SuperSTAR

Regressiemodellen in de accountantscontrole

Marit Schoonhoven

SuperSTAR

Regressiemodellen in de accountantscontrole

Marit Schoonhoven

Deloitte

Enterprise Risk Services/ Data Quality and Integrity

Laan van Kronenburg 2

1183 AS Amstelveen

Vrije Universiteit

Faculteit der Exacte Wetenschappen

Divisie Wiskunde en Informatica

Studierichting Bedrijfswiskunde en Informatica

De Boelelaan 1081a

1081 HV Amsterdam

Augustus 2004

Voorwoord

Ter afsluiting van mijn studie Bedrijfswiskunde en Informatica (BWI) heb ik zeven maanden stage gelopen bij Deloitte, bij de groep Data Quality and Integrity (DQI). Tijdens deze stage heb ik onderzoek gedaan naar het gebruik van regressiemodellen bij accountantscontroles.

Het programma STAR (Statistical Techniques for Analytical Review) maakt gebruik van regressie-analyse en wordt gebruikt bij accountantscontroles. STAR kan voor twee verschillende situaties controles uitvoeren. Doel van de stage was een controlemodel te ontwikkelen dat met beide situaties rekening houdt, zodat bij de controle meer informatie wordt gebruikt.

Gedurende de stageperiode heb ik niet alleen aan mijn afstudeerproject gewerkt; binnen DQI heb ik ook andere werkzaamheden mogen verrichten. Ik wil daarom Paul van Batenburg bedanken voor de grote vrijheid die ik heb gekregen tijdens mijn stage en de mede daardoor ontzettend leuke en leerzame periode. Daarnaast wil ik Angela Kroep, mijn begeleider binnen DQI, bedanken voor alle uitleg, het lezen van mijn stukken en het wegwijs maken binnen Deloitte. Ook wil ik alle DQI-collega's bedanken voor de uitleg en de gezelligheid.

Verder wil ik de accountants Marc van Gestel, Marc Fierens en Maurits de Graaf bedanken voor het aanleveren van data en alle uitleg op bedrijfseconomisch vlak.

Vanuit de Vrije Universiteit heeft Geurt Jongbloed mij begeleid. Ik wil hem bedanken voor de goede inhoudelijke ideeën, uitleg en het doorlezen van mijn stukken. Zijn enthousiasme heeft mij enorm gemotiveerd.

Marit Schoonhoven

Augustus 2004

Samenvatting

Deze scriptie gaat over de toepassing van regressiemodellen bij accountantscontroles.

Het door Deloitte ontwikkelde programma STAR maakt gebruik van regressiemodellen om voorspellingen te genereren en deze voorspellingen te confronteren met te controleren waarden. Hierdoor kan gegevensgerichte zekerheid worden verkregen.

STAR kan in twee verschillende situaties worden gebruikt: Het programma controleert of gegevens van één eenheid op verschillende momenten op basis van gecontroleerde historische gegevens van die eenheid, of het controleert gegevens in één periode van verschillende eenheden op basis van gecontroleerde gegevens uit die periode van andere eenheden. Nadeel is dat in het eerste geval geen rekening kan worden gehouden met periode-effecten en in het tweede geval geen rekening kan worden gehouden met eenheid-effecten.

Doel van de stage was een gecombineerd model (SuperSTAR) op te stellen, dat rekening kan houden met beide effecten. Tevens moest het model worden toegepast op data van een bestaande organisatie.

Het gecombineerde model dat tijdens de stageperiode is opgesteld, maakt ook gebruik van regressie-analyse en neemt alle winkels en perioden in één model op. Hierdoor kunnen eenheden onderling en perioden onderling met elkaar worden vergeleken en kunnen zowel eenheid- als periode-effecten worden geschat. Wanneer zowel winkel- als periode-effecten aanwezig zijn, kan met dit model op basis van meer informatie worden gecontroleerd, omdat met beide effecten tegelijk rekening wordt gehouden.

De toepassing van SuperSTAR bij de klant bleek een goede test aangezien uit de uitvoering van de twee huidige STAR-modellen bleek dat bij deze klant zowel eenheid- als periode-effecten aanwezig zijn. Het gecombineerde model sloot beter aan op de data. Bovendien hoefden bij het gecombineerde model achteraf minder aanvullende werkzaamheden worden uitgevoerd dan bij de afzonderlijke STAR-modellen. Deze modellen maten bepaalde effecten niet waardoor fouten in de boekwaarden te negatief werden afgeschilderd.

Inhoudsopgave

1. Inleiding

1.1 STAR

De accountant heeft bij de controle van een financiële verantwoording, bijvoorbeeld van een jaarrekening, verscheidene controlemiddelen ter beschikking. STAR is een tool waarmee kwantitatieve gegevens kunnen worden gecontroleerd.

STAR stelt aan de hand van eerder gecontroleerde data een regressiemodel op. Aan de hand van dit model kunnen voorspellingen worden gegenereerd voor ongecontroleerde waarnemingen. Daarnaast berekent STAR per te controleren waarneming een drempelwaarde, gebaseerd op controledoelstellingen die vooraf bepaald zijn door de accountant. Deze drempelwaarde wordt ook wel threshold genoemd en wordt gebruikt om van de betreffende waarneming het verschil tussen voorspelling op basis van het regressiemodel en de ongecontroleerde waarde te beoordelen. Indien dit verschil groter dan de threshold is, moet de accountant extra werkzaamheden uitvoeren om eventueel alsnog goed te kunnen keuren.

Het regressiemodel kan voor twee verschillende situaties worden opgesteld. Bij tijdreeksanalyse zijn de gegevens voor het regressiemodel in chronologische volgorde gerangschikt en geeft elke waarneming de resultaten weer voor één eenheid in een bepaalde periode (meestal een maand of een kwartaal). Het regressiemodel kan dan worden gebruikt bij de controle van gegevens van dezelfde eenheid gemeten in andere perioden. Bij cross-sectionele analyse geven de gegevens resultaten aan gemeten op één tijdstip bij verschillende eenheden, bijvoorbeeld bij filialen. In dat geval kan het regressiemodel worden gebruikt bij de controle van gegevens van andere filialen gemeten in dezelfde periode.

1.2. Probleemstelling

Voor aanvang van de stage was het nog niet mogelijk een combinatie van beide modellen te gebruiken. Binnen Deloitte zijn echter klanten waar de combinatie van beide modellen een grote meerwaarde kan leveren. Een situatie waarin een dergelijk gecombineerd model nodig is, is bijvoorbeeld wanneer de omzet van een filiaal moet worden voorspeld uit zowel historische gegevens van dit filiaal als uit gegevens van andere, vergelijkbare, filialen.

Het doel van de stageopdracht was tweeledig:

1. Het opstellen en implementeren van een wiskundig model en de threshold, waardoor combinatie van beide typen modellen mogelijk is.

2. Dit model toepassen bij de controle van één of meerdere klanten.

Het gecombineerde model wordt binnen Deloitte ook wel SuperSTAR genoemd. Deze termen zullen in de rest van de scriptie door elkaar heen worden gebruikt.

1.3. Indeling van het verslag

Hoofdstuk 2 is gewijd aan Deloitte. In hoofdstuk 3 wordt de accountantscontrole behandeld. Dit hoofdstuk is bedoeld om het onderwerp van de stage in een breder kader te kunnen plaatsen. Vervolgens wordt in hoofdstuk 4 uitleg gegeven over lineaire regressie. Deze theorie is nodig om de werking van STAR en SuperSTAR te kunnen begrijpen. Hoofdstuk 5 en hoofdstuk 6 zijn volledig gewijd aan STAR, respectievelijk SuperSTAR. In hoofdstuk 7 wordt beschreven in welke situaties STAR en SuperSTAR kunnen worden gebruikt en hoofdstuk 8 beschrijft een toepassing van SuperSTAR bij een klant in de consumentenbranche. Tenslotte worden in hoofdstuk 9 de conclusies gegeven.

2. Deloitte

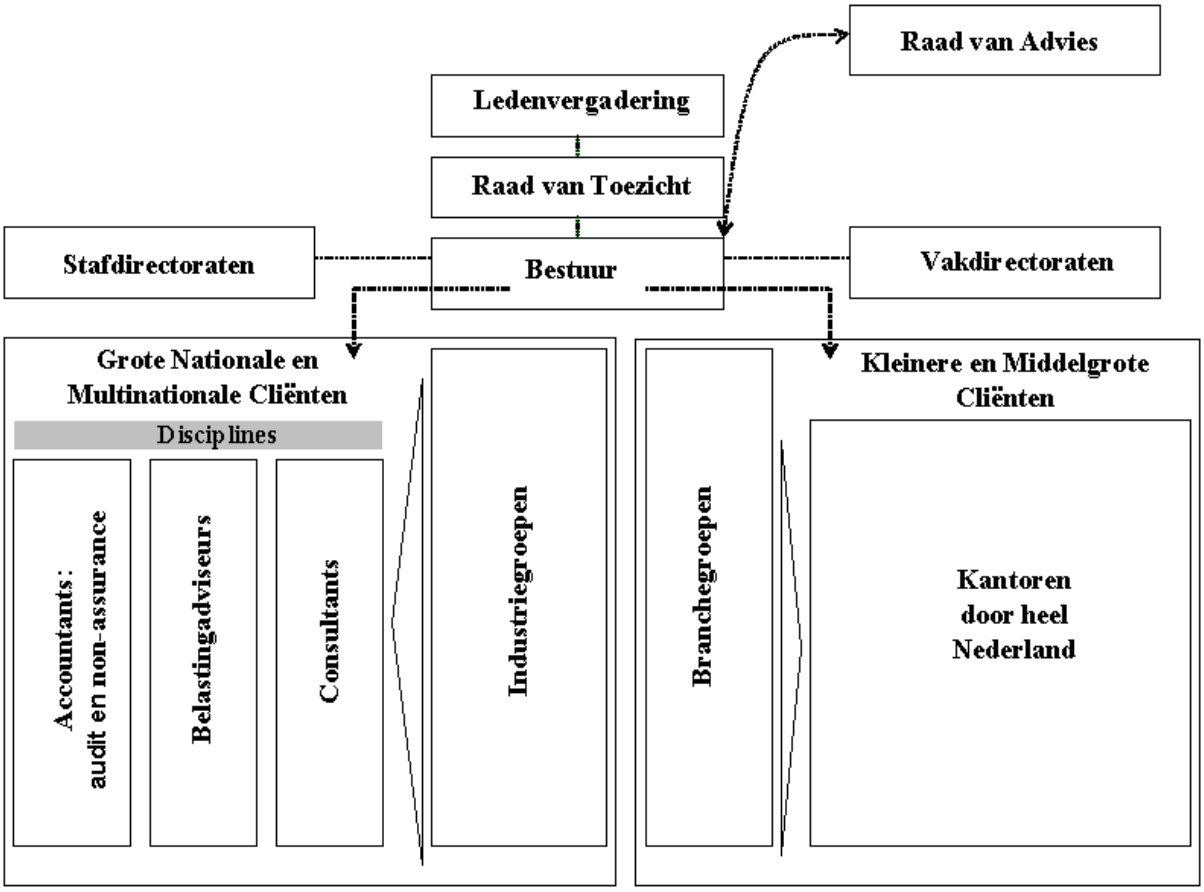
2.1. Deloitte Touche Tohmatsu

Deloitte Touche Tohmatsu telt 120.000 medewerkers verspreid over ongeveer 150 landen. Zij behoort naast KPMG, Price Waterhouse Coopers en Ernst & Young, tot de Big Four, de vier grootste accountantskantoren ter wereld.

Deloitte Touche Tohmatsu is gericht op de financieel zakelijke dienstverlening en wel op de volgende disciplines:

- Accountancy
- Belastingadvies
- Consultancy
- Financieel advies

Deloitte Nederland is een zelfstandige member firm van Deloitte Touche Tohmatsu. In Nederland werken ongeveer 6500 mensen bij Deloitte vanuit meer dan 50 kantoren. Deloitte Nederland is ontstaan door tal van overnames en fusies. Zo werd een uitgebreid landelijk netwerk van kantoren opgesteld. Het dienstenpakket is zowel op de internationale (corporate) markt als op de nationale markt gericht. De corporate market kent een indeling naar industrieën, de nationale markt is onderverdeeld in branchegroepen. Figuur 1 geeft de nationale organisatie van Deloitte weer.



