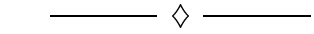
Med fokus på Dynamiske Modeller



Af Bo Eskerod Madsen

Januar 2002

Aalborg Universitet
Institut For Matematiske Fag · Fredrik Bajers Vej 7G
9220 Aalborg Øst

Det teknisk-naturvidenskabelige fakultet
Aalborg Universitet

Institut for matematiske fag

TITEL: Med fokus på Dynamiske Modeller

Af:
Bo Eskerod Madsen

VEJLEDERE:
Søren Lundbye-Christensen
Thomas Scheike

PERIODE:
1. september 2001 -
11. januar 2002

OPLÆG: 10 stk.

ANTAL SIDER: 121

SYNOPSIS:

Det overordnede mål med dette speciale er, at evaluere visse statistike teorier og metoder, samt illustrere dem ved hjælp af en række analyser.
Den teoretiske del af specialet omhandler dynaimske modeller, også kendt som state spece modeller. Opbygningen af den teoretiske del er udformet således, at der først introduceres en simpel dynamisk lineær model (DLM) med Gaussiske fordelinger. Dernæst introduceres den generaliserede lineære model (GLM), som er en udvidelse af den lineære model (LM) med Gassiske observationer. Ud fra disse to modeller udledes den dynamiske generaliserede lineære model (DGLM). Udvidelsen fra den DLM til den DGLM er tilsvarende udvidelsen fra den LM til den GLM. Således har den DGLM, ligesom den GLM, ikke nødvendigvis Gaussiske fordelinger. For de ovenstående modeller, indføres forskellige metoder til inferens.
Den teoretiske del af specialet er illustreret dels ved nogle teoretiske og simuleringsbaserede eksempler, og dels ved en dataanalyse. Dataanalysen handler om skizofrenitilfælde, og er udarbejdet i samarbejde med Center for Registerforskning ved Aarhus Universitet. I dataanalyserne anvendes såvel en GLM, som en DGLM.

Forord

Det overordnede mål med dette speciale er, at evaluere visse statistiske teorier og metoder, samt illustrere dem ved hjælp af en række analyser. Der har i specialet været lagt stor vægt på samarbejdet med Center for Registerforskning ved Aarhus Universitet, hvor den praktiske del af specialet er foregået. Af fortrolighedsmæssige årsager må data ikke tages med uden for centeret, hvorved alle dataanalyser er udført der, under vejledning af lektor Rodrigo Labouriau. Da Center for Registerforskning anvender statistikpakken SAS, er alle praktiske analyser lavet i dette program. Det har af denne grund ikke være muligt, at anvende de beskrevne estimationsmetoder, indført af Durbin and Koopman [2000], men i stedet er anvendt de metoder, som er implementeret i SAS pakken. Dog har jeg selv bla. implementeret visse algoritmer i forbindelse med det udvidede Kalmanfilter med iterationer for, at illustrere dele af teorien. Det er ikke oplagt, at anvende de beskrevne dynamiske modeller på de data, som er analyseret. Af pædagogiske grunde, og for at illustrere den beskrevne teori, er det dog valgt, at gøre dette alligevel.

Først i november måned har Søren Lundbye-Christensen desværre måttet sygemelde sig, og har dermed været nød til, at trække sig ud, som hovedvejleder på specialet. Herefter har Thomas Scheike overtaget hovedvejledningen. Af denne grund er der to hovedvejledere tilknyttet specialet. Som det fremgår af alle de mennesker der har været så søde, at give mig en hånd, har dette dog givet nogle udemærkede samarbejder på tværs.

Sætninger, definitioner, korollarer og metoder er nummereret fortløbende gennem hvert kapitel, mens ligninger har en egen nummerering gennem hvert kapitel. Kildehenvisninger inde i teksten angives ved kilde [år, evt. henvisning] , mens kilder til citater er angivet ved [kilde, år, evt. henvisning]. Hvis kildehenvisningen gives før et punktum, refererer den kun til den pågældende sætning, hvis den derimod gives efter et punktum, refererer den til det pågældende afsnit.
Litteraturlisten er ordnet alfabetisk efter forfatterens efternavn.
Den notation som er benyttet i specialet, er opsummeret i appendix E.
Ord inde i teksten, der er henvist til fra stikordsregisteret, er markeret med fed type.

Under udarbejdelsen af dette speciale er der blevet udleveret en disposition der er stort set ens med dispositionen til Larsson [2001]. Desuden har Larsson [2001] været anvendt som undervisningsmateriale i en studiekreds i forbindelse med specialet. Jeg har derfor valgt, at citere alt hvad de stort set viser på samme måde som jeg ville have gjort, og tage de ting med, hvor jeg viser noget anderledes.

Jeg har valgt ikke at trykke de algoritmer jeg har bygget i forbindelse med analyserne, da det ville fylde ganske betragteligt. Hvis nogen skulle ønske, at se disse algoritmer igennem vil jeg dog meget gerne sende dem.

Jeg vil gerne rette en stor tak til Thomas Scheike for, at ville overtage vejledningen af mit speciale så sent i perioden, og for den gode vejledning han har givet, Rodrigo Labouriau, fra Center for Registerforskning, for vejledning i forbindelse med opholdet på centeret, og til centerleder Preben Bo Mortensen for, at give mig muligheden for, at skrive mit speciale i samarbejde med Center for Registerforskning. Desuden har Bettina Kirk Øgendahl, fra Center for Registerforskning, været en god hjælp ved nogle ganske oplysende diskussioner omkring psykiatriske faktorer i skizofrenidiagnostisering, og Claus Dethlefsen, fra Aalborg universitet, har været i kærkommen hjælp, til at finde fejlen i mit programmel. Sidst men ikke mindst vil jeg gerne takke Søren Lundbye-Christensen for en altid god og engageret vejledning.

Aalborg, Januar 2002

Bo Eskerod Madsen

Indhold

1 Indledning

De modeller der betragtes i dette speciale, er overvejende til, at analysere processer der udvikler sig over tid. Data der er opsamlet i forbindelse med sådanne processer betegnes **longitudinelle data**, eller **tidsrækker**, hvis der kun er tale om en enkelt observation til hver tid. Da data for en proces der udvikler sig over tid, typisk er serielt korrelerede, er de klassiske modeller for ukorrelerede data, ikke passende til beskrivelse af longitudinelle data. Istedet kan anvendes **dynamiske modeller**.
De dynamiske modeller, som introduceres i dette speciale, betegnes i dele af litteraturen **state space modeller**. Disse modeller er opdelt, således at der er en **observationsligning** og en **systemligning**. Observationsligningen beskriver hvorledes observationerne er fordelt, ud fra nogle stokastiske dynamiske parametre. Deres udvikling over tid beskrives ved den tilhørende systemligning. Således kan strukturen for de beskrevne dynamiske modeller opfattes, som om der hele tiden kører en stokastisk underliggende **latent proces**. Til en given tid trækkes der en **tilstandsvektor** ud af denne latente proces. Ud fra tilstandsvektoren bestemmes de stokastiske parametre, som indgår i observationsligningen, hvorved der findes en fordeling af de observationer der foretages, til den givne tid. Ofte vil det være interessant, at kende den underliggende latente proces. Derfor beskæftiger en stor del af de metoder, som præsenteres i dette speciale, sig da også med estimation af denne latente proces.

For at kende forskel mellem de stokastiske og de ikke stokastiske parametre i modellerne, betegnes de ikke stokastiske parametre **hyperparametre**.

Inferens for dynamiske modeller indeholder følgende punkter.

Estimation af den latente proces er at bestemme den latente proces.

Dvs. tilstandsvektorerne der hører til de stokastiske parametre i modellen.

Estimation af ukendte hyperparametre er, at bestemme ukendte ikke stokastiske parametre i modellen.

Hypotesetest

Modelkontrol.

Der er udviklet mange metoder til estimation af såvel den latente proces, som ukendte hyperparametre, hvorimod hypotesetest for stokastiske dynamiske parametre, samt modelkontrol er et mere uudforsket område.

For at estimere den latente proces anvendes, til tiden t, følgende tre begreber.

Filtrering. Dette er, at estimere den latente proces til tiden t.

Prædiktion. Dette er, at estimere den latente proces, og observationsprocessen, senere end til tiden t.

Udglatning. Dette er, at estimere den latente proces op til tiden t.

De data der skal analyseres udgøres af to tidsrækker, hvoraf det kun er den ene tidsrække, som de præsenterede dynamiske modeller kan anvendes for. Af denne grund introduceres de teoretiske metoder i dette speciale for tidsrækker, dvs. univariate observationer.

I den teoretiske del, af specialet, introduceres først den dynamiske lineære model (**DLM**), hvor fordelingerne i observationsligningen og systemligningen er Gaussiske. Dernæst introduceres den generaliserede lineære model (**GLM**), som er en udvidelse af den lineære model (**LM**) med Gaussisk fordeling. Principperne for udvidelsen fra den LM til den GLM anvendes herefter til, at udvide den DLM til den dynamiske generaliserede lineære model (**DGLM**). I denne model gælder at fordelingerne, ligesom i den GLM, ikke nødvendigvis er Gaussiske.

Under de respektive modeller indføres metoder til estimeres den latente proces. Efter introduktionen af modellerne indføres alternative metoder til estimering af den latente proces, samt estimering af ukendte hyperparametre. Disse metoder bygger på importance sampling, og er introduceret i Durbin and Koopman [2000].

Af tidsmæssige årsager er hypotesetest for stokastiske dynamiske parametre, og modelkontrol for dynamiske modeller ikke behandlet i dette speciale. Der anvendes dog visse testmetoder for ikke dynamiske parametre. Disse

metoder hører dog ikke decideret til dynamiske modeller, men snarer til statiske modeller.

Den teoretiske del af specialet er illustreret dels ved nogle teoretiske og simuleringsbaserede eksempler, og dels ved en dataanalyse. Dataanalysen handler om skizofrenitilfælde, og er udarbejdet i samarbejde med Center for Registerforskning ved Aarhus Universitet. I dataanalyserne anvendes såvel en generaliseret lineær model, som en dynamisk generaliseret lineær model.

Resumé

Specialet er inddelt i fire dele om henholdsvis teori, illustrerende eksempler til teorien, anvendelse af teorien og til sidst appendiks, som omhandler forskelligt baggrundsinformation. De fire dele er beskrevet i det følgende.

Del I. Teori

Denne del indeholder den teoretiske baggrund for specialet.

I kapitel 2 gives en introduktion til dynamiske lineære modeller, og der præsenteres eksakte metoder til estimering af den latente proces. Der behandles filtrering, prædiktion og udglatning ved hjælp af Kalmanfilteret.

I kapitel 3 gives en introduktion til generaliserede lineære modeller. Der præsenteres metoder til estimation, hypotesetest, samt modelkontrol.

I kapitel 4 præsenteres dynamiske generaliserede lineære modeller. Disse modeller behandles generelt, og i et specialtilfælde, hvor systemligningen er Gaussisk. Der gennemgåes i begge tilfælde approksimative metoder til filtrering.

I kapitel 5 præsenteres estimering baseret på importancesampling. Dette kan anvendes til bla. at estimere den latente proces, og ukendte hyperparametre.

Del II. Illustrerende eksempler

Denne del indeholder illustrerende eksempler til den teori, som er præsenteret i første del.

I kapitel 6 illustreres den DGLM for henholdsvis poissonfordelte observationer og binomialfordelte observationer. For begge fordelinger illustreres hvorledes den generelle teori kommer til at se ud, for den DGLM med delvist specificeret systemligning, og den DGLM med Gaussisk systemligning.

Desuden gennemgås et simuleringsbaseret eksempel for den DGLM med binomialfordelte observationer og Gaussisk systemligning.

Del III. Anvendelse

Anvendelsen af teorien er illustreret ved behandling af et datasæt om skizofrenitilfælde i Danmark. Dataanalysen er udført i samarbejde med Center for Registerforskning ved Aarhus Universitet. I behandlingen ønskes en analyse af, hvorvidt fødselsmåneden og fødselsåret har indvirkning på sandsynligheden for at få tidlig skizofreni. Der analyseres data for personer født i årene 1955 til 1973.

I kapitel 7 gennemgåes baggrunden for dataanalysen, og der gives en kort introduktion til diagnosen for skizofreni, samt mulige faktorer der kan tænkes, at have indvirkning på skizofrenifrekvensen.

I kapitel 8 analyseres data på basis af den GLM. Her anvendes forskellige metoder, til at undersøge mulige afhængighedsstrukturer i data.

I kapitel 9 analyseres data på basis af den DGLM med Gaussisk systemligning.

Kapitel 10 indeholder en opsummering af den samlede dataanalyse, samt en videre diskussion af analyseresultaterne.

Del IV. Appendiks

Denne del indeholder en mængde resultater, som anvendes i de andre dele af specialet.

Appendiks A indeholder en introduktion til poissonfordelingen og binomialfordelingen. Desuden indeholder det en mængde anvendte egenskaber i forbindelse med disse fordelinger.

Appendiks B indeholder en introduktion til profil likelihoodfunktioner, samt en meget kort introduktion til likelihoodfunktioner.

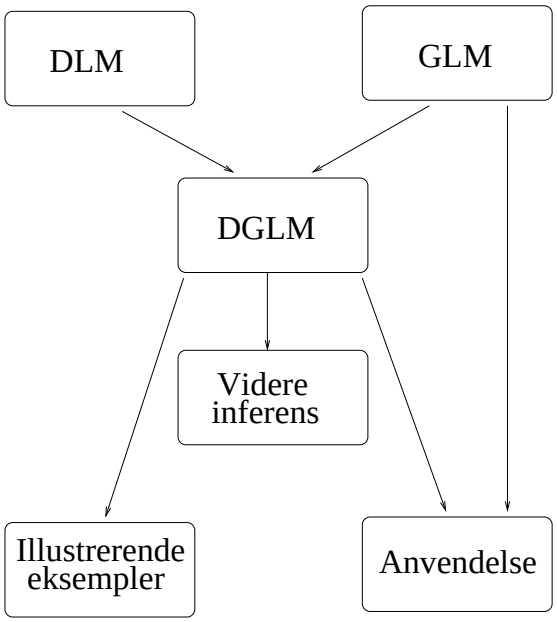
Appendiks C indeholder en gennemgang af de betegnelser der anvendes i analysetabellerne i anvendelsesdelen.

Appendiks D indeholder sætninger, som ikke passer ind under de andre appendiks.

Det sidste er appendiks E, der indeholder en gennemgang af den notation, som er anvendt i dette speciale.

Overordnet

Opbygningen af specialet er illustreret i figur 1.1. Denne figur illustrerer hvorledes de enkelte elementer i specialet hænger sammen. Det skal specielt bemærkes, at elementet indeholdene videre inferens ikke leder videre til andre dele. Dette skyldes, at der ikke har være mulighed for, at implementere algoritmerne i forbindelse med dataanalyserne. Emnet er dog behandlet alligevel, da det er repræsentativt for den type metoder, der arbejdes med at udvikle i denne computeralder.



Figur 1.1: Opbygningen af specialet. Boksene illustrer elementerne, mens pilene illustrerer sammenhængene.

Del I

Teori

2 Dynamiske lineære modeller

Dette kapitel indeholder en introduktion til den dynamiske lineære model (DLM). Dette er en dynamisk model med Gaussisk systemligning og observationsligning.

Da denne model har nogle gode beregningsmæssige egenskaber, tages der udgangspunkt i den for, at indføre mere avancerede modeller. Der introduceres ganske kort metoder til estimering af den latente proces. Disse metoder er bla. indført af Kalman, og betegnes dermed Kalmanfiltrering, Kalmanprædiktion og Kalmanudglatning.

Da [Larsson, 2001] har vist sætningerne rimelig korrekt efter West and Harrison [1997], undlades beviserne her. Istedet refereres direkte til West and Harrison [1997], hvor beviserne er nydeligt gennemgået. Dette kapitel skal dermed ikke ses som en ny gennemgang af teorien, men blot en præsentation af den, til senere brug.

2.1 Modellen

For at introducere den DLM, skal informationen først defineres.

Definition 2.1 **(Informationen $\mathbf{D_t}$)**
Lad D_0 betegne **begyndelsesinformationen** *til en dynamisk model for serien $\{Y_t\} = \{Y_1, \ldots, Y_r\}$. Størrelsen D_0 er da alt relevant information til tiden $t = 0$, dvs. den f.eks. indeholder kendte hyperparametre.*
For ethvert $t > 0$ betegner D_t informationen, som er en kombinationen af begyndelsesinformationen og realisationerne af $Y_1, \ldots, Y_t$. Dvs. D_t er informationen bestående af D_{t-1} og y_t. △

Ideen i den DLM er, som i alle dynamiske modeller, at betragte $\{Y_t\}$ som en serie observationer med stokastiske, dynamiske regressionsparametre $\{\boldsymbol{\theta}_t\}$. Herved kommer $\{Y_t\}$ til, at optræde som en observationsproces, relateret til en underliggende latent proces $\{\boldsymbol{\theta}_t\}$. For hvert tidstrin t tilføjes en stokastisk **udviklingsvariation** $\boldsymbol{\omega}_t$ til **tilstandsvektoren** $\boldsymbol{\theta}_t$ i den latente proces, hvorved denne udvikler sig som en første ordens Markovkæde. Hver gang der observeres, indgår ligeledes en stokastisk **observationsvariation** ν_t i observationen Y_t, hvorved denne bliver stokastisk, i relation til den latente proces. Den DLM er givet jvf. West and Harrison [1997, s. 102], som følger.

Definition 2.2 **(Den DLM)**
Lad der til enhver tid $t \in \mathbb{Z}_+$ være givet følgende hyperparametre.

- $\boldsymbol{F}_t$*, som er en kendt n-dimensional dynamisk regressionsvektor.*
- $\boldsymbol{G}_t$*, som er en kendt $n \times n$ evolutionsmatrix.*
- V_t*, som er en kendt positiv observationsvarians.*
- $\boldsymbol{W}_t$*, som er en kendt positiv definit $n \times n$ evolutionscovariansmatrix.*

For hvert $t \in \mathbb{Z}_+$ er den DLM, for observationerne $\{Y_t\}$ og de n-dimensionale stokastiske parametervektorer $\{\boldsymbol{\theta}_t\}$, defineret ved

$$\begin{aligned}
&\textit{Observationsligning:} && Y_t = \boldsymbol{F}_t^T \boldsymbol{\theta}_t + \nu_t, && \nu_t \sim \mathcal{N}(0, V_t),\\
&\textit{Systemligning:} && \boldsymbol{\theta}_t = \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, && \boldsymbol{\omega}_t \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{W}_t),\\
&\textit{Begyndelsesbetingelse:} && (\boldsymbol{\theta}_0 | D_0) \sim \mathcal{N}_n(\boldsymbol{m}_0, \boldsymbol{C}_0),
\end{aligned}$$

for givne værdier af $\boldsymbol{m}_0$ og $\boldsymbol{C}_0$.

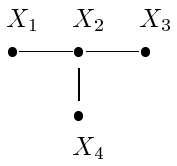
Der gælder endvidere, at $(Y_t|\boldsymbol{\theta}_t)$ er uafhængig af $Y_1, \ldots, Y_{t-1}, Y_{t+1}, \ldots, Y_r$, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_r$, at Y_t kun afhænger af $\boldsymbol{\theta}_t$ gennem λ_t, samt at $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_r, \boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r, (\boldsymbol{\theta}_0|D_0)$ er uafhængige. △

Bemærk at begyndelsesbetingelsen præsenterer viden om tilstandsvektoren $\boldsymbol{\theta}_0$ inden der foretages observationer. Begyndelsesinformationen D_0 i den

DLM er dermed givet ved $\boldsymbol{F}_t$, $\boldsymbol{G}_t$, V_t, $\boldsymbol{W}_t$, $\boldsymbol{m}_0$ og $\boldsymbol{C}_0\}$. Hvis der ikke foreligger nogen information, hvor ud fra middelværdivektoren og covariansmatricen til $t = 0$ kan angives, anvendes ofte $\boldsymbol{m}_0 = \boldsymbol{0}$ og $\boldsymbol{C}_0 = \kappa \boldsymbol{I}$, hvor κ er en stor konstant, f.eks. $\kappa = 10^6$ [Dethlefsen, 2001, s. 12].

Størrelsen $\lambda_t = \boldsymbol{F}_t^T \boldsymbol{\theta}_t$ vil blive betegnet **signalet**, da det er dette man ville observere, hvis der ikke var nogen observationsstøj ν_t.

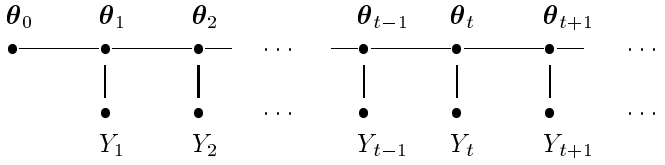
Strukturen af afhængighed i de præsenterede dynamiske modeller, kan beskrives mere overskueligt ved en korrelationsgraf. En sådanne graf er givet ud fra West and Harrison [1997, s. 98]. Grafen opbygges således, at et punkt angiver en stokastisk variabel, og en linje angiver afhængighed mellem to stokastiske variabler. Hvis der betinges med en stokastisk variabel, brydes afhængigheden i dette punkt i grafen. Herved angives betinget uafhængighed.
Lad f.eks. X_1, X_2, X_3 og X_4 være stokastiske variabler, og lad korrelationsgrafen for dem givet ved figur 2.1.



Figur 2.1: Korrelationsgraf for X_1, X_2, X_3 og X_4.

Tolkningen af grafen er, at X_1, X_3 og X_4 er betinget uafhængige, givet X_2.

Ud fra det ovenstående princip kan afhængigheden i den DLM karakteriseres ved korrelationsgrafen i figur 2.2.



Figur 2.2: Korrelationsgraf for den DLM.

Ud af denne ses, at den latente proces opfører sig som en første ordens Markovkæde, og at Y_t kun afhænger af de andre elementer gennem $\boldsymbol{\theta}_t$.

2.2 Kalmanfilteret

Dette afsnit omhandler estimering af den latente proces, samt prædiktion af observationsprocessen. Metoderne der anvendes er først introduceret i bla. Thiele [1880] (se endvidere Lauritzen [1981]), men metoderne blev først anvendt generelt efter, at være blevet introduceret af Kalman i Kalman [1960] og Kalman [1963]. Metoderne er, som følger.

Kalmanfiltrering : Dette er, at estimere tilstandsvektoren $\boldsymbol{\theta}_t$ til tiden t.
Kalmanprædiktion: Dette er, at estimere den latente proces $\{\boldsymbol{\theta}_t\}$, og observationsprocessen $\{Y_t\}$, senere end til tiden t.
Kalmanudglatning: Dette er, at estimere den latente proces $\{\boldsymbol{\theta}_t\}$ op til tiden t.

I de metoder som præsenteres i dette kapitel, antages at $\boldsymbol{F}_t, \boldsymbol{G}_t, V_t, \boldsymbol{W}_t, \boldsymbol{C}_0$ og $\boldsymbol{m}_0$ i den DLM, er kendte. I mange tilfælde kan de dog afhænge af ukendte hyperparametre $\boldsymbol{\psi}$, som skal estimeres. Metoder til estimation af sådanne ukendte hyperparametre vil blive gennemgået i kapitel 5.

2.2.1 Kalmanfilteret

Kalmanfilteret giver den betingede fordeling af tilstandsvektoren $\boldsymbol{\theta}_t$, givet alt relevant information D_t til tiden t, i den DLM. Filteret er formuleret som en metode til at opdatere fordelingerne i systemligningen og observationsligningen, hver gang der foretages en ny observation.

Sætning 2.3 (Kalmanfilteret)
Opdateringen i den den DLM, er givet ved følgende punkter.

***(a)* A posteriori til tid $t-1$:**

$$(\boldsymbol{\theta}_{t-1}|D_{t-1}) \sim \mathcal{N}_n(\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}),$$

hvor $\boldsymbol{m}_{t-1}$ og $\boldsymbol{C}_{t-1}$ er kendte funktioner af D_{t-1}.

***(b)* A priori til tid t:**

$$(\boldsymbol{\theta}_t|D_{t-1}) \sim \mathcal{N}_n(\underbrace{\boldsymbol{G}_t\boldsymbol{m}_{t-1}}_{\boldsymbol{a}_t}, \underbrace{\boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^{\mathrm{T}} + \boldsymbol{W}_t}_{\boldsymbol{R}_t}).$$

***(c)* 1-trinsprædiktion:**

$$(Y_t|D_{t-1}) \sim \mathcal{N}_m(\underbrace{\boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{a}_t}_{f_t}, \underbrace{\boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{R}_t\boldsymbol{F}_t + V_t}_{Q_t}).$$

***(d)* A posteriori til tid t:**

$$(\boldsymbol{\theta}_t|D_t) \sim \mathcal{N}_n(\underbrace{\boldsymbol{a}_t + \overbrace{\boldsymbol{R}_t\boldsymbol{F}_tQ_t^{-1}}^{\boldsymbol{A}_t}\overbrace{(y_t - f_t)}^{e_t}}_{\boldsymbol{m}_t}, \underbrace{\boldsymbol{R}_t - \boldsymbol{A}_tQ_t\boldsymbol{A}_t^{\mathrm{T}}}_{\boldsymbol{C}_t}).$$

Δ

[West and Harrison, 1997, s. 103-105].

Hvis observationerne foretages løbende, kan Kalmanprædiktion anvendes til, at estimere værdien af tilstandsvektoren, og observationerne ud i fremtiden. Hvis man til tiden t gerne vil estimere værdierne til tiden $t+k$, kan disse, jvf. West and Harrison [1997, s. 106-107], findes ved **Kalmanprædiktion**. Her er ligningerne for prædiktion af observationen og tilstandsvektoren til tiden $t+k$ givet rekursivt ved

(a) Tilstandsvektorfordelingen:

$$(\boldsymbol{\theta}_{t+k}|D_t) \sim \mathcal{N}_n\Big(\underbrace{\boldsymbol{G}_{t+k}\boldsymbol{a}_t(k-1)}_{\boldsymbol{a}_t(k)}, \underbrace{\boldsymbol{G}_{t+k}\boldsymbol{R}_t(k-1)\boldsymbol{G}_{t+k}^{\mathrm{T}} + \boldsymbol{W}_{t+k}}_{\boldsymbol{R}_t(k)}\Big),$$

med startværdier $\boldsymbol{a}_t(0) = \boldsymbol{m}_t$ og $\boldsymbol{R}_t(0) = \boldsymbol{C}_t$.

(b) Prædiktionsfordelingen:

$$(Y_{t+k}|D_t) \sim \mathcal{N}\Big(\underbrace{\boldsymbol{F}_{t+k}^{\mathrm{T}}\boldsymbol{a}_t(k)}_{f_t(k)}, \underbrace{\boldsymbol{F}_{t+k}^{\mathrm{T}}\boldsymbol{R}_t(k)\boldsymbol{F}_{t+k} + V_{t+k}}_{Q_t(k)}\Big).$$

2.2.2 Kalmanudglatning

I Kalmanfilteret estimeres den latente proces til tiden t, ud fra de informationer man har til denne tid. Dette estimat anvender således de observationer der foreligger op til og med tiden t. Efter at have observeret en serie af data $y_1, \ldots, y_r$, kan der opnås et bedre estimat for den latente proces til en given tid $t < r$ ved, at basere estimatet på alle observationerne for $t = 1, \ldots, r$. Dette kan, som nævnt, gøres ved, at anvende Kalmanfiltrering rekursivt fremad i tiden, og derefter anvende Kalmanudglatning rekursivt bagud i tiden.

