FORECASTING AF DEKOMPONERET REGNSKABSTAL

Forecasting decomposed accounting measures

Forfatter oplysninger

Jonathan Juhl Severinsen
Studienummer: 92441
Kandidatafhandling
Cand. Merc. FIR
Instituttet for Regnskab
Copenhagen Business School
Vejleder: Jeppe Christoffersen

Antal anslag: 162.454

Aflevering: 16-09-2019

Executive Summary

This project seeks to expand on the current literature, regarding the forecasts decomposed accounting measures. As shown by Fairfield and Yohn (2001), this part of the literature regarding forecasting and accounting, is not a well-covered subject. Therefore, this project has analyzed the effect of decomposing RNOA into PM and ATO and thereafter OPPL, NOA and GP, for the use of forecasting. This is done by creating five different models, that each of which has its own unique role in this project. The first model serves as a benchmark and the subsequent four models, seek to improve upon this. These four models are based on two different schools of thought. The first two of these four models are based on a traditional way of handling multiple linear regressions. The last two models are based on Plenborg, Petersen and Kinserdal (2017) and their way of forecasting value drivers. All five models are evaluated based on several key assumptions, that ensure the models are truly comparable. During the comparison of these five models it has been a focus to have a wide variety of accuracy measures, to thoroughly describe each model. It is then shown, that its possible to improve on the benchmark model, by decomposing RNOA into PM and ATO, and then focusing on forecasting these value drivers separately. However, this was the only model which was deemed an improvement on RNOA, but it was an improvement nonetheless.

Indhold

1. Indledning

Værdiansættelser, kreditvurderinger og budgetteringer, er blot nogle meget få eksempler på analyser baseret på virksomhedens regnskabstal. Det rapporterede regnskab giver brugeren et hav af informationer, som individet effektivt kan anvende til at få et kritisk indblik i virksomheden. Kreditvurdering og værdiansættelse, forsøger at skabe et indblik i fremtiden, og spå om virksomhedens indtjening om et bestemt antal år. Dette bruger aktieanalytikere, banker, og den interne finansielle afdeling, til at risikovurdere fremtidige projekter, opvejet mod fremtidig indtjening (Plenborg, Petersen, & Kinserdal, 2017). Et valg mellem forskellige investeringsmuligheder må ikke være reduceret til et gæt, men bør være baseret på statistiske teorier og modeller (Yohn, 2018). Statistiske analyser af regnskab kan være ekstremt tidskrævende, og det kan være svært at sige, hvor man skal starte, og hvor man skal stoppe. Kører en analytiker en fuld statistisk analyse af et regnskab, med en rationel model, der tager alle perspektiver og variabler med, bliver det hurtigt meget uoverskueligt. De væsentlige variabler at tage med i sådanne en analyse, er hvad denne kandidatafhandling søger at finde, med baggrund i den statiske videnskab, samt den begrænset litteratur på emnet (Fairfield & Yohn, 2001).
Den begrænsede mængde af litteratur på emnet, er lidt overraskende, da det er et utroligt vigtigt emne (Fairfield & Yohn, 2001). Dette projekt spekulerer, at følgende fire punkter er knyttet til den begrænsede mængde af litteratur på emnet:

- Virksomheder og deres medarbejder er ikke opmærksom på metodens potentiale i forhold til at forbedre deres analyser.
- De individer, som dette er relevant for, anvender allerede disse metoder.
- Der er ikke adgang til alt den data, som skal anvendes til en dybdegående analyse.
- Det er ikke et emne, som er en prioritet for individet.

Den sidste af de fire punkter, er den som dette projekt vil fokusere på. Der skal undersøges og teste, hvor væsentligt det er for en analyse, at der bliver dekomponeret, når der forecastes regnskabstal, og i begrænset forhold, hvor meget der skal dekomponeres.

Dette projekt vil tage udgangspunkt i nøgletallet ”return on net operating assets” (herefter RNOA). RNOA er blevet valgt på baggrund af den tilgængelige data, hvilket vil gennemgås senere i projektet. På trods af at mange andre nøgletal kan bruges til et lignende projekt, såsom ROE (Fairfield, Yohn, & Sweeney, 1996), så er RNOA ydermere valgt, grundet de operative poster den direkte dækker. De operative poster i en virksomhed,

er de poster, som er de primære drivkræfter bag værdiskabelse og derfor er vigtig at undersøge i en analytisk problemstilling.

Argumenterne fremlagt for læseren i denne del af projektet, bringer læseren videre til formålet bag selve projektet.

1.1 Formål

Formålet med dette projekt er at undersøge betydningen af dekomponeret regnskabstal, i en undersøgelse af statistisk præcision af forecasts. Til dette er der udarbejdet en problemformulering, som så vil brydes ned i tilhørende hypoteser, samt underspørgsmål. Disse underspørgsmål har til hensyn at give et indblik i problemformuleringen, samt hjælpe med at få en komplet besvarelse af den. Underspørgsmålene er lavet på baggrund af de opstillede hypoteser, for at teste teorien i forbindelse med problemformuleringen.

Helt lavpraktisk søger dette projekt at finde svar på, om det er bedst for en analytiker, investorer eller andre interessenter at tage den simple løsning og udelukket kigge på RNOA, eller om de kan få et bedre (mere statistisk præcist) resultat, ved at dekomponere RNOA. Hertil er det dog vigtigt at nævne, at dette projekt bliver udelukket udarbejdet, med et "outsider"-perspektiv. Dette betyder, at blandt andet variablerne, og deres grundlag, er valgt på baggrund af begrænset offentlige informationer (altså den data, som er tilgængelig). Ydermere, kan det ikke på baggrund af dette projekt vides, om andre resultater kan fremskaffes, med mere komplette datasæt. Senere i dette projekt, vil rammerne for denne undersøgelse blive gennemgået, med henblik på problemformuleringen, afgrænsninger, metoden, teorien osv.

1.2 Problemformulering

Denne kandidatafhandling bliver udarbejdet på baggrund af følgende problemformulering:

Forbedres præcisionen af et forecast af regnskabets nøgletal, hvis der bliver fokuseret på de variabler, som nøgletalende bliver dekomponeret i?

Denne problemformulering fokuserer på regnskabets nøgletal, hvor dette projekt tager udgangspunkt i RNOA, som det overordnede nøgletal. RNOA bliver i første omgang dekomponeret til variablerne profit margin (herefter PM), og aktivernes omsætningshastighed (herefter ATO), hvilket er baseret på Du-Pont pyramiden (se afsnit 3.3). Med præcision af forecast, menes om der gennem statistiske modeller og teorier, kan skabe et mere stabilt og pålideligt resultat. For at besvare problemformuleringen, bliver følgende hypoteser opbygget, på baggrund af egen inspiration, samt litteratur, der præsenteres i teoriafsnittet senere i projektet.

	Nul-hypoteser	Inspiration af:
Hypotese 1	H_0: *En simpel DuPont dekomponering af RNOA vil kunne forbedre den statiske præcision, til at forecaste nøgletallet.*	(Soliman, 2008), (Fairfield & Yohn, 2001).
Hypotese 2	H_0: *En dekomponering i to lag, er mere statistisk præcis, end en simple dekomponering, når det gælder forecasting af RNOA.*	Af egen inspiration.
Hypotese 3	H_0: *Et bedre resultat for forecastet af en simpel DuPont dekomponering af RNOA opnås, ved hjælp af en alternativ fremgangsmåde. **	Af egen inspiration, samt (Plenborg, Petersen, & Kinserdal, 2017).
Hypotese 4	H_0: *Kan en bedre model opnås, ved hjælp af fremgangsmåden fra hypotese 3, pålagt hypotese 2.*	Af egen inspiration, samt (Plenborg, Petersen, & Kinserdal, 2017).

*Tabel 1 - af egen inspiration. *Denne alternative fremgangsmåde (også kaldet den alternative metodik) præsenteres senere i projektet.*

Tabel 1 viser, de forskellige hypoteser, som dette projekt søger at teste, for at bedre kunne besvare problemformuleringen.

Hypotese 1 søger at få svar på, om en simpel DuPont dekomponering (Dette begreb uddybes i afsnit 3.3), af RNOA, kan forbedre dets forecast. Hypotese 1 søger derfor at få svar på selve kernen i projektet, og vil derfor kunne bane vej for en mere nuanceret analyse.

Hypotese 2 søger at uddybe på dekomponerings-komponenten i dette projekt, for at se om et endnu bedre resultat kan skabes.

Hypotese 3 tager kernen i problemformuleringen, som også bliver testet i hypotese 1, og undersøger om der kan forbedres på den, ved at tage et skridt tilbage, og anvende en alternativ fremgangsmåde. Til besvarelsen af hypotese 3, vil PM og ATO forecastes hver for sig, i to simple lineære regression, hvorefter RNOA estimeres, baseret på disse.

Hypotese 4 tager fremgangsmåden fra hypotese 3 ved at, som udgangspunkt, opdele modellen i tre simple lineære regressioner, og derefter bruges resultaterne til udregning af fremtidige værdier. Dette betyder at

hypotese 4, søger at forecaste GP, OPPL og NOA, separat og derefter estimere RNOA. Det vises senere i projektet, hvorfor GP her er overflødig, og variablen derfor udelukkes fra denne model.

Følgende underspørgsmål opstilles, for at svare på både de overstående hypoteser, samt problemformuleringen:

1. Er det en fordel, med fokus på statistiske præcision, at forecaste ATO og PM til at forklare RNOA, i modsætning til udelukket at forecaste RNOA?
2. Vil en yderligere dekomponering af ATO, samt PM kunne forbedre forecastet af regnskabstallene?
3. Vil den statistiske præcision forbedres, hvis der bliver forecastet, udelukket på value drivers, i forhold til underspørgsmål 1 og 2?
4. Hvis dette projekt udelukket skulle vælge en model, hvilken ville så ses, som at være den bedste?

Bemærk, at de overstående underspørgsmål gennemgås og refereres til, i overstående rækkefælge.

1.3 Afgrænsninger

Dette projekt vil, som udgangspunkt, afgrænse sig til kun at omhandle, hvad der er forstået ved problemformuleringen, underspørgsmålene, samt de relevante informationer, til test af hypoteserne.

Hertil betyder det for eksempel, at det ligger ude for projektets pladsmæssige og tidsmæssige begrænsninger, at teste alle dekomponerede variabler, for alle virksomheder, over alle industrier. Denne afgrænsning er også født af, at der bliver arbejdet med en stor mængde observationer, industrier og variabler, som gør hele denne proces eksponentielt mere vanskelig. Derudover holder dette projekt sig inden for det overordnede nøgletal RNOA's rammer, og de dertilhørende nødvendige dekomponerede nøgletal.

1.4 Struktur

Dette projekt, startede med en indledning, der præsenterede problemområdet, samt de overordnede tanker bag projektet. Dernæst blev der redegjort for, hvorfor dette projekt er interessant, samt hvor vigtigt disse former for analyser kan være, med andre ord, formålet med selve projektet. Med projektet introduceret, og dens formål opridset, er der banet vej for, at problemformuleringen, og dertilhørende hypoteser og underspørgsmål. Dette er efterfulgt af de tre afsnit omhandlende data, teori og metode.

Metodeafsnittet er efterfulgt af analysen, der søger at bygge videre på projektet, ved at gennemgå de opstillede modeller. Projektet tager derefter viden skabt i analysen, til at besvare de overstående underspørgsmål i diskussionen. Derefter kan hypoteserne gennemgås, så der endegyldigt kan svares på

problemformuleringen. Diskussionens næstsidste afsnit er et perspektiverings- og begrænsningsafsnit, der søger at uddybe på, hvilken emner kan tilføje værdi til projektet, hvorefter diskussionen bliver afrundet med en delkonklusion. Derefter bliver dette projekt afsluttet med en konklusion, som opridser de væsentligste punkter fra projektet.

2. Data

Dette afsnit vil omhandle dataet, som bliver anvendt i projektet, for at kunne give en fyldestgørende besvarelse af problemformuleringen. Dataet er hentet fra hjemmesiden Orbis, som har informationer om 310 millioner firmaer verden over (Orbis, 2019).

Datasættet består af næsten 2,9 millioner observationer over 152 variabler. Dette datasæt indeholder mange overflødige variabler, samt fejlagtige observationer. Selve datasættet er på alle danske virksomheder, der har udgivet regnskab mellem årene 1990-2013. Til dette projekt anvendes programmet RStudio til at udarbejde analysen, og i forbindelsen med afleveringen af dette projekt, vil både koden, samt datasættet afleveres. Læserne kan derfor genskabe resultaterne, som dette projekt når frem til.

For at skabe et mere overskueligt datasæt, fjernes alle overflødige variabler. Dernæst skabes en oversigt over variablen ”ARYEAR” (denne variable er årstal), hvilket viser et problem med datasættet. Dette problem ses i følgende tabel:

ARYEAR	Min.	1. kvartal	Median	Mean	3. kvartal	Maks.	NA
Årstal	-1,380e+06	2001	2006	2,031e+16	2009	3,702e+22	155.836

Tabel 2 - af egen inspiration.

Minimumsværdien viser år -1.380.000, og maksimumsværdien viser år 3,702e+22, hvilket er meget indlysende en fejl i datasættet. Derudover er der 155.836 observationer, som ikke har nogen værdi (NA). Baseret på tabel 2 ses det at, første og tredje kvartil, samt medianen, ser fornuftige ud. Dette hentyder til, at der er nogle ekstreme værdier i datasættet, men det viser sig kun at være 768 observationer. En kode bliver anvendt, til at slette observationer, som har et årstal over 2013, og samtidigt under 1990.
Dernæst testes for, om passiver og aktiver er lige store, for at fremhæve yderligere fejl i datasættet. Resultatet af balancetesten giver baggrund for at fjerne yderligere 10.452, observationer.

Senere i dette projekt vil udregning af NOA gennemgås. Hertil er der en række observationer, som ikke har nogen værdi, disse er blevet slettet. Dette betyder, at datasættet reduceres til 146.701 observationer, hvilket

stadig er en tilfredsstillende størrelse. Ligesom balancen blev testet tidligere i dette projekt, så vil der også blive kørt en test på RNOA. Hertil skal følgende relation gælde:

$$RNOA_t = \frac{OPPL_t}{NOA_t} = \frac{OPPL_t}{GP_t} * \frac{GP_t}{NOA_t}$$

En variabel er blevet konstrueret, som tester om produktet af PM og ATO giver det samme resultat, som OPPL divideret med NOA. Denne test virker ligegyldig, i det at matematikken stemmer, men på trods af dette, er der alligevel 51.264 observationer, hvor ovenstående relation ikke holder. Grunden til dette er højst sandsynligt afvigelser i datasættet, og der ses ingen anden udvej, end at slette disse observationer. Resultatet af dette, betyder at der nu er 95.437 observationer tilbage, hvilket stadig ses, som at være en betydelig mængde af observationer. Ydermere bliver der undersøgt og slettet de observationer uden værdi, som befinder sig i de relevante value drivers, dette reducerer datasættet til 95.365 observationer.

2.1 Ekstreme outliers

I dette afsnit, vil der blive fjernet de ekstreme outliers der er i datasættet. Først tages der udgangspunkt i RNOA, som er det essentielle nøgletal for dette projekt, og baggrunden for benchmark-modellen. Et residualplot over RNOA ses i bilag 1. På baggrund af bilag 1 ses, at der er ekstreme outliers i datasættet. For de to value drivers ATO og PM (Plenborg, Petersen, & Kinserdal, 2017), ses et ligeledes residualplot i bilag 2. For at kunne give et bedre overblik over observationerne for RNOA, og vurdere størrelsen af disse outliers er følgende tabel konstrueret:

	Min.	1. kvartil	Median	Mean	3. kvartil	Maks.
RNOA	-4.129	0,058	0,181	0,317	0,416	2.109

Tabel 3 - af egen inspiration

Disse outliers kan håndteres på flere forskellige måder. De to forskellige metoder dette projekt har overvejet at bruge, er taget fra Fairfield og Yohn (2001), og Soliman (2008). Soliman (2008) har en meget direkte fremgangsmåde når det kommer til håndteringen af outliers, hvilket er at fjerne den nederste og øverste 1 % af datasættet. Fairfield og Yohn (2001), har en lidt anderledes måde at håndtere outliers på. De sletter observationer, hvor ændringen af RNOA, PM og ATO, er stører end 0,5, eller for observationer, som har en RNOA med en absolut værdi, som er større end 1. Dette projekt finder ikke, at Fairfield og Yohn (2001) argumenterer overbevisende for deres lidt komplekse forslag, hvorimod forslaget fra Soliman (2008) virker intuitivt, fornuftigt og almindelig anvendt i praksis. Ydermere er det valgt, at outliers bliver fjernet i en "top-

down approach", hvilket i dette tilfælde, omhandler de value drivers (Plenborg, Petersen, & Kinserdal, 2017), som er indforstået med en simple dekomponering. Både bilag 1 og 2, viser hvordan der er nogle voldsomme outliers for både RNOA, ATO og PM. Efter anvendelsen af metoden fra Soliman (2008) på RNOA, PM og ATO reduceres datasættet til 90.942 observationer. I bilag 1 og 3 ses effekten af denne ændring i et residualplot for RNOA, PM og ATO. Nedenstående tabel, viser de værdier, som er gælder for den version af RNOA, der har fjernet for ekstreme outliers:

	Min.	1. Kvartil	Median	Mean	3. Kvartil	Max
RNOA	-2,7188	0,0650	0,1829	0,2907	0,4042	4,7016

Tabel 4 - af egen inspiration.

Tabel 4 viser en kraftig ændring i minimums- og maksimumsværdien for RNOA. Hertil er det stadig høje værdier, set på absolutte tal, men det er en stor forbedring, i forhold til tabel 3.

2.2 Datastrukturen

De kommende modeller vil for eksempel anvende RNOA fra år 2003, til at estimere RNOA for år 2004. Dette lyder simpelt, og for at lave disse modeller, vil de uafhængige variabler blive lagget, for at kunne håndtere sammenhængen mellem uafhængige og afhængige variabler. Ydermere skaber dette to problemstillinger, som skal korrigeres for på følgende måde.

Dette projekt er afhængig af, at der er en ubrudt række af årlige observationer i hele perioden for hver virksomhed. Hvis dette ikke er tilfældet, vil modellerne risikere ikke at forecaste på foregående kalenderår. Derudover risikeres også, at en virksomheds regnskabstal, anvendes til at beskrive en anden virksomheds regnskabstal.

Disse to problemstillinger, bliver løst ved at lagge alle variabler der skal anvendes i projektet på forhånd, og dernæst køre følgende tests:

$$ARYEAR.test = ARYEAR - l(ARYEAR)$$

Her skal det bemærkes, at $l(ARYEAR)$, betyder den laggede version af ARYEAR. Hvis resultatet for ovenstående ligning ikke er 1, slettes observationen. Dette betyder, at for de resterende observationer, er der ikke et spring større end 1 år.

Problemet med, at en virksomheds data anvendes til at forecaste den næste virksomheds nøgletal, rettes på en lignende måde. Hertil opstilles en "hvis-funktion", der tager CVRNR (virksomhedsnummeret) for observationerne, og sammenligner med $l(CVRNR)$ (igen, den lagged version af CVRNR). En succesfuld sammenligning printer resultatet 1, ellers resultatet 0. Dernæst slettes alle observationer, som har værdien 0.

Efter disse korrektioner, er datasættet reduceret til 39.455 observationer, og der kan nu sikres, at modellerne anvender år t, til at beskrive år $t + 1$. Derudover sikres der også, at ingen modeller anvender en virksomheds nøgletal, til at forklare en anden virksomheds nøgletal.

3. Teori

Dette afsnit vil omhandle de bærende artikler og teorier, som dette projekt vil anvende til besvarelse af problemformuleringen.

3.1 Fairfield & Yohn (2001)

Fairfield og Yohn skrev i år 2001, en artikel ved navn:

"Using Asset Turnover and Profit Margin to Forecast Changes in Profitability"

Hovedpunkterne I denne artikel vil blive gennemgået i dette afsnit, da denne artikel er essentiel for projektet, samt udgør inspiration til hypotese 1, som er vist i Tabel 1. I indledningen af dette projekt, blev Fairfield og Yohn (2001) diskuteret, i forbindelse med deres udsagn omkring mangel af information på det overordnede emne. Denne artikel vil hovedsageligt blive brugt, som en form for teoretisk inspiration. Dette projekt søger altså ikke at teste præcis det samme, som Fairfield og Yohn gør i deres artikel fra 2001. Hertil anvendes artiklen når det gælder opbygning af modeller, og centrale teoretiske ideer. Ydermere skal det nævnes, at den overordnede hypotese præsenteret af Fairfield og Yohn (2001) bliver testet i en modificeret tilstand, da den ligger parallelt, med dette projekts problemformulering.
Fairfield og Yohn's artikel fra 2001, tester om den, "fundamentale dekomponering" af RNOA er brugbar i en forecasting kontekst. Den "fundamentale dekomponering", som bliver beskrevet i artiklen (Fairfield & Yohn, 2001), svarer til hvad dette projekt har kaldt en "et-trins dekomponeringen" og "simple dekomponering". Dette betyder selvfølgelig, at RNOA bliver opdelt i ATO og PM. Fairfield og Yohn (2001), forklarer dernæst, at deres hypotese går ud på, at en ændring i ATO, og altså ikke PM, vil give information omkring den fremtidige rentabilitet i virksomheden. En stigning i ATO, øger produktiviteten i virksomheden, og derfor kan den øge den fremtidig rentabilitet. Ydermere vil en stigning i PM, på grund af virksomheden bliver mere efficient, også

resultere i forbedret fremtidig rentabilitet. Dog vil en stigning i PM på grund af regnskabsmæssige konservativisme ikke øge den fremtidige rentabilitet i virksomheden (Fairfield & Yohn, 2001). Fairfield og Yohn (2001) har derfor, som udgangspunkt, at der ikke er stor mulighed for, at PM kan give et godt indtryk i fremtidig rentabilitet. Ydermere påstår Fairfield og Yohn (2001) at det er ændringen i ATO, som analytikere og investorer skal fokusere på, når de har i sinde at forecaste profitabiliteten i en virksomhed, for at finde virksomhedens værdi (Fairfield & Yohn, 2001). Dette er på grund af, at ændringen i PM er vurderet til at være insignifikant (Fairfield & Yohn, 2001). Fairfield og Yohn (2001) har fundet disse resultater, ved at opstille fire modeller (model 0, 1, 2 go 3).

Modellerne søger at teste følgende fire aspekter (Fairfield & Yohn, 2001), der viser ændringen af RNOA:

- Model 0 er en naiv model, der forudser ingen ændring om 1 år i RNOA.
- Model 1 viser Fairfield og Yohn's (2001) tre kontrolvariabler, RNOA, ændringen i NOA og ændringen i RNOA.
- Model 2 bygger videre på model 1, ved at tillægge PM og ATO.
- Model 3 ændrer på model 2, ved at se på ændringen i PM og ATO, samt en variabel, hvor der tages produktet af disse to variabler.

Her er det specielt model 2 og model 3, som er interessante. Resultaterne fra dem er, at model 2 ikke viser en forbedring i forecasts, fremfor model 1, men model 3 giver en signifikant forbedring. Konklusionen på dette er, at ifølge Fairfield og Yohn (2001), forbedres et forecast af RNOA ikke, ved at dekomponerer i ATO og PM. Derimod forbedres et forecast af ændringen af RNOA, hvis der bliver dekomponeret i ændringen af både ATO, samt PM.

På baggrund af denne artikel, vil dette projekt søge at teste påstanden, om der kan forbedres på et forecast af RNOA, ved at dekomponere i PM og ATO. Derudover vil der anvendes en alternativ fremgangsmåde, for at se om en ny fremgangsmåde kan forbedre dette resultat. Denne fremgangsmåde præsenteres senere i projektet.

3.2 Soliman (2008)

For at kunne støtte artiklen fra Fairfield og Yohn (2001), samt problemformuleringen og de dertilhørende hypoteser, så vil følgende artikel fra Soliman (2008) anvendes:

" The Use of DuPont Analysis by Market Participants"

De grundlæggende punkter fra denne artikel, vil blive gennemgået i dette afsnit, samt artiklens relevans for dette projekt.

Denne artikel søger at tilføje til litteraturen, inden for emnet om analytikernes og investorernes reaktioner til DuPont komponenterne langs tre dimensioner (Soliman, 2008). Disse tre dimensioner, beskriver Soliman (2008), som at være følgende (Soliman, 2008):

- Genskabe de forrige dokumenteret forecasting evner og vurder om dette er robust og nødvendigt i forhold til andre forudsætninger i den nuværende litteratur.
- Udforskning af, hvorvidt dette bliver brugt af aktiemarkedets investorer.
- Til sidst undersøger denne artikel både nuværende forecasts og fremtidige forecast fejl.

Soliman (2008), anvender i sin analyse nøgletallet RNOA, hvilket bliver opdelt i både PM og ATO. Måden dette bliver opdelt på er simple, og ses i følgende formel:

$$RNOA = PM * ATO$$

Ifølge Soliman (2008), er PM og ATO regnskabssignaler, som måler forskellige aspekter, af virksomhedens opbygning af deres operationer. Hertil er PM ofte et resultat af, prisfastsættelsen, produktinnovation, produktpositionering, brandbevidsthed, ”first mover advantage” og ”market niches” (Soliman, 2008). Derudover er ATO defineret af Soliman (2008), som værende en måleenhed af, hvor efficient en virksomhed er med deres aktiver. Efficient brug af aktiver, sker gennem maksimering af anvendelsen af ejendomme, udstyr og efficient lagerprocesser (Soliman, 2008). Grunden til disse variabler og deres underliggende faktorer er vigtige, er for at få et indblik i, hvordan, og i hvilket omfang, konkurrence kan påvirke disse to variabler (Soliman, 2008). For eksempel, kan en høj PM i en branche, ofte tiltrække nye virksomheder til markedet, med hurtige imitationer af produkter, eller nye ideer til at udkonkurrere eksisterende virksomheder (Soliman, 2008). Dette vil resultere i, at en relativ høj PM på markedet, skrider mod et ”normalt niveau” for dette marked.

Soliman (2008) påpeger dernæst, hvordan Fairfield og Yohn (2001) beskriver ATO’s rolle i denne sammenhæng. Soliman (2008) fremhæver, at en virksomhed med høj RNOA, baseret på en høj værdi for ATO, er mindre truet af nye konkurrenter, end en virksomhed med høj RNOA, drevet af høj PM. Dette er grundet at, det er mere besværligt for en virksomhed at imitere en anden virksomheds efficiente brug af aktiver, da dette ofte kræver et større kapital og ændring for nuværende fabrikker og operationer (Soliman, 2008). Soliman (2008) påpeger i sit værk, at Fairfield & Yohn (2001), ikke tester ATO og PM, separeret fra den udeladte variable, $\Delta RNOA$.

Resultatet af, at Fairfield og Yohn (2001) ikke har kontrolleret for, $\Delta RNOA$'s statistiske beskrivende kræfter, kan sætte spørgsmålstegn, ved den signifikans, deres beskrivende variabler har (Soliman, 2008).

Soliman (2008) viser, at der til dels er overensstemmelse med den litteratur han vælger at teste, her iblandt Fairfield og Yohn (2001). Resultatet er, at ATO og PM er signifikant, når det gælder at beskrive udviklingen i RNOA (Soliman, 2008). Hertil skal det bemærkes, at dette er lidt i konflikt med tidligere litteratur af Fairfield og Yohn (2001), som viser at det kun er ændringen i ATO, som er signifikant til at forudsige udviklingen i RNOA, og altså ikke PM. Ydermere beviser denne artikel også, hvor nyttig DuPont analysen er, når det gælder undersøgelser, angående finansielle nøgletal (Soliman, 2008).

Dette projekt vil bruge Soliman (2008) til opbyggelsen af analysen, samt den allerede anvendte strukturering af de gennemgået hypoteser, underspørgsmål, samt problemformuleringen. Ydermere, vil artiklerne fra Soliman (2008) og Fairfield og Yohn (2001), også være et fokuspunkt, når det gælder forventningsafstemningen, af resultater fremover, både i den analytiske og diskuterende del af projektet. Derudover vil Soliman (2008), være baggrunden for næste afsnit, som omhandler DuPont pyramiden. Hvor Fairfield og Yohn (2001) bygger en solid base, vil deres konklusioner ikke kunne stå alene, uden korrektionerne, som Soliman (2008) bringer.

3.3 DuPont pyramiden

I dette afsnit vil der blive gennemgået DuPont pyramiden, til brug af dekomponering af regnskabstal. DuPont pyramiden kan også kaldes for en DuPont analyse, og i dette projekt vil DuPont pyramiden blive anvendt til at vise sammenhængen mellem de relevante nøgletal. Dette afsnit vil udelukkende fokusere på den del af DuPont pyramiden, som har relevans for projektet. Den essentielle del af DuPont pyramiden, bliver gennemgået i Soliman's (2008) værk, som er baseret på Penman og Nissim (2001). I bilag 4, ses en modificeret DuPont pyramide, baseret på de to tidligere nævnte kilder, som kan bruges til at visualisere dekomponeringen af RNOA.

Forholdet mellem de forskellige variabler, samt måden de bliver udregnet på, vil blive gennemgået senere i opgaven. Bilag 4 viser, at RNOA er baseret på de to variabler, PM og ATO. PM og ATO består hver af to variabler, hvor disse har GP (gross profit) tilfælles. Derudover er PM udledt af variablen OPPL (operating profit/loss, også kendt, som driftsresultatet eller EBIT) og ATO, som udregnes med variablen NOA (net operating assets). I dette projekt vil DuPont pyramiden fra bilag 4, blive beskrevet på to overordnede måder. Disse to måder er, som følger:

- Simpel dekomponering eller ”et-lags dekomponering” (dekomponering i et lag), i dette projekt defineret, som RNOA dekomponeret i PM og ATO.
- ”to-lags dekomponering” (dekomponering i to lag), i dette projekt defineret, som dekomponeringen af RNOA, til OPPL, GP og NOA.

Begreberne bliver anvendt for at gøre det nemmere at kommunikerer, hvilken del af bilag 4, som der arbejdes med.

3.4 Plenborg, Petersen & Kinserdal (2017) – Value drivers

Følgende afsnit, er baseret på kapitel 8, ved navn ”Forecasting”, af Plenborg, Petersen og Kinserdal (2017), i bogen:

” Financial Statement Analysis”

Dette afsnit vil gennemgå de mest relevante dele af kapitlet, og hvordan de hænger sammen med opgaven, her specielt de overstående hypoteser. Det vigtigste begreb fra dette kapitel, er ”value drivers”, som herefter vil bruges i projektet. Value drivers er delt op i to overordnede kategorier (Plenborg, Petersen, & Kinserdal, 2017):

- Strategic value drivers.
- Financial value drivers.

Strategiske value drivers, er handlinger en virksomhed kan tage, for at skabe værdi i virksomheden. Eksempler på strategic value drivers, er lanceringen af et nyt produkt, outsourcing af dele af virksomheden og indtræden på nye markeder (Plenborg, Petersen, & Kinserdal, 2017). Financial value drivers er finansielle forhold, hvor eksempler på dette er, væksttal, marginer og andre investeringsforhold. Strategic value drivers er ”input faktorer” og financial value drivers er set, som ”output faktorer”. Sammenhængen mellem strategic og financial value drivers, er at de strategiske og operative handlinger, påvirker de financial value drivers (Plenborg, Petersen, & Kinserdal, 2017). De strategic value drivers, er ikke relevante for dette projekt, da introduktionen af et nyt produkt, outsourcing af dele af virksomheder osv., ikke kan måles på det data, som projektet har til rådighed. Dette er på grund af, at projektet fokuserer sine resultater på regnskabstal spredt over en lang række virksomheder, hvilket betyder at der vil fokuseres på regnskabets value drivers, når dette begreb anvendes. Bemærk hertil, at når der bliver nævnt value drivers fremover, betyder det ikke nødvendigvis financial value drivers, da dette ikke gælder variabler, som NOA (Plenborg, Petersen, & Kinserdal, 2017).

