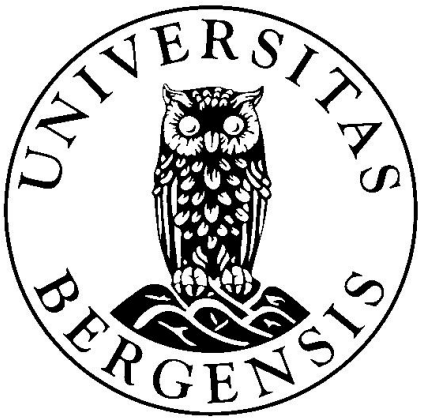
Negativ binomisk regresjon med modifiserte sannsynligheter for nullobservasjoner; ZINB og ZANB

Masteroppgave i statistikk

Mette Optun



Matematisk institutt

Universitetet i Bergen

6. november 2013

Takk

Jeg vil starte med å takke min veileder, Ivar Heuch, for alltid å stille opp med engasjement, godt humør og store doser med motivasjon. Og ikke minst har hans evne til å svare på spørsmål som ikke alltid har vært like godt formulert, vært uvurderlig.

Videre vil jeg takke mine medstudenter for en fin felles studietid med hyggelige pauser, som har inkludert noe faglig innspill, men mest bare kos. Spesielt vil jeg nevne min lesesalpartner, Marthe, som har bidratt til å gjøre studiehverdagen enda kjekkere.

Sist, men ikke minst, vil jeg takke min familie, og øvrige venner for all støtte og tålmodighet, spesielt de siste månedene med intens oppgaveskriving. Min kjære søster, Asta, må da særlig takkes, for all oppmuntring og til og med faglig hjelp hun har bidratt med.

Mette Optun
6. november 2013

Innhold

Innledning

Regresjonsmodeller av Poisson- og negativ binomisk fordeling blir ofte brukt til å utføre regresjonsanalyse på datasett innen medisin, biologi, økonomi og mange andre fagfelt. Det oppstår ofte tilfeller der andelen av verdien null i datasettet ikke samstemmer med den som er forventet når det antas at observasjonene enten er negativ binomisk (NB)- eller Poissonfordelt. Et eksempel på slike kan være antall innrapporterte skademeldinger et år fra en gruppe forsikringstakere, som grunnet egenandel vil inneholde flere observasjoner av null enn forventet. Antall dager en pasient er innlagt på sykehus, der kun innlagte pasienter er tatt med i beregningen, er et annet eksempel på at forventet antall nullobservasjoner ikke vil være lik forventet antall fra NB- eller Poisson fordelte data.

Hvis sannsynligheten for verdier av null er fjernet fra en fordeling, kalles denne nulltrunkert. Datasett med nullobservasjoner fra én fordeling og observasjoner større enn null fra en annen fordeling (nulltrunkert), kalles zero-altered (ZA) data. Når antall nullobservasjoner er større enn forventet og kommer fra to ulike fordelinger har man zero-inflated (ZI) data.

Ved regresjonsanalyse på datasett er vanlig NB- eller Poissonregresjon mye brukt. I tilfeller der antall nullobservasjoner ser ut til å avvike fra det som er forventet, er andre modeller foreslått og blitt hyppigere brukt i nyere tid. Eksempler på slike modeller er zero-inflated og zero-altered regresjonsmodell av antatt NB- eller Poissonfordelt data, ZINB, ZANB, ZIP og ZAP. Etter at Diane Lambert [11] skrev en artikkel om regresjonsmodellen ZIP i 1992, har ZI-modeller blitt mer anvendt. Modeller basert på ZA-fordelte data er også mer utbredt etter at Mullahy [13] i 1986 fikk publisert sin artikkel om modeller for regresjon av modifiserte datasett.

I en Poissonfordeling har man én parameter, intensiteten, og forventning og varians er like, noe som sjelden er tilfelle i et sett med observasjoner fra det virkelige liv. I en negativ binomisk fordeling har man to parametre, som fører til at datasett lettere kan tolkes når variansen er større enn forventningen.

Denne oppgaven vil fokusere på regresjonsmodeller som er basert på den negativ binomiske fordelingen, og spesielt på ZINB og ZANB der resultater fra bruk av modellene vil bli sammenlignet. Først vil den ta for seg en artikkel av Li [12], der hensikten var å modellere årlig tilvekst, mengde og sammensetning av forskjellige tresorter fra et skogsområde i nord Amerika som hadde

oppnådd en gitt størrelse, kalt ingrowth. Å finne en passende modell for slike data er et viktig ledd innen simulering av skogvekst. I studien i artikkelen ble det utført regresjon på dataene med modellene som hittil har blitt nevnt. Videre ble diverse måleverdier beregnet og tester utført for å kunne sammenligne og avklare hvilken modell som passet dataene best. Verdier som 2logL-, AIC og BIC, i tillegg til blant annet kjikvadrattester viste at NB-modellene var å foretrekke fremfor Poisson. Deretter ble resultater fra bruk av vanlig NB modell sammenlignet med resultater oppnådd med både ZINB og ZANB, og det ble konkludert med at ZINB totalt sett var den beste modellen i studien. I diskusjonsdelen av artikkelen ble forskjellen i resultater ved bruk av ZINB og ZANB kun forklart ved at modellene benytter ulike prosesser for tilpassing av nullobservasjoner. Valg av modell ut fra data ble også foreslått med forbehold om en viss formening om praktiske grunner til større antall nullobservasjoner enn forventet.

Li m.fl. sammenlignet ikke resultatene ved bruk av ZINB og ZANB videre, og det har heller ikke blitt funnet fagstoff der resultater fra de to modellene blir sammenlignet utover testverdiene. I en masteroppgave utført av Kaarstad [10] i 2011, ble modellene ZIP og ZAP sammenlignet ved detaljert analyse av regresjonsresultater. Her var hovedmålet blant annet å finne ut hvor stor betydning modellvalg hadde ved bruk på ulike tilfeller av datasett. I denne oppgaven er formålet å prøve på noe av det samme med regresjonsmodellene ZINB og ZANB. Før en introduksjon av disse modellene, er det valgt å inkludere en del teori om negativ binomisk fordelinger og regresjonsmodeller som tar utgangspunkt i disse. Dette vil gi et godt grunnlag for å forstå oppbyggingen i fordelingene bak, og til modellene ZINB og ZANB, når disse deretter vil bli grundig forklart. Til slutt vil det undersøkes om valg av de to modellene er avgjørende for resultat etter tilpassing av ulike datasett, og hvilke konsekvenser bruk av enten ZINB eller ZANB som regresjonsmodell når den andre er den korrekte kan gi. I tillegg vil det vurderes om det generelt er mulig å se hvilken modell som er mest riktig ut ifra egenskaper i det observerte datasettet.

Kapittel 1

Artikkel av Li og medarbeidere

I det følgende vil studien utført av Li m.fl [12] beskrives. Først vil det praktiske temaet, og hvordan de har tenkt å løse problemene statistisk belyses. Deretter vil resultatene de fikk og relevante konklusjoner som ble trukket oppsummeres. Til slutt vil det vurderes om det kan finnes flere interessante aspekter ved ZINB og ZANB som modeller for datasett.

1.1 Praktisk problemstilling

Å modellere årlig tilvekst av ingrowth er som nevnt en viktig faktor innen skogutvikling. Definisjonen på ingrowth er antall trær i et utvalg som har oppnådd en gitt størrelse over en bestemt periode. Denne størrelsen er vanligvis målt ved høyde, eller diameter ved brysthøyde (dbh). Antall nye trær som har oppnådd kravet for ingrowth i løpet av et år er en stokastisk prosess med påvirkning av klima og geografiske faktorer. Disse fører til store variasjoner av observert antall tilvekst av ingrowth mellom områder og derfor til større grad av usikkerhet ved modellering generelt. Antall ingrowth i et gitt område er også avhengig av andre faktorer, som blant annet hvilke tresorter som befinner seg der og sammensetningen av disse i både mengde og tetthet.

I studien har data fra en omfattende regional database blitt benyttet, der ulike typer kilder innen blant annet skogbruk og industri har samlet observasjoner fra forskjellige deler av den aktuelle skogsregionen i Nord-Amerika, kalt the Acadian Forest Region. Dette ble utført over lengre tid og totalt 33 587 observasjoner av blant annet utvalgsstørrelse, antall trær, og grunnflaten av trær i utvalgsområdet ble inkludert. Dataene ble hentet fra målinger som hadde blitt repetert med ulike tidsintervall, og for å oppnå et mål på et årlig antall observerte trær som passerte minimumskriteriet for ingrowth, ble det i studien brukt det totale antallet dividert med tidsintervallet.

Fra kildene ble det også registrert målinger av blant annet utvalgsstørrelse, antall trær, sammensetning av tresorter og grunnflaten av trær i utvalgsområdet. En av utfordringene i studien var at kildene hadde forskjellige kriterier på at et tre hadde oppnådd status ingrowth.

Hovedmålet med studien var å finne den beste modellen for et datasett bestående av årlig antall

ingrowth fra forskjellige utvalgsområder av ulike størrelser og med ulike sammensetninger av tresorter.

1.2 Modeller som foreslås

I artikkelen ble det først tatt opp tidligere utprøvde modeller for årlig antall tilvekst av ingrowth. *Statiske* modeller hadde tidligere blitt brukt til å predikere et konstant antall nye ingrowth fra observasjoner der et tilnærmet likt antall tilvekst av ingrowth ble observert over lengre perioder. *Dynamiske* modeller derimot, med ulike påvirkningsfaktorer som regresjonsvariabler, kunne predikere et antall fra observasjoner med større spredning i antall nye ingrowth, eller over kortere tidsintervall mer presist. Problemet i denne studien var den store mengden av nullobservasjoner i datasettene, og hverken statiske eller dynamiske modeller klarte å predikere verdien null.

Forfatterne kom deretter inn på mulighetene å bruke diskrete fordelinger i de statistiske modellene, der modifiserte versjoner kunne svare for et stort antall nullobservasjoner. Poissonfordelingen ble først nevnt, i tillegg til de tilhørende ZI- og ZA- fordelingene, ZIP og ZAP. Siden fordelingen i Poisson bare har én parameter og har lik forventning og varians, tok de også med negativ binomisk (NB) fordeling sammen med dens ZI- og ZA-fordelinger, altså modellene ZINB og ZANB. Negativ binomiske modeller har nemlig en tilleggsparameter som kunne forklare spredningen i observasjonene som var positive. I studien ble det utført regresjon med alle de seks forskjellige modellene.

1.3 Statistiske metoder

I studien ble regresjon utført, med årlig antall tilvekst av ingrowth som avhengig variabel. For de positive observasjonene, ble logaritmiske link-funksjoner brukt i alle modellene Poisson, ZIP, ZAP, NB, ZINB og ZANB.

Det ble brukt lineære prediktorer i alle modellene, som bestod av i) SBA = stand basal area (grunnflaten av trær i området), ii) prosentandel av trær i området, iii) antall trær per hektar, iv) skogbonitet (mål for arealets evne til å produsere trevirke) og v) minste diametermål ved brysthøyde. Sistnevnte, som var med på å avgjøre størrelsen på avhengig variabel, ble altså tatt med som forklaringsvariabel for å ta høyde for at det eksisterte ulike minstekrav til definisjonen ingrowth i observasjonene. I fordelingene som ligger til grunn for modellene ZIP, ZAP, ZINB og ZANB er det en felles parameter, π, som tilsvarer sannsynligheten for å få en nullobservasjon fra en binomisk fordeling. Denne sannsynligheten ble estimert med en logistisk modell med de samme forklaringsvariablene som nevnt tidligere.

I studien ble det samlet observasjoner over lang tid, gjentatt på de samme utvalgsområdene. I tillegg til de tidligere nevnte påvirkningsfaktorene og kombinasjoner av disse, var også andre krefter med på å påvirke årlig antall tilvekst av ingrowth i et utvalgsområde. Grunnet at en del av

disse ikke alltid var enkle å forklare, ble det lagt til parameterverdier, såkalte *random effects*, til konstantleddet i linkfunksjonen. Dette ble gjort for å dekke variabiliteten i de positive observasjonene, slik at faktorer som ikke ble tatt med som forklaringsvariable ble tatt høyde for i modellen. Maksimum likelihood estimat for de forskjellige parametrene ble funnet ved hjelp av statistisk programvare SAS.

For å evaluere hvilken modell som passet best, ble ulike måleverdier brukt. Noen av disse var Akaike's information criterion(AIC), Bayesisan information criterion (BIC) og -2log-likelihood, som alle indikerer bedre modelltilpasning med modellen som gir de laveste verdiene. Det ble også brukt likelihood ratio tester for å sammenligne Poissonmodeller med tilsvarende versjoner av NB-modeller. I tillegg ble Vuong-tester brukt for å finne ut hvilke av to ikke-nøstetet modeller som ga best uttelling i tilfellene der de andre målene ikke ga signifikante forskjeller.

En modell for å predikere sammensetninger av tresorter ble også inkludert i artikkelen, siden disse var med på å påvirke andelen av ingrowth og forklaringsvariablene som ble brukt i modellene. For å predikere denne sammensetningen ble prosentandelen av grunnflaten av trær med ingrowth i området brukt som avhengig variabel, mens total grunnflate i område, prosentandelen av grunnflaten av hver tresort i området, i tillegg til skogbonitet ble brukt som forklaringsvariabler. Til dette ble det brukt en logistisk modell.

1.4 Hovedresultater biologisk/statistisk

Av alle målingene i studien var 30.1% nullobservasjoner. Gjennomsnittsverdien av ingrowth i den positive delen av alle observasjonene, målt i antall per areal per år, var 22.8 med standardavvik på 34.1. Observasjonene varierte fra null til 299 antall nye ingrowht per areal per år. Ved å ta med random effects, så man ut ifra de ulike evalueringsmålene at alle de utprøvde modellene passet bedre til observasjonene.

Variansen var betraktelig større enn forventningen i observasjonene, og testverdiene -2logL, AIC og BIC var mye lavere for NB-modellene til sammenligning med Poisson-modellene. I tillegg viste likelihood ratio tester av NB vs Poisson, ZIP vs ZINB og ZAP vs ZANB signifikante forskjeller ved bruk av likelihood ratio tester. Negativ binomiske regresjonsmodeller med sine tilleggsparametre så med andre ord ut til å tilpasse observasjonene bedre enn Poissonmodellene.

I studien ble det konkludert med at NB-modellene passet relativt bra til observasjonene. Vanlig negativ binomisk modell viste seg å være dårlig til å predikere nok nullobservasjoner sammenlignet med ZINB og ZANB. Det ble vurdert at i denne studien var ZINB-modellen med random effects den beste av de som ble utprøvd.

Det ble funnen noen forskjeller i resultatene ved bruk av ZINB og ZANB: Forventet antall ingrowth mellom 1 og 19 var lavere ved bruk av ZANB enn ved bruk av ZINB, mens forventet

antall ingrowth mellom 20 og 59 var lavere ved bruk av ZINB enn ved bruk av ZANB. Ser man nærmere på de observerte hyppighetene fra tabell 3 i artikkelen, er det tydelig at i halvparten av intervallene av antall ingrowth mellom 1 og 19, så gir bruk av ZANB-modell et forventet antall nærmest det observerte antallet. I de resterende intervallene mellom 1 og 19 gir ZINB-modell et bedre forventet antall i forhold til det observerte. Når man derimot ser på intervallene mellom 20 og 60, så gir ZINB-modellen langt flere predikerte antall nærmere de observerte. Fra samme tabell gir ZANB-modellen det eksakt samme antallet av nullobservasjoner som er observert i studiet.

I artikkelen kom det frem at årsaken til de ulike forventningene i de to modellene var at nullobservasjonene ble modellert på forskjellige måter. I ZANB-modellen er det kun en binomisk prosess, mens i ZINB-modellen er det både en binomisk og en negativ binomisk prosess som styrer hvor mange nullverdier som forventes.

1.5 Videre tolkning og spørsmål etter artikkelen

Modellering av antall årlig tilvekst av ingrowth ble gjort på grunnlag av innsamlet data over en lengre tidsperiode. For å få et mål på årlig antall nye ingrowthtilfeller, ble antallet fra data observert i løpet av en periode dividert med tidslengden på perioden. Deretter ble det foretatt en avrunding til nærmeste hele tall i tillegg til standardisering i forhold til at måleenheten var antall ingrowth per hektar per år. Tradisjonelt har Gaussisk regresjon blitt brukt for å modellere ingrowth, men denne hadde vist seg å tilpasse observasjoner dårlig med tanke på spredning og skjevhet. Ved hjelp av avrunding ble det altså i denne artikkelen mulig å ta i bruk diskrete fordelinger som er mye brukt i regresjonssammenheng. Li m.fl. poengterte at det sannsynligvis var en viss korrelasjon mellom årlig tilvekst av ingrowth over et område med de samme ytre faktorene. Det kan likevel antas at avrundingen og benyttelse av årlig gjennomsnitt kan føre til feil, for eksempel at det i realiteten var færre nullobservasjoner noen år enn det som ble beregnet fra observasjonene. Dette vil i så tilfelle resultere i modeller som predikerer for få årlige nye ingrowthtilfeller.

I diskusjonsdelen i artikkelen ble problemet med valg av modell tatt opp, blant annet valg av ZINB eller ZANB, som ga de beste resultatene. Det ble nevnt at man bør ha kjennskap til biologiske faktorer i forhold til nullobservasjonene for å kunne velge riktig modell i et datasett. Hvis man grunnet mulige målefeil eller lignende kan ha fått for mange nullobservasjoner, så nevnes ZANB som et godt valg av modell. Hvis det i tillegg til målefeil vurderes naturlige grunner til for mange nullobservasjoner så ble ZINB foreslått som et bedre valg av modell. Grunnen til dette er hvordan fordelingene til ZINB og ZANB er oppbygd, nullobservasjonene styres bare av en binomisk prosess i ZANB-modellen. Det blir til slutt i artikkelen presisert at det i en annen studie kan vise seg at ZANB vil være en like god modell som ZINB. Modellene blir ikke sammenlignet utover dette.

I studien til Li m.fl bestod en stor andel av datasettet av nullobservasjoner. Det er ingen tvil om at det er viktig å fokusere på dette ved regresjonsanalyse. I denne oppgaven vil ZINB og ZANB som regresjonsmodeller for datasett bli nærmere belyst, og eventuelle andre egenskaper

ved datasett vurderes som avgjørende for valg mellom de to modellene. Ettersom det ikke har blitt funnet fagstoff der regresjonsresultater fra bruk av ZINB og ZANB blir sammenlignet mer detaljert, vil det blant annet bli undersøkt om det er mulig å finne konsekvensene av å bruke den ene modellen når den andre er den korrekte.

Kapittel 2

Negativ binomiske fordelinger og modeller

En negativ binomisk fordeling kan utledes på flere fremgangsmåter og med bruk av ulike parametre. Valg av utledning og parametrisering er avhengig av hvilke egenskaper ved fordelingen som ønskes å studeres videre. Dette kapitlet starter med å gjennomgå noen av utledningene som finnes i litteraturen. Deretter vil det bli fortalt litt om fordelingen og ulike utledninger historisk. Til slutt skal vi det bli vist hvordan de ulike utledningene blir brukt som utgangspunkt til å lage forskjellige regresjonsmodeller for å tilpasse datasett.

2.1 Negativ binomisk fordeling

Først vil utledningen som er den de fleste forbinder med en negativ binomisk fordeling beskrives. Deretter vil utledningen som leder til modellen som skal brukes videre i oppgaven, *NB2-modellen*, forklart. Til slutt vil det bli fortalt litt om noen av de andre utledningene som finnes.

2.1.1 Klassisk utledning

I den *klassiske utledningen* av en negativ binomisk fordeling blir det tatt utgangspunkt i en rekke Bernoulli trekninger,

$$x = \begin{cases} 0 \text{ med sannsynlighet } (1-p) \\ 1 \text{ med sannsynlighet } p, \end{cases} \tag{2.1}$$

der utfallet $x = 0$ kan kalles en fiasko og utfallet $x = 1$ for en suksess. Bernoullifordelingen er en binomisk fordeling med kun èn trekning.
I en negativ binomisk fordeling foretas trekninger fra en Bernoullifordeling inntil r suksesser, eller eventuelt r fiaskoer er oppnådd. Dersom den tilfeldige variabelen Y settes lik antall suksesser inntil r fiaskoer er oppnådd, så kan dette skje ved å først trekke $r-1$ fiaskoer på de første $y+r-1$ trekningene, og deretter trekke den r'te fiaskoen. Begge disse prosessene er binomiske,

altså uavhengige, og fører til følgende negativ binomisk sannsynlighetsfordeling for $y = 0, 1, ...$

$$\begin{aligned} P(y; p, r) &= P(\ r-1\ \text{ fiaskoer på } y+r-1\ \text{ forsøk }) P(\ \text{fiasko på ett forsøk}\ ) \\ &= \binom{y+r-1}{r-1} p^y (1-p)^{r-1}(1-p). \\ &= \binom{y+r-1}{r-1} p^y (1-p)^r. \end{aligned} \tag{2.2}$$

Forventningen og variansen til Y er

$$E(Y) = \frac{rp}{(1-p)}\ , \quad \text{Var}(Y) = \frac{rp}{(1-p)^2}. \tag{2.3}$$

Den tilfeldige variabelen Y kan også settes lik antall trekninger inntil r'te suksess eller r'te fiasko er oppnådd. I sistnevnte tilfelle må det først trekkes $r-1$ fiaskoer på totalt $y-1$ trekninger, påfulgt av r'te fiasko. Den negativ binomiske fordelingen vil da for $y = r, r+1, ...$ være på en annen form,

$$P(y; p, r) = \binom{y-1}{r-1} p^{y-r}(1-p)^r,$$

med den samme forventningen og variansen som i (2.3)
I forklaringen ovenfor er r et positivt heltall og fordelingen kalles da en *Pascal negativ binomisk*. En geometrisk fordeling er en negativ binomisk fordeling der antall suksesser før første fiasko oppnås blir målt, med andre ord der $r = 1$.

Navnet *negativ* binomisk kommer fra bruken av binomialteoremet med bruk av negativ eksponent. Denne gir

$$\begin{aligned} 1 &= (1-p)^r (1-p)^{-r} \\ &= (1-p)^r \sum_{y=0}^{\infty} \binom{-r}{y} (-p)^y \\ &= \sum_{y=0}^{\infty} (-1)^y \binom{-r}{y} p^y (1-p)^r \\ &= \sum_{y=0}^{\infty} (-1)^y \frac{(-r)(-r-1)\cdots(-r-y+1)}{y!} p^y (1-p)^r \\ &= \sum_{y=0}^{\infty} \binom{y+r-1}{y} p^y (1-p)^r. \end{aligned}$$

Uttrykket til slutt tilsvarer summen av forventningene i uttrykk (2.2). Her blir det i tillegg vist at den totale sannsynligheten i fordelingen er lik én.

2.1.2 Fordelingen bak NB2-modellen, en Poisson-Gamma mixture

En av de mest brukte fordelingene for å modellere datasett med uavhengige observasjoner er Poissonfordelingen. Denne antar at forventning og varians skal være like, i tillegg til at observasjonene er antatt å være uavhengige. Dette er sjelden tilfelle i praksis, og fører til at forventning og varians ikke blir like, som oftest blir variansen større enn forventningen. Det vil nå blir utledet en negativ binomisk fordeling i form av en såkalt *Poisson-Gamma mixture fordeling*, som heretter vil bli kalt *NB2-fordelingen* i oppgaven. Det er denne fordelingen som ligger til grunn for den mest brukte modellen av de negativ binomiske, *NB2-modellen*. Fordelen med å bruke denne vil bli forklart i kapittel 2.3.

I denne fremgangsmåten [9] er utgangspunktet at observasjonene i datasettet, y, er betinget Poissonfordelt gitt intensiteten λu, der u er Gammafordelt med fordeling $g(u)$ og forventning lik 1. Den betingete fordelingen for $y > 0$ er altså lik

$$f(y|u) = \frac{e^{-\lambda u}(\lambda u)^y}{y!},$$

og fordelingen til u med like parametre, ν, siden forventningen er lik 1;

$$g(u) = \frac{\nu^\nu}{\Gamma(\nu)} u^{\nu-1} e^{-\nu u}\, du.$$

Det er altså to parametre som styrer forventningen og variansen i den betingete fordelingen, λu. Setningen om total sannsynlighet gir oss den ubetingede fordelingen til y:

$$\begin{aligned}
f(y;\lambda,u) &= \int_0^\infty f(y|u)g(u)\, du \\
&= \int_0^\infty \frac{e^{-\lambda u}(\lambda u)^y}{y!} \frac{\nu^\nu}{\Gamma(\nu)} u^{\nu-1} e^{-\nu u}\, du \\
&= \frac{\lambda^y}{\Gamma(y+1)} \frac{\nu^\nu}{\Gamma(\nu)} \int_0^\infty e^{-(\lambda+\nu)u} u^{(y+\nu)-1}\, du \\
&= \frac{\lambda^y}{\Gamma(y+1)} \frac{\nu^\nu}{\Gamma(\nu)} \frac{\Gamma(y+\nu)}{(\lambda+\nu)^{y+\nu}} \\
&= \frac{\Gamma(y+\nu)}{\Gamma(y+1)\Gamma(\nu)} \frac{\lambda^y \nu^\nu}{(\lambda+\nu)^y (\lambda+\nu)^\nu} \\
&= \frac{\Gamma(y+\nu)}{\Gamma(y+1)\Gamma(\nu)} \left(\tfrac{\nu}{\lambda+\nu}\right)^\nu \left(\tfrac{\lambda}{\lambda+\nu}\right)^y \\
&= \frac{\Gamma(y+\nu)}{\Gamma(y+1)\Gamma(\nu)} \left(\tfrac{1}{\frac{\lambda}{\nu}+1}\right)^\nu \left(1-\tfrac{1}{\frac{\lambda}{\nu}+1}\right)^y
\end{aligned}$$

Ved å sette $\nu = \frac{1}{\alpha}$ og $\lambda = \mu$, oppnås den mye brukte negativ binomiske sannsynlighetsfordelingen

$$f(y;\mu,\alpha) = \binom{y+\frac{1}{\alpha}-1}{\frac{1}{\alpha}-1} \left(\tfrac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}} \left(\tfrac{\alpha\mu}{1+\alpha\mu}\right)^y \qquad ,\alpha > 0. \tag{2.4}$$

Parameteren α, som Hilbe kaller *overdispersion parameter* eller *heterogenity parameter*, blir heretter referert til som en dispersjonsparameter. Den antas å ha en konstant verdi, uavhengig av hver y-verdi i fordelingen. En negativ binomisk fordeling der r kan være alle positive reelle tall, som i denne utledningen, kalles for øvrig en *Polya negativ binomisk*, og er den som brukes mest i regresjonssammenheng. Ved sammenligning av de de to fordelingene i (2.2) og (2.4), viser disse at de er identiske sett bort fra ulik bruk av parametre; $r = \frac{1}{\alpha}$ og $p = \frac{\alpha\mu}{1+\alpha\mu}$, der dispersjonsparameteren, α, alltid er positiv. Ved å sette disse verdiene for r og p inn i forventningsuttrykket (2.3), får vi et uttrykk for forventningen og variansen til Y med hensyn på μ og α.

$$E(Y) = \frac{rp}{(1-p)} = \frac{\frac{1}{\alpha}\frac{\alpha\mu}{1+\alpha\mu}}{1-\frac{\alpha\mu}{1+\alpha\mu}} = \mu \tag{2.5}$$

$$\mathrm{Var}(Y) = \frac{rp}{(1-p)^2} = \frac{\frac{1}{\alpha}\frac{\alpha\mu}{1+\alpha\mu}}{\left(1-\frac{\alpha\mu}{1+\alpha\mu}\right)^2} = \frac{\mu}{1-\frac{\alpha\mu}{1+\alpha\mu}} = \mu + \alpha\mu^2 \tag{2.6}$$

Navnet *NB2* kommer av at det andre leddet i variansen er kvadrert, $\mathrm{Var}(Y) = \mu + \alpha\mu^2$. Det første leddet i det endelige uttrykket for variansen er lik forventningen, som er variansen fra Poissondelen av mixture-fordelingen. Det andre leddet er et produkt bestående av dispersjonsparameteren og forventningen, og tilsvarer variansen i gammadelen av mixture-fordelingen. Parameteren er altså med på å styre hvor mye større variansen er i forhold til forventningen og er således direkte knyttet til størrelsen på spredningen av observasjonene i et datasett. Jo større verdi av parameteren, uten at verdien av forventningen varierer, dess større spredning forekommer i observasjonene. Dersom parameterverdien er svært lav, $\alpha \to 0$, blir variansen i fordelingen tilnærmet lik forventningen, $\mathrm{Var}(Y_{\mathrm{NB2}}) = \mu + \alpha\mu^2 \to \mu$. Fordelingen tilsvarer da en Poissonfordeling.

Når α=1 blir fordelingen $f_{\mathrm{NB2}}(y;\mu,1) = \left(\frac{1}{1+\mu}\right)\left(\frac{\mu}{1+\mu}\right)^2$, og variansen er uavhengig av dispersjonsparameteren: $\mathrm{Var}(Y_{\mathrm{NB2}}) = \mu + \mu^2$. Dette tilsvarer en geometrisk fordeling.

Dersom den empiriske variansen i et datasett er større enn det som er forventet i forhold til antatt fordeling, har vi et datasett med *overdispersjon*. I et datasett som er antatt Poissonfordelt, har vi altså tilfelle av overdispersjon hvis variansen er større enn forventningen, siden disse er forventet å være like. En NB2-fordeling kan da være et bedre alternativ, siden denne tillater variansen å overskride forventningsverdien ved hjelp av dispersjonsparameteren. Det kan også forekomme tilfeller av overdispersjon i en negativ binomisk fordeling. Variansen i NB2-fordelingen blir da større enn forventet: $\mathrm{Var}(Y_{\mathrm{NB2}}) > \mu + \alpha\mu^2$. I tilfeller der den empiriske variansen er mindre enn det som er forventet fra en NB2-fordeling, $\mathrm{Var}(Y_{\mathrm{NB2}}) < \mu + \alpha\mu^2$, har vi et datasett med *underdispersjon*.

Ulike grunner kan føre til at overdispersjon oppstår i et datasett:

• Antakelser om at observasjonene er uavhengige er ukorrekte. Dette kan føre til at observasjonene oppstår i klynger, og således at variansen blir større enn forventet.
• Observasjonene i datasettet viser stor variasjon som den teoretiske fordelingen ikke klarer å dekke.

I senere kapitler vil tilfeller der bruk av NB2-modellen kan føre til enten under- eller overdispersjon grunnet avvik i det observerte datasettet bli studert.

2.1.3 Andre utledninger

Det finnes mange andre utledninger til en negativ binomisk fordeling. Boswell og Patil [14] viste for eksempel i et vitenskapelig bidrag som er utgitt på bokformat i 1970, tolv ulike utledninger av negativ binomisk fordelinger generert fra stokastiske modeller. I disse utledningene brukes den *sannsynlighetsgenererende funksjonen*, $E(Z^Y) = \sum_{y=0}^{\infty} p(y)z^y$, til å identifisere en negativ binomisk fordeling. Én av utledningene tilsvarer den klassiske utledningen som allerede er beskrevet. Fordelingen utledes også som blant annet en logaritmisk sum av tilfeldige variabler der antall ledd er poissonfordelt, som lineær fødsels- og dødsprosess, og som en poissonfordeling der intensiteten er en gammafordelt tilfeldig variabel.

I utledningen av NB2-fordelingen i forrige delkapittel ville den samme fordelingen blitt oppnådd ved å definere observasjonene som betinget Poissonfordelte gitt intensiteten λ, der λ er gammafordelt med en konstant formparameter lik $\frac{1}{\alpha}$ og en skalaparameter, som kan variere, lik $\alpha\mu$; $\lambda \sim \text{Gamma}(\frac{1}{\alpha}, \alpha\mu)$. Det finnes andre negativ binomiske fordelinger som kan utledes ved hjelp av en blanding av poisson- og gammafordelingen. En av disse er en *lineær negativ binomisk* fordeling, også kalt en *NB1-fordeling*. I denne utledningen settes observasjonene nok en gang som betinget Poissonfordelte gitt en intensitet som er gammafordelt, men i dette tilfellet kan formparameteren variere med observasjonene, mens skalaparameteren er en konstant verdi. Ved å sette $\lambda \sim \text{Gamma}(\mu, \frac{1}{\alpha})$, og integrere ut λ, oppnås fordelingen

$$f(y;\mu) = \binom{y+\mu-1}{\mu-1} \left(\tfrac{1}{1+\alpha}\right)^{\mu} \left(\tfrac{\alpha}{1+\alpha}\right)^{y}.$$

I dette tilfellet er $r = \alpha$ og $p = \frac{\alpha}{1+\alpha}$, og forventningen og variansen blir lik

$$E(Y_{\text{NB1}}) = \frac{rp}{1-p} = \mu\alpha \ , \quad \text{Var}(Y_{\text{NB1}}) = \frac{rp}{(1-p)^2} = \mu\alpha(1+\alpha).$$

Variansen er altså en lineær funksjon av forventningen, mens i NB2-fordelingen er variansen, $\mu(1+\alpha\mu)$, en kvadratisk funksjon av forventningen.

2.2 Negativ binomisk historisk

Blaise Pascal var en fransk matematiker som levde på 1600-tallet, og var i brevkorrespondanse med Pierre de Format med på å bidra til sannsynlighetsteori slik den er kjent i dag. Han skal blant annet ha utledet spesialtilfeller av den negativ binomiske fordelingen som ble utgitt i et verk

i 1679. Selv om fordelingen kalles *Pascal fordeling*, skal det ha vært Fermat som var den første til å benytte seg av den, i tilfeller der sannsynlighetene for suksess og fiasko er like.

Allerede tidlig på 1700-tallet skal en annen fransk matematiker, Pierre de Montmort, med kjennskap til Pascal og med andre matematikere som blant annet Jacob Bernoulli i sin bekjentskapskrets, ha brukt negativ binomisk fordeling til å finne en løsning av det klassiske sannsynlighetsproblemet *problem of points*. Dette skrivet kom ut i 1713 og er den første kjente publikasjonen av den negativ binomiske fordelingen. En mer moderne versjon av problemet og hans forslag til løsning på det er beskrevet i en matematisk historiebok, som ble utgitt i 1865 av en britisk matematiker, Isaac Todhunter. Den går ut på at to spillere ønsker å oppnå et gitt antall poeng, der kun ett poeng kan oppnås i hver omgang av spillet. Spiller A ønsker å oppnå m poeng før spiller B oppnår n poeng, og sannsynligheten for å få poeng i hver omgang er p for spiller A og $1-p=q$ for spiller B. Spillet vil da garantert være avgjort i løpet av $m+n-1=r$ omganger. Spiller A kan vinne hele spillet på enten m omganger, $m+1$ omganger, osv. opp til $(m+n-1)$ omganger. Den totale sannsynligheten for at spiller A vinner spillet blir da en sum av negativ binomiske sannsynligheter:

$$
\begin{aligned}
P_A = {} & P_A((m-1)\text{ poeng på de første }(m-1)\text{ omgangene}) \cdot P_A(\text{poeng på omgang } m) \\
& + P_A((m-1)\text{ poeng på de første } m \text{ omgangene}) \cdot P_A(\text{poeng på omgang } (m+1)) \\
& \vdots \\
& + P_A((m-1)\text{ poeng på de første }(m+n-2)\text{ omgangene}) \cdot P_A(\text{poeng på omgang } (m+n-1)) \\
= {} & \binom{m-1}{m-1} p^{m-1}p + \binom{m}{m-1} p^{m-1}qp + \binom{m+1}{m-1} p^{m-1}q^2p + \cdots + \binom{m+n-2}{m-1} p^{m-1}q^{n-1}p \\
= {} & \sum_{y=0}^{n-1} \binom{y+m-1}{m-1} p^m q^y \\
= {} & p^m(1 + mq + m(m+1)q^2 + \cdots + \frac{(r-1)!}{(m-1)!(n-1)!}q^{n-1})
\end{aligned}
$$

Den siste formelen her er den samme som står oppført i historieboken[17] s.97.

