfK, 2007) Deze definitie wordt in de gehele markt gebruikt. In de definitie van kijkdichtheid wordt rekening gehouden met "gedeeltelijk kijkgedrag". Niet slechts individuen

die een geheel traject (programma, kwartier of spot) hebben gekeken, maar elk waargenomen kijkmoment wordt betrokken in de berekeningen. Hierin is de lengte van “een kijkmoment” een seconde. De consequentie van deze denkwijze is dat niet meer wordt gedacht op zgn. dichotome wijze - dat wil zeggen van twee mogelijkheden: wel of niet - maar in een (vrijwel) continue schaal van mogelijkheden. (Stichting KijkOnderzoek, 2007b)

Het spotbereik is zo gedefinieerd dat:

- Het spotbereik, vermenigvuldigd met de gemiddelde contactfrequentie het aantal GRP's dient op te leveren, dat op haar beurt weer de som is van de spotkijkdichtheden (i.e. de kijkdichtheid van de reclameminuut)
- Door het dichotome karakter van het begrip contact - er is wel of geen contact geweest – een contactfrequentie een geheel getal is.

Het spotbereik kan worden verkregen door de formule 1) in te vullen per seconde, en hiervan het gemiddelde te nemen.

De contactfrequentieverdeling

De contactfrequentieverdeling is gedefinieerd als verdeling van de contacten op basis van bereik, uitgedrukt in een percentage personen per aantal contacten (Intomart GfK e.a., 2007).

$$f(c, K) = \frac{\sum_{r=1}^R w_r f_r(c, K)}{\sum_{r=1}^R w_r} \cdot 100\% \quad \text{voor } c = 0, 1, \dots, K \quad 3)$$

$f(c, K)$: percentage personen dat c contacten had op basis van K uitzendingen

w_r : gewicht van respondent r

$f_r(c, K)$: kans dat respondent r c contacten had op basis van K uitzendingen

Waarbij $f_r(c, K)$ wordt gegeven door de recursieve relatie

$$f_r(c, k) = f_r(c-1, k-1) \cdot p_{r,k} + f_r(c, k-1) \cdot (1 - p_{r,k})$$

$f_r(c, k)$: kans dat respondent r c contacten had na k uitzendingen

$p_{r,k}$: 'contact' kans van respondent r voor uitzending k

met startwaarden

$$f_r(1, 1) = p_{r,1}$$

$$f_r(0, 1) = 1 - p_{r,1}$$

De contactfrequentieverdeling kan voor programma's en spots berekend worden.

Gemiddelde contactfrequentie

De gemiddelde contactfrequentie is gedefinieerd als het gemiddelde aantal uitzendingen (contacten) dat de “mensen die hebben gekeken” heeft bereikt.

$$\bar{c} = \frac{\sum_{c=1}^K c \cdot f(c, K)}{\sum_{c=1}^K f(c, K)}$$

4)

$\bar{c}$: gemiddeld aantal contacten van de respondenten die minimaal één contact hadden
 $f(c, K)$: percentage personen dat c contacten had na K uitzendingen
(Intomart GfK e.a., 2007)

Gemiddelde contactfrequentie voor spots

De gemiddelde contactfrequentie van spots is ook op een tweede manier te berekenen omdat de contactfrequentie wordt berekend op basis van “het kansmodel”. Het kansmodel houdt in dat per seconde wordt gemeten, en vervolgens het gemiddelde wordt genomen.

$$\bar{c} = \frac{\sum_{c=1}^K c \cdot f(c, K)}{\sum_{c=1}^K f(c, K)} = \frac{\sum_{i=1}^K kdh_i}{\sum_{c=1}^K f(c, K)}$$

5)

$\bar{c}$: gemiddeld aantal contacten van de respondenten die minimaal één contact hadden
 $f(c, K)$: percentage personen dat c contacten had na K uitzendingen
 kdh_i : kijkdichtheid voor uitzending i
(Intomart GfK e.a., 2007)

De kijkdichtheden in de formule moeten worden berekend op basis van de periodesteekproef indien de uitzendingen meer dan één dag beslaan. Dit in tegenstelling tot de gerapporteerde

kijkdichtheden voor de afzonderlijke uitzendingen: deze worden op basis van de dagsteekproef voor de betreffende uitzending berekend.

De kijktijd

De kijktijd is gedefinieerd als de gemiddelde tijd die kijkers naar een programma of binnen een tijdvak naar een zender hebben gekeken, uitgedrukt in minuten.

$$kt_t = \frac{\sum_{r=1}^R w_r K_{r,t}}{\sum_{r=1}^R w_r}$$

6)

- kt_t : kijktijd in tijdvak t
 R : aantal respondenten in de steekproef
 w_r : gewicht van respondent r
 $K_{r,t}$: kijkduur van respondent r in tijdvak t

(Intomart GfK e.a., 2007)

De steekproef

De werving van huishoudens wordt gedaan op basis van een wervingsmatrix van maximaal 100 verschillende cellen, waar mogelijk uitgebreid met extra criteria. Elke cel in de wervingsmatrix beschrijft andere kenmerken van een huishouden/panel. Deze kenmerken zijn: regio, gezinscyclus, werkzaamheid en opleiding. Deze kenmerken zijn gebaseerd op “de gouden standaard”. Volgens deze gouden standaard is de indeling naar sociale klasse A, B1, B2, C en D weergegeven in tabel 2.

Indeling Sociale Klasse Gouden Standaard. (Per 01-01-2007 in Basis Ondervraging TV Panel).							
Hoogst genoten opleiding hoofdkostwinner:							
Beroep Hoofdkostwinner:	Lager Algemeen	Lager Beroeps	MAVO	MBO	HAVO/VWO	HBO	WO
Eigenaar bedrijf 10+	C	B2	B1	A	A	A	A
Eigenaar bedrijf 9-	C	B2	B1	A	A	A	A
Boer/tuinder	C	C	B2	B1	B1	A	A
Vrije beroepen	C	C	B2	B1	A	A	A
Hoger beroeps; leidinggevend	C	C	B2	B1	B1	B1	A
Hoger beroeps; niet-leidinggevend	C	C	B2	B1	B1	B1	A
Middelbaar beroeps; leidinggevend	C	C	B2	B1	B1	B1	A
Middelbaar beroeps; niet-leidinggevend	C	C	B2	B2	B1	B1	A
Elementair en Lager beroepsniveau	C	C	C	C	B2	B2	B1
VUT/Pensioen	D	C	B2	B1	B1	A	A
Werkloos/Arbeidsongeschikt/Bijstand	D	C	C	C	C	B2	B1
Student/overig	D	D	D	C	C	B2	B1

Tabel 2: indeling van sociale klassen volgens de gouden standaard (Stichting Kijkonderzoek 2007c)

Behalve cellen gebaseerd op deze kenmerken is er ook een aparte “allochtonencel” in de wervingsmatrix opgenomen, omdat allochtonen bijna niet voorkomen in de andere cellen in de steekproef. Dit komt omdat de respons om mee te doen aan het kijkonderzoek erg laag is onder allochtonen. De wervingsmatrix is te vinden in het appendix.

Weging

De eerste stap in de calculatie van de kijkcijfers is de dagelijkse weging van de steekproef. Weging is noodzakelijk als de verdeling van de steekproef op essentiële achtergrondvariabelen niet overeenkomt met de onderzoekspopulatie. Het panel is immers geen vaste groep personen, maar verandert per dag enigszins van samenstelling ten gevolge van opzeggingen, nieuwe aansluitingen en eventuele technische storingen. Door middel van weging van de steekproef kunnen de schommelingen in de panelsamenstelling ten opzichte van een aantal populatiegegevens worden gecompenseerd. Weging vindt elke nacht plaats na het binnenhalen en valideren van de kijkgegevens. In de weegprocedure wordt het panel voor een aantal variabelen gewogen naar de steekproefeis. Deze eis wordt jaarlijks opnieuw vastgesteld en is vanaf 2006 gebaseerd op de populatiegegevens uit de MOA Gouden Standaard (GfK Nederland). De omvang van de populatie wordt vastgesteld op basis van trendcijfers van het CBS (Intomart GfK, 2007).

De maximale weegfactor die is toegestaan, bedraagt 3. De som van alle weegfactoren is gelijk aan de totale populatiegrootte van alle personen van 3 jaar en ouder in Nederland.

In formule 7 staat hoe de weegfactoren worden berekend.

$$w_r = \left(\frac{\sum_{d=1}^D \sum_{r=1}^{R_d} \frac{w_{r,d}}{D}}{\sum_{d=1}^D \sum_{r=1}^{N_d} \frac{w_{r,d}}{D}} \right) \cdot \frac{\sum_{d=1}^D w_{r,d}}{D} = \left(\frac{\sum_{d=1}^D \sum_{r=1}^{R_d} w_{r,d}}{\sum_{d=1}^D \sum_{r=1}^{N_d} w_{r,d}} \right) \cdot \frac{\sum_{d=1}^D w_{r,d}}{D}$$

7)

- w_r : Periode weegfactor voor respondent r
- $w_{r,d}$: Weegfactor voor respondent r op dag d
- D : Aantal dagen in de periodesteekproef
- R : Totaal aantal respondenten in de dagsteekproef
- N : Aantal respondenten meegenomen in de periodesteekproef

(Intomart GfK e.a., 2007)

In formule 7 staat dat de weegfactor van elke respondent uit de periodesteekproef wordt berekend door middeling van de daggewichten van de respondent. Als deler geldt het totaal aantal dagen van de periodesteekproef. Hierbij gaat het dus niet alleen om de dagen waarop de betreffende uitzendingen vallen. De gemiddelde weegfactor wordt vermenigvuldigd met een correctiefactor om te corrigeren voor het wegvallen van respondenten in de steekproefperiode (zodat de som van de gewichten gelijk blijft aan de totale populatie). De correctiefactor is gelijk aan de som van de daggewichten van alle respondenten, ook de uiteindelijk niet geselecteerde respondenten, voor de betreffende dag en van alle dagen in de periodesteekproef, gedeeld door de som van de gewichten van alle geselecteerde respondenten in de periodesteekproef.

Relatie tussen kijkcijfers en reclame-inkomsten

We weten nu hoe de kijkcijfers worden gemeten en zijn samengesteld. Echter, de probleemstelling is niet om zoveel mogelijk kijkers te bereiken, maar om de reclame-inkomsten te maximaliseren. Natuurlijk hebben deze twee dingen met elkaar te maken, maar dat ze niet precies hetzelfde zijn wordt in dit hoofdstuk uitgelegd.

Opbouw spottarief

De tarieven van tv-reclamezendtijd worden uitgedrukt in een bedrag per spot, op basis van een lengte van 30 seconden. Dit tarief (op basis van 30 seconden) kan variëren van € 100 bij een reclameblok in de middag tot meer dan € 25.000 bij een belangrijke voetbalwedstrijd, afhankelijk van het aantal kijkers (STER, november 2007a).

In de reclamewereld gaat dus veel geld om. Enkele getallen: de grootste drie adverteerders zijn Unilever (242 miljoen euro), de Rijksvoorlichtingsdienst (159 miljoen), KPN (151 miljoen). De meest geadverteerde merknamen zijn: 1. KPN (48 miljoen) 2. Nivea (39 miljoen) en 3. L'Oréal Paris (36 miljoen). De top 3 bedrijfstakken bestaat uit 1. Auto's, 2. Banken en 3. Supermarkten (SPOT, 2007).

Omdat de probleemstelling de zenders sbs6, NET5 en Veronica betreft, zullen we bij de opbouw van het spottarief voornamelijk de opbouw van het spottarief bij SBS bespreken. Indien er grote verschillen zijn in de opbouw van het tarief met andere omroepen, zullen deze worden aangegeven.

Schematisch overzicht van de opbouw van een spottarief



Figuur 5 Opbouw van een spottarief (SBS, 2007)

Deze opbouw is exact hetzelfde voor RTL en vergelijkbaar voor de STER.

De basisprijs

In figuur 5 wordt gestart met de basisprijs. Over deze basisprijs wordt onderhandeld door de adverteerder en SBS. Om een indicatie te geven voor de hoogte basisprijs kan het publicatietarief worden gebruikt. In november 2007 was dit tarief 1659,86 euro per 30 seconden bij SBS, ongeveer 1200 euro bij RTL (hangt af van het totale budget), en 1000 euro bij de STER. Hierbij moet wel vermeld worden dat veel reclames korter duren, en dat de prijs niet lineair afloopt met het aantal seconden, zie hiervoor de spotlengte index die later besproken wordt.

De basisprijs is per 30 seconden, en per Gross Rating Point. Een Gross Rating Point staat voor 1% van een doelgroep in Nederland. De “standaard” doelgroep zijn personen van 20 tot 49 jaar. De behaalde GRP's worden bepaald met behulp van de kijkcijfers die gepubliceerd worden door Stichting KijkOnderzoek. Deze cijfers zijn op de minuut nauwkeurig. Voor reclameblokken bij SBS geldt dat bij de berekening voor de kijkcijfers van een enkele spot, het gemiddelde van de minuten dat het gehele reclameblok telt als kijkcijfer voor elke reclame in het betreffende reclameblok.

Alleen bij de STER is er ook de mogelijkheid om spots per stuk in te kopen (er kan in dat geval worden gekozen in welk reclameblok de spot wordt uitgezonden), en staat er per blok aangegeven hoeveel een reclamespot van 30 seconden kost. Voor maandag 3 december 2007 lopen de bedragen bijvoorbeeld van 150 euro om 2 uur 's middags (tussen de programma's “Tweenies” en “Dokter Hond”) tot 6140 euro om 20:20 (tussen “de Wereld draait door” en “Holland sport”). Ter vergelijking met programma's met zeer hoge kijkcijfers: Bij een Champions League wedstrijd met Nederlandse club, 13 december (PSV-Internazionale) is 20.000 euro per 30 seconden, terwijl de WK finale 2006, met Nederland in de finale, 100.000 euro had gekost. (STER, november 2007b)

De doelgroep index

Zoals boven vermeld, is de “standaard doelgroep” alle personen tussen de 20 en 49 jaar. Deze heeft dan ook index 100. De standaarddoelgroepen om te adverteren bij SBS zijn weergegeven in tabel 3.

Doelgroep		Index
Man/Vrouw	20-34 jaar	100
Man	20-34 jaar	102
Vrouw	20-34 jaar	96
Man/Vrouw	20-49 jaar	100
Man	20-49 jaar	101
Vrouw	20-49 jaar	97
Boodschappers	20-49 jaar	95
Boodschappers	20-34 jaar	95
Boodschappers	+ kind	97

Tabel 3: Verschillende reclamedoelgroepen bij SBS6 (SBS,2007).

RTL heeft nog enkele andere standaarddoelgroepen, deze zijn weergegeven in tabel 4.

35-49 jaar	97
15-25 jaar	111
20-49 jaar AB1	105
Mannen 20-49 jaar AB1	106

Tabel 4: verschillende doelgroepen voor RTL (RTL, 2007).

Bij RTL is het mogelijk om op afspraak andere doelgroepen overeen te komen, deze hebben een andere index. Alle doelgroepen die gemeten en gepubliceerd worden door Stichting Kijk Onderzoek, met de overeenkomstige populatie in Nederland zijn weergegeven in tabel 5.

Omschrijving	Afkorting	Populatie(x1000), inwoners in NL
Totaal 13 jaar en ouder	13+	13.551
3-12 jaar	39785	1.993
6-12 jaar	39788	1.380
13-19 jaar	13-19	1.377
20-34 jaar	203-4	3.238
35-49 jaar	35-49	3.799
50-64 jaar	50-64	2.999

65 jaar en ouder	65+	2.138
20-49 jaar	2049	7.037
Mannen 13 jaar en ouder	Man	6.689
Vrouw 13 jaar en ouder	Vrouw	6.862
Sociale klasse A, 13 jaar en ouder	A 13+	2.164
Sociale klasse B1, 13 jaar en ouder	B1 13+	4.379
Sociale klasse B2, 13 jaar en ouder	B2 13+	2.392
Sociale klasse C, 13 jaar en ouder	C 13+	3.904
Sociale klasse D, 13 jaar en ouder	D 13+	712
Sociale klasse AB1, 13 jaar en ouder	AB1 13+	6.543
Boodschapper, 13 jaar en ouder	BDS 13+	7.146
Boodschapper, 20-49 jaar	BDS20-49	3.793
Boodschapper met kind van 0 t/m 17 jaar	BDS+knd	1.888
Personen 6 jaar eo in huishouden met kinderen t/m 17 jaar	Huish+knd	6.352
Business to Business	BtoB	1.154
Man 20-34 jaar	M2034	1.631
Vrouw 20-34 jaar	V2034	1.607
Man 35-49 jaar	M3549	1.923
Vrouw 35-49 jaar	V35-49	1.876
Man 50 jaar en ouder	M50+	2.431
Vrouw 50 jaar en ouder	V50+	2.706
Sociale klasse AB1, 20-34 jaar	2034 AB1	1.598
Sociale klasse B2CD, 20-34 jaar	2034 B2CD	1.640
Sociale klasse AB1, 35-49 jaar	3549 AB1	1.959
Sociale klasse B2CD, 35-49 jaar	3549 B2CD	1.840
Sociale klasse AB1, 50 jaar en ouder	50+AB1	2.365
Sociale klasse B2CD, 50 jaar en ouder	50+B2CD	2.772
Sociale klasse AB1, Man 20-49 jaar	M2049AB1	1.795
Sociale klasse AB1, Vrouw 20-49 jaar	V2049AB1	1.762
Sociale klasse B2CD, Man 20-49 jaar	M2049B2CD	1.759
Sociale klasse B2CD, Vrouw 20-49 jaar	V2049B2CD	1.721
Totaal 6 jaar en ouder (inclusief gasten)	6+ inclusief	14.931

Tabel 5: Alle doelgroepen die worden gemeten bij het kijkonderzoek. (Stichting KijkOnderzoek, 2007d)

In tabel 5 staan alle doelgroepen waarvoor de kijkcijfers worden gemeten. Voor de uitleg van de sociale klassen verwijzen we naar paragraaf “de steekproef” waarin in tabel 2 de sociale klassen zijn weergegeven. In de laatste kolom staat het aantal inwoners van Nederland dat tot de genoemde doelgroep behoort.

Voor de doelgroep geldt dat deze wel moet worden goedgekeurd door de omroep waarvoor de reclame wordt uitgezonden. Een sluwe adverteerder van auto's zou anders kunnen zeggen dat zijn reclame voor oudere vrouwen (V50+) bedoeld is, en zijn reclames tijdens voetbalwedstrijden uitzenden. Op deze manier zouden weinig GRP's worden verrekend, terwijl er wel veel mannen (de eigenlijke doelgroep) worden bereikt. Voor RTL geldt zelfs een restrictie dat tijdens formule 1 wedstrijden alleen maar reclames voor de doelgroep “man” mogen worden uitgezonden.

De product index

Er bestaan verschillende inkoopopties voor reclames. Grofweg kunnen deze inkoopopties worden gesplitst in pakketten en ‘eigen inkoop’. Bij een pakket koopt een adverteerder een aantal GRP's van een bepaalde doelgroep. SBS deelt vervolgens de reclames in, en de reclames worden uitgezonden totdat het aantal GRP's is gehaald. Er zijn meerdere soorten pakketten te koop, elk pakket heeft andere kenmerken, zoals de uitzendtijden waarop de reclames worden uitgezonden. Een overzicht van deze pakketten staat in figuur 6.

Ook bestaat er de mogelijkheid tot eigen inkoop, bij SBS “Fixed Price” genoemd. Hierbij kan de adverteerder zelf aangeven in welke blokken de reclames uitgezonden dienen te worden. Bij SBS werkt het als volgt: de adverteerder kiest ongeveer 200% van het aantal GRP's aan reclamezendtijd. Uit deze 200% maakt SBS een selectie van blokken waarin de reclame daadwerkelijk wordt uitgezonden.

Een speciaal pakket, dat voor deze scriptie erg van belang is omdat het over kijkcijfers van films gaat, is het filmpakket. Maandelijks wordt door SBS de index voor dit filmpakket bekend gemaakt. Voor november 2007 is deze index 114. Filmpakketten kunnen alleen ingekocht worden voor de doelgroepen 20-34 en 20-49 jaar. Bij een filmpakket wordt de spot uitsluitend ingedeeld in blokken vóór, tijdens of na de speelfilms van SBS 6, NET5 en Veronica. De minimum looptijd van het filmpakket is 14 dagen, het minimum aantal GRP's 15.

	Sturing/Optimaliseren toegestaan	Product-index	Inkoop	Vast Schema	Budget vast op campagneniveau
KWALITATIEF					
Fixed Budget*	Bloksturing o.b.v. selectieve blokselectie (200%)	118	NET 5, SBS 6, Veronica	Ja	Ja
Fixed Price NET 5*	Bloksturing o.b.v. selectieve blokselectie (200%)	113	Per zender	Ja	Nee***
Fixed Price SBS 6*	Bloksturing o.b.v. selectieve blokselectie (200%)	110	Per zender	Ja	Nee***
Fixed Price Veronica*	Bloksturing o.b.v. selectieve blokselectie (200%)	108	Per zender	Ja	Nee***
Filmpakketten**	n.v.t.	per maand	NET 5, SBS 6, Veronica	Nee	Ja
KWANTITATIEF					
Super Prime Pakket	Tijdvaksturing 19.30-23.30 u	103	NET 5, SBS 6, Veronica	Nee	Ja
Super Prime Pakket X-tra		107		Nee	Ja
Prime Pakket	Tijdvaksturing 17.00-20.00 u + 23.00-24.00 u	92		Nee	Ja
Prime Pakket X-tra		96		Nee	Ja
Non Prime Pakket	Tijdvaksturing 07.00-19.00 u + 23.30-02.00 u	70		Nee	Ja
Non Prime Pakket X-tra		74		Nee	Ja
Non Prime Pakket Day	Tijdvaksturing 07.00-19.00 u	73		Nee	Ja
Non Prime Pakket Day X-tra		77		Nee	Ja
Day Time Spotpakket	Tijdvaksturing 07.00-17.00 u	n.v.t.		Nee	Ja

* Fixed Budget of Fixed Price in combinatie met GRP-pakket(ten) en/of Filmpakket

** in te kopen voor de maandelijks gepubliceerde doelgroepen

*** alle geleverde GRP's worden in rekening gebracht

Figuur 6: Mogelijke inkoopopties voor films (SBS, 2007)

In figuur 6 zijn alle inkoopopties weergegeven. Bij de kwalitatieve inkoopopties, is het verschil tussen Fixed Budget en Fixed Price dat bij Fixed budget er op alle kanalen van SBS blokken geselecteerd mogen worden, en dat in het geval van het overschrijden van het aantal “bestelde” GRP's, er geen verrekening plaatsvindt. In het geval van onderschrijden van het aantal GRP's

vindt er net als bij Fixed Price wel een verrekening plaats. Verrekening houdt in dat er betaald wordt voor het aantal GRP's dat gehaald is.

Zoals eerder gezegd is het grootste verschil tussen de pakketten (ook in prijs) het tijdstip van uitzenden van de reclames. Voor de "pakketsoorten" geldt dat er een maximaal toegestaan percentage van het totale campagnebudget is dat aan die "pakketsoort" mag worden besteed. Dit is 35% voor het superprime pakket en het superprime pakket X-tra, 100% voor het daytime spotpakket en 50% voor de andere pakketten. Als een adverteerder meer zendtijd in een bepaald tijdvak wil, heeft hij de optie tot aankoop van een X-tra pakket.

Voorbeeld: Adverteerder 1 maakt gebruik van een pakket, maar geen gebruik van de mogelijkheid tot aankoop van een X-tra pakket. Hij verdeelt zijn budget als volgt: 50% prime pakket, 50% non-prime pakket. Omdat de index van het prime pakket hoger is dan die van het non-prime pakket, zal hij minder GRP's in het prime pakket kunnen aankopen dan in het non-prime pakket.

Adverteerder 2 maakt wel gebruik van de mogelijkheid tot aankoop van een X-tra pakket. Hij verdeelt zijn budget als volgt: 50% + 50% (=100%) Prime pakket X-tra.

Maandindex

Per maand zijn de kosten per GRP verschillend. De maandindex voor SBS is vergelijkbaar met andere omroepen, en is te vinden in tabel 6.

Maand	januari	februari	maart	april	mei	juni
Index	56	71	88	113	127	117
Maand	juli	augustus	september	oktober	november	december
Index	75	75	125	126	127	100

Tabel 6: Maandindex (SBS, 2007)

Het grote verschil in indices in tabel 6 heeft te maken met de Sinterklaas en kerstinkopen, en de inkopen voor de zomer.

Marktindex

De marktindex wordt maandelijks door SBS gepubliceerd. Hiermee kan SBS inspelen op vraag en aanbod in commercials. De marktindex is een index tussen de 95 en 105.

Spotlengte index

In de reclamewereld geldt een “standaard” spotlengte van 30 seconden. Deze spotlengte heeft index 100. Een kortere spot kost meer per seconde, en een langere spot kost per seconde minder. De indices zijn weergegeven in tabel 7.

Spotlengte in seconden	k/grp-index		k/grp-index
5	40	40	130
10	50	45	145
15	60	50	160
20	75	55	170
25	85	60	180
30	100	>60	naar ratio van een 60' spot
35	115		

Tabel 7: Spotlengte index (SBS, 2007)

Overige inkoopopties

Een adverteerder kan ook nog enkele extra opties inkopen, die als doel hebben de effectiviteit van de reclame te vergroten. De belangrijkste inkoopopties zijn bijvoorbeeld de *meerlingcommercials*, dit zijn reclames die meerdere keren terugkomen tijdens een reclameblok. De toeslagen hiervoor zijn 0, 5, 10 en 15 procent voor reclames die respectievelijk 2, 3, 4 en 5 keer terugkomen in hetzelfde blok. Een andere optie is om de reclame helemaal aan het begin of aan het einde van een reclameblok te laten zien. Bij SBS zijn er vier voorkeursposities: eerste, tweede, voorlaatste en laatste in een reclameblok. De toeslagen hiervoor zijn respectievelijk 20, 15, 7.5 en 12.5 procent. Verder geldt er een toeslag voor kortere reclameblokken, omdat deze beter worden bekeken. Deze reclameblokken worden bij SBS powerbreaks en Umfeld+ breaks genoemd. Een powerbreak duurt ongeveer 2 minuten en heeft een toeslag van 15% en een Umfeld+ break duurt ongeveer 4 minuten en heeft een toeslag van 10%. Zowel powerbreaks als Umfeld+ blokken worden uitgezonden tijdens goed bekeken programma's.

Betrouwbaarheid

Nu we weten hoe de steekproef is opgebouwd, en hoe de kijkcijfers worden gemeten, kunnen we ook iets zeggen over de betrouwbaarheid van de gemeten kijkcijfers. Hierbij speelt niet alleen de grootte van de steekproef een rol, maar ook de wegingsfactoren en het “samen kijken” van gezinnen met een kastje. Den Boon en Wedel (1999) beschrijven in hun artikel hoe een vuistregel kan worden geconstrueerd om snel een betrouwbaarheidsinterval te berekenen. Op dit artikel is deze paragraaf voor een groot deel gebaseerd. Algemene formules en uitleg over betrouwbaarheidsintervallen is afkomstig uit Oosterhoff en van der Vaart (2003) en Bain & Engelhardt (1992)

Betrouwbaarheidsinterval van een enkele meting

We kunnen de kijkers die wel of niet kijken ($X_1, \dots, X_{2700}$ als er 2700 waarnemingen zijn) beschouwen als een onafhankelijke steekproef afkomstig uit een Bernoulli verdeling $X_i \sim \text{BIN}(1, p)$. In dat geval is de meest aannemelijke schatter voor p : $\hat{p} = \sum_{i=1}^n X_i / n$. Ook weten we dat $\sum_{i=1}^n X_i \sim \text{Bin}(n, p)$ verdeeld is.

