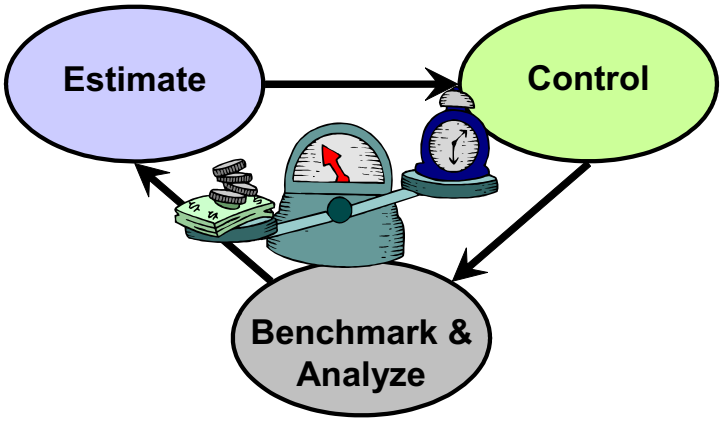
Data-analyse bij de Belastingdienst

Kasstromen en Functie Punt Analyse binnen
Belastingdienst/Centrum voor ICT



Liv Harkes
September 2008

Data-analyse bij de Belastingdienst

Kasstromen en Functie Punt Analyse binnen
Belastingdienst/Centrum voor ICT

Stageverslag
Liv Harkes
Begeleiding: Dr. S.A. Pot, Prof. Dr. A. Ran, Th.J.A. Aalbers RA, Drs. R.C. Dom

Vrije Universiteit Amsterdam
Faculteit der Exacte Wetenschappen
Bedrijfswiskunde & Informatica
De Boelelaan 1105
1081 HV Amsterdam

Stagebedrijf
Belastingdienst/Centrum voor ICT
John F. Kennedylaan 8
7314 PS Apeldoorn

Voorwoord

Het verslag dat voor u ligt, is het resultaat van een onderzoek binnen het Centrum voor Informatie & Communicatie Technologie van de Belastingdienst (B/CICT). Het vormt de afsluiting van mijn studie Bedrijfswiskunde en Informatica aan de Vrije Universiteit van Amsterdam.

In dit verslag worden drie probleemstellingen nader uitgewerkt. Het eerste deel van dit verslag beschrijft het onderzoek om te komen tot een betrouwbaar model voor de prognose van het kasresultaat van B/CICT. In het tweede deel wordt het gebruik van Functie Punt Analyse (FPA) binnen B/CICT geanalyseerd. In het derde deel wordt een betrouwbaar model opgesteld om de zuinigheid van automobielen te voorspellen. Doordat er meerdere probleemstellingen waren, was het voor mij nodig om naar verscheidene gebieden binnen de organisatie te kijken. Het onderzoek naar de derde probleemstelling is bovendien niet bij B/CICT uitgevoerd. Dit integrale karakter heeft ervoor gezorgd dat de afstudeeropdracht niet verveelde.

In het bijzonder wil ik mijn verschillende begeleiders binnen B/CICT, die mij hebben geholpen bij de totstandkoming van dit onderzoek en het verslag, mijn dank betuigen. Ten eerste mijn begeleidster Dori-Anne Aalbers. Dori-Anne heeft mij een grote mate van vrijheid gegeven om de aanpak en uitvoering van de opdracht uit te voeren. Tevens wil ik haar bedanken voor de nuttige feedback op het totale verslag. Verder wil ik Mark van de Streek bedanken voor zijn hulp bij het opzetten van de onderzoeken en Roy Dom voor het meedenken over verder onderzoek.
Graag wil ik ook Auke Pot en André Ran, mijn begeleiders van de Vrije Universiteit, bedanken voor de inzet en hulp gedurende de stage.

Ook wil ik graag de overige collega's van B/CICT hartelijk bedanken voor de gezellige sfeer, enthousiasme en openheid voor het stellen van vragen.

Liv Harkes
September 2008

Samenvatting

In dit verslag worden drie probleemstellingen nader uitgewerkt. Het eerste deel van dit verslag beschrijft het onderzoek om te komen tot een betrouwbaar model voor de prognose van het kasresultaat van B/CICT. In het tweede deel wordt het gebruik van Functie Punt Analyse (FPA) binnen B/CICT geanalyseerd. En in het derde deel wordt een betrouwbaar model opgesteld om de zuinigheid van automobielen te voorspellen. Door de eerste twee verschillende probleemstellingen was het voor mij nodig om naar verscheidene gebieden binnen de organisatie te kijken. De derde probleemstelling is bovendien niet bij B/CICT uitgevoerd. Deze laatste probleemstelling is uitgevoerd om te laten zien dat betrouwbare modellen kunnen bestaan in tegenstelling tot de eerste twee probleemstellingen.

Onderzoeksopzet Kasstroom
Door B/CICT worden, naar Belastingdienstbegrippen, substantiële kasuitgaven gedaan. Binnen B/CICT hanteert men voor de verantwoording van de apparaatuitgaven het *baten - lasten* stelsel. B/CICT dient zich extern echter om kasbasis te verantwoorden. Teneinde inzicht en grip op het kasresultaat te verkrijgen, zowel in realisatie als prognose, is het opstellen van een betrouwbaar kasstroomoverzicht onontbeerlijk.
De doelstelling van het onderzoek is de betrouwbaarheid van de kasstroomprognose te onderzoeken en vervolgens een model te ontwikkelen om betrouwbare kasprognoses te kunnen opstellen.
De volgende onderzoeksvraag is geformuleerd:

Zijn er statistische modellen om de variabelen van de kasstroom beter te kunnen voorspellen, die bijdragen aan een hogere betrouwbaarheid van de prognoses van het kasresultaat per jaareinde?
Zo ja welk statistisch model is het meest geschikt voor dit probleem?

Deze probleemstelling is op twee manieren onderzocht. Als eerste is met behulp van lineaire regressie onderzocht of bepaalde verklarende variabelen goede voorspellers zijn voor de totale kasprognose op basis van historische data. Daarna is met behulp van tijdreeks analyse geprobeerd de kasprognose te voorspellen.

Op basis van de gevonden determinatiecoëfficiënten blijkt dat meervoudige lineaire regressie (standaard methode) de werkelijk waarden van de afhankelijke variabelen beter door het model worden benaderd dan de andere modellen.
Op basis van een tijdreeks met 12 data punten kan geen betrouwbaar model gecreëerd worden. Voor uitgebreider onderzoek moet er meer data zijn.

Onderzoeksopzet Functie Punt Analyse
FPA is een methode om de functionele omvang van een informatiesysteem te meten. FPA meet deze functionele omvang door te kijken naar relevante functies en (logische) gegevensverzamelingen. De meeteenheid is de functiepunt (fp); de omvang van een systeem wordt uitgedrukt in een aantal functiepunten. FPA is een objectieve en (ISO) gecertificeerde methode voor de bepaling van de omvang van een systeem.
De doelstelling van het onderzoek is de betrouwbaarheid van de data van FPA te onderzoeken en vervolgens een model te ontwikkelen om de productiviteit van een project te voorspellen.
De volgende onderzoeksvraag is geformuleerd:

Zijn er statistische modellen om de variabelen van FPA te kunnen analyseren en om de productiviteit van een systeem te voorspellen, die bijdragen aan een hogere betrouwbaarheid? Zo ja welk statistisch model is het meest geschikt voor dit probleem?

Deze probleemstelling is op twee manieren benaderd en bestaat dus ook uit twee onderzoeken. Als eerste is met behulp van lineaire regressie onderzocht of bepaalde verklarende variabelen goede voorspellers zijn voor de totale productiviteit van een systeem op basis van historische data.
Op basis van de gevonden determinatiecoëfficiënten blijkt dat de werkelijk waarden van de afhankelijke variabelen slecht door de modellen worden benaderd.

Onderzoeksopzet Zuinigheid Automobielen
De doelstelling van het onderzoek is de data gegevens van automobielen te onderzoeken en vervolgens een model te ontwikkelen om de zuinigheid van een automobiel te kunnen voorspellen en analyseren.
De volgende onderzoeksvraag is geformuleerd:

Zijn er statistische modellen om de zuinigheid van een automobiel te kunnen voorspellen en analyseren?
Zo ja welk statistisch model is het meest geschikt voor dit probleem?

Deze probleemstelling is op twee manieren benaderd en bestaat dus ook uit twee onderzoeken. Als eerste is met behulp van lineaire regressie onderzocht of bepaalde verklarende variabelen goede voorspellers zijn voor de totale zuinigheid van automobielen op basis van historische data. Daarna is met behulp van een regressieboom-analyse geanalyseerd in hoeverre de zuinigheid van automobielen verklaard kunnen worden door de verklarende variabelen.
Op basis van de correlatie coëfficiënt blijkt dat de regressieboom-analyse een betere samenhang heeft tussen de werkelijke waarden en de regressiewaarden van MPG dan meervoudige lineaire regressie.

Inhoudsopgave

1 Inleiding

Dit verslag is het resultaat van een onderzoek binnen het Centrum voor Informatie & Communicatie Technologie van de Belastingdienst (B/CICT).

Er worden drie probleemstellingen nader uitgewerkt. Het eerste deel beschrijft het onderzoek naar een betrouwbaar model voor de prognose van het kasresultaat van B/CICT. In het tweede deel wordt het gebruik van Functie Punt Analyse (FPA) binnen B/CICT geanalyseerd. En in het derde deel wordt een betrouwbaar model opgesteld om de zuinigheid van automobielen te voorspellen.

Als eerste wordt het bedrijf beschreven waar de stage heeft plaats gevonden. Daarna wordt er in hoofdstuk 3 een algemene beschrijving gegeven over hoe een onderzoek moet opgezet worden. In hoofdstuk 4 wordt het onderzoeksopzet van kasstroom uitgewerkt. Hier wordt het onderzoeksprobleem besproken, aanpak van het onderzoek, onderzocht, de resultaten van het onderzoek en de conclusies. In hoofdstuk 5 wordt het onderzoeksopzet van functie punt analyse uitgewerkt. Hier wordt het onderzoeksprobleem besproken, aanpak van het onderzoek, onderzocht, de resultaten van het onderzoek en de conclusies. In hoofdstuk 6 wordt het onderzoeksopzet zuinigheid automobielen uitgewerkt. Hier wordt het onderzoeksprobleem besproken, aanpak van het onderzoek, onderzocht, de resultaten van het onderzoek en de conclusies.

2 Organisatiebeschrijving

Dit hoofdstuk bevat een beknopte beschrijving van de organisatie Belastingdienst/Centrum voor Informatie- en Communicatietechnologie waar ik mijn afstudeeropdracht heb uitgevoerd.

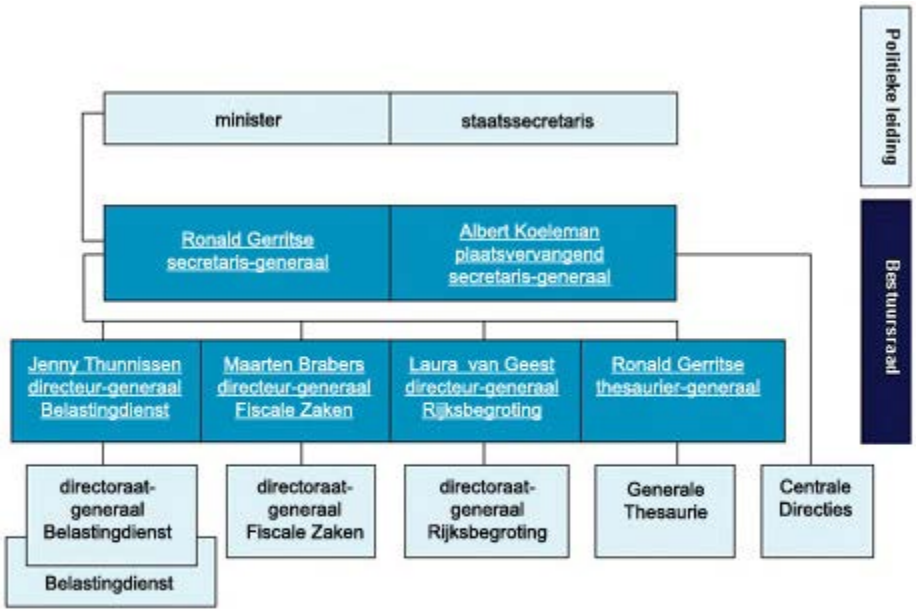
2.1 Belastingdienst/Centrum voor Informatie- en Communicatietechnologie

De Belastingdienst maakt onderdeel uit van het Ministerie van Financiën. Het Ministerie van Financiën bestaat uit vier Directoraten–generaal, namelijk:

1. De Generale Thesaurie
2. Het Directoraat-generaal van de Rijksbegroting
3. Het Directoraat-generaal voor Fiscale Zaken
4. Het Directoraat-generaal der Belastingen

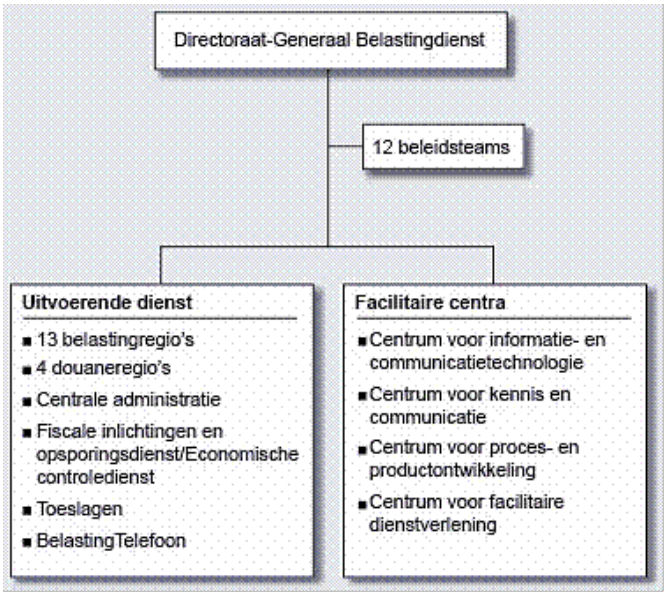
Het Directoraat-generaal der Belastingen is verantwoordelijk voor het besturen van de uitvoering van de belastingwetgeving, waaronder de douanewetgeving en de niet-fiscale wetgeving, waarvan de uitvoering aan de Belastingdienst is opgedragen. De Belastingdienst zorgt voor de heffing en de inning van het belastinggeld.

Figuur 1 Organisatieschema van het ministerie van Financiën



Het onderzoek zal worden uitgevoerd bij het Belastingdienst/Centrum voor Informatie- en Communicatietechnologie (B/CICT). Dit is een facilitair centrum van de Belastingdienst, dat de primaire processen van de Belastingdienst (registreren en distribueren, diensten verlenen, heffen, innen, controleren en douanetaken) ondersteunt middels automatisering. De centrale computer van B/CICT beheert gegevens van belastingplichtige particulieren, voertuigen en ondernemingen en verwerkt daarnaast miljoenen transacties. Daarnaast zorgt B/CICT er voor dat medewerkers van de Belastingdienst beschikken over alles wat zij op het gebied van automatisering nodig hebben. De figuur hieronder toont duidelijk de plaats van B/CICT binnen de structuur van de Belastingdienst onder het knopje 'Facilitaire centra'.

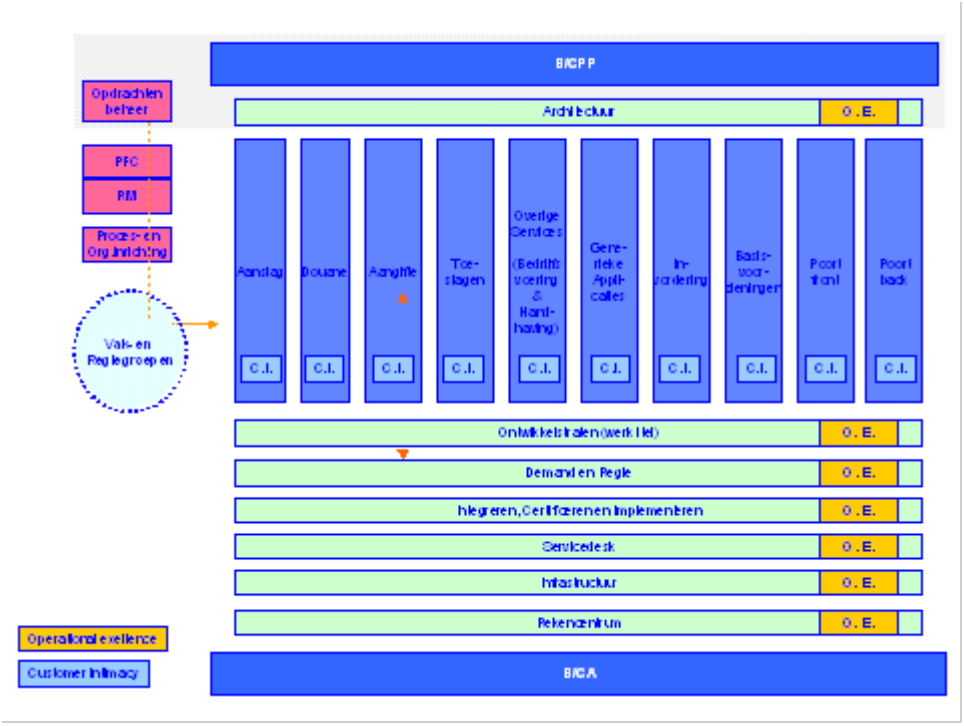
Figuur 2 Organisatiestructuur Belastingdienst



2.2 Organisatiemodel

Sinds 17 maart 2008 ziet de organisatiestructuur en de besturing van B/CICT er als volgt uit:

Figuur 3 Organisatiestructuur B/CICT



B/CICT kende als enige organisatieonderdeel binnen de Belastingdienst een matrixstructuur.

De hoofdlijnen van de nieuwe structuur zijn:

1. Korte lijnen naar de opdrachtgever en voor de klant herkenbare portefeuilles;
2. Voor de units die niet direct met de klant in contact staan en/of aan de Customer Intimacy (O.I.) Units een dienst leveren, geldt het principe van Operational Excellence (O.I.:de beste op de markt tegen de laagste prijs);
3. Een kleine en efficiënte staf.

Hierdoor is er een transparante organisatie-indeling ontstaan, die enerzijds gericht is op de Belastingdienst (Customer Intimacy) en anderzijds op goed presteren (Operational Excellence). De nieuwe organisatie is ingericht als een lijn/staf organisatie met managementteams die integraal verantwoordelijk zijn. De besturing gebeurt met behulp van collegiaal management bestaande uit twee bestuurslagen, conform de uitgangspunten van de Belastingdienst.

2.3 Stafunit Planning, Financiën en Control (PFC)

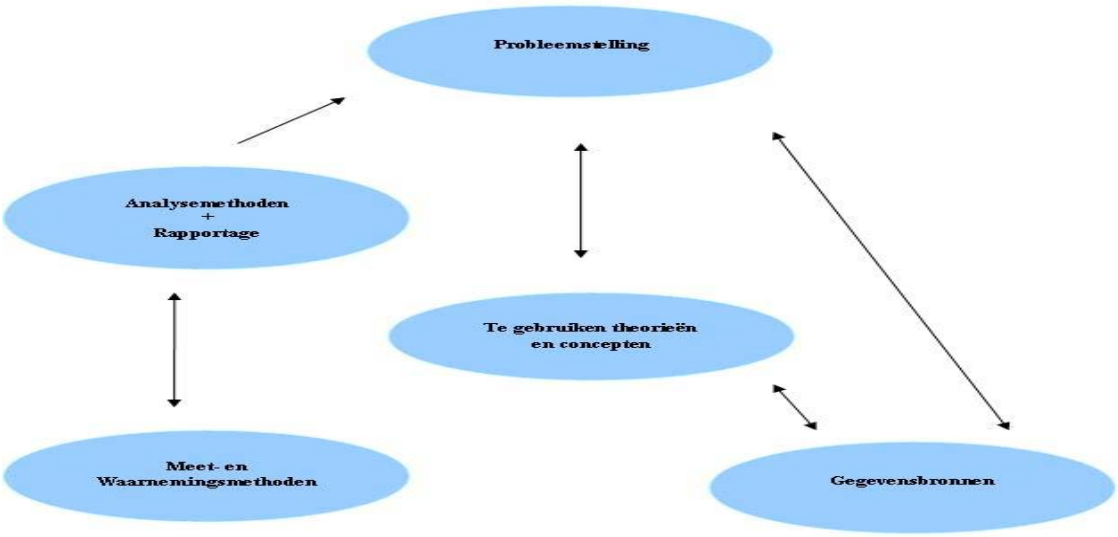
Het doel van de stafunit PFC is het ondersteunen van de besturing en de beheersing van het proces van klant tot klant. PFC tracht tijdig inzicht te geven in risico's van de organisatie en tevens bijsturingmaatregelen te formuleren. Op deze wijze wordt bijgedragen aan een efficiënte besturing van B/CICT. De sector PFC bestaat uit de volgende afdelingen: concern control, BP&R (beleid, planning en rapportage) en het expertisecentrum, waarvan onder andere de financiële administratie deel uit maakt.