Figuur 1: Organogram Deloitte Nederland

2.2. Enterprise Risk Services

Enterprise Risk Services (ERS) is opgericht als ondersteuningsafdeling voor de accountants van Deloitte Touche Tohmatsu. Momenteel biedt ERS naast ondersteuning aan de accountants ook diensten aan externe klanten. Het doel van ERS is het beschermen van de belangrijkste bedrijfsprocessen en het identificeren van bedrijfsrisico's voordat problemen optreden. Dit doel wordt bereikt door middel van:

- Het identificeren en meten van mogelijke risico's.
- Het bepalen in welke mate een organisatie klaar is om deze risico's te weerstaan.
- Het helpen opzetten van een intern controle systeem ter beheersing en controle van de risico's.

ERS kent een onderverdeling in de volgende competentiegroepen:

- Risk Consulting /Internal Audit (RC/IA)

- Control Assurance (CA)
- Data Quality and Integrity (DQI)
- Security Services Group (SSG)
- Web Services

2.3. Data Quality and Integrity

De stageopdracht is uitgevoerd binnen DQI. DQI is een competentiegroep binnen ERS. De doelstelling van DQI is het toepassen van wiskundige en statistische technieken en software vaardigheden om klanten te assisteren bij het oplossen van data gerelateerde vragen. De competenties van DQI zijn:

Data Quality

Het beoordelen en meten van de kwaliteit van de data, het beheersen van de kwaliteit van de data en het verbeteren van de kwaliteit door middel van corrigerende acties, zoals het opschonen van de data.

Business Metrics

Het identificeren en transformeren van de data in kennis en informatie die gebruikt kan worden bij het managen van bedrijfsrisico's. Hieronder valt ook de toepassing van benchmarking, performance scoring en datamining technieken.

Statistical Audit

Het ondersteunen van de accountant bij het opzetten, uitvoeren en evalueren van controlewerkzaamheden door toepassing van statistische methoden en technieken bij cijferanalyses, organisatietests en steekproefcontroles.

Het onderwerp van de stageopdracht valt binnen deze competentie.

Empirical Research

Het uitvoeren van statistisch onderzoek om specifieke vragen van klanten te kunnen beantwoorden. Hierbij kan worden gedacht aan statistische modelschatting en het testen van hypothesen.

3. De accountantscontrole

3.1. Doel van de controle

Doel van de accountantscontrole is het verkrijgen van zekerheid over de getrouwheid van een financiële verantwoording. In het geval van de jaarrekeningcontrole duidt getrouwheid op de mate van overeenstemming van de balans, winst- en verliesrekening en toelichting met de financiële werkelijkheid.

Om een oordeel over de getrouwheid van een jaarrekening te kunnen geven, moet een oordeel over de beweringen in de jaarrekening worden gegeven. Dit oordeel kan vanuit meerdere gezichtspunten worden bekeken. Dit worden controledoelstellingen genoemd. De primaire controledoelstellingen zijn juistheid en volledigheid. Bij controle van een boekhoudkundige grootheid op juistheid wordt gecontroleerd of de waarde van deze grootheid niet te hoog is. Bij controle op volledigheid wordt gecontroleerd of de waarde van de grootheid niet te laag is.

3.2. Materialiteit

Indien het verschil tussen de beweringen in de jaarrekening en de financiële werkelijkheid te groot is en het de oordeelsvorming van de gebruikers van de jaarrekening zou kunnen beïnvloeden, wordt het verschil materieel genoemd. Materialiteit heeft zowel betrekking op kwantitatieve beweringen zoals posten als op kwalitatieve beweringen zoals toelichtingen en andere informatie in de financiële verantwoording.

In deze scriptie worden methoden beschreven om gegevensgericht te controleren. Bij gegevensgerichte controles worden kwantitatieve beweringen beoordeeld. De materialiteit is dan uitgedrukt in een getal en geeft de maximaal toelaatbare afwijking aan tussen de ongecontroleerde waarde van de te controleren grootheid en de waarde voor deze grootheid die middels de controle is verkregen. In sommige gevallen wordt in plaats van de materialiteit de Monetary Precision (MP) gebruikt, een van de materialiteit afgeleid foutbedrag. Deze MP-waarde is zeventig tot negentig procent van de materialiteit en wordt gehanteerd om intensiever te controleren. Wanneer bij steekproeven deze MP-waarde wordt gehanteerd, kunnen iets meer fouten dan vooraf verwacht alsnog leiden tot een evaluatie die niet de materialiteit overschrijdt.

3.3. Het controleproces

Tijdens de interim-controle wordt de opzet, het bestaan en de werking van de administratieve organisatie en interne controle (AO/IC) beoordeeld. In deze fase wordt de controle-omgeving beoordeeld en wordt voor belangrijke bedrijfsprocessen naar specifiek getroffen maatregelen gekeken. Doel van deze werkzaamheden is te onderzoeken in hoeverre de accountant in de rest van het controleproces kan steunen op de AO/IC. In het laatste stadium van de interim-controle wordt bepaald welke aanvullende, gegevensgerichte werkzaamheden nog moeten

worden verricht om uiteindelijk een uitspraak te kunnen doen over de kwaliteit van de beweringen in de jaarrekening.

Er zijn twee soorten gegevensgerichte controlemiddelen:

- Cijferanalyses
- Detailwaarnemingen

STAR wordt gebruikt bij cijferanalyses. STAR controleert waarden op basis van een door de gebruiker ingevoerde MP-waarde en onbetrouwbaarheid α. Deze onbetrouwbaarheid is uitgedrukt in een R-waarde. De totstandkoming van deze R-waarde wordt in de paragraaf Steekproeven behandeld.

Detailwaarnemingen bestaan uit integrale controles, deelwaarnemingen en steekproeven. Bij integrale controles moet de hele populatie worden gecontroleerd, bij deelwaarnemingen hoeft slechts een gedeelte van de populatie te worden gecontroleerd. Ook in geval van steekproeven wordt slechts een selectie van de populatie gecontroleerd maar daarbij kan een statistisch onderbouwde uitspraak worden gedaan. De controles middels steekproeven en controles middels STAR bevatten overeenkomsten. De volgende paragraaf behandelt daarom de steekproefaanpak.

Steekproeven

Om een populatie te toetsen op een foutfractie p kunnen postensteekproeven of geldsteekproeven worden gebruikt. Bij een postensteekproef bestaat de populatie uit een aantal posten en is de foutfractie gedefinieerd als het percentage onjuiste posten. Bij een geldsteekproef bestaat de populatie uit euro's en is de foutfractie gedefinieerd als het percentage onjuiste euro's. Een populatie is acceptabel indien de foutfractie p kleiner is dan de bovengrens voor de foutfractie, p_m. Deze bovengrens wordt vastgesteld door de accountant en is gebaseerd op de materialiteit voor de jaarrekening en de toepassing waarvoor de steekproef wordt gebruikt.

De hypothesen zijn

$$\begin{aligned} H_0 &: p \geq p_m \\ H_1 &: p < p_m \end{aligned}.$$

De nulhypothese wordt verworpen indien in de steekproef 0 fouten worden aangetroffen en de steekproefgrootte n groot genoeg is zodat

$$(1\text{-}p_m)^n \leq \alpha .$$

Hierin is α de kans op het onterecht verwerpen van de nulhypothese en dus op het ten onrechte goedkeuren. In bovenstaande formule is de binomiale verdeling gebruikt. Door de ongelijkheid verder uit te werken wordt de volgende relatie gevonden

$$\begin{aligned}
&(1\text{-}p_m)^n \le \alpha \Leftrightarrow \\
&n\ln(1-p_m) \le \ln\alpha \Leftrightarrow \\
&n \ge \frac{\ln\alpha}{\ln(1-p_m)} \approx \frac{\ln\alpha}{-p_m} \Leftrightarrow \\
&np_m \ge -\ln\alpha
\end{aligned}.$$

De gelijkheid in de derde regel geldt alleen voor kleine waarden van p_m.

De relatie kan ook worden verkregen door de binomiale verdeling te benaderen door de Poisson verdeling

$$\begin{aligned}
&(1\text{-}p_m)^n \le \alpha \Leftrightarrow \\
&e^{-np_m} \le \alpha \Leftrightarrow \\
&-np_m \le \ln\alpha \Leftrightarrow \\
&np_m \ge -\ln\alpha
\end{aligned}.$$

Accountants drukken de onbetrouwbaarheid α van een toets uit in een R-waarde, met $R = -\ln\alpha$. De totstandkoming van deze R-waarde volgt uit bovenstaande afleidingen. Bij uitvoering van de steekproef volgt op deze manier de volgende relatie: Indien in een steekproef n en p_m zo worden gekozen dat geldt $np_m = R$ en er 0 fouten in de steekproef worden aangetroffen, wordt de nulhypothese met kans α onterecht verworpen.

4. Lineaire regressie

STAR stelt op basis van eerder gecontroleerde waarnemingen een lineair regressiemodel op. Aan de hand van dit regressiemodel kunnen voorspellingen en thresholds worden gegenereerd voor de ongecontroleerde waarnemingen.

In dit hoofdstuk wordt uitgelegd wat een lineair regressiemodel feitelijk is, hoe het kan worden opgesteld en aan welke veronderstellingen het model moet voldoen.

4.1. De regressievergelijking

Lineaire regressie draait om het zoeken van lineaire verbanden tussen een te verklaren variabele Y en één of meer verklarende variabelen $X_j\,(j=1,\ldots,k)$. Van deze verklarende variabelen wordt verondersteld dat deze een voorspellende waarde bezitten voor de te verklaren variabele. Doel is dus om een formule te vinden van de vorm

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + e\,.$$

Het aantal verklarende variabelen is in dit geval gelijk aan k. Wanneer $k=1$ is er één verklarende variabele en is er sprake van enkelvoudige lineaire regressie. Indien $k>1$ zijn er meerdere verklarende variabelen en is er sprake van meervoudige regressie. De coëfficiënten $b_0, b_1, \ldots b_k$ zijn de regressiecoëfficiënten en e is de foutenterm. De term lineair slaat op de lineairiteit van de modelvergelijking niet op de lineairiteit in de verklarende variabelen. Immers, een verklarende variabele mag ook een niet-lineaire functie zijn van een variabele.

4.2. Opstellen van het lineaire regressiemodel

Indien de schattingen $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k$ van de regressiecoëfficiënten $b_0, b_1, \ldots b_k$ zijn bepaald, kan voor iedere waarneming i de gerealiseerde waarde Y_i worden vergeleken met de voorspelde waarde $\hat{Y}_i$

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} + \ldots + \hat{b}_k X_{ki}\,.$$

Indien $b_0, b_1, \ldots b_k$ uit de data zijn geschat heet het verschil tussen realisatie en voorspelling het residu

$$r_i = Y_i - \hat{Y}_i\,.$$

De kwaliteit van de regressieresulaten wordt gemeten door deze residuen. Immers, hoe kleiner de residuen, des te beter de fit. Ordinary Least Squares kiest de waarden van de regressiecoëfficiënten zo dat de som van de gekwadrateerde residuen $(\sum r_i^2)$ minimaal is.

Deze fit is beschreven in de bijlage en kan worden gebruikt indien het model aan bepaalde statistische veronderstellingen voldoet. In paragraaf 4.3. worden de veronderstellingen beschreven.

