Symposium i anvendt statistik 2022-1

SEGES, Landbrug & Fødevarer

Symposium i
Anvendt Statistik

SYMPOSIUM I ANVENDT STATISTIK

2022-1

Redigeret af Peter Linde
på vegne af organisationskomiteen for
Symposium i Anvendt Statistik

Støttet af SAS Institute Inc.

SEGES, Landbrug & Fødevarer Symposium i Anvendt Statistik

Forord

Det er symposiets formål at fremme information om såvel anvendt statistik som statistisk databehandling. Symposiet er tværfagligt med særlig vægt på metodik, formidling og fortolkning af statistiske analyser. I år er SEGES, Landbrug & Fødevarer vært for symposiet, hvilket vi gerne vil takke for. Symposiet arrangeres af Symposium i Anvendt Statistik og SEGES, Landbrug & Fødevarer. Symposium i Anvendt Statistik er ansvarlig for det faglige program og økonomien.

Symposiet skulle oprindeligt være afholdt i januar 2022, men pga. Corona bliver det i stedet afholdt 29.-31. august 2022. Stadig i Axelborg, København. Symposiet har til formål at understøtte delingen af statistiske analyser, og alle forfattere har mulighed for at få publiceret både til januar og til august 2022. Til august vil alle indlæg, der fremlægges blive trykt i symposiebogen 2022-2. Den mindre publikation 2022-1 indeholder indlæggene fra de forfattere, der har brugt muligheden for også at få publiceret i januar 2022. Dette års indlæg spænder over mange forskellige fagområder og lægger derudover vægt på metoder og analyser. Som det er normalt ved videnskabelige indlæg, er bidragsyderne ansvarlige for indholdet af indlæggene, og spørgsmål herom kan rettes direkte til forfatterne.

Med symposiet tilstræbes det at skabe et forum for tværfaglig inspiration og dialog for at udbygge kommunikationen mellem personer, der arbejder med beslægtede metoder inden for forskellige fagområder.

Peter Linde, Organisationskomiteen

ISBN 978-87-989370-1-2

Trykt hos PRinfoTrekroner i 75 eksemplarer

Organisationskomiteen for Symposium i Anvendt Statistik 2022

Lisbeth la Cour
Økonomisk Institut
Copenhagen Business School
Porcelænshaven 16A
2000 Frederiksberg
llc.eco@cbs.dk

Peter Linde
Statistisk konsulent
Granparken 187
2800 Lyngby
Peter@Brede.dk

Anders Milhøj
Økonomisk Institut
Københavns Universitet
Studiestræde 6
1455 København K
Anders.Milhoj@econ.ku.dk

Esben Høg
Matematiske Fag
Aalborg Universitet
Fredrik Bajers Vej 7
9220 Aalborg Ø
esben@math.aau.dk

Gorm Gabrielsen
Institut for Finansiering
Copenhagen Business School
Solbjerg Plads 3
2000 Frederiksberg
stgg@cbs.dk

Sören Möller
Faculty of Health Sciences
Syddansk Universitet
J. B. Winsløws Vej 19
5000 Odense C
Moeller@health.sdu.dk

Helle M. Sommer
SEGES
Landbrug & Fødevarer
Axeltorv
1609 København V
hmso@seges.dk

Niels Kærgaard
Fødevare- og Ressourceøkonomi
Københavns Universitet
Rolighedsvej 25
1958 Frederiksberg
nik@life.ku.dk

Mogens Dilling-Hansen
Institut for Økonomi
Århus Universitet
8000 Århus C
dilling@econ.au.dk

Klaus Rostgaard
Kræftens Bekæmpelse
Strandboulevarden 49
2100 København Ø
klar@cancer.dk

Jørgen Lauridsen
Økonomisk Institut
Syddansk Universitet
Campusvej 55
5230 Odense M
jtl@sam.sdu.dk

Sara Armandi
SAS Institute
Købmagergade 7-9
1050 København K
Sara.Armandi@sdk.sas.com

Birthe Lykke Thomsen
Afdeling for Børn og Unge
Rigshospitalet
Blegdamsvej 9
2100 København Ø

Indholdsfortegnelse

Estimering af input-output koefficienter fra aggregeret data på virksomhedsniveau*

Rasmus Seneberg Zitthen, Arne Henningsen,
Simon Alexander Andreasen, Mads Frandsen
og Mathias Struck Jürgensen

Institut for Fødevare- og Ressourceøkonomi, Københavns Universitet

Januar 2022

Abstrakt: *Virksomheder producerer ofte multiple output, mens inputmængder i regnskabsdata ofte er aggregeret til virksomhedsniveau og andellen af hvert inputmængde som er anvendt i produktionen af hvert output er derfor ukendt. Eftersom mange analyser kræver data på inputmængderne for hvert af de producerede output, i særdeleshed analyser som er anvendt i både driftøkonomisk og politisk beslutningstagen. Dette kapitel opsummerer de væsentlige resultater i Zitthen et al. (2021), som vurderer og sammenligner forskellige metoder til estimering af input-output koefficienter for virksomheder som producerer multiple output, hvilket kan benyttes til at opnå output-specifikke inputmængder. Vi skelner mellem fire økonometriske metoder: (1) klassisk økonometriske metoder, (2) Random Coefficient Regression, (3) Entropy-baseret metoder og (4) Bayesiansk økonometri. Klassisk økonometrisk metode er ofte mangelfuld og forårsager urealistiske og upålidelige estimater. Random Coefficient Regression tager højde for at virksomheder i realiteten ikke har de samme input-output koefficienter. Entropy-baseret metoder og Bayesiansk økonometri kan levere mere pålidelige estimater, eftersom disse metoder tillader mere fleksible modelspecifikationer og kan inkorporere prior viden. Sammenlignet med Entropy-baseret metoder, har Bayesiansk økonometri den fordel, at kunne inkorporere prior viden på en mere transparent måde. Vi konkluderer at en kombination af Random Coefficient Regression og Bayesiansk økonometri ser ud som den mest egnede metode til at estimere input-output koefficienter og output-specifikke inputmængder.*

*Nærværende kapitel er et uddrag af Zitthen et al. (2021), hvori vi præsenterer den grundlæggende økonomiske teori bag estimering af output-specifikke inputallokeringer for landbrugsbedrifter. Herudover præsenterer Zitthen et al. (2021) den eksisterende litteratur på området og diskuterer styrkerne og svaghederne ved de mest brugte estimeringsmetoder.

1 Introduktion

Virksomheder i mange sektorer, herunder landbrugsbedrifter, bruger ofte multiple input til at producere multiple output. Ydermere går samme input ofte igen i produktionen af flere forskellige output. Dette ses f.eks. med et input som gødning, der både bruges i produktionen af byg, hvede, havre m.fl. De aggregerede inputmængder samt outputmængder opgøres i regnsskabsdata, mens de outputspecifikke inputallokeringer ofte udelades (Just et al., 1983). Viden om, hvor meget af et givet input kræves til produktion af et specifik output, er imidlertid essentiel for både udarbejdelsen af troværdige budgetkalkyler til driftøkonomiske beslutningstagen, men også ift. hvordan ændringer i de politiske rammevilkår påvirker erhvervet (Just et al., 1990; Louhichi et al., 2012). Sidstnævnte er især relevant i landbrugssektoren, da denne sektor ofte er yderst reguleret. Den fornødne indsigt kunne givetvis opnås gennem spørgeskemaer, i hvilke man spørger den enkelte landmand, hvilke og hvor store mængder af input, der er brugt til et givet output. Denne metode er dog både dyr og yderst tidskrævende. Herudover er det ej heller sikkert, at det enkelte landbrug har præcise informationer om inputallokeringen (Just et al., 1990). Gennem tiden, er der derfor blevet forslået en række simplere og mere hensigtsmæssige metoder til at finde de outputspecifikke inputallokeringer, hvoraf mange tager deres udgangspunkt i økonometriske estimeringer (f.eks. Errington, 1989; Heckelei et al., 2008; Just et al., 1983; Léon et al., 1999; Louhichi et al., 2012). Ofte benævnes information om input mængder, der bruges i produktionen af en enhed output eller en enheds produktionsaktivitet som henholdsvis input-output koefficienter, inputallokeringskoefficienter eller inputaktivitetskoefficienter (f.eks. Errington, 1989; Gocht, 2008).

2 Økonometriske metoder

Dette afsnit gennemgår økonometriske metoder, der er blevet brugt til at estimere input-output koefficienter og/eller outputspecifikke inputmængder. Vi udelader detaljerede udledninger og beskrivelser af de gennemgående estimeringsmetoder, men præsenterer de generelle ideer om disse metoder og vurderer og diskuterer deres styrker og svagheder.

I de sidste fire årtier er flere forskellige metoder blevet foreslået til at estimere input-output koefficienter og/eller outputspecifikke inputmængder (f.eks. Just et al., 1983; Errington, 1989; Just et al., 1990; Moxey and Tiffin, 1994; Lence and Miller, 1998; Léon et al., 1999; Louhichi et al., 2012; Fragoso and Carvalho, 2013; Lips, 2014). Vi kategoriserer de foreslåede metoder i fire kategorier: klassiske økonometriske metoder, Random Coefficient Regression, Entropi-baserede metoder og Bayesiansk økonometri.

Disse fire kategorier af estimeringsmetoder er kort forklaret, vurderet og diskuteret i henholdsvis afsnit 2.1, 2.2, 2.3 og 2.4.

Vores gennemgang af metoder omfatter ikke metoden til at estimere outputspecifikke inputmængder foreslået af De Loecker et al. (2016), som er baseret på et økonometrisk skøn der tager hensyn til bedrifters observationer som kun producerer et enkelt output. Metoden kan derfor kun anvendes, hvis der for hver af de betragtede output er et tilstrækkeligt stort antal bedrifter, som kun producerer dette output (og ingen andre output), hvilket sjældent er tilfældet i empiriske anvendelser. En anden ulempe ved denne metode er, at den forudsætter at fordelingen af input til output er den samme for alle input,[1] hvilket er en usandsynlig antagelse i mange empiriske anvendelser.[2]

Ikke-økonometriske metoder, især positiv matematisk programmering, kan også bruges til at opnå output-specifikke input-mængder (f.eks. Howitt, 1995), men i denne gennemgang overvejer vi kun økonometriske tilgange.

2.1 Klassiske Økonometriske Metoder

I Zitthen et al. (2021) diskuterer vi, om 'loven om én pris' gælder, dvs. om alle bedrifter står over for de samme priser for de samme input og output, når disse input og output har samme karakteristika på tværs af bedrifterne (dvs. input og output med forskellige karakteristika, f.eks. gødning med forskelligt næringsindhold eller hvede med forskelligt proteinindhold, kan have forskellige priser), og om priser tager højde for kvalitetsforskelle i input og output mellem bedrifter. Vi konkluderer at det ofte er rimeligt at antage, at 'loven om én pris' er opfyldt i vid udstrækning, hvorfor vi kan estimere regressioner i monetære værdier, som kan anses som mængdeindekser, der tager højde for kvalitetsforskelle.

Griliches (1963), Errington (1989), Hallam et al. (1999) m.fl. foreslår, at anvende regressionsmetoden Ordinary Least Squares (OLS) til at estimere f.eks. de samlede omkostninger for hvert input:

$$x_{li} = \sum_{j=1}^{J} a_{lj} y_{ji} + \varepsilon_{li} \; \forall \; l = 1, \ldots, L, \tag{1}$$

[1] Forfatterne er Emir Malikov taknemmelige for at gøre dem opmærksomme på metoden foreslået af De Loecker et al. (2016) samt for at have påpeget denne ulempe.

[2] F.eks. for en bedrift, der producerer korn og husdyr, vil det f.eks. betyde, at procentdelen af dens samlede gødningsinput, der bruges til kornproduktion, er den samme som procentdelen af dens samlede foderstofinput, der bruges til kornproduktion, og som følge heraf, at den procentdel af dens samlede gødningsinput, der bruges til husdyrproduktion, er den samme som den procentdel af dens samlede foderstofindsats, der bruges til husdyrproduktion.

hvor subscript $i = 1, \ldots, N$ angiver bedriften, x_{li} angiver omkostningerne ved input l for bedrift i, a_{lj} angiver den input-output koefficient for input l og output j, y_{ji} angiver omsætningen fra output j for bedrift i, L er det samlede antal af input, J er det samlede antal af output og ε_{li} er tilfældige fejltermer, der antages at være uafhængigt og identisk fordelt (iid).

De vigtigste fordele ved at anvende klassiske økonometriske metoder, såsom OLS, til at estimere input-output koefficienter er, at (1) de er enkle, (2) de kan give konfidensintervaller for input-output koefficienterne og (3) de er velegnet til at teste mere generelle økonometriske antagelser (Midmore, 1990). De to første punkter hænger sammen, og de gør, at man (med lidt baggrundsviden i økonometri) nemt kan opnå estimater af input-output koefficienterne. De opnåede estimater er dog kun gyldige, hvis visse antagelser er opfyldt, f.eks. en passende funktionel form af regressionsligningen. Dette fører os til punkt (3), som angiver, at forskellige standard testprocedurer kan bruges til at teste forskellige antagelser, der er nødvendige for at opnå middelrette og efficiente estimater med OLS og andre klassiske økonometriske metoder.

Anvendelsen af OLS og andre klassiske økonometriske metoder til at estimere inputallokering er imidlertid blevet stærkt kritiseret (f.eks. af Mittelhammer et al., 1981; Just et al., 1983; Midmore, 1990; Lence and Miller, 1998), f.eks. på grund af heteroskedasticitet, høj multikollinearitet, endogenitetsproblemer, ikke-homogenitet på tværs af observationer, potentiel ikke-linearitet og korrelation af fejlled mellem regressionsligninger.

F.eks. regressionligning (1) er ofte plaget af meget høj multikolinearitet, fordi mindre bedrifter ofte har lave omsætninger af alle de output de producerer, mens at større bedrifter ofte har høje omsætninger for de fleste output, de producerer. For at reducere problemet med multikolinearitet skelner Errington (1989) ikke mellem forskellige typer afgrøder, men bruger kun ét aggregeret "afgrøde"-output. Denne tilgang er dog ofte utilstrækkelig, fordi mange efterfølgende analyser kræver input-output koefficienter for individuelle afgrøder. En anden måde at håndtere multikolinearitet på er brugen af større datasæt, men denne mulighed er normalt umulig.

Frekventistiske økonometriske metoder, der forsøger at løse multikollinearitet, såsom Ridge-regression og Lasso-regression, synes at være uegnede til at estimere input-output koefficienter, fordi disse metoder giver estimater, der er biased mod nul. De eneste egnede metoder til at adressere multikollinearitet i estimeringen af input-output koefficienter synes at være entropi-baserede metoder og Bayesianske metoder, som er introduceret i henholdsvis afsnit 2.3 og 2.4.

Ud over multikolinearitet er regressionsligning (1) ofte plaget af væsentlig heteroske-

dasticitet, fordi variationen i inputmængderne eller af omkostningerne ved input stiger med bedriftens produktionsvolumen (Midmore, 1990; Léon et al., 1999). Man kunne adressere heteroskedasticiteten ved at estimere regressionsligningerne med metoden (Feasible) Weighted Least Squares ((F)WLS) regression. Men vi fandt ikke nogen undersøgelse, der bruger denne metode til at estimere input-output koefficienter. En anden metode, der adresserer heteroskedasticitet, er Random Coefficient Regression. Da vi ikke anser denne metode for at være en klassisk økonometrisk metode, diskuterer vi denne metode i afsnit 2.2.

I betragtning af at regressionsligningerne ofte er plaget af betydelig multikollinearitet og heteroskedasticitet, bliver estimaterne ofte meget upræcise. Den høje unøjagtighed medfører, at nogle estimerede input-output koefficienter er alt for høje, og nogle andre estimerede input-output koefficienter er alt for lave, f.eks. negative (Errington, 1989). Fra et praktisk synspunkt er meget upræcise estimater, f.eks. negative koefficienter eller alt for store estimater, ikke realistiske og ugyldiggør derfor fortolkningen af estimaterne. For at løse problemet med at have negative koefficienter kan man pålægge koefficienterne ikke-negativitetsbegrænsninger, når man estimerer modellen, men — som Moxey and Tiffin (1994) viser i en empirisk undersøgelse — dette fører blot til de koefficienter, der er negativ i ubegrænsede estimeringer til at være lig med nul i den begrænsede estimering.

Udover de økonometriske problemer, er der blevet stillet spørgsmålstegn ved estimering af ligning (1) med OLS-metoden, fordi antagelserne om nonjointness mellem produktionen af de forskellige output sandsynligvis er urealistiske i mange empiriske applikationer (Lence and Miller, 1998; Gocht, 2008).

Sammenfattende har klassiske økonometriske metoder vist sig at have mange svagheder (f.eks. Mittelhammer et al., 1981; Just et al., 1983; Errington, 1989; Just et al., 1990; Midmore, 1990; Moxey and Tiffin, 1994; Lence and Miller, 1998; Gocht, 2008).

2.2 Random Coefficient Regression

Modellen nævnt i afsnit 2.1 antager at der er ens input-output koefficienter på tværs af samtlige virksomheder. Den antagelse synes imidlertid relativt urealistisk, da forholdet mellem input og output ofte påvirkes af bedriftsspeciffike forhold, såsom produktionsstørrelse, produktionsgrenene på bedriften, jordtype og -bonitet, vejrforhold m.v. f.eks. i landbrugsbedrifter. Flere studier foreslår derfor at lempe denne antagelse og lade input-output koefficienterne varierer på tværs af bedrifterne (f.eks. Dixon et al., 1984; Hornbaker et al., 1989; Scandizzo, 1990; Dixon and Hornbaker, 1992; Wikström et al., 2011). Specifikt foreslår Dixon et al. (1984), Hornbaker et al. (1989) og Dixon and

Hornbaker (1992) brugen af Random Coefficient Regression (RCR). RCR tillader hver bedrift at have deres eget sæt af input-output koefficienter, som individuelt kan afhænge af de førnævnte observerede faktorer, såvel som ikke-observerede faktorer. Brugen af RCR kan implementeres i mange forskellige regressionsmodeller, herunder også dem nævnt i Zitthen et al. (2021). I det følgende beskrives RCR med udgangspunkt i ligning (1). Da regressionens underlæggende funktioner er identiteter[3], er det inkonsekvent at inkludere fejleddet ε_{li}, og fejleddet udelades ofte i estimeringen af RCR. For at tillade, at input-output koefficienterne a_{lj} varierer mellem bedrifterne tilføjes subscript i. Derved opnår vi følgende ligning:

$$x_{li} = \sum_{j=1}^{J} a_{lji} y_{ji} \; \forall \; l = 1, \ldots, L, \tag{2}$$

hvor input-output koefficienterne er estimeret som:

$$a_{lji} = B_{lj0} + \sum_{m=1}^{M} B_{ljm} \eta_{mi} + \kappa_{lji} \; \forall \; l = 1, \ldots, L, j = 1, \ldots, J, \tag{3}$$

hvor M er antallet af variable, der forklarer input-output koefficienterne, η_{mi} indikerer den m'de variabel, er forklarer input-output koefficienterne, B_{lj0} og B_{ljm} er koefficienterne, som skal estimeres og κ_{lji} er fejlledet, der fanger effekten af ikke-observerede variable på input-output koefficienterne. I ligning (3), kan vi se at samtlige input-output koefficienter afhænger af præcis de samme forklarende variable. Det er imidlertid muligt at bruge forskellige sæt af forklarende variable i estimeringen af de individuelle input-output koefficienter:

$$a_{lji} = B_{lj0} + \sum_{m \in \mathscr{M}_{lj}} B_{ljm} \eta_{mi} + \kappa_{lji} \; \forall \; l = 1, \ldots, L, j = 1, \ldots, J, \tag{4}$$

hvor $\mathscr{M}_{lj}$ indikerer sæt af variabler, som input-output koefficienten a_{lj} afhænger af.

Indsætter vi ligning (4) i ligning (2), får vi:

$$x_{li} = \sum_{j=1}^{J} \left(B_{lj0} + \sum_{m \in \mathscr{M}_{lj}} B_{ljm} \eta_{mi} + \kappa_{lji} \right) y_{ji} \; \forall \; l = 1, \ldots, L \tag{5}$$

$$= \sum_{j=1}^{J} B_{lj0} y_{ji} + \sum_{j=1}^{J} \sum_{m \in \mathscr{M}_{lj}} B_{ljm} \eta_{mi} y_{ji} + \sum_{j=1}^{J} \kappa_{lji} y_{ji} \; \forall \; l = 1, \ldots, L. \tag{6}$$

[3] Se Zitthen et al. (2021), for yderligere uddybning.

Foruden at lempe den ofte urealistiske antagelse om ens input-output koefficienter på tværs af bedrifter, har RCR også den store fordel, at metoden i høj grad også løser de OLS og andre klassiske økonometriske metoders problemer ift. heteroskedasticitet, fordi variansen af det sammesatte fejled, $\sum_{j=1}^{J} \kappa_{lji} y_{ji}$ kan samvarierer med produktionsomfanget y_{ji}. RCR lider dog under, at metoden øger problemerne med multikolinaritet, da de mange interaktionsled som bliver brugt som forklarende variable ($\eta_{mi} y_{ji}$) er korreleret med hinanden og med de 'enkelte' forklarende variable (y_{ji}).

2.3 Entropi-baserede Metoder

I jagten på mere plausible og pålidelige resultater på trods af multikollinaritet foreslås flere alternativer til de klassiske økonometriske metoder, herunder både entropi-baserede og bayesianske metoder (se afsnit 2.4). Idéen er, at inkorporere såkaldt *non-data information* (ofte kaldet prior information). Hermed forstås altså viden, som undersøgeren måtte have om emnet, inden studiet startes. Før har de entropi-baserede metoder været set som en selvstændig metodisk ramme, men det er senere vist, at de entropi-baserede også falder inden for den bayesianske økonometri. Metoderne er anseeligt mere komplekse end de klassiske metoder (OLS m.fl.), men har fået indpas, da de generelt giver væsentligt mere pålidelige resultater (f.eks. Moxey and Tiffin, 1994; Lence and Miller, 1998; Léon et al., 1999; Fragoso and Carvalho, 2013; Louhichi et al., 2018). I Zitthen et al. (2021) tager vi udgangspunkt i *Generalised Maximum Entropy (GME)*. Metoden blev først introduceret af Shannon (1948), og starter med de ikke-observerede sandsynligheder, $\rho = [\rho_1, \rho_2, \ldots, \rho_K]$:

$$H(\rho) = -\sum_{k=1}^{K} \rho_k \ln(\rho_k), \tag{7}$$

hvor $\rho_k \ln(\rho_k) \equiv 0$ for $\rho_k = 0$ og K er antallet af datapunkter, ofte også kaldet støttepunkter (Léon et al., 1999). Mere specifikt angiver, ρ_k sandsynligheden for at observere datapunkt k (Fragoso and Carvalho, 2013). Givet dette, når $H(\rho)$ sit maksimum når $\rho_1 = \rho_2 = \ldots = \rho_k = \frac{1}{K}$, hvilket er ækvivalent til, at sandsynlighederne er uniformt fordelt. Målet er nu, at maksimere $H(p)$, i hvilken der (hvis relevant) kan implementeres yderligere restriktioner. Idéen er derved, at de estimerede sandsynligheder for ρ_k vil trækkes mod en uniform fordeling under bibetingelse af de valgte restriktioner. Disse restriktioner vil her være en kombination af pågældende data samt undersøgerens prior information (Léon et al., 1999). I Zitthen et al. (2021) går vi yderligere i dybden med GME, samt alternativet Generalised Cross Entropy (GCE). Vi finder, at GME har mange fordele,

men også sine svagheder. Blandt fordelene er, at (1) der ikke kræves nogen antagelser ift. fejlleddets fordeling, (2) metoden løser udfordringen med multikolinaritet, (3) metoden kan benyttes på mindre datasæt og (4) metoden tillader ikke-lineære restriktioner og ulighedsrestriktioner (Louhichi et al., 2012; Fragoso and Carvalho, 2013). Den største udfordring ved GME fremkommer ved måden, hvorpå støttepunkterne er konstrueret. Dette fordi, at støttepunkterne har stor betydning for entropi-modellens resultater. Der kan tages højde for de nævnte udfordring ved brug af prior information (Louhichi et al., 2012) eller i stedet ved bruge af GCE (Lence and Miller, 1998). Ved brugen af prior information skal undersøgeren dog stadig bestemme såkaldte *støttebånd* og bredden herpå, hvilket igen kan have stor betydning for de endelige resultater.

2.4 Bayesiansk Økonometri

For at imødekomme udfordringerne fundet ved entropi-baserede metoder, undersøger vi i Zitthen et al. (2021) også brugen af bayesiansk økonometri til bestemmelse af input-output koefficienter. Herud over undersøges fordele og ulemper ved den Bayesianske metode. Den Bayesianske metode er baseret på ideen om at kombinere en prior forventet fordeling af de ukendte α-parametre med informationen givet af data baseret på Bayes' teorem (Moxey and Tiffin, 1994; Coelli et al., 2005, p. 231-234; Heckelei et al., 2008). I analysen behandler vi alle parametre i estimeringen som stokastiske variable (Gocht, 2008; Heckelei et al., 2008). Ydermere skelner vi mellem prior density for koefficienterne i form af en prior probabilty density function (pdf) angivet ved $p(\alpha)$, stikprøve informationen i form af en likelihood function[4] baseret på dataet, $L(\alpha, y)$, og posterior pdf, $h(\alpha \mid y)$. På baggrund af Bayes' teorem finder vi, at posterior pdf er proportional til prior pdf multipliceret med likelihood funktionen (Geweke, 1999; Coelli et al., 2005, p. 231-234; Gocht, 2008):

$$h(\alpha \mid y) \propto L(\alpha, y) \cdot p(\alpha). \tag{8}$$

Prior fordelingen kan enten specificeres som informativ eller som ikke-informativ prior. Som navnet indikerer bidrager en ikke-informativ prior ikke med forudgående viden om parametrene. I dette tilfælde er det kun data (og ikke den forudgående viden) der påvirker parameterestimaterne. I modsætning til indragelse af informativ prior, som bidrager med forudgående information om parameterestimaterne. En informativ prior er ofte en forudgående fordeling omkring en forventet middelværdi og en passende varians.

[4] I Heckelei et al. (2008) og Louhichi et al. (2018) er likelihood funktionen angivet som $L(\alpha \mid y)$, hvorimod Geweke (1999) og Gocht (2008) anvender $L(y \mid \alpha)$.

Oplysninger om en prior fordeling (f.eks. hvilken type af fordeling, middelværdi, varians) kan komme fra faglitteratur, undersøgelser eller tidligere studier (Louhichi et al., 2018).

I en empirisk undersøgelse fandt Moxey and Tiffin (1994), at den Bayesianske metode var meget god sammenlignet med den inequality-constrained least-squares (ICLS) metode. Moxey and Tiffin (1994) fremmer brugen af Bayesianske priors, når de arbejder med begrænsningsestimationer. Udover dette er det velkendt, at GME/GCE og den Bayesianske metode ligner hinanden meget og indeholder mange af de samme fordele såvel som ulemper (Heckelei et al., 2008; Louhichi et al., 2012). Nogle yderligere fordele ved den Bayesianske metode er blevet påpeget af Heckelei et al. (2008). For det første kan den Bayesianske metode angives således, at den svarer til GME/GCE-metoden[5], for det andet er de implementerede priors mere transparente, fordi prior pdf'er tildeles direkte til de ukendte. Endelig har denne metode færre variabler og kan derfor være mindre krævende i form af regnekraft (Heckelei et al., 2008).

I betragtning af disse fordele ser den Bayesianske metode ud til at være meget attraktiv til at estimere input-output koefficienter. Men da litteraturen kun indeholder en begrænset mængde studier, der bruger Bayesiansk økonometri på dette område, skal der udføres yderligere forskning i, hvordan man gør dette i praksis.