Met behulp van de centrale limietstelling kunnen we een betrouwbaarheidsinterval construeren. Uit de centrale limietstelling volgt:

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} Z \sim N(0,1)$$

8)

(Bain & Engelhardt, 1992)

In bovenstaande vergelijking is $\hat{p}$ de geschatte waarde van p en is p de “werkelijke” waarde. n is het aantal waarnemingen en $N(0,1)$ staat voor de standaardnormale verdeling. Er geldt voor grote n :

$$P \left[-z_{1-\frac{\alpha}{2}} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{1-\frac{\alpha}{2}} \right] = 1 - \alpha$$

9)

(Bain & Engelhardt, 1992)

In formule 9 is n de grootte van de steekproef, p de kijkkans, en is $z_{1-\frac{\alpha}{2}}$ het punt waarvoor geldt dat de oppervlakte van de standaardnormale verdeling van “min oneindig” tot $z_{1-\frac{\alpha}{2}}$ exact gelijk is aan $1-\alpha/2$. Vanwege de symmetrie van de normale verdeling is het getal $-z_{1-\frac{\alpha}{2}}$, dat getal waarvoor de oppervlakte van de standaardnormale verdeling van “min oneindig” tot $-z_{1-\frac{\alpha}{2}}$ exact gelijk is aan $-\alpha/2$. Deze formule willen we oplossen naar $-p_1$ en p_1 om zo een betrouwbaarheidsinterval (p_0, p_1) voor p te construeren. Omdat dit erg lastig is, wordt in de praktijk vaak de volgende aanname gemaakt:

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \rightarrow Z \sim N(0,1)$$

10)

(Bain & Engelhardt, 1992)

Bovenstaand resultaat geldt alleen als $n \rightarrow \infty$, en volgt uit de stelling van Slutsky.

Voor grote n geldt dus het volgende:

$$P \left[-z_{1-\frac{\alpha}{2}} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{1-\frac{\alpha}{2}} \right] = 1 - \alpha$$

11)

(Bain & Engelhardt, 1992)

Hieruit is p wel makkelijk op te lossen, het betrouwbaarheidsinterval voor p wordt gegeven door:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

12)

(Bain & Engelhardt, 1992)

In het geval van de kijkluchtheid kan deze direct worden ingevuld als waarde voor p , deze wordt daarom ook wel de “kijkkans” genoemd. Voor de waarde van $z_{1-\frac{\alpha}{2}}$ geldt dat bij een

betrouwbaarheidsniveau van 95% deze de waarde 1.96 heeft. Voor de volgende formules geldt dat voor p ook $\frac{1}{n}$ kan worden ingevuld, weer moet je dan de aanname maken dat $n \rightarrow \infty$.

Ook kunnen we in plaats van in absolute aantallen kijkers, kijken naar de betrouwbaarheidsmarges. Dit zijn betrouwbaarheidsintervallen die uitgedrukt zijn als percentage van de gemeten kijkdichtheid.

$$b_a' = \pm 1.96 \frac{\sqrt{\frac{p(1-p)}{n}}}{p} = \pm 1.96 \sqrt{\frac{1-p}{np}} \quad 13)$$

b_a = betrouwbaarheidsmarge bij 95% zekerheid

p = kijkdichtheid (variërend van 0,00-1,00)

n = steekproefomvang

Den Boon en Wedel (1999)

Voor kleine waarden van p kan de formule worden vereenvoudigd tot:

$$b_a' = \pm 1.96 \sqrt{\frac{1-p}{np}} = \pm 1.96 \sqrt{\frac{1}{np} - \frac{p}{np}} \approx \pm 1.96 \sqrt{\frac{1}{np}} \approx \pm 2 \sqrt{\frac{1}{np}} \quad 14)$$

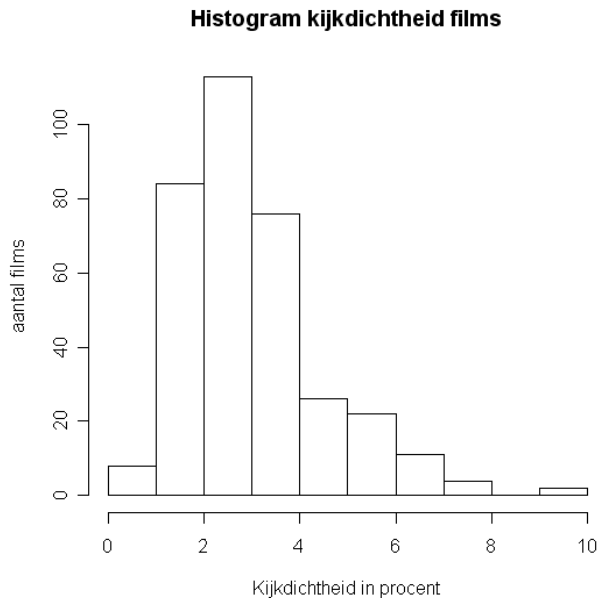
Den Boon en Wedel (1999).

Bij grote waarden voor n nadert p/np tot 0. Daardoor is bij benadering, bij kleine waarden voor p en grote waarden voor n, de betrouwbaarheidsmarge alleen maar afhankelijk van de kijkdichtheid p en de grootte van de steekproef n. Het product van n en p is gelijk aan het aantal kijkers (waarnemingen). Met andere woorden, bij een vaste grootte van de steekproef zoals het geval is bij het meten van kijkcijfers, is de marge alleen maar afhankelijk van het aantal kijkers.

Om een indruk te geven wat de kijkcijfers voor films ongeveer zijn, rekenen we voor een database met 350 SBS-films uit 2006 de statistieken gemiddelde, maximum, minimum en mediaan uit, en plotten tevens het histogram.

	Kijkdichtheid (%)
gemiddelde	3,03
maximum waarneming	9,60
minimum waarneming	0,30
mediaan	2,70

Tabel 8: verschillende statistieken voor de kijkdichtheid van SBS films



Figuur 7: histogram kijkdichtheid in procent, doelgroep 6 jaar en ouder

Stel dat we tijdens een film een enkele meting doen, en hiervan het betrouwbaarheidsinterval (p_0, p_1) opstellen. Dan krijgen we voor verschillende kijkdichtheden de volgende betrouwbaarheidsintervallen:

🎬 of kijkdichtheid	0,003	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,096
aantal kijkers met kastje	8	27	54	81	108	135	162	189	216	243	259
betrouwbaarheidsmarge	68,8%	37,5%	26,4%	21,4%	18,5%	16,4%	14,9%	13,7%	12,8%	12,0%	11,6%
p_0	0,001	0,006	0,015	0,024	0,033	0,042	0,051	0,060	0,070	0,079	0,085
p_1	0,005	0,014	0,025	0,036	0,047	0,058	0,069	0,080	0,090	0,101	0,107
betrouwbaarheidsmarge vereenvoudigde formule	70,3%	38,5%	27,2%	22,2%	19,2%	17,2%	15,7%	14,5%	13,6%	12,8%	12,4%

Tabel 9: betrouwbaarheidsintervallen voor enkele metingen bij verschillende kijkdichtheden

In tabel 9 is een steekproefgrootte gebruikt van 2.700 personen. Het betrouwbaarheidsniveau is 95%, dus $\alpha = 0.05$. In tabel 9 is 🎬 de kijkdichtheid op het moment van de meting. Met het aantal kijkers met kastje wordt bedoeld het aantal kijkers dat op het moment van de meting keek. De betrouwbaarheidsmarge is het percentage van de kijkdichtheid dat je bij de kijkdichtheid op

moet tellen om de bovengrens van het betrouwbaarheidsinterval uit te komen. p_0 is de ondergrens van het 95% betrouwbaarheidsinterval, p_1 is de bovengrens. Als laatste zijn de betrouwbaarheidsmarges weergegeven berekend met formule 14)

Bij zeer kleine aantallen kijkers mag bovenstaande formule overigens niet meer worden toegepast. Dit omdat de steekproefmarges dan asymmetrisch worden, en een normale benadering niet meer toepasbaar is. Toch is het geval van kijkdichtheid 0,003 in de tabel weergegeven om een indruk te geven welke kijkdichtheden kunnen voorkomen in de praktijk, en hoeveel kijkers met kastje er op dat moment kijken.

Verschil tussen enkele metingen en gesommeerde resultaten

Bij de berekening van de betrouwbaarheidsintervallen is er een groot verschil tussen oorspronkelijke uitkomsten en gesommeerde of gemiddelde uitkomsten. Oorspronkelijke uitkomsten zijn kijkcijfers die op zichzelf staan en niet zijn samengesteld uit andere gerapporteerde uitkomsten. De kijkdichtheid van een bepaalde minuut op een bepaalde dag van een bepaalde doelgroep is een oorspronkelijke uitkomst (Den Boon en Wedel, 1997). De kijkdichtheid van een programma (of reclameblok) van 5 minuten is het gemiddelde van alle minuutkijkdichtheden. Het resultaat is gemeten bij 5 verschillende minuten en bij 5 verschillende steekproeven. De steekproeven zullen onderling maar weinig verschillen, het kijkerspubliek zal in deze minuten nauwelijks gewisseld zijn. Als dit het geval is, maakt het voor de betrouwbaarheid nauwelijks uit of er over 1 minuut of over 5 minuten gerapporteerd wordt, tenzij het publiek tijdens deze 5 minuten sterk wisselt bijvoorbeeld door een aantrekkelijk aanbod op een andere zender. (Den Boon en Wedel, 1999)

In de praktijk worden kijkdichtheden vaak opgeteld of gemiddeld, bijvoorbeeld tot campagneresultaten (zie hiervoor ook de paragraaf “relatie tussen kijkcijfers en reclameinkomsten”). We noemen dit sommeren of aggregeren. De steekproefmarges nemen door optellen en middelen af, zodat de resultaten betrouwbaarder worden (Mulder, Snijders en Moors, 1992).

Den Boon en Wedel (1999) introduceren het begrip “statistische efficiëntie” om de betrouwbaarheid van een uitkomst aan te geven. De statistische efficiëntie is een maat voor een schatter, die betrekking heeft op de steekproefomvang in vergelijking tot een aselechte steekproef. Een statistische efficiëntie van 1 houdt in dat de steekproefmarge gelijk is aan een aselechte steekproef van dezelfde omvang. Een statistische efficiëntie van 0,8 houdt een steekproefmarge in die vergelijkbaar is met een aselechte steekproef met een omvang die 0,8 keer zo groot is (Den

Boon en Wedel, 1999). Bij een statistische efficiëntie van 0,8 zijn de betrouwbaarheidsintervallen dus groter dan bij een statistische efficiëntie van 1.

Formule 15 geeft de statistische efficiëntie van geaggregeerde metingen

$$SE_{\text{som}} = \frac{1}{1 + O \cdot C}$$

SE_{som} = Statistische efficiëntie van een som van twee metingen

O = overlap tussen beide substeekproeven

C = correlatie tussen het kijken naar beide spots in beide substeekproeven

15)

(Kish, 1965)

Voor kijkcijfers geldt vaak dat de overlap tussen beide substeekproeven bijna hetzelfde is (alle panelleden doen elke keer weer mee met het onderzoek), en dus is alleen de correlatie van belang. Je ziet aan de formule dat het optellen de betrouwbaarheidsintervallen verkleint, alleen als de correlatie van bijvoorbeeld twee reclamespots exact gelijk is aan 1, levert de formule een statistische efficiëntie van $\frac{1}{2}$, dus is de betrouwbaarheid niet toegenomen door te sommeren. Merk op dat de correlatie ook negatief kan zijn in bovenstaande formule, dit komt echter in de praktijk bijna nooit voor omdat reclameblokken meestal voor een bepaalde doelgroep bedoeld zijn. Ook is er, in het geval van twee reclames, vrijwel altijd een groep die beide reclames niet heeft gezien. In de praktijk is O een percentage. C kan met een statistiek worden geschat, bijvoorbeeld met de Cramer V statistiek.

Als we de effectieve steekproefgrootte n_{eff} definiëren als $n_{\text{eff}} = SE_{\text{som}} \cdot n$, kunnen we de formule voor de steekproefmarges voor de som van twee kijkdichtheden herschrijven als:

$$b_a' = \pm 2 \sqrt{\frac{1}{n_{\text{eff}} p}} = \pm 2 \sqrt{\frac{1 + O \cdot C}{np}}$$

16)

(Den Boon en Wedel, 1999)

In formule 16 is n_{eff} de effectieve steekproefgrootte, p de kijkdichtheid (kijkkans), n de steekproefgrootte (van beide steekproeven bij elkaar opgeteld), O de overlap tussen beide steekproeven, en C de correlatie tussen de twee programma's of reclamespots.

De formule kan worden uitgebreid voor meerdere spots. Als S het aantal spots is, kan de formule als volgt worden geschreven:

$$b_a' = \pm 2 \sqrt{\frac{1 + O \cdot C \cdot (S - 1)}{np}} \quad 17)$$

(Den Boon en Wedel, 1999)

De conclusie die uit deze formules kan worden getrokken, is dat de betrouwbaarheidsintervallen kleiner worden als kijkcijfers worden gesommeerd (en daarna eventueel gemiddeld).

Merk op dat je in alle bovenstaande formules de n steeds moet aanpassen als je met meerdere metingen werkt. Bij één reclame met een steekproefgrootte van 2700, is n gelijk aan 2700, bij twee reclames is n gelijk aan 5400 etc.

Behalve het sommeren van de resultaten en de correlatie, speelt ook clustering een rol bij het bepalen van de betrouwbaarheidsintervallen, een aspect dat tot nu toe achterwege is gelaten. Clustering speelt een rol als de waarnemingen niet onafhankelijk zijn. In het geval van kijkcijfers, kan dit ontstaan door programmatrouw. Als kijkers steeds naar dezelfde programma's kijken worden de reclamespots tijdens deze programma's steeds door dezelfde mensen bekeken.

Waar ook nog geen rekening mee is gehouden, is de weging in de steekproef. Weging is noodzakelijk als de verdeling van de steekproef op achtergrondvariabelen niet overeenkomt met de onderzoekspopulatie. Door weging worden de betrouwbaarheidsintervallen groter. Met andere woorden, de effectieve steekproefomvang neemt af.

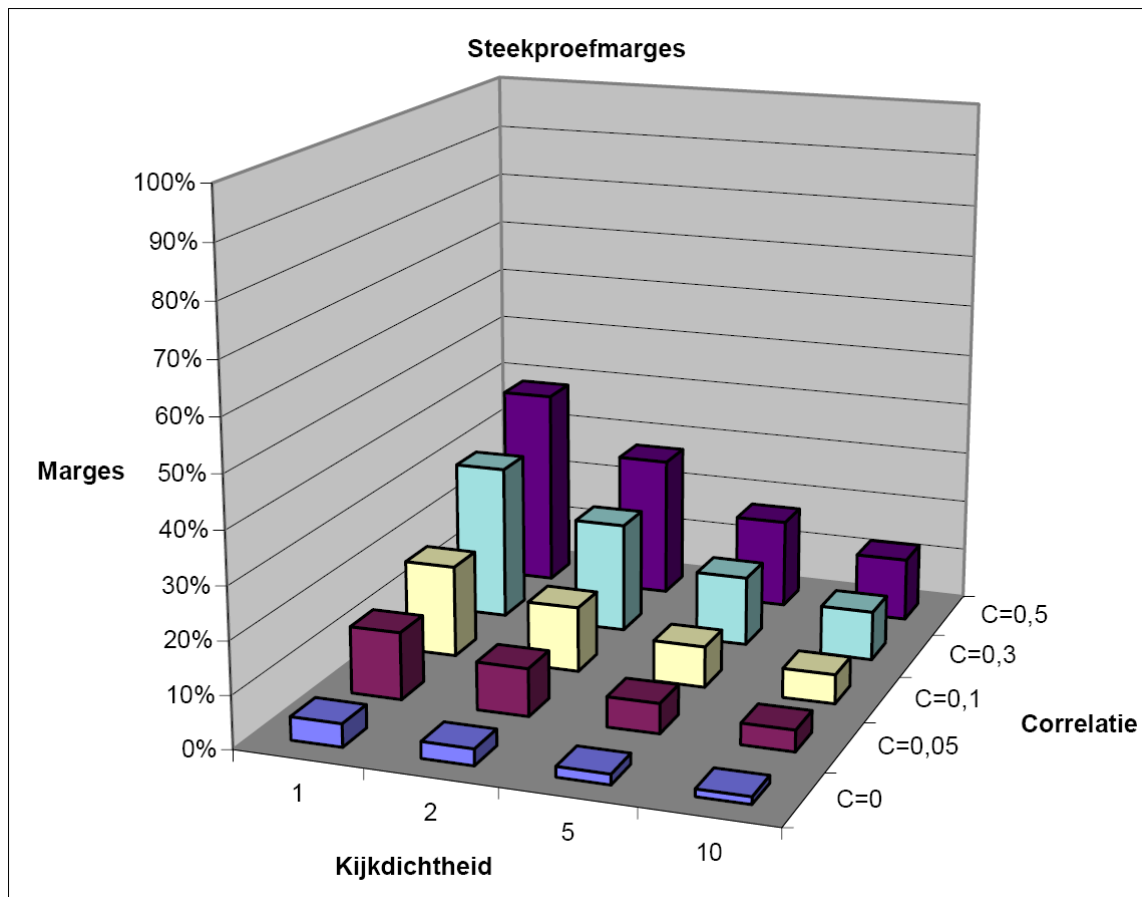
Het SCORE model

De Boon en Wedel (1999) hebben een vuistregel opgesteld voor het berekenen van de steekproefmarge, dat ze het SCORE-model noemen. Door formule 18 in te vullen kunnen de betrouwbaarheidsintervallen worden verkregen. De formule is als volgt:

$$\text{Marge}_{\text{som}} = \pm 2 \sqrt{\frac{1 + C \cdot O \cdot (S - 1)}{S \cdot R \cdot E}} \quad 18)$$

(Den Boon en Wedel, 1999)

In formule 18 staat S voor het aantal spots waarvan je een betrouwbaarheidsinterval wilt weten. C staat voor de correlatie tussen het kijken op het tijdstip van de verschillende metingen. O staat voor de overlap in de steekproef. R staat voor de gemiddelde kijkdichtheid van de spots en E staat voor de effectieve steekproefomvang. Deze is afhankelijk van de doelgroep, en van de wegingsfactoren en de clusterfactoren die op dat moment gelden voor de doelgroep.



Figuur 8: betrouwbaarheidsmarges t.o.v. de kijkdichtheid en de correlatie (Den Boon en Wedel, 1999)

Figuur 8 geeft de relatie tussen de kijkdichtheid (in procenten), de correlatie en de steekproefmarges weer. Het betreft een campagne van 150 spots. De effectieve steekproefomvang van de betreffende campagnedoelgroep is in dit geval 1357.

Figuur 8 kun je zien als een berekening uitgevoerd op een reclamecampagne van 150 spots, maar ook als een berekening die uitgevoerd is op een film van 150 minuten. De conclusie is dat

de kijkcijfers voor een enkele film vaak een hoge steekproefmarge zullen hebben, omdat de correlatie hoog zal zijn (mensen kijken een film vaak van begin tot eind, en kijken bijvoorbeeld niet alleen het laatste kwartier). Daarbij laat tabel 8 zien dat de kijkdichtheid voor films (vaak) laag is, waardoor de steekproefmarge groot zal zijn.

Een statistisch model voor het voorspellen van kijkcijfers

In dit hoofdstuk wordt eerst uitgelegd wat gegeneraliseerde lineaire modellen zijn. Daarna worden deze modellen gebruikt om de kijkcijfers van films te voorspellen. Het hoofdstuk is gebaseerd op de boeken Anderson e.a. (2004), de Gunst (2005), Dobson (2002) and McCullagh and Nelder (1989). Al deze boeken beschrijven of introduceren gegeneraliseerde lineaire modellen.

Als eerste wordt de “statistische” terminologie uitgelegd die gebruikt wordt in dit hoofdstuk: de responsvariabele is de “te verklaren” variabele. De verklarende variabelen worden gebruikt om de responsvariabele te voorspellen. Responsvariabelen en verklarende variabelen kunnen op verschillende schalen worden gemeten: binair betekent dat de variabele alleen maar twee verschillende waarden aan kan nemen. Multinomiaal betekent dat de variabele n verschillende waarden aan kan nemen. Ordinale variabelen zijn variabelen die een rangschikking hebben (bijvoorbeeld klein-groot, en jong- middelbare leeftijd- oud). Nominale variabelen worden ook wel discrete variabelen genoemd, of ook wel factoren. Continue variabelen zijn variabelen zoals lengte en tijd. Continue variabelen worden in het Engels soms ‘variates’ genoemd.

Het klassieke lineaire model

Een Gegeneraliseerd lineair model (GLM) is een soort lineair model. Van alle lineaire modellen is het klassieke lineaire model het meest bekend. Volgens dit model is de responsvariabele Y een som van zijn gemiddelde μ en een random (willekeurige) variabele ε .

$$Y_i = \mu_i + \varepsilon_i, \quad i = 1, \dots, n \quad 19)$$

Hierin is de aanname dat μ een lineaire combinatie is van de verklarende variabelen X en de ‘meetfout’ (in het engels: error term) ε een normaal verdeelde variabele is met verwachting 0 en variantie σ^2 . n is het aantal observaties, en i is een teller die het nummer aangeeft van de observatie.

Dit model kan (in dit voorbeeld met vier verklarende variabelen) als volgt worden opgeschreven:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i, \quad i = 1, \dots, n \quad 20)$$

Of in matrix notatie:

$$\underline{Y} = \underline{X} \cdot \underline{\beta} + \underline{\varepsilon} . \quad 21)$$

Als het aantal observaties gelijk is aan n , en het aantal verklarende variabelen gelijk is aan p , dan is $\underline{Y}$ een $n \times 1$ vector en $\underline{\beta}$ een $p \times 1$ vector. $\underline{X}$ is een $n \times p$ matrix en $\underline{\varepsilon}$ is een $n \times 1$ vector.

$\underline{X}$ wordt ook wel de design matrix genoemd, en $\underline{Y}$, $\underline{\beta}$ and $\underline{\varepsilon}$ zijn vectoren met componenten die corresponderen met respectievelijk de geobserveerde waarnemingen van de responsvariabele, de kolom vector van parameters en de vector van meetfouten/errors.

Om het verschil tussen het klassieke lineaire model en gegeneraliseerde model uit te leggen, schrijven we het klassieke lineaire model op de volgende manier:

$$\begin{array}{lll} \Omega : & Y \text{ is Normaal}(\mu, \sigma^2) \text{ verdeeld} & \text{i)} \\ & \eta = \underline{X} \cdot \underline{\beta} & \text{ii)} \\ & E[Y] = \mu = g^{-1}(\eta) = \eta & \text{iii)} \end{array} \quad 22)$$

(McCullagh and Nelder, 1989)

Van bovenstaand model kan je het eerste gedeelte (deel i)) beschouwen als de stochastische component. Het tweede gedeelte kun je beschouwen als de systematische component; de p covariaten worden gecombineerd om een lineaire voorspelling η te doen. De relatie tussen component i) en iii) wordt gelegd via de zogenoemde "link functie" g . Dit is de derde component van het model. In het geval van het klassieke lineaire model is dit gelijk aan de identiteitsfunctie. De link functie zal verderop worden behandeld.

De beperkingen van het klassieke lineaire model zijn dat de volgende aannames niet altijd gelden in de praktijk (Anderson e.a., 2004):

- De normale verdeling en de constante variantie gelden niet altijd
- De waarden van de responsvariabele kunnen soms alleen groter dan nul zijn (in het geval van aantallen bijvoorbeeld). De aanname dat de meetfouten normaal verdeeld zijn "overtreedt" deze restrictie.
- Als de responsvariabele niet-negatief is, dan gaat vaak intuïtief de variantie van Y naar nul als de verwachting van Y naar nul gaat. Met andere woorden, de variantie is een functie van de verwachting.

- De “additiviteit” van de effecten, beschreven in gedeelte ii) is niet altijd realistisch. Soms wordt het wel of niet bezitten van een eigenschap beter beschreven door een percentage te nemen, met andere woorden een multiplicatief model beschrijft de realiteit soms beter.

Het klassieke lineaire model is een speciaal geval van een GLM. GLMs bestaan uit een grotere verzameling modellen. Sommige restricties zoals de aanname van normaliteit, constante variantie en additiviteit worden losgelaten.

Algemene vorm van Generalized Linear Models (GLM)

De algemene vorm van een GLM wordt nu analoog aan de notatie van het lineaire model, als volgt opgeschreven. Aannames i), ii) en iii) worden:

$$\begin{array}{llll}
 \text{Y is afkomstig uit de exponentiële familie van verdelingen} & \text{i)} & & \\
 \Omega : \quad \eta = \mathbf{X} \cdot \underline{\beta} & \text{ii)} & 23) & \\
 & \text{5)} & & \\
 E[Y] = \mu = g^{-1}(\eta) & \text{iii)} & &
 \end{array}$$

(McCullagh and Nelder, 1989)

In dit model is alleen punt ii) hetzelfde als in het klassieke lineaire model. De relatie tussen component i) en iii) is via de link functie g . De linkfunctie moet differentieerbaar en monotoon zijn.

Y moet een kansverdeling hebben die afkomstig is uit de exponentiële familie. De exponentiële familie van verdelingen is als volgt gedefinieerd: Neem een stochastische grootheid Y waarvan de kansverdeling afhangt van een parameter θ . De verdeling is afkomstig uit de exponentiële familie als deze kan worden geschreven als: (Dobson, 2002):

$$f(y; \theta) = s(y)t(\theta)e^{a(y)b(\theta)} \quad . \quad 24)$$

Hierin zijn a , b , s , en t bekende functies. De formule kan ook worden geschreven als:

$$f(y; \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \quad . \quad 25)$$

Als $a(y) = y$, dan wordt de verdeling ook wel de “canonieke vorm” genoemd (in het Engels: canonical), of ook wel ‘standaard’ vorm. Ook wordt $b(\theta)$ soms de natuurlijke parameter van de verdeling genoemd. De canonieke vorm heeft vaak als voordeel dat er analytisch makkelijker mee te rekenen is dan een niet canonieke vorm.

Er zijn veel bekende verdelingen die behoren tot de exponentiële familie. Bijvoorbeeld de Poisson verdeling, de normale verdeling en de binomiale verdeling, kunnen alle drie worden geschreven in de canonieke vorm (Dobson, 2002).

Tabel 10 beschrijft de keuzes die gemaakt dienen te worden voor a, c en d om in te vullen in vergelijking 25, om deze verdelingen te krijgen.

Parameters van de exponentiële familie

Verdeling	Natuurlijke parameter	c	d
Poisson	$\log \theta$	$-\theta$	$-\log y!$
Normaal	μ / σ^2	$-\mu^2/2\sigma^2 - \frac{1}{2} \log(2\pi\sigma^2)$	$-y^2/2\sigma^2$
Binomiaal	$\log(p/(1-p))$	$n \log(1-p)$	$\log\left[\binom{n}{y}\right]$

Tabel: 10: Verschillende parameters van de exponentiële familie om bekende verdelingen te verkrijgen.