Før den første utledningen var på plass, var det enda flere som kom borti tilfeller av negativ binomisk fordelte observasjoner, uten å definere selve sannsynlighetsfordelingen. En som er verdt å nevne i den forbindelse, er den britiske matematikeren og statistikeren William Gosset. Med sitt pseudonavn *student*, fikk han i 1907 utgitt et skriv [16] i det kjente statistiske tidsskriftet *Biometrika*. Innholdet gjaldt resultater fra hans laboratoriearbeid på et bryggeri, som involverte opptelling av gjærceller ved hjelp av et instrument kalt et haemocytometer. Som første tilnærming brukte han Poissonfordelingen til å representere antall celler i et stort område ved å registrere antall forekomster i et mindre område, når en væske med celler ble uniformt fordelt utover platen i instrumentet. Han utførte opptelling med fire ulike løsninger av væske, og sammenlignet deretter resultatene av prediksjonene med de faktiske. I to av disse forsøkene fant han at binomialformelen hadde negative fortegn, og at variansen var større enn forventningen.

Den første formuleringen av en negativ binomisk fordeling er det derimot den britiske statistikeren Udny Yule som skal ha stått for. Denne formuleringen ble basert på en artikkel fra 1910 som handlet om antall dødsfall i en gruppe som var utsatt for en gitt sykdom.
I 1920 utledet Greenwood og Yule [7] den negativ binomiske fordelingen som sannsynligheten for å observere y fiaskoer inntil r suksesser var oppnådd- som er den samme utledningen som i uttrykk (2.2). Eggenberger og Polya kom i 1923 fram til en negativ binomisk fordeling med utgangspunkt i en Poissonfordeling med gammafordelt intensitet.

I løpet av det siste århundret har ulike utledninger av en negativ binomisk fordeling blitt utformet, som blant annet invers binomisk og fra geometriske rekker. Parallellt med nye utledninger har statistiske teknikker blitt utviklet for å kunne utføre regresjonsanalyse på datasett.

2.3 Negativ binomisk regresjonsmodell

I dette delkapitlet skal noen av de negativ binomiske regresjonsmodellene som finnes beskrives. Hilbe [9] lister opp over 20 forskjellige negativ binomiske regresjonsmodeller som kan brukes med moderne statistikkprogrammer. En av disse er NB2-modellen, som kan utledes som enten en Poisson-Gamma mixture modell eller som en eksponensiell modell. En annen modell som også tar utgangspunkt i en vanlig negativ binomisk fordeling er *NB-C-modellen.* Denne tar utgangspunkt i at fordelingen tilhører familien av generaliserte modeller, der fordelingen er på eksponensiell form;

$$f(y_i, \theta_i) = \exp\left(a(y)b(\theta) + c(\theta) + d(y)\right) \tag{2.7}$$

Her er θ den eneste parameteren i fordelingen, og a, b, c og d er kjente funksjoner.

Resten av regresjonsmodellene Hilbe nevner er modeller som blant annet er basert på justeringer av en negativ binomisk fordeling eller av regresjonsligningen.

Ulike programmeringsverktøy som for eksempel R, STATA og SAS har innebygde funksjoner som lar oss estimere parametre MED mange av de ulike negativ binomiske modellene som finnes. I denne oppgaven vil verktøyet R bli brukt. Innebygde pakker og funksjoner som kan brukes i ulike modeller vil bli nevnt eller beskrevet i løpet av dette kapitlet, og videre i oppgaven.

Oppgaven vil i videre kapitler bruke NB2-modellen, og det vil derfor bli fokusert på hvordan denne er bygget opp i dette kapitlet. Først vil regresjonsligninger bli satt opp og forklart. Deretter vil likelihoodfunksjonen og uttrykk for partiellderiverte av denne bli funnet. Videre vil NB-C-modellen bli beskrevet og sammenlignet med NB2-modellen. Til slutt vil det bli nevnt litt om noen av de andre modellen Hilbe forteller om.

2.3.1 NB2-modell

Vi antar at et datasett består av n uavhengige observasjoner av Y; $y_i, i = 1, \cdots, n$. For å kunne forklare variasjonen i observasjonene, innføres forklaringsvariabler som kan forårsake endringer i datasettet. Først blir forventningen til y_i funnet, uttrykt ved regresjonsparametre, ved hjelp av en lineær prediktor:

$$\boldsymbol{X}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_2 X_{iq}.$$

Her er $\boldsymbol{X}_i$ en vektor med de uavhengige forklaringsvariablene til y_i, og $\boldsymbol{\beta}$ er en vektor med tilhørende regresjonskoeffisienter og eventuelt et konstantledd β_0, som skal estimeres. I lineære modeller settes forventningen lik den lineære prediktoren, $\mu_i = \boldsymbol{X}_i^T\boldsymbol{\beta}$. I så tilfelle vil forventningen være lik konstantleddet hvis ingen forklaringsvariabler påvirker verdien av y_i.

Forventningen i en negativ binomisk fordeling må være positiv, og ved tilpassing av denne som en lineær funksjon, vil noen verdier av forklaringsvariablene, $\boldsymbol{X}_i$ kunne resultere i negative verdier av forventningen til y_i. For å ta høyde for slike tilfeller brukes derfor logaritmisk linkfunksjon i regresjonsligningen,

$$\log(\mu_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}, \tag{2.8}$$

som gir μ_i positive verdier for alle verdier av $\boldsymbol{X}_i$ og $\boldsymbol{\beta}$. Den inverse linkfunksjonen gir da forventningen uttrykt ved den lineære prediktoren.

$$\mu_i = \exp(\boldsymbol{X}_i^T\boldsymbol{\beta})$$

En regresjonsmodell med utgangspunkt i fordelingen i (2.4) og med linkfunksjon (2.8), blir kalt en *NB2 modell*. Ved bruk av NB2-fordelingen blir det blant annet mulighet til å estimere datasett der spredningen i observasjonene som opprinnelig er antatt Poissonfordelt er stor, og der det er observerbart at variansen ikke er lik, men heller større enn forventningen.

For å kunne utføre regresjonsanalyse må vi finne estimatene av regresjonskoeffisientene, det vil si verdiene av koeffisientene som gir observasjonene $y_1, ..., y_n$ størst mulig sannsynlighet gitt forventningsverdien. Vi trenger da en funksjon for sannsynlighetsfordelingen til disse observasjonene, som kalles likelihoodfunksjonen. Jeg finner denne ved å ta utgangspunkt i NB2-fordelingen (2.4), og jeg starter da med å sette fordelingen på eksponensiell form.

$$f(y;\mu,\alpha) = \exp\left(y\log\tfrac{\alpha\mu}{1+\alpha\mu} + \tfrac{1}{\alpha}\log(\tfrac{1}{1+\alpha\mu}) + \log\begin{pmatrix} y+\frac{1}{\alpha}-1 \\ \frac{1}{\alpha}-1 \end{pmatrix} \right) \tag{2.9}$$

Siden observasjonene er antatt å være uavhengige, kan vi bruke produktet av sannsynlighetene for hver observasjon for å finne den totale sannsynligheten for datasettet. Likelihoodfunksjonen til observasjonene y_i, med dispersjonsparameteren som en konstant verdi, og forventningen μ, som

varierer blir da

$$\begin{aligned} \mathrm{L}_{\mathrm{NB2}}(\mu_i; y_i, \alpha) &= \prod_{i=1}^{n} f(y_i; p, \alpha) \\ &= \prod_{i=1}^{n} \exp\left\{ \log\begin{pmatrix} y_i + \frac{1}{\alpha} - 1 \\ \frac{1}{\alpha} - 1 \end{pmatrix} - \frac{1}{\alpha}\log(1+\alpha\mu_i) + y_i \log\left(\tfrac{\alpha\mu_i}{1+\alpha\mu_i}\right) \right\}. \end{aligned}$$

Logaritmisk uttrykk av likelihoodfunksjonen til y_i er lettere å bruke videre ved estimering siden vi da får delt uttrykket opp i n ledd istedenfor faktorer.

$$\begin{aligned} \log\mathrm{L}_{\mathrm{NB2}}(\mu_i; y_i, \alpha) &= \log \prod_{i=1}^{n} f(y_i; \mu_i, \alpha) \\ &= \sum_{i=1}^{n} \left\{ y_i \log\left(\tfrac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \left(\tfrac{1}{\alpha}\right)\log(1+\alpha\mu_i) + \log\Gamma\left(y_i + \tfrac{1}{\alpha}\right) - \log\Gamma(y_i+1) - \log\Gamma\left(\tfrac{1}{\alpha}\right) \right\} \end{aligned}$$

Vi er ute etter å finne estimater av regresjonskoeffisientene til forklaringsvariablene i linkfunksjonen $\log(\mu_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}$. Disse estimatene er verdiene av koeffisientene som gir den største verdien av likelihoodfunksjonen, og kalles sannsynlighetsmaksimeringsestimater, SME. Med $\mu_i = \exp(\boldsymbol{X}_i^T\boldsymbol{\beta})$ får vi det endelige uttrykket for loglikelihoodfunksjonen som skal maksimeres.

$$\begin{aligned} \log\mathrm{L}_{\mathrm{NB2}}(\boldsymbol{\beta}; y_i, \alpha) = \sum_{i=1}^{n} \Bigg\{ & y_i \log\left(\tfrac{\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}\right) - \frac{1}{\alpha}\log(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})) \\ & + \log\Gamma\left(y_i + \tfrac{1}{\alpha}\right) - \log\Gamma(y_i+1) - \log\Gamma\left(\tfrac{1}{\alpha}\right) \Bigg\}. \end{aligned} \tag{2.10}$$

For å finne estimatene, må vi derivere uttrykket med hensyn på de aktuelle regresjonskoeffisientene en etter en, og deretter sette uttrykkene vi får lik null.

$$\begin{aligned} \frac{\partial}{\partial\beta_j}\log\mathrm{L}_{\mathrm{NB2}} = \sum_{i=1}^{n} \Bigg\{ & y_i \left(\frac{1}{\frac{\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}}\right)\left(\frac{\alpha X_{ij}\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})) - \alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})\alpha X_{ij}\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))^2}\right) \\ & - \left(\frac{1}{\alpha}\right)\frac{\alpha X_{ij}\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})} \Bigg\} \end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^{n} \left\{ y_i \frac{X_{ij}(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})) - \alpha X_{ij}\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})} - \frac{X_{ij}\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})} \right\} \\ &= \sum_{i=1}^{n} \frac{X_{ij}(y_i - \exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})} \end{aligned} \tag{2.11}$$

En vektor med disse partielle deriverte uttrykkene kalles *score-verdi.*

Hvis verdien til α er ukjent, må denne også estimeres. Vi finner da et uttrykk for den partiell-

deriverte av loglikelihoodfunksjonen med hensyn på denne.

$$\begin{aligned}\frac{\partial}{\partial\alpha}\text{logL}_{\text{NB2}} &= \sum_{i=1}^{n}\left\{ y_i\left(\frac{1}{\frac{\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}}\right)\left(\frac{\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))-\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))^2}\right)\right. \\ &\quad \left. +\frac{1}{\alpha^2}\log(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))-\frac{1}{\alpha}\frac{\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}+\Psi(y_i+\frac{1}{\alpha})-\Psi(\frac{1}{\alpha})\right\} \\ &= \sum_{i=1}^{n}\left\{\frac{1}{\alpha^2}\left(\log(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))+\frac{\alpha(y_i-\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}\right)+\Psi(y_i+\frac{1}{\alpha})-\Psi(\frac{1}{\alpha})\right\},\end{aligned}$$

der Ψ er den deriverte av $\log\Gamma()$.

Ved å derivere loglikelihoodfunksjonen enda en gang med hensyn på de forskjellige parametrene, får vi en matrise vi kaller *Hessian*. Den negative verdien av denne kaller vi observert informasjonsmatrise og forventet verdi av denne kalles forventet informasjonsmatrise. For å finne estimat som maksimerer likelihoodfunksjonen ved bruk av for eksempel en numerisk algoritme, kalt Newton-Raphson, trenger vi både scorevektoren og hessianmatrisen. Algoritmen blir gjennomgått i et senere delkapittel. Vi finner videre uttrykket for den andre deriverte med hensyn på de ulike parametrene.

$$\begin{aligned}\frac{\partial^2}{\partial\beta_j\partial\beta_k}\text{logL}_{\text{NB2}} &= \sum_{i=1}^{n}\left(\frac{-X_{ij}X_{ik}\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))-X_{ij}(y_i-\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})\alpha X_{ik}\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))^2}\right) \\ &= \sum_{i=1}^{n}\frac{-X_{ij}X_{ik}\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))}{(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))^2}\end{aligned}$$

Hvis dispersjonsparameteren må estimeres, trenger vi uttrykket for den andrederiverte av loglikelihoodfunksjonen med hensyn på denne også.

$$\begin{aligned}\frac{\partial^2}{\partial\beta_j\partial\alpha}\text{logL}_{\text{NB2}} &= \sum_{i=1}^{n}\frac{-X_{ij}(y_i-\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))^2},\ \text{og} \\ \frac{\partial^2}{\partial\alpha^2}\text{logL}_{\text{NB2}} &= \sum_{i=1}^{n}\left\{-\frac{1}{\alpha^3}\left(\frac{\alpha(1+2\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))(y_i-\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))-\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))}{(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))^2}\right.\right. \\ &\quad \left.\left. +2\log(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))\right)+\Psi'(y_i+\frac{1}{\alpha})-\Psi'(\frac{1}{\alpha})\right\}.\end{aligned}$$

Her er Ψ' den andrederiverte av $\log\Gamma()$.

Når vi nå har funnet første- og andreordens deriverte av likelihoodfunksjonen kan vi bruke numerisk optimering for å finne de estimerte parametrene vi ønsker. I en GLM-modell har vi en kanonisk linkfunksjon hvis $a(y) = y$ i fordelingen i (2.7). Vi ser av fordelingen i (2.9) at ved bruk av $\log(\mu_i)$ som linkfunksjon, så blir ikke fordelingen på kanonisk form. Metoden Newton-Raphson er da en

løsning for å finne parameterestimater og standardfeil. Algoritmen i metoden regner ut standardfeilene ved hjelp av observert informasjonsmatrise.

Vi kan også bruke GLM-metoden IRLS på NB2-modellen. Algoritmen i denne metoden bruker forventet informasjonsmatrise istedenfor observert informasjonsmatrise i beregningen av standardfeil. Vi må derfor justere denne algoritmen. I tillegg må dispersjonsparameteren innføres i modellen som en gitt konstant. En løsning på dette hvis vi ikke kjenner verdien av α, er å estimere den separat og deretter putte den inn i modellen. Pakkene *gamlss* og *pscl* i R lar oss bruke NB2-modellen til regresjonsanalyse.

2.3.2 NB-C-modell

I denne modellen tar vi utgangspunkt i fordelingen på eksponensiell form, (2.9).

$$f(y;\mu,\alpha) = \exp\left(y\log\tfrac{\alpha\mu}{1+\alpha\mu} + \tfrac{1}{\alpha}\log(\tfrac{1}{1+\alpha\mu}) + \log\begin{pmatrix} y+\frac{1}{\alpha}-1 \\ \frac{1}{\alpha}-1 \end{pmatrix} \right), \tag{2.12}$$

og at denne tilhører familien av generaliserte modeller. Navnet på modellen NB-C kommer av at vi har en negativ binomisk modell med kanonisk linkfunksjon. Fra teorien om GLM ser vi at den kanoniske linkfunksjonen er lik $\eta_i = g(\mu_i) = \log(\frac{\alpha\mu_i}{1+\alpha\mu_i}) = \boldsymbol{X}_i^T\boldsymbol{\beta}$. Et uttrykk for forventningen, uttrykt ved den lineære prediktoren blir da

$$\begin{aligned} \frac{\alpha\mu_i}{1+\alpha\mu_i} &= \exp(\boldsymbol{X}_i^T\boldsymbol{\beta}) \\ \alpha\mu_i &= \frac{1}{\exp(-\boldsymbol{X}_i^T\boldsymbol{\beta})-1} \\ \mu_i &= \frac{1}{\alpha(\exp(-\boldsymbol{X}_i^T\boldsymbol{\beta})-1)}. \end{aligned}$$

Ved å sette inn disse verdiene i loglikelihoodfunksjonen til NB2-modellen i (2.10), oppnår vi loglikelihoodfunksjonen til NB-C-modellen.

$$\begin{aligned} \log\mathrm{L}_{\text{NB-C}}(\boldsymbol{\beta};y_i,\alpha) = \sum_{i=1}^{n} \Big\{ & y_i(\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}) - \left(\tfrac{1}{\alpha}\right)\log(1+\alpha\left(\frac{1}{\alpha(\exp(-\boldsymbol{X}_i^T\boldsymbol{\beta})-1))}\right) + \log\Gamma\left(y_i+\tfrac{1}{\alpha}\right) \\ & -\log\Gamma(y_i+1) - \log\Gamma\left(\tfrac{1}{\alpha}\right)\Big\} \\ = \sum_{i=1}^{n} \Big\{ & y_i(\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}) + \left(\tfrac{1}{\alpha}\right)\log(1-(\exp(-\boldsymbol{X}_i^T\boldsymbol{\beta}) + \log\Gamma\left(y_i+\tfrac{1}{\alpha}\right) \\ & -\log\Gamma(y_i+1) - \log\Gamma\left(\tfrac{1}{\alpha}\right)\Big\} \end{aligned}$$

For kanoniske modeller som NB-C, er den observerte informasjonsmatrisen lik forventet informasjonsmatrise. Metoden IRLS kan altså brukes på denne modellen, og forventet informasjonsmatrise kan brukes for å finne standardfeilene.

Pakken **MASS** i R er en av pakkene som inneholder funksjoner for å estimere modeller som er basert på GLM-familier.

2.3.3 Andre regresjonsmodeller

Hilbe tar som nevnt for seg mange ulike typer negativ binomiske regresjonsmodeller. Jeg skal starte med å vise hva som skiller NB1-modellen fra NB2-modellen. I NB1-fordelingen er forventningen lik $E(Y_{\text{NB1}}) = \mu\alpha$. Med logaritmisk linkfunksjon får vi

$$\log(\mu_i\alpha) = \boldsymbol{X}_i^T\boldsymbol{\beta}.$$

Dette gir oss den inverse linkfunksjonen

$$\mu_i = \exp(\boldsymbol{X}_i^T\boldsymbol{\beta} - \log\,\alpha).$$

Når vi setter opp likelihoodfunksjonen med den inverse linkfunksjonen, ser vi at derivasjon fører til en score-vektor som er ulik den vi kom fram til i NB2-modellen. I programmeringsverktøyet R, kan funksjoner i pakkene **COUNT** og **gamlss** brukes til å estimere regresjonskoeffisientene i NB1-modeller.

I modellen *NB-H*, som står for heterogen negativ binomisk modell, blir dispersjonsparameteren, α, parametrisert med passende forklaringsvariabler for å se om noen av disse bidrar til større spredning i datasettet enn andre. Modellen kan brukes på de tre modellene jeg allerede har skrevet om, NB2, NB-C og NB1-modellen.

Modeller med endogene forklaringsvariabler er hittil mest brukt innen økonometri. En endogen forklaringsvariabel er avhengig feilleddet i regresjonsligningen og således også avhengig av responsvariabelen. Dette resulterer i forventningsskjeve estimater av regresjonskoeffisientene. Hilbe tar for seg negativ binomiske modeller med endogene variabler og sammenligner med NB2-modellen.

Når observasjoner i et datasett har blitt registrert over lang tid innen et gitt utvalgsområde, er sannsynligheten stor for at datasettet inneholder korrelerte observasjoner. Dette må tas høyde for i en regresjonsmodell, siden vi vanligvis bruker en likelihoodfunksjon som er basert på uavhengige observasjoner. Ulike typer endringer i loglikelihoodfunksjonen og regresjonsligningen kan gjøres for å rette opp i dette. Bruk av *fixed effects*, *random effects*, og *mixed effects* er eksempler på endringer i regresjonsligningen som bidrar til egne regresjonsmodeller for slike tilfeller.

Datasett med observasjoner fra ulike felt i det virkelige liv er i praksis aldri fordelt identisk med en gitt teoretisk fordeling, som den negativ binomiske. Men uten alt for store avvik mellom den observerte- og NB2- fordelingen, kan vi oppnå god tilpasning ved bruk av en regresjonsmodell som NB2-modellen. Når den observerte fordelingen avviker mer fra NB2-fordelingen, vil det i mange tilfeller resultere i en dårlig tilpasning ved bruk av NB2-modellen. I de neste kapitlene skal jeg se på

noen modifiserte negativ binomiske regresjonsmodeller som er laget for å kunne tilpasse datasett der den observerte fordelingen avviker fra en NB2-fordeling. Avviket ligger i hvordan nullobservasjonene i datasettet er fordelt. I de siste kapitlene skal jeg ta for meg to modeller som håndterer ekstra mange nullobservasjoner i datasettet som skal tilpasses. I neste kapittel skal jeg fortelle om en regresjonsmodell som kan brukes når datasettet ikke inneholder noen nullobservasjoner.

2.3.4 Metoder for maksimering av loglikelihoodfunksjon

En negativ binomisk fordeling kan estimeres ved maksimering av likelihoodfunksjonen eller som medlem av familien av generaliserte lineære modeller, GLM. I sistnevnte tilfelle brukes en iterativ metode, Iterated Reweighted Least Squares(IRLS), der vektede minstekvadrat tilpasses i hvert steg. Det kreves da at dispersjonsparameteren er gitt, eventuelt kan den estimeres og deretter puttes inn i den generaliserte lineære modellen.

En metode for å estimere parametre ved sannsynlighetsmaksimering, som også er mye brukt i modifiserte versjoner i programmeringsverktøyene som finnes, er *Newton-Raphson*. Algoritmen i i denne metoden er basert på en utvidelse av taylorrekken til loglikelihoodfunksjonen som skal maksimeres, $f(x)$. Ved utvidelse av denne rundt a, får vi

$$f(x) \approx f(a) + (x-a)f'(a) + \frac{1}{2}(x-a)^2 f''(a).$$

Vi har altså en kvadratisk tilnærming til funksjonen, og så lenge funksjonen er konkav, så vil vi finne et maksimeringspunkt med metoden. Dette gjøres ved derivering av den tilnærmede funksjonen, for så å sette uttrykket lik null og til slutt løse for x;

$$f'(a) + (x-a)f''(a) = 0,$$

som gir oss løsningen

$$x = a - \frac{f'(a)}{f''(a)}.$$

I algoritmen settes en startverdi av x, x_0, som tenkes å være i nærheten av den som gir oss den største verdien av funksjonen. Algoritmen vil da kjøre uttrykket

$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)}$$

iterativt, inntil den konvergerer til den rette verdien av x.

Kapittel 3

Nulltrunkert negativ binomisk

I noen datasett kan vi ha tilfeller der observasjoner ser ut til å tilhøre en spesifikk fordeling hvis man ser bort fra at noen observasjoner ikke finnes i dataene. Eksempler på slike tilfeller kan være undersøkelser med interesse i fordelingen av totalt antall skader for de som allerede har opplevd én eller flere skader fra før av et gitt år, og antall døgn på sykehus for en pasient som allerede er innlagt. En negativ binomisk fordeling kan ta alle hele verdier større eller lik null. Vi vil i tilfeller som disse eksemplene ha datasett som mangler null- og eventuelt andre observasjoner, og vi sier at fordelingen er *trunkert* for verdiene som mangler. Trunkering av nullobservasjoner må tas høyde for når vi skal tilpasse observasjonene med en modell, spesielt når forventningen i datasettet er lav, siden dette kan tyde på at en større andel nullobservasjoner er tatt bort.

Kapitlet starter med en beskrivelse av en nulltrunkert negativ binomisk fordeling og dens egenskaper. Deretter blir komponentene i regresjonsmodellen forklart, og brukt i uttrykket for likelihood-funksjonen som benyttes for å finne parameterestimater. Til slutt vil vi se hvor godt modellen klarer å tilpasset datasett ved endringer i regresjonskoeffisientene.

3.1 Nulltrunkert negativ binomisk fordeling

En *nulltrunkert negativ binomisk fordeling* er en negativ binomisk fordeling justert for at den ikke har verdier lik null, altså der summen av sannsynlighetene er lik én. Justeringen gjøres ved å dividere en negativ binomisk fordeling med sannsynligheten for at en observasjon er større enn null.
Hvis vi går ut fra (2.2) får vi følgende fordeling for trunkerte verdier av Y, $y_{\text{NBtrunc}} = 1, 2, 3, ...$;

$$f(y; p, r \mid y > 0) = \frac{\binom{y+r-1}{r-1} p^y (1-p)^r}{1-(1-p)^r}.$$

For å finne forventning og varians til en nulltrunkert negativ binomisk fordeling finner vi først den

momentgenererende funksjonen til Y_{Trunc}.

$$\begin{aligned}
M_{Y_{\text{NBtrunc}}}(t) &= E(e^{tY_{\text{Trunc}}}) \\
&= \frac{1}{1-(1-p)^r}\sum_{y=1}^{\infty} e^{ty}\binom{y+r-1}{r-1}(pe^t)^y \\
&= \frac{(1-p)^r}{1-(1-p)^r}\sum_{y=1}^{\infty}\binom{y+r-1}{r-1}p^r(e^t(1-p))^y \\
&= \frac{(1-p)^r}{(1-(1-p)^r)}\frac{1-(1-pe^t)^r}{(1-pe^t)^r}\sum_{y=1}^{\infty}\frac{1}{1-(1-pe^t)^r}\binom{y+r-1}{r-1}(pe^t)^y(1-pe^t)^r \\
&= \frac{(1-p)^r}{1-(1-p)^r}\left(\frac{1}{(1-pe^t)^r}-1\right).
\end{aligned}$$

Fra denne finner vi første- og andre-deriverte uttrykk.

$$\begin{aligned}
M'_{Y_{\text{NBtrunc}}}(t) &= \frac{(1-p)^r}{1-(1-p)^r}\frac{rpe^t}{(1-pe^t)^{r+1}} \\
M''_{Y_{\text{NBtrunc}}}(t) &= \frac{(1-p)^r}{1-(1-p)^r}\frac{rpe^t(1+rpe^t)}{(1-pe^t)^{r+2}}
\end{aligned}$$

Forventningen og variansen blir da

$$\begin{aligned}
E(Y_{\text{NBtrunc}}) &= M'_{Y_{\text{Trunc}}}(0) \\
&= \frac{rp}{1-p}\frac{1}{(1-(1-p)^r)} \\
&= E(Y_{\text{NB}})\frac{1}{1-P(Y_{\text{NB}}=0)} \\
Var(Y_{\text{Trunc}}) &= M''_{Y_{\text{NBtrunc}}}(0) - (M'_{Y_{\text{Trunc}}}(0))^2 \\
&= \frac{(1-p)^r}{1-(1-p)^r}\frac{rp(1+rp)}{(1-p)^{r+2}} - (E(Y_{\text{Trunc}}))^2 \\
&= \frac{rp}{(1-p)^2} + \left(\frac{rp}{1-p}\right) - \left(\frac{1}{1-P(Y_{\text{NB}}=0)}E(Y)\right)^2 \\
&= \frac{1}{1-P(Y_{\text{NB}}=0)}E(Y_{\text{NB}}^2) - \left(\frac{1}{1-P(Y_{\text{NB}}=0)}E(Y_{\text{NB}})\right)^2
\end{aligned}$$

Hvis vi går ut fra (2.4) får vi for $y_{\text{Trunc}} > 0$;

$$f(y_{\text{NBtrunc}};\mu,\alpha) = \frac{\binom{y+\frac{1}{\alpha}-1}{\frac{1}{\alpha}-1}\left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^y}{1-\left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}}}. \tag{3.1}$$

Denne har forventning og varians

$$E(Y_{\text{NBtrunc}}) = \frac{\mu}{1-(\frac{1}{1+\alpha\mu})^{\frac{1}{\alpha}}}, \tag{3.2}$$

$$\text{Var}(Y_{\text{NBtrunc}}) = \left(\frac{\mu^2+\mu+\alpha\mu^2}{1-(\frac{1}{1+\alpha\mu})^{\frac{1}{\alpha}}}\right) - \left(\frac{\mu}{(1-\frac{1}{1+\alpha\mu})^{\frac{1}{\alpha}}}\right)^2. \tag{3.3}$$

3.2 Nulltrunkert negativ binomisk regresjonsmodell

Hvis vi har et datasett som er negativ binomisk fordelt, og fjerner alle nullobservasjoner, må vi ta høyde for dette ved valg av regresjonsmodell. Når forventningen i det negativ binomisk fordelte datasettet, μ, er stor, vil andelen nullobservasjoner være lav. Hvis nullobservasjonene deretter fjernes fra datasettet, vil vi derfor oppnå tilnærmet forventningsrette estimater av regresjonskoeffisientene ved bruk av en negativ binomisk regresjonsmodell. I tilfeller der forventningen i datasettet er lav, vil vi ha en større andel nullobservasjoner og en vanlig negativ binomisk modell vil kunne resultere i forventningsskjeve estimater. Det naturlige startpunktet er å se på komponentene i en trunkert negativ binomisk regresjonsmodell.

3.2.1 Likelihoodfunksjoner

Siden en trunkert negativ binomisk fordeling ikke tilhører familien av generaliserte modeller, må vi bruke andre metoder enn IRLS for estimering når forventningen i et negativ binomisk datasett er lav, og deretter trunkeres for nullobservasjoner. Det blir igjen tatt utgangspunkt i maksimering av likelihoodfunksjonen for å estimere parametrene. Likelihood- og loglikelihoodfunksjonen til Y_{Trunc} blir med utgangspunkt i (3.1)

$$\text{L}_{\text{NBtrunc}}(\mu_i; y_i, \alpha) = \prod_{i;y_i>0} \begin{pmatrix} y_i+\frac{1}{\alpha}-1 \\ \frac{1}{\alpha}-1 \end{pmatrix} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \left(1-\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\right)^{-1}$$

$$\begin{aligned}\text{logL}_{\text{NBtrunc}}(\mu_i; y_i, \alpha) = \sum_{i;y_i>0} \Bigg\{ & y_i \log\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \left(\frac{1}{\alpha}\right)\log(1+\alpha\mu_i) - \log\left(1-\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\right) \\ & + \log\Gamma\left(y_i+\frac{1}{\alpha}\right) - \log\Gamma(y_i+1) - \log\Gamma\left(\frac{1}{\alpha}\right)\Bigg\}.\end{aligned}$$

Vi kan også her bruke logaritmisk linkfunksjon, $\log(\mu_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}$, og får da loglikelihoodfunksjonen der forventningen er erstattet med regresjonskoeffisientene som skal estimeres.

$$\begin{aligned}\text{logL}_{\text{NBtrunc}}(\boldsymbol{\beta}; y_i, \alpha) = \sum_{i;y_i>0} \Bigg\{ & y_i \log\left(\frac{\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}\right) - \left(\frac{1}{\alpha}\right)\log(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})) \\ & - \log\left(1-\left(\frac{1}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}\right)^{\frac{1}{\alpha}}\right) + \log\Gamma\left(y_i+\frac{1}{\alpha}\right) - \log\Gamma(y_i+1) - \log\Gamma\left(\frac{1}{\alpha}\right)\Bigg\}\end{aligned}$$

$$= \text{logL}_{\text{NB2}}(\boldsymbol{\beta}; y_i, \alpha) - \log\left(1-\left(\frac{1}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}\right)^{\frac{1}{\alpha}}\right) \tag{3.4}$$

Uttrykket for loglikelihoodfunksjonen i NB2-modellen ble funnet i (2.10).

Vi finner deretter score-verdiene som skal settes lik null for å oppnå sannsynlighetsmaksimerings-estimatene, og andrederivert av likelihoodfunksjonen, som gir oss hessian-matrisen.

$$\frac{\partial}{\partial\beta_j}\text{logL}_{\text{NBtrunc}} = \frac{\partial}{\partial\beta_j}\text{logL}_{\text{NB2}} - \frac{X_{ij}\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}.$$

$$\frac{\partial^2}{\partial\beta_j\beta_k}\text{logL}_{\text{NBtrunc}} = \frac{\partial^2}{\partial\beta_j\partial\beta_k}\text{logL}_{\text{NB2}} - \frac{X_{ij}X_{ik}\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{(1+\alpha\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}))^2}.$$

Hvis dispersjonsparameteren skal estimeres trenger vi også den partiellderiverte med hensyn på denne.

For å estimere parametrene som maksimerer likelihoodfunksjonen, blir numeriske metoder benyttet i ulike programmeringsverktøy. I oppgaven brukes verktøyet R [2], som inneholder flere pakker med funksjoner som gjør det mulig å utføre regresjonsanalyse på datasett som antas å være trunkerte negativ binomisk fordelte. Et eksempel er pakken er **VGAM**, der den nulltrunkerte fordelingen i VGAM-familien kalles positiv negativ binomisk fordeling. Et annet eksempel er pakken **gamlss**, som vil bli brukt i neste delkapittel. I denne oppgaven er brukermanualen fra den offisielle nettsiden [1] for å finne informasjon om pakken og hvilke endringer som kan gjøres i standardinnstillingene i funksjonen som brukes. Navnet *gamlss* står for *Generalized Additive Models for Location, Shape and Scale*, og pakken er laget for å kunne utføre regresjon på blant annet fordelinger som ikke er medlem av den eksponensielle familien. Vi kan bruke trunkerte verdier av gamlss-familiens negativ binomiske fordeling, *NBI*, til å utføre regresjonsanalyse, og denne vil bli benyttet i neste delkapittel. Navnet *NBI* i denne pakken tilsvarer NB2 i denne oppgaven. Funksjonen *gamlss* estimerer regresjonsparametrene ved maksimering av likelihoodfunksjonen.

3.3 Effekt av endringer i koeffisientverdier

Endringer i regresjonskoeffisienter og hvordan disse påvirker de forskjellige estimatene ved bruk av en nulltrunkert regresjonsmodell vil nå bli vist ved å simulere nulltrunkerte datasett der vi kjenner de eksakte verdiene til regresjonskoeffisientene. For å sammenligne hvor gode estimatene vi får er, skal vi se på en verdi kalt *mean squared error*, MSE, som viser til bedre estimat av en parameter i en modell jo lavere dens verdi er. Verdien MSE til en estimert parameter er definert som forventningen av de kvadrerte feilleddene, $\text{E}(\hat{\theta}-\theta)^2$, der $\hat{\theta}$ er den estimerte verdien av parameteren og θ er den faktiske parameterverdien. I denne oppgaven vil MSE-verdier til koeffisientestimatene bestå av gjennomsnittet av de kvadrerte feilleddene i hver av de n simuleringene som blir utført. Uttrykket for verdien blir altså $\text{MSE}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i-\theta_i)^2$.

I oppgaven er det laget en funksjon i R, *syntetiskNB2*, som simulerer nulltrunkerte negativ binomiske

datasett for deretter å utføre regresjon på disse. Programkoden for denne er lagt ved i Tillegg A. Funksjonen kan kjøres med ulike eksakte verdier av β_0 og β_1. Dispersjonssparameteren er satt kjent i alle tilfellene, med gitt verdi, $\alpha = 0.75$. Funksjonen starter med å simulerere datasett med 100 trunkerte negativ binomisk fordelte observasjoner. Dette gjøres ved å generere negativ binomiske observasjoner ved bruk av Poisson-Gamma fordelingen, der vi velger å bruke to forklaringsvariabler med tilhørende regresjonskoeffisienter i tillegg til et konstantledd;

$$\log \mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Forklaringsvariablene X_1 og X_2 settes uniformt fordelte, siden dette vil føre til flere nullobservasjoner enn ved bruk av andre fordelinger. Ved å i tillegg bruke sentrerte verdier av disse vil man lettere se konsekvensene av endringer i regresjonskoeffisientene. Sentrering av forklaringsvariabler er mye brukt ved regresjon, siden dette gjør tolkningen av konstantleddetet enklere, i tillegg til å gi ukorrelerte parameterestimater. Nullobservasjonene fjernes etterhvert som de forekommer, men andelen registreres for hvert simulerte datasett. Deretter brukes innebygde funksjoner i R til å finne estimater for regresjonsparametrene. Simuleringen av datasett og regresjon på dette utføres $n = 5000$ ganger i funksjonen, og resulterer i like mange estimat av hver regresjonskoeffisient. Gjennomsnittet av hver av disse blir brukt, som da blir et estimat for $E(\hat{\beta}_0)$, $E(\hat{\beta}_1)$ og $E(\hat{\beta}_2)$. I tillegg beregner og skriver funksjonen ut gjennomsnittet av andelen av nullobservasjoner før trunkering i de simulerte datasettene.