2.5 Prior

Som sagt kan oplysningerne om prior information komme fra faglitteratur, undersøgelser eller tidligere studier (Louhichi et al., 2018). For en empirisk anvendelse for danske landbrugsbedrifter forslår vi, at prior fordelinger af input-output koefficienterne specificeres således, at den forventede værdi er et vægtet gennemsnit af tidligere års estimerede input-output koefficienter:

$$a^*_{lj} \sim \Psi\left(\mu_{lj}, \sigma^2_{lj}\right) \tag{9}$$

$$\mu_{lj} = \sum_{t=1}^{T} \tau_t \hat{a}^*_{lj|t}, \tag{10}$$

hvor a^*_{lj} angiver input-output koefficienten for input l og output j for det gældende år, $a^*_{lj|t}$ er input-output koefficienten for input l og output j estimeret for t år siden, τ_t er et vægt for input-output koefficienten estimeret for t år siden, T er antallet af år som er anvendt til at beregne det vægtede gennemsnit, og $\Psi\left(\mu_{lj}, \sigma^2_{lj}\right)$ er en passende fordeling med en forventet værdig μ_{lj} og variansen σ^2_{lj}. Vi specificerer det vægtede gennemsnit

[5]Dette er også tilfældet ved brugen af faste input (Gocht, 2008).

således, at de seneste år vægtes højere end år længere tilbage:

$$\tau_t = \tau_1 (1-r)^{t-1} \ \forall \, t = 2, \ldots, T \tag{11}$$

$$\tau_1 = \left(\sum_{t=1}^{T-1} (1-r)^{t-1} \right)^{-1}, \tag{12}$$

hvor $r \in [0,1]$ er den relative sats, hvormed vægten falder, når man går et år tilbage i tiden, og ligning (12) normaliserer vægtene, så de summerer op til én. Vi tror, at et årligt fald i vægten med 10%, dvs. $r = 0,1$, er tilstrækkelig lille til at lade det vægtede gennemsnit være robust over for ændringer fra år til år og samtidig være tilstrækkeligt stort til at tage højde for tendenser over tid.

Vi beregner variansen af prior fordelingerne baseret på vægtede gennemsnits værdier af de kvadrerede residualer fra middelværdierne μ_{lj}:

$$\sigma_{lj}^2 = \psi_{lj} \sum_{t=1}^{T} \tau_t \left(\hat{a}_{lj|t}^* - \mu_{lj} \right)^2, \tag{13}$$

hvor ψ_{lj} er en tuning parameter der specificerer, hvor informativ prioren skal være, hvor en lille værdi angiver en meget informativ prior og en meget stor værdi indikerer en næsten ikke-informativ prior.

Vi forslår at bruge en log-normal fordeling til prior fordelingerne for input-output koefficienterne $\Psi\left(\mu_{lj}, \sigma_{lj}\right)$. En fordel ved log-normal fordelingen er at det udelukker negative estimater af input-output koefficienter. For at opnå en log-normal fordeling med forventet værdi μ_{lj} og varians σ_{lj}^2 sættes placeringsparameteren og skalaparameteren for log-normal fordelingen til:

$$\tilde{\mu}_{lj} = \ln \left(\frac{\mu_{lj}^2}{\sqrt{\mu_{lj}^2 + \sigma_{lj}^2}} \right) \tag{14}$$

$$\tilde{\sigma}_{lj}^2 = \ln \left(1 + \frac{\sigma_{lj}^2}{\mu_{lj}^2} \right), \tag{15}$$

således at:

$$\ln a_{lj}^* \sim \mathcal{N}\left(\tilde{\mu}_{lj}, \tilde{\sigma}_{lj}^2 \right), \tag{16}$$

hvor $\mathcal{N}\left(\tilde{\mu}_{lj}, \tilde{\sigma}_{lj}^2 \right)$ indikerer en normal fordeling med middel værdi $\tilde{\mu}_{lj}$ og variancen $\tilde{\sigma}_{lj}^2$.

3 Afsluttende bemærkninger

De økonometriske metoder som anvendes til estimering af input-output koefficienter og output-specifikke input mængder kan kategoriseres i: (1) klassisk økonometrisk metoder, (2) random coefficient regression, (3) entropi-basseret metoder, og (4) Bayesiansk økonometri.

Klassiske økonometriske metoder er meget restriktive og giver ofte upålidelige og usandsynlige estimater, som er forårsaget af multikollinearitet og andre statistiske og økonometriske problemer.

Random coefficient regression metoder har den enorme fordel, at de slækker på den restriktive antagelse om ens input-output koefficienter på tværs af bedrifter, som kræves af andre økonometriske estimeringsmetoder. Men givet de yderligere koefficienter i en random coefficient regression med observerede faktorer, der påvirker input-output koefficienterne, kan problemer med multikollinearitet blive endnu værre sammenlignet med klassiske økonometriske metoder.

Entropi-baserede og Bayesianske metoder har mange ligheder, og det blev vist, at entropi-baserede metoder er et særtilfælde af Bayesianske metoder. Begge typer metoder kan bruges til at løse ulemperne ved de klassiske økonometriske metoder, især ved at bruge forudgående viden til at løse problemer forårsaget af multikollinearitet. Den forudgående viden kan for eksempel komme fra faglitteraturen, eksisterende forskning eller undersøgelser. I tilfælde af årlige analyser kan priors være baseret på resultaterne opnået i de foregående år. Sammenlignet med entropi-baserede tilgange har Bayesianske tilgange den fordel, at de tager højde for priors på en mere gennemsigtig måde. Hverken eksisterende entropi-baserede metoder eller eksisterende Bayesianske metoder til at estimere input-output koefficienter slækker dog på den restriktive antagelse om lige input-output koefficienter på tværs af virksomheder.

Vi anbefaler på den baggrund, at estimere en random coefficient model med en Bayesiansk metode, da dette vil tillade for forskelle input-output koefficienter på tværs af bedrifter og for at bruge prior information for at opnå pålidelige og plausible estimater på trods af høj multikollinearitet og et stort antal koefficienter at estimere.

Litteratur

Coelli, T. J., Prasada Rao, D., O'Donnell, C. J., and Battese, G. E. (2005). *Econometric Estimation of Production Technologies*. Springer.

De Loecker, J., Goldberg, P. K., Khandelwal, A. K., and Pavcnik, N. (2016). Prices, markups, and trade reform. Econometrica, 84(2):445–510.

Dixon, B. L., Batte, M. T., and Sonka, S. T. (1984). Random coefficients estimation of average total product costs for multiproduct firms. Journal of Business & Economic Statistics, 2(4):360–366.

Dixon, B. L. and Hornbaker, R. H. (1992). Estimating the technology coefficients in linear programming models. American Journal of Agricultural Economics, 74(4):1029–1039. Errington, A. (1989). Estimating enterprise input-output coefficients from regional farm data. Journal of Agricultural Economics, 40:52–56.

Fragoso, R. and Carvalho, M. L. d. S. (2013). Estimation of cost allocation coefficients at the farm level using an entropy approach. Journal of Applied Statistics, 40(9):1893–1906.

Geweke, J. (1999). Using simulation methods for bayesian econometric models: inference, development, and communication. Econometric reviews, 18(1):1–73.

Gocht, A. (2008). Estimating input allocation for farm supply models. Contributed paper presented at the 107th EAAE Seminar “Modelling of Agricultural and Rural Development Policies” in Sevilla, Spain, January 29th – February 1st, 2008.

Griliches, Z. (1963). Estimates of the aggregate agricultural production function from cross-sectional data. Journal of Farm Economics, 45(2):419–428.

Hallam, D., Bailey, A., Jones, P., and Errington, A. (1999). Estimating input use and production costs from farm survey panel data. Journal of Agricultural Economics, 50(3):440–449.

Heckelei, T., Mittelhammer, R. C., and Jansson, T. (2008). A bayesian alternative to generalized cross entropy solutions for underdetermined econometric models. Technical report.

Hornbaker, R. H., Dixon, B. L., and Sonka, S. T. (1989). Estimating production activity costs for multioutput firms with a random coefficient regression model. American Journal of Agricultural Economics, 71(1):167–177.

Howitt, R. E. (1995). Positive mathematical programming. American Journal of Agricultural Economics, 77(2):329–342.

Just, R. E., Zilberman, D., and Hochman, E. (1983). Estimation of multicrop production functions. American Journal of Agricultural Economics, 65(4):770–780.

Just, R. E., Zilberman, D., Hochman, E., and Bar-Shira, Z. (1990). Input allocation in multicrop systems. American Journal of Agricultural Economics, 72(1):200–209.

Lence, S. H. and Miller, D. J. (1998). Estimation of multi-output production functions with incomplete data: A generalised maximum entropy approach. European Review of Agricultural Economics, 25(2):188–209.

Leon, Y., Peeters, L., Quinqu, M., and Surry, Y. (1999). The use of maximum entropy to stimate input-output coefficients from regional farm accounting data. Journal of Agricultural conomics, 50(3):425–439.

Lips, M. (2014). Calculating full costs for Swiss dairy farms in the mountain region using a maximum entropy approach for joint-cost allocation. International Journal of Agricultural Management, 3(3).

Louhichi, K., Espinosa, M., Ciaian, P., Perni, A., Ahmadi, B. V., Colen, L., and y Paloma, S. G. (2018). The eu-wide individual farm model for common agricultural policy analysis (ifm-cap v. 1): Economic impacts of cap greening. Technical report.

Louhichi, K., Jacquet, F., and Butault, J. P. (2012). Estimating input allocation from heterogeneous data sources: A comparison of alternative estimation approaches. Agricultural Economics Review, 13(389-2016-23472):83–102.

Midmore, P. (1990). Estimating input-output coefficients from regional farm data – a comment. Journal of Agricultural Economics, 41(1):108–111.

Mittelhammer, R. C., Matulich, S. C., and Bushaw, D. (1981). On implicit forms of multiproduct-multifactor production functions. American Journal of Agricultural Economics, 63(1):164–168.

Moxey, A. and Tiffin, R. (1994). Estimating linear production coefficients from farm business survey data: A note. Journal of Agricultural Economics, 45(3):381–385. Scandizzo, P. L. (1990). The estimation of input-output coefficients: Methods and problems. Ricerche Economiche, XLIV(4):455–474.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3):379–423.

Wikstrom, D., Peeters, L., and Surry, Y. R. (2011). Semiparametric cost allocation estimation. Paper presented at the XIIth International Congress of the European Association of Agricultural Economists (EAAE) in Zurich, Switzerland. http://purl.umn.edu/115742.

Zitthen, R. S., Henningsen, A., Andreasen, S. A., Frandsen, M., and Jurgensen, M. S. (2021). Estimating input allocation coefficients from aggregate firm-level data: A review of various econometric methods. Department of Food and Resource Economics: Unpublished paper.

Classification using binary and continuous variables

Guillermina Eslava and Gonzalo Pérez.
Faculty of Sciences, National Autonomous University of Mexico

Abstract

This work presents comparative results of some classification methods for the case of two populations and a set of binary and continuous variables. Methods used include linear and non-linear discriminant analysis, logistic discrimination, discriminant analysis based on conditional Gaussian distributions with a tree graph structure, and random forests. These methods are compared in terms of classification error rates on simulated data.

Key words. Conditional Gaussian distribution, Homogeneous mixed graphical models, Linear and non-linear discriminants, Logistic discrimination, Misclassification error, Mixed binary and continuous variables, Supervised classification, Random forests.

1 Introduction

Classification of labeled observations based on a set of measurements is a problem that has been developed within statistics in what is known as supervised classification. There are various methods that have been known for some time, like discriminant analysis in its different forms, parametric, non parametric, linear and non-linear, and logistic discrimination. More sophisticated techniques, such as probabilistic graphical models both with undirected or with directed graphs, corresponding to Markov and Bayesian networks, are parametric models which can also be used for classification. More recently, a surge of algorithmic methods has become popular for data-driven approach to classification, such as random forests, neural networks and more recently deep neural networks.

When dealing with numerical covariates, most classification methods apply. On the other hand, when all covariates are categorical, not all classification methods apply directly, as is the case of linear and quadratic discriminant analysis although they are often nonetheless successfully used. For a set of measurements which consist of both categorical and continuous variables, some methods become more difficult to apply, in particular those based on probabilistic graphical models. Although the theory for graphical models for mixed variables has been successfully developed, see e.g. Lauritzen (1996, Ch. 6), and some algorithms for model identification and model estimation exist, the use of these models in practical applications is still somewhat limited due to scarce availability of software. Note however, that some software has been available for some time, see e.g. Højsgaard et al. (2012, Ch. 5).

In this note we present a small simulation study where we consider classification between two groups on the basis of four binary and six continuous variables. We consider simulated datasets generated from a homogeneous conditional Gaussian distribution, with two independent paths as graph structure, one for the binary and one for the continuous variables. We apply three linear methods: linear discriminant analysis, naive discriminant analysis, and linear logistic regression. We also consider six non–linear classification methods: three modified discriminant functions, two logistic regressions with interaction terms, and random forests. We briefly present each method, and then provide the respective numerical results. This simulation study is part of ongoing research, where we explore other simulation settings and real data.

2 Classification methods

We consider the problem of classification between two well defined classes of individuals, Π_1 and Π_2, on the basis of p variables measured on a sample of individuals from each class. Let $C \in \{1, 2\}$ be the class variable and $\boldsymbol{x} = (x_1, \ldots, x_p)$ the random vector of p variables. Let $\pi_1 = P(C = 1)$ and $\pi_2 = P(C = 1)$ be the prior probabilities that an observed individual belongs to class Π_1 and Π_2, and $P(C = 1|\boldsymbol{x})$ and $P(C = 2|\boldsymbol{x})$ be the posterior probabilities, respectively.

2.1 Bayes rule

The Bayes classification rule is to choose the class with the higher posterior probability. That is assign an observation to Π_1 if

$$P(C = 1|\boldsymbol{x}) > P(C = 2|\boldsymbol{x}). \tag{1}$$

If one assumes that $\boldsymbol{x}$ has a density or probability function $f_c(\boldsymbol{x}|c) = f_c(\boldsymbol{x})$ in population c, $c = 1, 2$, the Bayes rule (1) is equivalent to assign an observation to Π_1 if

$$\log \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} - \log \frac{\pi_2}{\pi_1} > 0. \tag{2}$$

The rule (1) is optimal in the sense that minimizes the error rate or probability of misclassification $P(e)$ defined as

$$P(e) = \pi_1 P(2|1) + \pi_2 P(1|2), \tag{3}$$

where $P(i|j)$ denotes the probability of assigning an observation from population Π_j to Π_i.

The vector $\boldsymbol{x} = (x_1, \ldots, x_p)$ may contain only continuous, only discrete, or a mixture of both types of variables.

2.2 Discrimination for normal populations

If one assumes that $\boldsymbol{x}$ has a multivariate Gaussian density $f_c(\boldsymbol{x})$ with mean $\boldsymbol{\mu}_c$ and covariance $\boldsymbol{\Sigma}_c$ in population c, $c = 1, 2$, the left side of the equation (2) is the quadratic

discriminant function given by

$$\frac{1}{2}\log\frac{\mid \boldsymbol{\Sigma}_1^{-1} \mid}{\mid \boldsymbol{\Sigma}_2^{-1} \mid} + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - \log\frac{\pi_2}{\pi_1}. \quad (4)$$

Assuming $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, expresion (4) reduces to the following linear discriminant function:

$$\boldsymbol{x}^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \log\frac{\pi_2}{\pi_1}. \quad (5)$$

The naive classifier is obtained by assuming a Gaussian distribution on each population where $\boldsymbol{\Sigma}_c$ is a diagonal matrix with elements $\{\sigma_{1c}^2, ..., \sigma_{pc}^2\}$. The derived function obtained from (4) becomes:

$$\sum_{i=1}^{p}\left[\frac{1}{2}\log\frac{\sigma_{i2}^2}{\sigma_{i1}^2} - \frac{1}{2}\frac{(x_i-\mu_{i1})^2}{\sigma_{i1}^2} + \frac{1}{2}\frac{(x_i-\mu_{i2})^2}{\sigma_{i2}^2}\right] - \log\frac{\pi_2}{\pi_1}. \quad (6)$$

2.3 Logistic discrimination

Logistic regression can be used for classification when variables are both binary and continuous. It provides the posterior probabilities which are needed to use the Bayes discriminant rule (1), namely by specifying

$$P(C=1|\boldsymbol{x}) = \exp(\beta_0+\beta_1x_1+...+\beta_px_p)P(C=2|\boldsymbol{x}), \quad (7)$$
$$P(C=2|\boldsymbol{x}) = 1/[1+\exp(\beta_0+\beta_1x_1+...+\beta_px_p)]. \quad (8)$$

Logistic regression with interaction terms can similarly be used to compute the posterior probabilities.

2.4 Mixed graphical models

When both binary and continuous variables are present, we may consider specifying two conditionally independent graphical models, one for each of the two classes, resulting in a single classifier. Let $\Delta = \{i_1, \ldots, i_q\}$ be the set of q binary variables and $\Gamma = \{y_1, \ldots, y_r\}$ the set of r continuous variables. The set of $p = q + r$ variables $\boldsymbol{x} = (\boldsymbol{i}, \boldsymbol{y}) = (i_1, \ldots, i_q, y_1, \ldots, y_r)$ follows a conditional Gaussian density that satisfies the Markov properties with respect to an undirected marked graph $G_c = (V, E_c)$, with $V = \Delta \cup \Gamma$. The density function is

$$f_c(\boldsymbol{x}) = p_c(\boldsymbol{i})f(\boldsymbol{y}|\boldsymbol{i}, c) \quad (9)$$

where

$$p_c(\boldsymbol{i}) = P(\boldsymbol{x}_\Delta = \boldsymbol{i}|C=c) > 0 \text{ and } f(\boldsymbol{y}|\boldsymbol{i}, c) = N_{|\Gamma|}\Big(\boldsymbol{\mu}_c(\boldsymbol{i}), \boldsymbol{\Sigma}_c(\boldsymbol{i})\Big). \quad (10)$$

The Markov properties with respect to $G_c = (V, E_c)$ impose some restrictions on $p_c(\boldsymbol{i})$, $\boldsymbol{\mu}_c(\boldsymbol{i})$ and $\boldsymbol{\Sigma}_c(\boldsymbol{i})$. The density functions in (9), for populations $c = 1$ and $c = 2$, can be used in the Bayes rule given in (2). In practice, when using a conditional Gaussian density, the graph structure $G_c = (V, E_c)$ should be identified or estimated, and the parameters

$p_c(\boldsymbol{i})$, $\boldsymbol{\mu}_c(\boldsymbol{i})$ and $\boldsymbol{\Sigma}_c(\boldsymbol{i})$ should be estimated.

Here, we restrict the identification of the graph structure $G_c = (V, E_c)$ to be a decomposable graph, specifically a tree structure. For this structure, the estimation of the density is simple, considering that $p_c(\boldsymbol{i})$, $\boldsymbol{\mu}_c(\boldsymbol{i})$ and $\boldsymbol{\Sigma}_c(\boldsymbol{i})$, $c : 1, 2$, have exact maximum likelihood expressions.

2.5 Naive Classifier

The naive classifier is obtained by assuming independence among the p variables $x_1, \ldots, x_p$ in each population. This classifier can be obtained considering a mixed graphical model in each population with the empty graph as the interaction graph, see for example Figure 1b), which corresponds to the empty graph for four binary and six continuous variables. Considering $\boldsymbol{x} = (\boldsymbol{i}, \boldsymbol{y}) = (i_1, \ldots, i_q, y_1, \ldots, y_r)$, the left side of the equation (2) is given by

$$\sum_{j=1}^{q} \left[\log \frac{p_{j1}^{i_j}(1-p_{j1})^{1-i_j}}{p_{j2}^{i_j}(1-p_{j2})^{1-i_j}} \right] + \sum_{j=1}^{r} \left[\frac{1}{2} \log \frac{\sigma_{j2}^2}{\sigma_{j1}^2} - \frac{1}{2} \frac{(y_j - \mu_{j1})^2}{\sigma_{j1}^2} + \frac{1}{2} \frac{(y_j - \mu_{j2})^2}{\sigma_{j2}^2} \right] - \log \frac{\pi_2}{\pi_1}, \quad (11)$$

where $p_{jc} = P(i_j = 1|c)$, $j = 1, ..., q$, and $y_j|c \sim N(\mu_{jc}, \sigma_{jc}^2)$, $j = 1, ..., r$, $c = 1, 2$.

2.6 Random forests

Decision trees together with the use of the bootstrap method are the main ingredients of random forests. This methodology is used as a classifier, often successfully and with computational efficiency. Such method can be applied to data with both binary and continuous variables, and can additionally provide a measure of the importance of each of the variables in terms of classification performance. Random forests are not invariant to transformations of the covariates, and some of their parameters require tuning, for instance: the number of fitted trees, the number of variables allowed to participate on each one, the maximum depth of each tree, and the minimum number of observations at each terminal node of the trees.

3 Simulation study

We simulate data from a conditional Gaussian distribution in each population. The interaction graph is the same for both populations and is composed of two unconnected paths, one for four binary variables and another for six continuous variables, as shown in Figure 1d). The density function of $\boldsymbol{x} = (\boldsymbol{i}, \boldsymbol{y}) = (i_1, \ldots, i_4, y_1, \ldots, y_6)$ is given by

$$\begin{aligned} f_c(\boldsymbol{i}, \boldsymbol{y}) &= p_c(\boldsymbol{i}) f_c(\boldsymbol{y} \mid \boldsymbol{i}) \\ &= \frac{p_c(i_1, i_2) p_c(i_2, i_3) p_c(i_3, i_4)}{p_c(i_2) p_c(i_3)} f_c(y_1, ..., y_6 \mid i_1, i_2, i_3, i_4), \end{aligned} \quad (12)$$

where $f_c(y_1, ..., y_6 \mid i_1, i_2, i_3, i_4)$ is the density of a $N\Big(\boldsymbol{\mu}_c(\boldsymbol{i}), \boldsymbol{\Sigma}_c(\boldsymbol{i})\Big) = N\Big(\mathbf{0}, \boldsymbol{\Sigma}_c\Big)$, for $c = 1, 2$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$ with $\rho = 0.3$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ with $\rho = -0.3$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}^{-1} = \frac{1}{1-\rho_2^2} \begin{pmatrix} 1 & -\rho & 0 & 0 & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & 0 & 0 \\ 0 & 0 & -\rho & 1+\rho^2 & -\rho & 0 \\ 0 & 0 & 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & 0 & -\rho & 1 \end{pmatrix}. \tag{13}$$

The probability $p_c(\boldsymbol{i})$ is computed as in (12) considering that the bivariate probabilities associated with the connected binary variables in Figure 1d) are the same, that is, $p_c(i_1, i_2) = p_c(i_2, i_3) = p_c(i_3, i_4)$, where $p_1(0,0) = p_1(1,1) = 0.325$ and $p_1(0,1) = p_1(1,0) = 0.175$ for Π_1, and $p_2(0,0) = p_2(1,1) = 0.175$ and $p_2(0,1) = p_2(1,0) = 0.325$ for Π_2.

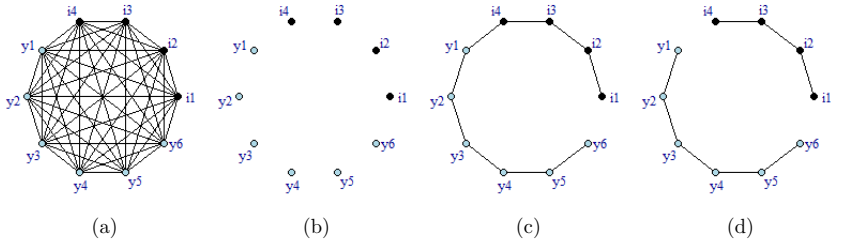


Figure 1: Interaction graph structures: a) complete: all interactions, b) empty: no interactions, c) path and d) two unconnected paths.

Table 1 and Figures 2 and 3 show the error rates obtained with 1000 training and test sets of size 50, 100 and 1000 in each population for nine methods, three linear and six non-linear. The computation of the errors was done with R (R Core Team, 2021), specifically with the following functions: *glm* to fit logistic regressions and, when selecting a model, *step* using BIC criterion; *naiveBayes* in *e1071* package (Meyer et al., 2020) for the naive classifier; *lda* and *qda* in *MASS* package (Venables and Ripley, 2002) for linear and quadratic discriminant analysis; *minForest* in *gRapHD* package (Abreu et al., 2010) to select a tree structure and various functions in *bnlearn* package (Scutari, 2017) to estimate the parameters and evaluate the discriminant function based on the CG distributions; and *randomForest* (Liaw and Wiener, 2002) to build a random forest using 1000 trees and tuning the *mtry* option.

Method	*Error rates %*					
	Training set			*Test set*		
Group sample size	50	100	1000	50	100	1000
Linear discriminant analysis LDA	37.1	41.0	47.2	50.2	49.8	50.0
Naive classifier	34.7	39.3	46.7	50.3	49.8	50.0
Logistic regression	37.1	41.0	47.2	50.2	49.8	50.0
CGD with tree structure	15.2	17.9	20.1	29.2	24.0	20.7
Quadratic discriminant analysis QDA	12.0	15.9	19.9	28.5	24.7	20.9
Best reduced Logistic with two-way interactions	0.0	15.2	20.3	35.8	26.7	20.9
Logistic with two-way interactions	0.1	9.6	19.6	38.4	29.4	21.2
LDA with two-way interactions	6.5	14.1	20.6	34.8	27.9	21.8
Random forests	0.0	0.2	0.0	37.3	33.0	23.3

Table 1: Training and test error rates for the simulated data based on 1000 training and test sets of size 50, 100 and 1000. The estimated Bayes error rate of 20.2% was computed using rule (1) with 20,000 observations from each population.

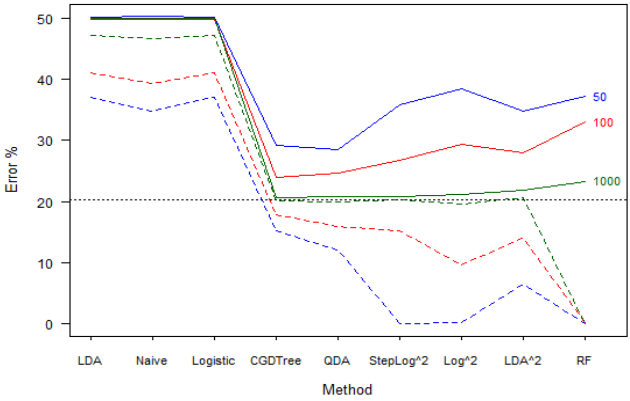


Figure 2: Estimated error rates for the simulated data based on 1000 training and test sets of size 50 (blue), 100 (red) and 1000 (green): test (solid lines) and training error (dashed lines). The dotted line shows the estimated Bayes error rate of 20.2%.

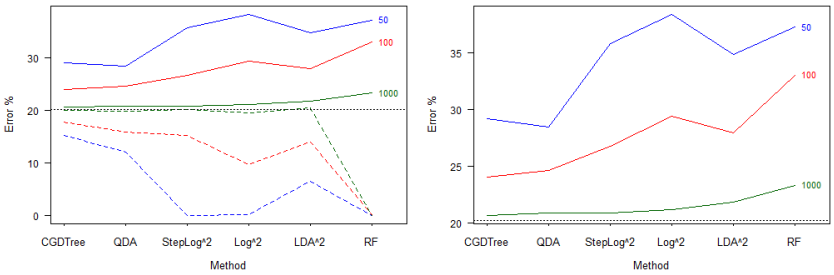


Figure 3: Estimated error rates for the simulated data based on 1000 training and test sets of size 50 (blue), 100 (red) and 1000 (green): test (solid lines) and training error (dashed lines). The dotted line shows the estimated Bayes error rate of 20.2%.

4 Discussion

The data set analysed here was generated from two homogeneous conditional Gaussian densities, one for each class. The simulation experiment was designed to show an instance where a linear discriminant function will not discriminate between the two populations. The main characteristics of the setting are the following.

i) Two conditional Gaussian densities $f_1(\boldsymbol{i}, \boldsymbol{y})$ and $f_2(\boldsymbol{i}, \boldsymbol{y})$ were used.

ii) There are $2^q = 2^4 = 16$ cells or different combinations of the values of the $q = 4$ binary variables in $\boldsymbol{i}$.

iii) The conditional Gaussian distribution for each population is homogeneous, this implies that its mean vector and covariance matrix do not depend on the cell value or location, i.e. $\boldsymbol{\mu}(\boldsymbol{i}) = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}(\boldsymbol{i}) = \boldsymbol{\Sigma}$ for all cells.

iv) $f_1(\boldsymbol{y} \mid \boldsymbol{i})$ and $f_2(\boldsymbol{y} \mid \boldsymbol{i})$ have mean vectors equal to zero and different covariance matrices, i.e. $\boldsymbol{\mu}_1(\boldsymbol{i}) = \boldsymbol{\mu}_2(\boldsymbol{i}) = \mathbf{0}$ and $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.

v) Although $f_1(\boldsymbol{y} \mid \boldsymbol{i})$ and $f_2(\boldsymbol{y} \mid \boldsymbol{i})$ have different covariance matrices $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, the marginal density function of each variable is the same in the two populations, i.e. $y_j \mid \boldsymbol{i} \sim N(0, 1)$, $j = 1, ..., 6$, for each cell in both populations.

vi) The two marginal probability distributions of the four binary variables in (9) are different, i.e. $p_1(\boldsymbol{i}) \neq p_2(\boldsymbol{i})$.

vii) The difference between $p_1(\boldsymbol{i})$ and $p_2(\boldsymbol{i})$ lies in the bivariate marginal distributions: $p_1(i_1, i_2) \neq p_2(i_1, i_2)$, $p_1(i_2, i_3) \neq p_2(i_2, i_3)$ and $p_1(i_3, i_4) \neq p_2(i_3, i_4)$. The univariate marginal probability functions are the same in both populations, i.e. $p_1(i_j) = p_2(i_j) = .5$ for $i_j \in \{i_1, i_2, i_3, i_4\}$.

viii) The graph structure associated with $f_1(\boldsymbol{i}, \boldsymbol{y})$ and $f_2(\boldsymbol{i}, \boldsymbol{y})$ is composed of two unconnected paths as shown in Figure 1d).