3 Beschrijving onderzoeksopzet algemeen

3.1 Inleiding

Bij het uitvoeren van mijn afstudeeropdracht is er gekozen voor de methodologie, die ontwikkeld is door De Leeuw (2001). Hij stelt dat methodologisch bekeken een aanpak van een onderzoek niet meer is dan een stel samenhangende beslissingen. Deze beslissingen worden door De Leeuw voorgesteld als de ballen van de ballentent (zie figuur 4). De verschillende ballen zijn aan elkaar gerelateerd en de beslissingen moeten dan ook in samenhang worden genomen. Zo dienen de theoretische begrippen (dat zijn de voornaamste denkinstrumenten) te worden gekozen met het oog op de probleemstelling. Maar anderzijds kunnen probleemstellingen niet worden geformuleerd zonder gebruik te maken van concepten. Gegevensbronnen moeten worden gekozen in het licht van de vraagstelling. Probleemstelling en theoretisch kader vormen samen het hart van de ballentent. De overige drie ballen vormen de aanpak in engere zin.

Figuur 4 De Ballentent



Bron: Leeuw, A.C.J. de, Een boekje over bedrijfskundige methodologie, Assen/Maastricht, Van Gorcum 2001, blz 88

3.2 Probleemstelling

De vraagstelling staat in verband met het uiteindelijke doel van het onderzoek: het vinden van oplossingen voor het probleem.

Probleemstelling bestaat uit:

- *Doel*
  Hier wordt vastgelegd voor wie het onderzoek gedaan wordt. Wat er voor hen uitkomt en waarom dat voor hen van belang is. Het gaat vooral om de relevantie van het onderzoek.
- *Vraagstelling*
  Hier wordt de hoofdvraag geformuleerd die bij die doelstelling aansluit.
- *Randvoorwaarden*
  Deze geven de beperkingen aan waaraan onderzoeksresultaten en methoden onderhevig zijn. Daaronder vallen ook de eisen en de voorwaarden die de klant ten aanzien van het onderzoek en de resultaten stelt.
  Een onderzoek dient nauwkeurig afgebakend te worden. Enerzijds is dit noodzakelijk omdat het gevraagde kennisproduct op een doelmatige en effectieve wijze moet worden geproduceerd. Anderzijds omdat voor een afstudeeropdracht maar een beperkte onderzoekstijd ter beschikking staat.

Via de rapportage worden de resultaten van het onderzoek aan de opdrachtgever gepresenteerd. Een belangrijk punt vormt ook de afweging van geheimhouding en vertrouwelijkheid. De randvoorwaarden omvatten daaromtrent de afspraken van het onderzoek.

3.3 Te gebruiken theorieën en concepten

Bij de keuze van de theoretische concepten spelen twee overwegingen een rol. Enerzijds moeten ze in staat stellen de problematiek werkelijk goed te begrijpen: duidelijk, precies en volledig aangeven waar het onderzoek om begonnen is. Anderzijds moeten ze bijdragen aan de onderzoekbaarheid. Ze moeten dus voldoen aan de eisen voor met name conceptuele definities en/of conceptuele modellen.

3.4 Gegevensbronnen

Om onderzoek te kunnen uitvoeren zijn gegevens nodig. Er zijn zes soorten bronnen waar gegevens vandaan gehaald kunnen worden: documenten, media, de werkelijkheid (het ‘veld’), de nagebootste werkelijkheid, databanken en de ervaring van de onderzoekers. Bronselectie is zeer belangrijk bij de opzet ven een onderzoek. Het gaat er immers om goede gegevens te verkrijgen waarmee de vraagstelling beantwoord kan worden.

3.5 Meet- en waarnemingsmethoden + Analysemethoden

De meetprocedure begint bij het conceptualiseren van de begrippen, loopt via de operationalisering in concrete begrippen en de concrete verzameling van de gegevens tot de verwerking van het al dan niet cijfermatig materiaal.

Als laatste stap in het onderzoek moet het verkregen materiaal worden geselecteerd, geanalyseerd en gerapporteerd.
Uit het ruwe materiaal wordt met verwerking en analyse het juiste en relevante materiaal afgeleid.
Hieruit volgen de conclusies en de aanbevelingen.
Enkele analysemethoden kunnen zijn:

- Statistische hulpmiddelen;
- Simulatie;
- Schematiseringen.

4 Uitwerking Onderzoeksopzet kasstroom

4.1 Probleemstelling

4.1.1 Doelstelling van het onderzoek

Door B/CICT worden, naar Belastingdienstbegrippen, substantiële kasuitgaven gedaan. Binnen B/CICT hanteert men voor de verantwoording van de apparaatuitgaven het *baten - lasten* stelsel. B/CICT dient zich extern echter op kasbasis te verantwoorden. Teneinde inzicht en grip op het kasresultaat te verkrijgen, zowel in realisatie als prognose, is het opstellen van een betrouwbaar kasstroomoverzicht onontbeerlijk.

Het is erg belangrijk om goede prognoses op te stellen van het kasresultaat per jaareinde. Als er afwijkingen zijn tussen de begroting en de prognoses c.q. tussen de prognoses in opeenvolgende perioden worden deze verklaard en geanalyseerd.

De doelstelling van het onderzoek is de betrouwbaarheid van de kasstroomprognose te onderzoeken en vervolgens een model te ontwikkelen om betrouwbare kasprognoses te kunnen opstellen.
In het huidige systeem van B/CICT worden er periodiek per variabele (exploitatieresultaat, afschrijvingen, mutaties balansposten, investeringen, kasresultaat) prognoses per ultimo jaar gemaakt. De daadwerkelijke realisatie blijkt na afsluiting van het jaar. Het is de bedoeling om per genoemde variabele periodieke stochastische variabelen te definiëren. Deze periodieke stochastische variabelen moeten leiden tot betrouwbaarder prognoses dan de huidige prognoses. Dit kan worden vastgesteld met correlaties tussen de huidige prognose en tussen de stochastische prognose.

4.1.2 Vraagstelling

De vraagstelling staat in verband met het uiteindelijke doel van het onderzoek. Het vinden van oplossingen voor het probleem om de kasstroom te voorspellen. De vraagstelling kan dan ook als volgt worden geformuleerd:

- Zijn er statistische modellen om de variabelen van de kasstroom beter te kunnen voorspellen, die bijdragen aan een hogere betrouwbaarheid van de prognoses van het kasresultaat per jaareinde? Deze prognoses moeten een goede indicatie geven van de uiteindelijke realisatie.
- Zo ja welk statistisch model is het meest geschikt voor dit probleem?

4.1.3 Randvoorwaarden

Randvoorwaarden:

- Het onderzoek beperkt zich tot B/CICT.
- Er wordt alleen gekeken naar de jaarverslagen met kasstroomoverzichten van de periode 1996-2007.
- De student zal periodiek/regelmatig naar de Vrije Universiteit gaan om met de begeleider te spreken over de stand van zaken en om de afbakening van het probleem.
- De student heeft tenminste 1 keer per week een gesprek met de Belastingdienst begeleider over de voortgang van het onderzoek.
- Historische data over de kasstroom zijn afkomstig uit de jaarrekeningen en de financiële administratie in SAP.
- Het systeem SAP waar de student mee zal leren omgaan, zal beschikbaar worden gesteld aan de student. Met behulp van dit systeem zal de student in staat zijn om de benodigde data uit de DataWarehouse en uit de jaarrekeningen te halen.
- Met behulp van SPSS zullen de data geanalyseerd worden middels statistische technieken.
- De opdrachtgever verwacht van de onderzoeker tijdens en na het onderzoek geheimhouding van de verkregen informatie.

4.2 Te gebruiken theorieën en concepten

Om de vraagstelling te beantwoorden, kunnen een aantal theorieën en concepten gebruikt worden:

- Correlatie
- Lineaire Regressie
  - Enkelvoudige Lineaire Regressie
  - Meervoudige Lineaire Regressie: Standaard Methode
  - Meervoudige Lineaire Regressie: Stapsgewijze Regressie
- Modellen voor Tijdreeks Analyse
  - Autocorrelatie
    - Toets voor Autocorrelatie
  - Exponentiële effening

Waarom worden deze theorieën en concepten gebruikt?

Correlatie:

- Of er samenhang is tussen de verschillende variabelen.

Enkelvoudige Lineaire Regressie:

- Of er een lineair verband is tussen twee grootheden *x* en *y.*

Meervoudige Lineaire Regressie: Standaard Methode:

- Of er een lineair verband is tussen alle verklarende variabelen x_i en één afhankelijke variabele *y.*

Meervoudige Lineaire Regressie: Stapsgewijze Regressie

- Of er een lineair verband is tussen alle verklarende variabelen x_i en één afhankelijke variabele *y.* Stapsgewijze regressie voegt alle variabelen stapsgewijs toe totdat de meest significante variabelen overblijven.

Autocorrelatie

- Met het toepassen van autocorrelatie wordt de onderlinge afhankelijkheid van opeenvolgende waarnemingen Y_t in een tijdreeks geanalyseerd.

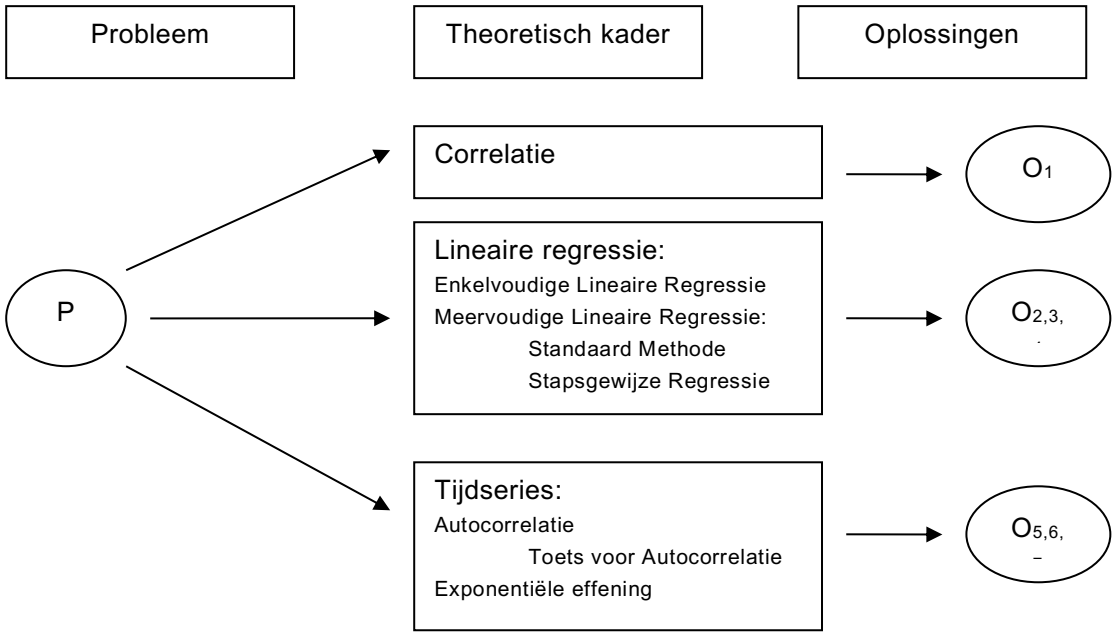
Toets voor Autocorrelatie

- Om te onderzoeken of de autocorrelatiecoëfficiënt significant van 0 afwijkt.

Exponentiële effening

- Met het toepassen van enkelvoudige exponentiële effening kan er voor het volgend tijdstip een voorspelling gedaan worden.

Figuur 5 Schema Theorieën en concepten



4.3 Gegevensbronnen

Voor het onderzoek naar het voorspellen van de kasstroom heb ik gebruik gemaakt van de jaarrekeningen uit de jaarverslagen van B/CICT van 1996 tot 2007. Deze jaarrekeningen bevatten de gegevens die nodig zijn voor de beantwoording van de in paragraaf 4.1 geformuleerde vraagstelling.

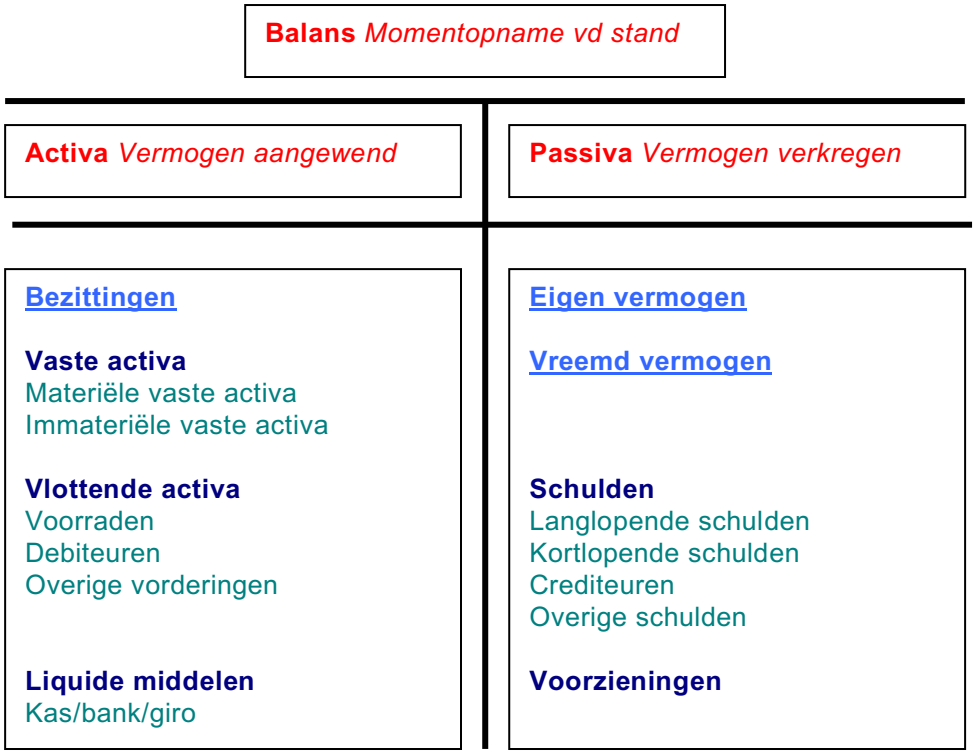
4.3.1 Jaarrekening

De jaarrekening is het stuk waarmee de leiding van een organisatie verantwoording aflegt over de financiële gevolgen van het door haar gevoerde beheer in het afgelopen boekjaar. De jaarrekening bestaat uit een balans, een resultatenrekening of winst- en verliesrekening over het afgelopen jaar, een toelichting op beide en het kasstroomoverzicht.

4.3.1.1 Balans

De balans is een overzicht van bezittingen en de schulden van de onderneming op een bepaald moment. De meest bekende presentatie van de balans is de T-vorm met aan de debetzijde (= activa; linkerkant) de bezittingen en aan de creditzijde (= passiva; de rechterkant) het eigen vermogen en de schulden. (zie figuur 6)

Figuur 6 Balans



De bezittingen van de onderneming worden activa of kapitaalgoederen genoemd. We onderscheiden twee soorten activa: duurzame activa (ook wel vaste activa) en vlottende activa.
De activa zijn in het verleden door de onderneming aangeschaft. Er is vermogen in geïnvesteerd. De waarde op de balans geeft aan hoeveel vermogen op dit moment in de activa is geïnvesteerd.

Duurzame activa gaan langer dan één productieperiode mee. Voorbeelden van duurzame activa zijn grond, gebouwen, machines en auto's. Tijdens hun gebruiksduur verminderen duurzame activa slechts geleidelijk in waarde als gevolg van slijtage. Het in de duurzame activa geïnvesteerde vermogen komt dan ook slechts geleidelijk vrij in de vorm van afschrijvingen. De afschrijvingen in een periode geven dus de waardevermindering van de duurzame activa in de betreffende periode weer.
Vlottende activa worden in het productieproces slechts eenmalig gebruikt, waarna het geïnvesteerde vermogen weer ter beschikking komt of vastligt in andere activa. De vlottende activa bestaan uit voorraden, vorderingen (op debiteuren) en liquide middelen en overige vorderingen.

Eigen vermogen wordt beschikbaar gesteld door de eigenaren van de onderneming en wordt voor onbepaalde tijd geïnvesteerd. Bij de persoonlijke ondernemingsvormen (eenmanszaak, vennotschap onder firma, commanditaire vennootschap, maatschap) is het eigen vermogen niet in afzonderlijke vormen onderverdeeld. Bij de NV en de BV bestaat het eigen vermogen uit aandelenkapitaal en reserves.

Vreemd vermogen wordt door de onderneming geleend van derden. Het wordt door de schuldeiser voor een bepaalde tijd aan de onderneming ter beschikking gesteld en moet daarna weer worden terugbetaald. Op grond van de looptijd van de lening (ofwel krediet) wordt onderscheid gemaakt tussen lang vreemd vermogen en kort vreemd vermogen. Lang vreemd vermogen heeft een looptijd van meer dan één jaar, kort vreemd vermogen van één jaar of korter.

4.3.1.2 De resultatenrekening

Een resultatenrekening of winst- en verliesrekening geeft een overzicht van de opbrengsten en kosten van een onderneming over een bepaalde periode, meestal een jaar. Dit overzicht eindigt met de over die periode behaalde winst of verlies.

Figuur 7 Resultatenrekening

Onder kosten verstaan we de in geld gemeten waarde van de opgeofferde productiemiddelen. Uit deze omschrijving kan worden afgeleid dat kosten niet hetzelfde zijn als uitgaven. Kosten en opbrengsten worden naar de winst- verliesrekening gebracht en hebben dus invloed op het resultaat over een periode. Ontvangsten en uitgaven doen de hoeveelheid liquide middelen toenemen respectievelijk afnemen. Zo is er bij een investering wel sprake van een uitgave, maar nog niet van kosten aangezien er immers nog niets is opgeofferd. De opoffering van dit productiemiddel bestaat uit de waardedaling van de machine die door afschrijving tot uitdrukking wordt gebracht. Deze afschrijvingen vormen de kosten en zijn geen uitgaven. Hetzelfde doet zich voor met opbrengsten en ontvangsten.

Opbrengsten zijn niet synoniem aan ontvangsten zoals kosten niet dezelfde betekenis hebben als uitgaven. Het onderscheid is met name van belang voor de financiële rapportage: opbrengsten en kosten bepalen het resultaat (winst of verlies) over een bepaalde periode, ontvangsten en uitgaven bepalen de liquiditeitspositie.

4.3.1.3 Het kasstroomoverzicht

De meeste grote ondernemingen nemen in hun jaarverslag als derde overzicht (naast balans en resultatenrekening) een kasstroomoverzicht op, ook wel “Staat van herkomst en besteding der middelen” genoemd. Het kasstroomoverzicht geeft een verklaring van de mutaties in de liquide middelen gedurende het boekjaar.

Het kasstroomoverzicht sluit aan bij de toegenomen belangstelling voor cash accounting. De kasmutatie is een objectief gegeven, terwijl de winstbepaling beïnvloed wordt door de keuze van de waarderingsgrondslagen en van schattingen ten aanzien van de grootte van de benodigde voorzieningen, levensduur van de activa, enzovoort.