Ideen i Kalmanudglatteren er, at korrigere Kalmanfilterets forudsigelser af hvorledes systemet burde opføre sig, med hvad der så er observeret. Kalmanudglatteren er præsenteret i West and Harrison [1997, s. 113-115], men i

dette speciale er dog anvendt notationen fra Kalmanfilteret. En tilsvarende notation er anvendt i Larsson [2001] og Dethlefsen [2001].

Sætning 2.4 (Kalmanudglatning)
Lad $\boldsymbol{B}_t = \boldsymbol{C}_t \boldsymbol{G}_{t+1}^{\mathrm{T}} \boldsymbol{R}_{t+1}^{-1}$, $r, t \in \mathbb{Z}_+$ *og* $t < r$ *i den DLM. Den udglattede fordeling af tilstandsvektoren til tiden* t *givet informationen til tiden* r*, er da givet rekursivt ved*

$$(\boldsymbol{\theta}_t | D_r) \sim \mathcal{N}_n\Big(\underbrace{\boldsymbol{m}_t + \boldsymbol{B}_t \left(\widetilde{\boldsymbol{m}}_{t+1} - \boldsymbol{a}_{t+1}\right)}_{\widetilde{\boldsymbol{m}}_t}, \underbrace{\boldsymbol{C}_t + \boldsymbol{B}_t \left(\widetilde{\boldsymbol{C}}_{t+1} - \boldsymbol{R}_{t+1}\right) \boldsymbol{B}_t{}^{\mathrm{T}}}_{\widetilde{\boldsymbol{C}}_t} \Big),$$

med startværdierne $\widetilde{\boldsymbol{m}}_r = \boldsymbol{m}_r$ *og* $\widetilde{\boldsymbol{C}}_r = \boldsymbol{C}_r$.

Δ

[West and Harrison, 1997, s. 114-115].

3 Generaliserede lineære modeller

I dette kapitel gives en introduktion til den generaliserede lineære model (GLM) indført af Nelder and Wedderburn [1972]. En mere dybdegående beskrivelse af den GLM er bla. givet i Dobson [1990] og McCullagh and Nelder [1989].

3.1 Den GLM

Den GLM kan betragtes som en udvidelse af den lineære model med Gaussisk fordeling. Denne model betegnes i dette speciale den lineære model (LM).
Udvidelsen af den LM til den GLM består i, at de uafhængige observationer ikke nødvendigvis er normalfordelte, men opfylder følgende model.

Definition 3.1 **(Den GLM)**
Lad $\boldsymbol{Y} = [Y_1, \ldots, Y_r]^{\mathrm{T}}$ *være en vektor af uafhængige stokastiske variabler, der følger en fordeling med tæthedsfunktion på formen*

$$f\left(y_i|\eta_i, V\right) = exp\left(\frac{y_i\eta_i - b(\eta_i)}{a_i(V)} + c(y_i, V)\right), \tag{3.1}$$

for $i = 1, \ldots, r$. *Der gælder, at* $a_i(\cdot)$, $b(\cdot)$ *og* $c(\cdot, \cdot)$, *er kendte reelle funktioner,* η_i *er en ukendt* **naturlig parameter**, *og* V *er en evt. ukendt* **dispersionsparameter**.

Lad der endvidere eksistere en **lineær prediktor** $\boldsymbol{\lambda}$, *som er en linearkombination af* p *forklarende variabler. Hvis* $\boldsymbol{x}_i$ *er den* p*-dimensionale vektor, som udgør værdierne af de forklarende variabler til den* i*'te observation, er*

$$\lambda_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}, \tag{3.2}$$

hvor $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^{\mathrm{T}}$ *er en ukendt vektor af regressionsparametre.*

Hvis $\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_r^{\mathrm{T}}$ *indsættes som rækker i en* nxp **designmatrix** $\boldsymbol{X}$, *kan* $\boldsymbol{\lambda}$ *udtrykkes ved*

$$\boldsymbol{\lambda} = \boldsymbol{X}\boldsymbol{\beta}.$$

Sammenhængen mellem (3.1), *og den lineære prædiktor i* (3.2), *er givet ved*

$$g(\mu_i) = \lambda_i \tag{3.3}$$

hvor $\mu_i = \mathbb{E}[Y_i]$, *for* $i = 1, \ldots, r$, *og* $g(\cdot)$ *er en kendt monoton og differentiabel* **link funktion**. △

Det skal bemærkes, at funktionen $b(\cdot)$ fra den ovenstående definition kaldes **kumulantfunktionen**, samt at funktionerne $a_i(\cdot)$, $b(\cdot)$ og $c(\cdot, \cdot)$ ikke vælges arbitrært, men skal opfylde kravene i det nedenstående.

NOTE: En tæthedsfunktionen givet på formen i (3.1), betegnes en tæthedsfunktion på eksponentiel form, i dette speciale. △

Ud af den ovenstående definition kan visse egenskaber for den GLM udledes. For at gøre dette indføres først følgende definition.

Definition 3.2 **(Kumulantfrembringende funktion)**
Den kumulantfrembringende funktion $K(t)$ *af en stokastisk variabel* Y, *er defineret ved*

$$K(t) = \ln\left(M(t)\right),$$

hvor

$$M(t) = \mathbb{E}\left[\exp\left(tY\right)\right]$$

er den **momentfrembringende funktion**. △

Ved anvendelse af at (3.1) er en tæthed opnås, at

$$\mathrm{K}_i(t) = a_i(V)^{-1}[b(ta_i(V) + \eta_i) - b(\eta_i)].$$

Ud fra generel teori for kumulantfrembringende funktioner, kan middelværdien findes ved den første afledede af $\mathrm{K}(t)$ evalueret i nul, og variansen kan findes ved den anden afledede af $\mathrm{K}(t)$ evalueret i nul. Dvs.

$$\mu_i = \mathrm{K}_i^{'}(0) = b'(\eta_i) \tag{3.4}$$

og

$$\mathbb{V}\mathrm{ar}\left[Y_i\right] = \mathrm{K}_i^{''}(0) = a_i(V)b''(\eta_i). \tag{3.5}$$

Fra det ovenstående kaldes $b^{''}(\eta_i)$ **variansfunktionen**, og betegnes, som en funktion af μ_i ved $\mathcal{V}(\mu_i)$.

3.1.1 Linkfunktioner

Linkfunktionen fra (3.3) kan ikke være en arbitrær funktion, da den skal afbillede værdimængden for μ_i på definitionsmængden for λ_i. Dette gør, at der normalt kun er nogle få linkfunktioner knyttet til en given fordeling.
En speciel linkfunktion, som giver nogle matematisk simplere beregninger, er den kanoniske linkfunktion, der er defineret, som følger.

Definition 3.3 **(Kanonisk linkfunktion)**
Hvis linkfunktionen er givet ved definition 3.1, er den kanoniske linkfunktion den link der opfylder, at $\eta_i = \lambda_i$, for alle i. △

Af (3.4) ses, at den kanoniske linkfunktion er givet ved $\lambda_i = g(\mu_i) = {b'}^{-1}(\mu_i)$.

3.1.2 Opsummering

Karakteristika for nogle ofte anvendte GLM'er er opsummeret i tabel 3.1.2.

3.2 Inferens for den GLM

Inferens for den GLM omfatter estimation af regressionsparametrene $\beta_1, \ldots, \beta_p$ og dispersionsparameteren V, hypotesetest, samt modelkontrol. Dette afsnit indeholder blot en kort gennemgang af metoderne. En mere detaljeret gennemgang findes bla. i McCullagh and Nelder [1989, afs. 2.5] og Dobson [1990, kap. 4-5].

	Gaussisk	Poisson	Binomial	Gamma
Notation	$\mathcal{N}(\mu, \sigma^2)$	$\text{Poi}(\mu)$	$\text{Bi}(r,p)/r$	$\text{Gamma}(\alpha, \alpha/\mu)$
Definitionsmængden for y	$(-\infty, \infty)$	$0(1)\infty$	$\frac{0(1)r}{r}$	$(0, \infty)$
Dispersionsparameter: V	σ^2	1	$1/r$	α^{-1}
Kumulantfunktion: $b(\eta)$	$\eta^2/2$	$\exp(\eta)$	$\ln(1+\exp(\eta))$	$-ln(-\eta)$
$c(y,V)$	$-\frac{1}{2}\left(\frac{y^2}{V} + \ln(2\pi V)\right)$	$-ln(y!)$	$ln(\binom{r}{ry})$	$\alpha\ln(\alpha y) - \ln(y) - ln(\Gamma(\alpha))$
$\mu(\eta) = \mathbb{E}[Y\|\eta]\ (= b'(\eta))$	η	$\exp(\eta)$	$\frac{\exp(\eta)}{(1+\exp(\eta))}$	$-\frac{1}{\eta}$
Kanonisk linkfunktion	identitet	$\ln(\cdot)$	$\text{logit}(\cdot)$	reciprok
Variansfunktion: $\mathcal{V}(\mu)$	1	μ	$\mu(1-\mu)$	μ^2

Tabel 3.1: Karakteristika for nogle univariate fordelinger fra den GLM, opsummeret ud fra McCullagh and Nelder [1989, Tabel 2.1].

3.2.1 Estimation

De følgende estimater bygger på log-likelihoodfunktionen (se appendiks B for en kort introduktion)

$$l\left(\boldsymbol{\beta}|\boldsymbol{y}\right) = \sum_{i=1}^{r} \ln\left(f\left(y_i|\eta_i, V\right)\right),$$

og findes ved hjælp af **Newton-Raphson metoden** eller **Fisher's scorings metode**.

Til at indføre disse metoder anvendes følgende definition.

Definition 3.4 **(Scorefunktionen)**
Lad $\boldsymbol{Y}$ være en stokastisk vektor, som afhænger af parametervektoren $\boldsymbol{\beta}$, så log-likelihoodfunktionen for $\boldsymbol{Y}$ er givet ved $l\left(\boldsymbol{\beta}|\boldsymbol{y}\right)$. Scorefunktionen $\boldsymbol{S}\left(\boldsymbol{\beta}\right)$ er da defineret ved

$$\boldsymbol{S}\left(\boldsymbol{\beta}\right) = \frac{\partial l\left(\boldsymbol{\beta}|\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}}.$$

Δ

MLE $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ findes ved at løse **scoreligningen** $\mathbf{S}\left(\boldsymbol{\beta}\right) = \mathbf{0}$. Ved at se på den j'te komponent $\mathrm{S}_j\left(\boldsymbol{\beta}\right)$, og ved at observere at (3.5) og (3.2) giver, at $\partial\mu_i/\partial\eta_i = \mathcal{V}(\mu_i)$ og $\partial\lambda_i/\partial\beta_j = x_{ij}$ opnås, at

$$\begin{aligned}
\mathrm{S}_j\left(\boldsymbol{\beta}\right) &= \frac{\partial l\left(\boldsymbol{\beta}|\boldsymbol{y}\right)}{\partial \beta_j} \\
&= \sum_{i=1}^{r} \frac{\partial l\left(\boldsymbol{\beta}|y_i\right)}{\partial \beta_j} \\
&= \sum_{i=1}^{r} \frac{\partial l\left(\boldsymbol{\beta}|y_i\right)}{\partial \eta_i} \frac{\partial \eta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial \beta_j} \\
&= \sum_{i=1}^{r} \frac{y_i - \mu_i}{a_i(V)} \frac{1}{\mathcal{V}(\mu_i)} \frac{1}{g'(\mu_i)} x_{ij} \\
&= \sum_{i=1}^{r} x_{ij} \{\mathbb{V}\mathrm{ar}\left[Y_i\right] g'(\mu_i)\}^{-1} (y_i - \mu_i),
\end{aligned}$$

for $j = 1, \ldots, p$. For at løse scoreligningen er det normalt nødvendigt, at anvende iterative metoder. Dvs. Newton-Raphson metoden eller Fisher's scorings metode.

Opdatering i Newton-Rapson metoden

Lad $\hat{\boldsymbol{\beta}}^{(m)}$ være den m'te approksimation af $\hat{\boldsymbol{\beta}}$ i Newton-Raphson metoden. Ved en første ordens Taylorudvikling af $\mathbf{S}\left(\hat{\boldsymbol{\beta}}\right)$ omkring $\hat{\boldsymbol{\beta}}^{(m)}$ opnås, at

$$\mathbf{S}\left(\hat{\boldsymbol{\beta}}\right) \doteq \mathbf{S}\left(\hat{\boldsymbol{\beta}}^{(m)}\right) + \frac{\partial \mathbf{S}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)}{\partial \hat{\boldsymbol{\beta}}}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(m)}\right).$$

Løsningen $\hat{\boldsymbol{\beta}}$ opnås for $\mathbf{S}\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{0}$, hvilket medfører, at

$$-\mathbf{S}\left(\hat{\boldsymbol{\beta}}^{(m)}\right) \doteq \boldsymbol{H}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(m)}\right),$$

hvor den stokastiske matrix

$$\boldsymbol{H}\left(\hat{\boldsymbol{\beta}}^{(m)}\right) = \frac{\partial \mathbf{S}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)}{\partial \hat{\boldsymbol{\beta}}}$$

kaldes Hessian matricen.

Under antagelse af at Hessian matricen er regulær, opnås det følgende rekursive trin i algoritmen ved

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} - \boldsymbol{H}^{-1}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)\mathbf{S}\left(\hat{\boldsymbol{\beta}}^{(m)}\right).$$

Denne algoritme gentages, ind til afstanden mellem $\hat{\boldsymbol{\beta}}^{(m)}$ og $\hat{\boldsymbol{\beta}}^{(m+1)}$ er tilstrækkelig lille.

NOTE: Algoritmen er ikke nødvendigvis konvergent. △

Fisher's scorings metode

Ud fra Newton-Raphson metoden fremkommer Fisher's scorings metode direkte ved, at erstatte Hessian matricen med dens middelværdi. I litteraturen anvendes ofte

$$\boldsymbol{\mathcal{I}}\left(\hat{\boldsymbol{\beta}}^{(m)}\right) = -\mathbb{E}\left[\boldsymbol{H}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)\right], \tag{3.7}$$

som kaldes **informations matricen**. Iterationerne er her givet ved

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \boldsymbol{\mathcal{I}}^{-1}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)\mathbf{S}\left(\hat{\boldsymbol{\beta}}^{(m)}\right), \tag{3.8}$$

som ofte er lettere at håndtere beregningsmæssigt end Newton-Raphson metoden.

NOTE: Hvis $g(\cdot)$ er den kanoniske link bliver Hessian matricen lig med den negative informationsmatrix, således at Newton-Raphson metoden og Fisher's scorings metode giver den samme algoritme. Δ

NOTE: Det skal bemærkes, at hvis $a_i(\cdot)$ er en arbitrær funktion, og V er ukendt, kan algoritmen fejle. Hvis $a_i(V)$ derimod er på formen

$$a_i(V) = \frac{V}{m_i},$$

hvor m_i er kendte vægte for $i = 1, \ldots, r$, går V ud i (3.8), hvorved $\hat{\boldsymbol{\beta}}$ kan estimeres uden at kende V. Δ

NOTE: Som startværdi til algoritmen kan $\hat{\boldsymbol{\mu}}_0 = \boldsymbol{y}$ normalt anvendes, men det er nødvendigt at kontrollere, at der ikke herved fremkommer udefinerede elementer, som for eksempel $\ln(0)$. Δ

Estimation af V

For at estimere V indføres først definitionen af deviansen.

Definition 3.5 **(Deviansen)**
Lad $\boldsymbol{Y}$ være en stokastisk vektor, som er antaget at tilhøre den GLM med $a_i(V) = V/m_i$, hvor m_i er kendt for $i = 1, \ldots, r$. Lad endvidere $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu}$. Deviansen $D(\boldsymbol{y}|\boldsymbol{\mu})$ for givne værdier af $\boldsymbol{\mu}$ og $\boldsymbol{y}$ er da defineret ved

$$D(\boldsymbol{y}|\boldsymbol{\mu}) = 2V\Big(l(\boldsymbol{y}|\boldsymbol{y}) - l(\boldsymbol{\mu}|\boldsymbol{y})\Big).$$

Δ

Iagttag at for en normalfordeling med identitetslinkfunktion, og p parametre, er deviansen $\mathrm{D}(\boldsymbol{y}|\boldsymbol{\mu})$ lig residualkvadratsummen, og dispersionsparameteren V er lig variansen, hvorved $\mathrm{D}(\boldsymbol{y}|\boldsymbol{\mu})/V \sim \chi^2(r-p)$. Ud fra dette opnås, under generelle regularitetsbetingelser beskrevet i Venables and Ripley [1997, s. 127], at

$$\hat{V} = \frac{\mathrm{D}(\boldsymbol{y}|\boldsymbol{\mu})}{r-p}$$

approksimativt er et centralt estimat for V i den GLM.

3.2.2 Hypotesetest

Hypotesetest behandles ud fra to metoder. Den første er baseret på deviansen, hvorfor den ikke kan anvendes i de tilfælde hvor deviansen ikke kendes. Af denne grund indføres endvidere en metode baseret på Wald's statistik, som kan anvendes i andre tilfælde.

Hypotesetest baseret på deviansen

Antag at M_0 er en basal GLM, som er fundet acceptabel til at beskrive data. For at reducere antallet af regressionsparametre i denne model, kan der anvendes test baseret på deviansen.

Lad M_0 være en model med p parametre, og lad M_1 være en undermodel af M_0 med $q < p$ parametre. Lad endvidere deviansen til M_i have notationen D_i.

Den reduktion der sker i deviansen, i kraft af at M_0 anvendes i stedet for M_1, kan under svage regularitetsbetingelser approksimeres godt ved

$$(\mathrm{D}_1 - \mathrm{D}_0) \dot{\sim} V\chi^2(p-q) \tag{3.9}$$

[Lindsey, 1997, s. 212-214]. Se endvidere Fahrmeir and Tutz [1994, s. 48] og McCullagh and Nelder [1989, s. 119] for yderligere diskussioner.

Ud fra (3.9), opnås herved, at en hypotese
H_1: M_1 er korrekt,
kan testes under hypotesen
H_0: M_0 er korrekt.

Hvis dispersionsparameteren V er kendt, kan der på grundlag af (3.9) laves en χ^2-test af H_1 under H_0.
Hvis derimod V er ukendt, og der er r observationer, kan der på grundlag af (3.9), opnås F-statistikken

$$\frac{\dfrac{\mathrm{D}_1 - \mathrm{D}_0}{p-q}}{\dfrac{\mathrm{D}_0}{r-p}} \dot{\sim} \mathrm{F}(p-q, r-p)$$

til at teste H_1 under H_0.

NOTE: De ovenstående statistikker kan endvidere anvendes til, at lave konfidensintervaller for parameterestimater. △

Hypotesetest baseret på Wald's statistik

En anden anvendt metode til hypotesetest er Wald's test. Denne test kan anvendes til at teste lineære hypoteser. Dvs. hypoteser af formen

$$\boldsymbol{B}\boldsymbol{\beta} = \mathbf{0},$$

hvor $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^{\mathrm{T}}$ er parametervektoren, og $\boldsymbol{B}$ er en konstant $k \times p$ matrix, af rang k, som angiver hypotesen. Hvis hypotesen f.eks. er, at $\beta_1 = 0$ er $\boldsymbol{B} = [1, 0, \ldots, 0]$.

Lad estimatet for parametervektoren være givet ved

$$\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}_p\Big(\boldsymbol{\beta}, \boldsymbol{J}(\hat{\boldsymbol{\beta}})\Big),$$

hvor $\boldsymbol{J}(\hat{\boldsymbol{\beta}})$ er regulær.
Hvis $\boldsymbol{Y} \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$, hvor $\boldsymbol{\Sigma}$ er regulær, gælder jvf. Seber [1984, s. 16], at

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} \sim \chi^2(k),$$

hvorved Wald's statistik kan opstilles ved

$$\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\left(\boldsymbol{B}\boldsymbol{J}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{B}^{\mathrm{T}}\right)^{-1}\boldsymbol{B}\hat{\boldsymbol{\beta}} \dot{\sim} \chi^2(k).$$

NOTE: Hvis $\hat{\boldsymbol{\beta}}$ er opnået ved maksimum likelihoodestimation, er den approksimative normalfordeling, af $\hat{\boldsymbol{\beta}}$, givet ved (B.1). △

Ud fra Wald's statistik kan der herved laves χ^2-test af lineære hypoteser ud fra samme princip, som hypotesetest baseret på deviansen.

En dybere indføring i Wald's test er bla. givet i Fahrmeir and Tutz [2001]. Her diskuteres endvidere approksimationer af fordelingen af $\hat{\boldsymbol{\beta}}$, under andre estimationsmetoder end maksimum likelihood.

3.2.3 Modelkontrol

For at lave modelkontrol for den GLM, kan der anvendes residualplots, ligesom ved den LM. Nogle af de residualer der ofte anvendes, er Pearsonresidualerne ($r^{(p)}$), og deviansresidualerne ($r^{(d)}$), som er defineret nedenfor.

Definition 3.6 **(Pearsonresidualet)**
Pearsonresidualet er givet ved

$$r_i^{(p)} = \frac{y_i - \mu_i}{\sqrt{\mathcal{V}(\mu_i)}},$$

hvor y_i er en observation, μ_i er middelværdien, og $\mathcal{V}(\cdot)$ er variansfunktionen, for $i = 1, \ldots, r$. △

***Definition 3.7* (Deviansresidualet)**
Lad der være r observationer, og lad d_i være givet således, at $D(\boldsymbol{y}|\boldsymbol{\mu}) = \sum_{i=1}^{r} d_i$. Deviansresidualet er da givet ved

$$r_i^{(d)} = sign(y_i - \mu_i)\sqrt{d_i},$$

hvor $sign(\cdot)$ angiver fortegnet af det givne udtryk, for $i = 1, \ldots, r$. △

Andre former for modelkontrol består i, at anvende metoder fra analyse af tidsrækker til, at analysere residualerne for f.eks. serielle korrelationer.

4 Dynamiske generaliserede lineære modeller

I dette kapitel indføres den dynamiske generaliserede lineære model (DGLM). Den DGLM kan dels opfattes som en udvidelse af den DLM, til at omfatte observationer fra en fordeling med tæthedsfunktion på samme form, som i den GLM. Dels kan den opfattes, som en udvidelse af den GLM, til en model med stokastiske parametre givet ud fra en proces, der udvikler sig som en første ordens markovkæde. Dette gør, at den DGLM kan betragtes, som værende en fælles overbygning på de to modeller der er introduceret indtil videre.

Sammen med den DGLM gennemgåes endvidere approksimative metoder til estimation af den latente proces, for den DGLM.

En mere dybdegående beskrivelse af den DGLM er bla. givet i West et al. [1985] og West and Harrison [1997, kap. 14], mens en mere dybdegående beskrivelse af et specialtilfælde af den DGLM, hvor fordelingen i systemligningen er Gaussisk, bla. er givet i Fahrmeir and Tutz [2001]. En mere detaljeret gennemgang af teorien er bla. givet i Larsson [2001]. Her gennem-

gåes teorien nogenlunde korrekt for den samme generelle DGLM som her, men dog i et specialtilfælde med kanonisk link, og dispersionsparameter lig en, for den DGLM, hvor fordelingen i systemligningen er Gaussisk. En mere generel, men mindre detaljeret gennemgang er at finde i Dethlefsen [2001].

4.1 Modelopstilling

For at kende forskel på den GLM og den DGLM, defineres den DGLM som en GLM med stokastiske, dynamiske parametre, men med rækker i designmatricen givet ved $\boldsymbol{F}_t^{\mathrm{T}}$, og dynamiske stokastiske parametre $\boldsymbol{\theta}_t$. Herved bliver notationen for den DGLM også i overensstemmelse med notationen for den DLM.

Den DGLM defineres jvf. gennemgangen i West et al. [1985], som følger.

Definition 4.1 **(Den DGLM)**
Lad $\boldsymbol{F}_t$ være en n-dimensional vektor, $\boldsymbol{G}_t$ være en $n \times n$ matrix, og $\boldsymbol{W}_t$ være en $n \times n$ covariansmatrix for $t = 1, \ldots, r$. Det antages, at $\boldsymbol{F}_t, \boldsymbol{G}_t$ og $\boldsymbol{W}_t$ er kendte eventuelt på nær en vektor $\boldsymbol{\psi}$ af ukendte hyperparametre.

Lad observationerne $\{Y_t\}$ være stokastiske variabler der følger en fordeling med tæthedsfunktion på formen

$$f\left(y_t|\eta_t, V_t\right) = \exp\left(\frac{y_t\eta_t - b(\eta_t)}{a(V_t)} + c(y_t, V_t)\right) \tag{4.1}$$

for $t = 1, \ldots, r$. Der gælder, at $a(\cdot)$, $b(\cdot)$ og $c(\cdot,\cdot)$, er kendte reelle funktioner, η_t er en ukendt **naturlig parameter**, *og V_t er en kendt* **dispersionsparameter**.

Lad der endvidere eksistere et **signal** *λ_t, som er givet ved*

$$\lambda_t = \boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{\theta}_t, \tag{4.2}$$

hvor $\boldsymbol{\theta}_t$ er en n-dimensional stokastisk parametervektor, der udvikler sig som en førsteordens Markovkæde, for $t = 1, \ldots, r$. Herved beskrives den latente proces ved systemligningen

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim [\boldsymbol{0}, \boldsymbol{W}_t], \tag{4.3}$$

hvor $[\boldsymbol{m}_1, \boldsymbol{M}_2]$ angiver, at fordelingen er delvist specificeret ved første moment $\boldsymbol{m}_1$, og andet moment $\boldsymbol{M}_2$.

Begyndelsesbetingelsen for den latente proces, er givet ved

$$(\boldsymbol{\theta}_0|D_0) \sim [\boldsymbol{m}_0, \boldsymbol{C}_0].$$

Sammenhængen mellem (4.1)*, og den lineære prædiktor* (4.2)*, er givet ved*

$$g(\mu_t) = \lambda_t \tag{4.4}$$

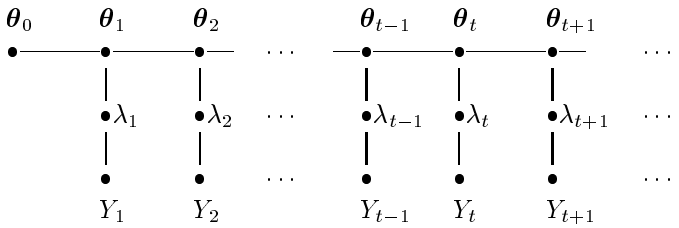
hvor $\mu_t = \mathbb{E}[Y_t]$*, for* $t = 1, \ldots, r$*, og* $g(\cdot)$ *er en kendt monoton og differentiabel* **link funktion**.