Bijvoorbeeld ingevuld voor de Poisson verdeling:

$$\begin{aligned}
 f(y; \theta) &= \exp[a(y)b(\theta) + c(\theta) + d(y)] \\
 &= \exp(y \log \theta - \theta - \log y!) = \theta^y e^{-\theta} / y!
 \end{aligned}
 \tag{26}$$

De exponentiële familie kan ook op een andere manier worden geschreven, zoals gedaan wordt in Anderson (2004) en de Gunst (2006):

$$f_i(y; \theta) = \exp[(y\theta_i - b(\theta_i))/(\phi / A_i) + c(y, \phi / A_i)] \tag{27}$$

(de Gunst, 2006)

Nu is θ_i in vergelijking 27) de parameter van de exponentiële familie die hoort bij Y_i . ϕ is in deze vergelijking een (vooraf bekende) schaal parameter die hetzelfde is voor alle Y_i . Deze parameter hoeft niet te worden geschat en telt niet als parameter van de exponentiële familie. Het symbool A_i in 27) staat voor een vooraf bekende constante. De vorm van f_i wordt bepaald door de functies b en c . Deze vorm van schrijven is nuttig omdat er kan worden aangetoond dat:

$$\begin{aligned} E[Y_i] &= \mu_i = b'(\theta_i) \quad i = 1, \dots, n \\ \text{Var}[Y_i] &= b''(\theta_i)\phi/A_i, \quad i = 1, \dots, n \end{aligned} \quad (28)$$

(de Gunst, 2006)

b en c moeten zó gekozen zijn dat f_i een kansdichtheid is, en zodat vergelijking 28) geldt.

Voor praktische doeleinden is het nuttig te weten dat een kansdichtheid die behoort tot de exponentiële familie de volgende eigenschappen heeft (Anderson e.a., 2004):

1. De verdeling is vastgelegd door verwachting en variantie
2. De variantie van Y_i is een functie van zijn gemiddelde

Bij deze schrijfwijze, als in formule 27), worden de keuzes voor b en c in tabel 11 weergegeven waarvoor je weer een aantal bekende kansverdelingen krijgt.

Parameters van the exponentiële familie

Verdeling	$b(\theta_i)$	ϕ	A_i	$c(y_i, \phi / A_i)$
Poisson	e^{θ_i}	1	1	$-\log(y_i!)$
Normaal	$\frac{1}{2}\theta_i^2$	σ^2	1	$-\frac{1}{2}(y_i / \sigma)^2 - \log \sigma \sqrt{2\pi}$
Binomiaal	$\log(1 + e^{\theta_i})$	1	n_i	$-\log \left[\binom{n_i}{n_i, y_i} \right]$

Tabel 11: Verschillende parameters van de exponentiële verdeling om bekende verdelingen te verkrijgen.

De link functie

Anderson e.a. (2004) zeggen het volgende over de link functie: In de praktijk proberen statistici die klassieke lineaire regressie toepassen de data te transformeren zodat de normaliteits, de constante variantie en de additiviteit van de effecten gelden. Gegeneraliseerde lineaire modellen daarentegen, hebben als enige eis dat er een link functie is die ervoor zorgt dat de additiviteits-eis geldt. In het klassieke lineaire model geldt dat Y additief afhangt van zijn covariaten, bij gegeneraliseerde modellen hoeft alleen te gelden dat een transformatie van Y , geschreven als $g(Y)$, additief afhangt van zijn covariaten. Deze link functie moet differentieerbaar zijn en monotoon. Er zijn een aantal veelgebruikte link functies die hun eigen naam hebben. Verschillende soorten link functies zijn weergegeven in tabel 12.

Link functies

	$g(x)$	$g^{-1}(x)$
Identiteit	x	x
Log-link	$\ln(x)$	e^x
Logit	$\ln(x/(1-x))$	$e^x/(1+e^x)$
Probit	$\Phi^{-1}(x)$	$\Phi(x)$
Reciproce	$1/x$	$1/x$

Tabel 12: Verschillende soorten link functies

In tabel 12 is $\Phi^{-1}(x)$ de inverse van de verdelingsfunctie van de standaard normale verdeling. De Logit en Probit link functies worden vaak gebruikt in combinatie met een error-functie die binomiaal verdeeld is, omdat deze linkfuncties het interval $(0,1)$ op de reële as transformeren. Als de logit link functie wordt gebruikt in combinatie met een binomiale verdeling spreekt men ook wel van een logistisch regressie model. Als de probit link functie wordt gebruikt heet dit een Probit regressie model. De log link functie wordt vaak gebruikt in combinatie met de Poisson verdeling, omdat het interval $(0, \infty)$ wordt getransponeerd op de reële as. De identiteitsfunctie wordt vaak gebruikt in combinatie met de normale verdeling. Zoals we eerder zagen, geeft deze combinatie het klassieke lineaire regressiemodel.

Een GLM is totaal vastgelegd door twee keuzes, namelijk de keuze voor de (parameters van de) exponentiële familie, en de keuze voor de link functie, omdat de systematische component (onderdeel ii) hetzelfde is voor alle GLMs.

Het schatten van de parameters

Nadat de vorm van een gegeneraliseerd lineair model is gekozen zoals beschreven in de vorige paragraaf, en gegeven een aantal observaties Y , worden de parameters β geschat door het maximaliseren van de likelihoodfunctie (of equivalent: het maximaliseren van de logaritme van de likelihoodfunctie). De essentie van deze methode is dat de parameters worden gevonden die, als de methode wordt toegepast op de gekozen modelvorm, met de grootste kans de geobserveerde data voortbrengen. De likelihood is gedefinieerd als het product van de kansen dat de geobserveerde data is waargenomen. Voor continue verdelingen wordt de kansdichtheid gebruikt in plaats van de kans. Gewoonlijk wordt de log van de likelihood gebruikt, omdat het makkelijker te werken is met een sommatie dan met een product. Maximum likelihood schatting zoekt de parameters die de log likelihoodfunctie maximaliseren (Anderson e.a. 2004). Bij de bovenstaande schattingsmethode dient te worden opgemerkt dat, als je deze methode vergelijkt met de kleinste kwadraten schatter van β , de kleinste kwadraten schatter in het algemeen minder goed functioneert. Alleen bij het klassieke lineaire model is dit niet het geval, omdat dan beide schattingsmethoden dezelfde schattingen geven (de Gunst, 2006).

We noteren de maximum likelihood schatter van β met $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)^T$. Het is de waarde die wordt verkregen door de log likelihoodfunctie van de exponentiële familie te maximaliseren naar β .

$$l(\theta) = l(\beta) = \sum_{i=1}^n \left(\frac{y_i \theta_i - b(\theta_i)}{\phi / A_i} + c(y_i, \phi / A_i) \right) \quad 29)$$

In bovenstaande formule is β de vector van onbekende constanten die via de linkfunctie wordt 'gelinkt aan' de parameter θ .

Als voorbeeld hoe dit wordt uitgerekend bekijken we hoe dit werkt voor de Poissonverdeling in combinatie met een log link functie. De kansdichtheid wordt gegeven door:

$$f(x; \mu) = e^{-\mu} \mu^y / y! \quad 30)$$

Dus de log likelihoodfunctie is:

$$l(y; \mu_i) = \sum_{i=1}^n \ln f(y_i; \mu_i) = \sum_{i=1}^n [-\mu_i + y_i \ln \mu_i - \ln(y_i!)] . \quad 31)$$

Als we hierin de logaritmische linkfunctie $\mu_i = \exp(\sum_j X_{ij}\beta_j)$ invullen, reduceert de likelihoodfunctie tot:

$$l(y; e^{X\beta}) = \sum_{i=1}^n \left[-\exp(\sum_{j=1}^p X_{ij}\beta_j) + y_i \sum_{j=1}^p X_{ij}\beta_j - \ln(y_i!) \right] \quad . \quad (32)$$

In de meeste gevallen bestaat er geen expliciete uitdrukking van het maximum en moet $\hat{\beta}$ numeriek worden bepaald. Nelder en Wedderburn (1972) beschrijven hiervoor een algemene methode die ze Fisher scoring noemen. Deze methode begint met een startoplossing β^0 en update deze oplossing tot β^1 op de volgende manier:

$$\beta^1 = \beta^0 + \{E_{\beta^0}(-\frac{\partial^2 l}{\partial \beta \partial \beta^T})\}^{-1} \frac{\partial l}{\partial \beta} \quad . \quad (33)$$

Waarbij de beide afgeleiden worden geëvalueerd in het punt β^0 en de verwachting wordt berekend alsof β^0 de 'ware parameter' is. De startoplossing β^0 kan worden verkregen door alle parameterwaarden op nul te zetten of de resultaten van een vorig gefitte GLM te gebruiken. De β^0 wordt dan vervangen door de β^1 en het updaten wordt weer herhaald. De methode stopt als $\beta^m - \beta^{m-1}$ klein genoeg is. Op dat moment wordt $\hat{\beta}$ gegeven door de laatste β^m .

Model Diagnose

Behalve de schattingen zelf die de likelihood maximaliseren, kan ook extra informatie worden berekend die de nauwkeurigheid van deze schattingen aangeeft. Om de nauwkeurigheid van een schatting van een parameter te bepalen, kan de asymptotische variantiematrix van de schatter $\hat{\beta}$ worden gebruikt. Deze wordt gegeven door de inverse van de Fisher informatiematrix:

$$\{E_{\beta^0}(-\frac{\partial^2 l}{\partial \beta \partial \beta^T})\}^{-1} = \phi(X^T W X)^{-1} \quad . \quad (34)$$

Hierin is W de n x n diagonaalmatrix met als i-de element:

$$w_i^0 = A_i \{g'(\mu_i^0)^2 b''(\theta_i^0)\}^{-1} \quad . \quad (35)$$

Als je deze variantie wilt gebruiken heb je wel schattingen nodig voor β , W en ϕ . Als schatter voor ϕ kan formule 36 gebruikt worden:

$$\hat{\phi} = \frac{\chi^2}{n - p - 1} \quad , \quad (36)$$

waarbij χ^2 de gegeneraliseerde Pearson chikwadraat statistiek is (de Gunst, 2006). De Pearson chikwadraat statistiek wordt gegeven door:

$$\chi^2 = \phi \sum_{i=1}^n \frac{\{Y_i - E_{\hat{\beta}} Y_i\}^2}{Var_{\hat{\beta}}(Y_i)} = \sum_{i=1}^n \frac{\{Y_i - b'(\hat{\theta}_i)\}^2}{b''(\hat{\theta}_i) / A_i} \quad . \quad (37)$$

Een andere manier om ϕ te schatten is door de “totale deviantie schatter” te gebruiken:

$$\hat{\phi} = \frac{D}{n - p} \quad . \quad (38)$$

(Anderson e.a., 2004)

De deviantie D wordt in de volgende paragraaf besproken.

Met behulp van deze schattingen voor de variantie, kan een (asymptotische), $(1-\alpha)100\%$ betrouwbaarheidsinterval voor β_j worden gemaakt. Zie hiervoor formule 39).

$$\beta_j \pm t_{n-p-1; (1-\alpha/2)} \sqrt{(\hat{\phi}((X^T \hat{W} X)^{-1})_{jj})} \quad . \quad (39)$$

(de Gunst, 2006).

Om te checken welke variabele wel of niet in het model moet, zou je de corresponderende t-ratio kunnen gebruiken. De t-ratio is de geschatte waarde gedeeld door de geschatte standaarddeviatie, en vergelijkt deze ratio met plus of min het $(1-\alpha/2)$ kwantiel van de t-verdeling met $(n-p-1)$ vrijheidsgraden. In de praktijk, echter, is een grafische interpretatie van de schattingen veel zinvoller. Een goede grafische interpretatie kan worden verkregen door een grafiek te maken met op de x-as de verklarende variabele en op de y-as de geschatte waarden van deze variabele. Vervolgens kan worden gekeken of de geschatte waarden “logisch” zijn, en of de betrouwbaarheidsintervallen niet te groot zijn.

De Deviantie

Een andere manier om het model te evalueren is om het te vergelijken met een algemener model dat het maximaal mogelijke aantal parameters heeft. Dit model wordt het verzadigde model genoemd (in het Engels: saturated model). Het is een gegeneraliseerd model met dezelfde kansverdeling en dezelfde link functie als het onderzochte model. Als er n observaties zijn Y_i , $i = 1, \dots, n$, allemaal met verschillende waarden voor de lineaire component $x_i^T \beta$ dan kan een verzadigd model worden gespecificeerd met n parameters, één parameter voor elke observatie. Dit model wordt ook wel het maximale of volledige model genoemd. Als sommige van deze observaties hetzelfde lineaire component of covariantie patroon hebben, met andere woorden ze corresponderen met dezelfde combinatie van waarden van de verklarende variabelen, dan worden ze replica's genoemd. In dat geval is het maximum aantal parameters dat kan worden geschat voor het verzadigde model gelijk aan het aantal unieke observaties, dit is dan kleiner dan n (Dobson 2002).

In het algemeen, is m het maximum aantal parameters is dat kan worden geschat, $\hat{\beta}_{\max}$ de parameter vector van het verzadigde model, en $\hat{\beta}$ de maximum likelihood schatter van het te onderzoeken model. De likelihoodfunctie voor het verzadigde model geëvalueerd in het punt $\hat{\beta}_{\max}$, $L(\hat{\beta}_{\max}; y)$, zal groter zijn dan elke andere likelihoodfunctie van deze observaties, met dezelfde kansverdeling en link functie, omdat het verzadigde model de meest complete beschrijving van de data is die mogelijk is.

De likelihood ratio is een goede maatstaf om de goodness-of-fit van het model te bepalen. In formulevorm wordt deze gegeven door:

$$\lambda = \frac{L(\hat{\beta}_{\max}; y)}{L(\hat{\beta}; y)}, \quad (40)$$

waarbij $L(\hat{\beta}; y)$ de maximum likelihood waarde is van de likelihood functie van het onderzochte model.

In de praktijk wordt vaak de logaritme van de likelihood ratio gebruikt, omdat hiermee makkelijker te rekenen is. Dit is het verschil tussen de likelihoodfuncties:

$$\ln \lambda = l(\hat{\beta}_{\max}; y) - l(\hat{\beta}; y) \quad (41)$$

Grote waarden van $\ln \lambda$ wijzen erop dat het model dat onderzocht wordt een slechte beschrijving van de data geeft, als je het vergelijkt met het verzadigde model. In de praktijk wordt nog vaker de statistiek $2 \ln \lambda$ gebruikt omdat blijkt dat $2 \ln \lambda$ chikwadraat verdeeld is. Deze statistiek wordt door Nelder en Wedderburn (1972) de *deviantie* (in het Engels: *deviance*) genoemd. Deze deviantie is hetzelfde als die in formule 38)

Bijvoorbeeld, voor de Poissonverdeling, zagen we in vergelijking (31) dat de log-likelihood functie wordt gegeven door:

$$l(\hat{\beta}_{\max}; y) = \sum_{i=1}^n [-y_i + y_i \ln y_i - \ln(y_i!)] \quad (42)$$

In bovenstaande formule zijn de μ_i vervangen door y_i , dit is omdat voor de maximum likelihood schatter, het model een parameter heeft voor elke observatie. De waarde van de parameter is gelijk aan de waarde van de observatie zelf. Stel dat het onderzochte model $p < n$ parameters heeft. De maximum likelihood schatter $\hat{\beta}$ kan gebruikt worden om de μ_i te schatten en dus zijn de geschatte waarden $y_i = \mu_i$ omdat $E(Y_i) = \mu_i$. De maximale waarde van de log-likelihood van het kleinere model is:

$$l(\hat{\beta}; y) = \sum_{i=1}^n [-\hat{y}_i + y_i \ln \hat{y}_i - \ln(y_i!)] \quad (43)$$

waarin $\hat{y}_i$ de geschatte waarde van y_i is volgens het kleinere model.

Dus de deviantie D wordt gegeven door:

$$D = 2[l(\hat{\beta}_{\max}; y) - l(\hat{\beta}; y)] \quad (44)$$

$$= 2 \left[\sum_{i=1}^n y_i \ln(y_i / \hat{y}_i) - \sum_{i=1}^n (y_i - \hat{y}_i) \right] \quad . \quad 45)$$

Voor de meeste modellen kan worden aangetoond dat $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$. In dat geval valt de laatste term in vergelijking 45 weg (Dobson, 2002).

De bijdrage van de i-de observatie aan de deviantie wordt soms het i-de deviantietoevoegsel (deviance increment in het Engels) genoemd. De deviantie D is een globale maat van de fit van het model op de data, en kan worden geïnterpreteerd als de gekwadrateerde som van de residuen in het klassieke lineaire regressiemodel. Als je de identiteitsfunctie als link functie gebruikt in combinatie met de normale verdeling, is de deviantie inderdaad gelijk aan de “residual sum of squares”

Een alternatieve maat voor de kwaliteit (van de fit) van het model is de Pearson chikwadraat statistiek χ^2 gegeven door formule 37.

Na ze te delen door ϕ , de schaalparameter uit formule 36, zijn de deviantie en de Pearson chikwadraat verdeling allebei chikwadraat verdeeld met n-p-1 vrijheidsgraden, in het geval van het klassieke lineaire model. Voor andere soorten GLMs zijn er alleen maar asymptotische resultaten voorhanden. Er moet gezegd worden dat deze asymptotische resultaten geen waarde hebben als D of χ^2 berekend worden aan de hand van een (te) kleine dataset (de Gunst 2006). Het grote voordeel van de deviantie en Pearson chikwadraat statistiek ligt in het feit dat ze handig zijn om twee geneste (“nested” in het Engels) modellen met elkaar te vergelijken. Twee modellen zijn genest als het ene model kan worden verkregen door een parameter aan het andere model toe te voegen. Of, andersom, het andere model kan worden verkregen door een parameter weg te halen uit het “grotere” model. Voor de selectie van een model kan men de deviantie D of de Pearson chikwadraat statistiek gebruiken op dezelfde manier als gedaan wordt met de “total sum of squares” in variantie analyse (AnoVa). Als hiervoor de deviantie gebruikt wordt, wordt deze techniek *deviantie analyse* genoemd (De Gunst, 2006).

De geschaalde deviantie kan worden verkregen door de deviantie te delen door de schaalparameter ϕ . Het verschil in “geschaalde deviantie” tussen de twee geneste modellen kan worden opgevat als een trekking uit een chikwadraatverdeling met een aantal vrijheidsgraden dat gelijk is aan het verschil in vrijheidsgraden tussen de twee modellen (waarbij het aantal vrijheidsgraden in een model is gedefinieerd als het aantal observaties minus het aantal

parameters). Hierdoor is het mogelijk om statistische toetsen uit te voeren. Deze toetsen kun je gebruiken om te meten of het toevoegen van een extra variabele het model goed genoeg verbeterd. De chikwadraat toets hangt af van de schaalparameter. Voor sommige kansverdelingen zoals de Poisson en Binomiale verdeling, is de schaalparameter bekend, maar voor andere verdelingen moet deze parameter geschat worden. Dit gaat ten koste van de betrouwbaarheid van deze tests (Anderson e.a., 2004). Maar in het algemeen, zelfs als asymptotische resultaten niet van toepassing zijn, is een sterke reductie van D bij het vergelijken van twee modellen, een indicatie dat het grotere model de voorkeur verdient (de Gunst, 2006).

Data analyse

In dit hoofdstuk zullen de modellen beschreven in het vorige hoofdstuk worden geschat met behulp van data van de Stichting Kijk en Luisteronderzoek. Ook zal worden gekeken naar welke verklarende variabelen een rol spelen bij het voorspellen van kijkcijfers voor films. Hierbij kun je denken aan variabelen als: het uitzendtijdstip, het aantal kijkers naar het programma voor de film, het soort film (actie, romantische komedie), de zender, de hoofdrolspeler (acteur), de internet movie database rating en de programma's van de concurrent.

Als te onderzoeken modellen worden de volgende modellen voorgesteld:

	Y is Normaal(μ, σ^2) verdeeld	i)	
$\Omega :$	$\eta = \mathbf{X} \cdot \underline{\beta}$	ii)	46)
	$E[Y] = \mu = g^{-1}(\eta) = \eta$	iii)	
	Y is Poisson(μ) verdeeld	i)	
$\Omega :$	$\eta = \mathbf{X} \cdot \underline{\beta}$	ii)	47)
	$E[Y] = \mu = e^{\eta}$	iii)	

Aan de hand van de deviantie zal beslist worden welke van deze twee modellen het beste fit op de data. De hypothese is dat het laatste model beter is, omdat in dit model alleen positieve waarden kunnen voorkomen ($E(Y)$ kan niet negatief zijn). Dit is ook bij kijkcijfers het geval. Ook neemt bij het tweede model de variantie toe als $E(Y)$ groter is, en dit is ook wat je zou verwachten bij kijkcijfers. Programma's waar maar weinig mensen naar kijken, zullen in absolute aantallen ook minder fluctuaties hebben in de kijkcijfers van week tot week, is de veronderstelling.

Met behulp van het softwarepakket R worden deze modellen geschat. De titels van de films die worden gebruikt voor deze analyse kun je vinden in bijlage 4.

De werkwijze hiervoor is het volgende stappenplan:

1. Begin met een start model dat leeg is, (zonder variabelen) of met enkele verklarende variabelen waarvan uit de verkennende data analyse is gebleken dat deze heel erg verklarend zijn.
2. Nu wordt het lineaire model geschat. Hoe de parameters worden geschat is in vorige hoofdstuk "Een statistisch model voor het voorspellen van kijkcijfers" besproken. Ook worden de betrouwbaarheidsintervallen uitgerekend, net als de deviantie van het model. De p-values van de variabelen worden uitgerekend. Dit wordt gedaan door het verschil in deviantie tussen het geschatte model te vergelijken met de deviantie van het model zonder deze variabele. Zie het einde van het vorige hoofdstuk over hoe dit precies in zijn werk gaat.
3. Aan de hand van de geschatte waarden en de p-values wordt een beslissing gemaakt om de variabelen in het model te laten, of er een uit te halen. Als er nog maar weinig verklarende variabelen zijn komt het praktisch nooit voor dat een variabele dient te worden weggelaten. Voor elke variabele die nog niet in het model zit, wordt getest of deze wel in het model zou moeten worden opgenomen. Dit wordt gedaan door een model te schatten inclusief de betreffende variabele. De p-value van de betreffende variabele wordt uitgerekend, alweer gebaseerd op het verschil in deviantie.
4. Als voor elke variabele die nog niet in het model zit de p-value is uitgerekend, wordt hiervan een lijst gemaakt en wordt deze lijst gesorteerd. De variabele met de laagste p-value is de beste kandidaat om te worden opgenomen in het model. Van deze variabelen worden de geschatte waarden bekeken, en de betrouwbaarheidsintervallen worden gecheckt. Als er een ordinale ranking is in een variabele, wordt ook gekeken of de geschatte waarde van de factoren een ordinale ranking heeft. Bijvoorbeeld: voor de variabele aantal bioscoopbezoekers bestaan de factoren "<30.000", "30.000-400.000" etc. Je verwacht dat de geschatte waarden voor deze factoren oplopen, en niet te veel fluctueren. De "beste" variabele wordt gekozen in het model. Als er geen geschikte variabelen meer zijn om te selecteren, is het model af. Anders worden stappen 2 t/m 4 herhaald.

Als het stappenplan wordt gevolgd, blijkt dat als meest verklarende de variabelen “Maandagavond SBS” en “Tijd Midden” zijn, met de laatste wordt bedoeld dat een film niet vroeg op de avond (voor 20:00) of laat op de avond (na 21:30) begint. Vervolgens is het aantal bioscoopbezoekers belangrijk, deze variabele kan worden gezien als maat voor kwaliteit van de film. Daarna is de variabele die aangeeft of de film een actiefilm is of niet, actiefilms worden beter bekeken dan niet-actiefilms. Vervolgens is er de Moviemeter rating (ook een maat voor de kwaliteit van de film). De laatste variabele is er een die aangeeft of een film op dinsdagavond op Veronica wordt uitgezonden. Uit de data blijkt dat er dan bijna niemand kijkt.

Schatten van het logistische regressiemodel

Het model dat als eerste wordt geschat is het multiplicatieve model, met een log-link functie en een Poisson verdeling voor de meetfouten. Dit model kan worden gezien als:

*Beginfactor * factor “SBSMaandagavond” * factor Tijd * factor bioscoopbezoekers **

In tabel 13 zijn de vermenigvuldigingsfactoren weergegeven:

Intercept	268.665
Tijdslot	vermenigvuldigingsfactor
SBS maandag 20-22 uur	1,917
Veronica dinsdag na 22 uur	0,792
Veronica zondag na 22 uur	0,867
Net 5 donderdag na 20-22 uur	1,299
Net 5 vrijdag na 22 uur	1,000
Veronica dinsdag 20-22 uur	1,299
Veronica zondag 18-20 uur	1,000
Veronica zondag 20-22 uur	1,299
Bioscoopbezoekers	
onbekend/kleiner dan 30.000	1,000
30.000 tot 400.000	1,116
groter dan 400.000	1,159
Genre	
Actie	1,219
Avontuur	1,165
Overig	1,000
Maand	
Oktober t/m februari	1,135
September, maart, april, mei	1,000
Kijkers vorige programma	
minder dan 150.000	1,000
150.000 tot 450.000	1,074

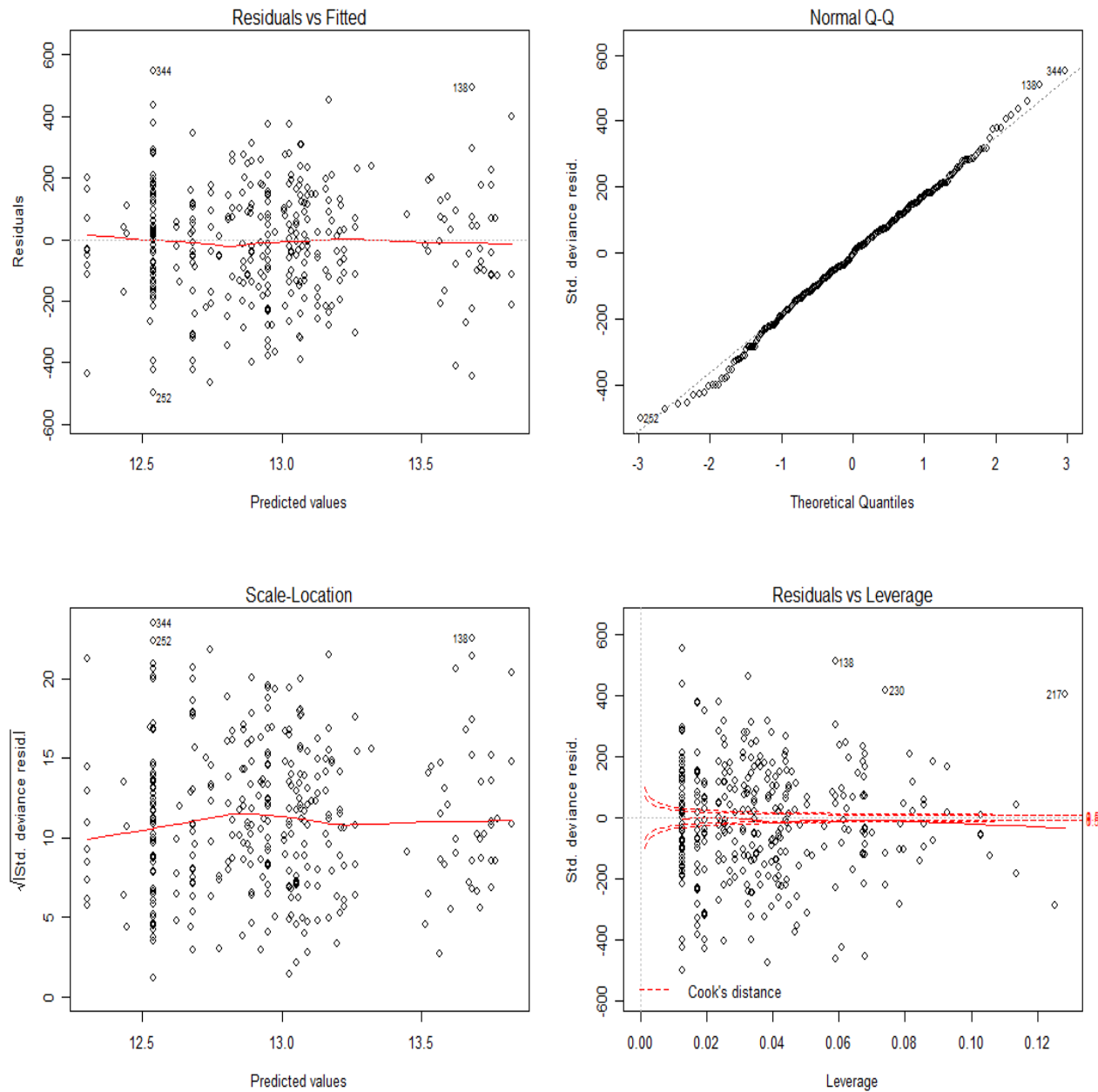
meer dan 450.000	1,160
onbekend	1,277
MovieMeter rating (0-3000)	
onbekend	0,910
kleiner dan 900	1,000
groter dan 900	1,077

Tabel 13: Geschatte waarden voor factoren in het logistische regressiemodel

Om aan de hand van tabel 13 een voorspelling te krijgen voor een bepaalde film ga je als volgt te werk: begin met het Intercept als startwaarde voor het aantal kijkcijfers. Vervolgens zoek je in de tabel de bijbehorende eigenschappen van de betreffende film op, en vermenigvuldigt de waarde met dit getal.