Wanneer meer dan één potentiële verklarende variabele beschikbaar is, moet worden uitgezocht welke verklarende variabelen moeten worden toegevoegd aan het regressiemodel. Het is immers niet nodig overbodige informatie op te nemen in het model. Er bestaan verschillende procedures om te bepalen welke verklarende variabelen moeten worden opgenomen in het regressiemodel. De Backward-procedure gaat uit van een volledig model, bepaalt de minst 'significante variabele', en voert nogmaals meervoudige regressie uit zonder deze variabele. Deze stappen worden herhaald tot alle geschatte regressiecoëfficiënten significant van nul verschillen. De Forward-procedure werkt net andersom: deze procedure bepaalt welke variabelen moeten worden toegevoegd. De procedure Stepwise wisselt toevoeging en verwijdering van verklarende variabelen af. Deze procedure is beschreven in de bijlage.

4.3. Modelaannames

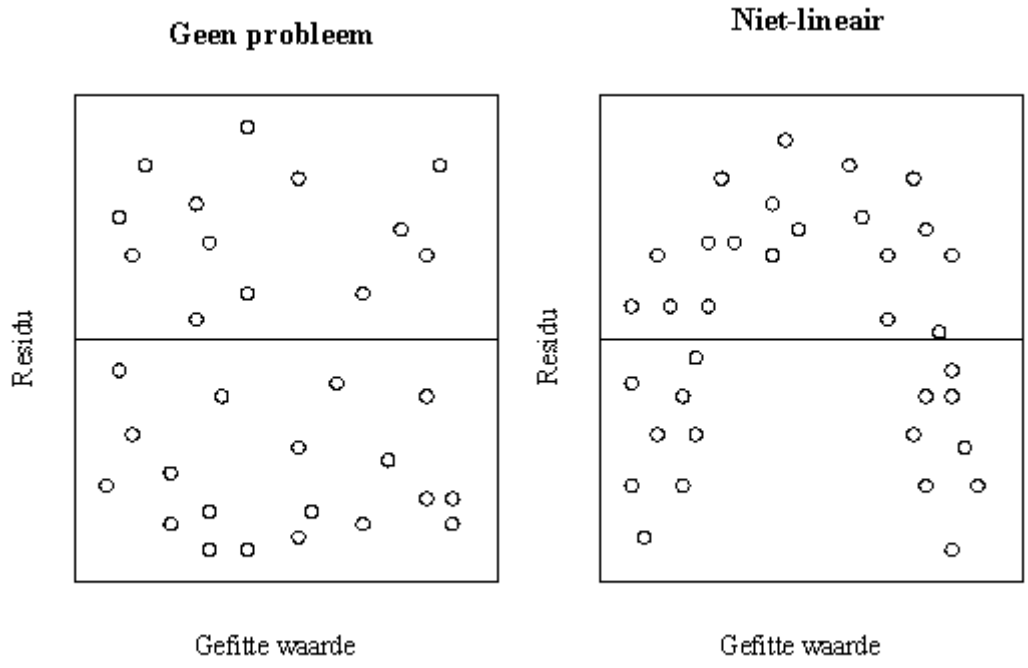
Verondersteld wordt dat het lineaire regressiemodel aan de volgende veronderstellingen voldoet:

- Het lineaire model is aanvaardbaar
- De fouten zijn onderling onafhankelijk
- De foutenterm heeft een constante variantie
- De foutenterm is normaal verdeeld met verwachtingswaarde nul

Er zijn grafische methoden en statistische toetsen mogelijk om het model op bovenstaande veronderstellingen te testen. De formele toetsen en aanpassingen die in STAR en SuperSTAR worden gehanteerd, zijn beschreven in hoofdstuk 5, respectievelijk hoofdstuk 6. In de rest van deze paragraaf worden per veronderstelling de grafische methoden beschreven waardoor tevens meer inzicht kan worden verkregen in de aanname.

Aanvaardbaarheid van het lineaire model

In een scatterplot kunnen de residuen tegen de voorspelde waarden worden uitgezet. Indien de punten ongeveer gelijkmatig verspreid liggen langs beide kanten van de horizontale lijn door nul, is aan de veronderstelling voldaan. Indien een patroon (zoals bijvoorbeeld een parabool of logaritmische curve) te zien is, is een lineaire functie niet de juiste manier om de gegevens te beschrijven.



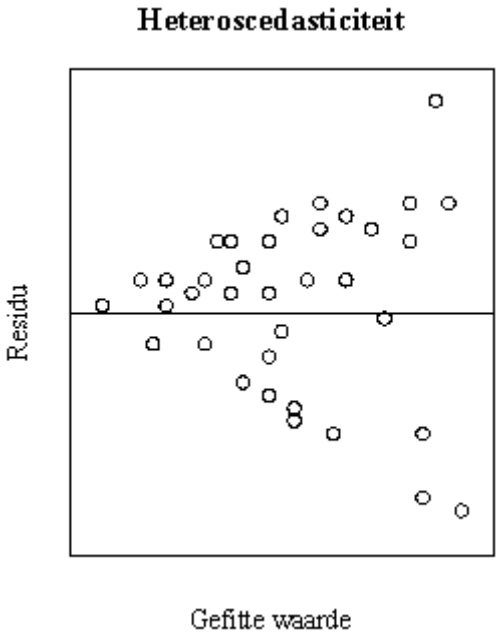
Figuur 2: Scatterplots residuen tegen gefitte waarden – De eerste plot suggereert dat er geen verandering nodig is in het huidige model, de tweede plot indiceert non-lineairiteit

Afhankelijkheid van de fouten

Indien bij tijdreeksanalyse een verband tussen opeenvolgende residuen is te zien (bijvoorbeeld een lange rij positieve residuen gevolgd door een lange rij negatieve residuen), zijn de residuen afhankelijk. Dit wordt ook wel autocorrelatie genoemd. Autocorrelatie kan tot gevolg hebben dat de voorspellingen minder goed zijn doordat het patroon niet in het model is opgenomen of dat de berekeningen van de standaard afwijkingen negatief worden beïnvloed.

Variantie van de fouten

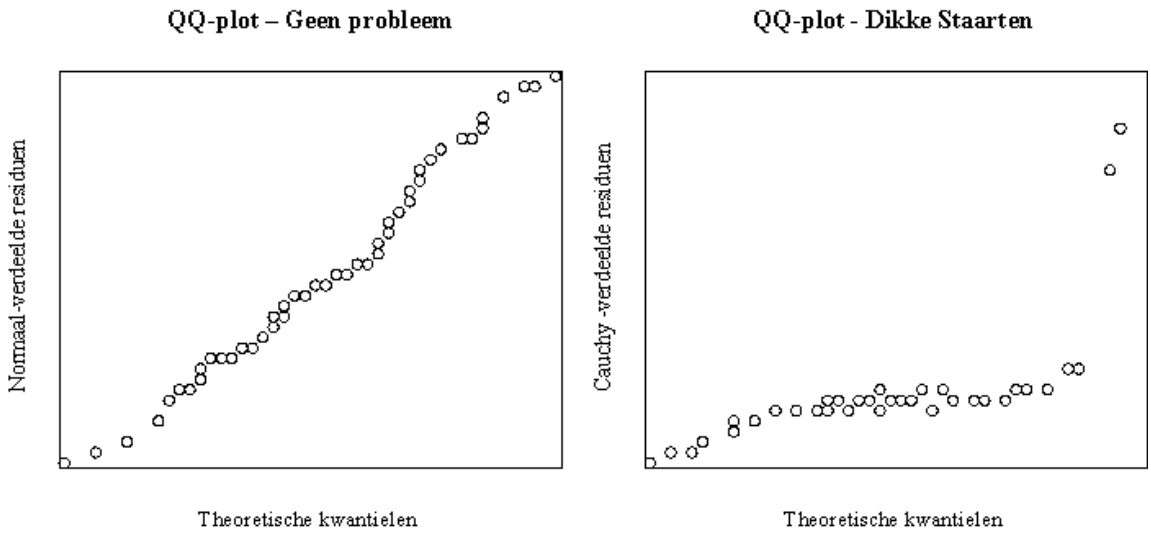
Wanneer de foutenterm een constante variantie heeft, wordt het model homoscedastisch genoemd. Als niet aan deze aanname is voldaan, is er sprake van heteroscedasticiteit. Er zijn verschillende vormen van heteroscedasticiteit. Eén ervan is wanneer de foutenterm afhankelijk is van het Y-niveau. Een scatterplot waarin de residuen zijn uitgezet tegen de voorspelde waarden kan helpen deze vorm van heteroscedasticiteit te ontdekken. Als de punten ongeveer even breed rond de nullijn verspreid liggen, kan worden aangenomen dat de foutenterm niet afhankelijk is van het Y-niveau. Indien een model heteroscedastisch is, kan de standaard fout van het model niet zonder aanpassing worden gebruikt.



Figuur 3: Scatterplot residuen tegen gefitte waarden – De plot suggereert dat de variantie van de foutenterm niet constant is.

Normaliteit van de residuen

Deze veronderstelling kan worden nagegaan door een QQ-plot van de residuen op te stellen. In de plot worden de residuen uitgezet tegen de theoretische kwantielen van de normale verdeling. Indien de residuen bij benadering normaal verdeeld zijn, geeft de QQ-plot een rechte lijn weer.



Figuur 4: QQ-plots – De eerste plot suggereert dat de foutenterm normaal is verdeeld, de tweede plot suggereert dat de verdeling van de foutenterm dikke staarten heeft en afwijkt van de normale verdeling.

Indien de residuen niet normaal zijn verdeeld zijn de schattingen voor de regressiecoëfficiënten niet optimaal en zijn de testen en betrouwbaarheidsintervallen ongeldig. Mogelijk kan de bootstrap-methode worden gebruikt bij de bepaling van de betrouwbaarheidsintervallen. Bij deze methode hoeft de verdeling van de residuen niet van tevoren bekend te zijn.

Tenslotte moet bij meervoudige regressie ook gekeken worden naar de correlaties tussen de te verklaren variabele en verklarende variabelen en de correlaties tussen de verklarende variabelen onderling. De correlaties tussen de te verklaren variabelen en de verklarende variabelen afzonderlijk laten zien hoe goed het regressiemodel zou zijn indien lineaire regressie wordt uitgevoerd van de te verklaren variabele ten opzichte van de verklarende variabele. Een te grote correlatie tussen verklarende variabelen onderling duidt op multi-collineariteit. Dit veroorzaakt problemen bij de schatting van de regressiecoëfficiënten.

5. STAR

In dit hoofdstuk komt de werking van STAR aan bod. Achtereenvolgens worden de input-data, het regressiemodel, de toetsing en aanpassing van het regressiemodel en de threshold behandeld.

5.1. Input-data

De input-data voor STAR ziet er als volgt uit:

	Te verklaren variabele	Verklarende variabele
Basis	400,00	146,08
	1.525,19	337,13
	1.313,77	305,58
	4.052,95	894,93
	2.207,44	481,89
	2.187,89	477,68
	2.805,96	628,78
	3.310,48	732,15
	141,34	45,06
	2.015,06	449,33
	1.906,91	422,28
Projectie	3.948,60	868,16
	3.249,30	707,73
	1.954,21	445,12
	3.826,16	847,60
	2.101,89	476,09
	4.308,34	941,07

Figuur 5 : Lay-out van de input-data voor STAR

De eerste kolom van de dataset bevat de waarden van de te verklaren variabele, de kolommen ernaast bevatten de waarden van de verklarende variabelen. Een rij in de dataset representeert een waarneming. In geval van tijdreeksanalyse zijn de gegevens in chronologische volgorde gerangschikt en geeft elke waarneming de resultaten van dezelfde eenheid gemeten in een verschillende periode weer. Bij cross-sectionele analyse geeft elke waarneming resultaten weer van een verschillende eenheid, gemeten in dezelfde periode.

De waarnemingen zijn opgesplitst in basiswaarnemingen en projectiewaarnemingen. De waarden van de te verklaren variabele behorend tot de projectiewaarnemingen moeten worden gecontroleerd. De waarden van de verklarende variabelen behorend tot de projectiewaarnemingen en de waarden van zowel de te verklaren variabele als van de verklarende variabelen behorend tot de basiswaarnemingen, zijn gecontroleerd.

5.2. Regressiemodel

STAR stelt op basis van n gecontroleerde basiswaarnemingen een regressiemodel op. Aan de hand van dit regressiemodel kunnen voorspellingen voor de te verklaren variabele in de projectieperiode of van de projectie-eenheid worden gegenereerd.