For å utføre regresjon på de simulerte nulltrunkerte datasettene blir pakken **gamlss** og en av dens underpakker, **gamlss.tr** brukt. Funksjonen *gen.trun* i sistnevnte pakke lar oss generere en trunkert fordeling fra fordelinger som tilhører *GAMLLS-familien*. Én av fordelingene i denne familien er NB2-fordelingen, som går under navnet *NBI* i gamlss-pakkene. Funksjonen tar som argumenter blant annet hvilken type trunkering som ønskes, hvilken fordeling som skal ligge til grunn for den trunkerte, og navnet som skal brukes på den trunkerte fordelingen. Videre benyttes funksjonen *gamlss* for å utføre regresjon på de simulerte datasettene ved hjelp av den genererte trunkerte fordelingen. I denne funksjonen brukes følgende argumenter:

1) *formula*, en formel som består av det simulerte datasettet som det skal utføres regresjonsanalyse på og tilhørende forklaringsvariabler.
2) *family*, der den genererte trunkerte fordelingen oppnådd fra funksjonen *gen.trun* benyttes.
3) *data*, som er en dataramme med alle verdiene i *formula*.

I tillegg har vi valgt å endre på kriteriet for konvergens i algoritmen som brukes i funksjonen. I argumentet *control* har vi satt opp verdien *c.crit*, som er et mål på endringer i deviansen som tillates, fra 0.001 til 0.05 for å få flere og raskere resultater, siden det utføres estimering av et stort antall simulerte datasett i funksjonen *syntetiskNB2*.

Vi starter med å beskrive hva som skjer med koeffisientestimatene når konstantleddet β_0 varierer, og de andre koeffisientene holdes konstant. De valgte gitte verdiene av regresjonskoeffisientene er

$\beta_1 = 0.75$ og $\beta_2 = -1.25$. Noen av resultatene fra funksjonen i R har jeg samlet i Tabell 3.1. Her er det tatt med verdier av β_0 fra 0.2 og oppover. For lavere verdier er andelen av trunkerte nullobservasjoner fra de negativ binomisk genererte observasjonene over 50%, og gjennomsnittet er svært lavt. Dette fører til at vi i løpet av 5000 simulerte nulltrunkerte datasett på 100 verdier vil oppnå datasett med for få ulike verdier til at de lar seg tilpasse med en nulltrunkert regresjonsmodell.

$\boldsymbol{\beta_0}$	**Andel nuller**	$\widehat{\boldsymbol{E}(\hat{\boldsymbol{\beta}}_\mathbf{0})}$	$\widehat{\boldsymbol{E}(\hat{\boldsymbol{\beta}}_\mathbf{1})}$	$\widehat{\boldsymbol{E}(\hat{\boldsymbol{\beta}}_\mathbf{2})}$	**MSE** $\hat{\boldsymbol{\beta}}_\mathbf{0}$	**MSE** $\hat{\boldsymbol{\beta}}_\mathbf{1}$	**MSE** $\hat{\boldsymbol{\beta}}_\mathbf{2}$
0.2	0.483	0.255	0.736	-1.235	0.065	0.241	0.265
0.5	0.420	0.529	0.737	-1.238	0.0502	0.225	0.238
1	0.321	1.006	0.756	-1.234	0.0339	0.202	0.207
2	0.172	1.986	0.741	-1.244	0.0211	0.183	0.186
5	0.0197	4.980	0.751	-1.261	0.0146	0.162	0.173
7	0.00439	6.982	0.744	-1.251	0.0139	0.168	0.170
10	0.000477	9.980	0.749	-1.246	0.0141	0.164	0.166
15	0.0000099	14.980	0.750	-1.249	0.0152	0.186	0.193
20	0	20.431	0.570	-0.935	0.207	0.230	0.291

Tabell 3.1: Estimat av forventet verdi av parameterestimater, og tilhørende MSE ved endringer i β_0. Andel nuller = andel trunkerte observasjoner.

Vi ser fra tabellen at vi får færre nullobservasjoner i de negativ binomiske genererte dataene jo større den eksakte verdien av konstantleddet er.

• Hvis β_0 er relativt lav, i dette eksemplet når $\beta_0 = 0.2$ og $\beta_0 = 0.5$, ser vi at andelene av nullobservasjoner i de genererte negativ binomiske datasettene er veldig store, og at koeffisientestimatene ikke ser ut til å være helt forventningsrette, spesielt for estimatene av konstantleddet. Ved disse verdiene av konstantleddet har antallet trunkeringer blitt for stort til at den trunkerte modellen klarer å tilpasse dataene godt nok.

• Når β_0 ligger innenfor en viss skala, mellom 1 og 15, har vi tilnærmet forventningsrette estimater for β_0, β_1 og β_2. Tilhørende MSE-verdier ser ut til å bli lavere etterhvert som konstantleddet øker fra verdien 1 og opp til et visst punkt, her når $\beta_0 = 7$ og vi har da en andel av nullobservasjoner på under 0.5%. Deretter stiger MSE-verdiene igjen, men de er fremdeles relativ lave når $\beta_0 = 15$. Andelen nullobservasjoner er da tilnærmet null, men ikke lik. Den nulltrunkerte modellen ser altså ut til å tilpasse datasettene godt innenfor disse verdiene av β_0.

• Når β_0 er større, her når $\beta_0 \geq 20$, er andelen av nullobservasjoner lik null i de negativ binomiske genererte observasjonene, og de positive verdiene varierer fra lave til ekstremt høye. Verdiene av MSE for estimatene er betraktelig større, dette gjelder spesielt MSE-verdien for konstantleddet. I tillegg er $\hat{\beta}_1$ og $\hat{\beta}_2$ da forventningsskjeve. Når antall nullverdier i de genererte negativ binomiske

observasjonene er lik null, vil vi få akkurat de samme verdiene i de 5000 simulerte nulltrunkerte datasettene.

En nulltrunkert negativ binomisk regresjonsmodell bruker en annen likelihoodfunksjon for å finne sannsynlighetsmaksimeringsestimater enn en vanlig negativ binomisk modell. Vi kan i dette tilfellet, som nevnt tidligere i delkapitlet, bruke en vanlig negativ binomisk regresjonsmodell for å oppnå forventningsrette estimater av regresjonskoeffisientene.

Vi skal nå se på effekten av endringer i koeffisienten til forklaringsvariablene, eksempelvis endringer i β_1. Her settes fremdeles $\beta_2 = -1.25$, mens konstantleddet settes lik $\beta_0 = 2.0$. Resultatene fra kjøring med funksjonen *syntetiskNB2* er satt inn i Tabell 3.2. For eksakte verdier av β_1 lavere enn -12 eller større enn 12 vil vi igjen ha en altfor stor andel av nullobservasjoner og for stor spredning i de positive observasjonene fra de negativ binomisk genererte dataene til at en nulltrunkert negativ binomisk regresjonsmodell vil passe de simulerte nulltrunkerte datasettene.

$\boldsymbol{\beta_1}$	**Andel nuller**	$\widehat{\boldsymbol{E(\hat{\beta}_0)}}$	$\widehat{\boldsymbol{E(\hat{\beta}_1)}}$	$\widehat{\boldsymbol{E(\hat{\beta}_2)}}$	**MSE** $\hat{\boldsymbol{\beta}}_0$	**MSE** $\hat{\boldsymbol{\beta}}_1$	**MSE** $\hat{\boldsymbol{\beta}}_2$
-12	0.351	1.979	-12.00	-1.251	0.0332	0.821	0.255
-10	0.324	1.984	-9.992	-1.252	0.0293	0.386	0.193
-5	0.230	1.983	-4.993	-1.242	0.0220	0.238	0.196
-2	0.180	1.980	-1.994	-1.248	0.0219	0.190	0.177
0	0.170	1.988	0.0060	-1.239	0.0202	0.182	0.180
2	0.181	1.981	1.993	-1.251	0.0219	0.193	0.183
5	0.230	1.982	4.994	-1.253	0.0225	0.254	0.183
10	0.323	1.986	9.988	-1.245	0.0301	0.386	0.189
12	0.350	1.988	12.01	-1.254	0.0321	1.328	0.280

Tabell 3.2: Estimat av forventet verdi av parameterestimater, og tilhørende MSE ved endringer i β_1. Andel nuller = andel trunkerte observasjoner.

Vi ser fra tabell 3.2 at endringer i den eksakte verdien til koeffisienten β_1 fremdeles gir tilnærmet forventningsrette estimater av koeffisientene i alle tilfellene.

• Når $\beta_1 = 0$ har vi den laveste MSE-verdien av estimatene til β_0, β_1 og β_2. Andelen av nullobservasjoner i de genererte negativ binomiske observasjonene er også på sitt laveste ved denne verdien.
• Når den eksakte verdien av β_1 blir mindre eller større enn null, får vi større MSE-verdi av estimatene til alle de tre regresjonsparametrene, påvirkningen på MSE-verdier er størst hos estimatene av β_1. Vi får da også større andel nullobservasjoner i de genererte observasjonene, endringene i MSE-verdiene ser ut til å øke jevnt med endringer i andelen av nullobservasjonerβ_0 i enten positiv eller negativ retning fra verdien null.

• Når $\beta_0 = \pm 12$ ser vi at MSE-verdiene øker betydelig, spesielt for $\hat{\beta}_1$. Vi har i dette tilfellet genererte negativ binomiske observasjoner med over 35% nullobservasjoner og resterende verdier fordelt med stor spredning. Dette kan, som vi ser fra tabellen, resultere i avvik ved estimering med nulltrunkert negativ binomisk regresjonsmodell.

Vi har nå sett at ved endringer i en gitt regresjonskoeffisient, så er det MSE-verdiene til estimatene av denne som varierer mest. For å lettere se effekten av endringer i en regresjonskoeffisient kan det være interessant å se på sannsynlighetsfordelingen til estimatene ved ulike verdier av den gitte koeffisienten. For å kunne tolke formen på disse fordelingene vil blant annet to begreper benyttes, nemlig skjevhet og kurtose. Pakken **moments** i R har innebygde funksjoner for de to verdiene.

Skjevhet er et mål på asymmetri rundt gjennomsnittet i en sannsynlighetsfordeling. For et utvalg med n observasjoner, $x_1, x_2, \ldots$, er utvalgsskjevheten i R definert som

$$\gamma = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3},$$

der s er det empiriske standardavviket. Hvis verdien γ er nær null, er fordelingen tilnærmet symmetrisk. Større skjevhet i absoluttverdi medfører skjevere fordeling. Når γ er negativ er halen på fordelingen lengst mot venstre, der verdiene er under gjennomsnittet. Når vi har positiv skjevhet er halen lengst mot høyre og vi har flere verdier over gjennomsnittet enn under. Hvis $|\gamma| > 1$, så har man en veldig skjev fordeling, når $\gamma = (0.5, 1)$ er fordelingen litt skjev, mens vi har en tilnærmet normal fordeling når $|\gamma| < 0.5$.

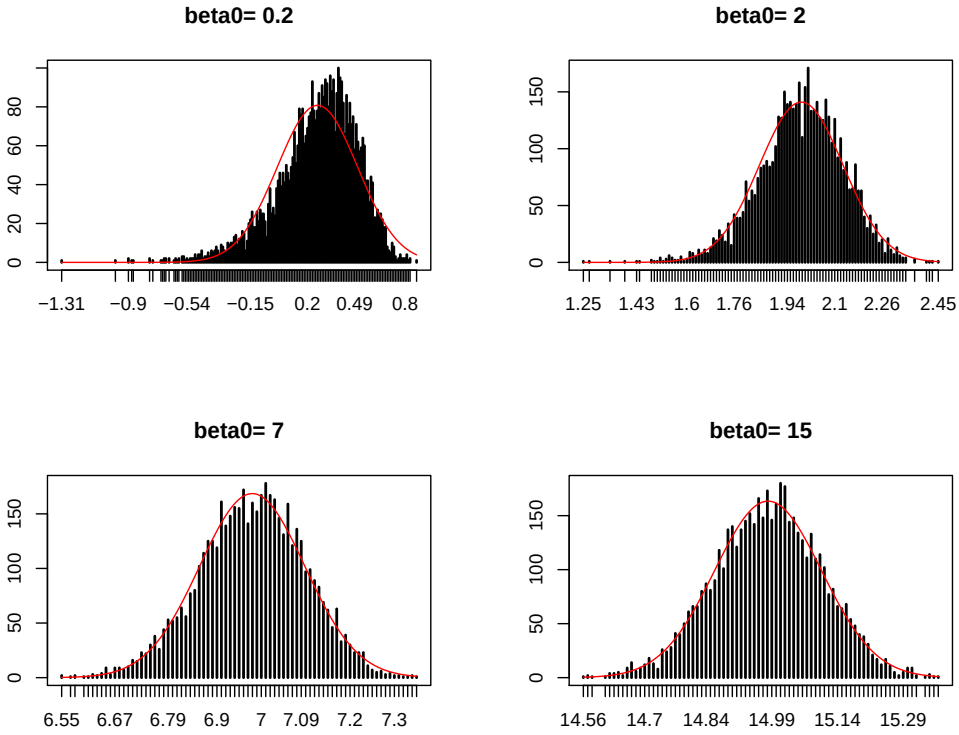
Verdien på kurtosen er et mål på haletykkelsen i en fordeling, altså et mål på mengden av observasjoner som er fordelt på verdiene som ligger et stykke unna gjennomsnittet. Pakken **moments** blir brukt til beregning av vanlig kurtose, men vil vil imidlertid trekke fra verdien 3, slik at vi får verdien som kalles *excess kurtose*;

$$\mathrm{K} = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4} - 3.$$

Denne kurtoseverdien er lettere å tolke siden en standard normal fordeling har excess kurtose lik 0. Den viser til en fordeling med tykkere hale når vi har en positiv verdi, og en mer tynnhalet fordeling når den er negativ.

Vi starter med å se på plot av fordelingen til $\hat{\beta}_0$ ved fire ulike verdier av β_0, der de andre koeffisientene holdes konstant gitt en verdi. Vi bruker den samme funksjonen, *syntetiskNB2*, for å oppnå estimater av koeffisientleddet. For å oppnå et analyse-vennlig plott av alle estimatene brukes avrunding av disse til to desimaler. Siden estimater oppnådd ved maksimering av likelihoodfunksjonen skal være asymptotisk normalfordelt, legges en normalfordelt kurve med lik forventning og standardavvik som observasjonene i datasettet til i plottene, for å lettere kunne se avvik. Plottene

i figur 3.1 viser fordelingen til estimerte verdier av konstantleddet β_0, ved ulike verdier av den eksakte gitte verdien. Verdiene til regresjonskoeffisientene β_1 og β_2 er satt til henholdsvis 0.75 og -1.25.



Figur 3.1: Fordelingen til $\hat{\beta}_0$ ved ulike verdier av β_0

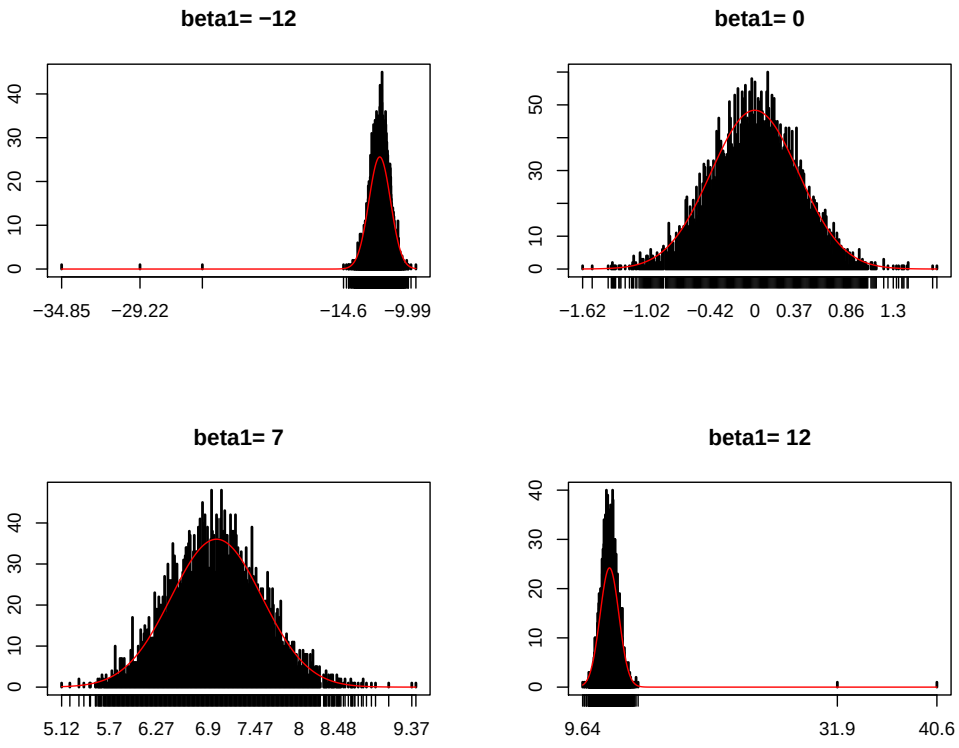
Vi kan se fra fordelingene at $\beta_0 = 0.2$ gir en skjevere fordeling av estimatene enn når verdien er større. Vi har estimater med større avvik fra gjennomsnittet på venstre side, og fordelingen ser ikke symmetrisk ut. Når $\beta_0 = 2$ ser vi fremdeles spor av skjevhet med lengre hale på venstre side av gjennomsnittet, men ikke i stor grad, og fordelingen ser mer symmetrisk ut. Ved de to største verdiene av konstantleddet ser fordelingene til estimatene ut til å være tilnærmet symmetrisk, og vi ser at de følger normalfordelingen i halene også. I tabellen 3.3 er verdiene for skjevhet og kurtose i de fire tilfellene ført opp.

	Skjevhet	**Kurtose**
$\beta_0 = 0.2$	-0.843	1.322
$\beta_0 = 2$	-0.369	0.560
$\beta_0 = 7$	-0.161	-0.072
$\beta_0 = 15$	-0.074	-0.085

Tabell 3.3: Skjevhet og kurtose til fordelinger av $\hat{\beta}_0$

Verdiene i tabellen indikerer også her at vi har en tendens til skjev fordeling av estimatene til konstantleddet ved lave verdier av den gitte verdien β_0. Skjevheten er negativ, altså har vi et større avvik fra gjennomsnittet og lengre hale på venstre side. Ved de største verdiene av konstantleddet er skjevheten nærmere null og fordelingene blir stadig mer symmetriske. Verdiene på kurtosen viser også at vi har tykkere haler i fordelingen av estimatene ved de to laveste verdiene av β_0. Ved å kjøre simulering- og regresjonsprosessen flere ganger med den samme gitte verdien av konstantleddet har formen på fordelingen til estimatene blitt observert å kunne variere noe. Dette gjelder mest kurtoseverdien. Alt i alt er tendensen den samme, lave verdier av β_0 gir litt skjeve fordelinger av estimater av konstantleddet og større kurtoseverdi sammenlignet med større verdier av β_0. Skjevhetsverdien er imidlertid et stykke under verdien 1 når $\beta_0 = 0.2$, altså er den ikke veldig skjev ved den verdien.

På samme måte som for endringer i den gitte verdien β_0, skal vi også se på fordelingene til estimatene av β_1 når dens eksakte verdi varierer. I figur 3.2 ser vi hvordan fordelingen til $\hat{\beta}_1$ endrer seg når β_1 varierer, her er de andre parametrene igjen satt med eksakte verdier, $\beta_2 = -1.25$ og $\beta_0 = 2$.



Figur 3.2: Fordelingen til $\hat{\beta}_1$ ved ulike verdier av β_1

Vi ser at i tilfellene der $\beta_1 = \pm 12$ har vi fordelinger som i første øyekast ikke ser direkte usymmetriske ut, men som inneholder et par lave verdier som ligger langt unna de andre. Når simulerings-

og regresjonsprosessen blir kjørt flere ganger med $\beta_1 = -12$ og $\beta_1 = 12$, forekommer noen få slike avvik. Ved de andre verdiene av β_1 følger fordelingene en mer normalfordelt form. Fordelingene er konstruerte som stolpediagrammer ved å avrunde estimatene ned til to desimaler. Mellomrommet mellom hver stolpe vises ikke i denne størrelsen av plottene, og det forklarer hvorfor det kan se ut som om estimatene har større total sannsynlighet enn 1. Her skal vi også se på verdiene av skjevheten og kurtosen i de fire tilfellene.

	Skjevhet	**Kurtose**
$\beta_1 = -12$	-6.04	185.5
$\beta_1 = 0$	-0.029	0.183
$\beta_1 = 7$	0.1404	0.051
$\beta_1 = 12$	11.23	357.5

Tabell 3.4: Skjevhet og kurtose til fordelinger av $\hat{\beta}_1$

Fra tabell 3.4 ser vi at verdiene for skjevhet og kurtose er lave, og at vi har en tilnærmet normalfordeling av estimatene når $\beta_1 = 0$ og $\beta_1 = 7$. Når $\beta_1 = 7$ har vi litt større spredning og flere verdier av estimatene i halen på fordelingen, noe som forklarer hvorfor MSE-verdien er litt større her enn når $\beta_1 = 0$. Når $\beta_1 = \pm 12$ er verdiene for skjevhet og kurtose ekstremt høye. Dette er på grunn av de få ekstremverdiene vi har i fordelingene. Hvis disse fjernes fra fordelingen oppnår vi for $\beta_1 = -12$ skjevhet- og kurtoseverdi på $S = -0.0695$ og $K = 0.0363$ og for $\beta_1 = 12$ blir disse verdiene $S = 0.1404$ og $K = 0.0221$. Estimatene er altså tilnærmet normalfordelte når vi ser vekk ifra få ekstremverdier av estimatene.

Hilbe sammenligner i sine eksempler bruk av forskjellige regresjonsmodeller på nulltrunkerte negativ binomiske data for å blant annet vise at man ikke bør bruke vanlig NB2 modell når fordelingen har en lav forventning og er uten nullobservasjoner. For å understreke betydningen av verdien til de ulike regresjonsparametrene har det i dette delkapitlet blitt fokusert på effekten av endringene av disse. Ved å endre på den eksakte verdien til konstantleddet så vi at de som førte til simulerte negativ binomiske datasett med andel nullobservasjoner større enn 0, men mindre enn 50%, førte til nulltrunkerte datasett der nulltrunkerte negativ binomisk regresjonsmodellen tilpasset datasettene godt. Verdiene av β_0 i intervallet (0.2,15) ga oss denne andelen. Dette er tall som fremkom ved eksakte, gitte verdier av β_1 og β_2 på henholdsvis 0.75 og -1.25. Ved endringer av den eksakte verdien til regresjonskoeffisienten β_1 ble de andre parameterverdiene satt til $\beta_0 = 2.0$ og $\beta_2 = -1.25$. Da førte verdier i intervallet $-12 \leq \beta_1 \leq 12$, som resulterte i rundt 35% eller færre nullobservasjoner i de genererte negativ binomiske observasjonene, til datasett der nulltrunkert negativ binomisk regresjonsmodell ga gode estimater av regresjonsparametrene.

Kapittel 4

ZINB OG ZANB

Innledningsvis i oppgaven ble det nevnt at vi ofte har datasett der andelen av nullobservasjoner ikke samstemmer med den forventede når observasjonene er antatt fordelt med en gitt fordeling. Dette kapitlet skal handle om de to regresjonsmodellene ZINB og ZANB, som er laget for å tilpasse datasett med slike avvik i observasjonene. Disse modellene håndterer også større spredning i observasjonene enn ved bruk av en vanlig NB-modell. Videre i oppgaven vil NB2-fordelingen bli brukt og referert til som negativ binomisk eller NB.

Først vil fordelingene og prosessene som modellene tar utgangspunkt i bli beskrevet. Deretter blir komponenter i regresjonsligninger etterfulgt av uttrykkene for likelihoodfunksjonene satt opp. Til teori om sannsynlighetsfordelingene og modellene er bøkene av Zuur m.fl.[20] og Hilbe [9] blitt benyttet. Vi vil også se på hvordan egenskapene i modellen påvirkes av endringer gitte verdier av regresjonskoeffisientene.

Algoritmer som blir brukt for å finne parameterestimater numerisk vil deretter bli beskrevet, i tillegg til noen av måleverdiene som brukes for å velge mellom ZINB og ZANB etter utført regresjon. Til slutt vil vi fortelle litt modellenes bakgrunn.

4.1 Fordelingene til ZINB og ZANB

4.1.1 Zero altered negativ binomisk fordeling

I en zero altered fordeling er sannsynligheten for at en vilkårlig observasjon i et datasett er null eller positiv, styrt av en binær prosess. Sannsynlighetsfordelingen i denne prosessen kan enten være binomisk, geometrisk eller negativ binomisk. Videre i oppgaven er denne antatt binomisk fordelt, tilsvarende uttrykk (2.1), og prosessen blir heretter referert til som den binomisk prosessen i ZANB. Sannsynlighen for at en vilkårlig observasjon er null settes lik π, $\mathrm{P}(y = 0) = \pi$, og vi får utrykket

$$\text{binomisk prosess: } \begin{cases} 1 & (y=0) \text{ med sannsynlighet } \pi \\ 0 & (y>0) \text{ med sannsynlighet } (1-\pi). \end{cases}$$

Denne prosessen styrer altså i tillegg sannsynligheten for at observasjonen er positiv. Sannsynlighetsfordelingen til de eksakte positive observasjonene er styrt av en negativ binomisk fordeling i ZANB-modellen.

Sannsynligheten for at en vilkårlig observasjon har en eksakt positiv verdi er betinget på at den ikke er en nullobservasjon;

$$P(Y = y_{\text{pos}}) = P(Y > 0) \cdot P(Y = y_{\text{pos}} \mid Y > 0).$$

Det må derfor benyttes en nulltrunkert negativ binomisk fordeling, som vist i uttrykk (3.1), i beregningen av disse. Fordelingen til ZANB blir

$$f_{\text{ZANB}}(y; \pi, \mu, \alpha) = \begin{cases} \pi & , y = 0 \\ (1-\pi) f_{\text{NBtrunc}}(y; \mu, \alpha) & , y > 0 \end{cases} . \tag{4.1}$$

Sannsynligheten for å få en bestemt verdi over null er altså lik produktet av sannsynligheten for å først i det hele tatt få en positiv verdi og sannsynligheten for å få den eksakte verdien i en nulltrunkert negativ binomisk fordeling.

Størrelsen på π, som kan være mellom null og èn, avgjør hvor mange nullobservasjoner vi har. Når sannsynligheten for å få en nullobservasjon ved trekning fra den binomiske fordelingen er lik sannsynligheten for å få en nullobservasjon fra en negativ binomisk fordeling, $\pi = P(Y_{\text{NB}} = 0)$, er fordelingen til ZANB lik en negativ binomisk fordeling. Ved trekninger fra fordelingen til ZANB når verdien til π ligger i intervallet $(0, P(Y_{\text{NB}} = 0))$, er det mulig å oppnå datasett med *færre* nullobservasjoner fra fordelingen til ZANB enn det man ville fått fra en NB fordeling. Det kan oppnås datasett ved trekninger fra fordelingen til ZANB som inneholder *flere* nullobservasjoner enn fra en NB-fordeling når $\pi = (P(Y_{\text{NB}} = 0), 1)$.

Modellen refereres ofte til som en type *hurdle modell*. Navnet *hurdle* kommer av at det kun kan oppnås en *spesifikk* positiv verdi fra fordelingen etter at et *hinder* er passert, nemlig å ikke først få en verdi lik null i den binære prosessen.

Forventningen og variansen til ZANB kan beregnes ved hjelp av de betingete momentene i den negativ binomiske prosessen. Uttrykk (3.2) gir

$$E(Y_{\text{ZANB}} \mid Y_{\text{ZANB}} > 0) = \frac{\mu}{1 - (\frac{1}{1+\alpha\mu})^{\frac{1}{\alpha}}} \quad ,$$

$$\text{Var}(Y_{\text{ZANB}} \mid Y_{\text{ZANB}} > 0) = \frac{\mu^2 + \mu + \alpha\mu^2}{1 - \left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}}} - \left(\frac{\mu}{(1 - \frac{1}{1+\alpha\mu})^{\frac{1}{\alpha}}}\right)^2 \text{ og}$$

$$E((Y_{\text{ZANB}} \mid Y_{\text{ZANB}} > 0)^2) = \frac{\mu^2 + \mu + \alpha\mu^2}{1 - \left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}}}$$

Ved hjelp av disse momentene, og setningene om dobbel forventning- og varians, kan vi oppnå den ubetingede forventningen og variansen i en ZANB-fordeling. Utfallsrommet i den binære prosessen kalles A i disse utregningene; $A = \{0 \Leftrightarrow Y_{\text{ZANB}} = 0, 1 \Leftrightarrow Y_{\text{ZANB}} > 0\}$.

$$\begin{aligned}
E(Y_{\text{ZANB}}) &= E_A[E(Y_{\text{ZANB}} \mid A)] \\
&= \sum_A E(Y_{\text{ZANB}} \mid A) \cdot P(A) \\
&= E(Y_{\text{ZANB}} \mid Y_{\text{ZANB}}{=}0)P(Y_{\text{ZANB}}{=}0) + E(Y_{\text{ZANB}} \mid Y_{\text{ZANB}}{>}0)P(Y_{\text{ZANB}}{>}0) \\
&= 0 \cdot \pi + \frac{\mu}{1 - (\frac{1}{1+\alpha\mu})^{\frac{1}{\alpha}}} \cdot (1 - \pi) \\
&= (1 - \pi) \cdot \frac{\mu}{1 - (\frac{1}{1+\alpha\mu})^{\frac{1}{\alpha}}} \\
\text{Var}(Y_{\text{ZANB}}) &= \text{Var}_A[E(Y_{\text{ZANB}} \mid A)] + E_A[\text{Var}(Y_{\text{ZANB}} \mid A)] \\
&= E_A[(E(Y_{\text{ZANB}} \mid A))^2] - (E_A[E(Y_{\text{ZANB}} \mid A)])^2 + E_A[\text{Var}(Y_{\text{ZANB}} \mid A)] \\
&= \sum_A (E(Y_{\text{ZANB}} \mid A))^2 P(A) - (E(Y_{\text{ZANB}}))^2 + \sum_A \text{Var}(Y_{\text{ZANB}} \mid A)P(A) \\
&= \sum_A (E(Y_{\text{ZANB}} \mid A)^2)P(A) - (E(Y_{\text{ZANB}}))^2 \\
&= (1 - \pi) \left(\frac{\mu^2 + \mu + \alpha\mu^2}{1 - (\frac{1}{1+\alpha\mu})^{\frac{1}{\alpha}}} \right) - (1 - \pi)^2 \left(\frac{\mu}{(1 - \frac{1}{1+\alpha\mu})^{\frac{1}{\alpha}}} \right)^2
\end{aligned}$$

4.1.2 Zero inflated negativ binomsk fordeling

Lambert [11] skrev en artikkel i 1992 om ZI-poisson regresjon, der hun også nevner at de samme egenskapene til ZI-fordelte data gjelder for andre fordelinger, som i dette tilfellet med negativ binomisk fordeling. Vi skal nå se på sannsynlighetsfordelingen i en zero inflated negativ binomisk modell.

I ZINB-modellen har vi også en todelt sannsynlighetsfordeling som skiller mellom nullobservasjoner og positive observasjoner, slik vi hadde i fordelingen til ZANB. Men i denne modellen blir nullobservasjoner styrt av *to* prosesser, framfor kun av én prosess som i ZANB. Den ene styrer andelen av nullobservasjoner som vi oppnår fra en negativ binomisk prosess. Den andre prosessen styrer mengden av de resterende nullobservasjonene som måtte forekomme i . Sistnevnte gruppe av nullobservasjoner kaller vi *strukturelle nuller*.

Sannsynligheten for at en vilkårlig observasjon, y, fra fordelingen til ZINB er en strukturell null settes lik ϕ. Vi har altså en

$$\text{binomisk prosess: } \begin{cases} 1 & (y{=}\text{strukturell 0}) \text{ med sannsynlighet } \phi \\ 0 & (y{\neq}\text{strukturell 0}) \text{ med sannsynlighet } (1 - \phi) \end{cases},$$

som styrer sannsynligheten for å oppnå en strukturell null.

Dersom den vilkårlige observasjonen ikke er en strukturell null, er den eksakte verdien, $y \geq 0$, styrt av en telleprosess; den negativ binomiske. Vi kan altså oppnå en nullobservasjon også i denne prosessen, men det er da betinget på at vi ikke fikk en strukturell null i den binomiske prosessen. Sannsynligheten for å få en vilkårlig verdi, $y \geq 0$, som ikke er en strukturell null blir

$$P(Y = y) = P(y \neq \text{strukturell } 0) \cdot P(Y_{\text{NB}} = y).$$

Vi har således både en binomisk og en negativ binomisk prosess som styrer sannsynligheten for nullobservasjoner. Den totale sannsynligheten for at en vilkårlig observasjon fra fordelingen til ZINB er en nullobservasjon blir da summen av sannsynligheten for å få en strukturell null, og sannsynligheten for å få en nullobservasjon fra en negativ binomisk prosess sett at vi ikke fikk en strukturell null. Sannsynligheten for å få en positiv verdi blir produktet av sannsynligheten for at vi ikke har en strukturell null og sannsynligheten for at vi får denne verdien i en negativ binomisk fordeling.

$$f_{\text{ZINB}}(y; \phi, \mu, \alpha) = \begin{cases} \phi + (1-\phi)\left(\frac{1}{\mu\alpha+1}\right)^{\frac{1}{\alpha}} & , y = 0 \\ (1-\phi) f_{\text{NB}}(y; \mu, \alpha) & , y > 0 \end{cases} \tag{4.2}$$

Uttrykket for $f_{\text{NB}}(y; \mu, \alpha)$ fant vi i (2.4). Når $\phi = 0$ er det ikke mulig å oppnå strukturelle nuller, og fordelingen til ZINB er identisk med en NB-fordeling. Når $\phi > 0$ kan vi få strukturelle nuller, og det blir da på bekostning av en NB-fordelt observasjon som enten er større eller lik null. Et datasett som er ZINB-fordelt kan altså inneholde flere nullobservasjoner enn et som er NB-fordelt.