Model diagnostiek en residu analyse logistische regressie model

Om iets te zeggen over de kwaliteit van het model, zijn in figuur 9 vier grafieken geplott.



Figuur 9: Residu analyse van het uiteindelijke logistische regressie model.

In figuur 9 is linksboven een grafiek te zien van de voorspelde waarden tegenover de residuen. Deze grafiek geeft geen reden tot bezorgdheid, omdat er geen verband zichtbaar is. Wel zijn de observaties met veel ontbrekende waarden goed te zien in de linkerkant van de grafiek.

In de grafiek rechtsboven wordt door middel van een QQ-plot getest of de gestandaardiseerde deviantie-residuen normaal verdeeld zijn. Deze deviantie-residuen behoeven enige uitleg. Een deviantie residu geeft aan hoeveel een individuele waarneming bijdraagt aan de totale deviantie D. Elke observatie draagt een hoeveelheid d_i bij, zodanig dat $\sum d_i = D$. In formulevorm wordt het deviantie residu gegeven door:

$$r_i^D = \text{sign}(y_i - \mu_i) \sqrt{d_i} \quad (48)$$

(McCullagh and Nelder, 1989)

Het gestandaardiseerde residu is uitgeschreven:

$$r_i^D = \text{sign}(y_i - \mu_i) \sqrt{2\omega_i \int_{\mu_i}^{y_i} (y_i - \xi) / V(\xi) d\xi} \quad (49)$$

(Anderson et al, 2004)

In formule 49 geeft het laatste gedeelte (onder de wortel) weer hoeveel een observatie bijdraagt aan de totale deviantie. $V(\xi)$ is de zogenaamde “variance function”, die de variantie van een observatie geeft. ω is de schaalparameter uit de eerder beschreven exponentiële familie. De deviantie residuen hebben verschillende nuttige eigenschappen. In het algemeen zijn ze vaker normaal verdeeld dan de ruwe residuen (die gedefinieerd worden als het verschil tussen de observatie en de voorspelde waarde door het model), omdat het deviantie residu corrigeert voor de scheefheid van verdelingen. Voor continue verdelingen is het mogelijk om de verdeling van de deviantie residuen te testen op normaliteit. Een grote afwijking van de normale verdeling is een goede indicatie dat de aannames over de verdeling in het model verkeerd is. (Anderson et al, 2004). In dit geval is er dus geen reden tot bezorgdheid omdat de QQplot een rechte lijn is.

In de grafiek linksonder is de relatie tussen de voorspelde waarden van het model en de gestandaardiseerde deviantie-residuen weergegeven. Deze grafiek geeft ook geen reden tot bezorgdheid omdat er geen relatie zichtbaar is tussen de voorspelde waarden en residuen.

In de laatste grafiek in figuur 9 is de relatie weergegeven tussen de leverage en de gestandaardiseerde deviantie-residuen. Ook de leverage heeft enige uitleg. De leverage is een maat die aangeeft hoeveel invloed een waarneming heeft op zijn eigen voorspelde waarde door het model. Het is een maat die aangeeft hoeveel effect een verandering in de waarneming heeft op de voorspelde waarde door het model. De leverage ligt altijd tussen de 0 en 1. De hoogste leverage heeft film nummer 217, dit is de film “Out for Justice”, een actiefilm met Steven Seagal. Dit is een van de weinige films op de dinsdagavond van Veronica die een beetje behoorlijk is bekeken (500.000 kijkers). Ook deze grafiek geeft geen reden tot bezorgdheid.

Schatten van het lineaire regressiemodel

Het model wat we nu gaan schatten is het volgende, en wordt ook wel het lineaire regressiemodel genoemd of in het engels “ordinary least squares” (OLS).

$$\begin{array}{lll} \Omega : & Y \text{ is Normaal}(\mu, \sigma^2) \text{ verdeeld} & \text{i)} \\ & \eta = X \cdot \beta & \text{ii)} \\ & E[Y] = \mu = g^{-1}(\eta) = \eta & \text{iii)} \end{array} \quad 50)$$

In tegenstelling tot het vorige model is dit model additief. De geschatte waarden zijn als volgt:

Intercept	216.704
Tijdslot	
SBS maandag 20-22 uur	402.161
Veronica dinsdag na 22 uur	-68.149
Veronica zondag na 22 uur	-73.316
Net 5 donderdag na 20-22 uur	114.763
Net 5 vrijdag na 22 uur	0
Veronica dinsdag 20-22 uur	114.763
Veronica zondag 18-20 uur	0
Veronica zondag 20-22 uur	114.763
Bioscoopbezoekers	
onbekend	0
kleiner dan 30.000	0
30.000 tot 200.000	35.818
200.000 tot 400.000	42.678
groter dan 400.000	74.561
Genre	
Actie	90.341
Avontuur	79.786
Overig	0

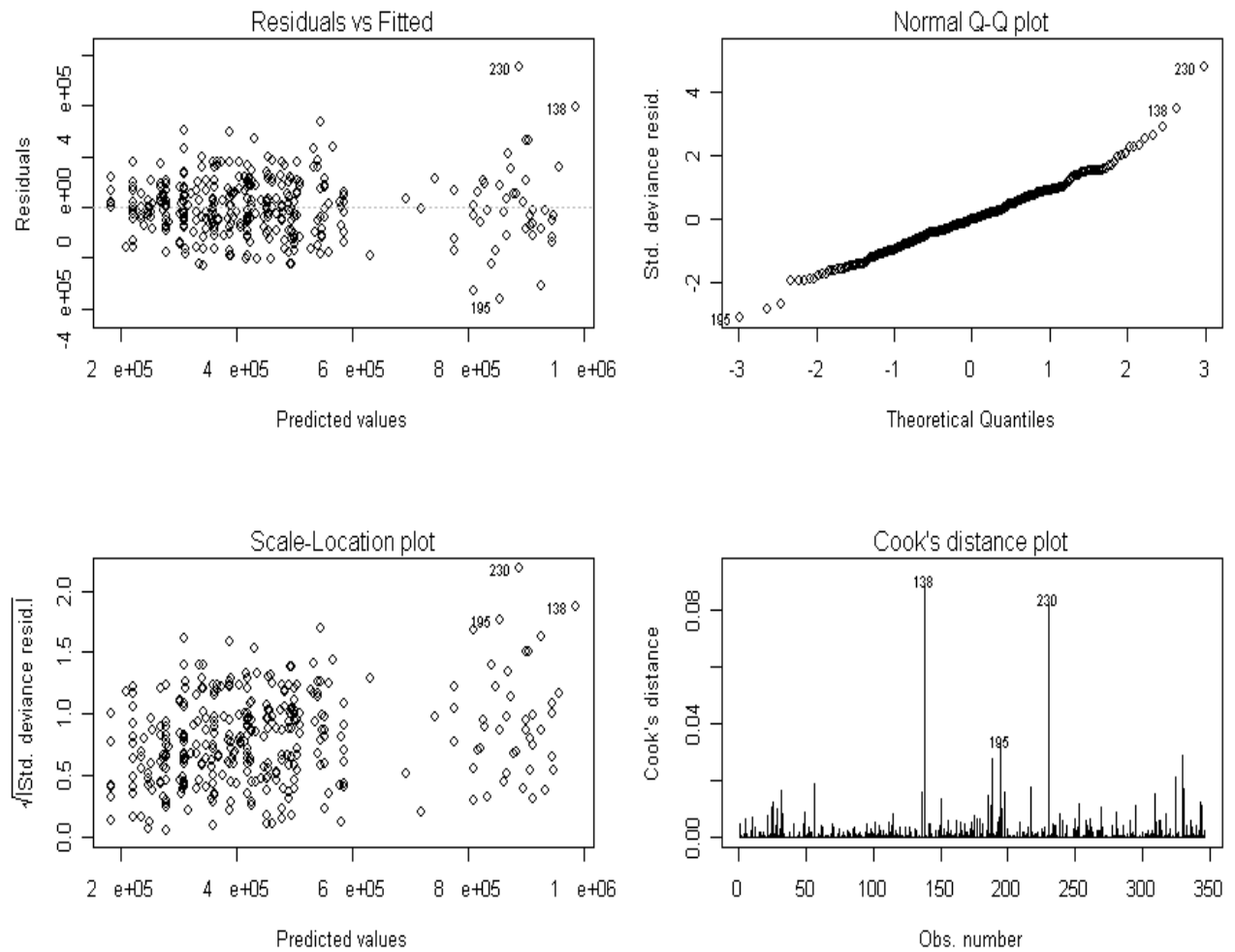
Maand	
Oktober t/m februari	56.882
September, Maart, April, Mei	0
Kijkers vorige programma	
minder dan 150000	0
150000 tot 450000	31.603
meer dan 450000	65.184
onbekend	116.231
MovieMeter rating (0-3000)	
onbekend/kleiner dan 900	0
groter dan 900	38.058

Tabel 14: Geschatte waarden voor het lineaire regressiemodel

Om vanuit tabel 14 de kijkcijfers voor een film te voorspellen begin je met de startwaarde (het intercept), en zoekt in de tabel de eigenschappen van de film op. Voor iedere eigenschap tel je het bijbehorende getal bij de kijkcijfers op.

Residu analyse lineaire regressiemodel

In figuur 10 zijn verschillende grafieken te zien die de residuen betreffen van het lineaire regressiemodel.



Figuur 10: Residu analyse voor het lineaire regressiemode

In figuur 10 is een residu analyse uitgevoerd voor het lineaire regressiemodel. Dit is ook aan de hand van de gestandaardiseerde deviantie-residuen gedaan, om goed te kunnen vergelijken met het vorige model. In het plaatje linksboven is goed te zien dat de residuen groter worden naarmate de voorspelde waarden groter worden. Dit is een teken dat er verkeerde aannames in het model zijn gedaan. In de grafiek rechtsboven in figuur 10 is ook te zien dat de gestandaardiseerde deviantie-residuen minder goed normaal verdeeld zijn dan in het geval van het logistische regressiemodel. In het grafiekje rechtsonder is te zien dat er uitbijters zijn die veel invloed hebben. In dit geval zijn dit de films: “Harry Potter and the Chamber of Secrets”, met 1.377.000 kijkers, en de film “Pirates of the Caribbean: The curse of the Black Pearl” met 1.438.000 kijkers. De residuen verder bestuderend, kan geconcludeerd worden dat dit model niet erg goed is in het voorspellen van hoge kijkcijfers.

Vergelijking MSE twee modellen

Behalve de plaatjes hierboven kun je de modellen ook vergelijken aan de hand van de Mean Square Error (MSE). Dit is de gemiddelde kwadratische fout, waarbij de fout is gedefinieerd als het verschil tussen de voorspelde waarde door het model en de “werkelijke waarde”.

	MSE	Wortel (MSE)
Logistische regressie	1,34E+10	115695
Lineaire regressie	1,35E+10	116238

Tabel 15: MSE van de twee geschatte modellen

In tabel 15 is de MSE uitgerekend. Om een gevoel te krijgen voor de getallen is de wortel van de MSE ook uitgerekend, omdat deze dezelfde dimensie heeft als het aantal kijkcijfers. Deze kan worden gezien als een soort standaarddeviatie. Het verschil is niet heel groot, maar het logistische regressiemodel is wel beter.

Conclusie en verbeteringen

De conclusie uit de residu analyse en uit tabel 15 is dat het logistische regressiemodel de voorkeur verdient. Dit is eigenlijk ook wat je intuïtief zou verwachten.

Een verbetering van het model zou je kunnen krijgen door extra verklarende variabelen toe te voegen. In de data die is gebruikt zat niet genoeg informatie over de programma's die de concurrentie uitzendt (heel veel missende waarden), terwijl dit wel erg verklarend zou kunnen zijn (denk aan voetbaluitzendingen, "Boer zoekt vrouw"). Dit probleem wordt wel deels verholpen door de tijdslotvariabelen: vaak worden door de concurrentie tijdens een bepaald tijdslot dezelfde programma's uitgezonden, waardoor de tijdslotvariabele de rol overneemt van de concurrentie variabelen. Andere verklarende variabelen die niet beschikbaar waren zijn: hoeveel een film al is uitgezonden op tv en (heel belangrijk) het aantal kijkcijfers dat de film had toen het de vorige keer werd uitgezonden. Voor een beginnende analyse hiervan, wordt de lezer naar de laatste paragraaf van dit werkstuk, "aankooprijks van een film", verwezen.

Een andere verbetering zou het gebruik van "variaten"(variabelen met continue waarden) in plaats van factoren kunnen zijn. In de modellen die nu zijn geschat, zijn variabelen als "aantal bioscoopbezoekers" en "MovieMeter Ranking" niet als variaat (continue waarden) gebruikt, maar "in stukjes gehakt" tot factoren. Doordat je van deze continue variabelen discrete variabelen maakt, verlies je informatie. De reden dat dit toch gedaan is, is dat er veel ontbrekende waarden in de gebruikte database zitten. Als je het model zou willen schatten met "variaten", zou je records met ontbrekende waarden weg moeten laten. Hierdoor zouden maar 170 films van de 346 overblijven. Dit model is wel geschat, maar had minder voorspellende waarde dan het model waar factoren zijn gebruikt. Dit probleem is op te lossen door de database beter te vullen.

Optimalisatie van reclameopbrengsten door middel van “scheduling”

Gegeven het statistische model, wordt hier beschreven hoe de films het beste kunnen worden “verdeeld” over de beschikbare tijd. In dit hoofdstuk zal hiervoor een algemene heuristiek worden beschreven. Het statistische model is hierbij het belangrijkste uitgangspunt.

Variabelen die een rol spelen bij de kostenkant van dit planningsprobleem zijn:

- De uitzendrechten (eenmalig) van een film,
- De uitzendkosten (per keer dat de film wordt uitgezonden) van een film

Variabelen die een rol spelen bij de opbrengstenkant van dit planningsprobleem zijn:

- De reclameopbrengsten die horen bij dit tijdstip, en de eigenschappen van de film.

Hoe de samenhang is tussen de reclameopbrengsten en de kijkcijfers is beschreven in het hoofdstuk ‘relatie tussen kijkcijfers en reclame inkomsten’. Om dit planningsprobleem op te lossen beginnen we met een makkelijk model, en breiden dit model vervolgens uit.

Als makkelijk voorbeeld beginnen we met de volgende aannames: dat er slechts 1 reclaimedoelgroep bestaat, en dat de reclame-inkomsten lineair afhangen van de kijkcijfers. Ook kan elke film 1 keer worden uitgezonden, en zijn de kijkcijfers onafhankelijk van de andere films die ingepland zijn.

In dit geval kan het probleem worden gezien als een “knapzakprobleem”. Je wilt de totale winst maximaliseren. Deze winst wordt verkregen door de totale reclame-inkomsten te verminderen met de totale kosten voor het uitzenden van de films. De restricties hierbij zijn dat je niet meer films uit kunt zenden dan het aantal “tijdsloten” dat je tot je beschikking hebt. Het aantal films dat je hierbij tot je beschikking hebt om uit te kiezen, wordt groter verondersteld dan het aantal tijdsloten. Het aantal films dat je tot je beschikking hebt worden in deze paragraaf “de portefeuille” genoemd. Voor de uitzendkosten van een film per uitzending, wordt verondersteld dat deze alleen een rol spelen in de “opbrengstenfunctie”, en er wordt dus geen maximum budget verondersteld, omdat de uitzendkosten van een film direct terugverdiend worden door de reclameopbrengsten tijdens de film. In formulevorm kan het probleem als volgt worden

beschreven:

$$\text{Maximaliseer } \sum_{t=1}^T \sum_{k=1}^K p_{k,t} x_{k,t} - w_k x_{k,t}$$

$$\text{Onder } \sum_{k=1}^K x_{k,t} = 1 \text{ voor } t = 1, \dots, T$$

$$\text{en } \sum_{t=1}^T \sum_{k=1}^K x_{k,t} = T$$

$$\text{en } x_{k,t} \in \{0,1\} \text{ voor } k = 1, \dots, K \text{ en } t = 1, \dots, T$$

51)

In formulering 51) stelt $p_{k,t}$ de verdiende reclameopbrengsten van film k op tijdstip t voor. In dit model is de veronderstelling gemaakt dat $p_{k,t}$ een lineaire relatie heeft met de kijkcijfers. $p_{k,t}$ kan dus worden verkregen door de kijkcijfers te voorspellen aan de hand van het eerder beschreven statistische model, en dit getal vervolgens met een constante te vermenigvuldigen. w_k zijn de uitzendkosten voor film k (deze worden verondersteld niet van het uitzendtijdstip t af te hangen). Deze worden bekend verondersteld. $x_{k,t}$ is de variabele die aangeeft of film k wel of niet wordt uitgezonden op tijdstip t . De restricties geven aan dat elke film slechts 1 keer mag worden uitgezonden en dat er in totaal T films moeten worden uitgezonden zodat alle tijdsloten “vol zitten”.

Een oplossingsheuristiek voor dit probleem is de volgende:

1. Voor elke film in de portefeuille, en voor elk uitzendtijdstip, worden de reclameopbrengsten uitgerekend. Hier worden de bijbehorende uitzendkosten afgetrokken. Dit getal wordt gedefinieerd als de winst.
2. Bovenstaande “winsten” worden gesorteerd van hoog naar laag. Dit noemen we een gesorteerde lijst.
3. De tijdsloten worden gevuld, dit gebeurt door bovenaan de gesorteerde lijst te beginnen, deze film (met de grootste winst) wordt geplaatst in bijbehorend tijdslot.
4. De geplaatste film, en alle films in de gesorteerde lijst die hetzelfde tijdslot hebben, of hetzelfde filmnummer worden uit de lijst verwijderd.
5. Als alle tijdsloten vol zitten, wordt gestopt, anders worden de stappen vanaf stap 3 herhaald.

Uitbreiding: Afhankelijkheid en het meerdere keren uitzenden van een film

Bovenstaande model kan uitgebreid worden, indien uit het statistische model blijkt dat de kijkcijfers van een film afhankelijk zijn van de andere ingeplande films. Je zou je kunnen voorstellen dat er een soort ‘verzadiging’ bij het kijkerspubliek optreedt indien er te vaak achter elkaar een film van hetzelfde genre wordt uitgezonden. Dit zou kunnen blijken uit verder statistisch onderzoek. Maar een nog belangrijker, en intuïtief veel duidelijkere, afhankelijkheid ontstaat als je de mogelijkheid in het model inbouwt om dezelfde film meerdere keren uit te zenden. Als dezelfde film vlak achter elkaar uitgezonden wordt, zou je verwachten dat dit (met name de tweede uitzending) ten koste gaat van de kijkcijfers en dat er dus minder mensen kijken. Zoals eerder is vermeld, missen in de dataset die gebruikt werd bij dit werkstuk, de variabelen “aantal kijkers vorige uitzending” en “tijd tussen nu en uitzending vorige keer”. Dit zijn twee variabelen die waarschijnlijk erg verklarend zijn.

Het model wordt nu:

$$\text{Maximaliseer } \sum_{t=1}^T \sum_{k=1}^K p_{k,t} x_{k,t} - w_k x_{k,t}$$

52)

$$\text{Onder } \sum_{t=1}^T \sum_{k=1}^K x_{k,t} = T$$

$$\text{en } x_{k,t} \in \{0,1\} \text{ voor } k = 1, \dots, K \text{ en } t = 1, \dots, T$$

In bovenstaande heuristiek hangt $p_{k,t}$ af van de overige films die ingepland zijn. Een speciaal geval is als op een ander tijdslot dezelfde film wordt uitgezonden, zodat de film meerdere keren wordt herhaald. De heuristiek moet nu worden aangepast, omdat het in de eerder beschreven heuristiek niet mogelijk was een film meerdere keren uit te zenden, en ook omdat de kijkcijfers nu afhankelijk zijn van elkaar. Het grootste verschil met de vorige heuristiek is dat elke keer nadat een film wordt geplaatst, de kijkcijfers van de overige films opnieuw moeten worden voorspeld.

De heuristiek wordt nu:

1. Voor elke film, en voor elk uitzendtijdstip dat nog beschikbaar is, worden de reclameopbrengsten uitgerekend. Hier worden de bijbehorende uitzendkosten afgetrokken. Dit getal wordt gedefinieerd als de winst.
2. De “winsten” worden gesorteerd van hoog naar laag.
3. De tijdsloten worden gevuld, dit gebeurt door bovenaan de gesorteerde lijst te beginnen, deze film (met de grootste winst) wordt geplaatst in bijbehorend tijdslot.

4. De geplaatste film, en alle films in de lijst die hetzelfde tijdslot hebben worden uit de lijst verwijderd.
5. Als alle tijdsloten vol zitten wordt de heuristiek gestopt, anders worden de stappen herhaald vanaf stap 1.

Uitbreiding: verschillende reclamedoelgroepen

In het model dat beschreven wordt in formule 51 en formule 52 is nog geen rekening gehouden met het feit dat de reclames worden ingekocht op verschillende doelgroepen. Wat zijn de reclameopbrengsten ($p_{k,t}$ in het eerder beschreven model) als er meerdere reclamedoelgroepen zijn? Dit wordt in deze paragraaf besproken.

Reclame-inkomsten als responsvariabele

Een methode om dit probleem aan te pakken is om de relatie tussen kijkcijfers achterwege te laten en de reclame-inkomsten als de responsvariabele te nemen in het statistische model. Deze aanpak heeft echter als nadeel dat je informatie verliest. Ten eerste verlies je informatie omdat je niet weet welke doelgroep wel of niet heeft gekeken. Ten tweede verlies je informatie omdat je niet weet of deze inkomsten tot stand zijn gekomen doordat er veel mensen hebben gekeken of doordat de film in trek is bij adverteerders.

Een ander nadeel kan zijn dat films die goed scoren voor veel kijkende doelgroepen, en GRP's opleveren, "bevoordeeld" worden bij het inplannen. In dat geval zou een probleem kunnen ontstaan: er worden alleen films geselecteerd voor deze "veelkijkende doelgroepen", en in de planning van films geen rekening wordt gehouden met reclames voor doelgroepen die minder kijken, maar waar wel reclames voor moeten worden uitgezonden. Dit probleem zal dan moeten worden opgelost door een geavanceerde planningsheuristiek te gebruiken.

Verskillende statistische modellen voor verschillende groepen

Een andere aanpak is om meerdere statistische modellen te maken. Elk model voorspelt voor elke doelgroep hoeveel kijkers (in GRP's) er naar de film zullen kijken. Met behulp van deze GRP's kunnen dan de reclameopbrengsten van een film worden berekend, per doelgroep. Voor een film is het altijd optimaal om reclames uit te zenden van de doelgroep die het meeste reclameopbrengsten oplevert (de doelgroep waarin de film het beste "scoort"). Vervolgens worden de films weer ingepland met behulp van het algoritme uit de vorige paragraaf.

Er zal bij het inplannen rekening moeten worden gehouden met de verdeling van de doelgroepen over de totale reclamezendtijd. Als je hier geen rekening mee houdt, is het mogelijk dat de heuristiek bepaalde reclamedoelgroepen te vaak of juist te weinig indeelt. Het is bijvoorbeeld bekend dat de meest gewilde reclamedoelgroep de leeftijd 20-49 is, terwijl er juist het meest televisie gekeken wordt door 50-plussers.

Voorbeeld: Vooraf maakt SBS de inschatting dat tijdens haar films de verdeling van de doelgroepen over de totale reclamezendtijd over een bepaalde periode de volgende is: 20% van de reclames zal zijn voor de doelgroep 6 jaar en ouder, 30% voor de doelgroep 20-49, 30% voor mannen tussen de 20-49, en 20% voor vrouwen tussen de 20-49. De optimale oplossing, zonder rekening te houden met deze restrictie, plant de reclames als volgt in: 20% van de reclames zal zijn voor de doelgroep 6 jaar en ouder, 0% voor de doelgroep 20-49, 30% voor mannen tussen de 20-49, en 50% voor vrouwen tussen de 20-49. Er is een probleem ontstaan.

Als dit probleem zich voordoet, zou je de volgende oplossingen kunnen onderzoeken:

- In de heuristiek zou kunnen worden ingebouwd dat wordt gestopt met het toewijzen van reclames aan een bepaalde doelgroep, als het maximum hiervoor in de planning is bereikt. In bovenstaand voorbeeld: als de 30% van de reclames voor mannen tussen de 20 en 49 jaar is bereikt tijdens het inplannen, wordt aan deze reclamegroep tijdens het verdere inplannen geen reclames meer toebedeeld. Nadeel van deze methode zou kunnen zijn dat het “gierig” is, en dat je op het einde van het toewijzen bijvoorbeeld overblijft met vrouwenfilms en alleen nog reclames hebt toe te wijzen van mannen tussen de 20-49 jaar.
- Een andere manier is om eerst de heuristiek te gebruiken zonder de restrictie, en achteraf te gaan schuiven met de reclamezendtijd. In bovenstaand voorbeeld zou je dan achteraf reclames van de doelgroep 20-49 jaar moeten gaan “ruilen” voor andere reclames, zodanig dat de totale reclameopbrengsten niet te veel afnemen.
- Een andere manier om dit probleem op te lossen is het aanpassen van de index van de doelgroepen, zodat het rendabel wordt een bepaalde doelgroep uit te zenden. Dit kan ook gebeuren in combinatie met een van de vorige twee genoemde punten.
- Een andere oplossing is het nieuw aankopen van films voor die hoge kijkcijfers hebben voor doelgroepen die in de huidige planning nog niet goed scoren. Dit is bijvoorbeeld het geval, als je gedwongen bent reclames voor een bepaalde doelgroep (neem bijvoorbeeld

de 50+ categorie), uit te zenden tijdens films die veel beter scoren bij andere doelgroepen. (Maar je moet deze reclames daar toch uitzenden omdat je ze “ergens kwijt moet”). In dat geval loont het misschien een film aan te kopen die heel goed scoort bij de 50+ doelgroep. Deze oplossing kan uiteraard ook in combinatie met de voorgaande oplossingen. Zie hiervoor ook de volgende paragraaf.