Hertil præsenterer Plenborg, Petersen og Kinserdal (2017), metodikken, hvor hver linje i regnskabet, bliver forecastet separat. Hver linje i regnskabet der bliver forecastet, bliver derfor omtalt value drivers. Dette projekt vil derfor følge denne alternative fremgangsmåde, hvor der vil fokuseres på de enkelte value drivers, som driver nøgletallet RNOA. I bilag 4 ses de seks value drivers, som dette projekt vil arbejde med.

Hypoteser 3 og 4, har forbindelse til dette kapitel, da det er her, den alternative fremgangsmåde bliver introduceret. Plenborg, Petersen og Kinserdal (2017), forecaster med en alternativ fremgangsmåde i forhold til andre værker, som er blevet gennemgået i dette projekt. Soliman (2008) og Fairfield og Yohn (2001), anvender multipel lineære regressioner, som deres forecasts, hvor Plenborg, Petersen og Kinserdal (2017), vælger at forecaste på hver enkelte value driver. Denne ændring i modeldesign, har dette projekt valgt at undersøge, for at se om dette vil kunne gøre en forskel, i forhold til præcisionen af de følgende modeller. DuPont pyramiden bliver anvendt, som værktøj til at identificere de value drivers, der skal anvendes, og derefter vil hver enkelt value driver forecastes individuelt, for at estimere RNOA. Den alternative fremgangsmåde opstilles ved siden af de mere ”traditionelle” modeller, hvilket vil blive uddybet i senere afsnit.

Et af argumenter for denne måde at forecaste på, er for at sikrer sig, at balancen vil stemme efter forecastet (Plenborg, Petersen, & Kinserdal, 2017). Dette projekt anser denne fremgangsmåde til at være specielt interessant, og vil derfor anvende fremgangsmåden, for at skabe nye modeller, og et helt nyt og unikt synspunkt på problemformuleringen.

3.5 Bowerman, O’Connell & Koehler (2005) og Osborne & Waters (2002) - Antagelser for lineære regressioner

Dette teoriafsnit er opdelt i to, på grund af begge kilders sammenhæng med emnet. Hertil introducere disse to kilder en samlet oversigt af antagelser, som de følgende modeller skal overholde.

Bowerman, O’Connell & Koehler (2005):

Bowerman, O’Connell og Koehler (2005) præsentere fire overordnede lineære antagelser. Disse fire antagelser bruges senere i analysen, hvor modellerne bliver forsøgt forbedret på baggrund af viden, præsenteret i dette afsnit. Dette afsnit er baseret på viden fra bogen ”Forecasting, Time series, and Regression” af Bowerman, O’Connell og Koehler (2005). For at sikrer de fremtidige modeller er bygget på pålidelige data, og de følgende hypotesetests kan anvendes, fremviser Bowerman, O’Connell og Koehler (2005), fire antagelser. Disse fire antagelser skal foretages, når der arbejdes med simpel og multipel lineære regressioner:

1. Ved en hvilken som helst værdi af de uafhængige variabler, så har populationen af de mulige fejlleds værdier en gennemsnitsværdi på 0 (Bowerman, O'Connell, & Koehler, 2005).
2. Anden antagelse omhandler den konstante variansantagelsen. Denne antagelse går ud på, at for en given værdi af den uafhængige variable, så er populationen af det potentielle fejlleds værdi af variansen ikke afhængig af de uafhængige variabler. Dette betyder at for de forskellige værdier for fejlleddene er det den samme varians der er tilknyttet alle værdier for den tilhørende uafhængige variable (Bowerman, O'Connell, & Koehler, 2005).
3. Tredje antagelse går ud på, at alle værdier for de uafhængige variabler og potentialet for fejlleddenes værdier er normalfordelt (Bowerman, O'Connell, & Koehler, 2005).
4. Den sidste antagelse fra Bowerman, O'Connell og Koehler (2005) omhandler uafhængigheden af fejlleddene. Dette betyder, at for en given værdi af et fejlled der beskriver den afhængige variable, er den uafhængig af andre fejlled til en hvilken som helst anden værdi af den afhængige variable.

De tre første antagelser for de simple og multiple lineære regressioner omhandler essentielt at fejlleddene er normalfordelt med en mean på nul, og en uafhængig variabel, hvis varians er uafhængige af dens værdi. Disse antagelser er relativt lette at tjekke, hvilket bliver set senere i projektet, når der bliver opstillet histogrammer for de forskellige uafhængige variabler. Bowerman, O'Connell og Koehler (2005) beskriver ikke tydeligt, hvordan der håndteres, hvis en af eller flere af disse antagelser ikke holder. Dertil bliver der inddraget artiklen fra Osborne og Waters (2002), som har nogle af de samme antagelser, som Bowerman, O'Connell og Koehler (2005). Grunden til, at dette projekt ikke bare udelukket bruger Osborne og Waters (2002), er fordi Bowerman, O'Connell og Koehler (2005), sætter de fire antagelser op på baggrund af fejlleddene. Dertil har Osborne og Waters (2002) nogle lidt andre antagelser, som er mere fokuseret på selve den lineære regression. Derfor er disse to værker sammensat under et afsnit i teorien, da de er meget ens, og det ene værk supplerer, hvor det andet værk er lidt mangelfuldt.

Osborne & Waters (2002)

Osborne og Waters (2002) påpeger, at der er mange videnskabelige artikler, som ikke følger eller tjekker for de fire antagelser. Dette gør, at de mange statistiske test ikke er pålidelige, da de er baseret på variabler, som ikke er egnet til tests. For at teste, at fejlleddene er normalfordelt, så foreslår Osborne og Waters (2002) en række forskellige metoder. Heriblandt P-P plots, visualisering af data plots, kurtosis, skew og histogrammer. Hvis en variabel viser sig, ikke at være normalfordelt, vil dette korrigeres for.

1. Første antagelse fra Osborne og Waters (2002) er antagelsen om et lineært forhold mellem de afhængige og uafhængige variabler. Hertil kan et residual plot observeres, hvor der kan observeres, hvilken form dataet har.
2. Anden antagelse er pålideligheden af estimaterne. Hertil er det multikollinearitet der bliver testet, hvilket betyder der bliver testet for, at i hvor høj grad de uafhængige variabler er afhængige af hinanden. Hertil er der fokus på de multiple lineære regressioner, som har mere end en variabel, og en åbenlys løsning til dette problem, er at fjerne variabler, som er meget højt korreleret med andre. Hertil vil en korrelationsmatrice opsættes, og støttes med en VIF-test.
3. Dernæst er tredje antagelse baseret på homoskedasticitet. Homoskedasticitet betyder, at variansen for fejlleddene er ens for alle uafhængige variabler. Dette er samtidigt antagelse fire, i forhold til Bowerman, O'Connell og Koehler (2005). Hvis variansen ikke er den samme for alle af de uafhængige variabler, kaldes det heteroskedasticitet. En lille hændelse af heteroskedasticitet har en minimal effekt på signifikanstesten, men en betydelig forekomst af heteroskedasticitet, kan kraftigt påvirke de lineære modeller (Osborne & Waters, 2002). Denne antagelse kan undersøges visuelt af et plot af de standardiseret fejlled ved siden af regressionens standardiseret forventede værdier. I denne test kan heteroskedasticitet tage flere forskellige former, og målet er, at residualerne skal være tilfældigt fordelt omkring en værdi af nul, for at åbne homoskedasticitet. En transformation af variablerne, kan være med til at opnå homoskedasticitet i dataet (Osborne & Waters, 2002).
4. Til sidst er fjerde antagelse for Osborne og Waters (2002), lignende den fra Bowerman, O'Connell og Koehler (2005), hvilket er antagelsen om normalitet. Da denne antagelse er gennemgået for Bowerman, O'Connell og Koehler (2005), vil den ikke gennemgås igen.

Med de overstående antagelser fra både Bowerman, O'Connell og Koehler (2005), samt Osborne og Waters (2002), kan dette projekt sikrer sig, at modellerne i dette projekt, er baseret på en solid og statistisk korrekt baggrund. De overstående rettelser for antagelserne, vil anvendes på de modeller, hvor det vurderes det er nødvendigt.

Opsamling på afsnittet

Dette afsnit kan være lidt indviklet med de to forskellige kilder, der hver har lignende antagelser, men stadig nogle forskelle. De er begge taget med for at kunne understrege det data, samt modellerne selv, er bygget på et pålideligt fundament. Derfor er denne afrunding med, for at klargøre, hvordan de overstående antagelser vil bruges. Dette vil selvfølgelig tage baggrund i alle antagelser fra både Bowerman, O'Connell og Koehler (2005),

samt Osborne og Waters (2002). Nedenstående er en gennemgang af, hvilke antagelser helt konkret, som modellerne skal opfylde:

- Der er et lineært forhold mellem afhængige og uafhængige variabler.
- Fejlleddene for de uafhængige og afhængige variabler er normalfordelte med en mean omkring nul.
- De uafhængige variabler er uafhængige af hinanden. Med andre ord testes der her for multikollinearitet i sættet.
- Variansen er konstant for alle fejlled, og fejlleddene er uafhængige af hinanden. Dette betyder, at der er homoskedasticitet i datasættet.

Efter hver model er blevet opsat med de relevante variabler, vil disse fire ovenstående antagelser blive testet. Dette er hovedsageligt for at sikre modellen er brugbar, ud fra et statistisk perspektiv, og dernæst for at forsøge at forbedre på modellen. Bemærk, at det er ikke alle antagelser, som er lige relevante for alle modeller. For eksempel, så er der ikke tale om multikollinearitet for en simple lineær regression.

4. Metode

I dette afsnit gennemgås de overordnede begreber, variabler, modeldesign og præcisionsmål for dette projekt. Dette projekt søger en høj grad af validitet og pålidelighed, hvilket er nået gennem valg af kilder og det anvendte data. Dette er blandt andet artikler i anerkendte tidsskrifter inden for de relevante felter, bøger og dokumenter fundet på websteder. Kilderne hertil er fundet blandt andet på hjemmesiderne Scopus og Google Scholar, hvor artiklerne er blevet vurderet, på baggrund af mængden af citationer, før de er blevet valgt. Derudover har dette projekt haft stort fokus på, at forfatterne af disse kilder, har en højere uddannelse, inden for emnet deres værker omhandler.

4.1 Variablerne og deres udregning

I dette afsnit vil der blive beskrevet, hvordan de forskellige variabler (også omtalt, som value drivers) bliver udregnet, her med hjælp fra bilag 4. Bilag 4 har de seks variabler:

- RNOA
- PM
- ATO
- OPPL
- GP

- NOA

For at besvare problemformuleringen vil RNOA, som udgangspunkt, udregnes på følgende måde:

$$RNOA_t = \frac{OPPL_t}{NOA_t}$$

Denne formel vil bruges til at udregne RNOA, til at forecaste den simple model, hvor der ikke bliver dekomponeret. Dernæst foretages en simple dekomponering af RNOA, for at opdele RNOA i de to value drivers, og det ser således ud:

$$RNOA_t = PM_t * ATO_t$$

Dette er en af de vigtige relationer, som dette projekt vil undersøge i dybden i analysen. PM og ATO bliver derefter dekomponeret i OPPL, GP og NOA, gennem følgende relationer:

$$PM_t = \frac{OPPL_t}{GP_t}$$

$$ATO_t = \frac{GP_t}{NOA_t}$$

PM og ATO bliver opdelt i tre yderlige value drivers, som bliver brugt i de efterfølgende modeller. I analysen fokuserer variablerne derfor på dækningsbidraget (GP/Gross profit) og netto operative aktiver (NOA), samt driftsresultatet (OPPL).

Da dette projekt tager udgangspunkt i offentlig tilgængelige data, er det bundet af lovgivningen for, hvad virksomheder skal oplyse i deres regnskab. Derfor er det første regnskabstal der beskriver salgstal, som dette projekt kan fremskaffe, er GP.

NOA bliver udregnet ved at trække operating liabilities fra operating assets:

$$NOA = operating\ assets - operating\ liabilities$$

Datasættet, som dette projekt arbejder med, har følgende poster til rådighed for at udregne operating assets:

- Anlægsaktiver i alt (FA – kode for variablen).
- Finansielle anlægsaktiver i alt (OFA).
- Varebeholdning (STOCK).
- Tilgodehavender fra salg og tjenesteydelser (DEBTO).

- Øvrige omsætningsaktiver (OCA).

Posterne fra datasættet, der bruges til udregning af operating liabilities, er som følger:

- Kortsigtede forpligtelser i alt (CL).
- Andre forpligtelser (inkl. Hensættelser) (ONCL).

Datasættet har to begrænsninger på hvordan NOA bliver udregnet. Fokuseres der først på aktiverne, kan anlægsaktiverne ikke deles op i immaterielle, materielle og finansielle anlægsaktiver i alt. Dette er på grund af en signifikant mangel på observationer under de immaterielle aktiver, derfor vælges der at anvende anlægsaktiver i alt, og dernæst fratrækkes finansielle aktiver. For operating liabilities er der et lignende problem. De kortsigtede forpligtelser i alt, ses gerne fratrukket de kortfristede lån, dog har dette ikke være muligt, igen på grund af mangel af observationer. Dette betyder, at NOA bliver udregnet på baggrund af de kortfristede forpligtelser i alt, og derefter tillægges de andre forpligtelser (inkl. Hensættelser). NOA udregnes mere præcist på følgende måde (kode for variabler ses ovenfor):

$$NOA_t = (FA_t - OFA_t + STOCK_t + DEBTO_t + OCA_t) - (CL_t + ONCL_t)$$

4.2 Modeldesign

I følgende afsnit præsenteres de forskellige modeller, som dette projekt benytter sig af, for at besvare problemformuleringen, samt hypoteserne og underspørgsmålene. Modellerne er baseret på teorien fra Soliman (2008), samt Fairfield og Yohn (2001). Koefficienterne og konstanterne er udregnet på baggrund af dette projekts tilknyttet datasæt.

4.2.1 Model 1:

Model 1 forudsiger niveauet af RNOA, udelukket ved hjælp af historisk data for RNOA. Model 1 ser ud, som følger:

$$RNOA_{t+1} = \alpha_0 + \beta_1 * RNOA_t$$

Model 1 er en simpel lineær regression, med α_0, som er konstanten og β_1, som koefficienten, der bestemmer den effekt det historiske data af RNOA har på fremtidige resultater. Denne notation anvendes i resten af projektet. Modellen har til hensigt at være et benchmark, for hele analysen. Resultatet for model 1, bruges derfor til at sammenligne, om der sker en forbedring af forecastet, ved hjælp af model 2, 3, 4 og 5. Modeller der ikke forbedre dette benchmark kan derfor ikke anbefales til fremtidig brug.

4.2.2 Model 2:

Model 2 anvendes til at undersøge hypotese 1, hvilket er at forbedre model 1, ved at anvende en simpel dekomponering. Dette betyder altså, at RNOA bliver opdelt i ATO og PM, til en multipel lineær regression. Model 2 ser således ud:

$$RNOA_{t+1} = \alpha_0 + \beta_1 * PM_t + \beta_2 * ATO_t$$

Som i model 1, er α_0 konstanten og β-værdierne er variablernes koefficienter, som bestemmer nøgletallenes påvirkning af resultatet.

4.2.3 Model 3:

Model 3 ser ud som følger:

$$RNOA_{t+1} = \alpha_0 + \beta_1 * OPPL_t + \beta_2 * GP_t + \beta_3 * NOA_t$$

Her er der foretaget en dekomponering i to lag, for at beskrive det fremtidige niveau af RNOA.

4.2.4 Model 4a og 4b:

Model 4 er essentielt opdelt i to modeller, hvor resultatet kombineres for at undersøge, om et forecast på de to value drivers, kan forbedre på de tidligere modeller. Model 4 (a og b), søger at svare på hypotese 3, og designet bag model 4, er at anvende den alternative fremgangsmåde til besvarelse af problemformuleringen. Dette afviger fra, hvad Soliman (2008) og Fairfield og Yohn (2001) gør, men er samtidig en interessent måde at undersøge problemformuleringen på. Ydermere, er model 4 den første model i dette projekt, som anvender den alternative fremgangsmåde. Model 4a ser således ud:

$$PM_{t+1} = \alpha_0 + \beta_1 * PM_t$$

Model 4b er derfor følgende:

$$ATO_{t+1} = \alpha_0 + \beta_1 * ATO_t$$

Estimaterne fra model 4a og 4b anvendes derefter i følgende model:

$$RNOA_{t+1} = PM_{t+1} * ATO_{t+1}$$

Produktet af model 4a og 4b giver altså resultatet $RNOA_{t+1}$, hvilket dernæst kan anvendes i udregningen af model 4's præcisionsmål.

4.2.5 Model 5a, 5b og 5c

Den samme fremgangsmåde, som bliver brugt til model 4a og 4b, vil blive anvendt i dette afsnit til at fremlægge model 5a, 5b og 5c. De følgende modeller er simple lineære regressioner, som fokuserer på at teste den statistiske præcision af forecasts af en to-lags DuPont dekomponering. Tre variabler skal anvendes, for at teste hypotese 4, og de er GP, OPPL og NOA. Først er der model 5a, som ser ud som følger:

$$OPPL_{t+1} = \alpha_0 + \beta_1 * OPPL_t$$

Model 5b er derfor således:

$$NOA_{t+1} = \alpha_0 + \beta_1 * NOA_t$$

Hvilket betyder at model 5c er:

$$GP_{t+1} = \alpha_0 + \beta_1 * GP_t$$

Disse tre modeller giver resultater, som vil blive brugt til følgende udregning:

$$RNOA_{t+1} = \frac{OPPL_{t+1}}{GP_{t+1}} * \frac{GP_{t+1}}{NOA_{t+1}}$$

Model 5a, 5b og 5c, vil blive benyttet og vurderes på en lignende måde, som model 4a og 4b, da den overordnede fremgangsmåde er ens. Grundtanken bag model 5 er altså at dekomponere i to lag, for at udregne en fremtidig værdi af RNOA. Her er det unødvendigt at have GP med som model, da RNOA kan udregnes på baggrund af OPPL og NOA. Derfor er udregningen, hvor der bliver omregnet til PM og ATO irrelevant. Dette betyder, at model 5c bliver udelukket, da modellen er overflødig.

Modellerne gennemgået i modeldesignet vil give et nuanceret og grundigt indblik i den teoretiske problemstilling, og derfor give resultater, som vil kunne besvare problemformuleringen, på en grundig og videnskabelig måde.

4.3 Introduktion for præcisionsmål

4.3.1 sMAPE

MAPE er en af de mest almindelige præcisionsmål anvendt i forecasting literaturen, og er en forkortelse af ”Mean Absolute Percentage Error” (Bruun, 2016). Følgende funktion anvendes til at udregne APE, som dernæst anvendes til MAPE:

$$APE_i = \left|\frac{A_i - F_i}{A_i}\right|$$

Her er A_i notation for den aktuelle værdi (Bruun, 2016), for dette projekt er det de aktuelle værdier af RNOA, og F_i, er den estimeret værdi en model har for RNOA. Derudover er i notation for en given observation. Bemærk her, at for både APE og derfor også MAPE er det baseret på absolutte værdier. MAPE er middelværdien af de summerede APE-værdier, ganget med 100 for at omregne til procent (Bruun, 2016), hvilket får MAPE til at se således ud:

$$MAPE_i = \frac{1}{n} * \sum APE_i * 100$$

På trods af dens popularitet i literaturen, har MAPE en del problemer tilknyttet den, som gør MAPE mindre præcis, når den aktuelle værdi af den afhængige variable er tæt på nul (Chen & Yang, 2004). Ifølge Makridakis (1993) er MAPE også plaget af, at MAPE ikke kan anvendes til sammenligning af naive modeller og random walks. Dertil har Makridakis og Hibon (2000) fremvist en modificeret version af MAPE, kaldet sMAPE, hvilket står for "Symmetric Mean Absolute Percentage Error". Den modificeret version af MAPE, udregnes på følgende måde (Makridakis & Hibon, 2000):

$$sMAPE_i = \frac{1}{n} * \sum \frac{|A_i - F_i|}{(|A|_i + |F_i|)/2} * 100$$

Forskellen mellem MAPE og sMAPE er ændringen i måden APE bliver udregnet på. For sMAPE er det udelukket i nævneren, hvor der tillægges modellens estimat til den aktuelle værdi. Derefter tages den nye værdi i nævneren og divideres med 2. Ved brug af sMAPE er det ifølge Makridakis og Hibon (2000), muligt at undgå store værdier for fejledende, da nævneren er blevet modificeret. Dette præcisionsmål forklarer, hvor meget modellen i gennemsnit vurdere forkert. Ydermere betyder dette at sMAPE bruges til at beskrive modellen der har de mindste fejlled i gennemsnit, hvilket ikke tager meget højde for enkelte store observationer. Modellen med den bedste værdi for sMAPE, kan derfor godt risikerer at have de mest ekstreme værdier, og hertil introduceres den subjektive del af diskussionen. Argumentet præsenteret her, er på baggrund af Makridakis (1993), og vil uddybes i diskussionen.

4.3.2 MSE

MSE står for "mean squared error" og selvom Bruun (2016) påstår at MAPE er en af de mest almindelige anvendte præcisionsmål, så påstår Diebold og Lopez (1996) at MSE er den klart mest anvendte præcisionsmål.

Dette betyder i realiteten ikke rigtigt noget, men det er en udemærke måde at beskrive pålideligheden for MSE. MSE udregnes på følgende måde:

$$MSE = \frac{1}{n} * \sum e_i^2$$

For den ovenstående formel er e en notation for fejlleddet, altså i dette tilfælde RNOA fratrukket modellens estimat for RNOA. Ydermere er n notation for antal observationer. MSE for en model skal være så lille en værdi, som muligt, for at have det bedst mulige resultat. MSE er et optimalt valg af præcisionsmål, når modellerne er baseret på normalfordelte fejlled, som er en af de vigtige antagelser for lineære regressioner (Chen & Yang, 2004). Derudover er MSE valgt, som præcisionsmål, på baggrund af dens hyppiged brug i litteratuen (Diebold & Lopez, 1996).

4.3.3 Interquartile range (IQR)

Interquartile range (herefter, IQR) er et præcisionsmål, der måler hvor stort intervallet er, hvor 50 % af dataet ligger (Whitley & Ball, 2001). Dette præcisionsmål håndterer ikke ekstreme værdier særlig godt, da de bliver ignoreret. Måden dette præcisionsmål udregnes på er som følger:

$$IQR = Q_3 - Q_1$$

Her er Q en notation for kvartil, hvor 1 og 3 repræsenterer kvartil 1 og kvartil 3. Kvartil 1 viser hvor grænsen for de første 25 % af dataet ligger, og repræsenterer derfor de laveste numeriske værdier i sættet. Samme tankegang gælder kvartil 3, hvor det her omhandler 75 % af datet, altså derfor udelukket de 25 % af dataet, hvilket er de største numeriske værdier. Trækkes kvartil 1 fra kvartil 3, fjernes "bunden" af dataet, og siden "toppen" ikke er inkluderet, så er IQR et estimat for de 50 % af dataet, som befinder sig i midten af datasættet. IQR anvendes i dette projekt, til at sammenligne de forskellige modelleres interval for, hvor langt intervallet er mellem grænserne for kvartil 1 og kvartil 3. IQR for de fem modeller vil sammenlignes med IQR for RNOA, hvilket hjælper med at illustrere forskellen på, spredningen af dataet for modellerne.

4.3.4 Justerede R^2-værdi

Den justerede R^2-værdi er modellens evne til at beskrive variationen af modellens afhængige variable. Da alle fem modeller skal sammenlignes, og modellerne er både simple og multiple lineære regressioner, er den justerede version af R^2-værdi anvendt. Dette præcisionsmål er valgt på baggrund af dens simple evne, til at beskrive, hvor god modellen er til at ramme residualerne. Ydermere vil der ikke være stor fokus på dette præcisionsmål, da det ikke nødvendigvis er ligeså dybdegående, som de andre præcisionsmål.

4.3.5 Beskrivende statistik for modellerne

For at give et helt standardiseret indblik i modellerne og deres estimater, vil der også være en hurtig gennemgang af modellernes beskrivende statistik. Hertil er det ikke noget nyt eller banebrydende, og begreberne er mean, median, varians, standard afvigelse, samt minimums- og maksimumsværdier.

4.3.6 Residual standard errors relative størrelse

Dette præcisionsmål er lavet på baggrund af dette projekt, og anvendes til at se den relative størrelse af modellernes residual standard error, i forhold til den afhængige variables middelværdi. Bemærk her, at dette præcisionsmål derfor ikke har en baggrund fra litteraturen, hvilket sætter spørgsmålstegn ved dens pålidelighed. Den relative størrelse af residual standard errors for modellerne vil derfor ikke have en reel vægt for konklusionen af den ”bedste” model, men den er taget med i projektet alligevel. Grunden til dette præcisionsmål er med, er for at se hvad residual standard error for en modellerne faktisk betyder. For eksempel så har model 1 en residual standard error på 0,4157 og for model 5a er den 140.300. Model 1’s afhængige variable er RNOA, og model 5a’s afhængige variable er OPPL, og det er derfor svært at sige ud fra deres residual standard errors, hvor stor en betydning disse begraber har. Præcisionsmålet bliver udregnet ved at tage residual standard error og dividere den med middelværdien for den afhængige variable. Ydermere giver dette præcisionmål en ide om, hvor stor en værdi den afhængige variable afviger i gennemsnit fra den sande værdi, i forhold til den afhængige variables middelværdi.

4.3.7 Residualer udenfor RNOA’s interval

Dette præcisionsmål er baseret på tabeller, som bliver fremvist i projektet. Bemærk her at for model 1, 2 og 4 tager dette præcisionsmål højde for residualer, som ligger udover intervallet for modellens estimerede værdier af RNOA. På grund af model 3 og 5’s ekstreme estimerede værdier, er den inverse version af dette præcisionmål anvendt, og vil tilsidesættes med de resterende modeller. Dette er på grund af, at for model 1, 2 og 4 er det den estimerede værdi, som er har det mindste interval, og for model 3 og 5, er det den afhængige variable der har det mindste interval. Igen vil dette præcisionsmål ikke have så stor vægt for konklusionen af projektet, men bliver stadig anvendt for at illustrere, hvor svært modellerne har ved at håndtere de ydrer værdier for dataet. Bemærk også for dette præcisionsmål, at der ikke er nogen decideret opbakning fra litteraturen. Dette betyder ikke at præcisionsmålet er ubrugeligt, men det skal bruges mere til at illustrere en given problematik, fremfor at være et definitivt mål for sammenligning af modeller.

4.3.8 RNOA plottet mod fejlled for de fem modeller

Det sidste præcisionsmål er en illustrering af fejlled, og skal ikke bruges, som et definitiv svar på en models evne til at forudsige RNOA. Dette præcisionsmål er derfor taget med, udelukket for at kunne visualisere, hvordan fejlleddene er, i forhold til RNOA. Hertil er det modellens fejlled på den ene akse af et plot, med RNOA på den anden akse.

I diskussionen, vil de overordnede præcisionsmål gennemgåes i forhold til modellerne fra analysen, som ses i modeldesignet. Fordelen ved at bruge så mange præcisionsmål, er at give en nuanceret konklusion på præcisionisen af modellerne, samt sikre at flere synspunkter bliver introduceret. Hertil vurderes de forskellige modeller, på baggrund af flere forskellige præcisionsmål, der kan fremhæve forskellige styrker ved modellerne. Dette er på baggrund af, at ofte anvendes kun en eller meget få præcisionsmål i en empirisk analyse (Chen & Yang, 2004), eller også ses der ikke på de forskellige synspunkter. Her kan et individ være overbevist om, at et præcisionsmål der måler de mindste gennemsnitslige errors er det bedste mål, hvor et andet individ kan have større vægt på mindre haler, altså færrest ekstreme værdier (Diebold, 1993). Her er det vigtigt at huske på, at der er ingen præcisionsmål, som kan kategoriseres som at være ”bedst” til at beskrive præcisionen af et forecast, men enkelte mål kan være gode til at beskrive forskellige synspunkter (Makridakis, 1993).

5. Analyse

5.1 Beskrivende statistik – De uafhængige variabler

I dette afsnit vil der blive fremlagt og gennemgået de seks uafhængige variablers beskrivende statistik. Hertil er der fokus på gennemsnittet (mean), median, variansen og standardafvigelsen. Formålet med at fremvise det beskrivende statistik for de uafhængige variabler, er at give et indblik i dem, før modellerne bliver gennemgået i de følgende afsnit. Ydermere anvendes nogle af disse begreber senere i projektet, mens de øvrige værdier samtidigt er tilgængelige for læseren, uden datasættet, samt koderne skal anvendes. Nedenstående i tabel 5, ses en oversigt over den beskrivende statistik, for de seks uafhængige variabler. Bemærk her, at tallene for LOPPL, LGP og LNOA er opgivet i tusinde, som datasættet. Ydermere er det opgivet for de laggede variabler, på grund af at det er de laggede variabler, som anvendes som uafhængige variabler i modellerne. Notationen for en lagged variable er et ”L” foran den uafhængige variabel, og dertil ses tabel 5:

	Mean	Median	Varians	Standardafvigelse
LRNOA	0,2851	0,1828	0,2108	0,4591
LPM	0,1843	0,1613	0,05	0,2237
LATO	1,8739	1,1764	6,5759	2,5644
LOPPL	11.526	833	21.266.960.229	145.832
LNOA	75.777	4.740	809.161.367.020	899.534
LGP	55.188	5.833	352.064.741.154	593.350,4

Tabel 5 - af egen inspiration.

Tabel 5 viser nogle voldsomme værdier for LOPPL, LNOA og LGP, hvor disse tre variabler, har utrolig høje varianser og standardafvigelser, i forhold til deres gennemsnit. Den gennemsnitlige spredning fra middelværdien for disse tre variabler er meget høje. Ydermere er de også langt højere end medianen, hvilket er et bekymrende syn. Disse værdier er så absurd høje, at man næsten skulle tro der var tale om en tastefejl, men det er ikke tilfældet. For at tage et dybere indblik i, hvorfor disse tal er så absurde, så vil der i følgende bilag 5 undersøges, om variablerne er normalfordelte.

bilag 5 viser, hvordan normalfordelingen ser ud for de forskellige variabler. Ydermere er det hurtigt synligt, hvad problemet kan være med datasættet. LRNOA, LPM og LATO ser relativt fornuftige ud, når det kommer til normalfordelingen, men der er problemer med LGP, LOPPL og LNOA. For LGP, LOPPL og LNOA, er normalfordelingen, hverken den ønskede ”klokkeform”, men også tungt vægtet i den ene side af fordelingen. Problemet her kan blandt andet løses ved at arbejde med den naturlige logaritme af disse variabler. Hertil er det vigtigt at bemærke, at tage den naturlige logaritme for disse variabler, ikke nødvendigvis er en løsning på problemet, da dette skaber et andet problem. På trods af dette, kan det stadig visualiseres, hvor stor en effekt på datasættet denne ændring kan have, men yderligere metoder skal anvendes. For at se effekten af denne ændring, henvendes der igen til bilag 5.