Forventningen og variansen i den binomiske delen av modellen er

$$E(Y_{\text{BIN}}) = \phi \text{ og } \text{Var}(Y_{\text{BIN}}) = \phi(1-\phi).$$

Forventningen og variansen i NB-delen ble funnet i uttrykk (2.5). Vi har altså

$$E(Y_{\text{ZINB}} \mid Y_{\text{BIN}} = 0) = \mu \text{ og } \text{Var}(Y_{\text{ZINB}} \mid Y_{\text{BIN}} = 0) = \mu + \alpha\mu^2.$$

Den ubetingede forventningen og variansen til ZINB finner vi ved å bruke setningene om dobbel forventning og dobbel varians, og her settes utfallsrommet i den binomisk prosessen lik D;
$D = \{0 \Leftrightarrow Y_{\text{ZINB}} \neq \text{strukturell null}, 1 \Leftrightarrow Y_{\text{ZINB}} = \text{strukturell null}\}$. Uttrykket for forventningen blir da følgende:

$$\begin{aligned} E(Y_{\text{ZINB}}) &= E_D[E(Y_{\text{ZINB}} \mid D)] \\ &= \sum_D E(Y_{\text{ZINB}} \mid D) P(D) \\ &= E(Y_{\text{ZINB}} \mid Y_{\text{BIN}} = 0) P(Y_{\text{BIN}} = 0) + E(Y_{\text{ZINB}} \mid Y_{\text{BIN}} = 1) P(Y_{\text{BIN}} = 1) \\ &= \mu(1-\phi) + 0 \cdot \phi \\ &= \mu(1-\phi) \end{aligned}$$

Uttrykket for variansen finner vi på tilsvarende måte:

$$\begin{aligned}
\mathrm{Var}(Y_{\mathrm{ZINB}}) &= \mathrm{Var}_D[E(Y_{\mathrm{ZINB}} \mid D)] + E_D[\mathrm{Var}(Y_{\mathrm{ZINB}} \mid D)] \\
&= E_D[(E(Y_{\mathrm{ZINB}} \mid D))^2] - (E_D[E(Y_{\mathrm{ZINB}} \mid D)])^2 + E_D[\mathrm{Var}(Y_{\mathrm{ZINB}} \mid D)] \\
&= \sum_D (E(Y_{\mathrm{ZINB}} \mid D))^2 P(D) - (E(Y_{\mathrm{ZINB}}))^2 + \sum_D \mathrm{Var}(Y_{\mathrm{ZINB}} \mid D) P(D) \\
&= (E(Y_{\mathrm{ZINB}} \,|\, Y_{\mathrm{BIN}} = 0))^2 P(Y_{\mathrm{BIN}} = 0) + (E(Y_{\mathrm{ZINB}} \,|\, Y_{\mathrm{BIN}} = 1))^2 P(Y_{\mathrm{BIN}} = 1) - (E(Y_{\mathrm{ZINB}}))^2 \\
&\quad + \mathrm{Var}(Y_{\mathrm{ZINB}} \mid Y_{\mathrm{BIN}} = 0) P(Y_{\mathrm{BIN}} = 0) + \mathrm{Var}(Y_{\mathrm{ZINB}} \mid Y_{\mathrm{BIN}} = 1) P(Y_{\mathrm{BIN}} = 1) \\
&= \mu^2(1-\phi) + 0^2 \cdot \phi - (1-\phi)^2\mu^2 + (\mu + \alpha\mu^2)(1-\phi) + 0 \cdot \phi \\
&= \mu(1-\phi)(1 + \mu(\phi + \alpha))
\end{aligned}$$

Vi ser også her at i tilfeller der sannsynligheten for strukturelle nuller er null, $\phi = 0$, så får vi lik forventning og varians som i en NB-fordeling. I Zuur m.fl [20] er uttrykkene for forventningen og variansen satt opp uten utregningen. Uttrykket for variansen i en ZINB-fordeling ser imidlertid ikke ut til å være korrekt.

Et datasett som er ZINB-fordelt, (4.2), inneholder parametrene ϕ, μ og α. Ved å generere observasjoner fra fordelingen der to av disse parametrene holdes konstant, og den tredje varierer, finner vi hver enkelt parameter sin effekt på den totale andelen av nullobservasjoner i datasettet.
• En økning i verdien av μ fører til færre nullobservasjoner. Dette er naturlig siden en økning av forventningen i NB-delen gir lavere sannsynlighet for nullobservasjoner fra denne prosessen.
• En økning i verdien av ϕ fører til flere nullobservasjoner, noe vi også kan se fra fordelingsfunksjonen siden forventningen til de strukturelle nullene øker.
• En økning i verdien av α fører til flere nullobservasjoner. Dispersjonsparameteren er kun en del av NB-prosessen i ZINB-fordelingen, og påvirker derfor kun utfallene av nullobservasjonene fra denne prosessen. I kap. 2 så vi at parameteren er direkte knyttet til spredningen av observasjonene i et datasett; vi får større spredning ved økning av parameterverdien, sett at vi har konstante gitte verdier av de andre regresjonsparametrene. Når forventningen i prosessen er satt konstant, fører større spredning i observasjonene til flere nullobservasjoner.

Ved sammenligning av sannsynlighetsfordelingene til ZANB og ZINB ,(4.1) og (4.2), ser man at disse er like når den totale sannsynligheten for nullobservasjoner i de to fordelingene er like.

$$\pi = \phi + (1-\phi)\left(\frac{1}{\mu\alpha + 1}\right)^{\frac{1}{\alpha}}.$$

Regresjonsmodellene ZINB og ZANB derimot, er to ulike modeller. I ZANB innføres ett sett med forklaringsvariabler for nullobservasjoner, og et annet sett for de positive observasjonene. I ZINB innføres også to sett med forklaringsvariabler, det ene for de strukturelle nullobservasjonene og det andre for NB-observasjonene, der både positive- og nullobservasjoner kan forekomme i sistnevnte.

4.2 Regresjonsmodeller

I programmeringsverktøyet R finnes ulike pakker med funksjoner som blant annet lar oss generere ZANB- og ZINB-fordelte datasett, og utføre regresjonsanalyse med modeller som tar utgangspunkt i disse fordelingene. Maksimering av likelihoodfunksjonen til den aktuelle fordelingen brukes ofte for å finne parameterestimatene. I dette delkapitlet skal regresjonsligninger, og deretter likelihoodfunksjonene som skal maksimeres, beskrives for modellene ZINB og ZANB.

4.2.1 ZANB-modell

Regresjonsmodeller som er basert på ZA-fordelte data er også kjent som *to-delte* modeller. Fordelingen til en zero altered negativ binomisk variabel, Y_{ZANB}, er delt i én binomisk og én nulltrunkert negativ binomisk prosess. Tilhørende forventninger i de to prosessene resulterer i to regresjons-ligninger. Vi har allerede sett på komponentene i en nulltrunkert negativ binomisk regresjonsmodell. Da ble logaritmisk linkfunksjon benyttet, $\log(\mu_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}$, som ga uttrykket

$$\mu_i = \exp(\boldsymbol{X}_i^T\boldsymbol{\beta}) \tag{4.3}$$

for forventningen. For binomiske modeller er logistisk regresjon mest brukt. Regresjonsligningen blir da

$$\text{logit}(\pi_i) = \log\frac{\pi_i}{1-\pi} = \boldsymbol{Z}_i^T\boldsymbol{\gamma},$$

der $\boldsymbol{Z}_i$ er en vektor med de uavhengige forklaringsvariablene til π_i som påvirker utfallet av forekomster av nullobservasjoner i den binomiske prosessen, og $\boldsymbol{\gamma}$ er vektoren med tilhørende regresjonskoeffisienter som skal estimeres. Det er ikke nødvendigvis de samme faktorene som avgjør utfallet av en nullobservasjon eller de eksakte positive observasjonene. Forklaringsvariablene i den binomiske prosessen kan altså være ulik de i den nulltrunkerte, og det benyttes derfor en annen notasjon for å skille disse. Forventningsuttrykket i den binomiske prosessen, uttrykt ved den lineære prediktoren blir $\boldsymbol{\gamma}_i^T\boldsymbol{Z}$.

$$\begin{aligned}
\frac{\pi_i}{1-\pi_i} &= \exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma}) \\
\pi_i &= \exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma}) - \pi_i\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma}) \\
\pi_i(1+\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})) &= \exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})
\end{aligned}$$

$$\pi_i = \frac{\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})}{1+\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})} \tag{4.4}$$

Heretter i oppgaven vil de to prosessene i ZANB-modellen bli referert til som den binomiske og den negativ binomiske prosessen. Det er da underforstått at NB-prosessen i ZANB-modellen tar utgangspunkt i en nulltrunkert negativ binomisk fordeling.

Siden observasjonene i fordelingen til ZANB også er antatt å være uavhengige, kan vi igjen bruke

produktet av sannsynlighetene for hver observasjon for å finne en funksjon for den totale sannsynligheten for datasettet. Vi har nå to faktorer i funksjonen, den ene for simultanfordelingen til alle nullobservasjonene og den andre for de positive verdiene. Likelihoodfunksjonen til observasjonene y_i, med konstant verdi av parameteren α, og forventningene π og μ, som varierer med observasjonene, blir i dette tilfellet

$$\begin{aligned} \mathrm{L}_{\mathrm{ZANB}}(\pi_i, \mu_i; y_i, \alpha) &= \prod_{i;y_i=0} f_{\mathrm{ZANB}}(y_i) \prod_{i;y_i>0} f_{\mathrm{ZANB}}(y_i) \\ &= \prod_{i;y_i=0} \pi_i \prod_{i;y_i>0} (1-\pi_i) f_{\mathrm{NBtrunc}}(y_i, \mu_i, \alpha). \end{aligned}$$

Uttrykket blir deretter delt i faktorer som kun består av forventningen i enten den binomiske eller den nulltrunkerte fordelingen. Vi ser da at den ene faktoren er lik likelihoodfunksjonen til en nulltrunkert negativ binomisk fordeling.

$$\mathrm{L}_{\mathrm{ZANB}}(\pi_i, \mu_i; y_i, \alpha) = \prod_{i;y_i=0} \pi_i \prod_{i;y_i>0} (1-\pi_i) \prod_{i;y_i>0} f_{\mathrm{NBtrunc}}(y_i, \mu_i, \alpha)$$

Ved å erstatte forventningene i de to prosessene, π_i og μ_i, med uttrykkene som inneholder de tilhørende lineære prediktorene, får vi likelihoodfunksjonen uttrykt ved regresjonsparametrene. Uttrykket for loglikelihoodfunksjonen til en nulltrunkert NB ble funnet i (3.4).

$$\mathrm{L}_{\mathrm{ZANB}}(\boldsymbol{\gamma}, \boldsymbol{\beta}; y_i, \alpha) = \prod_{i;y_i=0} \frac{\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})}{1+\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})} \prod_{i;y_i>0} \frac{1}{1+\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})} \cdot \mathrm{L}_{\mathrm{NBtrunc}}(\boldsymbol{\beta}; y_i, \alpha)$$

Loglikelihoodfunksjonen til ZANB blir da

$$\begin{aligned} \mathrm{logL}_{\mathrm{ZANB}}(\boldsymbol{\gamma}, \boldsymbol{\beta}; y_i, \alpha) &= \sum_{i;y_i=0} \big(\log(\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})) - \log(1+\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma}))\big) - \sum_{i;y_i=1,2,\ldots} \log(1+\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})) \\ &\quad + \mathrm{logL}_{\mathrm{NBtrunc}}(\boldsymbol{\beta}; y_i, \alpha) \\ &= \sum_{i;y_i=0} \boldsymbol{Z}_i^T\boldsymbol{\gamma} - \sum_{i;y_i\geq 0} \log(1+\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})) + \mathrm{logL}_{\mathrm{NBtrunc}}(\boldsymbol{\beta}; y_i, \alpha) \end{aligned} \tag{4.5}$$

Vi ser at de to første leddene i det endelige uttrykket for loglikelihoodfunksjonene inneholder alene den lineære komponenten i den binomiske prosessen, $\boldsymbol{Z}_i^T\boldsymbol{\gamma}$. Det siste leddet, den nulltrunkerte NB loglikelihoodfunksjonen, inneholder kun den lineære komponenten i den positive, nulltrunkerte prosessen. Funksjonen kan altså deles i to

$$\mathrm{logL}_{\mathrm{ZANB}}(\boldsymbol{\gamma}, \boldsymbol{\beta}; y_i, \alpha) = \mathrm{logL}_{\mathrm{ZANB_{BIN}}}(\boldsymbol{\gamma}; y_i) + \mathrm{logL}_{\mathrm{NBtrunc}}(\boldsymbol{\beta}; y_i, \alpha),$$

der maksimering kan utføres separat for de to prosessene. Verdiene av SME som vi oppnår ved optimering av disse, fører altså til den maksimerte verdien av hele uttrykket. I kapittel 3.2.1 fant vi første- og andreordens deriverte av likelihoodfunksjonen til en trunkert negativ binomisk modell

og en beskrivelse på hvordan vi ved hjelp av disse kan oppnå estimater for regresjonsparametrene $\boldsymbol{\beta}$. For å finne estimatene i den binomiske prosessen, må vi maksimere den delen av likelihood-funksjonen som inneholder tilsvarende regresjonsparametrene, $\boldsymbol{\gamma}$. Dette kan gjøres på samme måte som tidligere, ved å sette første- og andreordens deriverte uttrykk for likelihoodfunksjonen lik null, og deretter løse disse numerisk.

$$\begin{aligned}
\frac{\partial}{\partial\gamma_j}\text{logL}_{\text{ZANB}_{\text{BIN}}} &= \sum_{i;y_i=0} Z_{ij} - \sum_{i;y_i\geq 0}\frac{Z_{ij}\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})}{1+\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})} \\
\frac{\partial^2}{\partial\gamma_j\partial\gamma_k}\text{logL}_{\text{ZANB}_{\text{BIN}}} &= -\sum_{i;y_i\geq 0}\frac{Z_{ij}Z_{ik}\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})(1+\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})) - Z_{ij}\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})Z_{ik}\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})}{(1+\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma}))^2} \\
&= -\sum_{i;y_i\geq 0}\frac{Z_{ij}Z_{ik}\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma})}{(1+\exp(\boldsymbol{Z}_i^T\boldsymbol{\gamma}))^2}
\end{aligned}$$

4.2.2 ZINB-modell

I denne modellen har vi også to prosesser med forventninger som gir oss to ulike regresjonsligninger. I kapittel 2.3.2 ble komponentene i en negativ binomisk regresjonsmodell beskrevet. Det samme vil gjelde for NB-delen i ZINB, med bruk av logaritmisk linkfunksjon, $\log(\mu_i) = \boldsymbol{B}_i^T\boldsymbol{\rho}$. Her er $\boldsymbol{B}_i$ en vektor med forklaringsvariablene til observasjonene som er styrt av den negativ binomiske prosessen. I den binomiske prosessen, som styrer utfallet av strukturelle nuller, bruker vi igjen en logistisk linkfunksjon.

$$\text{logit}(\phi_i) = \log\frac{\phi_i}{1-\phi_i} = \boldsymbol{G}_i^T\boldsymbol{\omega},$$

der $\boldsymbol{G}_i$ er vektoren med forklaringsvariablene til de strukturelle nullene og $\boldsymbol{\omega}$ er vektoren med tilhørende regresjonsparametre. Uttrykkene for forventningene i ZINB-modellen blir

$$\mu_i = \exp(\boldsymbol{B}_i^T\boldsymbol{\rho}) \tag{4.6}$$

$$\phi_i = \frac{\exp(\boldsymbol{G}_i^T\boldsymbol{\omega})}{1+\exp(\boldsymbol{G}_i^T\boldsymbol{\omega})}. \tag{4.7}$$

Vi bruker også her produktet av sannsynlighetene for hver observasjon, gitt forventningsverdier, for å finne en funksjon for den totale sannsynligheten for datasettet. Likelihoodfunksjonen til observasjonene y_i, med α som en konstant gitt verdi, og forventningene ϕ og μ, som parametre, blir lik

$$\begin{aligned}
\text{L}_{\text{ZINB}}(\phi_i,\mu_i;y_i,\alpha) &= \prod_{i;y_i=0} f_{\text{ZINB}}(y_i)\prod_{i;y_i>0} f_{\text{ZINB}}(y_i,\mu_i,\alpha) \\
&= \prod_{i;y_i=0}\left[\phi_i + (1-\phi_i)\left(\tfrac{1}{\mu_i\alpha+1}\right)^{\frac{1}{\alpha}}\right]\prod_{i;y_i>0}(1-\phi_i)f_{\text{NB}}(y_i;\mu_i,\alpha)
\end{aligned}$$

Deretter settes uttrykkene for forventningene inn, med tilhørende lineære prediktorer, og forenkling utføres.

$$\begin{aligned}\mathrm{L}_{\mathrm{ZINB}}(\boldsymbol{\rho},\boldsymbol{\omega};y_i,\alpha) &= \prod_{i;y_i=0}\left[\frac{\exp(\boldsymbol{G}_i^T\boldsymbol{\omega})}{1+\exp(\boldsymbol{G}_i^T\boldsymbol{\omega})}+\left(\frac{1}{1+\exp(\boldsymbol{G}_i^T\boldsymbol{\omega})}\right)\left(\frac{1}{\alpha\exp(\boldsymbol{B}_i^T\boldsymbol{\rho})+1}\right)^{\frac{1}{\alpha}}\right]\prod_{i;y_i>0}\frac{f_{\mathrm{NB}}(y_i;\boldsymbol{\rho},\alpha)}{1+\exp(\boldsymbol{G}_i^T\boldsymbol{\omega})}\\ &= \prod_{i;y_i\geq 0}\frac{1}{1+\exp(\boldsymbol{G}_i^T\boldsymbol{\omega})}\prod_{i;y_i=0}\left(\exp(\boldsymbol{G}_i^T\boldsymbol{\omega})+\left(\frac{1}{\alpha\exp(\boldsymbol{B}_i^T\boldsymbol{\rho})+1}\right)^{\frac{1}{\alpha}}\right)\prod_{i;y_i>0}f_{\mathrm{NB}}(y_i;\boldsymbol{\rho},\alpha)\end{aligned}$$

Det endelige uttrykket for loglikelihoodfunksjonen til ZINB blir

$$\begin{aligned}\log\mathrm{L}_{\mathrm{ZINB}}(\boldsymbol{\rho},\boldsymbol{\omega};y_i,\alpha) = &\sum_{i;y_i\geq 0}-\log((1+\exp(\boldsymbol{G}_i^T\boldsymbol{\omega}))+\sum_{i;y_i=0}\log\left(\exp(\boldsymbol{G}_i^T\boldsymbol{\omega})+\left(\frac{1}{\alpha\exp(\boldsymbol{B}_i^T\boldsymbol{\rho})+1}\right)^{\frac{1}{\alpha}}\right)\\ &+\sum_{i;y_i>0}\log f_{\mathrm{NB}}(y_i;\boldsymbol{\rho},\alpha),\end{aligned}\tag{4.8}$$

der uttrykket i den siste summen er lik (2.10), bortsett fra at den lineære prediktoren er erstattet med $\boldsymbol{B}_i^T\boldsymbol{\rho}$. Leddet i midten av uttrykket for loglikelihoodfunksjonen inneholder den lineære prediktoren til både den binomiske- og NB- prosessen. Vi kan altså ikke estimere regresjonsparametrene $\boldsymbol{\rho}$ og $\boldsymbol{\omega}$ hver for seg, slik vi kunne med regresjonsparametrene i ZANB. Uttrykket for partiellderiverte av loglikelihoodfunksjonen i ZINB blir

$$\frac{\partial}{\partial\rho_j}\log\mathrm{L}_{\mathrm{ZINB}} = \sum_{i;y_i=0}\frac{-B_{ij}\exp(\boldsymbol{B}_i^T\boldsymbol{\rho})(1+\exp(\boldsymbol{B}_i^T\boldsymbol{\rho}))^{-1}}{1+\exp(\boldsymbol{G}_i^T\boldsymbol{\omega})(1+\alpha\exp(\boldsymbol{B}_i^T\boldsymbol{\rho}))^{\frac{1}{\alpha}}}+\sum_{i;y_i>0}\frac{B_{ij}(y_i-\exp(\boldsymbol{B}_i^T\boldsymbol{\rho}))}{1+\alpha\exp(\boldsymbol{B}_i^T\boldsymbol{\rho})},$$

der uttrykket i den siste summen ble funnet ved hjelp av (2.11).

Modeller med ZI-fordelte variabler blir også mye referert til som *mixture modeller*. Sistnevnte modell tar utgangspunkt i en blanding av n vektede fordelinger, $f(x;\boldsymbol{\lambda},\boldsymbol{\theta}) = \sum_{i=1}^n \lambda_i f_i(x;\theta_i)$, der λ_i angir hvor stor vekt fordelingen f_i skal ha, og har akkumulert verdi lik én; $\sum_{i=1}^n \lambda_i = 1$. Det har hittil i litteraturen vært mest bruk av mixture modeller i tilfeller der de vektede komponentene har den samme sannsynlighetsfordelingen, men med ulike verdier for parametrene. Hvis definisjonen av mixture modeller tillater bruk av ulike fordelinger for $f_i, i = 1\ldots n$, så er ZINB-modellen en mixture modell, med $\lambda_1 = \phi,\ \lambda_2 = (1-\phi), f_1(x) = 1$ og $f_2(x;\theta_2) = f_{\mathrm{NB}}(y;\mu)$.

4.2.3 Effekt av endringer i koeffisientverdier

Ulike verdier av regresjonskoeffisienter og konstantledd påvirker forventningsverdiene i prosessene, i tillegg til at andelen av nullobservasjoner et datasett forventes å inneholde vil variere. I oppgaven blir det senere utført regresjonsanalyse på simulerte datasett med bruk av kun èn forklaringsvariabel, som blir satt lik i begge prosesser i begge modellene. Vi vil da ha kun én regresjonskoeffisient i hver av regresjonsligningene i de to modellene. Dette gir et enklere sammenligningsgrunnlag for ZINB og ZANB, ved å sette verdiene til regresjonskoeffisientene i NB-prosessene lik hverandre i de to modellene, og tilsvarende med regresjonskoeffisientene i den binomiske prosessen.

Det vil også være enklere å se effekten endringer i koeffisientverdiene har på de ulike egenskapene i et datasett. Ved å sette enten konstantleddet eller regresjonskoeffisienten konstant gitt en verdi, vil man lettere kunne se hvordan egenskapene påvirkes ved endringer i den andre. Det blir i tillegg mulig å se hvilke tilfeller av endringer som gjør modellene ZINB og ZANB mer like eller ulike.

Gitt verdi av konstantledd

Når konstantleddet i de ulike prosessene har en gitt konstant verdi, for enkelthets skyld lik null, vil uttrykkene for forventningene i de ulike prosessene bestå av følgende ligninger.

	ZINB	ZANB
NB-prosess :	$\mu_{\text{ZINB}} = e^{\rho_1 X}$	$\mu_{\text{ZANB}} = e^{\beta_1 X}$
binomisk prosess :	$\phi = \frac{e^{\omega_1 X}}{1 + e^{\omega_1 X}}$	$\pi = \frac{e^{\gamma_1 X}}{1 + e^{\gamma_1 X}}$

Dersom forklaringsvariabelen har en positiv verdi, $X > 0$, vil en økning av verdien til en regresjonskoeffisient gi større forventet verdi i NB-prosessen i tilhørende modell. Hvis verdien til forklaringsvariabelen er negativ, vil en økning av koeffisientverdien resultere i lavere forventningsverdi i NB-prosessen. Dette er lett å se ved å sette $X = 1$ for positive verdier, og $X = -1$ for negative verdier av forklaringsvariabelen. Et eksempel med forventningsverdier fra NB-prosessen i ZINB-modellen følger i tabell 4.1.

	$\rho_1 = -1$		$\rho_1 = 0$		$\rho_1 = 1$
$X = -1$	e	$>$	1	$>$	$\frac{1}{e}$
$X = 1$	$\frac{1}{e}$	$<$	1	$<$	e

Tabell 4.1: Forventningsverdier i NB-prosess ved endringer av regresjonskoeffisienten i ZINB.

Det samme gjelder for tilsvarende endringer i regresjonskoeffisienten, β_1, i ZANB-modellen.

For den binomiske prosessen er endringer i forventningsverdien også like for de to modellene når verdien til regresjonskoeffisienten varierer. Dersom verdien til forklaringsvariabelen er negativ, vil negative verdier av regresjonskoeffisienten gi forventningsverdi under 0.5. Positive verdier av koeffisienten vil gi forventet verdi over 0.5. I tabell 4.2 blir endringene i ZINB vist ved ulike verdier for både negativ og positiv forklaringsvariabel. Identiske endringer i regresjonskoeffisienten, γ_1, i ZANB vil gi tilsvarende effekt på forventningen i den tilhørende binomiske prosessen.

	$\omega_1 = -1$	$\omega_1 = -0.5$	$\omega_1 = 0$	$\omega_1 = 0.5$	$\omega_1 = 1$
$X = -1$	0.731	0.622	0.5	0.378	0.269
$X = 1$	0.269	0.378	0.5	0.622	0.731

Tabell 4.2: Forventningsverdier i binomisk prosess ved endringer av regresjonskoeffisienten i ZINB.

Forventningen i den binomiske prosessen i ZANB, π, tilsvarer sannsynligheten for at en vilkårlig observasjon i et datasett er en nullobservasjon. Egenskapene i et ZANB-datasett kan derfor lett tolkes ved endringer i regresjonskoeffisientene. Eksempelvis vil en økning i koeffisientverdien γ_1 i tilfeller der forklaringsvariabelen er positiv, tilsvare at nullsannsynligheten i et datasett øker. Når koeffisientverdien i NB-prosessen, β_1, øker, vil forventningsverdien i de positive observasjonene vil bli større.

I ZINB-modellen tilsvarer forventningen i den binomiske prosessen, ϕ, sannsynligheten for å oppnå en strukturell nullobservasjon. For å se hvilke konsekvenser endringer i regresjonskoeffisienten har på den totale nullsannsynligheten, brukes uttrykk (4.2).

$$P(y_{\text{ZINB}} = 0) = \phi + (1 - \phi)\left(\frac{1}{\mu_{\text{ZINB}}\alpha + 1}\right)^{\frac{1}{\alpha}}.$$

Fra uttrykket ser vi at den totale nullsannsynligheten er avhengig av verdien til forventningen i NB-prosessen. Dersom forventningsverdien i NB-prosessen øker, vil sannsynligheten for å oppnå en nullobservasjon fra en NB-fordeling, $P(y_{\text{NB}} = 0) = \left(\frac{1}{\mu_{\text{ZINB}}\alpha+1}\right)^{\frac{1}{\alpha}}$, bli lavere, men alltid ligge i intervallet (0,1). Ved ulike endringer i forventingsverdiene i de to prosessene, vil egenskapene i et ZINB-datasett følge ulike trender. Ved å se på effekten av endringer i koeffisientverdiene når verdien på forklaringsvariabelen er negativ og positiv hver for seg, ser vi følgende ved hjelp av tabellene 4.1 og 4.2:

•$X = 1$: En økning i verdien til regresjonskoeffisienten gir som tidligere nevnt større forventningsverdi i NB-prosessen. Når μ_{ZINB} øker i intervallet $(0, 1)$, vil $P(y_{\text{NB}} = 0)$ få verdier i intervallet $(1,\ 0.5)$. Dersom verdien på $P(y_{\text{NB}})$ er stor, det vil si nær 1, vil den totale nullsannsynligheten i et datasett være stor uansett hvor stor koeffisientverdien ω_1 og således ϕ er; $P(y_{\text{ZINB}} = 0) \to 1$. Hvis verdien på $P(y_{\text{NB}} = 0)$ er nærmere 0.5, vil $P(y_{\text{ZINB}} = 0)$ variere mer med ulike verdier av ϕ. Stigende verdier av ϕ gir større total nullsannsynlighet, og dette oppnås når ω_1 vokser. Når μ_{ZINB} øker i intervallet $(1, \infty)$, vil verdien til $P(y_{\text{NB}} = 0)$ ligge i intervallet $(0.5,\ 0)$. Ved verdier av $P(y_{\text{NB}} = 0)$ nær 0, vil den totale nullsannsynligheten være tilnærmet lik sannsynligheten for strukturelle nuller: $P(y_{\text{ZINB}} = 0) \to \phi$. For større verdier av $P(y_{\text{NB}} = 0)$ vil $P(y_{\text{ZINB}} = 0)$ igjen variere mer med ulike verdier av ϕ, der større verdier av ϕ gir større total nullsannsynlighet.
•$X = -1$: En økning i ρ_1 gir lavere forventningsverdi for observasjoner som ikke er strukturelle nuller i et datasett. Her vil reduksjon i verdiene til μ_{ZINB} i intervallet $(1, 0)$ gi $P(y_{\text{NB}} = 0) \in (0.5, 1)$, og i intervallet $(\infty, 1)$ blir $P(y_{\text{NB}} = 0) \in (0, 0.5)$. Vi får igjen større total nullsannsynlighet i et datasett når ϕ blir større, det vil si når den eksakte verdien til ω_1 blir lavere.

Ved endringer i regresjonskoeffisienter kan vi altså få tilfeller av ZINB-fordelte datasett der den totale nullsannsynligheten blir større, samtidig som forventningsverdien til observasjonene som ikke er strukturelle nuller øker. Dette vil være tilfelle ved stigende verdier av ρ_1 og ω_1 når forklarings-

variabelen har en positiv verdi, og ved synkende verdier av regresjonskoeffisientene når $X < 0$. Når $X > 0$ kan vi oppnå ZINB-fordelte datasett med høy total nullsannsynlighet, og samtidig høy forventningsverdi i NB-del dersom begge regresjonskoeffisientene har positiv verdi, mens det samme kan være tilfelle med negative verdier av koeffisientene når $X < 0$. I disse tilfellene vil regresjonsmodellene ZINB og ZANB være mindre like, siden den totale nullsannsynligheten alltid blir lavere når forventningen i NB-prosessen øker ved bruk av ZANB.

Vi ser også direkte fra uttrykket for den totale nullsannsynligheten i et ZINB-datasett at stigende verdier av forventningen i NB-prosessen vil gi lavere verdier av $P(y_{\text{NB}} = 0)$, som vil gjøre $P(y_{\text{ZINB}} = 0)$ mer lik ϕ. Når $X = 1$ vil altså høye positive verdier av ρ_1 føre til at ZINB-modellen tilsvarer ZANB-modellen med $\rho_1 = \beta_1$ og $\omega_1 = \gamma_1$.

Gitt verdi av regresjonskoeffisient
Dersom regresjonskoeffisientene settes lik en gitt konstant verdi, har forklaringsvariabelen like stor påvirkning på variasjonen i et datasett, uavhengig av størrelsen på forventningene i de to prosessene. Når forklaringsvariabelen har en positiv verdi, vil stigende verdier av alle konstantleddene også her føre til større forventing i begge prosessene i modellene, i tillegg til større total nullsannsynlighet i et ZINB-fordelt datasett.

I tilfellene der $\omega_1 = 0$ og $\beta_1 = 0$, eller hvis verdien på forklaringsvariabelen er null, vil uttrykkene for forventningene i de ulike prosessene bestå av konstante verdier, uavhengig av observasjonene i datasettet. Sannsynligheten for å få oppnå en nullobservasjon, gitt forventningsverdiene, blir altså konstant i likelihoodfunksjonen til begge modellene. Det samme gjelder sannsynligheten for hver eksakte positive verdi, og de to regresjonsmodellene ZINB og ZANB vil da være like.

4.3 Algoritmer for å maksimere likelihoodfunksjoner

Det finnes ulike numeriske metoder for å finne parameterestimater som maksimerer likelihoodfunksjonen i de to modellene ZINB og ZANB. Videre i oppgaven vil det bli brukt to av disse, kalt *BFGS-* og *EM-algoritmen.*

Algoritmen BFGS har fått navnet etter de fire opphavspersonene Broyden, Fletcher, Goldfarb og Shanno, som uavhengig av hverandre kom fram til den i 1970. Metoden blir sett på som den mest effektive av typen quasi-Newton algoritmer, som bruker en tilnærming av den inverse Hessianmatrisen i utregningen. Dette gjør den numeriske kalkuleringen enklere og raskere. I kapittel 2.3.4 så vi på Newton-Raphson metoden, der beregninger av både første og andre derivert av likelihoodfunksjonen bruker for å finne maksimeringsverdien. Algoritmen i BFGS-metoden, det vil si dens tilnærming av hessian-matrisen i en quasi-Newton algoritme, fører til at den oppdaterte Hessianmatrisen alltid er positiv definit. Dette garanterer videre at funksjonen, loglikelihoodfunksjonen i denne oppgaven, kan økes i retningen av Newton-steget. Detaljert beskrivelse av algoritmen kan man finne i verket til Dennis og Schnabel [6].

En annen generell optimeringsmetode, som er ofte brukt i tilfeller med "missing values", er EM-algoritmen. Denne bruker lengre tid på å konvergering enn BFGS, men kan være mer stabil. Algoritmen består av to steg med følgende generelle handlinger;
E (expectation), som beregner den betingete forventningen til likelihoodfunksjonen, gitt observert data og nåværende estimat av parametrene, og
M (maximization), som maksimerer den betingete forventningen fra E-steget med hensyn på parametrene.

For hver repetisjon av stegene oppdateres estimatene, inntil algoritmen konvergerer ifølge et satt kriterie. I en ZI-modell kan metoden brukes med utgangspunkt i den uobserverte binomiske indikatoren W, som forteller om en observasjon kommer fra den binomiske- eller NB-prosessen. Når $W = 1$ vil en nullobservasjon være strukturell, og når $W = 0$ kommer den fra en negativ binomisk fordeling. Etter at en startverdi av W er gitt eller estimert, vil EM-algoritmen starte repitisjon av stegene. Estimatet av W blir således oppdatert i E-leddet ved hjelp av de foreløpige verdiene av parameterestimatene i den binomiske- og NB-prosessen. I M-steget blir parametrene i NB-prosessen estimert fra en vektet tilpasning av standardfordelingen til observasjonene, og parametrene i den binomiske prosessen oppnås ved å tilpasse en binomisk regresjonsmodell til foreløpige estimat av indikatoren W. Lambert [11] går igjennom detaljene ved bruk av EM-algoritmen i regresjonsmodellen ZIP.

4.4 Tester og måleverdier brukt til modellvalg av ZINB og ZANB

For å konkludere med hvilken modell som er best på å tilpasse datasett i praksis, er diverse tester og måleverdier tatt i bruk. Her skal vi se på noen av de som er mye brukt i artikler som omhandler modellvalg der både ZINB og ZANB har vært inkludert.

AIC og BIC
Akaike's information criterion, AIC, er en av måleverdiene som blir hyppig brukt ved studier der valg av den beste regresjonsmodellen skal tas. Testobservatoren følger som oftest formen

$$\text{AIC} = -2(\log(\text{L})) + 2p,$$

der p er antall frie parametre i modellen inkludert konstantledd, og log(L) er den naturlige logaritmen til maksimumsverdien i den tilpassede objektet. Det andre leddet i uttrykket gir en større AIC-verdi for hver parameter som er lagt til i modellen. Observatoren kan i noen programmeringsfunksjoner ta en annen form ;

$$\text{AIC} = \frac{-2(\log(\text{L})) + 2p}{n},$$

der n er antall observasjoner i modellen. Dette må tas hensyn til ved sammenligning av verdier fra to modeller. Lave AIC-verdier i forhold til antall observasjoner indikerer at sistnevnte uttrykk for testobservatoren er brukt.

Bayesian information criterion, BIC, blir ofte kalt Schwarz information criterion. Testverdien følger som oftest formen

$$\text{BIC} = -2(\log(\text{L})) + p\ \log(\text{n}),$$

der log(L), p og n representerer de samme verdiene som nevnt tidligere. Verdien BIC ble laget som en konkurrent til AIC; den har et strengere positivt andreledd, der verdien av p log(n) vil overgå verdien til 2p når $n \geq 8$.

Modellen med lavest AIC- og BIC-verdi indikerer best evne til tilpassing av datasett. For detaljert informasjon om måleverdiene henvises lesere til Akaike [3] og Schwarz [15] sine artikler.