Overigens zal het winstbegrip altijd een belangrijke functie blijven vervullen: voor het beoordelen van de mate van succes van een onderneming kan de jaarlijkse kastoename niet zonder meer gebruikt worden.

Het kasstroomoverzicht kan op twee manieren opgesteld worden. De eerste manier is de directe methode, waarbij een samenvatting van het kasboek van de onderneming gemaakt wordt. De tweede manier is de indirecte methode, waarbij de mutaties van de balansposten over een jaar als uitgangspunt worden genomen: er is sprake van een toename (herkomst) van liquide middelen, indien extra vermogen ter beschikking komt of indien in activa vastgelegd vermogen vrijkomt. Er is sprake van een afname (besteding) van liquide middelen indien vermogen wordt afgelost of vermogen wordt vastgelegd in activa. De post “liquide middelen” is de sluitpost van het kasstroomoverzicht. Het verschil tussen de stand van de liquide middelen op *t* en *t+1* is het kasresultaat.

Vrijwel alle ondernemingen maken gebruik van de indirecte methode. Deze heeft als voordeel dat er aansluiting blijft met balans en resultatenrekening.

Het kasstroomoverzicht kan op verschillende manieren ingedeeld worden. Een veelgebruikte rubricering is die in de volgende activiteiten.

- Operationele activiteiten: hieronder worden gerangschikt winst, afschrijvingen en mutaties in het nettowerkkapitaal.
- Financieringsactiviteiten: hieronder vallen de financieringstransacties op lange termijn, zoals aandelenmissies en het aantrekken en aflossen van lang vreemd vermogen.
- Investeringsactiviteiten: in deze rubriek komt de aanschaf en afstoting van duurzame productie middelen.

De Belastingdienst c.q. B/CICT heeft haar financiële administratie ingericht conform het baten-lastenstelsel; echter de verantwoording richting het Ministerie van Financiën moet op basis van het kasverplichtingenstelsel gebeuren. Via het kasstroomoverzicht wordt de vertaling van BLS (exploitatieresultaat) naar kas gemaakt.
Hieronder staat een overzicht hoe het kasstroomoverzicht is opgebouwd.

Figuur 8 Kasstroomoverzicht

Kasstroomoverzicht 2007 *x € 1.000*	
Kasstroom uit operationele activiteiten	
Prognose Exploitatieresultaat	
Prognose Afschrijvingen	
	Kasstroom uit bedrijfsoperaties
Mutatie in vorderingen (debiteuren)	
Mutatie overlopende activa (o.a. vooruitbetaalde kosten)	
Mutatie kortlopende schulden (crediteuren)	
Mutatie overlopende passiva	
Totaal mutatie balansposten	
	Kasstroom uit operationele activiteiten
Kasstroom uit investeringsactiviteiten	
Prognose investeringen - Uitbreiding en vervanging	
	Kasstroom uit investeringsactiviteiten
Prognose kasstroom 2007	
Kastarget B/CICT opgelegd door DGBel	
Investeringsbudget DGBel	
Inputbudget vaste telefonie (verwerkt in exploitatieresultaat)	
Kasoverschot /-tekort	

4.4 Meet- en Waarnemingsmethoden + Analysemethoden

In bijlage B worden de verschillende modellen beschreven die geschikt zijn om de toekomstige waarden van de jaarlijkse kasstroom te voorspellen. Ook wordt de gebruikte software voor het onderzoek beschreven.

Om de vraagstelling te beantwoorden worden de volgende modellen en software gebruikt:

- Correlatie
- Lineaire Regressie
  - Enkelvoudige Lineaire Regressie
  - Meervoudige Lineaire Regressie: Standaard Methode
  - Meervoudige Lineaire Regressie: Stapsgewijze Regressie
- Modellen voor Tijdreeks Analyse
  - Autocorrelatie
    - Toets voor Autocorrelatie
  - Exponentiële effening
- SPSS

Deze meet- en waarnemingsmethoden + analysemethoden worden nader uitgelegd in bijlage B.

4.5 Rapportage

[vertrouwelijk]

4.5.1 Conclusie

Zijn er statistische modellen om de variabelen van de kasstroom beter te kunnen voorspellen, die bijdragen aan een hogere betrouwbaarheid van de prognoses van het kasresultaat per jaareinde? Deze prognoses moeten een goede indicatie geven van de uiteindelijke realisatie. Zo ja welk statistisch model is het meest geschikt voor dit probleem?

Op basis van de gevonden determinatiecoëfficiënten blijkt dat meervoudige lineaire regressie de werkelijk waarden van de afhankelijke variabelen beter benadert dan enkelvoudige lineaire regressie en stapsgewijze meervoudige lineaire regressie .

Om de kasprognose te modelleren is gebruik gemaakt van exponentiële effening. Voor alle variabelen is een voorspelling gemaakt voor het volgende tijdstip. Deze voorspellingen zijn vergeleken met de huidige prognose door het verschil te berekenen tussen deze twee waarden. Uit deze berekening blijkt dat het verschil groot is.
Ook blijkt uit de autocorrelatiecoëfficiënten dat alle datapunten behalve lag 1 Mutatie in voorzieningen en lag 1 Totaal mutatie balansposten niet gecorreleerd zijn.

Op basis van een tijdreeks met 12 datapunten kan geen betrouwbaar model gecreëerd worden. Voor uitgebreider onderzoek moet er meer data zijn.

5 Uitwerking Onderzoeksopzet Functie Punt Analyse

5.1 Probleemstelling

5.1.1 Inleiding Functie Punt Analyse (FPA)

FPA is een methode om de functionele omvang van een informatiesysteem te meten. FPA meet deze functionele omvang door te kijken naar voor gebruikers relevante functies en (logische) gegevensverzamelingen. De meeteenheid is de *functiepunt (fp)*; de omvang van een systeem wordt uitgedrukt in een aantal functiepunten. FPA is een objectieve en (ISO) gecertificeerde methode voor de bepaling van de omvang of wijzigingen in de omvang van een systeem.

De *functiepunt* is de meeteenheid, waarin concreet en objectief de omvang van een te realiseren informatiesysteem kan worden uitgedrukt en waarmee systemen in omvang met elkaar kunnen worden vergeleken. Er zijn diverse toepassingsmogelijkheden, waarin FPA een rol speelt.

Beter en eerder begroten
De omvang - uitgedrukt in functiepunten - is één van de vijf fundamentele factoren die de planning van een project beïnvloeden. De andere factoren zijn de doorlooptijd, de kwaliteit, de productiekosten en de productiviteit van het ontwikkelen van software. Op grond van de functionele systeemspecificaties kan de systeemomvang (in aantal fp) reeds in een vroeg stadium worden vastgesteld. Bijvoorbeeld aan de hand van de definitie studie of het basisontwerp. Als we de omvang van het systeem weten en de overige projectkenmerken kan met behulp van planningstool QSM SLIM Estimate een schatting worden gemaakt van de planning. Dat kan op basis van ervaringscijfers uit de markt of op basis van referentieprojecten.

Beter beheersen van projecten
Wijzigingen in de systeemspecificaties kunnen in functiepunten worden uitgedrukt, waardoor ze beter en beheersbaar zijn. Het effect van deze wijzigingen kan via impactanalyses worden onderbouwd. Het effect van wijzigingen op de planning, alsmede de voortgang, kan via QSM SLIM Control worden ondersteund. Dit wordt ook wel scope management genoemd.

Beter communiceren tussen de betrokken partijen
Onduidelijke of onvolledige systeemspecificaties komen bij het maken van een FPA aan het licht.

Meten van de productiviteit
Het aantal besteedde uren gedeeld door het aantal gerealiseerde functiepunten geeft de productiviteit van een project aan. De productiviteit van een project is hoofdzakelijk afhankelijk van de omvang, de doorlooptijd, de kwaliteit en de beschikbaar gestelde resources (capaciteit projectteam en hulpmiddelen). De productiviteit wordt als volgt berekend:

Figuur 9 Formule Productiviteit

$$\text{Productiviteit} = \frac{\text{Omvang}}{\text{Doorlooptijd}^{4/3} \times \text{Inspanning}^{1/3}}$$

Omvang :in functiepunten of regels broncode
Doorlooptijd :in maanden
Inspanning: :in dagen

Meten van de kwaliteit van een systeem.
Het aantal fouten per functiepunt per tijdseenheid is een getal voor de kwaliteit van een ontwikkeld systeem.

Verbeteren van de kwaliteit van het ontwikkelproces.
Door het terugdringen van miscommunicatie en stuurmaatregelen op grond van productiviteits- en kwaliteitsmetingen, zoals hierboven beschreven, kan de kwaliteit van het ontwikkelproces worden verbeterd. Een andere veel gebruikte meeteenheid is aantal regels geproduceerde code (SLOC = Source Lines of Code).
Hiervoor is echter een gerealiseerd systeem nodig. Deze meeteenheid is dus alleen geschikt voor het bepalen van de productiviteit achteraf, niet voor begroten. Voor het vergelijken van SLOC met FPA worden in de praktijk omrekenfactoren gebruikt.

Bij het verzamelen van de projectgegevens wordt er gericht op 5 core metrieken van software ontwikkeling, te weten productiviteit, doorlooptijd, bestedingen, bevindingen en omvang. Verder zijn de ontwikkelstraat en gearingfactor van belang. In het onderstaande overzicht worden de metrieken beschreven.

Productiviteit:	Wordt uitgedrukt in een indexwaarde berekend met de volgende formule: $$productiviteit = \frac{omvang}{doorlooptijd^{4/3} * bestedingen^{1/3}}$$
Doorlooptijd:	Wordt uitgedrukt in maanden.
Bestedingen:	Wordt uitgedrukt in mensmaanden of dagen. Eén mensmaand heeft 13,75 dagen.
Bevindingen:	Door test gevonden fouten als maat voor de kwaliteit van het opgeleverde product.
Omvang:	Wordt uitgedrukt in functiepunten of effectieve regels broncode.
Gearingfactor:	Factor per ontwikkelstraat om functiepunten om te rekenen naar regels broncode. Hierdoor kunnen projecten uit verschillende ontwikkelstraten vergeleken worden

5.1.2 Doelstelling onderzoek

De doelstelling van het onderzoek is de betrouwbaarheid van de data van FPA te onderzoeken en vervolgens een model te ontwikkelen om betrouwbare data te kunnen opstellen.
In het huidige systeem van FPA wordt de variabele productiviteit van het ontwikkelen van software beïnvloedt door de variabelen omvang, doorlooptijd en inspanning. Het is de bedoeling om per genoemde variabele periodieke stochastische variabelen te definiëren. Deze periodieke stochastische variabelen moeten leiden tot betrouwbare data. Dit kan worden vastgesteld met correlaties.

5.1.3 Vraagstelling onderzoek

De vraagstelling staat in verband met het uiteindelijke doel van het onderzoek. Het vinden van oplossingen om de betrouwbaarheid van FPA data te analyseren en om de productiviteit van een systeem te voorspellen. De vraagstelling kan dan ook als volgt worden geformuleerd:

- Zijn er statistische modellen om de variabelen van FPA te kunnen analyseren en om de productiviteit van een systeem te voorspellen, die bijdragen aan een hogere betrouwbaarheid?
- Zo ja welk statistisch model is het meest geschikt voor dit probleem?

5.1.4 Randvoorwaarden

Randvoorwaarden:

- Het onderzoek beperkt zich tot B/CICT.
- Er wordt alleen gekeken naar de gegevens van circa 80 afgeronde ontwikkelopdrachten.
- De student zal periodiek/regelmatig naar de Vrije Universiteit gaan om met de begeleider te bespreken over hoe de stand van zaken is en om het probleem goed af te bakenen.
- De student heeft tenminste 1 keer per week een gesprek met de Belastingdienst begeleider over de voortgang van het onderzoek.
- Data over FPA kan beschikbaar worden gesteld met behulp van Quantitatieve Software Management (QSM) methodiek dat functiepuntanalyse ondersteunt.
- Het systeem QSM waar de student mee zal leren omgaan, zal beschikbaar worden gesteld aan de student. Met behulp van dit systeem zal de student in staat zijn om de benodigde data te verkrijgen.
- Met behulp van SPSS zullen de data geanalyseerd worden middels statistische technieken.
- De opdrachtgever verwacht van de onderzoeker tijdens en na het onderzoek geheimhouding van de verkregen informatie.

5.2 Te gebruiken theorieën en concepten

Om de vraagstelling te beantwoorden, kunnen een aantal theorieën en concepten gebruikt worden:

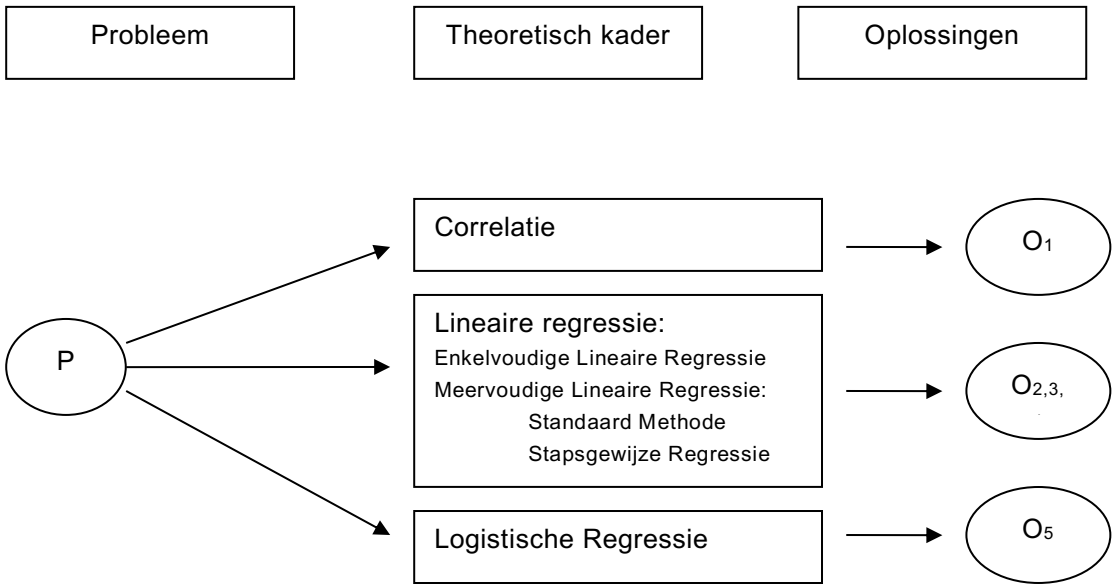
- Correlatie
- Lineaire Regressie
  - Enkelvoudige Lineaire Regressie
  - Meervoudige Lineaire Regressie: Standaard Methode
  - Meervoudige Lineaire Regressie: Stapsgewijze Regressie
- Logistische Regressie

Waarom worden deze theorieën en concepten gebruikt?
In paragraaf 4.2 wordt uitgelegd waarom Correlatie, Enkelvoudige Lineaire Regressie, Meervoudige Lineaire Regressie: Standaard Methode en Meervoudige Lineaire Regressie: Stapsgewijze Regressie gebruikt worden.

Logistische Regressie:

- Met logistische regressie wordt berekend hoe groot de kans is op één van de twee categorieën van een dichotome variabele, op basis van verklarende variabelen.

Figuur 10 Schema Theorieën en concepten



Deze theorieën en concepten zijn nader uitgelegd in bijlage B.

5.3 Gegevensbronnen

Met behulp van de tool QSM kunnen de benodigde data verkregen worden. Deze data zijn beschikbaar gesteld door B/CICT.

5.4 Meet- en waarnemingsmethoden + Analysemethoden

In bijlage B worden de verschillende modellen beschreven die geschikt zijn om de data van Functie Punt Analyse te analyseren en voorspellen. Ook wordt de gebruikte software voor het onderzoek beschreven.

Om de vraagstelling te beantwoorden worden de volgende modellen en software gebruikt:

- Correlatie
- Lineaire Regressie
    - Enkelvoudige Lineaire Regressie
    - Meervoudige Lineaire Regressie: Standaard Methode
    - Meervoudige Lineaire Regressie: Stapsgewijze Regressie
- Logistische Regressie
- QSM

5.5 Rapportage

[vertrouwelijk]

5.5.1 Conclusie

***Zijn er statistische modellen om de variabelen van FPA te kunnen analyseren en om de productiviteit van een systeem te voorspellen, die bijdragen aan een hogere betrouwbaarheid*?**
Zo ja welk statistisch model is het meest geschikt voor dit probleem?

Op basis van de gevonden slechte determinatiecoëfficiënten blijkt dat de werkelijk waarden van de afhankelijke variabelen slecht door de modellen worden benaderd.

Om de Labor Rate te modelleren is er gebruik gemaakt van logistische regressie. Hoe groot de kans is op één van de twee categorieën van een dichotome variabele Labor Rate, op basis van verklarende variabelen. De gevonden slechte R^2 van Nagelkerke duidt erop dat er weinig samenhang bestaat tussen Gearing Factor FP, Effective FP, ORAA Durations, ORAA Effort, Productiviteit PI met Labor Rate.

6 Uitwerking Onderzoeksopzet zuinigheid automobielen

6.1 Probleemstelling

6.1.1 Inleiding

Om de resultaten en conclusies in dit rapport goed te kunnen begrijpen, is het verstandig om de gebruikte variabelen te bestuderen. Deze zijn als volgt:

Model:	model auto (merk en type)
MPG:	aantal miles per gallon
Cylinders:	aantal cylinders
Horsepower:	paardenkracht (pk)
Weight:	gewicht auto
Accelerate:	snelheid van accelereren
Year:	bouwjaar
Origin:	continent van herkomst
Brand:	automerk

6.1.2 Doelstelling onderzoek

De doelstelling van het onderzoek is de data van automobielen te onderzoeken en vervolgens een model te ontwikkelen om de zuinigheid van een automobiel te kunnen voorspellen en analyseren.
De variabele MPG van een automobiel wordt beïnvloed door de variabelen cylinders, horsepower, weight, accelerate, year en origin.

6.1.3 Vraagstelling onderzoek

De vraagstelling staat in verband met het uiteindelijke doel van het onderzoek: het vinden van oplossingen om de zuinigheid van een automobiel te kunnen voorspellen en analyseren. De vraagstelling kan dan ook als volgt worden geformuleerd:

- Zijn er statistische modellen om de zuinigheid van een automobiel te kunnen voorspellen en analyseren?
- Zo ja welk statistisch model is het meest geschikt voor dit probleem?

6.1.4 Randvoorwaarden

Randvoorwaarden:

- Er wordt alleen gekeken naar de data van automobielen met circa 385 gegevens.

- Er wordt niet gekeken naar automobielen die op diesel rijden.
- De student zal periodiek/regelmatig naar de Vrije Universiteit gaan om met de begeleider te bespreken over hoe de stand van zaken is en om het probleem goed af te bakenen.
- Met behulp van Weka en SPSS zullen de data geanalyseerd worden middels statistische technieken en een beslissingsboom

6.2 Te gebruiken theorieën en concepten

Om de vraagstelling te beantwoorden, kunnen een aantal theorieën en concepten gebruikt worden:

- Lineaire Regressie
  - Meervoudige Lineaire Regressie: Standaard Methode
- Beslissingsboom

Waarom worden deze theorieën en concepten gebruikt?

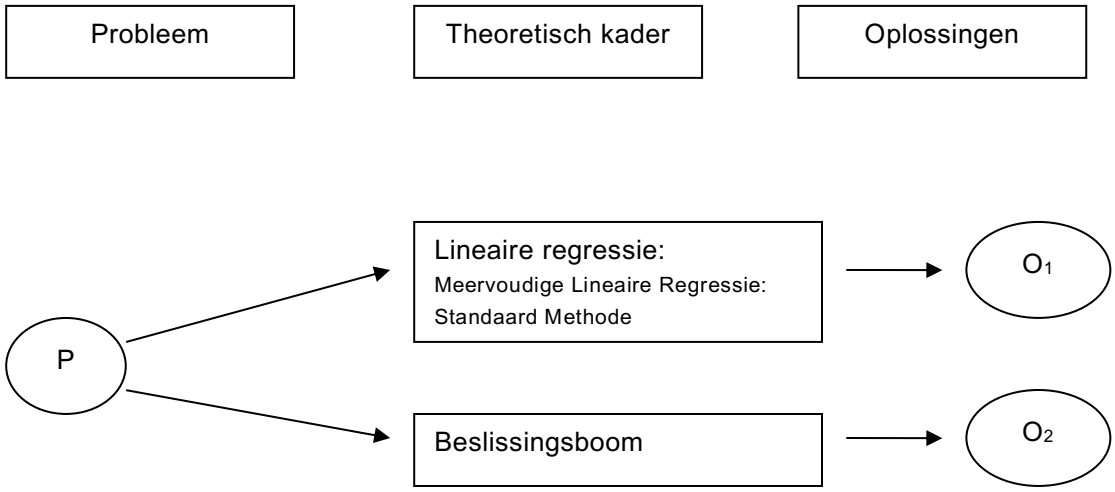
Meervoudige Lineaire Regressie: Standaard Methode:

- Of er een lineair verband is tussen alle verklarende variabelen x_i en één afhankelijke variabele *y*.

