Statistiske modellar for Poissonregresjon med modifiserte null-sannsyn, ZIP og ZAP

Masteroppgåve i Statistikk

Solveig Kårstad

Matematisk institutt
Universitetet i Bergen

6. mai 2011

Takk

Eg ynskjer først og fremst å takke min kjære ektemann som gjer kvar dag, både dei med passe mykje og dei med litt i overkant mykje studering, til gode og glade dagar. Du gjer kvardagane til dei aller beste dagane. Familien min må også nemnast.

I tillegg vil eg takke gjengen eg har studert i lag med desse åra, som har gjort mastertida kjekk og lærerik, ikkje minst med tanke på alle quizane me har hatt. Gunhild, mitt personlege oppslagsverk og trufaste kontorpartnar, må nemnast spesielt. Det same gjeld Jon Andreas, mentoren min i utdanning og arbeid.

Eg vil til slutt takka rettleiiaren min, som frå første stund har vist stor interesse for temaet i oppgåva. Ein alltid positiv og engasjert rettleiiar har gjort skrivinga av masteroppgåva til ei svært lærerik og positiv oppleving.

Solveig Kårstad,
6. mai 2011

Innhald

Innleiing

I løpet av tida mi som masterstudent vart eg gjennom ei bedrift involvert i ei praktisk regeresjonsanalyse med ulykker på oljeplattformar som responsvariabel. Svært mange av plattformane rapporterte om ingen ulykkestilfeller. Det vart difor fort klart at verken vanleg negativ binomisk eller vanleg Poissonfordeling var egna som regresjonsmodell, grunna at talet på 0-verdiane i responsvariabelen oversteig den mengda desse modellane forutset. Gjennom arbeidet med å finne ei løysing på problemet fekk eg kjenskap til modellane zero-inflated og zero-altered Poissonfordeling, vanlegvis forkorta som ZIP og ZAP. Dette er to regresjonsmodellar som bruker Poisson som grunnleggjande fordeling, men som tillet sannsynet for utfallet 0 å overstige verdien ordinær Poissonfordeling gir denne storleiken. På grunn av problem med innsamla data vart analysen av plattformane avslutta før arbeidet var kome ordentleg i gang. Eg hadde likevel allereie rokke å fatte interesse for ZIP og ZAP, og ynskte å setja meg meir inn i metodegrunnlaget for desse to modifiserte modellane. I lag med rettleiiar valte eg difor ZIP og ZAP som tema for masteroppgåva mi.

Dei to modellane vart første gang presenstert i si opprinnelige form i ein artikkel av John Mullahy i 1986. Modellane har sidan vorte meir og meir anvendt i praktisk regresjonsarbeid. Særlig voks interessa for ZIP etter at Diane Lambert i ein artikkel i 1992 vidareutvikla Mullahy sin ZIP-modellen til ein meir generell og anvendelig versjon. Lambert vert difor omtala om grunnleggjaren til ZIP, medan Mullahy framleis står som mannen bak ZAP-modellen. Det vert difor i masteroppgåva gitt ein noko grundig presentasjon av dei to artiklane.

Det har etterkvart vorte utgitt ein del artiklar som omhandlar ZIP og ZAP, men dei fleste ser på modellane i forhold til Poisson og ikkje i forhold til kvarandre. Dersom ZIP og ZAP er samanlikna skjer dette som oftast ved at det vert utført regresjon med begge modellane på same datasett, og så samanliknar ein i etterkant AIC-verdiane for å finne ut kven av dei to modellane

som forklarar best samanhengen mellom responsvariabel og forklaringsvariablar. Det har ikkje lukkast verken meg eller rettleiiar å finne godt fagstoff som omhandlar om korleis ein på førehand kan vite kven av modellane ein bør velje. Me ynskte difor gjennom arbeidet med oppgåva å finne ut om val av modell faktisk har betydning for påliteligheiten til estimeringa av regresjonskoeffisientane, kva konsekvensen er ved val av gal modell, og ikkje minst korleis me på førehand kan vita kva som er den korrekte modellen å bruke for i ein analysesituasjon.

Oppgåva innheld i hovudsak ein teoretisk del og ein del med praktisk analyse av ZIP og ZAP. I den første delen av oppgåva, kapittel 1-5, får lesaren ei teoretisk innføring i dei to modellane. Dette er naudsynt kunnskap for å kunne forstå analysen av modellane i den andre delen av oppgåva, kapittel 6-9. I kapittel 10 kjem me med forslag til vidare arbeid.

Både zero-inflated og zero-altered modell støttar også bruk av geometrisk og negativ binomisk som hovudfordeling i staden for Poisson. Me har valt å avgrense denne oppgåva til Poisson som grunnleggjande fordeling, grunna at dette er den enklaste versjonen om modellane, og difor greiast å byrje med dersom ein ikkje har særleg kjennskap til modellane på førehand. Sjølv om det kun er ZAP av dei to modellane som kan tilpassast situasjonar med færre 0-observasjonar enn det ordinær Poisson forutset, vel me å samanlikne ZIP og ZAP også for nokre slike praktiske tilfeller. Me ynskjer å sjå kor mykje betre ZAP er enn ZIP i desse situasjonane.

Kapittel 1

Grunnleggjande fordelingar

1.1 Grunnleggjande fordelinger

Lat oss først sjå på nokre diskret fordelingar me treng for å kunne utlede dei samansatte sannsynsmodellane ZIP og ZAP.

Den første er *Poissonfordelinga* som har punktsannsyn

$$f(y;\lambda) = \begin{cases} \frac{e^{-\lambda}\,\lambda^y}{y!}, & y = 0, 1, 2 \ldots \\ 0, & \text{ellers} \end{cases}$$

For Poissonfordelinga er $E(Y) = \text{var}(Y) = \lambda > 0$.

Me kjem også til å trenge den *trunkterte Poissonfordelinga*. Det er ei modifisert Poissonfordeling der det er gitt at verdiane er over 0. Ein kan ekskludere utfallet $y = 0$ frå den vanlige fordelingsfunksjonen ved å dividere funksjonen på $1 - e^{-\lambda}$, som er sannsynet at me får ein verdi over 0. Det gir oss punktsannsynet for den trunkerte Poissonfordelinga.

$$f(y;\lambda \mid y > 0) = \frac{\lambda^y\, e^{-\lambda}}{(1 - e^{-\lambda})\, y!}, \quad y = 1, 2, 3...$$

Den moment-genererande funksjonen til Y vert

$$\begin{aligned} M(t) = E(e^{tY}) &= \sum_{y=1}^{\infty} e^{ty} \, \frac{\lambda^y \, e^{-\lambda}}{y! \, (1 - e^{-\lambda})} \\ &= \frac{e^{-\lambda}}{(1 - e^{-\lambda})} \, \sum_{y=1}^{\infty} \frac{(\lambda e^t)^y}{y!} \\ &= \frac{e^{-\lambda}}{(1 - e^{-\lambda})} \, \left[\sum_{y=0}^{\infty} \frac{(\lambda e^t)^y}{y!} \, - 1 \right] \\ &= \frac{e^{-\lambda}}{(1 - e^{-\lambda})} \, \left[e^{\lambda e^t} - 1 \right] \\ &= \frac{e^{\lambda(e^t - 1)} - e^{-\lambda}}{1 - e^{-\lambda}} \end{aligned}$$

Me finn så den første- og andrederiverte til denne funksjonen, og set variablen t lik 0.

$$M'(t) = \frac{\lambda \, e^t \, e^{\lambda(e^t - 1)}}{1 - e^{-\lambda}}$$

$$M''(t) = \frac{(\lambda e^t)^2 \, e^{\lambda(e^t - 1)} + \lambda \, e^t \, e^{\lambda(e^t - 1)}}{1 - e^{-\lambda}}$$

$$M'(0) = \frac{\lambda \, e^0 \, e^{\lambda(e^0 - 1)}}{1 - e^{-\lambda}} = \frac{\lambda}{1 - e^{-\lambda}}$$

$$M''(0) = \frac{(\lambda e^0)^2 \, e^{\lambda(e^0 - 1)} + \lambda \, e^0 \, e^{\lambda(e^0 - 1)}}{1 - e^{-\lambda}} = \frac{\lambda^2 + \lambda}{1 - e^{-\lambda}}$$

Det gir oss

$$E(Y) = M'(0) = \frac{\lambda}{1 - e^{-\lambda}}$$

$$\text{var}(Y) = M''(0) - (M'(0))^2 = \frac{\lambda^2 + \lambda}{1 - e^{-\lambda}} - \frac{\lambda^2}{(1 - e^{-\lambda})^2}$$

For den trunkerte fordelinga er $E(Y) = \dfrac{\lambda}{1 - e^{-\lambda}} > 0$ og

$$\text{var}(Y) = \frac{\lambda^2 + \lambda}{1 - e^{-\lambda}} - \frac{\lambda^2}{(1 - e^{-\lambda})^2} > 0$$

.

Den siste sannsynsmodellen me treng for å utlede ZIP og ZAP er *bernoullifordelinga*. Den har kun to moglege utfall, og har punktsannsyn

$$f(y,p) = \begin{cases} p, & y = 1 \text{ (suksess)} \\ 1-p, & y = 0 \text{ (fiasko)} \end{cases}$$

Forventningsverdien til ein tilfeldig Bernoulli variabel er $E(Y) = p$, og variansen er $\text{var}(Y) = p(1-p)$. Bernoullifordelinga er identisk med binomisk fordeling med kun eit forsøk, og den vert difor ofte også kalla binomsk fordeling. Dette gjeld særlig fagstoff som omhandlar ZIP og ZAP-modellane, og vidare i denne oppgåva vil bernoullifordelinga difor verta omtalt som binomisk.

Kapittel 2

Zero inflated og zero altered Poissonfordeling

2.1 To modellar med modifisert null-sannsyn

Me kan no utlede dei to fordelingsfunksjonane som ligg til grunn i ZIP- og ZAP-regresjon, nemlig zero-inflated og zero-altered fordeling. Tanken bak både ZIP- og ZAP-modellen er at dei tillet sannsynet for å observere verdien 0 å følgje ei anna fordeling enn den opprinnelige Poissonfordelinga, medan det framleis er Poissonfordeling (trunkert for ZAP) som modellerer sannsynet for å observere utfallsverdiar større enn 0. ZAP-fordelinga støttar situasjonar med både for høgt og for lågt nullsannsyn i forhold til vanleg Poissonfordeling, medan ZIP kun kan tillegge sannsynet for observere ein 0 *meir* verdi enn ei tilsvarande ordinær Poissonfordeling.

Me tenkjer oss at me har ei vanleg teljefordeling som fylgjer eit visst punktsannsyn, med unntak av utfallet 0. Punktsannynet til fordelinga gir ikkje eit reelt bilde av sannsynet for å observere dette utfallet. Lat oss difor definere to ulike funksjonar for Y, $f_1(y, \theta_1)$ og $f_2(y, \theta_2)$, der θ_1 og θ_2 er to parametre, definert på alle reelle tall, som er med på å styre fordelingane. Me vel så bruke $f_1(y, \theta_1)$ for å modellere utfallet 0, medan me bruker $f_2(y, \theta_2)$ for å modellere sannsynet for å observere verdiar over 0. Me tek då hensyn til at 0-utfallet fylgjer eit anna punktsannsyn enn resten av observasjonane. Ved å setje saman dei to funksjonane dannar me eit samla punktsannsynet for heile tellefordelinga

$$f(y; \theta_1, \theta_2) = \begin{cases} f_1(0, \theta_1), & y = 0 \\ f_2(y, \theta_2), & y = 1, 2, 3 \dots \end{cases}$$

Fordelinga lyt framleis fylgje kriteria for sannsynsmodellar, og må kunne oppfylle

$$\sum_{y=0,1,2...} P(Y=y) = f_1(0,\theta_1) + \sum_{y=1,2,3...} f_2(y,\theta_2) = 1 \tag{2.1}$$

I vanleg Poissonfordeling er alle utfalla modellert med same sannsynsmodell. Då er $f_1(y,\theta_1) = f_2(y,\theta_2)$ for alle gyldige av Y, og

$$\sum_{y=0,1,2...} f_1(y,\theta_1) = \sum_{y=0,1,2...} f_2(y,\theta_2) = 1.$$

ZIP og ZAP er to ulike modifikasjoner av den vanlege Poissonfordelinga som tilfredstiller (2.1), men der $f_1(y,\theta_1) \neq f_2(y,\theta_2)$. Det som skil dei to modellane ZIP og ZAP er korleis dei modellerer sannsynet $P(Y=0)$, altså korleis dei definerer $f_1(y,\theta_1)$.

2.1.1 Zero-altered Poissonfordeling

ZAP vart første gang presentert i 1986 i ein artikkel av Mullahy [10]. Fordelinga slik den er kjent idag, skil seg noko frå modellen Mullahy la fram, men er i hovudsak berre gjort meir generell.

Tanken bak ZAP er at det ikkje er kun ein, men to prosessar som påverkar situasjonen datasettet er henta frå. Den eine prosessen har innverknad på om ein observasjon tek verdien 0 eller ikkje, medan den andre prosessen spelar inn på storleiken til dei observasjonane som har verdiar over 0. ZAP er difor bygd opp av to ledd med to ulike sannsynsfordelingar som modellerer ein kvar av dei to prosessane. I det første leddet bruker ein binomisk fordeling til å modellere sannsynet for nærværet av 0 mot fråværet av 0. Eit anna namn på ZAP-modellen er "hurdle-modellen", der hurdle er engelsk og tyder hinder. Namnet kjem av at ein i modellen må forsere hinderet, å observere ein annan verdi enn 0, for å kunne gå vidare til neste ledd. Dette neste leddet tek seg av sannsynsmodelleringa for storleiken til dei observasjonane som kom vidare. Til det bruker ein den opprinnelige fordelinga, Poisson. Men det er no gitt at observasjonane ikkje kan ha verdi 0, sidan dei har klart å forsere hinderet, og ein bruker difor ei *trunkert* Poissonfordeling i ledd to. Ein observasjon over 0 må altså først krysse hinderet og så verta observert som den spesifikke storleiken den er. Det gir oss at

$$\begin{aligned} &P(\text{observere ein spesifikk verdi over 0}) \\ &= P(\text{forsere hinderet}) \times P(\text{trunkert Poisson gir } Y=y) \end{aligned} \tag{2.2}$$

Dersom me definerer det binære utfallet at ein observasjon har verdien 0 som suksess mot alle andre verdiar som fiasko, og modellerer det med binomisk fordeling, får me at

$$P(Y \text{ har verdien } 0) = P_{\text{binomisk}}(\text{suksess}) = p \tag{2.3}$$
$$P(Y \text{ har verdi over } 0) = 1 - P_{\text{binomisk}}(\text{suksess}) = 1 - p \tag{2.4}$$

$1 - p$ er sannsynet for å krysse hinderet i det første leddet i ZAP. Setter me utrykka frå (2.3) og (2.4) inn i (2.2) får me

$$\begin{aligned} &P(\text{observere ein spesifikk verdi over } 0) \\ &= P_{\text{binomisk}}(Y_i \neq 0) \times P_{\text{Trunkert Poisson}}(Y_i = y) \end{aligned}$$

Dette gir oss følgjande fordelingsfunksjon for ZAP

$$f_{\text{ZAP}}(y; p, \lambda) = \begin{cases} p, & y = 0 \\ (1-p)\ \frac{e^{-\lambda}\,\lambda^y}{y!\,(1-e^{-\lambda})}, & y > 0 \end{cases} \tag{2.5}$$

ZAP tillegg sannsynet for å observere ein 0 meir eller mindre verdi enn det den vanlege Poissonfordelinga gjer. Sidan den samla sannsynsmengda må verta ein, vil ei endring av sannsynet for 0 måtte skje på bekostning av sannsynet for utfalla over 0. Dette tek leddet $(1 - p)$ seg av. Ei auke eller minsking av $P(Y = 0) = p$ vil redusere eller auke $(1 - p)$ med tilsvarande verdi. $(1 - p)$ inngår i uttrykket for $P(Y = y)$ for $y > 0$, og ei justering av $(1 - p)$ vil gjere at sannsynsmengda som vert tillagt eit utfall over 0 vil verta endra tilsvarande. Modellen er slik tilpassa situasjonar med fleire eller færre nullar enn det vanleg Poisson kalkulerer med.

Verdien til p styrer sannsynet for å observere 0-ar. I eit tenkt tilfelle der sannsynet for at ein variabel tek verdien 0 ikkje er større eller mindre enn det den vanlege Poissonfordelinga tilseier, vil p vera lik sannsynet ein vanleg Poissonfordeling gir utfallet 0, og $P_{\text{binomisk}}(Y = 0; p)$ vil bli lik $1 - P_{\text{Poisson}}(Y = 0; \lambda)$. Fordelinga for ZAP vert då

$$f_{\text{ZAP}}(y; p, \lambda) = \begin{cases} f_{\text{Poisson}}(y; \lambda), & y = 0 \\ f_{\text{Poisson}}(y; \lambda), & y > 0 \end{cases}$$

altså lik den vanlege Poissonfordelinga.

2.1.2 Zero-inflated Poissonfordeling

ZIP vart først verkeleg kjent i 1992 gjennom artikkelen til Lambert [8], som vert sett på som grunnleggjaren av modellen. Artikkelen førte til stor interesse rundt nullmodifiserte fordelingar, og ZIP har sidan vorte anvendt mykje

i praktisk analysearbeid. Som ZAP tillet også ZIP sannsynet for 0 å verta modellert med ei anna fordeling enn den brukt på dei andre moglege utfalla. ZIP har likevel ein noko annan innfallsvinkel. Medan ZAP behandlar alle 0-observasjonane som *ei* gruppe, deler ZIP desse observasjonane i to grupper. Den første gruppa innheld dei nullane som den vanlege Poissonfordelinga klarar å ta seg av. På engelsk vert dei kalla "count zeroes". Eg vil i oppgåva kalla dei vanlege 0-ar. Den andre gruppa inneheld dei ekstra 0-ane, dei som gjer at mengda med 0-ar overstig antallet den normale Poissonfordelinga kalkulerer med. Desse vert kalla strukturelle 0-ar. På same måte som for ZAP er det to prosessar som verkar inn på observasjonane. Men for ZIP har den eine prosessen innverknad på talet strukturelle nullar, medan den andre påverkar storleiken til resten av observasjonane. Det er vanleg å tenkje at dei to gruppene med nullar kjem frå to ulike kjelder, og at desse kjeldene vert påverka av kvar sin av dei prosessane i ZIP.

Lat oss definere sannsynet for å observere ein strukturell 0 som $\phi \in (0, 1)$. For å modellere det usynlege utfallet av strukturelle nullar mot resten av observasjonane bruker ein binomisk fordeling. Medan den binomiske fordelinga i ZAP har utfallsrom nærværet av 0 mot fråværet av 0, skil den binomiske fordelinga i ZIP kun dei strukturelle 0-ane frå alle dei andre observasjonane. Sannsynet for å observere ein annan verdi enn ein strukturell 0 vert då $(1-\phi)$.

Å observere ein verdi som ikkje er ein strukturell 0 er snitthendinga at verdien ikkje er ein strukturell 0 og at den får den spesifikke verdien me ynskjer. Det kan me skrive som

$$\begin{aligned} &P(A \cap B) = P(A) \times P(B \mid A) \\ &= P(Y \text{ er ikkje strukturell 0}) \times P(Y \text{ er verdien } y \mid Y \text{ er ikkje strukturell 0}) \end{aligned}$$

Det er kun dei strukturelle 0-ane som skil observasjonane frå å innfri forutsetningane til den vanlege Poissonfordelinga. Dei observasjonane vert luka bort i $P(A)$, og me kan difor anvende vanleg Poisson for $P(B \mid A)$. (Merk at ZAP her bruker trunkert fordeling). Sannsynet for at observasjonen har ein spesifikk verdi over 0 vert då

$$\begin{aligned} &P(Y = y) \\ &= \big(1 - P(Y \text{ er strukturell 0})\big) \times P(Y \text{ er verdien } y \text{ ved vanleg Poisson}) \\ &= (1 - \phi) \times P_{\text{Poisson}}(Y = y), \text{ for } y > 0) \end{aligned}$$

Sannsynet for å observere ein 0-verdi vert då summen av sannsyna for dei disjunkte utfalla at ein observerer ein strukturell 0 og at ein observerer ein

av dei vanlege 0-ane.

$$\begin{aligned} &P(Y=0) \\ &= P(Y \text{ er strukturell } 0) \\ &+P(Y \text{ er ikkje strukturell } 0) \times P(Y \text{ er } y \text{ ved vanleg Poisson}) \\ &= \phi + (1-\phi)P_{\text{Poisson}}(Y=0) \end{aligned}$$

Fordelingsfunksjonen til ZIP vert då

$$f_{\text{ZIP}}(y;\phi,\beta) = \begin{cases} \phi + (1-\phi)\, e^{-\lambda}, & y = 0 \\ (1-\phi)\, \frac{e^{-\lambda}\,\lambda^y}{y!}, & y > 0 \end{cases} \tag{2.6}$$

Samanlikna med vanleg Poisson er sannsynet for å observere ein 0 auka med ϕ. Ein må difor setje inn leddet $(1-\phi)$ i funksjonen for å redusere sannsynet for utfalla over 0 tilsvarande. Verdien til ϕ varierer etter mengda strukturelle 0-ar. Dersom $\phi = 0$ forsvinn leddet for desse "ekstra" 0-ane. Alle 0-observasjonane kjem då frå same kjelde, den vanlege Poissonfordelinga, og $1-\phi$, som avgjer nedjusteringa for dei andre observasjonane vert lik 1. Me står då igjen med kun vanleg Poissonfordeling. Sidan ϕ ikkje er gyldig for negative tal, kan ikkje ZIP tilpassa seg situasjonar med færre 0-observasjonar enn det ordinær Poisson føreset.

ZIP vert ofte omtala som ein "mixture model", sjølv om fordelinga skil seg noko frå den tradisjonelle bruken av denne definisjonen. Med "mixture model" tenkjer ein oftast på ein modell som er samansatt at to funksjonar frå same fordeling, men med ulike parameterverdiar. ZIP er derimot ein modell som kombinerer to heilt ulike fordelingar. Den eine er den vanlege telleprosessen i Poisson. Den andre er ei fordeling for ei klynge observasjonar med verdi 0. Den har kun eitt utfall og $P(y=0) = 1$. Ein brukar så binomisk fordelinga for å vekte dei to fordelingane når ein set dei saman til ZIP.

Det som skil ZIP og ZAP er hovudsakleg korleis dei angrip situasjonen med dei mange 0-observasjonane. I ZAP kjem 0-observasjonane frå *ei* kjelde, medan i ZIP kjem dei frå to kjelder.

Kapittel 3

Metodegrunnlag

3.1 Metodegrunnlag

Lat oss no sjå på ZIP og ZAP i regresjonssamanheng og sjå på metodegrunnlaget og likelihoodfunksjonane til desse modellane. Det neste naturlege steget vert då å innføre forklaringsvariablar og regresjonsparametrar i uttrykka for forventningane i fordelingane. For dette er artiklane til Mullahy [10] og Lambert [8] til god hjelp.

3.1.1 Zero-altered regresjonsmodell

Det er ikkje alltid dei same ytre faktorane som avgjer om ein observasjon får verdien 0 eller ikkje, som verkar inn på storleiken til observasjonane over 0. ZAP tillet oss å innføre forklaringsvariablar for begge desse prosessane. Det gjer ein ved å innføre regresjonsparametrar og kovariatar i uttrykka for begge forventningane. Både parametrane og forklaringsvariablane kan vera ulike for dei to prosessane, men ein kan også velje å setje dei like.

For den trunkerte Poissonregresjonen i ZAP er det mest vanleg å bruke dei same systematiske komponenta og den same linkfunksjonen som ved vanleg Poissonregresjon. Observasjonen y_i, av i alt n observasjonar, vert då modellert med linkfunksjonen $\eta = \log(\lambda_i)$. For y_i gir dette $\lambda_i = e^{\alpha+\beta_1 X_{i1}+...+\beta_q X_{iq}}$, der $\boldsymbol{X}_i = [X_{i1}, X_{i2}, ..., X_{iq}]$ er ein vektor med forklaringsvariablane tilhøyrande observasjon y_i, og $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta q]^T$ er ein vektor med regresjonsparametrane. Forventninga til trunkert Poisson er ulik forvenntinga til vanleg Poisson. (Sjå avsnittet om trunkert Poisson i 1.1). Me finn λ både i teljar og nemnar, og $\log(\lambda)$ som linkfunksjon vil ikkje skape lineær regresjon slik som for Poisson. ZAP-regresjon med dette valet av linkfunksjon er difor ikkje med

i klassen generaliserte lineære modellar.

Lat oss no sjå på sannsynet for nærværet av 0 mot fråværet av 0. Dette vert modellert med binomisk fordeling, som gir at $P(Y_i = 0) = p$ og $E(\mathbf{Y}) = \lambda$. (Sjå avsnittet om binomisk fordeling i 1.1). Når ein innfører regresjonsparametrar for p er det vanleg å bruke $\eta = \text{logit}(p) = \log(p/(1-p))$, som er den mest brukte linkfunksjonen for binomisk regresjon. Den enklaste versjonen av denne er å bruke kun eit interceptledd. Det gir $p = (e^{\nu})(1 + e^{\nu})$, der ν er definert som interceptparameter. Men i likhet med λ_i kan også p_i ta ulike forklaringsvariablar. Dette er faktorar som verkar inn på sannsynet for om ein observasjon er ein 0 eller ikkje. Dersom ein inkluderer desse faktorane som regresjonsparametre, vert modellen på forma

$$p_i = \frac{e^{\nu+\gamma_1 Z_{i1}+...+\gamma_q Z_{iq}}}{1 + e^{\nu+\gamma_1 Z_{i1}+...+\gamma_q Z_{iq}}}$$

Me bruker ulik notasjon for å skilje mellom kovariatmatrisene for dei to prosessane sidan desse kan vera ulike.

3.1.2 Zero-inflated regresjonsmodell

Også i ZIP har me to forventningar, ein frå binomisk fordeling og ein frå poissonfordelinga, og også her kan begge uttrykka ta forklaringsvariablar. For ZIP har det eine settet kovariater innverknad på om ein observasjon er ein struktuell 0, medan det andre settet avgjer storleiken til alle dei andre observasjonane. Det treng altså ikkje vera dei same ytre faktorane som påverkar dei to gruppene med 0-observasjonar. Den binomiske fordelinga gir oss sannsynet $P(\text{strukturell } 0) = \phi$ og forventning $E(\mathbf{Y}) = \phi$. Ein kan bruke dei same linkfunksjonane som for ZAP, og det er også for ZIP mest vanleg å bruke $\eta = \text{logit}(\phi_i)$ for den binomiske komponenten. Det gir $\phi_i = (e^{\iota+\omega_1 G_{i1}+...+\omega_q G_{iq}})/(1 + e^{\iota+\omega_1 G_{i1}+...+\omega_q G_{iq}})$, der $\boldsymbol{G}_i = [G_{i1}, G_{i2}, ..., G_{iq}]$ er vektoren med forklaringsvariablane , og $\boldsymbol{\omega} = [\omega_1, \omega_2, ..., \omega q]^T$ er vektoren med regresjonsparametrane. For å modellere dei andre observasjonane bruker ein Poissonfordeling, og som for ZAP er det er difor vanleg å bruke $\eta = \log(\lambda)$. For observasjon y_i gir dette $\lambda_i = e^{\delta+\rho_1 H_{i1}+...+\rho_q H_{iq}}$. Merk at her følgjer observasjonane vanleg Poisson og ikkje trunkert fordeling, sidan ein i binomisk regresjonen kun har teke vekk dei strukturelle nullane. Poissonregresjon i ZIP er difor lineær. Det kan vera lett å tenkja at ein treng å fastsette kven av dei to kjeldene 0-observasjonane kjem frå for å kunne bestemme kva kovariatar som er signifikante for å observere 0-ar. Men ved utledinga av likelihoodfunksjonen vil ein sjå at dette faktisk ikkje er naudsynt. Sjå uttrykk (3.6).

Dei ulike likelihoodfunksjonane til ZIP og ZAP vil føre til ulike parameterestimat, og ZIP og ZAP er to ulike regresjonsmodellar. Både kovariatmatrisa og koeffisientvektoren kan, og vil som regel, variere frå ZIP og ZAP, og det er difor brukt ulik notasjon for alle desse i dei to modellane. Eg har i resten av oppgåva valt å føre vidare tradisjonen med å bruke linkfunksjonen log for λ og logit for ϕ og p. Avvik frå dette er det kommentert i avsnitta det gjeld.

3.1.3 Forholdet mellom forventningane til dei to prosessane

Det er ikkje alltid det er dei same ytre faktorane som påverkar utfallet i den binære situasjonen og utfallet i Poissondelen av modellane. Kovariatane for dei tilhøyrande forventningane treng som nevnt over, difor heller ikkje å vera like. I fleire tilfeller er det likevel mest naturleg å anta at det er dei same ytre faktorane som verkar inn på begge situasjonane, og for ZAP vert då kovariatmatrisa $\boldsymbol{X}$ lik kovariatmatrisa $\boldsymbol{Z}$ (eller $\boldsymbol{G}$ lik $\boldsymbol{H}$ for ZIP). I desse tilfella krev regresjonsmodellen dobbelt så mange paramtre som ved vanleg Poissonregresjon, noko som gir mykje arbeid ved estimeringa av regresjonsparametrane. Nokre gonger er ein så heldig at ein kan anta at ikkje berre er $\boldsymbol{X}$ lik $\boldsymbol{Z}$ (eller $\boldsymbol{G}$ lik $\boldsymbol{H}$), men også at $\boldsymbol{p}$ (eller $\boldsymbol{\phi}$) relaterer til $\boldsymbol{\lambda}$ i eit fast mønster. Det betyr at $\boldsymbol{\beta}$ og $\boldsymbol{\gamma}$ (eller $\boldsymbol{\rho}$ og $\boldsymbol{\omega}$), dei to vektorane med regresjonskoeffisientar frå dei to komponentane i modellane, er avhengige. Det kan då vera til god hjelp å definere $\mathbf{p}$ (eller $\boldsymbol{\phi}$) som ein funksjon av $\boldsymbol{\lambda}$. Dersom ein ikkje veit noko spesifikt om korleis dei to forventningane avheng av kvarandre, er det vorte vanleg å bruke den enkle samanhengen $p_i = 1/(1+\lambda_i(\tau))$, der ein innfører τ definert som ein ukjent form-parameter med reell verdi. Då har me at

$$
\begin{aligned}
p_i = \frac{e^{\boldsymbol{\gamma Z}}}{1+e^{\boldsymbol{\gamma Z}}} = \frac{e^{\boldsymbol{\gamma X}}}{1+e^{\boldsymbol{\gamma X}}} &= \frac{1}{1+\lambda_i^{\tau}} \\
\frac{1+e^{\boldsymbol{\gamma X}}}{e^{\boldsymbol{\gamma X}}} &= 1+\lambda_i^{\tau} \\
\log(\frac{1}{e^{\boldsymbol{\gamma X}}}) &= \log(\lambda_i^{\tau}) \\
-\log(e^{\boldsymbol{\gamma X}}) &= \tau\log(\lambda_i) \\
-\boldsymbol{\gamma X} &= \tau\boldsymbol{\beta X} \\
\boldsymbol{\gamma} &= -\tau\boldsymbol{X}
\end{aligned}
$$

sidan $\log(\lambda_i) = \boldsymbol{\beta X}$. Det er vanleg å bruke det same forholdet for ZIP-modellen, $\phi_i = 1/(1+\lambda_i(\tau))$, og me får då desse linkfunksjonane for ZIP og

ZAP

$$\log(\boldsymbol{\lambda}) = \boldsymbol{\beta X} \text{ og } \text{logit}(\boldsymbol{p}) = -\tau\boldsymbol{\beta X} \text{ for ZAP} \tag{3.1}$$

$$\log(\boldsymbol{\lambda}) = \boldsymbol{\beta X} \text{ og } \text{logit}(\boldsymbol{\phi}) = -\tau\boldsymbol{\beta X} \text{ for ZIP.} \tag{3.2}$$

Parameteren τ er lik for alle observasjonane og styrer samanhengen mellom λ_i og p_i (eller λ_i og ϕ_i). Ein slepp då unna med kun eitt sett med regresjonskoeffisientar. Dette vil redusere talet på ukjente parametrar krafig, og akselerere estimeringa av dei betydelig. Desse modellane vert ofte kalla ZIP(τ) og ZAP(τ). Det kan også brukast andre funksjonar for samanhengen enn den i (3.1) og (3.2). Lambert gir i artikkelen [8] gir gode eksempler på dette for ZIP(τ), og eksempla ein finn der kan også brukast for ZAP(τ).

Det er likevel sjeldan me på førehand kan vita noko sikkert om $\boldsymbol{\beta}$ er avhengig av $\boldsymbol{\gamma}$, sjølv i dei tilfella der ein antek at kovariatmatrisene er like. Endringar i kovariatane frå observasjon til obervasjon vil ikkje nødvendigvis gi eit fast mønster for forholdet mellom dei to prosessane som påverkar responsvariabelen. Det mest utbredte i praksis er difor å halde seg til vanleg ZIP og ZAP, og denne tradisjonen vil bli følgt i resten av oppgåva, bortsett frå der det er noko anna som er spesifisert.

3.1.4 Likelihoodfunksjonane til ZIP og ZAP

For å kunne utføre regresjonsanalysen treng ein sannsynsmaksimeringsestimata (SME) for dei ukjente regresjonskoeffisientane, og desse kan ein finne ved bruk av vanleg regresjonsprosedyre med utgangspunkt i likelihoodfunksjonen for modellane. Me skal difor no utlede denne funksjonen for både ZIP og ZAP. Me tek då igjen utgangspunkt i artiklane til Mullahy [10] og Lambert [8]. Begge artiklane utleder i noko grad likelihoodfunksjonen til den aktuelle modellen, og me arbeider vidare utfrå modellgrunnlaget presentert der.