STAR gebruikt de procedure Stepwise voor de selectie van variabelen. De waarden van de regressiecoëfficiënten worden in eerste instantie middels Ordinary Least Squares bepaald (zie bijlage 1).

Bij tijdreeksanalyse geeft het regressiemodel het verband weer tussen de te verklaren variabele (Y_t) en k verklarende variabelen $(X_{jt}, j = 1,\ldots,k)$ en geldt dit verband voor één eenheid in verschillende perioden

$$Y_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + \ldots + b_k X_{kt} + e_t .$$

De index t geeft aan dat het een tijdreeksanalyse[1] betreft.

Een voorbeeld van een tijdreeks model is een model voor de omzet van een handelsonderneming. Binnen de onderneming zijn gecontroleerde gegevens bekend over de omzet en de kostprijs van de omzet per maand van twee jaar voorafgaand aan de controle. Deze gegevens behoren tot de basiswaarnemingen. Op basis van deze gegevens en de geschatte relatie daartussen kan een voorspelling gemaakt worden voor de omzet per maand van het te controleren jaar. De voorspelling voor maand t $(\hat{Y}_t)$ wordt verkregen door de waarden van de verklarende variabelen in maand t $(X_{jt}, j = 1,\ldots k)$ te substitueren in de geschatte regressiefunctie.

Bij cross-sectionele analyse geldt het model voor verschillende eenheden in dezelfde periode

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki} + e_i .$$

De index i geeft aan dat het een cross-sectionele analyse betreft.

Een voorbeeld van cross-sectionele analyse is wanneer de omzet van verschillende winkels van een bedrijf op hetzelfde tijdstip met elkaar wordt vergeleken. Op basis van gecontroleerde gegevens van verschillende winkels kan een verband tussen de omzet en bijvoorbeeld de oppervlakte van de winkel, of het aantal personeelsleden, worden geschat om de omzet van een andere winkel te voorspellen. De voorspelling voor winkel i $(\hat{Y}_i)$ wordt verkregen door de waarden van de verklarende variabelen van winkel i $(X_{ji}, j = 1,\ldots k)$ te substitueren in de geschatte regressiefunctie.

---

[1] Deze terminologie wijkt af van de geaccepteerde norm in de statistiek-literatuur

5.3. Toetsing en aanpassing van het regressiemodel

In deze paragraaf wordt beschreven hoe STAR het regressiemodel toetst en vervolgens eventueel aanpast. Voor de formules wordt verwezen naar Stewart en Stringer (1985).

Discontinuïteit van de regressiemodel

In geval van tijdreeksanalyse toetst STAR het lineaire model op continuïteit. Dit houdt in dat voor alle waarnemingen dezelfde onderliggende lineaire relatie moet gelden. Het tegenovergestelde, discontinuïteit, kan optreden tussen de basis- en projectieperiode en in de basisperiode zelf. STAR gebruikt een statistische toets om het model op beide vormen van discontinuïteit te controleren. Bij deze toets worden drie verschillende regressiemodellen opgesteld: twee voor de perioden waartussen mogelijk discontinuïteit kan zijn, en één voor de totale periode. Wanneer de som van de gekwadrateerde residuen van deze laatste regressiefunctie significant groter is dan de som van de gekwadrateerde residuen van de twee regressiefuncties samen, is sprake van discontinuïteit tussen de twee perioden. Indien sprake is van discontinuïteit in de basisperiode mag het model niet voor de controle worden gebruikt. Discontinuïteit tussen de basis- en projectieperiode wordt wel geaccepteerd. Immers, de reden van deze discontinuïteit kunnen fouten of ongebruikelijke transacties in de projectieperiode zijn die STAR juist moet ontdekken.

Autocorrelatie

STAR toetst middels een statistische test of opeenvolgende residuen afhankelijk zijn. Deze test is gebaseerd op de relatieve grootte van de som van de gekwadrateerde verschillen tussen opeenvolgende residuen. Indien de toetsingsgrootheid klein is, zijn de verschillen tussen opeenvolgende residuen significant kleiner dan wanneer de residuen onafhankelijk zouden zijn. Dit duidt op positieve correlatie. Wanneer de residuen afhankelijk zijn gebruikt STAR Generalised Least Squares om opnieuw de waarden van de regressiecoëfficiënten te bepalen.

Heteroscedasticiteit

STAR gaat uit van de veronderstelling dat wanneer heteroscedasticiteit optreedt, dit veroorzaakt wordt doordat de variantie van de foutenterm proportioneel gerelateerd is aan één van de verklarende variabelen. Om deze vorm van heteroscedasticiteit te ontdekken berekent STAR de correlatiecoëfficiënt tussen de grootte van de residuen en elke verklarende variabele en test vervolgens of de hoogste coëfficiënt significant is. Indien sprake is van heteroscedasticiteit, bepaalt STAR opnieuw de waarden van de regressiecoëfficiënten middels Weighted Least Squares. Bij deze methode wordt de data getransformeerd door iedere waarneming te delen door de waarde van de verklarende variabele die proportioneel gerelateerd is aan de variantie van de foutenterm. Op basis van deze getransformeerde data wordt een nieuw regressiemodel opgesteld middels Ordinary Least Squares.

Normaliteit

STAR toetst de onderliggende afwijkingen op scheefheid en kurtosis. Positieve (negatieve) scheefheid betekent dat het gemiddelde van de verdeling links (rechts) van de piek ligt. Kurtosis houdt in dat de verdeling extreem gepiekt of plat is. Ook wanneer uit de tests komt dat de fouten niet normaal zijn verdeeld, worden de thresholds berekend op basis van de veronderstelling dat de fouten normaal verdeeld zijn. In het artikel van Lam en Veall (2001) blijkt echter dat dit in sommige gevallen zeer onnauwkeurige thresholds kan opleveren.

5.4. Threshold

In Stewart en Stringer (1985) wordt de threshold-bepaling uitgelegd. De uitleg is echter niet in de vorm van hypothesetoetsen. In deze paragraaf wordt daarom aan de hand van hypothesetoetsen de threshold-bepaling uitgelegd. Hierbij is de notatie voor tijdreeksanalyse gehanteerd.

STAR berekent voor iedere projectiewaarneming de threshold. Deze threshold geeft bij controle op juistheid (volledigheid) de maximaal (minimaal) toelaatbare waarde $Y_t - \hat{Y}_t$ aan. Op deze wijze worden de volgende hypothesen bij controle op juistheid (volledigheid) getoetst

$$\begin{aligned} H_0 &: \text{De additieve fout is} \geq (\leq)\ MP\ (-MP) \\ H_1 &: \text{De additieve fout is} < (>) \text{ dan } MP\ (-MP) \end{aligned} . \qquad (1)$$

Bij deze toets kunnen de R -waarden 3.0, 2.0 en 0.7 worden gehanteerd. Bijbehorende onbetrouwbaarheden zijn respectievelijk 0.05, 0.135 en 0.5.

De berekening van de threshold wordt bemoeilijkt door het feit dat het aantal projectiewaarnemingen waarover een eventuele fout is verspreid ('spread of error' (SOE)), onbekend is. Indien de SOE bekend zou zijn is de kans op het onterecht verwerpen van bovengenoemde nulhypothese precies gelijk aan α. Dit wordt hieronder aangetoond voor toetsing op juistheid.

Voor SOE = 1:

De materiële fout zit in één projectiewaarneming.

Voor iedere projectiewaarneming t worden binnen het model

$$Y_t = b_0 + b_1 X_{t1} + \ldots + b_{tk} X_k + f_t + e_t . \quad (\text{met } f_t \text{ de parameter voor 'fraude'of fout})$$

de volgende hypothesen getoetst

$$H_0 : f_t \geq MP$$
$$H_1 : f_t < MP$$

(Bij STAR wordt de additieve dus fout vergeleken met de controletolerantie, in de steekproefaanpak wordt deze in een fractie p_m vertaald.)

Deze nulhypothese wordt verworpen indien $y_t - \hat{y}_t < c_{n,\alpha}(1)$, met $c_{n,\alpha}(1)$ de threshold voor projectiewaarneming t op basis van een SOE van 1

$$c_{n,\alpha}(1) = t_{n-k-1,\alpha} s_n \sqrt{1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}} + MP .$$

Hierbij is $t_{v;\alpha}$ het α-punt van de t-verdeling met v vrijheidgraden: $P(T_v \leq t_{v,\alpha}) = \alpha$ als $T_v \sim t_v$, $\mathbf{a}' = [1, x_{1t}, x_{2t}, \ldots, x_{kt}]$, $\mathbf{X}$ is een $n \times (k+1)$ matrix (waarin één rij de waarden van de verklarende variabelen van één basiswaarneming bevat) en s_n de standaardafwijking van het regressiemodel

$$s_n = \sqrt{\frac{\sum_{t=1}^{n} r_t^2}{n-k-1}} .$$

Indien de materiële fout in projectiewaarneming t zit, wordt deze niet gedetecteerd met kans α. Wanneer de kans op een materiële fout in waarneming t is gelijk is aan p_t met $\sum p_t = 1$, dan is de kans dat de nulhypothese in (1) onterecht wordt verworpen gelijk aan $\sum \alpha p_t = \alpha$.

Algemeen, voor SOE = l:

Verondersteld wordt dat de materiële fout uniform is verdeeld over l projectiewaarnemingen. Volgens Stewart en Stringer (1985) resulteert dit in de meest conservatieve thresholds.

Voor iedere projectiewaarneming t behorend bij indexverzameling $I_l \subset \{1,2\ldots,m\}$ met $|I_l| = l$ en m het aantal projectiewaarnemingen, worden binnen het model

$$Y_t = b_0 + b_1 X_{t1} + \ldots + b_{tk} X_k + f_t + e_t .$$

de volgende hypothesen getoetst

$$H_0 : f_t \geq MP/l$$
$$H_1 : f_t < MP/l$$

Deze nulhypothese wordt verworpen indien voor waarneming t geldt dat $y_t - \hat{y}_t < c_{n,\alpha}(l)$ en de nulhypothese in (1) wordt verworpen indien $y_t - \hat{y}_t < c_{n,\alpha}(l)$ voor $\forall t \in I_l$, met $c_{n,\alpha}(l)$ de threshold voor projectiewaarneming t op basis van een SOE van l

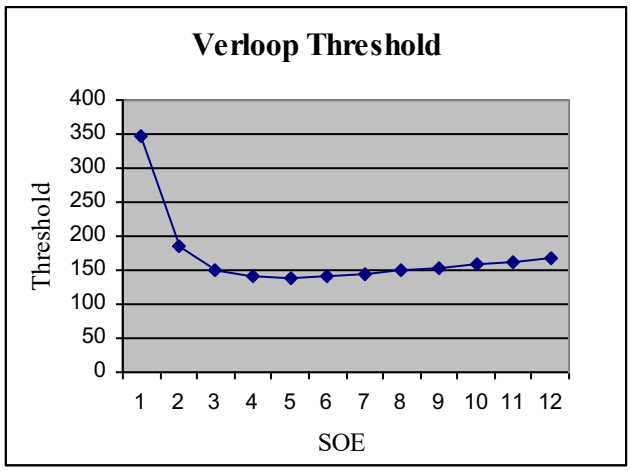
$$c_{n,\alpha}(l) = t_{n-k-1,\alpha_l} s_n \sqrt{1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}} + MP/l \,.$$

Wanneer de fout in een combinatie van l waarnemingen aanwezig is, wordt deze fout niet ontdekt met kans $(\alpha_l)^l$, onder de veronderstelling dat het wel of niet materieel zijn van waarnemingen ongecorreleerd is. Door α_l te kiezen als $\alpha_l = \sqrt[l]{\alpha}$, is de kans dat de fout in de l waarnemingen niet wordt ontdekt gelijk aan α , en dus is de kans dat de nulhypothese in (1) onterecht wordt verworpen gelijk aan α .