Det er tydeligt i bilag 5, at histogrammerne for LOPPL, LGP og LNOA ser mere fornuftige ud, efter de er blevet log transformeret. Hvad dette betyder for de følgende modeller, vil blive klargjort når det kommer til modellerne. Dette er i følgende afsnit, hvor modellerne fra modeldesignet bliver gennemgået, samt forsøgt forbedret i tilfælde der er statistisk grundlag for det. Disse afsnit vil følge den samme overordnede fremgangsmåde. Først vil modellen introduceres, og de statistiske parametre fra modellens output vil blive gennemgået. Dernæst vil de lineære antagelser blive gennemgået, for at se om modellen kan forbedres.

Grunden til en log transformation kan være problematisk, er at det er matematisk umuligt at tage den naturlige logaritme af et negativt tal og en nul-værdi, og derfor er det ikke altid hensigtsmæssigt at bruge på økonomiske variabler. For at løse denne matematiske problemstilling, kan alle variabler tillægges en konstant. Løsningen her flytter modellen kunstigt opad i forhold til y-aksen, som er den afhængige variable, men har derfor en effekt på interceptet (Ekwaru & Veugelers, 2018). Hertil er der ingen effekt på koefficienter, varians, t-statistikken osv. (Ekwaru & Veugelers, 2018).

5.2 Model 1: Benchmark

Model 1 er den omtalte benchmark-model, som selvfølgelig er den helt simple model, der udelukket baseres på baggrund af de historiske tal for RNOA. Baseret på modeldesignet, ser denne model ud, som følger:

$$RNOA_{t+1} = \alpha_0 + \beta_1 * RNOA_t$$

Princippet bag denne model, er at forudsige fremtidige værdier af RNOA, på baggrund af det nuværende niveau af RNOA. For at kunne udregne dette, bliver den uafhængige variable RNOA lagged en gang, så der kan konstrueres de nødvendige koefficienter. Nedenstående i illustration 1 ses outputtet for model 1:

```
Call:
lm(formula = RNOA ~ LRNOA)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9965 -0.1293 -0.0500  0.0801  5.2831

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.127358   0.002464   51.69   <2e-16 ***
LRNOA       0.479954   0.004559  105.28   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4157 on 39453 degrees of freedom
Multiple R-squared:  0.2193,    Adjusted R-squared:  0.2193
F-statistic: 1.108e+04 on 1 and 39453 DF,  p-value: < 2.2e-16
```

Illustration 1 - af egen inspiration.

Bemærk her, at det er lidt atypisk, at outputtet bliver fremvist på denne måde, hvor det bliver taget direkte fra statistikprogrammet, og lagt ind i projektet. Her er det mest normalt, når man kigger på litteraturen, at en tabel bliver fremstillet for at få det til at se bedre ud. Her har dette projekt valgt at give outputtet råt til læseren, for at alle er på samme side, så selv værdier i outputtet, som dette projekt ikke fremhæver, er tilgængelig, uden læserne skal ind og lave modellen selv. Alternativt kunne dette projekt, selv lave en tabel der fremhæver

værdierne fra outputtet, men så bliver der enten indsat malplaceret værdier der ikke bliver brugt, eller også kan læseren føle de mangler dele af analysen.

Model 1 kan nu forudsige niveauet for RNOA året efter en given værdi, men ikke specielt præcist. Ved hjælp af det givne output i illustration 1, så bliver model 1 lavet om til følgende ligning (afrundet til 4 decimaler):

$$RNOA_{t+1} = 0{,}1274 + 0{,}48 * RNOA_t$$

For at klarificere dette tages, for eksempel, den gennemsnitlige værdi af hele sættet for RNOA, fra tabel 5, og antages at dette er niveaut for RNOA i år 2010 for en given virksomhed, så kan der udregnes RNOA i år 2011 til at være:

$$RNOA_{2011} = 0{,}1274 + 0{,}48 * 0{,}2851 = 0{,}2642$$

For at fastsætte, hvor statistisk præcis denne model er, så den kan sætte et benchmark for resten af dette projekt, henvendes der til diskussionen. Dette bliver gjort gennem de tidligere nævnte præcisionsmål, men med det samme ses det på outputtet i illustration 1, at denne model har et "$Adjusted\ R-squared$" (herefter justerede R^2-værdi) på 0,2193, altså 21,93 %. Model 1 beskriver derfor 21,93 % af den afhængige variables variation.

Da model 1 er et benchmark for de andre modeller betyder dette, at dens evne til at forklarer variation i $RNOA_{t+1}$ ikke umiddelbart fortæller så meget relevant. Ydermere så er det også vigtigt at fokusere på, at problemformuleringen søger at finde en model der forbedrer på resultatet fra model 1.

Et logisk næste skridt i denne analyse, er at kigge på p-værdien for model 1. Til dette opstilles der nul-hypoteserne, der påstår at koefficienterne for interceptet og β_1 er lig med nul. Hvis det viser sig, at nul-hypoteserne ikke kan forkastes, så er koefficienterne ikke forskellig for nul, og variablen er derfor ikke signifikant. Til at teste dette, vil der blive brugt et signifikansniveau på 5 %, hvilket betyder at den tilknyttet p-værdi, skal være lavere end 0,05. Ydermere er det i denne del af analysen, der bliver undersøgt Fairfield og Yohn (2001) påstande om variable signifikans, det er dog til senere modeller, hvor PM og ATO bliver introduceret til modellerne. For illustration 1 er p-værdierne beskrevet i kolonnen markeret ”Pr(>|t|)”, derudover så ses en signifikanskode til højrer for værdien, samt den tilhørende kodes betydning, under kolonnen. Her ses at p-værdierne for interceptet og β_1 er godt under 0,05, hvilket betyder, at nul-hypoteserne forkastes, og koefficienterne er forskellige fra nul.

Dernæst bemærkes det, at residualernes standard error er på 0,4157, hvilket betyder at resultaterne i model 1 i gennemsnit kan fravige fra den sande lineære regression med 0,4157. Dette er en meget høj grad af fejl i model 1, og hænger sammen med, hvor dårligt modellen er til at beskrive residualerne. For at sætte dette i perspektiv, så sammenlignes residual standard error med den gennemsnitlige værdi på 0,2642 (bemærk her, at denne middelværdi er for RNOA, og ikke LRNOA). Dette betyder, at model 1 kan i gennemsnit fravige fra den sande lineære regression med 1,5734 gange den gennemsnitlige værdi af RNOA (se følgende udregning):

$$\frac{0{,}4157}{0{,}2642} = 1{,}5734$$

Før der med sikkerhed kan anvendes denne model til udregning af præcisionsmålene i diskussionen, skal der sikres at de lineære antagelser er opfyldt. Antagelserne gennemgås til sidst i hvert afsnit for modellerne, for blandt andet at vise, hvordan en mulig ændring, som disse antagelser bringer, kan forbedre modellen. Ydermere kan nogle af antagelserne først undersøges, efter modellen er sat op. For multipel lineære regressioner er dette multikollinearitet, og derudover kan det også være testen om lineæritet, samt homoskedaskicitet for både de simple og multiple modeller.

Model 1 bliver nu undersøgt, i forhold til antagelserne, som er gennemgået tidligere i projektet. Da der er tale om en simpel linear regression, vil der blive testet for lineæritet, homoskedasticitet og om LRNOA er normalfordelt. I Bilag 5 ses det, hvordan variablen LRNOA er normalfordelt, med en stor vægt af fejlleddene mean omkring nul, hvilket giver en smal og høj ”klokkeform”. Dernæst vil der testes for homoskedasticitet og lineæritet i samme illustration, som ses nedenfor:

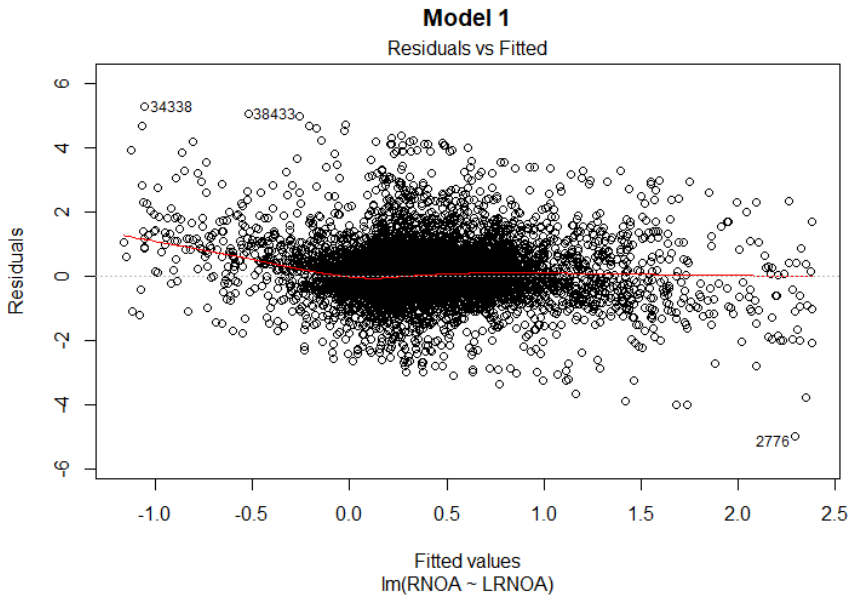


Illustration 2 - af egen inspiration.

Illustration 2 viser de fitted værdier for modellen, beskrevet på residualerne. Det vigtige ved illustration 2, er først og fremmest afstanden på residualerne i forhold til hinanden. Her skal det bemærkes at residualerne har en konstant afstand rundt om ”0-linjen”. Dette antyder at der er homoskedasticitet i sættet, specielt på grund af dataen ikke antager de almindelige tegn på hetroskedasticitet, som enten er formet, som en tragt eller en butterfly (Osborne & Waters, 2002). Dette efterlader model 1, kun med antagelsen om linearitet tilbage. Illustration 2 viser en rød linje, der stort set følger en lineær model. Model 1 ses derfor, på baggrund af illustration 2, at opfylde antagelsen om linearitet. Dette betyder at de tidligere hypotestests er pålidelige, og der ikke behøves mere arbejde, før model 1 kan anvendes til præcisionsmålene.

For at få et indblik i, hvor god model 1 er til at forudsige niveauet af fremtidige RNOA, så vil dette blive illustreret i følgende illustrationer, samt interval:

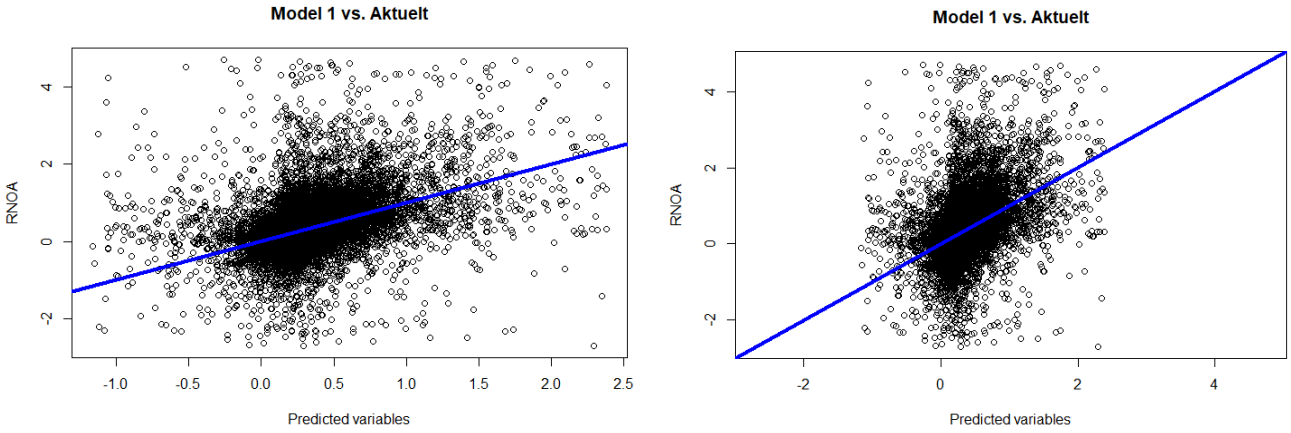


Illustration 3 - af egen inspiration.

Illustration 3 er opdelt i to plots, som begge er af model 1, hvor den eneste forskel er længden på x-aksen. For de to plots i illustration 3, er det resultatet for model 1 der er vist på x-aksen, og den aktuelle værdi af RNOA er vist på y-aksen. Ydermere ses en blå linje i begge plots, hvor linjen er baseret på samme funktion, hvilket er et skæringspunkt på nul og en hældning på en. Formålet med illustration 3, er altså at vise, hvor godt model 1 kan forudsige de aktuelle værdier for RNOA, her ville den perfekte model ligne den blå linje. Den perfekte model vil ramme den blå linje da dette betyder, at når model 1 forudsige en værdi, vil den matche præcist på den aktuelle værdi af RNOA, for alle observationer. Hertil ville en række nye problemer forekomme, da der godt kan være tale om en model, der er "for perfekt", da dette ikke ses, som at være realistisk at nå. Til venstre i illustration 3, ses et plot over alle residualerne, hvor x-aksen er tilpasset omkring den minimale og maksimale værdi, som model 1 estimerer. Til højre i illustration 3, ses et plot, hvor x-aksen og y-aksen matcher hinanden, hvilket giver den blå linje en 45 graders hældning. Ydermere er y- og x-aksens længder nu afhængige af minimums- og maksimumsværdien for RNOA. De to overstående plots er begge i illustration 3, for ikke at skabe forvirring omkring modellens evne til at forudsige RNOA, og derfor for at give et komplet indblik i modellen. Fremgangsmåden brugt til at illustrere model 1 på, vil blive brugt på alle modeller, gennem hele projektet. Billedet illustration 3 tegner af model 1, er tydeligt efter begge plots analyseres. Model 1 har svært ved at vurdere de mindste og de største værdier, da begge aksers længde er baseret på spændet for RNOA. Det er vigtigt at understrege her, at havde model 1 beskrevet RNOA perfekt, ville dens tendenslinje ligge præcis på den blå linje, men residualerne vil ikke nødvendigvis dække hele linjen. Med dette menes, at selvom der for illustration 3 er en mangel af residualer på de mindste og største værdier, så er det ikke til at sige ud fra

modellen, hvor mange residualer der ikke bliver beskrevet ordentligt. Da dette er lidt besværligt at forklare, vil det uddybes med følgende tabel:

	Estimeret værdier for model 1	RNOA	RNOA's ekstreme værdier udenfor model 1's interval
Minimumsværdier	-1,1629	-2,7021	113
Maksimumsværdier	2,3838	4,7011	318

Tabel 6 - af egen inspiration

Tabel 6 viser minimums- og maksimumsværdier for både de værdier model 1 forudsiger, samt den aktuelle værdi af RNOA. Her ses det, af for et datasæt på 39.455 observationer, er der 431 residualer, i model 1's ender, som ikke bliver beskrevet. Tabel 6, sammen med illustration 3, giver et godt indblik i, hvordan forholdet er mellem model 1's estimater, og de aktuelle værdier af RNOA. Selvom illustration 3 giver indblik i, hvor residualerne burde være, så er der ingen indblik i, hvor stor en vægt af residualerne, skal placeres i enderne, hvor model 1 har problemer med ens estimater. I forlængelse af problemstillingen, som tabel 6 har kastet lys over, så er det også besværligt at vurdere ud fra illustration 3, hvor mange observationer der ses i "midten". Illustration 3 viser en stor mængde residualer midt i de to plots, hvilket gør det svært at tolke på de øvrige residualer i plottet, da den reelle vægt af hver residual, bliver undermineret. Til at hjælpe denne visualisering, vil der nedenstående ses en ny version af de to plots i illustration 3, hvor der anvendes et mindre datasæt, med tilfældigt udvalgte observationer.

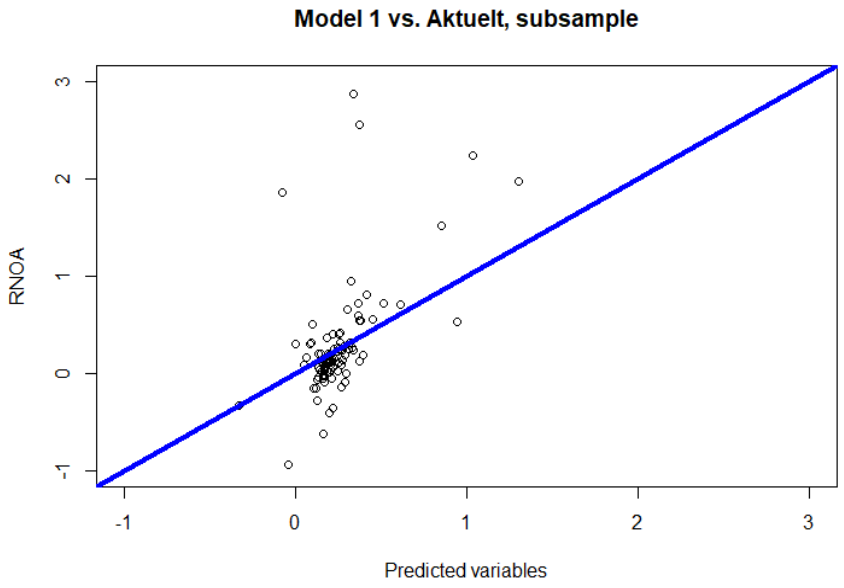


Illustration 4 - af egen inspiration.

Illustration 4 er baseret på 100 observationer, tilfældigt udvalgt ved hjælp af RStudio, hvor der er en ny minimumsværdi for RNOA på -0,9375 og maksimumsværdien på 2,8737. Hertil er der kun valgt at vises plottet der har ens længde for akserne, da den flot beskriver det, som er nødvendigt. Ydermere fravælges det andet plot også, så det illustrerede plot stås alene, og derfor kan forstørres lidt, dog ses det andet plot i bilag 6. Angående bilag 6, er der også en decil tabel, samt en visualisering af tætheden for residualerne. Selvom illustration 4 kun forstørres lidt, hjælper det en del med hvad plottet skal vise. Bemærk for illustration 4, at modellen stadig ikke er alt for god til at beskrive ekstreme værdier, da der for eksempel, ligger residualer højt på y-aksen, men midt på x-aksen. Herfra antages det at de 100 observationer anvendt i denne subsample skaber et repræsentativt billede af datasættet, og derfor kan anvendes til at hjælpe med at uddybe på illustration 3. Problemet med illustration 3, var besværet ved at vurdere den reelle vægt af observationerne, som var samlet i midten af plottene. Dette problem bliver løst ved hjælp af illustration 4, hvor et residual, kan repræsentere en procent af den anvendte subsample. Hertil ses det, at det er en betydelig lille del af residualerne, som rammer langt fra den blå linje, og størstedelen af observationerne er relativt tæt på hinanden.

For at få et indblik i, hvor kompakt dataet er, så ses følgende tabel over de første to og sidste to decilerne for RNOA:

Deciler	0 %	10 %	90 %	100 %
RNOA	-2,7021	-0,0328	0,6801	4,7011

Tabel 7 - af egen inspiration.

Tabel 7 viser, at 80 % af observationerne for RNOA ligger i intervallet $[-0{,}0328; 0{,}6801]$, hvilket er med til at understrege, hvor stor en vægt af RNOA, er på et relativt beskedent interval, mellem minimums- og maksimumsværdien. For at illustrere dette yderligere, og for at spare plads, er der i bilag 6 konstrueret, et komplet resumé af decilerne for RNOA, samt en illustration der markere og understreger dette interval. Dette interval vises mellem to røde linjer, hvor 80 % af residualerne ligger. Bilag 6 bekræfter derfor det tidligere omtalte problem, som illustration 4 hjælper med at løse. Fremgangsmåden med at vise modellernes estimerede værdier plottet mod deres afhængige variable, vil gennemgås for alle modeller. Ydermere vil der ikke blive gået lige så meget i dybden for de øvrige modeller, hvori dette betyder, at selve problemstillingen og tankerne bag valget af illustrationerne til at beskrive disse, ikke drøftes igen. Bemærk at, "Tæthedsplottet" der er illustreret i bilag 6, kan anvendes, som referencemateriale for øvrige modeller. Det er ikke selve "Tæthedsplottet" i bilag 6 der anvendes til dette, men konklusionen bag den, samt tabel 7. For at understrege denne pointe, så er det argumentet der omhandler, hvor stor en del af den afhængige variable for visualiseringen af disse plots, der er vægtet i intervallet $[-0{,}0328; 0{,}6801]$.

Sidste skridt for undersøgelse af model 1, og de fire modeller efterfølgende, er præcisionsmål, som kan bruges til at vurdere modellen. Ydermere er disse præcisionsmål vigtige til at kunne sammenlignes med andre modeller, for at kunne komme frem til et definitivt svar på, hvilke modeller er bedre end andre. Disse præcisionsmål vil dog ikke blive præsenteret i analysen, men først i diskussionen i tabel 19 og 20, hvor de reelt skal bruges. Dette gælder selvfølgelig for alle fem modeller.

5.3 Model 2: Simple dekomponering af RNOA

Efter benchmark-modellen er sat på plads, kan dette projekt fortsætte med model 2, som er den simple dekomponering af RNOA. Model 2 blev introduceret i modeldesignet, og er som følger:

$$RNOA_{t+1} = \alpha_0 + \beta_1 * LPM_t + \beta_2 * LATO_t$$

Model 2 har til formål at måle det fremtidige niveau af RNOA, hvor de uafhængige variabler er PM og ATO. Ydermere er model 2 også essensen i dette projekts problemformulering, i det den er den første model der tester om det er muligt at forbedre på model 1, gennem den mest simple form for dekomponering. Ydermere hjælper model 2 også med at kaste lys på påstande fra Fairfield og Yohn (2001), samt Soliman (2008).

Nedenstående ses illustration 5, der viser model 2 og dens output, hertil bruges den laggede variable af PM og ATO, som de uafhængige variable:

```
Call:
lm(formula = RNOA ~ LPM + LATO, data = Datasaet)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6125 -0.1442 -0.0503  0.0884  5.2946

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0537371  0.0033340   16.12   <2e-16 ***
LPM         0.4961861  0.0097700   50.79   <2e-16 ***
LATO        0.0635300  0.0008521   74.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4317 on 39452 degrees of freedom
Multiple R-squared:  0.1584,    Adjusted R-squared:  0.1584
F-statistic:  3713 on 2 and 39452 DF,  p-value: < 2.2e-16
```

Illustration 5 - af egen inspiration.

Model 2 beskriver det fremtidige niveau af RNOA, givet nogle koefficienter baseret på LPM og LATO, som vises i illustration 5. Ligningen for model 2 ser derfor således ud:

$$RNOA_{t+1} = 0{,}0537 + 0{,}4962 * LPM_t + 0{,}0635 * LATO_t$$

Den justerede R^2-værdi for model 2 er på 0,1584, hvilket svarer til 15,84 %, som ikke nødvendigvis er imponerende. Ydermere bemærkes det, at denne værdi er mindre end den af model 1.

Opstilles der en nul-hypotese til model 2, om variablerne er forskellige for nul, kan det ses, om alle modellens koefficienter er signifikante eller ej. Først ses det at p-værdierne for model 2, er ens med værdierne fra model 1, dette giver et indtryk om, at statistikprogrammet RStudio, ikke rapportere på disse utrolige lave værdier. Her bemærkes det også, at det, som bliver fremvist i outputtet for illustration 5, er symboliseret med ”<”, hvilket betyder at p-værdierne er så lave, at de er godt under signifikansniveauet på 0,05. Nul-hypotesen for model 2’s variabler forkastes og den alternative hypotese gælder. Ydermere er dette et tegn på at alle koefficienter for model 2 i hvert fald er signifikante på et niveau der mindst hedder 95 %. Dette var ved første blik lidt overraskende ifølge den litteratur, som bliver gennemgået i teoriafsnittet. Overraskelsen her er at, Fairfield og Yohn (2001), finder frem til at PM ikke er signifikant for forudsigelsen af fremtidige ændringer i RNOA. Her er der dog en fundamental forskel på litteraturen, og hvad model 2 vælger at undersøge. Model 2 undersøger niveaut af RNOA, hvor litteraturen kigger på ændringen i RNOA.

Illustration 5 viser, at model 2 har en residual standard error på 0,4317, hvilket igen er en høj grad af fejl i model 2. Dette betyder, at i gennemsnit, så kan model 2 fravige med 0,4317 for den sande lineære regression. Model 2 har ikke RNOA, som en uafhængig variable, RNOA er dog målet for denne model, og dette betyder samme udregning, som blev gjort ved model 1, kan anvendes her. Denne udregning er forholdet mellem gennemsnitsværdien af RNOA og den residuale standard error for model 2.

$$\frac{0{,}4317}{0{,}2642} = 1{,}634$$

Her fraviger model 2 i gennemsnit fra den sande linære regression, med en værdi der er 1,634 gange så stor, som den gennemsnitlige værdi af RNOA. Dernæst undersøges de lineære antagelser for model 2, for at se om de tidligere hypotesetests holder, og om modellen skal behandles. Model 2 har det store til forskel med model 1, at model 2 er en multipel lineær regression. Dette betyder at der nu bliver testet for normalfordeling, homoskedasticitet, linearitet, samt multikollinearitet. Hetil bliver antagelsen omkring multikollinearitet undersøgt først.

For at teste for multikollinearitet vil der tjekkes for korrelatione mellem de uafhængige variabler. Dette er for at se, om de uafhængige variabler LPM og LATO er forbundet med en lineær relation. Korrelationen mellem LPM og LATO er ifølge RStudio på -0,0864, hvilket betyder der er en meget lille negativ korrelation. Ydermere betyder dette, at når den ene værdi vokser, falder den anden. Da denne værdi er så lav, så betyder dette også, at variablerne stort set ikke følger hinanden lineært. For at sikrer signifikansen af denne korrelation, vil der testes for en p-værdien af korrelationstesten. P-værdien er oplyst i RStudio til igen at være den forsvindende lave værdi af ”<2,2e-16”, hvilket oplyser, at korrelationen uden tvivl bruges med 95 % sikkerhed. Model 3, som gennemgås i næste afsnit, vil også under multikollinearitet gennemgå en VIF-test for de uafhængie variabler. For sammenlignlighedens skyld, så vil model 2 derfor også fremvise en VIF-test, hvilket ser ud, som følger:

	LPM	LATO
VIF-scorer	1,0111	1,0111

Tabel 8 - af egen inspiration.

Tabel 8 viser en VIF-scorer på 1,0111 for begge variabler. En VIF-scorer under 4 (Pardoe, 2019), betyder at variablen ikke er ramt af multikollinearitet, og tabel 8 understreger derfor, hvad korrelationstesten viste for de uafhængige variabler i model 2. De næste tre skridt for model 2, er at teste for linearitet, homoskedasticitet og

normalfordelingen. Til dette vil der i RStudio blive konstrueret en lignende illustration for model 2, som der blev gjort for model 1. nedenstående ses denne illustration:

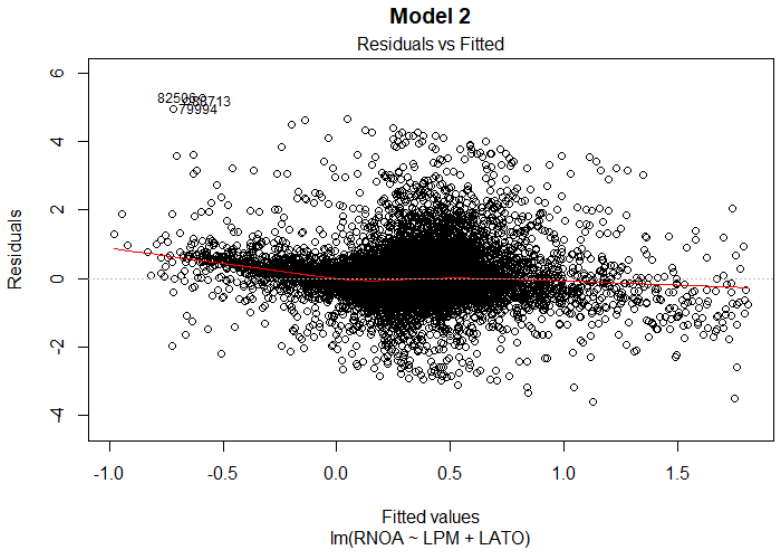


Illustration 6 - af egen inspiration.

Illustration 6 viser en lignende model, som illustration 2. Residualerne i illustration 6 har en lignende afstand fra ”0-linjen”, hvilket betyder der er konstant varians for model 2. Ydermere betyder dette, at model 2 kan antages, som at være homoskedasticitet. Dernæst kan det ses i bilag 5, at der igen er tale om et fejlleddene er normalfordelte. Dette er på baggrund af, at bilag 5 for LPM og LATO, har en ”klokkeform”, som både vises på normalfordelingens form, men også på grund af histogrammets søjler. Bilag 5 viser derfor, at både LPM og LATO har normalfordelte fejlled, omkring en mean på nul. Den sidste antagelse, som mangler at gennemgås for model 2, er antagelsen om linearitet. Hertil henvendes der til illustration 6, hvor den røde linje ses til at have en lineær form, samt den er relativ flad omkring ”0-linjen”. Med antagelsen om linearitet, homoskedasticitet, normalfordelingen og multikollinearitet gennemgået, kan det nu antages, at model 2 er pålidelig. Dette betyder, at model 2 kan gennemfører hypotesetests og kan derfor sammenlignes på baggrund af de kommende præcisionsmål.

Nedenstående ses illustration 7, hvilket er de to plots for de værdier model 2 estimerer, forklaret ved hjælp af RNOA’s aktuelle værdi:

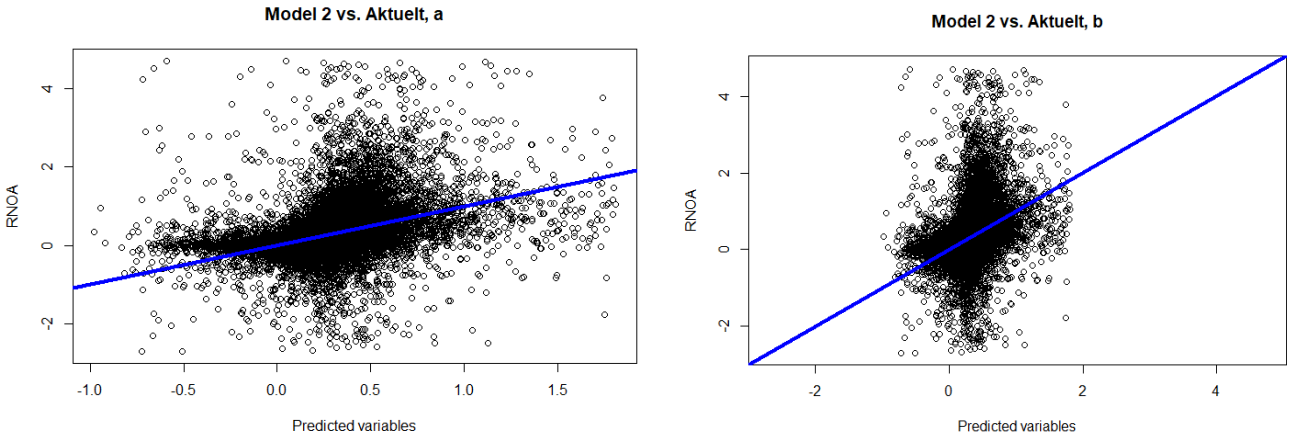


Illustration 7 - af egen inspiration.

Illustration 7 viser et lignende billede for model 2, som illustration 3 viser for model 1. Model 2 har svært ved at beskrive de værdier, som enten er relativt høje eller lave. For at give et bedre billede af, hvilket interval residualerne for illustration 7 ligger, så ses følgende tabel:

	Estimeret værdier for model 2	RNOA	RNOA's ekstreme værdier udenfor model 2's interval
Minimumsværdi	-0,9785	-2,7021	288
Maksimumsværdi	1,8093	4,7011	577

Tabel 9 - af egen inspiration.