Sidan likelihoodfunksjonen er sannsynsfordelingane for dei n y_i-observasjonane i eit datasett, vil den vera eit produkt bestående av så mange ledd med $P(Y_i = 0)$ som det er 0-observasjonar, og så mange ledd med $P(Y_i = y, \text{ for } y > 0)$ som det er observasjoner med verdi over 0. Ved regresjonsanalyse har me eit gitt datasett, og utfallet til y_i er difor kjent og konstant. Den generelle Likelihoodfunksjonen for både ZIP og ZAP blir då på forma

$$L = \prod_{i;\, y_i=0} f(y_i) \prod_{i;\, y_i\neq 0} f(y_i) \tag{3.3}$$

Lat oss no utlede dei konkrete likelihoodfunksjonane. Me byrjar med ZAP. Me tek utgangspunkt i $f_{\text{ZAP}}(y)$, sjå (2.5). Setter me denne inn i (3.3) får me

dette utrykket for L_{ZAP}

$$L(p_i, \lambda_i; y_i) = \prod_{i;\, y_i=0} p_i \prod_{i;\, y_i>0} (1-p_i)\frac{\lambda_i^{y_i}\, e^{-\lambda_i}}{y_i!\,(1-e^{-\lambda_i})}$$

Lat oss rekne vidare på uttrykket for å forenkle det og setje det på ei form som let oss enkelt innføre regresjonsparametrane i uttrykka for p_i og λ_i.

$$\begin{aligned} L_{\text{ZAP}}(p_i, \lambda_i; y_i) &= \prod_{i;\, y_i=0} p_i \prod_{i;\, y_i>0} (1-p_i)\frac{\lambda_i^{y_i}\, e^{-\lambda_i}}{y_i!\,(1-e^{-\lambda_i})} \\ &= \prod_{i;\, y_i=0} \frac{p_i}{1-p_i}(1-p_i) \prod_{i;\, y_i>0} (1-p_i)\frac{\lambda_i^{y_i}\, e^{-\lambda_i}}{y_i!\,(1-e^{-\lambda_i})} \\ &= \prod_{i;\, y_i=0,1,2,\ldots} (1-p_i) \prod_{i;\, y_i=0} \frac{p_i}{1-p_i} \prod_{i;\, y_i>0} \frac{\lambda_i^{y_i}\, e^{-\lambda_i}}{y_i!\,(1-e^{-\lambda_i})} \end{aligned}$$

Me kan no innføre regresjonsparametrar og bruker at $\dfrac{p_i}{1-p_i} = e^{\boldsymbol{\gamma}\,\boldsymbol{Z}_i}$ og $\lambda_i = e^{\boldsymbol{\beta}\,\boldsymbol{X}_i}$. I tillegg treng me at

$$\begin{aligned} \log\left(\frac{p_i}{1-p_i}\right) &= \boldsymbol{\gamma}\boldsymbol{Z}_i \\ \frac{p_i}{1-p_i} &= e^{\boldsymbol{\gamma}\boldsymbol{Z}_i} \\ p_i &= e^{\boldsymbol{\gamma}\boldsymbol{Z}_i} - p_i(e^{\boldsymbol{\gamma}\boldsymbol{Z}_i}) \\ p_i(1+e^{\boldsymbol{\gamma}\boldsymbol{Z}_i}) &= e^{\boldsymbol{\gamma}\boldsymbol{Z}_i} \\ 1-p_i &= 1 - \frac{e^{\boldsymbol{\gamma}\boldsymbol{Z}_i}}{1+e^{\boldsymbol{\gamma}\boldsymbol{Z}_i}} \\ 1-p_i &= \frac{1}{1+e^{\boldsymbol{\gamma}\boldsymbol{Z}_i}} \end{aligned} \tag{3.4}$$

Lat oss no gå tilbake til likelihoodfunksjonen og innføre regresjonsparametrane.

$$L_{\text{ZAP}}(\boldsymbol{\gamma}, \boldsymbol{\beta}; y_i) = \prod_{i;\, y_i=0,1,2,\ldots} \frac{1}{1+e^{\boldsymbol{\gamma Z}_i}} \prod_{i;\, y_i=0} e^{\boldsymbol{\gamma Z}_i} \prod_{i;\, y_i>0} \frac{\left(e^{\boldsymbol{\beta X}_i}\right)^{y_i} e^{-e^{\boldsymbol{\beta X_i}}}}{y_i!(1-e^{-e^{\boldsymbol{\beta X}_i}})}$$

$$\begin{aligned}
\log L_{\text{ZAP}}(\boldsymbol{\gamma}, \boldsymbol{\beta}; y_i) &= \log\left[\prod_{i;\, y_i=0,1,2,\ldots} \frac{1}{(1+e^{\boldsymbol{\gamma Z}_i})} \prod_{i;\, y_i=0} e^{\boldsymbol{\gamma Z}_i} \prod_{i;\, y_i>0} \frac{(e^{\boldsymbol{\beta X_i}})^{y_i} e^{-e^{\boldsymbol{\beta X}_i}}}{y_i!(1-e^{-e^{\boldsymbol{\beta X}_i}})}\right] \\
&= \sum_{\text{for alle } y_i} \log \frac{1}{(1+e^{\boldsymbol{\gamma Z}_i})} + \sum_{\text{for } y_i=0} \log\left(e^{\boldsymbol{\gamma Z}_i}\right) + \sum_{\text{for } y_i>0} \log \frac{\left(e^{\boldsymbol{\beta X}_i}\right)^{y_i} e^{-e^{\boldsymbol{\beta X}_i}}}{y_i!(1-e^{-e^{\boldsymbol{\beta X}_i}})} \\
&= \sum_{\text{for alle } y_i} \left[\log(1) - \log\left(1+e^{\boldsymbol{\gamma Z}_i}\right)\right] + \sum_{\text{for } y_i=0} \boldsymbol{\gamma Z}_i \\
&\quad + \sum_{\text{for } y_i>0} \left[y_i\ \log\left(e^{\boldsymbol{\beta X}_i}\right) + \log\left(e^{-e^{\boldsymbol{\beta X}_i}}\right) - \log(y_i!) - \log\left(1-e^{-e^{\boldsymbol{\beta X}_i}}\right)\right] \\
\log L_{\text{ZAP}}(\boldsymbol{\gamma}, \boldsymbol{\beta}; y_i) &= \sum_{\text{for alle } y_i} -\log\left(1+e^{\boldsymbol{\gamma Z}_i}\right) + \sum_{\text{for } y_i=0} \boldsymbol{\gamma Z}_i \\
&\quad + \sum_{\text{for } y_i>0} \left[y_i\ \boldsymbol{\beta X_i} - e^{\boldsymbol{\beta X}_i} - \log(y_i!) - \log\left(1-e^{-e^{\boldsymbol{\beta X}_i}}\right)\right]
\end{aligned}$$

Dette kan skrivast på forma

$$\log L_{\text{ZAP}}(\boldsymbol{\gamma}, \boldsymbol{\beta}; y_i) = \log L_1(\boldsymbol{\gamma}; y_i) + \log L_2(\boldsymbol{\beta}; y_i) \tag{3.5}$$

der L_1 kan sjåast på som likelihoodfunksjonen for det binære utfallet (0 mot høgare verdiar) og L_2 som likelihooden for den trunkerte Poissonmodellen. Då kan ein finne SME til $\boldsymbol{\gamma}$ og $\boldsymbol{\beta}$ ved separate maksimeringar av henholdsvis $\log(L_1)$ og $\log(L_2)$. På grunn av at likelihoodfunksjonnen kan skrivast som (3.5) kan ZAP regresjon utførast i to separate steg. Først ein logistisk regresjon med alle observasjonane og så ein trunkert Poissonregresjon med observasjonane $y_i > 0$. Sjå delkapittel 3.1.1. Dette vil gi same resultat som om ein gjer heile regresjonsanalysen i *ein* operasjon.

Lat oss no finne likelihoodfunksjonen til ZIP. Me tek utgangspunkt i $f_{\text{ZIP}}(y)$, sjå (2.6). Setter me denne inn i (3.3) får me dette utrykket for L_{ZIP}

$$L_{\text{ZIP}}(\phi_i, \lambda_i; y_i) = \prod_{i;\, y_i=0} \left(\phi_i + (1-\phi_i)e^{-\lambda_i}\right) \prod_{i;\, y_i>0} (1-\phi_i) \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$$

Me reknar så vidare på uttrykket for lettare å kunne setje inn regresjons-

parametrar i funksjonen.

$$
\begin{aligned}
L_{\text{ZIP}}(\phi_i, \lambda_i; y_i) &= \prod_{i;\, y_i=0} (1-\phi_i)\left(\frac{\phi_i}{1-\phi_i} + e^{-\lambda_i}\right) \prod_{i;\, y_i>0} (1-\phi_i) \frac{\lambda_i^{y_i}\, e^{-\lambda_i}}{y_i!} \\
&= \prod_{i;\, y_i=0,1,2,..} (1-\phi_i) \prod_{i;\, y_i=0} \left(\frac{\phi_i}{1-\phi_i} + e^{-\lambda_i}\right) \prod_{i;\, y_i>0} \frac{\lambda_i^{y_i}\, e^{-\lambda_i}}{y_i!}
\end{aligned}
$$

No innfører me regresjonsparametrar i uttrykka for ϕ_i og λ_i. Me har frå 3.1.2 at $(\phi_i)(1-\phi_i) = e^{\boldsymbol{\omega}\boldsymbol{G}_i}$ og $\lambda_i = e^{\boldsymbol{\rho}\boldsymbol{H}_i}$. Me får også frå mellomrekninga (3.4) at $1-\phi_i = 1/(1+e^{\boldsymbol{\omega}\boldsymbol{G}_i})$.

$$
L_{\text{ZIP}}(\boldsymbol{\omega}, \boldsymbol{\rho}; y_i) = \prod_{i;\, y_i=0,1,2,...} \frac{1}{1+e^{\boldsymbol{\omega}\boldsymbol{G}_i}} \prod_{i;\, y_i=0} \left[e^{\boldsymbol{\omega}\boldsymbol{G}_i} + e^{-e^{\boldsymbol{\rho}\boldsymbol{H}_i}}\right] \prod_{i;\, y_i>0} \frac{\left(e^{\boldsymbol{\rho}\boldsymbol{H}_i}\right)^{y_i} e^{-e^{\boldsymbol{\rho}\boldsymbol{H}_i}}}{y_i!}
$$

$$
\begin{aligned}
\log L_{\text{ZIP}}(\boldsymbol{\omega}, \boldsymbol{\rho}; y_i) &= \sum_{\text{alle } y_i} \log\left(\frac{1}{1+e^{\boldsymbol{\omega}\boldsymbol{G}_i}}\right) + \sum_{\text{for } y_i=0} \log\left(e^{\boldsymbol{\omega}\boldsymbol{G}_i} + e^{-e^{\boldsymbol{\rho}\boldsymbol{H}_i}}\right) \\
&\quad + \sum_{y_i>0} \log \frac{\left(e^{\boldsymbol{\rho}\boldsymbol{H}_i}\right)^{y_i} e^{-e^{\boldsymbol{\rho}\boldsymbol{H}_i}}}{y_i!} \\
&= \sum_{\text{alle } y_i} \left[\log(1) - \log\left(1+e^{\boldsymbol{\omega}\boldsymbol{G}_i}\right)\right] + \sum_{\text{for } y_i=0} \log\left(e^{\boldsymbol{\omega}\boldsymbol{G}_i} + e^{-e^{\boldsymbol{\rho}\boldsymbol{H}_i}}\right) \\
&\quad + \sum_{\text{for } y_i>0} \left[y_i \log\left(e^{\boldsymbol{\rho}\boldsymbol{H}_i}\right) + \log\left(e^{-e^{\boldsymbol{\rho}\boldsymbol{H}_i}}\right) - \log(y_i!)\right] \\
&= \sum_{\text{alle } y_i} -\log\left(1+e^{\boldsymbol{\omega}\boldsymbol{G}_i}\right) + \sum_{\text{for } y_i=0} \log\left(e^{\boldsymbol{\omega}\boldsymbol{G}_i} + e^{-e^{\boldsymbol{\rho}\boldsymbol{H}_i}}\right) \\
&\quad + \sum_{\text{for } y_i>0} \left[y_i \boldsymbol{\rho}\boldsymbol{H}_i + \log\left(e^{-e^{\boldsymbol{\rho}\boldsymbol{H}_i}}\right) - \log(y_i!)\right] \\
\log L_{\text{ZIP}}(\boldsymbol{\omega}, \boldsymbol{\rho}; y_i) &= \sum_{\text{alle } y_i} -\log\left(1+e^{\boldsymbol{\omega}\boldsymbol{G}_i}\right) + \sum_{\text{for } y_i=0} \log\left(e^{\boldsymbol{\omega}\boldsymbol{G}_i} + e^{-e^{\boldsymbol{\rho}\boldsymbol{H}_i}}\right) \\
&\quad + \sum_{\text{for } y_i>0} \left[y_i \boldsymbol{\rho}\boldsymbol{H}_i - e^{\boldsymbol{\rho}\boldsymbol{H}_i} - \log(y_i!)\right] \qquad (3.6)
\end{aligned}
$$

Regresjonskoeffisientane er dei einaste ukjente storleikane i uttrykket for likelihoodfunksjonen, og den kan difor maksimerast med hensyn på dei ukjente koeffisientane. Me treng altså ikkje å skilje mellom 0-observasjonane frå dei to kjeldene for å kunne finne SME for regresjonsparametrane og finne kva kovariater som er signifikante.

På grunn av at leddet for $y_i = 0$ innheld både $\boldsymbol{\omega}$ og $\boldsymbol{\rho}$ kan ikkje $\log L_{\text{ZIP}}$ skrivast på forma (3.5). Det gjer at ein må finne SME for ρ og ω i èin operasjon. Regresjonsanalysen kan difor ikkje utførast i to separate steg, slik det er mogleg for ZAP.

Likelihoodfunksjonane for ZIP(τ) og ZAP(τ) på forma (3.1) og (3.2), vil vera lik likelihoodfunksjonane til henholdsvis ZIP og ZAP med untak av at $\boldsymbol{\gamma Z}_i$ er bytta ut med $-\tau\boldsymbol{\beta X}_i$ og $\boldsymbol{\omega G}_i$ er bytta ut med $-\tau\boldsymbol{\rho H}_i$ pga. forholda i (3.1) og (3.2). Det gir

$$\log L_{\text{ZIP}}(\boldsymbol{\rho}, \tau; y_i) = \sum_{\text{for alle } y_i} -\log\left(1 + e^{-\tau\boldsymbol{\rho H}_i}\right) + \sum_{\text{for } y_i=0} \log\left(e^{-\tau\boldsymbol{\rho H}_i} + e^{-e^{\boldsymbol{\rho H}_i}}\right) + \sum_{\text{for } y_i>0} \left(y_i\, \boldsymbol{\rho H}_i - e^{\boldsymbol{\rho H}_i} - \log\left(y_i!\right)\right)$$

for ZIP(τ) og

$$\log L_{\text{ZAP}}(\boldsymbol{\beta}, \tau; y_i) = \sum_{\text{for alle } y_i} -\log\left(1 + e^{-\tau\boldsymbol{\beta X}_i}\right) + \sum_{\text{for } y_i=0} -\tau\boldsymbol{\beta X}_i + \sum_{\text{for } y_i>0} \left(y_i\, \boldsymbol{\beta X_i} - e^{\boldsymbol{\beta X}_i} - \log\left(y_i!\right) - \log\left(1 - e^{-e^{\boldsymbol{\beta X}_i}}\right)\right)$$

for ZAP(τ).

3.1.5 Algoritmar for å maksimere likelihoodfunksjonane

Vidare i analysen treng me SME for regresjonsparametrane i modellane. Dei kan ein finne ved bruk av vanleg prosedyre for maksimering av likelihoodfunksjonane med hensyn til parametrane. Det er fleire algoritmar tilgjengelig for dette. Dei mest brukte er Newton-Raphson-, quasi-Newton- og EM-algoritmen. Kven av dei som er den mest funksjonible varierer frå situasjon til situasjon. Newton-Raphson- og quasi-Newton-algoritmen er som regel raskare enn EM-algoritmen. Dei vert ofte brukt i ZAP-regresjon, men vil i fleire tilfeller for ZIP ikkje konvergere. Då er EM algoritmen eit godt verktøy. Den kan ofte også vera enklare å programmere enn Newton-algoritmane. For å finne SME for regresjonsparametrane i ZIP(τ) og ZAP(τ) vil ein ikkje kunne bruke EM algoritmen, grunna at ein ikkje enkelt kan finne estimat for $\boldsymbol{\rho}$ og τ sjølv om ein innfører ein uobservert variabel. Newton-Raphson metoden er derimot eit bra valg, og vil alltid konvergere for ZIP(τ) og ZAP(τ).

Det vil i oppgåva verta nytta quasi-Newton og EM algoritme, og eg vil no gi ei kort innføring i desse.

Newton-algoritmane

Newton-Raphson algoritmen har fått namn etter Isaac Newton og Joseph Raphson som kvar for seg har vore med på å utvikle metoden, [12] og [11]. Metoden er ein iterasjonsprosess som primært vert brukt for å finne røtene til ein funksjon. Den kan også brukast for å finne ekstrempunkta til ein funksjon ved at ein finn kva verdi av x som gir $f'(x) = 0$. Algoritmen krev då utrekning av både den første og andre deriverte til funksjonen for kvart iterasjonsledd. Quasi-Newton algoritmen er ein annan metode for å finne ekstrempunkt. Den bygg på Newton-Raphson, men bruker ikkje den andrederiverte til funksjonen. Ein slepp difor unna mykje kalkulering. Metoden vart først formulert av fysikaren W.C. Davidon i 1959 [3], men har vorte vidareutvikla av han sjølv og andre andre i ettertid.
Med Newton-Raphson algoritmen nyttar ein tangentane til funksjonen. Ein finn først eit punkt for $f(x)$ som ein trur ligg nær $f(x) = 0$. Dette punktet vert startverdiane i det første iterasjonsleddet, og ein merker det x_0. Tangenten til $f(x)$ i punktet har likning

$$y = f(x_0) + f'(x_0)(x - x_0) \tag{3.7}$$

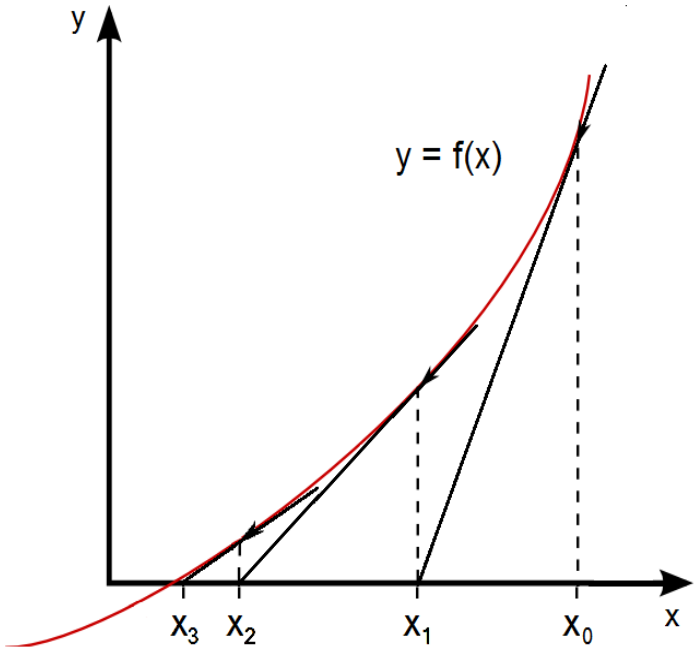
Tangenten vil krysse x-aksen i punktet $(x_1, 0)$, og for dette punktet får me at $0 = f(x_0) + f(x_0)(x_1 - x_0)$. Me løyser for x_1 og får $x_1 = x_0 - f(x_0)/f'(x_0)$. Dette gir oss den nye verdien for x som me tek utgangspunkt i for neste iterasjonsledd. Me finn då tangenten til punktet $f(x_1)$, som krysser x-aksen i $(x_2, 0)$. Slik går ein fram med nye iterasjonsledd heilt til det vert oppnådd konvergering, eller ein når ein satt maksimumsverdi for tal på iterasjonar. Det generelle uttrykket for x_n for kvart ledd er

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \tag{3.8}$$

Slik kan ein finne ein tilnærma verdi for x som gir $f(x) = 0$. Men me ynskjer å finne den verdien for x som gir maksimumspunktet til $f(x)$. Då treng me ikkje x-verdiane som gir $f(x) = 0$ men dei som gir $f'(x) = 0$. Framgangsmåten er den same, men me bruker no den deriverte funksjonen til x, $f'(x)$ istaden for den vanlege $f(x)$. Tangenten til denne funksjonen har då likning $y = f'(x_0) + f''(x_0)(x - x_0)$ som gir det generelle uttrykket for x_n

$$x_n = x_{n-1} - \frac{f'(x_{n-1})}{f''(X_{n-1})} \tag{3.9}$$

I vanleg Newton-Raphson algortime vert dette uttrykket brukt direkte og ein må difor finne både den første og den andrederiverte til $f(x)$ for kvart



Figur 3.1: *Newton-Raphson algoritme*

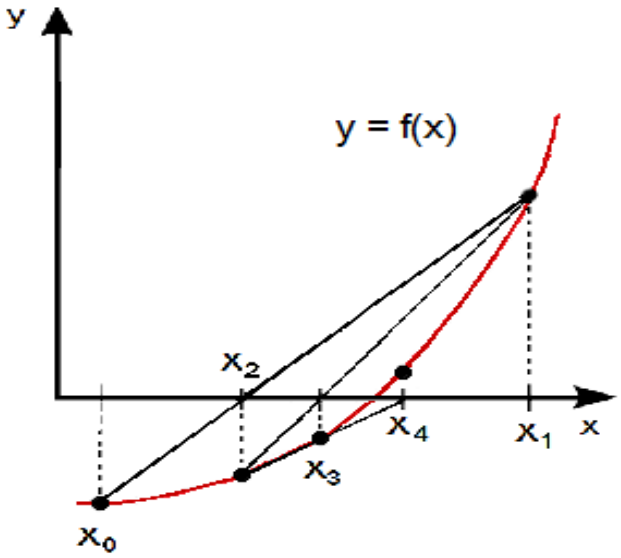
iterasjonsledd. Quasi-Newton bruker istaden eit tilnærma uttrykk for $f'(x)$.

$$f'(x_{n-1}) = \frac{f(x_{n-1}) - f(x_{n-2})}{x_{n-1} - x_{n-2}} \tag{3.10}$$

Dette kan setje inn i uttrykka (3.7) og (3.8). Her tek ein utgangspunkt i sekanten mellom to punkt istaden for tangenten til eit punkt. Sjå figur 3.2. Også for Quasi-Newton må ein overføre denne framgangsmåten til $f'(x)$ for å finne ekstrempunkta og ikkje røtene til $f(x)$. For å finne maksimumspunktet til ein funksjon får me då

$$x_n = x_{n-1} - f'(x_{n-1}) \frac{x_{n-1} - x_{n-2}}{f'(x_{n-1}) - f'(x_{n-2})}$$

Ein treng difor kun å finne den førstederiverte til f(x) med denne metoden. Dersom $\boldsymbol{x}$ er flerdimensjonal underestimerer Quasi-Newton-algoritmen matrisa med dei partiellderiverte av andre grad. Dette kjem av at metoden brukar kun ei tilnærming til den andrederiverte. Fleire modifiserte uttrykk for den hessiske matrisa er forsøkt formulert for å gjere estimatet meir lik korrekt verdi. BFGS (namngitt etter hovudpersonane bak modellen, Broyden, Fletcher, Goldfarb og Shanno) er ansett som den metoden som presterer best. Ved hjelp av den førstederiverte oppdaterer den estimatet til den hessiske matrisa etter kvar iterasjon, og modellen skaper slik ein tilnærming til korrekte verdiar.



Figur 3.2: *Quasi-Newton sekant algoritme*

For fleire detaljar rundt BFGS-metoden og Newton-algrotimane generelt vert lesar henvist til anna relevant litteratur, av t.d Dennis og Schnabel [5].

EM algoritmen

EM algorimen (The Expectation-Maximization Algorithm) er ein metode for å finne SME i tilfeller der ein ikkje har eit fullstendig datasett. Det kan t.d. vera i situasjonar med sensurert eller trunkert data, tilfeller der ein har samla observasjonar i grupper, eller i situasjonar der ein rett og slett har mista eller aldri har funne fleire av verdiane. Fordelinga er då avhengig av variablar som kun er indirekte observert gjennom dei andre direkte observerte variablane. Det kan i desse tilfella vera vanskeleg å finne SME for parametrane kun basert på det direkte observerte datasettet. Då kan det ofte vera til hjelp å auka datasettet ved å estimere verdiar for dei indirekte observerte variablane. Likelihoodfunksjonen vert då basert på ein funksjon som inneheld både indirekte og direkte variablar, og kan difor vera lettare å maksimera. Datasettet med kun dei direkte observerte variablane vert kalla det ukomplette datasettet og observasjonane vert kalla y. Datasettet som innheld dei direkte observerte variablane i tillegg til dei indirekte, vert kalla det komplette datasettet og er merka x.

EM algoritmen er ein iterativ metode som alternerer mellom to steg, E (expection-forventning)- steget og M (maximization-maksimerings)-steget. Det første algoritmen gjer er å bestemme byrjarverdiane for $\boldsymbol{\vartheta}$ ved gjetting. Desse verdiane vert så anvendt i det første E-steget. I E-steget vert dei uobserverte verdiane estimert ved hjelp av forventninga deira kalkulert med dei førebelse verdiane til $\boldsymbol{\vartheta}$. I M-steget maksimerer ein likelihooden til det komplette datasettet ved å halde verdiane for x funne i E-steget konstant. Ein finn då nye foreløbige SME for $\boldsymbol{\vartheta}$. Desse vert så brukt i eit nytt E-steg der ein finn nye forbetra verdiar for x. Prosessen vert så gjenteke heilt til ein oppnår konvergens. Verdiane ein då finn for $\boldsymbol{\vartheta}$ er dei som maksimerer likelihoodfunksjonen til $f(y \mid \boldsymbol{\vartheta})$.

Algoritmen er utleia og namngitt av Dempster, Laird og Rubin [4]. Den hadde vorte anvendt av andre i tidligare former, men vart gjennom artikkelen til Dempster at. el. etablert som eit viktig og populært verktøy i internasjonale statistiske kretser. I ZIP- og ZAP-regresjon er det for ZIP EM-algorimen er mest relevant. Me skal difor no sjå på EM algoritmen slik Lambert [8] anvender den for ZIP.

Ved maksimering av loglikelihooden for ZIP får Lambert eit problem i leddet for $y_i = 0$, då dette innheld både $\boldsymbol{\omega}$ og $\boldsymbol{\rho}$, sjå (3.6). Ho forutset difor vidare at me veit kven av 0-observasjonane som kjem frå den gruppa med vanlege 0-ar og kven som utgjer gruppa med dei strukturelle 0-ane. For å skilje mellom observasjonane frå dei to gruppene definerer ho ein variabel Z_i som får verdien 1 dersom observasjonen er ein strukturell 0, og verdien 0 ellers. Z vert då ein indirekte observert variabel. Det ukomplette datasettet er det opprinnelige som avheng av $\mathbf{y}$, medan det komplette datasettet inneheld både $\mathbf{y}$ og $\mathbf{z}$. Merk at $\mathbf{y}$ og $\mathbf{z}$ tilsaman utgjer variabelen kalla $\mathbf{x}$ i den generelle omtalen av EM algoritmen over. For å unngå å bruke same notasjon på denne variabelen som for kovariatmatrisa i ZAP vil eg bruke notasjonen q_i for Lambert sin z_i. Utleiinga av likelihoodfunksjonen til ZIP over viste at for å finne SME for regresjonskoeffisientaneein treng ein ikkje vite kva nullar som er strukturelle og ikkje. Det me treng er ein metode som klarar å maksimera likelihooden sjølv om leddet for $y_i = 0$ innheld både $\boldsymbol{\omega}$ og $\boldsymbol{\rho}$. Lambert bruker då heller ikkje EM algoritmen for å finne ut kva 0 som er strukturelle og ikkje, men for å dele opp likelihoodfunksjonen i to delar som kan maksimerast separat.

Lambert utleiiar vidare likelihoodfunksjonen for det komplette datasettet. Eg vil her gjengi utleiinga og ta med noko fleire mellomrekningar. Ein kan då bruke frå vanleg sannsynsrekning at $P(A \bigcap B) = P(A) \times P(B \mid A)$. Dette

gir oss ein samla fordelingsfunksjon felles for alle y_i-ane.

$$\begin{aligned} f(y_i, q_i; \phi_i, \lambda_i) &= f(q_i; \phi_i)\ f(y_i \mid q_i, \lambda_i) \\ &= \phi_i^{q_i}(1-\phi_i)^{1-q_i}\ (\frac{\lambda_i^{y_i}\ e_i^{\lambda}}{y_i!})^{1-q_i} \end{aligned}$$

Me kan no finne likelihoodfunksjonen for det komplette datasettet

$$\begin{aligned} L(\phi_i, \lambda_i; q_i, y_i) &= \prod_{i=1}^{n} f(y_i) \\ &= \prod_{i=1}^{n} \phi_i^{q_i}\ (1-\phi_i)^{1-q_i}\ (\frac{\lambda_i^{y_i} e_i^{\lambda}}{y_i!})^{1-q_i} \\ \log L(\phi_i, \lambda_i; q_i, y_i) &= \sum_{i=1}^{n} [q_i \log(\phi_i) + (1-q_i)\log(1-\phi_i) \\ &\quad + (1-q_i)\left(\log(\lambda_i^{y_i}) + \log(e^{-\lambda_i}) - \log(y_i!)\right)] \\ &= \sum_{i=1}^{n} [q_i \log(\phi_i) + (1-q_i)\log(1-\phi_i) \\ &\quad + (1-q_i)\big(y_i \log(\lambda_i) - \lambda_i - \log(y_i!)\big)] \end{aligned}$$

Me innfører så parameterverdiar for $\phi_i, (1-\phi_i)$ og λ_i.

$$\begin{aligned}
\log L(\omega_i, \rho_i; q_i, y_i) &= \sum_{i=1}^{n} \big[q_i \log(\frac{e^{\boldsymbol{\omega G}_i}}{1+e^{\boldsymbol{\omega G}_i}}) + (1-q_i)\log(\frac{1}{1+e^{\boldsymbol{\omega G}_i}}) \\
&\quad + (1-q_i)\left(y_i \ \log(e^{\boldsymbol{\rho H}_i}) - e^{\boldsymbol{\rho H}_i} - \log(y_i!)\right)\big] \\
&= \sum_{i=1}^{n} \big[q_i \ \left(\log(e^{\boldsymbol{\omega G}_i}) - \log(1+e^{\boldsymbol{\omega G}_i})\right) + (1-q_i)\left(\log(1) - \log(1+e^{\boldsymbol{\omega G}_i})\right) \\
&\quad + (1-q_i)\left(y_i \ \boldsymbol{\rho H}_i - e^{\boldsymbol{\rho H}_i} - \log(y_i!)\right)\big] \\
&= \sum_{i=1}^{n} \big[q_i \left(\boldsymbol{\omega G}_i - \log(1+e^{\boldsymbol{\omega G}_i})\right) - (1-q_i)\log\left(1+e^{\boldsymbol{\omega G}_i}\right) \\
&\quad + (1-q_i)(y_i \ \boldsymbol{\rho H}_i - e^{\boldsymbol{\rho H}_i}) - (1-q_i)\log(y_i!)\big] \\
&= \sum_{i=1}^{n} [q_i \ \boldsymbol{\omega G}_i - (q_i + 1 - q_i)\log(1+e^{\boldsymbol{\omega G}_i})] \\
&\quad + \sum_{i=1}^{n}(1-q_i)(y_i \ \boldsymbol{\rho H}_i - e^{\boldsymbol{\rho H}_i}) - \sum_{i=1}^{n}(1-q_i)\log(y_i!)] \\
&= \sum_{i=1}^{n} [q_i \ \boldsymbol{\omega G}_i - \log(1+e^{\boldsymbol{\omega G}_i})] + \sum_{i=1}^{n}(1-q_i)(y_i \ \boldsymbol{\rho H}_i - e^{\boldsymbol{\rho H}_i})] \\
&\quad - \sum_{i=1}^{n}[(1-q_i)\log(y_i!)]
\end{aligned}$$

$$\log L(\omega_i, \rho_i; q_i, y_i) = \log L_1(\boldsymbol{\omega}; \boldsymbol{y}, \boldsymbol{q}) + \log L_2(\boldsymbol{\rho}; \boldsymbol{y}, \boldsymbol{q}) - \sum_{i=1}^{n}(1-q_i)\log(y_i!) \tag{3.11}$$

Dette er likelihooden for det komplette datasettet, og denne kan brukast i M-leddet i EM-algoritmen for å finne SME for $\boldsymbol{\rho}$ og $\boldsymbol{\omega}$ for den opprinnelege likelihooden $L(\boldsymbol{\rho^k}, \boldsymbol{\omega^k}; \boldsymbol{y})$, sjå uttrykk (3.6). I E-ledda vert Q_i estimert med forventninga den har under dei førebelse verdiane for $\boldsymbol{\rho}$ og $\boldsymbol{\omega}$. I M-leddet maksimerer me likelihoodfunksjonane for det komplette datasettet med omsyn på $\boldsymbol{\rho}$ og $\boldsymbol{\omega}$. Dette gir oss nye estimat for $\boldsymbol{\rho}$ og $\boldsymbol{\omega}$ som ein bruker i eit nytt E-steg. Likelihoodfunksjonen kan no skrivast på forma (3.11), og $\log L_1$ og $\log L_2$ kan maksimerast separat i M-leddet i algoritmen. Då får ein eitt M-steg for kvar av $\boldsymbol{\rho}$ og $\boldsymbol{\omega}$. Når algoritmen konvergerer er verdiane funne i det siste itersjonsleddet SME for $\boldsymbol{\rho}$ og $\boldsymbol{\omega}$ for den opprinnelige likelihooden $L(\boldsymbol{\rho^k}, \boldsymbol{\omega^k}; \boldsymbol{y})$ (3.6). Lambert gir ein grundig gjennomgang av stega i algo-

ritmen, samt gode startverdiar for $\boldsymbol{\rho}$ og $\boldsymbol{\omega}$.

3.1.6 To ulike regresjonsmodellar

Sjølv om ZIP og ZAP har forskjellig strukturell oppbygging og angrip situasjonen med mange nullar på ulik måte, vil dei under visse omstende verta den same modellen. Ser ein på dei to modellane kun som sannsynsfordelingar, kan ZIP definerast som ei reparameterisering av ZAP, der $p = \phi + (1-\phi)\, e^{-\lambda}$. Dette fører til at i regresjonssituasjonar med kun intercept-ledd som regresjonsparameter vil modellane i teorien gi heilt like resultat. Dette skal me no bevise.

For å få situasjonen rundt datasettet heilt lik for begge modellane, gir me λ_i og sannsynet $P(y_i) = 0$ lik verdi for ZIP og ZAP. Dersom me let modellane kun ta intercept-ledd, vert vektoren med regresjonparametrar kun ein skalar og λ, p og ϕ vert fastsett som konstantar. Då vert $f(y_i)$ lik for alle y_i, og regresjonsmodellane blir lik fordelingsfunksjonane, sjølv om ein har innført regresjonsparametrar. Me har då

$$P(Y_i = y) = (1-\phi)\, \frac{\lambda^y\, e^{-\lambda}}{y!}\; y = 1, 2, \ldots \tag{3.12}$$

$$P(Y_i = 0) = \phi + (1-\phi)\, e^{-\lambda} \tag{3.13}$$

for ZIP, og

$$P(Y_i = y) = (1-p)\, \frac{\lambda^y\, e^{-\lambda}}{y!\,(1-e^{-\lambda})}\; y = 1, 2, \ldots \tag{3.14}$$

$$P(Y_i = 0) = p \tag{3.15}$$

for ZAP. Lat oss no setje $P(Y_i = 0)$ lik for begge modellane. Det vil seie

$$P_{\text{ZIP}}(Y_i = 0) = \phi + (1-\phi)e^{-\lambda} = p = P_{\text{ZAP}}(Y_i = 0)$$

Me løyser så denne for $1 - \phi$.

$$\begin{aligned}
\phi + (1-\phi)e^{-\lambda} &= p \\
\phi(1 - e^{-\lambda}) + e^{-\lambda} &= p \\
\phi &= \frac{p - e^{-\lambda}}{1 - e^{-\lambda}} \\
1 - \phi &= 1 - \frac{p - e^{-\lambda}}{1 - e^{-\lambda}} \\
1 - \phi &= \frac{1-p}{1-e^{-\lambda}}
\end{aligned}$$

Set me dette uttrykket for $(1-\phi)$ inn i (3.13) får me

$$P_{\text{ZIP}}(Y_i = y) = \frac{1-p}{1-e^{-\lambda}}\frac{\lambda^y e^{-\lambda}}{y!} = (1-p)\frac{\lambda^y e^{-\lambda}}{y!\ e^{-\lambda}} = P_{\text{ZAP}}(Y_i = y) \qquad (3.16)$$

Då vert fordelinga til ZIP lik fordelinga til ZAP. Dette viser at i situsjonar der ein innfører kun intercept-ledd som regresjonsparameter er det ikkje forskjell på ZIP og ZAP som regresjonsmodellar. Dette kjem av at ϕ, p og λ har same verdi for alle y_i. Sjølv om det er to ulike parametre p og ϕ, som vert estimert ulikt, vil verdien deira likevel alltid oppfylle (3.16). Ein står igjen med ein lik modell for både ZIP og ZAP.