Beslissingsboom:

- Met behulp van beslissingsboom kunnen de data geclassificeerd worden door ze te sorteren. Beginnend boven aan de boom (ter hoogte van de *root*) en naar onder toe werkend. Bij elke *node* van de boom wordt een attribuut getest. Afhankelijk van de waarde van het attribuut wordt een bepaalde tak (*branch*) naar beneden toe gevolgd totdat er een nieuwe node bereikt wordt. Daar wordt de bovenstaande procedure herhaald, tenzij het einde van de boom (*leaf*) bereikt is. In dit laatste geval kan het resultaat van de classificatie afgelezen worden.

Figuur 11 Schema Theorieën en concepten



Deze theorieën en concepten zijn nader uitgelegd in bijlage B.

6.3 Gegevensbronnen

De gebruikte data zijn van de volgende site gehaald:
http://mgtclass.mgt.unm.edu/Yourstone/Mgt%20701/Berk%20&%20Carey%20Data/.
De data bestaat uit een serie van 30 verschillende automerken uit Japan, Amerika en Europa.

Alle merken automobielen zijn voorzien van een nummer. Deze zijn als volgend ingedeeld:

0 = Datsun	10 = Mercedes	20 = Opel	30 = Volvo
1 = AMC	11 = Dodge	21 = Peugeot	
2 = Audi	12 = Fiat	22 = Plymouth	
3 = BMW	13 = Ford	23 = Pontiac	
4 = Buick	14 = Hi	24 = Renault	
5 = Cadillac	15 = Honda	25 = Saab	
6 = Capri	16 = Mazda	26 = Subaru	
7 = Chevrolet	17 = Mercury	27 = Toyota	
8 = Chevy	18 = Nissan	28 = Triumph	
9 = Chrysler	19 = Oldsmobile	29 = Volkswagen	

Er wordt aangenomen dat de gebruikte gegevens betrouwbaar en correct zijn.

6.4 Meet- en waarnemingsmethoden + Analysemethoden

In bijlage B worden de verschillende modellen beschreven die geschikt zijn om de zuinigheid van automobielen te kunnen en analyseren. Ook wordt de gebruikte software voor het onderzoek beschreven.

Om de vraagstelling te beantwoorden worden de volgende modellen en software gebruikt:

- Lineaire Regressie
  - Meervoudige Lineaire Regressie: Standaard Methode
- Beslissingsboom
- Weka

6.5 Rapportage

In deze paragraaf wordt een model ontwikkeld om de zuinigheid van een automobiel te kunnen voorspellen en analyseren. Hier wordt aangenomen dat de variabele MPG van een automobiel beïnvloed wordt door de variabelen cylinders, horsepower, weight, accelerate, year, en origin.
De data is bewerkt door de automobielen te classificeren in niet Amerikaans – Amerikaans door middel van 1 – 0 variabelen. Dit is als volgt:
Niet Amerikaans = 0
Amerikaans = 1
De verwachting is dat Amerikaanse automobielen meer benzine verbruiken dan niet Amerikaanse automobielen.

6.5.1 Meervoudige Lineaire Regressie: Standaard Methode

Onderzocht wordt of er een lineair verband is tussen alle verklarende variabelen x_i en één afhankelijke variabele y. Dit kan men vastleggen in een standaard meervoudig lineair regressiemodel als het verband rechtlijnig van aard is. De verklarende variabelen zijn cylinders, horsepower, weight, accelerate, year, en origin en de afhankelijke variabele is MPG.
Dit model is ontwikkeld met behulp van 10-fold cross-validation in Weka. Deze wordt geïllustreerd in de volgende figuur met de bijbehorende regressieresultaten.

Figuur 12 Regressieresultaten met MPG als afhankelijke variabele

```
Instances:    385
Attributes:   7
              MPG
              Cylinders
              Horsepower
              Weight
              Accelerate
              Year
              Origin
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Linear Regression Model

MPG = -0.006  * Weight + 0.7262 * Year -1.7783 * Origin - 13.0155

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9087
Mean absolute error                      2.4341
Root mean squared error                  3.1642
Relative absolute error                 37.8765 %
Root relative squared error             41.5368 %
Total Number of Instances              385
```

Aan de hand van de data is de correlation coefficient, mean absolute error, root mean squared error, relative absolute error, root relative squared error en total number of instances berekend.
Het model ziet als volgt uit:

MPG = –13.0155 – 0.006 * Weight + 0,7262 * Year – 1,7783 * Origin

De volgende tabel toont een gedeelte van de resultaten van de werkelijke waarden weight, year, origin en MPG. Met behulp van het lineaire model zijn de regressiewaarden van MPG berekend. In de laatste kolom van de tabel wordt het absolute verschil tussen de werkelijke waarden en regressiewaarden van MPG duidelijk gemaakt.

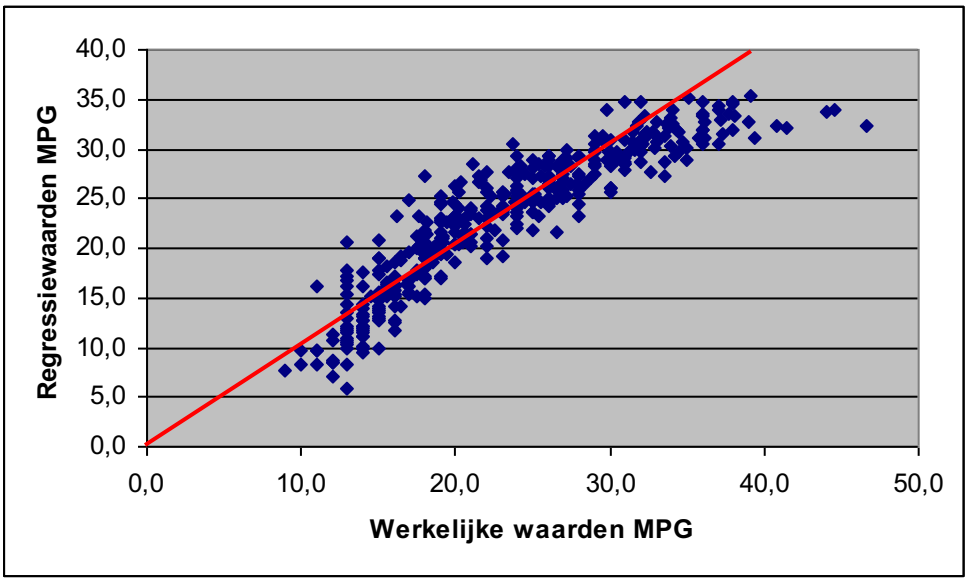
Figuur 13 Tabel resultaten werkelijke waarden en regressiewaarden

Weight	Year	Origin	MPG	Model MPG	Absolute verschil
4732	70	1	9,0	7,6	1,4
4376	70	1	10,0	9,8	0,2
4615	70	1	10,0	8,4	1,6
4997	73	1	11,0	8,2	2,8
4382	70	1	11,0	9,7	1,3
4633	72	1	11,0	9,7	1,3
3664	73	1	11,0	16,2	5,2
4951	73	1	12,0	8,5	3,5
4955	71	1	12,0	7,0	5,0
4906	73	1	12,0	8,8	3,2
4952	73	1	12,0	8,5	3,5
4456	72	1	12,0	10,8	1,2
4499	73	1	12,0	11,2	0,8
3821	73	1	13,0	15,3	2,3
4100	73	1	13,0	13,6	0,6
4699	74	1	13,0	10,8	2,2
4502	72	1	13,0	10,5	2,5
4464	73	1	13,0	11,4	1,6
4098	72	1	13,0	12,9	0,1
4274	72	1	13,0	11,8	1,2
3988	73	1	13,0	14,3	1,3
4055	76	1	13,0	16,1	3,1
4735	73	1	13,0	9,8	3,2
4422	72	1	13,0	11,0	2,0
3755	76	1	13,0	17,9	4,9
4746	71	1	13,0	8,3	4,7
3870	76	1	13,0	17,2	4,2
4294	72	1	13,0	11,7	1,3
4363	73	1	13,0	12,0	1,0
3169	75	1	13,0	20,7	7,7
4654	73	1	13,0	10,3	2,7
3940	76	1	13,0	16,8	3,8
5140	71	1	13,0	5,9	7,1
3672	73	1	14,0	16,2	2,2
4257	74	1	14,0	13,4	0,6
3086	70	1	14,0	17,5	3,5
4354	70	1	14,0	9,9	4,1
4209	71	1	14,0	11,5	2,5
4457	74	1	14,0	12,2	1,8
4154	71	1	14,0	11,8	2,2
Gem. abs. verschil					**2,6**

Zodat er duidelijk geanalyseerd kan worden of de regressiewaarden gelijk zijn met de werkelijke waarden

zijn alle resultaten in een grafiek gepresenteerd. Mocht de puntenwolk de vorm van een positieve schuine rechte lijn hebben met gelijke x-coördinaat en y-coördinaat dan is het regressiemodel betrouwbaar.

Figuur 14 Grafiek regressiewaarden MPG – werkelijke waarden MPG



6.5.1.1 Conclusie Meervoudige Lineaire Regressie: Standaard Methode

Onderzocht is op meervoudig lineair verband tussen de verklarende variabelen cylinders, horsepower, weight, accelerate, year, origin en de afhankelijke variabele is MPG.
Er is sprake van een significant regressiemodel met als verklarende variabelen weight, year, origin en afhankelijke variabele MPG. De gevonden **goede** correlatie coëfficiënt 0,9087 duidt erop de samenhang tussen de werkelijke waarden en de regressiewaarden van MPG. Hoe dichter deze correlatie coëfficiënt bij 1 ligt, hoe beter de werkelijke waarden van MPG benaderd worden door het model. Tevens blijkt uit de grafiek van figuur 33 dat de regressiewaarden redelijk gelijk zijn met de werkelijke waarden van MPG.

6.5.2 Beslissingsboom

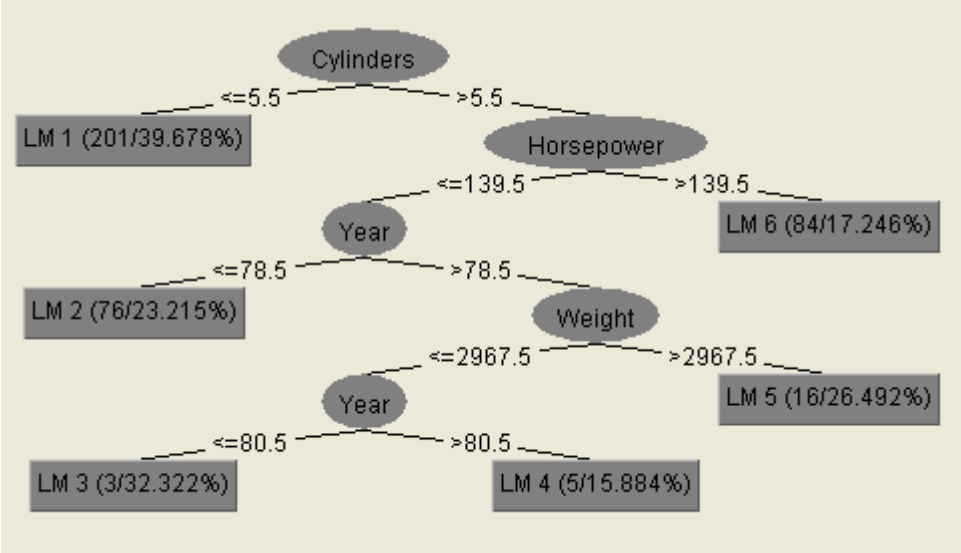
Met behulp van een statistische techniek, regressieboom-analyse, is geanalyseerd in hoeverre de zuinigheid van automobielen verklaard kan worden door de variabelen cylinders, horsepower, weight, accelerate, year, origin. Met andere woorden de verklarende variabelen zijn cylinders, horsepower, weight, accelerate, year, origin en de afhankelijke variabele is MPG. In een regressieboom worden de waarden van MPG optimaal gesplitst op basis van de verklarende variabelen cylinders, horsepower, weight, accelerate, year, origin.
Dit model is ontwikkeld met behulp van het commando trees.M5P in Weka. Dit wordt geïllustreerd in het volgende figuur met de bijbehorende regressieresultaten.

Figuur 15 Regressieresultaten met MPG als afhankelijke variabele

```
Instances:    385
Attributes:   7
              MPG, Cylinders, Horsepower, Weight, Accelerate, Year, Origin
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

M5 pruned model tree:
(using smoothed linear models)

Cylinders <= 5.5 : LM1 (201/39.678%)
Cylinders >  5.5 :
|   Horsepower <= 139.5 :
|   |   Year <= 78.5 : LM2 (76/23.215%)
|   |   Year >  78.5 :
|   |   |   Weight <= 2967.5 :
|   |   |   |   Year <= 80.5 : LM3 (3/32.322%)
|   |   |   |   Year >  80.5 : LM4 (5/15.884%)
|   |   |   Weight >  2967.5 : LM5 (16/26.492%)
|   Horsepower >  139.5 : LM6 (84/17.246%)

LM num: 1
MPG =    -0.0624 * Horsepower - 0.0073 * Weight + 0.8895 * Year - 1.3265 * Origin -
          17.4086
LM num: 2
MPG =    -0.0572 * Cylinders - 0.0224 * Horsepower - 0.0039 * Weight –
          0.0297 * Accelerate + 0.3795 * Year - 0.134 * Origin + 6.3416
LM num: 3
MPG =    -0.0572 * Cylinders + 0.0589 * Horsepower - 0.0031 * Weight +
          0.2635 * Accelerate - 0.1867 * Year - 0.134 * Origin + 39.2044
LM num: 4
MPG =    -0.0572 * Cylinders + 0.0589 * Horsepower - 0.0034 * Weight +
          0.2635 * Accelerate - 0.1411 * Year - 0.134 * Origin + 36.0392
LM num: 5
MPG =    -0.0572 * Cylinders + 0.0432 * Horsepower - 0.0048 * Weight +
          0.3903 * Accelerate + 0.2697 * Year - 0.134 * Origin + 4.87
LM num: 6
MPG =    -0.0664 * Cylinders - 0.0201 * Horsepower - 0.0021 * Weight –
          0.3552 * Accelerate + 0.4362 * Year - 0.134 * Origin - 0.446

Correlation coefficient                  0.9334
Mean absolute error                      1.949
Root mean squared error                  2.7212
Relative absolute error                  30.3279 %
Root relative squared error              35.7216 %
Total Number of Instances                385
```

De bijbehorende regressieboom wordt in de volgende figuur getoond.

Figuur 16 Regressieboom met MPG als afhankelijke variabele



Aan de hand van de data is de correlation coefficient, mean absolute error, root mean squared error, relative absolute error, root relative squared error en total number of instances berekend.
De modellen zien er als volgt uit:

- 201 van de 385 automobielen heeft het aantal cylinders kleiner of gelijk aan 5,5
  LM1: MPG = -0.0624 * Horsepower - 0.0073 * Weight + 0.8895 * Year - 1.3265 * Origin – 17.4086
- 76 van de 385 automobielen heeft het aantal cylinders groter dan 5,5. Het aantal paardenkracht kleiner of gelijk aan 139,5 en het bouwjaar is kleiner of gelijk aan 78,5.
  LM2: MPG = -0.0572 * Cylinders - 0.0224 * Horsepower - 0.0039 * Weight – 0.0297 * Accelerate + 0.3795 * Year - 0.134 * Origin + 6.3416
- 3 van de 385 automobielen heeft het aantal cylinders groter dan 5,5. Het aantal paardenkracht kleiner of gelijk aan 139,5. Het bouwjaar is kleiner of gelijk aan 80,5, maar groter dan 78,5. Het gewicht is kleiner of gelijk aan 2967,5.
  LM3: MPG = -0.0572 * Cylinders + 0.0589 * Horsepower - 0.0031 * Weight + 0.2635 * Accelerate - 0.1867 * Year - 0.134 * Origin + 39.2044

- 5 van de 385 automobielen heeft het aantal cylinders groter dan 5,5. Het aantal paardenkracht kleiner of gelijk aan 139,5. Het bouwjaar is groter dan 80,5. Het gewicht is kleiner of gelijk aan 2967,5.
  LM4: MPG = -0.0572 * Cylinders + 0.0589 * Horsepower - 0.0034 * Weight + 0.2635 * Accelerate - 0.1411 * Year - 0.134 * Origin + 36.0392
- 16 van de 385 automobielen heeft het aantal cylinders groter dan 5,5. Het aantal paardenkracht kleiner of gelijk aan 139,5. Het bouwjaar is groter dan 78,5. Het gewicht is groter dan 2967,5.
  LM5: MPG = -0.0572 * Cylinders + 0.0432 * Horsepower - 0.0048 * Weight + 0.3903 * Accelerate + 0.2697 * Year - 0.134 * Origin + 4.87
- 84 van de 385 automobielen heeft het aantal cylinders groter dan 5,5 en het aantal paardenkracht groter dan 139,5.
  LM6: MPG = -0.0664 * Cylinders - 0.0201 * Horsepower - 0.0021 * Weight – 0.3552 * Accelerate + 0.4362 * Year - 0.134 * Origin - 0.446

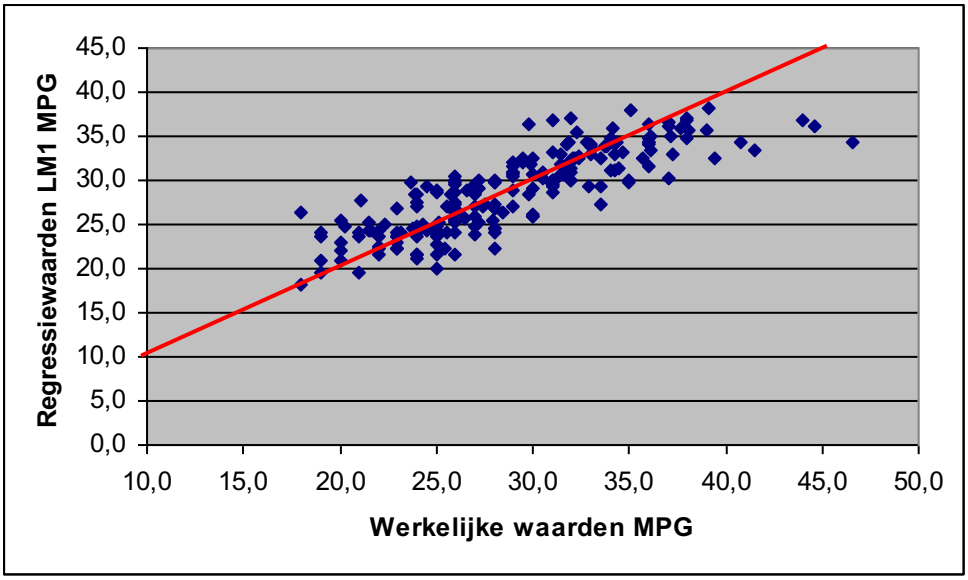
De volgende tabel toont de resultaten van de werkelijke waarden van horsepower, weight, year, origin en MPG. Met behulp van de lineaire modellen LM1 zijn de regressiewaarden van MPG berekend. In de laatste kolom van de tabel wordt het absolute verschil tussen de werkelijke waarden en regressiewaarden van MPG duidelijk gemaakt.