Der gælder endvidere, at $(Y_t|\boldsymbol{\theta}_t)$ *er uafhængig af* $Y_1, \ldots, Y_{t-1}, Y_{t+1}, \ldots, Y_r$, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_r$*, at* Y_t *kun afhænger af* $\boldsymbol{\theta}_t$ *gennem* λ_t *samt, at* $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r, (\boldsymbol{\theta}_0|D_0)$ *er uafhængige.* △

NOTE: Det skal bemærkes, at der i den DGLM anvendes $a(V_t)$, i modsæting til den tilsvarende $a_i(V)$ fra den GLM. Denne omskrivning anvendes, da den passer bedre ind i den nye modelopsætning. Den er blot et specialtilfælde af funktionen fra den GLM, da V_t er antaget at være kendt i den DGLM.

For at lette notationen anvendes ofte betegnelsen $\varphi_t = 1/a(V_t)$. △

Ligesom ved den DLM kan afhængighedsstrukturene ved den DGLM karakteriseres ved en korrelationsgraf. Dette er gjort i figur 4.1, hvor af det ses, at den eneste forskel i strukturene for den DLM og den DGLM er, at signalet nu er indskudt i figuren.



Figur 4.1: Korrelationsgraf for den DGLM.

4.2 Filtrering

For at estimere den latente proces indføres ligesom ved den DLM et filter. Dette filter bygger på konjugeret analyse og lineær Bayesiansk estimering. Det er introduceret i West et al. [1985], og kører efter en lidt anden strategi end Kalmanfilteret.
Metoden til at vise filteret for den DGLM bygger basalt på to ting. Den

første er, at anvende konjugeret analyse til, at approksimere fordelingen af den naturlige parameter. Konjugeret analyse går ud på, at finde en passende fordeling, således at opdateringen i filteret kan foretages udelukkende ved, at opdatere parametrene i fordelingen.
Det andet er, at anvende lineær Bayesiansk estimereing til, at approksimere a posteriorifordelingen af $\boldsymbol{\theta}_t$. Dette bygger på antagelsen, at middelværdien af $\boldsymbol{\theta}_t$ er givet ved en lineær transformation og forskydning af λ_t. Estimatet for middelværdien og variansen af $\boldsymbol{\theta}_t$, kan heraf findes ved at minimere middelkvadratfejlen af estimatet.

Under anvendelse af de ovenstående approksimationer, kan filteret for den DGLM gennemgåes på samme måde som Kalmanfilteret. Opdateringen bliver herved givet ud fra de følgende trin.

A posteriori til tiden t – 1

I opdateringen tages udgangspunkt i, at

$$(\boldsymbol{\theta}_{t-1}|D_{t-1}) \sim [\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}],$$

hvor $\boldsymbol{m}_{t-1}$ og $\boldsymbol{C}_{t-1}$ er kendte funktioner af D_{t-1}.

A priorifordelingerne

Lige som i Kalmanfilteret findes a priorifordelingerne direkte ud fra definitionen af den DGLM, ved

$$(\boldsymbol{\theta}_t|D_{t-1}) \sim \Big[\underbrace{\boldsymbol{G}_t\boldsymbol{m}_{t-1}}_{\boldsymbol{a}_t}, \underbrace{\boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^{\mathrm{T}} + \boldsymbol{W}_t}_{\boldsymbol{R}_t}\Big],$$

og

$$(\lambda_t|D_{t-1}) \sim \Big[\underbrace{\boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{a}_t}_{f_t}, \underbrace{\boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{R}_t\boldsymbol{F}_t}_{q_t}\Big].$$

Som nævnt approksimeres fordelingen af η_t med den konjugerede fordeling til (4.1). Ved at gøre dette opnås, at a priorifordelingen af η_t er af samme familie, som a posteriorifordelingen af η_t. Ud fra Bayes sætning D.2 ses, at en metode til, at finde den konjugerede fordeling til observationsfordelingen (4.1) er, at anvende en fordeling med tæthedsfunktion på samme form, som likelihoodfunktionen til den. Det antages således, at η_t har en konjugeret fordeling med en tæthedsfunktion givet ved

$$f(\eta_t|D_{t-1}) = d(r_t, s_t)\exp(r_t\eta_t - s_t b(\eta_t)), \tag{4.5}$$

hvor $d(\cdot,\cdot)$ er en kendt funktion, som giver normeringskonstanten. Størrelserne r_t og s_t vælges, så $\mathbb{E}\left[g(b'(\eta_t))|D_{t-1}\right] \doteq f_t$ og $\mathbb{V}\mathrm{ar}\left[g(b'(\eta_t))|D_{t-1}\right] \doteq q_t$. En dybere indføring i konjugeret analyse er bla. givet i West and Harrison [1997, s. 518-521].

A posteriorifordelingerne til tiden t

Efter observationen af y_t opdateres de ovenstående fordelinger. Den første der skal findes er a posteriorifordelingen af η_t. Denne findes ved, at indsætte (4.1) og (4.5) i Bayes sætning D.2. Herved opnås, at

$$f\left(\eta_t|D_t\right) = d(r_t^*, s_t^*)\exp\left(r_t^*\eta_t - s_t^* b(\eta_t)\right), \tag{4.6}$$

hvor

$$r_t^* = r_t + \varphi_t y_t \quad \text{og} \quad s_t^* = s_t + \varphi_t.$$

Her ud fra kan den approksimative a posteriorifordeling af λ_t findes ved

$$(\lambda_t|D_t) \dot{\sim} \left[\mathrm{f}_t^\star, \mathrm{q}_t^\star\right],$$

hvor

$$\mathrm{f}_t^\star = \mathbb{E}\left[g(b'(\eta_t))|D_t\right] \quad \text{og} \quad \mathrm{q}_t^\star = \mathbb{V}\mathrm{ar}\left[g(b'(\eta_t))|D_t\right].$$

Da $\boldsymbol{\theta}_t$ er betinget uafhængig af Y_t givet λ_t, opnås jvf. sætning D.1, at

$$\mathbb{E}\left[\boldsymbol{\theta}_t|D_t\right] = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\theta}_t|\lambda_t, D_{t-1}\right]|D_t\right], \tag{4.7}$$

og

$$\mathbb{V}\mathrm{ar}\left[\boldsymbol{\theta}_t|D_t\right] = \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left[\boldsymbol{\theta}_t|\lambda_t, D_{t-1}\right]|D_t\right] + \mathbb{V}\mathrm{ar}\left[\mathbb{E}\left[\boldsymbol{\theta}_t|\lambda_t, D_{t-1}\right]|D_t\right]. \tag{4.8}$$

Da covariansen mellem $\boldsymbol{\theta}_t$ og λ_t er givet ved

$$\mathrm{Cov}\left[\boldsymbol{\theta}_t, \lambda_t\right] = \mathrm{Cov}\left[\boldsymbol{\theta}_t, \boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{\theta}_t\right] = \boldsymbol{R}_t\boldsymbol{F}_t,$$

opnås ud fra den lineære Bayesianske estimering i sætning D.3, at

$$\mathbb{E}\left[\boldsymbol{\theta}_t|\lambda_t, D_{t-1}\right] \doteq \boldsymbol{a}_t + \boldsymbol{R}_t\boldsymbol{F}_t\frac{\lambda_t - f_t}{q_t}, \tag{4.9}$$

og

$$\mathbb{V}\mathrm{ar}\left[\boldsymbol{\theta}_t|\lambda_t, D_{t-1}\right] \doteq \boldsymbol{R}_t - \frac{\boldsymbol{R}_t\boldsymbol{F}_t\boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{R}_t}{q_t}. \tag{4.10}$$

Ved at indsætte (4.9) og (4.10) i (4.7) og (4.8) opnås herved, at

$$(\boldsymbol{\theta}_t|D_t) \dot{\sim} \left[\boldsymbol{a}_t + \boldsymbol{R}_t\boldsymbol{F}_t(\mathrm{f}_t^\star - f_t)/q_t, \boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{F}_t\boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{R}_t(\mathrm{q}_t^\star - q_t)/q_t^2\right]. \tag{4.11}$$

4.3 Den DGLM med Gaussisk systemligning

Et meget anvendeligt specialtilfælde af den DGLM, er det tilfælde, hvor der i stedet for delvise specificerede fordelinger i definitionen af den DGLM, anvendes Gaussiske fordelinger. Dette specialtilfælde er bla. behandlet i Fahrmeir and Tutz [2001]. Der tages da også udgangspunkt i modellen her fra, men samtidig anvendes nogle af ideerne fra Durbin and Koopman [2001]. Herved kan metoderne vises for den type model, som anvendes her, men noget simplere end det kan gøres direkte ud fra Fahrmeir and Tutz [2001], og med en lidt anden approksimationsprocedure. Til sammenligning med metoderne fra Durbin and Koopman [2001], har metoderne her den fordel, at de også gælder i de tilfælde, hvor der ikke anvendes kanonisk link, til gengæld omfatter modellen fra Durbin and Koopman [2001] flere forskellige typer af fordelinger.

I dette speciale betegnes denne model den DGLM med Gaussisk systemligning, og har følgende definition.

***Definition 4.2* (Den DGLM med Gaussisk systemligning)**
Lad $\boldsymbol{F}_t$ være en n-dimensional vektor, $\boldsymbol{G}_t$ være en $n \times n$ matrix, og $\boldsymbol{W}_t$ være en $n \times n$ covariansmatrix for $t = 1, \ldots, r$. Det antages, at $\boldsymbol{F}_t, \boldsymbol{G}_t$ og $\boldsymbol{W}_t$ er kendte eventuelt på nær en vektor $\boldsymbol{\psi}$ af ukendte hyperparametre.

Lad observationerne $\{Y_t\}$ være stokastiske variabler der følger en fordeling med tæthedsfunktion på formen

$$f\left(y_t|\eta_t, V_t\right) = \exp\left(\frac{y_t\eta_t - b(\eta_t)}{a(V_t)} + c(y_t, V_t)\right) \tag{4.12}$$

for $t = 1, \ldots, r$. Der gælder, at $a(\cdot)$, $b(\cdot)$ og $c(\cdot,\cdot)$, er kendte reelle funktioner, η_t er en ukendt **naturlig parameter**, *og V_t er en kendt* **dispersionsparameter**.

Lad der endvidere eksistere et **signal** *λ_t, som er givet ved*

$$\lambda_t = \boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{\theta}_t, \tag{4.13}$$

hvor $\boldsymbol{\theta}_t$ er en n-dimensional stokastisk parametervektor, der udvikler sig som en førsteordens Markovkæde, for $t = 1, \ldots, r$. Herved beskrives den latente proces ved systemligningen

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{W}_t). \tag{4.14}$$

Begyndelsesbetingelsen for den latente proces, er givet ved

$$(\boldsymbol{\theta}_0|D_0) \sim \mathcal{N}_n(\boldsymbol{m}_0, \boldsymbol{C}_0).$$

Sammenhængen mellem (4.12)*, og den lineære prædiktor* (4.13)*, er givet ved*

$$g(\mu_t) = \lambda_t \tag{4.15}$$

hvor $\mu_t = \mathbb{E}[Y_t]$*, for* $t = 1, \ldots, r$*, og* $g(\cdot)$ *er en kendt monoton og differentiabel* **link funktion**.

Der gælder endvidere, at $(Y_t|\boldsymbol{\theta}_t)$ *er uafhængig af* $Y_1, \ldots, Y_{t-1}, Y_{t+1}, \ldots, Y_r$, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_r$*, at* Y_t *kun afhænger af* $\boldsymbol{\theta}_t$ *gennem* λ_t*, samt at* $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r, (\boldsymbol{\theta}_0|D_0)$ *er uafhængige.* △

4.3.1 Estimation

Den DGLM med Gaussisk systemligning har den fordel, at den på mange felter ligner den DLM. Forskellene ligger to steder. Det første er, at der i den DGLM kan være en linkfunktion forskellig fra identitetsfunktionen mellem signalet λ, og middelværdien μ. Det andet er, at fordelingen af observationerne er på den eksponentielle form (4.12) i den DGLM, hvorimod den er Gaussisk i den DLM.

For at finde metoder til estimation af den latente proces i den DGLM med Gaussisk systemligning, anvendes ligheden med den DLM. Således indføres der, i modsæting til den generelle DGLM, ikke yderligere antagelser om fordelingen af systemligningen. Istedet tages der udgangspunkt i de to steder, hvor den DGLM med Gaussisk systemligning, og den DLM er forskellige fra hinanden. Her ud fra approksimeres den DGLM med Gaussisk systemligning ved en passende DLM, hvorved metoderne fra kapitel 2 kan anvendes til estimation. Ved denne approksimation anvendes således den samme systemligning, mens det er observationsligningen der approksimeres.

For at finde den approksimerende DLM, lineariseres således, at der approksimativt opnås samme middelværdi og varians for observationerne i den anvendte DGLM med Gaussisk systemligning, og den approksimerende DLM. Da middelværdien i den DLM er lig signalet, findes et estimat for middelværdien, ved en første ordens Taylorudvikling af $g^{-1}(\lambda_t)$ omkring et estimat $\hat{\lambda}_t$ for λ_t. Herved opnås, at

$$\mu_t = g^{-1}(\lambda_t) \doteq h_t + \ddot{h}_t(\lambda_t - \hat{\lambda}_t), \tag{4.16}$$

hvor

$$h_t = g^{-1}(\hat{\lambda}_t) \quad \text{og} \quad \ddot{h}_t = \frac{dg^{-1}(\hat{\lambda}_t)}{d\lambda_t}.$$

For at middelværdien i den approksimerende DLM skal være lig signalet λ_t, transformeres observationerne således, at dette er opfyldt. Ud fra (4.16) ses,

at den transformerede værdi er givet ved

$$\widetilde{Y}_t = \frac{Y_t - h_t}{\ddot{h}_t} + \hat{\lambda}_t.$$

For at finde variansen i den approksimerende DLM anvendes, at der jvf. (3.5) gælder, at $\mathbb{V}\mathrm{ar}\,[Y_t] = a(V_t)b''(\eta_t)$. Variansen af $(\widetilde{Y}_t|\hat{\lambda}_t)$ er dermed givet ved

$$\widetilde{V}_t = \mathbb{V}\mathrm{ar}\left[\widetilde{Y}_t|\hat{\lambda}_t\right] = \frac{\mathbb{V}\mathrm{ar}\left[Y_t|\hat{\lambda}_t\right]}{\ddot{h}_t^2} = \frac{a(V_t)b''(\varrho^{-1}(\hat{\lambda}_t))}{\ddot{h}_t^2},$$

hvor $\varrho(\cdot) = g(b'(\cdot))$.

Bemærk at der i specialtilfældet med kanonisk link specielt gælder, at

$$\widetilde{Y}_t = \frac{Y_t - b'(\hat{\lambda}_t)}{b''(\hat{\lambda}_t)} + \hat{\lambda}_t \quad \text{og} \quad \widetilde{V}_t = \frac{a(V_t)}{b''(\hat{\lambda}_t)}, \tag{4.17}$$

da $\lambda_t = \eta_t$ i dette tilfælde, hvorved $\varrho(\cdot)$ er identitetsfunktionen, $h_t = b'(\hat{\lambda}_t)$, og $\ddot{h}_t = b''(\hat{\lambda}_t)$.

Ud fra det ovenstående approksimeres den DGLM med Gaussisk systemligning til enhver tid t, ved en DLM med observationsligningen

$$\widetilde{Y}_t = \boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{\theta}_t + \nu_t, \quad \nu_t \sim \mathcal{N}(0, \widetilde{V}_t), \tag{4.18}$$

systemligning

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim \mathcal{N}(0, W_t), \tag{4.19}$$

og begyndelsesbetingelse

$$(\boldsymbol{\theta}_t|D_0) \sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{C}_0). \tag{4.20}$$

Ved filtrering foretages den ovenstående approksimation sekventielt gennem den anvendte tidsrække. Størrelsen $\hat{\lambda}_t$ kan herved bestemmes ud fra de filtrerede værdier til tiden $t-1$, ved $\hat{\lambda}_t = \boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{G}_t\boldsymbol{m}_{t-1}$, hvor $\boldsymbol{m}_{t-1}$ er den estimerede middelværdivektor for a posteriorifordelingen af $\boldsymbol{\theta}_{t-1}$. I dette tilfælde betegnes metoden, jvf. Fahrmeir and Tutz [2001, 352-354], det **udvidede Kalmanfilter**.

Ligesom ved den DLM, kan fordelingen af tilstandsvektoren $\boldsymbol{\theta}_t$ estimeres, under anvendelse af al tilgængelig information D_r. Dette gøres ved blot, at anvende Kalmanudglatteren fra sætning 2.4 med $\boldsymbol{C}_t, \boldsymbol{G}_t, \boldsymbol{R}_t, \boldsymbol{m}_t$ og $\boldsymbol{a}_t$ givet ved den approksimerende DLM, og det udvidede Kalmanfilter, for $t = 1, \ldots, r$.

Forbedring ved iteration

Alternativt kan metoden forbedres ved iteration. Dette gøres ved, til et givet trin, at køre Kalmanfilteret og Kalmanudglatteren på den approksimerende DLM, og derefter finde $\hat{\lambda}_t$, til det næste trin, ud fra de udglattede værdier. Denne procedure kan opstilles ved følgende iteration over $i = 0, 1, \ldots,$ for en tidsrække $\{y_t\}$ for $t = 1, \ldots, r$.

For i = 0

Lad $\hat{\lambda}_t^{(0)} = \boldsymbol{F}_t^{\mathrm{T}} \boldsymbol{G}_t \boldsymbol{m}_{t-1}$, hvor $\boldsymbol{m}_{t-1}$ er a posteriorimiddelværdien for $\boldsymbol{\theta}_t$ givet rekursivt ved det udvidede Kalmanfilter, for $t = 1, \ldots, r$.
Ved hjælp af det udvidede Kalmanfilter, og Kalmanudglatteren findes den stakkede vektor

$$\widetilde{\boldsymbol{m}}^{(0)} = \begin{bmatrix} \boldsymbol{m}_0 \\ \widetilde{\boldsymbol{m}}_1^{(0)} \\ \vdots \\ \widetilde{\boldsymbol{m}}_r^{(0)} \end{bmatrix},$$

hvor $\widetilde{\boldsymbol{m}}_1^{(0)}, \ldots, \widetilde{\boldsymbol{m}}_r^{(0)}$ er de udglattede middelværdier fundet ved Kalmanudglatning.

For i > 0

Lad $\hat{\lambda}_t^{(i)} = \boldsymbol{F}_t^{\mathrm{T}} \widetilde{\boldsymbol{m}}_t^{(i-1)}$. Find den approksimerende DLM for $\hat{\lambda}_t^{(i)}$ givet ved (4.18)-(4.20), for $t = 1, \ldots, r$. Anvend Kalmanfilteret og Kalmanudglatteren på den opnåede approksimerende DLM til, at finde den nye stakkede vektor

$$\widetilde{\boldsymbol{m}}^{(i)} = \begin{bmatrix} \boldsymbol{m}_0 \\ \widetilde{\boldsymbol{m}}_1^{(i)} \\ \vdots \\ \widetilde{\boldsymbol{m}}_r^{(i)} \end{bmatrix},$$

af udglattede middelværdivektorer.

Iterationen fortsættes ind til afstanden mellem $\widetilde{\boldsymbol{m}}^{(i)}$ og $\widetilde{\boldsymbol{m}}^{(i-1)}$ er tilstrækkelig lille.

Denne iterative approksimationsprocedure svarer til metoderne, der er gennemgået i Durbin and Koopman [2001, s. 191-198], dog anvendes her en anden model, end den som anvendes i Durbin and Koopman [2001].

5 Videre Inferens

I de metoder der er gennemgået i de forgående kapitler antages, at alle hyperparametre er kendte. Hvis dette ikke er tilfældet, må de estimeres før metoderne kan anvendes.

Der findes en del forskellige metoder til estimering af såvel ukendte hyperparametre, som de stokastiske parametre i dynamiske modeller. I dette kapitel gennemgås en af de nyere metoder der er baseret på importance sampling, og indført i Durbin and Koopman [2000]. Metoderne er endvidere gennemgået i Durbin and Koopman [2001], hvor der også gennemgås visse dynamiske modeller, samt nogle estimationsmetoder for disse.

Metoderne gennemgås i dette kapitel ud fra den samme terminologi og notation, som i Durbin and Koopman [2001, kap. 11]. Således indføres først basisideen i importancesampling, hvorefter denne ide anvendes til, at estimere ukendte hyperparametre.

5.1 Importance sampling

Lad i det følgende $\boldsymbol{y} = [y_1, \ldots, y_r]^{\mathrm{T}}$ være den stakkede vektor af observationer i en dynamisk model, og lad $\boldsymbol{\theta} = [\boldsymbol{\theta}_0^{\mathrm{T}}, \boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_r^{\mathrm{T}}]^{\mathrm{T}}$ være den stakkede vektor af stokastiske parametre i modellen. I den klassiske udgave handler importance sampling basalt om, at estimere den betingede middelværdi

$$\mathbb{E}\left[\boldsymbol{x}(\boldsymbol{\theta})|\boldsymbol{y}\right] = \int \boldsymbol{x}(\boldsymbol{\theta}) f\left(\boldsymbol{\theta}|\boldsymbol{y}\right) d\,\boldsymbol{\theta}, \tag{5.1}$$

hvor $\boldsymbol{x}(\cdot)$ er en arbitrær funktion af $\boldsymbol{\theta}$, og $f\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$ er tæthedsfunktionen for den betingede fordeling af $\boldsymbol{\theta}$ givet $\boldsymbol{y}$. Et estimat for den betingede middel-

værdi i (5.1) giver mulighed for, at estimere størrelser, som middelværdien og variansen af den latente proces, samt likelihoodfunktionen af ukendte hyperparametre. Dvs. at alle de estimater der skal findes i de behandlede dynamiske modeller, kan opnås ud fra (5.1). For at notetionen skal blive overskuelig, og i overensstemmelse med Durbin and Koopman [2001, kap. 11], er afhængigheden af eventuelle ukendte hyperparametre implicit i notationen.

En metode til, at opnå et estimat for (5.1) er, at simulere en mængde værdier $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$ ud fra en fordeling med tæthedsfunktionen $f(\boldsymbol{\theta}|\boldsymbol{y})$, og derefter approksimere middelværdien ved gennemsnittet af $\boldsymbol{x}(\boldsymbol{\theta}^{(1)}), \ldots, \boldsymbol{x}(\boldsymbol{\theta}^{(N)})$. I mange tilfælde kan $f(\boldsymbol{\theta}|\boldsymbol{y})$ dog være en funktion hvor fra det er besværligt at simulere. Ideen i importance sampling er her, at approksimere $f(\boldsymbol{\theta}|\boldsymbol{y})$ ved en tæthedsfunktion, som det er nemt at simulere fra, og derefter approksimere (5.1) ud fra simulerede værdier fra denne tæthedsfunktion. En sådanne tæthedsfunktion betegnes en **importance tæthedsfunktion**. Ofte anvendes en Gaussisk importance tæthedsfunktion, da det er nemt, at simulere fra en sådanne. Gennemgangen i Durbin and Koopman [2001, kap. 11] bygger da også på **Gaussisk importance sampling**, hvilket vil sige, at der anvendes en Gaussisk importance tæthedsfunktion. Gaussisk importance sampling har endvidere den fordel, at nogle af approksimationsmetoderne fra kapitel 4 kan anvendes til, at finde en passende Gaussisk importance tæthedsfunktion.

Lad i det følgende $\mathcal{G}(\cdot)$ være en Gaussisk importance tæthedsfunktion. Middelværdien i (5.1) kan da omskrives til

$$\mathbb{E}\left[\boldsymbol{x}(\boldsymbol{\theta})|\boldsymbol{y}\right] = \int \boldsymbol{x}(\boldsymbol{\theta}) \frac{f(\boldsymbol{\theta}|\boldsymbol{y})}{\mathcal{G}(\boldsymbol{\theta}|\boldsymbol{y})} \mathcal{G}(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} = \mathbb{E}_{\mathcal{G}}\left[\boldsymbol{x}(\boldsymbol{\theta}) \frac{f(\boldsymbol{\theta}|\boldsymbol{y})}{\mathcal{G}(\boldsymbol{\theta}|\boldsymbol{y})}\right], \tag{5.2}$$

hvor $\mathbb{E}_{\mathcal{G}}[\cdot]$ er middelværdien med hensyn til den Gaussiske importance tæthedsfunktion $\mathcal{G}(\cdot)$. Brøken $f(\boldsymbol{\theta}|\boldsymbol{y})/\mathcal{G}(\boldsymbol{\theta}|\boldsymbol{y})$ i (5.2) kan ud fra Bayes sætning D.2 omskrives til

$$\frac{f(\boldsymbol{\theta}|\boldsymbol{y})}{\mathcal{G}(\boldsymbol{\theta}|\boldsymbol{y})} = \frac{\mathcal{G}(\boldsymbol{y})\, f(\boldsymbol{\theta}, \boldsymbol{y})}{f(\boldsymbol{y})\, \mathcal{G}(\boldsymbol{\theta}, \boldsymbol{y})} = \frac{\mathcal{G}(\boldsymbol{y})}{f(\boldsymbol{y})} w(\boldsymbol{\theta}, \boldsymbol{y}), \tag{5.3}$$

hvor $w(\boldsymbol{\theta}, \boldsymbol{y}) = f(\boldsymbol{\theta}, \boldsymbol{y})/\mathcal{G}(\boldsymbol{\theta}, \boldsymbol{y})$ betegnes **importance vægten**. Ved at bemærke, at størrelsen $\mathcal{G}(\boldsymbol{y})/f(\boldsymbol{y})$ ikke afhænger af $\boldsymbol{\theta}$, kan denne findes ved, at betragte tilfældet, hvor $\boldsymbol{x}(\boldsymbol{\theta}) = 1$. I dette tilfælde giver (5.2) og (5.3), at

$$1 = \frac{\mathcal{G}(\boldsymbol{y})}{f(\boldsymbol{y})} \mathbb{E}_{\mathcal{G}}\left[w(\boldsymbol{\theta}, \boldsymbol{y})\right],$$

hvorved

$$\frac{\mathcal{G}(\boldsymbol{y})}{f(\boldsymbol{y})} = \frac{1}{\mathbb{E}_{\mathcal{G}}\left[w(\boldsymbol{\theta}, \boldsymbol{y})\right]}. \tag{5.4}$$

Dvs.

$$\mathbb{E}\left[\boldsymbol{x}(\boldsymbol{\theta})|\boldsymbol{y}\right] = \frac{\mathbb{E}_{\mathcal{G}}\left[\boldsymbol{x}(\boldsymbol{\theta})w(\boldsymbol{\theta},\boldsymbol{y})\right]}{\mathbb{E}_{\mathcal{G}}\left[w(\boldsymbol{\theta},\boldsymbol{y})\right]}. \tag{5.5}$$

Lad nu $\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(N)}$ være værdier der er simuleret ud fra en fordeling med importance tæthedsfunktionen $\mathcal{G}\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$. Den betingede middelværdi i (5.1) kan herved estimeres ud fra (5.5), ved gennemsnittet

$$\bar{\boldsymbol{x}} = \frac{\sum_{i=1}^{N}\boldsymbol{x}(\boldsymbol{\theta}^{(i)})w(\boldsymbol{\theta}^{(i)},\boldsymbol{y})}{\sum_{i=1}^{N}w(\boldsymbol{\theta}^{(i)},\boldsymbol{y})}, \tag{5.6}$$

hvor $\bar{\boldsymbol{x}}$ er estimatet for $\mathbb{E}_{\mathcal{G}}\left[\boldsymbol{x}(\boldsymbol{\theta})|\boldsymbol{y}\right]$.