Het aankoopbedrag van een film

Een deelvraag die werd gesteld in de probleemstelling was het bepalen van een geschikte aankoopprijs voor een film. Nu we een statistisch model hebben voor het bepalen van het aantal kijkers, en weten hoe we de films kunnen inroosteren kunnen we de aankoopprijs van een film bepalen.

Bijbehorend bij de periodeplanning zoals die gemaakt is in de vorige paragraaf, is een bedrag aan totale opbrengsten. Een film zal alleen toegevoegd (aangekocht) worden aan de bestaande “portefeuille” als de totale reclameopbrengsten hierdoor zullen stijgen. Het breakeven punt van de aankoopprijs is in dat geval het aantal euro’s dat de totale opbrengsten stijgt door het toevoegen van de film aan de portefeuille.

Het aantal euro’s dat de opbrengst in een periode omhoog gaat door het toevoegen van de film, kan worden geschat door het statistische model te gebruiken, en vervolgens de periodeplanning opnieuw te maken, inclusief de betreffende film. Echter, een film kan meerdere keren worden uitgezonden in de toekomst, en er zal dus ook een inschatting moeten worden gemaakt hoeveel de film de opbrengsten in de toekomst zal laten stijgen. Met andere woorden, er zal een inschatting moeten worden gemaakt hoe lang de film “de moeite waard” is om uit te zenden, en goed genoeg is om in de periodeplanningen in de toekomst terecht te komen.

Helaas missen in de filmdatabase die gebruikt is bij de statistische analyse, de variabelen “kijkcijfers vorige keer” en “aantal keer uitgezonden”. Wel staan er 15 dubbele films, die in 2005 en 2006 twee keer zijn uitgezonden in. Deze zijn in tabel 16 te zien. Deze tabel staat niet in de bijlage omdat het leuk is om een keer een tabel met filmtitels te zien:

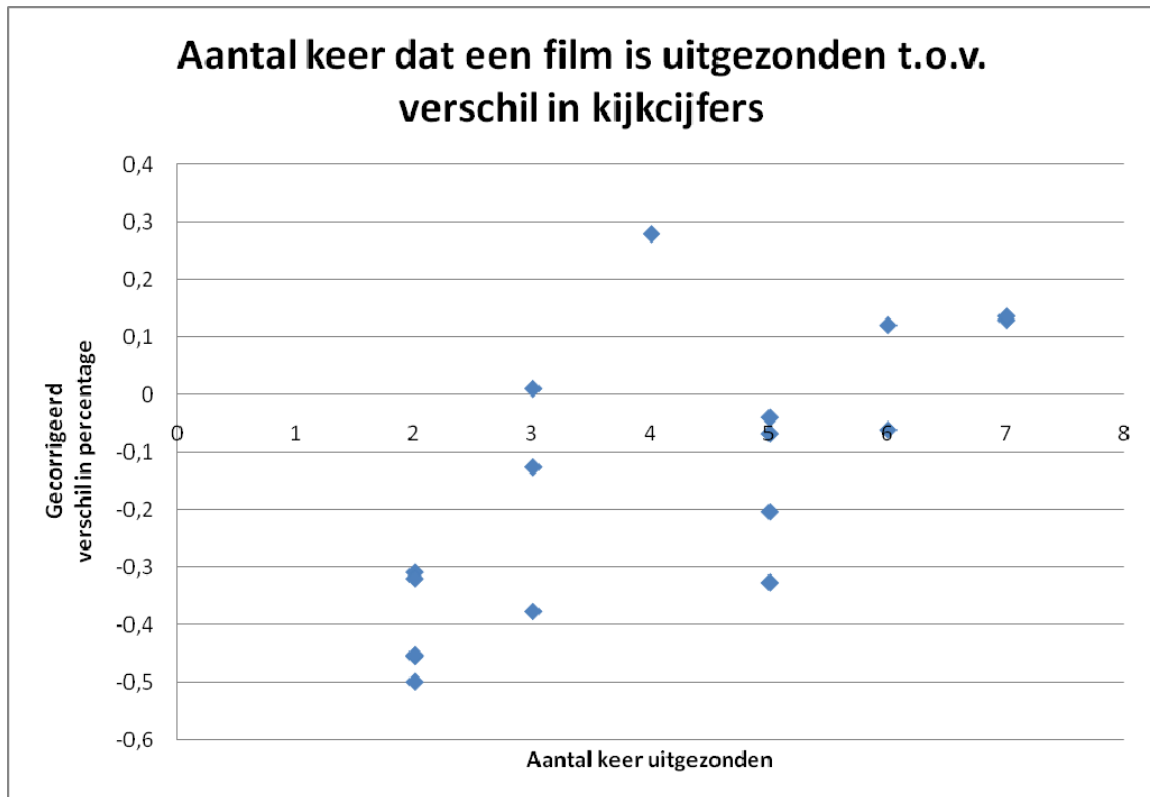
Programma (tussen haakjes het aantal keer uitgezonden)	Absoluut aantal kijkers	Voorspelde waarde door model
ACE VENTURA PET DETECTIVE (2)	510.000	426.611
ACE VENTURA PET DETECTIVE (3)	514.000	426.611
ACE VENTURA WHEN NATURE... (3)	378.000	442.059
ACE VENTURA WHEN NATURE... (4)	501.000	442.059
A PERFECT MURDER (6)	409.000	365.490
A PERFECT MURDER (7)	456.000	365.490

BASIC (1)	845.000	778.065
BASIC (2)	604.000	778.065
BONE COLLECTOR (6)	458.000	393.633
BONE COLLECTOR (7)	530.000	412.365
DEEP RISING (4)	510.000	418.221
DEEP RISING (5)	354.000	399.222
DEMOLITION MAN (4)	547.000	578.746
DEMOLITION MAN (6)	511.000	578.746
GANGS OF NEW YORK (1)	587.000	425.153
GANGS OF NEW YORK (2)	343.000	393.633
GHOST SHIP (1)	568.000	434.572
GHOST SHIP (2)	301.000	365.490
GOLDENEYE (4)	894.000	985.006
GOLDENEYE (5)	826.000	985.006
HARRY POTTER AND THE CHAMBER OF SECRETS (1)	1377.000	1.116.117
HARRY POTTER AND THE CHAMBER OF SECRETS (2)	917.000	1.013.857
LETHAL WEAPON 3 (4)	872.000	985.006
LETHAL WEAPON 3 (5)	448.000	601.046
NICK OF TIME (5)	452.000	456.907
NICK OF TIME (6)	449.000	399.222
RING (1)	448.000	441.534
RING (2)	215.000	408.800
THOMAS CROWN AFFAIR (4)	921.000	856.542
THOMAS CROWN AFFAIR (5)	364.000	474.771
TOMORROW NEVER DIES (2)	1162.000	955.379
TOMORROW NEVER DIES (3)	831.000	985.006
WHAT'S THE WORST THAT COULD HAPPEN (2)	365.000	425.422
WHAT'S THE WORST THAT COULD HAPPEN (3)	311.000	425.422

Tabel 16: kijkdichtheid van dubbel uitgezonden films.

In tabel 16 staat per film: de titel, hoeveel deze al op de Nederlandse tv te zien is geweest, het werkelijk aantal kijkers, en het voorspelde aantal kijkers. Leuk is het om bij de film “Thomas Crown Affair” het effect van het uitzendtijdstip te zien. Dit verschil is te verklaren doordat de film eerst prime time op maandag bij SBS6 werd uitgezonden en later prime time op donderdag op NET5. Ook zie je dat de kijkcijfers van de film heel erg goed te gebruiken zijn om de kijkcijfers voor de volgende keer te bepalen. Jammer dat deze variabele niet in de database zat!

Als gecorrigeerd wordt voor tijd, maand en het aantal kijkers van het vorige programma, zijn de kijkcijfers 14% lager voor de films die later uitgezonden zijn. Als je bovenstaande films bekijkt, krijg je sterk de indruk dat de verklarende variabele “tv-première” een goed verklarende variabele is, omdat voor alle films die voor het eerst worden uitgezonden (met een 1 achter hun titel) de kijkcijfers hoger uitvallen dan de voorspelling. Deze variabele zat helaas ook niet in de database, en is niet gebruikt bij de statistische analyse. Om een eerste indruk te geven van het verschil in kijkcijfers tussen films die achter elkaar uitgezonden zijn, plotten we de volgende grafiek:



Figuur 11: Aantal keer dat een film is uitgezonden t.o.v. het verschil in kijkcijfers met de vorige keer dat de film is uitgezonden

In Figuur 11 is op de x as te zien hoeveel een film is uitgezonden, en op de y-as is de toename of afname van het gecorrigeerde verschil in kijkcijfers. Er is gecorrigeerd voor uitzendtijdstip en uitzendmaand, en het aantal kijkers van het vorige programma. In bovenstaande grafiek is te zien dat vooral wanneer een film voor de 2^e keer wordt uitgezonden, de kijkcijfers een stuk lager zijn dan bij de tv-première. Hierna worden de kijkcijfers wel lager, maar is het effect minder groot, en in deze dataset worden de kijkcijfers zelf hoger als een film erg vaak is uitgezonden! Hoe de

kijkcijfers precies worden beïnvloed door het meerdere keren uitzenden van een film is een onderwerp voor verder onderzoek, omdat deze 15 datapunten te weinig zijn om iets uit te concluderen.

Als de relatie tussen het aantal keer uitzenden van een film en de kijkcijfers bekend is, kan ook een schatting worden gemaakt hoelang een film “de moeite waard is” om uit te zenden, en of de film in de toekomst in de periodeplanningen terecht komt. Ook voor de toekomstige periodeplanningen kan worden uitgerekend hoeveel de film de totale opbrengst van die planning verhoogt. Uit de (verdisconteerde) som hiervan volgt het breakeven punt van de aankoopprijs van de film. Uiteraard geldt, hoe verder de prijs van de film onder dit breakeven punt zit, hoe beter.

Conclusie

De probleemstelling was om te onderzoeken welke films het beste op welk moment uitgezonden zouden moeten worden. Het doel hierbij is het maximaliseren van de reclame-inkomsten. In dit werkstuk zijn de stappen beschreven die zouden moeten worden ondernomen om deze vraag goed te beantwoorden. Enkele van deze stappen zijn ook (gedeeltelijk) ondernomen.

De reclameopbrengsten tijdens een bepaalde film, kunnen aan de hand van twee statistische modellen worden voorspeld: een statistisch model dat de kijkcijfers voorspelt, en een ander model dat deze kijkcijfers “vertaalt” naar reclameopbrengsten.

De kijkcijfers van een film kunnen redelijk goed worden voorspeld aan de hand van de eigenschappen van een film. Uit de data analyse blijkt dat voor het voorspellen van kijkcijfers een logistisch regressiemodel beter werkt dan een lineair regressiemodel. Het verschil tussen deze modellen is, kort gezegd, dat het logistische regressiemodel een multiplicatief model is, dat wil zeggen dat de kijkcijfers met een percentage veranderen aan de hand van de eigenschappen van een film. In het geval van het lineaire regressiemodel veranderen de kijkcijfers met een absoluut aantal als een eigenschap wijzigt.

De belangrijkste eigenschappen om de kijkcijfers van een film te voorspellen zijn in volgorde van “belangrijkheid”: het uitzendtijdstip, of de film een actiefilm of avonturenfilm is, het aantal bioscoopbezoekers dat de film had, of de film in de winter (maanden oktober t/m februari) werd uitgezonden, het aantal kijkers naar het voorafgaande programma, en als laatste de Moviemeter rating. Verbeteringen van dit model kunnen worden gemaakt door een beter gevulde database, de toevoeging van concurrentiegegevens (programma's die door de concurrentie werden uitgezonden tijdens de film), en toevoeging van de kijkcijfers van de vorige keer dat een bepaalde film werd uitgezonden.

Het model dat de kijkcijfers “vertaalt” naar de reclameopbrengsten is voor een deel deterministisch. Dit komt doordat de kijkcijfers via een systeem van indexen (vermenigvuldigingsfactoren) worden omgerekend naar een reclameprijs. Hierbij moet je denken aan factoren die per tijdstip en per maand verschillen. Het statistische gedeelte van dit model bestaat voornamelijk uit keuzes die de adverteerder maakt qua inkoopopties. Tijdens welke films besluit de adverteerder de “dure” inkoopopties te gebruiken? Helaas waren hiervan bij het maken van deze scriptie geen data aanwezig, en daarom kon dit niet worden onderzocht.

Bij het inroosteren van de films in een uitzendschema, dient rekening worden gehouden te worden met de verschillende reclamedoelgroepen. Dit kan worden gedaan door een inschatting te maken over welk deel van de totale reclamezendtijd voor welke doelgroep zal zijn. Vervolgens dient een statistisch model te worden gemaakt per reclamedoelgroep. Het inroosteren van films kan worden gezien als een “knapzakprobleem” (met uitbreidingen), en hiervoor is in het hoofdstuk “*optimalisatie van reclameopbrengsten door middel van scheduling*” een heuristiek (stappenplan) beschreven.

De laatste vraag die beantwoord is, is hoe je een inschatting kunt maken wat een “goede” aankoopprijs is van een film. Bij een periodeplanning hoort een totale reclameopbrengst in euro's. Het breakeven punt van de aankoopprijs voor een film is de som van het aantal euro's dat een film de periodeopbrengsten, nu en in de toekomst, doet stijgen. Uiteraard geldt: hoe lager de aankoopprijs, hoe beter. De verwachting is dat een film de eerste keer dat hij wordt uitgezonden meer kijkers trekt dan de tweede keer, en de tweede keer meer dan de derde keer etc. Om dit te onderzoeken is een verkennende analyse gedaan aan de hand van 15 films. Dit aantal bleek echter te weinig om conclusies aan te verbinden, dus verder onderzoek zal nodig zijn.

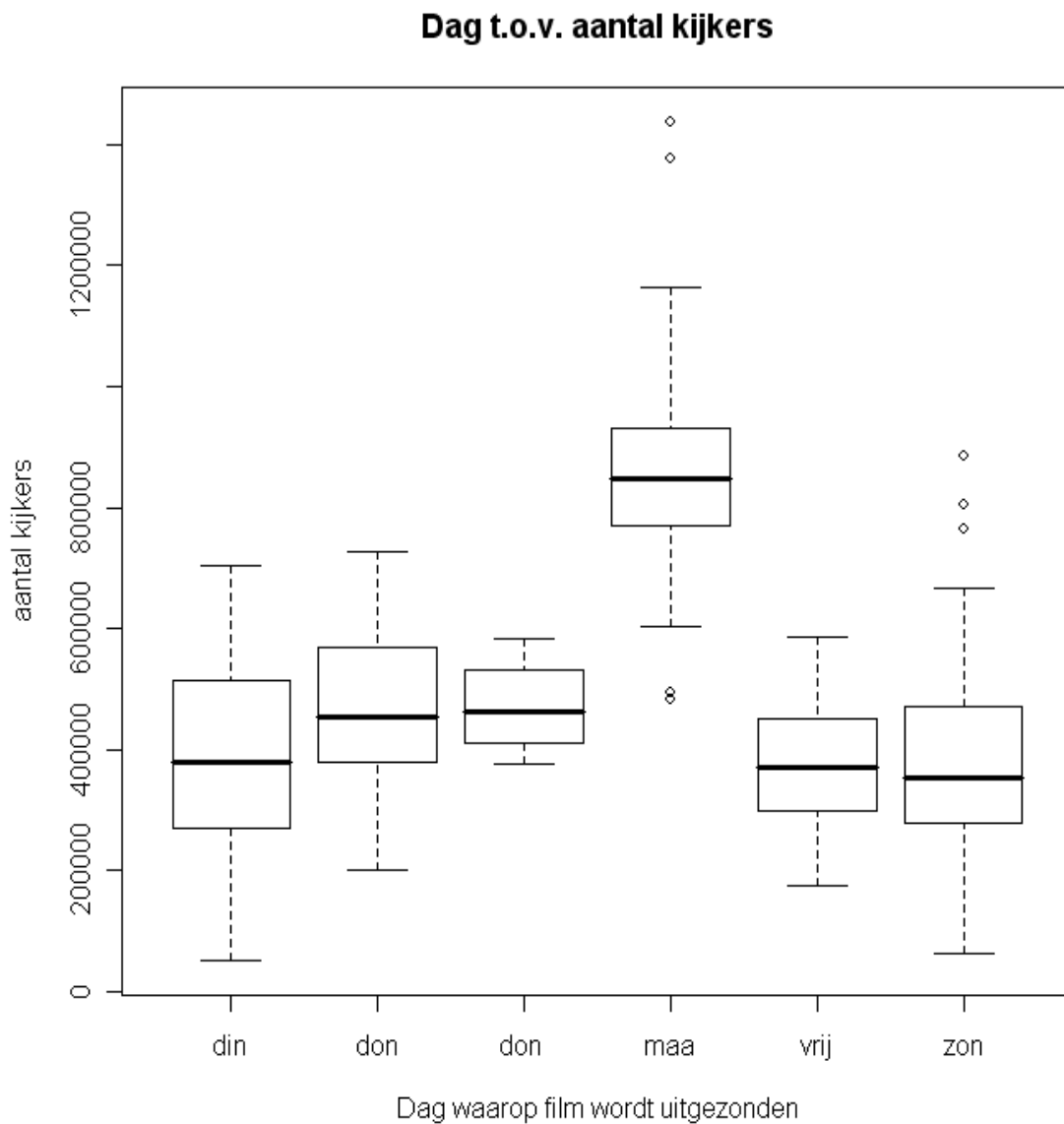
Literatuurlijst

- Intomart GfK, TV Times België, TV Times Nederland in samenwerking met SKO Nederland, (januari 2007) *Het kijkonderzoek, methodologische beschrijving*, Amstelveen.
- Stichting Kijkonderzoek, www.kijkonderzoek.nl, *kijkonderzoek, tv meting*, (december 2007a)
- Stichting Kijkonderzoek, (november 2007b) *SKO_Methoden: Berekening van het spotbereik via het kansmodel*
- STER, www.ster.nl, (november 2007a) *inkoopinformatie*
- SBS, www.adverterenbijbsbs.nl (2007), *Adverteren bij SBS*
- STER, www.ster.nl, (november 2007b) *dag tot dag schema Nederland 3*
- SPOT (Stichting Promotie Televisiereclame), (januari 2008), *Financieel Jaarverslag televisiereclame 2007*, Amstelveen.
- RTL, www.rtl.nl/service/rtlnederland/adverteren/televisie/spot, (2007) *Uw doelgroep, onze kijkers: inkoopbrochure spotzendtijd RTL Nederland 2007*
- Stichting Kijkonderzoek, (november 2007c) *SKO_Methoden: Indeling Sociale klassen*
- Stichting Kijkonderzoek (november 2007d) *SKO_Methoden: Commerciële doelgroepen*
- Oosterhoff, J, van der Vaart, A, (januari 2003), *Algemene statistiek*, Amsterdam
- Bain, L.J, Engelhard, M, (1992), *Introduction to probability and Mathematical Statistics*, Duxbury, California USA
- Wedel, M, den Boon, A.K, (1999) *Vuistregels voor kijkcijfers: het SCORE Model*, Groningen
- Muilwijk, J., Snijders, T.A.B. en Moors, J.J.A., (1992), *Kanssteekproeven*, Stenfert Kroese, Leiden.
- Kish, L, (1965), *Survey Sampling*. John Wiley, New York.
- Anderson, Feldblum, Modlin, Schrimacher, Schirmacher, Thandi, (May 2004), *A practitioners guide to Generalized Linear Models, A foundation for theory, interpretation and application*, Wattson Wyatt Worldwide
- Nelder, J.A, Wedderburn, R.W.M. (1972), *Generalized Linear Models*, Journal of the Royal Statistical Society A, 135, p.370-384
- Dobson, A, (2002), *An Introduction to Generalized Linear Models*, Chapman & Hall, Florida
- De Gunst (2005), *Statistical Models*, VU Bibliotheek, Amsterdam
- McCullagh and Nelder, (1989), *Generalized Linear Models*, 2nd Ed. Chapman & Hall

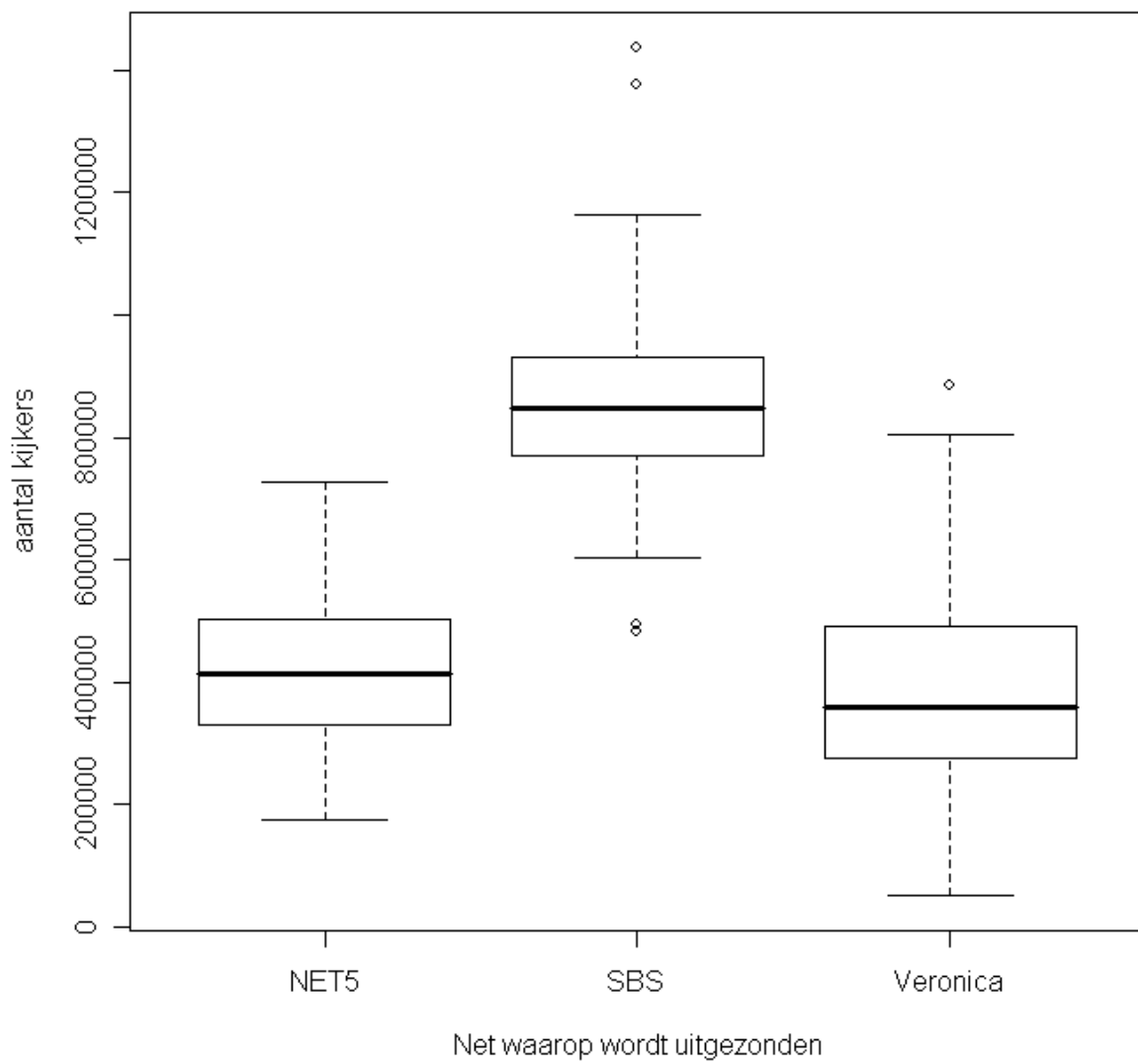
Appendix

Bijlage 1: Verkennende data analyse

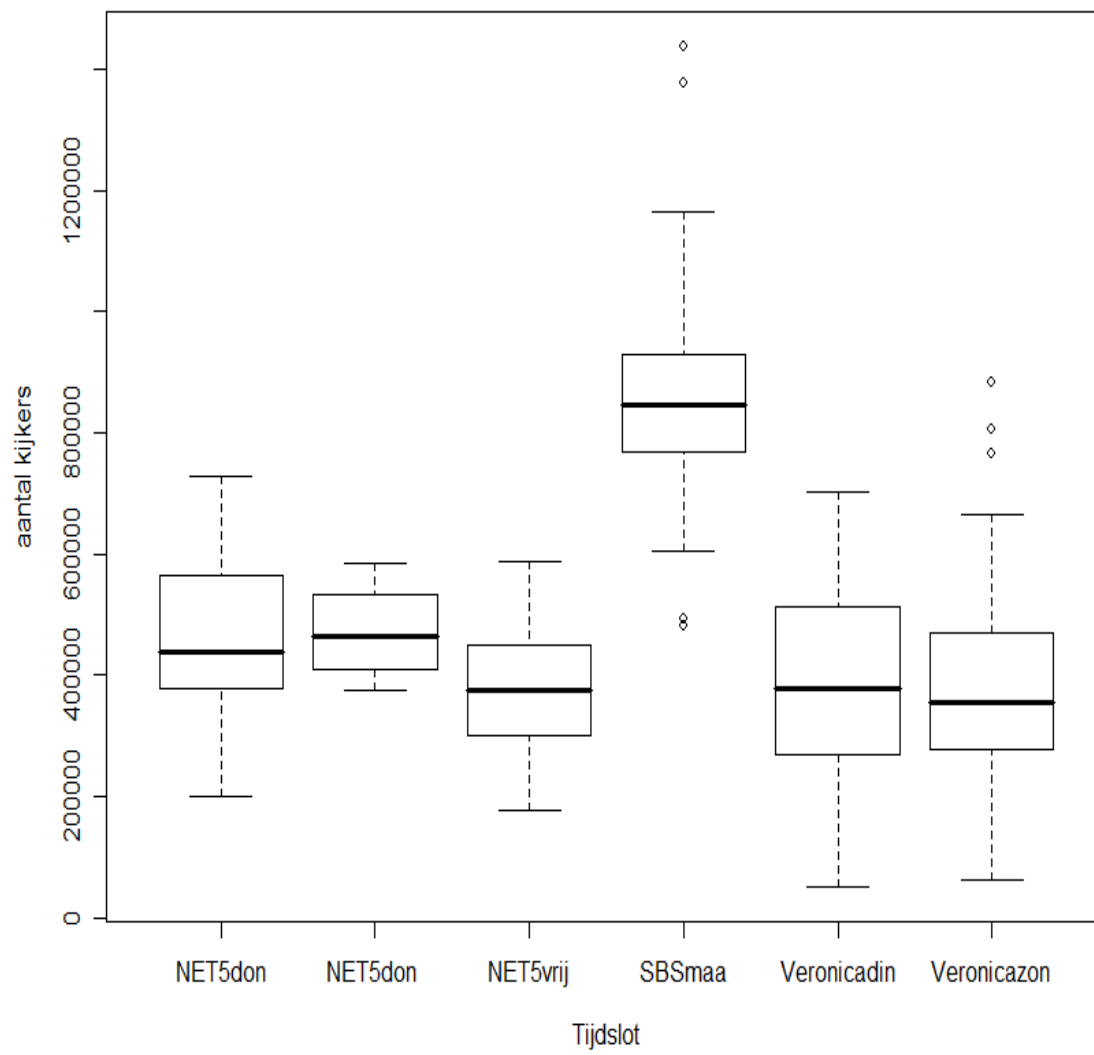
De volgende grafieken zijn gebruikt om een eerste indruk te krijgen van het verband tussen een verklarende variabele en de kijkcijfers. De verklarende variabele staat steeds op de x-as en de kijkcijfers, in absolute aantallen, op de y-as.