Bij testen op juistheid wordt bij oplopende SOE de threshold eerst kleiner, bereikt een minimum en wordt vervolgens groter. Dit kan worden uitgelegd aan de hand van de formule voor de threshold. De afgeleide van de eerste term naar l is groter dan de afgeleide tweede term in absolute waarde. Aangezien voor kleine waarden van de SOE de eerste term negatief is en de tweede term positief, wordt het verschil tussen de twee, de threshold, steeds kleiner (en is positief). Voor grotere waarden van de SOE is de linker term stijgend en gaat de tweede term naar 0, de threshold wordt dan dus steeds groter.

Aangezien de SOE onbekend is, is ook onbekend welke threshold moet worden gebruikt. STAR gaat daarom per projectiewaarneming uit van een SOE waarvoor geldt dat bij die SOE het ontdekken van een fout het moeilijkst is. Hierbij hoort zodoende de kleinste threshold. Er is geen maximum waarde gesteld voor deze SOE. Door gebruik te maken van deze meest conservatieve thresholds, is de kans op het onterecht verwerpen van de nulhypothese in (1) **maximaal** α .

Figuur 6 laat het verloop van de threshold zien bij een projectiewaarneming t met $s_n \sqrt{1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}} = 150$, $MP = 600$, $R = 3$ en het aantal vrijheidsgraden is 34. De threshold is gelijk aan het minimum van de grafiek.



Figuur 6: De threshold als functie van de SOE

De SOE waarbij de threshold het laagst is afhankelijk van twee belangrijke factoren: de relatieve MP $(MP / s_n \sqrt{1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}})$ en de R-waarde $(R = -\ln(\alpha))$. Hoe groter de relatieve MP en hoe lager de R-waarde, des te groter de SOE waarbij de threshold het laagst is. Dit wordt in de rest van deze paragraaf aangetoond. De formule van de threshold $(c_{n,\alpha}(l))$ wordt gedifferentieerd naar l en deze afgeleide wordt gelijk aan 0 gesteld. Op deze manier wordt die l gevonden waarvoor $c_{n,\alpha}(l)$ minimaal is

$$c_{n,\alpha}(l) = t_{n-k-1,\sqrt[l]{\alpha}}k + \frac{MP}{l} \quad (k = s_n\sqrt{1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}) \ (2).$$

Omdat de afgeleide van $c_{n,\alpha}(l)$ berekend moet worden, worden continue waarden van l bekeken $(l \geq 1)$. Voor grote waarden van n kan (2) geschreven worden als

$$c_{n,\alpha}(l) = u_{\sqrt[l]{\alpha}}k + \frac{MP}{l} \quad (\text{voor } u_{\sqrt[l]{\alpha}} \text{ geldt}: \int_{-\infty}^{u_{\sqrt[l]{\alpha}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \sqrt[l]{\alpha}).$$

De afgeleide van $c_{n,\alpha}(l)$ kan als volgt worden berekend

$$c_{n,\alpha}'(l) = (u_{\sqrt[l]{\alpha}})'k - \frac{MP}{l^2} \quad (\text{met } (u_{\sqrt[l]{\alpha}})' = (\Phi^{-1}(\sqrt[l]{\alpha}))' \times (\sqrt[l]{\alpha})' = \sqrt{2\pi}e^{\frac{1}{2}\alpha^{\frac{2}{l}}}\alpha^{\frac{1}{l}}\frac{-\ln\alpha}{l^2}) \Leftrightarrow$$

$$c_{n,\alpha}'(l) = \sqrt{2\pi}e^{\frac{1}{2}\alpha^{\frac{2}{l}}}\alpha^{\frac{1}{l}}\frac{-\ln\alpha}{l^2}k - \frac{MP}{l^2} \Leftrightarrow \quad .$$

$$e^{\frac{1}{2}\alpha^{\frac{2}{l}}}\alpha^{\frac{1}{l}} = \frac{MP}{k(-\ln\alpha)\sqrt{2\pi}}$$

Uit de laatste vergelijking volgt dat wanneer MP/k stijgt, l ook zal stijgen omdat α waarden tussen 0 en 1 aanneemt. Numeriek kan worden aangetoond dat indien de R-waarde daalt (en dus stijgt) α ook zal stijgen.

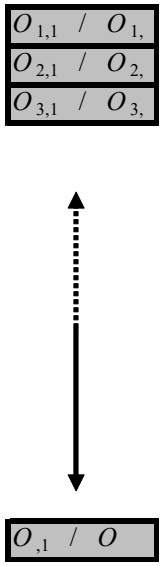
6. SuperSTAR

In dit hoofdstuk wordt een aantal mogelijkheden voor het gecombineerde model beschreven om controles bij verschillende eenheden op verschillende tijdstippen uit te kunnen voeren. Eén van de doelen van de stage was het gecombineerde model toe te passen op data van een bestaand bedrijf. Hiervoor is een bedrijf uit de consumentenbranche gekozen. Ter inleiding wordt in de volgende paragraaf een situatiebeschrijving van dit bedrijf gegeven.

6.1. Situatiebeschrijving van de klant

Bij het bedrijf waar SuperSTAR is toegepast, moet de omzet van verschillende filialen, gerealiseerd in meerdere maanden, worden gecontroleerd op volledigheid. Momenteel wordt deze controle handmatig uitgevoerd. Voor iedere maand wordt per filiaal de fractie maandomzet van de jaaromzet berekend. Tevens wordt op maandbasis de fractie totaal maandomzet van het totaal jaaromzet van alle filialen bepaald. Wanneer een fractie van een bepaald filiaal afwijkt van de fractie van alle filialen in de betreffende maand, moet een verklaring voor het verschil worden gezocht.

	Januari	Februari	…	December	Totaal per filiaal
Filiaal 1	$O_{1,1}$				$O_{1,}$
Filiaal 2	$O_{2,1}$				$O_{2,}$
Filiaal 3	$O_{3,1}$				$O_{3,}$
…					
…					
…					
…					
…					
…					
…					
…					
…					
…					
…					
…					
Totaal per maand	$O_{,1}$				O



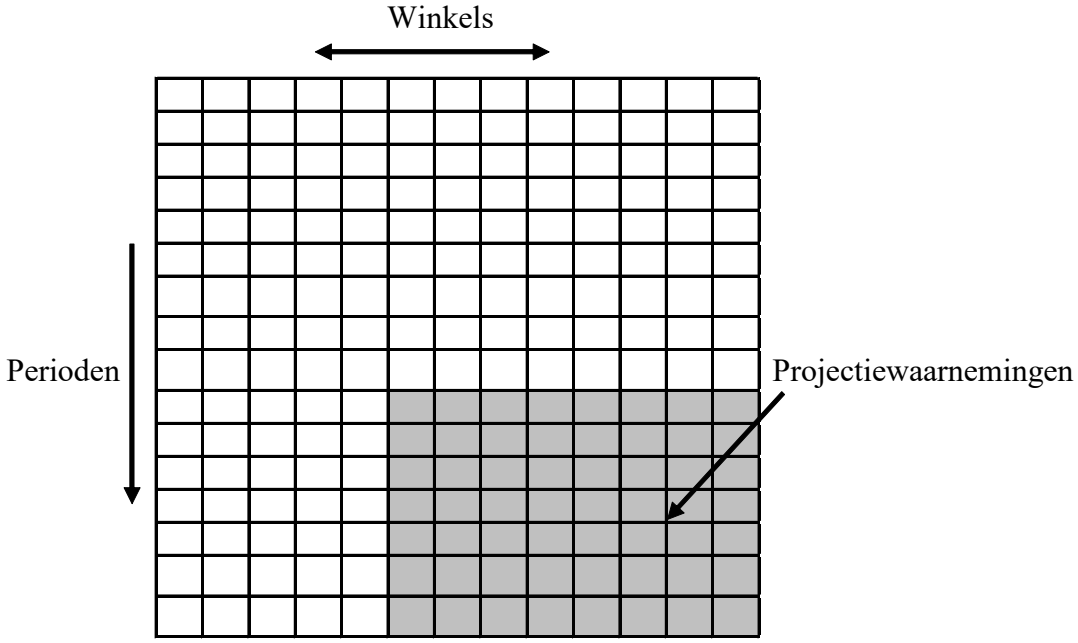
Figuur 7: Controle van de filiaal-omzetten gerealiseerd in januari. $O_{i,t}$ is de omzet van filiaal i in maand t, $O_{,t}$ is de totale omzet in maand t en $O_{i,}$ is de jaaromzet van filiaal i.

Middels STAR kan de omzet op een snellere en efficiëntere wijze worden gecontroleerd. In dat geval zijn één of meer verklarende variabelen nodig waarvan wordt verwacht dat deze goede voorspellers zijn voor de omzet. Afhankelijk van de beschikbare data kan vervolgens een tijdreeks- of cross-sectionele analyse worden uitgevoerd. Nadeel van apart uitvoeren van tijdreeks- of cross-sectionele analyse is dat periode-, respectievelijk winkel-effecten, niet

noodzakelijk worden meegenomen in de analyse. Zo is bijvoorbeeld te verwachten dat een winkel gesitueerd in een gebied waar veel concurrentie is, een lagere marge heeft dan een winkel die minder last heeft van concurrentie. Doel van het gecombineerde model is, indien noodzakelijk, rekening te houden met dergelijke winkel- en periode-effecten.

6.2. Mogelijkheden gecombineerd model

Om zowel eenheid- als periode-effecten in de analyse mee te kunnen nemen, zijn historische gegevens nodig van de te controleren filialen en gegevens van andere filialen gemeten in de basis- èn projectie-periode.



Figuur 8: Input-data voor SuperSTAR. Een hokje representeert een waarneming. Een kolom bevat waarnemingen van één winkel (eenheid), een rij bevat waarnemingen van één periode. Het grijze gedeelte bevat de ongecontroleerde waarnemingen.

Op basis van deze data kunnen constante verschillen en verschillen in verbanden tussen de te verklaren variabele en de verklarende variabelen, tussen winkels onderling en perioden onderling, in kaart worden gebracht. De regressievergelijking ziet er als volgt uit

$$Y_{it} = b_0 + b_{0t}^{cross} + b_{0i}^{tijd} + (b_{1t}^{cross} + b_{1i}^{tijd})X_{1it} + .. + (b_{kt}^{cross} + b_{ki}^{tijd})X_{kit} + e_{it}.$$

met

Y_{it} := De waarde van de te verklaren variabele bij eenheid *i* op tijdstip *t*

b_0 := De intercept

b_{0t}^{cross} := Het constante verschil tussen periode t en de intercept

b_{0i}^{tijd} := Het constante verschil tussen eenheid i en de intercept

b_{jt}^{cross} := De coëfficiënt van verklarende variabele j bij periode t

b_{ji}^{tijd} := De coëfficiënt van verklarende variabele j bij eenheid i

X_{jit} := De waarde van verklarende variabele j bij eenheid i op tijdstip t

e_{it} := De foutenterm bij eenheid i op tijdstip t.