The characteristics of the conditional Gaussian densities $f_1(\boldsymbol{i}, \boldsymbol{y})$ and $f_2(\boldsymbol{i}, \boldsymbol{y})$ for this setting were selected so that the difference between the two populations was due to interactions between pairs of variables. In this case a linear discriminant function will not discriminate between the two populations. Similar examples where linear discriminant

analysis does not work, have been shown for example by Krzanowski (1977). The fact that the two populations have the equal marginal means might not be interesting in most practical applications, though an interesting example is given in Bartlett and Please (1963) where they deal with the discrimination problem in the case of zero mean differences.

The numerical results in Table 1 for this particular example show that the three linear classifiers did not discriminate between the two populations, as expected by design. Test errors were around 50%.

Considering the non-linear classifiers we note the following.

i) The five parametric classifiers for sample size 1000 on each group performed well. Their test error was at most one percentage point from the estimated Bayes error of 20.2%. The nonparametric classifier, Random forests, also performed well, its test error was three percentage points higher than the estimated Bayes error.

ii) For sample size 100 on each group. The test error rate of each parametric classifier differs between three and nine percentage points from the Bayes error, and the test error for Random forests was about ten percentage points higher than the Bayes error.

iii) For the small sample size of 50 observations from each population, the quadratic discriminant function and the discriminant function based on conditional Gaussian densities had the best performance with a difference of about nine percentage points from the Bayes error rate. All other parametric classifiers together with Random forests had a difference of between 14 and 18 percentage points from the Bayes error.

iv) For the five parametric classifiers the difference between training and test errors for the sample size 1000 is less than two percentage points. However, for the sample size 50, the training error is very low, at most 15.2% and zero in some cases.

v) Random forests, for this example, had zero average training error for the three sample sizes.

We conclude saying that the use of a classification method should be accompanied with training and test errors, resampling methods can be used for real datasets, and more than one classifier should be used to better assess the performance of a specific classifier. As Taylor mentions when commenting about the comparison of classifiers in the discussion of Ripley (1994, p. 441):

> A comparison of methods ... is often difficult to interpret. Observed differences in goodness of result can arise from:
> (a) different suitabilities of the basic methods for given data sets,
> (b) different sophistications of default procedures for parameter settings,
> (c) different sophistication of the program user in selection of options and tuning of parameters and
> (d) the occurrance of effectiveness of processing of the data by the user.

Acknowledgements Guillermina Eslava gratefully acknowledges the hospitality of the Department of Applied Mathematics and Computer Science, Technical University of Denmark. This work was done while she was on Sabbatical leave from the Faculty of Sciences at the National Autonomous University of Mexico (UNAM), and gratefully acknowledges a six months grant from the program PASPA, DGAPA, UNAM.

References

Abreu, G., Edwards, D., Labouriau, R. (2010) High-Dimensional Graphical Model Search with the gRapHD R Package. *Journal of Statistical Software* 37(1), 1–18

Bartlett, M.S., Please, N.W. (1963) Discrimination in the case of zero mean differences. *Biometrika* 50, 17–21

Højsgaard, S., Lauritzen, S.L., Edwards, D. (2012) *Graphical Models with R.* Springer, New York

Krzanowski, W.J. (1977) The performance of Fisher's linear discriminant function under non-optimal conditions. *Technometrics*, 19, 191–200

Lauritzen, S.L. (1996) *Graphical Models.* Clarendon Press, Oxford

Liaw, A., Wiener, M. (2002) Classification and Regression by randomForest. *R News* 2(3), 18–22

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch F. (2020). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien

R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ripley, B. D. (1994) Neural Networks and Related Methods for Classification. *J. R. Statist. Soc. B.* 56, 409–456

Scutari, M. (2017) Bayesian Network Constraint-Based Structure Learning Algorithms: Parallel and Optimized Implementations in the bnlearn R Package. *Journal of Statistical Software*, 77(2), 1–20

Venables, W.N., Ripley, B.D. (2002) *Modern Applied Statistics with S.* Springer, New York

Inferens for den relative risiko, når estimatet er 0

Sören Möller[1,2] og Linda Juel Ahrenfeldt[3]

[1] Klinisk Institut, Syddansk Universitet (moeller@health.sdu.dk)
[2] Open Patient data Explorative Network, Odense Universitetshospital
[3] Epidemiologi, Biostatistik og Biodemografi, Institut for Sundhedstjenesteforskning, Syddansk Universitet
(lahrenfeldt@health.sdu.dk)

Bidraget er baseret på artiklen [9] i IJERPH i 2021 af de samme forfattere.

Introduktion

Relativ risiko (RR) er det foretrukne associationsmål i mange sundhedsvidenskabelige studier, herunder kliniske interventionsforsøg og epidemiologiske kohortestudier. Inferens for RR, specielt i form af konfidensintervaller, kan bestemmes ved velkendte standardmetoder, hvilke der bl.a. undervises i på indledende biostatistikkurser. De fleste af disse metoder bygger dog på normalfordelingstilnærmelser eller andre asymptotiske forhold, og kræver derfor at både stikprøven og prævalensen af udfaldet er tilstrækkelig stor. Disse krav er typisk opfyldt for det primære udfald i et studie, men RR bruges også jævnligt til at undersøge risikoen for alvorlige bivirkninger og andre sjældne sekundære udfald, det vil sige i situationer, hvor et lavt antal prævalente tilfælde, potentielt 0, er forventet. I sådanne tilfælde vil mange inferensmetoder for RR enten ikke være anvendelige eller resultere i misvisende resultater. Som et eksempel rapporterede artiklen [3] kofidensintervallet $(0; 0)$ for en relativ risiko, indikerende, at den undersøgte bivirkning var udelukket i eksponeringsgruppen, hvilket desværre skyldtes misvisende bestemmelse af konfidensintervallet med bootstrapping og ikke at udfaldet er en umulighed (vi kommenterede på den konkrete problemstilling i [10]).

I dette bidrag giver vi derfor et overblik over mulige tilgange til at håndtere sådanne situationer.

Metoder der ikke virker

Klassiske metoder til at bestemme konfidensintervaller for RR vil typisk udnytte at binomialfordelinger asymptotisk kan approksimeres af en normalfordeling. Dette betyder desværre at sådanne metoder ikke vil være brugbare i situationer med 0 observerede udfald, da approksimationen først vurderes at være acceptabel, når der er mindst 10 observationer i hver celle i 2×2-tabellen [2].

Tilsvarende vil metoder, der bruger maksimum likelihood-tilgange (f.eks. i forbindelse med binomialregression), ikke give relevante konfidensintervaller, da observationen med 0 udfald vil ligge på randen af parameterrummet, og derfor ikke vil kunne resultere i retvisende inferens.

Endvidere vil bootstrapping ikke kunne anvendes, da de 0 observerede udfald vil indebære, at der vil være 0 udfald i alle de bootstrappede stikprøver, og RR-estimatet derfor vil være 0 i alle iterationer, resulterende i et ukorrekt konfidensinterval på $(0; 0)$, hvilket gik galt i det ovenfor nævnte studie [3].

Metoder der virker

Modificerede antal

En klassisk lavpraktisk tilgang er kunstigt at øge det observerede antal udfald fra 0 til en værdi større end 0. Der er forskellige forslag til hvilken konstant, der skal lægges til, og om denne skal lægges til alle celler i 2×2-tabellen eller kun til den celle, der er 0 [6]. Desværre viser det sig, at de resulterende konfidensintervaller er følsomme overfor valg af konstant [4].

Vores anbefaling er, såfremt denne tilgang bruges, i stedet at flytte et ikke-udfald til at være et udfald i eksponeringsgruppen med 0 udfald, for at opretholde den samlede stikprøvestørrelse, samt bibeholde et samlet udfald, der er kompatibel med den antagede sandsynlighedsfordeling.

Metoder for odds ratio

Modsat metoder for RR kan metoder der bestemmer eksakte (kombinatoriske) konfidensintervaller for odds ratioen (OR) typisk håndtere 0 observerede udfald i den ene gruppe. Der findes en håndfuld forskellige metoder til at bestemme disse eksakte konfidensintervaller, men den mest udbredte er forslaget fra Baptista og Pike, som også er tilgængelig i softwarepakker [1,5]. En udfordring ved denne tilgang er, at OR er et andet associationsmål end RR, dog viser det sig at OR er en acceptabel approksimation for RR, såfremt udfaldet er sjældent i begge eksponeringsgrupper [8].

Bayesianske metoder

Som en tredje tilgang, kan bayesiansk estimering for RR også bruges. Den mest oplagte tilgang er at modellere sandsynligheden for udfald separat i de to grupper der sammenlignes, og så bestemme posteriorfordelingen for RR ud fra posteriorfordelingerne for disse to proportioner. Denne procedure vil typisk uden videre kunne håndtere 0 observerede udfald i en (eller begge) grupper. Resultaterne vil dog afhænge af de valgte priorfordelinger for proportionerne, hvor en Beta-fordelt prior vil være det naturlige valg, da denne både er teoretisk velbegrundet (som konjugeret prior til en binomialfordeling) og er let at anvende i praksis [7].

Konklusion

Vi har redegjort for, at det er muligt at rapportere retvisende konfidens- eller kredibilitetsintervaller for RR også ved 0 observerede udfald i eksponeringsgruppen. Vores opfordring er derfor, at sådanne intervaller rapporteres for at dokumentere (u)sikkerheden af estimaterne, men at intervallerne bestemmes med en

metode, såsom dem vi foreslår ovenfor, der kan håndtere 0 observerede udfald, sådan at misvisende resultater undgås.

Referencer

1. Baptista, J., and Pike, M. C. Algorithm as 115: Exact two-sided confidence limits for the odds ratio in a 2 × 2 table. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 26*, 2 (1977), 214–220.
2. Blyth, C. R., and Still, H. A. Binomial confidence intervals. *Journal of the American Statistical Association 78*, 381 (1983), 108–116.
3. Das, M. K., Arora, N. K., Poluru, R., Tate, J. E., Gupta, B., Sharan, A., Aggarwal, M. K., Haldar, P., Parashar, U. D., Zuber, P. L. F., Bonhoeffer, J., Ray, A., Wakhlu, A., Vyas, B. R., Iqbal Bhat, J., Goswami, J. K., Mathai, J., K, K., Bharadia, L., Sankhe, L., M K, A., Mohan, N., Jena, P. K., Sarangi, R., Shad, R., Debbarma, S. K., J, S., Ratan, S. K., Sarkar, S., Kumar, V., Maure, C. G., Dubey, A. P., Gupta, A., Sam, C. J., Mufti, G. N., Trivedi, H., Shad, J., Lahiri, K., R, K., Luthra, M., Behera, N., P, P., G, R., Kumar, R., Sarkar, R., A, S. K., Sahoo, S. K., Ghosh, S. K., Mane, S., Dash, A., Charoo, B. A., Tripathy, B. B., G, R. P., S, H. K., K, J., Sarkar, N. R., Arunachalam, P., Mohapatra, S. S. G., and Garge, S. Risk of intussusception after monovalent rotavirus vaccine (Rotavac) in Indian infants: A self-controlled case series analysis. *Vaccine 39*, 1 (01 2021), 78–84.
4. Dewey, M. E. Collated responses from r-help on confidence intervals for risk ratios, 2006.
5. Fagerland, M. W. Exact and mid-p confidence intervals for the odds ratio. *STATA JOURNAL 12*, 3 (2012), 505–514.
6. Gart, J. J. Alternative analyses of contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological) 28*, 1 (1966), 164–179.
7. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2011.
8. McNutt, L. A., Wu, C., Xue, X., and Hafner, J. P. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol 157*, 10 (May 2003), 940–943.
9. Möller, S., and Ahrenfeldt, L. J. Estimating Relative Risk when Observing Zero Events – Frequentist Inference and Bayesian Credibility Intervals. *IJERPH 18*, 11 (2021), 5527.
10. Möller, S., and Nielsen, S. Letter to the editor. *Vaccine* (online first 2021).

Google searches linked to Apple stock volatility

Niels Buus Lassen[a]

[a] *Centre for Business Data Analytics, Copenhagen Business School, Denmark*

ARTICLE INFO

Keywords: Investor behavior, Google searches, Stock markets,Investor sophistication, Decision making

ABSTRACT

The recent studies on social media that link news data to volatility show a Twitter buzz up is typically linked to higher volatility, while a general news media buzz is linked to lower volatility in the following month. This article demonstrates that Google searches influence Apple stock volatility in either on a weekly basis by analyzing the behavior of privateand professional investors in relation to Google searches and how this behavior links to Applestock volatility. To this end, this study employs the logic of sales modeling and, thus, contributes to the theoretical construction of the novel "investor journey model" by mapping Google searches onto investorbehavior, which is an under-researched field in the literature. Subsequently, the paper summarizes the main findings in this field and outlines future challenges in this research.

1. Introduction

Twitter data have been included in several models and shown to have predictive power for both stock price indexes and specific stock price movements (see, e.g., Bollen & Mao 2010; Jiao, Veiga, & Walther 2016; Li, van Dalen, & van Rees 2018.).

A logical explanation for the predictive power of social media data in terms of financial market behavior is the size of the big data from social media chats and also Google searches about the respective stocks. Stocks with large amount of social media data, are popular stocks people like to talk about on social media and also do Google searches about.

Text mining can identify patterns in the big data from the social media and Google searches. Statistical and machine learning methods can be tested to model the human behavior on social media and Google searches to financial market behaviors.

This approach is comparable to the sales of products and services, where big data from social media can be used to predict sales. The sales models that build on social media data work well if social media data are big enough or, inother words, if customers like to talk about a product or service on social media.Examples are H&M, Nike, and Apple (see, e.g., Lassen et al. 2017; Boldt et al. 2016). These are popular social media topics that produce big enough data to have predictive power for their sales.

All chat text on social media about products and services can be categorized into one of the phases in the Customer Infinity Model(Figure 1). These phases of customer behavior can then be modeled based on sales and provide logical explanation for why social media data can predict sales, if the data are big enough (see, e.g., Asur & Huberman 2010; Lassen et al. 2014).

Fig. 1 Customer Infinity Model

The logic for stock prices is comparable, based on which this article proposed the "investor journey model," which posits that all social media text or web searches can be categorized into one of the phases of this model, which can in turn be linked to Apple stock volatility, as well asother stocks with big enough data on social media and from web searches. This paper focuses on both the private and professional investor behavior related to Google searches.

More than 90% of all global web searches use Google; specifically, approximately 63% on the search engine, 23% on Google Images, 4% onYouTube, and 1% on Google Maps (BusinessInsider.com 2018). When amateur and professional investors search for information about the stocksthey are interested in, Google searches provide relevant stock information and form the majority of all web searches. The model proposed in this article has been tested on more than 60 Google searches before identifying good proxies for the amateur and professional investor behavior based on Google searches for Apple stocks.

The research questions this article tackles as follows:

RQ1: *Can Google search data predict stock price volatility?*

RQ2: *Which Google searches are creating ups and downs inApple stock price volatility?*

RQ3: *Can the identified ups and downs in Apple stock pricevolatility be linked to household and professional investor activity?*

The contribution of this article is identifying new patterns for investor behavior on web searches, and how these patterns link to ups and downs instock volatility theoretically explained through the proposed investor journey model, which relies on insights from related queriesand topics in Google Trends from more than 60 Apple-related Google searches. Apple is the example used in this article, and model may be applicable for similar big tech stocks with high volume google search data.

2. Literature review

One of the most notable articles modeling stock price volatility using Google Trends, is Preis et al. (2013), which mentions: "We suggest that *Google Trends* data and stock market data may reflect two subsequent stages in the decision making process of investors."

This article suggests that Apple stock-related Google Trends data follow an investor decision making journey, which affects Apple stock volatility. The proposed investor journey model is detailed in section 3.

Jiao, Veiga, and Walther (2016) find that a buzz up in coverage by traditional news media predicts subsequent decreases in volatility and turnover, while a buzz up in coverage by Twitter predicts increases in the subsequent volatility and turnover. However, they do not explain why the buzz in traditional newsmedia and Twitter coverage have these different effects on stock volatility.

Greenwich Associates published a report in 2015 based on interviews with 256 asset owners from more than 250 institutional investor organizations, which shows that institutional investors use Twitter in a very limited manner in their decision-making process, compared to LinkedIn, for example. Among the 256 interviewed institutional investors, LinkedIn was used by more than half and often played an important role in investor decision making. The interviewed institutional investors recognized the value of the Twitter news feed in seeking opinions or commentary on market events, but considered LinkedIn feeds to be better targeted, as they reflected their professional ties more closely.

This article does not include LinkedIn data, but only Apple-related Googlesearches due to their free availability through Google Trends. Specifically, LinkedIn data are difficult to access and expensive. Twitter data were also not considered due to their cost for this article.

Institutional and professional investors use information processed from Twitter by analytical companies such as Dataminr. For example, TheGlobeAndMail.com (2018) states: "Dr. Mohanram cautions that individual investors are not likely to be able to correctly replicate the conditions of the study to benefit from crowd-sourced opinion. It requires a certain amount of data crunching ability and sophistication [to] analyze a large sample of tweets, categorize them and aggregate them all in real time."

Private investors are using more Twitter unprocessed information in their decision-making processes such as by following investor gurus and searching for stock related information on Twitter. For example, MarketWatch.com's (2018) "Finance Twitter: The 50 most importantpeople for investors to follow" is one of many articles recommending private investors who to follow on Twitter to get smart insights on investing. Based on such recommendations, it becomes logical that the abundance of financial gurus on

Twitter is creating noise and increases stock volatility. Searching for stock related information on Twitter typically yields several credible news sources along with sources that require vetting. In short, the many news sources on Twitter create noise and increase stock volatility. Namely, private investors are more exposed to the risk of rumors, old news being perceived as new news, speculations, and manipulating info to drive stock prices up or down. All this noise can lead to irrational investor behavior by creating higher volatility when there is a buzz up in Twitter info for a stock.

Therefore, the logic of Twitter usage is that private investors are the main actors in the increased stock volatility after a buzz up in the Twitter infoabout a stock.

Traditional news media info about stocks are typically used by professional and institutional investors. Examples of reliable traditional news media channels are Financial Times, Bloomberg, Wall Street Journal, Reuters, CNBC, Forbes, and MarketWatch. They are associated with rational investor behavior and lower volatility after a news buzz up related to a specific stock.

For the Apple stock, around 60–70% of Apple stock trading is conducted by professional andinstitutional investors. For details, please refer to
https://finance.yahoo.com/quote/AAPL/holders/.

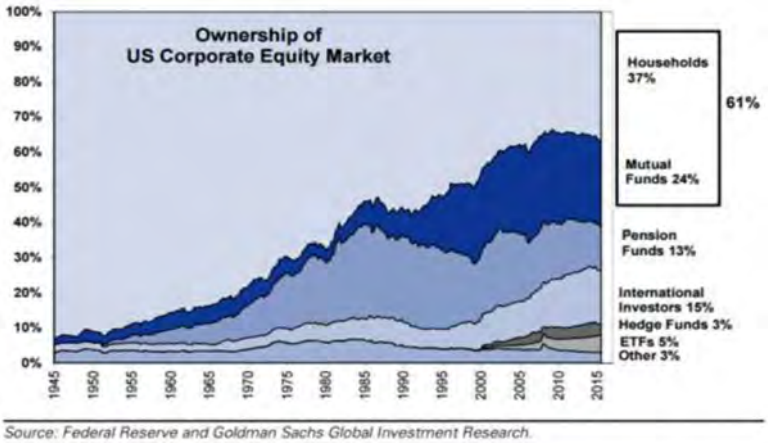


Fig. 2 Ownership of the US corporate equity market: USD 36 trillion as of Q2 2016 and includes USD 7 trillion of foreign equity holdings.

Based on Fig. 2, US households own approximately one third of the UScorporate equity market, meaning that the household investor activity in Apple stocks is estimated to be one third on average.

Many models for stock investor decision making are based on the prospect theory developed by Kahneman and Tversky (1979), which is a descriptive model of decision making under risk (see, e.g., Wakker & Levy 2015). Prospect theory has been criticized and several alternatives have been suggested (see, e.g., Nwogugu 2005; Levy et al. 2002).

Prospect theory assumes decision making has two phases: editing and valuation. The editing phase refers to investors scanning the information and forming their beliefs on eventual outcomes. This phase is heavily dependent on psychological biases. The second phase is the valuation phase, where agents value these eventual outcomes based on the beliefs in the first phase. This phase depends more on the risk preferences of investors. These two decision-making phases are distinct, albeit closely dependent on each other.

The model for stock investor decision making developed in this article isbased on the empirical observation of Google searches related to the Apple stock (stock symbol: AAPL). The model does contain some editing and valuation by investors but focuses on explaining how different types of investor use information before, during, and after the Apple Quarterly Reports and iPhone models are released.

3. The Investor Journey Model

The "investor journey model" is a conceptual model developed in this study for the analysis of investor behavior based on web searches and is applied to Apple stocks in this paper. Specifically, it has been developed based on related queries and topics on Google Trends from more than 60 Apple-related Google searches. It is thus the main analysis framework in this study, as described in Figs. 3 and 4.

		Household investors	Professional & Institutional investors
	Phase 1	Would I like to buy or sell the Apple stock? Or should I buy or sell Google, Amazon, Facebook, Microsoft, Nvidia or IBM instead?	Should I add or decrease Apple stocks in my portfolio? The Apple stock is here in a setup of portfolio diversification & risc profiles.
	Phase 2	Which stock or stocks do I proceed with? Which rumor news will affect this?	What is my analysis of Apple sales & results, before the Apple Quarterly Report? Any reliable Apple rumor news indicating level of sales before the Apple Quarterly Report?
Every end of Jan, Apr, July & Oct	Phase 3a	Apple Quarterly Report release. Above or under my expectations? Time to buy or sell Apple stock?	Apple Quarterly Report release. Above or under my analysis? Should I add or decrease Apple stocks in my portfolio?
Every Sep.	Phase 3b.	Would I like to buy or sell the Apple stock? before the Sep launch of new iPhone? Will the iPhone launch drive stock up or down?	What is my analysis of iPhone launch, before the Sep launch of new iPhone? What is the level of rumor news compared to earlier years? Will the iPhone launch drive stock up or down?

Fig. 3 The Investor Journey

Fig. 4 Timeline for the investor journey

I divide the Apple stock-related Google searches into two groups—household and professional investors—based on the following hypotheses.

Hypothesis 1: The Google searches for stock symbols AAPL, AMZN, and IBM are assumed to be mostly made by household investors, as professional investors are likely familiar these stock symbols and do not need to Google search them. Google searches for these stock symbols lead toSeekingAlpha.com and StockTwits.com, which have large groups of household investors among their readers.

Hypothesis 2: The Google searches for Apple rumor news are assumed to be largely done by professional investors, as the Apple rumor news sites use high vetting levels and due diligence for the sources.

4. Methodology

The analysis is based on the multiple regression modeling of more than 60Apple-related Google searches as input variables and investor behavior inthe form of Apple stock volatility as the dependent variable.

For the Apple and rumor news Google searches, professional investors largely consider the vetted news and rumor sites, while household investors largely Google the stock symbols of Appleand other big tech companies. These patterns were identified based on the related topics and queries for each Google search, but also on the assumption of the tendencythat household investors drive volatility up and professional investors drive it down. This argumentation of the patterns for household andprofessional investors will be further elaborated upon in the following sections.

The statistical software used is Oxmetrics 8.10, which mainly focuses on using the Autometrics functionality. Autometricsis a part of the PcGive module of Oxmetrics, being the automatic econometric model selection procedure that is available in PcGive. It is based on regression modeling under the general unrestricted model (GUM) framework. In Autometrics, the variable and model selection criteria are based upon the unique method developed by David Hendry and Jürgen Doornik, which performs well on gauge and potency. Gauge is the retention rate of irrelevant variables in the selected model and akin to size, because it accounts for the wrongly selected variables . Potency is the retention rate of the relevant variables in the selected model. It is also akin to size because it accounts for the variables that have been correctly selected (see Hendry et al. 2014).

The chosen target size for the dataset used in this study is 1%, which is the t-probability threshold for choosing and eliminating input variables. The1% level was chosen because of the 60+ Google searches selected as input variables in 1–4 time lags each; therefore, the large amount of input variables could be cut down to a reasonable number. After selecting among more than 60 Google searches with a target size of 1%, a target size of 5% was also tested when the input variables were reduced to a group of 10 predictors.

Another variable selection method that comes from machine learningis the least absolute shrinkage and selection operator (LASSO) method. It is a type of linear regression that uses shrinkage, where data values are shrunk towards a central point, such as the mean. LASSO adds the “absolute value of magnitude” of the coefficient as penalty term to the loss function, which will be minimized. This is called L1 regularization.

This regression works especially well for many inputvariables and multicollinearity and

can limit the input variables significantly. The input variables field is cut down by the described L1 regularization.

The analysis will examine if Apple-related Google searches are good proxies for investor behavior regarding the Apple stock. Both household and professional investor are reflected in Google searches on the investorjourney to buy or sell Apple stocks.

5. Data

5.1 Google searches during 2015–2020

The Google search data were collected from Google Trends (https://trends.google.com/trends/). Specifically weekly Google search data were collected from April 2015 to April 2020, as this is the longest period available on Google Trends with weekly Google search data.

It is possible to get obtain Google searches on Google Trends, but they are only available up to the 90 prior days. Weekly Google search data are available on Google Trends for the past 12 months or 5 years. For longer periods, starting from 2004, monthly date are available. The selected data were evaluated to be the most suitable dataset for modeling Applestock prices and volatility because it was the longest and most recent time period available with weekly data at the time of this study.

The Google search data are given in indexes from 0 to 100, as positive integers, available on Google Tends. An index 100 for one or more weeks would be the highest weekly search volume for the entire 5-year period. In this article, more than 60 Apple-related Google searches were extracted for the 5-year period and tested for their relationship to stock price and volatility. These searches were found by exploring Google Trends for the Apple stock symbol (AAPL) and products such as iPhone, iPad, MacBook iOS, or MacOS. Google Trends includes both related topics and queries, based on which I found stocks related to the Apple stock, which led to the idea of developing the investor journey model. Fig. 5 shows one data extract.

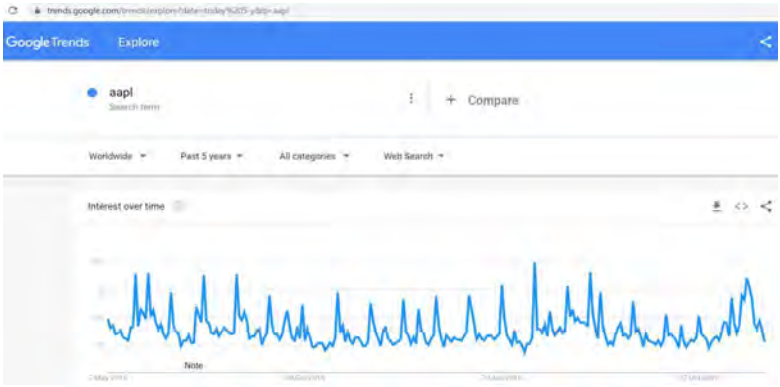


Fig. 5 Google search data for "AAPL" for the period 2015-2020

https://trends.google.com/trends/explore?date=today%205-y&q=AAPL

The Google searches can be extracted in sets of up to five, but I instead considered one search at a time, as for more than five searches extracting sets will create problems. That is, because the five searches will be indexed in a group of five, a second set will not be indexed against the first set. As such, unless there is an overall baseline, the highest of all Google searches will be identified for each set. However, the baseline can change during the research to include all the dataset. As such, the most practical approach is to extract Googlesearches one at a time and, in the end, the searches can be indexed together in one set of five or two sets of 10 for 5–10 input variables.

For example, for the Google search of the Apple stock symbol (AAPL) shownin Fig. 5, all weeks in the 5-year period are indexed around the datapoint with index 100 in week 31 of 2018. For up to 5 searches in the same Google Trends query, all weeks would be indexed around the highest search with index 100 in a given week. The final dataset from Google Trends included sets of five Google searches and all searches were indexed to the highest search index, "IBM," in the 5-year period.