Vuong-test Selv om AIC- og BIC-verdier viser forskjeller i to modeller, kan vi fremdeles ikke svare på hvilke av dem som er mest signifikant. I 1989 foreslo Quang Vuong [18] en test som også kan brukes på ikke-nøstete tilpassede objekter. Det vil si at vi for eksempel kan teste modellene ZIP og ZINB eller ZAP og ZANB mot hverandre, siden NB-fordelingen har dispersjonsparameteren som en tilleggsparameter. Observatoren i Vuong-testen bruker ikke bare informasjon om utfall av nullobservasjoner, men hele fordelingen, og viser til bedre modelltilpasning der verdier av individuell loglikelihood er signifikant størst. Den følger formen

$$V = \frac{\text{LR}_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n)}{(\sqrt{n\hat{\omega}_n^2})}.$$

Her er n lik antall observasjoner, og $\text{LR}_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n)$ representerer summen av individuelle sannsynlighetsforhold mellom to modeller som tar utgangspunkt i hver sin fordeling, f og g, og som gir tilhørende estimerte parametre $\hat{\boldsymbol{\beta}}_n$ og $\hat{\boldsymbol{\gamma}}_n$;

$$\text{LR}_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n)) = \sum_{i=1}^{n} \log \frac{f(Y_i \mid x_i; \hat{\boldsymbol{\beta}}_n)}{g(Y_i \mid x_i; \hat{\boldsymbol{\gamma}}_n)}.$$

Nevneren i uttrykket for testobservatoren, $\hat{\omega^2}_n$, representerer variansen til de individuelle sannsynlighetsforholdene mellom de to modellene;

$$\hat{\omega}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\log \frac{f(Y_i \mid x_i; \hat{\boldsymbol{\beta}}_n)}{g(Y_i \mid x_i; \hat{\boldsymbol{\gamma}}_n)} \right)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} \log \frac{f(Y_i \mid x_i; \hat{\boldsymbol{\beta}}_n)}{g(Y_i \mid x_i; \hat{\boldsymbol{\gamma}}_n)} \right)^2$$

Vuong viste at observatoren er standard normalfordelt, slik at den kan testes for signifikante forskjeller mellom modellene. Modellen som tar utgangspunkt i fordeling f er bedre når $V > 1.96$ og modellen med utgangspunkt i fordeling g bedre når $V < -1.96$.

4.5 ZINB og ZANB historisk

De første regresjonsmodellene med ZA-fordelte data ble utformet av Mullahy [13]. Han kalte disse for *hurdle*-modeller etter å ha tatt utgangspunkt i en artikkel av Cragg [5]. I modellene bruker han en binær prosess i modifiseringen av nullobservasjoner, slik det ble vist i utledningen av fordelingen til ZANB tidligere. Mullahy viser kun bruk av poisson- og geometrisk fordeling i spesifiseringen av sine hurdle-modeller, men han nevner i en fotnote (side 344) at modellen også kan benyttes med andre diskrete fordelinger, som den negativ binomiske. Navnet *zero altered* ble innført av Heilbron [8] i en artikkel om en modifisert Poisson-modell, som han kaller ZAP(λ, p). Denne upubliserte artikkelen har ikke blitt funnet, men den blir hyppigt referert til i litteratur om modellen.Både Mullahy og Heilbron beskriver i disse artiklene en annen modifisert modellklasse med utgangspunkt i diskrete fordelinger, såkalte *with-zeros*, WZ-modeller. Fordelingen bak disse modellene er lik en ZI-fordeling slik den er formulert i dag. Regresjonsmodellen, derimot, mangler den lineære prediktoren til de strukturelle nullobservasjonene.

Den første til å komplettere modellen til ZI-modellen vi bruker i dag, var Lambert i en artikkel utgitt i 1992 [11]. Her blir ZIP-modellen forklart i forbindelse med modellering av antall defekter av forskjellige slag på enheter ved produksjon av ulike elektriske tavler. Lambert deler denne modelleringen i to prosesser; tilfeller av enheter som ikke har noen defekter blir plassert i kategorien *perfekte tilstander*, mens *uperfekte tilstander* inneholder produserte enheter med minst én type defekt. Ved å innføre forklaringsvariabler for sannsynligheten i den perfekte tilstanden og for forventningen i den uperfekte, blir ZIP-modellen utformet. Forfatteren bruker altså Poisson som hovedfordeling for tilpassing av prosessen som inneholder et positivt antall defekter, og benytter de samme linkfunksjonene i begge prosessene som man bruker i dag. Resultatene fra ZIP-modellen sammenlignes med resultater ved bruk av en negativ binomisk modell på det samme datasettet, uten modifiseringer av nullobservasjoner. Dette blir gjort for å sjekke om nullobservasjonene i datasettet slik de er registrert kan være et resultat av overdispersjon. Forfatteren nevner videre at en ZI-versjon av den negativ binomiske modellen som blir brukt kan være en bedre modell i et annet tilfelle, men at denne ikke ga gode nok estimater for det aktuelle datasettet.

I tiden etter disse publikasjonene av Mullahy og Lambert har interessen for nullmodifiserte modeller vokst kraftig. De refereres ofte til som opphavspersonene til henholdsvis ZA og ZI modellene generelt, selv om modellene ikke er spesifisert for de mulige kombinasjonene av diskrete fordelinger i de ulike prosessene. I jakten på å finne andre personer som kan kreditteres for å ha funnet opp ZINB og ZANB, har vi kun funnet navn på personer som delvis har brukt disse kjente artiklene til videreføring av modellene, enten i teoribøker eller i artikler med praktiske anvendelser. I mange skriv refereres det til et overblikk over ZA og ZI modeller i en bok av Cameron og Trivedi [4].

Tidligere var programvaremetoder for blandede og todelte modeller i kompliserte og til dels merkelige funksjoner, og ulike pakker ga ulike resultater. De senere årene har modellene grunnet økt popularitet blitt forbedret og dette har videre ført til et stigende antall brukere. Modellene ser ut

til å hovedsakelig ha blitt brukt på datasett innen økologi, men er nå blitt så populære at ulike felt innen blant annet samfunnsvitenskap, forskning på trafikkskader, økonometri og psykologi også bruker disse med større frekvens.

Kapittel 5

Analysegrunnlag for sammenligning av regresjonsmodellene ZINB og ZANB ved bruk i praksis

Innledningsvis i oppgaven ble det poengtert at det ikke er lett å spore opp studier om detaljerte sammenligninger av modellene ZINB og ZANB. For å avgjøre hvilken modell som passer best til diverse datasett, er ulike test- og måleverdier tatt i bruk. I artikkelen til Li m.fl [12] ble alle testverdiene som ble nevnt i slutten av forrige kapittel tatt i bruk før en konklusjon ble foretatt.

I denne oppgaven skal vi prøve å konkludere på et annet grunnlag. Ved å allerede vite hvilken fordeling datasettet tilhører, kan det være interessant å se hvor godt de to modellene tilpasser både zinb- og zanb-datasett med ulike egenskaper. Forventningene i de to prosessene i begge modellene er avhengige av forklaringsvariabler og tilhørende regresjonskoeffisienter. Ved å endre på eksakte verdier av sistnevnte parametre, mens forklaringsvariablene holdes konstant, kan vi sammenligne evnen av tilpassing av datasett med ulike verdier av forventning i de negativ binomiske prosessene i fordelingen, og ikke minst med ulike totalandeler av nullobservasjoner. Vi skal finne ut om disse endringene gir utslag i evnen til å estimere regresjonsparametrene med de to modellene ZINB og ZANB, ved å allerede vite hva de korrekte verdiene av disse parametrene er.

I dette kapitlet vil grunnlaget for regresjonsanalysen i neste kapittel bli beskrevet. Først vil funksjonen som er laget, og brukt til å oppnå de nødvendige dataene for sammenligning av modellene, bli grundig beskrevet. Valg av verdier i denne, i tillegg til restriksjoner på verdiene som brukes for å oppnå antallet parameterkombinasjoner som blir brukt videre til analyse, vil deretter bli forklart. Videre vil bruk av pakker og innebygde funksjoner som er brukt bli gjennomgått, før valg av algoritme brukt i modellene blir forsvart.

5.1 Metodebeskrivelse

For å kunne sammenligne ZINB og ZANB som regresjonsmodeller i praksis, er det i oppgaven benyttet simulerte datasett fra de to underliggende fordelingene. Programmeringsverktøyet R er benyttet ved simulering, regresjonskjøring og beregninger generelt. Det blir forklart mer om hvilke pakker og funksjoner som er brukt til dette i delkapittel 5.2. Datasettene blir heretter referert til med små bokstaver, zinb og zanb, altså etter hvilken fordeling de er generert fra, for å lettere skille dem fra modellene ZINB og ZANB. Etter at regresjon med begge modeller er utført på datasettene, kan vi sammenligne resultatene for å prøve å konkludere med blant annet hvilken modell som bør brukes ved gitte egenskaper i et datasett.

5.1.1 Analysegrunnlag

I oppgaven er det laget en funksjon, kalt *fun*, som genererer zinb- og zanb- fordelte datasett i tillegg til å utføre regresjon på samtlige av disse med både ZINB- og ZANB-modellen. Koder for oppbygging av *fun* er lagt ved i Tillegg B. Funksjonen tar de eksakte verdiene av regresjonsparametrene som skal estimeres som parametre. Siden ZINB og ZANB totalt inneholder fire regresjonsligninger, er det for enkelthets skyld valgt å bruke kun én forklaringsvariabel, $\mathbf{X}$, med tilhørende regresjonskoeffisient i tillegg til et konstantledd. Forklaringsvariabelen er satt lik for binomisk og negativ binomisk del i begge modellene. Følgende forventningsuttrykk oppnås i de ulike prosessene, fra uttrykkene (4.3), (4.4) og (4.6).

ZINB

NB-prosess: $\mu_{\text{NB,ZINB}} = \exp(\rho_0 + \rho_1 X)$

binomisk prosess: $\mu_{\text{BIN,ZINB}} = \dfrac{\exp(\omega_0 + \omega_1 X)}{1 + \exp(\omega_0 + \omega_1 X)}$

ZANB

trunkert NB-prosess: $\mu_{\text{NB,ZANB}} = \exp(\beta_0 + \beta_1 X)$

binomisk prosess: $\mu_{\text{BIN,ZANB}} = \dfrac{\exp(\gamma_0 + \gamma_1 X)}{1 + \exp(\gamma_0 + \gamma_1 X)}$

Vi ser her at med bruk av én forklaringsvariabel har vi et enklere sammenligningsgrunnlag for modellene, med én regresjonskoeffisient og ett konstantledd som skal estimeres for hver av de fire prosessene. Ved bruk av ZANB på zinb-datasett vil resultatet av regresjonskjøring i tillaget funksjon vise om modellen evner å estimere de eksakte parametrene som er satt, nærmere bestemt om estimatene $\widehat{E(\hat{\beta}_0)}, \widehat{E(\hat{\beta}_1)}, \widehat{E(\hat{\gamma}_0)}$ og $\widehat{E(\hat{\gamma}_1)}$ er nær verdiene ρ_0, ρ_1, ω_0 og ω_1. På tilsvarende måte vil ZINB-modellen brukt på zanb-datasett vise om $\widehat{E(\hat{\rho}_0)} \to \beta_0$, $\widehat{E(\hat{\rho}_1)} \to \beta_1$, $\widehat{E(\hat{\omega}_0)} \to \gamma_0$ og $\widehat{E(\hat{\omega}_1)} \to \gamma_1$. Dispersjonsparameteren blir også estimert, så vi får totalt fem parameterestimat fra hver av modellene.

De to modellene estimerer de aktuelle regresjonsparametrene med utgangspunkt i ulike prosesser

og ved bruk av ulike likelihoodfunksjoner som vist i uttrykkene (4.5) og (4.8). I ZINB inneholder NB-prosessen eventuelle nullobservasjoner som ikke er strukturelle, i tillegg til alle positive observasjoner. I ZANB er NB-prosessen trunkert for nullobservasjoner, mens den binomiske prosessen tar hånd om alle nullobservasjonene. Ved å sette forklaringsvariablene like i alle prosessene, vil disse da forklare ulike egenskaper i de to modellene. Dette vil påvirke resultatene vi får ved regresjonskjøring med ulike eksakte verdier av regresjonsparametre.

Funksjonen starter med å regne ut verdiene som er nødvendige for å generere datasett med de to fordelingene, i tillegg til å kunne utføre regresjon med begge modellene. Disse består av forventningene i de to prosessene i hver av modellene, og er altså avhengige av de eksakte parameterverdiene vi setter i funksjonen. Deretter fortsetter den med å simulere datasett med hundre zinb-genererte variable. På disse blir det utført regresjon med både ZINB og ZANB, for deretter å lagre resultater i en matrise. Prosessen med simulering og regresjonskjøring utføres 10.000 ganger. Det er valgt å sette fem ulike verdier av forklaringsvariabelen, slik at datasettene på hundre verdier egentlig består av fem mindre, men like store datasett med ulike forklaringsvariable, forventningsverdier og nullsannsynligheter. For hvert simulerte datasett blir det registrert estimatverdier av de fem parametrene med de ulike modellene. For å finne et estimat for forventningen til hver regresjonsparameter, brukes gjennomsnittet av hvert enkelt parameterestimat i de 10.000 simuleringene. For hver simulering registreres også de kvadrerte feilleddene for hvert parameterestimat. Gjennomsnittet av disse etter 10.000 simuleringer utgjør MSE-verdiene til hvert enkelt parameterestimat, som danner grunnlaget til regresjonsanalysen som skal utføres videre i oppgaven.

Den samme prosedyren blir utført med zanb-genererte variabler.

Funksjonen lagrer til slutt en matrise med gjennomsnittet av estimatverdiene og tilhørende MSE-verdier for alle regresjonsparametrene, ved bruk av ZINB og ZANB på datasett som er zinb- og zanb-fordelte. I tillegg lagrer den en rekke andre opplysninger for hver parameterkombinasjon:
- Den eksakte parameterkombinasjonen som er brukt.
- Verdiene til forklaringsvariabelen.
- Forventningsverdien i de ulike prosessene.
- Totalantall nullobservasjoner i NB-del.
- Antall simuleringer av zinb- og zanb-datasett, n.
- Andel konvergerte tilfeller av n simulerte ved bruk av enten ZINB eller ZANB, på enten zinb- eller zanb-datasett.
- De n datasettene som skal tilpasses.

Konstantleddet og regresjonsparameteren i de binomiske prosessene påvirker to ulike egenskaper i et zinb- og zanb-datasett. Resultater fra den samme parameterkombinasjonen på datasett fra de to forskjellige fordelingene kan altså ikke sammenlignes.

For å kunne sammenligne ZINB og ZANB som modeller, er det behov for resultater fra funk-

sjonen ved ulike kombinasjoner av eksakte parameterverdier. Valg av disse verdiene er begrunnet i kapittel 5.1.1. Totalt ble det oppnådd resultater fra 3752 parameterkombinasjoner. Fordelingen av antall kombinasjoner med de ulike verdiene av dispersjonsparameteren ble som følger.

	$\alpha = 0.1$	$\alpha = 0.35$	$\alpha = 0.6$	$\alpha = 0.85$	$\alpha = 1.1$
Ant. parameterkombinasjoner	1798	1122	548	198	86

Tabell 5.1: Antall parameterkombinasjoner for hver verdi av α

5.1.2 Restriksjoner og valg av verdier

Hvert simulerte datasett i funksjonen *fun* består av 100 observasjoner. Dette antallet er satt fordi datasett brukt i praksis også kan bestå av dette antallet. Således vil en konklusjon kunne dekke slike tilfeller.

Verdiene til forklaringsvariabelen er satt til $(-2, -1, 0, 1, 2)$. I stedet for å bruke sentrerte verdier av forklaringsvariablene, som i kapittel 3.3, er disse nå satt symmetriske rundt null. Med de fem ulike verdiene oppnås like mange forventningsverdier i både binomisk- og NB-prosess, i tillegg til verdier for total sannsynlighet for nullobservasjoner. Disse verdiene varierer i ulik grad, avhengig av den eksakte parameterkombinasjonen som er bestemt. Jo lavere absoluttverdier av forklaringsvariabler som velges, desto mindre variasjon forekommer i de fem verdiene for forventning og sannsynlighet for nullobservasjoner. Eksempelvis så ville sammenligningsgrunnlaget videre blitt mer forutsigbart med verdier på forklaringsvariablene lik $(-0.2, -0.1, 0, 0.1, 0.2)$.

En stigning i den eksakte verdien av dispersjonsparameteren fører til at variasjonen i et datasett, i tillegg til den totale sannsynligheten for nullobservasjoner i et zinb-datasett blir større. I oppgaven er det valgt å kjøre funksjonen *fun* med kombinasjoner av eksakte verdier der samtlige kan variere innenfor det samme intervallet, symmetrisk om null. Dette er valgt fordi vi i kapittel 4.2.3 så at endringer i koeffisientverdier ga ulike tendenser for negative og positive verdier ved bruk på enten negative eller positive forklaringsvariabler. Ved økning av verdien på dispersjonsparameteren vil flere av kombinasjonene av de eksakte parameterverdiene gi veldig lave eller høye forventningsverdier i NB-del for både zinb og zanb, eller svært høy total nullsannsynlighet i zinb-datasett. Med tanke på at vi kun har 100 observasjoner i hvert datasett, er den største verdien av dispersjonsparameteren satt lik 1.1. For hver eksakte verdi av dispersjonsparameteren som er brukt, er det også satt en grense for hvor lav forventningsverdien i NB-prosessen, og for hvor høy den totale nullsannsynligheten i datasettene kan være. Disse er oppgitt i tabell 5.2. Restriksjonene er satt av hensyn til hva som er realistisk i et datasett med hundre observasjoner som er antatt negativ binomisk fordelte fra det virkelige liv. Den øvre grensen for den totale nullsannsynligheten er satt synkende for økende verdier av dispersjonsparameteren. Den nedre grensen for forventningsverdien i NB-delen er satt gradvis økende ved økende verdier av dispersjonsparameteren. Sistnevnte grenseverdier er satt etter testkjøringer med funkjsonen *fun* med bruk av EM-algoritmen ved tilpassing med ZINB, der lavere verdier førte til at konvergens gikk mye tregere for modellen. For den høyeste brukte

verdien av dispersjonsparameteren, $\alpha = 1.1$, er det valgt å være litt mer tålmodig; grenseverdien er satt ned noen desimaler i forhold til hva tidsforbruk til konvergens tilsa. Dette er gjort for å øke antallet av parameterkombinasjoner i hovedresultatene der $\alpha = 1.1$, fra 54 til 98 kombinasjoner.

Å kjøre funksjonen med alle kombinasjoner av regresjonsparametre for hver verdi av dispersjonsparameter er tidkrevene, spesielt ved bruk av EM-algoritmen. Det er derfor valgt fem ulike verdier av dispersjonsparameteren, og disse altså lave nok til å gi en del resultater med restriksjonene som er satt.

α	**Total nullsannsynlighet**	**Forventning i NB-del**
0.10	$<$0.71	$>$2.2
0.35	$<$0.68	$>$3.0
0.60	$<$0.65	$>$4.0
0.85	$<$0.62	$>$5.8
1.10	$<$0.59	$>$7.7

Tabell 5.2: Restriksjoner ved generering av datasett for ulike verdier av dispersjonsparameteren

Eksakte verdier av regresjonsparametrene er begrenset til verdiene $(-1.2, 1.2)$, med et intervall på 0.3. Vi får altså totalt 9 ulike verdier for hver av de fire regresjonskoeffisientene.

5.2 Bruk av programvare

For å generere zinb- og zanb-fordelte variable, og deretter utføre regresjon på datasett, finnes ulike programmeringsverktøy med tilhørende pakker og funksjoner. Noen av disse er R, stata, SAS og LIMDEP. I denne oppgaven er verktøyet R tatt i bruk. Til analyse og sammenligning av resultater er versjon 2.15.2 benyttet. For å få fortgang i et stort antall tidkrevende regresjonskjøringer har det også vært uvurderlig hjelp i eksterne servere med et større antall prosessorkjerner. Via egen pc har tilgangen kun vært på to kjerner, mens tilgjengelige servere tilbyr hele 32 av disse. Ved å i tillegg kjøre løkker i funksjonskoden parallelt, har det vært mulig å oppnå ønsket mengde resultater. Her har versjon 3.0.2 i R vært tilgjengelig.

Vi skal nå se på definerte pakker og funksjoner som er brukt for å oppnå resultater til videre analyse med funksjonen som er laget i oppgaven. Det vil også bli nevnt alternativer som kan brukes, begrenset til programmeringsverktøyet R. Til dette er det brukt dokumentasjon som finnes i funksjonsbeskrivelser i [2], i tillegg til en artikkel om praktisk bruk av disse [19].

5.2.1 Generering av data til zinb- og zanb-datasett

I funksjonen som er laget, genereres zinb- og zanbfordelte variabler ved hjelp av funksjonene *rzinegbin* og *rzanegbin* fra pakken **VGAM**. Førstnevnte funksjon tar følgende argumenter:
• munb = forventningen i NB-prosessen
• pobs0 = sannsynligheten for strukturelle nullobservasjoner
• size = theta, den inverse verdien av dispersjonsparameteren α.

For å generere zanb-fordelte variabler med funksjonen *rzanegbin* trengs
• munb = forventningen i den trunkerte NB-prosessen
• pstr0 = sannsynligheten for nullobservasjoner
• size = inverse av dispersjonsparameteren.

I tillegg velges ønsket mengde, n, av variabler som skal genereres i de to funksjonene.

5.2.2 Regresjon på datasett

Til å modellere datasettene er pakken **pscl** tatt i bruk for både ZINB og ZANB, med funksjonene *zeroinfl* og *hurdle*. Disse bruker mange av de samme argumentene:
• formula - spesifiserer komponentene i modellen, det vil si at det leses inn verdier fra det gjeldende datasettet med tilhørende forklaringsvariabler i både NB- og binomisk prosess. Denne er på formen $y \sim x1 + x2 \mid z1 + z2$ når NB-prosessen (trunkert i ZANB) har to forklaringsvariabler, x1 og x2, og den binomiske prosessen antas å ha to mulige forklaringsvariabler z1 og z2. Ved bruk av kun én forklaringsvariabel, som er satt lik i de to prosessene, trenger vi kun å skrive den opp én gang, $y \sim x$.
• data - tilsvarer en dataramme med tilsvarende verdier som i argumentet 'formula'. I den tillagde funksjonen i oppgaven er funksjonen *as.data.frame* benyttet for å oppnå korrekt form.
• dist - valg av hvilken fordeling som skal brukes til å finne tilpassede data til prosessen som ikke er binomisk. I denne oppgaven har vi valgt å bruke en negativ binomisk prosess, derfor kalt NB-prosess, og vi bruker derfor negbin i valget mellom denne, Poisson og geometrisk. I begge modellene blir det brukt log-link i NB-prosessen.

En av de største forskjellene med funksjonene er at vi i *hurdle* har et argument, zero.dist, der vi kan velge hvilken modell vi vil bruke til å tilpasse den binomiske prosessen i datasettet som skal testes. Denne kan settes som binomisk-, Poisson-, geometrisk-, eller negativ binomisk, og vi har valgt førstnevnte med logit-link siden denne er den mest brukte. I *zeroinfl* har vi ikke mulighet til å velge modell som brukes i den binomiske prosessen, denne er satt til å være binomisk med logit-link.

Det finnes flere valg i funksjonene som kan puttes inn som argumenter for å gjøre tilpassing og estimering enklere eller mer nøyaktige. I begge modellene blir parametrene estimert ved maksimering av likelihoodfunksjonen, med den innebygde optimeringsfunksjonen *optim*. Den numeriske estimeringen av kovariansmatrisen blir utført ved hjelp av Hessian-matrisen. I argumentet *hurdle.control*

og *zeroinfl.control* kan man blant annet velge hvilken metode som skal brukes ved maksimering av likelihoodfunksjonen i *optim*, maksimalt antall iterasjoner som skal utføres, og man kan sette startverdier for parametrene i begge prosessene og for dispersjonsparameteren. Uten gitte initialverdier vil begge modeller estimere disse ved bruk av *glm.fit*. I *zeroinfl.control* er det mulig å estimere startverdier med EM-algoritmen, i stedet for default-metoden glm.fit. Sistnevnte metode som kun kalles ' en gang for hver av regresjonsparametrene. Med EM-algoritmen kjøres iterasjoner inntil parametrene har konvergert til parameterverdiene som gir maksimert verdi av likelihoodfunksjonen. Med glm.fit brukes estimatene som oppnås etter første iterasjon med EM-algoritmen. I *hurdle.control* er det i tillegg mulig å velge separat estimering av de to prosessene i modellen, siden likelihoodfunksjonen kan deles opp etter disse, som i uttrykk (4.5).

Pakken **pscl** er den mest brukte til regresjonskjøring med ZINB og ZANB, siden den både er brukervennlig ved kjennskap til andre pakker i R, og gir ut viktig informasjon med mange av standardfunksjonene i R. Ved hjelp av funksjonen *coef* får vi hentet ut de fire regresjonsparametrene fra de to prosessene i modellene. Den inverse av dispersjonsparameteren blir estimert som en støyparameter, ved maksimering av likelihoodfunksjonen. Verdien av tilhørende parameterestimat kan hentes ut ved kall av *theta*. Av andre tilgjengelige standardfunksjoner fra pakken kan *print*, *summary*, *residuals* og *fitted* nevnes. I tillegg kan man bruke testfunksjonene *coeftest*, *waldtest* på det tilpassede objektet. Ved bruk av nøstete modeller kan man også ta i bruk *lrtest*.

5.2.3 Alternative pakker og funksjoner

I pakkene **gamlss** og **VGAM** er det også mulig å simulere datasett som er zinb-fordelte, og deretter utføre regresjon på disse. I førstnevnte pakke lar funksjonen *rZINBI* oss generere zinb-fordelte data, og ved hjelp av tilpassing med funksjonen *gamlss* kan vi hente ut estimerte verdier av regresjonsparametrene. Fordelingen ZINBI tilhører altså GAMLSS-familien, og er et av argumentene i modellen, i tillegg til datasettet med tilhørende forklaringsvariabler. I den samme pakken finner vi funksjonen *rZANBI*, der ZANBI tilhører GAMLSS-familien og tilsvarer ZANB-fordelingen. I pakken **VGAM** er funksjonene *zinegbinomial* og *zanegbinomial* laget for å tilpasse zinb- og zanb-datasett.
I begge pakkene kan derimot kun ett sett med forklaringsvariabler brukes. I tillegg er noen av funksjonene trege og kan ha store problemer med konvergens hvis ikke gode startverdier er satt, dette gjelder spesielt funksjonene fra pakken **VGAM**.

Det finnes flere mindre pakker i R med alternative versjoner av ZI- og ZA-regresjonsmodeller. Ingen av disse kan måle seg med **pscl** med tanke på brukervennlighet og mengden av standardmetoder som kan benyttes.

5.3 Valg av algoritme til å maksimere likelihoodfunksjoner

For å kunne sammenligne ZINB og ZANB som regresjonsmodeller, trenger vi et så korrekt og stabilt sammenligningsgrunnlag som mulig. Valg av algoritme til å produsere parameterestimater,

som videre brukes til å beregne tilhørende MSE-verdier, kan være avgjørende for videre resultater og konklusjoner.

Regresjonsmodellen ZANB har en todelt likelihoodfunksjon som gjør maksimering enkelt å fullføre numerisk. Bruk av BFGS-metoden fører til tilnærmet fullstendig konvergens av de 10.000 simuleringene som blir utført for alle parameterkombinasjoner, på både zinb- og zanb-datasett.

Likelihoodfunksjonen til ZINB er derimot ikke like enkel å optimere, siden regresjonsparametrene i de to ulike prosessene ikke kan estimeres separat. I oppgaven er det utført regresjonskjøring med bruk av både BFGS-algoritmen og EM-algoritmen i ZINB-modellen for alle parameterkombinasjonene, for å se om disse gir store forskjeller i MSE-verdiene til parameterestimatene. Ved bruk av BFGS fikk vi ingen parameterkombinasjoner der mer enn én av 10.000 simuleringer ikke hadde konvergert med bruk av begge modeller på begge typer datasett. Ved bruk av EM konvergerte minst 99.9,% av alle tilfellene ved bruk av ZANB. Ved bruk av ZINB derimot, hadde vi 32 parameterkombinasjoner der under 99% av simuleringene konvergerte på zinb-datasett (minste antall var 9668), og 155 av kombinasjonene ga under 99% konvergerte på zanb-datasett (minste antall var 8158).

Tabell 5.3 viser andelen av MSE-verdier som er lavere ved bruk av BFGS-algoritmen. Vi ser at andelen for verdiene i NB-del, på både zinb- og zanb-datasett er mindre enn 0.5, og kun 0.388 for konstantleddet der zinb-datasett er brukt. Til gjengjeld fikk vi lavere MSE-verdier av alle parameterestimater i binomisk del i over 80% av alle tilfellene ved bruk av BFGS. For MSE-verdiene til estimatene for dispersjonsparameteren ved bruk av zinb, er andelen 0.285, altså gir EM-algoritmen betydeligere lavere verdier.

NB-koef (zinb)	**NB-forkl (zinb)**	**BIN-koef (zinb)**	**BIN-forkl (zinb)**	**Disp (zinb)**
0.388	0.471	0.837	0.847	0.285
NB-koef (zanb)	**NB-forkl (zanb)**	**BIN-koef (zanb)**	**BIN-forkl (zanb)**	**Disp (zanb)**
0.491	0.495	0.841	0.855	0.468

Tabell 5.3: Andel parameterkombinasjoner med lavere MSE-verdier ved bruk av BFGS- i stedet for EM-algorimen i ZINB.

Det ble også sett på hvor stor andel av MSE-verdiene som lå under gitte grenser, $< 10, < 1, < 0.25, < 0.1, < 0.05$. Her var det ikke særlige forskjeller å detektere ved bruk av de to metodene når vi så på verdiene fra NB-delen eller fra dispersjonsparameteren. Andelen av verdiene av MSE i binomisk del derimot, viste store forskjeller. Tendensen var da den samme for MSE-verdier av estimat av både forklaringsparametrene og konstantleddene ved regresjonskjøring på både zinb- og zanb-datasett. Tabell 5.4 viser at EM-algoritmen gir mye større andel av MSE-verdier for estimatene av konstantleddet i binomisk del, for alle kriteriene som inneholder andeler.
For forklaringskoeffisienten gir EM størst andel ved kriteriene < 10, < 1.0 og < 0.1. For MSE-

verdiene under 0.05 er det omvendt, men andelene av verdier under kriteriet er riktignok relativt lave da.

		< 10	< 1.0	< 0.25	< 0.1	< 0.05
EM	BIN-konst (zinb)	0.937	0.512	0.183	0.026	0
	BIN-forkl (zinb)	0.980	0.706	0.313	0.100	0.014
BFGS	BIN-konst (zinb)	0.800	0.384	0.164	0.035	0
	BIN-forkl (zinb)	0.898	0.513	0.247	0.102	0.022
EM	BIN-konst (zanb)	0.839	0.441	0.082	0.003	0
	BIN-forkl (zanb)	0.963	0.6956	0.185	0.038	0.005
BFGS	BIN-konst (zanb)	0.733	0.365	0.078	0.008	0
	BIN-forkl (zanb)	0.855	0.553	0.149	0.038	0.009

Tabell 5.4: Andel MSE-verdier lavere enn gitte verdier i binomisk prosess

Algoritmen BFGS ga altså flere vesentlige lave MSE-verdier for parameterestimatene i den binomiske delen. Ut i fra disse tabellene er det ingen av algoritmene som er åpenbart best å bruke. Med funksjonene som er brukt for tilpasning, estimeres dispersjonsparameteren som en støyparameter. Forskjeller i tilhørende MSE-verdier ved bruk av metodene tillegges derfor ikke vekt når valg av algoritme skal foretas. Grunnet tidligere nevnte forskjeller i binomisk del, er det hovedsaklig verdier oppnådd med BFGS-algoritmen som er valgt å brukes videre i oppgaven. Valget av metode er likevel usikkert siden MSE-verdiene i NB-del er lavere ved bruk av EM-algorimen. Resultater videre i oppgaven som avviker ved bruk av BFGS- og EM-algoritmen vil derfor kommenteres.

Kapittel 6

Resultater fra sammenligning av ZINB og ZANB i praksis

I dette kapitlet vil ulike sammenligninger av resultater fra regresjonskjøring med ZINB og ZANB ved bruk av funksjonen *fun* bli presentert.

Til å starte med, vil det være naturlig å undersøke om valg av modell ved regresjonskjøring på de aktuelle parameterkombinasjonene faktisk spiller noen rolle. I så tilfelle, ønsker vi videre å finne ut om resultatene viser til at den korrekte modellen alltid bør brukes, altså om modellen ZANB er den beste å bruke på zanb-fordelte datasett, og tilsvarende om ZINB bør brukes på zinb-fordelte datasett. Deretter vil vi se hvor mye bedre den beste modellen er for de ulike parameterkombinasjonene, og således hvor store konsekvensene ved bruk av gal modell er. Vi vil også forsøke å se om valg av modell kan foretas ved å se på egenskapene i datasettet. slutt skal vi se om verdiene av egenskapene i datasettene kan være avgjørende for modellvalg.

Et datasett som er ZANB-fordelt kan inneholde færre nullobservasjoner enn en vanlig NB-fordeling ville ha gitt. Vi skal derfor også se hvilke konsekvenser bruk av ZINB gir i slike tilfeller, siden ZINB-fordelte datasett kun kan inneholde tilfeller med like mange, eller flere nullobservasjoner for enhver kombinasjon av parameterverdier.

Til slutt vil vi se på resultater fra regresjonskjøring med funksjonen *fun* ved en mye større verdi av dispersjonsparameteren.

Videre i oppgaven vil vi ofte referere til MSE-verdier til parameterestimatene for de to regresjonskoeffisientene som MSE-verdi for konstantleddet og forklaringsleddet.

6.1 Behov for modellvalg

Mange parameterkombinasjoner gir tilnærmet like resultater ved bruk av ZINB og ZANB, enten brukt på én av zinb- og zanb-datasettene, eller begge. I dette delkapitlet skal vi undersøke om bruk av de to modellene på zinb-og zanb-datasett faktisk gir forskjellige resultater. Her vil MSE-verdier fra de to modellene bli plottet mot hverandre, for å gi en indikasjon på eventuelle avvik ved modellene.

Ved å i tillegg sammenligne marginalfordelingene til MSE-verdiene ved bruk av de to modellene, vil vi også kunne se om ulike eksakte verdier av dispersjonsparameteren gir utslag i behovet for å velge én av modellene fremfor den andre. For å få et visuelt inntrykk, vil sammenligningene presenteres i plott, for deretter å bli kommentert. Vi vil se på resultatene i NB-prosess, binomisk prosses, og for dispersjonsparameteren, hver for seg. Til slutt vil resultatene samlet oppsummeres.