Tabel 9 viser at for de estimerede værdier for model 2, så er der 288 observationer, som ligger under den nedre grænse, og 577 over den øvre grænse. Dette betyder at for model 2 er der 865 observationer, som bliver undervurderet, og derfor er udenfor models estimeret interval. Igen er dette en mindre del af datasættet på 39.455 observationer. Hertil burde disse observationer, ligge i tomrummene illustreret i illustration 7 version b (til højre). Illustration 7 har samme fundamentale problem, som illustration 3, hvor det er umuligt ud fra de to plots at sige, hvor stor en vægt der ligger omkring midten af sættet. Hertil ses nedenstående illustration, hvor der er blevet lavet en visualisering på et datasæt af 100 observationer:

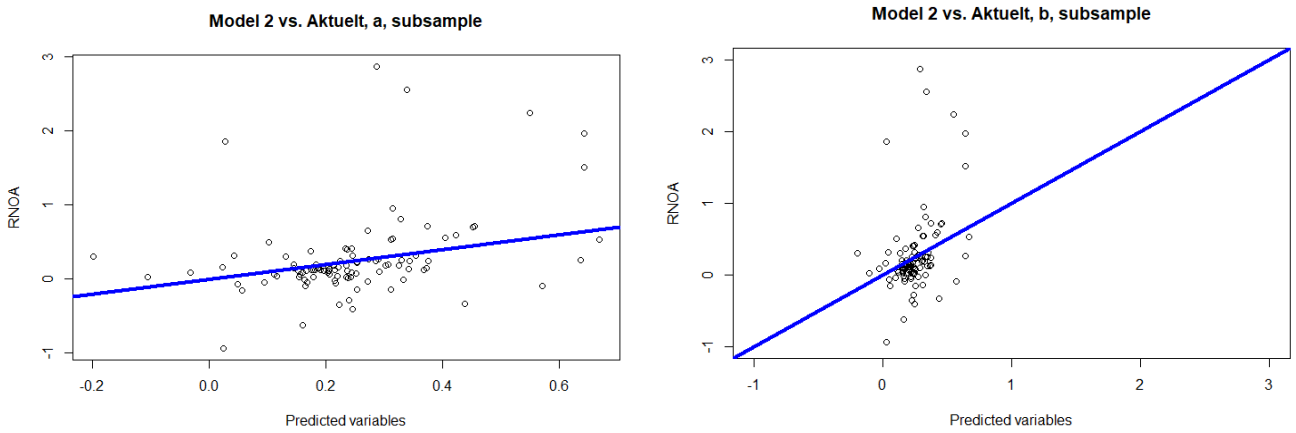


Illustration 8 - af egen inspiration.

For illustration 8 er der anvendt det præcis samme data, som er anvendt for de lignende plots for model 1. Illustration 8 består af version a (venstre) og b (højre). Det er ikke nødvendigvis det vigtigste at de 100 tilfældige udvalgte observationer er ens, men mere at det samme antal af tilfældigt udvalgte observationer er udvalgt. Denne påstand er baseret på, at de 100 tilfældigt udvalgte observationer, er en pålidelig repræsentation af datasættet, ligesom et andet tilfældigt sæt af 100 observationer, vil være repræsentativt. Bemærk derudover at illustration 8 har begge versioner af plottet, da versionen med en kortere x-akse kan give et bedre indblik i modellen. Version a af illustration 8 viser, hvordan residualerne er spredt rundt om den blå linje, med få værdier der fraviger dette. Derudover giver version a, også et indblik i, hvor stor en vægt af residualerne, som er placeret tæt på linjen, da de er spredt ud, og de lettere individuelt visualiseres.

Det gælder alle modeller, at sammenligningen mellem modellerne, sker i diskussionen, og der vil derfor ikke kommenteres her, om model 2 er en forbedring på model 1.

5.4 Model 3: To-lags dekomponering af RNOA

I dette afsnit analyseres model 3, der undersøger et fremtidigt niveau af RNOA, baseret på en to-lags DuPont dekomponering af RNOA. Model 3 fra modeldesignet ser ud, som følger:

$$RNOA_{t+1} = \alpha_0 + \beta_1 * OPPL_t + \beta_2 * NOA_t + \beta_3 * GP_t$$

Model 3 har altså de tre uafhængige variabler OPPL, NOA og GP. Outputtet for model 3 ses i illustration 9:

```
Call:
lm(formula = RNOA ~ LOPPL + LNOA + LGP, data = Datasaet)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9667 -0.1990 -0.0932  0.0990  4.4363

Coefficients:
                  Estimate      Std. Error t value Pr(>|t|)
(Intercept)  0.264784307506  0.002376575759 111.414  < 2e-16 ***
LOPPL        0.000000250250  0.000000031855   7.856 4.07e-15 ***
LNOA        -0.000000038970  0.000000005537  -7.038 1.99e-12 ***
LGP         -0.000000009190  0.000000008673  -1.060    0.289
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.47 on 39451 degrees of freedom
Multiple R-squared:  0.002169,  Adjusted R-squared:  0.002093
F-statistic: 28.58 on 3 and 39451 DF,  p-value: < 2.2e-16
```

Illustration 9 - af egen inspiration.

Illustration 9 viser en alarmerende justerede R^2-værdi, koefficienter for modellen, samt p-værdi for LGP. Først og fremmest, er den justerede R^2-værdi for model 3 på 0,0021, hvilket svarer til 0,21 %. Dette betyder, at model 3 i gennemsnit beskriver 0,21 % af variationen af den afhængige variable. Ydermere er koefficienternes værdier utrolig lave, hvilket skyldes tallene for LOPPL, LGP og LNOA er relativt høje, i forhold til RNOA. Det kan være grundet der bruges absolutte tal, til at beskrive satsen RNOA. P-værdien for hele modellen er stærkt signifikantniveau på 95 %. Koefficienternes p-værdier er alle signifikante på et niveau af 95 %, udover LGP, hvilket er godt over de acceptable 0,05, med værdien 0,289. Den høje p-værdi for LGP, betyder at LGP ikke er signifikant for model 3, og derfor forsøges der med en ny og modificeret model 3, hvor LGP er taget ud af ligningen. Illustration 10 viser nedenfor, outputtet for den nye modificeret model 3:

```
Call:
lm(formula = RNOA ~ LOPPL + LNOA, data = Datasaet)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9667 -0.1989 -0.0932  0.0990  4.4364

Coefficients:
                  Estimate      Std. Error t value Pr(>|t|)
(Intercept)  0.264692623782  0.002375003924 111.449   <2e-16 ***
LOPPL        0.000000235332  0.000000028575   8.235   <2e-16 ***
LNOA        -0.000000042184  0.000000004633  -9.106   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.47 on 39452 degrees of freedom
Multiple R-squared:  0.00214,   Adjusted R-squared:  0.00209
F-statistic: 42.31 on 2 and 39452 DF,  p-value: < 2.2e-16
```

Illustration 10 - af egen inspiration.

For begge versioner af model 3 rapporteres residual standard error, til at være på 0,47. Derudover ses koefficienterne også i de to modeller, til at være ændret i et meget begrænset niveau. Den justerede R^2-værdi, ændres også i en meget begrænset version.

Der kan være flere forskellige grunde til dette resultat, men en af dem kan være multikollinearitet i sættet, hvilket kan testes ved at undersøge korrelationen mellem de forskellige uafhængige variabler. Nedenfor ses tabel 10 der viser de forskellige værdier for korrelationen mellem variablerne.

Korrelationsmatrix	OPPL	GP	NOA
OPPL	1	0,8354	0,8231
GP	0,8354	1	0,8586
NOA	0,8231	0,8586	1

Tabel 10 - af egen inspiration.

Tabel 10 viser, at alle de uafhængige variabler har en stærk positiv korrelation med hinanden. Dette er et problem for modellen, da de forskellige variabler har en stærk lineær sammenhæng med hinanden.

Uden at på dette tidspunkt gå for meget ind i dette projekts diskussion, er det ingen hemmelighed, at model 3, både den ”rå” version, samt den modificerede, er forfærdelige. En af de tre variabler er vurderet til at være stærkt insignifikant, hvis man kigger på dens p-værdi. Derudover er det en overordnede model, som har en utrolig dårlig evne til at beskrive den afhængige variables varians. Det skal derfor ikke være nogen hemmelighed, at model 3 er stort set ubrugelig i dens nuværende form. En problematisk trend har derfor vist sig i analysen indtil nu, hvor der ikke bliver forbedret på benchmarket, og jo mere RNOA bliver dekomponeret, jo mere upræcist bliver resultatet. Til sidst er det også tydeligt at model 3 har nogle problemer med multikollinearitet, siden de uafhængige variabler er stærkt korreleret med hinanden. En yderligere undersøgelse af model 3, er udført, ved at kigge på de uafhængige variablers VIF-scorer. Følgende tabel viser resultatet af dette:

	GP	NOA	OPPL
VIF-scorer	5,2979	3,4452	3,5152

Tabel 11 - af egen inspiration.

Tabel 11 bekræfter påstanden om multikollinearitet i model 3, ved at fremhæve VIF-scoren for de forskellige variabler. Ydermere viser VIF-scoren, hvor meget en koefficients værdi er blevet ændret, på grund af

multikollinearitet. Udgangspunktet for en VIF undersøgelse, er at en værdi over 4 (Pardoe, 2019), betyder at variablen er ramt af multikollinearitet, og den skal derfor kasseres, hvilket også bekræfter tidligere signifikansniveau fra p-værdien. Hertil ses der også forskellen på VIF-scoren mellem model 2 og model 3, hvor tabel 11 hjælper med at understrege, at model 2 ikke er berørt af multikollinearitet. Konklusionen på dette ville umiddelbart være at fjerne den uafhængige variable LGP, og derfor er antagelsen om multikollinearitet taget højde for. Dette bekræfter tidligere antagelse om, at LGP ikke er signifikant for modellen.

Bilag 5 viser de tre uafhængige variabler til at være vægtet på venstre side af normalfordelingen, hvilket gør at antagelsen om normalfordeling ikke kan godkendes. Hertil skal de tre variablers data transformeres, så det kan bruges til pålidelige hypotesetests. Transformering af dataet kan også hjælpe med at forbedre på modellen, og kan være en grund til det ringe resultat. En helt simpel log-transformation af LGP, LOPPL og LNOA ses i bilag 5, hvor en normalfordeling fremkommer for fejlleddene, og den højt vægtede tendens disse tre variabler havde i venstre side er fjernet. Som nævnt tidligere, så skal bilag 5 udelukket bruges til at vise hvad en logaritmisk transformation kan gøre for data. Dette er fordi at den naturlige logaritmiske transformation ikke er særlig godt egnet til at håndtere regnskabstal, som LGP, LNOA og LOPPL. Den naturlige logaritmiske transformation gør det matematisk umuligt at tage den logaritmiske værdi af et negativt tal og nul. For empiriske analyser, med variabler der ikke går i nul eller har negative værdier, er det derfor meget normalt at tage den naturlige logaritme af tallet, for at transformere sit data. Dette gør at projektet står overfor fire fremgangsmåder, for at håndtere data fra de tre uafhængige variabler. Disse fire fremgangsmåder er, som følger:

1. Vælge ikke at lave nogen form for transformation på dataet for de tre uafhængige variabler. Resultatet af dette giver en model, hvor der ikke pålideligt kan testes på de forskellige parametre, med de samme tests, som er kørt på de forrige modeller.
2. Finde en alternativ måde at transformere datet på, hvor håndteringen af negative værdier, samt nul-værdier, ikke er matematisk umuligt.
3. Slette de værdier der ikke kan transformeres. Denne valgmulighed forhindrer at sammenligne model 3 med de øvrige modeller, med mindre de øvrige modeller også vælger at fjerne disse værdier. Ydermere betyder dette, at der ikke vil være nogle negative eller nulværdier tilbage i dataet. Alle observationer, som har negativ værdi for LGP, LOPPL og LNOA, bliver derfor fjernet, hvilket er en dårlig ide, i forhold til integriteten af projektet. Det er altså ikke hensigtsmæssigt at arbejde med regnskabstal, og i et projekt, som dette, slette alle negative og nulværdier.

4. Den sidste fremgangsmåde dette projekt ser, er at ligge en konstant til alle observationer for variablerne. Denne konstant har til formål at få alle observationer over en værdi på nul, og så længe konstanten er den samme, så er det udelukket interceptet der bliver ændret af det. Interceptet fratrækkes derfor konstanten efterfølgende, og den sande model viser koefficienterne for de uafhængige variabler er det samme. Dette er fordi, at koefficienterne baseres på forholdet mellem hinanden og den afhængige variable, og hvis modellen forskydes, bliver sammenhængen ikke rørt. (Ekwaru & Veugelers, 2018).

Fremgangsmåde 1 og 3 ses, som at være helt udelukket. Fjernes alle negative og nul-værdier er det ikke så meget antallet af observationer der mistes, men mere at der systematisk fjernes alle negative og nul-værdier. Af negative og nul-værdier har OPPL 5.107, NOA har 813 og GP har 56. Dette går imod integriteten for dette projekt, og at fjerne alle de år, hvor en virksomhed har haft underskud på LGP, LOPPL eller LNOA, vil påvirke pålideligheden af svaret på problemformuleringen. Vælges der at følge fremgangsmåde 1, så vil antagelsen om normalitet ikke være fulgt for model 3, og model 5 (hvor denne problemstilling også fremkommer), og hypotesetests ville ikke kunne gennemføres. Ultimativt er det kun fremgangsmåde 2 og 4, som ses at være realistiske for dette projekt. Hertil vil der blive anvendt fremgangsmåde 2, da det også viser sig at være en del problemer med at tage den naturlige logaritme af dataet. Hertil er det en ny transformering af dataet der bliver valgt, og den vil sammenlignes med fremgangsmåde 4.

For at transformere dette data med den naturlige logaritme, så skal de værdier der er lig med nul, eller er negative håndteres. For model 3 gælder det variablerne LOPPL, LNOA og LGP, som hver har deres minimum værdi, som ses i tabellen nedenfor:

Variable	LOPPL	LNOA	LGP
Minimumsværdi	$-3.001.000$	$-9.821.649$	-172.618

Tabel 12 - af egen inspiration.

Her ses det at den absolutte værdi af minimumsværdierne fra tabel 12 er relativt høje, specielt for LNOA og LOPPL. Dette betyder, at for at kunne lave en log transformation af alle tre variabler, skal der ligges konstanten $9.821.650$ til alle observationer. Dette er for at alle observationer, for alle tre variabler, sikres at være større end nul. Skæringspunktet for modellen bliver derfor lidt besværlig at tolke på, da det er blevet kunstigt oppustet af konstanten. Bemærk også forskellen på den negative værdi for LNOA og LGP, dette vil betyde at LGP bliver kunstigt oppustet med en værdi der er meget højere end dens mindste værdi. Dette er på grund af,

at alle tre uafhængige variabler skal tillægges den samme konstant, for ikke at ødelægge sammenhængen mellem observationerne. Hvis dette ikke skulle være nok til at overbevise enhver til at undgå en log transformation af dette data, så er der blevet konstrueret bilag 7, for at vise endnu en fejl ved transformationen. I bilag 7 vises der histogrammer over LOPPL, LNOA og LGP for log transformationen med konstanter. Derudover har de tre plots i bilag 7 også illustreret kurven for normalfordelingen. Bemærk her, at bilag 7 ikke er meget bedre end hvad der er vist i bilag 5, og variablerne ikke er normalfordelt. Dette betyder at en log transformation af dataet, med en konstant, ikke har gjort meget for at gøre variablerne normalfordelt.

Fremgangsmåde 1, 3 og 4 ses derfor ikke til at være valide strategier for at håndtere dataet, hvilket betyder at der skal anvendes fremgangsmåde 2. Af alternative måder at transformere data på, er der en del forskellige, som kan anvendes, hvor dette projekt vælger at anvende cube root transformationen. Cube root transformationen er en stærk transformation, som er lidt svagere end log transformationen (Cox, 2007). Cube root er stærkere end andre transformationer, som kvadratrodstransformationen, og derudover er cube root transformationen også et fornuftigt valg for data med negative værdier (Cox, 2007). Formlen for en cube root transformation er simple, og ses som følger:

$$Cube\ root(x) = x^{\frac{1}{3}}$$

Matematisk set, betyder dette at en negativ værdi kan forblive negativ, og dataet har derfor stadig et svar på underskud rapporteret for de gældende virksomheder. Fordelen ved cube root transformationen, fremfor log transformationen, er derfor styrken i at beholde underskud i dataet, hvilket log transformationen med en konstant vil korrigere for. Tages et kig på de nye histogrammer, baseret på cube root transformationen, ses resultatet i bilag 8.

Bilag 8 viser histogrammer over cube root transformationer for LOPPL, LNOA og LGP. På baggrund af bilag 8, er det tydeligt at se, at der er sket en forbedring, både når der bliver sammenlignet med bilag 5 og bilag 7. Dette beviser derfor, hvordan cube root transformationen er stærkere for dette data, i forhold til en log transformering, baseret på introduktionen af en konstant. Bilag 8 viser en normalfordelt LOPPL og LNOA, som er klar til at blive anvendt i en ny version af model 3. Hertil, vurderes dataet nu til at være normalfordelt, og følgende antagelser for model 3, vil derfor baseres på cube root transformationen.

En ny model kan nu konstrueres for model 3, hvilket er baseret på data der er transformeret med cube root, og søger at finde en fremtidig værdi af RNOA. Den nye model, kaldes ”Transformeret model 3”. Nedenstående ses

denne nye model i illustration 11, hvor dens output er til venstre og til højre ses model 3, men med en log transformering.

```
Call:
lm(formula = RNOA ~ Datasaet$cubeLOPPL + Datasaet$cubeLNOA +
    Datasaet$cubeLGP)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3170 -0.1755 -0.0795  0.0884  4.3399

Coefficients:
                    Estimate Std. Error t value  Pr(>|t|)
(Intercept)        0.2937301  0.0038354  76.584   < 2e-16 ***
Datasaet$cubeLOPPL 0.0096465  0.0002474  38.989   < 2e-16 ***
Datasaet$cubeLNOA -0.0072438  0.0002130 -34.008   < 2e-16 ***
Datasaet$cubeLGP   0.0012157  0.0002780   4.374 0.0000122 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.456 on 39451 degrees of freedom
Multiple R-squared:  0.06088,   Adjusted R-squared:  0.06081
F-statistic: 852.5 on 3 and 39451 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = RNOA ~ log.k.LOPPL + log.k.LNOA + log.k.LGP, data = Datasaet)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9670 -0.2001 -0.0928  0.0999  4.4350

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.95308    4.32735  -6.922 4.53e-12 ***
log.k.LOPPL   2.96953    0.37097   8.005 1.23e-15 ***
log.k.LNOA   -0.06942    0.02816  -2.465   0.0137 *
log.k.LGP    -1.02314    0.13347  -7.666 1.82e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4701 on 39451 degrees of freedom
Multiple R-squared:  0.001945,  Adjusted R-squared:  0.00187
F-statistic: 25.63 on 3 and 39451 DF,  p-value: < 2.2e-16
```

Illustration 11 - af egen inspiration.

Tidligere i dette afsnit er baggrunden bag beslutningen om ikke at anvende en log transformeret gennemgået. Til at understrege disse argumenter, vil der derfor tages et indblik i de statistiske forskelle, havde modellen anvendt en log transformation, og derfor testes fremgangsmåde 4. Alle p-værdier der er printet i illustration 11, er signifikante, for begge modeller, dog er p-værdierne lidt stærkere for cube root transformationen. Hertil antages igen, at nul-hypotesen forkastes, og den alternative hypotese accepteres. Derudover bedes der også bemærkes, at LGP nu er en koefficient, som er signifikant for modellen, og med over 95 % signifikansniveau. Med transformeringen af dataet, og den nye signifikans af LGP, vil antagelsen om multikollinearitet igen testes for modellen. Dernæst ses det at residual standard errors har en forskel på 0,0141, hvilket er en lille forskel, men det er i fordel for cube root modellen. Dog er den helt store forskel, forskellen på den justerede R^2-værdi, hvor cube root transformationen har en værdi på 0,0608, og log transformationen har en værdi på 0,0019. Den justerede R^2-værdi giver et godt indblik over forskellen på modellerne, den kan også sammenlignes med værdierne fra de tidligere versioner af model 3, hvor cube root transformationen er den klart stærkeste for dette datasæt.

Nedenstående ses illustration 12, hvor der igen bliver fremvist både versioner for log transformationen (højre), samt cube root transformationen (venstre):

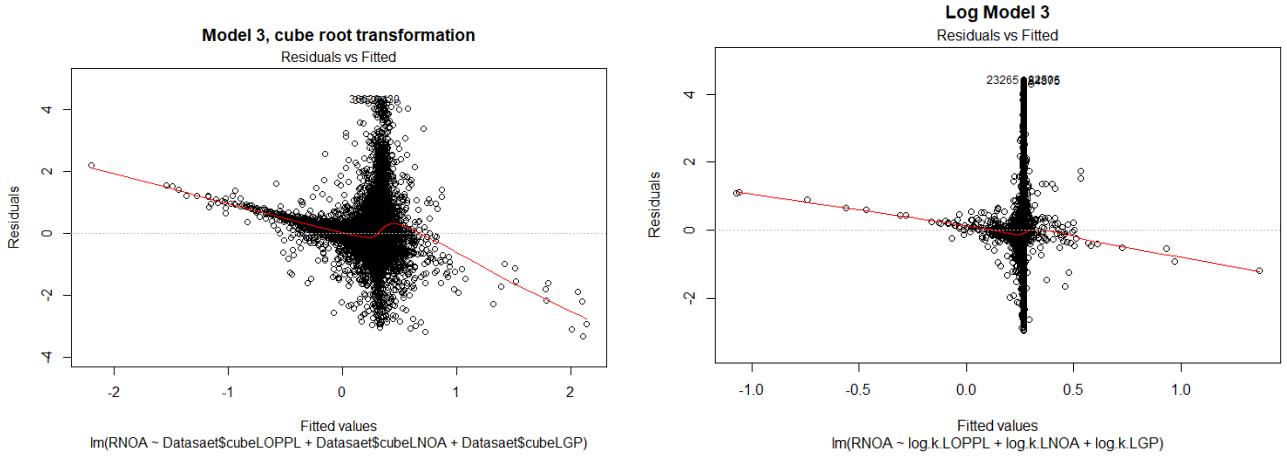


Illustration 12 - af egen inspiration.

Der er en tydelig forskel på de to måder at transformere data på, log transformationen kan ikke sprede dataet ud, så godt, som cube root transformationen. Hertil ses det i illustration 12, at residualerne er relativt spredt rundt om den fitted linje, hvilket tyder på en svagere homoskedasticitet. Ydermere betyder det at der kan forekomme heteroskedasticitet i sættet, men dataet er transformeret. Lineariteten for cube root transformations modellen er lidt mere tvivlsom, der er en tydelig linje i sættet, men den har problemer i halerne, med at ramme den såkaldte ”0-linje”. Ydermere ses det ikke, at i illustration 12, der optræder en tragtform, eller butterflyform i forhold til variansen af residualerne, hvilket igen tyder på homoskedasticitet i dataet. Antagelserne omkring linearitet, homoskedasticitet, og normalfordelingen er gennemgået, men med det nye transformeret data, skal multikollinearitet, som forklaret tidligere, gennemgås igen. Fremover vil der udelukket tages højde for model 3 med data der er cube root transformeret, da dette ses til at være den overlegne model. For at teste for multikollinearitet for cube root transformationen af model 3, ses nedenstående korrelationsmatrice for variablerne:

Korrelationsmatrix	LOPPL	LGP	LNOA
LOPPL	1	0,6844	0,6173
LGP	0,6844	1	0,8524
LNOA	0,6173	0,8524	1

Tabel 13 - af egen inspiration.

Tabel 13 viser hvor meget sammenhæng der er mellem de forskellige variabler. Denne tabel skal sammenlignes med tabel 10, hvor det ses at efter cube root transformationen er disse variabler ikke så højt korreleret. Den

ovenstående tabel 13 viser også at der stadig er en relativ høj korrelation mellem LNOA og LGP, men for at sikrer, at en af disse værdier ikke skal fjernes, vil følgende VIF-test blive gennemgået:

Nøgletal	LOPPL	LNOA	LGP
VIF-scorer	1,8962	3,6874	4,2934

Tabel 14 - af egen inspiration.

Tabel 14 viser at for LOPPL og LNOA har variabler en VIF-scorer der er under 4. Derimod er VIF-scoren for LGP over 4, hvilket betyder at der forekommer multikollinearitet i dataet (Pardoe, 2019). Den høje korrelation LGP og LNOA har vist i tabel 13, samt VIF-scoren i tabel 14, viser multikollinearitet i dataet, og LGP fjernes derfor fra modellen. Dette giver følgende nye output, tilsidesat med det nye plot, der beskriver linearitet, samt homoskedasticitet:

```
Call:
lm(formula = RNOA ~ cubeLOPPL + cubeLNOA)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1722 -0.1781 -0.0807  0.0897  4.3398

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3015232  0.0033972   88.76   <2e-16 ***
cubeLOPPL    0.0100627  0.0002284   44.05   <2e-16 ***
cubeLNOA    -0.0065455  0.0001410  -46.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4561 on 39452 degrees of freedom
Multiple R-squared:  0.06043,   Adjusted R-squared:  0.06038
F-statistic:  1269 on 2 and 39452 DF,  p-value: < 2.2e-16
```

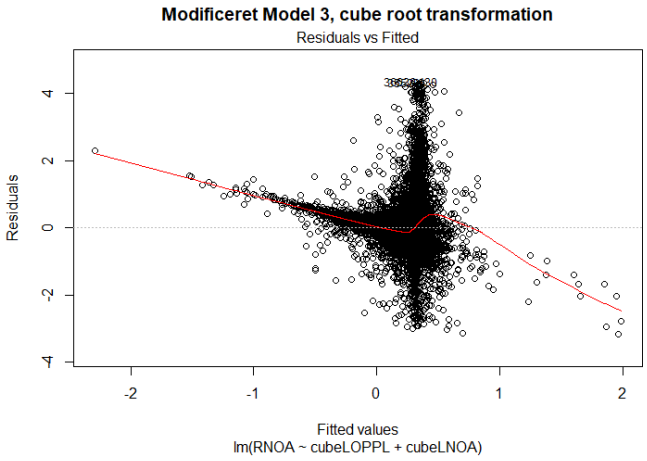


Illustration 131 - af egen inspiration.

Illustration 13 viser en modificeret transformeret model 3, som fremover i de senere afsnit, vil references til, som model 3, da denne model, anses, som at være den der bedst opfylder de lineære antagelser. Hertil ses en model, som er lidt svagere når det kommer til dens statistiske værdier, i forhold til før LGP blev fjernet. Ydermere er den nye residualstandard error på 0,4561, og med en justerede R^2-værdi på 0,0604. Til højre i illustration 13, ses en model, som antages at være lineær, med homoskedastiske varians på residualerne. Med den nye version af model 3, ses følgende tabel, der fremviser de nye VIF-scorer for modellen:

Nøgletal	LOPPL	LNOA
VIF-scorer	1,6157	1,6157

Tabel 15 - af egen inspiration.

Tabel 15 viser, at med den nye version af model 3, kommer der også VIF-scorer, der ligger langt under skæringspunktet på 4 (Pardoe, 2019).

Konklusionen på dette, er at alle antagelser nu er gennemgået for model 3, hvilket er resulteret i en forbedret model, der anvender cube root transformation, for at normalfordele variablerne. Den version af model 3, som vil bruges fremover, er derfor baseret på cube root transformeret data, og den ser ud, som følger:

$$RNOA_{t+1} = 0{,}3015 + 0{,}0101 * OPPL_t - 0{,}0065 * NOA_t$$

Værdierne for koefficienterne er relativt lave, men dette er på grund af de høje værdier for de uafhængige variabler, set i absolutte tal. For at se på, hvor betydelig den residual standard error er, i forhold til RNOA, ses følgende beregning:

$$\frac{0{,}4561}{0{,}2642} = 1{,}7263$$

Ovenstående beregning er, som ved de andre modeller, en beregning der viser, hvor meget model 3 i gennemsnit fraviger fra den sande værdi, med en værdi, der er 1,7263 gange højere end middelværdien.

En visualisering over hvad model 3 estimater, i forhold til den aktuelle værdi den forsøger at estimere, så illustreres det i nedenstående illustration:

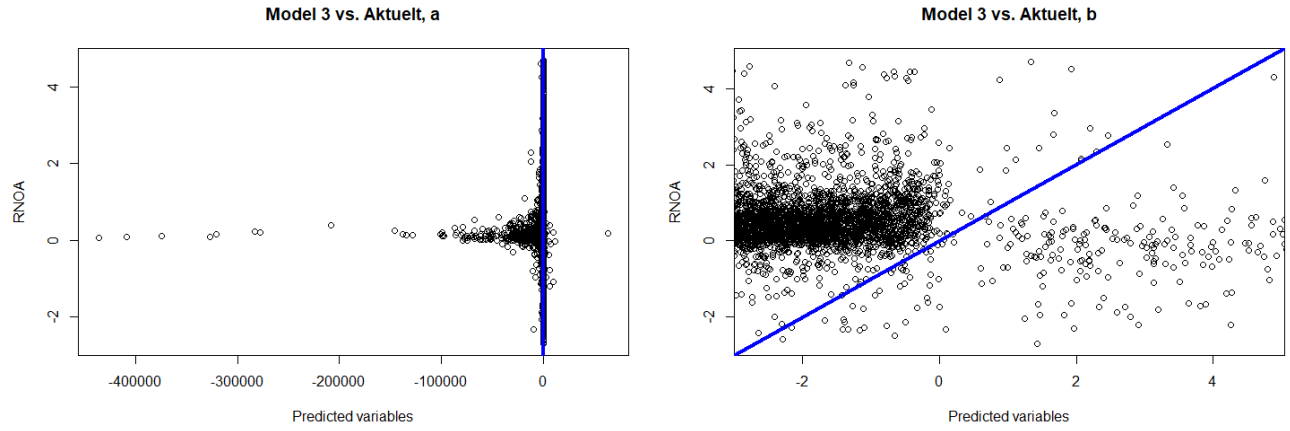


Illustration 142 - af egen inspiration.

Illustration 14 viser at model 3 har meget svært ved at beskrive RNOA, og der er nogle ekstreme værdier, hvor der forekommer en stor vægt af negative værdier. Bemærk for illustration 14 version b (højre), at det ikke er alle residualer der er beskrevet, da ændringen i længden af x-aksen, ikke ændrer residualernes placering. Illustration 14, version b viser derfor udelukket hvor residualerne burde være, da akselængderne er på samme interval, som RNOA. Ydermere viser illustration 14 også at da LNOA er den eneste variable med negativ koefficient, har LNOA en del residualer der ikke kan håndtere dette. Hertil vil denne problemstilling gennemgås ved hjælp af følgende tabel:

	Estimeret værdier for model 3	RNOA	Model 3's ekstreme værdier udenfor RNOA's interval
Minimumsværdi	-436.180,4	-2,7021	36.438
Maksimumsværdi	63.841,2	4,7011	447

Tabel 16 - af egen inspiration.