Variables	*N*	Mean	Median	St. dev.	Min	Max
AvgWeek-lyClose	260 weeks	164.5	158.1	54.3	91.9	323.6
Weekly Vol-ume, number of million shares	260 weeks	170.8	156.4	72.7	32.5	500.4
Weekly volatility	260 weeks	3.20%	2.64%	2.24%	0.51%	19.17%
First diff Log(avg-Wee kClose)	260 weeks	0.13%	0.24%	1.69%	-13.46%	12.76%

Table 1 List of weekly financial variables

Table 1 shows that the weekly volatility includes an extreme outlier of 19.17%. This was due to the COVID-19 pandemic from February 24 to April 12, 2020, were the weekly volatility ranged between 9% and 19.2%, peaking at 19.2% in week 11—March 9–15, 2020. In the same time window, the average trading volume was 300 million shares per week and stock price varied from USD 212 to USD 304.

Table 2 presents the descriptive statistics for the Google search data. As previously mentioned, all Google searches for 5 years have been downloaded from Google Trends as 260 weekly observations. Specifically, 62 Apple related Google searches were tested for modeling the Apple stock volatility, among which nine were significant and were chosen for further modeling. The two most important input variables, the Google searches for **"AAPL"** and **"AMZN"**, selected by SPSS LASSO and Autometrics are marked with **bold.** The Google searches for **"MacRumors"** and **"Apple rumors"** are also marked with **bold**, as they were additionally selected by Autometrics when the target size was changed from 1% to 5%. The target size in Autometrics is the t-probability threshold for choosing and eliminating input variables. There are two groups for the searches:

1. News and rumors searching/vetting for Apple stock
2. Apple and other related big tech stock searches

The Google search “Apple rumors” marked with green is the only variable considered with an overweight of professional investors and negative coefficient. All other Google searches are considered to have an overweight of household investors.

Google searches	*N*	Tested time lags	Mean	Median	St. dev.	Min	Max
First section, rumor, and news searches							
Apple Rumors	**260 weeks**	**1–4 weeks**	**0.9**	**1.0**	**0.4**	**0.5**	**3.0**
9to5mac	260 weeks	1–4 weeks	1.2	1.0	0.6	0.5	5.0
TheVerge	260 weeks	1–4 weeks	1.0	1.0	0.2	0.5	2.0
MacRumors	**260 weeks**	**1–4 weeks**	**2.7**	**2.0**	**1.5**	**1.0**	**12.0**
AppleInsider	260 weeks	1–4 weeks	0.7	0.5	0.3	0.5	2.0
Second section,Apple, and related big tech stock searches							
AAPL	**260 weeks**	**1–4 weeks**	**20.2**	**18.0**	**7.6**	**9.0**	**51.0**
AMZN	**260 weeks**	**1–4 weeks**	**14.9**	**13.0**	**9.7**	**2.0**	**48.0**
IBM	260 weeks	1–4 weeks	72.2	72.0	10.9	38.0	100.0

Table 2 Descriptive statistics for the Google searches

From Table 2, the search for “IBM” has the highest index number 100, all other Google searches being indexed to it. Had this search not been included, the search for AAPL would have been the main index, meaning all other Google searches would have been indexed after it, since it has the second highest index of 51. Excluding the Google search on IBM, would also have increased the index numbers for Google searches on MacRumors, Apple rumors, and AMZN, which could have changed their significance.

In Autometrics, the estimation sample cannot start before week 19 in 2015 because of the tested time lags for 1–4 weeks in the dataset. Therefore, the Autometrics estimation is conducted from week 19 in 2015 to week 42 in 201, covering 88% of the dataset. The last 26 weeks of the dataset from week 43 in 2019 to week 16 in 2020 are chosen as hold-out data for the forecast evaluation.

Fig. 6 Timeline for the dataset and visualization of the train/test split.

Fig. 7 shows time plots of the two tested dependent variables, namely the weekly volatility and first difference log(avg close), which is the stock price return. It also shows the time plots of the two most significant regressors, the Google searches for AAPL and AMZN and the highest Google search index in the 5-year period for IBM.

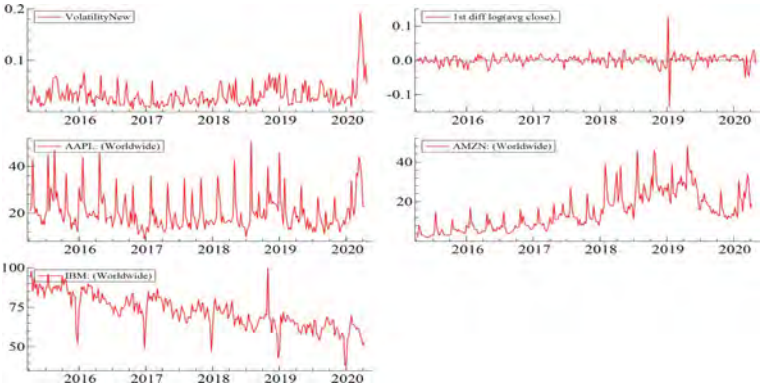


Fig. 7 Timeplots of the main variables

The selected Google searches developed as follows:

AAPL: Quarterly peaks around the ends of January, April, July, and October in each year, when Apple is releasing its quarterly reports to Nasdaq and on their investor site.

AMZN: Quarterly peaks around the ends of January, April, July, and October in each year, when Amazon is releasing its quarterly reports to Nasdaq and on their investor site.

IBM: Quarterly down peaks around the end of December. An explanation could be that IBM is not linked to the Christmas season, and the search for more Christmas-linked stocks overtakes the December searches. The downward pattern in the Google searches follows the decline in stock prices during 2015–2020. The main role of the Google searches for "IBM" is that it is the highest Google search index in the 5-year period for 10 out of 60 Google searches that were most significant for the weekly Apple stock volatility. Therefore, all last 10 Google searches in the last dataset tested are indexed after the IBM searches. The declining trend for IBM searches during 2015–2020 follows the declining trend in both IBM stock price and IBM's position in machine learning and AI, where IBM is not among the leaders.

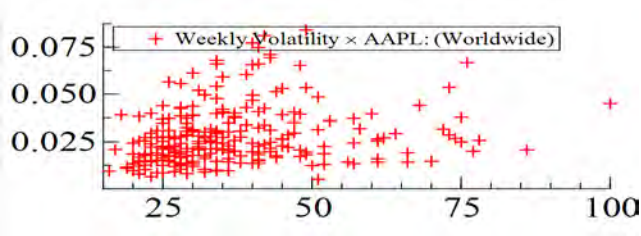


Fig. 8 x-y plot of weekly volatility x AAPL

Y axis is Weekly Volatility and X axis is AAPL Google searches. The x-y plot shows patterns of higher volatility for higher AAPL Google search indexes.

5.2 Apple announcement data

Apple releases its quarterly reports at the end of the month after a quarter closes, that is, approximately 1 month after the quarter closes, to stock holders and the media through the investor portal at Apple.com, https://investor.apple.com/investor-relations/default.aspx

There was a test of event variables for the iPhone launch in September; the quarterly Reports every end of January, April, July, October; and the Black Friday and Christmas sales. These event variables did not show any effect on either weekly volatility or stock price returns. These event variables were defined as 0/1 variables, taking 1 in the week the event occurred, and 0 otherwise.

There are patterns in some Apple-related Google searches before, during, and after the quarterly reports, which are considered as quarterly regular spikes in these Google searches. Refer to Fig. 7 for an example of these patterns.

Apart from the quarterly reports from Apple, the single most important event for Apple is the yearly iPhone launch in September. At the iPhone launch 2015, the weekly Apple stock volatility increased, but from 2016 onwards the iPhone launch has been a mean event based on weekly volatility. That also tells a story of the iPhone hype wearing off. The iPhone hype topped in 2012, and has been decreasing since then. Refer to the Google searches for iPhone from 2007-2021:

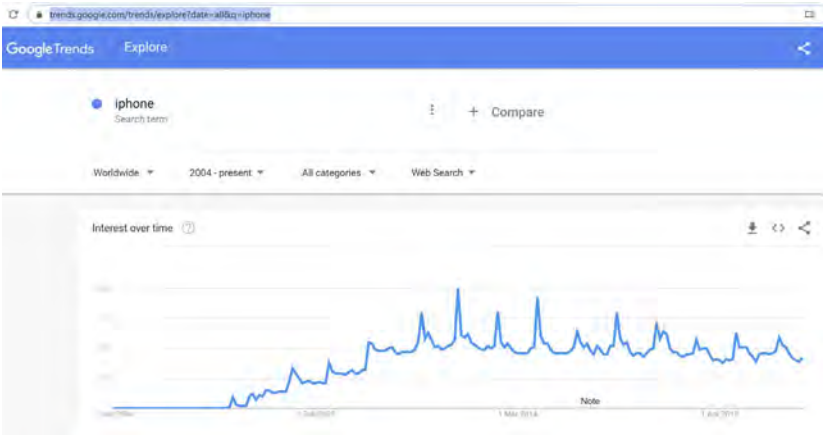


Fig. 8, iPhone Google searches, 2007-2021. https://trends.google.com/trends/explore?date=all&q=iphone

The first iPhone was launched in June 29, 2007 and, from 2012, the main iPhone launch has always been in September. There are patterns in some Apple-related Google searches before, during, and after the September iPhone launch, which are seen as yearly peaks in these searches. Refer to Fig. 7 for an example ofthese patterns.

There are also some announcement data on other Apple products, but the most important product announcement for Apple and Apple stock is theyearly iPhone launch.

5.3 Weekly return data

Nasdaq.com provides free data download for their listed stocks going back10 years at https://www.nasdaq.com/symbol/aapl/historical.
These data provide daily info for the selected period on Apple stockprice, namely open, high, low, close, and volume. These daily data from Nasdaq.com were the basis for calculating weekly Apple stock price volatility, based on formula (5).

These daily data were also used to calculate the average weekly stock priceand weekly volatility. The weekly average stock price was calculated based on the average of all the daily closing prices. Daily closing prices were also used to calculate the daily changes, forming the basis for calculating the weekly volatility. As the Nasdaq data only have daily close, high, low values for stock prices—with no daily average—it was chosen to use the close for the weekly calculations of the financial variables. These weekly financial variables were needed in the model together with the weekly Google searches.

5.4 Formulas

$\boldsymbol{P_t}$ is the weekly average stock price in week t, calculated as:

(1) $\boldsymbol{P_t} = (\boldsymbol{P_{1t}} + \boldsymbol{P_{2t}} + \boldsymbol{P_{3t}} + \boldsymbol{P_{4t}} + \boldsymbol{P_{5t}})/5$,

where $\boldsymbol{P_{it}}$ is the closing stock price for Apple on day i in week t.

Approximately 80% of the trading weeks, have 5 trading days—Monday to Friday—based on the above formula.
The remaining 20% of the trading weeks have 4 trading days, the formula being:

(2) $\boldsymbol{P_t} = (\boldsymbol{P_{1t}} + \boldsymbol{P_{2t}} + \boldsymbol{P_{3t}} + \boldsymbol{P_{4t}})/4$.

Very few trading weeks have 3 trading days, with the following formula:

(3) $\boldsymbol{P_t} = (\boldsymbol{P_{1t}} + \boldsymbol{P_{2t}} + \boldsymbol{P_{3t}})/3$.

The stock price return in this article is expressed as a first difference log variable,in line with the ARCH models. The use of stock price log returns has advantages over the arithmetic return (see, e.g.. Hudson & Gregoriou 2010). The first difference stock price return is expressed as:

(4) $\Delta \boldsymbol{log}(\boldsymbol{P_t}) = \boldsymbol{log}(\boldsymbol{P_t}) - \boldsymbol{log}(\boldsymbol{P_{t-1}})$.

The best explanatory variable for the weekly stock price return is theweekly stock price return from the previous week and two weeks before. Prior to volatility, I also investigated modeling stock price return. Theresults are available upon request. The

historical stock price return only creates a R^2 of 10% for the training data and 6% for the hold-out data. With the adding of the best three Google searches, R^2 increases to 17% for the training data and 10% for the hold-out data. The model for stock price return is not strong, the focus of this study being thus on modeling the weekly volatility for the Apple stock.

For the weekly average volatility, the model is much stronger. I follow Christiansen et al. (2012) and Paye (2012), and define weekly volatility as:

(5) $$\boldsymbol{\sigma_t = \sqrt{(r_{1t}^2 + r_{2t}^2 + r_{3t}^2 + r_{4t}^2 + r_{5t}^2)}},$$

where the *r*s are the five daily changes in stock price for a week with 5 trading days. For 80% of the weeks in the dataset with five trading days, there are five *r*s in the above formula. For 20% of the weeks in the dataset with 4 trading days, there are four *r*s in the formula. For the few weeks with 3 trading days,there are three *r*s in the formula. A classical weekly volatility formula is the standard deviation of five dailystock price changes, which is a variance. The above formula does not subtract the mean from each daily change, before considering each of the five daily returns in a week. The above formula is also not dividing with N - 1 before applying the square root. Therefore, compared to the classical volatility formula, the new formula can be interpreted as a measure of stock price fluctuations in a given week without using the weekly mean for daily changes.

All variables are now defined, so the final regression model for weeklyvolatility can be defined as:

(6) $$\sigma_t = \alpha_0 + \alpha_1 \sigma_{t-1} + \cdots . + \alpha_p \sigma_{t-p} + \sum^{N}{}_{i=} \quad \beta_i X_{it} + \varepsilon_t,$$

where X is the variable constructed from the Google search data and also the event variables for iPhone launch, Black Friday andChristmas sales, and quarterly reports. The event dummies are defined as0/1 variables, which take 1 in the week the event occurred, and 0 otherwise.

1. Results and Discussion

The initial test runs on 62 Google searches not indexed against each other, were mostly used to identify the most relevant Google searches. After the most relevant 10 Google searches were indexed against each other, both IBM SPSS LASSO and Autometrics picked only two relevant Google searches, that is, "AAPL" and AMZN," which are the stock symbols for Apple and Amazon, respectively.

Oxmetrics 8.10 was used to model the 62 Google searches as input variables, with weekly volatility as the dependent variable. Using the automatic model selection function in Oxmetrics 8.10, called Autometrics, the 62 Google searches were tested with time lags from 1 to 4 weeks, and 51 weekly seasonal dummies were also included in the modelling. Event dummies for Apple quarterly reports, iPhone releases, and Black Friday and Christmas sales were also tested.

The final model output in Table 1 is from the automatic model selection, single- equation dynamic modeling using Autometrics.

Software:	OxMetrics 8.10					IBM SPSS Statistics 26				
Variable selection method:	Autometrics					Lasso				
Variables selected	Coefficient	Std.Error	t-value	t-prob	Part.R^2	Coefficient	Std.Error	t-value	t-prob	Part.R^2
t-1 AAPL: (Worldwide)	0.00147	0.0890	16.5	0.000	0.5552	0.00132	0.0594	22.2	0.000	0.6901
t-4 AMZN: (Worldwide)	0.00037	0.0763	4.91	0.000	0.0995	0.00027	0.0715	3.71	0.000	0.0585
t-4 macrumors: (Worldwide)	0.00183	0.0008	2.39	0.018	0.0255					
t-4 apple rumors: (Worldwide)	-0.00659	0.0027	-2.45	0.015	0.0268					
Diagnostics tests:						**Diagnostics tests:**				
AR 1-7 test:	F(7,211) = 2.6487 [0.0121]*					AR 1-7 test: F(7,215) = 4.2173 [0.0002]**				
ARCH 1-7 test:	F(7,210) = 4.4896 [0.0001]**					ARCH 1-7 test: F(7,210) = 3.0202 [0.0048]**				
Normality test:	Chi^2(2) = 9.5215 [0.0086]**					Normality test: Chi^2(2) = 10.519 [0.0052]**				
Hetero test:	F(10,213) = 3.4856 [0.0003]**					Hetero test: F(4,219) = 7.9119 [0.0000]**				
Hetero-X test:	F(25,198) = 3.1901 [0.0000]**					Hetero-X test: F(5,218) = 6.7054 [0.0000]**				
RESET23 test:	F(2,216) = 6.7985 [0.0014]**					RESET23 test: F(2,220) = 5.8947 [0.0032]**				

Table 1. Model Output from Autometrics & Lasso

6.1 Diagnostic tests

The AR 1-7 test is a standard test of autocorrelation up to degree 7. It teststhe joint hypothesis that εˆt is uncorrelated with εˆt−j, for any choice of j, against the alternative that εˆt is correlated with εˆt−j. The null hypothesis of no autocorrelation between residuals can be rejected for a P-value of 1.21% and significance level of 2.5%. Therefore, there is formal evidence of littleautocorrelation between residuals. At a significance level of 1%, the null hypothesis is accepted, which makes it a borderline scenario for this test. In the LASSO model with just two predictors, there is clear rejection of thenull hypothesis and formal evidence of autocorrelation between residuals.

The ARCH 1-7 test is a standard ARCH test of the null hypothesis of no ARCH effect, that is, if the squared standardized residuals do not exhibit autocorrelation. The null hypothesis of no ARCH effect is rejected for P-values of 0.0001 and 0.0048. Therefore, there is formal evidence of the ARCH effect in the model.

Normality test. The null hypothesis of normality is rejected at a significance level of 1%, with P-values of 0.86% and 0.52%. Hence, there isno formal evidence of normality for this model.

Hetero and hetero-X tests. The null hypothesis of homoscedasticity can berejected for P-values of 0–0.03% at the 1% significance level. Hence, there is formal evidence of heteroscedasticity in this model.

RESET123 test. The regression specification error test has a null hypothesis of no squared and cubic terms in the regression model. The nullhypothesis can be rejected for P-values of 0.14% and 0.32% at the 1% significance level. Hence, there is formal evidence of mis-specification of the regression model from this test.

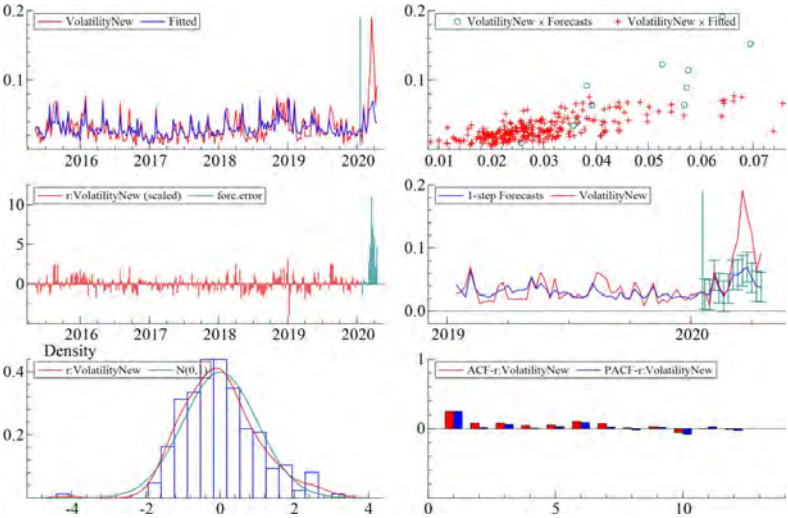


Fig. 9 Model Output from Autometrics

The residual plot shows a random pattern, suggesting a linear model wouldfit well the data. Residuals are also normally distributed, again suggesting the linear regression model is fitting the dependent variable well. The ACF and PACFplots show no autocorrelation. The COVID-19 peak in March 2020 is an outlier, but the model is not capturing this, as there were no COVID-19 data for training the model.

2 Forecasting evaluation

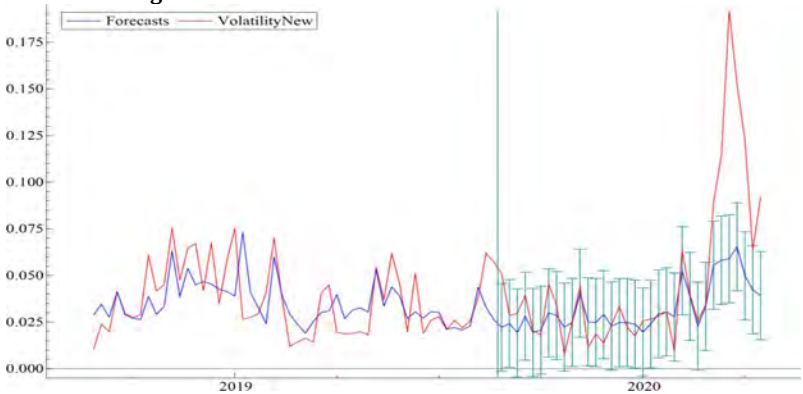


Fig. 10 Forecast graph, Y axis is Weekly volatility. 34 weeks out-of-sample

In Fig. 10, the blue line is the out-of-sample forecast from September 2019 to April 2020 for the last 34 weeks of the dataset. The estimation sample is from April 2015 to August 2019. The green band around the blue line is a 95% band marking +/- 2 forecast standard errors.

From late February 2020 to April 2020, the weekly volatility for the Apple stock took a hit due to the COVID-19 pandemic period, and the forecast model fail to forecast this peak. This is probably due to the COVID-19 pandemic not being captured by the Google searches in this model. The Oxmetrics output for the above 34 weeks dynamic forecast out-of- sample is in Appendix 1, including the COVID-19 pandemic period.

To test how the forecasting performed when excluding the COVID-19 pandemic period, a forecast scenario similar to the setting was tested for a 26-week forecast ending forecast in February 2020 before the pandemic hit the stock market. The difference in forecasting periods for the 34 weeks including the COVID-19 pandemic period and 26 weeks excluding it ensures identical training datasets from week 19 in 2015 to week 34 in 2019 under both forecasting scenarios.

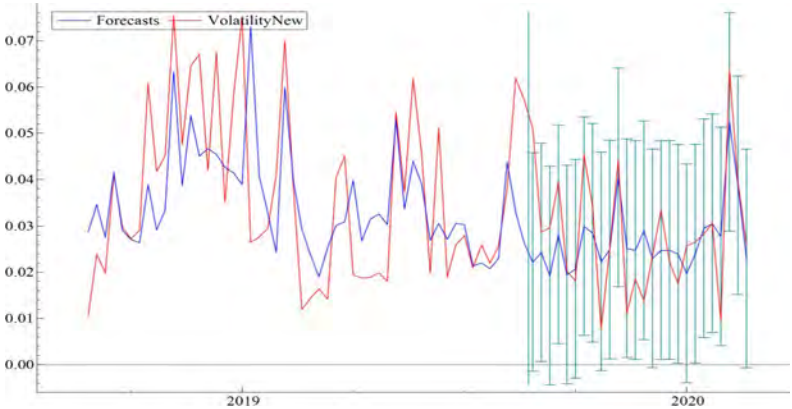


Fig. 11 Forecast graph, Y axis is Weekly volatility. 26 weeks out-of-sample

The Oxmetrics output for the above 26 weeks out-of- sample dynamic forecast is shown in Appendix 2. The Oxmetrics outputs for the two out-of-sample dynamic forecasts without and with the COVID-19 pandemic period are summarized in Table 2.

Excluding the COVID-19 pandemic period		Including the COVID-19 pandemic period	
One-step (ex post) forecast analysis 2019 (35)–2020 (8)		One-step (ex post) forecast analysis 2019 (35) –2020 (16)	
Training dataset 2015 (19)–2019 (34)		Training dataset 2015 (19)–2019 (34)	
Parameter constancy forecast tests:		Parameter constancy forecast tests:	
Forecast Chi^2(26) = 19.244 [0.8259]		Forecast Chi^2(34) = 291.67 [0.0000] **	
Chow F(26,218) = 0.73626 [0.8217]		Chow F(34,218) = 7.5341 [0.0000] **	
CUSUM t(25) = 0.4978 [0.6230]		CUSUM t(33) = 6.580 [0.0000]**	
RMSE =	0.010157	RMSE =	0.034577
MAPE =	38.524	MAPE =	40.205
mean(Error)=	0.001264	mean(Error)=	0.014414
SD(Error)=	0.010078	SD(Error)=	0.031434

Table 2. Autometrics Output excluding & including COVID-19.

Comparing the forecasts without and with the pandemic period, the former performs relatively well in terms of Chi^2, Chow, CUSUM, RMSE, MAPE, and mean(Error).

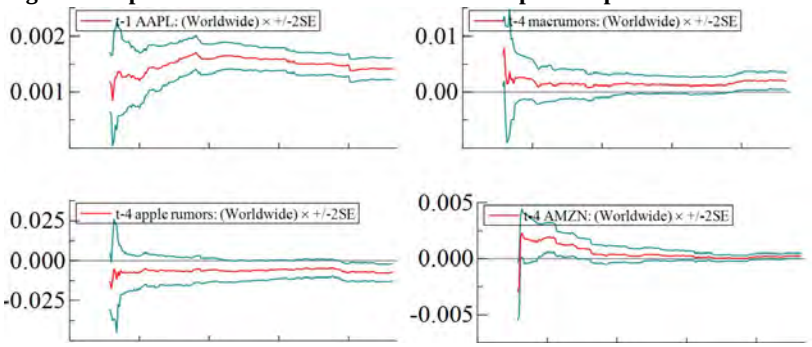
All forecasting was conducted using the model with four predictors selected by Autometrics. For comparing the two used variable selection methods in this article—Autometrics and LASSO—Table 3 presents both methods and their forecasting KPIs.

Software:	OxMetrics 8.10	IBM SPSS Statistics 26
Variable selection method:	**Autometrics**	**Lasso**
Variables selected: t-1 AAPL, t-4 AMZN, t-4 macrumors & t-4 apple rumors		t-1 AAPL & t-4 AMZN
Forecasting excluding the COVID-19 pandemic period		**Forecasting excluding the COVID-19 pandemic period**
One-step (ex post) forecast analysis 2019 (35)–2020 (8)		One-step (ex post) forecast analysis 2019 (35)–2020 (8)
Training dataset 2015 (19)–2019 (34)		Training dataset 2015 (19)–2019 (34)
Parameter constancy forecast tests:		Parameter constancy forecast tests:
Forecast Chi^2(26) = 18.646 [0.8510]		Forecast Chi^2(26) = 19.189 [0.8284]
Chow F(26,218) = 0.71147 [0.8482]		Chow F(26,222) = 0.73701 [0.8210]
CUSUM t(25) = 0.6245 [0.5380] (zero forecast innovation mean)		CUSUM t(25) = 0.3433 [0.7342]
Forecasting including the COVID-19 pandemic period		**Forecasting including the COVID-19 pandemic period**
One-step (ex post) forecast analysis 2019 (35)–2020 (8)		One-step (ex post) forecast analysis 2019 (35)–2020 (8)
Training dataset 2015 (19)–2019 (34)		Training dataset 2015 (19)–2019 (34)
Parameter constancy forecast tests:		Parameter constancy forecast tests:
Forecast Chi^2(34) = 288.54 [0.0000]**		Forecast Chi^2(34) = 278.85 [0.0000]**
Chow F(34,218) = 7.4183 [0.0000]**		Chow F(34,222) = 7.6773 [0.0000]**
CUSUM t(33) = 6.659 [0.0000]** (zero forecast innovation mean)		CUSUM t(33) = 6.536 [0.0000]**

Table 3. Forecasting output from Autometrics & SPSS

7.1 Autometrics recursive graphs

Fig. 12. Graphs of the coefficients for the most important predictors



In Fig. 12, the Google searches for AAPL have the most significant pattern. The coefficient on the Google search AMZN t - 4 is close enough to zero so that it is not worth analyzing it. However, the IBM SPSS LASSO has a much higher partial R^2 on 8% for this variable compared to only 2% in Autometrics. The coefficients for the Google searches for MacRumors t - 4 and Apple rumors t - 4 are larger than forAAPL, but these Google searches are also relative small compared to the AAPL searches

7.2 IBM SPSS 26 output

Model Summary

Multiple R	R Square	Adjusted R Square	Regularizatio n "R Square" (1-Error)	Apparent Prediction Error	Expected Prediction Error		
					Estimate[a]	Std. Error	N[b]
,725	,525	,514	,475	,525	,560	,077	256

Penalty ,420
Dependent Variable: VolatilityNew
Predictors: t-1 AAPL: (Worldwide) t-1 9to5mac: (Worldwide) t-1 theVerge: (Worldwide) t-1 macrumors: (Worldwide) t-1 IBM: (Worldwide) t-1 appleinsider: (Worldwide) t-1 AMZN: (Worldwide) t-1 apple rumors: (Worldwide) t-1 SSNLF: (Worldwide) t-2 AAPL: (Worldwide) t-2 9to5mac: (Worldwide) t-2 theVerge: (Worldwide) t-2 macrumors: (Worldwide) t-2 IBM: (Worldwide) t-2 appleinsider: (Worldwide) t-2 AMZN: (Worldwide) t-2 apple rumors: (Worldwide) t-2 SSNLF: (Worldwide) t-3 AAPL: (Worldwide) t-3 9to5mac: (Worldwide) t-3 theVerge: (Worldwide) t-3 macrumors: (Worldwide) t-3 IBM: (Worldwide) t-3 appleinsider: (Worldwide) t-3 AMZN: (Worldwide) t-3 apple rumors: (Worldwide) t-3 SSNLF: (Worldwide) t-4 AAPL: (Worldwide) t-4 9to5mac: (Worldwide) t-4 theVerge: (Worldwide) t-4 macrumors: (Worldwide) t-4 IBM: (Worldwide) t-4 appleinsider: (Worldwide) t-4 AMZN: (Worldwide) t-4 apple rumors: (Worldwide) t-4 SSNLF: (Worldwide)

a. .632 Bootstrap estimate (50 bootstrap samples).

b. If N is smaller than the number of active (training) cases, this is due to excluding cases from estimation of the expected prediction error for reason(s) explained in the warnings table.