Wanneer één type verklarende variabele beschikbaar is en deze wordt gebruikt om zowel eenheid- als periode-effecten te schatten, verondersteld wordt dat er geen constante verschillen aanwezig zijn, het aantal winkels gelijk is aan m en het aantal perioden gelijk is aan n, ziet het model in matrixnotatie er als volgt uit

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n} \\ \vdots \\ \vdots \\ Y_{m1} \\ Y_{m2} \\ \vdots \\ Y_{mn} \end{pmatrix} = \begin{pmatrix} x_{111} & 0 & \cdots & 0 & x_{111} & 0 & \cdots & 0 \\ x_{112} & 0 & \cdots & 0 & 0 & x_{112} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ x_{11n} & 0 & \cdots & 0 & 0 & 0 & \cdots & x_{11n} \\ 0 & x_{121} & \cdots & 0 & x_{121} & 0 & \cdots & 0 \\ 0 & x_{122} & \cdots & 0 & 0 & x_{122} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & x_{12n} & \cdots & 0 & 0 & 0 & \cdots & x_{12n} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & x_{1m1} & x_{1m1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & x_{1m2} & 0 & x_{1m2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & x_{1mn} & 0 & 0 & \cdots & x_{1mn} \end{pmatrix} \times \begin{pmatrix} b_{11}^{tijd} \\ b_{12}^{tijd} \\ \vdots \\ b_{1m}^{tijd} \\ b_{11}^{cross} \\ b_{12}^{cross} \\ \vdots \\ b_{1n}^{cross} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1n} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2n} \\ \vdots \\ \vdots \\ e_{m1} \\ e_{m2} \\ \vdots \\ e_{mn} \end{pmatrix}.$$

Figuur 9: Het model in matrixnotatie

De waarden van de regressiecoëfficiënten kunnen in eerste instantie middels Ordinary Least Squares worden bepaald. Er kan een zoekprocedure, bijvoorbeeld Stepwise, worden gebruikt om te onderzoeken welke verklarende variabelen in het model moeten worden opgenomen. Deze procedure bepaalt dan ook voor welke eenheden constante verschillen ten opzichte van

de intercept een rol spelen en voor welke perioden constante verschillen ten opzichte van de intercept waarneembaar zijn. Naast de intercept b_0 kan niet voor iedere periode b_{0t}^{cross} en voor iedere eenheid b_{0i}^{tijd} worden meegenomen; de kolommen van X zijn dan lineair afhankelijk waardoor de regressiecoëfficiënten niet kunnen worden geschat. Een andere mogelijkheid is de intercept b_0 uit het model weg te laten en voor iedere winkel b_{0i}^{tijd} en voor iedere periode b_{0t}^{cross} op te nemen.

Het is mogelijk dat periode- en/ of eenheid-effecten uit het model kunnen worden weggelaten. Hieronder wordt een aantal voorbeelden gegeven van simpelere varianten van het gecombineerde model.

Geen verschillen tussen eenheden en perioden

Wanneer de modellen voor verschillende perioden en eenheden niet significant afwijken, kunnen de data van alle perioden en eenheden op één hoop worden gegooid. De eventuele periode- en eenheid-effecten moeten in dit model via de verklarende variabelen komen. Het model ziet er als volgt uit

$$Y_{it} = b_0 + b_1 X_{1it} + .. + b_k X_{kit} + e_{it}.$$

Alleen verschillen tussen perioden

De coëfficiënten zijn per periode verschillend, maar daarnaast zijn geen specifieke eenheid-effecten waarneembaar. In dit geval is het niet nodig gebruik te maken van SuperSTAR; voor iedere te controleren periode kan een cross-sectionele analyse worden uitgevoerd

$$Y_i = b_0 + b_1 X_{1i} + .. + b_k X_{ki} + e_i.$$

Verschillen tussen perioden en constante verschillen tussen winkels

Het bovenstaande model kan worden uitgebreid door constante verschillen tussen eenheden in kaart te brengen middels dummy-variabelen. Nu zijn de verschillende perioden opgenomen in één model om daarnaast ook constante verschillen tussen eenheden te kunnen meten

$$Y_{it} = b_0 + b_{0t}^{cross} + b_{0i}^{tijd} + b_{1t}^{cross} X_{1it} + .. + b_{kt}^{cross} X_{kit} + e_{it}.$$

6.3. Toetsing en aanpassing van het model

Het gecombineerde regressiemodel wordt op andere wijze getoetst dan het STAR-regressiemodel. Het toetsen op en aanpassen van het gecombineerde model op discontinuïteit en autocorrelatie is achterwege gelaten. Immers, bij autocorrelatie kan een bepaald patroon worden gemeten op basis van data van één eenheid. Doel van SuperSTAR is juist afwijkende periode-effecten te meten die voor alle eenheden gelden. Wat discontinuïteit betreft, het

gecombineerde model wordt juist gebruikt om extreme periode-effecten, geldend voor alle eenheden, te schatten.

In het gecombineerde model wordt getoetst of de foutenterm normaal is verdeeld en of er sprake is van heteroscedasticiteit. Hieronder wordt uitleg omtrent deze toetsing en aanpassing gegeven.

Heteroscedasticiteit

Er zijn verschillende vormen van heteroscedasticiteit. STAR gaat uit van een vorm waarbij de grootte van één verklarende variabele proportioneel gerelateerd is aan de variantie van de foutenterm. Deze veronderstelling kan voor SuperSTAR niet worden gehanteerd gezien de keuze van de verklarende variabelen. Een meer voor de hand liggende aanname is een andere vorm van heteroscedasticiteit waarbij de variantie van de foutenterm gecorreleerd is met $E(Y_i)$. De Lagrange Multiplier test toetst op deze vorm van heteroscedasticiteit. De hypothesen zijn

$$\begin{aligned} H_0 &: \text{De variantie van de foutenterm hangt niet van } E(Y_i) \text{ af} \\ H_1 &: \text{De variantie van de foutenterm hangt van } E(Y_i) \text{ af} \end{aligned} .$$

Om deze toets uit te voeren moet een hulp-regressiemodel worden opgesteld die het verband weergeeft tussen de residuen en gefitte waarden van het oorspronkelijke model

$$r_i^2 = \delta_0 + \delta_1 \hat{y}_i^2 + v_i .$$

De R^2 van het hulp-regressiemodel wordt aangeduid middels R_{LM}^2. Onder H_0 heeft $n \times R_{LM}^2$ een χ_1^2 verdeling. De nulhypothese wordt bij significatieniveau α verworpen wanneer $n \times R_{LM}^2 > \chi_{1,1-\alpha}^2$.

Een voorbeeld van een andere mogelijke vorm van heteroscedasticiteit is dat de variantie van de foutenterm per eenheid verschillend is. Er moet nog onderzoek worden gedaan naar de aanwezigheid van deze en andere vormen van heteroscedasticiteit in SuperSTAR.

Correctie van heteroscedasticiteit

Indien de variantie van de foutenterm afhangt van $E(Y_i)$, kan iedere waarneming door $E(Y_i)$ worden gedeeld

$$\frac{Y_i}{E(Y_i)} = b_0 \frac{1}{E(Y_i)} + b_i \frac{X_i}{E(Y_i)} + \frac{e_i}{E(Y_i)} .$$

Aangezien $E(Y_i)$ niet bekend is, wordt in plaats van $E(Y_i)$ $\hat{Y}_i$ gebruikt. De regressiecoëfficiënten van dit model kunnen opnieuw middels Ordinary Least Squares worden bepaald.

Een tweede mogelijkheid voor aanpassing van het model is de te verklaren variabele te transformeren. Middels de Box-Cox methode kan worden onderzocht welke transformatie het best kan worden uitgevoerd.

Normaliteit van de fouten

De toetsen die STAR gebruikt om de fouten op normaliteit te onderzoeken, zijn niet geschikt voor het gecombineerde model. De dataset bij SuperSTAR is namelijk groot, waardoor bij formele methoden zelfs kleine afwijkingen van normaliteit worden ontdekt. Een andere meer flexibele maar grafische manier is het opstellen van een QQ-plot.

Een niet normaal verdeelde foutenterm levert met name problemen op bij het opstellen van de threshold. Daarom wordt in dat geval bij de bepaling van de threshold gebruik gemaakt van de bootstrap methode. Deze methode komt in de volgende paragraaf aan bod.

6.4. De threshold

Verondersteld is dat het totaal van projectiewaarnemingen moet worden getoetst op de aanwezigheid van een materiële fout. Bij een vooraf vastgestelde MP- en R-waarde zijn de hypothesen bij toetsen op juistheid (volledigheid)

$$\begin{aligned} H_0 &: \text{De additieve fout is} \geq (\leq)\ MP\ (-MP) \\ H_1 &: \text{De additieve fout is} < (>) \text{ dan } MP\ (-MP) \end{aligned}.$$

In principe kan dezelfde manier van toetsen worden aangehouden als bij STAR, beschreven in paragraaf 5.3. Indien van tevoren bekend is dat de SOE gelijk is aan l wordt voor iedere projectiewaarneming een threshold opgesteld, behorend bij een SOE van l en een uniforme foutspreiding

$$c_{n,\alpha}(l) = t_{n-k-1,\alpha_l} s_n \sqrt{1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}} + MP/l\,.$$

Bovenstaande nulhypothese wordt dan met kans α onterecht verworpen. Aangezien de SOE onbekend is, is ook onbekend welke threshold moet worden gebruikt. Daarom moet per projectiewaarneming worden onderzocht bij welke SOE het ontdekken van een fout het moeilijkst is. Door gebruik te maken van deze meest conservatieve thresholds, is de kans op het onterecht verwerpen van bovenstaande nulhypothese **maximaal** α.

STAR gebruikt bovenstaande formule voor $c_{n,\alpha}(l)$ ook wanneer $y_t - \hat{y}_t - f_t$ onder de nulhypothese niet normaal is verdeeld terwijl in de formule gebruik wordt gemaakt van de t-verdeling en dus wordt verondersteld dat de foutenterm wel normaal is verdeeld. In het

artikel van Lam en Veall (2001) wordt beschreven dat dit zeer onnauwkeurige thresholds kan veroorzaken en dat, in tegenstelling tot hetgeen dikwijls wordt gedacht, deze onnauwkeurigheid groeit in plaats van vermindert wanneer het aantal waarnemingen groeit. In het artikel wordt de Bootstrap Percentile-methode genoemd als een methode om wel nauwkeurige thresholds te genereren. Daarom is in het geval van een niet-normaal verdeelde foutenterm deze laatste methode gebruikt voor de threshold-bepaling in het gecombineerde model. Bij de bootstrap-methode hoeft de verdeling van de residuen niet van tevoren bekend te zijn, er wordt getrokken uit de residuen zelf.

De k-de bootstrap sample $(y_t^{*k}, \mathbf{x}_t), t = 1, \cdots, n+1, \ldots, n+m$ (met n het aantal basiswaarnemingen, m het aantal projectiewaarnemingen en $\mathbf{x}_t$ de t-de rij van de $\mathbf{X}$-matrix) wordt berekend door bij iedere $\mathbf{x}_t$ een y_t^{*k} te berekenen

$$y_t^{*k} = \mathbf{x}_t\mathbf{b} + \text{trekking uit } n \text{ residuen} \times \left(\frac{n}{n-k-1} \right).$$

Deze trekking uit de n residuen behorend tot de basiswaarnemingen is random en met teruglegging. Vervolgens wordt een bootstrap $\mathbf{b}^{*\mathbf{k}}$ geschat middels Ordinary Least Squares. De voorspellingsfout voor projectiewaarneming w $(w = 1, \ldots, m)$, e_w^{*k}, is

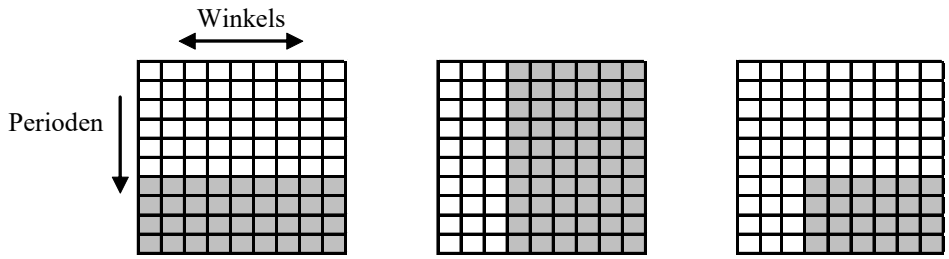
$$e_w^{*k} = y_w^{*k} - \mathbf{x_t}\mathbf{b}^{*\mathbf{k}}.$$

De thresholds worden nu berekend op basis van de formule voor de threshold-bepaling beschreven in paragraaf 5.3., waarbij $t_{n-k-1,\alpha_l} s_n \sqrt{1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}$ wordt vervangen door het α_l-kwantiel van de empirische verdeling van e_w^{*k}. Per projectiewaarneming kan op deze manier voor iedere waarde van de SOE de threshold worden bepaald. De threshold behorend bij een SOE waarvoor geldt dat de foutfractie het moeilijkst te ontdekken is, wordt gebruikt. Bij deze foutfractie is de threshold het laagst.