5.1.1 Estimation af hyperparametre

En metode til estimation af ukendte hyperparametre, kan opnås ved, at tage udgangspunkt i (5.4). Hvis de ukendte hyperparametre i den betragtede dynamiske model betegnes $\boldsymbol{\psi}$, opnås ud fra (5.4), at likelihoodfunktionen

$$L\left(\boldsymbol{\psi}|\boldsymbol{y}\right) = \mathcal{G}\left(\boldsymbol{y}\right)\mathbb{E}_{\mathcal{G}}\left[w(\boldsymbol{\theta},\boldsymbol{y})\right], \tag{5.7}$$

hvor $L\left(\boldsymbol{\psi}|\boldsymbol{y}\right)$ er givet ved $f\left(\boldsymbol{y}\right)$. Lad $\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(N)}$ være værdier der er simuleret ud fra en fordeling med importance tæthedsfunktionen. Et simuleringsbaseret estimat for $L\left(\boldsymbol{\psi}|\boldsymbol{y}\right)$ kan da opnås ud fra (5.7), ved

$$\hat{L}\left(\boldsymbol{\psi}|\boldsymbol{y}\right) = L_{\mathcal{G}}\left(\boldsymbol{\psi}|\boldsymbol{y}\right)\frac{\sum_{i=1}^{N}w(\boldsymbol{\theta}^{(i)},\boldsymbol{y})}{N}, \tag{5.8}$$

hvor $L_{\mathcal{G}}\left(\cdot\right) = \mathcal{G}\left(\boldsymbol{y}\right)$ er likelihoodfunktionen af $\boldsymbol{\psi}$ for den Gaussiske importance tæthedsfunktion. Det kan bemærkes, at approksimationen af likelihoodfunktionen, givet ved (5.8), blot er en vægtet udgave af den Gaussiske likelihoodfunktion $L_{\mathcal{G}}\left(\boldsymbol{\psi}|\boldsymbol{y}\right)$.

5.1.2 Den DGLM med Gaussisk systemligning

For den DGLM med Gaussisk systemligning, som er gennemgået i forgående kapitel, er metoderne i importancesampling særligt simple. Da det er denne model der sigtes mod i dette speciale, gennemgås teorien specielt for den.

Det første der kan bemærkes for den DGLM med Gaussisk systemligning er, jvf. Durbin and Koopman [2001, s. 191], at $f\left(\boldsymbol{\theta}\right) = \mathcal{G}\left(\boldsymbol{\theta}\right)$. Heraf følger, at udtrykket for importancevægten bliver særlig simpelt, da vægten, ud fra Bayes sætning D.2 og uafhængighedsstrukturerne i den DGLM, er givet ved

$$w(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f\left(\boldsymbol{\theta},\boldsymbol{y}\right)}{\mathcal{G}\left(\boldsymbol{\theta},\boldsymbol{y}\right)} = \frac{f\left(\boldsymbol{\theta}\right)f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)}{\mathcal{G}\left(\boldsymbol{\theta}\right)\mathcal{G}\left(\boldsymbol{y}|\boldsymbol{\theta}\right)} = \frac{f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)}{\mathcal{G}\left(\boldsymbol{y}|\boldsymbol{\theta}\right)} = \frac{f\left(\boldsymbol{y}|\boldsymbol{\lambda}\right)}{\mathcal{G}\left(\boldsymbol{y}|\boldsymbol{\lambda}\right)}, \tag{5.9}$$

hvor $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_r]^{\mathrm{T}}$ er den stakkede vektor af signalerne. Importancevægten i (5.9) har den fordel, at udtrykkene i den er givet ud fra definitionen af den DGLM med Gaussisk systemligning, og ud fra det udvidede Kalmanfilter med iterationsproceduren introduceret i afsnit 4.3.1 (denne benævnes blot det udvidede Kalmanfilter i dette kapitel).

Det eneste der mangler for at kunne anvende importance sampling til estimation i den DGLM med Gaussisk systemligning er, at finde importance tæthedsfunktionen $\mathcal{G}(\boldsymbol{\theta}|\boldsymbol{y})$ til at simulere fra. Denne kan ligeledes findes ud fra det udvidede Kalmanfilter. For at gøre dette anvendes metoden fra Durbin and Koopman [2001, afs. 11.9]. Her benyttes, at $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r$ jvf. definition 4.2 er uafhængige, og givet ved

$$\boldsymbol{\omega}_t = \boldsymbol{\theta}_t - \boldsymbol{G}_t \boldsymbol{\theta}_{t-1}$$

for $t = 1, \ldots, r$. Da fordelingen af $\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_r$ givet $\boldsymbol{y}$ er givet ved det udvidede Kalmanfilter, og $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r$ er uafhængige, kan fordelingen af $\boldsymbol{\omega} = [\boldsymbol{\omega}_1^{\mathrm{T}}, \ldots, \omega_r^{\mathrm{T}}]^{\mathrm{T}}$ givet $\boldsymbol{y}$ herved opnås. Ud fra denne kan der så simuleres, og hvis der ønskes, kan de simulerede værdier af $\boldsymbol{\theta}$ udregnes ud fra systemligningen

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t.$$

Inferens for dynamiske modeller omfatter også hypotesetest for dynamiske stokastiske parametre og modelkontrol. Da der ikke er forsket særlig meget i disse områder, og tiden jo løber, arbejdes dog ikke videre med emnerne her.

Del II

Illustrerende eksempler

6 Eksempler til den DGLM

Dette kapitel indeholder illustrerende eksempler til de to versioner af den DGLM fra kapitel 4. Formålet med dette kapitel er, at vise hvorledes modellerne kommer til, at se ud for to konkrete fordelinger. De anvendte fordelinger er poissonfordelingen og binomialfordelingen. Eksemplet med poissonfordelingen vil kun blive gennemgået teoretisk, mens eksemplet med binomialfordelingen også vil blive illustreret ved et simuleret datasæt.

Der vil være en del ting som er gentaget under begge fordelinger. Disse kommentarer er medtaget begge steder for, at eksemplerne rimeligvis kan læses hver for sig.

6.1 Den DGLM med poissonfordelte observationer

Det antages i dette afsnit, at $Y_t \sim \text{Poi}(\mu_t)$ hvorved der, jvf. appendiks A, gælder, at

$$f\left(y_t|\eta_t, V_t\right) = \exp\Big[y_t \underbrace{\ln\left(\mu_t\right)}_{\eta_t} - \underbrace{\exp\left(\eta_t\right)}_{b(\eta_t)} + \underbrace{\ln\left(y_t!\right)}_{c(y_t, V_t)}\Big], \tag{6.1}$$

for $y_t \in \mathbb{N}$. For alle andre værdier af y_t er $f\left(y_t|\eta_t, V_t\right) = 0$. Det kan bemærkes, at $\varphi_t = 1/a(V_t) = 1$.

Der anvendes i dette afsnit kanonisk link, hvorved

$$\ln(\mu_t) = \eta_t = \lambda_t = \boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{\theta}_t.$$

Modellen gennemgåes først i det generelle tilfælde, hvor systemligningen og begyndelsesinformationen er delvist specificeret ved

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim [\boldsymbol{0}, \boldsymbol{W}_t]$$

og

$$(\boldsymbol{\theta}_0|D_0) \sim [\boldsymbol{m}_0, \boldsymbol{C}_0],$$

for kendte momenter $\boldsymbol{m}_0$ og $\boldsymbol{C}_0$.

6.1.1 Filtrering

Opdateringen i filteret for den DGLM vises ud fra den generelle teori i afsnit 4.2.

A priorifordelingerne

A priorifordelingen af $\boldsymbol{\theta}_t$ følger direkte af den generelle teori, og er dermed givet ved

$$(\boldsymbol{\theta}_t|D_{t-1}) \sim \Big[\underbrace{\boldsymbol{G}_t\boldsymbol{m}_{t-1}}_{\boldsymbol{a}_t}, \underbrace{\boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^{\mathrm{T}} + \boldsymbol{W}_t}_{\boldsymbol{R}_t}\Big].$$

Da der anvendes kanonisk link er $\eta_t = \lambda_t$ for $t = 1, \ldots, r$. Dermed kan λ_t anvendes alle steder på η_t's plads.

Jævnfør afsnit A.1 antages, at $(\lambda_t|D_{t-1})$ er fordelt ved den konjugerede fordeling til poissonfordelingen, dvs. log-gammafordelingen. A priorifordelingen af λ_t er dermed givet ved

$$(\lambda_t|D_{t-1}) \sim \text{log-gamma}(\alpha_t, \beta_t), \tag{6.2}$$

hvor tæthedsfunktionen jvf. (A.10) er givet ved

$$f(\lambda_t|D_{t-1}) = \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)}\exp(\alpha_t\lambda_t - \beta_t\exp(\lambda_t)). \tag{6.3}$$

En metode til at bestemme α_t og β_t i (6.2) er givet i West et al. [1985, s. 79] og West and Harrison [1997, s. 529-530]. Den går ud på, at approksimere størrelserne således, at modus i tæthedsfunktionen (6.3) er lig $f_t = \boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{a}_t$, og krumningen af logaritmen til tæthedsfunktionen (6.3), evalueret i modus, er lig $1/q_t$, hvor $q_t = \boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{R}_t\boldsymbol{F}_t$.

Modus bestemmes ved at finde maksimum af (6.3). Da den afledte af (6.3) er givet ved

$$\frac{d\, f\,(\lambda_t|D_{t-1})}{d\,\lambda_t} = \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)}\exp\left(\alpha_t\lambda_t - \beta_t\exp\left(\lambda_t\right)\right)\left(\alpha_t - \beta_t\exp\left(\lambda_t\right)\right)$$

opnås, at

$$f_t = \ln\left(\frac{\alpha_t}{\beta_t}\right).$$

Krumningen findes ved minus den anden afledede af logaritmen til tæthedsfunktionen (6.3), givet ved

$$-\frac{d^{\,2} f\,(\lambda_t|D_{t-1})}{d\,\lambda_t^2} = \beta_t\exp\left(\lambda_t\right).$$

Evalueret f_t giver dette, at

$$\frac{1}{q_t} = \alpha_t.$$

Ved at løse de to ligninger opnås, at

$$\alpha_t = \frac{1}{q_t} \quad \text{og} \quad \beta_t = \frac{1}{q_t\exp\left(f_t\right)}.$$

Alternativt kan der anvendes nummeriske metoder til, at bestemme α_t og β_t således, at f_t er lig middelværdien af (6.2), og q_t er lig variansen af (6.2).

A posteriorifordelingerne

Da $\varphi_t = 1$ opnås a posteriorifordelingen af λ_t direkte fra (6.4) i den generelle teori, ved

$$(\lambda_t|D_t) \;\sim\; \text{log-gamma}\left(\alpha_t^*, \beta_t^*\right), \tag{6.4}$$

hvor

$$\alpha_t^* = \alpha_t + y_t \quad \text{og} \quad \beta_t^* = \beta_t + 1.$$

Momenterne til a posteriorifordelingen er da jvf. (A.9) givet ved

$$\begin{aligned} \mathrm{f}_t^\star &= \mathbb{E}\left[\lambda_t|D_t\right] &&= \gamma\left(\alpha_t + y_t\right) - \ln\left(\beta_t + 1\right) \\ \mathrm{q}_t^\star &= \mathbb{V}\mathrm{ar}\left[\lambda_t|D_t\right] &&= \gamma'\left(\alpha_t + y_t\right). \end{aligned}$$

Den approksimative a posteriorifordeling af $\boldsymbol{\theta}_t$ er herefter givet direkte ud fra (4.11) i den generelle teori, ved

$$(\boldsymbol{\theta}_t|D_t) \,\dot{\sim}\, \left[\boldsymbol{a}_t + \boldsymbol{R}_t\boldsymbol{F}_t(\mathrm{f}_t^\star - f_t)/q_t, \boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{F}_t\boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{R}_t(\mathrm{q}_t^\star - q_t)/q_t^2\right].$$

6.1.2 Gaussisk systemligning

I denne model er det antaget, at systemligningen og begyndelsesbetingelsen er fuldstændigt specificeret ved

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{W}_t)$$

og

$$(\boldsymbol{\theta}_0|D_0) \sim \mathcal{N}_n(\boldsymbol{m}_0, \boldsymbol{C}_0),$$

for kendte momenter $\boldsymbol{m}_0$ og $\boldsymbol{C}_0$.

Da filtreringen i den DGLM med Gaussisk systemligning opnås ved, at approksimere modellen med en DLM, er de eneste størrelser der skal findes i dette tilfælde $\widetilde{Y}$ og $\widetilde{V}$. Eftersom der anvendes kanonisk link i dette eksempel, ses af (4.17), at

$$\widetilde{Y}_t = \frac{Y_t - b'(\hat{\lambda}_t)}{b''(\hat{\lambda}_t)} + \hat{\lambda}_t \quad \text{og} \quad \widetilde{V}_t = \frac{a(V_t)}{b''(\hat{\lambda}_t)}.$$

Ud fra dette ses, at $b'(f_t)$, $b''(f_t)$ og φ_t er det eneste, der skal benyttes for at udlede den approksimerende DLM. Jvf. (6.1) er $\varphi_t = 1$ og $b(f_t) = \exp(f_t)$. Heraf fås, at

$$b'(f_t) = \exp(f_t) \quad b''(f_t) = \exp(f_t).$$

Estimatet for fordelingen af den latente proces kan herved opnås ved den iterative procedure beskrevet i afsnit 4.3.1.

6.2 Den DGLM med binomialfordelte observationer

Det antages i dette afsnit, at $X_t \sim \text{Bi}(n_t, p_t)$ hvorved der jvf. appendiks A gælder, at

$$f(x_t) = \binom{n_t}{x_t} p_t^{x_t}(1-p_t)^{n_t - x_t}, \tag{6.5}$$

for $x_t = 0, 1, \ldots, n_t$. For alle andre værdier af x_t er $f(x_t) = 0$.
Jvf. appendikset, gælder ækvivalent, at

$$f(y_t|\eta_t, V_t) = \exp\Big[n_t\big(y_t \underbrace{\text{logit}(p_t)}_{\eta_t} - \underbrace{\ln(1+\exp(\text{logit}(p_t)))}_{b(\eta_t)}\big) + \underbrace{\ln\left(\binom{n_t}{y_t}\right)}_{c(y_t, V_t)}\Big], \tag{6.6}$$

hvor $\varphi_t = 1/a(V_t) = n_t$, og $y_t = x_t/n_t$.

Der anvendes i dette afsnit kanonisk link, hvorved

$$\text{logit}\,(\mu_t) = \eta_t = \lambda_t = \boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{\theta}_t.$$

Modellen gennemgåes først i det generelle tilfælde, hvor systemligningen og begyndelsesinformationen er delvist specificeret ved

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim [\boldsymbol{0}, \boldsymbol{W}_t]$$

og

$$(\boldsymbol{\theta}_0|D_0) \sim [\boldsymbol{m}_0, \boldsymbol{C}_0],$$

for kendte momenter $\boldsymbol{m}_0$ og $\boldsymbol{C}_0$.

6.2.1 Filtrering

Opdateringen i filteret for den DGLM vises ud fra den generelle teori i afsnit 4.2.

A priorifordelingerne

A priorifordelingen af $\boldsymbol{\theta}_t$ følger direkte af den generelle teori, og er dermed givet ved

$$(\boldsymbol{\theta}_t|D_{t-1}) \sim \Big[\underbrace{\boldsymbol{G}_t\boldsymbol{m}_{t-1}}_{\boldsymbol{a}_t}, \underbrace{\boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^{\mathrm{T}} + \boldsymbol{W}_t}_{\boldsymbol{R}_t}\Big].$$

Da der anvendes kanonisk link er $\eta_t = \lambda_t$ for $t = 1, \ldots, r$. Dermed kan λ_t anvendes alle steder på η_t's plads.

Jvf. afsnit A.2 antages, at $(p_t|D_{t-1})$ er fordelt ved den konjugerede fordeling til binomialfordelingen, dvs. betafordelingen. A priorifordelingen af p_t er dermed givet ved

$$(p_t|D_{t-1}) \sim \text{Beta}(\alpha_t, \beta_t),$$

med tilhørende tæthedsfunktion

$$f\,(p_t) = \frac{1}{B\,(\alpha_t, \beta_t)} p_t^{\alpha_t - 1}(1 - p_t)^{\beta_t - 1}, \tag{6.7}$$

for $0 < p_t < 1$. For $p_t \leq 0$ eller $p_t \geq 1$ er $f\,(p_t) = 0$. Størrelsen $\mathrm{B}\,(\cdot, \cdot)$ er betafunktionen givet ved (A.13).

Parametrene α_t og β_t til a priorifordelingen for p_t, opnås tilsvarende metoden anvendt under poissonfordelingen.

Middelværdien af $\lambda_t = \text{logit}\,(p_t)$ er jvf. opdateringen i filteret givet ved f_t, og variansen er givet ved q_t. Under den ovenstående antagelse for a priorifordelingen af λ_t opnås jvf. West and Harrison [1997, s. 529-530], at

$$f_t = \gamma\,(\alpha_t) - \gamma\,(\beta_t) \quad \text{og} \quad q_t = \gamma'\,(\alpha_t) + \gamma'\,(\beta_t)\,, \tag{6.8}$$

hvorved f_t og q_t kan approksimeres ved

$$f_t \doteq \ln\left(\frac{\alpha_t}{\beta_t}\right) \quad \text{og} \quad q_t \doteq \frac{1}{\alpha_t} + \frac{1}{\beta_t}. \tag{6.9}$$

Ved at invertere (6.9) bestemmes α_t og β_t til

$$\alpha_t = \frac{1 + \exp\,(f_t)}{q_t} \quad \text{og} \quad \beta_t = \frac{1 + \exp\,(-f_t)}{q_t}.$$

Alternativt kan der anvendes nummeriske metoder til, at bestemme α_t og β_t således, at f_t er lig middelværdien af den antagende a priorifordeling af λ_t, og q_t er lig den tilhørende varians.

A posteriorifordelingerne

Den approksimative a posteriorifordeling for p_t opnås ved Bayes sætning

$$f\,(p_t|D_t) \propto f\,(x_t|p_t)\,f\,(p_t|D_{t-1})\,.$$

De to tætheder på højresiden tilhører henholdsvis en binomialfordelingen og en betafordeling, hvorved $f\,(p_t|D_t)$, ifølge afsnit A.2.1, er betafordelt. Ud fra (6.5) og (6.7) opnås herved, at

$$(p_t|D_t) \;\sim\; \text{Beta}(\alpha_t^*, \beta_t^*),$$

hvor parametrene er givet ved

$$\alpha_t^* = \alpha_t + x_t \quad \text{og} \quad \beta_t^* = \beta_t + n_t - x_t.$$

Momenterne til den approksimative a posteriori for λ_t, kan herefter bestemmes ved (6.8), til

$$\begin{aligned} \mathrm{f}_t^\star &= \mathbb{E}\,[\lambda_t|D_t] \quad = \gamma\,(\alpha_t + x_t) - \gamma\,(\beta_t + n_t - x_t) \\ \mathrm{q}_t^\star &= \mathbb{V}\mathrm{ar}\,[\lambda_t|D_t] = \gamma'\,(\alpha_t + x_t) + \gamma'\,(\beta_t + n_t - x_t)\,. \end{aligned}$$

Den approksimative a posteriorifordeling af $\boldsymbol{\theta}_t$ er herefter givet direkte ud fra (4.11) i den generelle teori, ved

$$(\boldsymbol{\theta}_t|D_t) \dot{\sim} \left[\boldsymbol{a}_t + \boldsymbol{R}_t\boldsymbol{F}_t(\mathrm{f}_t^\star - f_t)/q_t, \boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{F}_t\boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{R}_t(\mathrm{q}_t^\star - q_t)/q_t^2\right].$$

6.2.2 Gaussisk systemligning

I denne model er det antaget, at systemligningen og begyndelsesbetingelsen er fuldstændigt specificeret ved

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{W}_t)$$

og

$$(\boldsymbol{\theta}_0|D_0) \sim \mathcal{N}_n(\boldsymbol{m}_0, \boldsymbol{C}_0),$$

for kendte momenter $\boldsymbol{m}_0$ og $\boldsymbol{C}_0$.

Da filtreringen i den DGLM med Gaussisk systemligning opnås ved, at approksimere modellen med en DLM, er de eneste størrelser der skal findes i dette tilfælde $\widetilde{Y}$ og $\widetilde{V}$. Eftersom der anvendes kanonisk link i dette eksempel, ses af (4.17), at

$$\widetilde{Y}_t = \frac{Y_t - b'(\hat{\lambda}_t)}{b''(\hat{\lambda}_t)} + \hat{\lambda}_t \quad \text{og} \quad \widetilde{V}_t = \frac{a(V_t)}{b''(\hat{\lambda}_t)}.$$

Ud fra dette ses, at $b'(f_t)$, $b''(f_t)$ og φ_t er det eneste, der skal benyttes for at udlede den approksimerende DLM. Jvf. (6.6) er $\varphi_t = n_t$ og $b(f_t) = \ln(1 + \exp(f_t))$. Heraf fås, at

$$b'(f_t) = \frac{\exp(f_t)}{1 + \exp(f_t)} \quad b''(f_t) = \frac{\exp(f_t)}{(1 + \exp(f_t))^2} \quad .$$

Estimatet for fordelingen af den latente proces kan herved opnås ved den iterative procedure beskrevet i afsnit 4.3.1.

Illustration

Den anvendte iterationsprocedure der er beskrevet til det udvidede Kalmanfilter i afsnit 4.3.1 adskiller sig fra teorien i Fahrmeir and Tutz [2001]. Da den desuden anvendes senere under dataanalysen i dette speciale, illustreres den her ved et simuleret datasæt.

Datasættet er simuleret ud fra den ovennævnte DGLM med binomialfordelte observationer, Gaussisk systemligning og kanonisk link. Der anvendes $F_t = G_t = W_t = 1$ og $n_t = 300$, for $t = 1, \ldots, 100$. Som begyndelsesværdier er der anvendt $m_0 = 0$ og $C_0 = 1$.

Den latente proces estimeres henholdsvis ved det udvidede Kalmanfilter, og ved det udvidede Kalmanfilter med iterationsproceduren beskrevet i afsnit

4.3.1. Som stopkriterier for iterationsproceduren anvendes, at

$$\left| \frac{\widetilde{m}_t^{(i)} - \widetilde{m}_t^{(i-1)}}{\widetilde{m}_t^{(i-1)}} \right| < 10^{-6},$$

for alle $t = 1, \ldots, 100$, hvor $\widetilde{m}_t$ er middelværdiestimatet for tilstandsvektoren. Indekset i angiver hvilken gang iterationen kører. Hvis $\widetilde{m}_t^{(i-1)} = 0$ anvendes kriteriet

$$\left| \widetilde{m}_t^{(i)} - \widetilde{m}_t^{(i-1)} \right|.$$

I det anvendte tilfælde konvergerer algoritmen ved 5 iterationer, hvilket stemmer meget godt over ens med Durbin and Koopman [2000], hvor der angives, at der normalt ikke skal anvendes mere end 10 iterationer for, at opnå konvergens.

Ved det øverste plot i figur 6.1 illustreres estimatet givet ved det udvidede Kalmanfilter med den beskrevne iterationsprocedure. Sammen med estimatet er der angivet punktvise intervaller givet ved

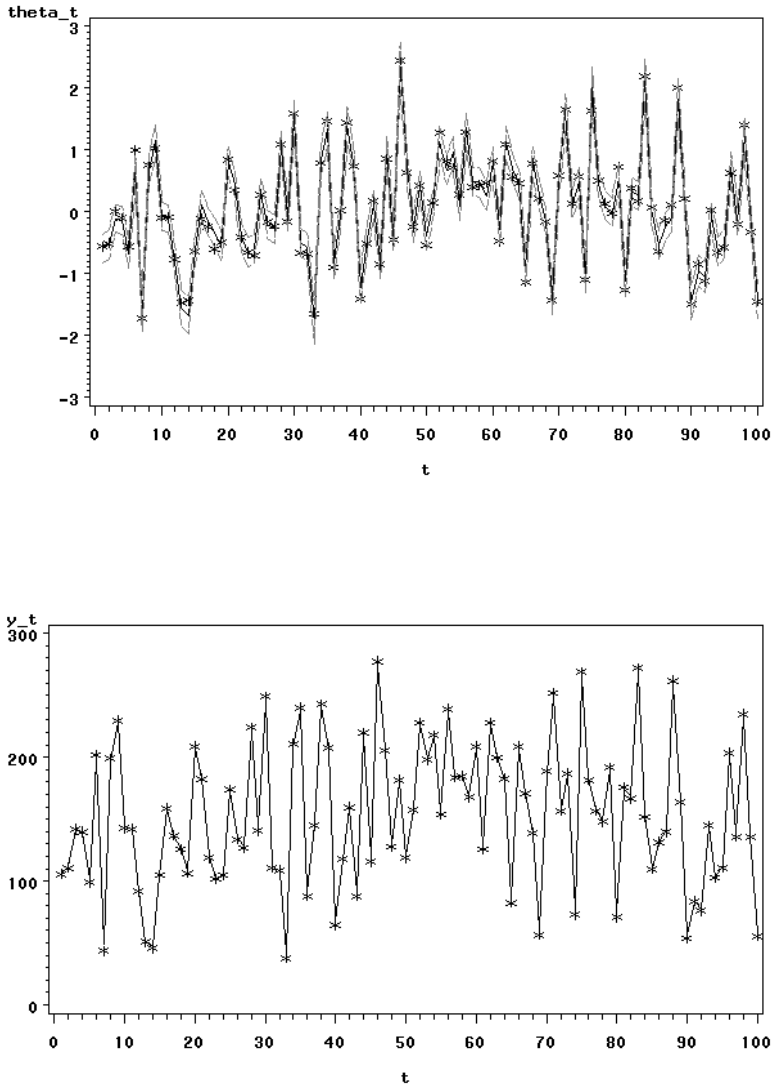
$$\widetilde{m}_t \pm 1.96\sqrt{\widetilde{C}_t},$$

hvor $\widetilde{C}_t$ er variansestimatet for θ_t, for $t = 1, \ldots, 100$.