Table 4: Model summary from SPSS Lasso

The LASSO model in IBM SPSS 26 has an R^2 on 52.5%, which is very comparable to the one in Autometrics, around 50%. The above lambdian model yielded 0.42, which is optimized after the 19% bootstrap test set (50 bootstrap samples out of the 260 weekly observations).

Final model

The automatic model selection in Autometrics results in two significant Google searches linked to the weekly volatility of the Apple stock and no effects from the seasonal and event dummies or historical values of volatility. The significant Google searches are AAPL t - 1 and AMZN t - 4.

The Google search **"AAPL"** is time lagged 1 week, being the most significant search linked Apple stock volatility, with a partial R^2 of 49% in Autometrics and 50% in SPSS 26 LASSO.

The Google search AAPL at t - 1, which is the Google search for the Applestock symbol on Nasdaq, is driving up volatility 1 week after thesearches. This is because it is assumed the Google searches for AAPL are mainly done by private investors, under the assumption that professional investors have a lower need to Google search the Apple stock symbol AAPL. These assumptions cannot be proven but are logical.

To verify this conjecture, I interviewed Henrik Ekman, Independent Investment Consultant, former Head of Equities at Maj Invest, on January 28, 2021. He confirmed that professional portfolio analysts are using Google searches to find indications of sales going up and down for the stocks they are analyzing. They also use Google searches in general for information gathering for the stocks currently in their portfolio. While I did not find out any specifics on the difference between private and professional investors in terms of Google searches for stock symbols, this confirmed the professional investors' general use of Google searches, as shown in this articles Investor Journey Model.

The LASSO algorithm was run in IBM SPSS 26 on the exact same dataset, with weekly volatility as the dependent variable and two significant input variables. **AAPL was** time lagged 1 week as the most significant input variable, similar to the Autometrics method, with a similar partial R^2 of 50%. For the second significant input variable, LASSO also chose **AMZN** t – 4, with a partial R^2 of 8%. In Autometrics, AMZN t - 4 had a similar partial R^2.

Autometrics was also tested with a target size of 5% instead of the 1% target size for all other tests. The target size is the t-probability threshold for choosing and eliminating input variables.

The target size for 5% also results in the choice of the MacRumors t - 4 Google search with a partial R^2 of 2.7% and a positive coefficient and Apple rumors t - 4 with a partial R^2 of 3.1% and negative coefficient. Given these relative small partial R^2, further analysis is not necessary. The positive coefficient on MacRumors indicates more private investors conduct these Google searches, while the negative coefficient on Apple rumors indicates more professional investors rely on these Google searches. MacRumors.com has good reputation for vetting Apple rumors, but the Google searches for Apple rumors lead also to MacRumors.com, 9to5mac.com, AppleInsider.com, and TheVerge.com; this wider mix of Apple rumor news sites are used more by professional investors compared to just MacRumors.com. However, given the small partial R^2 of around 3% for these Apple rumor Google searches, this analysis should of course be interpreted cautiously.

Conclusions

More than 60 Apple-related Google searches were tested in this study aspredictors for weekly Apple stock volatility. Under the framework of the newly proposed investor journey model, I analyzed and explained why a buzz in some Apple-related Google searches will dampen the weekly Apple stock volatility and why a buzz in the other searches will increase the weekly volatility for the Apple stock.

A buzz up in the Google search for “AAPL” will increase the Apple stock price volatility in the following week. This is the most significant pattern, since the partial R^2 for the Google search “AAPL” is 44% in Autometrics and 50% in SPSS’s LASSO.

A buzz up in Google search “AMZN” will also increase Apple stock price volatility after 4 weeks. However, the effect is small compared to AAPL, as the partial R^2 for this Google search is just 2% in Autometrics and 8% in SPSS LASSO. With the small partial R^2 of 2–8% for these AMZN Google searches, the results of this analysis should be interpreted with caution.

When the target size changed from 1% to 5% in Autometrics, which is the t-probability threshold for choosing and eliminating input variables, MacRumors t - 4 and Apple rumors t - 4 also become significant, but their partial R^2 are 2.7% and 3.1%, respectively, so their effect is quite small. MacRumors has a positive coefficient and will increase the Apple stock price volatility after 4 weeks. Apple rumors has a negative coefficient and will decrease the Apple stock price volatility after 4 weeks.The explanation could be that an increase in Google searches 4 weeks before this information is available will result in investors buying and selling more. However, it could also be a random pattern,

considering the small partial R^2 of around 3% for these Apple rumor news Google searches.

The most predictive Google search for Apple stock volatility was AAPL t-1. The Google searches for AAPL stock symbol are mostly done by private investors, which explains why their buzz up increases volatility.

When the COVID-19 pandemic disrupted the stock market during February–April 2020, both private and professional investors panicked, which is why and the proposed model could not capture the disruption, as the model was not trained with data from this period.

Further research ideas would be to model the Amazon stock based on Google searches, as the AMZN Google searches show a better visual correlation with the Amazon stock price compared to the Apple stock and AAPL Google searches. Perhaps the stock price return for the Amazon stock has a better potential for being modeled and predicted with Google searches. However, there is the need to test whether the Amazon stock price volatility modeling based on Google searches is stronger than that for Apple in this article. That could require an additional model.

The novel investor journey model presented in this article will thus enable further analyses linking big social data to investor behavior on the financial markets.

References

Asur, S., Huberman, B. A., 2010. Predicting the future with social media. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, pp. 492–499.

Boldt, L. C., Vinayagamoorthy, V., Winder, F., Schnittger, M., Ekran, M., Mukkamala, R. R., Lassen, N. B., Flesch, B., Hussain, A., Vatrapu, R., 2016.Forecasting Nike's sales using Facebook data. 2016 IEEE International Conference on Big Data (Big Data), pp. 2447–2456.

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. *Journal of Computer Science, 2*(1), pp. 1–8.

BusinessInsider.com, April 2018. How Google retains more than 90% of market share". http://uk.businessinsider.com/how-google-retains-more-than-90-of-market-share-2018-4?r=US&IR=T

Buus Lassen, N., la Cour, L., Vatrapu, R., 2017. Predictive analytics with social media data. In Sloan, L. and Quan-Haase, A. (eds). *The SAGE Handbook of Social MediaResearch Methods*, Chapter 20, pp. 328–341, Sage.

Christiansen, C., Schmeling, M., Schrimpf, A., 2012. A comprehensive look atfinancial volatility prediction by economic variables. *Journal of Applied Econometrics*, *27*(6), pp. 956–977.

Diebold, F., Mariano, R., 1995. Comparing predictive accuracy. *Journal ofBusiness & Economic Statistics, 13*(3), pp. 253–263. doi:10.2307/1392185

Federal Reserve Board, 2016. Triennial Survey of Consumer Finance (SCF),https://www.federalreserve.gov/econres/files/BulletinCharts.pdf

Greenwich Associates, 2015. Social media influencing investment decisions at globalinstitutions.

https://www.greenwich.com/press-release/social-media-influencing-investment- deci-sions-global-institutions

Hendry, D., Castle, J., Doornik, J., Johansen, S., Pretis, F., 2014. Model selection with big data. Oxmetrics Conference, September 2014.

http://www.timberlake.co.uk/media/wysiwyg/pdf/David%20F.%20Hendry.pdf

Hendry, D. F., Doornik, J.A., 2014. Empirical model discovery and theoryevaluation: Automatic model selection methods in Econometrics.

Hudson, R., Gregoriou, A., 2010. Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns. Available at SSRN: https://ssrn.com/abstract=1549328 or http://dx.doi.org/10.2139/ssrn.1549328

Jacobsen, P. Ø., Ringberg, T., Bjerre M., CBS & CBS BIG DATA LAB RF2016- 2017 + inter-views 2018 inspired from: Customer Infinity Model – Per Østergaard Jacobsen, Mogens Bjerre, Claus Andersen 2015https://cbs-executive.dk/wp-content/uploads/2019/05/Ten-Deadly-Mar-keting-Sins-30- april-2019-report-and-conclusions.pdf

Jiao, P., Veiga, A., Walther, A., September 25, 2018. Social media, news media and the stock market. Available at: https://ssrn.com/abstract=2755933 or http://dx.doi.org/10.2139/ssrn.2755933

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. *Econometrica, 47*, pp. 263–291.

Lassen, N. B., Madsen, R. and Vatrapu, R., 2014,. Predicting iPhone sales from iPhone tweets. In *2014 IEEE 18th International Enterprise Distributed Object Computing Conference*, IEEE, pp. 81–90.

Levy, M., Levy, H., 2002. Prospect theory: Much ado about nothing?*Management Science*, *INFORMS*, *48*(10), pp. 1334–1349. DOI: 10.1287/mnsc.48.10.1334.276

Levy, H., 2015. *Stochastic Dominance: Investment Decision Making underUncer-tainty.* Springer.

Li, T., van Dalen, J., van Rees, P. J., 2018. More than just noise? Examining the information content of stock microblogs on financial markets. *Journal of InformationTechnology*, *33*(1), pp. 50–69.

MarketWatch.com, 2018. Finance Twitter: The 50 most important people for investors to follow.

https://www.marketwatch.com/story/finance-twitter-the-50-most-important-people-for- inves-tors-to-follow-2018-12-13

Nwogugu, M., 2005. Towards multi-factor models of decision making and risk: Acritique of Prospect Theory and related approaches, part I. *Journal of Risk Finance*, *6*(2), pp. 150–162. https://doi.org/10.1108/15265940510585815

Paye, B.S., 2012. “Déjà vol”: Predictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics*, *106*(3),pp. 527–546.

Preis, T., Moat, H. S., Stanley, H. E., 2013. Quantifying trading behavior in financial markets using Google Trends. *Nature, Scientific Reports*, *3*, p. 1684. DOI:10.1038/srep01684 (2013).

TheGlobeAndMail.com, 2018. How Twitter can help institutional investors market better trading. https://www.theglobeandmail.com/business/careers/business-education/article-how-twitter-can-help-institutional-investors-make-better-trading/

Wakker, P. P., 2010. *Prospect Theory: For Risk and Ambiguity*. Cambridge University Press.

Appendix 1. Forecast output from Oxmetrics, September 2019–April 2020, including the COVID-19 pandemic period.

```
1-step (ex post) forecast analysis 2019(35) - 2020(16)
Parameter constancy forecast tests:
Forecast  Chi^2(34) =   291.67 [0.0000]**
Chow      F(34,218) =   7.5341 [0.0000]**
CUSUM     t(33)     =    6.580 [0.0000]**  (zero forecast innovation mean)

Dynamic (ex ante) forecasts for VolatilityNew (SE based on error variance only)
   Horizon      Forecast          SE        Actual         Error   t-value        -2SE        +2SE
  2019(35)     0.0221757     0.01181     0.0510110     0.028835     2.443  -0.0014354     0.045787
  2019(36)     0.0242622     0.01181     0.0286650    0.0044028     0.373   0.00065115    0.047873
  2019(37)     0.0192232     0.01181     0.0297024     0.010479     0.888  -0.0043879     0.042834
  2019(38)     0.0281562     0.01181     0.0393877     0.011231     0.951   0.0045452     0.051767
  2019(39)     0.0194324     0.01181     0.0202125   0.00078013     0.066  -0.0041787     0.043043
  2019(40)     0.0207155     0.01181     0.0181770   -0.0025385    -0.215  -0.0028956     0.044327
  2019(41)     0.0299563     0.01181     0.0452540     0.015298     1.296   0.0063453     0.053567
  2019(42)     0.0284997     0.01181     0.0341143    0.0056147     0.476   0.0048886     0.052111
  2019(43)     0.0223038     0.01181    0.00787080    -0.014433    -1.223  -0.0013072     0.045915
  2019(44)     0.0248636     0.01181     0.0253074   0.00044384     0.038   0.0012525     0.048475
  2019(45)     0.0405084     0.01181     0.0441805    0.0036721     0.311    0.016897     0.064119
  2019(46)     0.0251504     0.01181     0.0111957    -0.013955    -1.182   0.0015393     0.048761
  2019(47)     0.0247585     0.01181     0.0185567   -0.0062018    -0.525   0.0011474     0.048370
  2019(48)     0.0290163     0.01181     0.0138208    -0.015195    -1.287   0.0054052     0.052627
  2019(49)     0.0229785     0.01181     0.0235305   0.00055208     0.047 -0.00063259    0.046590
  2019(50)     0.0247348     0.01181     0.0334344    0.0086996     0.737   0.0011237     0.048346
  2019(51)     0.0247689     0.01181     0.0222302   -0.0025387    -0.215   0.0011579     0.048380
  2019(52)     0.0239338     0.01181     0.0175467   -0.0063871    -0.541   0.00032276    0.047545
   2020(1)     0.0197210     0.01181     0.0257094    0.0059884     0.507  -0.0038901     0.043332
   2020(2)     0.0239680     0.01181     0.0265275    0.0025596     0.217   0.00035692    0.047579
   2020(3)     0.0294778     0.01181     0.0282960   -0.0011817    -0.100   0.0058667     0.053089
   2020(4)     0.0305566     0.01181     0.0306042   4.7624e-05     0.004   0.0069455     0.054168
   2020(5)     0.0277304     0.01181    0.00949557    -0.018235    -1.545   0.0041193     0.051341
   2020(6)     0.0525129     0.01181     0.0638056     0.011293     0.957    0.028902     0.076124
   2020(7)     0.0388066     0.01181     0.0394267   0.00062008     0.053    0.015196     0.062418
   2020(8)     0.0229329     0.01181     0.0259554    0.0030225     0.256 -0.00067814    0.046544
   2020(9)     0.0334589     0.01181     0.0340971   0.00063827     0.054   0.0098478     0.057070
  2020(10)     0.0554885     0.01181     0.0890427     0.033554     2.842    0.031877     0.079100
  2020(11)     0.0581828     0.01181      0.114265     0.056082     4.750    0.034572     0.081794
  2020(12)     0.0590218     0.01181      0.191717      0.13270    11.240    0.035411     0.082633
  2020(13)     0.0653553     0.01181      0.152223     0.086868     7.358    0.041744     0.088966
  2020(14)     0.0498154     0.01181      0.122598     0.072782     6.165    0.026204     0.073426
  2020(15)     0.0423422     0.01181     0.0638132     0.021471     1.819    0.018731     0.065953
  2020(16)     0.0391658     0.01181     0.0919330     0.052767     4.470    0.015555     0.062777
   mean(Error) =     0.014404    RMSE =     0.034577
   SD(Error)   =     0.031434    MAPE =       40.205
```

Appendix 2. Forecast output from Oxmetrics, September 2019–February2020, excluding the COVID-19 pandemic period.

1-step (ex post) forecast analysis 2019(35) - 2020(8)
Parameter constancy forecast tests:
Forecast Chi^2(26) = 19.244 [0.8259]
Chow F(26,218) = 0.73626 [0.8217]
CUSUM t(25) = 0.4978 [0.6230] (zero forecast innovation mean)

Dynamic (ex ante) forecasts for VolatilityNew (SE based on error variance only)

Horizon	Forecast	SE	Actual	Error	t-value	-2SE	+2SE
2019(35)	0.0221757	0.01181	0.0510110	0.028835	2.443	-0.0014354	0.045787
2019(36)	0.0242622	0.01181	0.0286650	0.0044028	0.373	0.00065115	0.047873
2019(37)	0.0192232	0.01181	0.0297024	0.010479	0.888	-0.0043879	0.042834
2019(38)	0.0281562	0.01181	0.0393877	0.011231	0.951	0.0045452	0.051767
2019(39)	0.0194324	0.01181	0.0202125	0.00078013	0.066	-0.0041787	0.043043
2019(40)	0.0207155	0.01181	0.0181770	-0.0025385	-0.215	-0.0028956	0.044327
2019(41)	0.0299563	0.01181	0.0452540	0.015298	1.296	0.0063453	0.053567
2019(42)	0.0284997	0.01181	0.0341143	0.0056147	0.476	0.0048886	0.052111
2019(43)	0.0223038	0.01181	0.00787080	-0.014433	-1.223	-0.0013072	0.045915
2019(44)	0.0248636	0.01181	0.0253074	0.00044384	0.038	0.0012525	0.048475
2019(45)	0.0405084	0.01181	0.0441805	0.0036721	0.311	0.016897	0.064119
2019(46)	0.0251504	0.01181	0.0111957	-0.013955	-1.182	0.0015393	0.048761
2019(47)	0.0247585	0.01181	0.0185567	-0.0062018	-0.525	0.0011474	0.048370
2019(48)	0.0290163	0.01181	0.0138208	-0.015195	-1.287	0.0054052	0.052627
2019(49)	0.0229785	0.01181	0.0235305	0.00055208	0.047	-0.00063259	0.046590
2019(50)	0.0247348	0.01181	0.0334344	0.0086996	0.737	0.0011237	0.048346
2019(51)	0.0247689	0.01181	0.0222302	-0.0025387	-0.215	0.0011579	0.048380
2019(52)	0.0239338	0.01181	0.0175467	-0.0063871	-0.541	0.00032276	0.047545
2020(1)	0.0197210	0.01181	0.0257094	0.0059884	0.507	-0.0038901	0.043332
2020(2)	0.0239680	0.01181	0.0265275	0.0025596	0.217	0.00035692	0.047579
2020(3)	0.0294778	0.01181	0.0282960	-0.0011817	-0.100	0.0058667	0.053089
2020(4)	0.0305566	0.01181	0.0306042	4.7624e-05	0.004	0.0069455	0.054168
2020(5)	0.0277304	0.01181	0.00949557	-0.018235	-1.545	0.0041193	0.051341
2020(6)	0.0525129	0.01181	0.0638056	0.011293	0.957	0.028902	0.076124
2020(7)	0.0388066	0.01181	0.0394267	0.00062008	0.053	0.015196	0.062418
2020(8)	0.0229329	0.01181	0.0259554	0.0030225	0.256	-0.00067814	0.046544

mean(Error) = 0.0012644 RMSE = 0.010157
SD(Error) = 0.010078 MAPE = 38.524

The New Statutory Audit Framework in Europe: Consistency of Implementation Rationale and Audit Fee Dependence in Denmark?

Claus Holm, Department of Economics and Business Economics, Aarhus University, email hoc@econ.au.dk.

Abstract: Individual EU Member States had options on how they implement the new statutory audit framework in Europe. They could introduce stricter rules or apply certain exemptions where deemed appropriate. Denmark exemplifies Member States with a traditionally high level of non-audit services provided by its auditors. The aim of this study is to contrast the *minimum implementation rationale* observed in the Danish implementation process with an ex ante examination of fee dependency. The audit reform introduced 'cap' and 'blacklist' measures on non-audit fees, which implied a regulator-determined condition of non-independence. In a sample with 2,858 observations, Denmark is compared with Finland, Germany, Sweden and the UK to determine the consistency of implementation rationale and audit fee dependence. The findings support the regulators' concern that auditors of public interest entities with high levels of non-audit services are more likely to have self-interest threats. Findings also suggest, that inconsistent with the implementation rationale, Denmark stands out as a country with greater challenges related to the auditors' provision of non-audit services.

Introduction

A major conflict is at the heart of the provision of non-audit services to public interest entities (PIEs). In addition to the statutory audit, audit firms have traditionally provided an array of services based on their experience and expertise within business-related matters. The value of knowledge spillovers amongst various services provided by audit firms has been stressed by both corporations and the accounting profession (Francis, 2006; Daugherty and Tervo, 2008). Audit market regulators have taken the opposite stance, advocating the better interest of the financial statement users. Regulators overseeing the major financial markets (including the US and European Union) have strenuously questioned the possibility of audit firms retaining independence from, especially, the large PIE clients when these clients pay for both the statutory audit and non-audit services (DeFond and Zhang, 2014; Quick, 2012). In Europe, the recent audit reform has amended the statutory audit Directive (2014/56/EU), and the associated EU Regulation 537/2014 prohibits a number of non-audit services and further cap the volume of allowed non-audit services to PIEs.

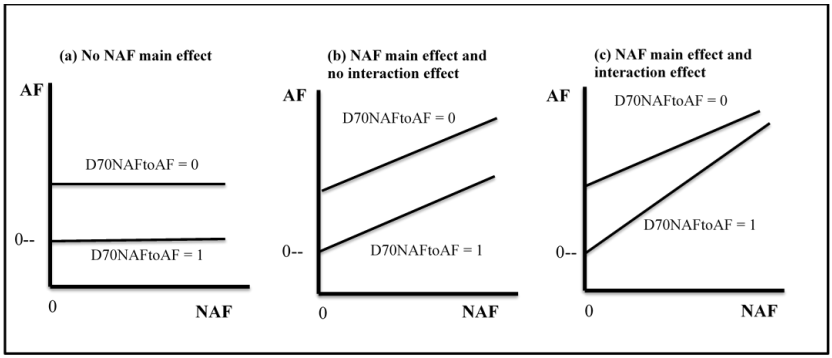
The aim of this study is to contrast the arguments from the Danish implementation process with an examination of fee dependence related to the provision of non-audit services to PIEs prior to the regulation. The aim responds to a general concern for regulatory implications of 'mindless argumentation' across different settings (countries). Denmark is one of

the 28 (27) Member States in which the audit reform was mandated by the European Commission to take effect no later than 16 June 2016. It is possible to find support for an argument that Denmark would be a prime candidate to implement the allowed options of Regulation 537/2014 enforcing *stricter* audit requirements to enhance independence and audit quality. The audit market in Denmark reportedly displays many of the preconditions for auditor independence problems, including increased audit fee competition after the abolition of the joint audit requirement in 2005 (Langer, 2015), infrequent auditor changes and infrequent modified audit reports (Holm and Thinggaard, 2016), absence of strict liability environment (Holm and Warming-Rasmussen, 2008) and a tradition for lax prescriptive regulation, thus allowing liberal access for audit firms to provide consultancy services (Lesage et al., 2017; Ratzinger-Sakel and Schönberger, 2015; Holm and Thinggaard, 2014). Disciplinary rulings related to negligent auditor independence are rare and only one case is mentioned in the support material for the proposed revision of the Danish audit legislation. While this suggests a culture of compliance management, it could also be interpreted as the professions' success in impression management (Holm and Zaman, 2012; Humphrey et al., 2011). In their study, Ratzinger-Sakel and Schönberger (2015, 71) refer to Portugal and Denmark as the countries with exceptionally high non-audit fee ratios amongst the EU Member States. High non-audit fee ratios may reflect particular independence problems involving self-interest threats.

The empirical tests in this study consider the possible fee dependency before the regulation. Specifically, an interaction effect approach is applied to identify whether and to what extent unwanted fee relationships existed. Figure 1 illustrates alternative projected relationships - *ceteris paribus* - between the audit fee (AF) and non-audit fee (NAF). Based on prior research, the absence of a direct relationship (main effect) between AF and NAF is considered indicative that the audits are without independence problems pertaining to economic bonding. The loss-leader argument alone would imply that the absence of a negative association supports this argument while the possibility that the audit firm may retain cost savings from knowledge spillovers implies that a significant main effect in either direction potentially imposes such concern. To highlight the potential differences in the projected relationships, the intersections for the two groups above and below the 70 percent cap in Figure 1 are drawn as being different. Because companies belonging to the group where the cap is exceeded (D70NAFtoAF = 1) by default acquire both audit services and non-audit services, the intersection is depicted as close to zero, whereas the other group (D70NAFtoAF = 0) will have some companies that do not acquire NAF and, thus, have a somewhat higher intersection with the AF axis

.

Figure 1. Projected relationships between AF and NAF



In examining the importance of the cap measure, the existence of a NAF main effect is not the prime issue. The major interest is to see if there is a difference in slope between the two groups. The concern is whether the association between AF and NAF is stronger when a certain threshold is reached. The regulators' suggestion is that above the threshold the company–auditor relationship is tainted with an independence problem. The group of companies with a critical non-audit fee proportion (D70NAFtoAF = 1) will have a higher slope coefficient if the association between AF and NAF is stronger. A significant, stronger association (interaction effect) will suggest that the regulators' threshold is a good indicator of possible independence problems for this group. Hence, the following interaction effect is hypothesised for the pooled international sample:

> **H1:** There is a stronger positive relationship between audit and non-audit fees for companies with a proportion of NAF to AF above the cap level than for companies below the cap level.

The saliency of the *minimum implementation rationale* in the Danish implementation process is contested by examining the influence of the proposed 70% cap measure. In the public hearing process, it was argued that this rationale also was favoured by Denmark's neighbour countries and, thus, by regulators deemed appropriate due to similar circumstances for the transposition of the audit reform into national legislation. However, the preconditions for auditor independence problems seem to vary across the Member States as noted in the introduction. Thus, in an international sample, Denmark is compared with four other countries (Finland, Germany, Sweden and the UK) to determine whether independence issues are similar in nature and/or more pronounced in Denmark than in other Member States otherwise comparable on different dimensions. To test the hypothesis, the same fee model and period are applied as for the international sample. The H2 hypothesis is stated as a null (no difference):

H2: In Denmark, the relationship between audit and non-audit fees for companies with a proportion of NAF to AF above the cap level is similar to the relationship in other countries.

Methodology: Dataset, Model and Variables

The **dataset** contains company financial information and auditor-related data derived from Datastream, Orbis and EUR Business Research. For the years 2010–2013, the total sample population is 2,858 firm-year observations (after excluding financial institutions and observations due to missing variables). The sample observations are based on the four years and the five countries considered. The two largest subsamples are derived from the UK with a total of 915 observations and Germany with 808 observations. In comparison, the number of listed companies in each of the three Scandinavian countries is smaller. Sweden is the largest with 530 observations while the sample for Denmark holds 309 observations and Finland holds 296 observations. Fee dependence is examined using the core **audit fee determinants model** which has evolved from the seminal study by Simunic (1980). In the model audit fees (AF) are regressed on non-audit fees (NAF) as well as control variables, which in prior studies have been considered key determinants of audit fees.

$$AF = \beta_0 + \beta_1 NAF + \beta_2 D70NAFtoAF + \beta_3 NAF^* D70NAFtoAF + \sum \beta_i CONTROL\ VARIABLES_i + \sum \beta_j FIXED\ EFFECTS_j + \varepsilon \quad (1)$$

The **variables** identified in model (1) control for differences in client, auditor and engagement attributes. To test whether the main effect association between AF and NAF is stronger when the regulatory threshold is reached, the NAF variable is interacted with the indicator variable D70NAFtoAF; the critical non-audit fee proportion of 70 percent of audit fees (the cap measure). All fee measures are transformed using the natural logarithm to annual fees measured in thousand euros. The dependent variable in all the models is the fee for the statutory audit (AF) which in equation (1) is specified as a function of NAF (fee for non-statutory audit services); that is, the sum of all fees not related directly to the statutory audit (for example fee for other assurance services, fee for tax services and other consultancy services). The CONTROL VARIABLES can be subdivided into three types of attributes consistent with the typology used in the meta-analysis of prior fee studies by Hay (2013). First, the models include a number of variables that can be identified as *client attributes* related to size, complexity, performance and and risk: the variable SIZE is measured as the natural logarithm of total assets in thousand euros; SQRTSUB is the square root of the number of subsidiaries. GEOSEG is the natural logarithm of the number of geographical segments of the company; PB is the price-to-book ratio; INVREC is the ratio of the sum of inventory and receivables to total assets; ROA is the ratio of EBIT to total assets; LOSS is an indicator variable with a value of one if net income is negative; LEVERAGE

is the ratio of debt to total assets; CAtoCL is the ratio of current assets to current liabilities. The models also include a set of industry indicators to control for industry-fixed effects. The industry indicators are based on the industry classification GICS (Global Industry Classification Standard) sectors. Second, the models include *auditor attributes* related to auditor quality: BIG4 is an indicator variable with the value one if a Big Four audit firm conduct the audit; TENURE is an indicator variable with the value of one if the auditor has conducted the statutory audit of the company for five years or more. Third, the models include *engagement attributes:* In addition to the control variables, it should be noted that the primary exploratory variable NAF can be classified as belonging to this category; BUZYSEASON is an indicator variable with the value of one if the company uses the calendar year as the financial year.