6.1.1 Resultat for NB-prosess

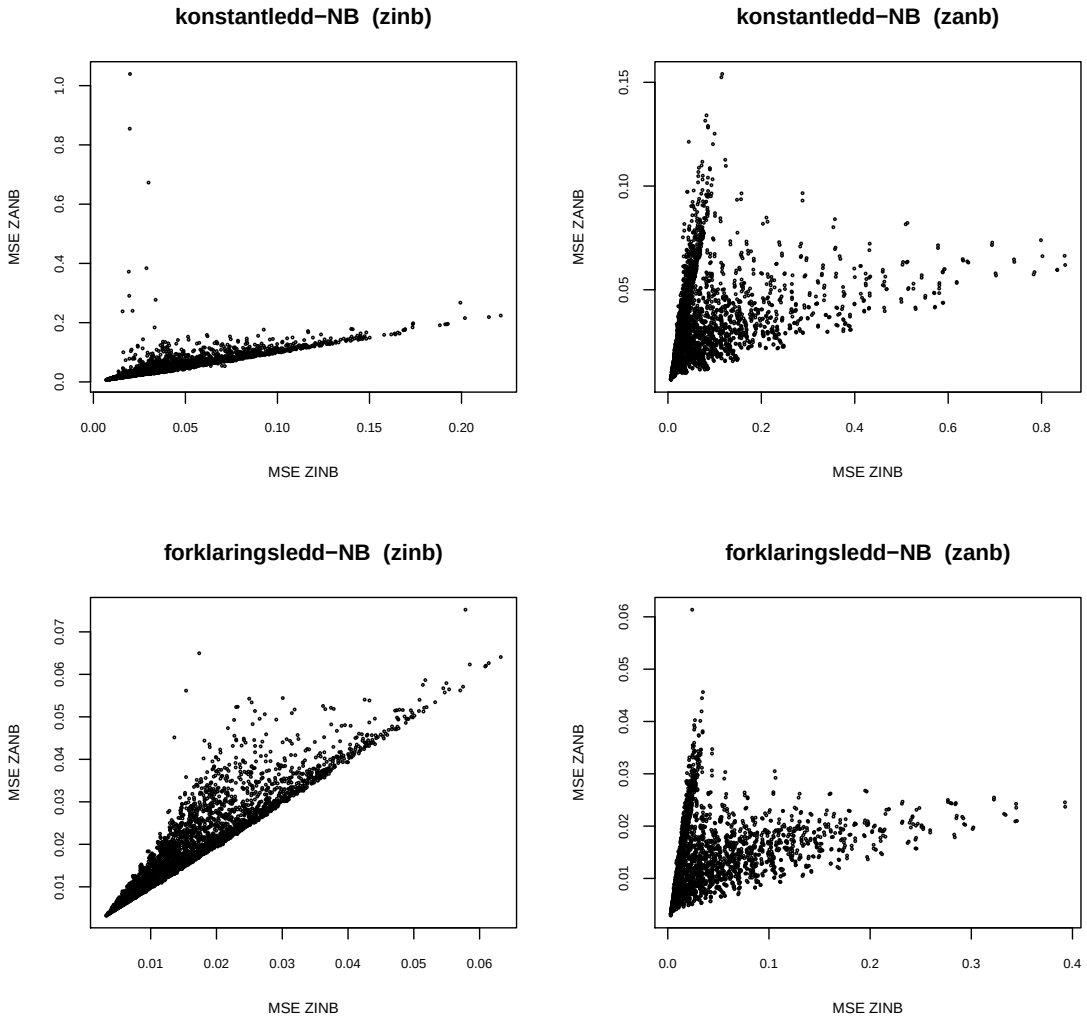
Vi starter med å se om modellene gir forskjellige resultat for konstantleddet og forklaringskoeffisienten i NB-prosessen til de genererte datasettene. I figur 6.1 er MSE-verdiene ved bruk av ZINB og ZANB plottet mot hverandre, med regresjon på zinb-datasett til venstre, og zanb-datasett til høyre. Hvis modellene hadde gitt identiske resultater for hver av parameterkombinasjonene, ville plottene ha vist seg som punkter i en rett linje, fra koordinatene ved minste MSE-verdi til koordinatene ved høyeste verdi. Vi ser fra samtige av plottene at mange av punktene ligger langs denne linjen. Men vi ser også klart at bruk av feil modell i forhold til fordelingen til datasettene det er utført regresjon på, gir en del punkter i plottene som ligger ligger nærmere aksen til MSE-verdiene av den feile modellen.

Figur 6.2 viser de marginale fordelingene til MSE-verdiene ved tilpassing av zinb-datasett med bruk av ZINB og ZANB i samme plot, for henholdsvis konstantleddet og forklaringskoeffisienten Her ser vi at de marginale fordelingene følger samme trend. Parameterkombinasjonene langs x-aksen er blant annet rangert etter stigende eksakte verdier av dispersjonsparameteren. Skillet på sistnevnte verdier kan sees fra plottet, men for de største verdiene kan tabell 5.1 være til hjelp.

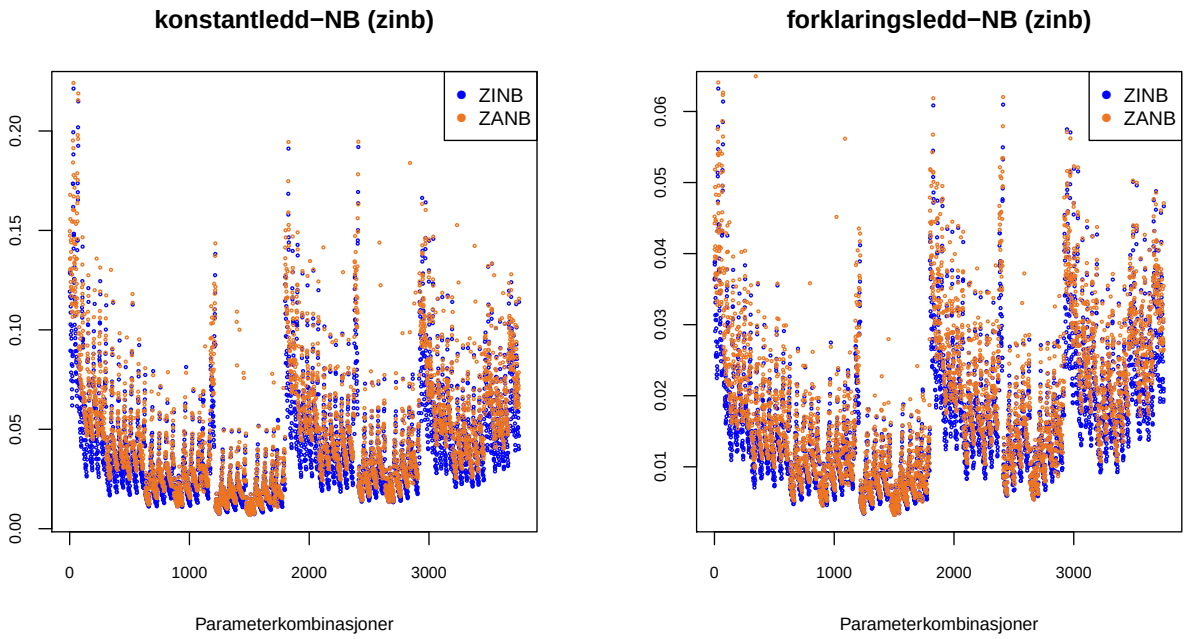
Vi ser at de to modellene gir ulike MSE-verdier for både konstantleddet og forklaringsleddet i mange av parameterkombinasjonene, for samtlige eksakte verdier av dispersjonsparameteren. De ulike MSE-verdiene som er oppnådd i NB-prosessen er relativt lave i alle tilfellene. For konstantleddet ligger de fleste MSE-verdiene under 0.20, og for forklaringsleddet er høyeste verdi lavere enn 0.08.

Ved regresjonskjøring på zanb-datasett ser vi større forskjeller i den marginale fordelingen av verdiene, slik figur 6.3 viser. Bruk av ZINB gir da en del tilfeller av høye MSE-verdier i forhold til ved bruk av ZANB, for begge regresjonsparametrene. Modellen ZINB gir MSE-verdier opp mot

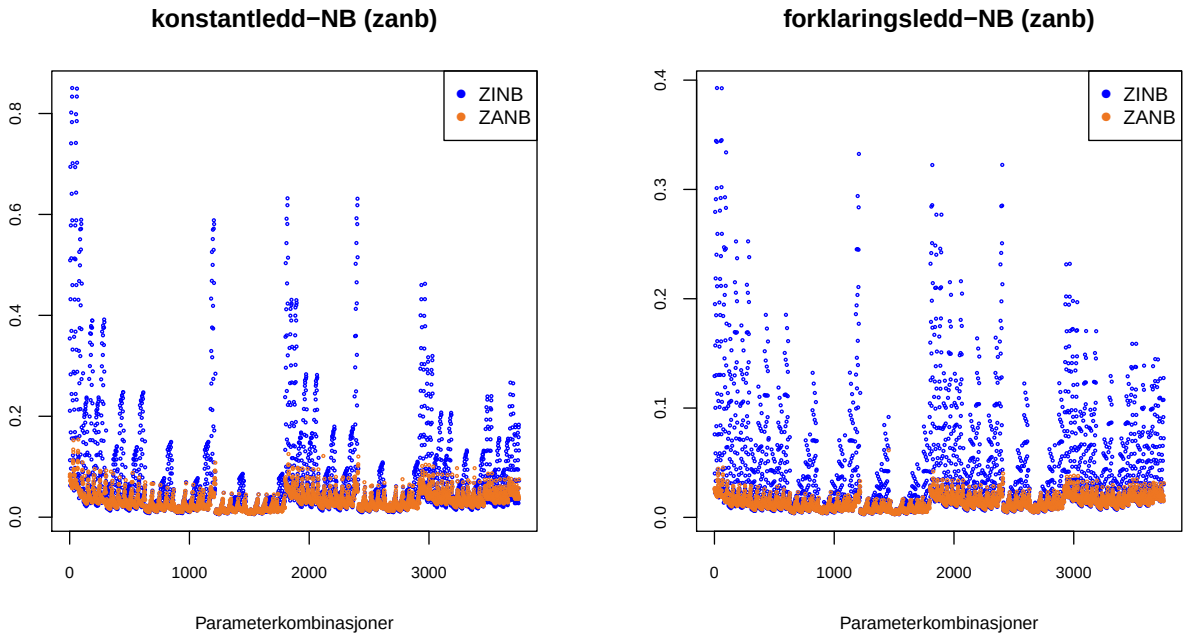
0.85 og 0.39 for henholdsvis konstantleddet og forklaringsleddet. De tilsvarende høyeste verdiene ved bruk av ZANB er 0.15 og 0.06.



Figur 6.1: MSE-verdier ved bruk av ZINB og ZANB i den negativ binomiske prosessen av modellene, plottet mot hverandre



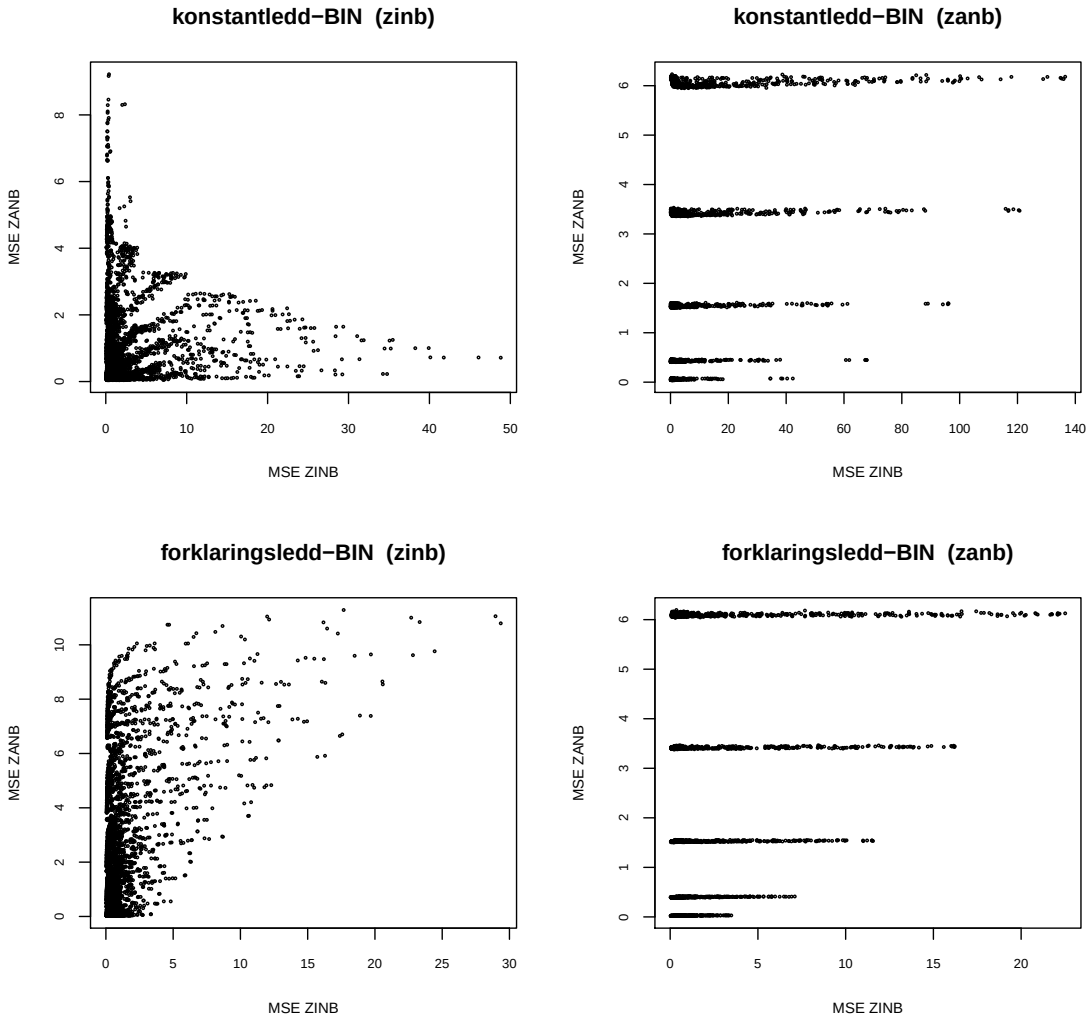
Figur 6.2: Marginal fordeling av MSE-verdier for NB-del ved zinb-fordelte datasett.



Figur 6.3: Marginal fordeling av MSE-verdier for NB-del ved zanb-fordelte datasett.

6.1.2 Resultat for binomisk prosess

I den binomiske prosessen viser det seg også at bruk av ZINB og ZANB gir ulike MSE-verdier for begge regresjonsparametrene ved tilpassing på både zinb- og zanb-datasett. Figur 6.4 viser at MSE-verdiene fra de to modellene plottet mot hverandre, tar ulike former, der samtlige er langt unna å gi en rett linje mellom den laveste og høyest MSE-verdien. Regresjonskjøring på zanb-datasett med modellen ZANB gir MSE-verdier som er sentrert rundt fem forskjellige verdier, mens bruk av ZINB gir verdier som varierer i en større skala.

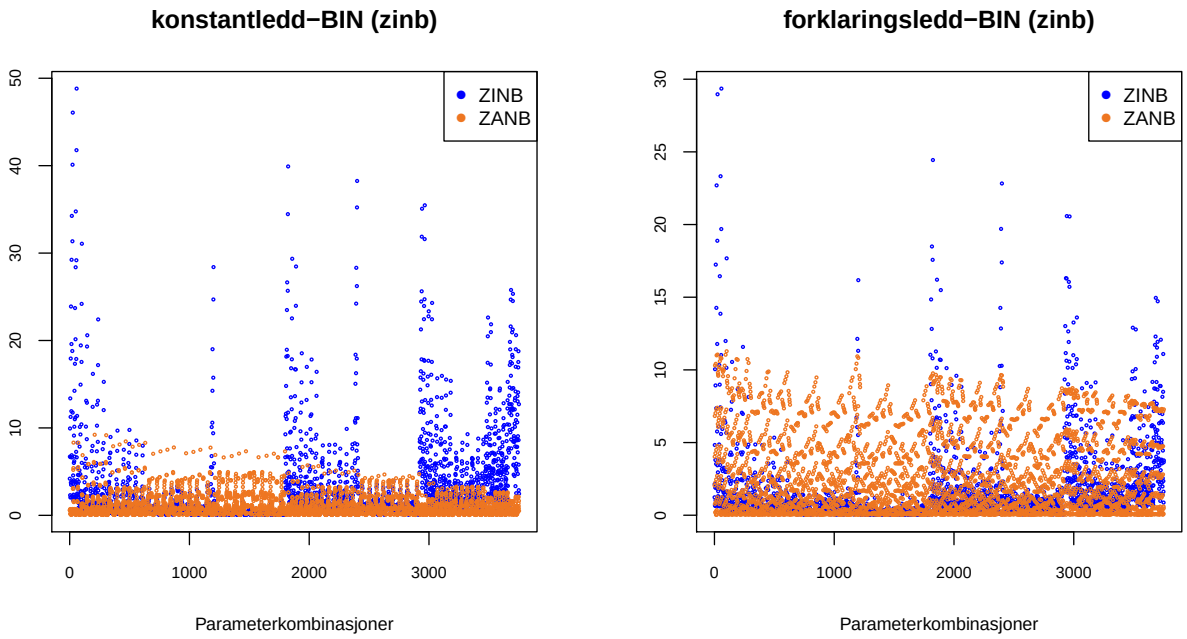


Figur 6.4: MSE-verdier ved bruk av ZINB og ZANB i den binomiske prosessen av modellene, plottet mot hverandre

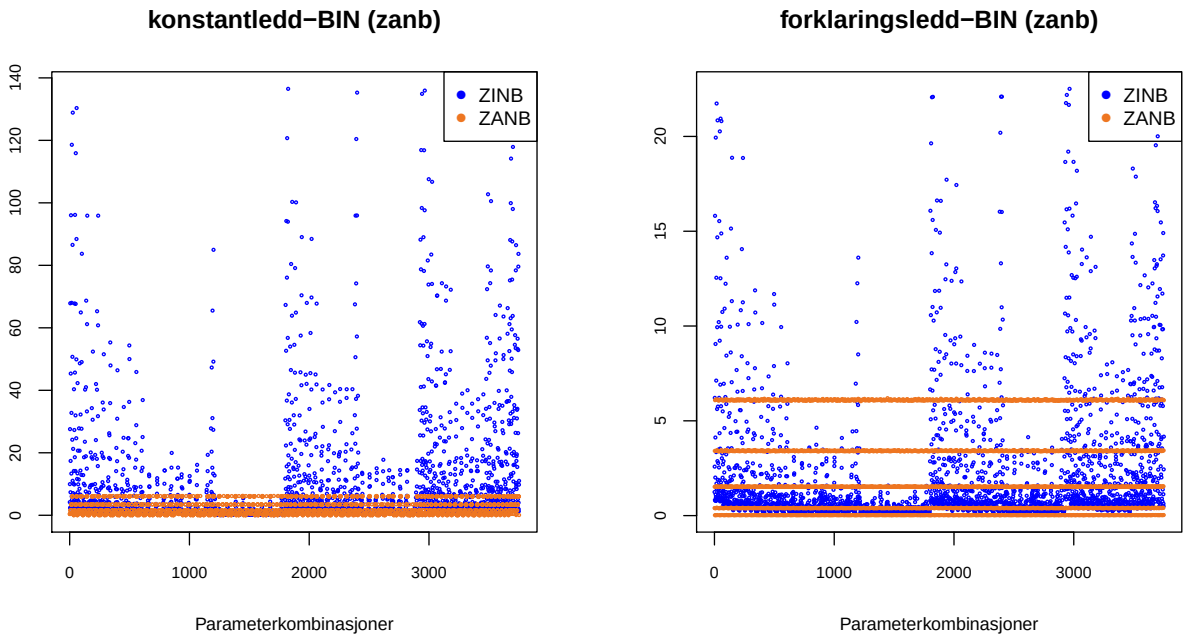
Figur 6.5 viser at modellene brukt på zinb-datasett som oftest gir lave MSE-verdier. Bruk av ZINB skiller seg mest ut fra plottene; vi ser at en del parameterkombinasjoner gir mye større verdier enn ved bruk av ZANB. For konstantleddet er høyeste MSE-verdi ved bruk av ZINB og ZANB henholdsvis 48.8 og 9.2. For forklaringsleddet er tilsvarende høyeste verdier 29.4 og 11.3. Vi ser

derimot også klart at noen kombinasjoner gir lavere verdier ved bruk av ZINB.

Resultatene etter bruk av modellene på zanb-datasett gir MSE-verdier som avviker i enda større grad. I figur 6.6 ser vi det samme mønsteret ved bruk av ZANB som vi så i figur 6.4, nemlig at modellen gir MSE-verdier som er sentrert rundt fem forskjellige verdier. Dette er ikke like synlig i plottet av marginalfordelingene til verdiene for konstantleddet. Grunnen til forskjeller i skala på de to plottene i 6.6 er at ZINB, også ved bruk på zanb-datasett, gir høye verdier for en del av parameterkombinasjonene. For konstantleddet og forklaringsleddet er høyeste verdi oppnådd med ZINB henholdsvis 136.5 og 22.5. Ved bruk av ZANB er tilsvarende høyeste verdier 6.23 og 6.19.



Figur 6.5: Marginal fordeling av MSE-verdier for NB-del ved zinb-fordelte datasett.



Figur 6.6: Marginal fordeling av MSE-verdier for NB-del ved zinb-fordelte datasett.

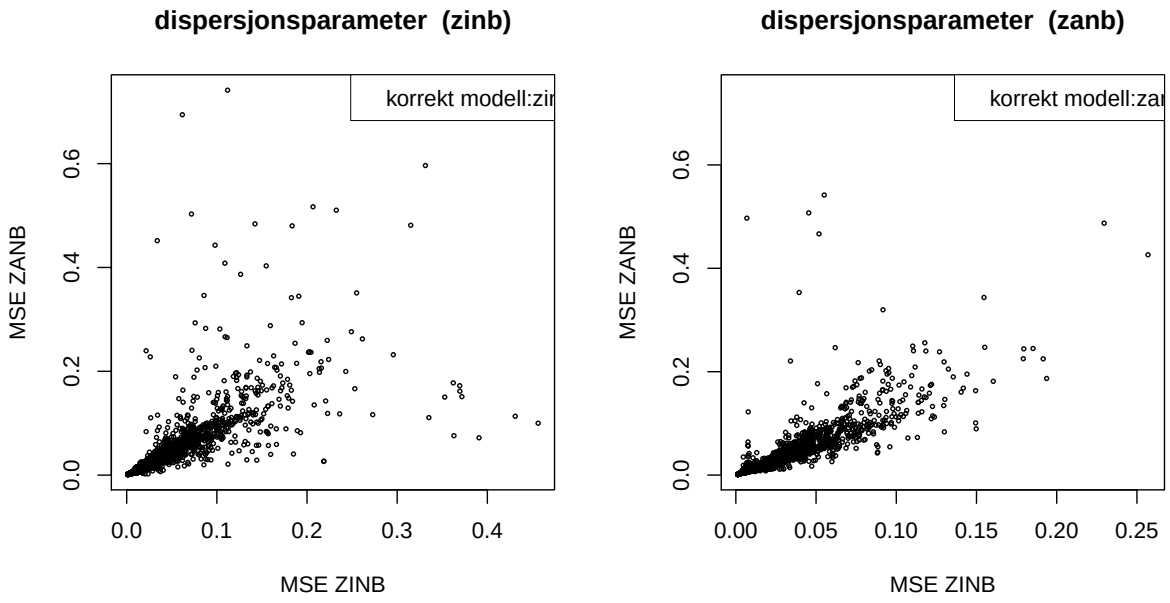
6.1.3 Resultat for dispersjonsparameter

Av de 3752 parameterkombinasjonene som brukes i regresjonsanalysen i oppgaven, fører noen til veldig høye MSE-verdier for dispersjonsparameteren ved bruk av ZANB-modellen. Dette er grunnet noen få tilfeller av de 10.000 simulerte datasettene, der modellen ikke har klart å tilpasse dataene godt nok. Disse gir videre utslag på gjennomsnittsverdien som brukes videre, siden de resterende verdiene av MSE ligger under 1. Ved å se på de aktuelle simulerte datasettene som fører til at disse høye estimat- og MSE-verdiene forekommer, ser vi noen klare tendenser. Datasettene består da ofte av en stor andel nullobservasjoner i forhold til sannsynligheten for totalantallet beregnet for parameterkombinasjonen. I tillegg består datasettet ofte enten av få verdier, eksempelvis kun av $0, 1$ og 2, eller av svært mange verdier med kun ett tilfelle av hver i en lang hale. I funksjonene som brukes for å tilpasse datasett i oppgaven, blir den inverse av dispersjonsparameteren, $\theta = \frac{1}{\alpha}$, estimert som en støyparameter. Det antas da at θ ($\log \theta$) ved sentralgrenseteoremet er asymtotisk normalfordelt. Ekstremverdier i datasettet ser ut til å gjøre estimering av dispersjonsparameteren svært ustabil med ZANB-modellen.

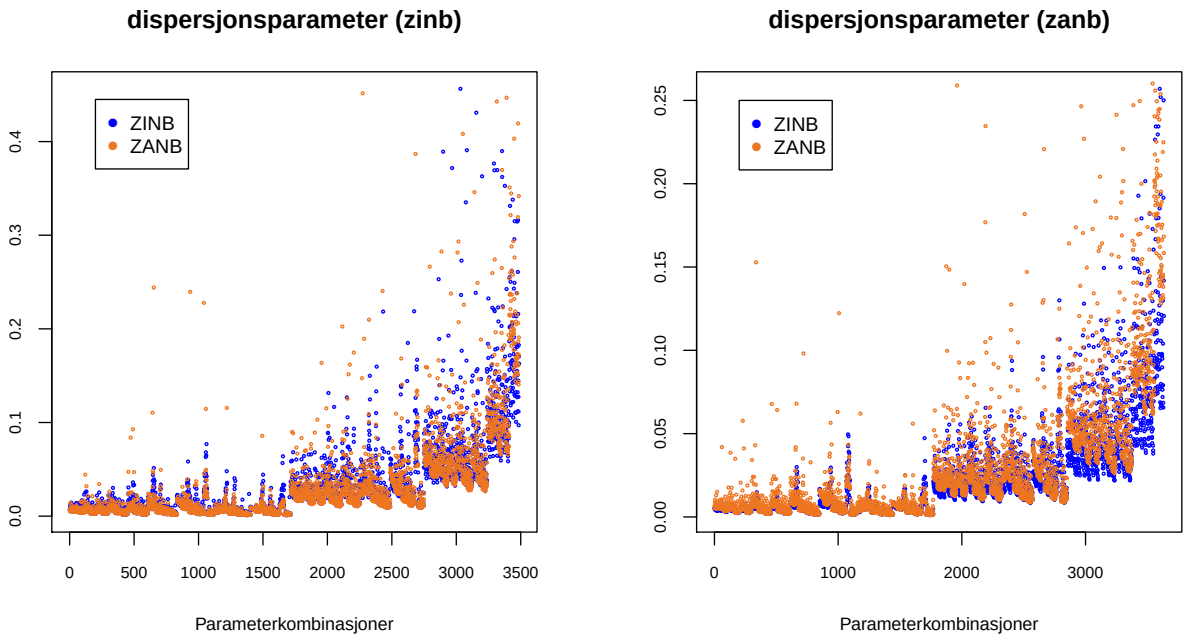
Parameterkombinasjonene som fører til tilfeller av veldig høye estimatverdier av dispersjonsparameteren ved bruk av enten ZINB eller ZANB, har like fortegn på de gitte eksakte verdiene av regresjonskoeffisienten som påvirker forklaringsvariabelen i de to prosessene. I kapittel 4.2.3 så vi at i slike tilfeller kan vi ha zinb-fordelte datasett der både total nullsannsynlighet og forventningsverdi i NB-delen er høye.

I Figur 6.7 er MSE-verdiene for dispersjonsparameteren ved bruk av de to modellene plottet mot hverandre. Parameterkombinasjoner som ga MSE-verdi > 1 er ikke tatt med her, for å enklere få et visuelt inntrykk av forholdet mellom verdiene. Majoriteten av kombinasjonene med MSE-verdi > 1 hadde enda større verdier, der de høyeste var $8.0 \cdot 10^{12}$ ved tilpassing av zinb-datasett og $9.0 \cdot 10^{8}$ ved tilpassing av zanb-datasett. Fra plottet til venstre er 265 punkter utelatt ved tilpassing av zinb-datasett. I disse kombinasjonene førte tilpassing av datasettene til få, men minst èn av 10.000 estimerte parameterverdier som var mye større enn de resterende og som dermed førte til ekstremt høye MSE-verdier. Av tilsvarende grunn førte bruk av ZANB på zanb-datasett til utelatelse av 119 punkter i plottet til høyre i figur 6.7. Vi ser at bruk av både ZINB og ZANB på begge typer datasett gir lave MSE-verdier for de fleste parameterkombinasjonene, og at mange av disse har verdier som ikke ligger langt fra en linje fra origo til et punkt med like verdier i koordinatene. Men vi ser også mange punkter som ligger utenfor en slik linje, og disse er spredt mot begge aksene. I tillegg kommer punktene som ikke er tatt med i plottene, der bruk av ZANB har gitt klart større verdier enn ZINB i alle tilfellene. Figur 6.8 viser de marginale fordelingene av MSE-verdier for de samme parameterkombinasjonene som i figur 6.7, også her for å lettere få et visuelt inntrykk av resultatene ved bruk av de to modellene. Vi har altså 3487 verdier fra bruk av ZINB og ZANB på zinb-datasett, i plottet til venstre, og 3633 verdier ved bruk av modellene på zanb-datasett, til høyre. Plottene viser relativt like mønster for verdiene ved bruk av ZINB og ZANB. En økning i gitte verdier av dispersjonsparameteren ser ut til å gi større MSE-verdier ved tilpassing med begge modellene på begge typer datasett. Ulikheter i verdiene modellene gir er også mer synlige

for større gitte verdier av parameteren. Bruk av ZINB på zinb-datasett gir MSE-verdier < 0.311 og på zanb-datasett gir modellen MSE-verdier < 0.318, for samtlige av parameterkombinasjonene som er tatt bort i plottene.



Figur 6.7: MSE-verdier for dispersjonsparameteren ved bruk av ZINB og ZANB, plottet mot hverandre



Figur 6.8: Marginal fordeling av MSE-verdier for NB-del ved zinb-fordelte datasett.

6.1.4 Oppsummering

Vi har nå sett at ZINB og ZANB gir ulike MSE-verdier i mange tilfeller for alle regresjonsparametrene og for dispersjonsparameteren. Ulikhetene var mest synlig når vi så på parametrene i NB-delen til zanb-datasett, og for parametrene i binomisk del generelt. I tillegg ga noen av parameterkombinasjonene svært høye MSE-verdier for dispersjonsparameteren. I kombinasjonene der MSE-verdiene til de ulike parametrene var lave ved bruk av begge modellene, var det ikke like lett å se fra figurene om verdiene som ble oppnådd med ZINB og ZANB var ulike nok til å utgjøre en særlig forskjell i praksis. Dette vil vi se nærmere på neste delkapittel, siden resultatene hittil viser at valg av modell er avgjørende for hvor god tilpassing av datasettene vi vil oppnå.

Ved bruk av EM-algorimen fikk vi tilsvarende mønster i tilsvarende plotter som vi har sett i figurene hittil. Den største synlige forskjellen var at enda større MSE-verdier ble oppnådd ved tilpassing med ZINB, for alle parametrene i den binomiske delen.

6.2 Valg av ZINB og ZANB med hensyn på fordeling i datasettet

Vi vil nå undersøke om valg av ZINB og ZANB generelt bør tas med hensyn på hvilken fordeling datasettet tilhører. Fra alle figurene som ble presenterte i kapittel 6.1, ser vi at majoriteten av kombinasjonene gir lave MSE-verdier for alle regresjonsparametrene i tillegg til dispersjonsparameteren, både ved bruk av ZINB og ZANB for tilpassing av zinb- og zanb-datasett. Det er vanskelig å se hvilken av modellene som gir de laveste verdiene i disse tilfellene.

Ved å se på differansene i MSE-verdier bruk av modellene gir for hver parameterkombinasjon, kan vi se hvilken modell som gir de laveste, og hvor store forskjellene er. I mange tilfeller vil modellene gi verdier som er svært nære, og som således gir tilnærmet like gode resultater. Resultater blir derfor vist med ulike kriterier for at en MSE-verdi viser til god tilpassing for de ulike parametrene. Modellene vil ofte bli referert til som den beste i forhold til lavest MSE-verdi til regresjonsparameteren som blir studert.

Vi vil også i dette delkapitlet se på resultatene i NB- og binomisk prosess hver for seg, i tillegg til MSE-verdiene for dispersjonsparameteren. Figurene fra forrige delkapittel vil også bli brukt videre i analysen.

6.2.1 Resultat for NB-prosess

For å finne ut om den korrekte modellen er den beste å bruke på datasettene for parametrene i NB-prosessen, starter vi med å se på resultatene etter tilpassing av zinb- og zanb-datasett hver for seg.

Vi ser først på MSE-verdiene som er oppnådd ved tilpassing av zinb-datasett. I figur 6.2 ser

vi at det finnes kombinasjoner der MSE-verdier ved bruk av ZINB er lavere enn ved bruk av ZANB, både for konstantleddet og forklaringskoeffisienten. Dette ser ut til å gjelde for alle de eksakte verdiene av dispersjonsparameteren. Det er en klar tendens til større MSE-verdier ved bruk av ZANB, mens det ikke er like lett å lese fra plottene om bruk av ZINB gir lavere MSE-verdier for alle parameterkombinasjonene, eller hvor stor variasjon i verdiene modellene gir i hvert enkelt tilfelle. Fra plottene ser vi at de fleste MSE-verdiene ligger under 0.1 for konstantleddet, og under 0.03 for forklaringsleddet. Begge modellene gir riktignok en del MSE-verdier som er en del større enn disse; opp mot 0.25 for konstantleddet og 0.07 for forklaringsleddet.

For å finne ut hvilken modell som gir den beste, eller generelt god tilpasning av datasettene, skal vi se på differansen i MSE-verdier som oppnås med modellene for de ulike parametrene. Kriteriet for at en MSE-verdi viser til god tilpassing er uvisst, men vi vil vise hvordan andelen av parameterkombinasjoner der modellene er best, eller like gode, varierer med ulike kriterier. Tabell 6.1 viser andelen av parameterkombinasjoner der de ulike modellene ved bruk på zinb-datasett er best, eller like gode i følge gitte kriterier. Dette er gjort ved å se på MSE-verdiene som oppnås for konstantleddet og forklaringskoeffisienten hver for seg. Ved det strengeste kriteriet, $= 0$, ser vi på andel kombinasjoner der hver av modellene generelt gir den laveste MSE-verdien. Videre vil eksempelvis kriteriet < 0.0001 vise til like god tilpassing ved bruk av modellene hvis absoluttverdien til MSE-verdiene er mindre enn 0.0001.

Vi ser at ZINB gir lavest MSE-verdi for konstantleddet i 75.2% av parameterkombinasjonene ved det strengeste kriteriet, og at ZANB således gir lavest verdi i de resterende kombinasjonene. Ved å øke intervallet for MSE-verdier som viser til god tilpassing ser vi hvordan disse andelene minker, og at vi får en økende andel av parameterkombinasjonene der modellene er like gode. Ved kriteriet < 0.005 ser vi at det er svært få kombinasjoner der ZANB er den beste modellen for konstantleddet sin del. Andelen der ZINB er den beste for det samme kriteriet er derimot 33.9%. For forklaringskoeffisienten gir ZINB lavere MSE-verdi i hele 80.6% av parameterkombinasjonene ved det strengeste kriteriet. Ved kriteriene < 0.001 og < 0.005 ser vi at bortimot ingen av parameterkombinasjonene gir best tilpassing ved bruk av ZANB. Her er tilsvarende andel der ZINB er den beste modellen henholdsvis på 49.0% og 19.9%.

På tilsvarende måte viser tabell 6.2 andelen av parameterkombinasjonene der de to modellene brukt på zanb-datasett gir lavest MSE-verdier for parametrene i NB-prosessen. For konstantleddet gir ZANB lavest verdi i 52.7% av tilfellene og for forklaringsleddet er tilsvarende andel 57.4%. Ved tilsvarende økning i intervallet for hva som utgjør en fullgod MSE-verdi ser vi at andelen ikke synker like raskt som ved tilpassing av zinb-datasett. Ved kriteriet $< 0.005\%$ er ZANB best i 39.0% av parameterkombinasjonene for konstantleddet, og i 41.3% for forklaringsleddet. Tilsvarende er ZINB best i 17.2% av kombinasjonene for konstantleddet og i 2.6% av disse for forklaringsleddet.

Ved bruk av de gitte kriteriene i tabellene, ser vi at regresjonsparametrene i NB-prosessen generelt blir best tilpasset ved bruk av ZINB på zinb-datasett og ved bruk av ZANB på ZANB-datasett.

Ved tilpassing av zanb-datasettene er ikke forskjellen på bruk av de to modellene like store i NB-prosessen som de er ved tilpassing av zinb-datasett. Ved bruk av EM-algorimen ga ZINB enda bedre tilpassing enn ZANB ved tilpassing av zinb-datasett, mens på zanb-datasett var resultatene mer lik de vi fikk ved bruk av BFGS-algoritmen. Tendensene av hvilken modell som ga best resultater i NB-prosessen var riktignok den samme ved bruk av de to ulike algoritmene.