ZIP og ZAP utgjer likevel ikkje same regresjonsmodell i meir generelle situasjonar. Dette ser me dersom me innfører kovariatar i modellane. Då får ein også fleire regresjonparametrar, og kriteria til parameterestimata vert strengare. Då må regresjonsparametrane som ein funksjon av kovariatane danne lineære grafar. T.d må $\nu+\gamma_1 Z_{i1}+...+\gamma_q Z_{iq}$ danne ei rett linje for dei ulike verdiane i vektoren Z_i. Det er ikkje alltid ein samstundes klarer å setje $P(Y=0)$ lik for ZIP og ZAP. Merk at for ZIP har λ_i også innverknad på sannsynet for $y=0$, og ikkje berre for observasjonane over 0. Det gjer kravet om lineære regresjon endå strengare og vanskelegare å få til. Skal modellane vera like må ein oppfylle det lineære kriteriet samstundes som at λ og $P(Y=0)$ må kunne ta same verdi for begge modellane. Klarar me å finne ein situasjon der dette ikkje vert oppfylt, har me vist at ZIP og ZAP ikkje er den same modellen, og ikkje treng å gi same resultat på analysen. Lat oss sjå på eit slikt tilfelle. For at me framleis skal ha lik situasjon for ZIP og ZAP, er dei fire kovariatmatrisene satt like og kalla $\mathbf{X}$. I tillegg er $P(Y=0)$ framleis lik for begge. Regresjonsparametrane vert estimert ulikt for modellane, og ein skil desse frå kvarandre med ulik notasjon.

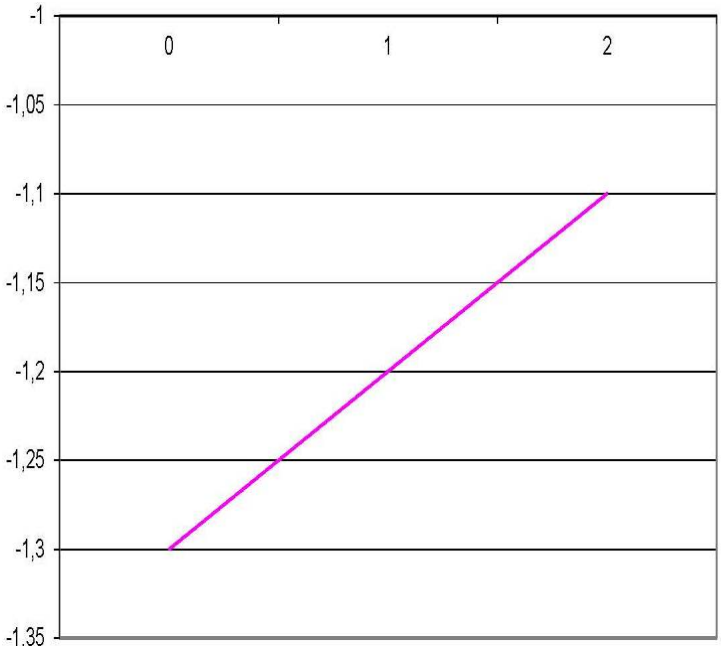
Me innfører no ein forklaringsvariabel for kvar av dei to komponentane i modellane. Det gir oss at $\log(\phi_i)/(1-\phi_i) = \iota+\omega X_i$, $\log(p_i)/(1-p_i) = \nu+\gamma X_i$, $\log(\lambda_i) = \delta + \rho X_i$ og $\log(\lambda_i) = \alpha + \beta X_i$. Alle desse må vera lineære funksjoner av dei ulike verdiane til $\mathbf{X}$. For at modellane skal bli like må i tillegg $P_{\text{ZIP}}(Y_i=0)$ kunne settast lik $P_{\text{ZAP}}(Y_i=0)$. Det vil seie at dersom me finn verdier for ν, γ, ι, ω og δ som oppfyller at $\log(p_i)(1-p_i) = \nu + \gamma X_i$ er lineær for $\mathbf{X}$ og $P_{\text{ZIP}}(Y_i=0) = P_{\text{ZAP}}(Y_i=0)$, men som ikkje oppfyller at $\log(\phi_i)(1-\phi_i) = \iota + \omega X_i$ er lineær for $\mathbf{X}$, held ikkje alle kriteria over. Då er ikkje ZIP og ZAP like i alle situasjonar, og kan ikkje sjåast på som same model.

Me byrjar med å sjå på ZAP og fastset først ν = -1,3 og γ = 0,1. Verdiane er valt vilkårlig, men oppfyller kravet om at alle sannsyna i modellane, $P(Y_i = 0)$ og $\phi = P(Y_i =$ strukturell 0), ligg mellom 0 og 1. Då får me desse verdiane for $\log(p_i)(1 - p_i) = \nu + \gamma X_i$

X_i	$\log(p_i)(1 - p_i)$
0	-1,3
1	-1,2
2	- 1,1

Tabell 3.1: *Verdiar for* $\log(p_i)(1 - p_i)$ *for ZAP*

Me har kun èin forklaringsvaribel og $\nu + \gamma X_i$ er på formelen til ei rett linje for alle val av ν og γ og X. Sjå figur 3.3.



Figur 3.3: *ZAP gir lineær samanheng for reg. koef.,* $\log(p_i/(1 - p_i))$

No treng me å finne verdien til $p_i = P_{\text{ZAP}}(Y_i = 0)$ i dømet vårt for å kunne setje den lik $P_{\text{ZIP}}(Y_i = 0)$. Det gjer me med å løyse $\log(p_i)(1 - p_i) = \nu + \gamma X_i$

med hensyn på p_i.

$$\begin{aligned}
\log \frac{p_i}{1-p_i} &= \nu + \gamma X_i \\
\frac{p_i}{1-p_i} &= e^{\nu+\gamma X_i} \\
p_i &= e^{\nu+\gamma X_i} - p_i\, e^{\nu+\gamma X_i} \\
p_i(1+e^{\nu+\gamma X_i}) &= e^{\nu+\gamma X_i} \\
p_i &= \frac{e^{\nu+\gamma X_i}}{1+e^{\nu+\gamma X_i}}
\end{aligned}$$

Verdiane for $\mathbf{p}$ vert då som vist i tabell 3.1.6.

X_i	$p_i = P(y_i = 0)$
0	0,374630521
1	0,431012761
2	0,498960659

Tabell 3.2: $P(y = 0)$ *for ZAP ved ulike verdier av* X

Me set så desse verdiane lik $P_{\text{ZIP}}(y = 0)$ og løyser denne med hensyn på $\boldsymbol{\phi}_i$.

$$\begin{aligned}
p_i = P_{\text{ZIP}}(Y_i = 0) &= \phi_i + (1-\phi_i)\, e^{-(\delta+\rho X_i)} \\
p_i &= \phi_i + e^{-(\delta+\rho X_i)} - \phi_i\, e^{-(\delta+\rho X_i)} \\
p_i - e^{-(\delta+\rho X_i)} &= \phi_i - \phi_i\, e^{-(\delta+\rho X_i)} \\
p_i - e^{-(\delta+\rho X_i)} &= \phi_i\, (1 - e^{-(\delta+\rho X_i)}) \\
\frac{p_i - e^{-(\delta+\rho X_i)}}{1 - e^{-(\delta+\rho X_i)}} &= \phi_i
\end{aligned}$$

Me fastset så vilkårlig $\delta = 1$ og $\rho = 2$, for å finne verdiane til $\boldsymbol{\phi}$.

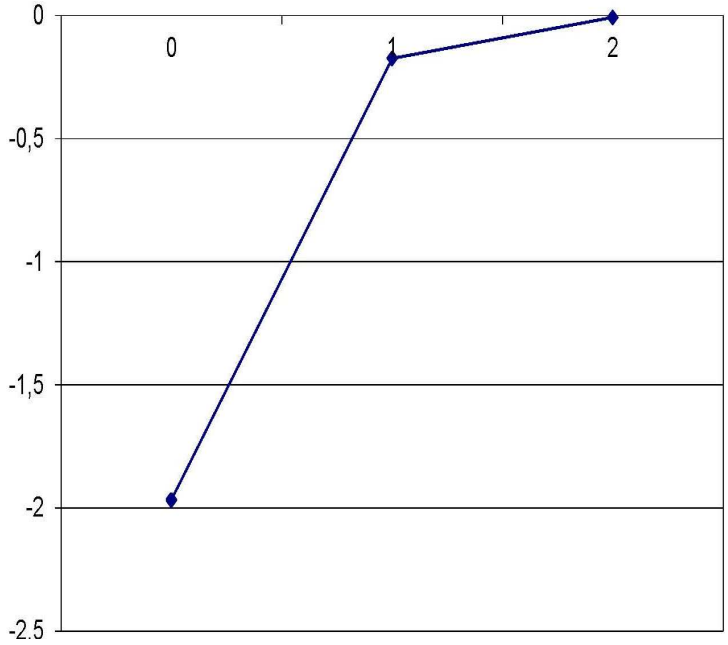
X_i	ϕ_i
0	0,01068005
1	0,401200278
2	0,495561781

Tabell 3.3: *Verdiar for* $\boldsymbol{\phi}$ *for ulik* X_i

Me kan no utfrå ϕ_i finne $\log(\phi_i/(1-\phi_i))$ for $\mathbf{X}$ direkte, utan å trenge verdiar for ι og ω.

X_i	$\log(\phi_i/(1-\phi_i))$
0	-1,96676347
1	-0,173920365
2	-0,007710179

Tabell 3.4: *Verdiar for* $\log(\phi_i/(1-\phi_i))$ *for ulik* X_i



Figur 3.4: *ZIP gir ikkje lineær samanheng for reg.koef.,* $\log(\phi_i/(1-\phi_i))$

Figur 3.4 viser $\log(\phi_i/(1-\phi_i))$ som ein funksjon av $\mathbf{X}$). Ein ser tydeleg at punkta ikkje fell på ei rett linje. Det vil seie at $\log(\phi_i/(1-\phi_i))$ ikkje er lineær som ein funksjon av $\mathbf{X}$. Då er ikkje det lineære kriteriet møtt, og me har funne ein situasjon der ZIP og ZAP ikkje utgjer same regresjonsmodell. Me kunne fått ZIP lineær også, men då ville me måtte gå bort frå kriteriet at den skal vera ZAP. Dette viser at ZIP og ZAP kan gi ulike resultat i regresjonsanalyser dersom regresjonsvektorane har dimensjon større enn 1. Skilnaden på modellane er korleis dei ser på nullane, som igjen gjer utslag for SME for $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}$ og $\boldsymbol{\rho}$. Desse avgjer kva forklarinsvariablar som vert signifikante, og kva resultat analysen gjev.

Utrekningane og verdiane er funne ved prøve og feile-metoden av ulike verdier i eit arbeidsark laga i excell.

Kapittel 4

Zero-inflated og zero-altered Poissonfordeling i historisk samanheng

4.1 ZIP og ZAP i historisk samanheng

Det har dei siste åra vokse fram ei betydeleg interesse for regresjonsmodellar som tillet og kalkulerer med fleire nullar enn det Poissonfordelinga forutset. Denne interessen har hovedsakleg vokse fram frå artiklane til Mullahy [10] og Lambert [8]. Dei står som pionerar på området, og vert nemnt i dei fleste artiklar og bøker som omhandlar emnet. Eg vil difor her sjå nærare på desse to artiklane og modellane dei omhandlar.

Mullahy [10] står som den første til å ha lagt fram og utvikla ZIP og ZAP modellane. Han tek utgangspunkt i "the double hurdle modell" (dobbelt hinder modellen) som vart utvikla av Cragg i 1971 [2], og det er denne modellen namnet hurdle er henta frå. Cragg sin modell byggjer på Tobit-fordelinga, som er ei fordeling for sensurerte normalfordelte variablar. Fram til då var det denne fordelinga som var mest anvendt for datasett med ekstra mange 0-observasjonar. Med utgangspunkt i Tobit-fordelinga utviklar Mullahy ein modell som tillet sannsyna til utfalla over 0 å avhenge av andre parametrar og forklaringsvariablar enn for sannsynet til utfallet 0. Modellen er forklart med at ein må forsere to hinder for å observere ein verdi som er over 0, og det er dette namnet to hinder- modellen kjem frå. Det første hinderet vert omtalt som eit krav for å få delta,(dvs. ikkje ha verdien 0), medan det andre hinderet er graden av deltaking (å vere den spesifikke verdien over 0).

Men tobit-fordelinga, og difor også Cragg sin double hurdle-modell er ei fordeling for normalfordelte, kontinuerlege variablar. Mullahy tek tak i ei bekymring som hadde vakse fram om at regresjon basert på kontinuerlege fordelinger som tobit ville gi ukonsekvente og ukorrekte estimat i regresjonsanalyse for diskret teljedata. Han tek difor med seg dei grunnleggande tankane i modellen til Cragg, og formulerer ein modell som tek seg av situasjonar med mange nullar, men som passar for diskret telledata. Han adopterer også delar av namnet, men omtalar kun det første hinderet til Cragg som eit hinder.

Hurdle-modellen Mullahy utvikla er same modell som idag vert kalla ZAP. Mullahy gir i artikkelen ei kort forklaring av modellen og set opp likelihoodfunksjonen på den generelle forma

$$\prod_{t\in\Omega_0}[1-\Phi_1(\theta_1)]\prod_{t\in\Omega_1}[\phi_2(y,\theta_2)\Phi_1(\theta_1)/\Phi_2(\theta_2)] \tag{4.1}$$

Mullahy bruker i artikkelen noko utradisjonell notasjon. Med $t \in \Omega_i$ omtalar han alle observasjonar med verdi frå og med i og oppover. $\phi_2(y,\theta_2)$ er den opprinnelege sannsynsfordelinga og er definert som ei høgresensurert fordeling ved 0, medan Φ_2 er definert som den kummulative fordelinga til ϕ_2. Merk at Mullahy her utelater variabelen y i den kummulative fordelinga. θ_i er parametrar i fordelingsfunksjonane. Den 0-trunkerte fordelinga til ϕ_2 vert $\big(\phi_2(y,\theta_2)\big)/\big(\Phi_2(\theta_2)\big)$. (Merk forskjellen på sensurert og trunkert fordeling.) Φ_1 er sannsynet for å krysse hinderet. Sannsynet for å observere ein 0 blir då $(1-\Phi_1)$. Sidan det samla akkumulerte sannsynet må bli ein, er det trunkerte sannsynet for $P(Y=y)$ for $y>0$ nedjustert med hensyn på dette. Sannsynet for å ikkje krysse hinderet, $(1-\Phi_1)$, treng difor ikkje vera lik den sannsynsmengda den opprinnelege fordelinga ϕ_2 ville gitt $P(Y=0)$. Det vil seie at Φ_1 og Φ_2 kan vera to ulike fordelingar, og sannsynet for å krysse hinderet treng ikkje vera lik det akkumulerte sannsynet $\sum_{y=1,2,\ldots}\phi(Y=y;\theta_2)$. Dersom ein set $\Phi_1=\Phi_2$, det vil seia å setje sannsynet å forserer hinderet lik det akkumulerte sannsynet for verdiane over 0 modellert med opprinneleg fordeling, vil (4.1) bli på forma

$$\prod_{t\in\Omega_0}[1-\Phi_2(\theta_2)]\prod_{t\in\Omega_1}[\phi_2(y,\theta_2)] \tag{4.2}$$

Då ender me opp med den opprinnelege fordelinga ϕ_2 der $P(Y=0) = 1-\sum_{y=1,2,\ldots}\phi(Y=y;\theta_2) = \phi_2(Y=0;\theta_2)$.

(4.2) er på forma til likelihoodfunksjonen for den høgre sensurerte tobit fordelinga, der alle negative verdiar vert notert som ein 0-observasjon. $(1-$

$\sum_{y=1,2,...} P_{\text{Tobit}}(Y = y)$ er difor det samla sannsynet for både dei observasjonane som opprinneleg var negative og 0-obervasjonane. Men for teljedata er ingen observasjonar negative i utgangspunktet, og $(1 - \sum_{y=1,2,...} P(Y = y) = P(Y = 0)$. Det vil seie at for ϕ_2 brukt i hurdle-modellen er ikkje sensureringa aktiv. (4.2) er difor også på forma til alle fordelingar for teljedata. Det Mullahy vil poengtere er då heller ikkje sensureringa, men samanhengen mellom sin eigen modell og Cragg sin modell som tek utganspunkt i Tobit-fordelinga. Det er fleire måtar Φ_1 kan bli lik Φ_2 på. For ZAP slik modellen er allment anerkjent i dag og vert omtala i denne oppgåva, skjer dette ved at $p = P_{\text{bernoulli}}(Y = 0)$ vert lik $\left(1 - \sum_{Y=1,2,...} P_{\text{Poisson}}(Y = y)\right)$. Då vert (4.2) likelihoodfunksjonen til den vanlege Poissonfordelinga. Merk at for ZAP er ikkje parametrane $\theta_1 = p$ og $\theta_2 = \lambda$ like sjølv om Φ_1 og Φ_2 er det. Mullahy påpeikar at dette er mogleg, og opnar slik opp for ei generell form av hurdlemodellen, lik den moderne ZAP. Han vel likevel i den vidare formuleringa av hurdle-modellen å legge inn ei forutsetning om at når Φ_1 vert lik Φ_2 må dette komme av at parametrane ϕ_1 og ϕ_2 er like. Det vil seie ar hurdle-modellen kjem på forma til den opprinnelege modellen kun grunna like parameterverdiar. Då må ϕ_1 og ϕ_2 følgje samme opprinnelege fordeling. Dersom fordelinga er Poisson vert sannsynet for 0 $P_{\text{Poisson}}(y = 0) = e^{-\lambda}$ og $P(\text{forsere hinderet}) = (1 - e^{-\lambda})$. I (4.1) treng likevel ikkje parametrane ha same verdi, og Mullahy kallar dei λ_1 og λ_2. Ein får då ein hurdle modell og ikkje ein vanleg modell. Andre har seinare utvikla modellen til Mullahy vidare til den me kjenner som ZAP. I ZAP treng ikkje sannsynet for å krysse hinderet ha noko samanheng med fordelinga til utfalla over 0.

I artikkelen introduserer Mullahy også ein modell han kallar "the with-zero model". Det er denne som no er kjent som ZIP, og både sjølve modellen til Mullahy og strukturen og tankane bak er lik ZIP slik den er omtalt i denne oppgåva, med untak av at han ikkje innfører regresjonsparametrar for forventninga til dei struktuelle nullane. I staden definerer han ϕ som ein konstant som styrer kor mykje ekstra sannsyn den opprinnelege fordelinga skal til utfallet 0. Det er difor i Mullahy sin "with-zero model" kun eit sett med regresjonsparametrar.

Mullahy bruker i sin artikkel Poisson og geomterisk fordelingar, men poengterer at det også er mogleg å anvende idèane på andre diskret fordelingar.

Seinare i artikkelen tek han opp situasjonen der ein har kun intercept ledd i regresjonen, og poengterer at modellane då vil verte til samme modell. Han held som nevnt over ϕ som konstant, men innfører parametrar for λ_1 og λ_2.

Han set $\lambda_1 = \beta + \alpha$ og $\lambda_2 = \beta$, og let med det eit interceptledd skille mellom dei. Dette er nyttig når han seinare testar om det er overdispersjon med hensyn til 0-ane. Då testar han for hurdle-modellen om α er lik 0, og for With-zero modellen om ϕ er lik 0. Dersom dette er tilfellet vil modellane bli lik den opprinnelege Poisson eller geometriske fordelinga. For å gjere denne analysen bruker han Score, Hausman of information matrix tester. Score-testen fungerer godt i dei flest tilfella, medan Hausman-testen er særleg bra ved testing opp mot geometrisk fordeling.

Mullahy illustrerar modellane med ein analyse av inntak av kaffi, te og mjølk for vaksne mellom 20 og 64 år busette i USA. Datasettet er innhenta i to omgangar, våren 1979 og våren 1980 av the National Survey og Personal Health Practices and Consequences. Han brukar kun geometrisk og ikkje Poissonfordeling i analysen, og utfører regresjonen med vanleg geometrisk, hurdle og with zeros modellane. Han anvender alle tre modellane to gonger, først med bruk av kun et interceptledd, og deretter med fleire kovariatar. Analysen gir mykje likt resultat for alle modellane, og den viser at ved bruk av kun interceptleddet gir hurdle og With-zero modellane heilt same estimatverdiar.

Mullahy sin artikkel kan virka svært tunglest og vanskeleg å forstå. Han brukar noko uvanleg notasjon, samt at han forutset at lesaren kjenner til Craggs "double hurdle-modell". Han er også noko kort i forklaringane sine, noko som krev innsats av lesaren og gjer det vanskeleg for lesarar uten same bakgrunn å setje seg inn i modellane. Artikkelen er også noko mangelfull med tanke på dagens ZAP og ZIP sidan han ikkje innfører regresjonsparametre for "with zero-modellen", og at sannsynet for 0 i ZAP er avgrensa til å avhenge av Poissonfordelinga. Likevel førte artikkelen til interesse for modellane og auka auka bekymring for å bruke modifiserte kontinuerlege fordelingar for teljedata.

Det har særleg vore ei stadig auka interesse for ZIP. Mykje av grunnen til dette er Lambert sin artikkel frå 1992 [8] der ho forklarar ZIP ved hjelp av ein regresjonsanalyse av manglar ved lodding av komponentar på elektriske tavler. Det var ofte anteke at slike datasett fylgde vanleg Poissonfordeling, men Lambert påpeikar at fleire innhenta datasett har fleire 0-observasjonar enn det Poisson tilseier. Ho meinar at ei forklaring kan vera uobserverbare endringar i omgivnadane som fører til at prosessen skiftar fram og tilbake mellom to stadier. I det ho kallar det perfekte stadiet er tavlene bortimot heilt resistante mot feil, medan i mangel-stadiet er feil meir vanleg men ikkje absolutt. Det perfekte stadiet fører til ei auka mengd 0-ar. Dette er eit prak-

tisk døme på den grunnleggjande tanken i ZIP at 0-ane kjem frå to kilder.

Sjølv om det også før hadde vore skrive artiklar som omtalte ZIP, er Lambert kjent som den første til å innføre kovariatar for sannsynet for 0-observasjonar. Ho brukar $\log(\lambda)$ og $\text{logit}(p)$ som linkfunksjonar, og dette har i etterkant vorte det tradisjonelle valet for ZIP. I forhold til Mullahy innfører ho også eit nytt aspekt ved å la λ og p kunne avhenge av kvarandre, og gir fleire døme på forholdet mellom dei, samt fleire moglege linkfunksjonar.

Artikkelen er noko lettare skrive enn den av Mullahy, og utleiingane og forklaringane er gode og til dels detaljerte. Særleg er utleiinga av likelihoodfunksjonen og estimeringa av SME for parametrane utfyllande. Ho går grundig gjennom bruken av EM-aloritmen for å finne SM-estimat, og viser med simuleringer for både ZIP og ZIP(τ) at EM er ein meir robust metode for ZIP enn Newton-Raphson. Artikkelen gir også ei god innføring i standardavvik og konfidensintervall for estimata, samt noko innføring i korleis ein skal tolke resultata av ein ZIP-regresjon. Den gode innføringa og formuleringa er antakeleg ein stor del av grunnen til at artikkelen er svært mykje referert til i liknande litteratur, og at interesan for ZIP stadig er aukande i fleire fagområde også utanom statistikk. Dette til tross for at ein bør kjenne til sjølve fordelingsfunksjonen til ZIP på førehand. Ein treng også noko kjenskap til generell statistikk.

Både namnet og modellen i artikkelen til Lambert er lik den som lærebøker og nyare artiklar bruker, og den som har vorte presentert tidlegare i denne oppgåva. Eg vil difor ikkje gå grundigare inn på detaljar om dette her.

I den praktiske anvendinga ser Lambert på mengda feil gjort ved lodding av komponenter på elektriske tavler. Datasettet er henta frå eit forsøk ved AT&T Bell Laboratories. For dei fleste tavlene oppstod det ingen feil, men der det vart funne feil vart det ofte funne fleire. To sett av kovariatar gav samme forventning for tilfelle av feil. Men det perfekte stadiet var meir sannsynleg under det eine settet av kovariatar, medan forventninga til tal på feil i det uperfekte stadiet var mindre under det andre settet med kovariatar. Ut frå dette kunne ein ikkje berre sjå kva sett av kovariater som gir minst forventning av feil, men også kva kovariatar som fører til dette restultatet. Ho utfører også negativ binomisk-regresjon på datasettet, og viser at ZIP taklar betre både under og overdispersjon. Ho poengterer likevel at ein zero-inflated negativ binomisk modell nok ville gitt endå betre resultat sidan den tek seg at overdispersjon grunna mange nullar og høge verdiar separat. Både ZIP og ZIP(τ) gav bedre treff for datasettet enn vanleg Poisson, og sjølv om ZIP(τ)

gav noko dårlegare resultat enn ZIP er forskjellen så liten at ein ut frå eit ønske om enklast moleg modell kunne valt ZIP(τ). Men sidan ein ikkje kunne vite noko sikkert om ein slik sammenheng, og sidan resultata falt bra saman med tidlegare resutltat, valgte Lambert å bruke ZIP.

Mullahy er den som vert omtalt som å først ha formulert dei to modellane, og det ser slik ut til modellane først oppstod i fagområdet statistikk i økonomisamanheng. Namnet zero-altered Poisson vart innført av Heilbron i 1989 [7] som anvendte ZAP på eit datasett som omhandla helserisiko blant homofile menn. Både ZIP og ZAP har har sidan vorte nytta også i mange ulike fagområde, t.d sosialvitenskap, psykologi og forskning på trafikkulykker. Det har vore ei stadig veksande interesse for begge modellane, men den mest kjente og brukte av dei to er ZIP. Mykje av grunnen til dette ser ut til å vera artikkelen til Lambert.

Kapittel 5

Modellane brukt som verktøy praktisk analysearbeid

5.1 ZIP og ZAP anvendt i praktiske situasjonar

Interessa og bruken av ZIP og ZAP har stadig auka, og modellane vert no anvendt i analyser i mange artiklar og studiar der ein bruker regresjonsmodellar. Eg vil no gi eit par døme på dette, samt sjå med noko kritisk blikk på anvendinga av modellane.

Eit eksempel der ZIP er nytta er ein forskningsartikkel av Manh m. fl [9] frå 2010. Etter store malariaepedemier i Vietnam tidleg på 1990-talet har tilfella av sjukdomen vorte kraftig redusert, hovedsakleg som følgje av strengare malariakontrollprogram og ein betre sosialøkonomisk situasjon. Likevel framstår malaria som eit stort problem i mange område, til tross for eit godt utbetra helsetilbod og gratis medisinering. Vietnam vert i studiet inndelt i åtte sosial-økologiske soner, og i tidsrommet januar 2007 til desember 2008 vart det for kvar sone notert månadlege tal på tilfeller av malaria. Talet på tilfelle vert så modellert med ZIP-regresjon opp mot forklaringsvariablar som jungel-relatert arbeid og fattigdom. Målet er å identifisere risikoområda. Valet av ZIP som analyseverktøy vert begunna med at over 80 prosent av observasjonane vart rapportert som ingen tilfelle av sjukdomen, samt at 0-rapportane kan komme frå to kilder eller prosessar. Dei strukturelle 0-ane i studiet, også kalla sanne nullar, kjem frå område der det ikkje var mogleg med overføring av sjukdomen, medan dei andre nullane, kalla tilfeldige nullar, kjem frå område der overføring kunne ha skjedd, men som ikkje hadde nokre innrapporterte tilfeller. Om det hadde vore mogleg å fjerne dei strukturelle nullane frå datasettet før utføring av analysen, kjem ikkje klart fram

i artikkelen.

I dette tilfellet er modellen valt med omhu. Men det kan i andre tilfelle virke som valet av modell er meir tilfeldig. Fleire av studia nemner ikkje noko om begrunninga for valet av modell, anna enn at det er fleire 0-observasjoner enn det vanleg Poisson kalkulerer med. Men dette kan både ZIP og ZAP ta seg av. Det kan tyde på at valet ofte er gjort med for lite kunnskap om dei to modellane. Eit døme på dette er Goodrich m. fl [6] si analyse frå 2010 av hjarte og karsjukdomar blant personar med bipolar liding. Det har vist seg at pasientar med bipolar liding ofte fører ein livsstil med mindre fysisk aktivitet, meir usunne matvanar og meir ruspåverknad enn resten av befolkninga. Risikofaktorar som høgt blodtrykk, høgt kolesterolnivå og type 2 diabetes oppstår i gjennomsnitt 14 år tidlegare blant personar med diagnosen bipolar liding samanlikna med menneske uten noko psykisk diagnose. I alt 298 krigsveteranar deltok i studiet som samla inn data over to år. Det vart blant anna prøvd å finne ein samanheng mellom kor mange gonger ein pasient nyttar eit tilbod om konsultasjonar med fokus på god livsstil, og faktorar som alkoholvaner, bruk av narkotika, BMI, kjønn og etnisk bakgrunn. For denne delen av studiet vart det anvendt ZIP-regresjon, kort begrunna med at det var mange pasienter registert med ingen konsultasjonar. Det kjem ikkje fram noko i beskrivinga av datasettet som tyder på at 0-ane kjem frå to kjelder. Alle pasientane hadde tilgang til tilbodet, så alle tok eit bevisst val om å ikkje nytte det. Artikkelen omtalar også ZIP-modellen noko feilaktig. Den skriv at logit-delen undersøkjer sannsynet for å ikkje delta på nokon konsultasjon, medan Poissondelen tek seg av dei som har møtt minst ein gong. Dette stemmer betre med dei grunnlegjande tankane i ZAP modellen, og gir inntrykk av at valet av modell er basert på for liten kunnskap og innsikt i ZIP. Ein kan lett tenkje seg at forfattaren av studiet ikkje har høyrt om ZAP. Det står i det heile lite i artikkelen om analysemetoden, grunna at artikkelen også er retta mot lesarar utan statistisk bakgrunn. Det er heller ikkje oppført noko relevant litteratur i referansane.

Andre artiklar viser god innsikt i den eine av modellane, men lite eller ukorrekt innsikt i den andre. Det er då oftast ZIP forfattarane har god kjennskap til. Dette gjeld t.d Shankar m.fl [13] si analyse frå 1997 av frekvensen på trafikkulykker opp mot veggeometri, fartsregulering og trafikkmønster. Han poengeter at ZIP tek seg av fleire nullar enn vanleg Poisson, og at dette høge talet på 0-observasjoner kjem frå to kjelder. Analysen er mykje lik Lambert [8] si anvending av modellen, og det er denne artikkelen forfattarane av ulykkesanalysen hentar tankane sine frå. Det vert innsamla tal på ulykker for kvar vegstrekning i studiet, og dei ulike vegstrekningane inngår i

ein av to stadier. I det eine er strekningane heilt trafikksikre, og det er garantert at det ikkje skjer nokon ulykker. I det andre stadiet er vegane usikre, og ulykkesfrekvensen følgjer vanleg Poissonfordeling. 0-observasjonane i dette stadiet kan kome av at fleire av ulykkene ikkje er store nok til å bli rapportere inn til myndighetene. I tillegg kan det vera mange nesten-ulykker som vert korrekt rapportert som ikkje-ulykker, men som likevel indikerer ein farleg vegstrekning. Ein strekning kan difor vera farleg sjølv om det ikkje har vorte rapportert om nokre ulykker. Dette fører til ei stor mengd 0-observasjonar. Forfattarane poengterer at ZIP skil mellom dei to stadia, og begrunner valet av modellen godt. Artikkelen gir likevel grunn til å fatte mistanke om forvirring angåande ZAP. Dei omtalar ZAP som ei hovudgruppe med modellar som skil mellom dei to kildene 0-rapporteringa kjem frå, der ZIP er ein modell som inngår i denne ZAP-gruppa. Men det er akkurat dette synet på 0-ane som *skil* ZIP og ZAP. ZIP vert i artikkelen omtala som ein versjon av ZAP, og sjølv om ein, som vist i delkapittel 3.1.6, kan sjå på ZIP som ein reparametrisert modell av ZAP, er dette i faglitteraturen tradisjonelt omtala som to forskjellige og likestilte modeller. Forfattarane viser i alle høve at dei ikkje veit at ZAP behandlar alle 0-observasjoane som *ei* gruppe.

Sjøv om val av modell i mange studier er teke på korrekt grunnlag og med god kunnskap, viser døma over ein negativ tendens ved praktisk anvending av ZIP og ZAP. Det virkar som det er forvirring og lite kunnskap om modellane, og at fagpersonane ikkje alltid kjenner til både ZIP og ZAP. Dette gjeld nok særleg ZAP. Mykje av grunnen til dette er truleg at artikkelen til Lambert [8] førte til ei stor interesse for ZIP, medan ZAP havna litt i bakgrunnen. Det er difor mogleg at fleire av forskarane med lite kunnskap om modellane vel ZIP ut frå at den er mest brukt, og kanskje slik framstår som den mest anerkjente av dei to modellane. I tilfella der ZAP er valt uten god nok kunnskap kan det kome av at det er ein enklare modell enn ZIP. Det er i det heile lite litteratur om korleis ein på førehand bør velgje mellom ZIP og ZAP, og konsekvensar val av gal modell gir. Det skal seiast at bruken av modellane i fleire tilfelle kan vera meir bevisst enn kva som kjem fram i artiklane, sidan artiklane ofte ikkje er skrivne med fokus på metodeval og analysemetoden. Ein ser likevel klar grunngiving for dei kritiske tankane me her presenterer.

Som regel vert avgjersla om kva modell som er den beste i dei praktiske situasjonane teke ved at analysen vert utført med både ZIP og ZAP, og at ein i etterkant fastsett kva modell som gav best treff opp mot datasettet. Dette skaper mykje arbeid. Ellers vert ofte bruken av ZIP begrunna med at 0-ane kjem frå to kjelder, men det er ikkje alltid ein kan sjå dette ut frå analysesituasjonen. Det er også i arbeidet med denne oppgåva funne mange

tilfelle av at ZAP estimerer betre for eit ZIP-fordelt datasett, og omvendt. Kva dette kan kome av, og om valet av modell bør takast på eit anna grunnlag enn om nullane kjem frå ei eller to kjelder, skal me no prøve å finne ut i analysedelen av oppgåva der me tek modellane i bruk i praksis. Me ynskjer også å finne ut kor ulike estimat modellane gir for det dame datasettet. Det er svært lite fagstoff å finne som omhandlar dette, og artiklar om emnet er til tider vage og nokre gonger også ukorrekte.