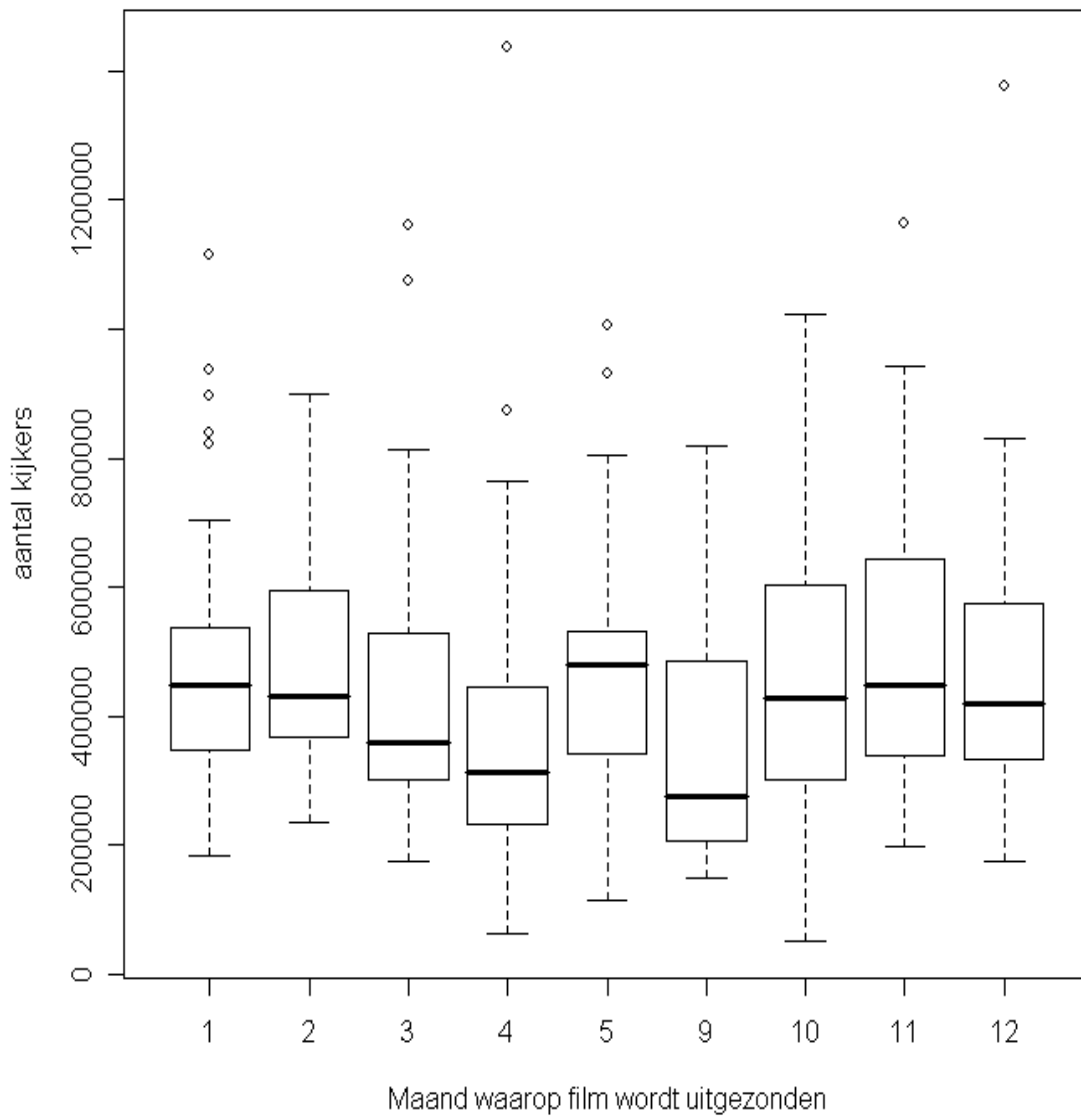
Net t.o.v. aantal kijkers



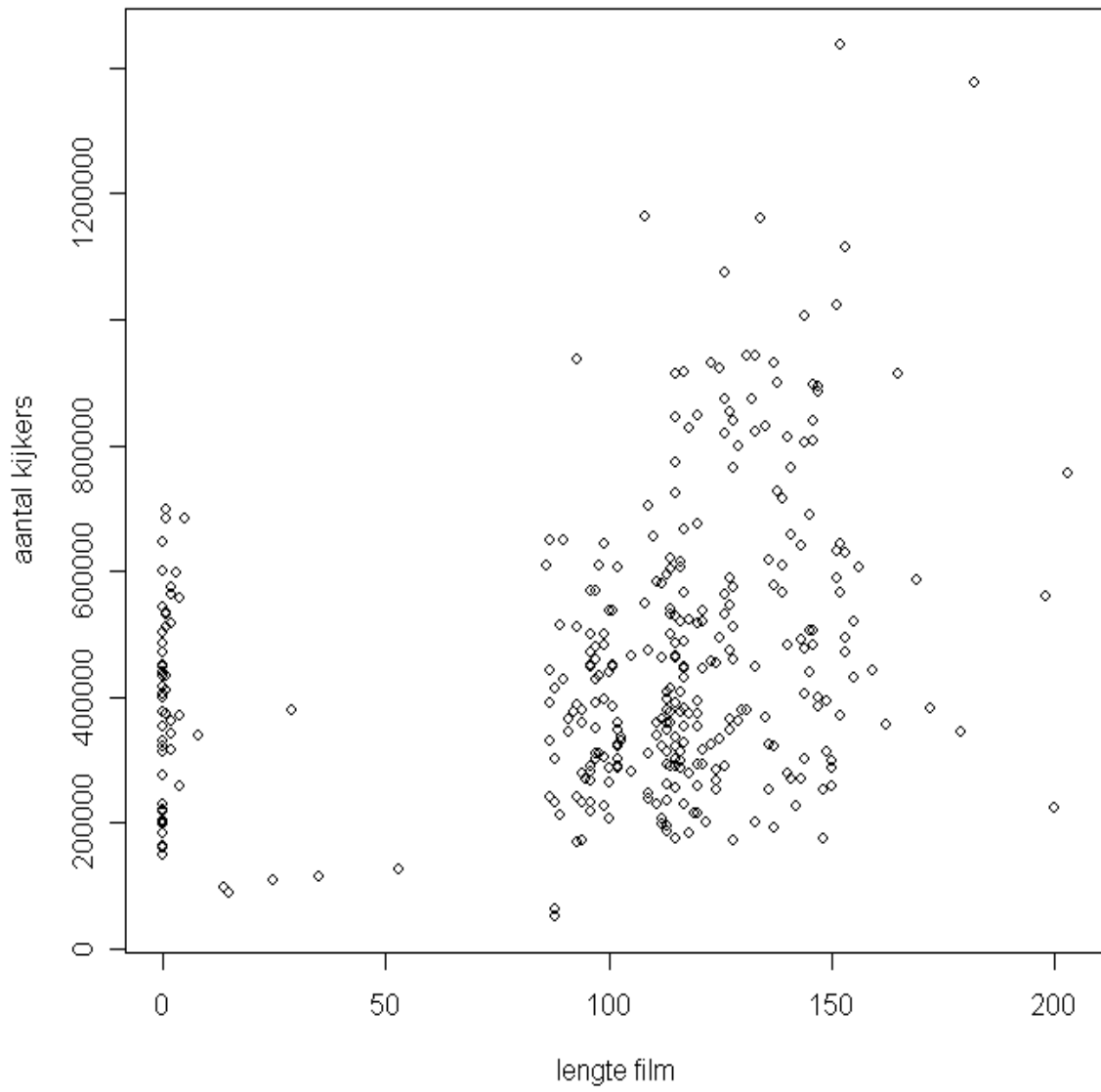
Tijdslot t.o.v. aantal kijkers



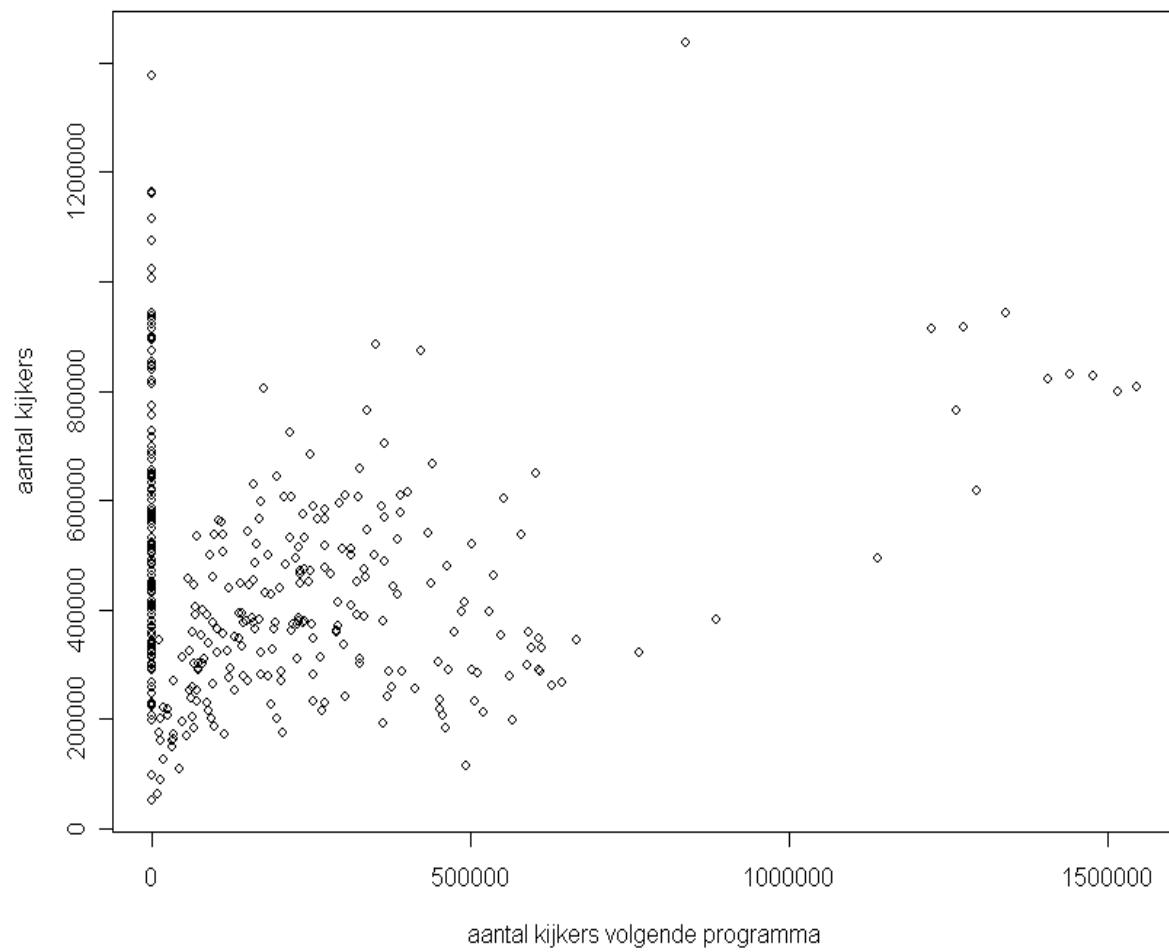
Maand t.o.v. aantal kijkers



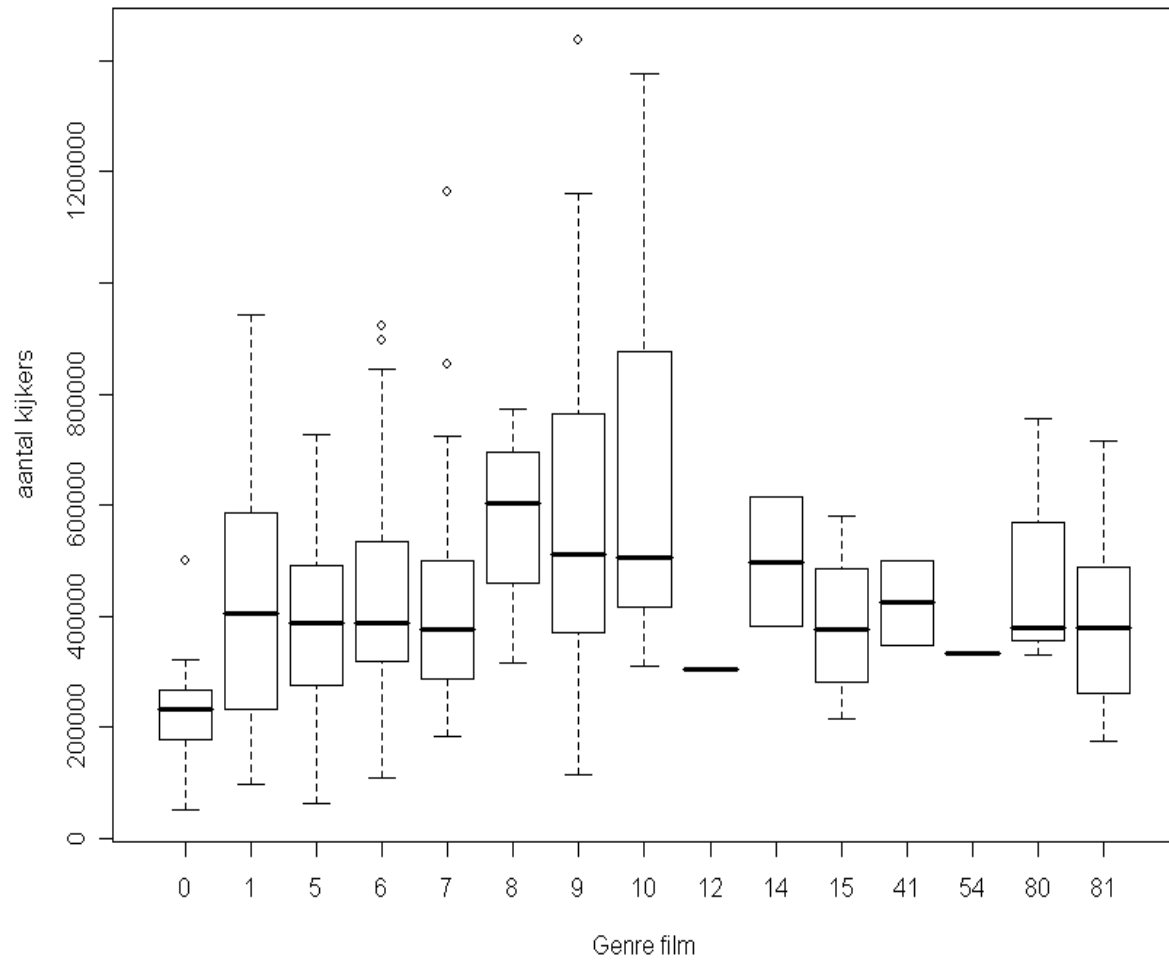
Lengte van de film in minuten t.o.v. aantal kijkers



Aantal kijkers volgende programma t.o.v. aantal kijkers

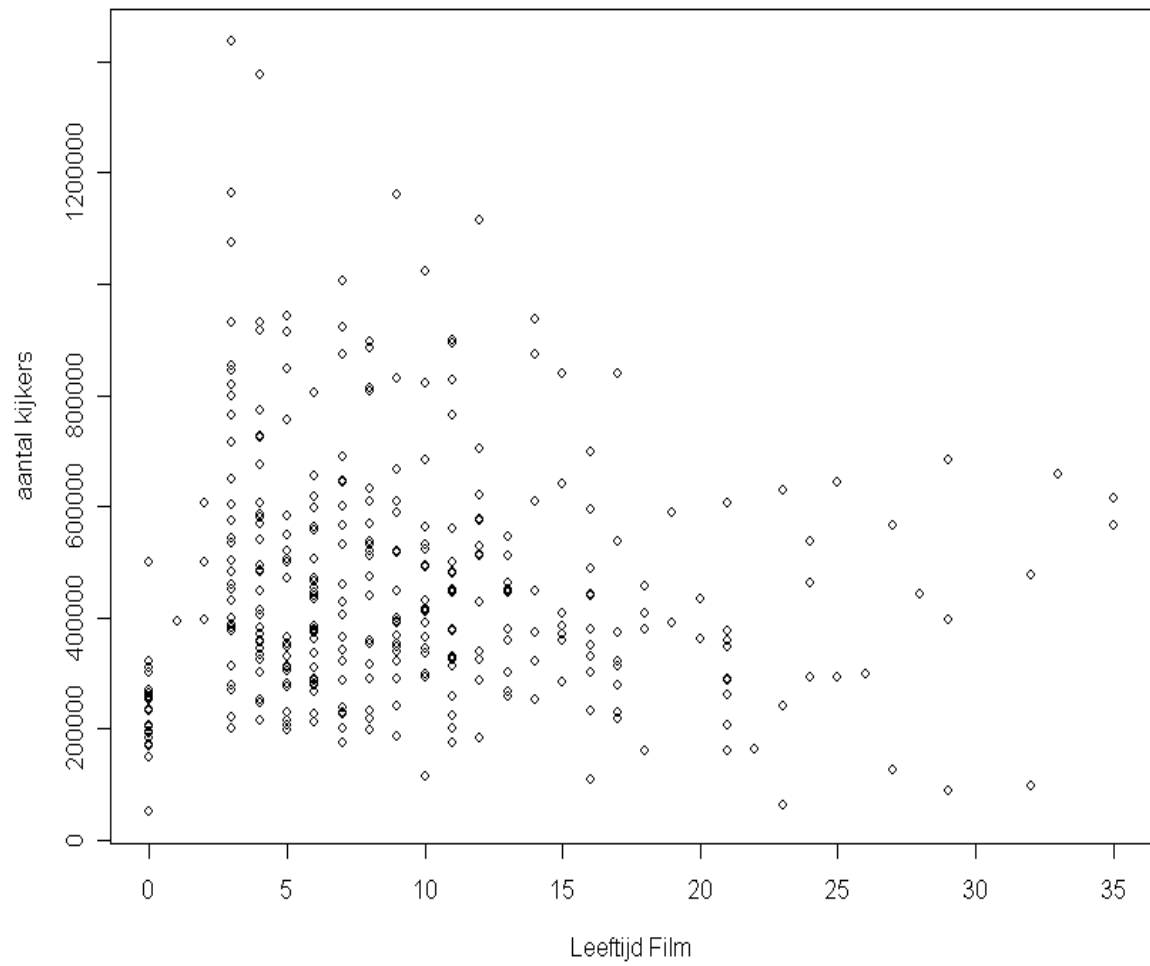


Genre t.o.v. aantal kijkers

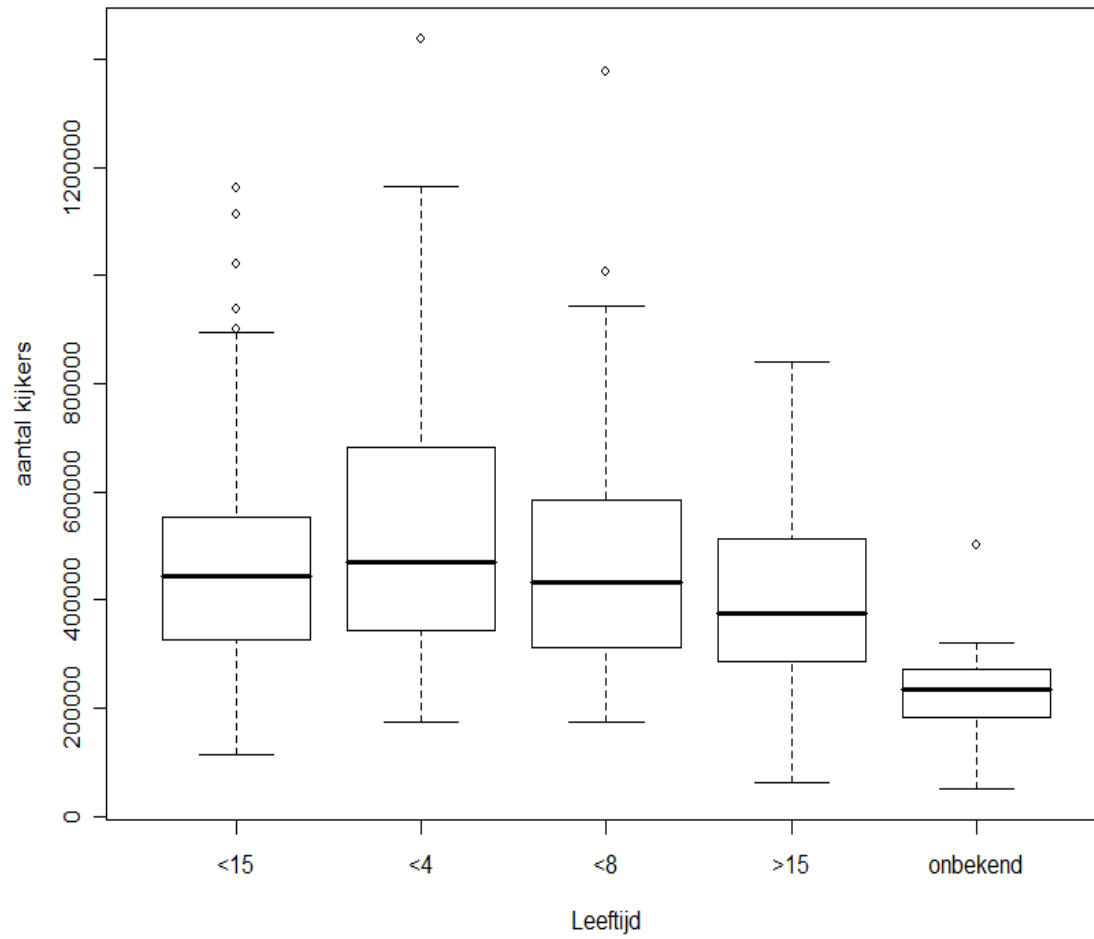


Opmerking: 8= romantiek, 9= actie, 10 = avontuur

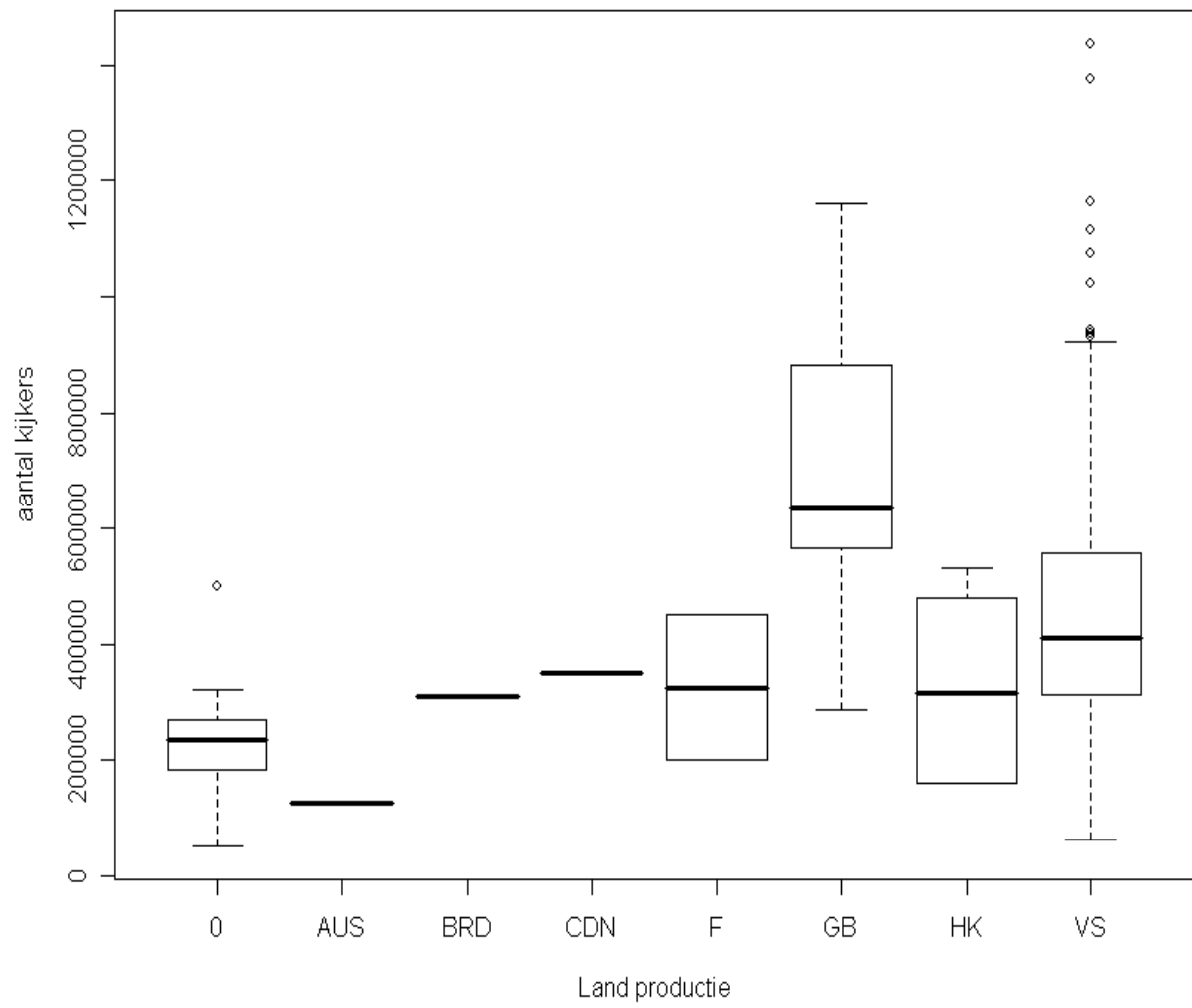
Leeftijd film t.o.v. aantal kijkers



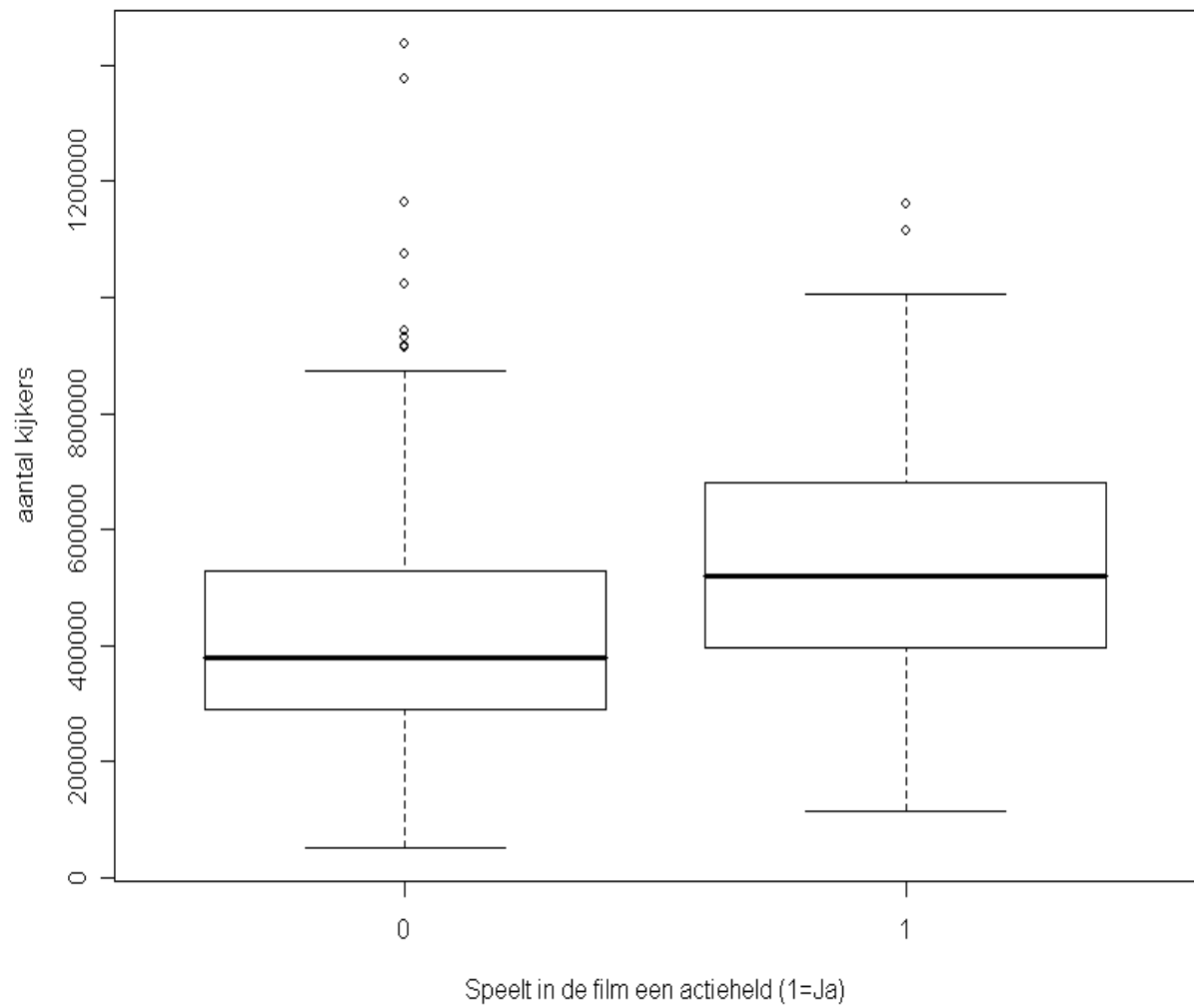
Leeftijd in categorieën t.o.v. aantal kijkers