Analyses

The descriptive statistics for the pooled international sample are provided in Table 1. The top part of the table provides an overview of the distribution of the continuous variables which have been winsorised at the 1 and 99 percent levels to restrict unnecessary sensitivity to outliers. Mean and standard deviation of the continuous variables are shown for the total sample (pooled by year and countries). The following may be noted in terms of comparison of the individual countries (not shown in table). In terms of SIZE, the Danish sample has the smallest companies and the UK has the largest. Leverage is, on average, about 52 percent, with Denmark and Sweden slightly below average. The ROA is not consistent across the five countries. That is, the average ROA is 5.6 percent for the pooled sample but ranging from 1.4 percent in Denmark to 9.7 percent in UK for the period 2010–2013. Complexity, as measured by SQRTSUB and GEOSEG, seems to be the highest in Finland while the companies from Germany have low complexity with respect to these client attributes. In relation to the indicator variables shown in the lower part of Table 1, some variation is discernible. The proportion of BIG4 audits is high (on average 84.3 percent). The German percentage is notably lower with 58.2 percent, whereas the Finnish companies in the sample are almost exclusively audited by the big four audit firms. In terms of LOSS, Denmark has the highest proportion (33.3 percent for the period) and the UK has the lowest (11.3 percent). Amongst the five countries, most audits are conducted with the calendar year as the financial year (BUZYSEASON). Denmark is slightly above average while UK as the largest audit market in the sample has a considerably lower proportion of 41.1 percent calendar year audits.

Table1. Descriptive statistics for Denmark and the pooled sample in the period 2010-2013

Variables	Denmark (N=309)		Pooled sample (N=2,858)						
Continuous	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Min**	**p25**	**Median**	**p75**	**Max**
SIZE	12,139	1,972	12,764	2,137	7,126	11,185	12,577	14,128	18,697
LEVERAGE	0,504	0,196	0,525	0,188	0,051	0,402	0,541	0,657	0,951
INVREC	0,251	0,177	0,224	0,181	0,001	0,071	0,192	0,342	0,888
ROA	0,014	0,155	0,056	0,130	-0,919	0,029	0,068	0,112	0,364
CAtoCL	1,883	1,688	1,941	1,899	0,147	1,076	1,478	2,076	23,148
PB	2,034	2,432	2,591	2,893	0,288	1,029	1,723	3,083	27,328
SQRTSUB	3,872	3,361	4,606	3,729	0,000	2,000	3,606	5,831	24,938
GEOSEG	1,448	0,630	1,421	0,614	0,693	0,693	1,609	1,792	4,159
Indicators	**n**	**pct**	**n**	**pct**					
BIG4	292	94,5%	2408	84,3%					
TENURE	214	69,3%	2036	71,2%					
LOSS	103	33,3%	581	20,3%					
BUZYSEASON	238	77,0%	2082	72,8%					

"N" is the number of sample observations for the variable, "n" is the number of positive indicators for the variable (1:0).

Table 2 provides further descriptive statistics related to the fee variables (number of observations, mean and standard deviation). The variables included in the top part of Table 2 are audit fee (AF) and the total non-audit fee (NAF) as well as the three individual components of the latter, i.e. fees for audit-related services (FARS), fees for tax-related services (FTAX) and fees for other services (FOS). In addition, Table 2 shows fees for the combined tax and other services (FTAXOS). Comparing the Danish sample of 309 observations with the sample of 2,549 observations from other countries demonstrates that on average, Danish audit fees are lower (the reported two-sample *t*-test finds a significant difference of means). On average, companies from Denmark have higher non-audit fees (NAF) than those in the other countries. The composition of NAF demonstrates that this difference is caused by having higher fees for tax services and for other services (both significantly higher), whereas fees for audit-related services in Denmark are significantly lower in comparison to FARS for companies in the other countries. In the lower part of Table 2, the indicator variable D70NAFtoAF shows that for 46.6 percent of the observations from Denmark, the ratio of fees for non-audit services to audit fees is above 70 percent. This is a significantly higher proportion than for the group of other countries where the percentage is 29.8. When FARS is excluded from the ratio, the indicator variable D70FTAXOS shows that the proportion of observations above 70 percent drops to 38.5 for Denmark and 19.8 for other countries. Finally, the indicator D70FOS shows that the similar proportion drops to 19.1 and 7.6 percent, respectively, above 70 percent. These three indicators suggest that Denmark, overall, has more companies above the critical proportion of audit fees when compared to the other group of countries in the period 2010–2013. This is evidenced regardless of whether total NAF or the individual fee components are considered.

Table 2. Descriptive fee statistics for Denmark and group of other countries. Univariate tests of difference.

Variables	Denmark			Other Countries			Two-sample t-test of means	
Continuous	**N**	**Mean**	**Std. Dev.**	**N**	**Mean**	**Std. Dev.**	**Difference**	**t-stat**
AF	309	5,258	1,354	2.549	5,682	1,443	0,424	5.1636***
NAF	309	4,837	1,690	2.253	4,519	2,243	-0,318	-2.9713***
FARS	309	1,833	1,958	2.253	2,418	2,461	0,585	4.7614***
FTAX	309	3,391	2,071	2.253	2,833	2,487	-0,558	-4.3274***
FOS	309	3,924	1,840	2.253	2,760	2,487	-1,163	-9.9378***
FTAXOS	309	4,653	1,717	2.253	3,976	2,377	-0,676	-6.1601***
Variables	**Denmark**			**Other Countries**			**Two-sample test of proportions**	
Indicators	**N**	**n**	**pct**	**N**	**n**	**pct**	**Difference**	**z-stat**
D70NAFtoAF	309	144	46,6%	2549	759	29,8%	-0,168	-6.0082***
D70FTAXOStoAF	309	119	38,5%	2253	447	19,8%	-0,187	-7.4188***
D70FOStoAF	309	59	19,1%	2253	171	7,6%	-0,115	-6.5954***

"N" is the number of sample observations for the variable, "n" is the number of positive indicators for the variable (1:0).

Table 3 shows the multivariate audit fee models constructed to examine the hypothesised fee dependences in H1 and H2. The OLS models are effective with an adjusted R-square of 0.88 for the pooled sample model, 0.94 for the sample of Danish companies and 0.87 for the group of other countries. The models are estimated based on robust standard errors; however, the adjusted R-squares are reported to compare explanatory power to previous research, i.e. based on OLS regression before the control for robust standard errors (Wooldridge, 2009). The model is not affected by serious multicollinearity amongst variables (max VIF of 2.92; not shown in table). First, Table 3 provides the results of H1 for the international sample.

Table 3. Multivariate audit fee model

	Expected	Pooled sample		Denmark		Other countries	
	sign	**Coef.**	**t**	**Coef.**	**t**	**Coef.**	**t**
NAF	?	0,1528	9.80***	0,2756	5.12***	0,1505	9.39***
D70NAFtoAF	?	-0,9898	-8.31***	-0,4864	-2.40*	-1,0633	-7.9***
NAF*D70NAFtoAF	?	0,0931	4.55***	0,0094	0.24	0,1032	4.51***
TENURE	+	0,0146	0.49	0,1146	1.81†	0,0027	0.09
BIG4	+	0,0757	1.65†	0,0305	0.32	0,0796	1.63
SIZE	+	0,4219	25.38***	0,3678	8.59***	0,4233	24.16***
LEVERAGE	+	0,2016	1.95†	-0,0049	-0.02	0,2284	2.07*
INVREC	+	0,1508	1.29	0,6237	3.34***	0,0861	0.68
ROA	-	-0,0622	-0.49	-0,1069	-0.53	-0,0860	-0.63
LOSS	+	0,0464	1.37	-0,0643	-1.00	0,0584	1.54
CAtoCL	-	-0,0230	-2.33*	-0,0646	-2.26*	-0,0194	-1.91†
PB	+	0,0099	2.41**	0,0112	0.85	0,0095	2.20*
SQRTSUB	+	0,0426	6.78***	0,0485	2.96**	0,0416	6.31***
GEOSEG	+	0,1838	7.66***	0,2095	3.36***	0,1651	6.48***
BUZYSEASON	+	0,1142	2.85**	0,0826	0.89	0,1286	2.93**
CONSTANT	?	-1,2185	-5.65***	-1,4049	-2.93**	-1,2528	-5.41***
YEAR FE		Y		Y		Y	
INDUSTRY FE		Y		Y		Y	
COUNTRY FE		Y		N		Y	
N		2858		309		2549	
F-value		250.49***		106.22***		224.85***	
Adj. R-squared		0,8785		0,9382		0,8737	

The reported F-Ratio is the robustly estimated variance matrix (Wald test). Two-tailed p-values are indicated by *** 0.001, **0.01, *0.05, † 0.10

D70NAFtoAF is an indicator variable with the value of one if the ratio of fees for non-audit services to audit fees is above 70 percent, and zero otherwise

The pooled five-country sample demonstrates that the fee dependence is best depicted as illustration (c) in Figure 1. *Ceteris paribus*, the intercepts for the two groups are different as suggested by the significantly negative coefficient for D70NAFtoAF (−0.9898). The significantly positive interaction term NAF*D70NAFtoAF suggests support for H1, i.e. *that there is a stronger positive relationship between audit and non-audit fees for companies with a proportion of NAF to AF above the cap level than it is for companies below the cap level*. The latter supports regulators' concern that companies with high levels of non-audit services may have an economic dependence problem or at least face a problem with independence in appearance.

Second, the results shown in the separate models for the sample of Danish companies and sample of companies for the remaining four countries suggest that H2 may be rejected. While the interaction term NAF*D70NAFtoAF is positive and significant (*t*-statistic is 4.51) for the group of other countries, this term is not significant for Denmark; i.e. *the relationship between audit and non-audit fees for companies with a proportion of NAF to AF above the cap level in Denmark is not similar to the relationship in other countries*. The interaction effect is present across all countries except for Denmark. Here, the positive relationship between AF and NAF is consistent both below and above the future cap level at 70 percent, thus indicating a general economic bonding across all companies despite this threshold. One possible implication would be that the cap is set too high in the Danish context.

To **further test the robustness** of H1 and H2, an alternative approach is applied to the generic audit fee model. The results (not shown in table) are derived from testing a model similar to a Difference-in-Difference design with control for covariates. The 'treatment effect' is proxied by the cap measure D70NAFtoAF. The period considered is before the introduction of the cap measure. Therefore, the treatment identifies those companies who under the future policy could be labelled as 'transgressors'. In a classical DID design, the model tests whether a particular treatment (policy change) can be observed when comparing the groups over time. However, the design used here study tests whether Denmark stands out as compared with the group of countries (instead of testing pre- and post-treatment). The covariates in the audit fee model (control variables) are introduced to reduce confounding (omitted variable) bias and to reduce residual variance in the model. In the model, the exploratory variables are (1) D70NAFtoAF measuring the main effect of critical proportion (above and below), (2) DK measuring a main country effect (Denmark vs other countries) and (3) D70NAFtoAF*DK measuring the interaction effect of critical proportion on DK. The OLS model is effective with an adjusted R-square of 0.87 for the model comparing Denmark (DK) with the group of other countries (pooled sample) and also effective when considering the comparison of Denmark with the individual countries (0.87 to 0.89). Overall, the model supports the main findings that the NAF dependency is different for the groups currently above and below the critical proportion and that Denmark stands out by exhibiting the same dependency above as below the 70 percent cap. The fee comparison of

companies from Denmark with companies from the individual countries suggests that Denmark is more similar to Germany and Sweden than to Finland and UK.

Conclusion

In this study, implications of the regulatory harmonisation process are explored using Denmark as an example of an EU Member State with a tradition for allowing higher levels of non-audit services. The implementation process demonstrates that a *minimum implementation rationale* has been dominant in Denmark. The audit reforms' introduction of the 'blacklist' and 'capping' measures related to the auditors' provision of the non-audit services has established a predefined threshold of 'assumed non-independence'. In a sample with 2,858 observations, Denmark is compared with Finland, Germany, Sweden and the UK to determine the consistency of implementation rationale and audit fee dependence. The findings support the regulators' concern that auditors of public interest entities with high levels of non-audit services are more likely to have self-interest threats. Findings also suggest that the implications of the harmonisation process will be different across countries. Inconsistent with the implementation rationale, Denmark stands out as a country with greater challenges related to the auditors' provision of non-audit services.

References

Daugherty, B. E. and Tervo, W. A. (2008) 'Auditor changes and audit satisfaction: Client perceptions in the Sarbanes-Oxley era of legislative restrictions and involuntary auditor change', *Critical Perspectives on Accounting,* 19**,** 931-951.

DeFond, M. L. and Zhang, J. (2014) 'A review of archival auditing research', *Journal of Accounting & Economics,* 58**,** 275-326.

Francis, J. R. (2006) 'Are Auditors Compromised by Nonaudit Services? Assessing the Evidence', *Contemporary Accounting Research,* 23**,** 747-760.

Hay, D. (2013) 'Further Evidence from Meta-Analysis of Audit Fee Research', *International Journal of Auditing,* 17**,** 162-176.

Holm, C. and Thinggaard, F. (2014) 'Leaving a joint audit system: conditional fee reductions', *Managerial Auditing Journal,* 29**,** 131-152.

Holm, C. and Thinggaard, F. (2016) 'Paying for Joint or Single Audits? The Importance of Auditor Pairings and Differences in Technology Efficiency', *International Journal of Auditing,* 20**,** 1-16.

Holm, C. and Warming-Rasmussen, B. (2008) An Account of Accountants - Audit Regulation and the Audit Profession in Denmark. IN Quick, R., Turley, S. and Willekens, M. (Eds.) *Auditing, Trust & Governance - Regulation in Europe.* EARNet, Routledge.

Holm, C. and Zaman, M. (2012) 'Regulating audit quality: Restoring trust and legitimacy', *Accounting Forum,* 36**,** 51-61.

Humphrey, C., Kausar, A., Loft, A. and Woods, M. (2011) 'Regulating Audit beyond the Crisis: A Critical Discussion of the EU Green Paper', *European Accounting Review,* 20**,** 431-457.

Langer, M. W. (2015) 'Revisorkrigen koster revisorer årstab på 200 mio', *Økonomisk Ugebrev***,** 1 and 11-15.

Lesage, C., Ratzinger-Sakel, N. V. S. and Kettunen, J. (2017) 'Consequences of the Abandonment of Mandatory Joint Audit: An Empirical Study of Audit Costs and Audit Quality Effects', *European Accounting Review,* 26**,** 311-339.

Quick, R. (2012) 'EC Green Paper Proposals and Audit Quality', *Accounting in Europe,* 9**,** 17-38.

Ratzinger-Sakel, N. V. S. and Schönberger, M. W. (2015) 'Restricting Non-Audit Services in Europe – The Potential (Lack of) Impact of a Blacklist and a Fee Cap on Auditor Independence and Audit Quality', *Accounting in Europe,* 12**,** 61-86.

Simunic, D. A. (1980) 'The Pricing of Audit Services: Theory and Evidence', *Journal of Accounting Research,* 18**,** 161-190.

Wooldridge, J. M. (2009) *Introductory Econometrics: A Modern Approach,* Ohio, USA, South-Western Cengage Learning.

Is stress regionally persistent?

Jørgen T. Lauridsen, Department of Economics, University of Southern Denmark, jtl@sam.sdu.dk

United Nations Sustainable Development Goal 8:

Promote sustained, inclusive and sustainable economic growth, full and productive employment *and decent work for all* (https://sdgs.un.org/)

Abstract

Stress is known to be a widespread disease, which is caused by several reasons, and which is known to be intriguing to remedy or prevent. The present paper applies spatial regression based on spatially aggregated data for 98 Danish municipalities with repeated observation for the three years 2010, 2013 and 2017. The present paper confirms what is known by demonstrating that unhealthy lifestyle and social burden increase the local level of stress, that stress is intriguing to prevent, and that stress is increasing over time from 2010 to 2017. Furthermore, the paper adds new knowledge by showing a strong local persistency, in the sense that the local level of stress is increased by high stress level in the surrounding neighborhood. Apart from indicating persistency, such spatial patterns are furthermore indicative of an increasing geographical inequality in stress, in the sense that deprived neighborhoods with high stress levels are even further deprived. Complexity of stress as a persistent and growing problems is demonstrated, which leads to a policy recommendation of not only considering stress as a micro-level phenomenon between individuals and their micro-environment, but also consider implications of the macro-level in terms of regions and neighborhoods.

Keywords:

Stress; spatial spillover; spatio-temporal models

JEL classifications:

C13, C21, C23, I10, I12, I14

1. Introduction

The incidence and prevalence of work and life stress and derived diseases is increasing and among the major and increasing threats against health and quality of life. Thus, from being number 10 on the 2001 list of such threats, stress and related diseases is expected to jump to a second place by 2020 (WHO, 2001).

For the case of Denmark, the proportion with a high stress level rose from 2010 to 2017, except for males above 75 and females above 65 (DHA, 2018). In 2017, the proportion

was 29 percent for female and 21 percent for male, while the highest figures were 40.5 percent for females aged 16-24, 47 percent for unemployed, 55.6 percent for disability retired, and 54.6 percent for other groups outside the labor market (op. cit.).

Stress is obviously related to workplace conditions like lack of control over own work situation and workload, and it is further strengthened by a number of other negative features including job uncertainty; uncertainty about roles and responsibilities; burnout; lack of fairness and broadness; lack of collegial support, recognition and collaboration; lack of work and life balance; and negative experience alike sexual harassment, bullying, violence and unwanted attention (NFA, 2018).

However, stress is not only caused by workplace conditions. To the contrary, stress is well known to be related to several socioeconomic and lifestyle factors including unemployment, job loss and job uncertainty (Rugulies et al. 2010). Poor neighborhood conditions like criminality, highly trafficked roads, poor school conditions, lack of opportunities for leisure activities and good social interaction causes stress (Maggi et al. 2005). Lifestyle factors like alcohol consumption, smoking habits and drug abuse are known to be related to stress (Kriegbaum et al. 2011a-b, Kjøller and Rasmussen 2006) alike physical inactivity and obesity (NHA, 2018). Childhood poverty is known as a determinant for stress later in life (Klein et al. 2007,Yanos et al. 2007,Evans and Kim 2007). Lack of education is related to stress, as the proportion of stressed among those with only basic school is almost double the corresponding figure for high educated (DHA, 2018). Civil status matters, as married and cohabitated feel less stressed than divorced and widowed (DHA, 2018). Although stress is not in itself a disease, it is known to be strongly related to several diseases like hypertension (Liu et al. 2017), cardiovascular diseases (Kivimaki et al. 2012, Nabi et al. 2013, Rosengren et al. 2004), depression (Rugulies et al. 2006) and longstanding illness (NHA, 2018).

Although the preventive healthcare system should play an increasing active role, it is surprising to see indications of the opposite. Thus, for the case of Denmark, the proportion of stressed who consulted a GP and were advised to take it easy dropped slightly from 2010 to 2017 (NHA, 2018).

Given that several of the above conditions obviously relates to local environment and neighborhood, one should readily expect local environment and neighborhood characteristics to matter as sources for regional variation in stress. However, apart from simple descriptive studies merely just summarizing regional variation in work and life related stress, the author was not able to find studies of relevance. As a single exception, Wang et al. (2015) relates socioeconomic inequality to stress level for 21 Chinese cities in a multilevel study. For the case of Denmark, descriptive figures by 2017 showed that more

inhabitants of the regions of Zeeland and Southern Denmark than of the Capitol Region were stressed, while the opposite were the case for the regions of Central Denmark and Northern Jutland (DHA, 2018).

2. Methods

The point of departure is a linear regression model defined for the *N*=98 Danish municipalities in a single year by

(1) $$y_t = X_t\beta + \upsilon_t, \quad \upsilon_t \sim N(0, \sigma^2 I)$$

where X_t is an *N* by *K* dimensional matrix of the *K* explanatory variables, y_t an *N* dimensional vector of the stress rates in the municipalities, and β a *K* dimensional coefficient vector measuring the effects of the explanatory variables on the stress rate. The term υ_t is a residual term, which represents the fertility rates when controlled for the explanatory factors of X_t and may be denoted the residual stress rate.

Operationally, endogenous (learning) spatial spillover is controlled for by adding the average of y_t in the neighbourhood municipalities (denoted by y_t^W) as an explanatory variable in (1) to obtain the *spatially autoregressive* (SAR) specification (Anselin, 1988)

(2) $$y_t = y_t^W\lambda + X_t\beta + \upsilon_t,$$

where λ is a parameter specifying the magnitude of spill-over, formally restricted to the interval between (-1) and (+1), but for most practical purposes restricted to be positive.

Alternatively, any kind of spatial clustering, including observed as well as unobserved exogenous spatial spillover, may be controlled for by applying the spatially autocorrelated (SAC) specification (Anselin 1988)

(3) $$y_t = X_t\beta + \varepsilon_t, \ \varepsilon_t = \lambda W\varepsilon_t + \upsilon_t.$$

The SAC approach will be applied to investigate whether spatial spillover of stress rates is merely ascribed to exogenous structures rather than being of an endogenous learning nature.

One further methodological problem needs attention for the SAR as well as the SAC specification. While pooled data for *T*=2 years are applied, the residual stress rates across years for any municipality are correlated. Also, the variance of the residual stress rate within each year may potentially vary across years. Thus, between any two years, the covariance of the residual stress rate reads as

(4) $$E(\upsilon_t'\upsilon_s) = \sigma_{ts}^2 \quad t, s = 1,..,T.$$

To obtain efficient estimates of β, we apply Feasible Generalised Least Squares (F-GLS) estimation as suggested by Zellner (1962) to obtain Seemingly Unrelated Regression (SUR) estimates for β. By integrating (4) into any of (1) through (3), SUR, SAR-SUR and SAC-SUR specifications are obtained.

3. Data

Table 1 provides an overview of the data used for the study. Data are collected for 98 municipalities for the years 2010, 2013 and 2017 from three different sources, The Danish Health Profile, The Danish Ministry of Taxation, and the Key Figure Base at The Ministry of the Interior. The outcome variable is percentage feeling highly stressed, defined as belonging to the upper 20 percent on the Cohen's Perceived Stress Scale (PSS) with 10 items (Cohen et al. 1983, Eskildsen et al. 1015). According to theory, literature and availability, a variety of explanatory variables affecting stress were included. As an indication of change over time, dummies for 2013 and 2017 are defined. Health behaviour includes physical activity, alcohol, unhealthy food habits and obesity, while health is captured by Self-Assessed Health (SAH), ill physical health and ill mental health. Contact to primary healthcare is measured by percentage reporting been in contact with a general practitioner (GP) within last 12 months. Social relations are represented by one variable (being unwanted alone), and age structure by percentages of 25-64 and 65+ year. Quality of workplace and control over own work situation is measured by educational level. Income level is measured by tax deductible income per inhabitant. Finally, presence or absence of social distress and neighbourhood deprivation is captured by a variety of variables (urbanisation, social housing, unemployment, non-Western inhabitants, criminal activity, social benefit receivers, and peripheral area).

Table 1. Description of data.

Variable	Description	Mean	SD
Stress [1]	% feeling highly stressed (upper 20% on the PSS-10 scale)	13.29	2.76
PhysAct [1]	% reporting being physically active in leisure time	27.19	3.30
DailySmoker [1]	Percentage reporting being daily smokers	18.71	3.58
Alcohol [1]	% reporting drinking too much (21/14 glasses weekly for M/F)	8.47	2.33
UheFood [1]	% reporting easting unhealthy food daily	15.08	3.77
Obese [1]	% reporting being obese (BMI>30)	51.22	5.66
SAH [1]	% reporting good / very good Self Assessed Health	83.85	2.80
IllPhysHealth [1]	% reporting ill physical health	11.39	2.45
IllMentalHealth[1]	% reporting ill mental health	10.68	2.18
GP	% reporting been in contact with GP within last 12 months		
UnwantedAlone[1]	% reporting being unwanted alone daily	5.62	1.09
P25_64 [3]	% population between 25 and 64	51.19	1.88
P_65 [3]	% population 65 and above	19.69	3.87
Educ [3]	% of population with higher education	24.48	8.76
Income [3]	Tax deductible income per inhabitant (100,000 DKK, 2010 level)	1.72	0.34
Urban [3]	% population living in urban area	83.22	12.77
SocHous [3]	% of population living in social housing	18.05	12.01
Unemp [3]	% of population unemployed	4.01	1.12
NonWest [3]	% of population being from non-Western country	3.10	1.92
SiCrime [3]	Simple crimes reported per 1,000 inhabitants	48.48	19.82
ViCrime [3]	Violence crimes reported per 1,000 inhabitants	1.57	0.62
SocBen [3]	% population receiving social benefits	4.12	1.25
Peripheral [2]	Indicator for peripheral municipality	0.15	0.36
year13	Indicator for year 2013 (reference year 2010)	0.33	0.47
year17	Indicator for year 2017 (reference year 2010)	0.33	0.47

Source: [1] The Danish Health Profile (www.danskernessundhed.dk), [2] The Danish Ministry of Taxation (www.skm.dk), and [3] the Key Figure Base (www.im.dk)

4. Results

Table 2 shows results from two different spatial models. First and foremost, the significantly positive spillover parameters demonstrates a strong local persistency, in the sense that the local level of stress is increased by high stress level in the surrounding neighborhood. Apart from indicating persistency, such spatial patterns are furthermore indicative of an increasing geographical inequality in stress, in the sense that deprived neighborhoods with high stress levels are even further deprived.

Table 2. Estimated spatial models

	SUR			SAC-SUR			SAR-SUR		
Variable	**Coef**	**Stderr**	**P**	**Coef**	**Stderr**	**P**	**Coef**	**Stderr**	**P**
Constant	18.037	0.041	0.04	17.902	8.731	0.04	19.257	8.564	0.02
PhysAct	0.035	0.433	0.43	0.020	0.041	0.62	0.03	0.043	0.49
DailySmoker	0.192	<0.001	<0.01	0.169	0.046	<0.01	0.185	0.046	<0.01
Alcohol	0.132	0.009	0.01	0.096	0.052	0.06	0.136	0.050	0.01
UheFood	0.002	0.949	0.94	-0.015	0.037	0.68	0.026	0.035	0.46
Obese	0.005	0.872	0.87	0.003	0.034	0.92	0.013	0.032	0.68
SAH	-0.145	0.042	0.04	-0.147	0.067	0.02	-0.134	0.069	0.05
IllPhysHealth	-0.131	0.086	0.08	-0.121	0.073	0.09	-0.125	0.074	0.09
IllMentalHealth	0.592	<0.001	<0.01	0.567	0.057	<0.01	0.577	0.060	<0.01
GP	-0.013	0.699	0.69	-0.015	0.035	0.67	-0.013	0.033	0.69
UnwantedAlone	0.210	0.010	0.01	0.182	0.078	0.02	0.192	0.080	0.02
P25_64	-0.103	0.171	0.17	-0.057	0.078	0.46	-0.178	0.075	0.02
P_65	-0.209	<0.001	<0.01	-0.192	0.051	<0.01	-0.239	0.049	<0.01
Educ	0.017	0.494	0.49	0.015	0.025	0.55	0.028	0.024	0.24
Income	1.186	0.017	0.01	0.960	0.504	0.05	1.063	0.484	0.03
Urban	-0.002	0.878	0.87	0.001	0.011	0.93	-0.008	0.011	0.48
SocHous	0.009	0.358	0.35	0.015	0.010	0.13	0.011	0.009	0.25
Unemp	0.256	0.007	0.01	0.221	0.100	0.02	0.253	0.093	0.01
NonWest	-0.076	0.910	0.91	-0.042	0.713	0.95	-0.266	0.659	0.69
SiCrime	0.002	0.715	0.71	0.002	0.006	0.72	0.004	0.006	0.55
ViCrime	0.333	0.010	0.01	0.379	0.125	<0.01	0.408	0.127	<0.01
SocBen	0.029	0.785	0.78	-0.034	0.105	0.74	0.003	0.102	0.97
Peripheral	-0.415	0.061	0.06	-0.364	0.212	0.08	-0.459	0.218	0.03
year13	1.422	<0.001	<0.01	1.405	0.376	<0.01	1.286	0.328	<0.01
year17	3.583	<0.001	<0.01	3.62	0.574	<0.01	2.771	0.586	<0.01
Spillover				0.359	0.123	<0.01	0.195	0.044	<0.01
LogL	-144.0			-133.2			-134.3		

Furthermore, the results confirm – with small variations across specifications – what is known by demonstrating that unhealthy lifestyle and social burden increase the local level of stress, that stress is intriguing to prevent, and that stress is increasing over time from 2010 to 2017. Complexity of stress as a persistent and growing problems is thus demonstrated, which leads to a policy recommendation of not only considering stress as a micro-level phenomenon between individuals and their micro-environment, but also consider implications of the macro-level in terms of regions and neighborhoods.