Kapittel 6

Praktisk analyse av ZIP og ZAP som regresjonsmodellar

6.1 Praktisk analyse av ZIP og ZAP

Me har til no sett på ZIP- og ZAP-modellane utfrå artiklar og anna teoretisk fagstoff. Det er no ynskjeleg å samanlikne dei to modellane gjennom ein meir praktisk og analytisk innfallsvinkel. Det kan tenkjast at sjølv om dei to modellane brukar ulike likelihoodfunksjonar, vil maksimeringa av funksjonen likevel gi svært like verdiar for estimeringa av regresjonskoeffisientane. Me ynskjer difor å finne ut om val av modell i praktisk anvending har noko å seia for estimeringa av koeffisientane og stabiliteten til estimata, og i tilfelle kva konsekvensane er ved val av feil modell. Sjølv om ZIP kun klarar å modifisere nullsannsynet med hensyn til for "mange" nullar, ynskjer me å samanlikne modellane også i tilfelle med lågare nullsannsyn enn det vanleg Poisson forutser. Det er interessant å sjå kor mykje betre ZAP er enn ZIP i desse situasjonane.

Ved å sjå på artiklar om utførte studium ser ein at valet mellom ZIP og ZAP ofte vert teke ut frå ei meining om korleis 0-observasjonane i datasettet har oppstått. Dette bygg på dei grunnleggjande tankane bak modellane, og ein forutset at det er modellen med same fordeling som datasettet som vil vera den beste. Det er likevel ikkje alltid i praktisk analysearbeid klart om 0-observasjonane kjem frå ei eller to kilder, og kva som er den korrekte fordelinga til datasettet. I tillegg har testkøyringar vist at det ikkje alltid er den korrekte modellen som gir estimat nærast dei korrekte verdiane for dei fire regresjonskoeffisientane. Me ynskjer difor å utføre ei grundigare analyse for å finne ut om ein bør ta valet mellom ZIP og ZAP-modellane på eit

anna grunnlag enn fordelinga i datasettet. Det er naturleg å tenkje seg at kor bra modellane estimerer kan ha samanheng med visse kombinasjonar av regresjonskoeffisientane, samt vera avhengig av forventninga til Poissondelen og det samla nullsannsynet i datasetta. Me ynskjer difor å finne ut om strukturelle tendensar i datasettet kan avgjere kven av dei to modellane ein bør velje for å få mest korrekte estimat. Ein viktig del av analysen vil vera å finne ut for kva verdiar av dei strukturelle faktorane ZIP og ZAP estimerer svært ulikt, og for kva verdiar valet av modell er av mindre betydning. Det vil vise konsekvensane av valet av modell i ulike praktiske situasjonar, men også fortelje oss kva som gjer at modellane er svært like i nokre tilfeller, og svært ulike i andre.

Ut frå modellgrunnlaget til ZIP ser me at dersom me har same forklaringsvariabel for dei to komponentane i modellen, kan ein får tilfelle der variabelen trekker i motsatt retning for det samla nullsannsynet og forventninga til Poissondelen. Eit døme på dette er når høg verdi av forklaringsvariabelen gir både mange nullobservasjonar og høge verdiar blant observasjonane over null. Dette kan truleg gjere estimeringa til ZIP ustabil, og det vert i analysen sett på kor godt ZAP kan konkurrere med ZIP i desse situasjonane.

Det er viktig å ta med i vurderiga av val av modell at ZIP og ZAP utfører ein noko ulik analyse av datasettet. ZIP fortel oss kva innverknad forklaringsvariablane i den binomske delen har på talet av strukturelle nullar, medan koeffisientestimata for Poissondelen viser kva innverknad dei andre forklaringsvariablane har på storleiken til resten av utfalla. Til samanlikning analyserer ZAP det eine settet med forklaringsvariablar opp mot 0-observasjonane og det andre settet mot alle dei andre utfalla. Resultata modellane gir ut fortel oss difor ikkje heilt den same informasjonen om datasettet. Det er likevel som regel betre å bruke ein modell med pålitelge svar, som gir svar på noko litt annan enn me opprinneleg er ute etter, enn ein metode der me ikkje kan stole på resultata.

Me ser i analysedelen også på situasjonar der talet på 0-observasjonar er for få til at vanleg Poissonregresjon er eit godt val. Sjølv om det av ZIP og ZAP kun er den sistnemnte som kan tilegne det samla nullsannsynet mindre verdi enn det vanleg Poissonfordeling ville gjort for det same datasettet, ynskjer me å sjå kor godt ZIP klarar å konkurrere med ZAP i desse tilfella. Me tek difor også med slike situasjonar i analysen, men dei utgjer kun ein liten del av grunnlaget for analysen vår, og der dei ikkje er relevante er dei ikkje teke med i verken presentasjonen av resultata eller grunnlaget for tolkninga av resultata.

Oppgåva vidare er lagt opp slik at ein først får ei innføring i den generelle metoden for analysearbeidet, begrunning av metoden og spesifisering av forutsetningar og val. Det vil også verta gitt ei innføring i praktisk implementering av programvarene som vert brukt for regresjonsutføringa i oppgåva. Deretter går me gjennom dei ulike underpunkta i analysen. Lesaren får då først ei innføring i metode og analysegrunnlag for det spesifikke punktet, før dei tilhøyrande resultat vert presentert. Det vert då lagt hovudvekt på tendensar me finn som er viktige for sjølve tolkningsarbeidet. Det er også lagt med tabellar og diagram som i tillegg viser meir detaljerte resultat. Den samla analysedelen vert så avslutta med ei felles tolking av alle resultata, og til slutt ein kort konklusjonen på hovedspørsmåla me har stilt i oppgvåva.

All genering av datasett og regresjonssimuleringar er utført i `R` versjon `2.9.2`.

Innføring av nokre nyttige begrep og forkortingar brukt vidare i oppgåva

Vidare i oppgåva vil omgrepet best estimat verta nytta om det estimatet med minst tilhøyrande MSE. Begrepet den beste modellen vil sameleis verta brukt om modellen som gir det beste estimatet. Modellen som har utganspunkt i same punktsannsyn som fordelinga til datasettet me utfører regresjon på, vil ofte verta omtalt som den korrekte modellen i situasjonen, medan den andre vil verta omtalt som ukorrekt modell.

Dei forklaringsvariablane og regresjonskoeffisientane som i ZIP- og ZAP-modellane inngår i forventninga til det binære utfallet med omsyn til nullane vil verta kalla den binomiske delen av regresjonsmodellane. Sameleis vil forklaringsvariablane og regresjonskoeffisientane som inngår i forventninga til den trunkerte Poissonfordeling i ZAP-modellen, og den vanlege Poissonfordelinga i ZIP-modellen, verta omtalt som Poissondelen av modellane.

Regresjonskoeffisienten til interceptleddet i den tilpassa funksjonen vert ofte forkorta med koef. int. Koeffisienten til forklaringsvariabelen vert forkorta koef. fokl. Dette gjeld både for binomisk og Poisson-del av regresjonsmodellane.

6.2 Metode og analysegrunnlag

Målet med analysearbeidet er å sjå nærare på, no med praktisk innfallsvinkel, kor ulike ZIP- og ZAP-modellane er og i kva tilfeller dei er svært ulike eller like. Me ynskjer særleg å sjå på kor ulik estimeringa er for den korrekte modellen i forhold til den ukorrekte modellen. Til dette treng me mange ZIP-fordelte datasett og mange ZAP-fordelte datasett. På desse vert det utført både ZAP- og ZIP-regresjon, og resultata vert etterpå samanlikna. Me ynskjer å sjå på alle "naturleg" situasjonar. Det vil seie kombinasjonar av verdiane for dei fire regresjonskoeffisientane som gir nullsannsyn og forventnng i Poissondelen av modellane i tråd med praktiske situasjonar modellane tradisjonelt vert bruk i.

Ved bruk av modellane i vanleg forskningsarbeid er oftast regresjonskoeffisientane til forklaringsvariablane dei mest interessante sidan det er dei som avgjer kor stor innverknad forklaringsvariabel har på responsvariabelen. I denne oppgåva ynskjer me å gå ned på eit meir statistisk-teoretisk plan. Me ser difor på regresjonskoeffisientane for både intercept-leddet og forklaringsvariablane for begge komponentane i modellane. Det er likevel i nokre delar av analysen fokusert mest på koeffisientane til forklaringsvariablane. Dei fire estimata vert analysert kvar for seg, men også gjennom ulike samansetningar.

Datasetta me utfører regresjon på er anten ZIP- eller ZAP-fordelte. Dei korrekte verdiane for regresjonskoeffisientane ligg i strukturen for den sanne fordelinga ved at dei er innført som ein del av forventningane i både Poisson- og binomisk del av punktsannsynet. Når me utfører regresjon på datasettet er det kun den korrekte modellen som nyttar heilt riktig likelihoodfunksjon for å estimere koeffisientane. Den andre modellen brukar feil metodegrunnlag i forhold til forutsetningane for responsvariabelen. Me ynskjer å finne ut kor godt modellane klarar å finne tilbake til dei korrekte verdiane som ligg i strukturen i datasettet, og slik gi gode estimat. Me ynskjer særleg å finne ut om modellen som bruker feil likelihoodfunksjon klarar dette like godt som den korrekte modellen. Hovudgrunnlaget i analysen vår vert å utføre ZIP- og ZAP-regresjon i ulike praktiske situasjonar, for både ZIP- og ZAP som korrekt modell. For å samanlikne resultata frå dei ulike regresjonskøyringane har me valt å samanlikne modellane gjennom estimata til regresjonskoeffisientane. Det er estimata som ligg til grunn for det praktiske resultatet i regresjonsanalyser. Kor korrekte desse er avgjer om styrken og p-verdien modellane gir ut er pålitelege.

6.2.1 Metodebeskriving

Generering av datasett

For å kunne måle kor godt ZIP og ZAP-regresjonsmodellane klarar å estimera regresjonskoeffisientane treng me datasett der me veit dei korrekte verdiane for desse parametrane, i tillegg til den sanne fordelinga til datasettet. Det er difor til bruk i analysearbeidet laga ein funksjon i `R`, kalla `genererdatasett()`, som genererer det ynskja datasettet for oss for både ZIP og ZAP. (Sjå B.1 for programkode). Som parametrar til funksjonen vert det lest inn satte korrekte verdiar for regresjonsparametrane, samt kor mange observasjonar me vil ha i datasettet. I heile analysen, bortsett frå der noko anna er kommentert, er det brukt 50 observasjonar. Me må også gi inn ein vektor med verdiane til forklaringsvariablane. I denne oppgåva er det valt å bruke kun èin forklaringsvariabel, og den er gjort identisk for både Poissondel og binomisk del. Det er valt fem ulike verdiar for forklaringsvariabelen, og for kvar at dei fem verdiane er det genenerert ti observasjonar. Heile datasettet på 50 observasjonar består altså av fem identisk fordelte små datasett på ti observasjonar kvar. Det er lik mengd observasjonar for kvar verdi for forklaringsvariabelen, og verdiane er også symmetriske om 0. Dette er sjeldant tilfellet i praktiske situasjonar, men er her valt for at me til dømes skal kunne lese ut det samla nullsannsynet og forventninga til Poissondelen for heile datasettet på 50 observasjonar. Dette er naudsynt for å kunne finne ut om ulike kombinasjonar av liten eller stor forventning og lågt eller høgt nullsannsyn i datasettet har innverknad på differansen i ZIP- og ZAP estimata.

For sjølve genereringa av datasetta brukar me funksjonane `rzipois()` og `rzapois()` frå pakken VGAM i `R`. Funksjonane tek som parametrar forventning for Poissondelen, i tillegg til det samla sannsynet for 0-observasjonar for `rzapois()` og sannsynet for strukturelle nullar for `rzipois()`.
`genererdatasett()` reknar difor ut dei nødvendige parametrane som trengst i `rzipois()` og `rzapois()` ut frå verdiane til kovariatvektoren og dei korrekte regresjonskoeffisientane, før sjølve genereringa finn stad. Funksjonen `genererdatasett()` lagar til slutt eit datasett med to kolonner, ei med dei 50 tilfeldige observasjonane, og ei med dei tilhøyrande verdiane for forklaringsvariabelen. Desse datasetta vert så grunnlaget for det vidare analysearbeidet.

6.2.2 Regresjonskøyring og uthenting av estimat og resultat frå regresjonskøyringa

Det er også for arbeidet med regresjonskøyringane og uthenting av nødvendige verdiar frå dei tilpassa objekta laga eit dataprogram for bruk i *R*. (Sjå B.2 for programkode). Me skal samanlikna estimatverdiar for dei to modellane for ulike parameterkombinasjonar, og for både ZIP og ZAP som korrekt og ikkje korrekt modell. I programmet fastset ein først verdiar for alle regresjonsparametrane. Deretter vert det med bruk av funksjonen `genererdatasett()` laga eit ZAP-datasett med utgangspunkt i dei korrekte verdiane til dei aktuelle parametrane. På dette datasettet vert det så utført både ZIP- og ZAP-regresjon, og det vert henta ut verdiar for dei fire regresjonskoeffisientane for kvar av dei to modellane. Det vert så generert eit ZIP-fordelt datasett, som me også køyrer både ZIP- og ZAP-regresjon på. For at me skal vite at tendensane me finn for parameterkombinasjonane ikkje kun kjem av vilkårlegheit, utfører programmet genering av datasetta og regresjon 10 000 gonger for kvar parameterkombinasjon. Det vil seie at me får 10 000 estimatverdiar for kvar koeffisient frå både ZIP- og ZAP-regresjonen, og for både ZIP og ZAP som korrekt datasett.

For å få eit samla tal på kor bra estimata frå alle 10 000 simuleringane er, bruker ein storleiken gjennomsnittleg kvadrert avvik, MSE (mean squared error). Det er eit mål på differansen mellom ein estimert verdi og den sanne verdien til storleiken som vert estimert. For estimatet $\hat{\theta}$ er den definert som

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

Den gir oss forventninga til alle avvika kvadrert, og kan i statistisk modellering verta sett på som eit mål på kor bra ein modell predikerer eit datasett. I analysen vår bruker me ein estimator for den teoretiske målestorleiken MSE. For å lettare skilje dette estimatet frå sjølve estimata til regresjonskoeffisientane, vil me i resten av oppgåva omtala estimatverdien for MSE som vanleg MSE. Estimatoren for MSE-verdiane i analysen vår er

$$\text{MSE} = \sum_{\text{alle observasjonar}} (\text{estimert verdi} - \text{observert verdi})^2$$

Me finn for kvar simulering differansen mellom estimatet og den korrekte verdien og kvadrerer dette avviket. Summen av alle dei kvadrerte avvika vert dividert på mengda simuleringar. MSE tek ikkje berre omsyn til kor bra modellen gjennomsnittleg er for dei 10 000 simuleringane, men også stabiliteten til estimatverdiane. Dersom regresjonsmodellen er ustabil kan det

få store konsekvensar i praktiske situasjonar sidan ein då ikkje nyttar simuleringar men gitte datasett.

I tillegg til MSE-verdiane tek også programmet vare på verdiane til forventninga i Poissondelen, det samla nullsannsynet og den aktuelle parameterkombinasjonen. Det summerer også talet på simuleringar der ein ikkje fikk ut resultat frå regresjonsanalysen. Dette er tilfeller der korrelasjonsmatrisa for forklaringsvariblane er singulær slik at metoden ikkje klarar å kalkulere den inverse, eller der iterasjonen av parameterestimata ikkje konvergerer. Desse verdiane vert så i lag med tilhøyrande MSE-verdiar for begge modellane samla i to nye stort datasett. Det eine datasettet inneheld resultat henta frå ZIP- og ZAP-regresjon på ZIP-fordelte datasett, og det andre innheld resultat frå henholdsvis regresjon på dei ZAP-fordelte datasetta. Desse to nye datasetta som vert brukt vidare i analysen kallar me ZIP- og ZAP-datasetta. Dei har altså ikkje namn etter fordelinga i desse to nye datasettet, men etter fordelinga som ligg til grunn i datasetta det er utført regresjon på. Både ZIP- og ZAP-datasettet innheld resultat frå både ZIP- og ZAP-regresjon. Dei datasetta som regresjonen vert utført på vil ofte verta omtalt som observasjonsdatasetta, og fordelinga til responsvariabelen vert også omtalt som fordelinga til heile datasettet det vert utført regresjon på.

6.2.3 Vidare undersøkjingar på regresjonsresultata

Dei innsamla resultata frå regresjonsutføringane vert analysert vidare for å finne svar på det me spør om i oppgåva. Me ser først på kor bra modellane estimerer som korrekt regresjonsmodell opp mot den ukorrekt modellen. Då vert det både sett på om den korrekte modellen alltid er best, og kor stor forskjellen er for MSE-verdiane til dei to modellane. Deretter undersøkjer me den tilsvarande relative differansen. Me undersøkjer også om det er mogleg å finne nokon samanheng mellom den relative differansen i MSE-verdiane og, etter tur, verdiane til regresjonskoeffisientane, det samla nullsannsynet og forventninga til Poissondelen. Til slutt vert det undersøkt om større datasett vil gi til likare MSE-verdiar for ZIP og ZAP.

6.2.4 Kommentar til val av verdiar

Regresjonsparametrar

Dei korrekte verdiane til regresjonskoeffisientane er begrensa til intervallet -2,5 og 2,4. Regresjonskoeffisientane styrer sannsynet for talet på 0-observasjonar, samt forventninga til Poissondelen. Dei ytre rammene for samla

nullsannsyn er satt til 0,2 og 0,81, og 1 og 12,2 for forventninga for Poissondelen. Krava er innført for å halda analysen innanfor det ein kan anta som "vanlege" praktiske situasjonar der ZIP- og ZAP-modellane vert nytta. Eit datasett med kun 50 observasjonar og mengd nullar på over 80% er ikkje ein "naturleg" analysesituasjon med Poisson som grunnfordeling. Det same gjeld dersom forventninga vert for høg. Desse situasjonane fell difor utanfor avgrensingane av analysen gitt i innleiinga av oppgåva. Testkøyringar har vist at i tilfelle med forventning i Poissondelen under 1, er det svært ofte ein ikkje får ut resultat av regresjonsanalysen for begge modellane. Det same gjeld med nullsannsyn under 0,2. (Grunnen for dette vil kome fram i tolkningsdelen av oppgåva.) Me har difor valt å bruke parameterkombinasjonar som gjer at nullsannsyn og forventning i Poissondelen ligg innanfor dei gitte krava. Sjølv om det samla nullsannsyn alltid er over 0,2 vil det saman med låg forventning til Poissondelen gi datasett med færre 0-observasjonar enn det vanleg Poissonfordeling forutset, som er ein del av analysen i oppgåva. Det er brukt god variasjon i samansetjinga av koeffisientverdiane både generelt og innanfor kvar kombinasjon av høg, låg og middels forventning mot høg, låg og middels nullsannsyn. Det er også teke med mange tilfelle med ulike kombinasjonar av forteikna til koeffisientane til forklaringsvariabelen.

Verdiane til forklaringsvariabelen er satt til 1, -0,5, 0, 0,5 og 1. Dei noko låge verdiane er valt grunna at det gir konsentrert nullsannsyn og forventning i Poissondelen for dei ulike verdiane for forklaringsvariabelen. Dette gjer det lettare å finne ut kva praktiske situasjonar som gjer ZIP- og ZAP-modallane like og ulike. At forklaringsvariabelen er identisk for både Poissondelen og den binomiske delen, gjer at me lettare kan styre nokre delar av analysen og tolkinga av resultata. Til dømes kan me sjå på tilfelle for ZIP der variablen dreg i ulik retning for nullsannsyn og storleiken til observasjonane over null.

Storleik på datasetta

Antall observasjonar i datasetta det vert utført regresjon på kjem også av ynskjet om å sjå på "vanlege" praktiske situasjonar for ZIP og ZAP. Ofte har ein ikkje nok observasjonar tilgjengeleg for eit større datasett, og det er også grunn til å tru at det er i tilfelle med så små datasett at skilnaden på modellane vert størst, og difor i vår oppgåve er mest interessante. Det er likevel mot slutten av oppgåva utført analyse på datasett med 1000 observasjonar. Dette er kommentert der det gjeld.

6.2.5 Kommentar til dei to analysedatasetta

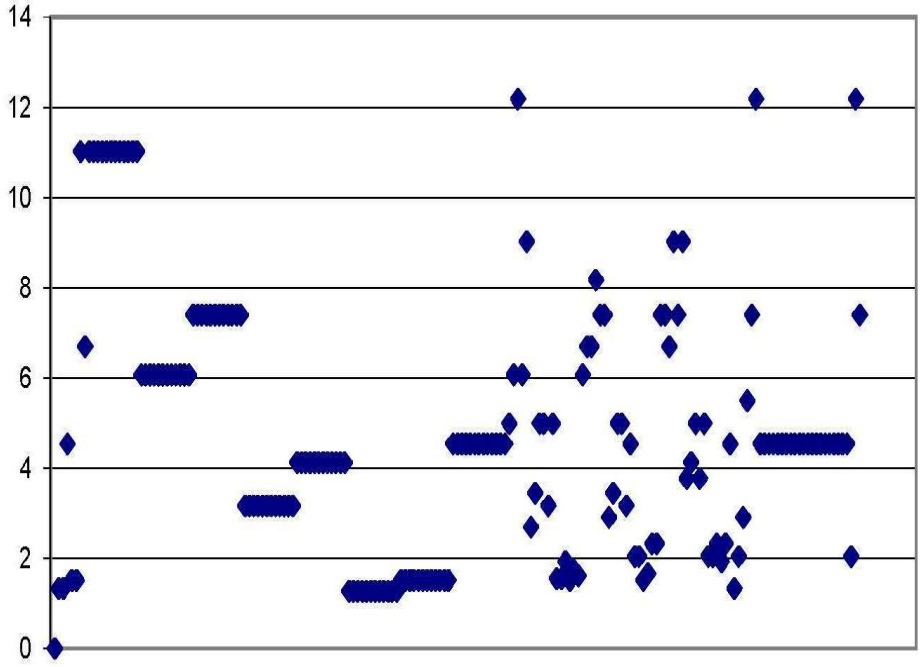
Det er i utgangspunktet dei same parameterkombinasjonane som ligg til grunn i både ZIP- og ZAP-datasettet. Å få to identiske ZIP- og ZAP-datasett som er heilt rettferdige å samanlikne er likevel umogleg. Då treng ein parameterkombinasjonar der både regresjonskoeffisientane, samla nullsannsyn og forventninga i Poissondelen er heilt lik for både ZIP og ZAP. Dette er umogleg sidan dei same koeffisientverdiane gir ulikt samla nullsannsyn og forventning i Poissondelen for fordelingane. Sjå kapittel 2.1. Dette gjer at ein ikkje kan samanlikne resultata for kombinasjonane frå ZIP- og ZAP-datasettet parvis. Ein kan heller ikkje ukritisk samanlikne ZIP- og ZAP som korrekte modellar, men ein kan sjå på kor stor forskjell det er for MSE-verdiane mellom korrekt og ukorrekt modell i dei to analysedatasetta. ZIP- og ZAP-datasetta er store datasett, og figurane 6.1 til 6.4 viser også at den generelle fordelinga med tanke på nullsannsyn og forventning er svært lik. Ein kan difor i dei fleste delane av analysen samanlikna dei to datasetta dersom ein ser på alle kombinasjonane under eitt. Der ein får svært like resultat for dei to analysedatasetta, er ZIP- og ZAP-datasetta slått saman i presentasjonen av resultata. Sjølve analysen er likevel alltid utført separat.

Opprinneleg var det like mange parameterkombinasjonar for begge dei to store analysedatasetta. Men i tilfelle der me for meir enn 300 av simuleringane ikkje fekk ut regresjonresultat for både ZIP og ZAP, vart kombinasjonen teke ut av analysen. Dette gjeld for fleire tilfelle for ZIP-datasettet enn ZAP-datasettet. Det er likevel nok kombinasjonar for begge datasetta til å utgjere eit solid analysegrunnlag. I alt har me 186 kombinasjonar i ZAP-datasettet, og 164 kombinasjonar i ZIP-datasettet. Det er ein klar tendens til at det er parameterkombinasjonar med forventning for Poissondelen ned mot 1 som ikkje klarar å gi ut resultat for nok simuleringar.

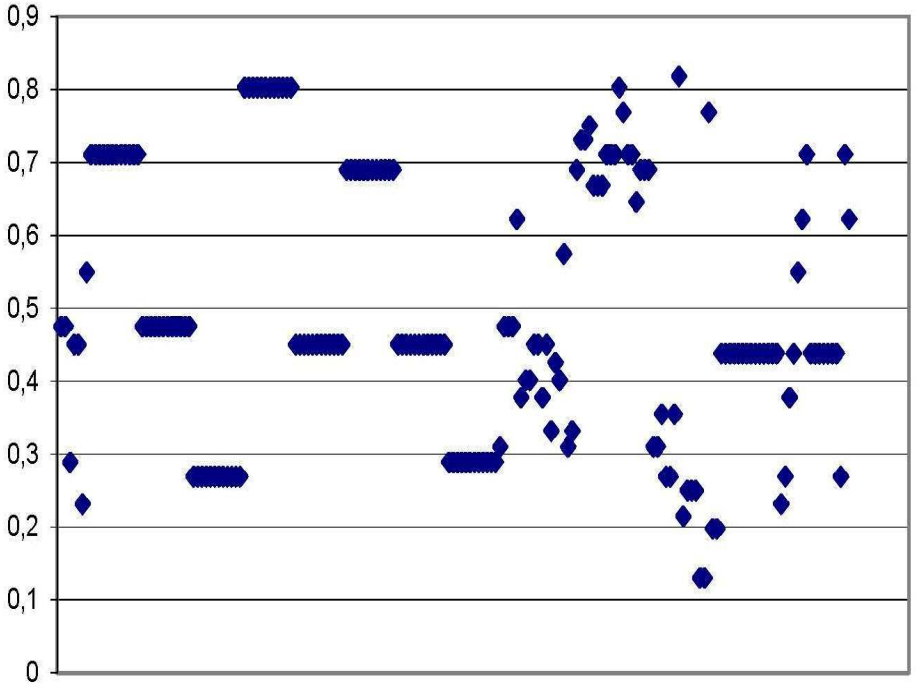
6.2.6 Merknader til visuell framstilling av resultata

I framstillinga av resultata i analysen er det nytta mykje diagram og spreiingsplott. Resultata består i dei fleste tilfella av mange verdiar, og ei visuell framstilling gir eit godt første inntrykk. Figurane er i lag med tabellar og skrfitleg framstilling også med på å gi ei god forståing ved grundigare studering.

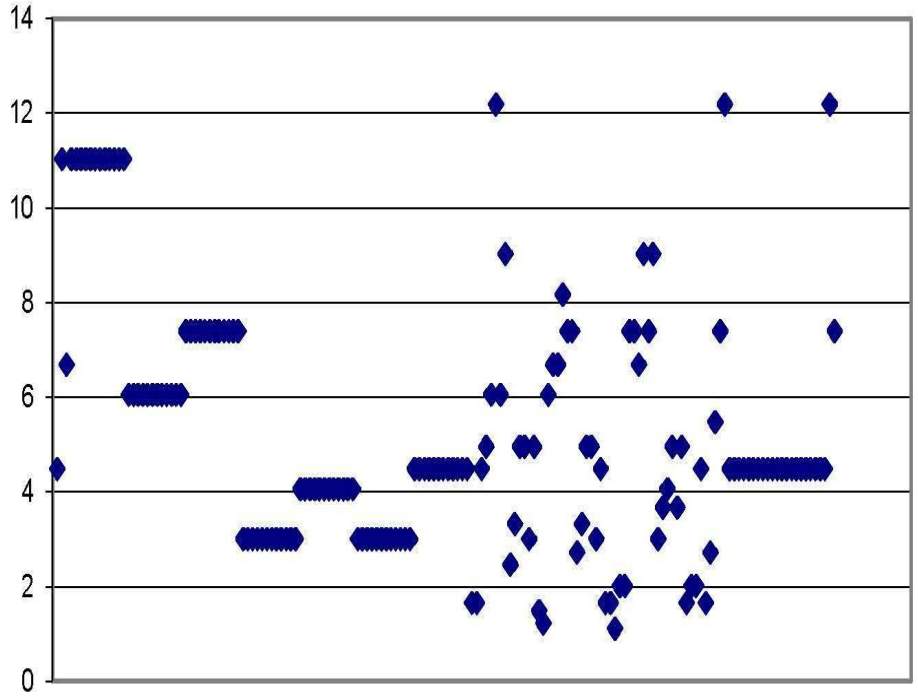
Det er også nytta ein del plotting av MSE-verdiar og relativ differanse for dei to modellane mot kvarandre, og opp mot andre storleikar i resultata. Dette er ein god måte å få fram eventuell tilknytning. Det er ellers brukt røyrdiagram, som på ein god måte viser både samla tal på aktuelle kombinasjonar, samt



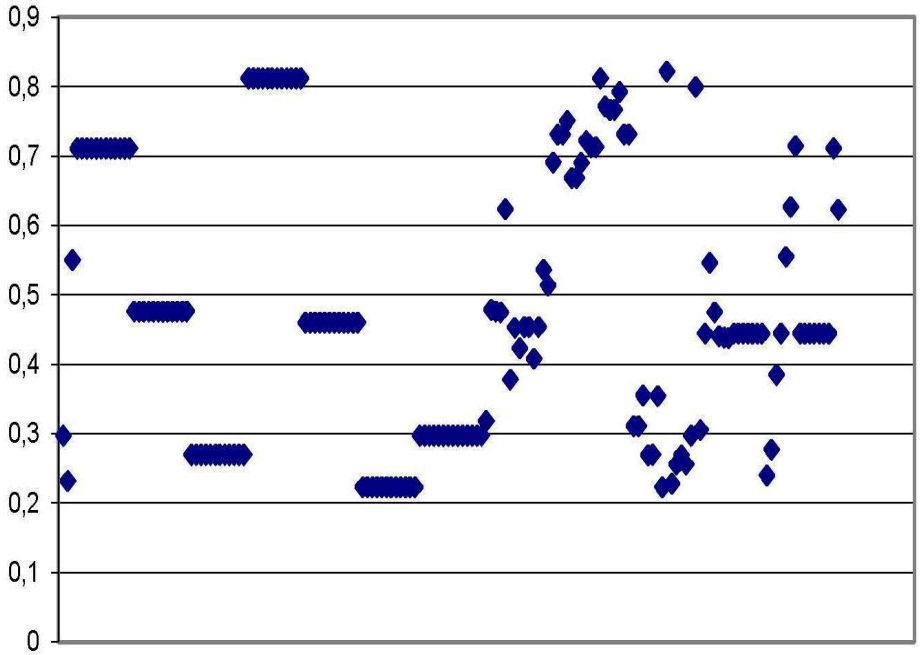
Figur 6.1: *Fordeling av forventning i Poissondelen for ZAP-analysedatasettet*



Figur 6.2: *Fordeling av samla nullsannsyn i ZAP-analysedatasettet*



Figur 6.3: *Fordeling av forventning i Poissondelen for ZIP-analysedatasettet*



Figur 6.4: *Fordeling av samla nullsannsyn i ZAP-analysedatasettet*

interessante fordelingar av dei.

Kapittel 7

Aktuell programvare i rekneverktøyet R

7.1 Praktisk implementering av programvarene brukt i regresjonen i analysen

For regresjonsutføringa i oppgåva har me nytta rekneverktøyet `R`, versjon `2.9.2`. `R` inneheld fleire pakkar som støttar regresjon med zero-inflated og zero-altered datasett. I dette kapittelet vil me sjå nærare på den praktiske implementering av regresjonsmodellane ZIP og ZAP i dei aktuelle programvarene i `R`. Artikkelen [17] gir saman med dei offisielle nettsidene til `R` [1] ei god innføring i metodegrunnlaget og moglege val for innstillingar i funksjonane pakkane tilbyr.

7.1.1 Pscl-pakken

Den mest brukte programvara for ZIP- og ZAP-regresjon i `R` er pscl-pakken (Jackman 2008) med funksjonen `hurdle()` for å kalle zap-regresjon og `zeroinfl()` for å kalle zip-regresjon. Denne pakken er designa slik at både den predikerte funksjonen og det tilhøyrande returnerte objektet så langt som mogleg er gjort lik standarden for regresjonsobjekt i `R`. Ein hovudforskjell er at me for kvar av dei predikerte verdiane eit standard regresjonsobjekt inneheld, får ut to verdiar for `hurdle()` og `zeroinfl()`, ein for kvar av dei to komponentane i modellane. Sidan pscl-pakken har gjort objekta så lik standarden i `R`, kan ein gjere bruk av flesteparten av dei mange tilleggsfunksjonane som er tilgjengelege for andre regresjonsobjekt, til dømes `print()`, `summary()`, `fitted()` og `residuals()`. Ein kan også utføre testar som `coeftest()`, `waldtest` og `lmtest()` på dei returnerte objekta. Dette

gjer pakken svært brukarvenleg, og den integrerer seg lett som eit funksjonibelt analyseverktøy i `R`. Det er denne pakken som er valt til bruk i denne analysen, og eg vil difor gi ei kort innføring av dei tekniske detaljane i `zeroinfl()` og `hurdle()`.

Teknisk oppbygging av zeroinfl() og hurdle()-funksjonane i pscl-pakken

Den tekniske oppbyggina er svært lik for `zeroinfl()` og `hurdle()`. Nokre forskjellar finn ein likevel, grunna at funksjonane har utgangspunkt i to ulike fordelingar og likelihoodfunksjonar. Grunntanken i sjølve utføringa av regresjonen er derimot svært lik, og det same gjeld oppbygginga av det tilpassa objektet som inneheld resultata frå regresjonsanalysen. Både `hurdle()` og `zeroinfl()` støtter bruk av forklaringsvariablar for begge komponentane i modellen. I `hurdle()` kan ein spesifisere kva linkfunksjon ein vil nytte for kvar av dei to kovariatmatrisene, medan for `zeroinfl()` er linkfunksjonen for regresjonen på det binære utfallet fastsatt til logit. `zeroinfl()` tillet også kun binomisk fordeling for modelleringa av struktuelle nullar mot resten av observasjonane. For `hurdle()` kan ein velje mellom binomisk, høgresensurert Poisson og negativ binomisk fordeling for den binære modelleringa.

For både `zeroinfl()` og `hurdle()` er det gjennom argumentet formula ein angir kva forklaringsvariablar ein ynskjer å teste for korrelasjon i dei to delane i regresjonsmodellen. Argumentet er på forma $y \sim x1+x2|z1+z2$ der forklaringsvariablane bak det loddrette skiljet vert brukt for predikering av forventning til det binære utfallet, medan forklaringsvariablane framfor vert brukt til å predikerer forvetninga til resten av utfalla. Talet på forklaringsvariablar treng ikkje vere det same for begge forvetningane, og kovariatmatrisene kan vera identiske eller ulike.