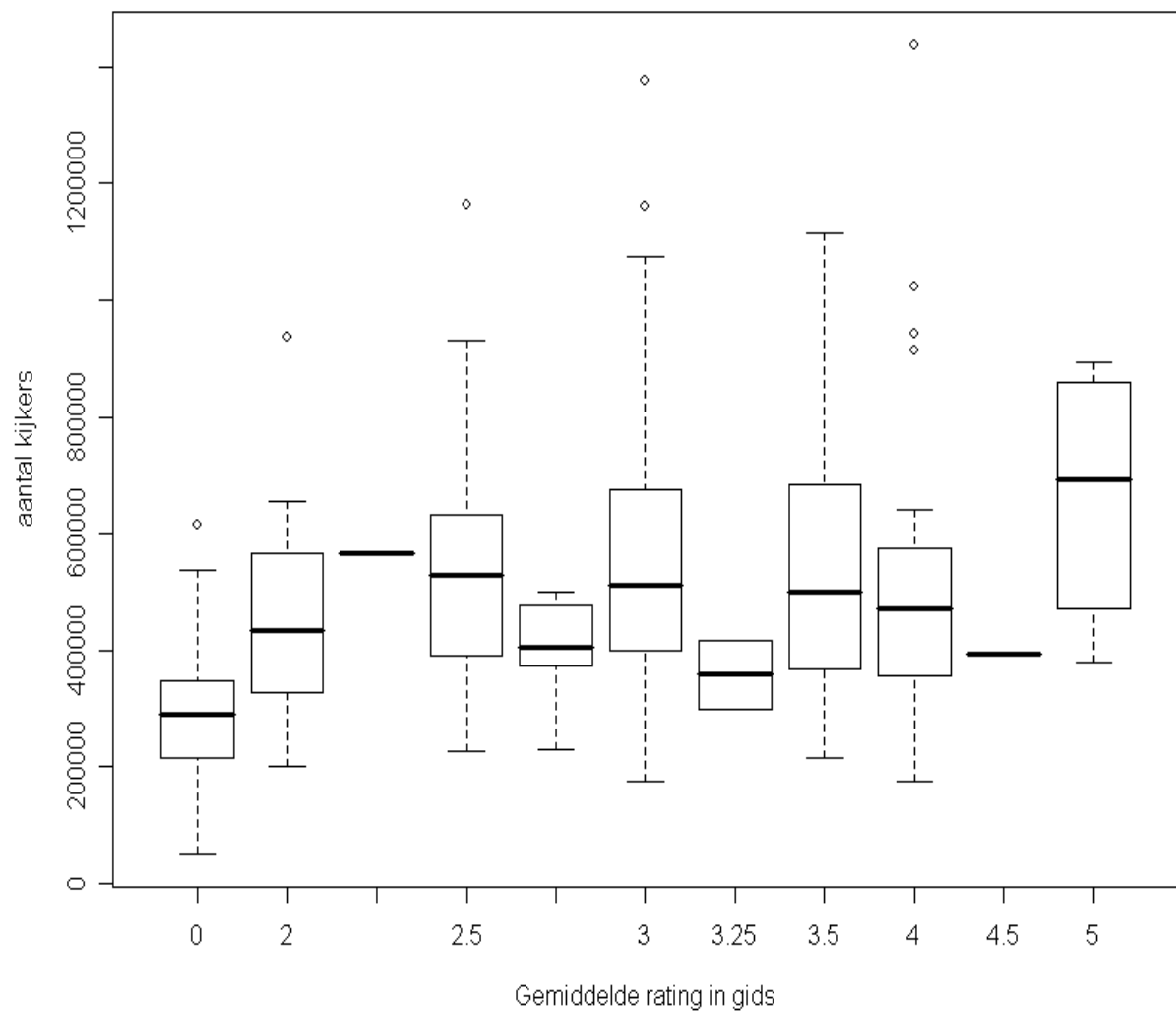
Land productie film t.o.v. aantal kijkers



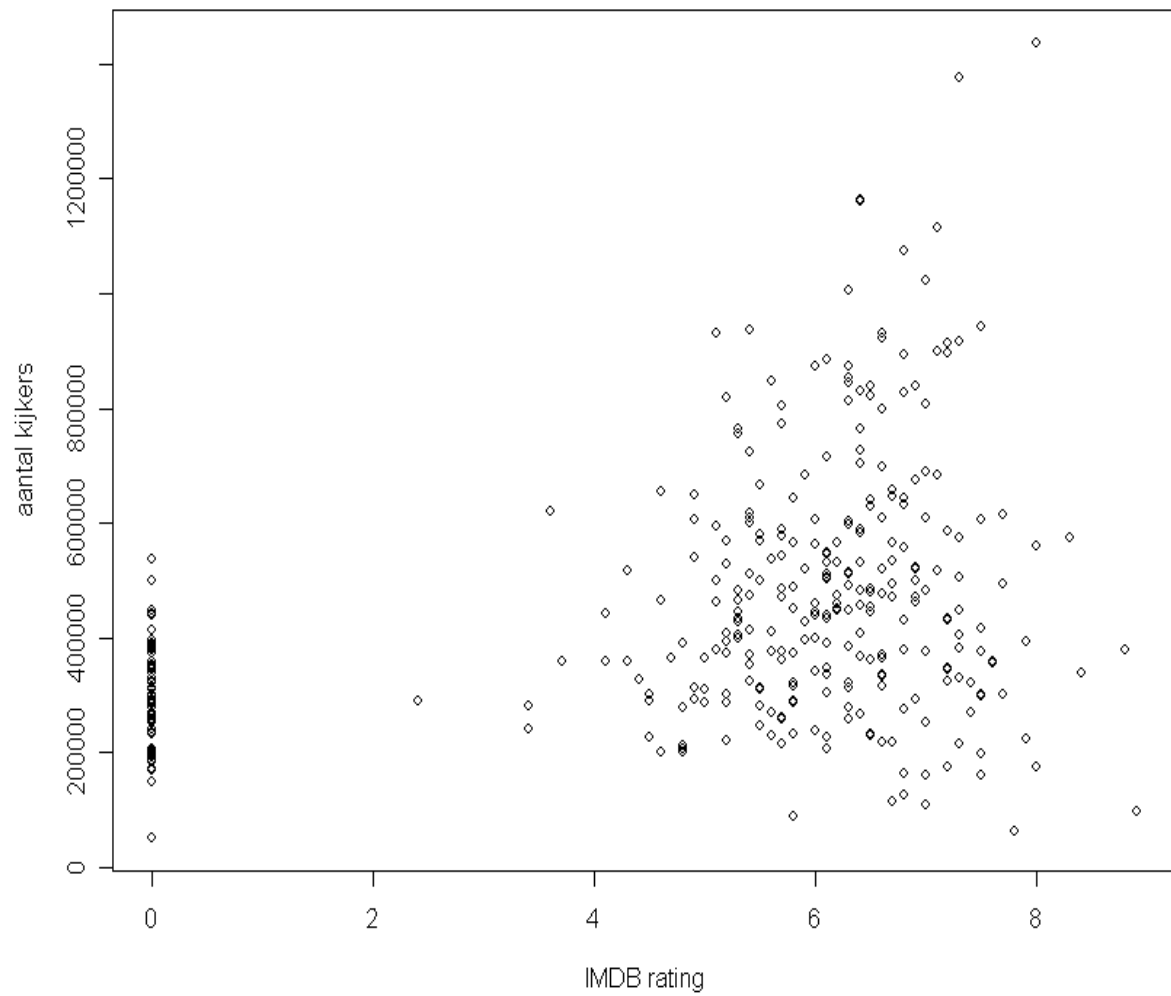
Actiehield t.o.v. aantal kijkers



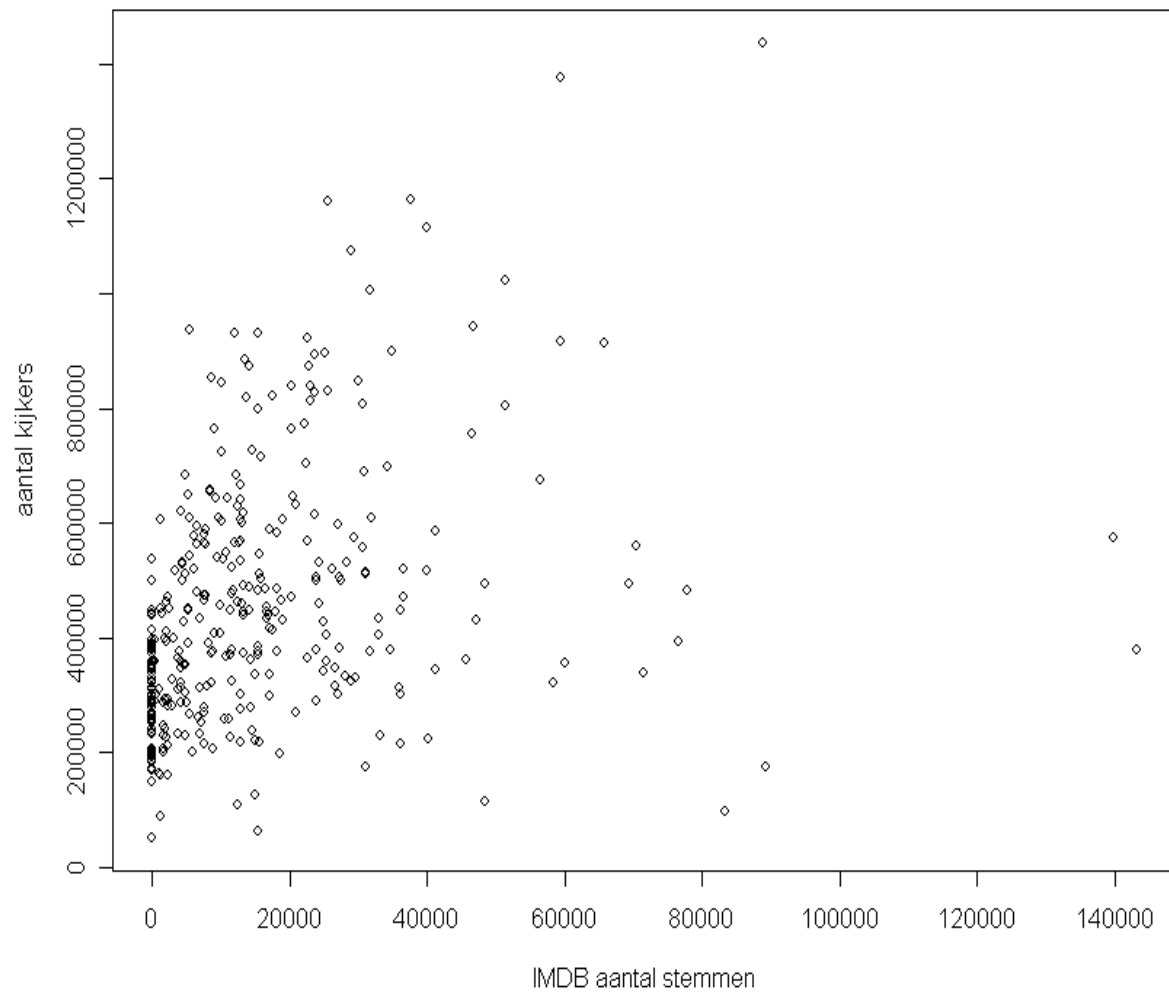
Gemiddelde rating in gids t.o.v. aantal kijkers



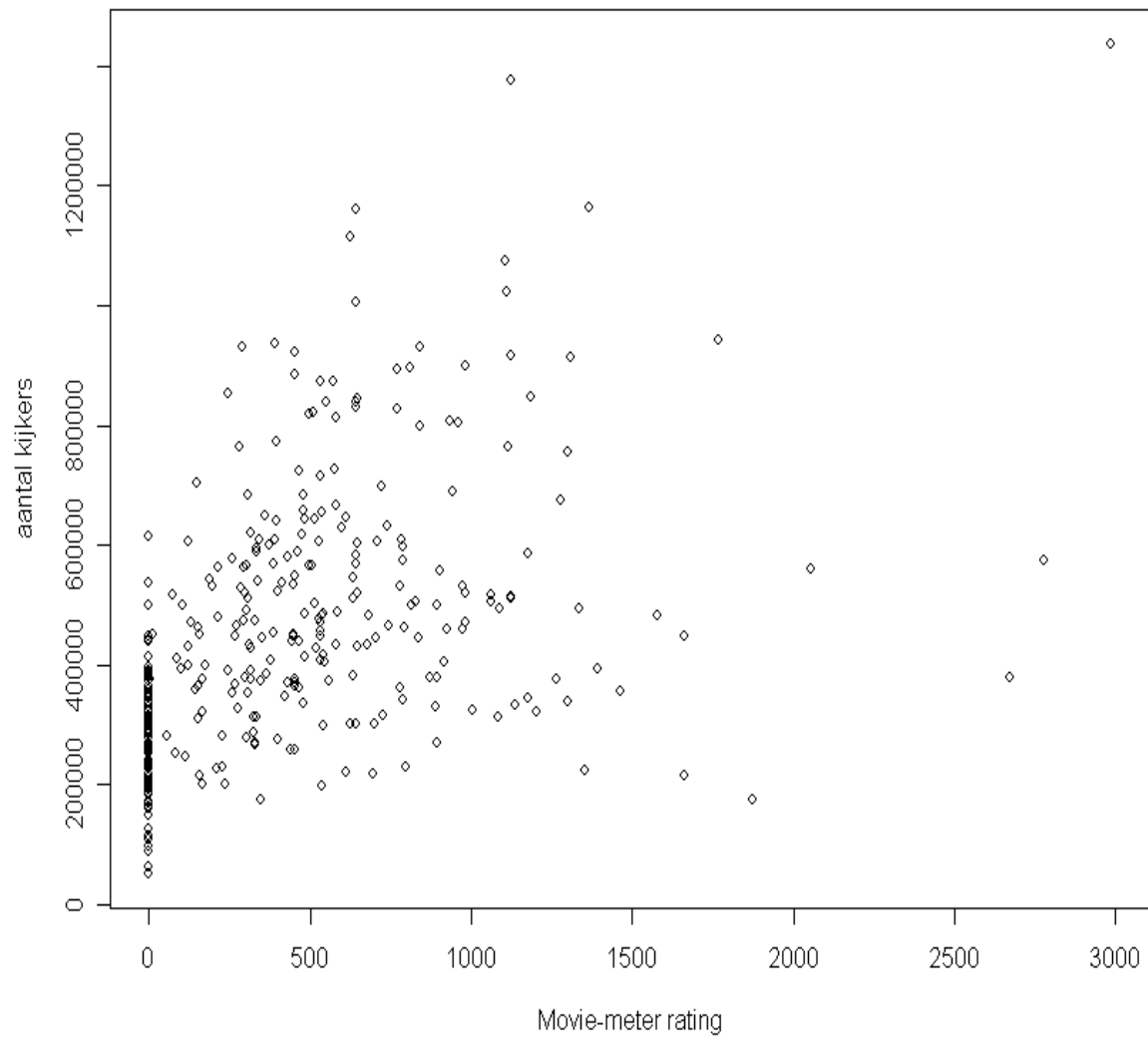
Rating film volgens IMDB.com t.o.v. aantal kijkers



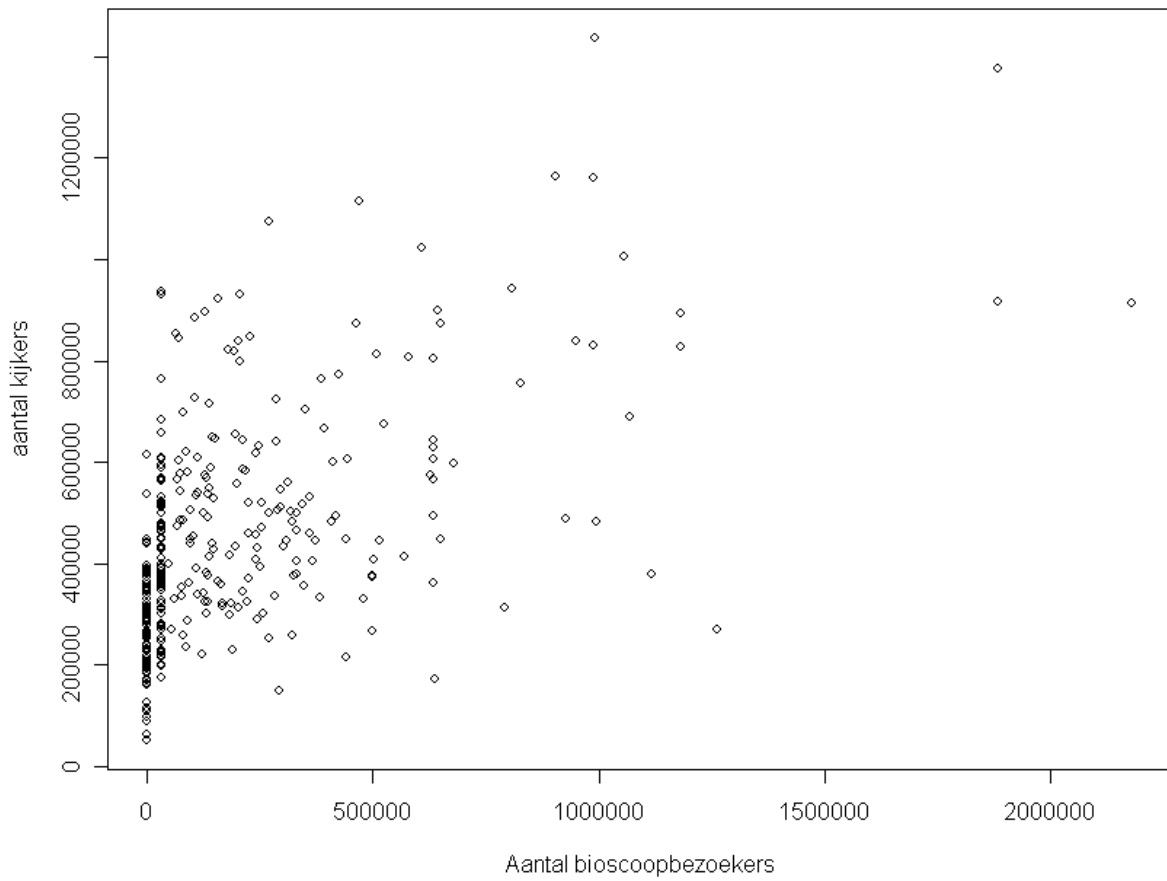
Aantal stemmen op IMDB.com t.o.v. aantal kijkers



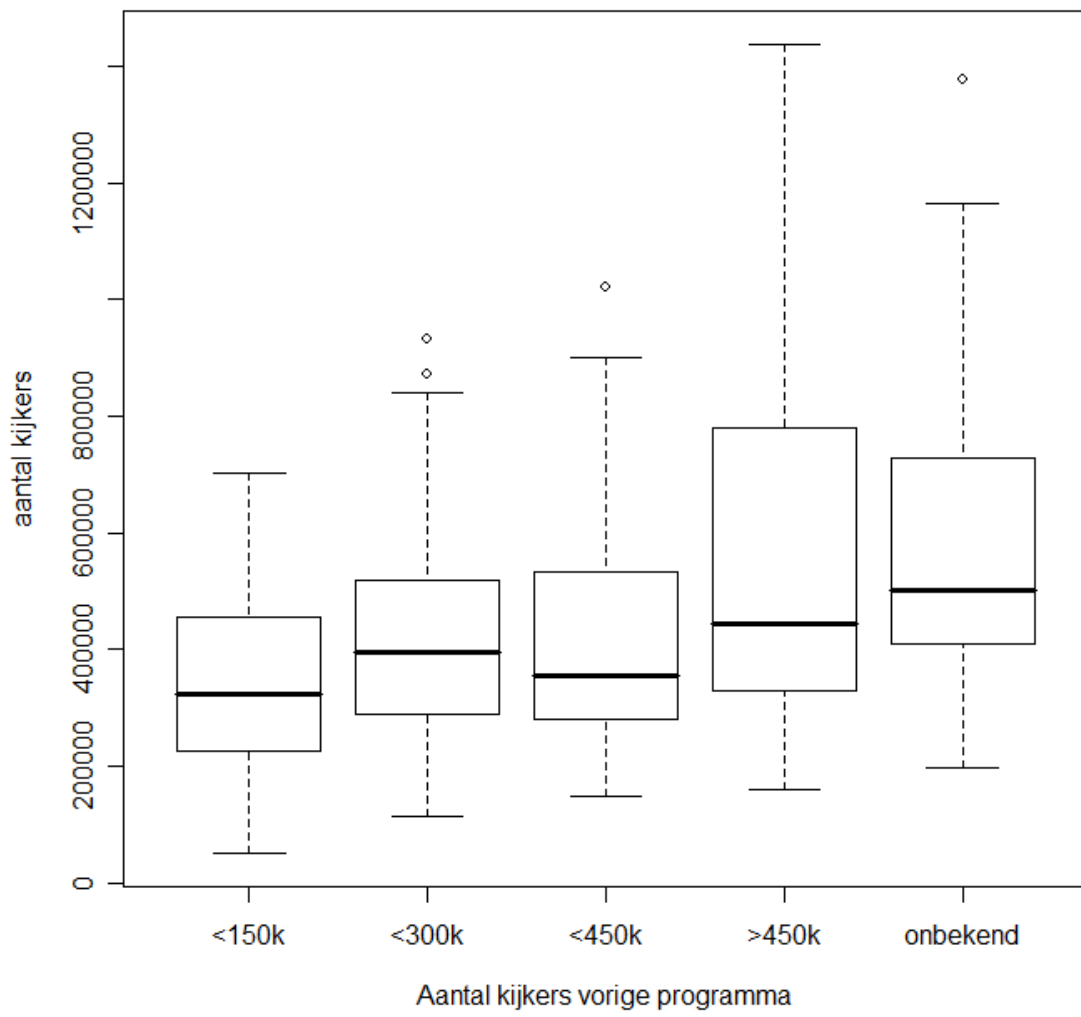
Moviemeter rating t.o.v. aantal kijkers



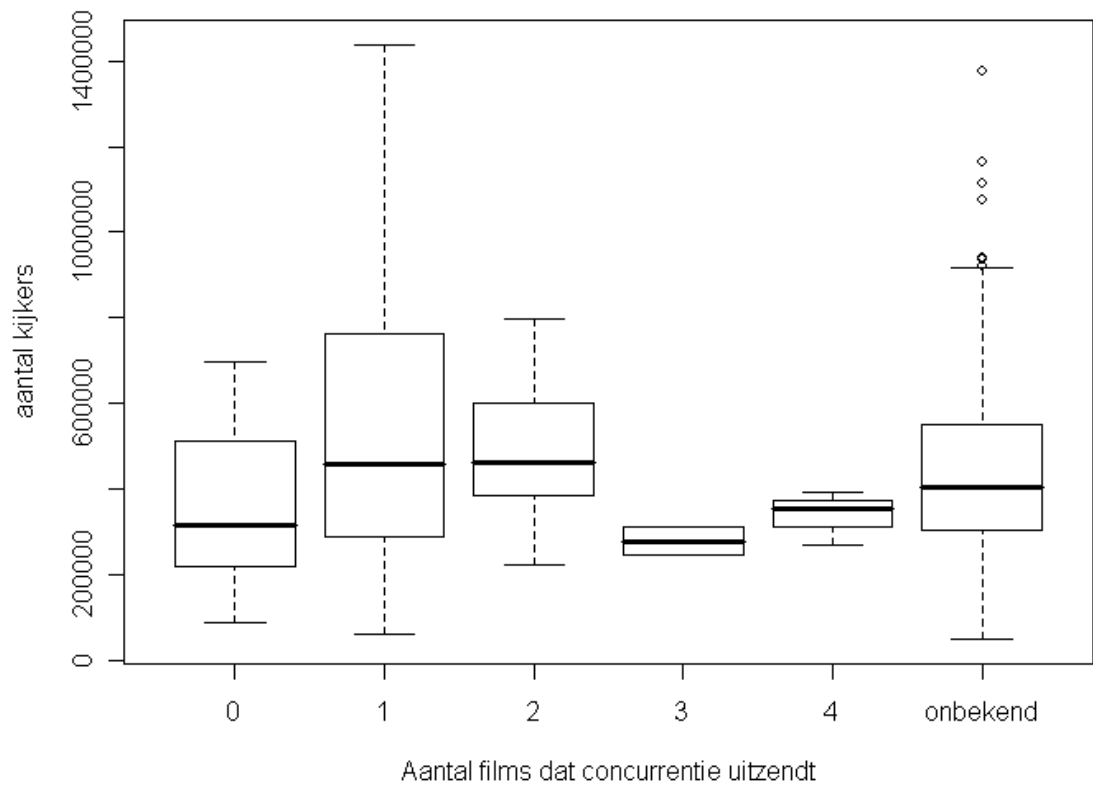
Aantal bioscoopbezoeken t.ov. aantal kijkers