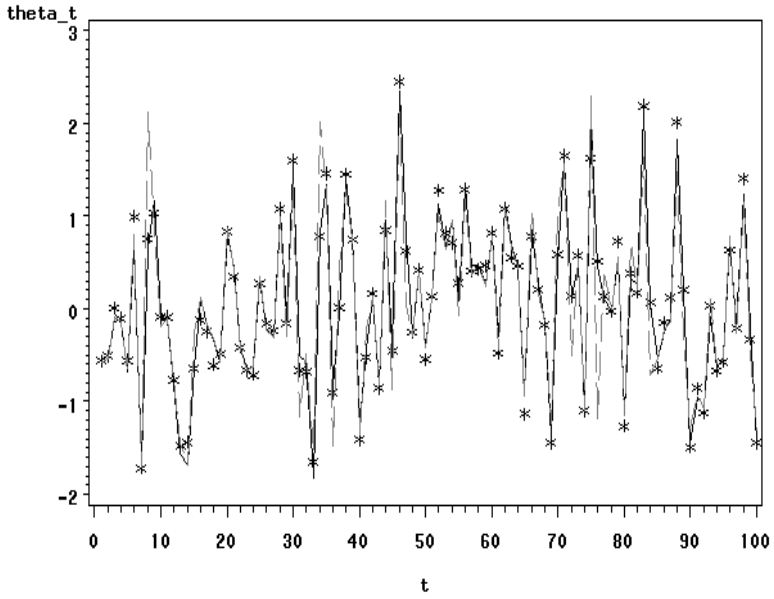
Af figuren ses, at den latente proces estimeres ret godt ved den anvendte metode. Samtidig ses, at de punktvise intervaller er ret smalle, hvilket stemmer godt over ens med det præcise estimat.

Ved det nederste plot i figur 6.1 illustreres estimationen i samme skala, som observationerne $x_1, \ldots, x_{100}$. Dette estimat er opnået ved, at anvende den inverse logit-funktionen på $\widetilde{m}_1, \ldots, \widetilde{m}_{100}$. Heraf ses, at estimatet ligeledes følger observationerne ret præcist, hvilket dog også er, at forvente, da det er dem der er anvendt til at lave estimatet. Dette tjener således mere til at evaluere approksimationerne end til, at illustrere dem.

Til sidst sammenlignes estimatet opnået udelukkende ved det udvidede Kalmanfilter, og det udvidede Kalmanfilter med den beskrevne iterationsprocedure. Dette gøres ved, at plotte estimaterne fra de to metoder sammen. Illustrationen heraf er at finde i figur 6.2, hvor det ses, at iterationsproceduren forbedrer estimatet noget. Det skal dog bemærkes, at en stor del af denne forbedring skyldes, at Kalmanudglatteren er anvendt i metoden med iteration.



Figur 6.1: Øverst: Den estimerede middelværdi for den latente proces (fuldtoptrukken linje) plottet sammen med 95% punktvise intervaller (grå stiplede linjer). De simulerede værdier af tilstandsvektoren er markeret ved stjerner.
Nederst: Den estimerede middelværdi for de simulerede værdier $x_1, \ldots, x_{100}$ (fuldtoptrukken linje) plottet sammen med $x_1, \ldots, x_{100}$ markeret ved stjerner.



Figur 6.2: Middelværdien for den latente proces estimeret ved henholdsvis det udvidede Kalmanfilter (stiplet linje), og ved det udvidede Kalmanfilter med iteration (fuldtoptrukken linje). De simulerede tilstandsvektorer er markeret ved stjerner.

Del III

Anvendelse

7 Baggrund for dataanalysen

Anvendelsesdelen af dette speciale omfatter en analyse af data med relation til skizofreni i Danmark. Dataanalysen er udført hos Center for Registerforskning ved Aarhus Universitet, under vejledning af lektor Rodrigo Labouriau.

Anvendelsesdelen er inddelt i fire kapitler. Det første kapitel indeholder en kort beskrivelse af sygdommen skizofreni, en introduktion til den problemstilling der ønskes belyst, samt en introduktion til det datamateriale der anvendes. Det andet kapitel indeholder en dataanalyse baseret på den GLM, der er introduceret i kapitel 3. Det tredie kapitel indeholder en dataanalyse baseret på den DGLM, der er introduceret i kapitel 4. Det fjerde kapitel er en kort opsummering af anvendelsesdelen, samt en samlet diskussions af analysen baseret på den GLM, og analysen baseret på den DGLM.

7.1 Skizofreni

Dette afsnit indeholder en kort gennemgang af hvad sygdommen skizofreni er, samt lidt om de undersøgelser der er udarbejdet omkring risikofaktorer for sygdommen.
En mere dybdegående redegørelse for sygdommen og især dens behandlingsmuligheder er bla. givet i Gronemann [1995] og Sørensen [2000].

7.1.1 Sygdommen skizofreni

Skizofreni er en alvorlig og ret hyppig sindslidelse. I det anvendte datasæt er den samlede frekvens for tidlig skizofreni på ca. 0.23%. Op imod halvdelen af alle patienter på de danske psykiatriske institutioner har ifølge [Gronemann, 1995, s. 18] fået diagnosen skizofreni (dette tal stemmer jvf. Rodrigo Labouriau godt over ens med tallene for det anvendte register).

På nuværende tidspunkt findes der ikke effektive behandlingsmuligheder for sygdommen, og man kender ikke præcist hvilke faktorer der kan have indvirkning dens opståen og udvikling. Diagnosen skizofreni bliver typisk stillet mens patienten er ung. Således får over 67% af skizofrenipatienterne diagnosen første gang i en alder af 15-25 år [Westergaard et al., 1999, s. 995].

I Sørensen [2000] indføres en diagnose af skizofreni baseret på diagnostiseringssystemet ICD10. En kort gennemgang af denne diagnose kan gives ved, at mindst et af følgende punkter skal være opfyldt for, at der er tale om skizofreni.

(1) Mindst et første rangs symptom, hvilket vil sige et af følgende punkter:

Tankepåvirkningsoplevelser. Den skizofrene oplever at nogen eller noget påvirker individets tanker, eller stjæler tankerne.

Tredje-persons hørehallucinationer. Det vil sige, at den skizofrene forestiller sig stemmer, som ikke er der.

Styringsoplevelser. Den skizofrene oplever, at få påført handlinger, viljeimpulser eller følelser fra omverdenen.

Legemlige påvirkningsoplevelser. Den skizofrene oplever, at blive påvirket fysisk af omverdenen.

Derealisation. Den skizofrene oplever omverden uvirkelig. F.eks. kan patienten opleve at alle andre er lavet af plastic.

(2) Vedvarende bizarre vrangforestillinger. Dvs. fuldstændig umulige, og for samfundet uacceptable vrangforestillinger.

(3) Mindst to af følgende punkter skal være opfyldt

Vedvarende hallucinationer med vrangforestillinger

uden affektivt indhold. Dvs. hallucinationer med vrangforestillinger, uden tilknyttede sygeligt følesesmæssigt indhold.

Sproglige tankeforstyrrelser.

Katoni. Fysiske reaktioner som kan give sig udslag i stivhed, rastløshed eller lignende.

Negative symptomer. Dette kan være, at den skizofrene opbygger sin egen logik og tankeverden, som ofte kan være svær at forstå for udenforstående. Det kan også være, at den skizofrene kan have meget svært ved at beslutte sig, at den skizofrene udviser en meget stærke, eller meget svage føleser for sine omgivelser, og lignende.

Desuden skal disse symptomer have en varighed på mindst en måned for, at der kan være tale om skizofreni. Diagnosen skizofreni udelukkes, hvis der er diagnosticeret en primær affektiv sindslidelse, som f.eks. depression, eller hvis der er tale om organisk ætiologi, som f.eks. kan skyldes indtagelse af stoffer. Grunden til dette er, at disse faktorer kan fremkalde nogle af de samme symptomer, som skizofreni.

En dybere indførsel til skizofreni er bla. givet i Sørensen [2000]. En fuldstændig gennemgang af ICD10 diagnostiseringssystemet for skizofreni er at finde i WHO [1992, F20-F29]. En tilsvarende gennemgang af ICD8 systemet er at finde i WHO [1967].

Nogle af de behandlingsmuligheder man har i dag, er social-psykologisk behandling, medikamentel behandling, psykoterapi og kropsterapi. Da der ikke bliver arbejdet med selve behandlingen i dette speciale, vil disse metoder dog ikke blive diskuteret yderligere her.

7.1.2 Risikofaktorer

For at undersøge baggrunden for at nogle mennesker udvikler skizofreni, er der lavet en mængde undersøgelser af hvilke faktorer, der kan have indvirkning på sygdomsfrekvensen. Nogle af de faktorer der har være genstand for undersøgelse, er som følger.

Biologiske faktorer

- **Familiær disposition.** Dette dækker over hvor vidt skizofrenifrekvensen afhænger af, om den nærmeste familie har fået diagnosen skizofreni. Specielt er der lavet en del undersøgelser for tvillinger. Blandt de faktorer som er undersøgt, giver familiær disposition den største forøgelse af skizofrenifrekvensen [Mortensen et al., 1999].

Sociale faktorer

- **Socialklasse.** Dette dækker over, om der er en sammenhæng mellem hvilken social klasse man tilhører, og risikoen for at udvikle skizofreni.

- **Household crowding.** Dette dækker over hvor tæt man bor. Herunder antallet af personer i husstanden.
- **Søskendeflok.** Denne parameter relaterer dels til størrelsen af den søskendeflok, man kommer fra, og dels til aldersintervallet til den nærmeste ældre eller yngre søskende. I Westergaard et al. [1999] påvises det, at aldersintervallet til den nærmeste søskende, samt størrelsen af søskendeflokken, kan have indvirkning på skizofrenifrekvensen, hvorimod nummeret i søskendeflokken ikke kan påvises at have nogen indvirkning.

Graviditets og fødsesrelaterede faktorer

- **Fødselstidspunkt.** Dette dækker over hvilket tidspunkt på året den skizofrene er født. Der er påvist en større risiko for at udvikle sygdommen, hvis man er født i februar eller marts, og en lavere risiko hvis man er født i august eller september [Mortensen et al., 1999].
- **Komplikationer ved graviditet og fødsel.**
- **Influenza og andre infektioner hos moderen.** Der har været undersøgelser af, om influenza og infektioner under graviditeten kan have indvirkning på skizofrenifrekvensen. Hvorvidt dette er tilfældet er dog stadig et åbent spørgsmål [Mortensen, 2001, s. 4718].
- **Udendørstemperatur.** Udendørstemperaturen seks måneder før fødslen har være forslået som parameter [Kendel and Adams, 1991].
- **Amning.** Dette er en parameter for om man er blevet ammet eller ej.

Demografiske faktorer

- **Urbanisering.** Dette er en parameter for om man er opvokset på landet, eller i givet fald hvilken størrelse by man er opvokset i. Således betegner en højere grad af urbanisering, at man er født og opvokset i en større by. Der er påvist en markant sammenhæng mellem en højere grad af urbanisering, og en større skizofrenifrekvens [Torrey et al., 2000] og [Mortensen et al., 1999].
- **Geografisk ophobning.** Dette dækker over, at der i visse geografiske områder, kan være en større risiko for at udvikle skizofreni end i andre. Fænomenet kan dog delvist forklares ved urbanisering [Torrey et al., 2000].

Andre faktorer

- **Kost** Denne parameter baserer sig på om mennesker, som f.eks. har været udsat for hungerkatastrofer i fosterlivet, kan have en øget risiko for at udvikle skizofreni.
- **Giftstoffer** Dette dækker over hvor vidt man udsættes for forskellige giftstoffer som bly, sprøjtegifte og lignende.
- **Hovedtraumer.** Dette er traumer fremkommet som følge af f.eks. bilulykker eller vold.
- **Smitte fra husdyr.** Dette dækker over infektioner overført fra husdyr.

Ovenstående liste er baseret på Mortensen [2001]. For en dybere introduktion kan bla. anvendes Jablensky [1997]. Som det kan ses af listen, forskes der på mange områder for, at finde de faktorer, som ligger til grund for udviklingen af skizofreni. Det er dog stadig uvist, hvad der præcist er tale om, og om der er mulighed for at forebygge nogle af disse faktorer.

7.2 Overordnet problemstilling

Den overordnede problemstilling er opsat af Preben Bo Mortensen, centerleder ved Center for Registerforskning. Problemstillingen er, at undersøge om der kan være påvirkninger knyttet til fødselsmåneden, der kan øge risikoen for senere, at udvikle **tidlig skizofreni**. Med tidlig skizofreni menes patienter, der første gang får diagnosen skizofreni, i en alder af 15 til og med 25 år. Således betragtes skizofrenifrekvenser, givet ved antallet af skizofrene, der er født i en given måned og et givet år, i forhold til det totale antal børn født i denne måned og dette år. De spørgsmål der konkret ønskes besvaret, er som følger.

(1) Er der afhængighed mellem skizofrenifrekvenserne og fødselsmåneden og året ?

(2) Hvis der er afhængighed, hvilke faktorer kan da have en indflydelse på skizofrenifrekvensen?

7.3 Databeskrivelse

De anvendte datakilder er det danske CPR-nummer register, samt det danske psykiatriske register.

Betingelser for dataudvælgelse

CPR-nummer registeret, og dermed det psykiatriske register, er oprettet d. 1/4 1969, og det datasæt der er anvendes går frem til 1/1 1999. Da datasættet først er taget ud senere er der ikke "reporting delay", dvs. alle personer der har fået skizofrenidiagnosen inden 1/1 1999, er med i datasættet. Efter som der er valgt, at se på personer med tidlig skizofreni, er der følgende kriterier for personerne i det anvendte datasæt.

- Ud fra den anvendte definition af tidlig skizofreni, betragtes udelukkende personer, der har fået diagnosen skizofreni første gang i en alder af 15 til og med 25 år.
- For at undgå strukturelle censorering, skal personerne være født mellem d. 1. januar 1955 og d. 1. januar 1973. Dette valg er begrundet i, at en person født i 1955 bliver 15 år i 1970, og en person født i 1973 bliver 25 år i 1998.
- Da der betragtes personer der har fået diagnosen skizofreni efter deres 15 års fødselsdag, medtages kun personer som har overlevet ind til denne dag.
- Da der ønskes en undersøgelse af danske patienter, medtages kun personer født i Danmark. Der medtages ikke personer født på Grønland eller Færøerne. Dette valg begrunder sig i at nogle af de faktorer, der muligvis kan have en indvirkning på skizofrenifrekvensen, er forskellige for selve Danmark og Grønland/Færøerne.
- Personer som er hjemløse, og derfor ikke bliver registreret i det psykiatriske system udelades. Det skal bemærkes, at denne gruppe er meget lille i Danmark. Den udgør kun ca. 0.25% af hele befolkningen mellem 15 og 25 år.

Censoreringer

Da der er personer der er "droppet ud"af registeret, kan det ikke helt udgås at have censoreringer i datasættet. Således er der overvejet følgende årsager til, at personer er "droppet ud".

- Personer i datasættet der er bortrejs fra landet, i en alder af 15 til og med 25 år, uden at have fået konstateret skizofreni.
- Personer i datasættet der er døde, i en alder af 15 til og med 25 år, uden at have fået konstateret skizofreni.

Da det ikke vides, om de ovenstående grupper af personer har en anden skizofrenirisiko end alle andre, udgør de en mulig kilde til censoreringer. Grupperne udgør samlet 4.3% af de betragtede personer. Grundet den begrænsede tid, arbejdes der ikke videre med denne problemstilling.

Variable

De variabler, som indgår i det anvendte datasæt, har følgende betegnelser:

t angiver tiden i måneder, efter 1. januar 1955. Således svarer $t = 1$ til den første måned, januar 1955, og $t = 216$ svarer til den sidste måned, december 1972. Tiden t angives, i analysen, enten ved måned og år, eller ved antallet af måneder efter 1. januar 1955.

g angiver personens køn, hvor henholdsvis kvinder og mænd betegnes Fem. og Masc.

n_{tg} angiver antallet af observerede børn, af et givet køn g, der er født i Danmark i en given måned t. I hele den betragtede periode, udgør dette 1513951 personer fordelt på 735069 kvinder og 778882 mænd.

s_{tg} angiver antallet af observerede personer, af et givet køn g, der er født i Danmark i en given måned t, som senere har fået diagnosen tidlig skizofreni.
Der er en niveauforskel, mellem skizofrenifrekvensen for henholdsvis kvinder og mænd, i datasættet. Således er der for de 216 måneder samlet diagnosticeret 1086 skizofrene kvinder, og 2398 skizofrene mænd. Skizofrenifrekvensen er herved henholdsvis 0.0015 for kvinder og 0.0031 for mænd. Denne niveauforskel kan meget vel hænge sammen med, at kvinder typisk får diagnosticeret skizofreni senere end mænd [Häfner and Heiden, 1997, s. 143]. Dette kan dog ikke påvises ud fra det datasæt, som anvendes her.

m angiver fødselsmåneden jan,...,dec kodet som 1,...,12.

Desuden forventes at der er et knækpunkt i data, som følge af at man d. 1/1 1994 gik fra ICD8 til ICD10 diagnostiseringssystemet. Samtidig med denne overgang gennemførtes endvidere en efteruddannelse. Disse ting tilsammen forventes, at give en højere skizofrenifrekvens for de personer, som er diagnosticeret efter ICD10 systemet i stedet for ICD8 systemet.

NOTE: Hvis der ikke er angivet andet, anvendes 5% signifikansniveau i de følgende analyser. △

7.4 Analyseprogram

Til analysen af anvendelsesdelen i dette speciale, anvendes statistikprogrammet SAS. Dette kommer ganske naturligt af, at det var det program der var tilgængeligt ved Center for Registerforskning. Den eneste ulempe ved dette var, at programpakken ikke anvendes på statistikstudiet ved AAU, og derfor først skulle læres. En kort forklaring af udskriften fra analyserutinerne er givet i appendiks C. En mere dybdegående forklaring af de enkelte rutiner er bla. givet i SAS manualen SAS [1999].

Dataanalysen baseret på den GLM, bygger på GENMOD proceduren, og dataanalysen baseret på den DGLM, bygger på GLIMMIX makroen, som igen er en overbygning til MIXED proceduren. Desuden anvendes IML delen af SAS under begge analyseformer til, at bygge rutiner, som ikke er implementeret direkte i SAS. Dette omfatter f.eks. dele af Fischer's eksakte test for hvid støj, test for ekstra temporal variation, og det udvidede iterative Kalmanfilter.

8 Dataanalyse baseret på den GLM

For at analysere datasættet for skizofrenifrekvensen, anvendes i dette kapitel en generaliseret lineær model (GLM). I den GLM anvendes udelukkende faste parametre. Den GLM har den fordel, at der findes en stor mængde værktøjer, til hypotesetest og modelkontrol, hvilket ikke på samme måde er tilfældet for dynamiske modeller.
Den teoretiske baggrund for dataanalysen er beskrevet i kapitel 3.

8.1 Modelopstilling

Den første del af analysen består i, at opstille en model for skizofrenifrekvensen givet ved s_{tg}/n_{tg}.

8.1.1 Grundlæggende modelopstilling

Ud fra definitionen af binomialfordelingen i appendiks A, kan antallet af skizofrenitilfælde s_{tg} til tiden t, og kønnet g, beskrives ved en binomialfordeling. Størrelsen s_{tg} kan herved opfattes, som en realisation af den stokastiske

variabel

$$Z_{tg} \sim \text{Bi}(n_{tg}, p_{tg}), \quad g(p_{tg}) = \boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{\beta}$$

hvor p_{tg} er sandsynligheden for, at få skizofreni for et givet køn, til en given tid, $g(\cdot)$ er linkfunktionen, $\boldsymbol{x}_t^{\mathrm{T}}$ er den t'te række i designmatricen $\boldsymbol{X}$, og $\boldsymbol{\beta}$ er parametervektoren. Disse størrelser er givet ved definition 3.1 af den GLM.

Denne model kan omformuleres til i stedet, at beskrive skizofrenifrekvensen, som realisationer af en stokastisk variabel. Dette gøres ved, at opfatte skizofrenifrekvensen, til tiden t og kønnet g, som en realisation, $y_{tg} = s_{tg}/n_{tg}$, af den stokastiske variabel $Y_{tg} = Z_{tg}/n_{tg}$, med tæthedsfunktionen givet ved (A.11).

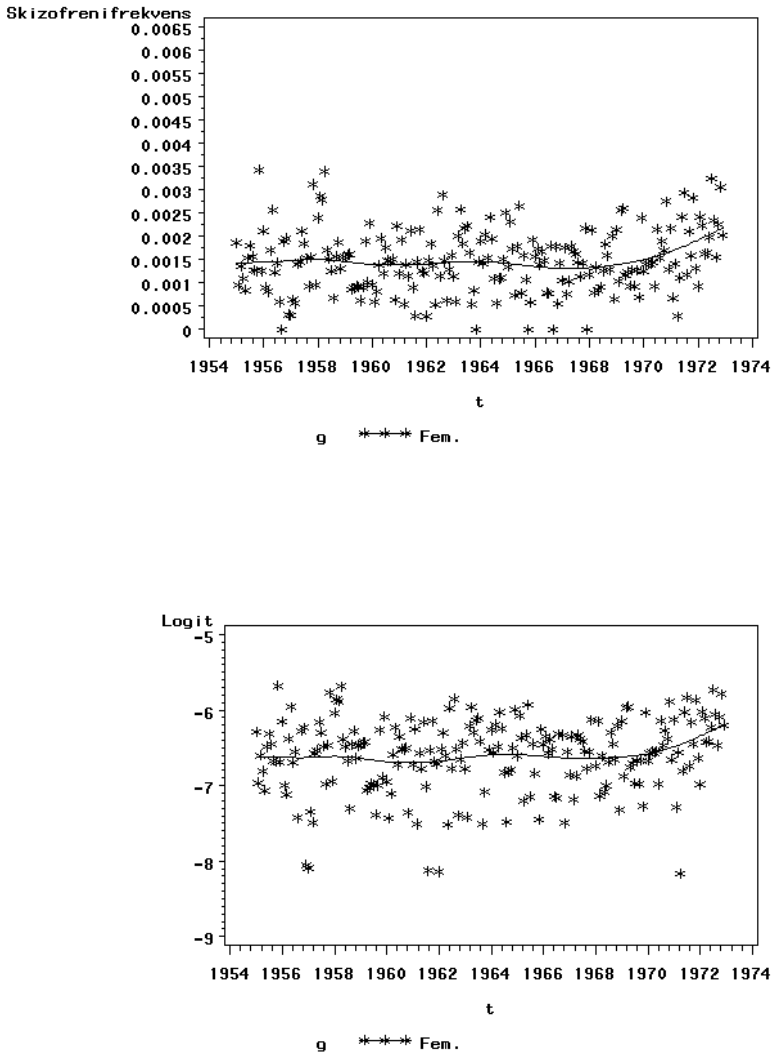
8.1.2 Indledende analyse

Det næste trin er nu, at se på hvorledes data ser ud. Da den kanoniske link for binomialfordelingen er givet ved logit $(\cdot)$, undersøges hvor vidt denne linkfunktion er anvendelig. Dvs. om der kan anvendes logistisk regression.
Da figurerne for henholdsvis mænd og kvinder, igennem hele analysen, ligner hinanden meget i struktur, vises kun figurerne for kvinder.

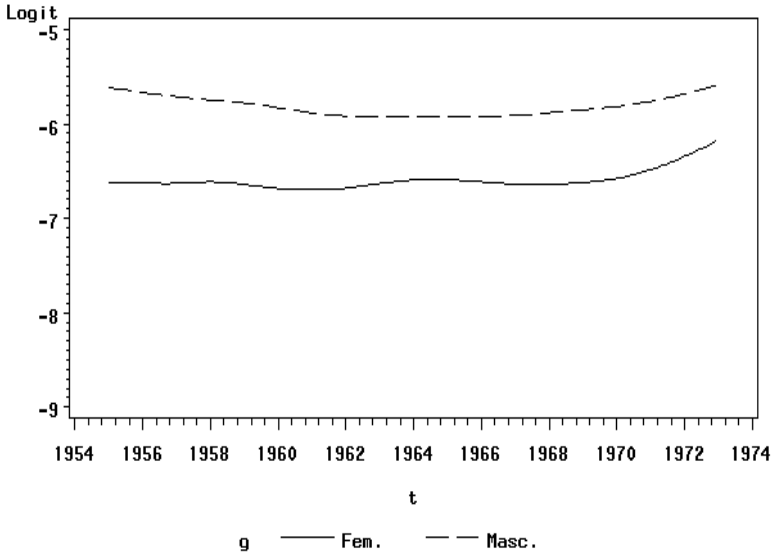
Af figur 8.1 ses, at mønsteret i data et meget ens i rå skala, og i logistisk skala. Desuden ses, at den udglattede kurve for logistisk skala tilnærmelsesvis består af rette linjer, hvilket indikerer, at der kan anvendes lineær regression til parameterestimation.
Da logistisk link, som det også ses senere, passer udemærket til at modellere disse data, vælges denne linkfunktion. Efterfølgende i denne analyse vil den anvendte skala derfor også være logistisk.
De udglattede kurver for henholdsvis kvinder og mænd er plottet sammen i figur 8.2. Heraf ses, at formen på kurverne er meget ens, blot tydeligt adskilt af niveauforskellen, i skizofrenifrekvensen for kvinder og mænd.

Det ses endvidere af de tre figurer af skizofrenifrekvensen mod tiden, at det ser ud som om, der er et knæk (change point) i skizofrenifrekvensen omkring 1968. Dette knæk er forventet, og kan umiddelbart forklares ved, at man i januar 1994, som nævnt, gik fra ICD8 til ICD10 diagnosticeringssystemet. Herved har personer født efter 1. januar 1968 en mulighed for, at få diagnosen tidlig skizofreni, under diagnostiseringssystemet ICD10. Det skal igen bemærkes, at der også blev gennemført en efteruddannelse i forbindelse med dette skift. Inden undersøgelsens start, blev det da også forventet, at skiftet ville give en stigning i antallet af personer der fik diagnosen tidlig skizofreni. Af denne årsag anvendes dette knækpunkt i den indeledende fase af modelopstillingen, men senere analyseres fænomenet nærmere.

Det sidste indledende plot, i figur 8.3, viser boxplot af skizofrenifrekvenserne i hver måned, for henholdsvis kvinder og mænd. Heraf ses, at medianen i



Figur 8.1: Skizofrenifrekvenserne for kvinder, samt en udglattet kurve for disse, i henholdsvis rå skala (øverst), og logistisk skala (nederst).