For model 1 og 2 var der fremvist en lignende tabel, dog med en fundamental ændring. Tabel 16 søger at svare på samme problem, som for tidligere modeller, dog er fremgangsmåden ændret lidt. For model 1 og 2 var det residualer udenfor intervallet for modellernes estimeret værdier, hvor det omvendte er sandt for model 3. Dette betyder at i stedet for, at modellen underestimere, i forhold til den aktuelle værdi af RNOA, så for model 3 overvurderes den aktuelle værdi af RNOA, og meget i nogle tilfælde. Tabel 16 viser blandt andet en minimums- og maksimumsværdi, der ligger langt over hvad den burde for modellen. Hertil ses det også i illustration 14 version a (venstre), hvor den blå linje er lodret, på grund af længden af x-aksen. Ignoreres mange af de helt ekstreme værdier, og forstørres der på x-aksen, ses det også at residualerne ikke nødvendigvis ligger godt op ad den blå linje. Tabel 16 viser også at model 3 har 36.438 residualer under minimumsværdien på RNOA, og 447 værdier over maksimumsværdien på RNOA. Dette betyder at der er 36.885 residualer, som ikke befinder sig indenfor det aktuelle interval for RNOA, sammenlignet med hele datasættet på 39.455. Illustration 14 og tabel 16 afspejler den utrolig lave justerede R^2-værdi på 6,04 %. I bilag 9 ses et lignende billede for de to plots fra illustration 14, men lavet på baggrund af de 100 tilfældige udvalgte observationer. Ydermere, på det ene plot med samme længde på akserne (tilpasset til RNOA), er der kun tre residualer på plottet.

5.5 Model 4: a og b, den nye metodik

De foregående modeller har ikke været nogen succes, derfor blev denne alternative fremgangsmåde introduceret. Her bliver der foretaget en simple dekomponering af RNOA, men variablerne deles op i hver deres model. Da tankerne bag dette er gennemgået, ses model 4a og 4b nedenfor, med model 4a til venstre og 4b til højre.

$$PM_{t+1} = \alpha_0 + \beta_1 * PM_t \quad , \quad ATO_{t+1} = \alpha_0 + \beta_1 * ATO_t$$

Bemærk her, at der ikke måles det nuværende niveau af RNOA, men det er de underliggende niveauer for RNOA's to value drivers PM og ATO. Hvis de foregående modeller giver nogen indikation for disse modeller, så er det at jo færre variabler jo bedre, dette kan måske være med til at forklare de skuffende resultater fra model 2 og 3. Nedenstående ses illustration 15, som er outputtet for model 4a og 4b, her er det med 4a til venstre, og 4b til højre:

```
Call:
lm(formula = PM ~ LPM)

Residuals:
     Min       1Q   Median       3Q      Max
-1.82240 -0.06646  0.00115  0.07939  1.89541

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.060263   0.001234   48.83   <2e-16 ***
LPM         0.606254   0.004259  142.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1892 on 39453 degrees of freedom
Multiple R-squared:  0.3394,    Adjusted R-squared:  0.3393
F-statistic: 2.027e+04 on 1 and 39453 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = ATO ~ LATO)

Residuals:
    Min      1Q  Median      3Q     Max
-31.687  -0.473  -0.271   0.139  31.953

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.593897   0.012284   48.35   <2e-16 ***
LATO        0.665542   0.003868  172.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.97 on 39453 degrees of freedom
Multiple R-squared:  0.4287,    Adjusted R-squared:  0.4287
F-statistic: 2.961e+04 on 1 and 39453 DF,  p-value: < 2.2e-16
```

Illustration 15 - af egen inspiration.

Både model 4a og 4b søger at forudsige en fremtidig værdi af den afhængige variable, ved at lagge den uafhængige variable en gang. Dette er samme metode, som blev brugt i model 1, for benchmarket, da der igen er tale om simple lineære regressioner. Outputtet fra illustration 15 giver følgende ligning for model 4a:

$$PM_{t+1} = 0{,}0603 + 0{,}6063 * PM_t$$

Og model 4b:

$$ATO_{t+1} = 0{,}5939 + 0{,}6655 * ATO_t$$

Resultatet af disse ligninger danner en forventet værdi af variablen i den kommende periode. Produktet af disse resultater, giver den forventede værdi af RNOA i den kommende periode. Disse to modeller er, som nævnt tidligere baseret på en teoretisk tilgang, præsenteret af Plenborg, Petersen og Kinserdal (2017). Spørgsmålet er her, om denne alternative fremgangsmåde er bedre end det resultat model 1 fremviser. For de to modeller gælder der også to justerede R^2-værdier. For model 4a, bliver den justerede R^2-værdi 0,3393, eller 33,93 % og model 4b er den justerede R^2-værdi på 0,4287, altså 42,87 %. For både model 4a og 4b, så produceres der høje justerede R^2-værdier, hvilket som udgangspunkt er et godt tegn.

For at sikrer at model 4a og 4b er relevante modeller, skal nul-hypotesen for signifikansen af koefficienterne testes. Dette er gjort, som tidligere i projektet, hvor der bliver fokuseret på p-værdier for modellen. P-værdien for koefficienterne for model 4a og 4b er den samme. Ydermere er disse p-værdier så utrolig lave igen, at koefficienterne vurderes, som at være signifikante, med et signifikansniveau på 95 %. Dette betyder, at nul-hypotesen bliver forkastet, og den alternative hypotese accepteres, altså at koefficienterne er signifikante.

Efter koefficienterne er vurderet signifikante, er det næste skridt for at teste modellens statistiske præcision, at se på modellernes residual standard error. For model 4a, har modellen en residual standard error på 0,1892, og PM har et mean på 0,172. Dette betyder, at i gennemsnit, så fejlvurdere denne model med en værdi på 0,1892 af den sande lineære regression. I forhold til den gennemsnitlige værdi af PM, så svarer det til:

$$PM: \quad \frac{0{,}1892}{0{,}172} = 1{,}1$$

Altså fraviger model 4a fra den sande lineære regression, med en værdi på 1,1 gange større end den gennemsnitlige værdi af sættet. For model 4b er der en standard residual error på 1,97, og ATO har en mean på 1,841. Laver man samme regnestykke med ATO, ser det ud som følger:

$$ATO: \quad \frac{1{,}97}{1{,}841} = 1{,}0701$$

Dette betyder altså, at model 4b fraviger i gennemsnit fra den sande lineære regression med en værdi der er 1,0701 gange større end gennemsnitsværdien af ATO.

I model 2 blev der gennemgået, hvordan begge uafhængige variabler, LPM og LATO, er normalfordelte. Ydermere, da der er tale om to forskellige simple lineære regressioner så er multikollinearitet ikke et problem for model 4a og 4b. Dette betyder, at der kun mangler at testes for homoskedasticitet og linearitet for model

4a og 4b, før der kan bekræftes at hypotesetests er lavet på en pålidelig baggrund. Nedenstående ses illustration 16, hvor til venstre ses model 4a (PM) og til højre ses model 4b (ATO):

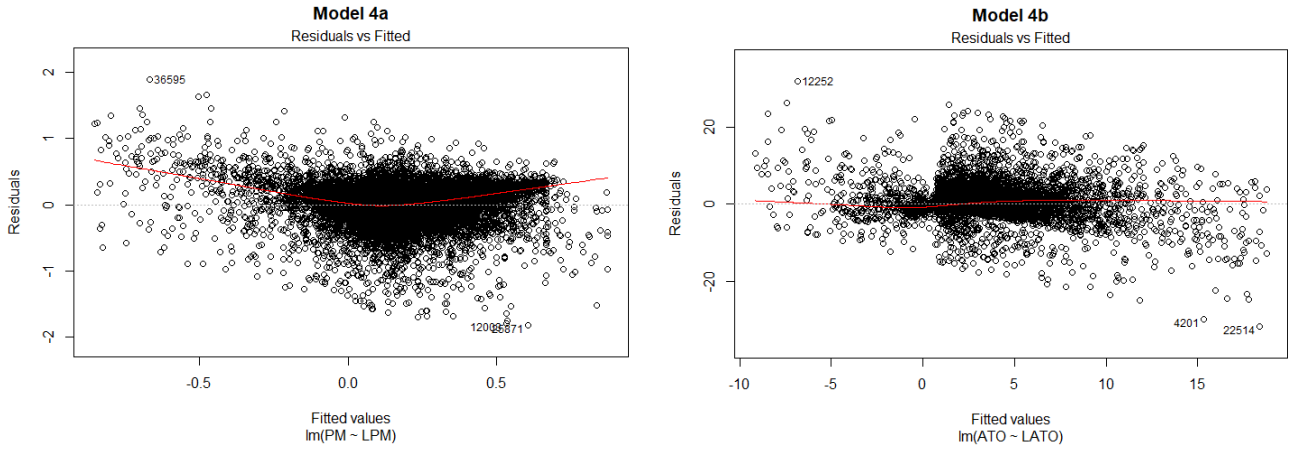


Illustration 163 - af egen inspiration.

Model 4a og 4b har en lignende tendens med variansen i deres variabler. Variablerne ses at have en konstant længde på aksen omhandlende residualer. Der er derfor både en mangel på en tendens til residualerne er tragtformet eller følger en butterflyform. Ifølge Osborne og Waters (2002) betyder manglen af disse to typer af trend på de ovenstående modeller, at der også er mangel på heteroskedasticitet. Mangel på heteroskedasticitet, betyder naturligvis at for begge modeller, er der tale om homoskedasticitet for variansledende. For model 4a er den røde linje relativ tæt på "0-linjen", udover i halerne. Selvom disse haler for model 4a eksistere, så på baggrund af en fin længde på variansen og et normalfordelt sæt, vurderes model 4a til at være lineær, og hypotesetests kan derfor forekomme. For model 4b, ses en linje der ligger meget nær "0-linje", hvilket kraftigt tyder på linearitet for modellen. Ud fra dette, kan det konkluderes at for model 4a og 4b er antagelserne omkring en lineær model godkendt, og sammenligninger med andre modeller kan anvendes på baggrund af de pålidelige modeller.

En visualisering af præcisionen af model 4 ses i følgende illustration:

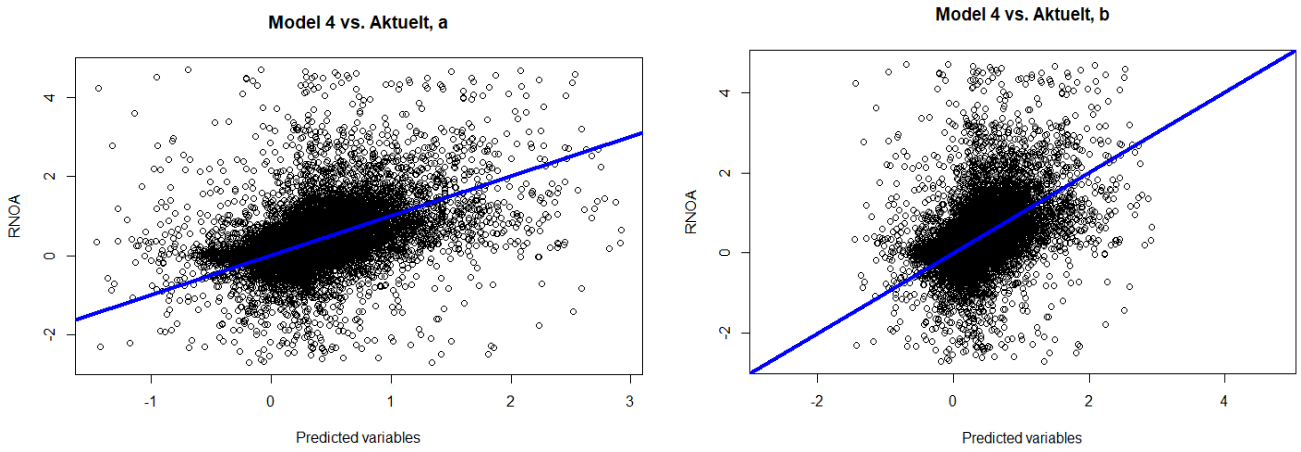


Illustration 174 - af egen inspiration.

Illustration 17 viser et lignende billede, som er set for model 1 og 2, og er samtidigt en stor forbedring for model 3. For at kunne sige mere om illustration 17, ses følgende tabel:

	Estimeret værdier for model 4	RNOA	RNOA's ekstreme værdier udenfor model 2's interval
Minimumsværdi	-1,4576	-2,7021	136
Maksimumsværdi	2,9243	4,7011	164

Tabel 16 - af egen inspiration.

I illustration 17 version a (til venstre) ses en x-akse der er længere end for de tidligere modeller, hvilket tyder på en model, der er bedre til at estimere de mindste og største værdier. Dette understøttes af tabel 17, der informerer at det totale antal af residualer, som er udenfor intervallet for de estimeret værdier for model 4, er på præcis 300. Metoden til at måle antallet af estimater udenfor modellens forventede værdier, er brugt i hele dette projekt, og resultatet fra model 4 giver det bedste resultat hidtil. Dette kan ikke enestående anvendes til at sammenligne modellerne, hvor der her skal flere præcisionsmål til. For at få et bedre indblik i illustration 17, ses følgende to plots, der er baseret på det tidligere nævnte subsample:

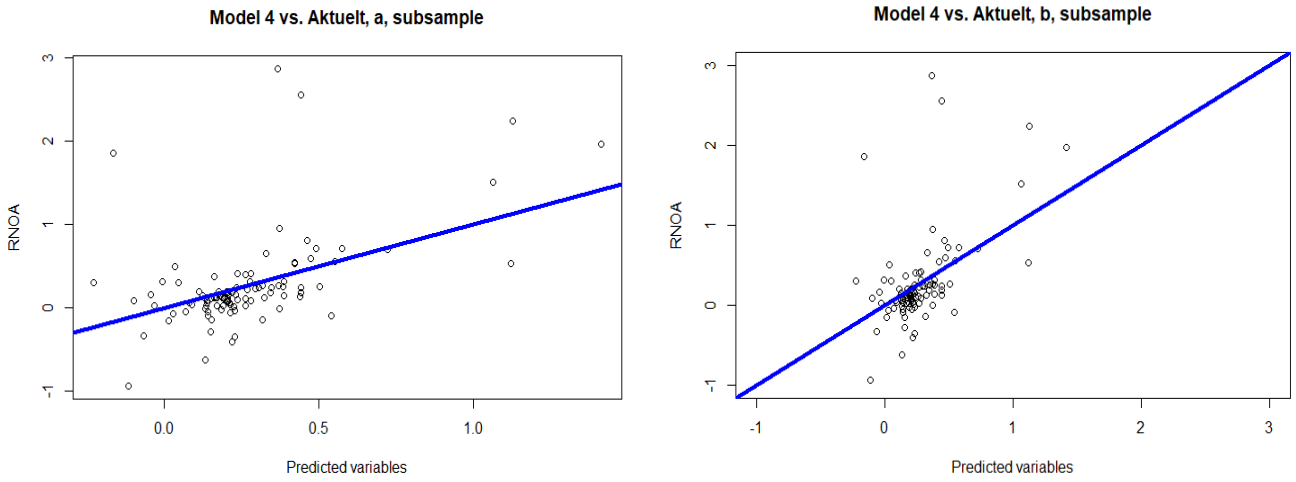


Illustration 18 - af egen inspiration.

Illustration 18 viser samme overordnede resultat, som for de tilsvarende plots for model 1 og 2. Her er der illustreret 100 residualer, hvor de fleste er centreret omkring den blå linje, med en lille håndfuld residualer, der afviger fra linjen.

5.6 Model 5: a og b: To-lags DuPont dekomponering, med ny metodik

Næste skridt for at svare på problemformuleringen, er at tage den alternative fremgangsmåde introduceret i model 4, og pålægge den en to-lags DuPont dekomponering. Som vist i modeldesignet, giver dette to relevante simple lineære regressioner, her ses model 5a og 5b i rækkefølge:

$$OPPL_{t+1} = \alpha_0 + \beta_1 * OPPL_t$$

$$NOA_{t+1} = \alpha_0 + \beta_1 * NOA_t$$

Disse to modeller ses nedenfor i illustration 19, med model 5a (venstre) og 5b (højre). Hertil anvendes de cube root transformationen af tallene, på baggrund af fund i model 3:

```
Call:
lm(formula = OPPL ~ cubeLOPPL)

Residuals:
     Min       1Q   Median       3Q      Max
-1301160   -30065    -3601    15055 11823116

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -51970.17     919.63  -56.51   <2e-16 ***
cubeLOPPL     6054.26      55.27  109.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140300 on 39453 degrees of freedom
Multiple R-squared:  0.2332,    Adjusted R-squared:  0.2332
F-statistic: 1.2e+04 on 1 and 39453 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = NOA ~ cubeLNOA)

Residuals:
     Min       1Q   Median       3Q      Max
-8579897  -168916    65529   201998 74882154

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -505348       6325   -79.9   <2e-16 ***
cubeLNOA       26637        209   127.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 859000 on 39453 degrees of freedom
Multiple R-squared:  0.2917,    Adjusted R-squared:  0.2917
F-statistic: 1.625e+04 on 1 and 39453 DF,  p-value: < 2.2e-16
```

Illustration 19 - af egen inspiration.

De to simple lineære regressioner, ser derfor ud som følger, her er der igen tale om cube root transformationen af de uafhængige variabler:

$$OPPL_{t+1} = -51.970{,}17 + 6.054{,}26 * OPPL_t$$

$$NOA_{t+1} = -505.348 + 26.637 * NOA_t$$

Disse er selvfølgelig i rækkefølgen med 5a øverst og 5b nederst. Første indtryk af disse modeller, er de relativt høje værdier for den justerede R^2-værdi. Model 5a har en justerede R^2-værdi på 0,2332, hvilket er 23,32 %, dernæst har model 5b en justerede R^2-værdi på 0,2917, som svarer til 29,17 %. Disse værdier er relativt fornuftige, hvilket er et godt udgangspunkt for modellen. Ydermere kan man se alle p-værdier for alle koefficienter, og derfor også for modellerne, er langt under 0,05, hvilket igen betyder at alle koefficienter er signifikante på et niveau af 95 %. Opstilles der her en nul-hypotese, som der er gjort for de andre modeller, forkastes den, hvilket betyder den alternative hypotese accepteres og koefficenterne er forskellige for nul. Som de yderligere modeller, så vil residual standard errors for modellerne blive sammenlignet med middelværdien for koefficienterne. Denne udregning er, som diskuteret tidligere, for at give en ide om, hvor langt modellen i gennemsnit fejlvurdere den sande lineære model, i forhold til dens gennemsnitlige værdi. For model 5a ser det således ud:

$$OPPL: \quad \frac{140.300}{12.568} = 11{,}1633$$

Og for model 5b:

$$NOA: \quad \frac{859.000}{83.040{,}39} = 10{,}3444$$

Residual standard error for model 5a er på 140.300, og med et gennemsnit på 12.568 fraviger den sande værdi i gennemsnit med 11,1633 gange så stor en værdi end gennemsnittet. For model 5b, så falder denne værdi til 10,3444, udregnet fra en standard residual error på 859.000 og en mean på 83.040,39. Værdierne for model 5a og 5b, viser at de to modeller, har en tendens til at i gennemsnit tage fejl med en relativt høj værdi, i forhold til variablernes gennemsnit.

Baseret på resultater fra model 3, som nævnt tidligere, er det set at de uafhængige variabler er normalfordelt, når der anvendes cube root transformationen. Ydermere er der tale om simple lineære regressioner, hvilket betyder, at der ikke skal gennemgås antagelsen omkring multikollinearitet. Hertil er det linearitet og homoskedasticitet der skal gennemgåes. Nedenstående ses illustration 20, der til venstre viser model 5a, og til højre ses 5b, der er baseret på cube root transformationen:

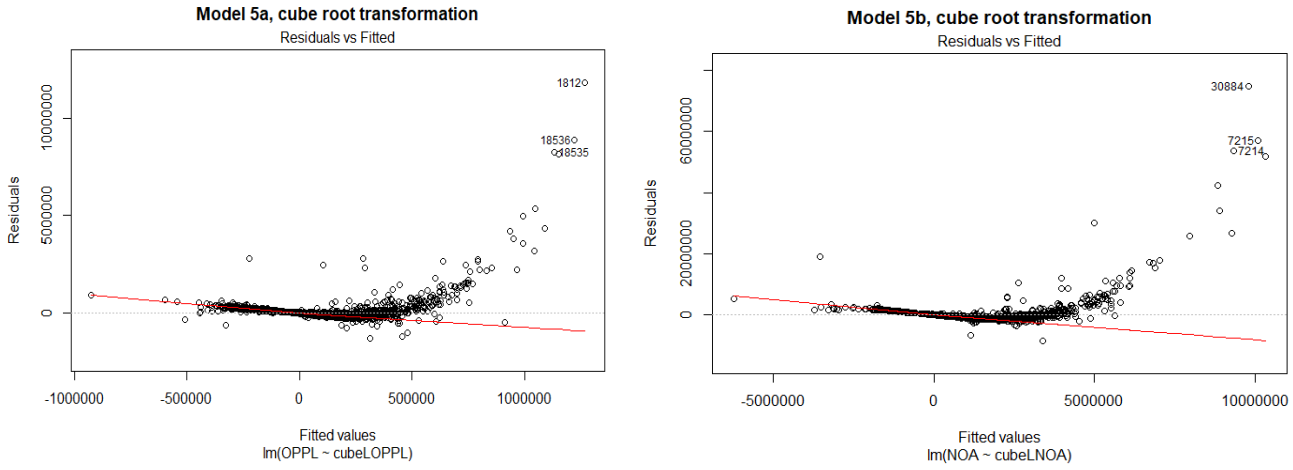


Illustration 205 - af egen inspiration.

Det essentielle for illustration 20 er, hvordan residualernes forhold er til den røde linje, samt den røde linjes forhold til "0-linjen". For begge modeller ses en rød linje der passer meget godt til "0-linjen". Dette betyder at modellerne vurderes begge, som at være lineare. Den del af illustrations 20, der ses at være mest foruroligene, er højre side af modellerne. Her ligger der en del residualer over den røde linje og "0-linjen", men der understreges her, at langt størstedelen af residualerne ligger lige omkring linjen, og omkring nul på det der traditionelt kaldes x-aksen. Datasættet er på over 39.400 observationer, hvilket betyder at residualerne der relativt tydeligt kan ses, og som er længere fra linjen end de andre residualer, udgør en mindre del af antallet af

residualer. For illustration 20 ses der ikke en tragt- eller butterflyform, hvilket igen antyder at dataet er homoskdastisk. Hertil skal det siges at der kan være en tendens til heteroskedasticitet i dataet, men dette er ikke nødventigvis katastrofalt for antagelsen om homoskedasticitet.

Model 5a og 5b ses derfor, som at være baseret på normalfordelt data, homoskadsticitet i dataet, samt en stærk lineær tendens. Ydermere, da dette omhandler simple lineære regressioner, så er antagelsen om multikollinearitet ikke relevant. Dette betyder at alle antagelser angående model 5a og 5b er godkendt, og der kan nu fortsættes med tests der måler præcisionen af modellerne, til sammenligning med andre modeller.

Når det kommer til at visualisere præcisionen af modellerne, er det ikke nogen hemmelighed at model 3 har et skuffende resultat. Model 5 er en forlængelse af model 3, i den forstand, at model 5 bryder model 3 op i flere simple lineære regerssioner. Nedenstående ses illustration 21, der viser de estimerede værdier for model 5, plottet mod de aktuelle værdier for RNOA:

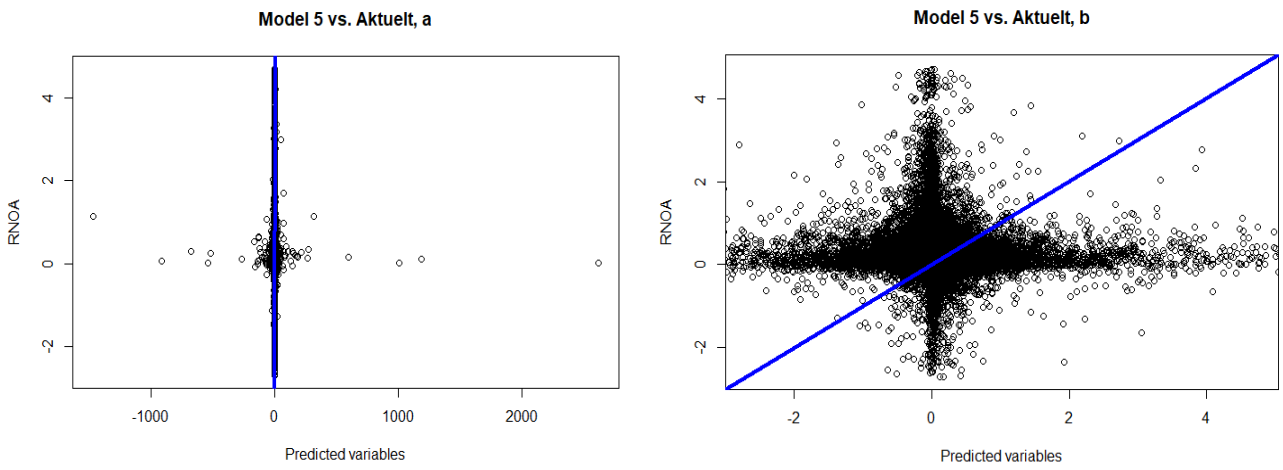


Illustration 21 - af egen inspiration.

Illustration 21 viser samme overordnede problem, som model 3 gør. Bemærk igen, at version b (til højre) af illustration 21, er akserne sat til at være samme længde, i forhold til RNOA, og derfor er ikke alle residualer i plottet. Igen er dette gjort for at give den ønskede 45-graders hældning på den blå linje. Illustration 21 kan derfor ikke bruges til meget andet, end at konkludere at model 5 ikke er god til at håndtere større absolutte værdier for nogle variabler. Nedenstående ses en tabel, hvis formål er at viderebygge på illustration 21, ved at undersøge de værdier der er udenfor RNOA's interval:

	Estimeret værdier for model 5	RNOA	Model 5's ekstreme værdier udenfor RNOA's interval
Minimumsværdi	-1.469,3494	-2,7021	470
Maksimumsværdi	2.625,2345	4,7011	298

Tabel 17 - af egen inspiration.

Illustration 21 får residualerne fra model 5, til at ligne resultatet fra model 3, hvor tabel 18 giver et helt andet billede. Tabel 18 følger samme metode, som tabel 16, hvor der er fokus på residualer udenfor RNOA's interval, eftersom de yderste værdier tilhører de estimerede værdier for model 5. Derudover viser tabel 18 også at der er 768 residualer, som er udenfor hvad illustration 21 version b viser. Dette betyder også at illustration 21, har en tendens til at overdrive med 768 residualer, sammenlignet med de 36.438 residualer fra model 3, som ligger udenfor RNOA's inteval. Konklusionen på illustration 21 og tabel 18 er ikke at modellen er god, dette gælder specielt når udtalelsen er på baggrund af model 3. For at kunne vurdere model 5, skal der flere præcisionsmål til, samt at anvende model 5 på de 100 tilfældige udvalgte observationer, hvilket ses nedenfor:

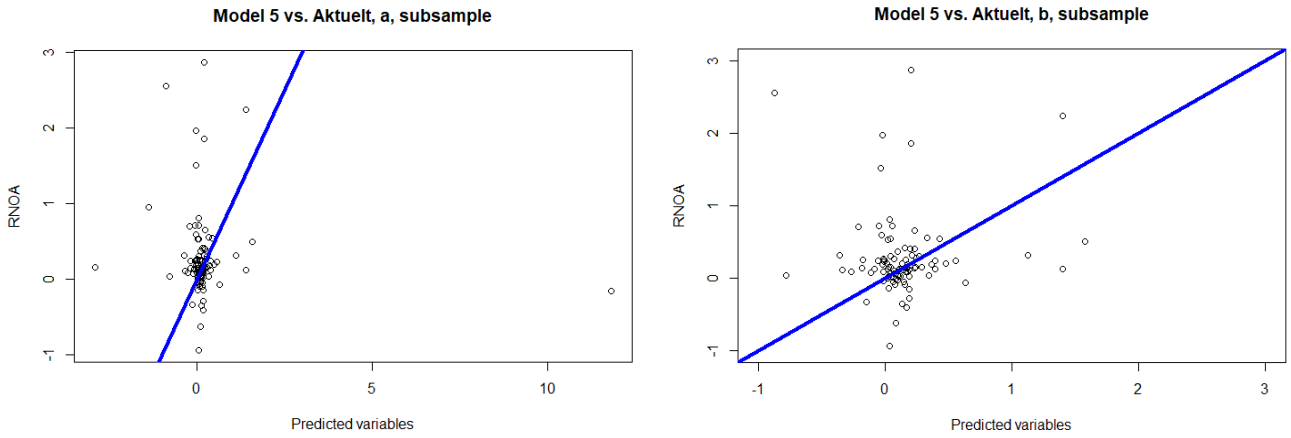


Illustration 226 - af egen inspiration.

På baggrund af illustration 22 ses at der stadig er en betydelig mængde residualer rundt om den blå linje. Der er dog nogle mere ekstreme værdier, som ikke er set for model 1, 2 og 4. Bemærk også, at i illustration 22, version b, er det ikke alle residualer, der er vist. Dette er igen på baggrund af manipulationen med x-aksen.

5.7 Delkonklusion – analysen

For at kunne have en fyldestgørende besvarelse af problemformuleringen, underspørgsmålene og hypoteserne, har det været nødvendigt at have en dybdegående analyse. Analysen er baseret på tidligere afsnit, heriblandt teori og metode, hvor modeller, nøgletal, data og vigtige teoretiske værker er gennemgået. På baggrund af modeldesignet og forarbejdet lavet på dataet, er der opsat fem modeller, som bedst muligt kan svare på problemformuleringen, på baggrund af et benchmark. Dette benchmark er model 1, som er den første model der gennemgås, og bliver basen for sammenligning af de fremtidige modeller. For alle fem modeller er der gennemgået en række antagelser, for at kunne finde frem til den mest optimale model. Model 1 viser sig at have relativt fornuftige resultater. Model 1's vigtigste funktion, er at den skal anvendes til at sammenlignes med andre modeller, og derfor anvendes af de andre modeller til at forbedre på. For alle modeller er der gennemgået en række forskellige værdier, til at kunne vurdere dem, hvor dette kan opdeles til to formål. Dette er den tidligere nævnte evne til at sammenligne, og for at sikre, at den optimale model er fundet frem til. Her kigges der blandt andet på residual standard error, justerede R^2-værdi, p-værdier, plots, samt visualiseringer på en subsample osv. Model 1 er rimelig ligetil, og det samme er model 2, hvor der her bliver introduceret, hvordan der behandles for en multipel lineær regression, hvor model 1 selvfølgelig er en simple lineær regression. Dernæst er der model 3, som er en yderligere dekomponering af model 2, og derfor stadig en multipel lineær regression. Dette skaber tre uafhængige variabler, hvilket er OPPL, NOA og GP, og model 3 ser ikke umiddelbart ud til at være en god model. Hertil prøves der at forbedres på modellen, ved at se på to overordnede antagelser for lineære regressioner. Første antagelse der anvendes for at forbedre på model 3, er at gennemgå multikollinearitet for at sikrer den høje p-værdi for LGP er med til at forklare dette. Resultatet af en modificeret version af model 3, forbedrer kun modellen utroligt lidt, og der undersøges derfor for normalfordelingen af fejlleddene. Hertil vises det, forskellen på de to populære typer af transformering af data, hvilket viser det bedste resultat for cube root transformationen. Efter den rette transformation er anvendt, tages et skridt tilbage for igen at se om LGP er blevet signifikant efter transformationen. Selvom LGP er signifikant, tages den ud af modellen, baseret på dens høje scorer i den efterfølgende VIF-test. Det antages nu at den optimale version af model 3 er fundet, men den viser dog stadig et skuffende resultat, hvor modellen har utroligt svært ved at gætte indenfor RNOA's aktuelle interval.