Figuur 17 Tabel resultaten werkelijke waarden en regressiewaarden LM1

Horsepower	Weight	Year	Origin	MPG	LM1 MPG	Absolute verschil
90	2124	73	0	18,0	26,4	8,4
112	2933	72	0	18,0	18,2	0,2
97	2330	72	0	19,0	23,6	4,6
85	2310	73	1	19,0	24,0	5,0
88	3270	76	0	19,0	20,8	1,8
112	2868	73	0	19,0	19,6	0,6
91	2582	73	0	20,0	23,0	3,0
90	2408	72	1	20,0	22,1	2,1
88	2279	73	0	20,0	25,4	5,4
102	3150	76	0	20,0	20,8	0,8
103	2830	78	0	20,3	24,9	4,6
72	2401	73	1	21,0	24,2	3,2
86	2226	72	1	21,0	23,7	2,7
87	2979	72	0	21,0	19,5	1,5
95	2515	78	0	21,1	27,7	6,6
110	2720	77	0	21,5	24,4	2,9
110	2600	77	0	21,5	25,2	3,7
115	2795	78	0	21,6	24,4	2,8
72	2408	71	1	22,0	22,3	0,3
94	2379	73	0	22,0	24,3	2,3
86	2395	72	1	22,0	22,5	0,5
76	2511	72	0	22,0	23,6	1,6
98	2945	75	0	22,0	21,7	0,3
88	2890	79	1	22,3	24,9	2,6
95	2694	75	0	23,0	23,7	0,7
83	2639	75	1	23,0	23,5	0,5
86	2220	71	1	23,0	22,8	0,2
88	2957	75	0	23,0	22,2	0,8
78	2592	75	1	23,0	24,2	1,2
97	2506	72	0	23,0	22,3	0,7
54	2254	72	0	23,0	26,8	3,8
105	2745	78	1	23,2	24,1	0,9
100	2420	80	0	23,7	29,8	6,1
85	2855	78	1	23,8	24,5	0,7
97	2405	78	0	23,9	28,4	4,5
90	2430	70	0	24,0	21,5	2,5
97	2545	75	0	24,0	24,7	0,7
75	2108	74	0	24,0	28,3	4,3
92	2865	82	1	24,0	27,5	3,5
97	2489	74	0	24,0	24,2	0,2
Gem. abs. verschil						**2,5**

Zodat er duidelijk geanalyseerd kan worden of de regressiewaarden gelijk zijn met de werkelijke waarden zijn alle resultaten in een grafiek gepresenteerd. Mocht de puntenwolk de vorm van een positieve schuine rechte lijn hebben met gelijke x-coördinaat en y-coördinaat dan is het regressiemodel betrouwbaar.

Figuur 18 Grafiek regressiewaarden LM1 MPG – werkelijke waarden MPG



De volgende tabel toont de resultaten van de werkelijke waarden van cylinders, horsepower, weight, accelerate, year, origin en MPG. Met behulp van de lineaire modellen LM2 zijn de regressiewaarden van MPG berekend. In de laatste kolom van de tabel wordt het absolute verschil tussen de werkelijke waarden en regressiewaarden van MPG duidelijk gemaakt.

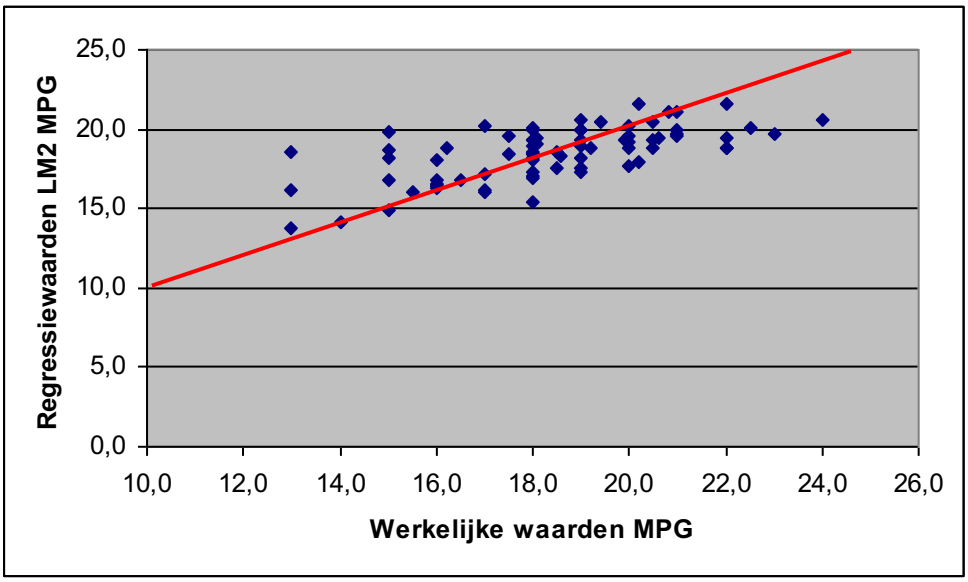
Figuur 19 Tabel resultaten werkelijke waarden en regressiewaarden LM2

Cylinders	Horsepower	Weight	Accelerate	Year	Origin	MPG	LM2 MPG	Absolute verschil
8	130	4098	14,0	72	1	13,0	13,8	0,8
8	129	3169	12,0	75	1	13,0	18,6	5,6
8	130	3870	15,0	76	1	13,0	16,1	3,1
8	137	4042	14,5	73	1	14,0	14,2	0,2
6	100	3336	17,0	74	1	15,0	18,2	3,2
6	72	3158	19,5	75	1	15,0	19,8	4,8
6	72	3432	21,0	75	1	15,0	18,7	3,7
6	110	3730	19,0	75	1	15,0	16,8	1,8
8	130	4295	14,9	77	1	15,0	14,9	0,1
8	120	3962	13,9	76	1	15,5	16,0	0,5
6	105	3439	15,5	71	1	16,0	16,6	0,6
6	100	3278	18,0	73	1	16,0	18,0	2,0
6	100	3781	17,0	74	1	16,0	16,5	0,5
6	110	3632	18,0	74	1	16,0	16,8	0,8
6	105	3897	18,5	75	1	16,0	16,2	0,2
6	133	3410	15,8	78	0	16,2	18,9	2,7
6	120	3820	16,7	76	0	16,5	16,8	0,3
6	100	3329	15,5	71	1	17,0	17,1	0,1
6	110	3907	21,0	75	1	17,0	16,0	1,0
8	110	4060	19,0	77	1	17,0	16,1	0,9
6	125	3140	13,6	78	0	17,0	20,1	3,1
6	95	3193	17,8	76	1	17,5	19,6	2,1
6	110	3520	16,4	77	1	17,5	18,4	0,9
6	97	2774	15,5	70	1	18,0	19,0	1,0
8	130	3504	12,0	70	1	18,0	15,4	2,6
6	88	3139	14,5	71	1	18,0	18,2	0,2
6	100	3288	15,5	71	1	18,0	17,3	0,7
6	110	2962	13,5	71	1	18,0	18,4	0,4
6	88	3021	16,5	73	1	18,0	19,3	1,3
6	100	2789	15,0	73	1	18,0	20,0	2,0
6	100	2945	16,0	73	1	18,0	19,4	1,4
6	105	3121	16,5	73	1	18,0	18,6	0,6
6	105	3613	16,5	74	1	18,0	17,0	1,0
6	95	3785	19,0	75	1	18,0	16,9	1,1
6	97	2984	14,5	75	1	18,0	20,1	2,1
6	105	3459	16,0	75	1	18,0	18,0	0,0
6	78	3574	21,0	76	1	18,0	18,4	0,4
6	120	3410	15,1	78	1	18,1	19,0	0,9
8	139	3205	11,2	78	1	18,1	19,4	1,3
6	110	3645	16,2	76	1	18,5	17,5	1,0
Gem. abs. verschil								**1,4**

Zodat er duidelijk geanalyseerd kan worden of de regressiewaarden gelijk zijn met de werkelijke waarden

zijn alle resultaten in een grafiek gepresenteerd. Mocht de puntenwolk de vorm van een positieve schuine rechte lijn hebben met gelijke x-coördinaat en y-coördinaat dan is het regressiemodel betrouwbaar.

Figuur 20 Grafiek regressiewaarden LM2 MPG – werkelijke waarden MPG



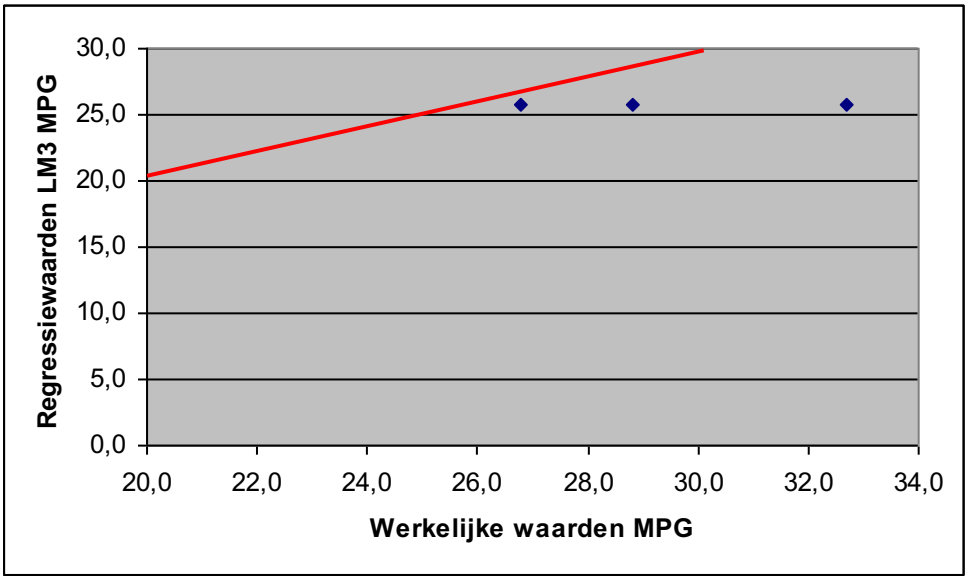
De volgende tabel toont de resultaten van de werkelijke waarden van cylinders, horsepower, weight, accelerate, year, origin en MPG. Met behulp van de lineaire modellen LM3 zijn de regressiewaarden van MPG berekend. In de laatste kolom van de tabel wordt het absolute verschil tussen de werkelijke waarden en regressiewaarden van MPG duidelijk gemaakt.

Figuur 21 Tabel resultaten werkelijke waarden en regressiewaarden LM3

Cylinders	Horsepower	Weight	Accelerate	Year	Origin	MPG	LM3 MPG	Absolute verschil
6	115	2700	12,9	79	1	26,8	25,8	1,0
6	115	2595	11,3	79	1	28,8	25,7	3,1
6	132	2910	11,4	80	0	32,7	25,7	7,0
Gem. abs. veschil								**3,7**

Zodat er duidelijk geanalyseerd kan worden of de regressiewaarden gelijk zijn met de werkelijke waarden zijn alle resultaten in een grafiek gepresenteerd. Mocht de puntenwolk de vorm van een positieve schuine rechte lijn hebben met gelijke x-coördinaat en y-coördinaat dan is het regressiemodel betrouwbaar.

Figuur 22 Grafiek regressiewaarden LM3 MPG – werkelijke waarden MPG



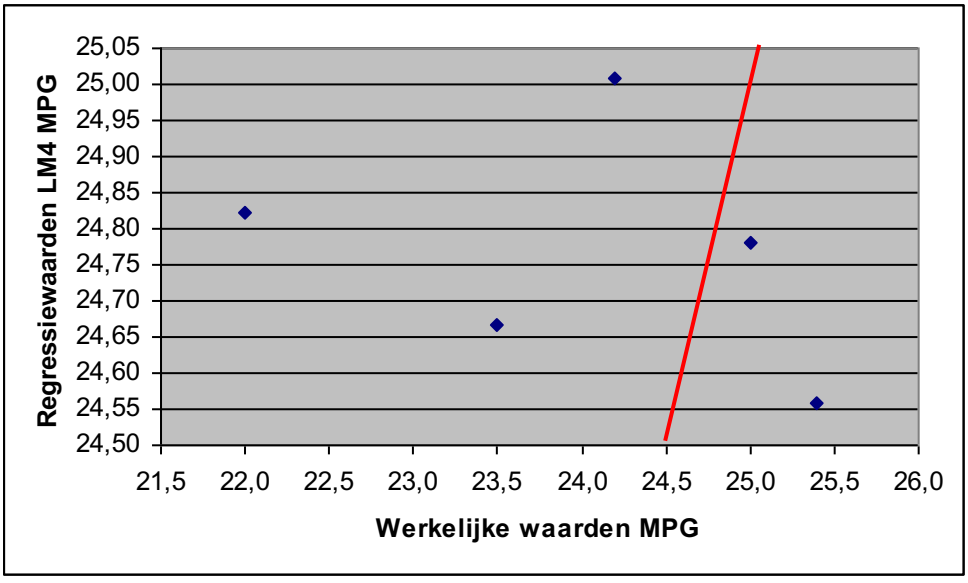
De volgende tabel toont de resultaten van de werkelijke waarden van cylinders, horsepower, weight, accelerate, year, origin en MPG. Met behulp van de lineaire modellen LM4 zijn de regressiewaarden van MPG berekend. In de laatste kolom van de tabel wordt het absolute verschil tussen de werkelijke waarden en regressiewaarden van MPG duidelijk gemaakt.

Figuur 23 Tabel resultaten werkelijke waarden en regressiewaarden LM4

Cylinders	Horsepower	Weight	Accelerate	Year	Origin	MPG	LM4 MPG	Absolute verschil
6	112	2835	14,7	82	1	22,0	24,8	2,8
6	110	2725	12,6	81	1	23,5	24,7	1,2
6	120	2930	13,8	81	0	24,2	25,0	0,8
6	110	2945	16,4	82	1	25,0	24,8	0,2
6	116	2900	12,6	81	0	25,4	24,6	0,8
Gem. abs. verschil								**1,2**

Zodat er duidelijk geanalyseerd kan worden of de regressiewaarden gelijk zijn met de werkelijke waarden zijn alle resultaten in een grafiek gepresenteerd. Mocht de puntenwolk de vorm van een positieve schuine rechte lijn hebben met gelijke x-coördinaat en y-coördinaat dan is het regressiemodel betrouwbaar.

Figuur 24 Grafiek regressiewaarden LM4 MPG – werkelijke waarden MPG



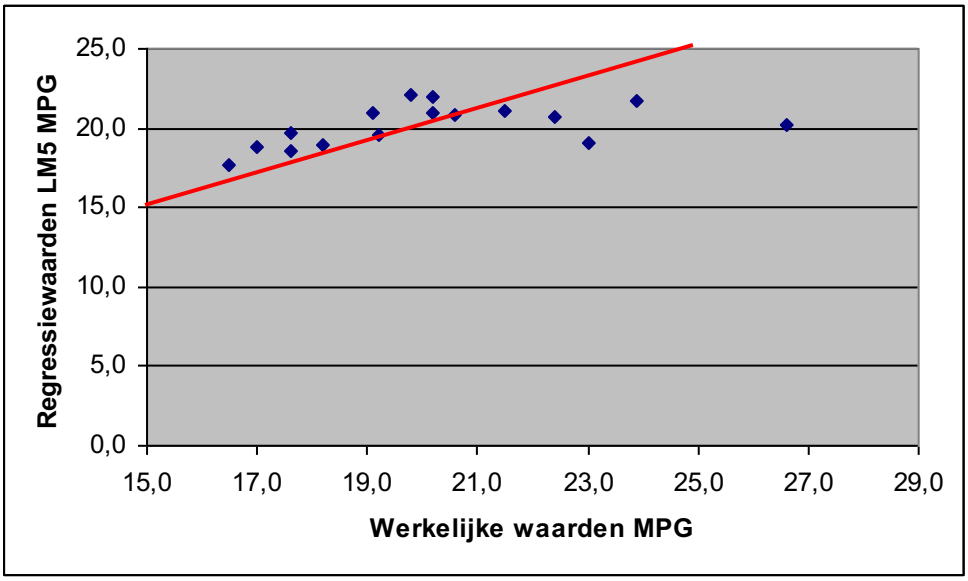
De volgende tabel toont de resultaten van de werkelijke waarden van cylinders, horsepower, weight, accelerate, year, origin en MPG. Met behulp van de lineaire modellen LM5 zijn de regressiewaarden van MPG berekend. In de laatste kolom van de tabel wordt het absolute verschil tussen de werkelijke waarden en regressiewaarden van MPG duidelijk gemaakt.

Figuur 25 Tabel resultaten werkelijke waarden en regressiewaarden LM5

Cylinders	Horsepower	Weight	Accelerate	Year	Origin	MPG	LM5 MPG	Absolute verschil
8	138	3955	13,2	79	1	16,5	17,7	1,2
8	130	3840	15,4	79	1	17,0	18,8	1,8
6	85	3465	16,6	81	1	17,6	19,8	2,2
8	129	3725	13,4	79	1	17,6	18,5	0,9
8	135	3830	15,2	79	1	18,2	19,0	0,8
6	90	3381	18,7	80	1	19,1	20,9	1,8
8	125	3605	15,0	79	1	19,2	19,5	0,3
6	85	2990	18,2	79	1	19,8	22,1	2,3
6	88	3060	17,1	81	1	20,2	22,0	1,8
6	90	3265	18,2	79	1	20,2	21,0	0,8
6	110	3360	16,6	79	1	20,6	20,8	0,2
6	115	3245	15,4	79	1	21,5	21,1	0,4
6	110	3415	15,8	81	1	22,4	20,8	1,6
8	125	3900	17,4	79	1	23,0	19,1	3,9
8	90	3420	22,2	79	1	23,9	21,7	2,2
8	105	3725	19,0	81	1	26,6	20,2	6,4
Gem. abs. verschil								**1,8**

Zodat er duidelijk geanalyseerd kan worden of de regressiewaarden gelijk zijn met de werkelijke waarden zijn alle resultaten in een grafiek gepresenteerd. Mocht de puntenwolk de vorm van een positieve schuine rechte lijn hebben met gelijke x-coördinaat en y-coördinaat dan is het regressiemodel betrouwbaar.

Figuur 26 Grafiek regressiewaarden LM5 MPG – werkelijke waarden MPG



De volgende tabel toont de resultaten van de werkelijke waarden van cylinders, horsepower, weight, accelerate, year, origin en MPG. Met behulp van de lineaire modellen LM6 zijn de regressiewaarden van MPG berekend. In de laatste kolom van de tabel wordt het absolute verschil tussen de werkelijke waarden en regressiewaarden van MPG duidelijk gemaakt.