		$= 0$	< 0.0001	< 0.0005	< 0.001	< 0.005
konstantledd	ZINB best	75.2 %	73.2 %	66.3 %	59.8 %	33.9 %
	ZANB best	24.8 %	21.9 %	14.7 %	9.1 %	0.6 %
	Like gode	0 %	4.9 %	19.0 %	31.1 %	65.5 %
forklaringsledd	ZINB best	80.6 %	75.4 %	59.3 %	49.0 %	19.9 %
	ZANB best	19.4 %	12.6 %	13.3 %	0 %	0 %
	Like gode	0 %	12.0 %	27.4 %	51.0 %	80.1 %

Tabell 6.1: Andel av parameterkombinasjonene der modellene, tilpasset på zinb-datasett, gir lavest MSE-verdier i NB-del, ved ulike kriterier for når MSE-verdier viser til fullgod tilpassing.

		$= 0$	< 0.0001	< 0.0005	< 0.001	< 0.005
konstantledd	ZANB best	52.7 %	51.0 %	47.1 %	44.7 %	39.0 %
	ZINB best	47.3 %	46.0 %	42.8 %	38.6 %	17.2 %
	Like gode	0 %	3.0 %	10.1 %	16.7 %	43.8 %
forklaringsledd	ZANB best	57.4 %	53.0 %	48.8 %	47.3 %	41.3 %
	ZINB best	42.6 %	39.5 %	29.2 %	20.5 %	2.6 %
	Like gode	0 %	7.5 %	22.0 %	32.2 %	56.1 %

Tabell 6.2: Andel av parameterkombinasjonene der modellene, tilpasset på zanb-datasett, som gir lavest MSE-verdier i NB-del, ved ulike kriterier for når MSE-verdier viser til fullgod tilpassing.

6.2.2 Resultat for binomisk prosess

I forrige delkapittel viste blant annet figurene 6.4 og 6.6 at bruk av ZANB på zanb-datasett gir MSE-verdier som er sentrerte rundt fem ulike verdier i den binomiske prosessen. Sistnevnte verdier er tilnærmet lik $(0.025, 0.4, 1.5, 3.4, 6.1)$. Andel parameterkombinasjoner som gir disse verdiene er henholdsvis (12.9%, 24.6%, 22.9%, 20.4%, 19.2%) for konstantleddet, og (10.7%, 21.6%, 22.0%, 22.3%, 23.4%) for forklaringsleddet. Bruk av ZANB på zinb-datasett gir derimot varierende verdier, i intervallene $(0.04, 9.22)$ og $(0.02, 11.28)$ for henholdsvis konstantleddet og forklaringsleddet. Figurene viser også at det er flest av de lavere MSE-verdiene.

Ved bruk av ZINB oppnår vi MSE-verdier som varierer i en enda større grad. Majoriteten av verdiene er lave, men en del er også svært høye. Ved tilpassing av zinb-datasett gir modellen

MSE-verdier<1 i over 50% av parameterkombinasjonene og MSE-verdier<10 i over 93% av kombinasjonene for konstantleddet. For de resterende tilfellene oppnås verdier helt opp mot MSE$=49$. I figur 6.4 ser vi at parameterkombinasjoner som gir høye verdier oppnådd med ZINB, gir relativt lave verdier av MSE-verdi ved tilpassing med ZANB. Når vi ser på forklaringskoeffisienten er tendensen litt annerledes. Her får vi en enda større andel av lave verdier, MSE-verdi<1 i 70.6% av parameterkombinasjonene, og MSE<10 i 98.0% av kombinasjonene. Resten av tilfellene varierer opp mot MSE$=30$. Figur 6.4 viser at disse høye MSE-verdiene forekommer i parameterkombinasjoner som gir høye MSE-verdier ved bruk av ZANB.

For å se om den den korrekte modellen er den beste å bruke i den binomiske prosessen, vil vi også her se på differansen i MSE-verdiene modellene gir. De fleste kriteriene for at en MSE-verdi er fullgod er her satt større, ($=0$, <0.1, <0.6, <2.0, <4.5, <8.0), for å se på eventuelle endringer i andelen av parameterkombinasjonene med lavest MSE-verdi de to modellene gir. I figur 6.3 ser vi andelene som er oppnådd ved tilpassing av zinb-datasett med ZINB og ZANB. For konstantleddet er andel kombinasjoner som gir lavest MSE-verdi ved bruk av ZINB lavere (40.1 %) enn andelen ZANB gir (59.9 %). Tendensen er den samme for alle kriteriene som er satt. For forklaringsleddet er ZINB best ved alle kriteriene, modellen gir lavere MSE-verdier i 71.2 % av parameterkombinasjoner ved det strengeste kravet. Ved tilpassing av zanb-datasett ser vi fra figur 6.4 at den feile modellen er best på å tilpasse koeffisienten i forklaringsleddet. Modellen ZINB gir lavere MSE-verdier i 63.8 % av parameterkombinasjonene. Dette gjelder alle kriteriene som er satt, bortsett fra <8.0. Til gjengjeld er det ikke realistisk å sette en så høy tolleransegrense for en MSE-verdi som gir fullgod tilpassing. For konstantleddet gir ZANB lavere verdi i 55.8% av parameterkombinasjonene og modellen er den beste for alle kriteriene.

Ved bruk av EM-algoritmen for tilpassing med ZINB får vi enda sterkere indikasjoner på at ZANB er den beste modellen å tilpasse zinb-datasett med for konstantleddet sin del. Her gir ZANB lavere MSE-verdier i 70.4% av parameterkombinasjonene. For forklaringsleddet gir ikke bruk av modellene like store forskjeller som de gjorde ved bruk av BFGS-algoritmen; 53.7% av kombinasjonene gir lavere MSE-verdier ved bruk av ZINB. Ved tilpassing av zanb-datasett gir bruk av EM-algoritmen en svakere indikasjon på at ZINB er den beste modellen for forklaringsleddet enn BFGS. Men modellen gir fremdeles en stor nok andel av parameterkombinasjonene der MSE-verdien er lavere, 55%. For konstantleddet gir ZANB enda flere kombinasjoner med lavest MSE-verdi, 63%

		$= 0$	< 0.1	< 0.6	< 2.0	< 4.5	< 8.0
konstantledd	ZINB best	40.1 %	36.4 %	24.9 %	9.6 %	1.8 %	0.1 %
	ZANB best	59.9 %	56.1 %	42.5 %	24.3 %	13.3 %	7.3 %
	Like gode	0 %	7.5 %	32.6 %	66.1 %	84.9 %	92.6 %
forklaringsledd	ZINB best	71.2 %	66.7 %	53.9 %	33.9 %	15.1 %	1.1 %
	ZANB best	28.8 %	23.8 %	12.1 %	3.9 %	1.5 %	0.4 %
	Like gode	0 %	9.5 %	34.0 %	62.2 %	83.4 %	98.5 %

Tabell 6.3: Andel av parameterkombinasjonene der modellene, tilpasset på zinb-datasett, som gir lavest MSE-verdier i binomisk del, ved ulike kriterier for når MSE-verdier viser til fullgod tilpassing.

		$= 0$	< 0.1	< 0.6	< 2.0	< 4.5	< 8.0
konstantledd	ZANB best	55.8 %	52.6 %	38.6 %	26.7 %	20.0 %	15.2 %
	ZINB best	44.2 %	41.5 %	33.4 %	18.6 %	5.9 %	0 %
	Like gode	0 %	5.9 %	28.0 %	54.7 %	74.1 %	84.8 %
forklaringsledd	ZANB best	36.2 %	32.8 %	20.9 %	11.6 %	5.5 %	2.4 %
	ZINB best	63.8 %	60.7 %	52.9 %	34.8 %	16.4 %	0 %
	Like gode	0 %	6.5 %	26.2 %	53.6 %	78.1 %	97.6 %

Tabell 6.4: Andel av parameterkombinasjonene der modellene, tilpasset på zanb-datasett, som gir lavest MSE-verdier i binomisk del, ved ulike kriterier for når MSE-verdier viser til fullgod tilpassing.

6.2.3 Resultat for dispersjonsparameter

Figurene i forrige delkapittel, 6.7 og 6.8, viste at ZINB og ZANB gir MSE-verdier for dispersjonsparameteren som er relativt like. Vi vil derfor nå se på hvilken modell som gir lavest verdi ved de samme kriteriene som ble brukt i analysen i NB-prosessen; tilpassing er fullgod når absolutt-verdien til MSE-verdien er $= 0$, < 0.0001, < 0.0005, < 0.001, < 0.005 og < 0.01.. Figur 6.5 viser andel parameterkombinasjoner der modellene gir lavest MSE-verdie for zinb- og zanb-datasett. Her er kombinasjonene som ga ekstremverdier også tatt med. Vi ser at den korrekte modellen i forhold til datasettene faktisk gir lavest andel. For zinb-datasett er ZINB best i kun 32.4%, og for zanb-datasett er ZANB best i 27.8% av parameterkombinasjonene. Ved tilpassing av zinb-datasett gir ZINB derimot større andel av kombinasjoner der modellen gir lavere MSE-verdier enn ZANB når kriteriet er satt til < 0.005 og < 0.01.

		$= 0$	< 0.0001	< 0.0005	< 0.001	< 0.005	< 0.01
zinb-datasett	ZINB best	32.4 %	24.7 %	22.7 %	21.2 %	16.1 %	13.4 %
	ZANB best	67.6 %	55.2 %	40.5 %	32.6 %	15.3 %	8.7 %
	Like gode	0 %	20.1 %	36.8 %	46.2 %	68.6 %	77.9 %
zanb-datasett	ZANB best	27.8 %	17.6 %	11.2 %	8.6 %	2.8 %	1.5 %
	ZINB best	72.2 %	68.0 %	61.1 %	54.7 %	34.4 %	23.1 %
	Like gode	0 %	14.4 %	27.7 %	36.7 %	62.8 %	75.4 %

Tabell 6.5: Andel av parameterkombinasjonene der modellene, tilpasset på zinb- og zanb-datasett, som gir lavest MSE-verdier for dispersjonsparameteren, ved ulike kriterier for når MSE-verdier viser til fullgod tilpassing.

6.2.4 Oppsummering

Ved hjelp av tabellene 6.1- 6.5, har vi sett at den korrekte modellen ikke alltid gir flest parameterkombinasjoner med lavest MSE-verdi. Dette var spesielt åpenbart for parametrene i den binomiske prosessen, der ZANB ga bedre resultater for konstantleddet ved tilpassing av zinb-datasett, og der ZINB ga bedre resultater for koeffisienten i forklaringsleddet ved tilpassing av zanb-datasett.

6.3 Relativ differanse

Modellene ZINB og ZANB gir som tidligere nevnt tilnærmet like MSE-verdier i mange av parameterkombinasjonene for de ulike parametrene, mens i noen kombinasjoner er det stor forskjell på verdiene. For å ta hensyn til hvilken størrelse MSE-verdiene består av ved sammenligning av modellen, tar vi i bruk den relative differansen.

6.3.1 Relativ differanse med hensyn på korrekt bruk av modell

For å finne ut hvor mye bedre den korrekte modellen er for de ulike parametrene i hver kombinasjon, bruker vi den relative differansen av MSE-verdiene som er oppnådd ved bruk av modellene. Ved tilpassing av zinb-datasett ser vi først på uttrykket

$$\frac{\text{MSE}(\text{estimat}_{\text{ZINB}}) - \text{MSE}(\text{estimat}_{\text{ZANB}})}{\text{minste verdi av}(\text{MSE}(\text{estimat}_{\text{ZINB}}), \text{MSE}(\text{estimat}_{\text{ZANB}}))}.$$

Når vi ser på hvor mye bedre ZANB er på å tilpasse parametrene i zanb-datasett i forhold til ZINB bruker vi uttrykket

$$\frac{\text{MSE}(\text{estimat}_{\text{ZANB}}) - \text{MSE}(\text{estimat}_{\text{ZINB}})}{\text{minste verdi av}(\text{MSE}(\text{estimat}_{\text{ZINB}}), \text{MSE}(\text{estimat}_{\text{ZANB}}))}.$$

Med disse uttrykkene vil vi oppnå negative verdier for parameterkombinasjoner når den korrekte modellen gir den laveste MSE-verdien for parameterestimatet. Jo nærmere absoluttverdien av den relative differansen er verdien null, dess likere er MSE-verdiene i de to modellene. Ved bruk av

relativ differanse blir det tatt hensyn til størrelsesordenen på MSE-verdiene. Verdier av MSE på 0.5 og 0.75 har absolutt differanse lik 0.25, og relativ differanse lik 0.5. For større verdier av MSE, eksempelvis 5 og 5.25, kan den absolutte differansen være like stor, men gi lavere relativ differanse, 0.05. Vi ser nok en gang på resultatene for parameterestimatene i NB-prosessen, den binomiske prosessen, og for estimatene dispersjonsparameteren separat. dispersjonsparameteren.

Tabell 6.6 viser resultatene ved bruk av summary-funksjonen i R på den relative differansen for hver av parameterestimatene i de 3752 parameterkombinasjonene. Fra tabellen ser vi de samme tendensene som vi så i figurene i kapittel 6.1, og i tabellene i kapittel 6.2. For å gi et eksempel på dette, ser vi på den relative differansen til MSE-verdiene for estimatene av konstantleddet i NB-prosessen der zinb-datasett har blitt tilpasset. Her er minste verdi av relativ differanse lik -51.2, mens den er tilnærmet lik -0.25 ved grensen til første kvartil av de sorterte verdiene. Ved en nærmere undersøkelse av verdiene av den relative differansen ser vi at kun ti av parameterkombinasjonene ga verdi lavere enn -5.0. De fleste av disse tilsvarer punktene som ligger mest synlig ovenfor linjen som viser til like MSE-verdier ved bruk av modellene, i plottet øverst til venstre i figur 6.1. Vi ser at de fleste verdiene er negative, til og med i grenseverdien ved tredje kvartil; altså er den korrekte modellen ZINB best i over 75% av parameterkombinasjonene. Dette kom vi også frem til i tabell 6.1, som viste at ZINB ga lavest MSE-verdier i 75.2% av parameterkombinasjonene. Videre ser vi at den største verdien av relativ differanse er 0.33630. Den er positiv, det vil si at ZANB gir lavere MSE-verdi enn ZINB, men den har samtidig en relativt lav absoluttverdi. Dette kan vi såvidt skimte fra plottet øverst til høyre i figur 6.1; eventuelle punkter under tidligere nevnte linje, viser også til relativt lave MSE-verdier ved bruk av ZINB.

For de andre parametrene kan vi på tilsvarenede måte sammenligne verdiene i tabell 6.6 med figurene 6.1 - 6.8 og tabellene 6.1 - 6.5, og dermed se at resultatene stemmer overens.

For dispersjonsparameteren var laveste verdi av relativ differanse for parameterestimatene oppnådd ved tilpassing av zinb-datasett, og høyeste verdi ved tilpassing av zanb-datasett såpass høye i absoluttverdi at summary-funksjonen ga verdien $0.000e+00$ for de verdiene som ligger i grensene til første og tredje kvartil. Verdiene ble da oppnådd på egenhånd.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
NB-Konst.ledd zinb	-51.24000	-0.25490	-0.05287	-0.24270	-0.00038	0.33630
NB-Forkl.ledd zinb	-2.739000	-0.277400	-0.060940	-0.176000	-0.007696	0.064440
NB-Konst.ledd zanb	-13.01000	-1.514000	-0.008822	-1.139000	0.133700	1.723000
NB-Forkl.ledd zanb	-15.57000	-2.41400	-0.02582	-1.61000	0.07100	1.54600
Bin-Konst.ledd zinb	-44.6600	-1.0940	0.8657	4.8070	5.5730	159.5000
Bin-Forkl.ledd zinb	-76.6900	-6.6350	-1.3580	-5.9120	0.2018	71.2900
Bin-Konst.ledd zanb	-573.8000	-4.5650	-0.3314	-5.7010	1.5570	44.5900
Bin-Forkl.ledd zanb	-112.8000	-0.8874	1.3830	2.1330	6.9400	78.4500
Disp-zinb	1.178e+14	0.01271	0.02403	-6.271e+10	0.11053	8.72098
Disp-zanb	-1.73974	-0.00389	0.17291	1.646e+6	0.45254	5.993e+9

Tabell 6.6: resultater fra summary(): Relativ differanse for MSE-verdier til de ulike parameterestimatene ved bruk av ZINB og ZANB.

6.3.2 Relativ differanse med hensyn på den beste modellen for hver parameterkombinasjon

For å finne ut hvor mye bedre den beste modellen er på å tilpasse de ulike datasettene, og således hvor store konsekvensene ved bruk av den dårligste modellen blir, ser vi på den relative differansen til den beste modellen for alle parameterkombinasjonene. Vi bruker da følgende uttrykk for både zinb- og zanb-datasett.

$$\frac{\text{MSE(best)} - \text{MSE(dårligst)}}{\text{MSE(best)}}$$

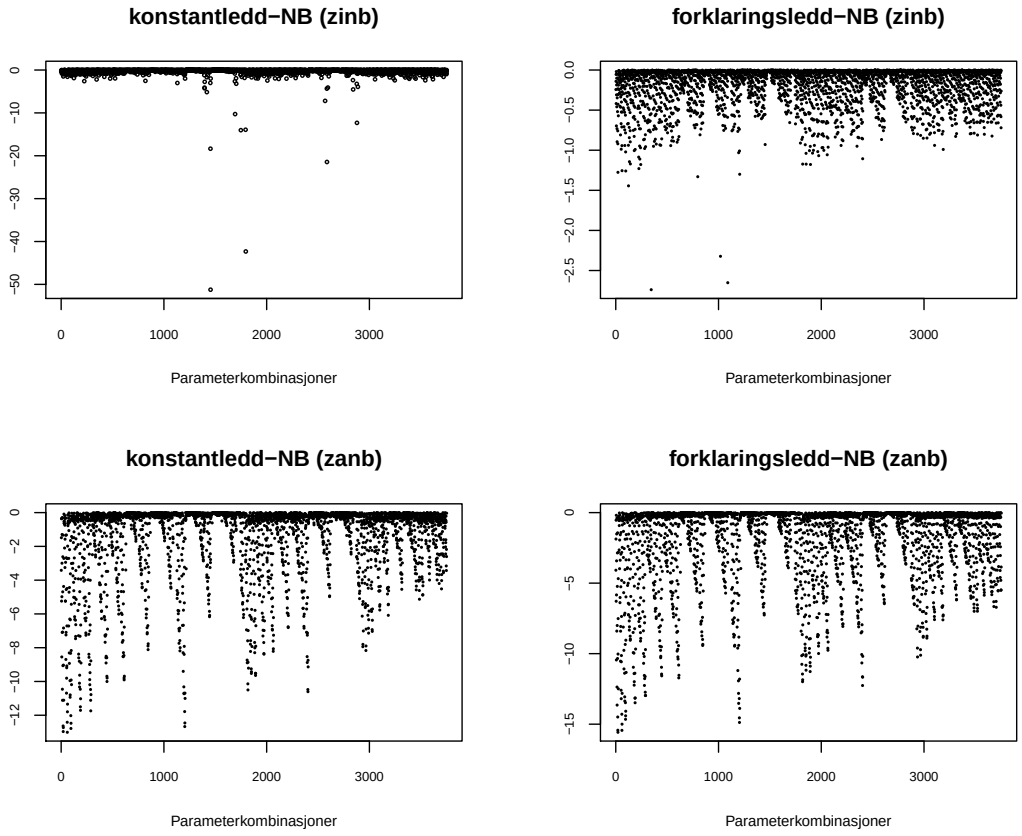
Verdien på den relative differansen til den beste modellen vil derfor være negativ, og vil vise til større forskjell ved bruk av modellene jo lavere den er. Vi undersøker forskjellene på den beste og den dårligste modellen for parametrene i de ulike prosessene hver for seg.

Resultater fra NB-prosessen

Figur 6.9 viser den relative differansen for MSE-verdiene til den beste modellen i forhold til den dårligste modellen, for hver parameterkombinasjon. De ulike plottene viser resultatene for konstantleddet og forklaringsleddet i NB-prosessen ved tilpassing av zinb- og zanb-datasett. Vi ser at de fleste parameterkombinasjonene gir differanse nær null, det vil si at det da ikke er stor forskjell ved bruk av ZINB og ZANB. Ved å se nærmere på skalaen på y-aksene, ser vi at det er minst forskell ved bruk av modellene for forklaringsleddet i binomisk del. Ved tilpassing av zanb-datasett ser vi flere parameterkombinasjoner der forskjellene på den beste og den dårligste modellen er mye større enn ved tilpassing av zinb-datasett.

Ulike verdier av den eksakte verdien til dispersjonsparameteren ser ikke ut til å endre stort på resultatene som oppnås. Det kan se ut som høyere eskakte verdier av den gir litt mindre forskjell

på den relative differansen, spesielt for parametrene ved tilpassing av zanb-datasett.



Figur 6.9: Relativ differanse av MSE-verdi for den beste modellen for parametrene i NB-prosessen, for hver parameterkombinasjon.

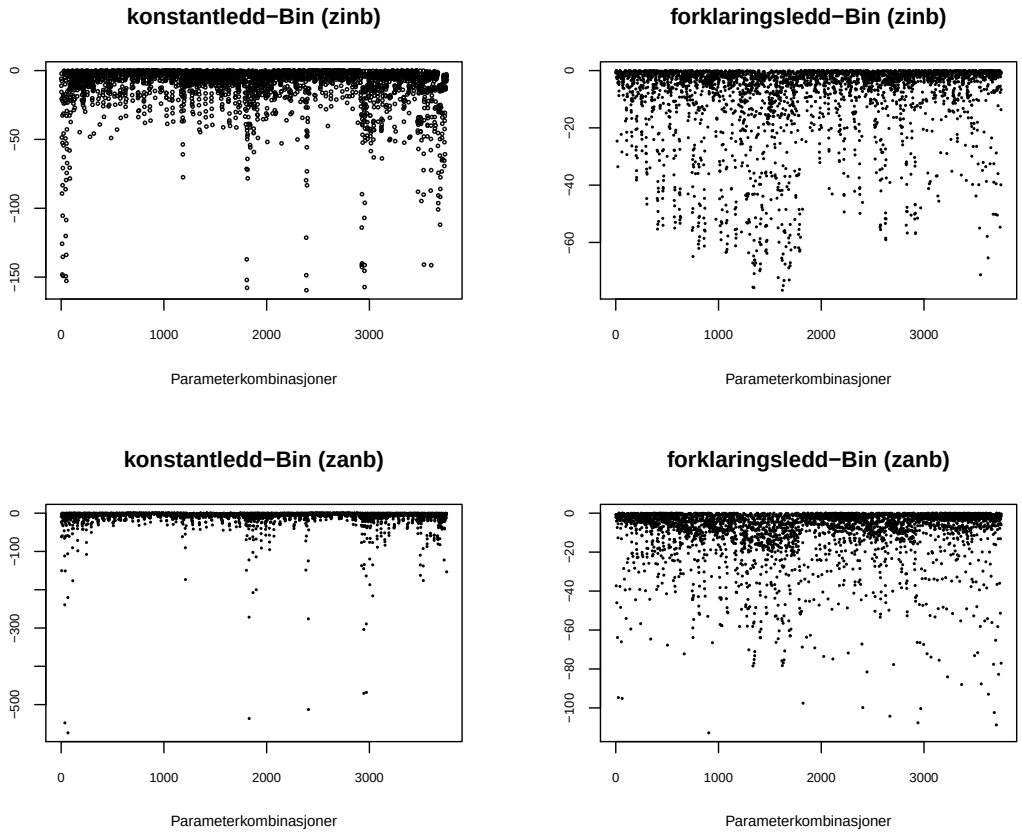
Vi skal se litt nærmere på den relative differansen av MSE-verdiene for hver av parametrene. Tabell 6.7 viser hvor stor andel av parameterkombinasjonene der absoluttverdien av den relative differansen ligger i ulike intervaller. Fra tabellen ser vi at over 80% av kombinasjonene ligger i intervallet $[0.01, 1)$ for både konstantleddet og forklaringsleddet når datasettet er zinb-fordelt. Det er få verdier av den relative differansen som er store, kun 3.3% av kombinasjonene gir større verdi enn én for konstantleddet, den største av disse er -51.2. For forklaringsleddet får vi enda lavere andel der absoluttverdien til den relative differansen er større enn 1 (0.8%). Ved zanb-fordelte datasett ser vi en større andel av kombinasjonene der absoluttverdier større enn én; 29.5% og 35.0% for henholdsvis konstantleddet og forklaringsleddet. Her er de laveste verdiene av den relative differansen på -13.0 for konstantleddet og -15.6 for forklaringsleddet, så den største forskjellen på modellene er ikke like stor som for konstantleddet ved zinb-fordelte data.

		$[0, 0.001)$	$[0.001, 0.01)$	$[0.01, 0.1)$	$[0.1, 1)$	$[1, \infty)$
zinb	konst.ledd	1.2 %	11.8 %	45.4 %	38.3 %	3.3 %
	forkl.ledd	0.9 %	14.5 %	41.9 %	41.9 %	0.8 %
zanb	konst.ledd	0.7 %	4.7 %	23.9 %	41.2 %	29.5 %
	forkl.ledd	0.7 %	6.4 %	25.6 %	32.3 %	35.0 %

Tabell 6.7: Andel av parameterkombinasjoner der absoluttverdien av den relative differansen som ligger i ulike intervaller for parametrene i NB-prosessen.

Resultater fra den binomiske prosessen

For parametrene i den binomiske prosessen er den relative differansen til den beste modellen mot den dårligste for hver parameterkombinasjon plottet i figur 6.10. Her ser vi at forskjellen på den beste og den dårligste modellen gir mye større utslag enn for parametrene i NB-prosessen. Konsekvensene ved bruk av den dårligste modellen vil atså være større. Ulike eksakte verdier av dispersjonsparameteren ser ikke ut til å gi et annerledes mønster for differansene.



Figur 6.10: Relativ differanse av MSE-verdi for den beste modellen for parametrene i den binomiske prosessen, for hver parameterkombinasjon.

Andelen av parameterkombinasjoner der absoluttverdien av den relative differansen ligger i ulike intervaller er vist i tabell 6.8. Her ser vi at andelen er over 70% i intervallet $[1, \infty)$ for begge parametrene ved tilpassing av både zinb- og zanb-datasett. De laveste verdiene av relativ differanse oppnådd for zinb-datasett er -159 og -77 for henholdsvis konstantleddet og forklaringsvariabelen. Dette viser til en veldig stor forskjell på beste og dårligste modell. For zanb-datasett var tilsvarende laveste verdier -574 og -113, og viser dermed til enda større forskjeller.

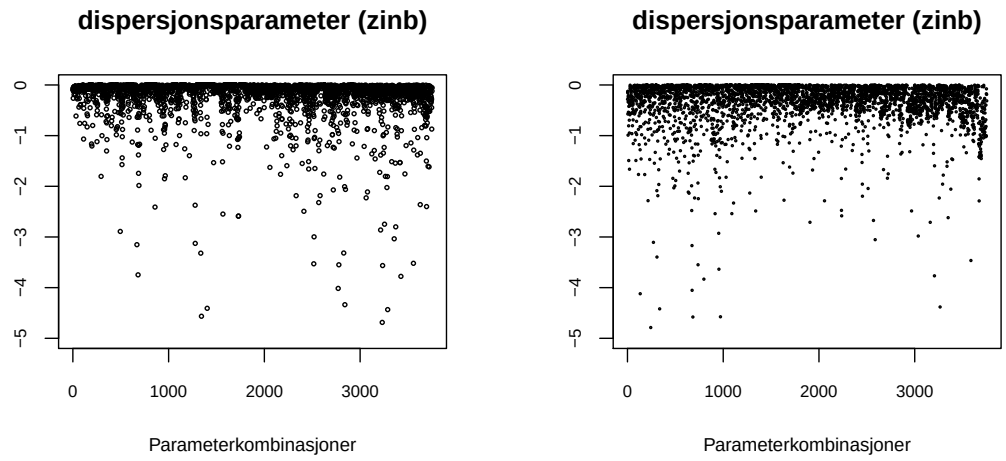
		$[0, 0.001)$	$[0.001, 0.01)$	$[0.01, 0.1)$	$[0.1, 1)$	$[1, 10)$	$[10, \infty)$
zinb	konst.ledd	< 0.1 %	0.4 %	2.7 %	23.0 %	48.8 %	25.0 %
	forkl.ledd	0.3 %	0.5 %	4.4 %	23.3 %	48.0 %	23.5 %
zanb	konst.ledd	0 %	0.5 %	4.1 %	25.0 %	52.2 %	18.2 %
	forkl.ledd	0 %	0.3 %	3.2 %	19.8 %	51.4 %	25.3 %

Tabell 6.8: Andel av parameterkombinasjoner der absoluttverdien av den relative differansen som ligger i ulike intervaller for parametrene i den binomiske prosessen.

Resultater for dispersjonsparameteren

Figur 6.11 viser den relative differansen av MSE-verdiene for den beste modellen for dispersjonsparameteren. Plottene viser kun verdiene som er større enn -5, for å få et visuelt inntrykk av forskjellene for verdiene som ikke er ekstremt store. I plottet der zinb-fordelte datasett er tilpasset er totalt 274 punkter utelatt. Disse varierte i intervallet $(-1.2e+14)$, der over 87% var lavere enn -1000. I plottet til høyre er 140 punkter utelatt, hvorav over 87% var lavere enn -1000, og laveste verdi var $-6.0e+09$.

For å få et totalinntrykk av hvor mye bedre den beste modellen er for hver parameterkombinasjon ser vi igjen på andelen av kombinasjoner der absoluttverdien til den relative differansen ligger i ulike intervaller. Fra tabell 6.9 ser vi at nær 90% av kombinasjonene gir absoluttverdier i intervallet [0.001,1) der både zinb- og zanb-datasett har blitt tilpasset. Det ser med andre ord ut til at modellene gir ut tilnærmet like resultater, bortsett fra i de tilfellene der absoluttverdien er ekstremt stor. Sistnevnte tilfeller er konsekvens av parameterestimater som avviker betraktelig fra majoriteten av de resterende estimatene etter 10.000 simuleringer.



Figur 6.11: Relativ differanse av MSE-verdi for den beste modellen for dispersjonsparameteren.

		$[0, 0.001)$	$[0.001, 0.01)$	$[0.01, 0.1)$	$[0.1, 1)$	$[1, 10)$	$[10, \infty)$
zinb	konst.ledd	3.9 %	11.6 %	40.6 %	33.1 %	3.7 %	7.1 %
zanb	forkl.ledd	0.7 %	8.8 %	26.9 %	54.0 %	6.3 %	3.3 %

Tabell 6.9: Andel av parameterkombinasjoner der absoluttverdien av den relative differansen som ligger i ulike intervaller for dispersjonsparameteren.

Oppsummering

Ved tilpassing av zanb-datasett så vi i tabell 6.7 at en relativt liten del av parameterkombinasjoner ga store forskjeller ved bruk av den beste og den dårligste modellen for parametrene i NB-prosessen. Valg av modell vil derfor ikke nødvendigvis være avgjørende for hvor god tilpassing vi får av disse parametrene. Fra figur 6.10 og tabell 6.8 så vi at bruk av den dårligste modellen kan gi store konsekvenser for parameterestimatene i den binomiske prosessen. I tillegg så vi fra figurene 6.9, 6.10 og 6.11 at tendenser til mindre og større forskjeller på den relative differanse ikke ga klare ulike trekk for de ulike verdiene av dispersjonsparameteren. Ved å rangere verdiene av den relative differansen for de ulike parametrene, var det en klar tendens til lave verdier av forventningen i NB-prosessen ved store absoluttverdier.

6.4 Modellvalg med hensyn på ulike egenskapsverdier

Et av hovedmålene i oppgaven var å undersøke om modellvalg mellom ZINB og ZANB kan bli tatt på et annet grunnlag enn det som hittil har blitt gjort i praksis. Det mest praktiske vil være om man kan se hvilken modell som passer best ved å se på egenskapene i datasettene. Vi har allerede sett at valg av modell gir ulik tilpassing for parametrene i modellene og at den korrekte modellen ikke alltid gir størst andel av lavest MSE-verdi for parameterkombinasjonene som er valgt. Ved

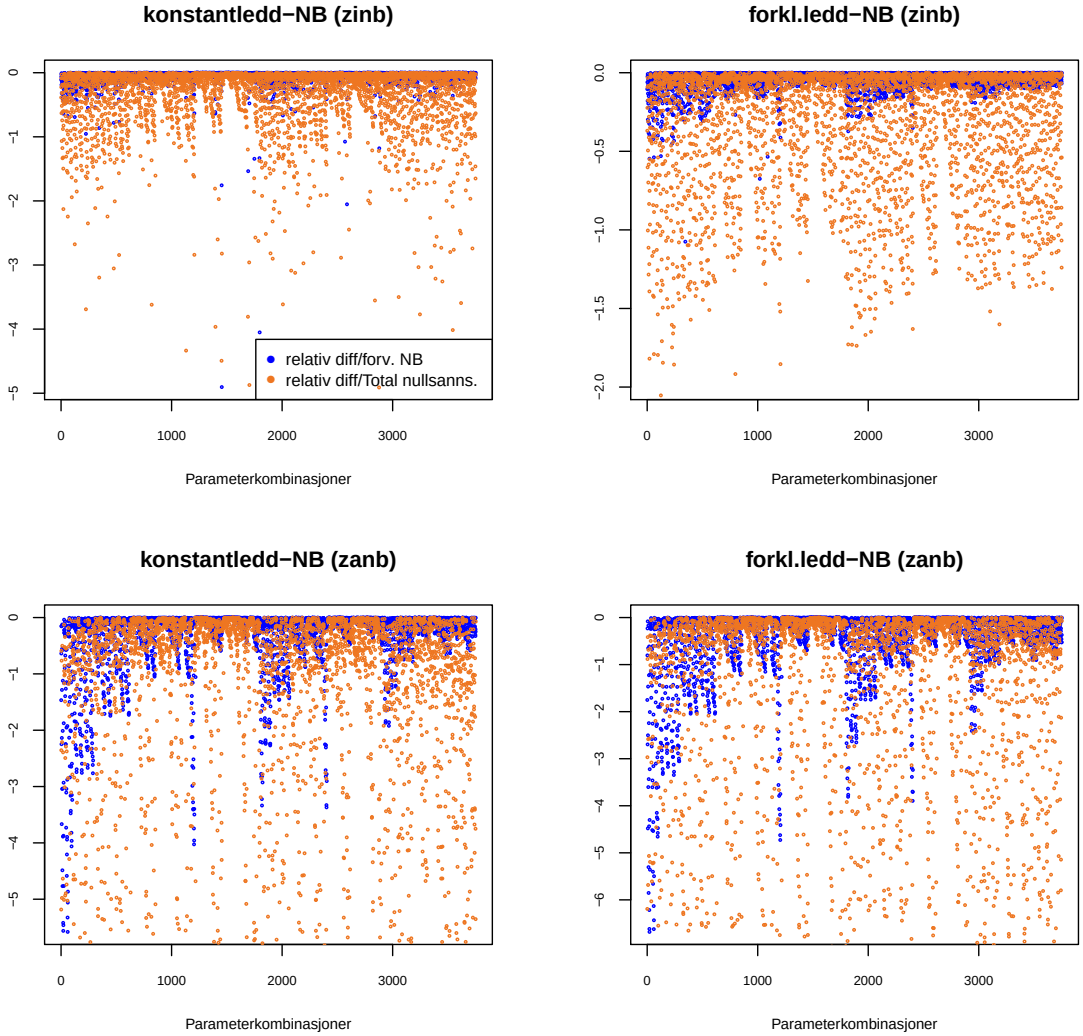
tilpassing av zinb-datasett ga ZANB bedre resultater for konstantleddet, og ved tilpassing av zanb-datasett ga ZINB bedre resultater for koeffisienten i forklaringsleddet i den binomiske prosessen. For estimatverdiene til dispersjonsparameteren var den gale modellen best i klart flere tilfeller for både zinb- og zanb-datasett. Tabell 6.9 viste derimot at den dårligste modellen ga tilnærmet like gode resultater i nesten 90% av parameterkombinasjonene.