For `hurdle()` kan ein velje om dei to komponentane i likelihoodfunksjonen skal verta maksimert separat eller som ein samla funksjon. På grunn av kryssleddet i (3.6) er ikkje dette eit val for `zeroinfl()`. For begge funksjonane vert regresjonskoeffisientane funne ved maksimering av likelihoodfunksjonen ved bruk av Quasi-Newton metoden BFGS. Brukar kan velge å gi inn startverdiar til iterasjonsprossessen. Dersom dette ikkje er gjort vert startverdiane funne ved hjelp av metoden "iteratively reweighted least squares", som vert kalla ein gang for kvar av dei to delane i modellen. For `zeroinfl()` kan ein velje at programmet skal bruke EM-algoritme for maksimeringa av likelihooden. Dersom det ikkje vert lese inn startverdiar, vil EM-algoritmen

bruke dei same startverdiane som "iteratively reweighted least squares". EM-algortimen held fram til det vert oppnådd konvergering, eller det vert nådd ei satt øvre grense for tal på iterasjonar. Default er satt til 10 000. Sjølv om EM algoritmen maksimerar likelihoodfunksjonen, vert BFGS metoden i tillegg kalla ein gang til slutt for å kalkulere kovariansmatrisa. Dersom ein vel bort EM-algoritmen byrjar BFGS-metoden med startverdiane som er funne med "iteratively reweighted least squares". Den set også den uobserverte indikatorvariabelen som avgjer struktuell 0 mot dei andre utfalla lik for alle 0-observasjonane. For både `hurdle()` og `zeroinfl()` vert standardavvika til dei estimerte koeffisientverdiane finne ved hessianske matrisa. Der det er tillete å setje eigne val for funksjonen, tildømes maksimeringmetode, øvre grense for tal på iterasjonar og startverdiar, vert desse fastsett med funksjonen `hurdle.control()` og
`zeroinfl.control()` som er parametrar i hovedfunksjonane.

7.1.2 Andre tilgjengelege pakkar

Det er også tilgjengeleg andre pakkar i `R` som støttar zero-inflated regresjon. ZIGP (Erhardt 2008) tillet innføring av forklaringsvariablar for overdispersjon i vanleg forventning, i tilegg til overdispersjon med hensyn til nullobservasjonar. Ein annan tilgjengeleg pakke er zicounts (Mwalili 2007). Diverre skil grensesnittet på begge desse pakkane seg frå standarden i `R` og tilbyr liten grad av ekstra funksjoner til bruk på dei returnerte objekta. Det er difor tungvint å utføre vidare analyse. Dei to pakkane gamlss (Stasinopoulos og Rigby 2007) og VGAM (Yee 2008) retunerer objekt som er meir tilpassa standarden i `R`, og sistnemnde pakke tilbyr også zero-altered regresjon. Ulempa med desse pakkane er at dei kun tillet eitt sett med forklaringvariablar.

Dei fleste pakkane, også pscl, tillet negativ binomisk og geometrisk fordeling i tillegg til Poisson som val for hovedfordeling i regresjonen. For meir informasjon om pakkane og fleire tekniske detaljar vert lesar vist til dei ulike pakkane sine sider på det offisielle heimeområdet til `R` [1]

7.1.3 Val av innstillingar for metode i zeroinfl() og hurdle() i vår analyse

For maksimering av likelihoodfunksjonen har me valt å bruke default for både `hurdle.control()` og `zeroinfl.control()`. For `hurdlel()` er dette BFGS(Quasi-Newton), medan det for `zeroinfl()` er EM-algorimen. Valet

er teke ut frå at det i teoridelen i oppgåva kjem fram at algoritmane både er dei tradisjonelt mest anvendte og anerkjente for henholdsvis ZIP og ZAP. At me nyttar EM-algoritmen for ZIP-regresjonen er også begrunna med at testkøyringar viser at ein då får færre tilfeller uten regresjonsresultat enn med BFGS. Det er ofte i dei situasjonane ZIP og ZAP estimerer ulikt at vanleg BFGS ikkje klarar maksimeringa, og det er turleg desse resultata som er dei viktigaste i vår analyse. Sjølv om BFGS i nokre tilfeller gir noko riktigare estimat, gir dei to algroitmane oftast like verdiar for `zeroinfl()` , og samanlikna med estimatet frå `hurdlel()` er tendensen i MSE-differansen lik for både BFGS og EM-algoritmen.

Val av Poisson som hovedfunksjon i regresjonen er sjølvsagt sidan det er zero-inflated og zero-altered Poissonfordeling me ser på. Som fordelingsfunksjon for det binære utfallet brukar me binomisk, som tradisjonelt sett er den mest brukte, og som er den me hovudsakleg bruker i framstillinga av modellane i teoridelen av oppgåva. For linkfunksjonane bruker me log for forventninga i vanleg og trunkert Poisson, og logit for binomisk.

Kapittel 8

Resultat og detaljert analysemetode

Me vil no gå punktvis gjennom analysearbeidet i oppgåva med meir detaljert innføring i metoden, før dei tilhøyrande resultata vert presentert.

8.1 Ulike koeffisientestimat i praksis

8.1.1 Metode og analysegrunnlag

Sjølv om me i teoridelen har funne ut at ZIP og ZAP som regresjonsverktøy brukar ulikt metodegrunnlag, kan modellane likevel ved praktisk anvending gi så like estimat at val av modell ikkje er av stor betydning. Det er difor naturleg å starte analysearbeidet med å bekrefte at val av modell faktisk har betydning. Me plottar parvis MSE-verdiane for ZIP-modellen mot MSE-verdiane for ZAP-modellen for alle parameterkombinasjonane. Dette gjer me for alle dei fire regresjonskoeffisientane kvar for seg. Dersom modellane har tendens til å estimerer like verdiar, vil punkta i plotta ligge rundt ei rett linje med stigningstal ein. Ein kan også samanlikne dei marginale fordelingane til MSE-verdiane ved å sjå på spreiinga til punkta på dei to aksane. Dette gir oss eit førsteinntrykk av den generelle differansen i MSE-verdiane til dei to modellane. I tillegg vil plotta også vise eventuelle mønster i differansen i MSE for dei to modellane.

8.1.2 Resultat

Ingen av figurane 8.1 til 8.4 har punkt som fell rundt ei rett linje med stigningstal ein. Alle dei fire koeffisientane viser derimot ein tendens til at

punkta ligg nær ein av dei to aksane. Der den eine modellen har svært høg MSE-verdi har den andre relativ låg MSE-verdi. Alle dei fire regresjonskoeffisientane får for dei fleste parameterkombinasjonane likevel MSE ned mot origo for begge modellane. Av utsnitta av plotta ser me at for MSE-verdiar under 0,2 for Poissondel og under 0,3 for binomsk del har punkta også tendens til å falle på eit noko rett linje med stigningstal ein.

Av dei marginale fordelingane til MSE-verdiane ser me at ZIP generelt har noko lågare verdiar enn ZAP for Poissondelen av modellane, sjølv om maksimumsverdien er høgare for ZIP enn ZAP for koef. forkl. For den binomiske delen ligg dei aller fleste MSE-verdiane til ZIP høgare enn for ZAP, ofte svært mykje høgare. ZIP har maksimumsverdi på 543,5 for koef.int og 775,2 for koef.forkl. Tilsvarande verdiar for ZAP er under 6,5. Sjølv om dei fleste parameterkombinasjonane gir MSE under 2 for Poissondel og 3 for binomisk del for begge modellane, er dette noko høge tal i regresjonssamanheng.

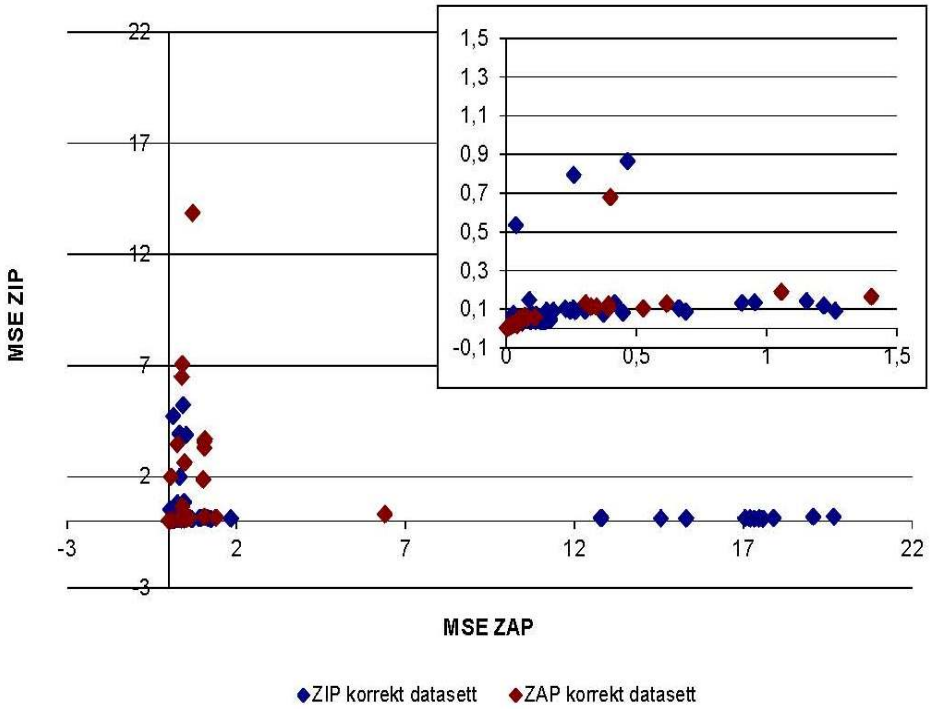
8.1.3 Oppsummering

ZIP og ZAP som regresjonsmodellar estimerar i mange tilfelle ulike verdiar for regresjonskoeffisientane, og det er ofte svært stor differanse i MSE-verdiane for modellane. Dette gjeld særleg for den binomiske delen av modellane. For Poissondelen ser me ein tendens til at ZIP generelt har noko lågare MSE-verdiar, men for den binomiske delen av modellane har ZAP klart betre MSE-verdiar enn ZIP.

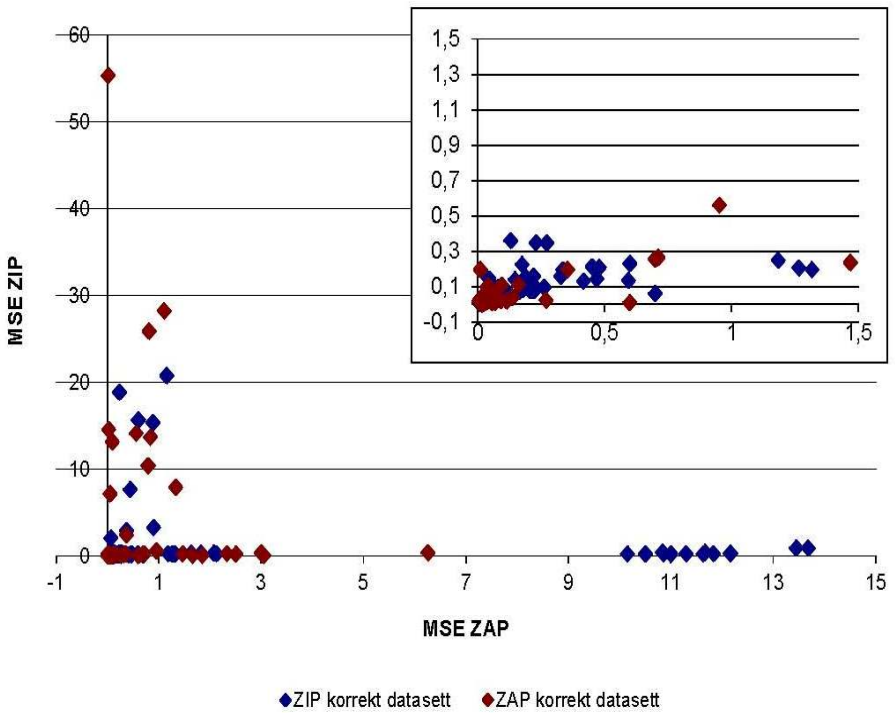
8.2 Konsekvens av korrekt fordeling og analyse av Poisson og binomsk del

8.2.1 Metode og analysegrunnlag

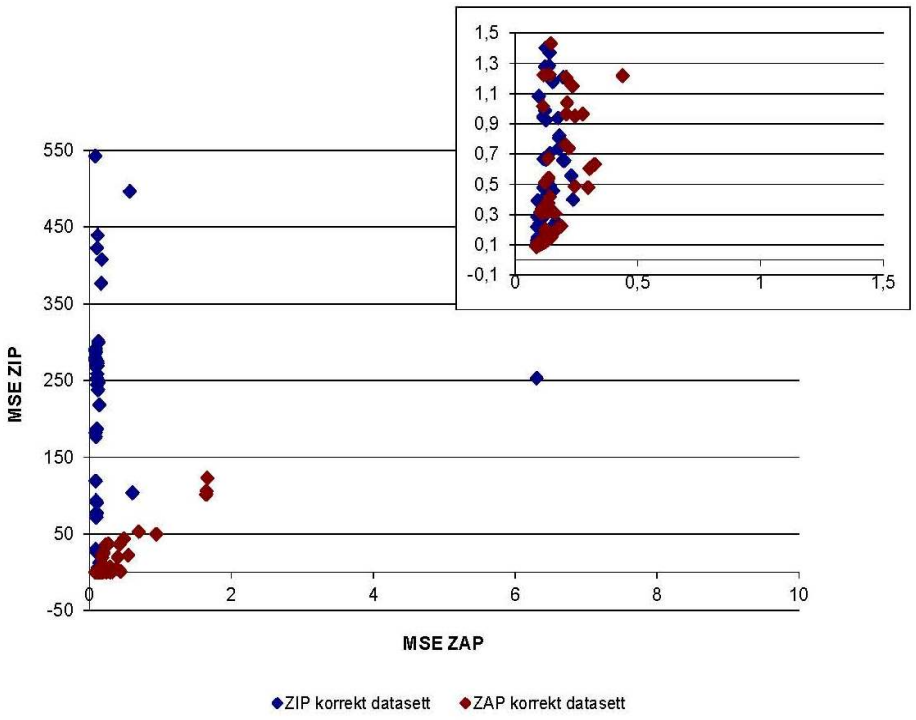
Me har no funne ut at valet mellom ZIP og ZAP som regresjonsmetode har betydning for estimeringa av regresjonskoeffisientane. Det er naturleg å tenkje seg at det hovudsakleg er den korrekte modellen som gir dei mest korrekte koeffisientestimata. Men analysen så langt viser ein tendens til at ZAP er ein generelt betre modell enn ZIP for den binomiske delen, medan modellane er meir like for Poissondelen. Sjølv om det i figurane 8.1 til 8.4 frå førre del av analysen er ulik mengd tilfelle frå ZIP- og ZAP-datasetta, gir resultata grunn til å tru at det ikkje alltid er den korrekte modellen som gir dei beste koeffisientestimata. Er dette tilfellet sår det tvil om kor pålitelege ZIP og ZAP eigentleg er som regresjonsmodellar. Men for denne oppgåva som er



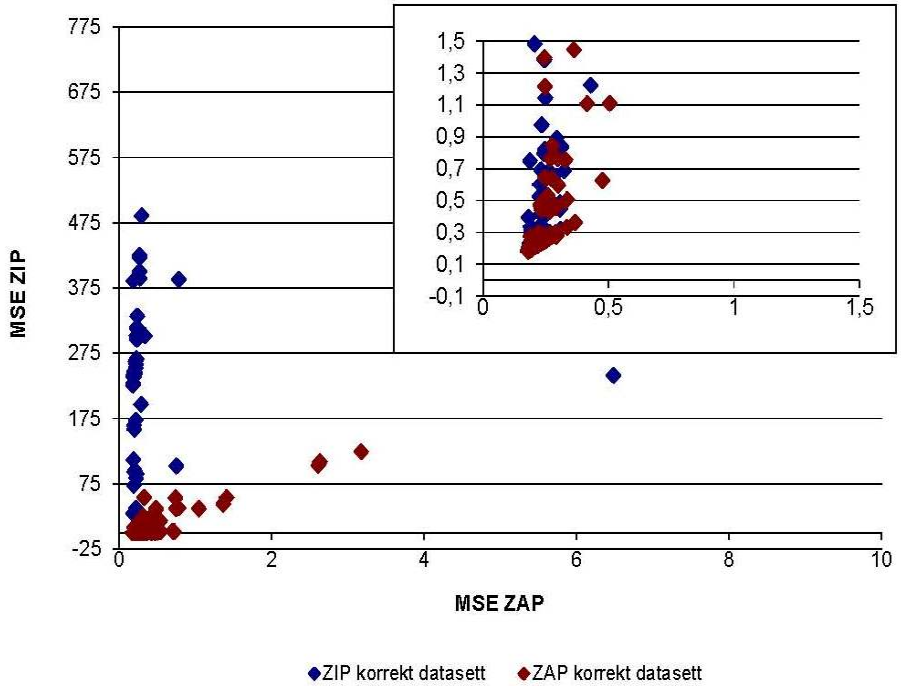
Figur 8.1: *MSE for koef. int. Poisson frå ZIP og ZAP-regresjon på same observasjonsdatasett plotta mot kvarandre*



Figur 8.2: *MSE for koef. forkl. Poisson frå ZIP og ZAP-regresjon på same observasjonsdatasett plotta mot kvarandre*



Figur 8.3: *MSE for koef. int. binomisk frå ZIP og ZAP-regresjon på same observasjonsdatasett plotta mot kvarandre*



Figur 8.4: *MSE for koef. forkl. binomisk frå ZIP og ZAP-regresjon på same observasjonsdatasett plotta mot kvarandre*

avgrensa til å samanlikne ZIP og ZAP med kvarandre, ikkje opp mot andre modellar, har det hovedsakleg innverknad på korleis me bør leggje opp det vidare arbeidet med analysen. Dersom ein antar at den korrekte modellen alltid gir lågaste MSE, kan det føre til svært gal konklusjon av modellane i praksis. Eit døme på dette er dersom det er stor differanse i MSE-verdiane for modellane og ZAP som korrekt modell feilaktig også vert sett på som den beste modellen. Då vil avstaden mellom estimatverdiane vera korrekt, men ZAP vil feilaktig verta sett på som ein mykje betre modell enn ZIP i den situasjonen, sjølv om det motsatte er riktig. Det er difor som eit neste steg i analysen naturleg å sjå på kor godt ZIP og ZAP estimerer parametrane for Poisson- og binomisk del som både korrekt og ukorrekt modell samanlikna med den andre modellen som korrekt og ukorrekt.

For nokre parameterkombinasjonar gir ZIP og ZAP så like estimat at det i regresjonssamanheng vil vera naturleg å kalla modellane like gode. Men kva ein legg i omgrepet like gode er relativt. Talet på tilfelle der me kan seie at valet mellom ZIP og ZAP som regresjonsmodellar ikkje har betydning, er altså avhengig av kva ein definerer som lik verdi i analysen vår. Me har difor valt å utføre samanlikningane av MSE-verdiane med fire ulike krav til kva ein ser på som like gode modellar. Me innfører då eit minstekrav til differansen mellom MSE for ZIP og ZAP for den konkrete koeffisienten. Kun dei tilfella der differansen er større enn minstekravet vert parameterkombinasjonen notert som ulike MSE-verdiar. For kvar parameterkombinasjon vil altså regresjonskoeffisientane verta kategorisert som ZIP best, ZAP best eller at modellane gir like gode estimat. Minstekrava for differansen er satt til 0, 0,0025, 0,01 og 0,25. Det tilsvarar absoluttverdiar for gjennomsnittlege avvik frå korrekt koeffisient verdi på 0, 0,05, 0,1 og 0,5. Med minstekrav på 0 ser ein på kva modell som er best, uavhengig av kor *mykje* betre den er. 0,1 kan sjåast på som det naturlege valet for å kalle modellane like i praksis, og det er i tillegg teke med to ytterpunkt for å utvide analysen. 0,0025 er eit ganske strengt krav, og ein kan tenkje seg at ein så liten differanse har lite å seie ved reint praktisk anvending av modellane, men i teoretisk statistisk samanheng er dette likevel interessant. Me gjer lesaren merksam på at sjølv om talet på tilfeller der ein modell er best avtek med stigande minstekrav, fører ikkje dette til fleire tilfelle for den andre modellen som best. Det er derimot fleire parameterkombinasjonar som vert kategorisert som at modellane har like MSE-verdiar. Tabellane for ZIP- og ZAP-regresjonen (t.d. 8.2.2 og 8.2.2), må difor sjåast i lys av kvarandre.

For å finne svar på kva modell som er best for dei to komponentane binomisk og Poisson, og om den korrekte modellen alltid er den beste, avgjer me

først for kvar parameterkombinasjon kven av dei to modellane som har lågast MSE-verdi. Me samanliknar så talet på kombinasjonar der ZIP er best med taetl på gonger ZAP er best. Dette vert gjort for dei fire regresjonskoeffisientane separat, og med både ZIP og ZAP som korrekt modell. Regresjonskoeffisientane vert analysert både kvar for seg og gjennom ulike samansetnadar. Det er til dømes ysnkjeleg å finne ut om den same modellen ofte er best for begge parametrane i den eine komponenten, eller kanskje også for alle fire koeffisientane. Då vil den modelllen vera den klart beste for den spesifikke parameterkombinasjonen. Me ser på resultata frå dei to store analyesdatasetta separat, men studerer også mønster på tvers av datasetta.

Auka av tilfelle som ved innføring av eit høgare minstekrav fell innanfor kategorien at modellane er like, fortel kor mange kombinasjonar som har differanse for MSE i intervallet mellom det gamle og det nye differansekravet. Det gir oss eit førsteinntrykk av kor ulike estimeringa til ZIP og ZAP er. Denne forskjellen ser me vidare på i neste steg i analysen.

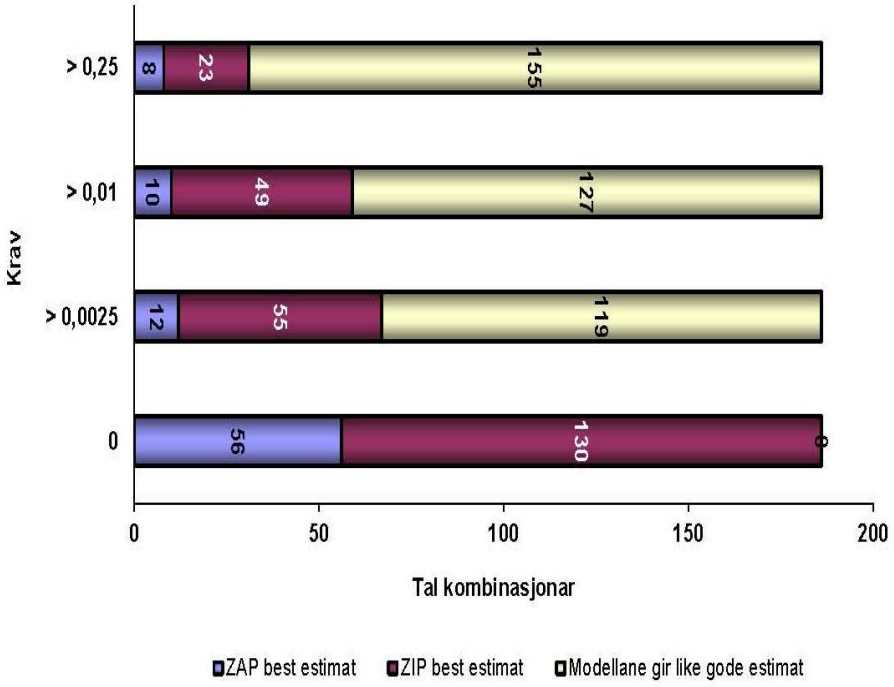
8.2.2 Resultat

Resultata me har funne er vist som tabellar og røyrdiagram under. Røyrdiagramma (figur 8.5 til 8.12) viser kor mange av parameterkombinasjonane som for dei ulike differansekrava til MSE er kategorisert som ZAP best, ZIP best og at modellane er like gode. Det er eit diagram for kvar regresjonskoeffisient, med ulike diagram for ZIP og ZAP som korrekt modell. Ut frå røyrdiagram 8.7 og 8.8 ser ein heilt tydeleg at ZAP som korrekt modell er svært god på den binomiske komponenten, særleg i forhold til ZIP på same observasjonsdatasett. Med minstekrav 0 er ZAP betre enn ZIP for 90,9% av kombinasjonane for koef. int og heile 96,2% for koef. forkl. Ved høgare differansekrav avtek talet på tilfeller der ein av modellane vert kategorisert som best, men sjølv med krav på 0,25 er ZAP best modell for over halvparten av kombinasjonane for begge parametrane i binomisk del. For ZIP ser me at allereie ved minstekrav på 0,0025 ikkje er modellen best for nokon kombinasjon. Dette er kanskje ikkje overraskande sidan observasjonsdatasetta for desse resultata er ZAP-fordelt. Men ser ein på ZIP-datasettet ser ein heilt klart den same tendensen. Sjå figur 8.11 og 8.12. Med minstekrav på 0 er ZAP som ukorrekt modell best for 88,4% for koef. int. og framleis 92,7% koef. forkl. Dette er overraskande med tanke på at observasjonsdatasetta her er ZIP-fordelt. Ser me på ZIP på binomsk del er den noko betre som korrekt enn ukorrekt modell. Men skilnaden er liten. Den kan kome av at regresjonen her er gjort på nettopp ZIP-fordelte datasett, men det kan også vera eit resultat av at ZIP og ZAP datasetta ikkje er heilt rettferdige å samanlikna.

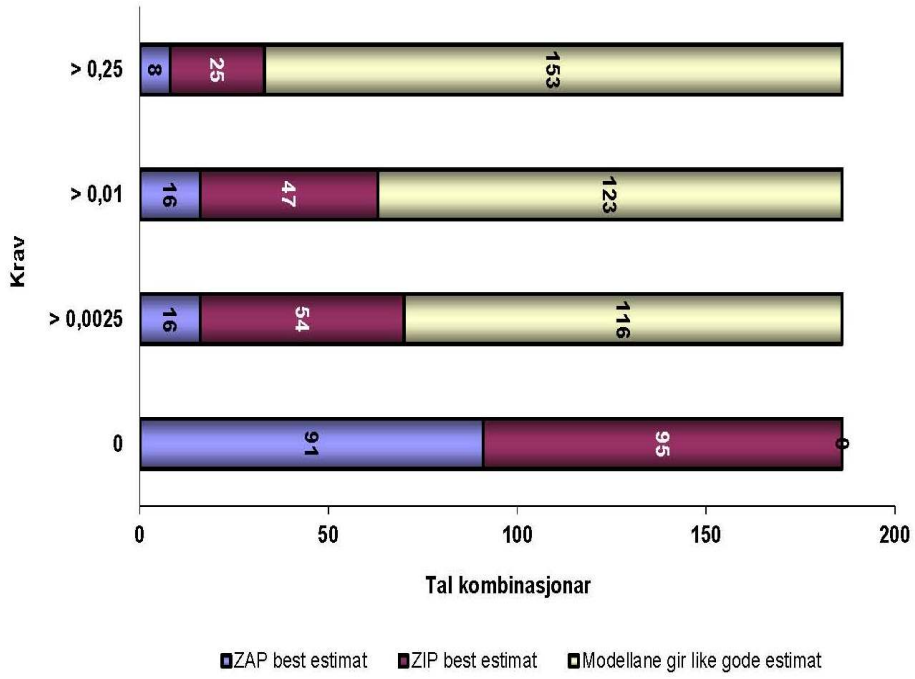
Sjå relevant avsnitt i 6.2.5 i teoridelen. Dette kan også vera årsaka til kvifor modellane for nokre av koeffisientane faktisk gir betre estimat som ukorrekt modell enn som korrekte modell. ZAP er i dei fleste tilfella best for begge dei to estimata i binomisk del. Dette ser me for alle differansekrava til MSE, og for ZAP som både ukorrekt og korrekt modell.

For Poissondelen har ZIP lågast MSE fleire gonger enn ZAP, men ingen av av modellane er generelt sett klart den beste. Ser me først på ZIP-datasettet og koef.int.,figur 8.9, er ZAP kun betre enn ZIP med differansekrav 0,25, og då for kun 11 mot 7 kombinasjonar av i alt 164 moglege. For alle andre minstekrav til MSE er ZIP best. Dette gjeld også for alle dei fire differansekrava med ZAP som korrekt modell, figur 8.5. For koef. forkl ser me først på resultata frå ZAP som korrekt modell. Figur 8.6 viser at med differansekrav på 0 er modellane best i bortimot like mange tilfelle, medan for dei tre andre krava er ZIP best for betydeleg fleire parameterkobinasjonar enn ZAP. For ZIP som korrekt modell er ZIP og ZAP like gode i svært mange av kombinasjonane med differansekrav over 0. Sjå figur 8.10. Resultata for koef. forkl. med ZIP som korrekt datasett skil seg noko frå resten av resultata for Poissondelen. Med differansekrav 0 er ZAP best modell i langt fleire komnbinasjonar enn ZIP. For kravet 0,25 har ZAP kun nokre fleire tilfelle enn ZIP som best modell for koeffisienten til forklaringsvariabelen, medan for dei to resterande krava har ZIP litt fleire tilfeller enn ZAP som best modell. Samla resultat viser at modellane for Poissondelen er like gode for svært mange av kombinasjonane, med ein tendens til at ZIP er betre i noko fleire tilfeller enn ZAP der ein modell er best.

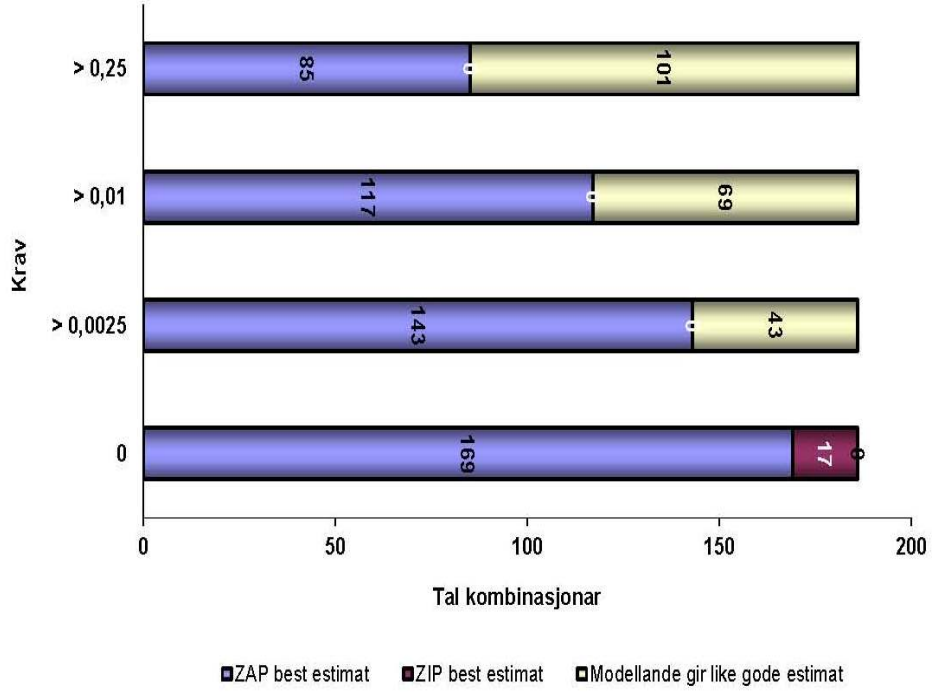
Tabell 8.2.2 til 8.2.2 viser kven av ZIP og ZAP som er best modell når me ser på fleire regresjonskoeffisientar samstundes. Ser me på dei to regresjonskoeffisientane i Poissondelen under eitt finn me at for ZAP-datasettet er ZIP betre enn ZAP for ein mykje større prosentdel av parameterkombinasjonane for alle differansekrav. Dette gjeld både når modellen må ha lågast MSE for begge estimata og kun for minst det eine, sjå tabellane 8.2.2 og 8.2.2. Dette stemmer godt overeins med at ZIP som ukorrekt modell alltid er best for dei to koeffisientane i Poissondelen kvar for seg. Tabellane 8.2.2 og 8.2.2 viser derimot at på ZIP-datasettet er ZAP ein større konkurrent til ZIP. Ser me på mengd tilfelle der modellen er best for minst ein av estimata, er ZIP betre for kun nokre prosent fleire parameterkombinasjonar for differansekrava 0,01 og 0,0025. For krav på 0,25 og 0 er ZAP ein generelt litt betre modell enn ZIP. Me ser at ZIP estimerer generelt oftare betre estimat enn ZAP for Poissondelen, og særleg med ZAP-fordelte observasjondataett. Men i mange av analysesituasjonane er ikkje ZIP best modell for mange fleire tilfelle enn



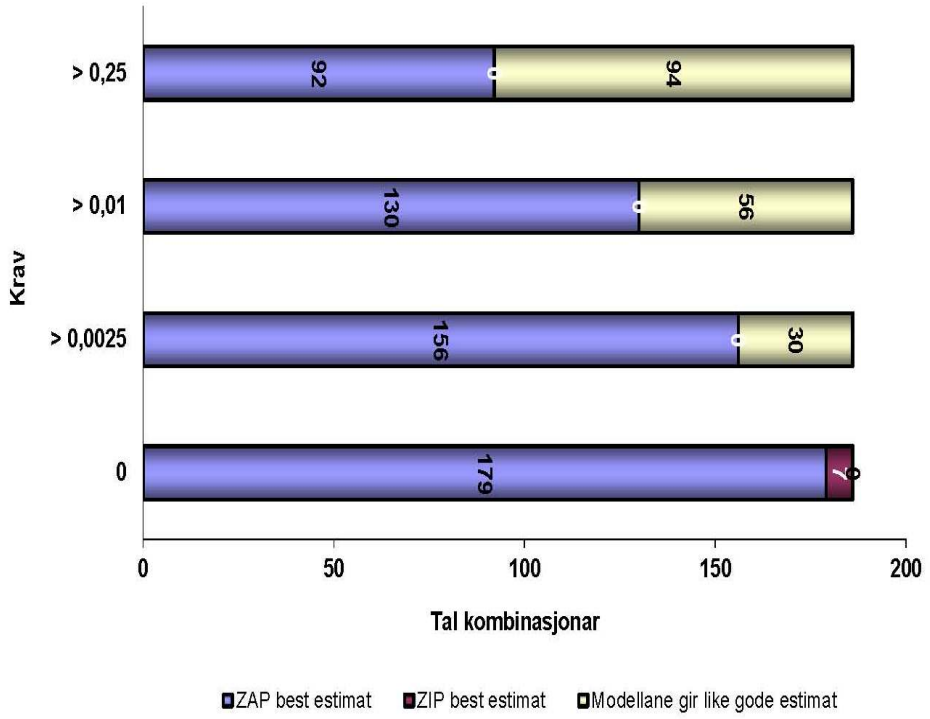
Figur 8.5: *Mengd parameterkombinasjonar der ZIP og ZAP er kategorisert som best modell for dei ulike differansekrava til MSE, koef. int. Poissondel, ZAP korrekt modell*