7. Gebruik Super(STAR)

In dit hoofdstuk wordt beschreven hoe op basis van een dataset, bestaande uit waarnemingen van verschillende eenheden gemeten op verschillende tijdstippen, controles middels STAR of SuperSTAR kunnen worden uitgevoerd.

De te controleren waarnemingen kunnen op drie verschillende manieren onderdeel zijn van de dataset:



Figuur 10: Drie datasets. Een hokje representeert een waarneming. Het grijze gedeelte bevat de ongecontroleerde waarnemingen.

In het eerste geval zijn onvoldoende gegevens bekend van gecontroleerde eenheden in de te controleren periode. Het efficiëntst in dit geval is voor iedere te controleren eenheid een tijdreeksanalyse uit te voeren. In het tweede geval zijn onvoldoende gegevens beschikbaar van gecontroleerde perioden van de te controleren eenheden. Nu kunnen de waarnemingen het meest efficiënt worden gecontroleerd door voor iedere te controleren periode een cross-sectionele analyse uit te voeren. Bovenstaande werkwijze is alleen toepasbaar wanneer in geval één en twee geen periode-, respectievelijk winkeleffecten aanwezig zijn. Indien deze effecten wel aanwezig zijn, is het verstandig een gedeelte van de data eerst op een andere manier te controleren en de data vervolgens hetzelfde te behandelen als in geval drie. Een andere mogelijkheid is ‘doen alsof de ongecontroleerde waarnemingen van winkels of perioden die nodig zijn om een periode- respectievelijk winkeleffect te meten, gecontroleerd zijn’. Daarna kan de data hetzelfde worden behandeld als in geval drie. Nadeel van deze methode is dat wanneer veel fouten in de waarnemingen aanwezig zijn, de geschatte winkel- of periode-effecten erg kunnen afwijken van werkelijke effecten.

In het derde geval zijn voldoende gecontroleerde waarnemingen beschikbaar om zowel eenheid- als periode-effecten te kunnen meten. Als maatstaaf kan worden gehanteerd dat van tenminste dertig eenheden waarnemingen nodig zijn van zowel de basis- als de projectieperiode, en van iedere te controleren eenheid minimaal 24 eerder gecontroleerde waarnemingen. De ongecontroleerde waarnemingen kunnen in dit geval op drie verschillende manieren worden gecontroleerd: door tijdreeksanalyses uit te voeren voor iedere te controleren eenheid, door cross-sectionele analyses uit te voeren voor iedere te controleren periode of door het toepassen van SuperSTAR. Het is in dit geval verstandig de drie verschillende analyses uit te voeren en vervolgens het resultaat te evalueren. Bij deze evaluatie moet worden gelet op effecten die in de STAR-modellen niet zijn opgenomen en in

SuperSTAR wel, en of deze effecten invloed hebben op de controleresultaten. Bij effecten in tegengestelde richting van de controlerichting wordt een fout minder snel ontdekt dan zou moeten: het tegengestelde effect kan een fout corrigeren. Bij effecten in dezelfde richting als de controlerichting kan het effect veroorzaken dat de afwijking slechter lijkt dan eigenlijk het geval is. In het volgende hoofdstuk komt deze situatie bij een beschrijving van een toepassing van SuperSTAR, naar voren.

8. Toepassing SuperSTAR

8.1. Aanpak

Het gecombineerde model is getest op reële data van een bedrijf uit de consumentenbranche. Doel van de test was te onderzoeken wat de voordelen van SuperSTAR zijn ten opzichte van de aparte STAR-modellen. Uitgaande van een correcte dataset, is bij zowel het gecombineerde model als bij de aparte STAR-modellen gekeken naar hoe nauwkeurig de voorspellingen zijn en wat de resultaten zijn van de controle.

De analyses zijn uitgevoerd in een statistisch pakket, R. Allereerst is een lineair model opgesteld waarin alle potentiële verklarende variabelen zijn opgenomen. Dit model is vervolgens getest op en gecorrigeerd voor heteroscedasticiteit. In het gecorrigeerde model is daarna de variabele selectie uitgevoerd waarbij gebruik is gemaakt van een soort branch en bound algoritme. Dit algoritme zoekt op slimme wijze door de ruimte van mogelijke modellen. Tenslotte zijn de voorspellingen en thresholds gegenereerd. Voor de threshold-bepaling is de bootstrap-methode gebruikt.

8.2. De data

Voor de toepassing is een dataset gebruikt met gegevens van 35 winkels. Gedurende de maanden oktober 2000 tot en met maart 2003 zijn van deze winkels de omzet, de kosten van de omzet en de huurkosten gemeten en gecontroleerd. Het gecombineerde model is gebruikt om de omzet van een selectie van vijf winkels in zes perioden te controleren. Dit houdt dus in dat er in totaal 30 x 35 = 1050 waarnemingen zijn waarvan er dertig moeten worden gecontroleerd. (In de praktijk is het natuurlijk efficiënter een groter gedeelte van de data te controleren.) Bij de controle is een MP-waarde gebruikt van 50000 en een R-waarde van 3.

8.3. Het gecombineerde model

Het volgende gecombineerde model is gebruikt

$$Y_{it} = b_0 + b_{0t}^{cross} + b_{1i}^{tijd} X_{1it} + b_{1t}^{cross} X_{1it} + e_{it}.$$

met

b_0 := De intercept

b_{0t}^{cross} := Constante voor het verschil tussen periode t en de intercept

b_{1t}^{cross} := De coëfficiënt geldend voor periode t van de verklarende variabele kosten omzet

b_{1i}^{tijd} := De coëfficiënt geldend voor winkel i van de verklarende variabele kosten omzet

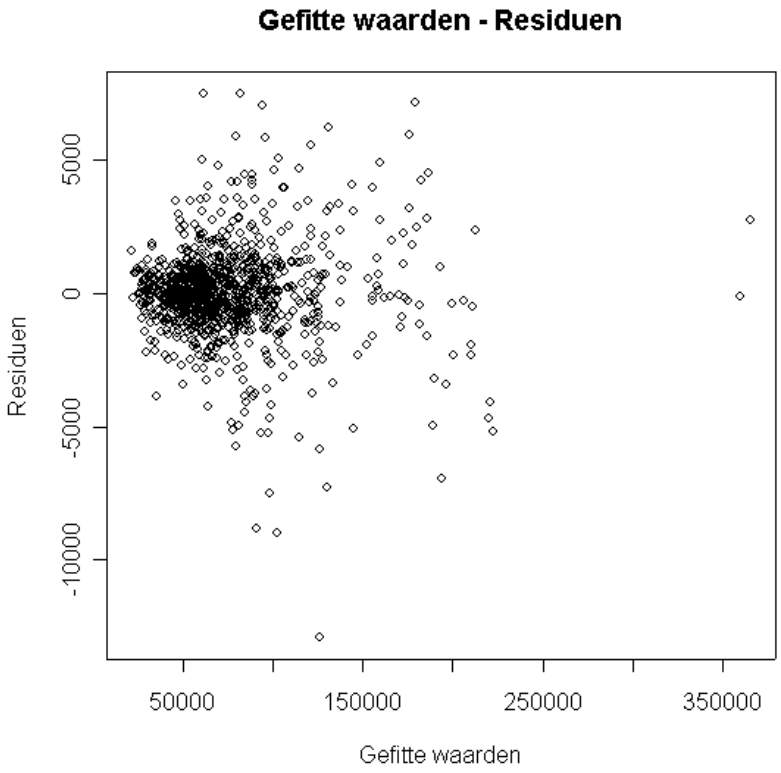
e_{it} := De foutenterm bij winkel i op tijdstip t.

Alleen deze variabelen bleken waardevol te zijn.

Het is mogelijk in eerste instantie meer potentiële verklarende variabelen in het model op te nemen. Het branch en bound algoritme zorgt er immers voor dat alleen de significante variabelen in het model blijven. Wanneer van tevoren al een idee aanwezig is van welke variabelen noodzakelijk zijn en welke minder, is het verstandig niet teveel variabelen toe te voegen. Dit scheelt een hoop rekentijd.

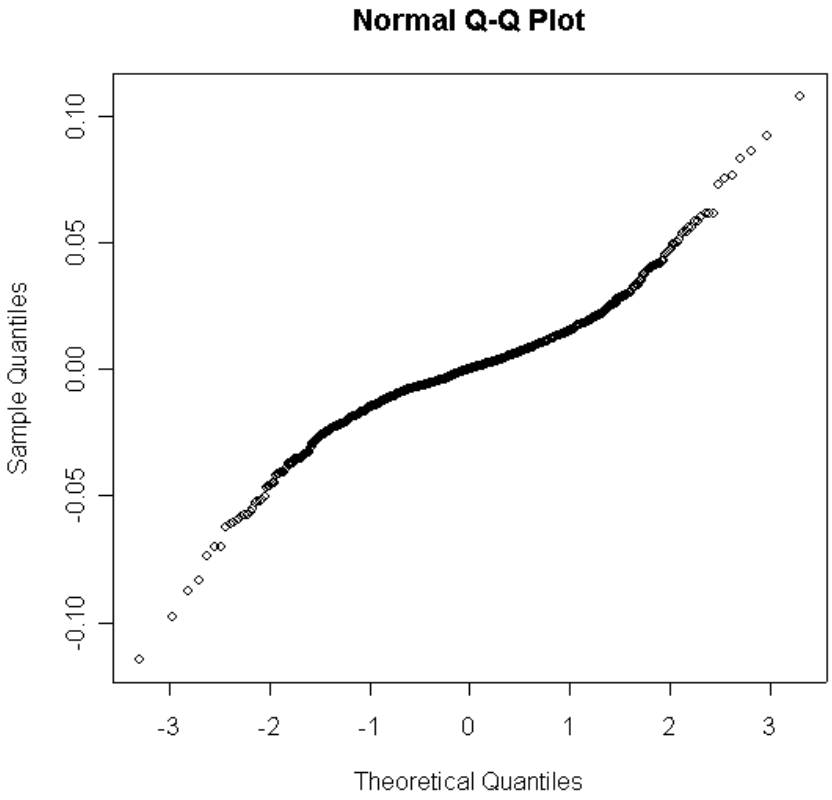
8.4. Toetsen modelaannames

In onderstaande plot zijn de residuen uitgezet tegen de gefitte waarden:



Figuur 11: Scatterplot residuen tegen gefitte waarden

In deze plot is moeilijk te zien of de foutenterm gecorreleerd is met $E(Y_i)$. De Lagrange Multiplier test geeft wel duidelijkheid: de foutenterm hangt af van $E(Y_i)$. Daarom is iedere waarneming in de dataset gedeeld door Y_i en op basis van deze dataset is opnieuw een lineair model opgesteld. Vervolgens is in dit gecorrigeerde model de variabele selectie uitgevoerd. Hieronder wordt van dit nieuwe model de QQ-plot van de residuen weergegeven.



Figuur 12: QQ-plot van de residuen

Deze plot vertoont geen rechte lijn en dus is aangenomen dat de foutenterm niet normaal is verdeeld. Daarom wordt bij de threshold-bepaling gebruik gemaakt van de bootstrap-methode. Er moet rekening worden gehouden met het feit dat voor de afzonderlijke STAR-modellen een zelfde plot geldt. STAR doet echter geen aanpassingen wanneer de foutenterm niet normaal is verdeeld.

8.5. Resultaten

De dertig projectiewaarnemingen zijn op drie verschillende manieren gecontroleerd: er zijn vijf tijdreeksanalyses uitgevoerd (voor iedere winkel één), zes cross-sectionele analyses (voor iedere periode één) en een analyse middels SuperSTAR. Onderstaande tabel geeft de resultaten van de drie analyses weer.