I forrige delkapittel fant vi et mål på hvor mye bedre den beste modellen er i forhold til den dårligste for hver av parametrene, for hver parameterkombinasjon. Ved bruk av den relative differansen vil vi nå se om det er en sammenheng mellom denne og verdiene til egenskapene i fordelingene modellene tar utgangspunkt i, for hver parameterkombinasjon.

Dersom ZINB er den beste modellen i en kombinasjon, blir den totale nullsannsynligheten annerledes enn hvis ZANB er den beste. Vi skal se om forholdene (relativ differanse)/(forventning i NB-prosess) og (relativ differanse)/(total sannsynlighet for nullobservasjoner) gir like trender.

6.4.1 Resultater for NB-prosess

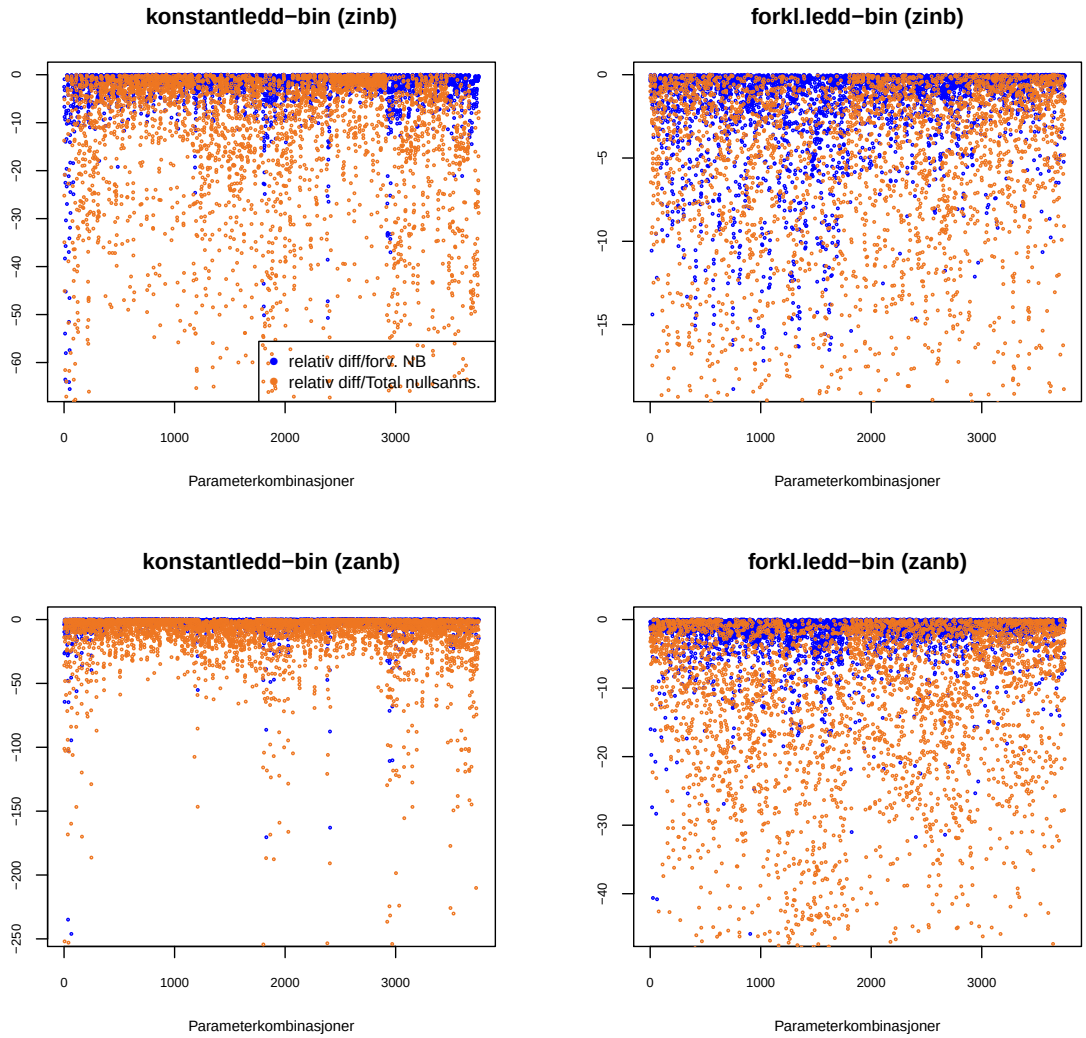
Figur 6.12 viser plott av forholdet mellom relativ differanse for den beste modellen, og forventningen i de to prosessene for alle parameterkombinasjonene. Vi ser tilfeller av like trender i alle plottene. For konstantleddet ved zinb-fordelte datasett ser vi at klynger av det ene forholdet gir den samme tendensen for det andre forholdet. I tillegg ser vi at lavere verdier av det ene gir tilsvarende for det andre forholdet. For forklaringsleddet ved tilpassing av zinb-datasett ser vi like spor av opphopning av verdiene. For parametrene ved tilpassing av zanb-datasett er likhetstrekkene de to ulike verdiene gir ikke like lette å se fra figurene, men de er likevel merkbare nok til å kunne se at det er sammenheng mellom en god del av parameterkombinasjonene.



Figur 6.12: Relativ differanse av MSE-verdi/Forventningsverdier for parametrene i NB-prosessen.

6.4.2 Resultater for binomisk prosess

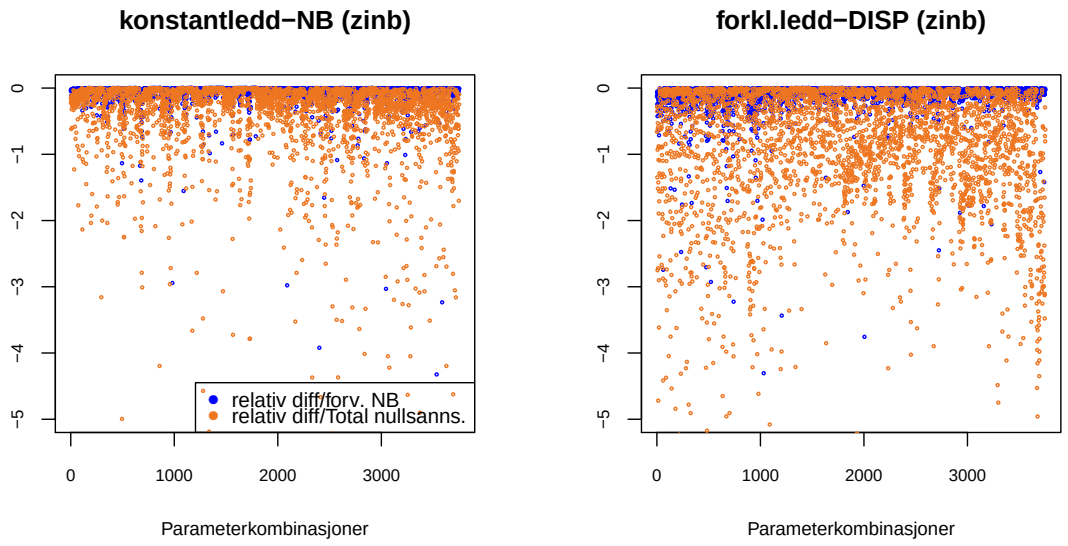
Figur 6.13 viser tilsvarende forholdsverdier som vi så på i NB-prosessen. Her ser vi at konstantleddet ved tilpassing av både zinb- og zanb-datasett gir like tendenser for forholdene. Plottene til høyre er ikke trekkene like enkle å spore. Ved å skalere y-aksen for så og se nærmere på plottene, var det en tendens til at parameterkombinasjoner som ga lave verdier av forholdet (relativ differanse av MSE-verdier)/(forventning i NB) ga merkbart lavere verdier av forholdet (relativ differanse av MSE-verdier)/(total sannsynlighet for nullobservasjon).



Figur 6.13: Relativ differanse av MSE-verdi/Forventningsverdier for parametrene i den binomiske prosessen.

6.4.3 Resultater for dispersjonsparameteren

Det kan også være interessant å se hvordan forholdet mellom den relative differansen og de to ulike egenskapene endrer seg for dispersjonsparameteren, siden denne også estimeres med de to modellene. Fra figur 6.14 ser vi at verdiene er tilsynelatende like for mange av parameterkombinasjonene.



Figur 6.14: Relativ differanse av MSE-verdi/Forventningsverdier for dispersjonsparameteren.

6.4.4 Oppsummering

Selv om det ikke alltid er like lett å se at det er en sammenheng mellom de to forholdene fra plottene i figurene 6.12 - 6.14, så viser de fleste like trekk for mange av parameterkombinasjonene. Vi ser altså en sammenheng mellom forventningen i NB-prosessen og den totale nullsannsynligheten i forhold til bruk av den beste modellen.

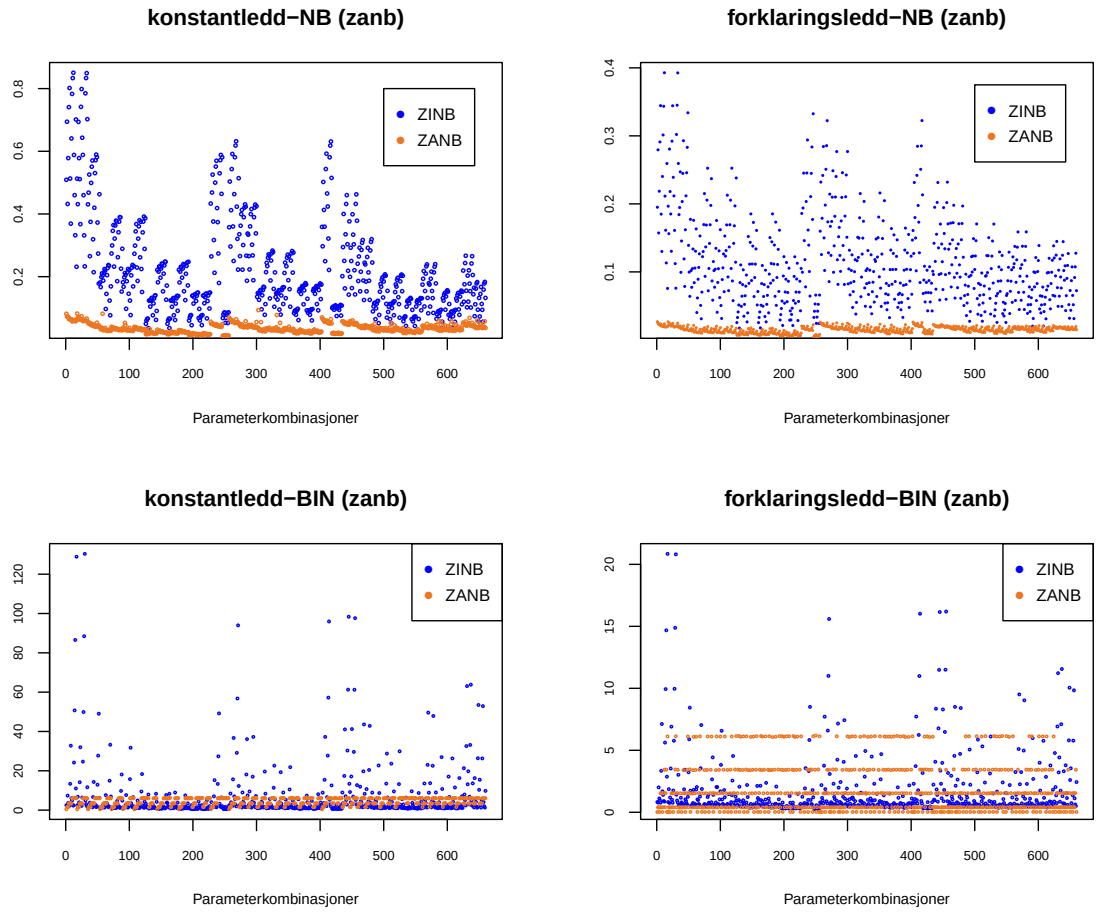
6.5 Færre nullobservasjoner i ZANB enn NB

I kapittel 4.1 så vi at i tilfeller der sannsynligheten for å oppnå en nullobservasjon fra et ZANB-datasett ligger i intervallet $(0, P(y_{\text{NB}} = 0))$, kan vi få datasett med færre nullobservasjoner enn det en NB-fordeling ville ha gitt. Dette kan ikke forekomme i et ZINB-datasett, noe vi lett ser fra uttrykk (4.2). Det vil derfor være interessant å se hvor store forskjeller bruk av ZINB og ZANB gir ved tilpassing av zanb-datasett der dette er tilfelle. Fra hovedresultatene oppnås det 660 parameterkombinasjoner der zanb-datasettene har lavere sannsynlighet for nullobservasjoner enn et vanlig negativ binomisk datasett vil ha. Vi sammenligner nok en gang MSE-verdiene bruk av modellene gir i de ulike prosessene og for dispersjonsparameteren. Figur 6.15 viser marginalfordelingen av MSE-verdiene oppnådd for parametrene i de to prosessene ved tilpassing av zanb-datasett med de to modellene. Vi ser klart fra plottene at ZANB gir mye lavere MSE-verdier for parametrene i NB-prosessen. I den binomiske prosessen ser vi igjen at ZANB gir verdier som er sentrert rundt de fem ulike verdiene som tidligere nevnt. Det er også synlig at ZINB faktisk gir MSE-verdi som er lavere enn de ZANB gir i mange av de 660 parameterkombinasjonene.

For å se hvor ulike resultater de to modellene gir ved tilpassing av zanb-datasettene, ser vi igjen

på differansen i MSE-verdiene de to modellene gir for parametrene, for hver av de 660 parameterkombinasjonene. For parametrene i NB-prosessen var ZANB den beste modellen for samtlige av parameterkombinasjonene. I den binomiske prosessen var ZANB den beste modellen for konstantleddet i kun 32.6% av kombinasjonene, mens samme modell ga lavere verdier i 53.6% av tilfellene for forklaringskoeffisienten. For dispersjonsparameteren ga ZANB lavere MSE-verdier i kun 1.7% av parameterkombinasjonene.

Bruk av EM-algoritmen ga tilnærmet like resultater som bruk av BFGS. Den største forskjellen var i andelen kombinasjoner der modellene ga lavest MSE-verdi for konstantleddet i den binomiske prosessen. Her var ikke forskjellen på modellene like store; ZANB ga lavere verdier i 45.8% av kombinasjonene, ,



Figur 6.15: Marginalfordelingene av MSE-verdier ved bruk av ZINB og ZANB ved tilpassing av zanb-datasett der nullsannsynligheten er lavere enn $P(y_{\text{NB}} = 0)$

Kapittel 7

Konklusjon og videre arbeid

7.1 Konklusjon

Et av målene i denne masteroppgaven har vært å finne ut om valg av ZINB og ZANB er avgjørende for å oppnå gode resultater ved tilpassing av ulike datasett, i denne oppgaven zinb- og zanb-datasett. I tillegg ønsket vi å se hvor store konsekvenser bruk av gal modell ville gi. Fra resultatene så vi at valg av modell gir ulike resultater for samtlige parameterestimater. Bruk av ZINB på zinb-datasett ga klart best resultater for parametrene i NB-prosessen, mens ZANB var best, men ikke like overlegent, på zanb-datasett. Den relative differansen viste derimot at konsekvensene sjeldent var store ved bruk av den dårligste modellen for denne prosessen. I de tilfellene absoluttverdien av differansen var stor, var det en klar tendens til lave forventningsverdier i NB-prosessen. Resultatene fra den binomiske prosessen var de som var mest overraskende, her viste det seg at den gale modellen var bedre i mange tilfeller, i tillegg til at absoluttverdien av den relative differansen ofte var rimeligt stor. Det ser med andre ord ut som at valg av korrekt modell i forhold til fordelingen i datasettet kan føre til store negative konsekvenser.

Et annet mål i denne oppgaven har vært å finne ut om valg av modellene ZINB og ZANB bør tas på et annet grunnlag enn hva som har vært vanlig i tidligere studier. Vi tok utgangspunkt i en artikkel av Li m.fl [12], der det ble foreslått at innblikk i økologiske eller biologiske krefter som påvirket antallet nullobservasjoner på ulikt vis kunne gi utslag i det beste valget av de to modellene. Bortsett fra dette, med henvisning til litteratur om modellene brukt innenfor fagfeltet, bestod modellvalg av testverdier, og da spesielt Vuong-testen for valg mellom ZINB og ZANB. I denne oppgaven har vi fokusert på resultater fra regresjonskjøring på ZINB- og ZANB-fordelte datasett, med bruk av begge modellene på disse. Vi så at det var en sammenheng mellom MSE-verdiene til den beste modellen og verdiene til egenskapene i de simulerte datasettene. Kaarstad [10] kom også i sin oppgave om ZIP og ZAP frem til at valg av modell ga størst utslag for MSE-verdien til regresjonskoeffisientene i binomisk del.

7.2 Videre arbeid

Hvis det hadde vært mer tid til rådighet, ville det blant annet være interessant å se nærmere på hvilke kombinasjoner av forventningsverdier av egenskapene i datasettene som leder til bruk av den beste modellen.

I oppgaven er det generert 100 observasjoner for hvert datasett som blir tilpasset. Det kunne også ha vært interessant å se hvor gode ZINB og ZANB vil være ved tilpassing av datasett som er enten mindre, eller mye større. Når datasettet består av færre observasjoner, kan det bli vanskeligere for modellene å tilpasse disse. Siden vi har hele fire regresjonsparametre som estimeres for hver modell, kan det tenkes at forskjellene ved bruk av ZINB og ZANB ville bli større, og muligens annerledes i disse tilfellene. Dersom datasettet består av mange observasjoner, er det naturlig å tro at modellene lettere vil finne fram til de eksakte parameterverdiene. I tillegg ville det ha vært interessant å se om parametre fra de to prosessene lar seg påvirke i ulik grad ved tilpassing av store datasett ved bruk av modellene.

Selv om det er sjeldent at et negativ binomisk fordelt datasett i praksis vil passe utenfor restriksjonene som ble satt i tabell 5.2, ville det ha vært interessant å se hvor godt modellene klarte å tilpasse datasettene i disse tilfellene. Både EM-algoritmen og BFGS klarte å gi ut resultater ved lave forventningsverdier i NB-del og høye totale nullsannsynligheter. Dette ble det ikke tid til å analysere i løpet av masteroppgaven. Det ble heller ikke tid til detaljert undersøkelse av påvirkningen på MSE-verdiene til parametersetimatene ulike verdier av dispersjonsparameteren eventuelt gir. I tillegg ville det være interessant å se hvordan større eksakte verdier av dispersjonsparameteren ville ha påvirket tilpassing av datasettene.

Tillegg A

Nulltrunkert negativ binomisk

A.1 Regresjonsanalyse på simulert nulltrunkert NB2- datasett

```
library(gamlss.tr)
library(gamlss)
library(COUNT)

# Setter faste eksakte verdier og definerer vektorer som skal brukes til å
# lagre resultater fra regresjonskjøring.
alfa=0.75
mean_y1 = rep(0,5000)
andel_nuller = rep(0,5000)
beta0_hat_utenNuller_truncNB2 = rep(0,5000)
beta1_hat_utenNuller_truncNB2 = rep(0,5000)
beta2_hat_utenNuller_truncNB2 = rep(0,5000)
se_beta0_hat = rep(0,5000)
se_beta2_hat = rep(0,5000)
se_beta1_hat = rep(0,5000)

# Nedjusterer konvergens-kriteriet
con1 <- gamlss.control(c.crit=0.05)

# Funksjon som genererer datasett og utfører regresjon på disse
# ved ulike eksakte verdier av beta0, beta1 og beta2. Lagrer til
# slutt en liste med resultater.
syntetiskNB2 = function(beta0, beta1,beta2){
  # Utfører 5000 simuleringer av nulltrunkerte datasett og utfører
  # regresjon på hver av disse.
```

```
for(i in 1:5000){
   y_tot=0
   antall=0
   ant_pos=0
   antall_nuller=0
   antall_y_pos = 0
   y_pos=rep(0,100)
   x1_sentrert=rep(0,100)
   x2_sentrert=rep(0,100)

   # genererer 100 nulltrunkerte NB2-data.
   repeat{
    x1 = runif(1)
    x2 = runif(1)
    x1_sentr = x1 - 0.5
    x2_sentr = x2 - 0.5
    xb = beta0 + beta1*x1_sentr + beta2*x2_sentr
    inv_alfa = 1/alfa
    exb = exp(xb)
    xg = rgamma(1, alfa, alfa, inv_alfa)
    xbg = exb*xg
    y1 = rpois(1, xbg)

    antall = antall + 1
    y_tot[antall]=y1
    if(y1>0){
      antall_y_pos = antall_y_pos + 1
      y_pos[antall_y_pos]=y1
      x1_sentrert[antall_y_pos]=x1_sentr
      x2_sentrert[antall_y_pos]=x2_sentr}
    else{antall_nuller=antall_nuller+1}
    if(antall_y_pos==100){
      break}
   }

 # Lagrer andel av nuller i datasettet for hver simulering.
 andel_nuller[i] = antall_nuller/antall

 # Putter data i en dataramme og definerer navn på tilpassede data.
 # Utfører deretter regresjon på datasettet i datarammen.
 sdata = data.frame(y_pos, x1_sentrert, x2_sentrert)
```

```
   gen.trun(0, "NBI", type="left", name="lefttr")
   lt0nb = gamlss(y_pos~x1_sentrert+x2_sentrert, data= sdata, family= NBIlefttr,
           control=con1)

   # Lagrer verdier av parameterestimat og kvadratavvik i vektorer for
   # hver simulering.
   beta0_hat_utenNuller_truncNB2[i] = lt0nb$mu.coefficients[1]
   beta1_hat_utenNuller_truncNB2[i] = lt0nb$mu.coefficients[2]
   beta2_hat_utenNuller_truncNB2[i] = lt0nb$mu.coefficients[3]

   se_beta0_hat[i] = (beta0_hat_utenNuller_truncNB2[i] - beta0)^2
   se_beta1_hat[i] = (beta1_hat_utenNuller_truncNB2[i] - beta1)^2
   se_beta2_hat[i] = (beta2_hat_utenNuller_truncNB2[i] - beta2)^2
  }

  # Beregner og lagrer gjennomsnittet av kvadratavvikene.
  mse_beta0 = mean(se_beta0_hat)
  mse_beta1 = mean(se_beta1_hat)
  mse_beta2 = mean(se_beta2_hat)

  # Lagrer en liste med resultater som kan hentes ved kjøring av
  # funksjonen
  ret = list()
  ret$gj.sn_andel_nuller = mean(andel_nuller)
  ret$gj.sn_beta_hat_truncNB2 = c(mean(beta0_hat_utenNuller_truncNB2),
        mean(beta1_hat_utenNuller_truncNB2),mean(beta2_hat_utenNuller_truncNB2))
  ret$mse = c(mse_beta0, mse_beta1, mse_beta2)
  return(ret)
}
```

Tillegg B

ZINB og ZANB

```
library(MASS)
library(VGAM)
library(pscl)
library(foreach)
library(doParallel)
# Ved tilgang til flere prosessor-kjerner
registerDoParallel(cores=32)

# Antall simuleringer for hver parameterkombinasjon settes.
n <- 10000

# Verdiene til forklaringsvariabelen settes.
x1 <- c(-2,-1,0,1,2)



# Funksjon som genererer datasett og utfører regresjon på disse med
# parameterkombinasjoner som argumenter. Her står konstNB for
# konstantleddet i NB-del, forklBIN er regresjonskoeffisienten som
# påvirker forklaringsvariabelen i binomisk del, osv..
# Lagrer resulater i en liste til slutt.

fun <- function(alfa,konstNB,forklNB,konstBIN,forklBIN){

  # Finner vektorene med forventet verdi i binomisk og NB-prosess i tillegg
  # til total sannsynlighet for nullobservasjoner for de fem ulike verdiene
  # av forklaringsvariabelen.
  phi <- array(dim=length(x1))
  P_null <- array(dim=length(x1))
```

```
mu <- exp(konstNB + forklNB*x1)
for(i in 1:length(x1)){
 # Strukturelle nullobservasjoner i zinb-datasett/
 # Total nullsannsynlighet i zanb-datasett
 phi[i]=exp(konstBIN+forklBIN[i])/(1+exp(konstBIN + forklBIN[i]))
 # Total nullsannsynlighet i zinb-datasett
 P_null[i]=phi[i] + (1-phi[i])*(1/(mu[i]*alfa+1))^(1/alfa)
}

#-- Genererer zinb-fordelte datasett og utfører regresjon med modellene
#-- ZINB, ZANB og NB2.

ZINB_est <- foreach(i=1:n, .combine=rbind)%dopar%{

     # Genererer datasett med 100 observasjoner og ordner disse i dataramme
     # med tilhørende bruk av forklaringsvariabler.
     y_zinb <- rzinegbin(100, munb=mu, pstr0= phi, size=1/alfa)
     yzinbdata <- data.frame(y_zinb, x1)
     yzinbdataframe <- as.data.frame(yzinbdata)

     # Utfører regresjon på ZINB-datasettet med de to modellene
     ZINB <- try(zeroinfl(y_zinb~ x1|x1, data=yzinbdataframe, dist= "negbin",
                 control=zeroinfl.control(BFGS=TRUE)), silent=TRUE)  ## evnt. EM=TRUE
     ZANB <- try(hurdle(y_zinb~ x1|x1, data=yzinbdataframe, dist= "negbin",
                 link="logit"), silent=TRUE)


    # Henter ut og lagrer estimerte verdier av regresjonskoeffisientene fra
    # tilpasset objekt, i tillegg til å lagre kvadrerte feilledd av disse.
    # Lagrer også alle data i datasett for hver simulering.
     zi <- array(NA,127)
     zi[1] <- try(ZINB$coef$count[1],silent=T)
     zi[2] <- try((ZINB$coef$count[1]-konstNB)^2, silent = T)
     zi[3] <- try(ZINB$coef$count[2],silent = T)
     zi[4] <- try((ZINB$coef$count[2]-forklNB)^2, silent = T)
     zi[5] <- try(ZINB$coef$zero[1], silent = T)
     zi[6] <- try((ZINB$coef$zero[1]-konstBIN)^2, silent = T)
     zi[7] <- try(ZINB$coef$zero[2], silent = T)
     zi[8] <- try((ZINB$coef$zero[2]-forklBIN)^2, silent = T)
     zi[9] <- try(1/ZINB$theta, silent = T)
     zi[10] <- try((1/ZINB$theta-alfa)^2, silent = T)
```

```
    zi[11] <- try(ZANB$coef$count[1], silent = T)
    zi[12] <- try((ZANB$coef$count[1]-konstNB)^2, silent = T)
    zi[13] <- try(ZANB$coef$count[2], silent = T)
    zi[14] <- try((ZANB$coef$count[2]-forklNB)^2, silent = T)
    zi[15] <- try(ZANB$coef$zero[1], silent = T)
    zi[16] <- try((ZANB$coef$zero[1]-konstBIN)^2, silent = T)
    zi[17] <- try(ZANB$coef$zero[2], silent = T)
    zi[18] <- try((ZANB$coef$zero[2]-forklBIN)^2, silent = T)
    zi[19] <- try(1/ZANB$theta,silent = T)
    zi[20] <- try((1/ZANB$theta-alfa)^2, silent = T)
    zi[21] <- length(which(y_zinb==0))

    for(j in 22:123){
     zi[j]=y_zinb[j-23]
    }

    zi
   }


# Ordner de lagrede resultatene i en matrise der de kan leses uavhengig
# av om de har konvergert eller ei.
ZI_PAREST = matrix(nrow=n, ncol=dim(ZINB_est)[2])
for(i in 1:dim(ZINB_est)[2]){
  # I tilfelle ikke konvergert => setter resultat lik NA
  ZINB_est[,i]<-as.numeric(as.character(ZINB_est[,i]))
  # Setter de numeriske verdiene fra konverterte tilfeller inn i matrisen.
  ZI_PAREST[,i]<-as.numeric(as.character(ZINB_est[,i]))
}


#-- Genererer zanb-fordelte datasett og utfører regresjon med modellene
#-- ZINB, ZANB og NB2.

ZANB_est <- foreach(i=1:n, .combine=rbind)%dopar%{
    y_zanb <- rzanegbin(100, munb=mu, pobs0= phi, size=1/alfa)
    yzanbdata <- data.frame(y_zanb, x1)
    yzanbdataframe <- as.data.frame(yzanbdata)

    # Utfører regresjon på ZANB-datasettet med de to modellene
    ZINB <- try(zeroinfl(y_zanb~ x1|x1, data=yzanbdataframe, dist= "negbin",
```

```
            control=zeroinfl.control(BFGS=TRUE)), silent=TRUE)   ## evnt. EM=TRUE
 ZANB <- try(hurdle(y_zanb~ x1|x1, data=yzanbdataframe, dist= "negbin",
            link="logit"), silent=TRUE)


# Henter ut og lagrer estimerte verdier av regresjonskoeffisientene fra
# tilpasset objekt, i tillegg til å lagre kvadrerte feilledd av disse.
# Lagrer også alle data i datasett for hver simulering.
za <- array(NA,127)
 za[1] <- try(ZINB$coef$count[1],silent=T)
 za[2] <- try((ZINB$coef$count[1]-konstNB)^2, silent = T)
 za[3] <- try(ZINB$coef$count[2],silent = T)
 za[4] <- try((ZINB$coef$count[2]-forklNB)^2, silent = T)
 za[5] <- try(ZINB$coef$zero[1], silent = T)
 za[6] <- try((ZINB$coef$zero[1]-konstBIN)^2, silent = T)
 za[7] <- try(ZINB$coef$zero[2], silent = T)
 za[8] <- try((ZINB$coef$zero[2]-forklBIN)^2, silent = T)
 za[9] <- try(1/ZINB$theta, silent = T)
 za[10] <- try((1/ZINB$theta-alfa)^2, silent = T)
 za[11] <- try(ZANB$coef$count[1], silent = T)
 za[12] <- try((ZANB$coef$count[1]-konstNB)^2, silent = T)
 za[13] <- try(ZANB$coef$count[2], silent = T)
 za[14] <- try((ZANB$coef$count[2]-forklNB)^2, silent = T)
 za[15] <- try(ZANB$coef$zero[1], silent = T)
 za[16] <- try((ZANB$coef$zero[1]-konstBIN)^2, silent = T)
 za[17] <- try(ZANB$coef$zero[2], silent = T)
 za[18] <- try((ZANB$coef$zero[2]-forklBIN)^2, silent = T)
 za[19] <- try(1/ZANB$theta,silent = T)
 za[20] <- try((1/ZANB$theta-alfa)^2, silent = T)
 za[21] <- length(which(y_zanb==0))

 for(j in 22:123){
  za[j]=y_zanb[j-23]
 }

 za
}


# Ordner de lagrede resultatene i en matrise der de kan leses uavhengig
# av om de har konvergert eller ei.
```

```
  ZA_PAREST = matrix(nrow=n, ncol=dim(ZANB_est)[2])
  for(j in 1:dim(ZANB_est)[2]){
   # I tilfelle ikke konvergert => setter resultat lik NA
   ZANB_est[,j]<-as.numeric(as.character(ZANB_est[,j]))
   # Setter de numeriske verdiene fra konverterte tilfeller inn i matrisen.
   ZA_PAREST[,j]=as.numeric(as.character(ZANB_est[,j]))
  }


### Estimater av forventede parameterestimater og MSE-verdier til
### parameterestimatene. (samlet)

 # Matrise for endelig resultat
 EST_MSE = matrix(nrow=(dim(ZINB_est)[2]-101), ncol= 2)

 for(k in 1:dim(EST_MSE)[1]){
 EST_MSE[k,1]= mean(ZI_PAREST[,k],na.rm=T)
 EST_MSE[k,2]= mean(ZA_PAREST[,k],na.rm=T)
 }

 # Finner hvilke av simuleringene som ikke har konvergert for hver
 # av modellene brukt på zinb-datasett.
 y_zi_konv_zi <- which(is.na(ZI_PAREST[,1]))
 y_zi_konv_za <- which(is.na(ZI_PAREST[,11]))

 # Finner hvilke av simuleringene som ikke har konvergert for hver
 # av modellene brukt på zanb-datasett.
 y_za_konv_zi <- which(is.na(ZA_PAREST[,1]))
 y_za_konv_za <- which(is.na(ZA_PAREST[,11]))



### Samler verdier og resultat til analyse
ret = list()
 ret$alfa = alfa
 ret$parametre = c(konstNB,forklNB,konstBIN,forklBIN)
 ret$forkl_var = x1
 ret$mu <- mu
 ret$mean_mu <- mean(mu)
 ret$phi <- phi
 ret$mean_phi <- mean(phi)
```

```
  ret$P_null <- P_null
  ret$mean_P_null <- mean(P_null)
  ret$tot_antallsim <- n
  ret$Yzinb_andelkonvZINB = length(which(!is.na(ZI_PAREST[,1])))/n
  ret$Yzinb_andelkonvZANB = length(which(!is.na(ZI_PAREST[,11])))/n
  ret$Yzanb_andelkonvZINB = length(which(!is.na(ZA_PAREST[,1])))/n
  ret$Yzanb_andelkonvZANB = length(which(!is.na(ZA_PAREST[,11])))/n
  ret$EST_MSE <- EST_MSE

 # Lagrer disse verdiene og resultatene for hver parameterkombinasjon
  key <- c(alfa, konstNB, forklNB, konstBIN, forklBIN)
  filnavn <- paste("res", key[1], key[2], key[3], key[4], key[5], ".Rdata", sep="")
  save(ret, file=filnavn)

}
```

Bibliografi

[1] *GAMLSS- Generalized Additive Models for Location, Scale and Shape.* http://www.gamlss.org/, Mai 2013.

[2] *The R project for statistical computing.* http://www.r-project.org, Mai 2013.

[3] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium of Information Theory*, Edited by B.N. Petrov and F. Csaki. Akademiai Kiado, Budapest:267–281, 1973.

[4] A. C. Cameron and P. K. Trivedi. *Regression Analysis of Count Data.* Cambridge University Press, New York, 1998.

[5] J. G. Cragg. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, 39:829–844, 1971.

[6] J.E. Dennis and R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Classics in applied mathematics. Society for Industrial and Applied Mathematics, 1983.

[7] M. Greenwood and G. U. Yule. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, 83(2):255–279, 1920.

[8] D. C. Heilbron. Generalized linear models for altered zero probabilities and overdispersion in count data. *Technical report*, Department of Epidemiology and Biostatistics, University of California, San Francisco, 1989.

[9] J. M. Hilbe. *Negative Binomial Regression.* Cambridge University Press, New York, 2011.

[10] S. Kaarstad. *Statistiske modellar for Poissonregresjon med modifiserte null-sannsyn, ZIP og ZAP.* Masteroppgave, Matematisk institutt, Universitetet i Bergen, Bergen, 2011.

[11] D. Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14, 1992.

[12] R. Li, A.R. Weiskittel, and J.A. Kershaw Jr. Modeling annualized occurrence, frequency, and composition of ingrowth using mixed- effects zero-inflated models and permanent plots in the acadian forest region of north america. *Canadian Journal of Forest Research*, 41:2077–2089, 2011.

[13] J. Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33:341–365, 1986.

[14] G. P. Patil. *Random counts in scientific work, volume 1.* The Pennsylvania State University Press, University Park, Pennsylvania, 1970.

[15] G. Schwarz. Estimating the Dimension of a Model. *Ann. Stat*, 6:461–464, 1978.

[16] "Student". On the error of counting with a haemacytometer. *Biometrika*, 5:255–279, 1907.

[17] I. Todhunter. *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace.* Macmillan and co, 1865.

[18] Q. H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57:307–333, 1989.

[19] A. Zeileis, C. Kleiber, and S. Jackman. Regression models for count data in r. *Journal of Statistical Software*, 27:1–25, 2008.

[20] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith. *Mixed Effects Models and Extensions in Ecology with R, Statistics for Biology and Health.* Springer, 2009.