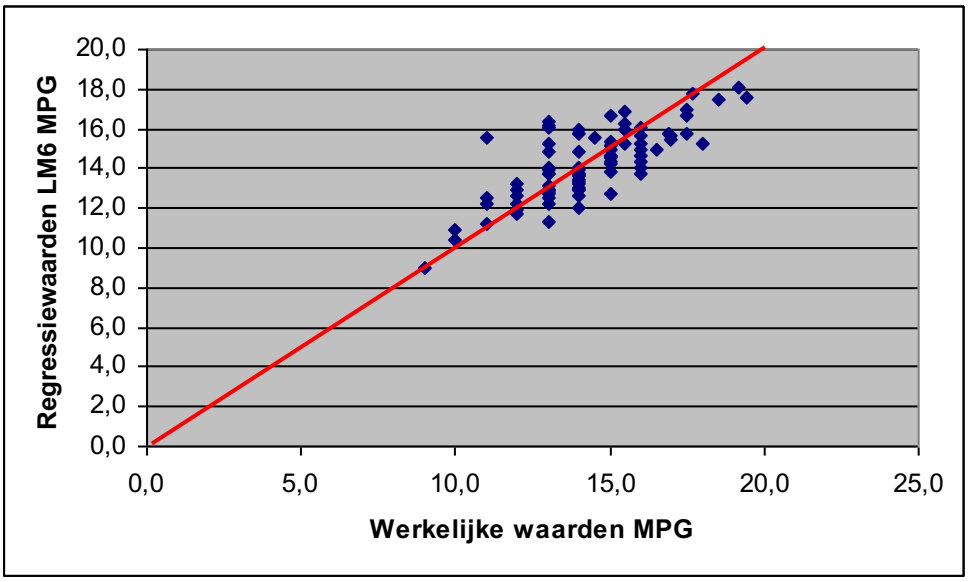
Figuur 27 Tabel resultaten werkelijke waarden en regressiewaarden LM6

Cylinders	Horsepower	Weight	Accelerate	Year	Origin	MPG	LM6 MPG	Absolute verschil
8	193	4732	18,5	70	1	9,0	9,0	0,0
8	200	4376	15,0	70	1	10,0	10,9	0,9
8	215	4615	14,0	70	1	10,0	10,4	0,4
8	150	4997	14,0	73	1	11,0	12,2	1,2
8	180	3664	11,0	73	1	11,0	15,5	4,5
8	208	4633	11,0	72	1	11,0	12,5	1,5
8	210	4382	13,5	70	1	11,0	11,2	0,2
8	160	4456	13,5	72	1	12,0	12,9	0,9
8	167	4906	12,5	73	1	12,0	12,6	0,6
8	180	4955	11,5	71	1	12,0	11,8	0,2
8	180	4499	12,5	73	1	12,0	13,2	1,2
8	198	4952	11,5	73	1	12,0	12,3	0,3
8	225	4951	11,0	73	1	12,0	11,9	0,1
8	140	4294	16,0	72	1	13,0	12,8	0,2
8	145	3988	13,0	73	1	13,0	14,8	1,8
8	145	4055	12,0	76	1	13,0	16,3	3,3
8	150	4699	14,5	74	1	13,0	13,1	0,1
8	150	4464	12,0	73	1	13,0	14,1	1,1
8	150	3755	14,0	76	1	13,0	16,2	3,2
8	150	3940	13,2	76	1	13,0	16,1	3,1
8	155	4502	13,5	72	1	13,0	12,9	0,1
8	158	4363	13,0	73	1	13,0	13,8	0,8
8	165	4274	12,0	72	1	13,0	13,7	0,7
8	170	4746	12,0	71	1	13,0	12,2	0,8
8	170	4654	13,0	73	1	13,0	12,9	0,1
8	175	3821	11,0	73	1	13,0	15,3	2,3
8	175	4100	13,0	73	1	13,0	14,0	1,0
8	175	5140	12,0	71	1	13,0	11,3	1,7
8	190	4422	12,5	72	1	13,0	12,8	0,3
8	215	4735	11,0	73	1	13,0	12,6	0,4
8	140	4638	16,0	74	1	14,0	12,9	1,1
8	148	4657	13,5	75	1	14,0	14,1	0,1
8	150	3672	11,5	73	1	14,0	15,9	1,9
8	150	4257	15,5	74	1	14,0	13,7	0,3
8	150	4457	13,5	74	1	14,0	14,0	0,0
8	150	4237	14,5	73	1	14,0	13,7	0,3
8	150	4096	13,0	71	1	14,0	13,6	0,4
8	150	4077	14,0	72	1	14,0	13,7	0,3
8	153	4154	13,5	71	1	14,0	13,3	0,7
8	153	4129	13,0	72	1	14,0	13,9	0,1
Gem. abs. verschil								**1,0**

Zodat er duidelijk geanalyseerd kan worden of de regressiewaarden gelijk zijn met de werkelijke waarden zijn alle resultaten in een grafiek gepresenteerd. Mocht de puntenwolk de vorm van een positieve schuine

rechte lijn hebben met gelijke x-coördinaat en y-coördinaat dan is het regressiemodel betrouwbaar.

Figuur 28 Grafiek regressiewaarden LM6 MPG – werkelijke waarden MPG



6.5.2.1 Conclusie beslissingsboom

In paragraaf 6.5.2 is er met behulp van een statistische techniek, regressieboom-analyse, is geanalyseerd in hoeverre de zuinigheid van automobielen verklaard kunnen worden door cylinders, horsepower, weight, accelerate, year, origin. Met behulp van de regressieboom zijn de volgende modellen:

- Een regressiemodel, aantal cylinders kleiner of gelijk aan 5,5. Met als verklarende variabelen horsepower, weight, year, origin en afhankelijke variabele MPG.
- Een regressiemodel, aantal cylinders groter dan 5,5. Het aantal paardenkracht kleiner of gelijk aan 139,5 en het bouwjaar is kleiner of gelijk aan 78,5. Met als verklarende variabelen cylinders, horsepower, weight, accelerate, year, origin en afhankelijke variabele MPG.
- Een regressiemodel, aantal cylinders groter dan 5,5. Het aantal paardenkracht kleiner of gelijk aan 139,5. Het bouwjaar is kleiner of gelijk aan 80,5, maar groter dan 78,5. Het gewicht is kleiner of gelijk aan 2967,5. Met als verklarende variabelen cylinders, horsepower, weight, accelerate, year, origin en afhankelijke variabele MPG.
- Een regressiemodel, aantal cylinders groter dan 5,5. Het aantal paardenkracht kleiner of gelijk aan 139,5. Het bouwjaar is groter dan 80,5. Het gewicht is kleiner of gelijk aan 2967,5. Met als verklarende variabelen cylinders, horsepower, weight, accelerate, year, origin en afhankelijke variabele MPG.
- Een regressiemodel, aantal cylinders groter dan 5,5. Het aantal paardenkracht kleiner of gelijk aan 139,5. Het bouwjaar is groter dan 78,5. Het gewicht is groter dan 2967,5. Met als verklarende variabelen cylinders, horsepower, weight, accelerate, year, origin en afhankelijke variabele MPG.
- Een regressiemodel, aantal cylinders groter dan 5,5 en het aantal paardenkracht groter dan 139,5. Met als verklarende variabelen cylinders, horsepower, weight, accelerate, year, origin en afhankelijke variabele MPG.

De gevonden **goede** correlatie coëfficiënt 0,9334 duidt erop de samenhang tussen de werkelijke waarden en de regressiewaarden van MPG (hoe dichter deze correlatie coëfficiënt bij 1 ligt, hoe beter de werkelijke waarden van MPG benaderd worden door het model).

6.5.3 Conclusie

Zijn er statistische modellen om de zuinigheid van een automobiel te kunnen voorspellen en analyseren?
Zo ja welk statistisch model is het meest geschikt voor dit probleem?

Op basis van de correlatie coëfficiënt blijkt dat de regressieboom-analyse een betere samenhang heeft tussen de werkelijke waarden en de regressiewaarden van MPG dan meervoudige lineaire regressie. Daarnaast blijkt ook dat de gemiddelde absolute fout kleiner is. De gemiddelde absolute fout is een hoeveelheid die is gebruikt om te meten hoe dicht de regressiewaarden zijn tot de uiteindelijke waarden. Het is een sommatie van het absolute verschil tussen de regressiewaarden en de uiteindelijke waarden gedeeld door het aantal waarnemingen. Dit betekent hoe kleiner de gemiddelde absolute fout hoe betrouwbaarder het model bij gelijke aantal waarnemingen. Om deze redenen gaat de voorkeur naar de regressieboom-analyse.

7 Referenties

Boer, P. de, Brouwers, M.P., Koetzier, W. (1998); *Basisboek Bedrijfseconomie*; Wolters Noordhoff

Buijs, A. (1997)*; Statistiek om mee te werken*; Educatieve Partners Nederland

Buijs, A. (1998); *Statistiek om mee verder te werken*; Stenfert Kroese

Gunst, M. de, (2006)*; Statistical Models*; Collegedictaat; Vrije Universiteit, Amsterdam

Gunst, M. de, Vaart van der A.W. (2005); *Statistische Data Analyse*; Collegedictaat Vrije Universiteit, Amsterdam;

Howitt, D., Cramer, D (2004); *Statistiek met SPSS 11 voor Windows*; Pearson Education

Leeuw, A.C.J. de (2001); *Bedrijfskundige Methodologie*; van Gorcum

Vaart, A. van der (2003); *Algemene Statistiek*; Collegedictaat; Vrije Universiteit

Vocht, A. de (2007); *Basishandboek SPSS 15 voor Windows*; Bijleveld Press

Witten, I., Frank E. (2005); *Data Mining; Practical Machine Learning Tools and Techniques*; Morgan Kaufmann Publishers

B/CICT, (2008), Intranet

Eindrapportage FPA en QSM (2007), B/CICT

Jaarverslagen (1996-2007) met Kasstroomoverzichten, B/CICT

8 Bijlagen

8.1 Bijlage A: Lijst afkortingen

B/CICT	Belastingdienst Centrum voor Informatie Technologie
B/CPP	Belastingdienst/Centrum voor proces- en productontwikkeling
B/CA	Belastingdienst/Centrale administratie
PFC	Sector Planning, Financiën en Control
ON	Sector Ontwikkeling
FPA	Functie punt analyse
CI	Customer Intimacy
BLS	Baten/lasten-stelsel
QSM	Quantitatieve Software Management
OLS	Ordinary Least Squares
Weka	Waikato Environment for Knowledge Analysis
SAP	Systeme, Anwendungen, Produkte in der Datenverarbeitung of Systems, Applications, Products
SPSS	Statistical Package for the Social Sciences

8.2 Bijlage B: Meet- en Waarnemingsmethoden + Analysemethoden

In deze bijlage worden de verschillende modellen beschreven en de software die gebruikt is voor het onderzoek naar de Kasstroom en de Functie Punt Analyse.

8.2.1 Toetsingprocedure

Een onderzoek moet leiden tot een samenvattende uitspraak of conclusie. Daarin wordt een antwoord gegeven op de gestelde onderzoeksvraag. Die uitspraak wordt meestal gedaan op grond van steekproefonderzoek, maar betreft de populatie, waaruit de steekproef is getrokken. Wetenschappelijke conclusies hebben veelal de vorm van een statistische toets. In een statistische toets wordt eerst een zo concreet mogelijke *(nul)hypothese* geformuleerd over de populatie en vervolgens wordt de juistheid van deze hypothese getoetst. Het resultaat van de toets is de uitspraak, dat de hypothese op grond van de resultaten van de steekproef kan worden verworpen, of juist niet kan worden verworpen. Kenmerkend voor een statistische toets is dat wordt aangegeven hoe groot de kans is dat de hypothese ten onrechte wordt verworpen.

Bij een toetsingsprobleem worden de volgende stappen genomen:

1. Formuleer een nulhypothese (H_0) en een alternatieve hypothese (H_1).
2. Kies een waarde voor α, dit is het significantieniveau oftewel fout van de eerste soort. Bij geldigheid van de nulhypothese mag er voor de steekproefgrootheid een kans zijn op een uitkomst in het kritieke gebied *Z*, die hoogstens α bedraagt.
3. Bepaal met welke toetsingsgrootheid er gewerkt moet worden.
4. Bereken het kritieke gebied *Z*.
5. Bepaal de uitkomst van de toetsinggrootheid en bekijk met behulp van het kritieke gebied *Z* of de nulhypothese al dan niet verworpen moet worden. Als beslisregel geldt:
    - Als de gevonden waarde in *Z* ligt dan wordt H_0 verworpen,
    - Als de gevonden waarde niet in *Z* ligt dan wordt H_0 aangenomen.
6. Geef een formulering van de conclusie.

Hieronder volgen de definities *van hypothese*, *toetsingsgrootheid*, *kritiek gebied Z*, *z-waarde* en het *significantie niveau*.
Deze informatie komt uit *Statistiek om mee te werken* [Buijs (1997)].

8.2.1.1 Hypothese

De hypothese die we formuleren en toetsen in een statistische toets wordt de nulhypothese H_0 genoemd. De nulhypothese moet bij het begin van het onderzoek worden geformuleerd. Deze hypothese heeft de vorm van een *exacte*, *kwantitatieve* uitspraak over een parameter van de populatie, waaruit de steekproef is getrokken. De alternatieve hypothese (H_1) heeft de vorm van een ontkenning van de nulhypothese. In formule is dat:

H_0: populatieparameter = waarde
H_1 :populatieparameter ≠ waarde

Op deze manier geformuleerd kan de waarde van de populatieparameter zowel groter als kleiner zijn dan de waarde onder de nulhypothese. Omdat de populatiewaarde zowel groter als kleiner dan de *waarde* kan zijn, wordt er *tweezijdig getoetst*. Een meer specifieke alternatieve hypothese is bijvoorbeeld:

H_1: populatieparameter > waarde

Nu wordt er *rechtseenzijdig* getoetst (populatiewaarde kan alleen > waarde zijn). De vorm van de alternatieve hypothese bepaalt dus of er *eenzijdig of tweezijdig getoetst* wordt.

8.2.1.2 Toetsinggrootheid

De populatie wordt onderzocht door het nemen van een steekproef. Als een steekproef genomen is, dan moeten de verkregen resultaten geanalyseerd worden. Er wordt dan een grootheid opgesteld waarmee de toets wordt uitgevoerd. Dat is de zogenaamde *toetsingsgrootheid*. De bedoeling is dat de informatie uit de steekproef zo goed mogelijk tot uitdrukking komt in een waarde van de toetsinggrootheid.

Veel gebruikte toetsingsgrootheden zijn het steekproefgemiddelde $\bar{x}$ en de steekproeffractie $p = \frac{k}{n}$.

8.2.1.3 Kritiek gebied en voorspellingsinterval

Door een 95% voorspellingsinterval voor de toetsingsgrootheid op te stellen, kan worden gezegd dat de toetsingsgrootheid met kans 0.95 een waarde in dat interval zal aannemen. Deze collectie uitkomsten worden als normaal 'gekwalificeerd', gegeven het feit dat H_0 juist is. Wat dan overblijft, is een kans 0.05. Deze kans wordt bij een toets aangegeven met α. Dit is dus de kans om een waarde buiten het voorspellingsinterval te vinden, in de situatie dat de nulhypothese correct is. Als in dat gebied een uitkomst aangetroffen wordt, dan wordt H_0 verworpen.
De verzameling uitkomsten die leiden tot verwerpen van de nulhypothese wordt het kritieke gebied *Z* genoemd. De overige uitkomsten, die dus niet leiden tot verwerping van H_0, duidt men aan als het acceptatiegebied.

In termen van de standaard normale verdeling luidt het kritieke gebied:

$Z=\{x|x<\mu-z\sigma$ of $x>\mu+z\sigma\}$ bij *tweezijdig toetsen*

$Z=\{x|\ x>\mu+z\sigma\}$ bij *rechtseenzijdig toetsen*

Waarbij μ de verwachting van de toetsingsgrootheid, σ de standaarddeviatie van de toetsingsgrootheid en *z* de z-waarde is. Het begrip *z-waarde* wordt in de volgende paragraaf besproken. Als de gevonden waarde in *Z* ligt dan wordt H_0 verworpen. Als de gevonden waarde niet in *Z* zit dan wordt H_0 aangenomen.

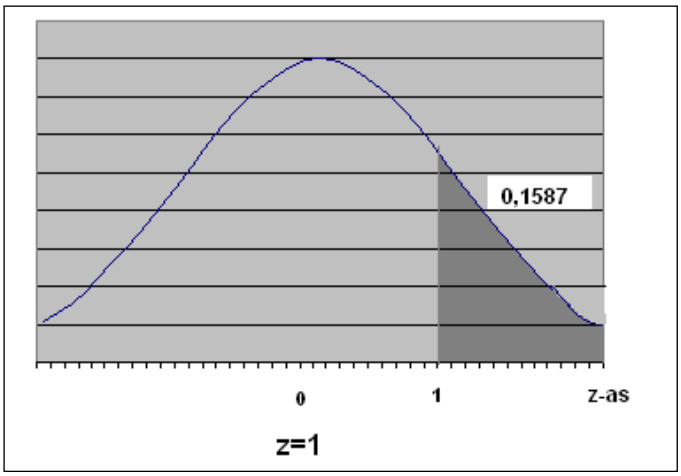
8.2.1.4 Z-waarde

De normale verdeling met $\mu=0$ en $\sigma=1$ noemt men de standaard normale verdeling. Een variabele *z* die een standaard normale verdeling volgt, schrijft men als:

$z\sim N(\mu=0, \sigma=1)$

In de literatuur zijn tabellen bekend van de standaard normale verdeling waar de rechteroverschrijdingskans uit gelezen kan worden. In grafiek 1 is te zien hoe voor $z = 1$ (de *grenswaarde*) de kans 0,1587 toebehoort (deze kans is ook uit de tabel van de standaard normale verdeling bij *z-waarde* 1 af te lezen). Dit is de kans om een waarneming te doen die groter is dan 1. De oppervlakte van het staartgedeelte onder de curve ter rechterzijde van de grenswaarde $z = 1$ bedraagt dus 0,1587.

Figuur 29 Kans bij z = 1



Voor een normale verdeling met willekeurig normaal verdeelde μ en σ moet een transformatie worden uitgevoerd. Deze transformatie is een voorbeeld van *standaardisatie* of *normalisatie* van de data. Door standaardisatie zijn de waarden van verschillende variabelen beter te vergelijken. Ook kan door deze transformatie gebruik worden gemaakt van de tabel van de standaard normale verdeling waar vervolgens de kansen opgezocht kunnen worden.

Voor een grenswaarde g in het oorspronkelijke probleem leidt dit tot een grenswaarde in de standaard normale verdeling van:

$$z = \frac{g - \mu}{\sigma}$$

Dit wordt meestal aangeduid als de *z-waarde* die bij een bepaalde grens g hoort. Voor deze z-waarde kunnen vervolgens kansen opgezocht worden in de tabel van de standaard normale verdeling.

8.2.1.5 Significantieniveau

Om tot een verdeling te komen in een kritiek gebied en een acceptatie gebied is een criterium nodig. Dit wordt weergegeven door een kans *α*. Dit wordt *het betrouwbaarheidsniveau* of *fout van de eerste soort* genoemd. Deze wordt vooraf gekozen. Bij geldigheid van de nulhypothese mag er voor de steekproefgrootheid een kans zijn op een uitkomst in het kritieke gebeid *Z*, die hoogstens α bedraagt.

Als eenmaal de waarde voor *α* vaststaat, is het betrekkelijk eenvoudig om voor een kansverdeling aan te geven wat het kritieke gebied *Z* is. Als de steekproef een uitkomst toont die tot *Z* behoort, wordt de nulhypothese verworpen. In formule wordt α weergegeven door:

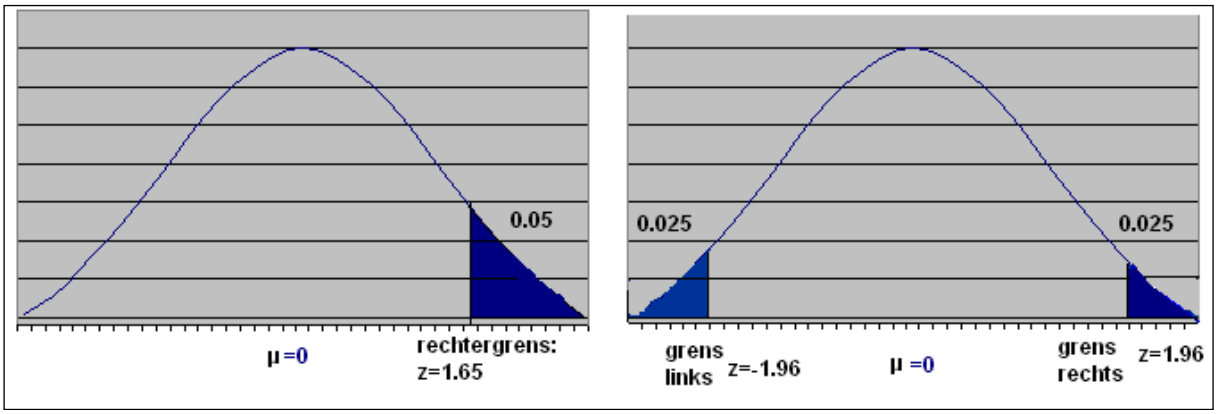
$$P(x \in Z \mid H_0) < \alpha$$

Dus *α* is de kans om ten onrechte de nulhypothese te verwerpen, want de onderzochte variabele heeft bij geldigheid van H_0 een kans van hoogstens *α* op een uitkomst in *Z*. Een significantieniveau van *α = 0.05* is gebruikelijk en wordt in dit onderzoek ook gebruikt. Het houdt in, dat gemiddeld genomen één op de twintig keer de nulhypothese ten onrechte wordt verworpen.