Tabel 1: Resultaten van de drie analyses

			Tijdreeks		Cross-sectie		SuperSTAR	
Winkel	Periode	Y_i	$Y_i - \hat{Y}_i$	Threshold	$Y_i - \hat{Y}_i$	Threshold	$Y_i - \hat{Y}_i$	Threshold
1	1	29696	-1160	-1928	982	-4065	-141	-1571
1	2	40805	3607	-1955	347	-5549	281	-2083
1	3	49677	-2827	-2198	-27	-3136	-1085	-2702
1	4	25558	-387	-1941	-405	-2953	-1751	-1644
1	5	25700	21	-1942	38	-4363	-260	-1582
1	6	26989	-1621	-1930	929	-4663	-389	-1647
2	1	98657	-1873	-4969	1423	-3951	482	-3152
2	2	133977	10694	-5020	245	-5391	-2183	-4255
2	3	188570	-13549	-5731	4992	-7015	-1913	-5075
2	4	99725	-1274	-4969	1281	-2859	-484	-3321
2	5	88877	188	-4977	1016	-4198	-653	-3127
2	6	97433	-2069	-4969	3109	-4485	-1835	-3312
3	1	178513	-4412	-10921	-10270	-5118	-795	-5294
3	2	243617	20502	-10926	-13310	-6680	-6516	-6937
3	3	379297	-21939	-11506	-7073	-10708	8412	-8994
3	4	206774	-979	-10913	-6506	-4608	4673	-6343
3	5	196161	1869	-10913	-6820	-5959	68	-5981
3	6	226231	-7157	-10943	-10397	-6531	1958	-7039
4	1	54479	-3983	-4704	-1928	-3891	-2089	-1988
4	2	91758	7003	-4873	-234	-5311	430	-3020
4	3	130180	-14984	-6002	-1425	-5517	322	-4061
4	4	67910	-1471	-4742	-357	-2785	523	-2364
4	5	65954	-1102	-4729	-911	-4133	-132	-2346

4	6	76566	-4790	-4855	-357	-4435	-1002	-2607
5	1	106606	-4033	-7051	-223	-4020	-533	-3491
5	2	136543	14757	-7135	-5148	-5380	2999	-4282
5	3	202840	-15190	-8963	8504	-7299	6113	-5556
5	4	100540	-988	-7019	1557	-2862	1041	-3511
5	5	96114	3318	-7019	3933	-4226	7534	-3156
5	6	107837	-5879	-7069	299	-4586	-258	-3585

Uit de analyses is gebleken dat zowel aparte tijdreeksanalyses als aparte cross-sectionele analyses in dit geval niet voldoende toereikend zijn de data te beschrijven en de controle uit te voeren.

Voor de tijdreeksanalyses geldt dat alle winkels het in periode drie minder goed doen dan in de andere perioden. In deze periode is een groot negatief verschil tussen realisatie en schatting. De afwijkingen zijn zelfs zo groot dat deze bij iedere winkel de threshold overschrijdt.

Voor de cross-sectionele analyses geldt dat de informatie dat winkel drie het ten opzichte van de rest slecht doet, niet kan worden meegenomen. Bij tijdreeksanalyse is dit effect niet aanwezig aangezien winkel drie apart wordt bekeken. In dit geval is echter bij iedere cross-sectionele analyse een negatief effect voor winkel drie waarneembaar. Ook hier zijn de afwijkingen zo groot dat in vijf van de zes gevallen de verschillen de threshold overschrijden.

Het gecombineerde model sluit beter aan op de data. De gemiddelde gekwadrateerde afwijking is met 90 procent en 70 procent gedaald ten opzichte van de gemiddelde gekwadrateerde afwijking bij tijdreeksanalyse, respectievelijk cross-sectionele analyse. Natuurlijk moet wel worden opgemerkt dat voor het gecombineerde model meer variabelen worden gebruikt, maar het aantal waarnemingen is ook groter. Bij twee waarnemingen overschrijdt het verschil de threshold, maar de verschillen tussen threshold en afwijking zijn in dit geval summier.

Kortom, in dit geval hoeven door gebruik van SuperSTAR minder aanvullende werkzaamheden worden verricht dan men van tevoren bij hantering van de oude STAR-modellen had verwacht. SuperSTAR is ook nuttig bij het omgekeerde effect: wanneer sprake is bij toetsing op volledigheid van positieve effecten kunnen fouten over het hoofd worden gezien wanneer de oude STAR-modellen worden gebruikt. SuperSTAR corrigeert deze positieve effecten en gebruikt zodoende meer informatie bij de controle.

9. Conclusies

Nadeel van apart uitvoeren van tijdreeks- of cross-sectionele analyse in STAR is dat geen rekening kan worden gehouden met periode-, respectievelijk eenheideffecten. Een andere tekortkoming van STAR is dat dit programma uitgaat van het feit dat de foutenterm normaal is verdeeld. Aan deze veronderstelling is in de praktijk vaak niet voldaan.

Doel van de stage was een gecombineerd model op te stellen dat met zowel eenheid- als periode-effecten rekening kan houden en de thresholds bij dit gecombineerde model te bepalen.

Tijdens de stage is een gecombineerd model opgesteld in de vorm van een lineair regressiemodel. Het gecombineerde model moet worden getest op en aangepast voor heteroscedasticiteit. In het gecombineerde model wordt verondersteld dat indien heteroscedasticiteit optreedt, dit komt doordat de variantie van de foutenterm gecorreleerd is met $E(Y_i)$. Nadeel van deze veronderstelling is dat met andere vormen van heteroscedasticiteit geen rekening wordt gehouden. Zo kan het mogelijk zijn dat de variantie van de foutenterm per eenheid verschillend is. Mogelijk kan in de toekomst onderzoek naar toetsing op en aanpassing van het model voor deze en andere aannames worden verricht.

Wanneer ‘voldoende’ data aanwezig is om het gecombineerde model toe te kunnen passen (hoofdstuk 7) is het verstandig de drie verschillende analyses uit te voeren (cross-sectioneel, tijdreeks en gecombineerd) en het resultaat vervolgens te evalueren. Bij deze evaluatie moet worden gelet op effecten die in de STAR-modellen niet zijn opgenomen en in SuperSTAR wel, en of deze effecten invloed hebben op de controleresultaten.

Uit de toepassing van het gecombineerde model (hoofdstuk 8) bleek dat, wanneer periode- en eenheid-effecten aanwezig zijn, mogelijk met het gecombineerde model nauwkeuriger kan worden gecontroleerd. Afhankelijk van de richting van de effecten ten opzichte van de controlerichting, kan dit zelfs veroorzaken dat de hoeveelheid aanvullende werkzaamheden minder is.

Nadeel van het gecombineerde model is dat het niet altijd kan worden gebruikt omdat de ‘juiste’ data niet altijd beschikbaar zijn. De hoeveelheid benodigde data neemt toe ten opzichte van de benodigde hoeveelheid bij de oude STAR-modellen (hoofdstuk 7) en de verklarende variabelen moeten in staat zijn specifieke winkel- en periode-effecten te beschrijven.

Literatuuropgave

BATENBURG VAN, P.C., EMANUELS, J.A., ETTEMA, H.C.J., Afwegingen bij de toepassing van de statistische steekproef in de accountantscontrole, 1998

BOLLE, E.A.W., LENOIR, J.M.H., LOON VAN, J.N.M., Wiskundige Statistiek, Van Loghum Slaterus, 1974

DELOITTE & TOUCHE ENTERPRISE RISK SERVICES, Syllabus Steekproeven in de Accountantscontrole, 1998

DIGGLE, P.J., LIANG, K., ZEGER, S.L., Analysis of Longitudinal Data, Clarendon Press, 1996

EDWARDS, A.L., Multiple Regression and the Analysis of Variance and Covariance, W.H. Freeman and Company, 1979

EVERIT, B.S., An Introduction to Latent Variable Models, Chapman and Hall, 1984

FARAWAY, J.J., Practical Regression and Anova using R, 2002, http://books.pdox.net/Math/Practical%20Regression.pdf

GREENE, W.H., Econometric Research, Prentice Hall, 1993

GUNST DE, M.C.M., Statistische modellen, Vrije Universiteit Amsterdam, 1999

LAM, J.P., VEAL, M.R., Bootstrap Prediction Intervals for Single Period Regression Forecasts, McMaster University Canada, 2001

MENDENHALL, W., SCHEAFFER, R.L., WACKERLY, D.D., Mathematical Statistics with Applications, Duxbury Press, 1996

STEWART, R.S., STRINGER, K.W., Statistical Techniques for Analytical Review in Auditing, John Wiley & Sons, 1985

VAART VAN DER, A.W., Handleiding S-Plus, Vrije Universiteit Amsterdam, 1999

Websites:

Slides Econometrie 1 Universiteit Gent, http://fetew.ugent.be/fineco/econometrie1/files/Slides/4_2%20afwijkingen%20heteroscedasticiteit.pdf

Slides Econometric Methods, Manchester Metropolitan University, http://www.hlss.mmu.ac.uk/eco/Albertson/EMET/EMET07_Heteroscedastic_Errors.pdf

Bijlage 1 De procedure Stepwise

STAR gebruikt de procedure Stepwise voor de selectie van de verklarende variabelen. Deze procedure werkt als volgt.

1. Bereken de correlatiecoëfficiënt R_j tussen Y en elke ingevoerde X_j variabele

$$R_j = \frac{\sum x_{jt} y_t}{\sqrt{\sum x_{jt}^2 \sum y_t^2}} \text{ met } x_{tj} = (X_{tj} - \bar{X}_j) \text{ en } y_t = (Y_t - \bar{Y}).$$

De X_j variabele met de hoogste correlatiecoëfficiënt is kandidaat voor toevoeging.

2. Bereken de volgende F-ratio

$$F = \frac{R_j^2 / 1}{(1 - R_j^2)/(n - k - 1)}.$$

Indien deze F-ratio significant is wordt de kandidaat-variabele aan het model toegevoegd. De waarden van de regressiecoëfficiënten b_j voor de regressiefunctie worden middels Ordinary Least Squares bepaald. In matrixnotatie

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Stop de procedure indien de ratio niet significant is. Er zijn dan geen significante variabelen. Ga naar stap 7.

3. Wanneer alle opgegeven verklarende variabelen zijn opgenomen in het regressiemodel, stop de procedure en ga naar stap 7. Bereken in het andere geval de partiële correlatiecoëfficiënt tussen Y en iedere nog niet toegevoegde X_j variabele na het verwijderen van de invloed van de toegevoegde variabelen.

Indien de residuen resulterend van een regressie tussen Y en de toegevoegde verklarende variabelen $X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$ door e_t worden aangeduid en de residuen resulterend van de regressie tussen X_j en de toegevoegde verklarende variabelen door g_t, dan is de partiële correlatiecoëfficiënt tussen Y en X_j na verwijdering van de invloed van de toegevoegde verklarende variabelen

$$R_{Yj \cdot 12\ldots(j-1)(j+1)\ldots k} = \frac{\sum e_t g_t}{\sqrt{\sum e_t^2 \sum g_t^2}}.$$

De X_j variabele met de hoogste coëfficiënt is kandidaat voor toevoeging.

4. Bereken de volgende F-ratio

$$F = \frac{R^2_{Yj\cdot 12\ldots(j-1)(j+1)\ldots k} / 1}{(1 - R^2_{Yj\cdot 12\ldots(j-1)(j+1)\ldots k}) / (n-k-1)}.$$

Voeg de kandidaat variabele toe indien de F-ratio significant is. Indien de F-ratio niet significant is, stop de procedure en ga naar stap 7.

5. Indien meer dan twee verklarende variabelen in het regressiemodel zijn opgenomen, ga verder met de Backward-procedure in stap 6. Anders, ga verder met de Forward-procedure in stap 3.

6. Bereken voor iedere toegevoegde verklarende variabele X_j in het regressiemodel de partiële correlatiecoëfficiënt na verwijdering van de invloed van de andere variabelen in het regressiemodel. Bereken de F-ratio van de X_j variabele met de laagste coëfficiënt. Indien deze significant is ga naar stap 3. Indien de F-ratio niet significant is, is de variabele overbodig geworden en moet deze worden verwijderd. Ga dan naar stap 5.

7. De Stepwise procedure eindigt indien alle overbodige variabelen zijn verwijderd en niet toegevoegde variabelen insignificant zijn.