Model 4 introduceres med en ny metodik, som model 5 også anvender, hvilket er den såkaldte alternative fremgangsmåde præsenteret af Plenborg, Petersen og Kinserdal (2017). Hertil bliver to simple lineære regressioner anvendt, til at hver value driver har en separat lineær regression, hvorefter resultatet af disse

regressioner anvendes til at finde et estimat for RNOA. Model 4 giver umiddelbart et godt estimat for RNOA, og rammer for det meste godt indenfor intervallet for den aktuelle værdi for RNOA. Effekten af denne alternative fremgangsmåde bliver også understøttet til dels af model 5, som anvender uafhængige variabler der er baseret på model 3's uafhængige variabler. Model 3 og model 5 kan derfor direkte sammenlignes, hvor det er tydeligt at der allerede her er en forbedring på model 3. Model 5 er langt bedre til at ramme indenfor intervallet for RNOA. Hertil skal det understreges, at model 5 ikke nødvendigvis er en god model, udelukket på det kriterie at den er bedrer end model 3.

Konklusionen på dette, er at der er gennemgået fem modeller, hvor det nu antages at for alle fem modeller er der fundet den optimale model, som vil anvendes fremover. Disse modeller er baseret på antagelser præsenteret tidligere i projektet, som er understøttet af teoretiske undersøgelser, samt illustrationer, for at optimere modellerne. Dette bringer projektet videre til næste afsnit, hvilket er diskussionen, hvor der vil uddybes på analysen, samt en gennemgang af præcisionsmål.

6. Diskussion

Efterfulgt af analysen følger diskussionen, som søger at definitivt at svare på problemformuleringen, underspørgsmålene og gennemgå de dertilhørende hypoteser. Dette vil gøres på baggrund af den viden skabt i analysen, samt yderligere punkter der bliver fremhævet i denne del af projektet. Den overordnede struktur af diskussionen er som følger:

- Præcisionsmål fra metodeafsnittet.
- Sammenligning af modellerne.
- Gennemgang og besvarelse af underspørgsmål, hypoteser og problemformuleringen for projektet.
- Begrænsninger og perspektivering.
- Afrunding, samt delkonklusion.

Som set i overstående liste, vil der i følgende afsnit være en kort rekapitulering af præcisionsmålene fra metodeafsnittet, som dette projekt vil bruge til at evaluere modellerne. Hertil ses en overordnede gennemgang af de enkelte præcisionsmål, som blev udregnet i analysen, og hvordan disse vil anvendes til fortolkning af modellerne. Derudover udregnes de resterende præcisionsmål fra metodeafsnittet, så disse kan anvendes til sammenligning af modellerne. Dette projekt vægter højt, værdien i en bred vifte af væsentlige præcisionsmål, for at have flere forskellige synsvinkler på modellernes evne til at forudsige RNOA. En bred vifte af

præcisionsmål er også vigtigt for dette projekt for at se et nuanceret billede af modellernes styrker. Ydermere er dette også for at undgå en fejl fra dele af litteraturen, hvor en stor kritik er deres værker udelukkende vurdere modeller baseret på en eller få præcisionsmål (Chen & Yang, 2004). Dernæst vil de fem modeller sammenlignes, hvor dette vil være på baggrund af både præcisionsmål, samt yderligere illustrationer, som vil gennemgås og forklares i det tilsvarende afsnit. Efterfulgt af udregningerne af præcisionsmålene, vil projektet anvende den viden, til at besvarer underspørgsmål, hypoteser og problemformuleringen. Denne del af projektet er den essentielle del, som alle tidligere afsnit kulminerer til, og er efterfulgt af et begrænsnings- og perspektiveringsafsnit. Indholdet i begrænsnings- og perspektiveringsafsnittet, har til formål at give et indblik i, hvad der kunne tilføjes af værdi til dette projekt, hvis yderligere fordybning i emnet føles at være nødvendigt. Dette betyder, at essensen i begrænsnings- og perspektiveringsafsnittet er at få indblik i teori og metoder, der kan hjælpe med at understøtte undersøgelserne, som ligger bag problemformuleringen. Til sidst i diskussionen vil der være en afrunding af afsnittet, for at kunne opsummere de vigtigste punkter og lave en glidende overgang til konklusionen, som vil afrunde projektet.

6.1 Præcisionsmål for forecasts

På baggrund af viden skabt i analysen, er der fundet frem til fem overordnede modeller, med hver deres forskellige fremgangsmåde til at forudsige RNOA. Hertil er der antaget, at der i løbet af afsnittet, er fundet frem til den optimale model, gennem blandt andet gennemgang af de lineære antagelser, introduceret af Osborne og Waters (2002), samt Bowerman, O'Connell og Koehler (2005). Med antagelsen om, at der er fundet frem til de mest optimale versioner af modellerne, kan de forskellige præcisionsmål anvendes til at vurdere og sammenligne modellerne. Følgende er en liste over de forskellige præcisionsmål, samt sammenligningsmetoder der vil anvendes i efterfølgende afsnit:

- sMAPE.
- MSE.
- Interquartile range (IQR).
- Justerede R^2-værdi.
- Beskrivende statistik for modellerne.
- Residual standard errors relative størrelse (Residual std. Error over mean).
- Residualer udenfor estimaters/RNOA's interval.
- RNOA plottet mod residualer for de fem modeller.

6.2 Sammenligning af modellerne, baseret på præcisionsmålene

Baseret på resultaterne fra analysen, og præcisionsmålene fra metodeafsnittet, vil dette afsnit søge at sammenligne og vurdere de forskellige modeller. Hertil vil der fremhæves de forskellige modellers styrker, i forhold til hinanden, og hvilke modeller, som er bedrer til at beskrive RNOA, i forhold til andre. Nedenstående ses en tabel, hvor de forskellige præcisionsmål ses, udover RNOA plottet mod fejlled, som vil illustreres senere i dette afsnit. Her vises en blå markering, hvilket er det bedst vurderet resultat:

	sMAPE.	MSE.	IQR.	Justerede R^2-værdi.	Residual std. Error over mean.	Residualer udenfor estimater/RNOA's interval.
Model 1.	78,36 %	0,1728	0,1458	21,93 %	1,5734	431
Model 2.	80,13 %	0,1863	0,1536	15,84 %	1,634	865
Model 3.	198,53 %	34.427.930	108,4303	6,04 %	1,7263	36.885
Model 4.	78,20 %	0,1731	0,2269	a: 33,93 % b: 42,87 %	a: 1,1 b: 1,0701	300
Model 5.	80,13 %	372,6123	0,1952	a: 23,32 % b: 29,17 %	a: 11,1633 b: 10,3444	768

Tabel 18 - af egen inspiration.

Dernæst ses en lignende tabel, hvor den afhængige variable også er med for sammenligningens skyld, hvori den beskrivende statistik er:

	Mean.	Median.	Varians.	Standardafvigelse.	Minimumsværdi.	Maksimumsværdien.
RNOA.	0,2642	0,1706	0,2214	0,4705	-2,7021	4,7016
Model 1.	0,2643	0,2151	0,0486	0,2204	-1,1629	2,3838
Model 2.	0,2641	0,2439	0,0350	0,1872	-0,9785	1,8093
Model 3.	-492,1	-30,4	34.186.449	5.846,918	-436.180,4	63.841,2
Model 4.	0,2924	0,2434	0,0789	0,2808	-1,4576	2,9243
Model 5.	0,1270	0,0710	372,2841	19,2947	-1.469,3494	2.625,2345

Tabel 19 - af egen inspiration.

De følgende afsnit vil gennemgå indholdet i tabel 19 og tabel 20, hvor en celle for næsten alle præcisionsmål er markeret med blå, for at visualisere, hvilket modeller der performer bedst. Hertil er der forskellige grundlag for, hvordan disse er vurderet, hvilket vil gennemgås for det relevante præcisionsmål. Begrundelsen for at den justerede R^2-værdi og residual standard error over mean ikke har en markering, er besværet med at vurdere dem, da model 4 og 5 er opdelt i to simple lineære regressioner. Dette gør, at model 4 og 5 ikke giver en entydig scorer, for flere præcisionsmål.

6.2.1 sMAPE

I metodeafsnittet blev sMAPE beskrevet, og hvordan den vil bruges i dette projekt. Kort fortalt forklarer sMAPE, hvor meget at en model i gennemsnit er fra den aktuelle værdi. Hertil bliver der taget højde for de lave værdier i RNOA kan have en stor effekt på resultatet, hvilket har ledet projektet frem til Chen og Yang's (2000) version MAPE. Tages model 1, som eksempel, så går MAPE fra at være 401,09 % til at være 78,36 % for sMAPE. Dette er tydeligvis en meget stor forbedring for selve præcisionsmålet, og det understreger også, hvad blandt andet Makridakis og Hibon (2000) fremlagde i deres værk, angående herunder sMAPE. Det skal her understreges, at det stadig er høje værdier for sMAPE, på trods af den transformering, som blev gennemgået for selve præcisionsmålet. Hertil er det vigtigt at fremhæve synspunktet introduceret tidligere i dette projekt, hvor det i forhold til problemformuleringen, ikke nødvendigvis handler om at have en god model, men mere om der kan forbedres på benchmarket. Model 4 er den model, som performer bedst under sMAPE, men kun med en forbedring af 0,16 procentpoint, over benchmarket. Derudover ses det at model 2 og model 5 har præcis samme værdi for sMAPE, og er kun meget få procentpoint fra model 1 og model 4. Til sidst ses det igen, hvor dårligt model 3 performer under sMAPE, alt tyder på at model 3 ikke er en model der er god, og dette er mere eller mindre et gennemgående tema i projektet. Uden at have model 3 med i tankerne, så er de resterende fire modeller meget tæt på hinanden, i forhold til sMAPE, hvilket understreger tidligere argumenter om, fejlen ved, at udelukket anvende et enkelt præcisionsmål. Altså, havde der kun været valgt et præcisionsmål, er det meget muligt det enten have været MAPE (og derfor sMAPE) eller MSE, hvis der ses på litteraturen (Bruun, 2016) (Diebold & Lopez, 1996). Hertil vil der ikke kunne laves et dybdegående indblik i modellerne, da de relative tætte værdier for sMAPE, gør det svært at konkludere på. Model 4 er snævert den bedst performende model for sMAPE, hvilket betyder den i gennemsnit har den korteste afstand fra dens sande værdi af residualer, i forhold til tendenslinjen.

6.2.2 MSE

Diebold og Lopez (1996) påstår at MSE er den mest anvendte præcisionsmål i litteraturen, og i forlængelse af MSE er der præcisionsmålet RMSE. RMSE er ”root mean squared error”, dette betyder at RMSE er gennemsnittet af fejlleddene, og på grund af dens lighed til sMAPE, er denne ikke taget med.

For modellerne i dette projekt er det model 1 der performer bedst baseret på MSE. MSE skal have den mindst mulige værdi, og selvom model 4 har en værdi tæt på, så er model 1 en lille forbedring. Set i forhold til model 1 og model 4, så har model 2 også en meget fornuftig MSE, men her stopper det. Model 3 og model 5 har meget dårlige MSE-værdier, hvor værdierne er meget høje og her begynder der at vise sig en bekymrende trend, når der tages præcisionsmålene for modellerne i denne rækkefælge. Dette vil der vendes tilbage til i næste afsnit af diskussionen, men det ser ud til, at model 3 og model 5 har svært ved at beskrive RNOA, ud fra en række præcisionsmål.

6.2.3 IQR

I tabel 19 er der markeret den bedste performende model under IQR, til at være model 4. Dette er på baggrund af RNOA’s værdi for IQR, som er på 0,3. Her vurderes det, at for IQR, er det vigtigste, at modellen er så tæt, som muligt på den sande IQR, for RNOA. Om IQR er høj eller lav, det betyder ikke så meget for dette projekt i første omgang, da der skal sammenlignes med det interval, som modellerne søger at efterligne. IQR viser, hvor de midterste 50 % af estimaterne befinder sig, altså et meget stort interval betyder at dataet er spredt ud over midten, hvor et snævert interval peger på en stor vægt af værdier, indenfor et lille interval. Dette præcisionsmål søger at vurdere modeller, i forhold til hvordan de estimerede værdier er distribueret. Tilsammen med præcisionsmålet for residualer udenfor RNOA’s og estimaternes interval, giver IQR et billede af, om modellerne kan håndtere intervallet for RNOA, og derfor findes den model der kan håndtere de ydrer værdier, samt fokusset på de 50 % af værdier der befinder sig i ”midten” af dataet. Derudover ses det også i tabel 19, at model 1 har et meget snævert interval, hvor 50 % af observationerne, ligger indenfor et spænd af 0,1458.

Igen understreges der her, at model 3 performer utroligt dårligt, men interessant nok, ses det at model 5 har en væsentlig forbedring i forhold til model 3. Dette tyder på, at i nogle tilfælde, hjælper det præcisionen, at der skiftes fremgangsmåde, når der forecasts OPPL og NOA, til at finde RNOA.

6.2.4 Justerede R^2-værdi

Dette præcisionsmål er ikke så brugbart alene, hvilket blandt andet er problematisk på grund af fremgangsmåden, der blev introduceret af Plenborg, Petersen og Kinserdal (2017). Dette er på grund af model 4 og 5 i teorien er delt op i to seperate simple lineære regressioner. Derfor er der ikke i tabel 19, nogen markering for den model, som performer bedst. På trods af dette, så er værdierne for alle modeller, udover model 3, relativt fornuftige. Dette er på baggrund af Soliman (2008), hvor hans bedste model er på en justerede R^2-værdi på 19,8 %. Dette tyder på, at den justerede R^2-værdi, generelt er lav i dette felt, hvilket er logisk til en vis grad, da der er tale om forecasts af regnskabstal. For model 4 og 5, ses fire værdier, der er relativt høje, hvilket tyder på at RNOA er udregnet på en baggrund af modeller der er relativt gode til at forklarer variationen i den afhængige variable.

6.2.5 Residual standard error over mean

Dette præcisionsmål anvendes til at kunne vise, hvor stor en del residual standard error for en model er, i forhold til middelværdien af den afhængige variable. Dette præcisionmål hjælper med at illustrere, hvor stor en værdi residual standard error er, for model 5, når den afhængige variable for model 5a og 5b kan være meget høje. Ydermere betyder dette at det kan være svært at sammenligne residual standard error mellem model 5a og 5b, med model 1. For dette præcisionsmål er der igen ikke markeret en model til at være den mest optimale, da det er besværligt med ændringen i fremgangsmådden for model 4 og model 5. For model 1, 2 og 3, som alle er simple lineære regressioner, er værdierne næsten ens, og for model 4 ses det, at residual standard error ikke er meget højere end mean for den afhængige variable. For model 5a og 5b, er der residual standard errors, der er meget højrere end mean, hvilket betyder at modellen i gennemsnit afviger fra den sande værdi af den afhængige variable med henholdsvis 11,1633 og 10,3444 gange større værdi end middelværdien.

6.2.6 Residualer udenfor intervallet for estimattet/RNOA

Dette præcisionsmål har to nuancer, hvor der er en forskel på fortolkningen fra model 1, 2 og 4, mod model 3 og 5. Bemærk her den fundamentale forskel på de to grupper, dette præcisionsmål kan opdeles i. Hertil er det for modeller, hvis estimater der har grænser for minimus- og maksimumsværdi indenfor RNOA's interval, samt for modeller med estimater udenfor dette interval. Dette præcisionsmål er, som det tidligere præcisionsmål, konstrueret i forbindelse med dette projekt, for at kunne få et bestemt indblik i modellernes estimater. Hertil søger dette præcisionsmål at finde de residualer, udenfor modellernes interval for estimaterne, for at se, hvilke modeller har problemer med at vurdere halerne for den afhængige variable. For dette præcisionsmål er model

4 den klart bedste, hvilket betyder model 4 har et interval der passer bedst til de aktuelle værdier for RNOA's interval. Dernæst er model 1 den anden bedste under dette præcisionsmål, og bemærk at model 5 beskriver bedre intervallet for den afhængige variable RNOA, end model 2. Til sidst er der model 3, som er helt forfældelig, hvilket er berørt tidligere i opgaven, og de fleste af estimaterne, er højere eller lavere end RNOA's minimums- og maksimumsværdier.

6.2.7 Beskrivende statistik

For det beskrivende statistik vil der ikke blivet gået for meget i dybden med de forskellige modeller, men dertil fremhæves de modeller, som performer bedst for de forskellige begreber. For mean, median, minimums- og maksimumsværdien, er det ikke deres størrelse der er indikation for den model, som performer bedst på disse begreber, men modellernes relation til den aktuelle værdi af RNOA. Ydermere ses det for variansen og standard afvigelse, at der her er fokus på den model, med de mindste værdi, i disse begreber. Dette betyder, at for eksempel for mean, er der taget højde for den model, som er tættest på mean for RNOA. For dette projekt er det ikke nødvendigvis interressant at have en høj mean eller median for nogen modeller, men der er større interesse for modeller der bedst repræsentere RNOA's aktuelle værdier. Derudover når det kommer til varians og standard afvigelse, omhandler det den model med den mindste risiko, set på disse to begreber, og derfor vælges der her den model, som har den mindste værdi på dette præcisionsmål. Model 1 og 2 er den model, som er tættest på RNOA's mean, samtidigt med at model 2 også har de mindste værdier for varians, og derfor også standard afvigelse. Model 1 har ydermere den median, som er tættest på median for RNOA. Model 4 er den model, som er tættest på det interval, som gælder for RNOA's mindste og største værdi, hvilket giver god mening set på baggrund af det tidligere afsnit. For det beskrivende statistik er det svært at vælge en model, som er bedre end de andre, selvom vurderingen for den bedste model er spredt, så udelukker det ikke modeller for nogen variabler. Dette betyder, at selvom model 2 har de bedste værdier for varians og standard afvigelse, betyder det ikke at model 4 er svag på disse begreber. Derfor vil der på baggrund af dette, ikke vælges en model der er "bedst".

6.2.8 RNOA plottet mod residualer for de fem modeller

Dette afsnit vil illustrere fem plots, hvor der vil visualiseres de fem modellers residualer, i forhold til RNOA. Dette vil give en beskrivelse af, hvordan residualerne er plottet i forhold til udviklinge af RNOA. Her er RNOA repræsenteret på x-aksen og residualerne på y-aksen. Residualerne er selvfølgelig den aktuelle værdi af RNOA, fratrukket modellens estimat. Dette betyder at plottene beskrevet i dette afsnit skal ligge på en lige linje med en værdi af nul, på y-aksen. Ligger en residual med en værdi af nul, så er modellens estimat og den akutelle

værdi for RNOA ens, for den givende værdi af RNOA. Illustrationer af disse plots ses i bilag 10 og 11, hvor bilag 10 fokusere på data fra hele sættet, og bilag 11 er det førhen anvendte subsample. Ydermere er subsamplet i bilag 11 anvendt, igen på grund af mængden af observationer, gør det besværligt at se vægten af den enkelte observation. For både bilag 10 og bilag 11 gælder det, at model 3 og 5 har samme problemer, som tidligere. Disse problemer er med længden af y-aksen, hvor de meget esktreme værdier for modellerne, som forlænger y-aksen. Bemærk at i bilag 11 er der to ekstra versioner af model 5 og 3. Dette er af samme grund, som der var tidligere i projektet, hvor residualerne har så ekstreme værdier, at her bliver y-aksen manipuleret, så det ligner der er mange residualer, der ligger omkring nul. Det skal derfor understreges, at når der bliver tilpasset model 3 og 5 en ens længde på y-aksen, forsvinder en helt del residualer for model 3, og få for model 5. For model 3 befinder der sig 90 residualer over en værdi af 3, og 4 residualer under værdien -0,6. Dette er to værdier, som er taget umiddelbart så y-aksen samtidigt passer bedre med de andre modeller, og samtidigt illustrerer, hvor ringe model 3 og 5 er. For model 5 er der 2 residualer over værdien 3, og 7 residualer under værdien -0,6. Igen er model 5 en forbedring på model 3, men igen betyder det ikke at model 5 er nogen speciel god model. Model 1, 2 og 4 ser umiddelbart ud til at have en fornuftig fordeling af deres residualer, med en stor del af disse residualer, der er omkring en værdi af nul. På trods af de meget ekstreme værdier i model 5, er der stadig en del observationer omkring nul, men modellen er ekstrem risikabel, med nogle af de helt yderligere residualer.

6.3 Gennemgang af problemformuleringen, underspørgsmål og hypoteser

Gennemgangen af problemformuleringen, underspørgsmålene og hypoteserne, vil foregå først med besvarelse af underspørgsmålene, hvilket leder projektet videre til hypoteserne. På baggrund af besvarelsen af underspørgsmålene og hypoteserne, vil problemformuleringen besvares. Disse besvarelser er baseret på viden skabt i analysen, samt resultaterne fra præcisionsmålene.

6.3.1 Underspørgsmål 1: Simple dekomponering af RNOA

Underspørgsmål 1, ses i afsnit 1.2 og er som følger:

1. Er det en fordel, med fokus på statistisk præcision, at forecaste ATO og PM til at forklare RNOA, i modsætning til udelukket at forecaste RNOA?

Dette underspørgsmål søger at teste, om model 1 kan forbedres på, ved at anvende model 2. Fairfield og Yohn (2001) mener ikke at model 2 burde forbedre modellen, og Soliman (2008), mener at den burde kunne forbedre den statistiske præcision. Her er der en konflikt i litteraturen, hvor dette projekt vil tage et basalt blik på de to synspunkter. Soliman (2008) påstår at begge variabler er signifikante, hvor Fairfield og Yohn (2001)

viser at PM ikke er signifikant. Dette projekt har gennem analysen testet påstanden fra Fairfield og Yohn (2001), og fundet en signifikant model for model 2. Analysen konkluderer at PM og ATO begge er signifikante, både gennem p-værdien, men også i forhold til VIF-testen. Der er derfor ikke tale om insignifikant uafhængige variabler, og samtidigt ikke problemer med multikollinearitet. Soliman (2008) har derfor ret i hans påstand, at PM og ATO begge er signifikante, men i forhold til underspørgsmål 1, så er model 1 og 2 meget tætte når det gælder præcisionsmål. Model 1 performer bedst på sMAPE, MSE, justerede R^2-værdi, residual standard error over mean, median, minimums- og maksimumsværdier og residualer udenfor RNOA's interval. Dertil performer model 2 bedst på IQR, mean, varians og standardafvigelse. Model 2 har derfor en bedre varians og standardafvigelse end model 1, og dens IQR er tættere den for RNOA, men de resterende præcisionsmål performer den, som sagt ringere end model 1. Hertil er mange af præcisionsmålene relativt tætte for model 1 og 2, men skal der vælges en model for dette underspørgsmål, vurderes det at være model 1. Dette betyder at der ikke kan forbedres på benchmark modellen, ved hjælp af en simple dekomponering af RNOA, i en multipel lineær regression.

6.3.2 Underspørgsmål 2: To-lags dekomponering af RNOA

Underspørgsmål 2 søger at teste, om model 2 kan forbedres ved at dekomponere PM og ATO i variablerne OPPL, GP og NOA. Dette sker ved at svare på følgende spørgsmål:

2. Vil en yderligere dekomponering af ATO, samt PM kunne forbedre forecastet af regnskabstallene?

Dette underspørgsmål omhandler model 3, og på baggrund af svaret fra underspøgsmål 1, så søger dette underspørgsmål at teste, om model 3 er bedre end model 1, og derfor også model 2. Både i analysen og tidligere afsnit i diskussionen, er der gennemgået, hvor dårlig model 3 rent faktisk er. For alle præcisionsmål i tabel 19 og tabel 20, performer model 3 langt værre end model 2 og model 1. Det er kun for residual standard error over mean, at denne model har en relativ fornuftig værdi, men hertil er der tilknyttet et stort problem. Selvom residual standard error er højt for denne model, så er mean for model 3 også alt for højt. Dette præcisionsmål er derfor meget upålidelig for denne model, hvilket ikke betyder så meget, da det er tydeligt på de resterende præcisionsmål, at model 3 er en dårlig model. Der kan derfor ikke forbedres, hverken på model 1 eller model 2, ved at lave en dekomponeringen af PM og ATO i en multipel lineær regression. Model 3 er dermed forkastet og må direkte frarådes.

6.3.3 Underspørgsmål 3: Den alternative fremgangsmåde

Model 1 vurderes til at være bedre end model 2, og model 3 vurderes til at være den klart dårligste model. Den alternative fremgangsmåde fra Plenborg, Petersen og Kinserdal (2017). Underspørgsmål 3 tager derfor et nyt synspunkt på de tre første modeller, og forecaster derfor de value drivers hver for sig, og derefter bliver RNOA udregnet. Tidligere i projektet blev underspørgsmål 3 fremvist, og den er som følger:

3. Vil den statistiske præcision forbedres, hvis der bliver forecastet, udelukket på value drivers, i forhold til underspørgsmål 1 og 2?

Model 5 er en stor forbedring på model 3, for alle præcisionsmål, men som nævnt tidligere i dette projekt, er dette ikke en indikation for at model 5 er en god model. For at uddybe lidt på denne påstand, sammenlignes model 5 med model 1 og 2. Model 5 har præcis den samme sMAPE, som model 2, men en meget højere MSE. På trods af dette er IQR for model 5 bedre end model 2, samt præcisionsmålet, hvor antallet af estimater for modellerne over/undervurdere det aktuelle interval for RNOA. Dette giver nogle ekstreme værdier, hvor der blandt andet kan ses en RNOA på 2.625,2345, hvilket er en urealistisk høj værdi. På baggrund af præcisionsmålene, vurderes model 5 til at være en stor forbedring på model 3, men stadig en ringere statistisk model, end model 2 og model 1.

Model 4 er lidt mere interessant, hvor denne model har det interval, som bedst passer til intervallet for RNOA. Der er derfor færrest værdier for model 4, hvor der over/undervurderes på de yderligere værdier, samt det mest passende IQR og dertil de tætteste helt konkrete minimums- og maksimumsværdier. Derudover har model 4 også den bedste performance på sMAPE, hvilket betyder at model 4 i gennemsnit har den mindste afvigelse af dens forecasts, i relation til RNOA. Her skal det dog bemærkes, at selvom model 4 har den bedste performance på sMAPE, så er den meget tæt på værdien for model 1. Samme argument kan derfor også anvendes til MSE, hvor model 4 ikke performer bedst, men er kun overgået af model 1, og med en meget lille værdi. Model 4 kan derfor vurderes, som at være en bedre model, end model 1, ud fra flere forskellige præcisionsmål, og er samtidigt også en forbedring på model 2. Ydermere signalerer dette, at den alternative fremgangsmåde med at forecaste value drivers, og dernæst udregne RNOA på baggrund af dette, vil forbedre modellerne.

6.3.4 Underspørgsmål 4: Bedst performende model

Sidste underspørgsmål omhandler den model, som dette projekt vurderer, som værende den model der performer bedst. Hertil er underspørgsmål 4 ligetil, og er som følger:

4. Hvis dette projekt udelukket skulle vælge en model, hvilken ville så ses, som at være den bedste?

Ifølge Chen og Yang (2004) er det bedste forecast valgt ud fra en subjektiv vurdering af de forskellige præcisionsmål. De forskellige præcisionsmål i denne opgave, giver en objektiv vurdering af den model, som performer bedst på det enkelte mål, og derefter tages et subjektivt valg, for hvilken model er bedst. Det er op til individet der skal vurdere modellerne, for at markere en model, som "bedst", og dertil kan flere forskellige biases være med i overvejelsesprocessen. Når det er sagt, så er det også tydeligt, at hvis en model er bedst på alle valgte præcisionsmålene, så bliver den model den bedste, rent objektivt set, baseret på de subjektive valgte præcisionsmål. Baseret på de tidligere underspøgsmål, tabel 19, samt 20 og bilag 10 og 11, ses det at model 4 performer bedst på næsten alle præcisionsmål der kan sammenlignes. Her har model 4 mindst sMAPE, næstmindst MSE, og det mest beskrivende interval i forhold til den aktuelle værdi af RNOA. Derfor ses model 4, som at være den bedste model for der er fremvist i dette projekt.

6.3.5 Hypotese 1

For alle hypoteser, gælder det at de ligger tæt op ad underspørgsmålene, så der vil ikke blive gået for meget i dybden hertil. Nul-hypotesen for hypotese 1 er:

"H_0: *En simple DuPont dekomponering af RNOA vil kunne forbedre den statiske præcision, til at forecaste nøgletallet."*

Hertil ses den overordnede problemstilling i underspørgsmål 1, og der vises her, at en simple DuPont dekomponering af RNOA ikke er nok til at forbedre på benchmarket. Derfor forkastes nul-hypotesen, og den alternativ hypotese accepteres, hvilket er at et forecast af RNOA ikke forbedres, udelukket ved at dekomponere til PM og ATO. Denne hypotese er inspireret af Fairfield og Yohn (2001), samt Soliman (2008), hvor Soliman (2008) kritiserer Fairfield og Yohn (2001), for at påstå PM ikke er signifikant for denne model. Hertil vises at PM er signifikant for model 2, men ikke at model 2 er en forbedring af benchmarket, som Soliman (2008) viser.

6.3.6 Hypotese 2

Hypotese 2 lyder, som følger:

"H_0: *En dekomponering i to lag, er mere statistisk præcis, end en simple dekomponering, når det gælder forecasting af RNOA."*

Denne hypotese søger at teste, om model 2 kan forbedres, ved at dekomponere PM og ATO. Denne tankegang er fundamentet for model 3. Dog viser analysen og besvarelsen af underspørgsmål 2, at en yderligere dekomponering ikke hjælper med den statistiske præcision. Derimod skaber det en meget ringere model, som ikke kan anbefales. Nul-hypotesen for hypotese 2 forkastes, og den alternative hypotese accepteres, hvilket er at en dekomponering i to lag, ikke forbedrer den statiske præcision for modellerne.

6.3.7 Hypotese 3

Dette projekt vælger at gå i retning af Plenborg, Petersen og Kinserdal (2017), når det gælder en yderligere undersøgelse af potentiel forbedring af de valgte præcisionsmål. Denne alternative fremgangsmåde skaber følgende hypotese:

"H_0: *Et bedre resultat for forecastet af en simple DuPont dekomponering af RNOA opnås, ved hjælp af den alternative fremgangsmåde."*

Hypotese 3 bliver besvaret i første del af underspørgsmål 3, og dette er baggrunden for model 4. På baggrund af besvarelsen for underspørgsmål 3 og 4, vurderes model 4 til at være den bedste i dette projekt, og derfor accepteres nul-hypotesen for hypotese 3.

6.3.8 Hypotese 4

Sidste hypotese for dette projekt lyder som følger:

"H_0: *Kan en bedre model opnås, ved hjælp af fremgangsmåden fra hypotese 3, pålagt hypotese 2."*

Hypotese 4 skaber model 5, som derfor er en forlængelse af model 3, hvor model 4 er en forlængelsen af model 2. Hypotese 4 forkastes, på baggrund af, at model 5 er en forbedring af model 3, men ikke en forbedring af model 4. For denne nul-hypotese kunne accepteres, skulle model 5, være en forbedring af model 4, hvilket er vist i dette projekt, til ikke at være sandt.

6.3.9 Problemformuleringen

Problemformuleringen for dette projekt vil i dette afsnit besvares, på baggrund af besvarelserne fra underspørgsmålene, samt hypoteserne. Problemformuleringen er følgende:

Forbedres præcisionen af et forecast af regnskabets nøgletal, hvis der bliver fokuseret på de variabler, som nøgletalende bliver dekomponeret i?