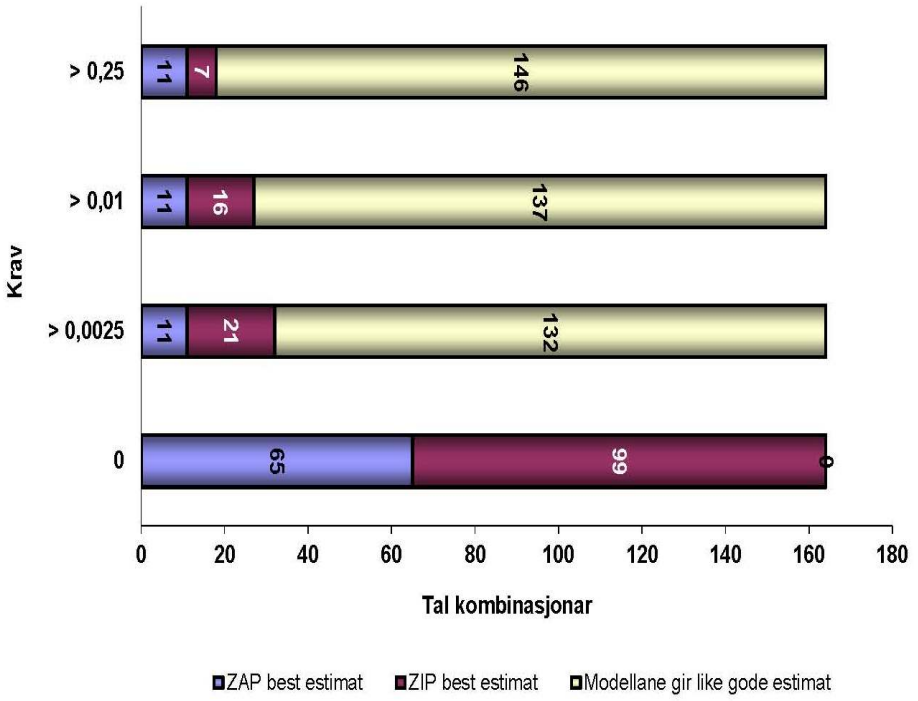
Figur 8.6: *Mengd parameterkombinasjonar der ZIP og ZAP er kategorisert som best modell for dei ulike differansekrava til MSE, koef. forkl. Poissondel, ZAP korrekt modell*



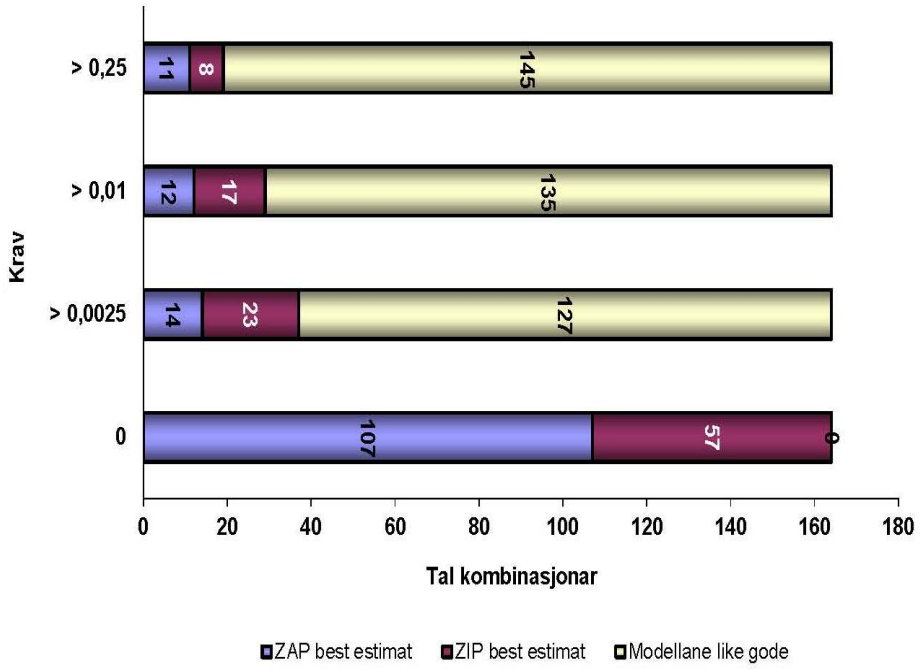
Figur 8.7: *Mengd parameterkombinasjonar der ZIP og ZAP er kategorisert som best modell for dei ulike differansekrava til MSE, koef. int. binomisk del, ZAP korrekt modell*



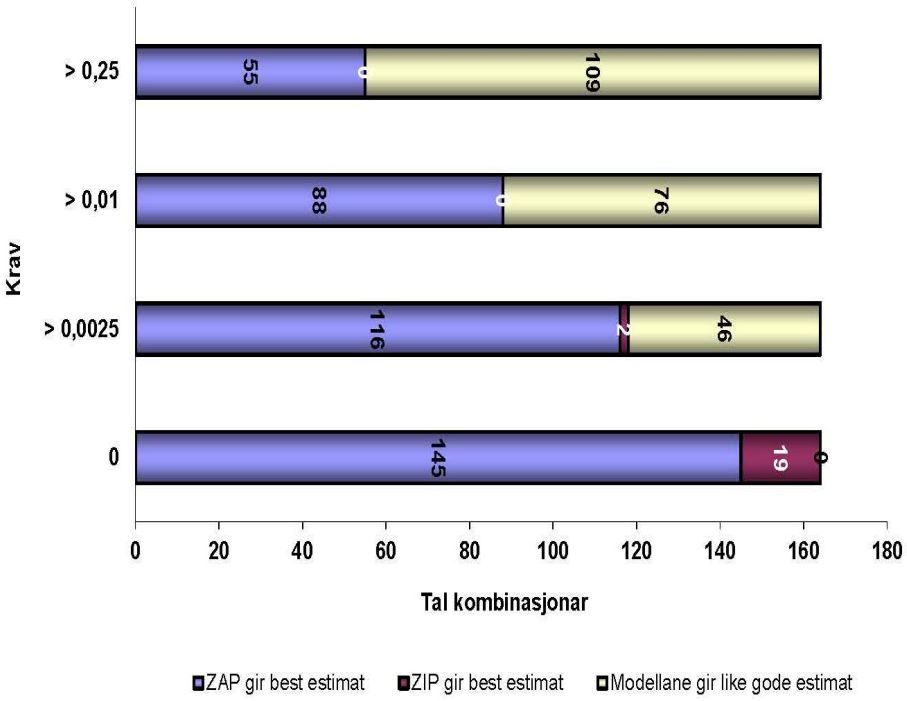
Figur 8.8: *Mengd parameterkombinasjonar der ZIP og ZAP er kategorisert som best modell for dei ulike differansekrava til MSE, koef. forkl. binomisk del, ZAP korrekt modell*



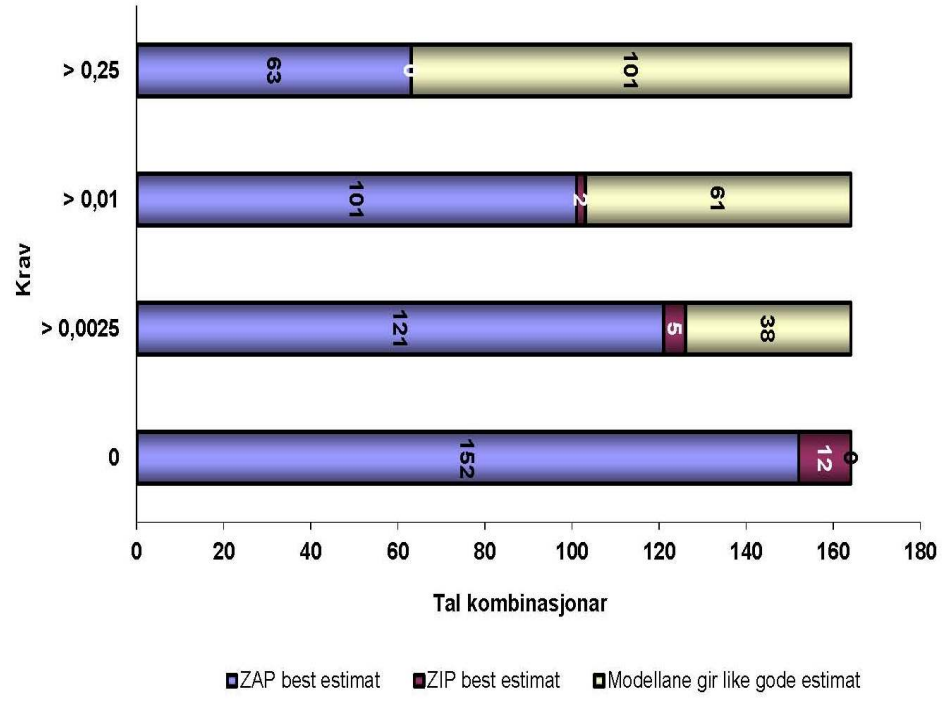
Figur 8.9: *Mengd parameterkombinasjonar der ZIP og ZAP er kategorisert som best modell for dei ulike differansekrava til MSE, koef. int. Poissondel, ZIP korrekt modell*



Figur 8.10: *Mengd parameterkombinasjonar der ZIP og ZAP er kategorisert som best modell for dei ulike differansekrava til MSE, koef. forkl. Poissondel, ZIP korrekt modell*



Figur 8.11: *Mengd parameterkombinasjonar der ZIP og ZAP er kategorisert som best modell for dei ulike differansekrava til MSE, koef. int. binomisk del, ZIP korrekt modell*



Figur 8.12: *Mengd parameterkombinasjonar der ZIP og ZAP er kategorisert som best modell for dei ulike differansekrava til MSE, koef. forkl. binomisk del, ZIP korrekt modell*

ZAP, og for differansekrav på 0 og 0,25 er ZAP også nokre gonger best for størst prosentdel.

Minstekrav for differanse	0	0,0025	0,01	0,25
Alle fire estimata	23,7%	6,5%	5,4%	3,8%
Minst ein av koef. forkl.	100,0%	83,9%	69,9%	49,5%
Begge koef. forkl.	45,2%	8,6%	8,6%	4,3%
Minst ein for Poisson del	48,9%	8,6%	8,6%	4,3%
Begge for Poisson del	30,1%	6,5%	5,4%	4,3%
Minst ein for binomisk del	96,2%	83,9%	69,9%	49,5%
Begge for binomisk del	90,9%	76,9%	62,9%	45,7%

Tabell 8.1: *Prosentdel parameterkombinasjonar der ZAP som korrekt modell er best ved å sjå på fleire regresjonskoeffisientar samstundes, med bruk av ulike differansekrava for MSE*

Minstekrav for differanse	0	0,0025	0,01	0,25
Alle fire estimata	0,0%	0,0%	0,0%	0,0%
Minst ein av koef. forkl.	54,8%	29,0%	25,3%	13,4%
Begge koef. forkl.	0,0%	0,0%	0,0%	0,0%
Minst ein for Poisson del	69,9%	31,2%	28,0%	13,4%
Begge for Poisson del	51,1%	27,4%	23,7%	12,4%
Minst ein for binomisk del	9,01%	0,0%	0,0%	0,0%
Begge for binomisk del	3,8%	0,0%	0,0%	0,0%

Tabell 8.2: *Prosentdel parameterkombinasjonar der ZIP som ukorrekt modell er best ved å sjå på fleire regresjonskoeffisientar samstundes, med bruk av ulike differansekrava for MSE*

Verken ZIP eller ZAP er klart best som korrekt modell. Det er heller ingen av modellane som i mange parameterkombinasjonar er best for alle fire koeffisientane. ZIP er faktisk ikkje best for alle estimata samtidig verken som korrekt eller ukorrekt modell. Tendensen til kven av modellane som er best for Poisson og binomisk del går på tvers av dei to store analysedatasetta. Det ser difor ikkje ut til at det er av særleg betydning om responsvariabelen i observasjonsdatasetta er ZIP- eller ZAP-fordelte. Dette til tross for at modellane som regel er mykje betre som korrekt enn ukorrekt modell. Fordelinga til observasjonsdatasetta har difor innverknad på kor bra modellen estimerer, men er ikkje avgjerande for kven av modellane som generelt gir lågast MSE-verdiar for dei to komponentane i ZIP og ZAP.

Minstekrav for differanse	0	0,0025	0,01	0,25
Alle fire estimata	0,0%	0,0%	0,0%	0,0%
Minst ein koef. forkl.	38,4%	17,1%	11,6%	4,9%
Begge koef. forkl	3,7%	0,0%	0,0%	,,0%
Minst ein i Poisson del	61,0%	14,0%	10,4%	,,9%
Begge for Poisson del	34,1%	12,8%	9,8%	4,3%
Minst ein for binomsik del	15,2%	4,3%	1,2%	0,0%
Begge for binomsk del	3,7%	0,0%	0,0%	0,0%

Tabell 8.3: *Prosentdel parameterkombinasjonar der ZIP som korrekt modell er best ved å sjå på fleire regresjonskoeffisientar samstundes, med bruk av ulike differansekrava for MSE*

Minstekrav for differanse	0	0,0025	0,01	0,25
Alle fire estimata	31,1%	6,7%	6,7%	6,1%
Minst ein koef. forkl.	96,3%	73,8%	61,6%	39,0%
Begge koef. forkl.	61,6%	8,5%	7,3%	6,1%
Minst ein i Poisson del	65,9%	8,5%	7,3%	6,7%
Begge i Poisson del	39,0%	6,7%	6,7%	6,7%
Minst ein i binomisk del	96,3%	77,4%	62,2%	38,4%
Begge i bonimsik del	84,8%	67,1%	53,0%	33,5%

Tabell 8.4: *Prosentdel parameterkombinasjonar der ZAP som ukorrekt modell er best ved å sjå på fleire regresjonskoeffisientar samstundes, med bruk av ulike differansekrava for MSE*

Ved å innføre stigande minstekrav til differansen til MSE-verdiane ser me at mengda kombinasjonar der ZIP og ZAP vert kategorisert som like veks mykje hurtigare for Poissondelen enn for den binomiske delen. Røyrdiagramma 8.9 og 8.10 viser at allereie med krav på 0,0025 er over 77 % av parameterkombinasjonane kategorisert som like for Poissondelen med ZIP som korrekt modell. Tilsvarande verdi er 62 % for ZAP-datasettet. Det fortel oss at MSE-verdiane for koeffisientane i Poissondelen er svært like for dei to modellane for dei fleste kombinasjonane. Det er likevel fleire tilfelle med MSE-differanse på over 0,25. For den binomiske delen gjeld dette svært mange kombiasjonar. For desse tilfella veit me lite om den faktiske skilnaden i estimata frå ZIP- og ZAP-regresjonen. Me veit heller ikkje kor stor den relative differansen til MSE-verdiane er. Dette er det neste steget i analysen.

8.2.3 Oppsummering

Ein ser at både for ZIP og ZAP fordelt datasett er ZIP ein del betre for estimeringa av parametrane i Poissondelen, medan ZAP er heilt klart bedre for estimering av regresjonskoeffisientane i binomisk del. Modellane estimerer likare for Poisson enn binomisk del, men valet av regresjonsmodell mellom ZIP og ZAP har også her heilt klart betydning for MSE-verdiane til estimata. Differansen i MSE-verdiane for modellane er for størsteparten av kombinasjonane liten, med det er også mange tilfelle, særleg for binomisk del, der differansen er svært stor.

8.3 Relativ differanse for MSE-verdiane til estimata

8.3.1 Metode og analysegrunnlag

Me har no funne ut at også i praktisk anvending har valet mellom ZIP- og ZAP-som regresjonsmodell betydning for MSE til estimata til regresjonskoeffisientane. Me ynskjer no å sjå meir på kor bra den beste modellen er i forhold til den dårlegaste modellen. Storleiken (MSE til den beste modellen) / (MSE til den dårlegaste modellen) er i den samanhengen ein svært bra observator. Den gir oss den relative differansen for dei parvise MSE-verdiane til ZIP og ZAP. Ein skilnad på MSE-verdiane på 0,5 har større betydning dersom den beste modellen har MSE på 0,2 enn dersom den har MSE på 20. MSE/MSE fortel kor ulikt modellane estimerer i forhold til verdien til den dårlegaste modellen. Med ein MSE/MSE på over 0,75 er MSE-verdiane svært

like, og over er dei ,9 bortimot heilt like. MSE/MSE-verdi på 0,5 fortel om svært ulike MSE-verdiar, og med MSE/MSE under 0,1 er estimata til ZIP og ZAP totalt ulike. Me vil i analysen alltid bruke den lågaste MSE-verdien som teljar og den høgaste som nemnar, uavhengig av kva som er modell som er korrekt. Det er også alltid brukt ein MSE-verdi frå ZIP-regsjon og ein frå ZAP-regresjon. Me vil difor berre bruke MSE/MSE som namn på storleiken.

Me ser først på dei fire regresjonskoeffisientane separat, og finn MSE/MSE for kvar parameterkombinasjon. Merk at MSE for begge regresjonsmodellane inngår i denne storleiken. Ein deler difor ikkje resultat opp etter nytta regresjonsmodell, men etter kva modell som er best. Figur 8.13 til 8.16 viser resultata for kvar modell som best som korrekt og best som ukorrekt modell. Lengda på røyra er samla mengd parameterkombinasjonar der modellen har lågast MSE for den aktuelle koeffisienten. Oppdeling av kvart røyr fortel korleis parameterkombinasjonane er fordelt på interval av verdiar for MSE/MSE.

8.3.2 Restultat

Me ser først på tilfelle der ZAP er den beste modellen. Resultata er vist i figur 8.13 og 8.16. For ZAP-datasettet er MSE/MSE-verdien for koef. int. Poissondel over 0,9 for 44 av 56 kombinasjonar, og for 75 av 91 kombinasjonar for koef. forkl. For ZIP-datasettet er tendensen den same. For binomsk del er talet på kombinasjonar med MSE/MSE over 0,9 betydeleg lågare både for ZAP som korrekt og ukorrekt modell. For ZAP-datasetttet er kombinasjonane mykje meir fordelt mellom intervallkategoriane for binomisk del enn for Poissondelen. For binomisk del fell dei aller fleste innanfor kategorien $> 0{,}9$, medan for Poissondeelen inneheld dei to ytterkategoriane $> 0{,}9$ og $< 0{,}1$ tilsaman størstedelen av parameterkombinasjonane.

I tilfella der ZIP er den beste modellen er det for alle fire regresjonskoeffisientane svært mange kombinasjonar som har MSE/MSE på over 0,9, og få under 0,5. Sjå figur 8.14 og 8.15. For koef. int. for Poissondelen er 69 av 130 moglege kombinasjonar for ZAP-datasettet og heile 76 av 99 moglege kombinasjonar for ZIP-datasettet over 0,9. Det er tilsvarande berre 41 av 130 moglege og 22 av 99 moglege kombinasjonar som er under 0,5. For binomisk del er alle dei 55 kombinasjonane der ZIP er best for minst ein av koeffisientane alle i kategorien over 0,9. For ZIP som best modell er den relative differansen for MSE svært liten, og for binomisk del er den alltid over 0,9.

Kategoriane i midten for MSE/MSE har få kombinasjonar for alle røyrene på alle fire diagramma. Dette kjem av at modellane aldri er svært dårlege for

same parameterkombinasjon. Dette ser me også frå figur 8.1 til 8.4 i starten av analysen.

8.3.3 Oppsummering

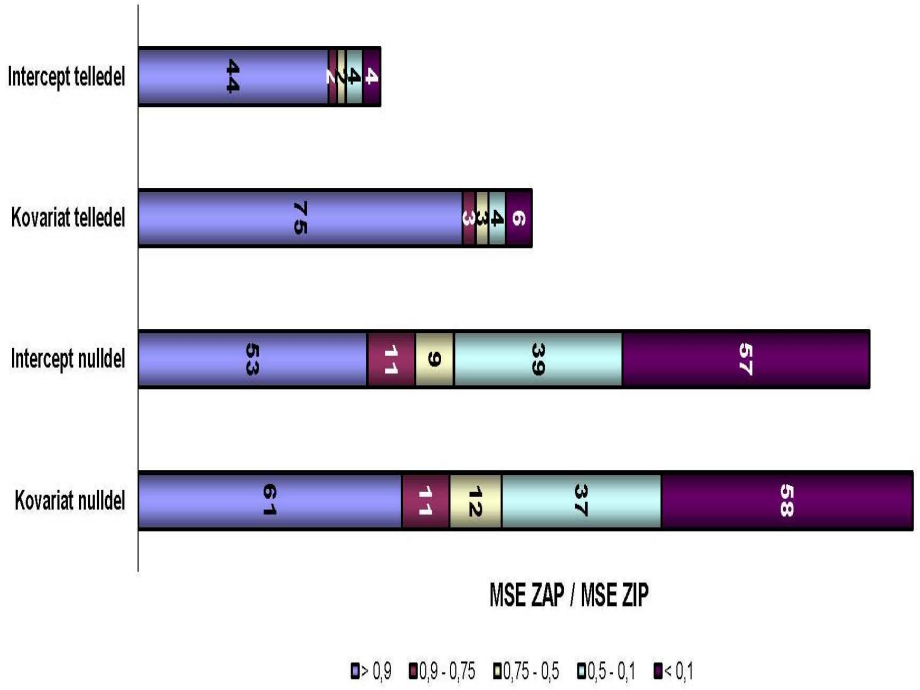
For Poissondelen er det ei betydeleg stor mengd av parameterkombinasjonane som har MSE/MSE-verdi over 0,9. Modellane estimerer difor svært likt for Poissondelen. Eit unntak er med ZIP som best modell og ZAP som korrekt modell. Då ser me ein tendens til større relativ differanse. For den binomiske delen er MSE-verdiane svært like for alle tilfella med ZIP som best modell, medan for ZAP som best modell, har godt over ein tredjedel av kombinasjonane svært ulike estimat for ZIP- og ZAP-regresjon.

8.4 Analyse av strukturelle komponentar mot relativ differanse i MSE-verdiane

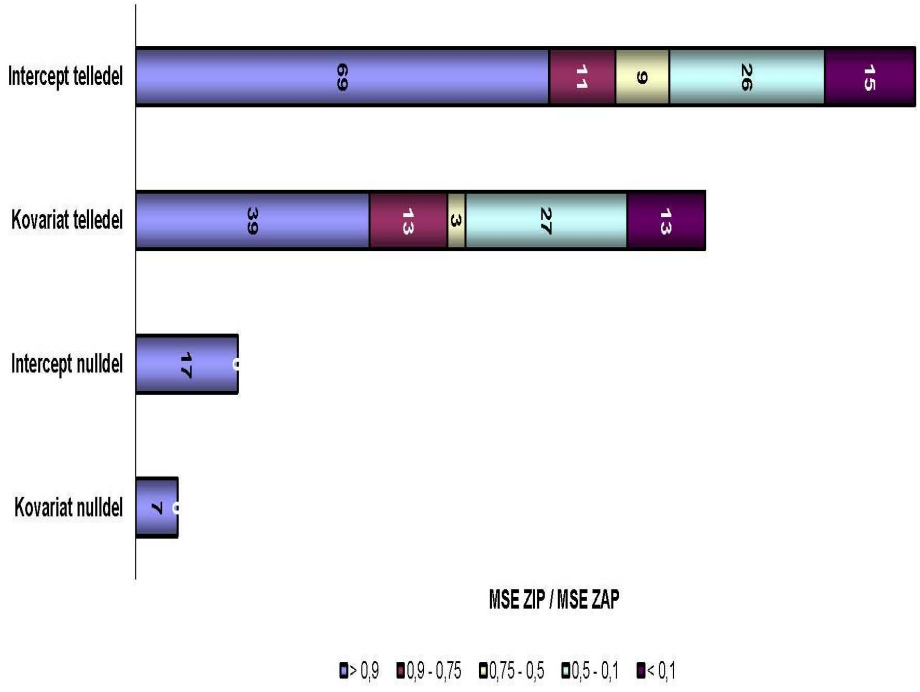
8.4.1 Metode og analysegrunnlag

Me har no sett at valet mellom ZIP og ZAP som regresjonsmodell har betydning for kor gode estimata for regresjonskoeffisientane er. For fleirtalet av kombinasjonane er differansen i MSE-verdiane for modellane liten, men det også mange tilfelle der dei er svært ulike. Det overraskande i resultata våre er at det ikkje alltid er den korrekte modellen som gir dei beste estimata. Me ynskjer difor å finna ut om valet av modell bør takast på eit anna grunnlag enn fordelinga i observasjonsdatasetta. Analysen så lagt i oppgåva indikerer at ZIP og ZAP har styrken sin på kvar sin komponent i modellane. Det er svært sannsynleg at forventninga i Poissondelen og det samla nullsannsynet har innverknad på differansen i MSE-verdiane for dei to modellane. Dersom dette stemmer kan ein truleg ta valet av modell ut frå mengda 0-observasjonar og/eller tendens til høge eller låge verdiar i dei resterande observasjonane i observasjonsdatasettet. Me ynskjer i denne delen av analysen å finne ut om dette stemmer.

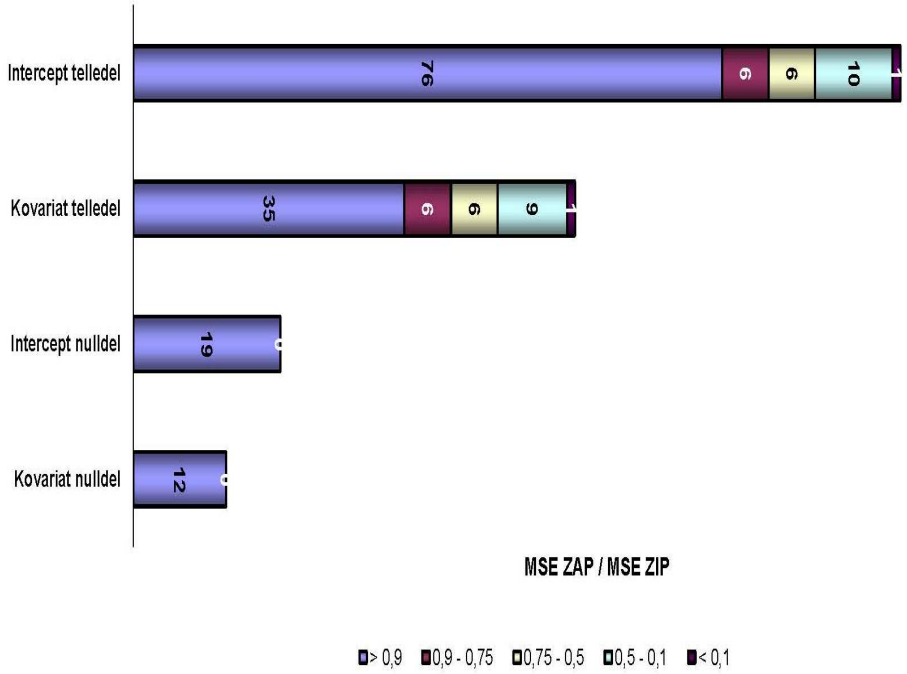
For å finne ut om det samla nullsannsynet og forventninga til Poissondelen har innverknad på kor ulike ZIP og ZAP estimerer, plottar me MSE/MSE opp mot dei tilhøyrande verdiane til dei to faktorane. Dette vert gjort for kvar av dei fire regresjonskoeffisientane separat. Dei sanne verdiane til regresjonskoeffisientane er med på bestemme forventninga for begge komponentane i ZIP og ZAP. Me tester difor også om det er samanheng mellom MSE/MSE og visse kombinasjonar av koeffisientane. Resultata for det samla nullsannsynet



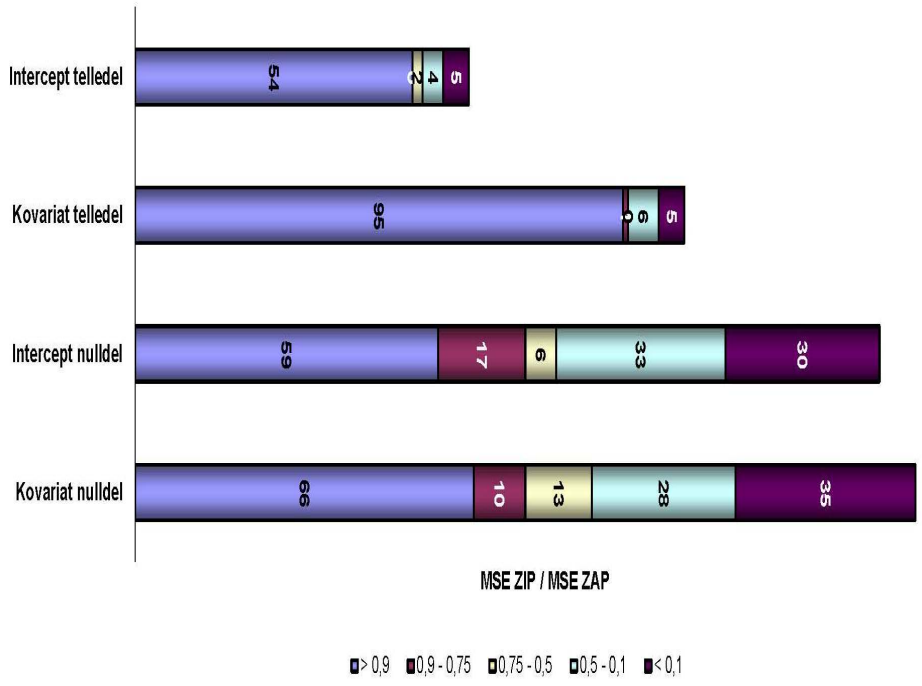
Figur 8.13: *Mengd parameterkombinasjonar som fell innanfor dei ulike intervalla for MSE/MSE for ZAP best og korrekt modell for dei ulike regresjonskoeffisientane*



Figur 8.14: *Mengd parameterkombinasjonar som fell innanfor dei ulike intervalla for MSE/MSE for ZIP best og ukorrekt modell for dei ulike regresjonskoeffisientane*



Figur 8.15: *Mengd parameterkombinasjonar som fell innanfor dei ulike intervalla for MSE/MSE for ZIP best og korrekt modell for dei ulike regresjonskoeffisientane*



Figur 8.16: *Mengd parameterkombinasjonar som fell innanfor dei ulike intervalla for MSE/MSE for ZAP best og ukorrekt modell for dei ulike regresjonskoeffisientane*

og forventninga i Poissondelen er presentert i same diagram. Dette er gjort for å også vise ein evt. samanheng mellom MSE/MSE og ulike kombinasjonar av nullsansyn og forventning. For å forstå kvifor ulike verdiar av dei to faktorane gjer ZIP og ZAP like eller ulike i praksis, treng me også å vite kva dei to faktorane har å seie for kor bra modellane estimerer utan samanlikning av regresjonsmodellane. Me ser difor i tillegg på MSE-verdiane til ZIP og ZAP opp mot dei to faktorane kvar for seg.

Dersom ein forklaringsvariabel virkar motsett inn på Poisson-delen og den binomiske delen av regresjonen vil dette truleg skape problem for ZIP. Eit døme på dette er når høg verdi for ein forklaringsvariabel gir både høgt nullsannsyn og høg forventning i Poissondelen. Dette skjer når regresjonskoeffisientane til forklaringsvariabelen har like forteikn (Sjå bevis i tillegg A). Sidan begge komponentane i ZIP inngår i både det samla nullsannsynet og forventninga i Poissondelen, må ZIP finne eit estimat som stemmer overeins med innverknaden forklaringsvariabelen har for begge desse faktorane. Dette kan gjere iterasjonsprossesen i maksimeringa av likelihoodfunksjonen svært ustabil. Me ynskjer difor å sjå om slike situasjonar gjer estimata til ZIP og ZAP spesielt ulike. Til dette ser me på to parameterkombinasjonar der alle koeffisientane har parvis same absoluttverdi, men der regresjonskoeffisientane til forklaringsvariablane har like forteikn for den eine kombinasjonen i paret, og ulike for den andre. Sidan verdiane til forklaringsvariabelen er sentrert rundt gjennomsnittet, har ikkje forteiknet til koef. forkl. for dei to komponentane innverknad på verken forvetning i Poissondel eller samla nullsannsyn. Dei bestem kun om det er høg eller låg verdi for forklaringsvariabelen som fører til høg eller låg verdi for nullsannsyn og forvetning i Poissondel. Også her brukar me storleiken MSE/MSE som mål på kor like MSE-verdiane for ZIP og ZAP er. Det er nytta 23 identiske paramterekombinasjonar frå kvar av dei to store analysedatasetta, dvs 46 par ialt.

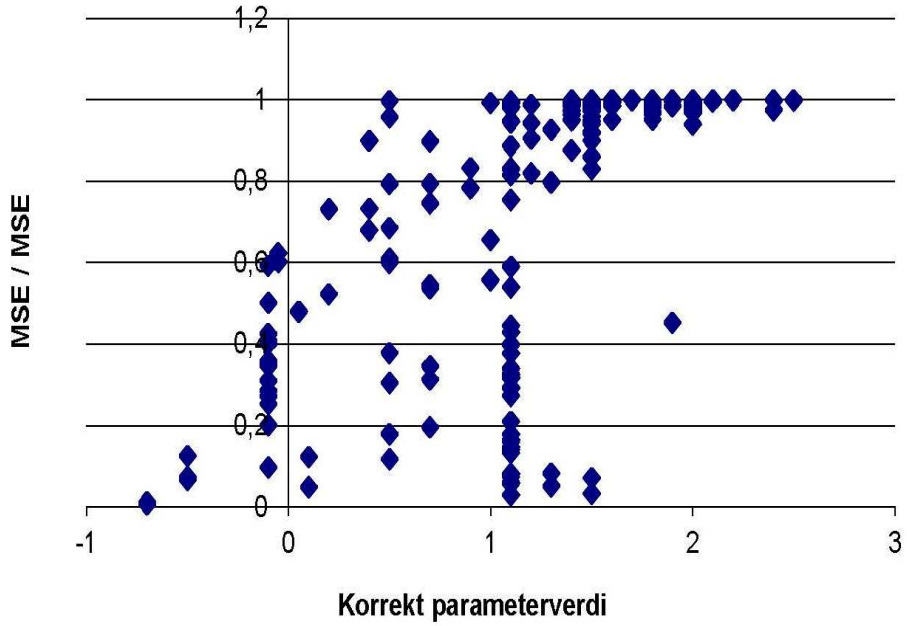
Der MSE/MSE er plotta mot forventning for Poissondelen og samla nullsannsyn er verdiane til desse faktorane rekna ut i samsvar med om det er ZIP eller ZAP som er korrekt modell. Resultata frå ZIP- og ZAP-datasetta var bortimot identiske, og resultata er difor presentert samla. Unntaket er figur (8.25) til (8.28) der MSE/MSE frå dei to samanliknan kombinasjonane er plotta parvis. Me vil i denne delen av analysen også sjå på korleis ZIP og ZAP estimerer i forhold til kvarandre i situasjonar der det er færre 0-observasjonar i datasetta enn det vanleg Poisson forutsetter. Dette er hovudsakleg parameterkombinasjonar med både låg forventing i Poissondelen og lågt nullsannsyn.