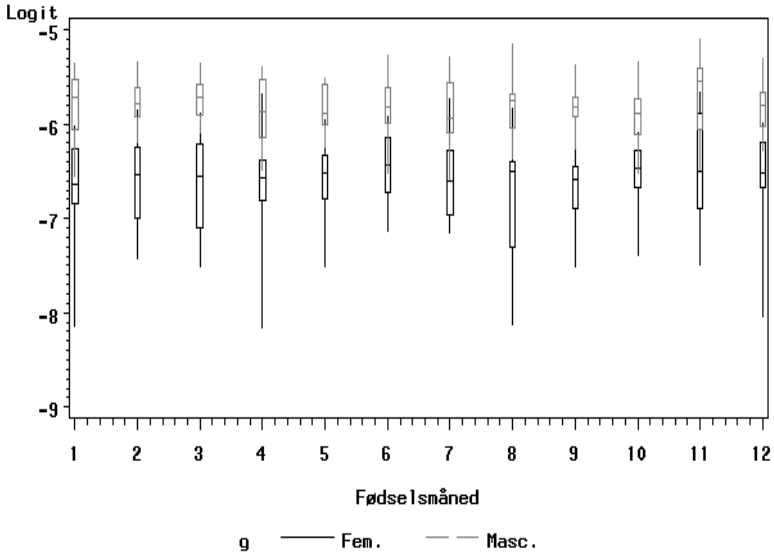


Figur 8.2: Udglattede kurver af skizofrenifrekvenserne for henholdsvis kvinder og mænd.

data ikke svinger det store over året.

8.1.3 Modelopstilling

Der opstilles først en grundmodel indeholdende alle de mulige forklarende variabler. Da der muligvis er et knækpunkt i data, indføres der en indikatorvariabel I_k, som er 1 før knækpunktet og 0 i knækpunktet og senere. Grundemodellen kommer herefter til at indeholde de forklarende variable køn i form af g, tid i form af t, knækpunktet i form af I_k, samt alle vekselvirkninger mellem disse. Heraf er grundmodellen giver ved



Figur 8.3: Skizofrenifrekvenserne i hver måned for henholdsvis kvinder og mænd. De lodrette streger går mellem den største og den mindste værdi. De vandrette streger angiver henholdsvis 25%, 50% (medianen) og 75% af data.

$$\begin{aligned}\text{logit}\,(p_{tg}) = \beta_0 + \beta_t t + \beta_g I_g + \beta_k I_k + \beta_{tk} I_k t \\ + \beta_{tg} I_g t + \beta_{gk} I_g I_k + \beta_{tgk} I_g I_k t,\end{aligned} \tag{8.1}$$

hvor

t er tiden,

I_k er en indikatorvariabel for knækpunktet (der gælder, at $I_k = 1$ før knækpunktet, og $I_k = 0$ i knækpunktet og derefter),

I_g er en indikatorvariabel for g (der gælder, at $I_g = 1$ for kvinder, og $I_g = 0$ for mænd),

β_0 er et fast referenceniveau der betegnes interseptet.

Tilsvarende er de resterende β'er koefficienter til de forklarende variable, og vekselvirkningerne mellem disse. Bemærk at alle β-værdierne er faste

størrelser, hvorved β_t ikke afhænger af tiden, men blot er koefficienten til tiden.

NOTE: Der er anvendt Pearsonresidualer til analyserne i dette speciale. Dette valg er gjort ud fra, at selv om Pearsonresidualerne "blot" er de rå residualer, skaleret med den estimerede standardafvigelse på data, kan deres fordeling, som det senere fremgår af figur 8.9, approksimeres godt ved en standard normalfordeling. Andre typer af residualer kan meget vel anvendes i stedet, men da fordelingen af Pearsonresidualerne kan approksimeres godt ved en standard normalfordeling, er disse fint anvendelige. F.eks. er værdierne af deviansresidualerne meget tæt på værdierne af Pearsonresidualerne, men et normal-fraktilplot af deviansresidualerne afviger dog mere fra referencelinjen, end det er tilfældet for Pearsonresidualerne.
En definition af Pearsonresidualerne er givet i definition 3.6, og en definition af deviansresidualerne er givet i definition 3.7. Δ

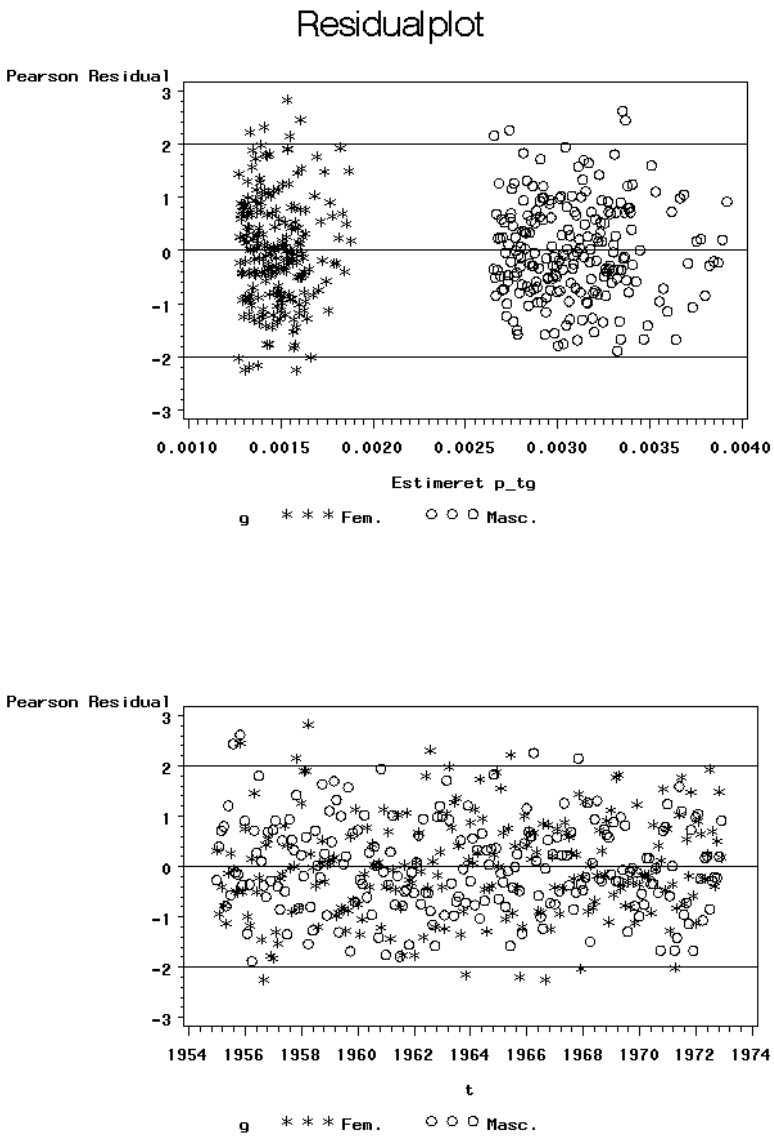
Pearsonresidualerne til grundmodellen er plottet i figur 8.4. Da værdierne af residualerne for henholdsvis kvinder og mænd ikke ser ud til, at afhænge af hverken den estimerede skizofrenisandsynlighed eller tiden, ser modellen ud til, at kunne forklare variationen i data rimeligt. Det skal specielt bemærkes, at da residualerne ikke ser ud til, at afhængige af middelværdien, er her en indikation af, at den anvendte linkfunktion passer godt.

Herefter testes hvor vidt nogle af de forklarende variable i grundmodellen kan udelades. Resultatet af denne analyse er opsummeret i tabel 8.1. (Se appendiks C for en gennemgang af metoderne.)

Koefficient	Type 1 P-værdi	Type 3 P-værdi
β_t	0.9378	0.0061
β_g	$<$.0001	0.0014
β_k	0.0003	0.0005
β_{tk}	$<$.0001	$<$.0001
β_{tg}	0.0576	0.1478
β_{gk}	0.4829	0.3976
β_{tgk}	0.4625	0.4625

Tabel 8.1: Dette er en opsummering af de estimerede koefficienter til grundmodellen.

Ud fra den ovenstående analyses ses, at grundmodellen (8.1) kan testes ned til en simplere model givet ved



Figur 8.4: Residualerne til grundmodellen, plottet mod henholdsvis den estimerede middelværdi, og tiden.

$$\text{logit}\,(p_{tg}) = \beta_0 + \beta_t t + \beta_g I_g + \beta_k I_k + \beta_{tk} I_k t. \tag{8.2}$$

Estimationen af koefficienterne er opsummeret i tabel 8.2. (Se appendiks C for en gennemgang af metoderne.)

Koefficient	Estimat	Standard-afvigelse	LR 95% konfidensgrænser	
β_0	-6.8637	0.3488	-7.5504	-6.1829
β_t	0.0061	0.0018	0.0025	0.0098
β_g	-0.7365	0.0366	-0.8086	-0.6651
β_k	1.1871	0.3509	0.5021	1.8778
β_{tk}	-0.0078	0.0019	-0.0115	-0.0040

Tabel 8.2: Opsummering af de estimerede værdierne til koefficienterne til modellen givet ved (8.2).

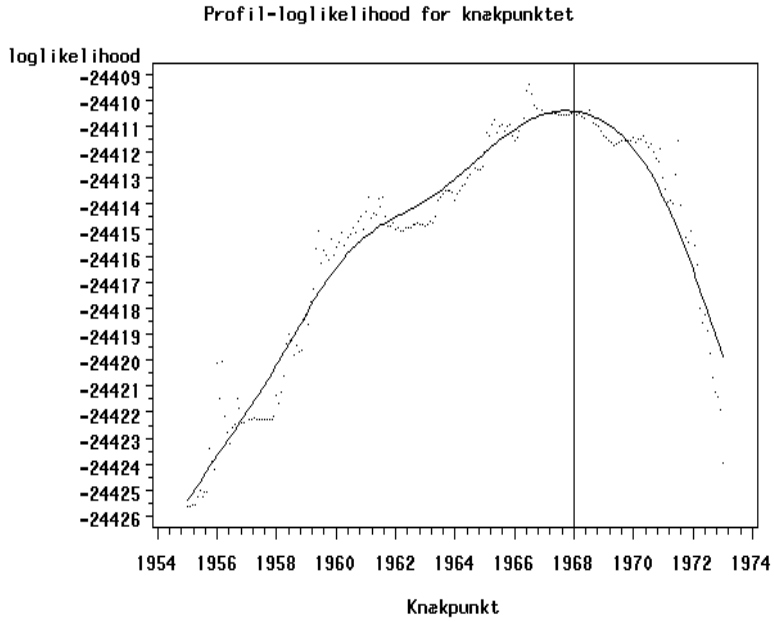
NOTE: De forklarende variable, som testes væk ved den ovenstående analyse, kan også testes væk samlet. Dette er undersøgt ved alle følgende analyser i denne anvendelsesdel, men det nævnes ikke med mindre den samlede test er signifikant. △

Residualplot til modellen givet ved (8.2), har samme struktur som plotene til grundmodellen, hvorved denne model også ser ud til, at kunne forklare variationen i data rimeligt.

8.1.4 Knækpunkt

Det sidste trin i modelopstillingen er at se på placeringen af knækpunktet. For at estimere dette findes log-likelihooden, for den anvendte model (8.2), til alle mulige knækpunkter, hvorved der opnås en profil likelihood for disse. (Se appendiks B for en introduktion til profil likelihood.)

I figur 8.5 er profil likelihooden til hvert knækpunkt plottet sammen med en udglattet kurve af punkterne. Heraf ses, at toppunktet ligger meget tæt på januar 1968, hvilket stemmer særdeles godt overens med overgangen fra ICD8 til ICD10. Det skal bemærkes, at de største værdier for profil likelihooden ligger midt i 1966, men da dette blot er et tjek på knækpunktet, anvendes dette punkt ikke. Hvis der laves tilsvarende kurver for profil likelihooden, for hvert køn hver for sig, opnås et tilsvarende resultat.



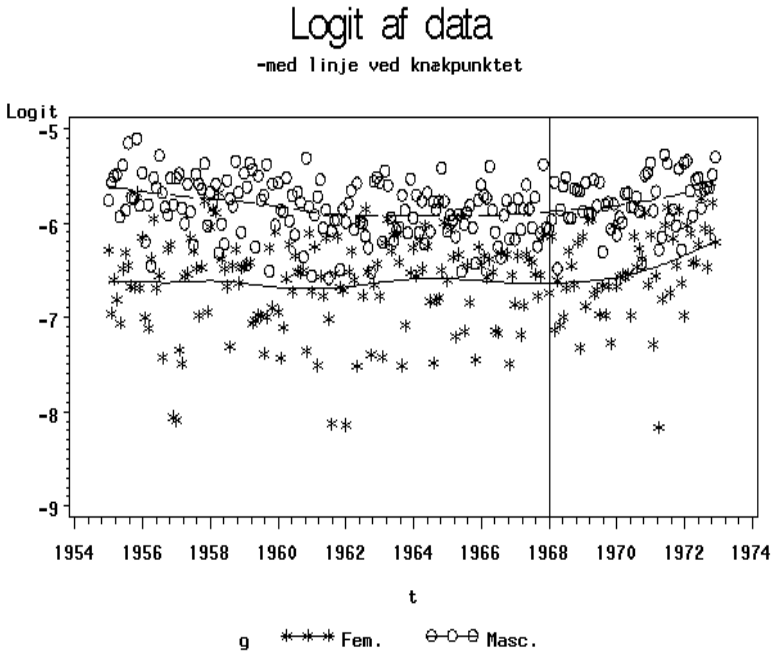
Figur 8.5: Punkterne til profil likelihooden, plottet sammen med en udglattet kurve for disse.

I figur 8.6 plottes skizofrenifrekvenserne, med en referencelinje ved 1968, hvorved det gerne skulle være mere tydeligt, at data begynder at stige efter dette punkt.

Efter placeringen af knækpunktet er undersøgt nærmere, kan modellen omparametriseres. Dette gøres for, at undersøge hvorvidt der er et spring ved januar 1968 (mellem $t = 156$ og $t = 157$), eller om der blot er et knæk i regressionslinjen. Denne omparametrisering laves ved, at indføre de forklarende variable $t1$, hvor værdierne ændre sig før knækpunktet, og $t2$, hvor værdierne ændrer sig efter knækpunktet. Disse er givet ved

$$t1 = \begin{cases} t & \text{for } t < 157 \\ 156 & \text{for } t \geq 157, \end{cases}$$

og $t2 = t - t1$.



Figur 8.6: Skizofrenifrekvenserne, med en referencelinje ved knækpunktet i januar 1968.

Dette giver modellen

$$\text{logit}\,(p_{tg}) = \beta_0 + \beta_{t1} t1 + \beta_{t2} t2 + \beta_g I_g + \beta_k I_k, \tag{8.3}$$

hvor β_k udelukkende giver størrelsen på springet i knækpunktet. En test, svarende til de tidligere tests, af denne model giver, at springet i knækpunktet ikke er signifikant, hvorved den nye anvendte model er givet ved

$$\text{logit}\,(p_{tg}) = \beta_0 + \beta_{t1} t1 + \beta_{t2} t2 + \beta_g I_g. \tag{8.4}$$

Estimationen af koefficienterne er opsummeret i tabel 8.3 (se appendiks C for en gennemgang af metoderne). Modelkontrol for denne model giver et resultat der er helt tilsvarende resultaterne for de tidligere modeller.

Koefficient	Estimat	Standard-afvigelse	LR 95% konfidensgrænser	
β_0	-5.6796	0.0400	-5.7585	-5.6018
β_{t1}	-0.0016	0.0004	-0.0023	-0.0008
β_{t2}	0.0065	0.0012	0.0042	0.0089
β_g	-0.7366	0.0366	-0.8086	-0.6651

Tabel 8.3: Opsummering af de estimerede værdierne til koefficienterne til den nye anvendte model.

Det ses, at resultaterne stemmer godt over ens med resultaterne for model (8.2). Det skal bemærkes, at β_0 for denne model, svarer til $\beta_0 + \beta_k$ fra den tidligere model, β_{t1} for denne model, svarer til $\beta_t + \beta_{tk}$ fra den tidligere model, og β_2 for denne model, svarer til β_t fra den tidligere model. De lavere standardafvigelser på interseptet og koefficienten til tiden, må tilskrives, at modellen er omparametriseret, og testet ned, således, at den ikke længere indeholder I_k, som forklarende variabel.

NOTE: I denne modelopstilling betragtes knækpunktet ikke som en parameter der skal estimeres. Derimod undersøges blot, om der er rimelighed i at det eksisterer, og har den angivne placering. Hvis knækpunktet opfattes, som en parameter der skal estimeres er de angivne standardafvigelser lidt for optimistiske, og dermed for små. △

8.2 Eventuel ekstra temporal variation

Efter at have opstillet en model for skizofrenifrekvensen, og lavet basal inferens, arbejdes der nu med, om der er yderligere variationer over tid. Den første del af denne analyse ser på, om der er nogle systematiske variationer i data, som ikke er forklaret af den anvendte model (8.4). Derefter analyseres for om der er nogle stokastiske variationer i data, som ikke er forklaret af den anvendte model.

Den teori der anvendes til analyserne for ekstra temporal variation, er for en stor del basale metoder til analyse af tidsrækker. Af tidsmæssige årsager er en teoretisk beskrivelse af disse metoder ikke behandlet i dette speciale, men en sådanne er bla. givet i Diggle [1990].

8.2.1 Systematisk variation

Denne analyse ser på, om der er nogle systematiske variationer i data, som ikke er forklaret af den anvendte model (8.4).

Analyse ved periodogrammer

For at undersøge om der er nogle faste svingninger i data, laves først periodogrammer for residualerne til den anvendte model (8.4).

Princippet i periodogrammer er, at undersøge om der er nogle systematiske variationer i data, på formen

$$\alpha \cos(\omega t) + \beta \sin(\omega t),$$

hvor $\omega \in [0; 2\pi]$ og $t = 1, \ldots, 216$.
For at gøre dette, indføres funktionen

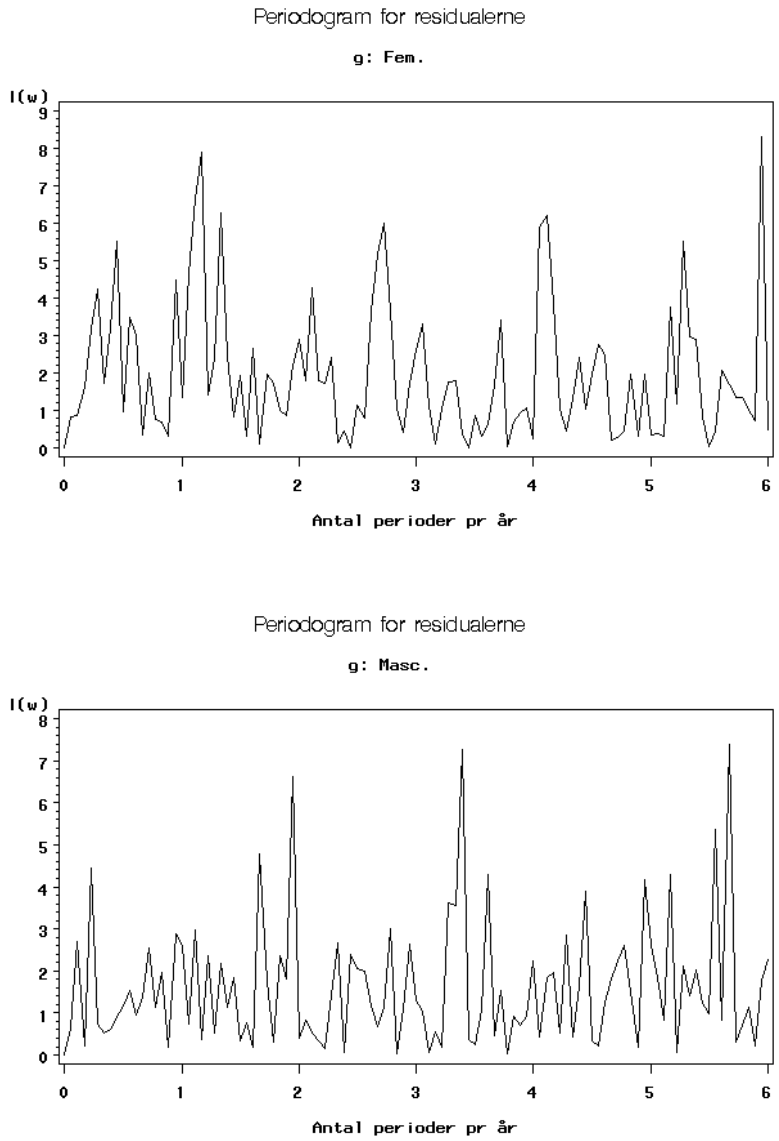
$$I(\omega) = \frac{1}{216} \left[\left\{ \sum_{t=1}^{216} r_t \cos(\omega t) \right\}^2 + \left\{ \sum_{t=1}^{216} r_t \sin(\omega t) \right\}^2 \right],$$

hvor ω og t er givet som ovenfor, og $r_1, \ldots, r_{216}$ er residualerne, som ønskes undersøgt for harmoniske svingninger.
Som det fremgår af Diggle [1990, 47-53], angiver store værdier af $I(\omega)$, at harmoniske svingninger med de givne frekvenser ω, stemmer godt over ens med variationen i data.

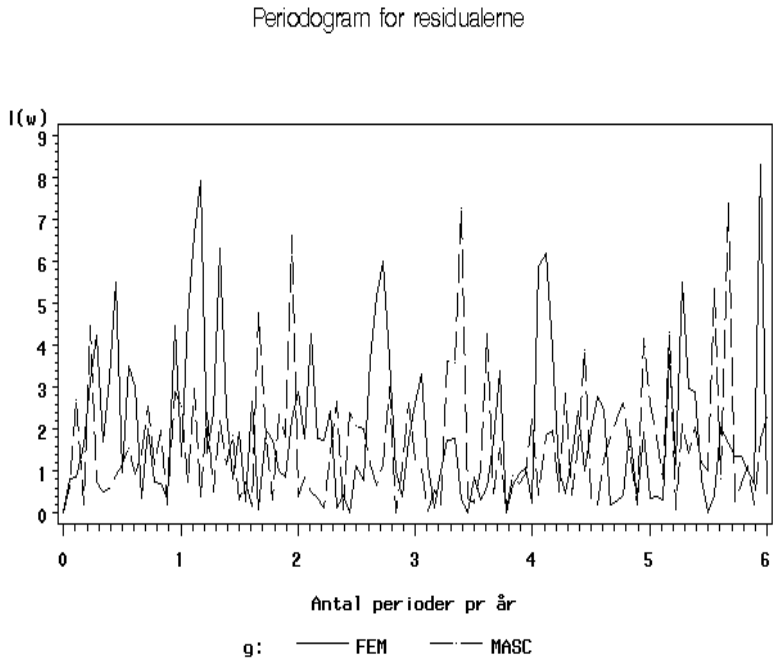
Da der tidligere er påvist en forskel i skizofrenifrekvensen for kvinder og mænd, opdeles residualerne således, at de to køn undersøges hver for sig. Periodogrammerne for henholdsvis kvinder og mænd er plottet i figur 8.7.

Der er tilsyneladende ikke nogen entydige frekvenser, som forklarer variationen i residualerne godt, men der er dog nogle toppe for begge køn. I figur 8.8 er periodogrammerne for hver af kønnene plottet sammen, hvorved man ser, at de ikke topper samtidigt. Dette underbyger den anvendte opdeling i køn. En yderligere opdeling, således at residualerne for hvert køn, også opdeles i en gruppe før knækpunket og en gruppe efter knækpunktet, kunne endvidere anvendes. Denne opdeling har dog den ulempe, at antallet af værdier efter knækpunktet ikke er særlig stort. En analyse med denne opdeling, har da også givet et resultat, der svarer fuldstændig til resultatet, for en opdeling kun med hensyn til kønnene. Derfor vil denne gennemgang af analysen kun anvende opdelingen i forhold til køn.

Da det ikke er klart ud fra periodogrammerne, om der ligger nogle harmoniske svingninger i residualerne, testes der herefter for sådanne. Metoden er, at udvide den anvendte model (8.4) med nogle sinus og cosinus led, der er



Figur 8.7: Periodogram for residualerne, for henholdsvis kvinder (øverst), og mænd (nederst).



Figur 8.8: Periodogram for residualerne til de to køn, plottet sammen.

på samme form, som dem der anvendes til at lave periodogrammet. Herved kan der testes om en sum af sinus- og cosinus-funktioner, med udvalgte frekvenser, kan forklare en del af variationen i data. Dette svarer til, at teste om visse frekvenser fra periodogrammet er signifikante.

Den model der testes bliver herved på formen

$$\begin{aligned}\operatorname{logit}(p_{tg}) = \beta_0 &+ \beta_{t1} t1 + \beta_{t2} t2 + \beta_g I_g \\ &+ \alpha_{k1}\cos(\omega_{k1} t) + \alpha_{k2}\cos(\omega_{k2} t) + \cdots \\ &+ \beta_{k1}\sin(\omega_{k1} t) + \beta_{k2}\sin(\omega_{k2} t) + \cdots ,\end{aligned} \tag{8.5}$$

hvor $\omega_{ki} = 2\pi ki/216$, og ki er antallet af svingninger i hele datasættet. Dvs. $ki \in \mathbb{Z}_+$, og $ki \leq 108$.

Testen udføres ved, at der først laves konfidensintervaller, og beregnes testværdier, for hvert led hver for sig. Dernæst anvendes den χ^2-test, som

er beskrevet i afsnit 3.2.2, til at teste den tidligere anvendte model (8.4), imod model (8.5) der desuden indeholder alle de udvalgte frekvenser.
En sådanne test af henholdsvis $1/4, 1/2, 1$ og 2 svingninger pr. år, giver, at ingen af leddene i sig selv er signifikante. Desuden giver testen af den tidligere anvendte model (8.4), imod modellen på formen (8.5), indeholdene $1/4, 1/2, 1$ og 2 svingninger pr. år, en P-værdi på 0.75. Dvs. der heller ikke er signifikans for alle frekvenserne samlet. Ligeledes opnås ikke signifikans for en mængde andre udvalgte frekvenser, eller serier af frekvenser, som er testet på en tilsvarende måde.
Hvis der yderligere tilføjes forklarende variable bestående af vekselvirkninger mellem kønnet, og sinus/cosinus leddene, opnås samme resultat.

Efterfølgende analyseres for, om fødselsmåneden m, kan forklare noget af variationen i data, som en ustruktureret sæsonvariation. Dette gøres ved, at tilføje den forklarende variabel m, samt vekselvirkningen mellem m og g i den anvendte model. Herved opnås modellen

$$\text{logit}\,(p_{tg}) = \beta_0 + \beta_{t1} t1 + \beta_{t2} t2 + \beta_g I_g + \beta_m m + \beta_{gm} I_g m. \tag{8.6}$$

En type 1 test af denne model giver en P-værdi på 0.88, for at β_{gm} er 0, og en P-værdi på 0.34, for at β_m er 0. Dvs. der ikke er signifikans for, at fødselsmåneden har nogen indvirkning på variationen i data.

Konklusionen på de ovenstående analyser er, at ingen af de undersøgte tilfælde påviser nogen yderligere systematiske variationer.