5. Conclusion

The study confirms what is known by demonstrating that unhealthy lifestyle and social burden increase the local level of stress, that stress is intriguing to prevent, and that stress is increasing over time from 2010 to 2017. Furthermore, the paper adds new knowledge by showing a strong local persistency, in the sense that the local level of stress is increased by high stress level in the surrounding neighborhood. Apart from indicating persistency, such spatial patterns are furthermore indicative of an increasing geographical inequality in stress, in the sense that deprived neighborhoods with high stress levels are even further deprived. Complexity of stress as a persistent and growing problems is demonstrated, which leads to a policy recommendation of not only considering stress as a micro-level phenomenon between individuals and their micro-environment, but also consider implications of the macro-level in terms of regions and neighborhoods.

References

Anselin L. 1988. *Spatial Econometrics: Methods and Models*. Kluwer Academics, Dordrecht

Kjøller M, Rasmussen NK. 2006. *Sundhed og sygelighed i Danmark 2006 og udviklingen siden 1987* [Danish Health and Morbidity Survey 2005 and trends since 1987]. Copenhagen: National Institute of Public Health.

Maggi S, Irwin LG, Siddiqi A, Poureslami I, Hertzman E, Hertzman C. 2005. *Knowledge Network for Early Child Development. Analytic and Strategic Review Paper: International Perspectives on Early Child Development.* World Health Organization's Commission on the Social Determinants of Health.

Klein H, Elifson KW, Sterk CE. 2007. Childhood neglect and adulthood involvement in HIV-related risk behaviors. *Child Abuse Negl* 31(1):39-53.

Yanos PT, Czaja SJ, Widom CS. 2010. A prospective examination of service use by abused and neglected children followed up into adulthood. *Psychiatr Serv* 61(8):796-802.

Evans GW, Kim P. 2007. Childhood poverty and health: cumulative risk exposure and stress dysregulation. *Psychol Sci* 18(11):953-957.

Rugulies R, Thielen K, Nygaard E, Diderichsen F. 2010. Job insecurity and the use of antidepressant medication among Danish employees with and without a history of prolonged unemployment: a 3.5-year follow-up study. *J Epidemiol Community Health* 64(1):75-81.

Kriegbaum M, Christensen U, Osler M, Lund R. 2011a. Excessive drinking and history of unemployment and cohabitation in Danish men born in 1953. *Eur J Public Health* 21(4):444-8 https://doi.org/10.1093/eurpub/ckq152.

Kriegbaum M, Larsen AM, Christensen U, Lund R, Osler M. 2011b. Reduced probability of smoking cessation in men with increasing number of job losses and partnership breakdowns. *J Epidemiol Community Health* 65(6):511-6, https://doi.org/10.1136/jech.2009.100446.

DHA (The Danish Health Authority). 2018. *Danskernes Sundhed – Den Nationale Sundhedsprofil 2017*. Copenhagen: Danish Health Authority,

https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwiNg-IXSg6vnAhUx-uqQKHY0NBQsQFjABegQIBhAB&url=https%3A%2F%2Fwww.sst.dk%2Fda%2Fudgivelser%2F2018%2F~%2Fmedia%2F73EADC242CDB46BD8ABF9DE895A6132C.ashx&usg=AOvVaw33zxXFg-SIcH9dixAITGpAp.

Liu MY, Li N, Li WA, Khan H. 2017. Association between psychosocial stress and hypertension: a systematic review and metaanalysis. *Neurol Res* 39: 573-580.

Kivimaki M, Nyberg ST, Batty GD, Fransson EI, Heikkila K, Alfredsson L, et al. 2012. Job strain as a risk factor for coronary heart disease: a collaborative meta-analysis of individual participant data. *Lancet* 380: 1491-7.

Nabi H, Kivimaki M, Batty GD, Shipley MJ, Britton A, Brunner EJ, et al. 2013. Increased risk of coronary heart disease among individuals reporting adverse impact of stress on their health: the Whitehall II prospective cohort study. *Eur Heart J* 34: 2697-705.

Rosengren A, Hawken S, Ounpuu S, Sliwa K, Zubaid M, Almahmeed WA, et al. 2004. Association of psychosocial risk factors with risk of acute myocardial infarction in 11119 cases and 13 648 controls from 52 countries (the INTERHEART study): casecontrol study. *Lancet* 364: 953-62.

Rugulies R, Bultmann U, Aust B, Burr H. 2006. Psychosocial work environment and incidence of severe depressive symptoms: Prospective findings from a 5-year follow-up of the Danish work environment cohort study. *Am J Epidemiol* 163: 877-87.

Cohen S, Kamarck T, Mermelstein R. 1983. A global measure of perceived stress. J *Health Soc Behav* 24: 385-96.

Eskildsen A, Dalgaard VL, Nielsen KJ, Andersen JH, Zachariae R, Olsen LR, et al. 2015. Cross-cultural adaptation and validation of the Danish consensus version of the 10-item Perceived Stress Scale. *Scand J Work Env Health* 41: 486-90.

WHO. 2001. *Mental Health - A Call for Action by World Health Ministers*. Geneva: WHO. https://www.mhinnovation.net/resources/mental-health-call-action-world-health-ministers.

NFA - The National Research Center for the Working Environment. 2018. *Fakta om Arbejdsmiljø og Helbred 2018*. Copenhagen: NFA, https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=2ahUKEwjruciOn6vnAhWFoFwKHZTIDm4QFjABegQIBBAB&url=https%3A%2F%2Fnfa.dk%2F-%2Fmedia%2FNFA%2FArbejdsmiljodata%2FFakta-om-Arbejdsmiljo-og-Helbred-2018.ashx%3Fla%3Dda&usg=AOvVaw1VK33REKD0fz3pmxwPloqW.

Wang H, Yang XY, Yang T, Cottrell RR, Yu L, Feng Y, JiangS. 2015. Socioeconomic inequalities and mental stress in individual and regional level: a twenty one cities study in China. *International Journal for Equity in Health* 14:25, https://doi.org/10.1186/s12939-015-0152-4.

Zellner A. 1962. An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests of Aggregation Bias. *Journal of the American Statistical Association* 58: 977-992.

PTSD in school-age children: A nationwide prospective birth cohort study

Mogens Nygaard Christoffersen and Anne Amalie Elgaard Thorup

(Revision 10th December 2021)

Abstract

Traumatic childhood events are one of the few identifiable and to some extent preventable causes of psychiatric illness. Children exposed to severely stressful events may react with PTSD and this may impact their daily life level of function, their future development and mental health.

The traumatic stress model suggests that child maltreatment, community violence, accidents, and other traumas are regarded as an additive environmental factor, which can over-weight protective, compensatory factors and interact with predisposition for complex developmental disorders.

The study is based on prospective panel data including the whole population of children born in Denmark from 1984 to 1994 followed from age 7 to age 18 (N=679,000). Risk factors (child maltreatment, community violence, individual vulnerabilities) and compensatory factors (social support) are analyzed by the discrete time-Cox-model.

We found a lifetime prevalence in the range of 2.3 % PTSD in school-age children (n=15,636). Child maltreatment, victims of violent crime, parental violence and individual vulnerabilities especially autism (OR 7.08) and ADHD (OR 10.7) were predicative of PTSD. Our results were consistent with the traumatic stress model. Substance abuse was associated with PTSD.

Many potentially traumatic events were not reported in hospital or administrative records, consequently, PTSD may be underestimated in this study.

Introduction

Traumatic childhood events are one of the few identifiable causes of psychiatric illness (Kerns, Newschaffer, & Berkowitz, 2015). When children and adolescents are exposed to traumatic events considered as extreme life stressors outside the range of normal human experience, the societal, interpersonal, and psychological consequences can be considerable (Keane & Barlow, 2004; Koss, Koss, & Woodruff, 1991; Kulka et al., 1990; Toth, Gravener-Davis, Guild, & Cicchetti, 2013). Post-traumatic stress disorder (PTSD) in school-age children acquire that children are exposed to or witness a stressor that the individual perceives as threatening to physical and/or psychological integrity of self (Blacker, Frye, Morava, Kozicz, & Veldic, 2019). PTSD is defined from a specific set of well described symptoms (e.g. flashbacks, nightmares, avoidance, memory lapses, emotional numbing and hypervigilance, DSM-5: 309.81 or WHO: F43.1) that persists for more than a month and impacts the individual's functioning (American Psychiatric Association, 2013; Kerns et al., 2015).

The purpose of this study is to investigate both factors of resilience and constraining factors' influence on PTSD symptoms in school-age children. Anxiety is very common as a comorbid condition to neurodevelopmental disorders like autism and Attention-Deficit Hyperactivity Disorder (ADHD) (Green & Ben-Sasson, 2010; Tannock, 2009) and is also very common in combination with e.g. depression, eating disorders and psychosis.

Methodological challenges

(a) A relatively low prevalence of potential risk factors such as child maltreatment and family violence.

(b) A low prevalence of individual extreme stressors such as vehicular accidents, industrial accidents, war, rape, torture, terrorism, natural disasters.

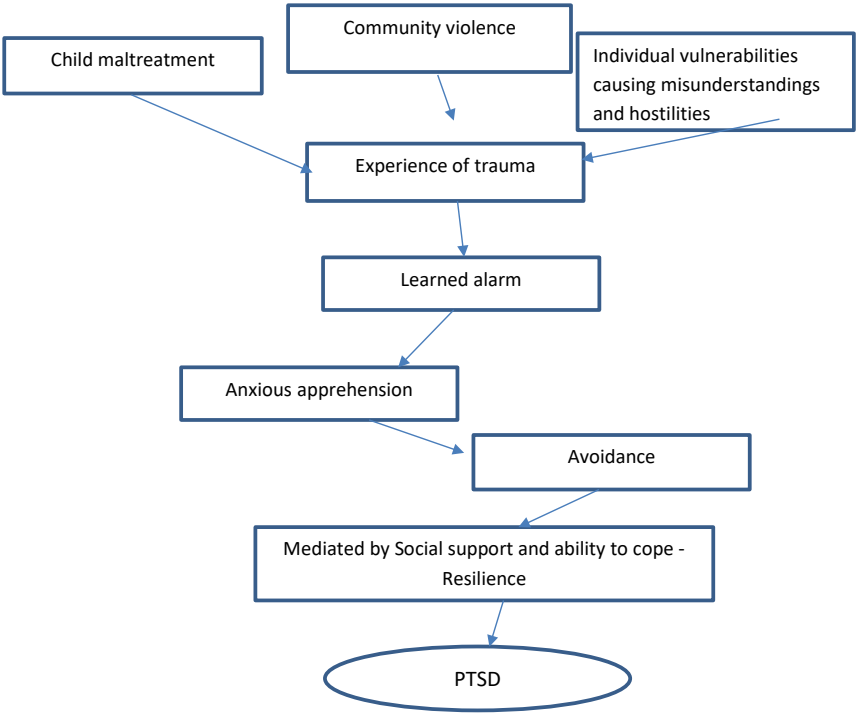
(c) The presence of other confounding factors (e.g. poverty, hostile environmental factors, bullying in school, and individual vulnerability).

(d) Unknown potential risk factors (e.g. genetic risk factors or childhood traumas that are not registered) and unknown protective factors (e.g. social support).

Theoretical issues

The aim of the study is to contribute to the understanding of the etiology of PTSD in seven to eighteen years old children and adolescents. We want to investigate the prevalence of PTSD and identify risk and protective factors for, and underlying causes of incident PTSD. The present study is inspired by Mowrer's two-factor learning theory (see Figure 1) which posits that fear is learned through associations and individuals will do whatever is necessary to escape and avoid cues that stimulate these aversive emotional memories. Learned alarms emerge from one special chain of events and occur during exposure to situations that resemble the traumatic event. The anxiety is moderated by adequate coping skills and social support. Identification of the precipitating event or proximal cause of PTSD is relatively simple, based on the theoretical descriptions of this form of anxiety (Keane & Barlow, 2004).

Figure 1. A model of the etiology of PTSD



This hypothetical model is informed by prior theoretical models proposed by Keane, Barlow, Mowrer and De Bellis (Keane & Barlow, 2004; Mowrer, 1960)

Many factors affect the development of PTSD: genetic vulnerabilities; frequency and intensity of trauma; developmental stage at which trauma occurs; and comorbid psychiatric and substance use disorders (Blacker et al., 2019).

Experiences of child maltreatment

Child maltreatment is a severe form of interpersonal trauma to which the child is exposed on a daily basis or in periods (De Bellis, Michael D., 2001; Ethier, Lemelin, & Lacharité, 2004).

Early preadoptive child maltreatment

Before being adopted, many children have been exposed to some kind of adversity, although most of the adopted children show psychological adjustment within the normative range (Bimmel, Juffer, Van IJzendoorn, & Bakermans-Kranenburg, 2003; Juffer & Van Ijzendoorn, 2005).

Social support as a mediating factor for creating resilience

The ecological maltreatment model suggests that maltreatment can emerge when multiple risk factors out-weight possible protective, compensatory, and buffering factors where applicable (Cicchetti & Lynch, 1993) in the family.

Living in a disadvantaged area

Three primary types of trauma are often used: *community violence* (e.g. war, inner city gang-related violence), *accidents* (e.g. transportation accidents, natural disasters), and *family or individual trauma* (e.g. interpersonal trauma, child abuse and neglect) (Foy et al., 1996). In some inner-city areas in USA children have a high risk of being exposed to community violence. Parents and other adults can provide valuable support, but are limited in how much they can offset the effects (Luthar & Goldstein, 2004).

Individual vulnerability

One of the basic assumptions of developmental traumatology research is that the biological stress system's regulation will be based on the nature of the stressors, frequencies, chronicity of the stressors, and individual differences i.e. genetic vulnerabilities (De Bellis, M. D., 2001). For example, individual genetic and biologic differences that lead to neurodevelopmental deviations such as Attention-Deficit Hyperactivity Disorder (ADHD) and Autism Spectrum Disorder (ASD) can produce an overwhelming amount of stress or conflicts with peers and family members, the immediate neighborhood and other social settings. It is well-known that certain disabilities cause higher prevalence of social misunderstandings and environmental conflicts in schoolchildren i.e. in children with ADHD, ASD or mild mental retardation. Police reports on violence and sexual assaults are also seen more often in these vulnerable groups, when controlled for other potential risk factors (Christoffersen, Mogens N., 2019; Christoffersen, Mogens Nygaard, 2020). Social isolation, family stress and poor communication skills increase risk of maltreatment in children with disabilities (Sullivan & Knutson, 2000). Therefore, children with ASD and ADHD are at an increased risk of experiencing trauma-related symptoms due to daily stressors such as social confusion, peer rejection, punishment, being reprimanded, difficulties coping with changes, and due to their difficulties in regulating of emotions and coping with stress (Kerns et al., 2015).

Children with ASD have profound social cognitive disabilities which make it difficult for them to deal with symbols, including language symbols of thoughts and feelings. Some children with ASD describe that they live in an environment in which they are unable to make themselves understood (Peeters & Gillberg, 1999). Aggressiveness and hostilities from environment may cause anxiety in the children and adolescents who find it incomprehensible. We hypothesize that ADHD and ASD may increase the risk for encountering traumatic events and for developing PTSD.

Methods: Research design

The study is prospective and based on longitudinal panel data from Denmark including the whole population sample, which provides information about the chronological sequence of potential causes before first-time registration of PTSD. The primary outcome

is incident cases diagnosed PTSD in the secondary sector, (i.e. hospital based mental health care). The outcome is binary – either it occurs, or it does not occur.

The study examines which risk factors preceded the first-time recorded PTSD diagnosis in registers set up by hospitals. We define a risk factor as a correlate that is shown to precede the outcome of interest according to Kraemer and colleagues. Risk is as a probability for an event (or an outcome) within a specified population (Kraemer, Lowe, & Kupfer, 2005). Data consists of administrative records with indicators of child maltreatment, pre-adoptive risk, environmental stressors in general, and indicators of various individual vulnerabilities. Administrative data are derived from numerous independent agencies such as police records, hospital records, and educational records, records from the criminal statistic register, records from social assistance register, and records from unemployment statistics.

Study population

National birth cohorts of children born 1984-1994 age 7-18 are followed (N=679,000) until a first time PTSD confirmed in a psychiatric ward.

The data are used for indication of experiences of child maltreatment, pre-adoptive risks or individual vulnerabilities, traumatic stress in the family, family lack of support, disadvantaged neighborhood, and demographic variables. Data are available from 1980 and forward for both children and their parents. Data on victims of sexual assault or violent assault are only available after 2001 by way of an administrative register of police records. This was decisive for the decision to limit the study to include the window from 2001 to 2012 where potentially victims are tracked.

Risk factors (covariates): Indicators of traumatic stress in the family

Children's developmental disabilities and other disabilities are considered individual vulnerabilities. The types of disability are based on a database mandated, compiled and maintained by Danish hospitals in accordance with the international statistical classification (ICD-10) of diseases and health-related problems (WHO, 1994). We classified disabilities into 14 main groups, which did not cover all disabilities. The categories did not include disabilities, which could be consequences of maltreatment such as internalizing disorders, depression, and other emotional disorders.

For the sake of our analysis, some disabilities that may affect the adolescent's ability to communicate, were grouped under speech disability (i.e. developmental disorders of speech and language), while other disabilities such as Cerebral Palsy was kept in a separate category.

Data analysis

Our model allows that individuals can have multiple disabilities. When reviewing the effect of a specific type of vulnerability in the regression analysis, the reference group would be the person-years without that specific type of disability. The model allows for changing covariates and changing disability over time. The study takes advantage of

analysis of covariance and multiple regression statistical analysis methods so that inter-relationship between several predicative variables and first-time PTSD can be examined simultaneously.

Potential demographic risk factors such as living in a disadvantaged area, long-term parental unemployment, parental teenage-motherhood, non-Danish citizens, and gender are included.

The data is analyzed by the discrete time-Cox-model (Allison, 1982). A discrete-time model treats each individual history as a set of independent observations. Only incident cases are included into the model in order to characterise the sequence of events characterising the causes and circumstances of the event. Individuals' history is broken up into 12 sets of discrete time-units (age 7 to 18 years) in which an event either did or did not occur. Each individual is observed until either an event occurs, or the observation is censored by reaching the age limit, because of death, or because the individual is lost to observation for other reasons e.g. immigration. Consequently, individuals are excluded from the case group and controls after the first event.

The person-years at risk were constructed for the total birth cohort. Pooling the non-censored years of all individuals, the person-years, made the numbers at risk (N= 4,917,535). It has been shown that the maximum likelihood estimator can be obtained by treating all the time units for all individuals as though they were independent, when studying first-time events (Allison, 1982).

The model allows for changing covariates over time. The risk covariates such as experiences of psychosocial maltreatment, stressful life events or individual differences are divided into three types for the purpose of this study. The Type I covariates are those that are taken to be indicative throughout the risk period, irrespective of when the covariate was notified e.g. parental substance abuse or child's diagnosis of autism. Covariates of Type II, in contrast, identify the presence of that factor in the year prior to the event e.g. parental long-term unemployment during a calendar year, or moving into a disadvantaged housing area, until moving out. Finally, the Type III covariates act on the following year and all the subsequent years when observed the first time e.g. family separation or brain injury.

Finally, for each age-groups (age=8, 9,..,17) a constant (Dummy) is estimated. This allows for standardization relating to age. Maximum likelihood estimators for the regression models are then calculated on the basis of pooling all the person-years.

Mediator analysis

Social support is hypothesized as a mediator (Figure 1) between traumatic events and PTSD. The method uses three regression equations to test for the statistical significance of a mediator effect (Baron & Kenny, 1986).

[Fig. 2 presents the three regression equations. In case of a simple linear regression, the regression coefficient (or the total effect) C′ will be the sum of the indirect effect A × B and the direct effect C. It can be proven algebraically without any reference to time

sequence or cause and effect (Wonnacott & Wonnacott, 1990). Sobel (1982) provided an approximate significance test for this purpose but used a bootstrap approach to obtaining confidence intervals, according to Preacher and Hayes (Iacobucci, 2008; Preacher & Hayes, 2004; Sobel, 1982). We assume that the independent variable (childhood trauma) is active at t1, the mediator (social support/lack of social support) is effective at t2 while the outcome variable (PTSD) is measured at t3. The sequential ordering (t1< t2< t3) can be tested.]

Statistics

We will describe schoolchildren with a PTSD diagnosis in comparison with their contemporaries. In order to evaluate the risk factors' contribution to the number of persons diagnosed with PTSD, attributable fractions (AF) are calculated (Greenland, 2008).

Attributable fractions express the reduction in incidence of PTSD that would be achieved if the population had not been exposed at all compared with the current exposure pattern (Greenland & Drescher, 1993). The estimated AF of a certain risk factor depends on two parameters, only. One parameter is the strength of the risk factor measured by OR or Relative Risk (RR). The other parameter is the current exposure of the risk factor in the population. The estimated AF is calculated solely on the basis on these two parameters (Levin, 1953; Woodward, 1999). AF is only defined when OR and Relative Risk is more than 1.

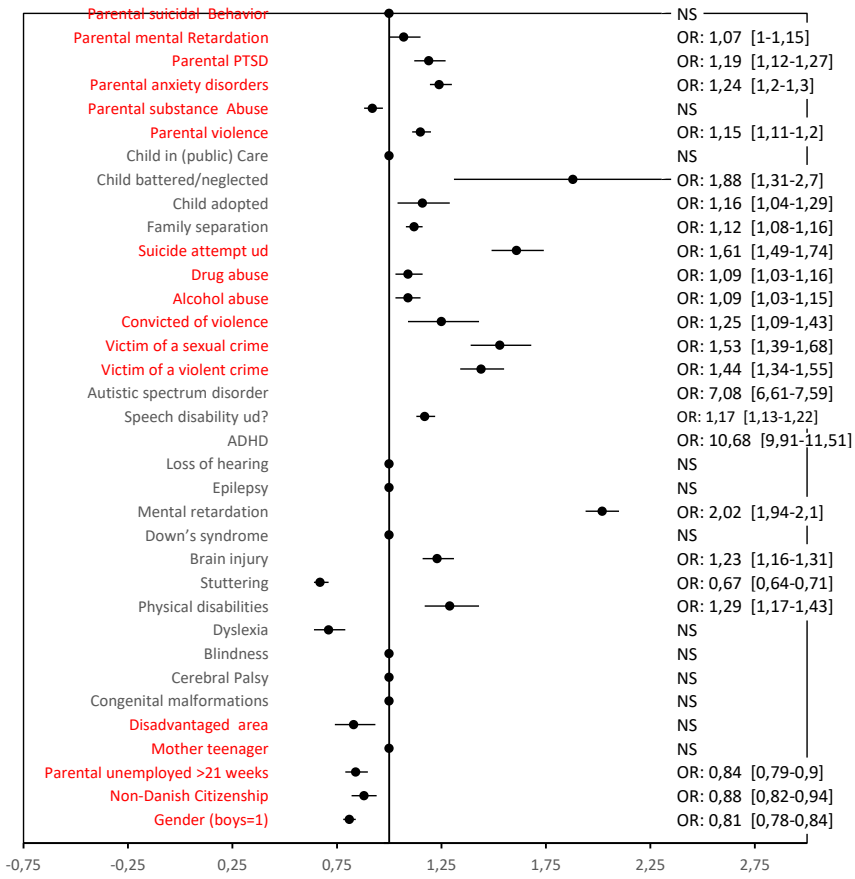
Results: Test of the trauma stress model

The lifetime prevalence of PTSD in school-age children diagnosed according to ICD-10 in a hospital ward was 2.3 per cent (N=15,636 following the eleven birth cohorts from seven to eighteen years.

The aim of the study was to illuminate the etiology of PTSD in children and adolescent by using Danish Register data to evaluate associations between adverse environmental factors and traumatic events and later PTSD in school-age children. We found a lifetime estimation of PTSD in school-age children in the range of 2.3%. The PTSD diagnosis is confirmed only in psychiatric wars, and it is likely that less severe cases will not be included in the present study.

We found an increased risk for PTSD among children diagnosed with neurodevelopmental disorders like ASD, ADHD, children with low levels of social support. Children exposed to child maltreatment, and children being adopted showed an increased risk of PTSD. These estimates are most likely in the lower end due to hidden data on anxiety disorders, and the hazards such as domestic violence, child maltreatment, conflicts and bullying between school children and other potential risk factors that are underreported or not reported to the registers. The results have similarities with previous studies but results also indicated structural divergences.

Figure 2. Forrest Plot of subgroups, post-traumatic stress disorder, adjusted Odds Ratio (OR).



Discussion

We conclude that the high prevalence of PTSD among school-age children with ADSH or ASD must be understood both as a co-morbid phenomenon and as a secondary condition that develop more easily due to a low threshold for PTSD in school-age children with ASD or ADHD because of their problems with understanding and interpreting all kinds of stimuli from the surroundings.

Our study of disadvantaged areas did not corroborate the American findings, maybe because disadvantaged areas are smaller and the society is more supportive? Rates of exposure to violence are high among children in United States (Luthar & Goldstein, 2004). These studies have methodological difficulties separating the individual risk factors in the family from characteristics of local areas. When other potential risk factors such as family violence, being a victim of violence, were taken into account, disadvantaged areas in isolation were not associated with school-age children's PTSD. This may be explained by aspects of the Danish welfare system like higher levels of security and smaller societal differences between rich and poor.

To obtain data for all important outcomes of severe constrains on school-age children, it is necessary to examine long-term follow up studies based on observational data. Non-randomized population studies might be the only way to study effects of for example maltreatment, environmental reactions to individual vulnerabilities and anxiety.

Our results were consistent with the described traumatic stress model. Children exposed to maltreatment or other kinds of traumatic stress (e.g. pre-adoptive deprivation, sexual abuse, violence in the family, community violence) will almost always react with fear, re-experiencing of the trauma and will be likely to avoid stimuli associated with the trauma. Interpersonal trauma where the perpetrator is a family member increases the risk of PTSD and erodes social support, since the child lives with the person who represents the trauma. Parents' own anxiety symptoms or other mental health problems may reduce their capacity to support children. The lack of available social support for resilience and the exposure to traumatic stress were associated with diagnosed PTSD in school-children.

Limits of the study

The results indicate that children with ASD and children with ADHD have a seven to ten times higher risk of being diagnosed with PTSD, when other potential constrains were taken into account. We suppose that these figures estimate the association because Berkson's bias may influence data (Berkson, 1946; Schwartzbaum, Ahlbom, & Feychting, 2003). We have a reason to believe that many children with PTSD do not get sufficient treatment for the disorder, that may become chronic, from psychiatric ward, municipality early interventions or general practitioners, and PTSD may in some cases only be reported in connection with elucidating decreasing function or increasing of symptoms of ADHD or ASD.

One drawback to the present study is that the data only provide information on diagnoses on personal vulnerabilities and lack information on existing and received help and social support. We use family dissolution as in indicator on lack of social support, but family cohesion is not the same as social support. There is a need for further studies because it is crucial to understand the possible positive impact of social support and adapt treatment strategies to counterstrike adverse outcomes.

Cognitive Behavioral Therapy (CBT) is recommended as first choice treatment of schoolchildren, Trauma-focused cognitive-behavioral therapy has demonstrated positive outcomes in reducing symptoms of PTSD (de Arellano, Michael A Ramirez et al., 2014). Individually information on these measures are not made available in administrative databases for the use of research.

Prospective longitudinal studies on large probability samples offer the best way to study predictors of environmental stressors causing trauma and possible PTSD in school-age children. The present study is based on a huge sample with comprehensive information about potential risk factors for all individuals. This allows for disentangling of the predictors' influence on the risk of developing PTSD. Unknown potential risk factors are the Achilles' heel of the strategy. The administrative data-system gives little or no knowledge of individual measures and local initiatives in primary sector taken to address potential traumatic life-events and anxiety disorders in school-age children.

Another key psychosocial factor is individual vulnerabilities (e.g. ADHD, ASD) causing misunderstandings, hostilities or just negative feedback and experiences from the surroundings. The positive role of psychoeducation and other social and educational interventions in children with neurodevelopmental disorders is corroborated by a systematic review (Montoya, Colom, & Ferrin, 2011), but ongoing research is needed to find means of changing the psychosocial environmental conditions by way of psychoeducation.