8.4.2 Resultat

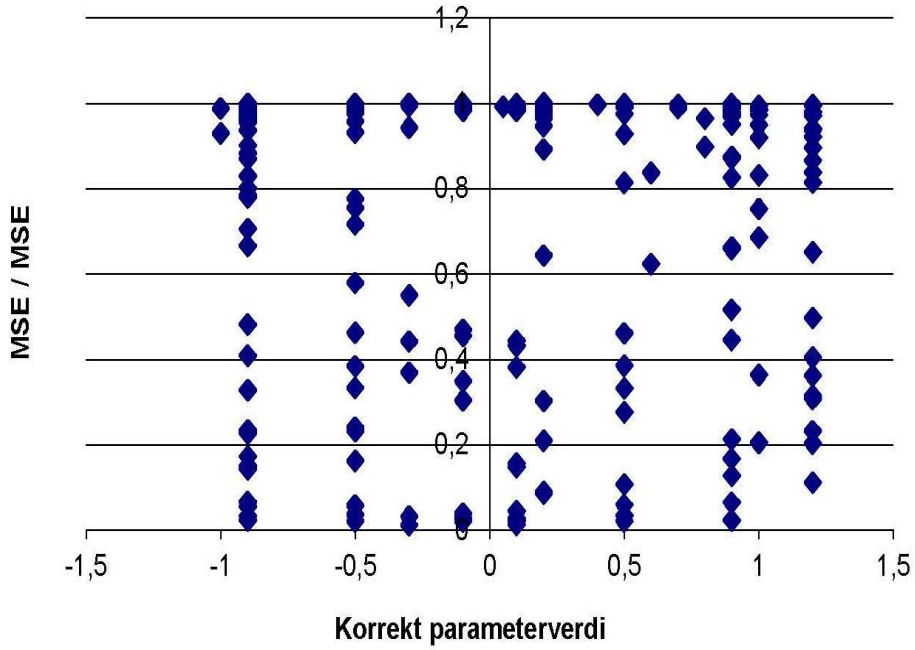
Me ser først på om det er nokon samanheng mellom storleiken til MSE/MSE og den korrekte verdien til den tilhøyrande regresjonskoeffisienten. Dei relevante resultata er vist i figur 8.17 til 8.20. Her er korrekt verdi plotta opp mot MSE/MSE for den aktuelle regresjonskoeffisienten. For koef. int. for Poissondelen er det ein klar samanheng mellom stigande korrekt verdi for koeffisienten og MSE/MSE. Særleg er det ein klar tendens til at korrekt verdi over 1,5 gir svært lik estimering med ZIP og ZAP. For koef. int. i binomisk del ser me at det for korrekt verdi mellom -1 og 1 ikkje er nokon tendens til verken stor eller liten MSE/MSE, medan korrekte koeffisientverdiar lågare enn -1 gir svært ulike estimat, og korrekte verdiar høgare enn 1 gir relativt ulike estimat for modellane. For koeffisientane til forklaringsvariablane viser ikkje plotta noko mønster, bortsett frå to kombinasjonar for binomisk del med -2,4 og -1,4 som korrekt verdi, der begge MSE/MSE er under 0,001.

Det me no har funne ut om samanhengen mellom korrekt koeffisientverdi og MSE/MSE stemmer godt overeins med resultata med fant om forventning Poissondel og samla nullsansyn verkar inn på om ZIP og ZAP gir ulike estimat. Desse er vist i figur 8.21 til 8.24, der MSE/MSE for estimatet til regresjonskoeffisienten er plotta opp mot den korrekte verdien. Alle figurane viser svært tydeleg at forventninga i Poissondelen har stor innverknad på om modellane estimerer ulikt, der høg forventning gir svært liten relativ differanse. Dette gjeld for alle koeffisientane. For binomisk del ser me ein proporsjonal samanheng mellom den relative differansen i MSE-verdiane og forvetningsverdien i Poissondelen. Høg forventningsverdi fører til like estimat, medan låg forventning fører til svært ulik estimering med ZIP og ZAP. Også for Poissondelen har forventninga stor betydning, men her ser me eit noko anna mønster i resultata. For MSE/MSE over 0,8 er forventninga alltid høg, og dess høgare den er dess likare estimering får me for ZIP og ZAP. Men for MSE/MSE under 0,8 snur tendensen. Her ser me ei gradvis auke i forventninga når MSE/MSE synk i verdi ned mot 0.

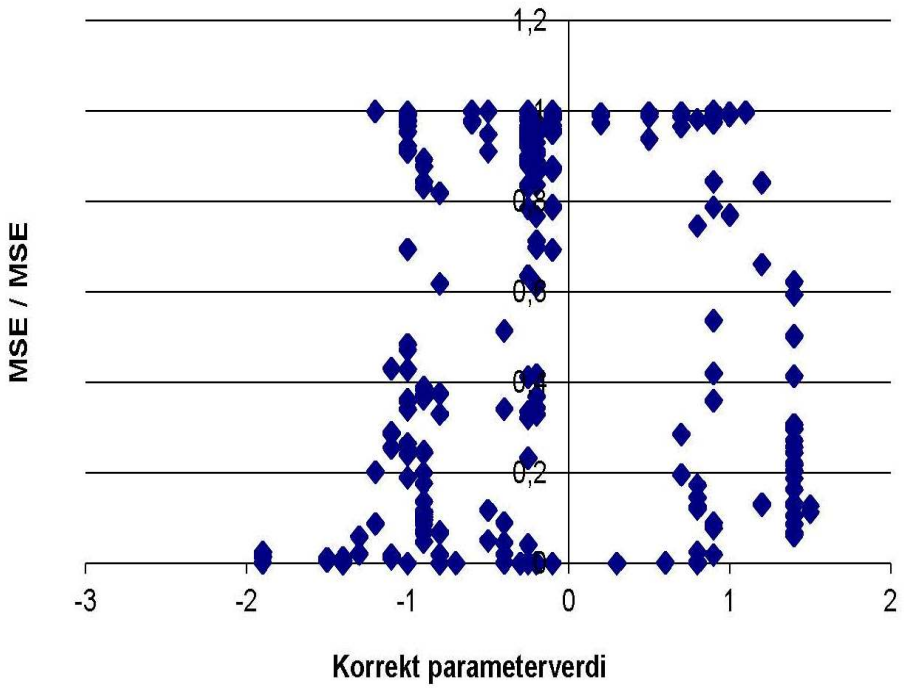
Også det samla nullsannsynet har innverknad på den relative differansen mellom MSE-verdiane for dei to modellane. For Poissondelen ser ein at dess høgare nullsannsyn, dess større relativ differanse. Men for MSE/MSE over 0,9 endrar dette seg noko, og for desse tilfella har me alle verdiar for samla nullsannsyn. For den binomiske delen, sjå figur 8.23 og 8.24, er det særleg interessant at ser på MSE/MSE-verdiar nært opp mot 1 og ned mot 0 ligg nullsannsynet i begge tilfella som ei klynge rundt 0,5. For MSE/MSE-verdiar mellom 0,1 og 0,9 er det derimot god variasjon for det samla nullsannsynet.



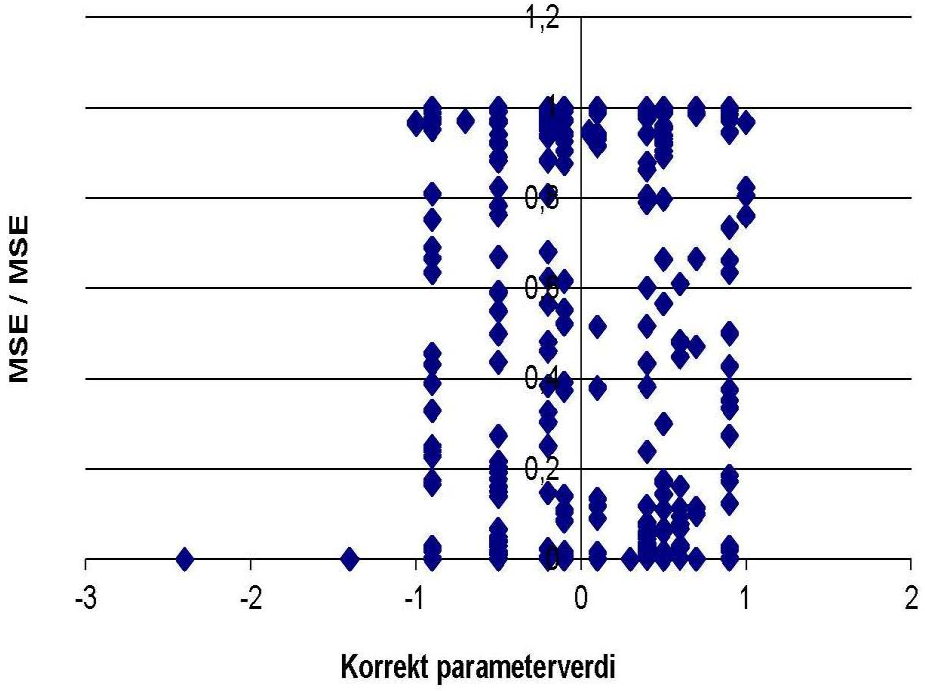
Figur 8.17: *MSE/MSE for estimata mot korrekt verdi for koef. int. Poissondel*



Figur 8.18: *MSE/MSE for estimata mot korrekt verdi for koef. forkl. Poissondel*



Figur 8.19: *MSE/MSE for estimata mot korrekt verdi for koef. int. binomisk*



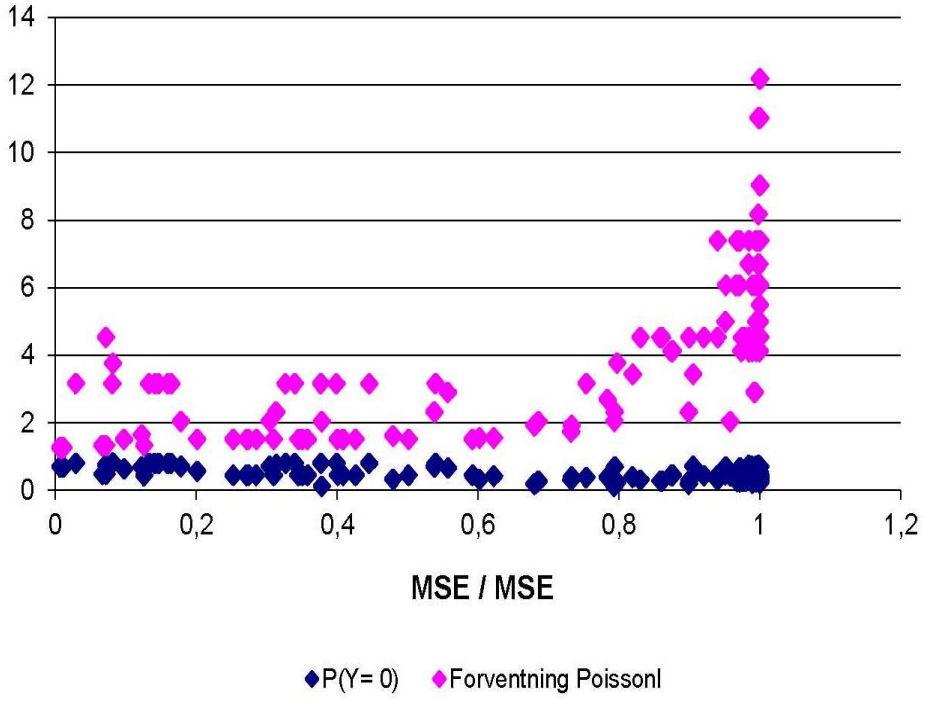
Figur 8.20: *MSE/MSE for estimata mot korrekt verdi for koef. forkl. binomisk*

Dei tendensane me ser for samanhengen mellom MSE/MSE og samla nullsannsyn og forventning i Poissondelen stemmer overeins med resultata me finn for analysen av MSE/MSE mot korrekt koeffisientverdi, der interceptleddet har mest betydning for storleiken til dei to sktrukturelle faktorane.

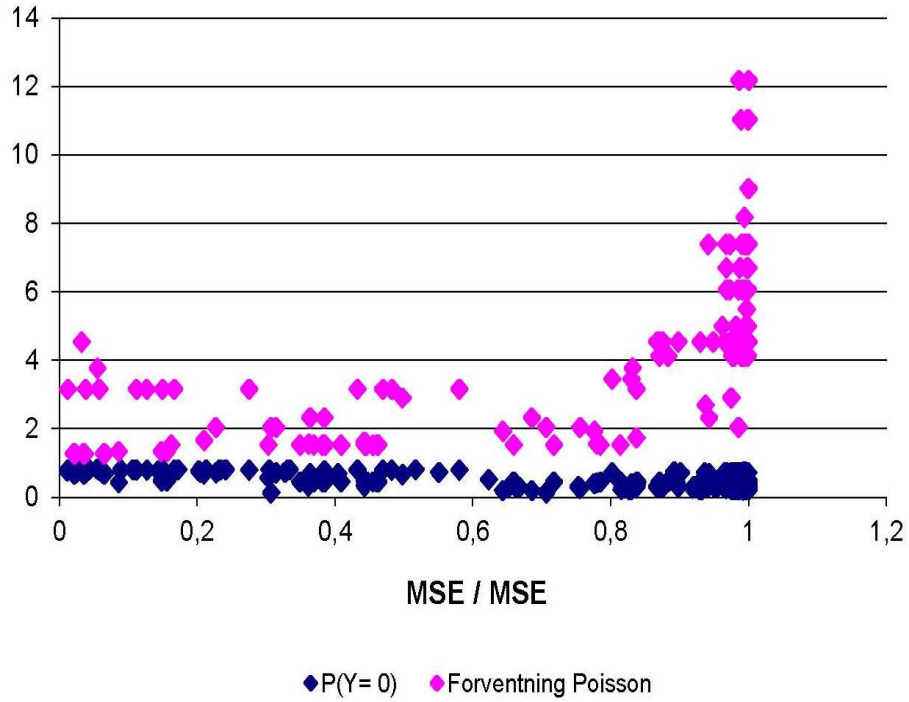
I analysen ser me også på MSE-verdiane til ZIP og ZAP separat både generelt og opp mot det samla nullsannsynet og forventning i Poissondelen. Me finn då at ZIP som antatt har svært høge MSE-verdiar for den binomiske delen, medan resultata frå Poissondelen er mykje lågare. Høgt nullsannsyn fører til dårlig estimering, og det same gjeld låg forventning i Poissondelen. For ZAP er det kun nullsannsynet som har innverknad på den binomiske delen. Forventninga til utfalla over 0 er den av dei to faktorane med størst betydning for estimeringa i Poissonkomponeten i modellen, men nullsannsynet har også innverknad. ZAP gir gode estimat for den binomiske delen, og relativt gode estimat for Poissondelen.

Det ser også ut til at kombinasjonen av verdiane til samla nullsannsyn og forventning i Poissondelen verkar inn på MSE/MSE for ZIP og ZAP. For Poissondelen er estimata alltid svært like ved høg forventning i Poissondelen, uavhengig av nullsannsynet for parameterkombinasjonen. Dersom forventningsverdien derimot er under 4 har nullsannsynet også innverknad, grunna at forventningsverdien her er meir spredt for dei ulike MSE/MSE. Då fører høgare samla nullsannsyn til meir ulik estimering. Me ser også at det er i tilfella der ein har høgt nullsannsyn samtidig med forventning Poissondel rundt 4 me har den største relative differansen. Dette gjeld for regresjonskoeffisienten til både interceptleddet og forklaringsvariabelen. For den binomiske delen er det forvetninga fotilr Poissondelen som har mest innverknad for alle MSE/MSE. Me ser likevel at nullsannsynet har noko betydning sidan både stor og liten realtiv differanse kun har nullsannsyn rundt 0,5.

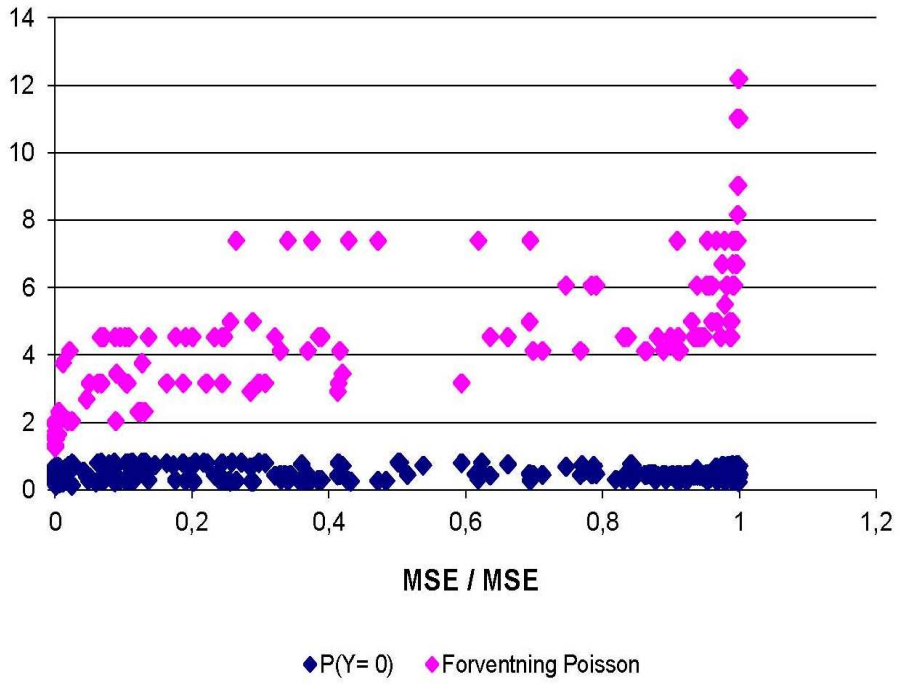
Figur 8.25 til 8.28 viser at dersom ein forklaringsvariabel dreg i ulik retning for forventning i Poissondel og samla nullsannsyn fører dette til større relativ differanse mellom MSE-verdiane enn for tilsvarande parameterkombinasjon der den trekk i same retninga. Mange av kombinasjonspara har lik MSE/MSE for begge tilfella, men mange par har også svært ulik relativ differanse for like og ulike forteikn. For Poissondelen er det rundt halvparten som har lik MSE-verdi. Forskjellen i den relative differansen for dei andre para varierar, og nokre gonger får me også høgast verdi for MSE/MSE i kombinasjonen med like forteikn. For binomisk del er talet på kombinasjonar med lik MSE noko lågare. Den halvdelen av punkta som ligg mot venstre i alle dei fire figurane er regresjon på ZAP-fordelte observasjonsdatasett, medan halvdelen til høgre



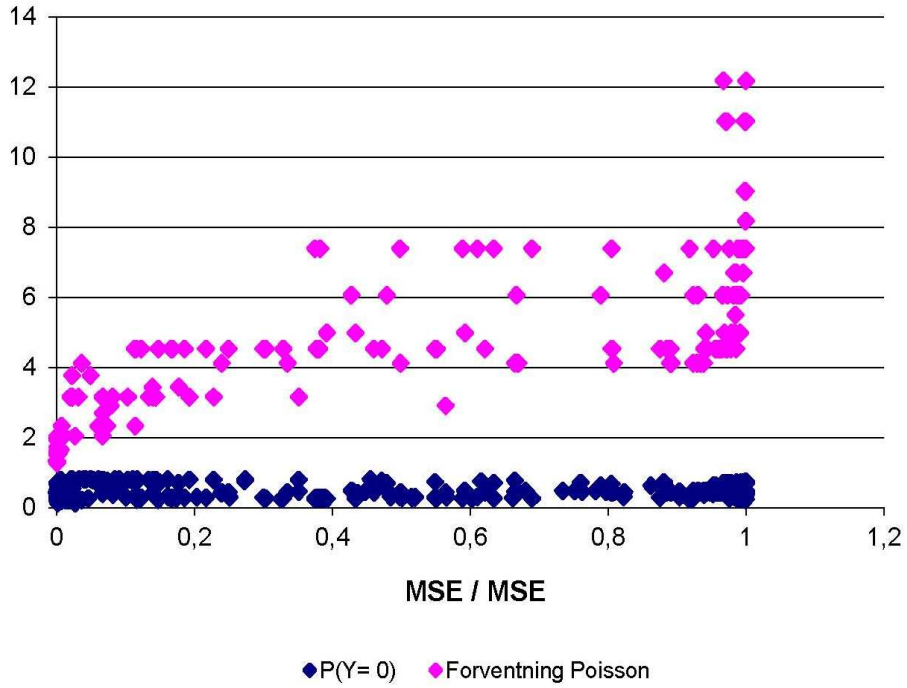
Figur 8.21: *Påverknad av $P(Y = 0)$ og forventning Poissondel på MSE/MSE for estimata til koef. int. Poissondel*



Figur 8.22: *Påverknad av $P(Y = 0)$ og forventning Poissondel på MSE/MSE for estimata til koef. forkl. Poissondel*



Figur 8.23: *Påverknad av $P(Y = 0)$ og forventning Poissondel på MSE/MSE for estimata til koef. int. binomisk del*



Figur 8.24: *Påverknad av $P(Y = 0)$ og forventning Poissondel på MSE/MSE for estimata til koef. forkl. binomisk del*

kjem frå ZIP-fordelte observasjonsdatasett. Resultata frå datasetta er svært like.

Dei observasjonsdatasetta der me i vår analyse får færre 0-observasjonar enn det ordniær Poisson forutset har som regel både lågt nullsannsyn og svært låg forventning i Poissondelen. Ein kan difor kjenne igjen dei tilhøyrande parameterkombinasjonane i plotta av MSE/MSE mot dei to strukturelle faktorane. I figur 8.21 ser ein at dei aktuelle parameterkombinasjonane har MSE/MSE på mellom 0,3 og 0,6 for koef. int. for Poissondelen. Tilsvarande intervall for koef. forkl. er 0,6 til 0,8. Sjå figur 8.22. ZIP klarar seg i ei viss grad bra for Poissondelen. Figur 8.23 og 8.24 viser plot for den binomiske komponenten til modellane. Her har alle dei aktuelle parameterkombinasjonane MSE/MSE-verdi heilt ned mot 0. Alle dei aktuelle resultata frå denne delen av analysen viser som anteke på førehand at ZIP er svært dårleg på å modellerer det binæret utfallet når mengda 0-observasjonar er færre enn i vanleg Poissonfordeling.

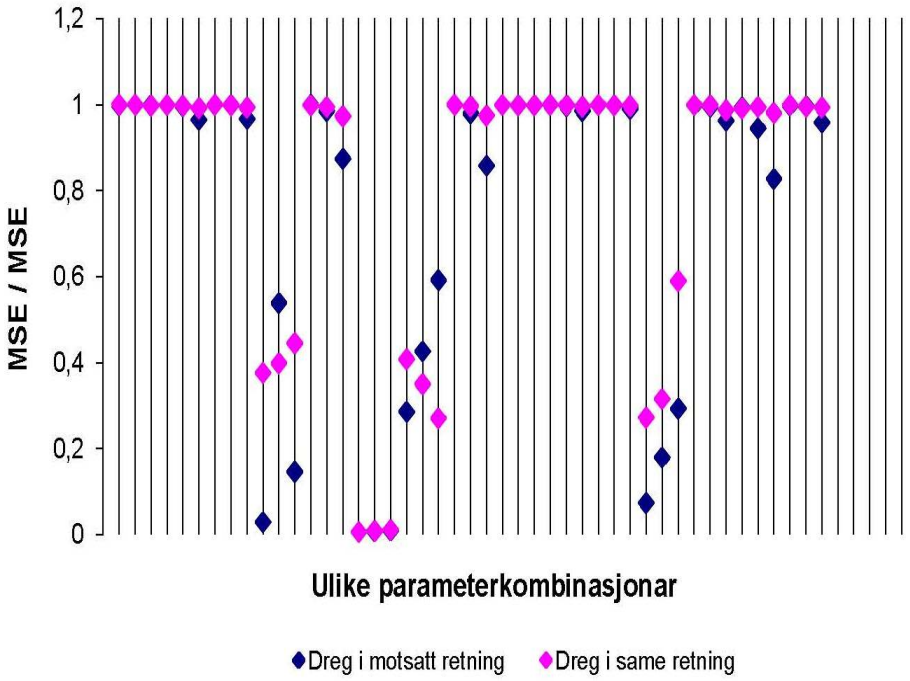
8.4.3 Oppsummering

Der er klar samanheng både mellom forventning for Poissondelen og den relative differansen for MSE, og mellom samla nullsannsyn og ulik MSE-verdiar. For binomisk del er det forventning Poissondel som har mest betydning, medan for Poissondelen har samla nullsannsyn også innverknad. Visse kombinasjonar av desse faktorane gir svært lik eller ulik estimering ved samanlikning av ZIP- og ZAP-regresjon. Motsatt dragning på nullsannsyn og forventning i Poissondelen for same verdi av forklaringsvariabel har mest innverknad for den binomiske delen, og kor stor betydninga er varierar. I dei situasjonane der me har færre 0-observasjonar enn det vanleg Poisson forutset, er ZIP svært dåleg på estimeringa for binomisk del.

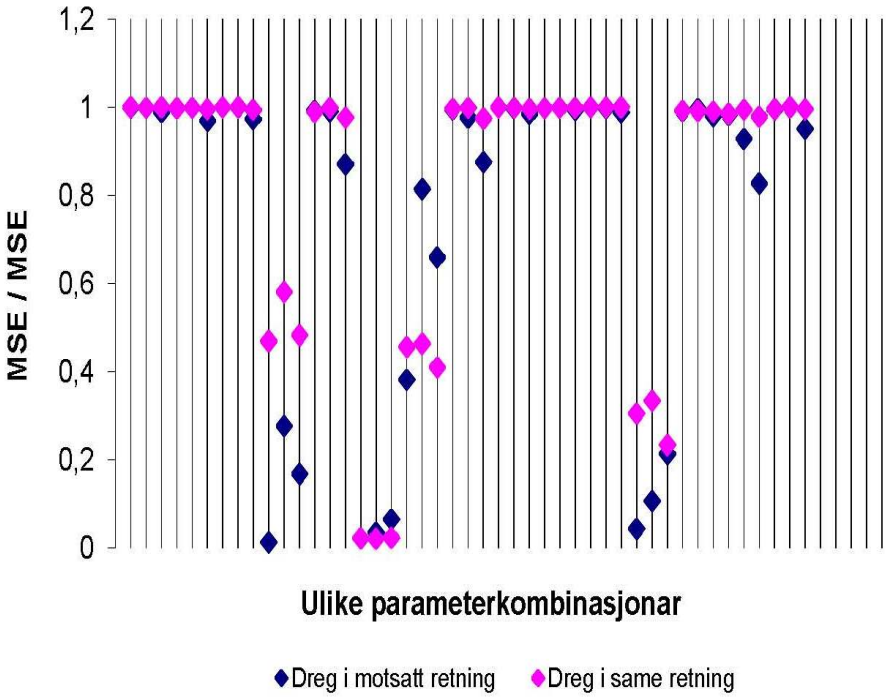
8.5 Regresjonsanalyse på store datasett

8.5.1 Metode og analysegrunnlag

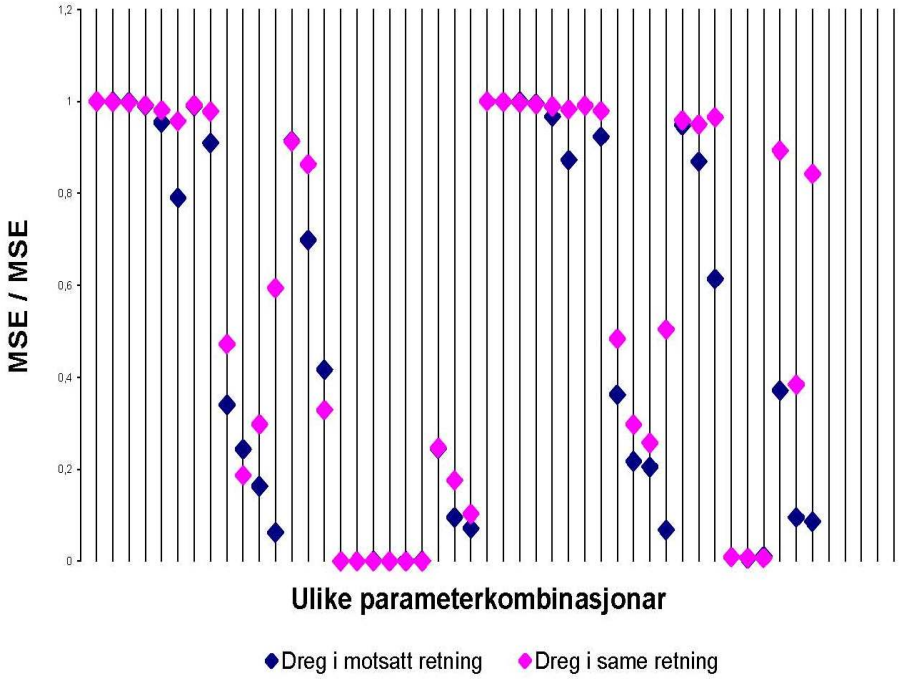
Analysen så langt er gjort med observasjonsdatasett på 50 responsvariablar. Det er grunn til å tru at dersom ein auker talet på observasjonar, vil regresjonskøyring med ZIP og ZAP gi likare estimering. Med større datasett har modellane fleire observasjonar som kan vise samanhengen mellom forklaringsvariabelen og responsvariabelen. Me ynskjer difor å samanlikne MSE/MSE



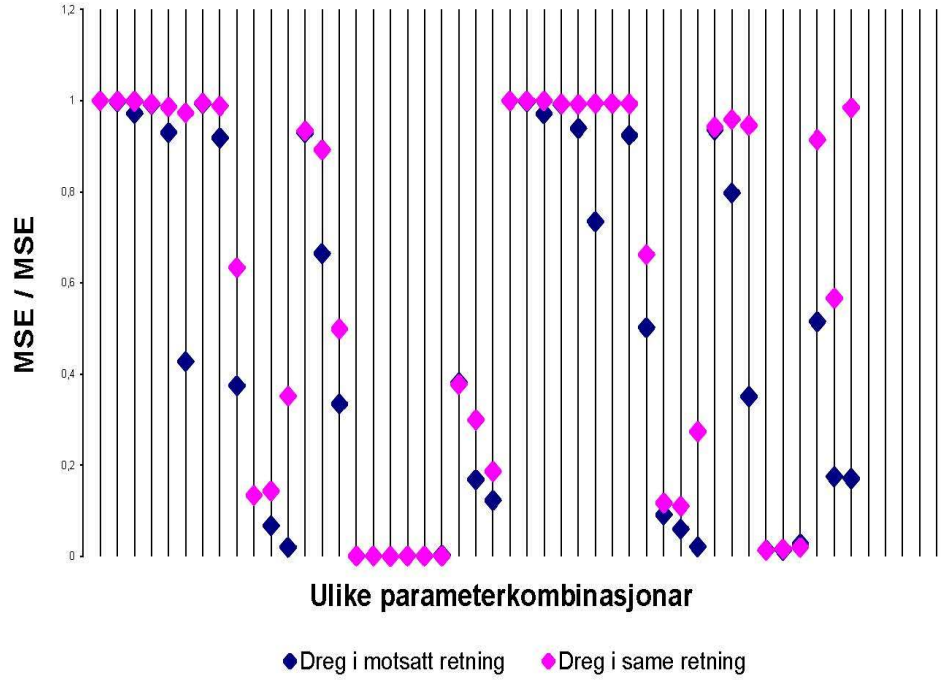
Figur 8.25: *MSE/MSE for parameterkombinasjon med like forteikn for koef. int. Poissondel plotta parvis med MSE/MSE for parameterkombinasjon med like forteikn*



Figur 8.26: *MSE/MSE for parameterkombinasjon med like forteikn for koef. forkl. Poissondel plotta parvis med MSE/MSE for parameterkombinasjon med like forteikn*



Figur 8.27: *MSE/MSE for parameterkombinasjon med like forteikn for koef. int. binomisk del plotta parvis med MSE/MSE for parameterkombinasjon med like forteikn*



Figur 8.28: *MSE/MSE for parameterkombinasjon med like forteikn for koef. forkl. binomisk del plotta parvis med MSE/MSE for parameterkombinasjon med like forteikn*

frå datasettet med 50 observasjonar med tilsvarande verdiar frå eit datasett med langt fleire observasjonar, men med same parameterkombinasjon.

For å få tydeleg fram ein evt. tendens til likare estimering med ZIP og ZAP ved store datasett, vel me å sjå på parameterkombinasjonar som med 50 observasjonar har svært stor relativ differanse. Av kombinasjonane med MSE/MSE-verdi under 0,25 plukkar me ut 24 kombinasjonar til bruk i analysen. Desse er valt med tanke på å ha med tilfelle frå både ZIP- og ZAP-datasettet, samt å nytte heile variasjonsbredda i forventning Poissondel og samla nullsannsyn bland dei aktuelle kombinasjonane. Det er også teke med tilfelle der koeffisientane til forklaringsvariablane har same forteikn og motsette forteikn. For dei 25 kombinasjonane generer me nye datasett, no med 1000 observasjonar. Med ei så stor auke i observasjonar vil kome klart fram dersom storleiken på datasettet har innverknad på MSE/MSE. Det vert så utført både ZIP- og ZAP-regresjon med 10 000 simuleringar for dei 24 kombinasjonane, og dei nye regresjonsresultata vert til slutt samanlikna med dei tilsvarande resultata frå det mindre datasettet for kvar kombinasjon.

8.5.2 Resultat

Oversikt over MSE-verdiane for kvar frå resgresjon med 50 og 1000 observasjonar for kvar av dei aktuelle parameterkobinasjonane er samla i tabell (8.5.2).

Strukt. faktor		Int. Poisson		Forkl. Poisson		Int. binomisk		Forkl. binomisk	
forv	nullsann	50 obs.	1000 obs.	50 obs.	1000 obs.	50 obs.	1000 obs.	50 obs.	1000 obs.
1,3	0,48	0,0723	0,2237	0,1484	0,0937	0,0003	0,0002	0,0006	0,0011
3,2	0,8	0,0289	0,9999	0,0124	0,9989	0,2442	0,6066	0,1342	0,9196
3,2	0,8	0,0808	1	0,0386	0,9982	0,2216	0,5992	0,1032	0,9236
3,2	0,8	0,1329	0,9938	0,0579	0,9873	0,1044	0,4271	0,0667	0,4138
3,2	0,8	0,1466	0,9593	0,1671	0,9552	0,0625	0,2165	0,0201	0,1165
3,2	0,8	0,1411	0,9622	0,1499	0,9604	0,0676	0,2164	0,0231	0,1179
1,3	0,69	0,0083	0,1512	0,0368	0,0496	0,0005	0,0016	0,0008	0,0065
1,3	0,69	0,0105	0,1512	0,0643	0,0496	0,0005	0,0016	0,0007	0,0065
1,3	0,69	0,0113	0,6743	0,0229	0,6745	0,0004	0,0008	0,0006	0,0033
1,3	0,69	0,0111	0,6782	0,0228	0,6784	0,0004	0,0007	0,0006	0,0031
1,3	0,69	0,0103	0,1537	0,0663	0,0498	0,0005	0,0016	0,0007	0,0066
3,2	0,8	0,1596	0,9502	0,1118	0,9802	0,1058	0,1345	0,0321	0,0881
2	0,71	0,1784	0,9668	0,2271	0,7534	0,0193	0,0209	0,0109	0,024
1,5	0,65	0,0974	0,9586	0,1624	0,9277	0,0013	0,0041	0,0022	0,0287
3,8	0,82	0,0814	0,9818	0,0557	0,978	0,1273	0,5269	0,0495	0,3534
1,3	0,44	0,1256	0,1136	0,0863	0,5512	0,0002	0,0004	0,0005	0,0007
3	0,81	0,0735	1	0,0439	0,9977	0,2172	0,6043	0,0912	0,9911
3	0,81	0,0569	1	0,0288	0,9974	0,2044	0,6063	0,0847	0,9973
3	0,81	0,179	0,9993	0,1069	0,9937	0,2051	0,4546	0,0605	0,5653
3	0,81	0,2087	0,9982	0,2411	0,9959	0,1283	0,4538	0,0428	0,5625
1,6	0,77	0,1174	0,9641	0,1435	0,9409	0,0785	0,0649	0,0292	0,0904
1,1	0,79	0,0491	0,9997	0,0896	0,9909	0,0249	0,0428	0,0061	0,9137
2	0,73	0,1956	0,9624	0,205	0,9442	0,1456	0,0752	0,0575	0,0907

Tabell 8.5: *MSE/MSE for dei fire regresjonskoeffisientane for regresjon på datasett med 50 og 1000 obsevasjonar*

Av resultata i tabellen 8.5.2 ser ein tydeleg at storleiken på observasjonsdatasettet har innverknad på kor likt ZIP og ZAP estimerer regresjonskoeffisientane. Særleg for estimata i Poissondelen vert den relative differansen mykje mindre med eit større datasett. For koef. int. har me no svært høg MSE/MSE-verdi for alle kombinasjonane med forventning over 2,0. Også fem av dei ti tilfella med forventning under 0,2 får ganske lik estimering for interceptleddet i Poissondelen. Alle tilfella der me no har liten relativ differanse for koef. int. har me også liten relativ differanse for koef. forkl.