Problemformuleringen er relativ åben, hvilket hjælper med at understrege påstanden fra Chen og Yang (2004), hvor det er subjektivt, i forhold til de præcisionsmål der er valgt. Ydermere er der også to forskellige muligheder for denne problemformulering, hvor de forskellige fremgangsmåder for de fem modeller, skal tages i betragtning. På baggrund af underspørgsmålene og hypoteserne, ses det at der ikke forbedres på den statistiske præcision, ved simpelt at fokusere på dekomponering af RNOA, og derfor fokuseres der på den alternative fremgangsmåde. For begge modeller, hvor denne fremgangsmåde anvendes, forbedres modellerne i forhold til den model de er en forlængelse af. Hertil har model 5 forbedret en del på en dårlig model, men det er model 4 der er interessant. Model 4 er vurderet i dette projekt, som den bedste model, og den forbedrer på model 2, som ligger omkring niveauet af model 1.

Resultatet af de tidligere afsnit, viser at der kan forbedres på den statistiske præcision af et forecast af regnskabets nøgletal (RNOA), når dette nøgletal bliver dekomponeret. Hertil skal det bemærkes, at for at få en forbedret model, skal dekomponeringen følge den alternative fremgangsmåde præsenteret af Plenborg, Petersen og Kinserdal (2017).

6.4 Begrænsning og perspektivering

Begrænsnings- og perspektiveringsafsnittet søger at give eksempler på, hvilke emner der er interessante i forhold til at uddybe dette projekt. Dette er på baggrund af den tids- og pladsmæssige begrænsninger, der er for dette projekt. Ydermere vil dette afsnit forsøge at holde sig kortfattet, af samme grund, som dette afsnit er en del af projektet.

6.4.1 Mean reversion

Mean reversion er et begreb der beskriver den tendens, som blandt andet RNOA har, hvor de ekstreme værdier for RNOA, søger at nærme sig middelværdien, over en given tidshorisont (Yohn, 2018). For eksempel, virksomheder med relativ høj RNOA, har en tendens til at have en aftagende værdi af RNOA i fremtiden. Måden dette begreb anvendes på, er når en virksomhed har en overnormal værdi af RNOA i forhold til industriens middelværdi, diktere mean reversion at RNOA falder i fremtiden. Dette er et interessant begreb at introducere i dette projekt, for at nuancere projektet.

6.4.2 Yderligere præcisionsmål

For et projekt som dette, kan der næsten altid uddybes i den statistiske analyse, hvor der blandt andet kan anvendes flere præcisionsmål. Dette projekt har valgt præcisionsmål ud fra Chen og Yang (2004), der fremhæver nødvendigheden i at have mange præcisionsmål, der dækker flere forskellige synspunkter. Der

kunne derfor gøres analysen endnu mere solid, ved at tage flere præcisionsmål med. Et eksempel på et yderligere præcisionsmål er den tilgang Fairfield og Yohn (2001) har, hvor de forskellige modeller, bliver sammenlignet i forhold til hinandens residual error. Hertil bliver der undersøgt, hvordan de forskellige modeller forbedrer på den sammenlignet model.

6.4.3 Flere modeller

Dette projekt kunne uddybe på problemformuleringen, ved at tage endnu flere modeller med i analysen. En af mulighederne, er at prøve en "tre-lags DuPont dekomponering", hvor OPPL, NOA og GP bliver dekomponeret, og dernæst kan der forecastes på disse værdier. Hertil kan der igen testes, om Plenborg, Petersen og Kinserdal (2017)'s fremgangsmåde til forecasting, kan forbedre på disse modeller. En anden måde dette projekt kunne inkorporere flere modeller på, er at tage fremgangsmåde fra blandt andet Fairfield og Yohn (2001), hvor der blandt andet bliver brugt ændringen af RNOA, som den afhængige variable. Tages fremgangsmåden anvendt i model 4 og 5, og anvendes på modellerne fra Fairfield og Yohn (2001), kunne det også give et interessant udgangspunkt for en ny diskussion.

6.4.4 Markeds-, industri- og virksomhedsfaktorer

I litteraturstudiet for dette projekt blev der fundet en interessant artikel af Jackson, Plumlee og Rountree (2018), hvor der undersøges forskellige faktorer der kan påvirke en virksomheds profitabilitet. Disse faktorer er bestående af et markeds-, industri- og virksomhedskomponent (Jackson, Plumlee, & Rountree, 2018). Hertil konkluderer Jackson, Plumlee og Rountree (2018), at der kan forbedres på et forecast, ved at tilføje komponenter for markedet, industrien og virksomheden. Dog konkluderes også, at denne metode er meget dataintensiv og der er høj risiko for potentiel støj. Dette er specielt interessant for dette projekt, da der ikke er taget højde for disse tre komponenter. Derfor findes muligheden for, at denne metode kan tilføje værdi til de fem modeller.

6.5 Delkonklusion: diskussionen

Diskussionen for dette projekt fokuserede på at sammenligne og vurdere de forskellige modeller gennemgået i analysen. Samtidigt med dette, sætter diskussionen op til, at underspørgsmål, hypoteser og problemformuleringen kan besvares, på baggrund af præcisionsmålene. Hertil kan dette projekt konkludere, at model 1 performer marginalt bedre end model 2. Det interessante i diskussionen, er når model 4 og 5 vurderes og sammenlignes med de øvrige modeller. Model 5 forbedrer meget på den skuffende model 3, og selvom model 5 ikke er perfekt, er det imponerende, hvordan den nye fremgangsmåde, ændrer så meget for

modellen. Dernæst ses det, at model 2 er relativ tæt på model 1 for en hel del af præcisionsmålene, og introduktionen af model 4, kan forbedre på model 2. Dette fremhæver model 4, som den bedste model i dette projekt, på baggrund af underspørgsmål 4. Besvarelserne af underspørgsmålene og hypoteserne, er fundamentet for problemformuleringen. Ud fra diskussionen og analysen, ses det at der kan forbedres på forecastet af RNOA, gennem både dekomponering af den afhængige variable (RNOA), og dernæst med fokus på forecasts af disse value drivers.

7. Konklusion

Forecasting af regnskabstal er nødvendigt for en række forskellige funktioner for virksomheder. Her anvender analytikere disse forecasts, til at lave et velbegrundet gæt på, hvordan fremtidens regnskabstal ser ud. Disse resultater anvendes til en række forskellige analyser, som blandt andet budgettering, investering, kreditvurderinger osv. Dette kan let være meget betydelige beslutninger, der derfor tages på baggrund af risikable modeller. Derfor er det vigtigt for individet der anvender disse metoder, at tallene er så præcise, som muligt. På trods af dette, er der utrolig lidt viden og litteratur på dette område, i forhold til hvor vigtigt nogle af disse beslutninger er (Fairfield & Yohn, 2001). Der kan være flere forskellig grunde til, hvorfor der ikke er en stor mængde af litteratur på dette emne. Dette kan muligvis være på grund af anvendelsen i praksis ikke dikterer, at det er nødvendigt. Altså at i praksis anvendes der basale metoder, og der bliver ikke fokuseret så meget på at gå i dybden med disse analyser. Problemformuleringen for dette projekt er formet på baggrund af at gå i dybden med en benchmark-model, og undersøger om der kan forbedres på denne model. For at bedre kunne besvare problemformuleringen, opsættes hypoteser, som besvares gennem underspørgsmål. Disse hypoteser er inspireret af litteraturen og problemformuleringen selv.

For dette projekt tages der udgangspunkt i nøgletallet RNOA, hvilket bliver anvendt, som baggrund for de fem opstillede modeller. RNOA er valgt, som det overordnede nøgletal, blandt andet på baggrund af det tilgængelige data, samt dens repræsentation i litteraturen, på dette emne. Disse fem modeller er inspireret af Fairfield og Yohn (2001), Soliman (2008) og Plenborg, Petersen og Kinserdal (2017), og søger at svare problemformuleringen, fra flere forskellige vinkler. Før disse modeller kan gennemgås i det anvendte dataprogram, RStudios, er det vigtigt det rigtige data findes, bearbejdes og dertil undersøges, hvilke variabler kan bruges til at udregne RNOA. Datasættet starter på omkring 2,7 millioner observationer, og efter alle problemer hertil er korrigeret for, er der ca. 39.500 observationer tilbage. Der er derfor mange observationer i dette datasæt, som ikke kan anvendes i dette projekt. Hertil er det blandt andet urealistiske år, ikke udfyldte

værdier for nogle variabler, og ikke sammenhængende årstal. Efter en grundig reduktion af dataet, ses slutpopulationen, til at være mere en rigelig, til formålet for dette projekt.

For at kunne designe de rigtige modeller til at besvare grundigt på problemformuleringen, er der introduceret en alternativ fremgangsmåde, for at kunne forecaste value drivers. Denne alternative fremgangsmåde sammenlignes med de mere "traditionelle" metoder at forecaste den dekomponeret version af RNOA. Hertil er det blandt andet tankegangen fra Soliman (2008) og Fairfield og Yohn (2001), der anvendes.

De fem modeller dette projekt opstiller på baggrund af litteraturen, bliver analyseret og forberedt, gennem de lineære antagelser, for at sikrer at modellerne kan sammenlignes på samme grundlag. Disse antagelser ses opfyldt for alle modeller, og derfor antages, at de mest optimale modeller er fundet, for hver af de fem modeller. Efterfulgt af, at de fem modeller har deres optimale form, skal de kunne sammenlignes på flere forskellige præcisionsmål. Hertil tages der baggrund i viden fra Chen og Yang (2004), der fremhæver fordelen ved at have flere forskellige præcisionsmål. Derfor er der valgt præcisionsmål, som forsøger at fremhæve forholdet mellem residualerne og den aktuelle værdi af RNOA. Derudover er der også præcisionsmål, der fokuserer mere på intervallet for modellernes estimeret værdier, for at give et mere nuanceret indblik i modellerne. Med en bred vifte af præcisionsmål, kan dette projekt give et mere pålideligt svar på problemformuleringen.

Model 1 er derfor modellen dette projekt søger at forbedre på. Analysen viser at model 1 er relativ god i forhold til de øvrige modeller, målt på de gældende præcisionsmål. Dette ses gennem gennemgangen af præcisionsmålene. På baggrund af disse præcisionsmål, vurderes model 2, ikke at være en forbedring på model 1, men udelukket med et marginalt forhold. Model 3 er dernæst introduceret, og den er uden tvivl den ringeste model i dette projekt. For projektet betyder dette, at svaret på problemformuleringen indtil dette punkt, er at der ikke kan forbedres på forecastet, ved at dekomponere RNOA. En helt ny tankegang bliver dertil introduceret, hvor der bliver fokuseret på, at forecaste de enkelte value drivers hver for sig, og dernæst kalkulere RNOA ud fra disse simple lineære regressioner. Denne alternative fremgangsmåde er meget interessant, da den formår at forbedre på de to modeller den pålægges. Altså er model 4 en direkte forbedring af model 2, og model 5 er en direkte forbedring af model 3. Model 2 var, som sagt, ikke vurderet til at kunne forbedre på model 1, men med introduktionen af model 4, ses en klar forbedring på model 1, ud fra en række præcisionsmål. Hertil vurderes model 4 til at være den bedste model i dette projekt, og derfor kan der forbedres på et forecast af RNOA, ved at dekomponere det til PM og ATO, og forecaste dem hver for sig. Dog er

det bemærkelsesværdigt at model 5 er en meget stor forbedring på model 3, selvom model 5 ikke er vurderet til at være bedre end model 1.

Problemformuleringen for dette projekt har derfor lidt af en nuance, når det kommer til at besvare den. Det er vurderet at en simple DuPont dekomponering af RNOA baseret på multipel lineær regression ikke forøger den statistiske præcision. Brydes denne multipel lineære regression op i to lineære regressioner, med fokus på dens value drivers, ses en forbedring af den statistiske præcision, i forhold til benchmarket. Dette betyder derfor, at den statistiske præcision af et forecast af RNOA, kan forbedres ved at fokusere på de variabler, som RNOA dekomponeres i.

8. Bibliografi

Bowerman, B. L., O'Connell, R. T., & Koehler, A. B. (2005). Part II: Regression Analysis. I B. L. Bowerman, R. T. O'Connell, & A. B. Koehler, *Forecasting, Time Series and Regression* (s. 79-279). Curt Hinrichs.

Bruun, M. (2016). Using Time-series Properties of Financial Ratios to Forecast Earnings. I *Eassys on Earnings Predictability* (s. 40-42). Frederiksberg.

Chen, Z., & Yang, Y. (14. Marts 2004). Assessing Forecast Accuracy Measures.

Cox, N. J. (25. Juli 2007). *Transformation: an introduction*. Hentet fra bc.edu: http://fmwww.bc.edu/repec/bocode/t/transint.html

Diebold, F. X. (1.. December 1993). On the Limitations of Comparing Mean. *Journal of Forecasting*, s. 641-642.

Diebold, F. X., & Lopez, J. A. (1996). Forecast Evaluation and Combination. *Handbook of Statistics Vol. 14*, s. 241-268.

Ekwaru, J. P., & Veugelers, P. J. (26.. Marts 2018). The Overlooked Importance of Constants Added in Log Transformation of Independent Variables with Zero Values: A Proposed Approach for Determining an Optimal Constant. *Statistics in Biopharmaceutical Research* , s. 26-29.

Fairfield, P. M., & Yohn, T. L. (2001). Using Asset Turnover and Profit Margin to Forecast Changes in Profitability. *Review of Accounting Studies*, s. 16.

Fairfield, P. M., Yohn, T. L., & Sweeney, R. J. (Juli 1996). Accounting Classification and the Predictive Content of Earnings. *The Accounting Review*, s. 337-355.

Jackson, A. B., Plumlee, M. A., & Rountree, B. R. (6. Juni 2018). *Decomposing the market, industry and firm components of profitability: implications for forecasts of profitability.* Hentet fra Springer Science+Business Media, LLC.

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting 9*, s. 527-529.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting 16*, s. 451-476.

Orbis. (2019). Hentet fra Bureau van Dijk: https://www.bvdinfo.com/en-gb/our-products/data/international/orbis

Osborne, J. W., & Waters, E. (Januar 2002). Four Assumptions of Multiple Regression That Researchers Should Always Test. *Practical Assessment*, s. 5. Hentet fra https://www.researchgate.net/profile/Jason_Osborne2/publication/234616195_Four_Assumptions_of_Multiple_Regression_That_Researchers_Should_Always_Test/links/59302555aca272fc55e144da/Four-Assumptions-of-Multiple-Regression-That-Researchers-Should-Always-Tes

Pardoe, I. (5. august 2019). *12.4 Detecting Multicollinearity Using Variance Inflation Factors*. Hentet fra Penn State Science: https://newonlinecourses.science.psu.edu/stat501/node/347/

Penman, S. H., & Nissim, D. (2001). Ratio Analysis and Equity Valuation: From Research to Practice. *Review of Accounting Studies*, s. 116.

Plenborg, T., Petersen, C., & Kinserdal, F. (2017). Forecasting. I T. Plenborg, C. Petersen, & F. Kinserdal, *Financial Statement Analysis* (s. 251-294). Fagbokforlaget.

Soliman, M. T. (2004). *Using Industry-Adjusted DuPont Analysis to Predict Future Profitability.* Stanford.

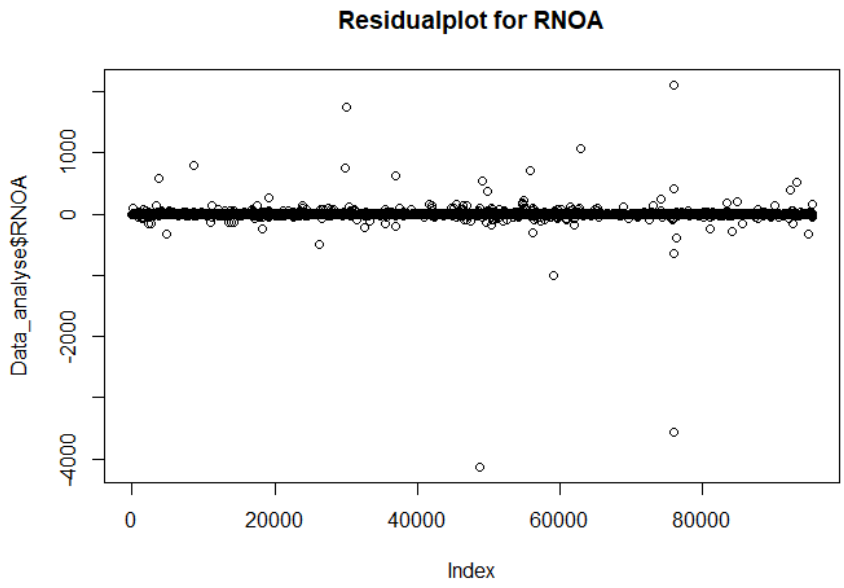
Soliman, M. T. (2008). The Use of DuPont Analysis by Market Participants. *The Accounting Review*, s. 823-853.

Whitley, E., & Ball, J. (29. November 2001). *Statistics review 1: Presenting and summarising data.* Hentet 3. September 2019 fra biomedcentral: https://ccforum.biomedcentral.com/track/pdf/10.1186/cc1455

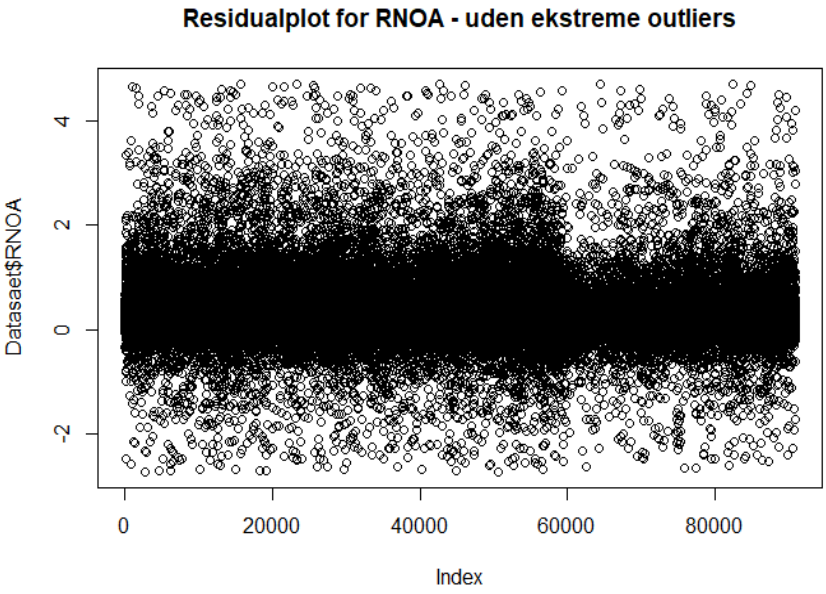
Yohn, T. L. (2018). Research on the use of financial statement information for forecasting profitability. *Accounting and Finance*, s. 1-19.

Bilag 1: Residualplots for RNOA

Residual plot for RNOA:



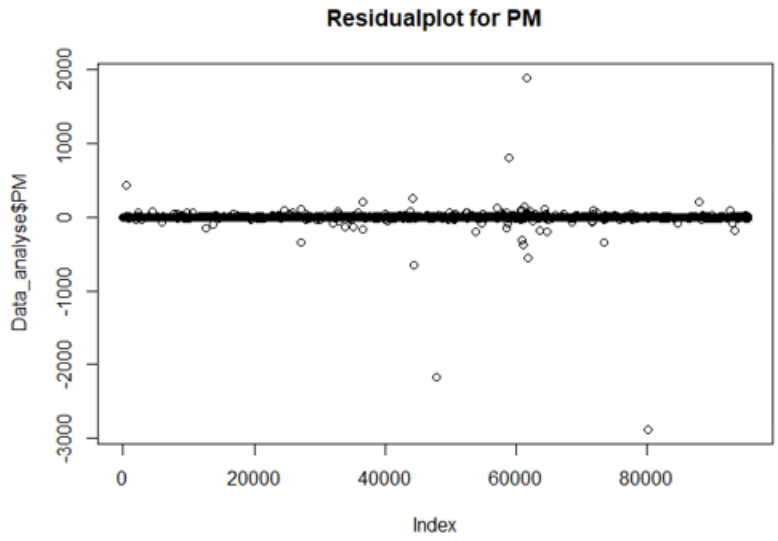
Residual plot for RNOA, uden ekstreme outliers:



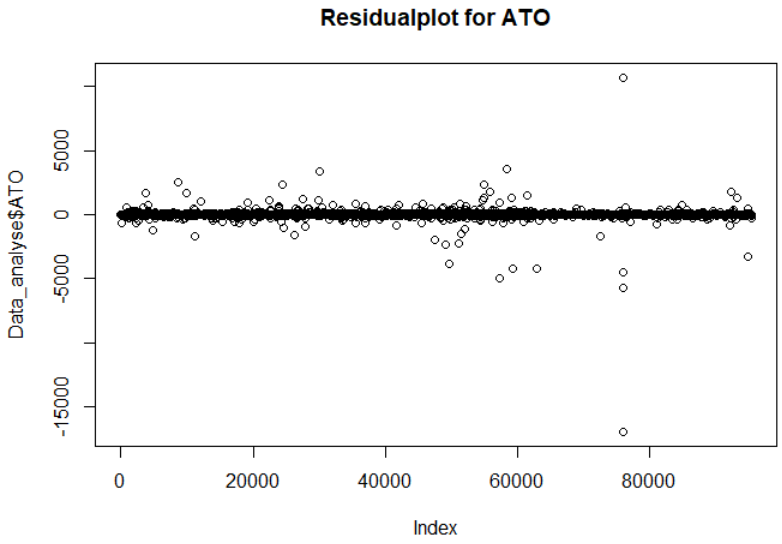
Bilag 1 - af egen inspiration.

Bilag 2: Residualplot for PM og ATO, før outliers er fjernet

Residualplot for PM:



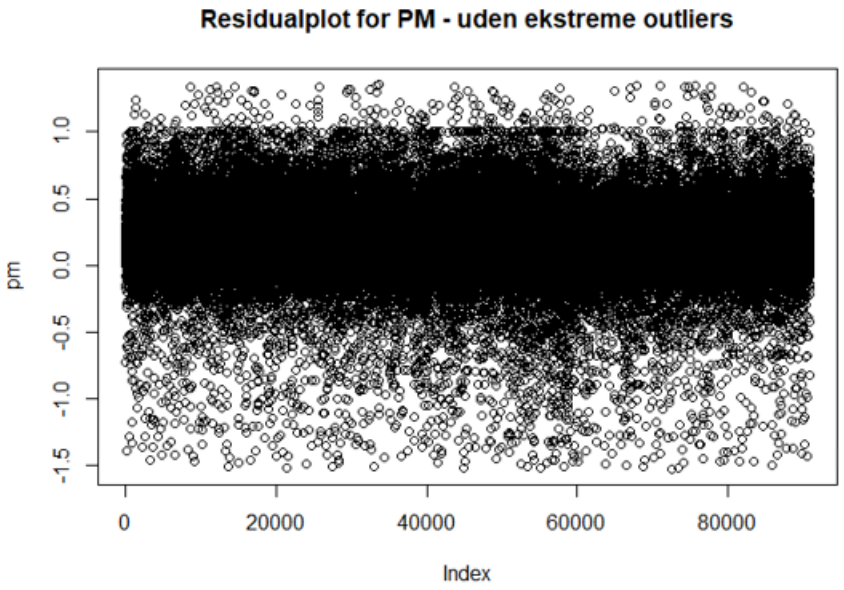
Residualplot for ATO:



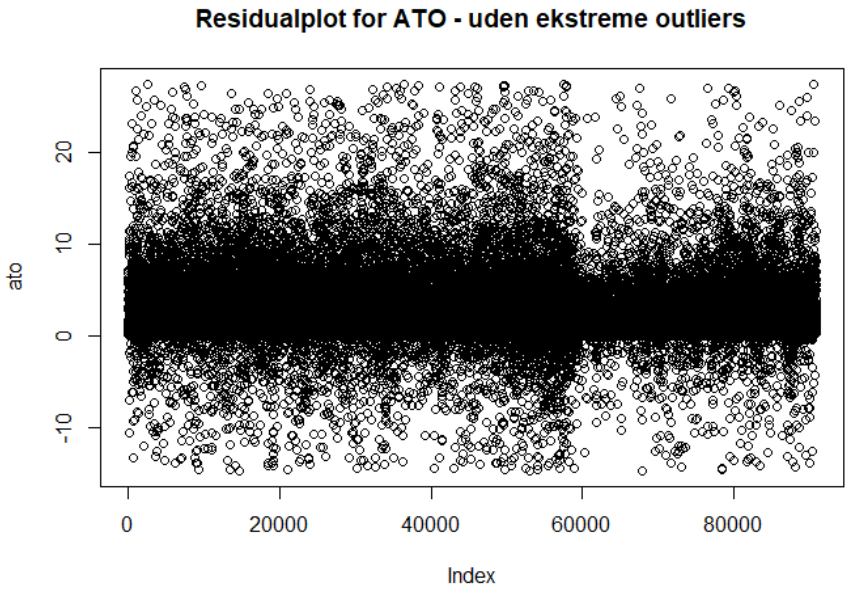
Bilag 2 - af egen inspiration.

Bilag 3: Residualplot for PM og ATO – uden ekstreme outliers

Residualplot for PM:

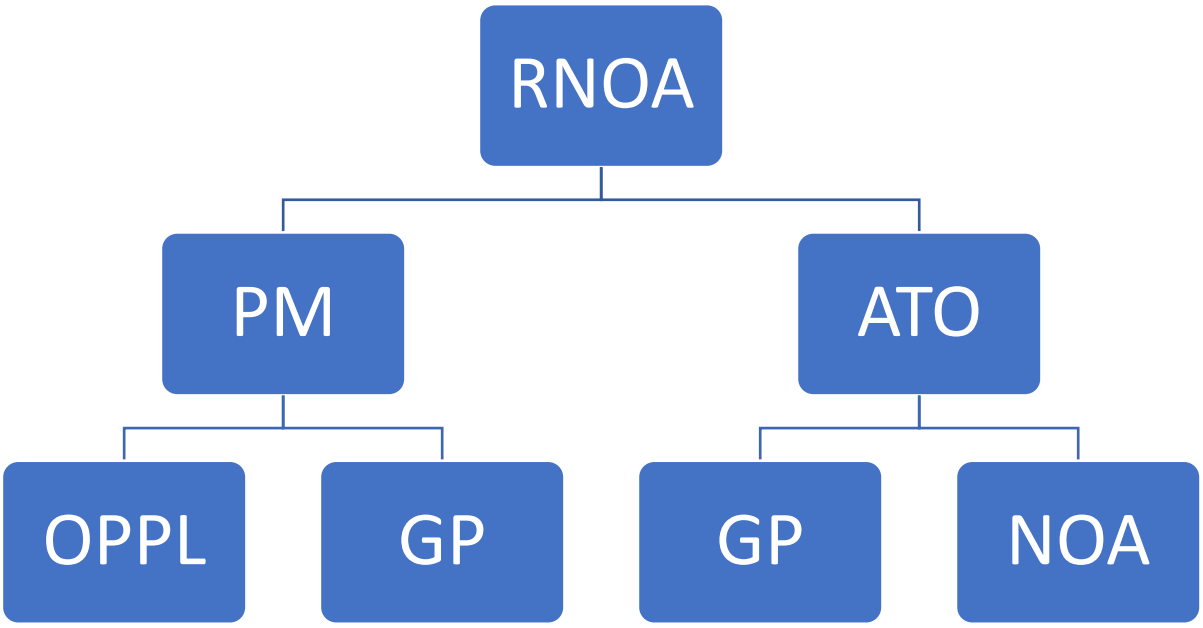


Residualplot for ATO:



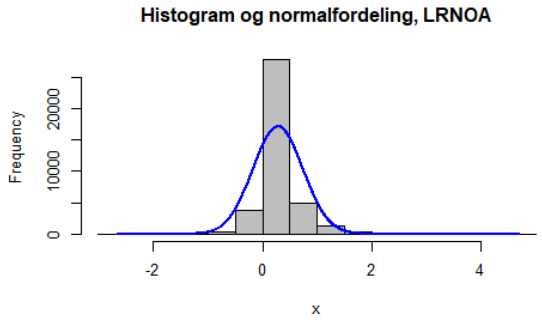
Bilag 3 - af egen inspiration.