8.2.2 Stokastisk variation

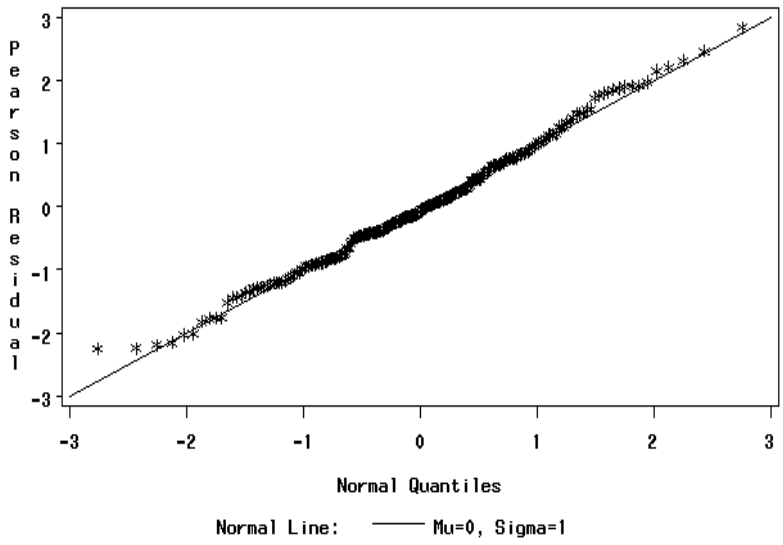
Den anden del af analysen for systematiske variationer, som ikke er forklaret af den anvendte model (8.4), omfatter analyse for stokastiske variationer.

Korrelogram

Den første del af denne analyse omfatter korrelogrammer for residualerne til den anvendte model (8.4). Da det tidligere er påvist, at der er en forskel mellem kønnene, anvendes opdeling efter køn ligeledes i dette afsnit. Fortolkningen af korrelogrammet baserer sig på, at residualerne er approksimativt standard normalfordelt. Derfor undersøges dette først.

Undersøgelsen udføres ved, at lave et normal-fraktilplot af residualerne. I figur 8.9 ses normal-fraktilplottet for kvinder. Normal-fraktilplottet for mænd ser tilsvarende ud.

Heraf ses, at residualerne tilsyneladende er approksimativt standard normalfordelt. Som det nævnes tidligere, underbygger dette valget af Pearson-residualerne til modelkontrol.

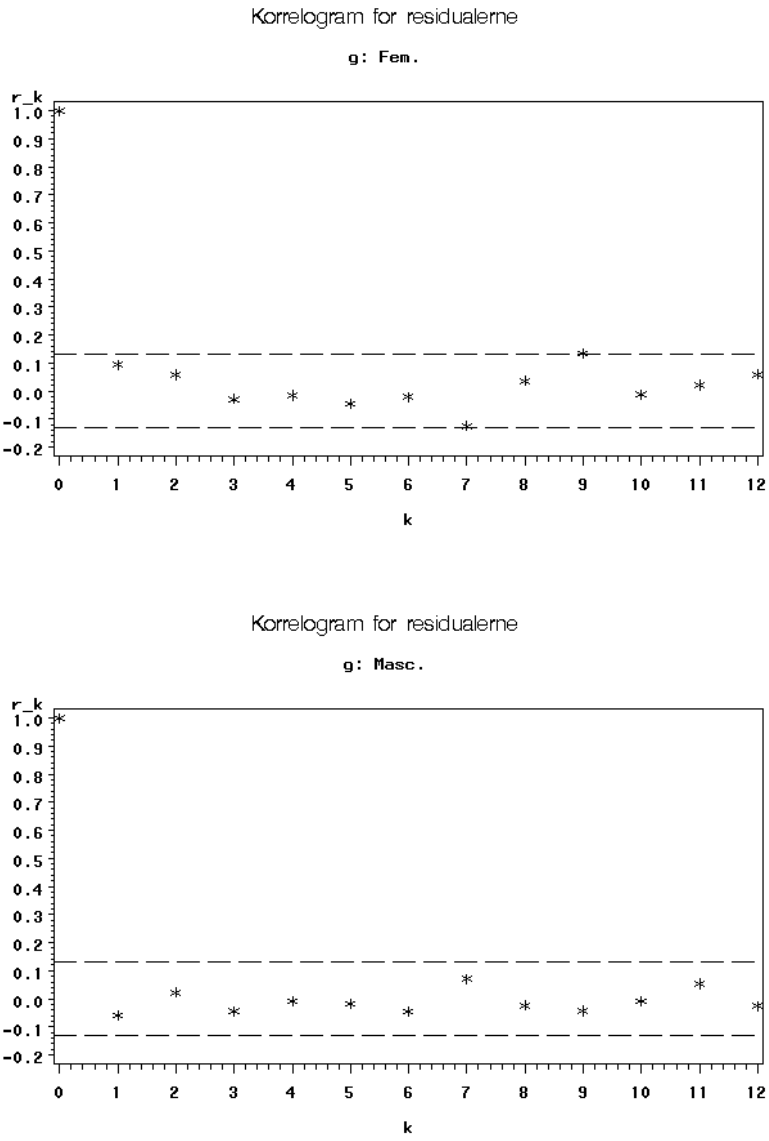


Figur 8.9: Normal-fraktilplot af residualerne for kvinder, med en referencelinje for middelværdien (Mu) lig nul, og en standardafvigelsen (Sigma) lig en.

Da fraktilplottet for residualerne ser pænt ud, udregnes en empiriske autokorrelation ρ_k, for henholdsvis kvinder og mænd, og der laves korrelogrammer for disse.

Der er lavet korrelogrammer for henholdsvis kvinder og mænd i figur 8.10. Da der gælder, at $\rho_k \dot{\sim} \mathcal{N}(0, 1/n)$, hvor n er antallet af observationer, kan $\pm\sqrt{2/216}$ anvendes, som approksimative 95% konfidensgrænser [Diggle, 1990, s. 39]. Disse signifikansgrænser er markeret på korrelogrammerne som stiplede linjer.

Af korrelogrammerne ses, at de empiriske autokorrelationskoefficienter holder sig inden for grænserne for mænd, men for kvinder er der nogle ekstreme værdier i $k = 7$ og $k = 9$, der ligger lige på grænsen. Desuden er ρ_1 og til



Figur 8.10: Korrelogram for kvinder (øverst) og mænd (nederst). Autokorrelationskoefficienten i måneder er givet ved k. De stiplede linjer angiver $\pm\sqrt{1/108}$.

dels også ρ_2 høje for kvinder, hvilket tyder på at der måske kan være en form for korrelation i tid. Dette er dog kun en meget svag indikation, som kun kan anvendes som et fingerpeg i videre undersøgleser.

Det skal bemærkes, at hvis analyser køres for begge køn under ét, opnås et tilsvarende resultat. I dette tilfælde er værdierne i korrelogrammet dog forholdsvis tættere på nul, da forskellene i autokorrelationsstrukturen mellem de to køn, til sammen giver en mere flad struktur.

Residualplots

En yderligere visuel undersøgelse af, om der er positiv seriel korrelation, opnås ved at lave scatterplots af residualerne til tiden t, mod residualerne til tiden $t-k$, hvor $k = 1, \ldots, 12$. Disse scatterplots er at finde i figur 8.11.

Heraf ses, at der ikke umiddelbart ser ud til, at være nogen tydelige mønstre. Det kan dog se ud som om, der er et ovalt mønster i plottet af residualerne, for kvinderne, til tiden t mod residualerne til tiden $t-1$, og $t-9$. Dette stemmer godt over ens med iagttagelserne fra korrelogrammerne.
Residualplottene for mænd har ikke et nær så tydeligt mønster, men ellers ligner de plottene for kvinder.

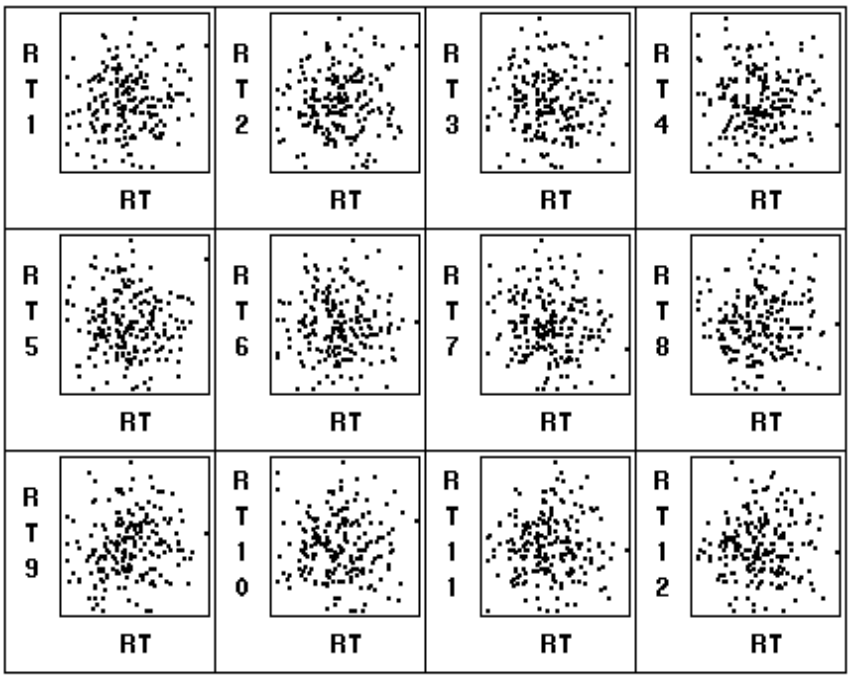
Test for hvis støj

Den sidste del af denne undersøgelse går på, om residualerne kan opfattes som værende hvid støj. Til dette formål anvendes henholdsvis Fisher's eksakte test og Bartlett's Kolmogorov-Smirnov test. Disse tests er beskrevet i Diggle [1990, s. 98-101]. Da P-værdierne, givet i tabel 8.4 er store, afviser dette ikke, at residualerne er hvid støj.

	Kvinder	Mænd
Fisher's eksakte test	0.86	0.70
Bartlett's Kolmogorov-Smirnov test	0.22	0.76

Tabel 8.4: P-værdier til tests for hvid støj.

Det skal bemærkes, at hvis den analyse der her er kørt for kønnene hver for sig, køres for begge køn under ét, eller for en opdeling med hensyn til både køn og periode, opnås et tilsvarende resultat.



Figur 8.11: Scatterplot af residualerne til tiden t mod residualerne til tiden $t-k$, hvor $k = 1, \ldots, 12$, for kvinder.

8.3 Diskussion

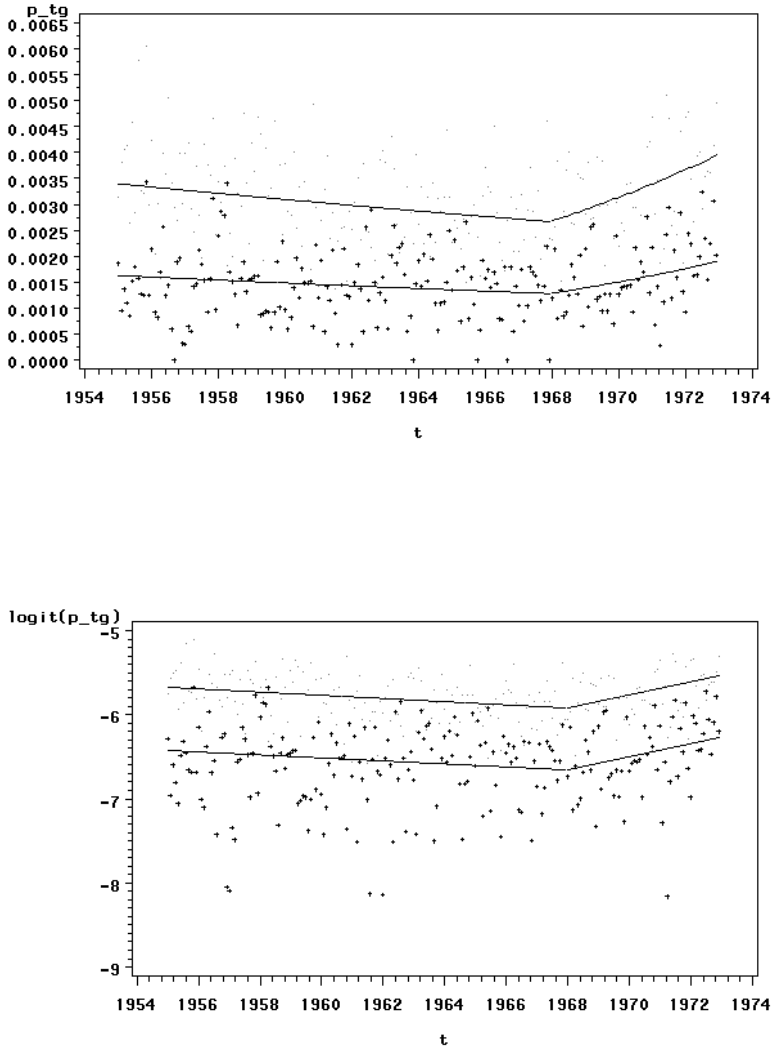
Ud fra de ovenstående analyser må konkluderes, at den anvendte model (8.4) forklarer variationen i dataene rimeligt. Der er, på basis af analyserne, ingen indikation af, at der er yderligere systematiske variationer i data. Derimod indikerer undersøgelsen for stokastiske variationer, at der måske kan være visse stokastiske strukturer, som ikke er forklaret ved den anvendte model. Dette er dog ikke oplagt.
Den anvendte model denne analyse ender op med, er således givet ved

$$\text{logit}\,(p_{tg}) = \beta_0 + \beta_{t1} t1 + \beta_{t2} t2 + \beta_g I_g, \tag{8.7}$$

hvor estimationen af koefficienterne er opsummeret i tabel 8.3.

Hvis de estimerede værdier for koefficienterne indsættes i (8.7), opnås den modellerede skizofrenisandsynlighed p_{tg}. Dette er illustreret i figur 8.12, hvor p_{tg} er givet i henholdsvis rå og logistisk skala. Det ses, at p_{tg} udgøres af fire rette linjer i logistisk skala. Dette skyldes, at estimatet er opnået ved lineær regression med logit link. I rå skala ses, at der ikke længere er tale om rette linjer, men formen er meget lig formen i logistisk skala. Dette illustrerer, at der ikke er en lineær sammenhæng mellem de to skala.

I figuren er den estimerede skizofrenisandsynlighed for mænd de øverste linjer, og den estimerede skizofrenisandsynlighed for kvinder, er de nederste. Værdien til januar 1955 for mænd, er givet ved β_0. Forskellen mellem kønnene er givet ved β_g. Hældningen af linjerne før knækpunktet, er givet ved β_{t1}, og hældningen af linjerne efter knækpunktet, er givet ved β_{t2}. Det ses endvidere, at skizofrenisandsynligheden er svagt faldene frem til knækpunktet, hvorefter den begynder at stige. Dette indikerer den stigning der kan være sket med overgangen fra ICD8 til ICD10 diagnostiseringssystemet. Samtidig indikeres, at der har været et fald i skizofrenifrekvensen frem til dette tidspunkt. Ud fra verserede hypoteser på Center for Registerforskning, kan dette fald i antallet af diagnosticerede skizofrene meget vel skyldes en nedskæring på det psykiatriske område op gennem tiden. Herved er der blevet færre sengepladser til rådighed for skizofrene patienter, og dermed diagnosticeret færre. Noget sådanne er der dog ikke basis for at konkludere ud fra disse analyser.



Figur 8.12: Den modellerede skizofrenisandsynlighed p_{tg} i henholdsvis rå skala (øverst) og logistisk skala (nederst). Data er plottet i de respektive skala, med kvinder markeret ved sorte plusser, og mænd markeret ved grå punkter.

9 Dataanalyse baseret på den DGLM

I dette kapitel gennemgås resultaterne for en dataanalyse baseret på den DGLM. Denne analyse gennemgåes dels for, at undersøge data nærmere, og dels for at illustrere anvendelsen af den beskrevne teori for den DGLM. Til estimation af ukendte hyperparametre anvendes GLIMMIX makroen i SAS.

De beskrevne metoder fra kapitel 5 anvendes ikke, da der af tidsmæssige årsager ikke har være muligt, at nå implementeringen af disse algoritmer i SAS. Jvf. [Durbin and Koopman, 2001, s. 202-203] er der dog meget lille forskel på, at anvende det udvidede Kalmanfilter, med iterationsmetoden beskrevet i afsnit 4.3.1, og importance sampling til, at estimere middelværdien af den latente proces i den DGLM med Gaussisk systemligning.

Den teoretiske baggrund, for dataanalysen, er beskrevet i kapitel 4, og gennemgået specielt for binomialfordelingen i kapitel 6.

9.1 Modelopstilling

I den model der opnås ved dataanalysen i foregående kapitel, bygger testresultaterne på, at modellen er statisk, og data er ukorrelerede. Da den model der ønskes anvendt i dette kapitel ikke anvender disse krav, tages der udgangspunkt i grundmodellen (8.1) i stedet for den anvendte model (8.7). Dog omparametriseres, så de forklarende variable bliver på samme form som (8.7). Den grundmodel der anvendes i denne analyse bliver herved, at antallet af skizofrenitilfælde s_{tg} beskrives ved den stokastiske variabel Z_{tg}, hvor

$$(Z_{tg}|\boldsymbol{\theta}_{tg}) \sim \text{Bi}(n_{tg}, p_{tg}),$$

med

$$\begin{aligned}\text{logit}\,(p_{tg}) &= \lambda_{tg} = \boldsymbol{F}_{tg}^{\text{T}}\boldsymbol{\theta}_{tg} \\ &= \beta_0 + \beta_{t1}t1 + \beta_{t2}t2 + \beta_g I_g + \beta_k I_k + \beta_{t1g} I_g t1 \\ &\quad + \beta_{t2g} I_g t2 + \beta_{gk} I_g I_k + \varepsilon(t,g),\end{aligned}$$

hvor $\boldsymbol{\beta} = [\beta_0, \beta_{t1}, \beta_{t2}, \beta_g, \beta_k, \beta_{t1g}, \beta_{t2g}, \beta_{gk}]^{\text{T}}$ er regressionsparametrene i modellen, $\varepsilon(t,g)$ er en dynamisk parameter, som afhænger af tiden og kønnet, og $t1$, $t2$ og I_g er de forklarende variabler defineret i afsnit 8.1.

Det ses, at $\boldsymbol{F}_{tg}^{\text{T}} = [1, t1, t2, I_g, I_k, I_g t1, I_g t2, I_g I_k, 1]$, og $\boldsymbol{\theta}_{tg} = [\boldsymbol{\beta}^{\text{T}}, \varepsilon(t,g)]^{\text{T}}$.

Da der ikke er nogen forhåndsviden om fordelingen af $\varepsilon(t,g)$, antages denne at være Gaussisk. Her ud fra opstilles en DGLM med Gaussisk systemligning, for de anvendte skizofrenidata. Strukturen af den latente proces bestemmes dels ved, at se på analysen for ekstra temporal stokastisk variation i afsnit 8.2.2, og dels ud fra hvilken struktur der kan forventes, at være i data. Ud fra figur 8.10 kan det se ud som om, der er en positiv korrelation mellem en given måned, og den forgående måned. Dette stemmer godt over ens med strukturen for mulige risikofaktorer, fra afsnit 7.1.2, som f.eks. influenza under graviditeten. Af denne grund anvendes en stationær AR(1) struktur til, at beskrive $\varepsilon(t,g)$.

Korrelogrammet for mænd ligner, som tidligere diskuteret, udpræget hvid støj, hvorfor der også anvendes to forskellige AR(1) modeller for kvinder og mænd.

Da en model med negativ korrelation ville afhænge af hvilken tidsopdeling der anvendes, betragtes kun modeller med positiv korrelation. En stationær AR(1) model med positiv korrelation kan for et givet køn g specificeres ved, at

$$\varepsilon(t,g) = \rho_g \varepsilon(t-1,g) + \omega_{tg}, \quad \omega_{tg} \sim \mathcal{N}(0, v_g),$$

for $t = 1, \ldots, 216$. Der gælder endvidere, at $\omega_{1g}, \ldots, \omega_{216g}$ er uafhængige, og $0 \leq \rho_g < 1$, da der skal være positiv korrelation.

Da processen er stationær, er variansen af $\varepsilon(t,g)$ givet ved en konstant σ_g^2, for alle $t = 1, \ldots, 216$. Dvs. v_g er givet ved

$$v_g = (1 - \rho_g^2)\sigma_g^2.$$

Jvf. Fahrmeir and Tutz [2001, s. 332-337 og s.345-346] kan den DGLM med Gaussisk systemligning, lige så vel anvendes i det tilfælde, hvor covariansmatricen $\boldsymbol{W}_{tg}$, fra definition 4.2, er singulær. Da regressionsparametrene ikke er dynamiske, vil denne model have en sådanne form.

Den anvendte model for et givet køn g, kan herved opskrives ud fra definition 4.2 af den DGLM med Gaussisk systemligning, som følger.

Model:

Observationsligning

Tæthedsfunktionen for skizofrenifrekvensen er givet ved

$$f\left(y_{tg}|\eta_{tg}, V_{tg}\right) = \exp\Big[n_{tg}\big(y_{tg}\underbrace{\operatorname{logit}\left(p_{tg}\right)}_{\eta_{tg}} - \underbrace{\ln\left(1+\exp\left(\eta_{tg}\right)\right)}_{b(\eta_{tg})}\big) + \underbrace{\ln\left(\binom{n_{tg}}{y_{tg}}\right)}_{c(y_{tg},V_{tg})}\Big], \tag{9.1}$$

hvor $\varphi_{tg} = 1/a(V_{tg}) = n_{tg}$, og $y_{tg} = z_{tg}/n_{tg}$, for $t = 1, \ldots, 216$.

Der eksisterer endvidere et **signal** λ_{tg}, som er givet ved

$$\operatorname{logit}\left(p_{tg}\right) = \lambda_{tg} = \boldsymbol{F}_{tg}^{\mathrm{T}}\boldsymbol{\theta}_{tg}, \tag{9.2}$$

for $t = 1, \ldots, 216$.

Systemligning

Den latente proces er givet ved systemligningen

$$\boldsymbol{\theta}_{tg} = \boldsymbol{G}_{tg}\boldsymbol{\theta}_{(t-1)g} + \begin{bmatrix}\mathbf{0}\\ \omega_{tg}\end{bmatrix}, \quad \boldsymbol{\omega}_{tg} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{W}_g),$$

hvor

$$\boldsymbol{G}_{tg} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & \rho_g \end{bmatrix} \quad \text{og} \quad \boldsymbol{W}_g = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & v_g\end{bmatrix},$$

med $v_g = (1 - \rho_g^2)\sigma_g^2$.

Begyndelsesbetingelse

Begyndelsesbetingelsen for den latente proces, er givet ved

$$(\boldsymbol{\theta}_{0g}|D_{0g}) \sim \mathcal{N}_n(\boldsymbol{m}_{0g}, \boldsymbol{C}_{0g}),$$

hvor

$$\boldsymbol{\theta}_{0g} = \begin{bmatrix} \boldsymbol{\beta} \\ \varepsilon(0,g) \end{bmatrix},$$

og begyndelsesinformationen D_{0g} er givet ved hyperparametrene $\boldsymbol{F}_{tg}$, $\boldsymbol{G}_{tg}$, $\boldsymbol{m}_{0g}$, $\boldsymbol{C}_{0g}$ og $\boldsymbol{W}_{tg}$, med

$$\boldsymbol{m}_{0g} = \begin{bmatrix} \boldsymbol{\beta}_{begynd} \\ 0 \end{bmatrix} \quad \text{og} \quad \boldsymbol{C}_{0g} = \begin{bmatrix} cc_g & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & cc_g & 0 \\ 0 & \cdots & 0 & \sigma_g^2 \end{bmatrix},$$

hvor $\boldsymbol{\beta}_{begynd}$ er begyndelsesværdierne for regressionsparametrene, og cc_g er den tilhørende varians.

Der gælder endvidere, at $(Y_{tg}|\boldsymbol{\theta}_{tg})$ er uafhængig af $Y_1, \ldots, Y_{t-1}, Y_{t+1}, \ldots, Y_{216}$, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_{216}$, at Y_{tg} kun afhænger af $\boldsymbol{\theta}_{tg}$ gennem λ_{tg}, samt at $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_{216}, (\boldsymbol{\theta}_0|D_0)$ er uafhængige. Δ

Det kan bemærkes, at begyndelsesvariansen σ_g^2, for $\varepsilon(0,g)$ angiver, at der startes et tilfældigt sted i en proces med konstant varians σ_g^2.

9.2 Estimation

For at kunne køre det udvidede Kalmanfilter, skal de ukendte hyperparametre først estimeres. I dette afsnit estimeres derfor først de ukendte hyperparametre, og derefter den latente proces.

9.2.1 Estimation af ukendte hyperparametre

Ud fra det ovenstående modelopsæt ses, at de ukendte hyperparametre udgøres af σ_g og ρ_g. Da dataanalysen er blevet udført i SAS, er GLIMMIX makroen i dette program anvendt til, at lave maksimum likelihood estimation af disse ukendte hyperparametre. Detaljer om algoritmen fremgår ikke klart af manualen, hvorfor der er prioriteret ikke at gå i detaljer med det.

Estimationen giver, at $\sigma_g^2 = 0$ for mænd, hvilket vil sige, at der ikke er en latent proces, af den givne form, for mænd. For kvinder er $\sigma_g^2 = 0.01442$, og $\rho_g = 0.6778$.

Da der, jvf. estimationen, ikke er nogen latent proces, af den angivne form, for mændene, er den følgende analyse udelukkende for kvinderne. Grundmodellen kun for kvinderne bliver herved

$$\text{logit}\,(p_t) = \lambda_t = \boldsymbol{F}_t^{\mathrm{T}}\boldsymbol{\theta}_t = \beta_0 + \beta_{t1}t1 + \beta_{t2}t2 + \beta_k I_k + \varepsilon(t),$$

hvor det er implicit, at data kun omfatter kvinder. Der anvendes, i den følgende analyse af data for kvinder, ikke et indeks for køn på nogen af størrelserne.

9.2.2 Estimation af $\boldsymbol{\theta}_t$

Efter de ukendte hyperparametre er estimeret, kan tilstandsvektoren $\boldsymbol{\theta}_t$ estimeres for kvinderne. Dette gøres ved hjælp af det udvidede Kalmanfilter, med iterationsmetoden beskrevet i afsnit 4.3.1. Som stopkriterier anvendes, at

$$\left| \frac{\widetilde{m}_{tj}^{(i)} - \widetilde{m}_{tj}^{(i-1)}}{\widetilde{m}_{tj}^{(i-1)}} \right| < 10^{-6},$$

for alle $j = 1, \ldots, n$ og $t = 1, \ldots, 216$, hvor $\widetilde{\boldsymbol{m}}_t$ er middelværdiestimatet for tilstandsvektoren, og n er antallet af elementer i $\boldsymbol{\theta}_t$. Indekset i angiver hvilken gang iterationen kører. Hvis $\widetilde{m}_{tj}^{(i-1)} = 0$ anvendes kriteriet

$$\left| \widetilde{m}_{tj}^{(i)} - \widetilde{m}_{tj}^{(i-1)} \right|.$$

I de anvendte tilfælde konvergerer algoritmen ved under 8 iterationer, hvilket stemmer meget godt over ens med Durbin and Koopman [2000], hvor der angives, at der normalt ikke skal anvendes mere end 10 iterationer for, at opnå konvergens.

I dette kapitel betegnes det ovennævnte filter blot ved det udvidede Kalmanfilter.