8.2.1.6 Keuze z in formule van het kritieke gebied

Bij rechtseenzijdig toetsen levert de waarde *α = 0.05* een *z*-waarde van *1.65*. Dit volgt uit de tabel van de normale verdeling. Bij tweezijdig toetsen moet aan beide kanten van de verdeling een gebied van met oppervlakte 0.025 gezocht worden. De waarde van 0.025 levert *z=1.96* in de tabel van de normale verdeling. In de formule voor het opstellen van het kritieke gebied moet dus afhankelijk van of er eenzijdig of tweezijdig getoetst wordt deze waardes voor *z* genomen worden.

Figuur 30 Links: z-waarde bij eenzijdig toetsen Rechts: z-waarde bij tweezijdig toetsen



8.2.1.7 P-waarde

Bij elke toets kan ook een overschrijdingskans berekend worden. De p-waarde of overschrijdingskans is de kans dat in de verdeling gegeven door de nul-hypothese de waarde van de toetsinggrootheid wordt overschreden. In dit onderzoek wordt de waarde van 5% aangehouden als grens (α); is de p-waarde kleiner dan 5%, dan spreekt men van een **significante** uitkomst en wordt de nul hypothese verworpen.

8.2.2 Correlatie

Een correlatiecoëfficiënt geeft de richting en sterkte aan van de samenhang tussen twee variabelen. Correlatiecoëfficiënten kunnen waarden aannemen tussen -1 (een perfecte negatieve correlatie) en +1 (een perfecte positieve correlatie). Een waarde rond nul geeft aan dat de variabelen geen lineaire samenhang vertonen.
Een positieve correlatie betekent, dat een hoge waarde van de ene variabele gepaard gaat met een hoge waarde van de andere variabele. Bij een negatieve correlatie is het omgekeerde het geval: hoge uitkomsten op de ene variabele gaan gepaard met lage waarden op de andere.
Pearson's correlatiecoëfficiënt is de meest gebruikte maat voor samenhang tussen continue variabelen. De toepassing van Pearson's correlatiecoëfficiënt veronderstelt, dat de samenhang lineair is.

Een Pearson's correlatiecoëfficiënt gelijk aan nul wil niet zeggen, dat de variabelen onafhankelijk zijn. Er kan bijvoorbeeld een niet-lineaire samenhang zijn.

Pearson's correlatiecoëfficiënt wordt berekend met de formule:

$$r = \frac{\sum\{(y_1 - \bar{y}_1)(y_2 - \bar{y}_2)\}}{\sqrt{\sum\{(y_1 - \bar{y}_1)^2 (y_2 - \bar{y}_2)^2\}}}$$

De uitkomsten van de variabelen zijn y_1 en y_2. De gemiddelden van de uitkomsten zijn $\bar{y}_1$ en $\bar{y}_2$. De waarden die *r* kan aannemen liggen tussen -1 en +1.

8.2.3 Lineaire Regressie

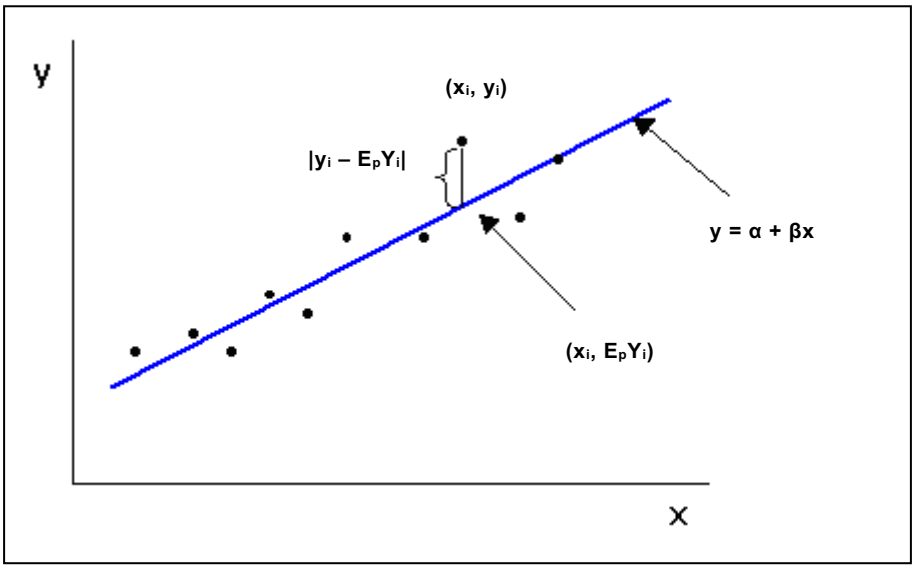
In dit onderzoek vormt het lineaire regressiemodel de grondslag. Daarom zal in het kort de theorie van lineaire regressie uitgelegd worden.

Theorie achter lineaire regressie
Wanneer men verbanden tussen twee grootheden *x* en *y* wil onderzoeken, waarbij *x* als verklarende variabele kan worden beschouwd en *y* als afhankelijke variabele, kan men dit verband goed vastleggen in een lineair regressiemodel als het verband rechtlijnig van aard is. Dit kan men makkelijk nagaan door een spreidingsdiagram te maken. Ook wanneer er bijvoorbeeld een kwadratisch verband is, kan men lineaire regressie toepassen, waarbij een kwadratische term als extra factor wordt toegevoegd aan het model.
Ook kan men eventuele niet-lineaire verbanden terugbrengen tot een lineair verband door het uitvoeren van een of meerdere transformaties (logaritme, wortel, etc.) op de variabelen.
Het model ziet er bij enkelvoudige lineaire regressie als volgt uit:

$$y = \alpha + \beta x + \varepsilon$$

met α de constante, β de hellingshoek en ε de term voor de storingen. De fout ε_i is de verticale afstand tussen de observatie y_i en het punt op de regressiecurve behorende bij die observatie. Deze storingen zijn onafhankelijk van elkaar en normaal verdeeld met verwachting 0 en variantie σ^2.

Figuur 31 Voorbeeld Lineaire Regressie in tweedimensionale ruimte



In geval van twee variabelen, de afhankelijke variabele en één onafhankelijke variabele, wordt de regressiefunctie beschreven door een rechte lijn.

De parameters worden geschat met de zogenaamde gewone *kleinste kwadraten* methode (Eng. Ordinary Least Squares (OLS)). Hierbij wordt

$$Q(c,d) = \sum_{i=1}^{n} [y_i - (c + dx_i)]^2$$

geminimaliseerd met *n* het aantal waarnemingen en *c* en *d* variabel. De optimale waarden voor *c* en *d* worden berekend door de eerste orde condities op te lossen:

$$\partial Q / \partial c = 0$$
$$\partial Q / \partial d = 0$$

Uit deze twee condities volgen de optimale waarden van *c* en *d* (respectievelijk *a* en *b*):

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Deze waarden (*a* en *b*) zijn de OLS schattingen van respectievelijk α en β en hiermee kan de regressielijn van *y* op *x* opgesteld worden:

$$\hat{y} = a + bx$$

Dit model kan vrij eenvoudig uitgebreid worden tot een *meervoudig lineair regressiemodel* wanneer de grootheid *y* afhangt van meerdere verklarende variabelen.
Dit standaard *meervoudig lineair regressiemodel* kan verder stelselmatig uitgebreid en aangepast worden totdat het meest geschikte model is gebouwd.

8.2.3.1 Stapsgewijze Regressie

Als het meervoudig lineair regressiemodel gebouwd is, kan de methode *stapsgewijze regressie* toegepast worden. Stapsgewijze regressie voegt alle variabelen stapsgewijs toe totdat de meest significante variabelen overblijven. Deze methode wordt hieronder verder beschreven.

Met behulp van stapsgewijze regressie kan bepaald worden welke variabelen wel en niet in de regressievergelijking thuishoren. De volgende stappen worden genomen:

- Begin met 1 variabele en kijkt of de p-waarde kleiner is dan α.
  Zo niet, dan wordt de variabele niet opgenomen.

- Zo wel, dan wordt de variabele opgenomen . Een tweede variabele wordt dan toegevoegd, waarvan de p-waarde met α vergeleken wordt.
  Is deze kleiner dan α, dan wordt ook deze opgenomen en wordt een derde variabele toegevoegd.

Zo gaat dit door tot er geen variabelen meer gevonden kunnen worden met een p-waarde kleiner dan α. Bij elke stap wordt steeds de variabele met de kleinste p-waarde toegevoegd. De variabele die al in het model zitten worden verwijderd als ze niet significant worden na het toevoegen van andere variabelen.

8.2.3.2 Determinatiecoëfficiënt

De determinatiecoëfficiënt is een maat voor het deel van de variatie in de waarnemingen, dat door het lineaire regressiemodel wordt verklaard. Als eerste wordt de spreiding gedefinieerd:

Totale spreiding	=	Verklaarde Spreiding	+	Onverklaarde spreiding
SST	=	SSR	+	SSE
$\sum(Y_i - \overline{Y})^2$	=	$\sum(\hat{Y}_i - \overline{Y})^2$	+	$\sum(\hat{Y}_i - Y_i)^2$

De determinatiecoëfficiënt is gedefinieerd als:

$$R^2 = \frac{SST - SSE}{SST} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

waarbij:

- SST := de totale kwadraatsom.
- SSE := de gekwadrateerde verschillen tussen de werkelijk gemeten waarden Y_i en de aan de hand van het model voorspelde waarden $\hat{Y}_i$.
- n := het aantal observaties.
- $\overline{Y}_i$:= het gemiddelde van de geobserveerde Y_i waarden.

SST geeft de fout weer wanneer het gemiddelde van de observaties als voorspelling van de waarnemingen genomen wordt. SSE geeft de fout weer indien voor de voorspelling van de waarnemingen het regressiemodel gebruikt wordt.
R^2 geeft dus het percentage van de reductie van de fouten weer indien het regressiemodel gebruikt wordt. De waarde van R^2 ligt steeds tussen 0 en 1. Hoe dichter R^2 bij 1 ligt, hoe beter de werkelijke waarden van de afhankelijke variabele benaderd worden door het model. Als R^2 gelijk is aan 0, dan wil dit zeggen dat het model geen enkele toegevoegde waarde heeft. Het model past dan even goed als het model waarin geen enkele verklarende variabele voorkomt.

8.2.3.3 F-toets en t-toets

De *F-toets* geeft informatie over de standaard schattingsfout. Er wordt gekeken of er een verband is tussen de regressielijn en de puntenwolk en hoe groot de schattingsfout met andere woorden er wordt beoordeeld of de afstand van de punten in de puntenwolk niet te sterk afwijken van de regressielijn. Als de overschrijdingskans *p* groter is dan 0,05 is er geen sprake van een zinvolle regressielijn.
De *t-toets* onderzoekt de betrouwbaarheid van β. Als de overschrijdingskans *p* groter is dan 0,05 is er geen verband tussen *x* als verklarende variabele en *y* als afhankelijke variabele .

8.2.4 Modellen voor tijdreeksanalyse

8.2.4.1 Wat is een tijdreeks?

Kenmerk van tijdreeksanalyse is dat op regelmatige tijdstippen t=1,2,3,…, T een waarde Y_t van een

variabele wordt waargenomen waardoor een reeks Y_1, Y_2, Y_3,……, Y_T ontstaat. Door zo'n reeks in een grafiek te plaatsen, ontstaat de mogelijkheid om een beeld te vormen van het historisch verloop van zo'n variabele. Een dergelijke reeks wordt een *tijdreeks* of een *historische reeks* genoemd. In beginsel zijn de tijdsintervallen tussen twee opeenvolgende waarnemingen even lang.
Voor het analyseren van tijdreeksen zijn methoden ontwikkeld. De klassieke decompositiemethode is een methode waarmee een tijdreeks als het ware wordt ontbonden in componenten. Hierbij gaat het vooral om de trend en de seizoencomponent.
Met behulp van een grafiek met de waargenomen uitkomsten Y_t, van een tijdreeks, kan snel een indruk worden verkregen van het verloop van de reeks.

Doel bij klassieke tijdreeksanalyse is het opsporen van een aantal componenten waaruit de uitkomsten Y_t van een tijdreeks opgebouwd kunnen worden.

Het lange-termijngedrag van een tijdreeks wordt aangeduid met de term trend of trendmatige ontwikkeling.
Om de trend te kunnen bepalen, is het de bedoeling dat in de waargenomen reeks de periodieke afwijkingen en de eventuele toevallige factoren zoveel mogelijk geëlimineerd worden, zodat een duidelijk beeld van de trend ontstaat.

Vaak wordt de trendlijn niet alleen gebruikt om achteraf een beschrijving te geven van de lange-termijnontwikkeling van een bepaalde variabele, maar wordt de lijn ook benut om voorspellingen te doen voor de nabije toekomst.

8.2.4.2 Autocorrelatie

Een heel belangrijk onderdeel bij het analyseren van tijdreeksen betreft het bestuderen van de onderlinge afhankelijkheid van opeenvolgende waarnemingen Y_t in zo'n reeks. Hierbij kan het zijn dat een waarneming Y_t een verband toont met de waarneming van één periode eerder, maar het kan ook zijn dat waarnemingen die verder weg liggen een invloed op Y_t hebben. Een eenvoudige manier om dit te onderzoeken, is het toepassen van autocorrelatie.

Hierbij wordt in beginsel van dezelfde formule gebruik gemaakt als bij gewone correlatie, waar de grootheid r de mate van lineaire samenhang van de variabelen X en Y aangaf. Nu wordt er gewerkt met de variabelen Y_t en Y_{t-1}. De 'getallenparen' zijn nu (Y_2, Y_1), (Y_3, Y_2) enz.

Wanneer de autocorrelatie wordt berekend van een variabele op twee opeenvolgende tijdstippen, spreekt men van een vertragingsfactor één (*lag* = 1). Uiteraard kan ook autocorrelatie voorkomen tussen waarden die meer dan één tijdseenheid uit elkaar liggen. Zo is de autocorrelatie voor ieder gewenst tijdsinterval te berekenen (lag = 2, 3, . . . , k).

De gebruikelijke formule voor de autocorrelatiecoëfficiënt luidt:

$$r_1 = \frac{\sum_{t=2}^{n}(Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} = \frac{\sum_{t=2}^{n} y_t y_{t-1}}{\sum_{t=1}^{n} y_t^2}$$

De coëfficiënt r_1 wordt de autocorrelatiecoëfficiënt met time-lag 1 genoemd. Op een soortgelijke manier kan de autocorrelatiecoëfficiënt met een time-lag van k perioden gedefinieerd worden. In formule:

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} = \frac{\sum_{t=k+1}^{n} y_t y_{t-k}}{\sum_{t=1}^{n} y_t^2}$$

De autocorrelatiecoëfficiënten die berekend zijn op basis van beschikbare (historische) gegevens vormen een belangrijk hulpmiddel bij het opsporen van bepaalde patronen in een tijdreeks.

Vaak worden berekende autocorrelatiecoëfficiënten aangegeven in een speciale grafische voorstelling: het autocorrelogram. Het belang van zo'n autocorrelogram is dat hiermee op het gebied van tijdreeksanalyse vrij snel de onderliggende structuur van deze reeks onderkend kan worden.

Een reeks met een trend of met een seizoenspatroon zal een autocorrelogram voortbrengen dat een zeer herkenbare vorm heeft. Een belangrijk thema bij tijdreeksanalyse is echter of bij een stationaire reeks, dat is een reeks waarin trend- en seizoeninvloeden verwijderd zijn, nog steeds bepaalde afhankelijkheden bij opeenvolgende waarnemingen geconstateerd kunnen worden door middel van autocorrelatiecoëfficiënten.

8.2.4.3 Toets voor Autocorrelatie

Evenals bij gewone correlatiecoëfficiënten wordt bij autocorrelatie een onderscheid gemaakt tussen populatie- en steekproefwaarden. Het idee is hierbij dat een verzamelde reeks gegevens ten gevolge van steekproeftoeval een waarde voor een autocorrelatiecoëfficiënt r_k kan tonen die iets afwijkt van een populatiecoëfficiënt p_k. Zo zal bij een populatiecoëfficiënt $p_k=0$ niet elke verzamelde reeks Y_t-waarden een r_k tonen die exact gelijk aan 0 is. Om te onderzoeken of een r_k significant van 0 afwijkt, is een toets ontworpen die als volgt is opgezet.

De hypothesen zijn:

$$H_0 : \rho_k = 0$$
$$H_1 : \rho_k \neq 0$$

De grootheid r_k is de toetsinggrootheid waarvoor aangetoond is dat deze bij benadering een normale verdeling volgt met als standaarddeviatie $1/\sqrt{n}$. Meestal wordt deze toets uitgevoerd met $\alpha=0.05$, waarbij gewerkt wordt met $z=2$ (eigenlijk $z=1.96$). Als een berekende autocorrelatiecoëfficiënt vervolgens buiten het gebied $(-2/\sqrt{n}\ ,\ 2/\sqrt{n})$ valt, dan wordt H_0 verworpen.

8.2.4.4 Toets voor witte ruis

Er zijn verschillende toetsen voor witte ruis bekend. Eén daarvan is *de Box-Ljung* toets. Deze bepaalt de (gewogen) som van de eerste autocorrelatiecoëfficiënten. Deze som volgt bij benadering een Chi-kwadraat verdeling.

De *Ljung-Box* test is gebaseerd op de autocorrelatie plotjes. In plaats van het toetsen van witte ruis op elke willekeurige lag, toetst het de overall randomness gebaseerd op een aantal lags. De *Ljung-Box* toets kan als volgt gedefinieerd worden:

H_0 : De data is random (witte ruis).
H_1 : De data is niet random (geen witte ruis).

De toetsinggrootheid is:

$$Q_{LB} = T(T+2)\sum_{j=1}^{M}\frac{r_j^2}{T-j}$$

met T de grootte van de steekproef, r_j de autocorrelatie at lag j, en M is het aantal lags dat gebruikt wordt . Als betrouwbaarheid wordt $\alpha=0.05$ genomen.

De witte ruis hypothese wordt verworpen als $Q_{LB} > X^2_{1-\alpha,M}$ met $X^2_{1-\alpha,M}$ Chi-kwadraat verdeeld bij M vrijheidsgraden.

8.2.4.5 Exponentiële Effening (= Exponential Smoothing)

Bij enkelvoudige exponentiële effening wordt bij een tijdreeks $\{Y_t \mid t = 1,...,T\}$ stap voor stap een reeks trendwaarden $\{\overline{Y}_t \mid t = 1,...,T\}$ berekend. Voor het berekenen van een trendwaarde $\overline{Y}_t$ zijn twee gegevens van belang, namelijk de vorige trend waarde $\overline{Y}_{t-1}$ en het nieuwe gegeven $\overline{Y}_t$.
De nieuwe trendwaarde wordt dan bepaald als een gewogen gemiddelde van Y_t en $\overline{Y}_{t-1}$ volgens de formule:

$$\overline{Y} = \alpha Y_t + (1-\alpha)\overline{Y}_{t-1} \text{ met } 0 < \alpha < 1$$

Hierbij is α een willekeurig te kiezen getal tussen 0 en 1, dat bepalend is voor de gevoeligheid die de reeks trendwaarden vertoont voor nieuwe waarnemingen. Deze factor α wordt wel de *effeningsconstante* genoemd. Om een begin te kunnen maken met de berekeningen wordt in de eerste periode gesteld dat de geëffende waarde $\overline{Y}_1$ gelijk is aan de eerste waarneming Y_t.
Door de basisformule enkele malen achter elkaar toe te passen, kunnen we analyseren hoe de geëffende waarde $\overline{Y}_t$ afhangt van eerdere waarnemingen.