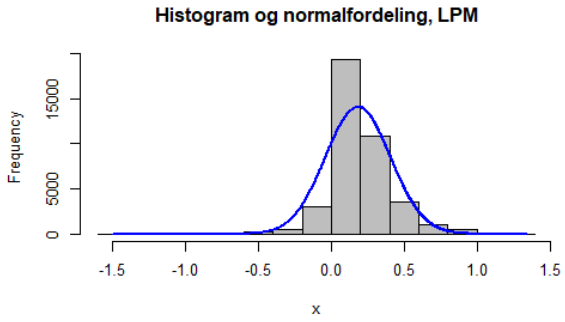
Bilag 4: DuPont Pyramide

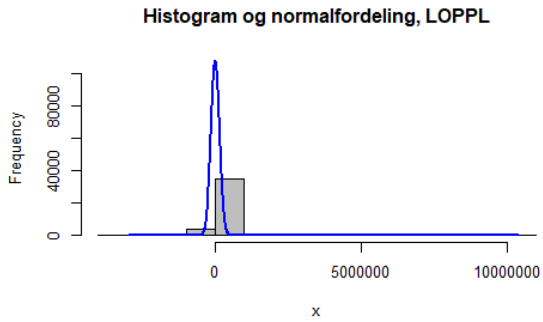


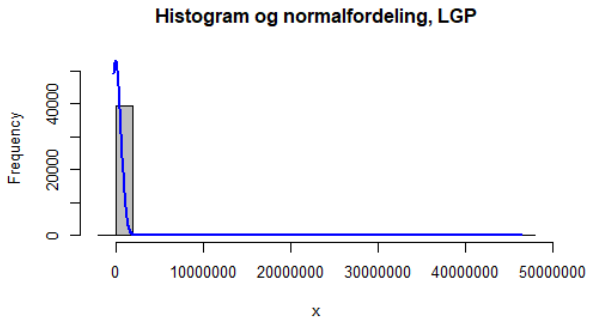
Bilag 4 - af egen inspiration. Samt Soliman (2008)

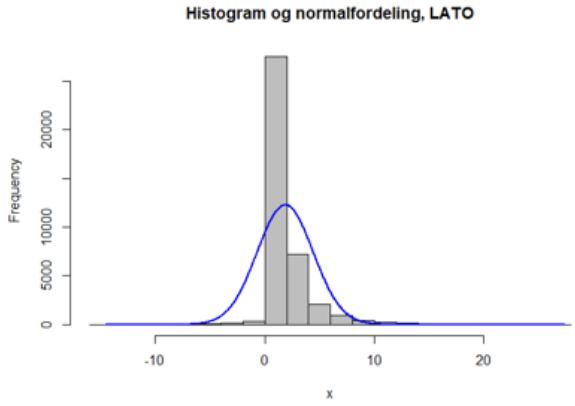
Bilag 5: Histogrammer over uafhængige variabler

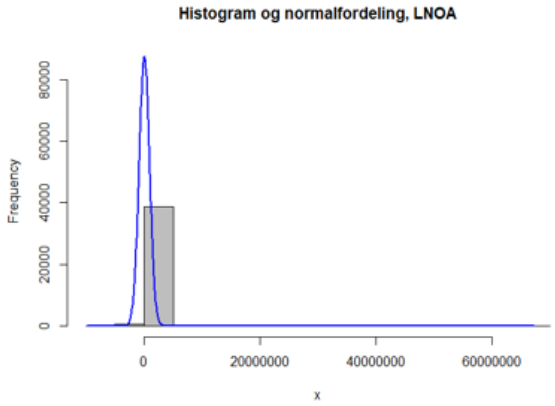


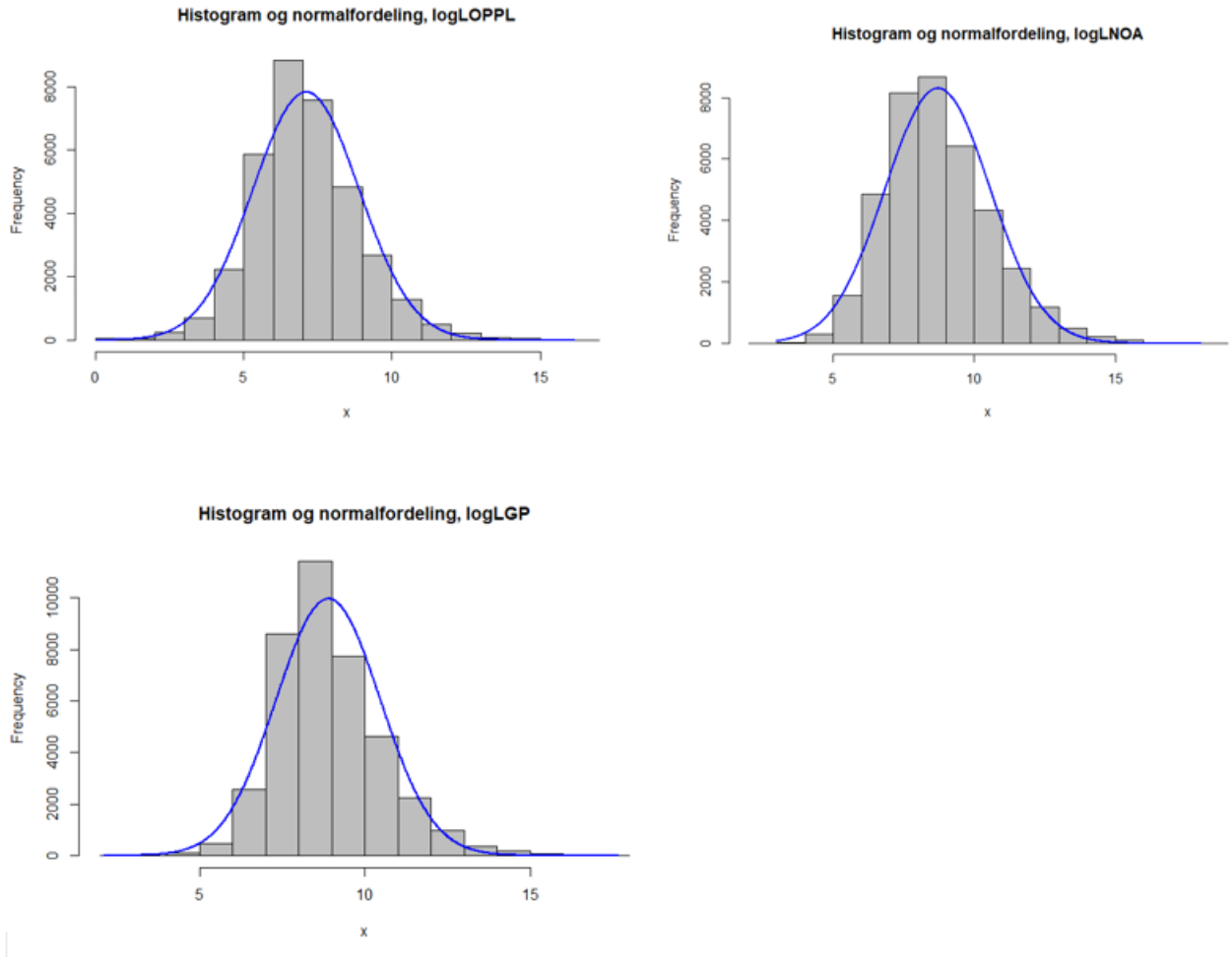








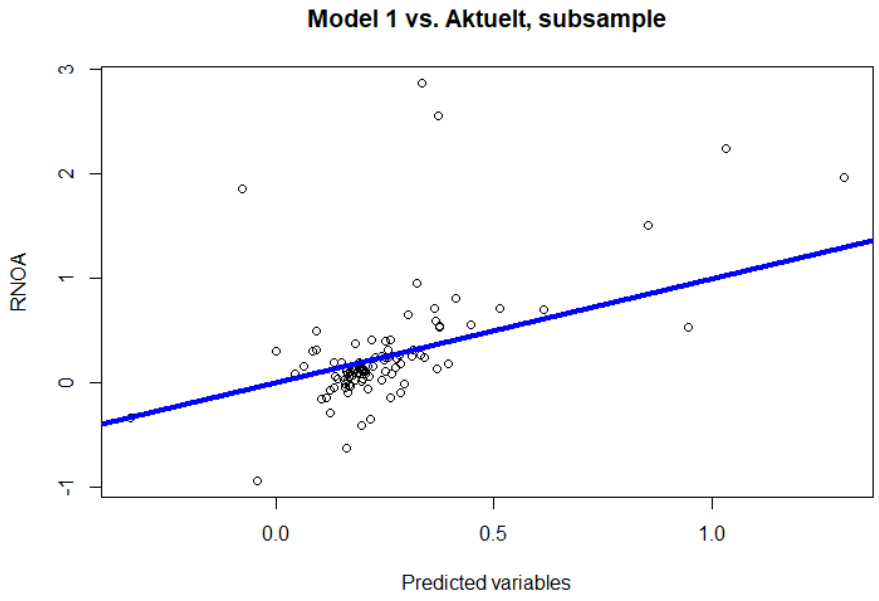




Bilag 5 - af egen inspiration.

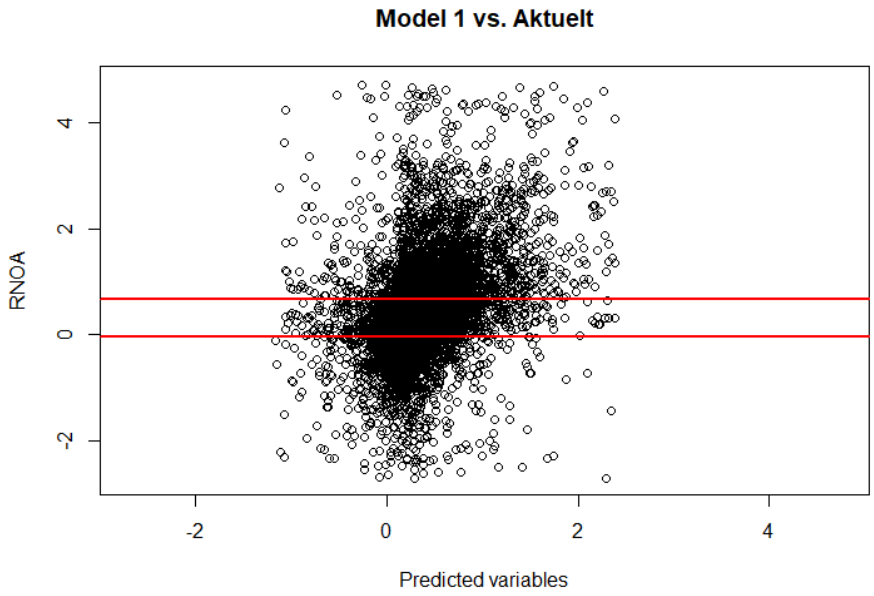
Bilag 6: Deciler for RNOA, samt tæthed for model 1 og plottet subsample:

Plot for subsample af model 1:



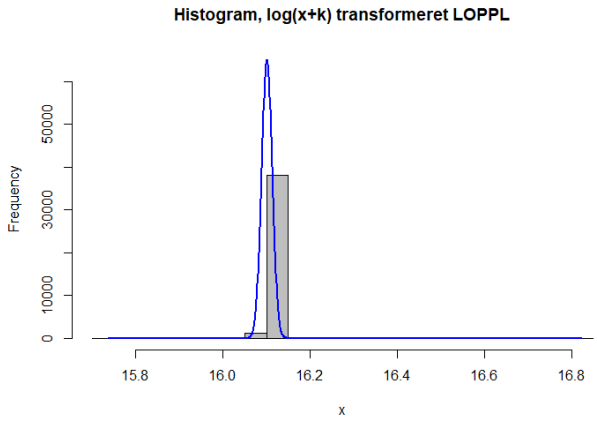
Deciler samt "Tæthedsplot" for model 1

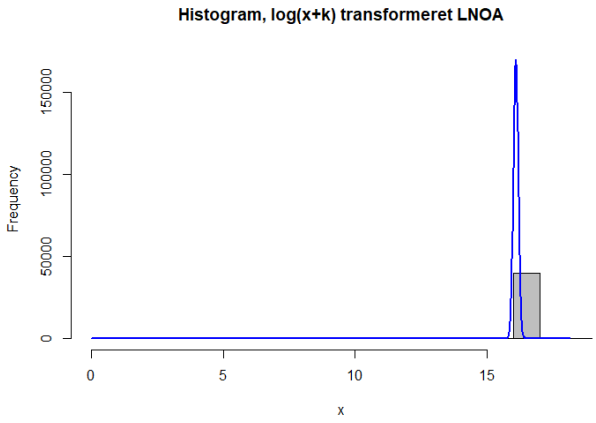
Decil	RNOA
0 %	-2,7021
10 %	-0,0328
20 %	0,0394
30 %	0,0827
40 %	0,1275
50 %	0,1706
60 %	0,2317
70 %	0,3116
80 %	0,4377
90 %	0,6801
100 %	4,7011

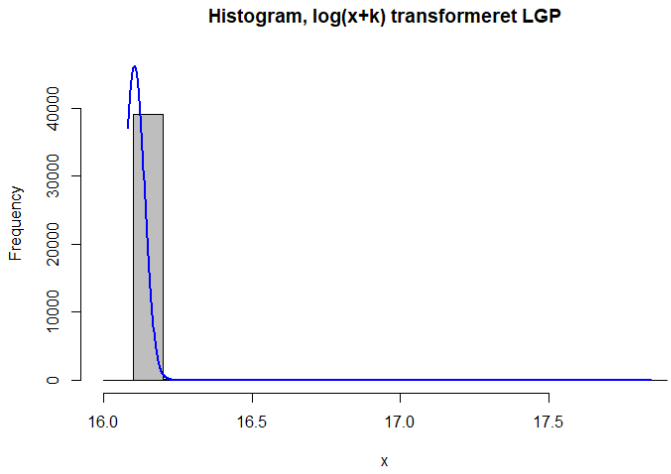


Bilag 6 - af egen inspiration.

Bilag 7: Histogrammer over log(x+k) af LOPPL, LNOA og LGP

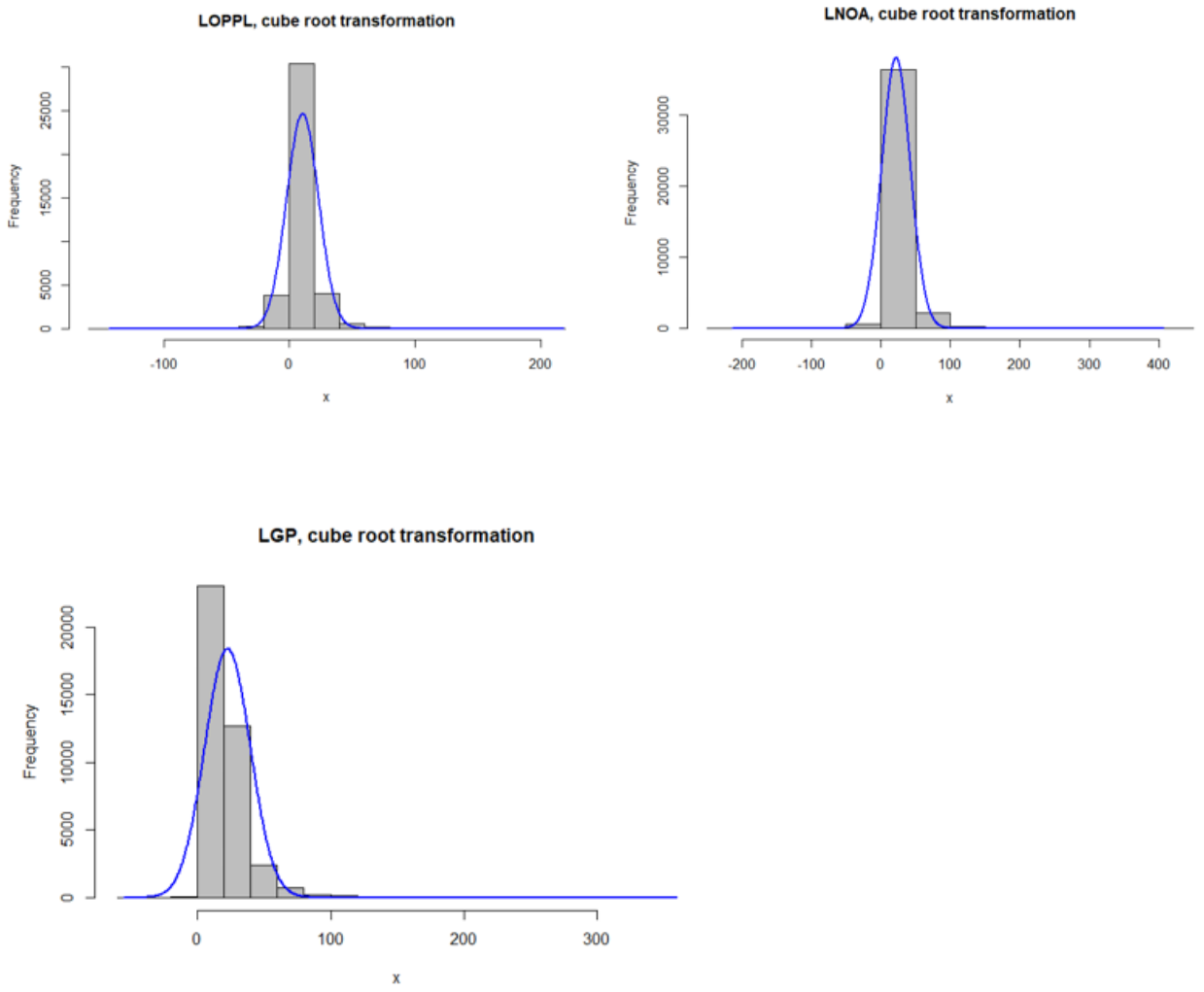






Bilag 7 - af egen inspiration.

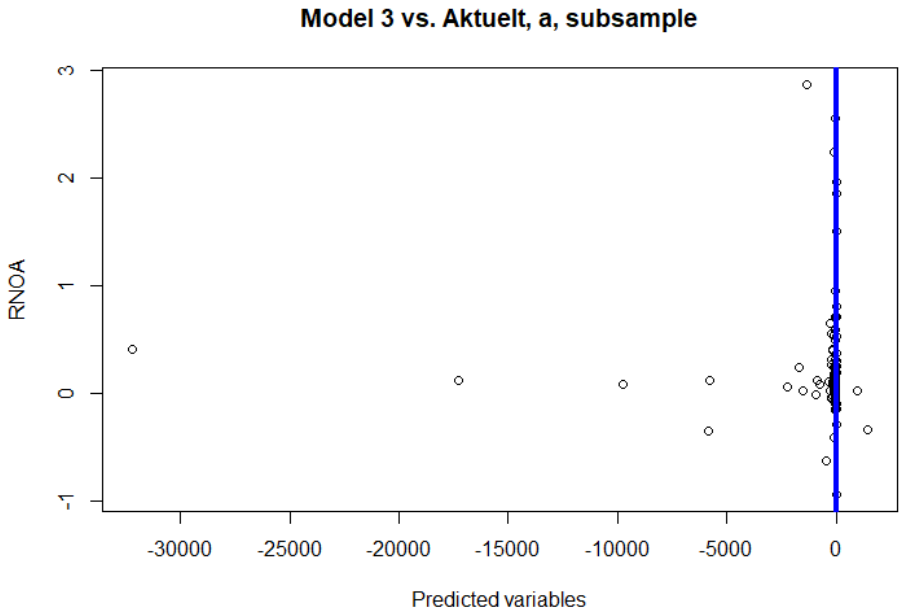
Bilag 8: Histogrammer over cube root transformation af LOPPL, LNOA, og LGP.



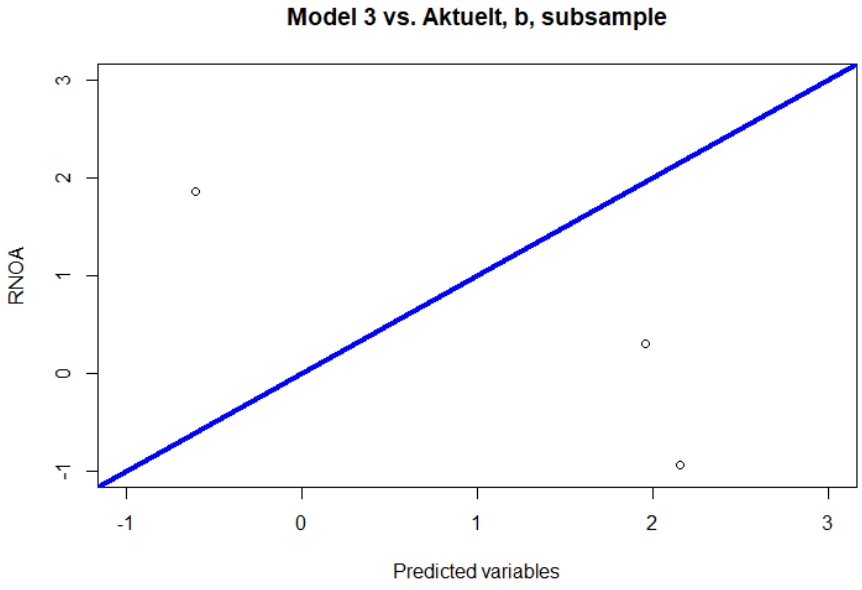
Bilag 8 - af egen inspiration.

Bilag 9: Visualisering af model 3's præcision

Forskellige værdier på akserne:

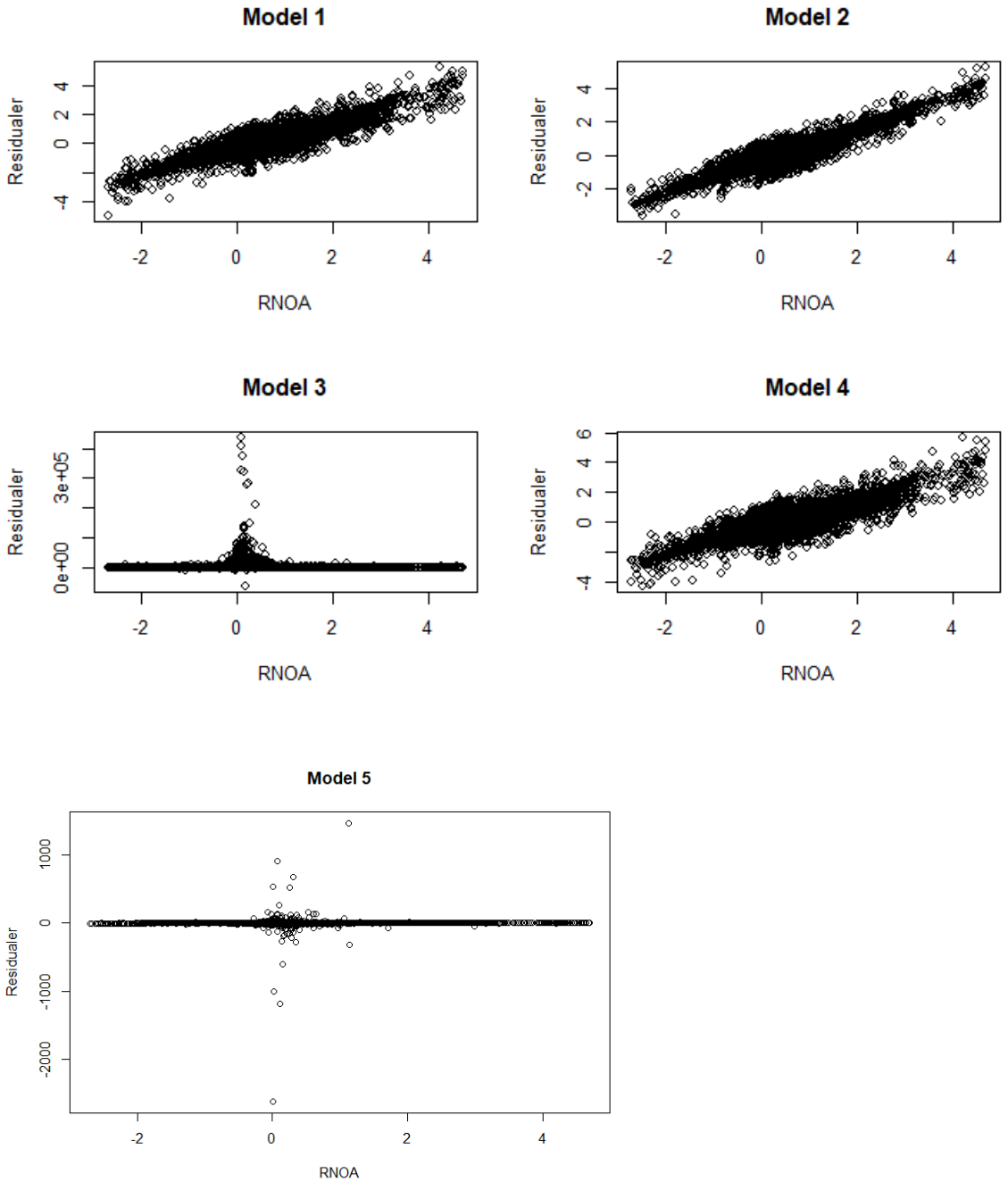


Samme længde på akserne:



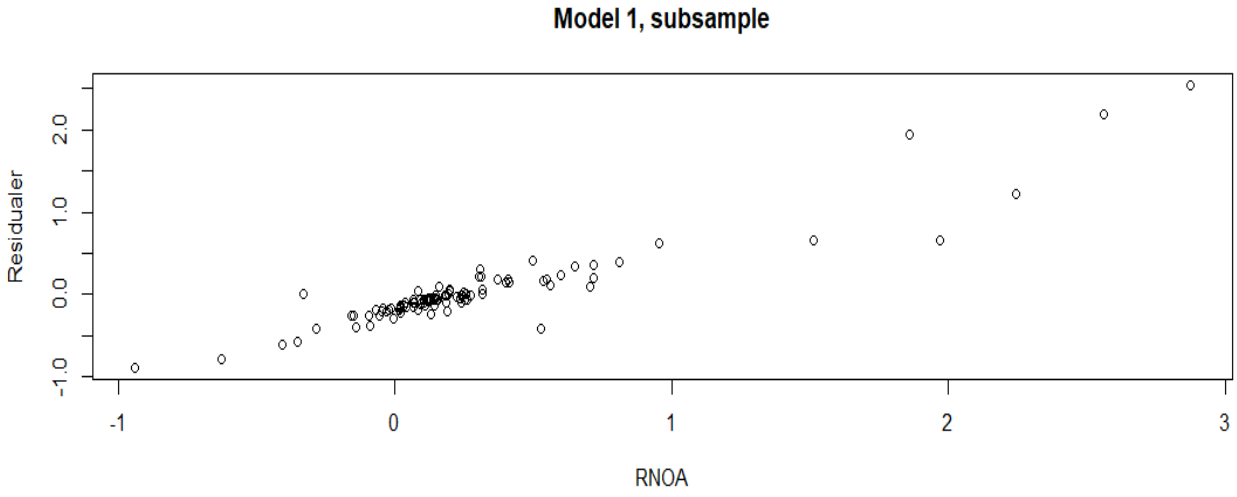
Bilag 9 - af egen inspiration.

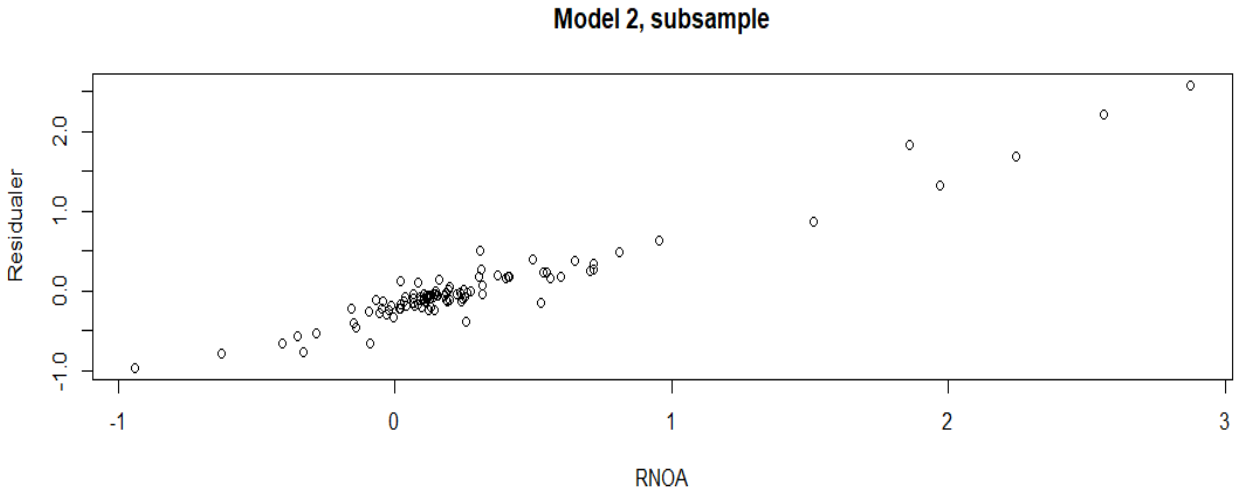
Bilag 10: RNOA plottet mod residualerne for de fem modeller

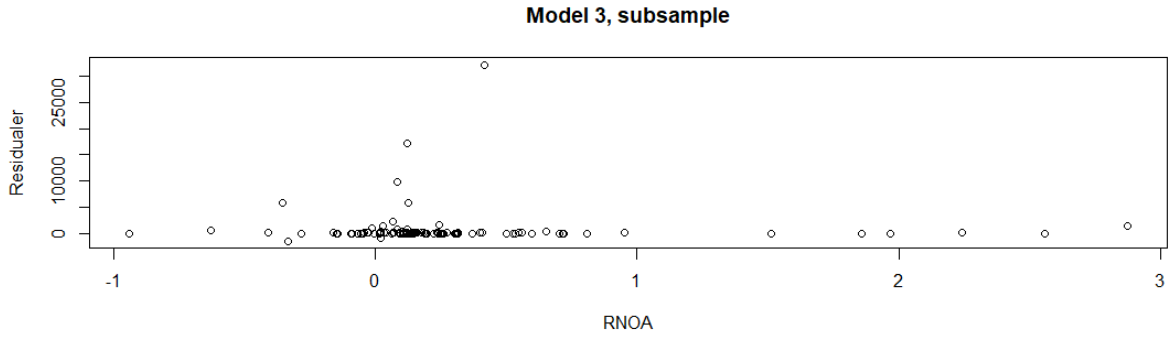


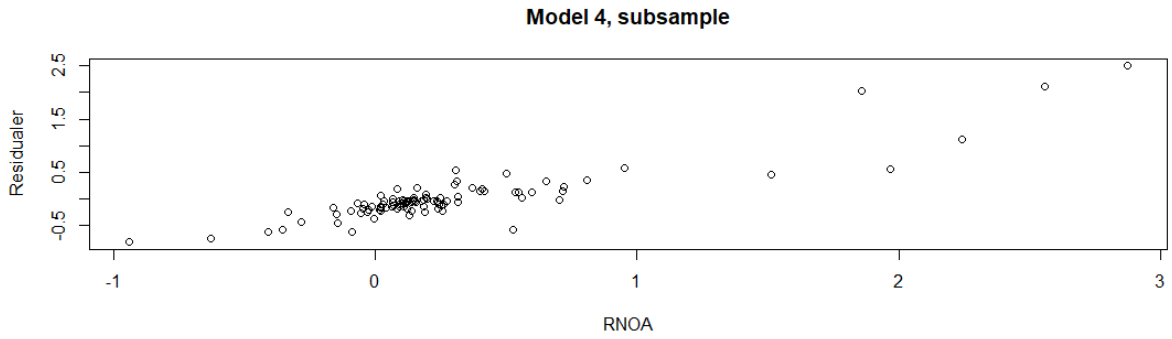
Bilag 10 - af egen inspiration.

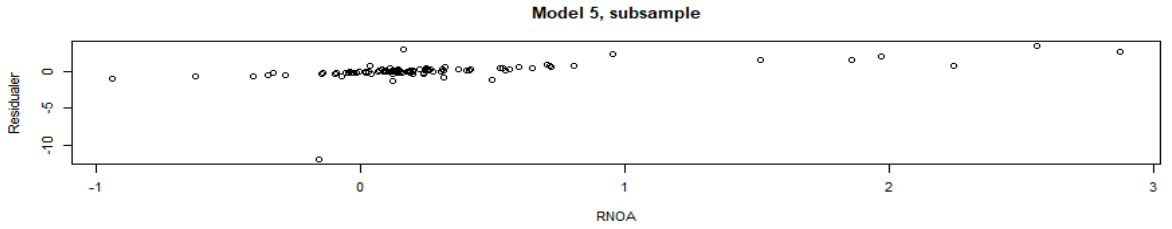
Bilag 11: Subsample af RNOA plottet mod residualerne for de fem modeller

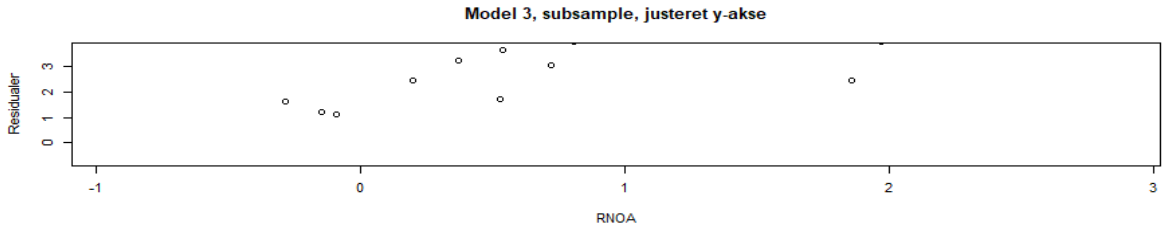


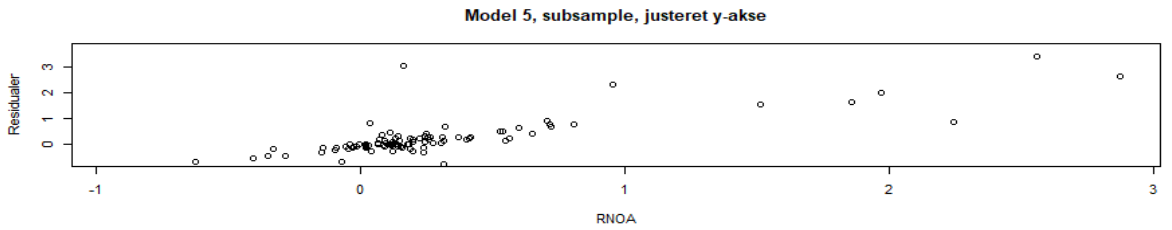












Bilag 11 - af egen inspiration.