Extreme group analysis of patient cost of antibiotic prescribing among general practitioners

Bootstrapping of confidence intervals in subgroups of a sample of General Practitioners

Work in progress: Please do not cite or circulate without permission

Troels Kristensen[1], Charlotte Ejersted[2], Jens Søndergaard[3]
[1] Danish Centre for Health Economics (DaCHE), University of Southern Denmark
[2] Department of Endocrinology, Odense University Hospital
[3] Research Unit of General Practice, University of Southern Denmark

Background: Both high and low empathy among GPs may influence patient care. For instance, low empathy may have cost due to decreased patient satisfaction and patients frequenting the GPs less often or even switching GP. Nevertheless, very little is known about the magnitude and variation in antibiotic prescribing-profiles among GPs with high versus low empathy levels.

Aim: To make profiles of antibiotic prescribing for GPs with high versus low empathy and estimate uncertainties in statistic metrics in these groups of GPs. In addition, to explore alternative approaches to bootstrapping of confidence intervals (CIs) for descriptive statistic metrics.

Methods: This study applies extreme group analyses (EGA) to explore patient costs of antibiotic prescribing among subgroups of GPs with high and low empathy from a stratified random sample of 464 Danish GPs. The dataset includes combined survey-data on GP empathy and drug register-data merged via the GP's authorization number. Antibiotics were divided into subcategories of penicillin, non-penicillin antibiotics, antifungals as well as broad and narrow spectrum antibiotics based on the Anatomical Therapeutic Chemical classification (ATC). The GPs in the top decile and bottom decile of the empathy score distribution, were identified to make profiles of their antibiotic prescribing in terms of patient cost. The profiles included the means, coefficient of variation (CV), variation index (VI) and mean differences (DIs) in cost between the subgroups. Next, the uncertainty of the costs was estimated via CIs for all antibiotic categories. These CIs were estimated in 3 alternative ways: 1) A bootstrap procedure using the predefined extreme groups samples (n1, n2) only, but still for all antibiotic categories, 2) One single bootstrap procedure using the full sample of GPs (n=464) estimating CIs for all antibiotic categories and 3) A serial bootstrap procedure for each separate antibiotic category. This allows us to compare the CIs for the three alternative approaches.

Results: There were relatively few differences in patient cost of antibiotic prescribing across GPs with extreme empathy levels in this sample of GPs. However, it was a trend in the data that the high empathy group had lower patient cost of antibiotic prescribing for most categories of penicillins than the low empathy group. Bootstrapping of CIs based on predefined subsamples rather than the entire sample resulted in larger CIs. Restricted application of data from predefined extreme groups may lead to wider and biased CIs in extreme group analysis.

Conclusion: This study shows that extreme group analysis and related GP profiles can be used to explore antibiotic prescribing behavior among GPs. One generated hypothesis is that high empathy GPs prescribe some penicillin categories in a different way that results in lower patient costs than low empathy GPs. Included CIs based on alternative approaches reveals the level of uncertainty in the estimated patient cost across the included antibiotic categories.

1. Introduction:

It is well known that antibiotic prescribing behavior is very important for patient in both primary and secondary care as well as the society in general[1]. One central reason is that unwarranted variation in terms of so called "under prescribing" and "over prescribing" may be harmful for patients and society. "Under prescribing" occur when patients receive no or too little of one or more categories of drugs such as penicillins or other non-penicillin antibiotics. "Over prescribing" means that the patient receives more antibiotic prescriptions than needed from a clinical or biomedical point of view[2,3]. In practice, it is usually best to use specific antibiotics or narrow spectrum antibiotics to avoid use of broad spectrum antibiotics if a narrow spectrum substitute exists. Another best way is to avoid antibiotics if other better or equivalent options exists. In cases where no antibiotic bacterial resistance measures have been done, antibiotics may be avoided by careful examination of patients as symptoms may be related to non-infectious disease or virus. To understand and address unwarranted variation in antibiotic prescribing it is important to explore circumstances that may influence prescribing of antibiotics and by using microbial diagnostic methods. For instance, it has been hypothesized that variation in GP empathy may influence GP behavior and thus prescribing of antibiotics[6]. The reason is that both high and low empathy have been shown to directly influence other elements of patient care. Both directly in terms of unwarranted variation for the individual patient and indirectly, in terms of patient opportunity cost and negative externalities in terms of antibiotic resistance for society. Therefore, it is relevant to understand the association between GP empathy and their antibiotic prescribing profiles. However, before this type of confirmatory hypothesis can be tested, there is a need to undertake hypothesis generating exploratory analysis, which calls for the latter analysis. So far, very little empirical research has been conducted around empathy levels and drug prescribing[6]. Subgroup analysis such as extreme group analysis has been stated to be well suited to undertake exploratory analysis[5,7]. Besides it helps us understand the profiles of antibiotic prescribing among GPs with high and low empathy levels. The present study focuses on direct patient prescribing cost and not least the approaches to estimation of uncertainty related to these patient costs. From a resource perspective cost or expenditure to patients is standard way to measure resource use.

Accordingly, the aim of this study is to make profiles of antibiotic prescribing patient costs for GPs with high versus low empathy and quantify uncertainties in statistic metrics based on the available sample of GP survey and register data[8]. Furthermore, it is an aim of this symposium paper to explore alternative approaches to bootstrapping of confidence intervals (CIs) for descriptive statistic metrics.

2. Method

Extreme groups analysis was employed to explore the nature of antibiotic prescribing patient costs among subgroups of GPs with high and low empathy in a sample of 464 general practitioners[4,5]. The patient cost of antibiotic prescribing was measured in terms of the pharmacy retail pris (abbreviated AUP in Danish) aggregated over all fillings. The price is set as the cost price (AIP) plus profit margin and value added tax (TVA) and represents the direct cost to consumers.

The GPs empathy was measured via the Jefferson Scale of empathy for Health care professionals[9]. The dataset includes combined survey-data on GP empathy and drug register-data merged via the GP's authorization number for the year 2017. Antibiotics were divided into categories of penicillin, non-penicillin antibiotics, antifungals as well as broad and narrow spectrum antibiotics. The GPs in the top decile and bottom decile of the empathy score distribution, were identified to make profiles of their costs of antibiotic prescribing. The profiles included the mean costs, coefficient of variation (CV), variation index (VI) and mean differences (DIs) between the groups. Next, the uncertainty of the costs was estimated via CIs for all cost categories.

2.1 Estimation of confidence intervals in the bottom and top decile:

This study has used bootstrapping to estimate uncertainty around cost of antibiotic prescription metrics in terms of CIs similar to other health economic studies[10,11]. The idea is that the bootstrapped empirical cumulative density function based on our data sample, gives a good sense of what the true unknown (population) distribution is[12]. In addition to extreme group analysis (EGA) of GPs with high and low empathy, the idea was to quantify and explore the effect of three alternative bootstrapping approaches on CIs:

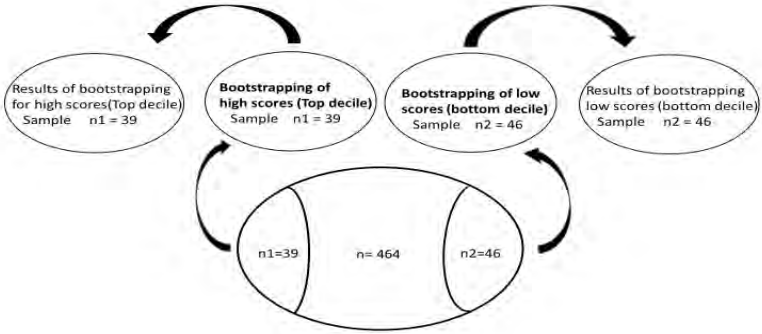
1. Bootstrapping based on subgroups of GPs with high and low empathy[13]. Here, the two extreme groups were predefined as top and bottom deciles and identified before the bootstrap procedure was implemented as one simultaneous or parallel procedure for estimation of CIs for all variables, see illustration in Figure 1.
2. Bootstrapping using the entire sample to calculate extreme groups including parallel estimation on related statistic metrics and their confidence intervals, see Figure 2. This was based on a program adopted in Stata. The idea was to calculate all subgroup statistic metrics in one bootstrap procedure rather than several independent standard procedures for each variable or antibiotic category. The parallel estimation of all statistic metrics will restrict variation to the variation created by the by the specific bootstrap procedure. This is different from a procedure where the bootstrap is performed in a serial way for each statistic metric. The latter is expected to increase variation because the results for the statistic metrics will be based on multiple bootstrap procedure – in particular when the start seed option for the bootstrap is set to random.
3. Independent serial bootstrapping of each individual variable according to approach 1 (for the pre- calculated extreme top and bottom decile groups) and approach 2 (using the entire sample for bootstrapping). In these cases, the statistic metrics for each variable will be based on separate bootstrapping procedures where the variation will be less restricted and CIs for metrics expected to be wider compared to approach 1 & 2.

The bootstrap replicates, which increases accuracy was set to 1000. The seed for the bootstrap was set to zero (random) to allow for differences between 1 and 3 as well as 2 and 3.

Re 1: Bootstrapping approach 1 – using the extreme groups only:

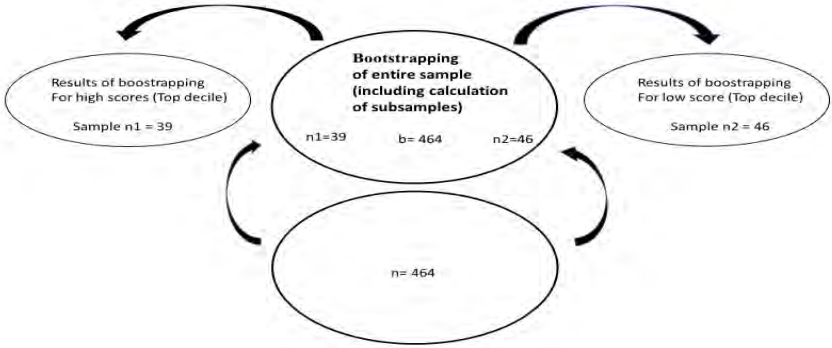
Figure 1 and Figure 2 illustrate the applied approach to bootstrapping of the CI for statistic metrics based on the survey sample. Figure 1 shows the restricted version of the bootstrap procedure where the subsamples were predefined before the bootstrap procedure was applied on the subsamples n1=39 and n2=46 rather than the entire sample.

Figure 1 Bootstrapping approach 1 based on predefined extreme groups



Re 2: Bootstrapping approach 2 using the entire sample

Figure 2 illustrates the applied approach to bootstrapping of the CI for descriptives statistics based on the entire survey sample (n=464). The bootstrap procedure was based on a stata program which calculates the top an bottom deciles of the sample. This program was bootstrapped with respect to the desired statistics metrics such as the mean costs, coefficient of variation, variation index and all antibiotic categories in such a way that all statistic metrics was calculated based on a bootstrap proecure with 1000 replications.



Besides the following examples of the approaches 1, 2 and 3, this study only reports results based on approach 2.

2.1.1 Example: Bootstrapping of CIs for descriptive statistics in extreme groups

In this example the yearly prescribing patient cost related to Macrolides JO1F are applied to illustrate the alternative approaches illustrated in Figure 1 and Figure 2 to bootstrap CIs for statistic metrics in extreme group analysis. In Table 1, the mean and CV metrics was chosen to show examples including related bootstrapped CIs for each of the three approaches. The first approach uses predefined subgroups for bootstrapping, the second the entire sample and the third serial bootstrapping for each variable:

Table 1 Alternativ uncertainty measures of cost of macrolides prescriptions

	High-empathy GPs † (n1=39)			
Bootstrapping approach:	**Mean**	**[95% CI]**	**CV**	**[95% CI]**
Approach 1 predefined subgroup:				
Macrolides (B) J01F	3,911.6	[2,878.7; 4,944.4]	0.85	[0.67;1.02]
Approach 2 using the entire sample:				
Macrolides (B) J01F	3,935.6	[3,039.7; 4,831.6]	0.82	[0.68;0.94]
Approach 3 serial approach:				
Predefined subgroups:				
Macrolides (B) J01F:	3911,6	[2900.3;4922.9]	0.85	[0.54;0.79]
Using entire sample:				
Macrolides (B) J01F:	3935,6	[3047.6;4823.6]	0.81	[0.68;0.96]

These three alternative approaches may be applied in practice to estimate uncertainty of statistic metrics in top decile and bottom decile groups of GPs as part of an EGA. The applied illustrative example for Macrolides (J01F) for the different approaches is based on data from the high empathy group and includes the 95% CIs. According to approach 2, which uses the entire sample of GPs to estimate the CIs, the mean yearly Macrolides J01F patient prescription cost was DKK 3,935.6 within a 95% CI of [3,039.7;4.831.6] and the coefficient of variation was 0.82 within a 95% CI of [0.68;0.94]. In contrast, approach 1 only uses the reduced predefined subgroup (n1=39) to estimate the mean and related CI. Compared to approach 2, this shows a similar mean and a broader CI [2,878.7;4.944.4] corresponding to larger uncertainty, when a smaller subgroup is used to bootstrap CIs. This was also the case for the CV and the related CI when approach 1 and approach 2 were compared. Both the CV [0.85] and the 95% CI [0.67;1.02] in approach 1 reflected larger variation an uncertainty than in approach 2 with CV [0.82] and 95% CI [0.68;0.94].

The GP profile characteristics in the subsamples with high and low empathy will be reported in a related future study. So far, the characteristics of the entire sample have been described in a previous study[9]. Therefore, this study only focuses on the cost of antibiotic prescribing in the reported subgroups.

3. Results

The results of the EGA of yearly prescribing patient cost in Table 2 and Table 3 show the estimated yearly mean antibiotic prescriptions patient costs and within subgroup variation in terms of the CV for the groups of high empathy GPs (n1= 39, Table 1) versus the same yearly mean cost and CV for the subgroup of low empathy GPs (n2=46, Table 3). In Table 2, the results including bootstrapped 95% CIs for the mean and CV are divided into subcategories of penicillin, non-penicillin antibiotics, antifungals and broad versus narrow spectrum antibiotics.

Table 2 yearly cost of antibiotic prescriptions for high empathy GPs

	High-empathy GPs † (N=39)			
Antibiotic category:	**Mean**	**[95% CI]**	**CV**	**[95% CI]**
Penicillins:				
Extended spectrum (B) JO1CA	28,855.8	[23,457.6; 34,254.1]	0.68	[0.48;0.88]
Beta-lactamase sensitive (N) J01CE	6,652.0	[5,721.6; 7,582.3]	0.53	[0.42;0.63]
Beta-lactamase resistant (N) J01CF	6,129.0	[3,712.4; 8,545.6]	1.47	[0.94;1.99]
Combinations with beta lactamase inhibitors (B) J01CR	2,162,4	[1,459.2; 2,865,6]	1.24	[0.87;1.61]
All penicillins: J01C	48,546.5	[41,191.7; 55,901.4]	0.56	[0.42;0.69]
Non-penicillin antibiotics :				
Tetrayclines (B) J01A	8,887.6	[1,555.9; 16,219.4]	3.04	[1.04;5.06]
Cephalosporins J01D	32.6	[0.84; 64.3]	3.66	[1.37;5.94]
Sulfonamides J01E	10,175.7	[5,971.5; 14,379.8]	1.52	[1.25;1.78]
Macrolides (B) J01F	3,935.6	[3,039.7; 4,831.6]	0.82	[0.68;0.94]
Quinalones (B) J01M	594.8	[392.3; 797.3]	1.23	[0.85;1.62]
Other antibiotics J01X	3,815.5	[602.0; 7,029.0]	3.10	[1.97;4.23]
All other antibiotics	27,441.8	[17,760.4; 37,123.2]	1.35	[0.92;1.77]
Antifungals, J02	4,983.8	[2,705.1; 7,262.5]	1.72	[1.25;2.20]
Total all antibiotics:	91,188.8	[74,855.0;107,522.6]	0.67	[0.53;0.80]
Narrow spectrum (N)	34,202.4	[26,109.0; 42,295.9]	0.88	[0.65;1.11]
Broad spectrum (B)	42,977.2	[33,478.6; 52,475.7]	0.83	[0.50;1.15]

Narrow spectrum antibiotic was defined via the following ATC-codes: J01CE, J01CF, J01DB, J01DF, JO1EA, J01EB, J01FA, J01FF, J01XA, J01XC, J01XD, J01XE, J01XX. Broad spectrum antibiotics included: J01AA, J01CA, J01CR, J01DC, J01DD, J01DH, J01EE, J01GB, J01MA, J01MXB.

Overall, the patient cost results in Table 2 and Table 3 indicate that there is a trend that high empathy GPs prescribe under circumstances which means their patients have lower cost of antibiotic prescribing compared to the bottom decile of low empathy GPs. In Table 2 and Table 3 the within category variation in terms of the CV indicates a trend towards a slightly higher variation in the patient cost of antibiotic prescribing. Both for the group of penicillins and the group of non-penicillin antibiotics. Among both the high and low empathy groups the lowest within category variation was in the group of penicillins. The largest variation was in the categories of non-penicillin antibiotics with lower activity in terms of direct patient costs such as macrolides.

Table 3 yearly cost of antibiotic prescriptions for low empathy GPs

		low-empathy GPs † (N=46)		
Antibiotic category:	**Mean**	**[95% CI]**	**CV**	**[95% CI]**
Penicillins:				
Extended spectrum (B) JO1CA	40,489.8	[29,195.5; 51,784.2]	0.85	[0.64;1.07]
Beta-lactamase sensitive (N) J01CE	8,343.0	[6,469.0; 10,217.0]	0.71	[0.56;0.85]
Beta-lactamase resistant (N) J01CF	5,971.4	[4,003.6; 7,939.1]	1.05	[0.53;1.58]
Combinations with beta lactamase inhibitors (B) J01CR	2,746.9	[1,988.9; 3,504.8]	0.90	[0.71;1.09]
All penicillins: J01C	64,852.2	[49,822.6; 79,881.8]	0.71	[0.555;0.86]
Non-penicillin antibiotics :				
Tetrayclines (B) J01A	4,775.0	[3,304.2; 6,245.7]	0.97	[0.72;1.22]
Cephalosporins J01D	21.6	[0.4; 42.9]	3.16	[1.19;5.12]
Sulfonamides J01E	6,959.7	[3,391.0; 10,528.5]	1.61	[1.27;1.96]
Macrolides (B) J01F	4,093.5	[2,832.2; 5,354.8]	0.99	[0.78;1.21]
Quinalones (B) J01M	1,143.1	[572.6; 1,713.6]	1.63	[1.06;2.20]
Other antibiotics J01X	6,903.1	[389.8; 13,416.5]	2.93	[1.83;4.02]
All other antibiotics	23,896.0	[15,495.3; 32,296.7]	1.12	[0.86;1.38]
Antifungals, J02	4,990.0	[2,568.6; 7,711.5]	1.56	[1.13;1.98]
Total all antibiotics:	106,931.5	[80,334.4;133,528.7]	0.76	[0.63;0.89]
Narrow spectrum (N)	34,478.6	[24,836.2; 44,121.1]	0.90	[0.65;1.15]
Broad spectrum (B)	53,985.5	[40,138.3; 67,833.0]	0.78	[0.62;0.94]

Table 4 shows an analysis of the difference between the high and low empathy GPs. This means the group differences between the patient costs of each antibiotic category based on Table 2 and Table 3. The latter differences are measured in terms of a) a variation index (VI), b) the mean group difference and c) tests of the differences between the costs of antibiotic prescribing. The variation index (VI) indicates a trend that high empathy GPs have about 20-30% lower patient cost of antibiotic prescribing of most penicillins than low empathy GPs. Differently, the variation picture was more fluctuating for non-penicillin antibiotics where the high empathy GPs also had higher yearly cost of prescribing for tetracyclines, cephalosporins and sulfonamides. High empathy GPs seems to prescribe less broad- spectrum antibiotics. However, the above mentioned trends were not confirmed by the Mann-Whitney (ranksum) test of group difference and t-test of the group mean difference and related confidence intervals. Only the group of all penicillins was boarder-line significant according to the Mann-Whitney/t-test. In the e-return list of the bootstrap procedure in stata there was a Z0-test-score for the mean difference of each antibiotic category in Table 4. The last column in Table 4 includes this Z-score for the Mean difference.

Table 4 yearly cost of antibiotic prescriptions – high versus low empathy group differences

			Group difference measures			
Penicillins:	VI	[95% CI]	Mean	[95% CI]	MannW/t-test	Z0
Extended spectrum (B) JO1CA	0.71	[0.47;0.95]	-11,634.0	[-23,846.1; 578.1]	0.0763/0.0325	0.0258*
Beta-lactamase sensitive (N) J01CE	0.80	[0.58;1.01]	-1,691.0	[-3,762.1; 380.1]	0.2169/0.0760	0.0000***
Beta-lactamase resistant (N) J01CF	1.03	[0.51;1,54]	157.6	[-2,863.6; 3,178.8]	0.4482/0.9743	0.0000***
Combinations with beta lactamase inhibitors (B) J01CR	0.79	[0.44;1.14]	-584.5	[-1,620.6; 451.7]	0.1065/0.3233	0.0517
All penicillins: J01C	0.75	[0.55;0.95]	-16,305.6	[-32,523.4; -87.9]	0.0720/0.0301	0.0258*
Non-penicillin antibiotics:						
Tetrayclines (B) J01A	1.86	[0.16;3.56]	4,112.6	[-3,385.5;11,610.8]	0.4588/0.2594	0.1661
Cephalosporins J01D	1.51	[-2.10;5.11]	10.9	[-26.8; 48.7]	0.67600.8307	0.0181*
Sulfonamides J01E	1.46	[0.32;2.60]	3,215.9	[-2,390.9; 8,822.7]	0.4588/0.2594	.0.0129*
Macrolides (B) J01F	0.96	[0.58;1.34]	-157.8	[-1,671.7; 1,356.0]	0.9367/0.8202	0.0232*
Quinalones (B) J01M2	0.52	[0.17;0.87]	-548.3	[-1,151.3; 54.7]	0.2552/0.1221	0.0439*
Other antibiotics J01X	0.55	[-0.80;1.91]	-3,087.6	[-10,355.1; 4,179.8]	0.5139/0.6925	0.0621
All other antibiotics	1.15	[0.54;1.76]	3,545.8	[-9,168.2;16,259.7]	0.7177/0.3727	0.0414*
Antifungals, J02	1.00	[0.24;1.76]	-6.2	[-3,327.9; 3,315.4]	0.2627/0.9478	0.0439*
Total all antibiotics:	0.85	[0.59;1.12]	-15,742.7	[-46,579.7;15,094.2]	0.4588/0.4104	0.0336*
Narrow spectrum (N)	0.99	[0.62;1.36]	-276.2	[-12,658.3;12,105.9]	0.8255/0.8720	0.0026**
Broad spectrum (B)	0.80	[0.52;1.07]	-11,008.5	[-27,656.6; 5,639.7]	0.2270/0.2486	0.0078**

The differences between the groups have been tested: The Mann-Whitney (ranksum) test was applied. *: p <0.05, **: p <0.01 ***: p <0.001. Furthermore, 95% CIs for the mean group differences have been estimated. All reported confidence intervals are based on bootstrapping (1000 reps) of the observations in the high and low empathy groups respectively.

4. Discussion

This exploratory pilot study of profiles of antibiotic prescribing patient cost among GPs indicate that both high and low empathy GPs prescribe categories of antibiotics in a similar way. This may be interpreted as good news for GP patients. Nevertheless, based on the results one hypothesis may be that high empathy GPs prescribe some penicillin categories in a different way that results in lower patient costs than low empathy GPs. Another hypothesis may be that high empathy GPs prescribe antibiotics in a more homogenous way than the group of low empathy GPs. One explanation could be that high empathy GPs communicate better with their patients and therefore are able to explain to them that they do not need the more expensive and broad spectrum antibiotics. Thus, the results reveal that it may be meaning full to conduct further confirmatory research with respect to behavioral research agendas since empathy could have an important influence on patient and subsequently societal costs.

The applied extreme group analysis has strength and weaknesses[5]. However, it is a recognized method to undertake exploratory pilot research and considered well suited to make profile pictures of GP behavior.

Furthermore, it is a strength of this study that the estimated CIs for the profiles allow the readers to know the level of uncertainty of the patient cost in each ATC-category.

To assess the sensitivity of the uncertainty to the applied bootstrapping approach this study has used at least three alternative bootstrapping approaches to estimate CIs in subgroup analysis. This experience and the examples included in Table 1 shows that CIs based on bootstrapping on the predefined or reduced subgroups yields broader CIs. Furthermore, serial bootstrapping does not appear to give systematically different results than the preferred approach 2 which is based on the entire sample. Finally, the results demonstrate the chosen approach may have an impact on the results in this type of research and should be investigated further.

Patient cost of antibiotic prescribing was measured in terms of retail prices. The prices did not include any administrative fees for fillings and any final drug preparation fees and can only be perceived as a proxy for patient costs. For instance, changes in patient subsidies and profit margins may have had influence on the patient retail prices in 2017. Besides the applied cost metric did not include other potential patient opportunity cost due to other patient cost such as lost income.

The aim of this study has not been to explain variation in antibiotic prescribing patient costs. It is clear, that fluctuations in the investigated antibiotic patient cost may be determined by a range of patient characteristics such as differences in patient morbidity in terms of co-morbidities/multimorbidity and socioeconomic patient characteristics. Nevertheless, these limitations do not preclude EGA or subgroup analysis of profiles of antibiotic prescribing patient cost across antibiotic categories among GPs with high and

low empathy levels according to the Jefferson Scale of Empathy for health care professionals.

5. Conclusion

This study show that extreme group analysis and related GP profiles can be used to explore antibiotic prescribing behavior among GPs. One generated hypothesis is that high empathy GPs prescribe some penicillin categories in a different way that results in lower patient costs than low empathy GPs. Included CIs based on alternative approaches reveals the level of uncertainty in the estimated patient cost across the included antibiotic categories.

Literature

1. Colgan R, Powers JH. Appropriate antimicrobial prescribing: approaches that limit antibiotic resistance. *Am Fam Physician*. 2001;64(6):999-1004.
2. Cole A. GPs feel pressurised to prescribe unnecessary antibiotics, survey finds. *BMJ*. 2014;349:g5238. doi:10.1136/bmj.g5238
3. Alber K, Kuehlein T, Schedlbauer A, Schaffer S. Medical overuse and quaternary prevention in primary care–A qualitative study with general practitioners. *BMC family practice*. 2017;18(1):1-13.
4. Feldt LS. The use of extreme groups to test for the presence of a relationship. *Psychometrika*. 1961;26(3):307-316. doi:10.1007/BF02289799
5. Preacher KJ, Rucker DD, MacCallum RC, Nicewander WA. Use of the extreme groups approach: a critical reexamination and new recommendations. *Psychol Methods*. 2005;10(2):178-192. doi:10.1037/1082-989X.10.2.178
6. Yuguero O, Ramon Marsal J, Esquerda M, Vivanco L, Soler-González J. Association between low empathy and high burnout among primary care physicians and nurses in Lleida, Spain. *Eur J Gen Pract*. 2016;23(1):4-10. doi:10.1080/13814788.2016.1233173
7. Burke JF, Sussman JB, Kent DM, Hayward RA. Three simple rules to ensure reasonably credible subgroup analyses. *BMJ*. 2015;351:h5651. doi:10.1136/bmj.h5651
8. Hodges S, Klein K. Regulating the costs of empathy: The price of being human. *The Journal of Socio-Economics*. 2001;30:437-452. doi:10.1016/S1053-5357(01)00112-3
9. Charles JA, Ahnfeldt-Mollerup P, Søndergaard J, Kristensen T. Empathy Variation in General Practice: A Survey among General Practitioners in Denmark. *Int J Environ Res Public Health*. 2018;15(3). doi:10.3390/ijerph15030433
10. Eakin BK, McMillen DP, Buono MJ. Constructing Confidence Intervals Using the Bootstrap: An Application to a Multi-Product Cost Function. *The Review of Economics and Statistics*. 1990;72(2):339-344. doi:10.2307/2109725
11. Campbell MK, Torgerson DJ. Bootstrapping: estimating confidence intervals for cost-effectiveness ratios. *QJM*. 1999;92(3):177-182. doi:10.1093/qjmed/92.3.177
12. Efron B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*. 1979;7(1):1-26. doi:10.1214/aos/1176344552
13. Breivik Ø, Aarnes OJ. Efficient bootstrap estimates for tail statistics. *Natural Hazards and Earth System Sciences*. 2017;17(3):357-366. doi:10.5194/nhess-17-357-2017

What is SAS® Viya® for Learners?

SAS® Viya® for Learners delivers free access to advanced analytics software for teaching and learning. It is a suite of cloud-based software that supports the entire analytics life cycle – from data, to discovery, to deployment – and lets you code in SAS, Python or R.

Sign Up Today!

As a student or an academic educator, you can easily get access by following this link:
www.sas.com/viya-learners

sas