Voor de trendwaarde $\overline{Y}_{t-1}$ geldt:
$$\overline{Y}_{t-1} = \alpha Y_{t-1} + (1-\alpha)\overline{Y}_{t-2}$$

Als we dit invullen in de vergelijking voor Y_t dan vinden we:
$$\overline{Y}_t = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + (1-\alpha)^2 \overline{Y}_{t-2}$$

Nu kan $\overline{Y}_{t-2}$ ingevuld worden en vervolgens $\overline{Y}_{t-3}$ enzovoort. Dit leidt uiteindelijk tot de volgende vergelijking:
$$\overline{Y}_t = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \alpha(1-\alpha)^3 Y_{t-3} + ...$$

Aan deze laatste uitdrukking is te zien dat een waarneming van oudere datum een kleinere coëfficiënt heeft dan een recent gegeven. Ook nemen de coëfficiënten exponentieel af. Vandaar de naam *exponential smoothing* voor deze methode.

De keuze van α speelt een niet onbelangrijke rol bij het werken met de methode van exponentiële effening. Indien α klein wordt gekozen (bijvoorbeeld α = 0,1), dan krijgt de nieuwe waarneming een betrekkelijk klein gewicht, terwijl de oude geëffende waarde een groot gewicht krijgt (namelijk 0,9). Het gevolg hiervan is dat de reeks $\overline{Y}_t$ een zeer stabiel verloop zal vertonen. Naarmate α groter wordt gekozen, zal de reeks $\overline{Y}_t$ meer gelijkenis vertonen met de reeks waargenomen uitkomsten Y_t (voor α = 1 vallen ze samen).

Exponentiële effening kan ook gebruikt worden bij het doen van voorspellingen (op korte termijn). Als we de geëffende waarde op tijdstip *t* aangeven met $\overline{Y}_t$, dan wordt de voorspelde waarde F voor tijdstip $t+1$ gegeven door $F_{t+1} = \overline{Y}_t$.

Als voorspelling voor het eerstkomende tijdvak (en eventueel enkele daaropvolgende perioden) wordt dus de meest recente uitkomst Y_t van de geëffende waarden genomen. Deze vorm van voorspellen mag echter niet klakkeloos worden gebruikt, omdat hij in feite gebaseerd is op de veronderstelling dat de waargenomen Y-waarden voortdurend rond een vaste waarde schommelen. We spreken in zo'n geval van het *horizontale model* $\overline{Y}_t = \mu + \varepsilon_t$.
Men zou kunnen stellen dat de uitkomsten $\overline{Y}_t$ schattingen zijn van μ. Problemen ontstaan er als we bij de tijdreeks een gedrag waarnemen dat afwijkt van het horizontale model.

8.2.5 Logistische regressie

Wanneer de invloed wordt nagegaan van één of meerdere verklarende variabelen *x* op een afhankelijke variabele *y*, kom je al snel bij lineaire regressieanalyse uit. Zo'n regressie model gaat er van uit dat de afhankelijke variabele continue van aard is, dus gemeten is op interval- of rationiveau. Het komt echter regelmatig voor dat de afhankelijke variabele van een ander meetniveau is, bijvoorbeeld dat er sprake is van een nominale variabele met slechts enkele categorieën. Lineaire regressieanalyse is dan niet mogelijk. Om toch de invloed van allerlei verklarende variabelen op een nominale variabele te kunnen nagaan, zijn er verschillende analysetechnieken ontwikkeld. De techniek die het meest bij lineaire regressieanalyse aansluit, is logistische regressieanalyse. Logistische regressie analyse is geschikt voor een afhankelijke variabele die dichotoom van aard is. Er zijn maar twee categorieën.
Men is dus geïnteresseerd in de voorspelling van de kans dat een verklarende variabele in de categorie 'wel' of 'niet' valt. Een gewone lineaire regressieanalyse zal over het algemeen wel de juiste richting van de β-coëfficiënten opleveren. Maar de schatting is niet helemaal correct, omdat enkele belangrijke regressie assumpties geschonden worden, zoals de normaliteitsassumptie en de assumptie van homoscedasticiteit. Het grootste probleem is evenwel dat de door lineaire regressie voorspelde kansen groter kunnen zijn dan 1 en kleiner dan 0. Dergelijke kansen zijn niet te interpreteren. Het is daarom aan te raden om logistische regressie analyse te gebruiken wanneer je te maken hebt met een dichotome afhankelijke variabele. Zoals gezegd gaat het logistische model uit van kansen, of beter gezegd van kansverhoudingen: odds. De odds is de kans om 'wel' (P_{wel}) gedeeld door de kans 'niet' (P_{niet}). Een odds heeft een bereik van 0 tot oneindig. Omdat men liever met een variabele rekent die loopt van min oneindig naar plus oneindig, wordt de natuurlijke logaritme[2] van de odds genomen. Deze wordt de log odds of logit genoemd. Kans, odds en logit zijn dus eigenlijk drie manieren om hetzelfde te zeggen. Met de volgende verklarende X_1, X_2 enz. Dan ziet het logistische model er in formulevorm als volgt uit:

$$\ln \frac{P_{wel}}{P_{niet}} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

Dit model lijkt sterk op een gewoon regressiemodel: α is het intercept, β_1 is de parameter die het effect van X_1 aangeeft, β_2 de parameter die het effect van X_2 aangeeft enz. Het logistische model kan ook omgezet worden in een kansmodel. De kans op 'wel' is:

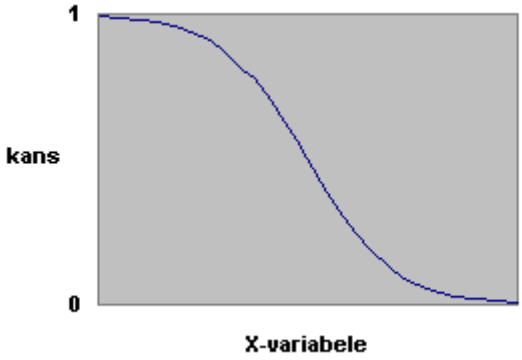
$$P_{wel} = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\ldots}}{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\ldots}+1}$$

En de kans op 'niet' is:

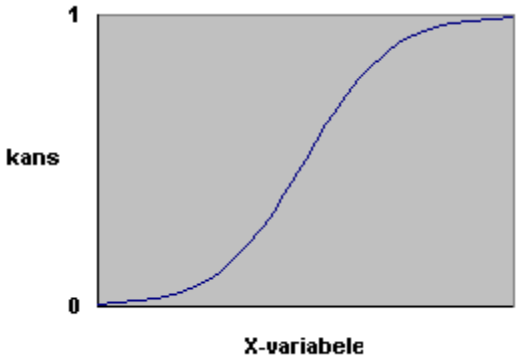
$$P_{niet} = \frac{1}{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\ldots}+1}$$

Aan deze formules is af te lezen dat de kansen P_{wel} en P_{niet} bij elkaar opgeteld gelijk zijn aan 1. Verder is te zien dat de kansen P_{wel} en P_{niet} afhankelijk zijn van de variabelen X_1, X_2 enz., maar dat deze afhankelijkheid niet lineair is. Een logistische regressielijn ziet er dus niet als een rechte lijn uit, maar als een S-vormige curve. Hieronder zijn twee logistische curve getekend: figuur 52 voor een positief effect van de verklarende variabele X en figuur 51 voor een negatief effect van X. Deze curves beschrijven het niet-lineaire verband tussen de kans en de verklarende variabele X. Het effect van de variabele X op de kans dat Y voorkomt is het grootst bij de middenwaarden van X.

Figuur 32 Logistische curve negatief effect van X



Figuur 33 Logistische curve positief effect van X



8.2.6 Datamining

Wat is datamining?
Het begrip data kan omschreven worden als hetgeen verzameld en opgeslagen wordt. Kennis daarentegen is datgene dat mensen helpt om weldoordachte beslissingen te nemen. Het extraheren van kennis uit data wordt datamining genoemd. Het ultieme doel van datamining is om waardevolle kennis te vergaren: het extraheren van impliciete en vooraf ongekende, maar toch potentieel nuttige informatie. Hiertoe wordt gegrepen naar statistische en mathematische methoden, maar ook naar complexe technologieën die betrekking hebben op patroonherkenning.
De data worden voorbereid met behulp van statistiek: selectie van de data (sampling). Statistiek is ook achteraf belangrijk voor het evalueren van de verkregen kennis. Statistiek biedt het voordeel dat kan omgegaan worden met ontbrekende data door een beroep te doen op schattingstechnieken.
Na de voorbereiding van de data zorgen relationele databanken voor reductie van de data door sterke associatie-regels te formuleren.

Vervolgens kan men een datamining techniek toepassen om nuttige informatie te zoeken in de data: classificatie of clustering. In beide gevallen zouden de data ingedeeld moeten worden in drie groepen.
- Trainingsset: uitgebreid aantal data dat gebruikt wordt om het algoritme te trainen in classificatie of clustering. Hoe beter de trainingsgegevens, hoe beter de uiteindelijke prestaties van het algoritme. De tijd nodig voor de training neemt echter lineair toe met het aantal trainingsdata.
  Een belangrijke karakteristiek voor de evaluatie van een techniek is zijn prestatie bij een kleine trainingsset.
- Validatieset: data waarmee men nagaat hoe goed het algoritme is in classificeren of clusteren van nog niet geziene data. Voor dit doeleinde wordt vaak een deel van de data uit de trainingsset achter de hand gehouden.
- Testset: totaal nieuwe data die het algoritme dient te classificeren of clusteren. Cruciaal is dat de data representatief zijn voor het probleem dat men wil oplossen. Pas wanneer het algoritme in staat is om goede resultaten te geven voor de testset is het klaar voor gebruik.

Deze verdeling van de data in groepen wordt soms met de voeten getreden wanneer er te weinig data beschikbaar zijn om elke set te voorzien van voldoende gegevens.

Wat is classificatie?
Classificatie is het proces dat een dataset verdeelt in niet-overlappende groepen, zodanig dat de leden van elke groep zoveel mogelijk op elkaar lijken en tegelijk zo verschillend mogelijk zijn van de leden van andere groepen.
Classificatie houdt in dat het datamining algoritme getraind wordt met gevallen waarbij het antwoord gegeven is. Wanneer het datamining algoritme vervolgens een nog niet eerder gezien geval dient te classificeren, zal het vaak een goede gooi kunnen doen naar het antwoord. Dit gedrag noemt men generaliseren.

Enkele belangrijke classificatietechnieken

Decision trees

Uit de naam decision tree kan reeds afgeleid worden dat het resultaat van het algoritme weergegeven wordt in de vorm van een boomstructuur. Deze structuur kan omgezet worden naar een tekst van de vorm 'als… dan … anders…', maar de boomstructuur is voor de meeste gebruikers makkelijker te interpreteren.

Een decision tree is als het ware een kaart van redeneringprocessen.

Decision trees zijn erg goed in het oplossen van classificatieproblemen. Ze classificeren gevallen door ze te sorteren, beginnend boven aan de boom (ter hoogte van de *root*) en naar onder toe werkend. Bij elke *node* van de boom wordt een attribuut getest. Afhankelijk van de waarde van het attribuut wordt een bepaalde tak (*branch*) naar beneden toe gevolgd totdat er een nieuwe node bereikt wordt. Daar wordt de bovenstaande procedure herhaald, tenzij het einde van de boom (*leaf*) bereikt is. In dit laatste geval kan het resultaat van de classificatie afgelezen worden.

Het is erg belangrijk om de opeenvolging van de attributen in de nodes goed te kiezen. De beste keuze wordt bepaald met behulp van algoritmes. De bovenste node moet het meest discriminerend zijn. Dit wil zeggen dat deze node de meeste verschillen aanduidt in de trainingsdata. Hoe meer discriminerend een attribuut is, hoe meer informatie er uit kan gehaald worden.

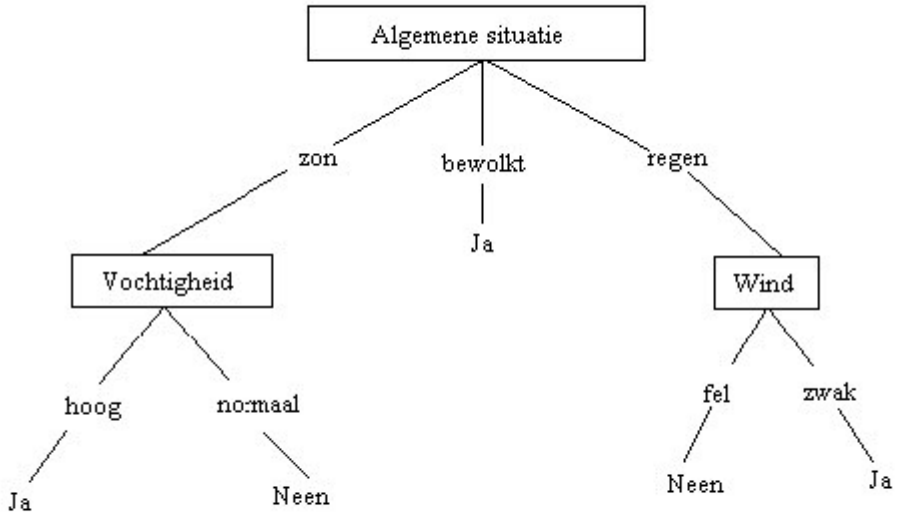
Aan decision trees zijn enkele nadelen verbonden.

1. Ze kunnen enkel omgaan met discrete waarden.
2. Aangezien de trainingsdata slechts een deel zijn van alle mogelijke gevallen, is het mogelijk dat er takken toegevoegd worden die de prestaties verbeteren voor de trainingsdata, maar tegelijkertijd de prestaties doen afnemen in geval van nieuwe data. Dit moet te allen tijde vermeden worden. Om hieraan tegemoet te komen wordt meestal de voorkeur gegeven aan relatief korte en eenvoudige decision trees.
3. De meeste decision tree systemen zijn niet incrementeel. Dit houdt in dat ze niet in staat zijn om continu de verworven kennis te onderhouden en *up to date* te houden, wat echter essentieel is om efficiënt te kunnen leren. Het belang hiervan ligt vooral in situaties waar de kennis aangepast dient te worden aan de dynamica van de echte wereld.
   De theoretische benadering wordt nu verduidelijkt aan de hand van een voorbeeld. Op basis van het weer wordt beslist of iemand op een bepaalde dag tennis gaat spelen of niet. De attributen en hun mogelijke waarden zijn de volgende:

   - Algemene situatie: zon – bewolkt – regen
   - Vochtigheid: hoog – normaal
   - Wind: fel – zwak
   - Temperatuur: koel – matig – warm

Een mogelijke decision tree wordt weergegeven in figuur 5. Omdat het attribuut temperatuur weinig discriminerend is en zorgt voor weinig informatiewinst, wordt het niet in de boom opgenomen.

Figuur 34 Een decision tree voor tennis spelen



8.2.7 SPSS

Het programma SPSS (oorspronkelijk Statistical Package for the Social Sciences) staat tegenwoordig voor Superior Performing Software Systems en is een professioneel software pakket voor statistische analyse dat een universeel toepassingsgebied kent.

SPSS kan gegevens lezen uit vrijwel ieder bestandsformaat en deze gebruiken voor het genereren van tabellen en grafieken van distributies of trends voor beschrijvende statistiek en voor statistische analyse.

Met SPSS kan data worden:

gemaakt	-	gegevensinvoer
bewerkt	-	bijvoorbeeld het samenvoegen van mensen in een aantal leeftijdscategorieën
geanalyseerd	-	bijvoorbeeld het berekenen van de samenhang tussen twee variabelen of het bepalen van de statistische significantie van een verschil tussen twee groepen

Het programma is opgebouwd uit verschillende modules. De 'Base Module' is onmisbaar om het programma te laten draaien. De basis kan worden uitgebreid met onderdelen die meer geavanceerde analyses mogelijk maken, zoals speciale vormen van regressie-analyse (SPSS Regression Models), tabellen (SPSS Tables) en SPSS Trends voor trendanalyses.

8.2.8 QSM SLIM tooling

De QSM product familie bestaat uit SLIM Estimate, SLIM Control, SLIM Metrics en SLIM Datamanager. De genoemde modules vormen een geïntegreerd geheel, maar zijn ook afzonderlijk van elkaar te gebruiken. Quantitatieve Software Management (QSM) biedt een inzichtelijk open model, waarbij met een beperkt aantal gegevens krachtige analyses uitgevoerd kunnen worden.

SLIM Estimate biedt de mogelijkheid om ten behoeve van nieuwe projecten, op basis van de omvang en productiviteitsgegevens, een optimaal projectplan samen te stellen waarbij onzekerheden en randvoorwaarden meegenomen kunnen worden. Naast de omvang van het te ontwikkelen software product blijkt dat vooral de doorlooptijd een belangrijke invloed heeft op de kosten en de kwaliteit van het eindproduct. Het SLIM model wordt ondersteund met een historische database van ca. 7200 projecten. Van alle typen software (administratief, scientific, control, telecom, embedded) zijn ervaringscijfers voorhanden.

SLIM Control biedt de mogelijkheid om de voortgang van een bestaand plan te volgen. Er zijn slechts een beperkt aantal gegevens nodig (gerealiseerde omvang, inspanning, mijlpalen en gevonden fouten) om vast te kunnen stellen of een project volgens plan verloopt. Een projectplan kan worden geïmporteerd vanuit SLIM Estimate, maar kan ook los van SLIM Estimate worden ingebracht. Zo kunnen ook lopende projectplannen ingebracht worden en kan de status van projecten vastgesteld en geanalyseerd worden. De actuele gegevens die nodig zijn om de status van het project vast te stellen, kunnen tevens gebruikt worden om een inschatting te maken van het verdere verloop (forecasting) en voor de opbouw van een historische database.

SLIM Metrics biedt de mogelijkheid om eigen historische projecten te vergelijken met projecten in de QSM industriedatabase (benchmarking). SLIM Metrics maakt gebruik van SLIM datamanager die ook gebruikt wordt in SLIM Estimate en Control om historische projectgegevens op te slaan en te gebruiken. Met SLIM Metrics kunnen de eigen historische gegevens uitvoerig geanalyseerd worden. SLIM Metrics bevat vele mogelijkheden om de gegevens statistisch te onderzoeken en grafieken en rapporten te maken. Slim Metrics kan gebruikt worden als het dashboard voor softwareontwikkeling en –onderhoud en voor het zelf uitvoeren van een benchmark.

SLIM Datamanager is een product dat standaard wordt meegeleverd bij de hiervoor genoemde SLIM producten. SLIM Datamanager biedt de mogelijkheid om historische projectgegevens in te voeren en om daarmee een eigen historische database op te bouwen. SLIM Datamanager heeft een groot aantal vooraf gedefinieerde variabelen waarvoor ook overeenkomstige gegevens in de QSM database aanwezig zijn. Tot slot kunnen variabelen gedefinieerd worden waarmee in SLIM Metrics analyses kunnen worden uitgevoerd.

SLIM Masterplan is een product dat standaard wordt meegeleverd bij de hiervoor genoemde SLIM producten. SLIM Masterplan biedt de mogelijkheid portfolio analyses uit te voeren op lopende en nog te starten projecten. Een resource manager kan hiermee bijvoorbeeld zien wanneer een tekort of overschot aan personeel ontstaat of de financieel manager kan hiermee zien hoe de kosten van software ontwikkeling verdeeld zijn over de tijd.

8.2.9 Weka (Waikato Environment for Knowledge Analysis)

Weka is vrije data mining-software geschreven in Java. Het is ontwikkeld aan de Universiteit van Waikato te Nieuw-Zeeland. Weka is een werkomgeving voor het uitvoeren van de benodigde stappen bij data mining, waaronder het voorbewerken van de data en het opbouwen van een voorspellend model. Het bevat algoritmen en hulpmiddelen voor clusteren, classificatie, regressie-analyse, visualisatie en feature selection (het selecteren van attributen / variabelen).

8.3 Bijlage C: Data onderzoek Kasstroom

[vertrouwelijk]

8.4 Bijlage D: Autocorrelogram onderzoek Kasstroom

[vertrouwelijk]

8.5 Bijlage E: Autocorrelatie + Box-Ljung toets onderzoek Kasstroom

[vertrouwelijk]

8.6 Bijlage F: Partiële autocorrelatie onderzoek Kasstroom

[vertrouwelijk]

8.7 Bijlage G: Exponentiële effening Sums of Squared Errors onderzoek Kasstroom

[vertrouwelijk]

8.8 Bijlage H: Data onderzoek Functie Punt Analyse

[vertrouwelijk]