For estimata i den binomiske komponenten av modellane finn me ikkje ein like klar tendens til at større datasett gir likare estimering. For interceptparameteren er de kun kombinasjonane med forventningsverdi i Poissondelen på over 3,0 som har fått ein betydeleg mindre relativ differanse, og estimata er framleis er svært ulike. For koef. forkl. er det 5 kombinasjonar som med 1000 observasjonar får MSE/MSE mellom 0,3 og 0,55. 6 kombinasjonar får MSE/MSE-verdi over 0,78. Med untak av eit tilfelle har alle desse høgare forventning for Poissondelen enn 3,0. Det er likevel godt over halvparten av dei 24 kombinasjonane som ikkje har høgare MSE/MSE-verdi enn 0,096. Kombinasjonen med størt relativ differanse har framleis MSE/MSE under 0,0003 for koef. forkl. Poisson.

For dei aller fleste av dei 24 aktuelle kombinasjonane er nullsannsynet nokså høgt, med unntak av to tilfelle med nullsannsyn under 0,5. Men desse har forventning for Poissondel under 2,0 og ville truleg også med 1000 observasjonar fått stor relativ differanse uavhengig av nullsannsynet. Det er difor vanskeleg for oss å seia noko om kor mykje det samla nullsannsynet har å seia for om eit større datasett gir likare estimering.

Resultata viser ingen tendens til at forteikna på forklaringsvariablane har noko å seie for om eit større datasett gjer modellane likare i praksis.

8.5.3 Oppsummering

Storleiken på observasjonsdatsettet har mykje å seie for kor like ZIP og ZAP estimerer regresjonskoeffisientane i Poissondelen av modellane. Dette gjeld særleg for datasett med forventning for Poissondelen over 2,0, men også fleire av kombinasjonane med lågare forventning får likare MSE ved fleire observasjonar. For estimeringa av koeffisientane i binomisk del har storleiken kun ein liten innverknad på MSE/MSE.

Kapittel 9

Tolking av analysedata og konklusjon

Me vil no tolke resultata som er presentert i kapittel 7 ut frå den strukturelle oppbygginga til metodegrunnlaget til ZIP og ZAP. Me ynskjer å finne ut kva som avgjer om MSE-verdiane vert like eller ulike for dei to modellane, samt finne ut om me bør ta valet av regresjonsmodell ut frå eit anna grunnlag enn fordelinga til observasjonsdatasettet. Etter ein drøfningsdel vert det gitt ein konklusjon på spørsmåla me har stilt i denne oppgåva.

9.1 Samanheng resultatdata og strukturell oppbygging av ZIP og ZAP

Me har i analysearbeidet til no funne ut at ZAP er mykje betre enn ZIP på estimeringa av koeffisientane i binomisk del, medan modellane er meir likt gode for Poissondelen. Det har også kome tydeleg fram at det samla nullsannsynet og forventninga i Poissondelen har innverknad på den relative differansen i MSE-verdiane. Me skal no finne årsaka til dette ved å samanlikne tendensane i resultatdata med korleis det samla nullsannsynet og forventninga til utfalla over 0 verkar inn på ZIP og ZAP som gode regresjonsmodellar. Dette vil vera med å avgjere om val av regresjonsmodell kan takast ut frå strukturelle tendensar me ser i observasjonsdatasetta.

Det er kun den av ZIP- og ZAP-modellane som har same grunnleggjande fordeling som observasjonsdatasettet som bruker korrekt likelihoodfunksjon i estimeringa av regresjonskoeffisientane. Den forutset difor korrekt strukturell oppbygging av datasettet når den prøver å finne tilbake til dei korrekte

verdiane for koeffisientane. I praktiske situasjonar som forskningsarbeid vil ein ofte ha datasett som ikkje heilt følgjer verken ZIP eller ZAP fordeling. Ein av modellane er likevel nærast korrekt punktsannsyn. I vårt studium følgjer observasjonsdatasetta alltid ei av fordelingane heilt, grunna at me genererer datasetta utfrå den gitte fordelinga. Det er naturleg å tenkje seg at den korrekte modellen som brukar det riktige metodegrunnlaget vil finne lettast tilbake til dei sanne regresjonskoeffisientane, og gi rettast og mest stabile estimat. Både ZIP og ZAP for dei aller fleste koeffisientane er betre som korrekt enn ukorrekt modell, men resultata viser overraskane nok at den korrekte modellen ikkje alltid er den beste.

Det kjem tydeleg fram av resultata presentert i kapittel 7 at ZIP verken som korrekt eller ukorrekt modell klarar å finne tilbake til korrekte koeffisientverdiar for den binomiske komponenten i modellen. Det er også tydeleg at høgt nullsannsyn og låg forventning for Poissondelen samtidig fører til spesielt høge MSE-verdiar for ZIP. ZIP forutset at ved låg forventning for utfalla over null, kjem 0-observasjonane i datasettet frå to kjelder. For ZAP-fordelte observasjonsdatasett vil ikkje dette vera tilfellet, og det er difor ikkje vanskeleg å tenkje seg at ZIP som ukorrekt modell vil vera dårleg på estimeringa for den binomske delen. Men i fleire av desse tilfella gir ZIP dårlege estimat for den binomske delen samtidig som den gir gode estimat for Poissondelen. Modellen klarar altså bra å finne den korrekte samanhengen mellom forklaringsvariabelen og den trunkerte Poissonfordeling i ZAP, sjøl om i ZIP har dei to komponentane innverknad på kvarandre. Dei 0-observasjonane ZIP fastset som vanlege nullar føyer seg utan store problem inn som ein del av Poissonfordelinga, sjølv om dei i den korrekte ZAP-fordelinga i datasettet ikkje er det. Observasjonane stemmer likevel sjeldant heilt med det korrekte mønsteret mellom forklaringsvariabelen og observasjonane over 0, og dess fleire vanlege nullar dess vanskelegare vil det vera for ZIP som ukorrekt modell å finne dei korrekte koeffisientverdiane. Låg verdi for λ fører til at ZIP modellerer mange nullobservasjonar som ein del av den eigentlege trunkerte Poissonfordelinga. Det er då fleire observasjonar som ikkje passar heilt inn i forhold til det sanne mønsteret mellom koeffisient og forklaringsvariabel, og det fører til ustabile estimat og høg MSE-verdi. Resultata viser likevel at ZIP generelt sett er ein like god modell som ZAP på Poissondelen dersom me ser på både ZIP- og ZAP-datasettet under eitt.

ZIP er også som korrekt modell svært dårleg på å modellere det binære utfallet i regresjonen. Ut frå resultata i analysen ser me at det samla nullsannsynet ikkje har stor innveknad på dette. Derimot har verdien til λ, forventninga til Poissondelen i ZIP, stor betydnng. Låg λ fører til svært høg

MSE-verdi. Dette kjem av at låg λ gir mange 0-observasjonar i datasettet. Mange vanlege nullar gir ikkje store problem for ZIP for Poissondelen. Derimot skaper dei vanskar når ZIP skal modellere det binære utfallet i datasettet, sjølv når ZIP er korrekt modell. Dei vanlege nullane er ikkje avhengige av forklaringsvariabelen som inngår i logit-linken i modellen, men av forklaringsvariabelen til log-linken. Verdien til forklaringsvariabelen i den binomiske komponenten varierar difor fritt for dei vanlege nullane. Medan dei strukturelle nullane er avhengige av forklaringsvariabelen i den binomske komponenten, og varierar fritt for forklaringsvariabelen i Poissondelen. Dersom det er for mange vanlege nullar i forhold til strukturelle vil det vera vanskeleg for ZIP å sjå samanhengen mellom dei strukturelle nullane og tilhøyrande verdi til forklaringsvariabelen, og modellen får problem med å finne ut kva som er korrekt verdi for regresjonskoeffisienten. Det er altså ikkje talet på 0-observasjonar, men forholdet mellom dei to gruppene med nullar som avgjer kor bra ZIP klarar å estimere den binomiske delen av modellen. ZIP har behov for ei tilstrekkeleg mengd strukturelle nullar, som vil vise den faktiske innverknaden til forklaringsvariabelen, i forhold til mengda vanlege nullar, som derimot skjuler innverknaden. Er det for stor del vanlege nullar vil estimeringa vera ustabil og me observerer høge MSE-verdiar. Forholdet mellom dei to gruppene med nullar har altså større betydning for kor bra ZIP klarar å modellere den binomske delen av modellen enn om ZIP faktik er korrekt modell eller ikkje.

For ZAP som regresjonsmodell er det heile mykje enklare. For modelleringa av det binære utfallet er det kun nullsannsynet av dei to strukturelle faktorane som har innverknad. Det er difor kun èin linkfunksjon med tilhøyrande forklaringsvariablar å ta omsyn til i maksimeringa av likelihoodfunksjonen til Poissonkomponenten. Dersom ZIP er det korrekte datasettet vil det vera ulik verdi for forklaringsvariablane i dei to gruppene med nullar. Dette gjer det vanskelegare for ZAP å finne det korrekte mønsteret, sidan dei observasjonane som eigentleg høyrer til Poissondelen ikkje følgjer mønsteret mellom dei strukturelle nullane og forklaringsvariabelen. For binomisk del er ZAP difor noko betre som korrekt enn ukorrekt modell. Men også med ZIP-fordelte observasjonsdatasett er det kun èin linkfunksjon ZAP estimerer for kvar komponent, og me ser at ZAP generelt er svært mykje betre enn ZIP på å modellere det binære utfallet, uavhengig av fordelinga i observasjonsdatasetta. For Poissondelen inkluderer ZAP kun observasjonane over null, og sjølv på ZIP-datasett der 0-observasjonane som eigentleg tilhøyrer Poissonkomponenten er tekne bort, vil dei resterande observasjonane vise den sanne innverknaden for dei ulike verdiane til forklaringsvariabelen. Det samla nullsannsynet har her kun indirekte innverknad ved at eit veldig høgt nullsannsyn

på små datasett kan for ZAP føre til for få observasjonar i Poissondelen. I regresjon treng ein eit visst antall observasjonar for å klare å få stabile og korrekte estimat. Unngår ein svært høge nullsannsyn klarar ZAP seg bra på Poissondelen både som korrekt og ukorrekt modell.

Diskusjonen over viser også kvifor ZIP og ZAP som regresjonsmodellar er svært like og ulike i akkurat dei situasjonane dei er det. I tilfelle med høg nok verdi for λ tillet ikkje ZIP nokon nullar i Poissondelen, men kategoriserar alle 0-observasjonane som strukturelle nullar. I modelleringa av det binære utfallet vert då alle nullane plassert i same gruppe, medan den andre gruppa inneheld kun observasjonar med verdi over 0. Dei to gruppene vert då like for ZIP og ZAP, uavhengig av det samla nullsannsyet. I tillegg er gruppa med observasjonane som inngår i Poissondelen identisk for dei to modellane. Høg nok forventningsverdi i Poissondelen fører difor til like situasjonar for ZIP og ZAP for begge komponentane i modellane, og dei estimerer i desse tilfella svært likt.

Der forventninga i Poissondelen er så låg at ZIP kategoriserer nokre observasjonar som vanlege nullar, har også det samla nullsannsynet innverknad på Poissondelen. Det er dei vanlege nullane som skil Poissondelen for ZIP og ZAP, og som gjer estimata ulike. Dersom me held forventninga i Poissondelen konstant, vil ei auke i samla nullsannsyn føre til fleire vanlege nullar for ZIP. Resultata våre viser at i desse situasjonane er det ofte den korrekte modellen som er den beste for Poissondelen. Dette kjem av at det er den modellen som bruker korrekt metodegrunnlag og som difor les strukturen best. I tillegg har ein ikkje i Poissonkomponenten dei to gruppene med 0-observasjonar som ofte øydelegg for ZIP-modellen. For tilfella med middels stor og stor differanse i MSE-verdiane til dei to modellane finn me god variasjon i forventninga til Poissondelen. Dette kjem av at det samla nullsannsynet har meir betydning enn forventningsverdien i dersom forventninga er for låg til at ZIP "unngår"vanlege nullar.

Drøftinga over omhandlar ikkje situasjonane der talet på 0-observasjonar er mindre enn det tilsvarande vanleg Poisson forutset. I dei tilfella fastset ZIP verdien til ϕ lik 0, og modellen har heller ikkje nok 0-observasjonar til å møte antatt mengd nullar i Poissondelen. Me står då tilbake med same problem som ved bruk av ordinær Poissonfordeling. I motsetning til ZIP klarar ZAP å tilpasse nullsannsynet også til å situasjonar med “for få” nullar. Dette kjem av at ZAP nyttar trunkert fordeling for Poissonkomponenten, og er difor ikkje avhengig av nullsannsynet for denne delen av modellen. ZAP treng likevel ei viss mengd 0-observasjonar som kan vise samanhengen responsvariabel og

forklaringsvariabel for den binomiske regresjonen. Det er lågt nullsannsyn samtidig med låg forventning for Poissondelen som gir den aktuelle situasjonen.

9.2 Val av regresjonsmodell

Ut frå resultata ser ein heilt klart at det samla nullsannsynet i lag med forventninga i Poissondelen har meir innverknad på kven av ZIP og ZAP som er best modell, enn fordelinga til observasjonsdatasettet. Begge modellane er som regel betre som korrekt enn ukorrekt, men ZAP er likevel generelt ein betre modell enn ZIP når me ser på begge komponentane under eitt. For estimeringa av koeffisientane i binomisk del er ZIP med nokre få unntak alltid dårlegare enn ZAP, og estimata er svært ustabile med mange høge MSE-verdiar for både koef. int og koef. forkl. ZIP er difor ikkje påliteleg som regresjonsverktøy uavhengig av fordelinga til datasettet. For Poissondelen er ZIP noko betre enn ZAP, men differansen i MSE-verdiane er ikkje stor, og med høg forventning er estimata svært like. Dersom ein i regresjonsanalysen er interssert i begge komponentane i modellane er difor ZAP alltid det klart beste valet. Dette er i stekt kontrast til tendensen me har funne i praktiske døme, der det som regel alltid er ZIP av dei to modellane som vert nytta i studia.

Ei auke i talet på observasjonar i datasetta me utfører regresjon på fører til generelt mykje likare estimering for ZIP og ZAP for Poissondelen. ærleg gjeld dette tilfelle med forventing over 3,0. Storleiken på datasetta har mindre innverknad på den binomiske delen av modellane, der er ZAP framleis ein klart betre modell enn ZIP. Ein bør difor også for store datasett generelt velje å nytte ZAP som regresjonvverktøy.

Kva resultata til modellane fortel oss er noko ulikt frå ZIP og ZAP, sidan dei analyserer forklaringavariablane opp mot ulike grupperingar av observasjonane. Det er difor ikkje sikkert at begge modellane gir svar på akkurat det me ynskjer å finne ut med regresjonsanalysen. Det er likevel alltid betre å nytte den modellen som gir oss pålitelege og korrekte resultat, enn modellen som gir oss den informasjonen me ynskjer, men der me ikkje kan stole på at det me lærer om variablane i analysen er sant.

Det kan tenkjast at i nokre praktiske analysestudium og forskningsarbeid er kun dei observasjonane for responsvariabelen med verdi over 0 av reell interesse. Dette kan t.d. vera tilfeller der 0-observasjonane kjem av feil i

innsamlingsprosedyren av data. Då er ikkje estimeringa av den binomiske delen av betydning, og me ser kun på estimata til koeffisientane i Poissondelen. Resultata frå analysen vår viser at dersom ein kun er interessert i observasjonane med verdi over 0, bør ein velje den av ZIP og ZAP som har lik fordeling som observsasjonsdatasettet. Men sidan den relative differansen for MSE-verdiane i dei fleste tilfella vil vera liten, vil ikkje konsekvensen for estimatverdiane vera stor ved val av gal modell i desse situasjonane.

I parameterkombinasjonar med færre nullar enn det ordinær Poissonfordeling foutset, er ZAP betydeleg betre enn ZIP for begge komponentane i modellane. Særleg for den binomiske delen gir ZIP svært ustabile estimatverdiar. I siuasjonar der me har for få 0-observasjonar til at ein bør nytte vanleg Poissonfordeling, bør ein difor alltid nytte ZAP og ikkje ZIP som regresjonsverktøy.

9.3 Konklusjon

Eit av hovudmåla med denne oppgåva var å finne ut om valet mellom ZIP og ZAP som analyseverktøy også ved praktisk anvending hadde betydning for kor gode estimat me får for regresjonskoeffisientane, og kor stor konsekvensen er ved val av gal modell. Me har funne ut at modellane i mange tilfeller gir svært like koeffisientestimat for Poissondelen, særleg i tilfeller med høg forventning. Med svært høg forventning har difor ikkje val av modell nokon praktisk betydning verken for koef. int. eller koef. forkl. for Poissondelen. Val av modell er derimot av stor betydning for MSE-verdien for koeffisientane i binomisk del. Her er den relative differansen til MSE-verdiane i dei aller fleste tilfella stor.

Me ynskte også å finne ut om val av modell burde takast utfrå eit anna grunnlag enn fordelinga til observasjonsdatasettet, t.d strukturelle tendensar i datasetta. Konklusjonen er at dersom ein kun er interesert i koeffisientane til Poissondelen, bør ein velje den modellen som har same fordeling som responsvariabelen. Men i tilfelle med svært høg forventning har val av modell svært liten betydning. Dersom du i tillegg er interessert i den binomiske delen av modellane bør ein alltid velje ZAP som analysemodell. Nullsannsynet har liten innverknad på kva modell ein bør velje, men har betydning for mykje betre den beste modellen er enn den med dårlegast MSE. Dette same gjeld for storleiken på observasjonsdatasetta.

I tilfelle med for få 0-observasjonar til at ordinær Poissonregresjon er eit bra val, bør ein i valet av ZIP og ZAP alltid velge ZAP.

Kapittel 10

Avslutning og forslag til vidare arbeid

Me har i denne oppgåva teke utgangspunkt i to ariklar av Mullahy og Lambert og samanlikna zero-inflated og zero-altered Poissonfordeling. Dette er gjort både i teoretisk samanheng og ved hjelp av praktisk analysearbeid. Me har sett på korleis me i dei fleste tilfella ut frå reponsvariabelen kan avgjere kva modell som gir mest korrekt estimat for regresjonskoeffisientane. Særleg har me hatt fokus på korleis samla nullsannsyn og forventninga i Poissondelen har innverknad på kor like modellane er i praksis.

Kva me kallar gode og dårlege estimat og MSE-verdiar i denne analysen er sett i samanheng med den totale variasjonsbreidda i avviket estimtata har frå dei korrekte regresjonskoeffisientane. Me ser i oppgåva på kva betydning valet av modell har for estimatverdiane og tilhøyrande MSE-verdi, men kva differansen i estimata har å seie i praksis for konklusjonen om signifikans for forklaringsvariabelen har me ikkje sett på. Ein MSE-verdi på 0,5 vert i vår analyse sett på som liten, grunna at me at også har MSE-verdiar på over 500. Men med tanke på kva verdiar koeffisientane til forklaringsvariablane pleier å ligge på i regresjonanalyse, er det sannsynleg at også ein så liten differanse vil få konsekvens for om modellane konkluderer med signifikans eller ikkje. Kva valet av modell har å seie for dei meir praktiske resultata er difor ein naturleg og svært interessant del av analysen rundt ZIP og ZAP-modellane. Ei masteroppgåve er dog diverre tidsmessig avgrensa, og det vart grunna mykje nødvendig arbeid med å få god oversikt over metodegrunnlaget til modellane både før, under og etter utføringa av den praktiske analysen ikkje funne tid til dette. Dette er likevel svært viktige spørsmål med tanke på kva modell ein bør velje, og vil vera eit naturleg steg med tanke på vidare arbeid. Det vil då vera naturleg å sjå på kor mange tilfelle av dei 10 000

simuleringane der både ZIP- og ZAP-regresjon på same datasettet fører til signigikans for forklaringsvariabelen. Dette vil vera interessant å sjå på opp mot dei ulike strukturelle tendensane i datasetta som me har funne har innverknad på differansen i modellane. Dersom ein primært er interessert i kun Poissondeleni, kan ein sjå på ZIP og ZAP som ein måte å "korrigere" ein situasjon med Poissonregresjon ved å enten trunkere ved 0 for ZAP, eller modellere dei strukturelle nullane separat som i ZIP. Det vil då vera rettferdig å samanlikna p-verdiane for Poissondelen til dei to modellane. For samanlikninga av signifikante resultat for binomisk del må ein bruke ei meir overordna samanlikning, grunna at ZIP og ZAP bruker to svært ulike uttrykk for samla nullsannsyn. Ein lyt i tillegg passe på at modellane for begge komponentane har likt nivå i styrkefunksjonen til hypotesetesten for verdien til koef. forkl.

Eit anna naturleg val med tanke på vidare arbeid er å innføre fleire forklaringsvariablar i analysesituasjonen. Det er også interessant å sjå på om tendensane me har funne i resultata også stemmer for zero-inflated og zero-altered regresjon med negativ binomisk som hovudfordeling. Moglegvis vil ZAP også då vera like suveren på den binomiske delen, men det kan også tenkjast at den ekstra parameteren i negativ binomisk fordeling samanlikna med Poisson gjere modellane meir like for begge komponentane. Zero-inflated og zero-altered Poissonmodell vert meir og meir populære og stadig oftare nytta i praktisk analysearbeid, men me ser eit behov for meir innsikt i og betre grunnleggjande forståing av dei to modellane.

Tillegg A

utledning av forteiknsinnverknaden i ZIP

I analysedelen i oppgåva ser me på situasjonar med same forklaringsvariabel for både Poissondelen og den binomiske delen. For ZIP kan det oppstå tilfelle der same verdi for forklaringsvariabelen vil gi høgt samla nullsannsyn samtidig som høg forventning i Poissondelen. Me vil her utleie prov for at dette skjer når me har like forteikn for dei to tilhøyrande regresjonskoeffisientane. Verdien til forklaringsvariabelen vel me å setje lik ein, og regresjonskoeffisentane for intercept-ledda held me konstant.

For å få høg forventning i Poissondelen treng me høg verdi for λ. Frå (3.1.2) får me $\lambda_i = e^{\delta+\rho H_i} = e^{\delta+\rho}$ ved innføring av kun ein forklaringsvariabel og ved å setje $H = 1$. Det viser at stigande verdi for ρ fører til auka verdi for λ. Ein lyt difor ha positivt forteikn for ρ for å få høg forventning i Poissondelen.

Lat oss no sjå på kva forteikn som må til for å få stigande samla nullsannsyn. Ved å setje verdien til begge forklaringsvariablane lik ein får me frå (3.1.2)

$$\begin{aligned} P(Y=0) &= \frac{e^{\iota+\omega G_i}}{1+e^{\iota+\omega G_i}} + (1 - \frac{e^{\iota+\omega G_i}}{1+e^{\iota+\omega G_i}}) \left(\exp(-e^{\delta+\rho})\right) \\ &= \frac{e^{\iota+\omega}}{1+e^{\iota+\omega}} + (1 - \frac{e^{\iota+\omega}}{1+e^{\iota+\omega}})(\exp(-e^{\delta+\rho}) \end{aligned}$$

Alle ledda i uttrykket vil alltid vera positive. Det gjer at det første leddet og den første faktoren i det andre leddet arbeider mot kvarandre. Ein lyt difor finne ut kva ledd som er det sterkaste og som lyt bestemme forteiknet til ω for å skape den aktuelle situasjonen. Me ynskjer høgast mogleg forventning i Poissondelen samtidig med høgast mogleg samla nullsannsyn. Frå Poissondelen veit me at ρ må vera positiv. Dess høgre verdi for ρ, dess høgare verdi

for λ. Men dette vil føre til at det siste leddet i uttrykket gir mindre og mindre bidrag til høg verdi for $P(Y = 0)$. For å få det samla nullsannsynet så høgt som mogleg samtidig med stigande λ, må me få det første leddet så stort som mogleg. Då treng me stigande verdi for ω. Me har då at like forteikn for ω og ρ gir høgt samla nullsannsyn samtidig med høg forvetning i Poissondelen, og same verdi for forklaringsvariabelen trekk i motsett retning for dei to komponentane.

Tillegg B

Programkode

B.1 Generering av ZIP- og ZAP-fordelte observasjonsdatasett

```
generer <- function(fordeling, antObsPr, x_forkl, lambdaint,
lambdaforkl, nullkompint, nullkompforkl) {

# fordeling avgjer om ein bruker ZIP- eller ZAP-fordeling
# antObsPr gir tal på observasjonar for kvar parameterkomb.
# x_forkl er vektor med verdiane til forklaringsvariabelen
# lambdaint, lambdaforkl, nullkompint og nullkompforkl
# gir verdien til dei fire regresjonskoeffisientane

antObs <- antObsPr*length(x_forkl)  #Antall observasjonar

#Vektorar til lagring av respons- og tilhøyrande forklaringsvariabel
y <- array(dim = antObs)
x <- array(dim = antObs)

#--------Genererer datasettet------------

k <- 1  #Tilpasser x- og y-vektorane i forhold til kvarandre

for ( i in 1: length(x_forkl)) {
p <- c(k:(k+antObsPr-1))

#Reknar ut og lagrar verdien av lambdakomponenten
lambda <- exp(lambdaint + lambdaforkl*x_forkl[i])
```

```
#Reknar ut og lagrar verdien av nullkomponenten (p og phi)
nullkomp <- exp(nullkompint + nullkompforkl*x_forkl[i])
     /(1+ exp(nullkompint + nullkompforkl*x_forkl[i]))

#Genererer y for denne kombinasjonen av kovariatane
if(fordeling == "zip"){y[p]<-rzipois(antObsPr,lambda,nullkomp)}
if(fordeling == "zap"){y[p]<-rzapois(antObsPr,lambda,nullkomp)}

#Lagrar den brukte kombinasjonen av kovariatar
x[p] <- x_forkl[i]

k = k+antObsPr
}

#Lagar og returnerer matrise av datasettet
matrisedata <- cbind(y,x)
return(matrisedata)
}
```

B.2 Simulering med ZIP- og ZAP-regresjon

```
library(VGAM)
library(pscl)
seedet <- 1
set.seed(seedet)

# -- Fastsetting av variablar --
#Vektoren med verdiane til forklaringsvariabelen
x_variabel = c(-1,-0.5,0,0.5,1)

#Regresjonskoeffisientane
lambda_int <- -.5 #Reg. koeff intercept Poisson-delen
lambda_forkl <- 0.1 #Reg. koeff forkl. Poisson-delen
null_int <- -0.1 #Reg. koeff intercept Bernoulli-delen
null_forkl <- -2.4 #Reg. koeff forkl. Bernoulli-delen

#Variablar som styrer simuleringa
antObsPr <- 10  #Antall obs. for kvar kombinasjon av kovariatane
antSim <- 5  #Antall simuleringar i analysen
form <- y ~ x | x  #Formula til regresjonsfunksjonen

# -- Finn lambda, nullkomponent og nullsannsyn (zip) --
lambda <- array(dim = length(x_variabel))
```

```
nullkomp <- array(dim = length(x_variabel))
nullsann <- array(dim = length(x_variabel))

for(j in 1:length(x_variabel)) {
lambda[j] <- exp(lambda_int + lambda_forkl*x_variabel[j])
nullkomp[j] <- exp(null_int + null_forkl*x_variabel[j]) /
     (1+ exp(null_int + null_forkl*x_variabel[j]))
nullsann[j] <- nullkomp[j] + (1-nullkomp[j])*exp(-lambda[j])
}

# -- Simuleringsdelen --

#Vektorar for lagring av regresjonkoeffisentane
zapT <- matrix(nrow = antSim, ncol =20)
zipT <- matrix(nrow = antSim, ncol = 20)

#Matriser for å lagre resultata frå denne simuleringa
Resultat <- matrix(nrow = 20, ncol = 2)
sim_var <- array(dim = 10)

## Køyrer to løkker, ei med ZAP-datasett og ei med ZIP-datasett

for (j in 1:antSim) {
fordeling <- "zap"

for(w in 1:2) {

# Lagar datasett
datasettet <- generer(fordeling, antObsPr, x_variabel,
lambda_int, lambda_forkl, null_int, null_forkl)
frame_datasett <- as.data.frame(datasettet)

# Køyrer regresjon
if(is(zip_reg,"try-error")) {next}
zap_reg <- try(hurdle(formula = form, dist ="poisson",
     link = "logit", data = frame_datasett), silent=T)
if(is(zap_reg,"try-error")) {next}
zip_reg <- try(zeroinfl(form, dist ="poisson", frame_datasett,
     control = zeroinfl.control(EM = TRUE)), silent=T)
pois_reg <- try(glm(formula = y ~ x, family = poisson,
     data = frame_datasett), silent=T)
if(is(pois_reg,"try-error")) {next}
```

```
# Hentar ut resultat frå denne simuleringa
if(fordeling == "zap") {
zapT[j,1] <- zap_reg$coef$count["(Intercept)"]
zapT[j,2] <- (zap_reg$coef$count["(Intercept)"] - lambda_int)^2
zapT[j,3] <- zap_reg$coef$count["x"]
zapT[j,4] <- (zap_reg$coef$count["x"] - lambda_forkl)^2
zapT[j,5] <- -zap_reg$coef$zero["(Intercept)"]
zapT[j,6] <- ((-zap_reg$coef$zero["(Intercept)"]) - null_int)^2
zapT[j,7] <- -zap_reg$coef$zero["x"]
zapT[j,8] <- ((-zap_reg$coef$zero["x"]) - null_forkl)^2
zapT[j,9] <- zip_reg$coef$count["(Intercept)"]
zapT[j,10] <- (zip_reg$coef$count["(Intercept)"] - lambda_int)^2
zapT[j,11] <- zip_reg$coef$count["x"]
zapT[j,12] <- (zip_reg$coef$count["x"] - lambda_forkl)^2
zapT[j,13] <- zip_reg$coef$zero["(Intercept)"]
zapT[j,14] <- (zip_reg$coef$zero["(Intercept)"] - null_int)^2
zapT[j,15] <- zip_reg$coef$zero["x"]
zapT[j,16] <- (zip_reg$coef$zero["x"] - null_forkl)^2
zapT[j,17] <- pois_reg$coef["(Intercept)"]
zapT[j,18] <- (pois_reg$coef["(Intercept)"] - lambda_int)^2
zapT[j,19] <- pois_reg$coef["x"]
zapT[j,20] <- (pois_reg$coef["x"] - lambda_forkl)^2
}

if(fordeling == "zip") {
zipT[j,1] <- zap_reg$coef$count["(Intercept)"]
zipT[j,2] <- (zap_reg$coef$count["(Intercept)"] - lambda_int)^2
zipT[j,3] <- zap_reg$coef$count["x"]
zipT[j,4] <- (zap_reg$coef$count["x"] - lambda_forkl)^2
zipT[j,5] <- -zap_reg$coef$zero["(Intercept)"]
zipT[j,6] <- ((-zap_reg$coef$zero["(Intercept)"]) - null_int)^2
zipT[j,7] <- -zap_reg$coef$zero["x"]
zipT[j,8] <- ((-zap_reg$coef$zero["x"]) - null_forkl)^2
zipT[j,9] <- zip_reg$coef$count["(Intercept)"]
zipT[j,10] <- (zip_reg$coef$count["(Intercept)"] - lambda_int)^2
zipT[j,11] <- zip_reg$coef$count["x"]
zipT[j,12] <- (zip_reg$coef$count["x"] - lambda_forkl)^2
zipT[j,13] <- zip_reg$coef$zero["(Intercept)"]
zipT[j,14] <- (zip_reg$coef$zero["(Intercept)"] - null_int)^2
zipT[j,15] <- zip_reg$coef$zero["x"]
zipT[j,16] <- (zip_reg$coef$zero["x"] - null_forkl)^2
zipT[j,17] <- pois_reg$coef["(Intercept)"]
zipT[j,18] <- (pois_reg$coef["(Intercept)"] - lambda_int)^2
```

```
zipT[j,19] <- pois_reg$coef["x"]
zipT[j,20] <- (pois_reg$coef["x"] - lambda_forkl)^2
}

fordeling <- "zip"
}
}

# Finn endelige resultat
for(m in 1:20) {
Resultat[m,1] <- mean(zapT[,m], na.rm = TRUE)
Resultat[m,2] <- mean(zipT[,m], na.rm = TRUE)
}
Resultat_avrund <- round(Resultat, digits = 10)

#Hentar ut variablar som styrer simuleringane
sim_var[1] <- mean(lambda)
sim_var[2] <- mean(nullsann)
sim_var[3] <- mean(nullkomp)
sim_var[4] <- lambda_int
sim_var[5] <- lambda_forkl
sim_var[6] <- null_int
sim_var[7] <- null_forkl
sim_var[8] <- antObsPr
sim_var[9] <- antSim - sum(!is.na( zapT[,1]) )
sim_var[10] <- seedet

# -- Skriv ut resultata --

sim_var
x_variabel
Resultat_avrund
```

Litteratur

[1] The r project for statistical computing. `http://www.r-project.org/index.html`, Mai 2011.

[2] 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(5):829–844, 1971.

[3] William C. Davidon. Variable metric method for minimization. *SIAM Journal on Optimization*, 1(1):1–17, 1991.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, 39(1):1–38, 1977. With discussion.

[5] J. E. Dennis, Jr. and Robert B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*, volume 16 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. Corrected reprint of the 1983 original.

[6] D.E. Goodrich, Z. Lai, E. Lasky, A.R. Burghardt, and A.M. Kilbourne. Access to weight loss counseling services among patients with bipolar disorder. *Journal of affective disorders*, 126(1-2):75–79, 2010.

[7] D 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.

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

[9] Thieu NQ Hung NM Hung LX Hay SI Hien TT Wertheim HF Snow RW Horby P. Manh BH, Clements AC. Social and environmental determinants of malaria in space and time in viet nam. *International Journal for Parasitology*, 2010.

[10] John Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986.

[11] Isaac Newton. *The method of fluxions and infinite series : with its application to the geometry of curve-lines.* London : Printed by H. Woodfall, 1736.

[12] Joseph Raphson. *Analysis Aequationum universalis.* Londinum, 1690.

[13] V. Shankar, J. Milton, and F. Mannering. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. *Accident Analysis & Prevention*, 29(6):829–837, 1997.

[14] G. Solinas, G. Campus, C. Maida, G. Sotgiu, M.G. Cagetti, E. Lesaffre, and P. Castiglia. What statistical method should be used to evaluate risk factors associated with dmfs index? evidence from the national pathfinder survey of 4-year-old italian children. *Community Dentistry and Oral Epidemiology*, 37(6):539–546, 2009.

[15] S. Ullah, C.F. Finch, and L. Day. Statistical modelling for falls count data. *Accident Analysis & Prevention*, 42(2):384–392, 2010.

[16] T.W. Yee. Vglms and vgams: an overview for applications in fisheries research. *Fisheries Research*, 101(1-2):116–126, 2010.

[17] A. Zeileis, C. Kleiber, and S. Jackman. Regression models for count data in r. *Journal of Statistical Software*, 27(8):1–25, 2008.

[18] 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.* Springer Verlag, 2009.