Aantal kijkers vorige programma t.o.v. aantal kijkers



Aantal Films Concurrentie t.o.v. aantal kijkers



Bijlage 2: Een deel van de R-code gebruikt tijdens het schatten van het statistische model

```
VijfdeModel <- glm(Abs ~ SBSmaa20 + TijdMidden + BiosBezoekFactor + genActie +
MmeterFactor, family = poisson, data=db)
anova.glm(TweedeModel, DerdeModel, VierdeModel, VijfdeModel)
```

Analysis of Deviance Table

```
Model 1: Abs ~ SBSmaa20 + TijdMidden
Model 2: Abs ~ SBSmaa20 + TijdMidden + BiosBezoekFactor
Model 3: Abs ~ SBSmaa20 + TijdMidden + BiosBezoekFactor + genActie
Model 4: Abs ~ SBSmaa20 + TijdMidden + BiosBezoekFactor + genActie + MmeterFactor
```

Resid. Df Resid. Dev Df Deviance

```
1 343 13960588
2 339 12666114 4 1294474
3 338 12183123 1 482991
4 334 11691409 4 491714
```

```
> summary.glm(VijfdeModel)
```

Call:

```
glm(formula = Abs ~ SBSmaa20 + TijdMidden + BiosBezoekFactor +
genActie + MmeterFactor, family = poisson, data = db)
```

Deviance Residuals:

```
Min 1Q Median 3Q Max
-523.264 -129.230 -7.122 113.935 564.556
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	12.744037	0.0002819	45204.22	<2e-16 ***
SBSmaa20	0.5031033	0.0002278	2208.07	<2e-16 ***
TijdMidden	0.2046611	0.0002581	793.09	<2e-16 ***
BiosBezoekFactor<200000	0.1341248	0.0002592	517.41	<2e-16 ***
BiosBezoekFactor<400000	0.1220090	0.0002857	427.04	<2e-16 ***
BiosBezoekFactor>=400000	0.1705226	0.0002897	588.56	<2e-16 ***
BiosBezoekFactoronbekend	0.0761314	0.0005720	133.09	<2e-16 ***
genActie	0.1436679	0.0001779	807.53	<2e-16 ***
MmeterFactor<600	-0.0567850	0.0002322	-244.59	<2e-16 ***
MmeterFactor<900	-0.0163363	0.0002686	-60.82	<2e-16 ***
MmeterFactor>=900	0.0594681	0.0002975	199.91	<2e-16 ***
MmeterFactoronbekend	-0.308453	0.0005621	-548.71	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 34843388 on 345 degrees of freedom
Residual deviance: 11691409 on 334 degrees of freedom
AIC: 11696537
```

Number of Fisher Scoring iterations: 4

```
> add1(VijfdeModel,scope= ~ SBSmaa20 + TijdMidden + NaDeZomer + WinterMaanden +
Verzon20 + Veronicadin22 + Veronicadin20 + NET5vrij22 + NET5don20 + TijdVroeg+ TijdLaat
+ LengteFactor + VorProgFactor + VolProgFactor+ genRomantiek + genActie
+genAvontuur + LeeftijdFactor + X.4 + X.8 + X.15 + Land +
Actiehield + GemGidsFactor + IMDBfactor + IMDBStemfact +
++ GeweldFactor + MmeterFactor + BiosBezoekFactor + X..30000 + X.200000 +
X.400000 , test ="Chisq",data=db)
```

Single term additions

Model:

Abs ~ SBSmaa20 + TijdMidden + BiosBezoekFactor + genActie + MmeterFactor

Df	Deviance	AIC	LRT	Pr(Chi)
<none>	11671676	1	1696537	
NaDeZomer	1 11643764	11668627	27912 < 2.2e-16 ***	
WinterMaanden	1 11648376	11673239	23300 < 2.2e-16 ***	
Verzon20	1 11659658	11684521	12018 < 2.2e-16 ***	
Veronicadin22	1 11490575	11515439	181101 < 2.2e-16 ***	
Veronicadin20	1 11641153	11666016	30523 < 2.2e-16 ***	
NET5vrij22	1 11460390	11485254	211286 < 2.2e-16 ***	
NET5don20	1 11666608	11691472	5068 < 2.2e-16 ***	
TijdVroeg	1 11461867	11486731	209809 < 2.2e-16 ***	
TijdLaat	1 11579290	11604154	92386 < 2.2e-16 ***	
LengteFactor	3 11494602	11519469	177074 < 2.2e-16 ***	
VorProgFactor	4 11211961	11236831	459714 < 2.2e-16 ***	
VolProgFactor	5 10442409	10467280	1229267 < 2.2e-16 ***	
genRomantiek	1 11664049	11688912	7627 < 2.2e-16 ***	
genAvontuur	1 11438601	11463464	233075 < 2.2e-16 ***	
LeeftijdFactor	4 11350896	11375766	320780 < 2.2e-16 ***	
X.4	1 11661129	11685993	10546 < 2.2e-16 ***	
X.8	1 11627857	11652720	43819 < 2.2e-16 ***	
X.15	1 11630256	11655119	41420 < 2.2e-16 ***	
Land	6 11358033	11382906	313643 < 2.2e-16 ***	
Actiehield	1 11656363	11681226	15313 < 2.2e-16 ***	
GemGidsFactor	2 11613172	11638038	58504 < 2.2e-16 ***	
IMDBfactor	4 10986615	11011484	685061 < 2.2e-16 ***	
IMDBStemfact	4 11385373	11410242	286303 < 2.2e-16 ***	
GeweldFactor	10 10961008	10985889	710668 < 2.2e-16 ***	
X..30000	0 11671676	11696537	0	
X.200000	0 11671676	11696537	0	
X.400000	0 11671676	11696537	0	

--

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Warning messages:

1: In add1.glm(VijfdeModel, scope = ~SBSmaa20 + TijdMidden + NaDeZomer + :
using the 345/346 rows from a combined fit

2: In pchisq(q, df, lower.tail, log.p) : NaNs produced

```
> ZesdeModel <- glm(Abs ~ SBSmaa20 + TijdMidden + BiosBezoekFactor + genActie +
MmeterFactor + Veronicadin22, family = poisson, data=db)
> summary.glm(ZesdeModel)
```

```
Call:
glm(formula = Abs ~ SBSmaa20 + TijdMidden + BiosBezoekFactor +
genActie + MmeterFactor + Veronicadin22, family = poisson,
data = db)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-497.057 -120.490  -7.113  114.483  548.516
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)      12.7447072  0.0002818 45219.58 <2e-16 ***
SBSmaa20          0.5023178  0.0002279 2204.41 <2e-16 ***
TijdMidden        0.2040028  0.0002579  790.88 <2e-16 ***
BiosBezoekFactor<200000 0.1332318  0.0002593  513.89 <2e-16 ***
BiosBezoekFactor<400000 0.1214686  0.0002857  425.17 <2e-16 ***
BiosBezoekFactor>=400000 0.1727669  0.0002897  596.26 <2e-16 ***
BiosBezoekFactoronbekend 0.0381854  0.0005782   66.04 <2e-16 ***
genActie          0.1434767  0.0001779  806.64 <2e-16 ***
MmeterFactor<600   -0.0567051  0.0002322  -244.24 <2e-16 ***
MmeterFactor<900   -0.0161847  0.0002686   -60.25 <2e-16 ***
MmeterFactor>=900   0.0586634  0.0002974   197.23 <2e-16 ***
MmeterFactoronbekend -0.2449146  0.0005792  -422.84 <2e-16 ***
Veronicadin22      -0.2375901  0.0005742  -413.75 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)
```

```
Null deviance: 34843388 on 345 degrees of freedom
Residual deviance: 11510170 on 333 degrees of freedom
AIC: 11515301
```

```
Number of Fisher Scoring iterations: 4
```

```
anova.glm(VijfdeModel,ZesdeModel, test="Chisq")
Analysis of Deviance Table
```

```
Model 1: Abs ~ SBSmaa20 + TijdMidden + BiosBezoekFactor + genActie + MmeterFactor
Model 2: Abs ~ SBSmaa20 + TijdMidden + BiosBezoekFactor + genActie + MmeterFactor +
Veronicadin22
```

```
Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1    334  11691409
2    333  11510170  1  181239      0
```

Bijlage 3: De wervingsmatrix

1 GrStd/gg1/wrk0/opl1
2 GrStd/Hkw<=34;gg2+//opl23
3 GrStd/Hkw<=34;gg1/wrk1/opl12
4 GrStd/Hkw<=34;gg1//opl3
5 GrStd/Hkw<=49;knd0/wrk0/opl23
6 GrStd/Hkw3549;gg1/wrk1/opl3
7 GrStd/Hkw3549;gg2+;knd0//opl2
8 GrStd/Knd012/wrk1/opl12
9 GrStd/Knd012/wrk1/opl3
10 GrStd/Knd1317//
11 GrStd/Hkw3549;gg1/wrk1/opl12
12 GrStd/Hkw>=50;knd0/wrk1/opl12
13 GrStd/Hkw>=50;gg2+;knd0/wrk0/opl2
14 GrStd/Hkw>-50;knd0/wrk0/opl23
15 GrStd/Knd012/wrk0/
16 GrStd/Hkw<=34|Hkw>=50;gg2+//opl1
17 GrStd/Hkw>=35;;knd0/wrk1/opl3
18 West/Hkw<=34|>=50;knd0//opl1
19 West/Hkw<=34;knd0//opl2
20 West/Hkw3549;gg1//opl12
21 West/Hkw<=49;;knd0//opl3
22 West/Hkw>=50;gg2+/wrk0/opl3
23 West/Hkw3549;;knd0/wrk1/opl3
24 West/Hkw3549;gg2+;knd0//opl12
25 West/Knd012/wrk1/opl2
26 West/Knd012/wrk1/opl3
27 West/Knd1317//opl12
28 West/Hkw>=50;gg1/wrk0/opl2
29 West/Hkw>=50;gg1|gg2+;knd0/wrk1/opl12
30 West/Hkw>=50;gg2+/wrk1/opl2
31 West/hkw<=34|Knd1317//opl13
32 West/Hkw<=34;gg2+;knd0/wrk1/opl23
33 West/Hkw>=50;gg2+/wrk0/opl1
34 West/Knd012/wrk0/
35 West/Hkw>=50;knd0/wrk1/opl3
36 West/Hkw>=50;gg1/wrk0/opl3
37 West/Hkw>=50;gg2+/wrk0/opl2
38 West/Knd012/wrk1/opl1
39 Nrd;;knd0/wrk0/
40 Nrd/Hkw<=49;gg1/wrk1/opl12
41 Nrd/Hkw<=34;gg1/wrk1/opl3
42 Nrd/Hkw3549;;knd0//
43 Nrd/Knd012//opl12
44 Nrd/Knd012//opl3
45 Nrd/Knd1317//
46 Nrd/Hkw>=50;;knd0/wrk0/opl23
47 Nrd/Hkw>=50;;knd0/wrk1/opl12
2/2
48 Nrd/Hkw>=50;gg2+;knd0/wrk0/opl2
49 Nrd/Hkw<=34;gg2+;knd0//
50 Nrd/Hkw<=34|>=50;gg2+;knd0/wrk0/opl1
51 Nrd/Hkw>=35;;knd0/wrk1/opl3
52 Oost/gg1/wrk0/opl1

53 Oost/Hkw<=34;;knd0/wrk0/opl23
 54 Oost/Hkw<=49;gg1/wrk1/opl12
 55 Oost/Hkw<=34;gg1/wrk1/opl3
 56 Oost/Hkw>=35;gg1//opl3
 57 Oost/Hkw3549;gg2+;knd0//opl12
 58 Oost/Knd012/wrk1/opl1
 59 Oost/Knd012/wrk1/opl2
 60 Oost/Knd012/wrk1/opl3
 61 Oost/Knd012/wrk1/opl3
 62 Oost/Hkw>=50;gg1/wrk0/opl2
 63 Oost/Hkw>=50;;knd0/wrk1/opl12
 64 Oost/Hkw>=50;gg2+/wrk1/opl2
 65 Oost/Hkw>=35;;knd0/wrk0/opl23
 66 Oost/Hkw>=50;gg2+/wrk0/opl3
 67 Oost/Hkw<=34;gg2+;knd0/wrk1/
 68 Oost/Knd012/wrk0/
 69 Oost/Hkw<=34|>=50;gg2+;knd0/wrk0/opl1
 70 Oost/Hkw>=35;;knd0/wrk1/opl3
 71 Zuid;gg1/wrk0/opl1
 72 Zuid/Hkw<=34;gg1//opl12
 73 Zuid/Hkw3549;gg1/wrk1/opl12
 74 Zuid/Hkw<=34;gg1//opl3
 75 Zuid/Hkw3549;;knd0//opl23
 76 Zuid/Hkw3549;gg2+;knd0//opl12
 77 Zuid/Knd012/wrk1/opl12
 78 Zuid/Knd012/wrk1/opl3
 79 Zuid/Knd1317//opl12
 80 Zuid/Hkw>=50;gg1/wrk0/opl2
 81 Zuid/Hkw>=50;gg2+/wrk1/opl2
 82 Zuid/Hkw>=50;gg2+/wrk0/opl2
 83 Zuid/Knd1317//opl3
 84 Zuid/Hkw>=50;;knd0/wrk0/opl3
 85 Zuid/Hkw<=34;gg2+;knd0/wrk1/opl1|3
 86 Zuid/Knd012/wrk0/
 87 Zuid/Hkw<=34;gg2+//
 88 Zuid/Hkw>=35;;knd0/wrk1/opl3
 89 Zuid/Hkw>=35;;knd0/wrk1/opl3
 90 Dren/Hkw<=49;;knd0//
 91 Dren/hkw<=49;knd0|Knd012//
 92 Dren/Hkw>=50;gg1|knd017//
 93 Dren/Hkw>=50;gg2+;knd0//
 94 Zee/Hkw<=49;;knd0//
 95 Zee/Hkw<=49;knd0|Knd012//
 96 Zee/Hkw>=50;gg1//
 97 Zee/Hkw>=50;;knd0//opl3
 98 Allochtoon
 3/2

Legenda Wervingsmatrix:

De volgorde van de naamgeving van de cellen is als volgt:

Regio/Gezinscyclus/Werkzaamheid/Opleiding

"|" = EN relatie

"//" = alle categorieën van deze variabelen vallen binnen deze cel

Regio:

GrStd 3 grote steden+randgemeenten

West overig west (excl. 3 grote std en randgem)

Nrd Noord

Oost Oost

Zuid Zuid

Dren Drenthe

Zee Zeeland

Gezinscyclus:

Hkw<=34;GG1 Alleenstaand leeftijd <=34

Hkw3549;GG1 Alleenstaand leeftijd 35-49

Hkw>=50;GG1 Alleenstaand leeftijd 50+

Hkw<=34;gg2+;Knd-0 Volwassen gezin (zonder kinderen), leeftijd
hoofdkostwinner <=34

Hkw3549;gg2+;Knd-0 Volwassen gezin (zonder kinderen), leeftijd
hoofdkostwinner 35-49

Hkw>=50;gg2+ Volwassen gezin (zonder kinderen), leeftijd
hoofdkostwinner 50+

Knd012 Huishouden met jongste kind tussen de leeftijd 0-
12

Knd1317 Huishouden met jongste kind tussen de leeftijd
13-17

Werkzaamheid:

Wrk0 0 uren werkzaam per week

Wrk1 > 0 uren werkzaam per week

Opleiding:

Opl1 Laag opleidingsniveau

Opl2 Midden opleidingsniveau

Opl3 Hoog opleidingsniveau

Bijlage 4: Filmtitels van de geanalyseerde films

Film Naam	Film Nummer
10 THINGS I HATE ABOUT YOU	1
28 DAYS	2
3000 MILES TO GRACELAND	3
40 DAYS AND 40 NIGHTS	4
8 MILE	5
8MM	6
A LIFE LESS ORDINARY	7
A PERFECT MURDER	8
A PERFECT MURDER	9
A SIMPLE PLAN	10
A VIEW TO A KILL	11
ABOUT A BOY	12
ABSOLUTE POWER	13
ACCIDENTAL SPY	14
ACE VENTURA PET DETECTIVE	15
ACE VENTURA PET DETECTIVE	16
ACE VENTURA WHEN NATURE...	17
ACE VENTURA WHEN NATURE...	18
AIRPLANE!	19
ALMOST HEROES	20
ALONG CAME A SPIDER	21
AMERICAN OUTLAWS	22
ANALYZE THAT	23
ANALYZE THIS	24
ANATOMIE	25
ANOTHER 48 HRS	26
ARRIVAL	27
ASSASSINS	28
ASSOCIATE	29
BACK TO THE FUTURE 3	30
BACKDRAFT	31
BAD BOYS	32
BASIC	33
BASIC	34
BASIC INSTINCT	35
BEAUTY SHOP	36
BEST OF THE BEST 4 WITHOUT WARNING	37
BEVERLY HILLS NINJA	38
BIG DADDY	39
BIG HIT	40
BIG MOMMA'S HOUSE	41
BILL & TED'S EXCELLENT ADVENTURE	42
BLACK RAIN	43
BLACK SHEEP	44
BLOOD WORK	45
BLOODSPORT	46
BLOWN AWAY	47
BLUE STREAK	48
BODYGUARD	49
BOILER ROOM	50

BONE COLLECTOR	51
BONE COLLECTOR	52
BREAKDOWN	53
BREAKFAST CLUB	54
BREWSTER'S MILLIONS	55
BRUCE ALMIGHTY	56
BULLETPROOF	57
CABLE GUY	58
CADET KELLY	59
CAMOUFLAGE	60
CASINO	61
CATCH ME IF YOU CAN	62
CELL	63
CHANGING LANES	64
CHASE	65
CLASS OF 1999	66
CLIFFHANGER	67
CLUELESS	68
COLD CREEK MANOR	69
COLD MOUNTAIN	70
COLLATERAL DAMAGE	71
CONCIERGE	72
COYOTE UGLY	73
D2 THE MIGHTY DUCKS	74
DANTE'S PEAK	75
DEATH WARRANT	76
DEEP RISING	77
DEEP RISING	78
DEMOLITION MAN	79
DEMOLITION MAN	80
DENNIS THE MENACE	81
DEVIL'S ADVOCATE	82
DIAMONDS ARE FOREVER	83
DIE HARD WITH A VENGEANCE	84
DOMESTIC DISTURBANCE	85
DRAGONHEART	86
DREAMCATCHER	87
DROP ZONE	88
DUPLEX	89
EIGHT LEGGED FREAKS	90
ELF	91
EMPIRE	92
END OF DAYS	93
ENEMY AT THE GATES	94
ENEMY OF THE STATE	95
ENTRAPMENT	96
EQUILIBRIUM	97
ERIN BROCKOVICH	98
ESCAPE FROM L.A.	99
EVENT HORIZON	100
EXTREME MEASURES	101
FALLEN	102
FAST AND THE FURIOUS	103
FIERCE CREATURES	104

FIGHTING TEMPTATIONS	105
FINAL DESTINATION	106
FIRST BLOOD	107
FLETCH	108
FLINTSTONES IN VIVA ROCK VEGAS	109
FOR YOUR EYES ONLY	110
FORREST GUMP	111
FREAKY FRIDAY	112
FRIGHTENERS	113
GALAXY QUEST	114
GANGS OF NEW YORK	115
GANGS OF NEW YORK	116
GEORGE AND THE DRAGON	117
GET CARTER	118
GET SHORTY	119
GHOST SHIP	120
GHOST SHIP	121
GHOSTBUSTERS 2	122
GO	123
GODFATHER PART 2	124
GOLDEN CHILD	125
GOLDENEYE	126
GOLDENEYE	127
GOOD WILL HUNTING	128
GOONIES	129
GRIND	130
GUEST HOUSE PARADISO	131
HALLOWEEN THE CURSE OF MICHAEL MYERS	132
HAND THAT ROCKS THE CRADLE	133
HANNIBAL	134
HARD RAIN	135
HARD WAY	136
HARRY POTTER AND THE CHAMBER OF SECRETS	137
HARRY POTTER AND THE CHAMBER OF SECRETS	138
HARRY POTTER AND THE PHILOSOPHER'S STONE	139
HARRY POTTER AND THE PHILOSOPHER'S STONE	140
HAUNTED MANSION	141
HAUNTED MANSION	142
HEAT	143
HEAVYWEIGHTS	144
HELD UP	145
HOLLOW MAN	146
HOW TO LOSE A GUY IN 10 DAYS	147
HUDSON HAWK	148
INDECENT PROPOSAL	149
INSOMNIA	150
ITALIAN JOB	151
JACKASS THE MOVIE	152
JERRY MAGUIRE	153
JOAN OF ARC	154
JOE SOMEBODY	155
JOHN Q	156
JUNGLE 2 JUNGLE	157
KINDERGARTEN COP	158

KNIGHTRIDER 2000	159
KNOCK OFF	160
L.A.CONFIDENTIAL	161
LAKE PLACID	162
LARA CROFT TOMB RAIDER THE CRADLE OF...	163
LAST STAND AT SABER RIVER	164
LEGALLY BLONDE	165
LETHAL WEAPON 2	166
LETHAL WEAPON 3	167
LETHAL WEAPON 3	168
LETHAL WEAPON 4	169
LIFE	170
LITTLE NICKY	171
LITTLE VAMPIRE	172
LIVE AND LET DIE	173
LIVING DAYLIGHTS	174
LIZZIE MCGUIRE MOVIE	175
LONG KISS GOODNIGHT	176
LOOK WHO'S TALKING	177
LOOK WHO'S TALKING TOO	178
MAD MAX	179
MAJOR LEAGUE	180
MAN IN THE IRON MASK	181
MAN WHO KNEW TOO LITTLE	182
MAN WITH THE GOLDEN GUN	183
MARS ATTACKS	184
MASK	185
MASK OF ZORRO	186
MATCHSTICK MEN	187
MATRIX RELOADED	188
MAXIMUM RISK	189
MCHALE'S NAVY	190
MEAN GIRLS	191
MEAN MACHINE	192
MEET JOE BLACK	193
MERCURY RISING	194
MINORITY REPORT	195
MISSION IMPOSSIBLE	196
MISSION IMPOSSIBLE	197
MISSION IMPOSSIBLE 2	198
MOONRAKER	199
MURDER AT 1600	200
MY FAVORITE MARTIAN	201
NAT. LAMPOON'S CHRISTMAS VACATION 2	202
NECESSARY ROUGHNESS	203
NEGOTIATOR	204
NICK OF TIME	205
NICK OF TIME	206
NINE LIVES	207
NO GOOD DEED	208
NOTTING HILL	209
NOWHERE TO RUN	210
OCEAN'S ELEVEN	211
OCEAN'S ELEVEN	212

OCTOPUSSY	213
ON DEADLY GROUND	214
OPERATION DUMBO DROP	215
OSMOSIS JONES	216
OUT FOR JUSTICE	217
OUTBREAK	218
PARENT TRAP	219
PASSENGER 57	220
PASSENGER 57	221
PATRIOT	222
PAY IT FORWARD	223
PAYCHECK	224
PEARL HARBOR	225
PELICAN BRIEF	226
PERFECT STORM	227
PHENOMENON	228
PHILADELPHIA	229
PIRATES OF THE CARIBBEAN THE CURSE OF...	230
PLEDGE	231
PLUTO NASH	232
POINT BREAK	233
POLICE STORY	234
POLICE STORY 2	235
PRETTY WOMAN	236
PRIMAL FEAR	237
PRIMAL FEAR	238
PROOF OF LIFE	239
QUICK CHANGE	240
RAMBO FIRST BLOOD PART 2	241
REAL MCCOY	242
RECRUIT	243
RECRUIT	244
RED CORNER	245
RED DRAGON	246
REIGN OF FIRE	247
RENAISSANCE MAN	248
REPLACEMENT KILLERS	249
REPLACEMENTS	250
REPOSSESSED	251
RIGHT STUFF	252
RING	253
RING	254
ROAD RAGE	255
ROBIN HOOD PRINCE OF THIEVES	256
ROBINSON CRUSOE	257
ROCK	258
ROCK STAR	259
ROLLERCOASTER	260
ROMEO MUST DIE	261
RONIN	262
ROOKIE	263
RUMBLE IN THE BRONX	264
RUNAWAY BRIDE	265
RUSH HOUR	266

SAINT	267
SCARY MOVIE	268
SCHINDLER'S LIST	269
SCHOOL OF ROCK	270
SCREAM 2	271
SCREWED	272
SCROOGED	273
SECOND STRING	274
SECONDHAND LIONS	275
SEE SPOT RUN	276
SGT BILKO	277
SHANGHAI KNIGHTS	278
SHANGHAI NOON	279
SHOWTIME	280
SIGNS	281
SISTER ACT	282
SKULLS	283
SLEEPY HOLLOW	284
SMALL SOLDIERS	285
SMOKEY AND THE BANDIT	286
SMOKEY AND THE BANDIT 2	287
SMOKEY AND THE BANDIT 3	288
SOUTH PARK BIGGER LONGER & UNCUT	289
SOUTHERN COMFORT	290
SPACE COWBOYS	291
SPACE JAM	292
SPECIES	293
SPIES LIKE US	294
SPY WHO LOVED ME	295
STIGMATA	296
STREET FIGHTER	297
SUPERMAN	298
SWITCHBACK	299
TAKING OF BEVERLY HILLS	300
THE IN-LAWS	301
THING	302
THOMAS CROWN AFFAIR	303
THOMAS CROWN AFFAIR	304
TIME MACHINE	305
TIMECOP	306
TIMELINE	307
TOMORROW NEVER DIES	308
TOMORROW NEVER DIES	309
TOP GUN	310
TRAINING DAY	311
TREMORS	312
TROJAN WAR	313
TRUE LIES	314
TRUE ROMANCE	315
TURNER & HOOG	316
TUXEDO	317
TWELVE MONKEYS	318
TWO WEEKS NOTICE	319
UNCLE BUCK	320

UNDER SUSPICION	321
UNDERCOVER BLUES	322
UNIVERSAL SOLDIER	323
UNSAID	324
US MARSHALS	325
VALENTINE	326
VAMPIRE IN BROOKLYN	327
VANILLA SKY	328
VANILLA SKY	329
VERTICAL LIMIT	330
VERTICAL LIMIT	331
VIRTUOSITY	332
WATERBOY	333
WEEKEND AT BERNIE'S	334
WEIRD SCIENCE	335
WHAT A GIRL WANTS	336
WHAT WOMEN WANT	337
WHAT'S THE WORST THAT COULD HAPPEN	338
WHAT'S THE WORST THAT COULD HAPPEN	339
WHEELS ON MEALS	340
WHILE YOU WERE SLEEPING	341
WHO AM I	342
WILD AT HEART	343
WILLY WONKA AND THE CHOC	344
WITNESS	345
WORLD IS NOT ENOUGH	346