For, at estimere værdierne køres først det udvidede Kalmanfilter på data. Som startværdier for regressionsparametrene anvendes estimaterne, givet ved den generelle model fra GLM-analysen, kørt kun for kvinderne. Herved opnås begyndelsesværdierne givet i tabel 9.1. Bemærk at denne analyse ikke giver helt de samme resultater som analysen fra foregående kapitel, da analysen derfra omfatter både kvinder og mænd, mens denne analyse kun omfatter kvinderne.

Parameter	Begyndelsesværdi
$\boldsymbol{\beta}_0$	-6.6015
$\boldsymbol{\beta}_{t1}$	-0.0006
$\boldsymbol{\beta}_{t2}$	0.0091
$\boldsymbol{\beta}_k$	0.0929

Tabel 9.1: Begyndelsesværdier givet ved de estimerede regressionsparametre fra GLM-analysen.

Da der ikke er nogen grund til, at være meget usikker på disse begyndelsesværdier, anvendes begyndelsesvariansen $cc = 1$.

Det udvidede Kalmanfilter er kodet op ved hjælp af IML-pakken i SAS. Ved at køre dette opnås det udglattede middelværdiestimat $\widetilde{\boldsymbol{m}}_t$, og det udglattede covariansmatrixestimat $\widetilde{\boldsymbol{C}}_t$, for tilstandsvektoren $\boldsymbol{\theta}_t$. Her ud fra laves en Wald test (beskrevet i afsnit 3.2.2) af, om regressionsparametrene er lig nul. Desuden opsættes approksimative konfidensgrænser, ved

$$\hat{\beta}_i \pm 1.96\sqrt{\widetilde{C}_{ii}},$$

hvor $\hat{\boldsymbol{\beta}}$ er de estimerede regressionsparametre fra $\widetilde{\boldsymbol{m}}_1$, og $\widetilde{C}_{ii}$ er det tilhørende variansestimat. Denne metode svarer til de anvendte metoder i West and Harrison [1997, kap. 10 og kap. 14]. Det skal bemærkes, at de udglattede estimater for regressionsparametrene er stort set konstante over tid, hvorved størrelsen af estimatet ikke afhænger af hvilket tidspunkt det er taget til.

Ud fra det ovenstående opnås en P-værdi for β_k på 0.61, hvorved denne testes væk. Herefter testes modellen

$$\text{logit}\,(p_t) = \beta_0 + \beta_{t1}t1 + \beta_{t2}t2 + \varepsilon(t).$$

Dette giver en P-værdi for β_{t1} på 0.29, hvorved denne testes væk. Herefter testes modellen

$$\text{logit}\,(p_t) = \beta_0 + \beta_{t2}t2 + \varepsilon(t). \tag{9.3}$$

De opnåede estimater er givet i tabel 9.2. Samtidig opnås estimatet for middelværdien af $\varepsilon(t)$, illustreret ved figur 9.1. Sammen med estimatet er der plottet 95% punktvise intervaller, givet ved

$$\widetilde{m}_{t3} \pm 1.96\sqrt{\widetilde{C}_{t33}},$$

hvor m_{t3} er middelværdiestimatet for $\varepsilon(t)$, og $\widetilde{C}_{t33}$ er det tilhørende variansestimatet.

Koefficient	Estimat	Standard-afvigelse	95% interval		P-værdi
β_0	-6.5700	0.0405	-6.6495	-6.4906	$<$0.0001
β_{t2}	0.0060	0.0020	0.0021	0.0100	0.0029

Tabel 9.2: Estimerede værdier opnået ved det udvidede Kalmanfilter. P-værdien er givet ved en type 3 analyse (se appendiks C for en gennemgang af de anvendte betegnelser).

Af de ovenstående analyser, og tabel 9.2, ses at det kun er interseptet og $t2$, som ikke kan testes væk, hvilket er i overensstemmelse med analyser af en tilsvarende GLM-model, uden $\varepsilon(t)$ leddet. Bemærk at denne model ikke er den analyserede model kapitel 8, da modellen der fra også indeholder mændene.

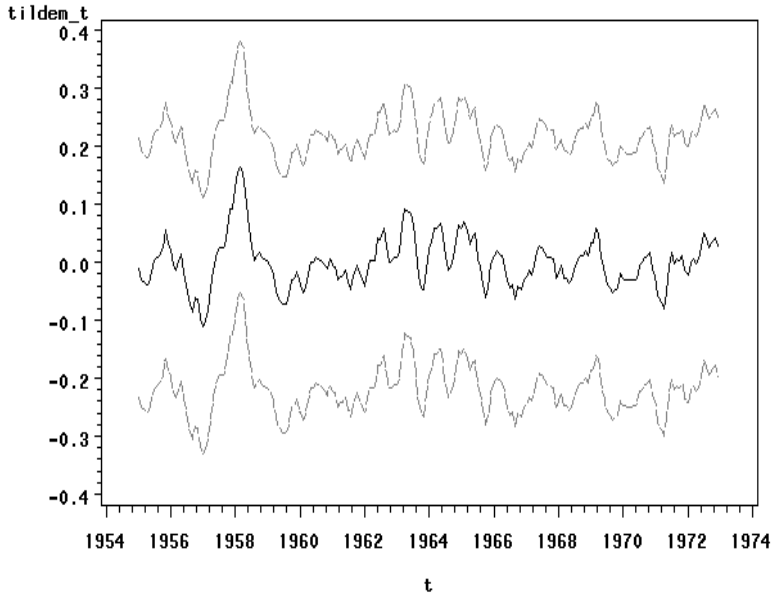
Jvf. Fahrmeir and Tutz [2001, s. 354] opnås ikke bedre estimater ved det iterative udvidede Kalmanfilter, hvis begyndelsesværdierne i en binomial model er nogenlunde nøjagtigt bestemt. Dette er heller ikke tilfældet i denne model, hvor estimaterne dog heller ikke forbedres mærkbart, ved at køre den anvendte iterationsprocedure, beskrevet i afsnit 4.3.1, mere end en gang.

Hvis estimationen udføres med en begyndelsesvarians for regressionsparametrene, på henholdsvis 10 og 100, opnås et resultat som er stort set identisk med den angivne.

9.3 Analysekritik

Som tidligere nævnt er der ikke forsket særlig meget i modelkontrol for dynamiske modeller. Af denne grund, og af tidsmæssige årsager, er der valgt ikke, at gå særlig dybt ind i dette emne. Visse kritiske betragtninger omkring modellens og dermed analyseresultaternes anvendelighed, vil dog være på sin plads.

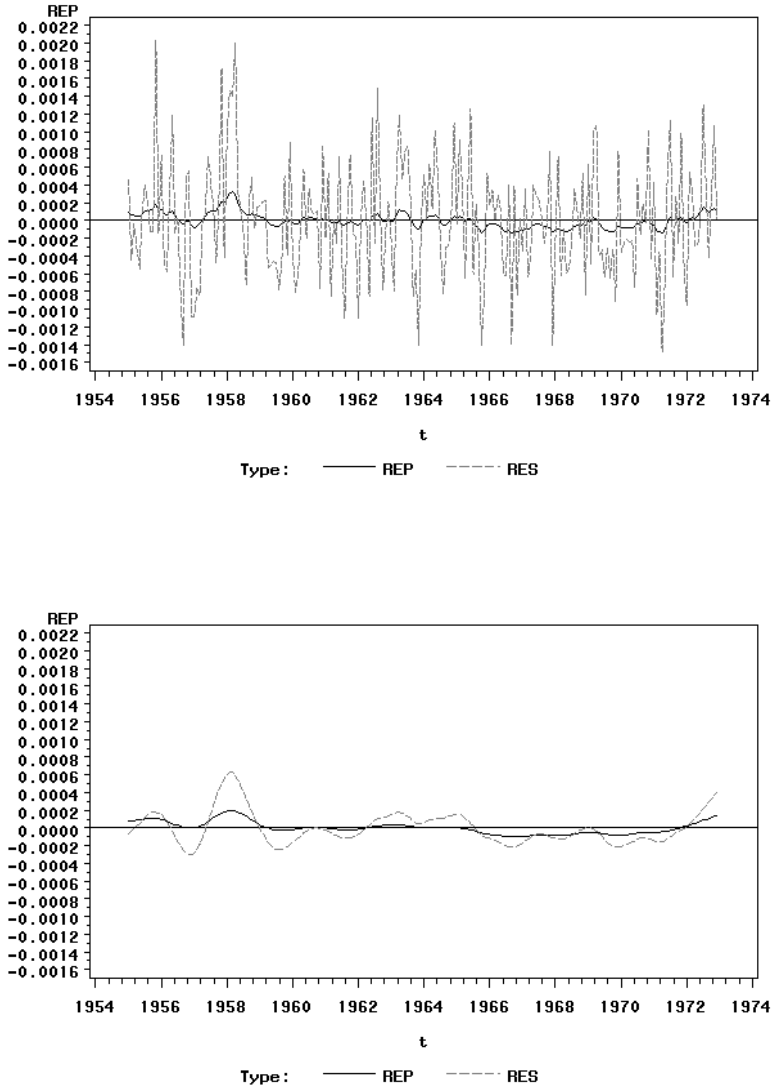
Da de punktvise intervaller for den estimerede middelværdi, i figur 9.1, indeholder nul til enhver tid t, indikeres at processen ikke er signifikant forskellig fra nul. En decideret test af om dette er tilfældet hører dog under de ikke særlig udforskede og ikke behandlede emner, hypotesetest for stokastiske dynamiske parametre, og modelkontrol for dynamiske modeller.
Det er endvidere ikke klart ud fra figuren, om $\varepsilon(t)$ forklarer noget af variationen i data. For at undersøge dette sammenlignes den estimerede middelværdi af $\varepsilon(t)$ med de rå residualer for en tilsvarende GLM-analyse uden $\varepsilon(t)$ leddet. Da logit-funktionen ikke er lineær, er den estimerede middelværdi for $\varepsilon(t)$, og rå-residualerne ikke direkte sammenlignlige. For at kunne



Figur 9.1: Den estimerede middelværdi for $\varepsilon(t)$ (fuldtoptrukken linje) plottet sammen med 95% punktvise intervaller (grå stiplede linjer).

sammenligne størrelserne udregnes derfor forskellen mellem modellen med $\varepsilon(t)$, og modellen uden. Denne størrelse er givet ved $rep = \breve{p}_t - \ddot{p}_t$, for $t = 1, \ldots, 216$, hvor $\ddot{p}_t$ er skizofrenisandsynligheden, estimeret ved modellen uden $\varepsilon(t)$, og $\breve{p}_t$ er skizofrenisandsynligheden, estimeret ved modellen med $\varepsilon(t)$. Sammenligningen mellem rå-residualet og rep er illustreret i figur 9.2. Af den øverste figur ses, at $\varepsilon(t)$ ikke forklarer en særlig stor del af variationen i residualet, men dog en del. Af den nederste figur ses, at formen i middelværdiestimatet for $\varepsilon(t)$, og rå-residualerne følges ad, hvilket underbygger, at $\varepsilon(t)$ forklarer en del af variationen i data. Det kan endvidere bemærkes, at middelværdiestimatet for $\varepsilon(t)$, mere eller mindre, er en udglattet kurve for rå-residualerne.

For at evaluere de estimerede regressionsparametre, sammenlignes resultaterne for de forskellige estimationsmetoder. Den første er GLM-analysen fra foregående kapitel, den anden er det udvidede Kalmanfilter, og det tredje er



Figur 9.2: Effekten af $\varepsilon(t)$ (fuldtoptrukken linje) plottet sammen med det rå residual for modellen uden $\varepsilon(t)$ (stiplet linje). Øverst er værdierne plottet forbundet med linjer, mens en udglattet kurve for dem er plottet nederst.

GLIMMIX makroen i SAS, som også estimerer regressionsparametrene. De tre metoder når alle frem til en slutmodel der ligesom (9.3) kun indeholder regressionsparametrene β_0 og β_{t2}. Værdierne for de tre metoder er samlet i tabel 9.3.

	GLM-analysen		DGLM-analysen		SAS-analysen	
Koefficient	Estimat	Std.afv.	Estimat	Std.afv.	Estimat	Std.afv.
β_0	-6.5701	0.0348	-6.5700	0.0405	-6.5711	0.0405
β_{t2}	0.0060	0.0017	0.0060	0.0020	0.0061	0.0020

Tabel 9.3: Sammenligning af estimaterne, og standardafvigelserne, for regressionsparametrene, ved de tre anvendte metoder.

Af tabel 9.3 ses, at estimaterne for regressionsparametrene er meget tæt på hinanden, for de tre forskellige metoder. Forskellen er ikke større end, at for et hvilket som helst af estimaterne ligger de to andre estimater inden for en standardafvigelse af det givne estimat. Det skal dog bemærkes, at standardafvigelsen på de tre metoder er lidt forskellige. Dette må henvises til forskelle i metoderne.

For at se nærmere på valget af en AR(1) model, anvendes en mere generel model i GLIMMIX. For at indføre denne model er det først nødvendigt, at omskrive den anvendte model til GLIMMIX syntaks. Dette gøres ganske simpelt ved, at samle elementerne til alle tiderne i en vektor. Modellen bliver herved

$$\text{logit}\,(\boldsymbol{p}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{9.4}$$

hvor $\text{logit}\,(\boldsymbol{p}) = [\text{logit}\,(p_1), \ldots, \text{logit}\,(p_{216})]^{\mathrm{T}}$, $\boldsymbol{X}$ er designmatricen, $\boldsymbol{\beta}$ er vektoren af regressionsparametre og $\boldsymbol{\varepsilon} = [\varepsilon(1), \ldots, \varepsilon(216)]^{\mathrm{T}}$. Strukturen af $\varepsilon(t)$ angives ved, at definere strukturen af covariansmatricen for $\boldsymbol{\varepsilon}$. Covariansstrukturen for en AR(1) proces bliver herved

$$\mathbb{V}\text{ar}\,[\boldsymbol{\varepsilon}] = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \\ \rho & \ddots & \ddots & \ddots & \vdots \\ \rho^2 & \ddots & \ddots & \ddots & \rho^2 \\ \vdots & \ddots & \ddots & \ddots & \rho \\ & \cdots & \rho^2 & \rho & 1 \end{bmatrix}.$$

En udvidelse af denne model er en TOEP(k), hvor covariansstrukturen er

$$\mathbb{V}\mathrm{ar}\left[\boldsymbol{\varepsilon}\right] = \begin{bmatrix} \sigma^2 & \sigma_1 & \cdots & \sigma_{k-1} & 0 & \cdots & 0 \\ \sigma_1 & \ddots & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & 0 \\ \sigma_{k-1} & & \ddots & \ddots & \ddots & & \sigma_{k-1} \\ 0 & \ddots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & \ddots & \sigma_1 \\ 0 & \cdots & 0 & \sigma_{k-1} & \cdots & \sigma_1 & \sigma^2 \end{bmatrix}.$$

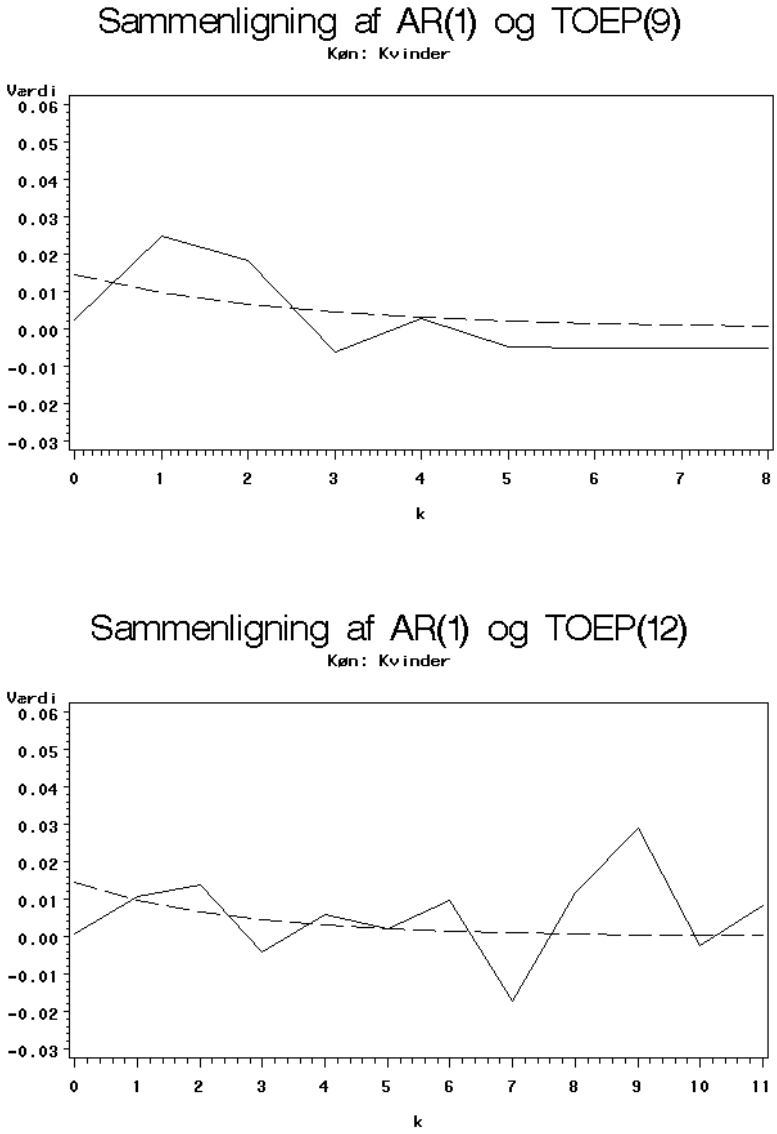
Ud fra disse parametre kan værdierne af σ^2 sammenlignes, og $\sigma^2\rho^j$, fra AR(1) modellen kan sammenlignes med σ_j fra TOEP(k) modellen. Dette gøres for, at undersøge hvorvidt covariansstrukturen af den anvendte AR(1) model er for simpel, i forhold til den mere avancerede TOEP model. Det har ved hjælp af GLIMMIX, været muligt at estimere parametre op til en TOEP(12) model. Sammenligningen mellem modellerne er illustreret ved figur 9.3.
Heraf ses, at TOEP værdierne svinger omkring AR(1) værdierne. Af det øverste plot ses, at værdierne for modellerne starter forskelligt ud, men ellers stemmer nogenlunde godt over ens i et tidsspænd af op til 8 måneder. Af den nederste figur ses, at σ_9 er langt større end forventet i AR(1) modellen. Denne struktur går igen i korrelogrammet for kvinder i figur 8.10, hvor strukturen også tyder på en høj korrelation ved et tidsspænd på 9 måneder.

Det skal bemærkes, at TOEP modellen ikke kan opskrives ved den DGLM-model, som anvendes i dette speciale. Derfor arbejdes der ikke videre med denne model.

NOTE: Hvis der anvendes en tilsvarende TOEP model for mænd, estimeres variansen til at være nul, ligesom ved AR(1) modellen. △

9.4 Diskussion

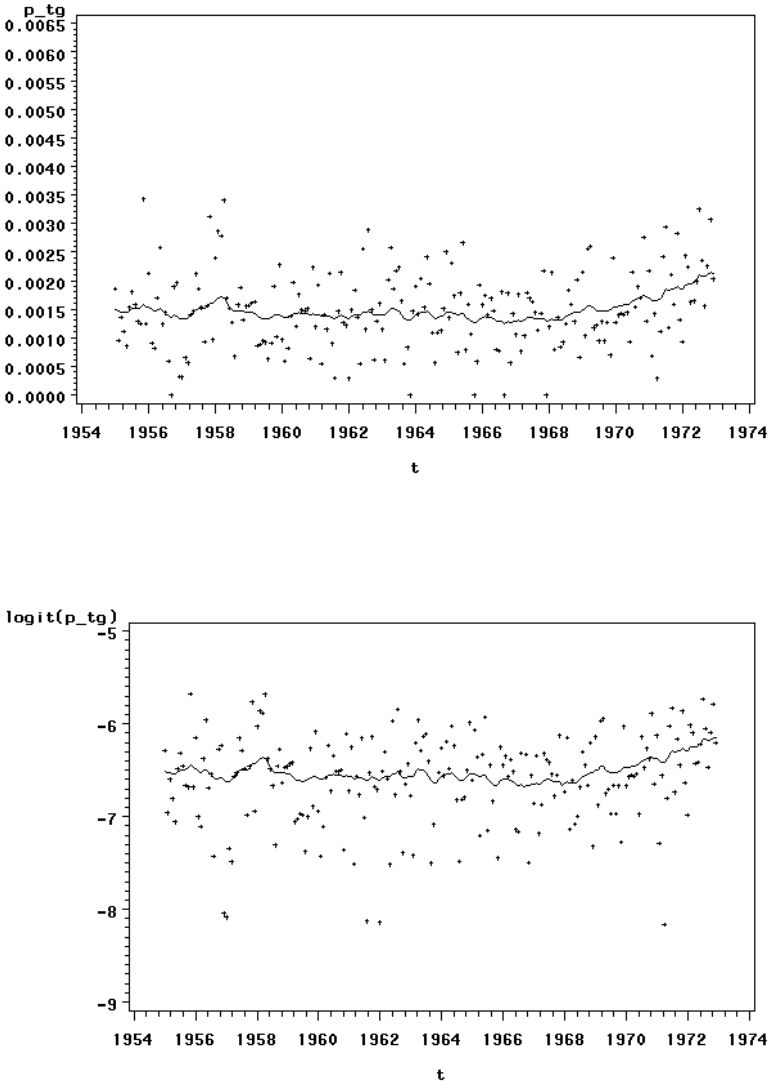
Af figur 9.1 fremgår hvilken form middelværdien $\varepsilon(t)$ er estimeret til at have. Et sådanne plot vil kunne anvendes ved overvejelse af hvilke forklarende variable, der kunne tænkes, at have indvirkning på skizofrenifrekvensen for kvinder. Dvs. hvis en mængde forklarende variable med den angivene struktur blev indkluderet i modellen, ville de kunne forklare effekten af $\varepsilon(t)$-leddet.



Figur 9.3: Kurverne for den estimerede covarians for en afstand på k måneder. Den fuldtoptrukne graf er værdierne for TOEP modellen, og den stiplede er værdierne for AR(1) modellen. Sammenligningen er lavet for henholdsvis en TOEP(9) (øverst), og en TOEP(12) (nederst).

Den model der er opnået ved denne analyse er givet ved (9.3). Den skizofrenifrekvens den modellerer er illustreret ved figur 9.4. Fortolkningen af plottet i logistisk skala, er tilsvarende fortolkningen ved GLM-analysen. Således er værdien til januar 1955, er givet ved β_0. Hældningen af regressionslinjen før knækpunktet, er 0, og hældningen af regressionslinjen efter knækpunktet, er givet ved β_{t2}. Det skal dog bemærkes, at de modellerede niveauer ikke længere udgøres af rette linjer, men, i sagens natur, har udsving bestemt ved den estimerede middelværdi af $\varepsilon(t)$.

Af figur 9.3 ses, at AR(1) modellen for den latente proces ikke kan forkastes, men der kan undemærket tænkes, at være andre modeller, som beskriver den latente proces bedre. En sådanne model behøver ikke nødvendigvis, at være en DGLM, hvorved det ville være nødvendigt, at anvende andre metoder end de, som er beskrevet i dette speciale, for at modellere den latente proces.



Figur 9.4: Den modellerede skizofrenisandsynlighed p_t, for kvinder, i rå skala (øverst), og i logistisk skala (nederst). Data er markeret ved sorte plusser.

10 Opsummering & Diskussion

Den gennemgåede dataanalyse er udført i samarbejde med Center for Registerforskning ved Aarhus universitet. Undersøgelsens formål er at klarlægge, hvorvidt måneden og året en person er født i, har indvirkning på skizofrenifrekvensen. Desuden er det, at undersøge om der er nogle faktorer, som kan forklare antallet af skizofrene født i de givne måneder.

Da specialet har til formål, at evaluere gennemgåede teorier og metoder, er et primært formål med analyserne endvidere, at illustrere teorien i del I.

Analyserne indledes med en analyse baseret på en GLM. Denne analyse når frem til, at en model baseret på et intersept, tiden, kønnet og et knækpunkt giver en rimelig forklaring på variationen i data. Det angivne knækpunkt er bestemt ud fra overgangen fra ICD8 til ICD10 diagnostiseringssystemet pr. 1. januar 1994. Da definitionen af tidlig skizofreni omfatter personer der får diagnosen skizofreni inden udgangen af deres 25. år, giver dette er knækpunkt ved 1. januar 1968. Dette skyldes, at en person født efter knækpunktet herved har en mulighed for, at få diagnosen tidlig skizofreni efter ICD10 systemet. Placeringen af knækpunktet er endvidere kontrolleret ved lave en profil likelihood for placeringen. Denne analyse kan ikke afvise den angivne placering af knækpunktet.

Der undersøges endvidere for ekstra temporale variationer i data. Denne analyse finder ingen klare indikationer af at der skulle være sådanne. Tværtimod forkastes alle hypoteser om systematisk variation. Dvs.hypoteserne om at der er yderligere variationer, enten i form af serier af harmoniske svingnin-

ger, eller ustrukturerede effekter, som skyldes fødselsmåneden. Tilsvarende findes der heller ikke nogen klare ekstra temporale stokastiske variationer. Der er dog en lille indikation af sådanne i korrelogrammet for kvinderne.

Det skal bemærkes at hvis der sammenlignes med de mulige faktorer, som er introduceret i afsnit 7.1.2, kan der ikke i denne analyse påvises den årstidsvariation, som har været fundet i andre undersøgelser.

Efterfølgende analyseres data ved en DGLM med Gaussisk systemligning. I denne analyse betragtes en model, hvor den latente proces modelleres ved en AR(1) proces. Denne model viser sig dog kun at være anvendelig for kvinderne, hvorved det kun er denne del af data, som betragtes i den videre modellering.

Fra afsnit 7.1.2 ses, at en mulig faktor, som evt. kunne forklare den anvendte AR(1) proces, er influenza under graviditeten.

Ved hjælp af GLIMMIX makroen i SAS estimeres ukendte hyperparametre. Ud fra disse estimeres den latente proces ved det udvidede Kalmanfilter, med iterationsproceduren beskrevet i afsnit 4.3.1. Det udvidede Kalmanfilter er kodet op i IML i SAS, direkte ud fra den beskrevne teori. Algoritmen er testet ved simulerede datasæt for henholdsvis binomialfordelte og poissonfordelte observationer. Desuden har Claus Dethlefsen været så venlig, at køre et simuleret poissonfordelt datasæt igennem en tilsvarende algoritme han har kodet op i R, specielt for poissonfordelte observationer. Denne test gav det samme, som den algoritme der er anvendt her.

Den anvendte algoritme estimerer således både de ikke dynamiske regressionsparametre, og den dynamiske AR(1) proces. Det skal således bemærkes, at de punktvise intervaller for den latente AR(1) proces indeholder nul til enhver tid. Den estimerede middelværdi for processen, bør derfor tages med en pæn portion skepsis.

For at give en kritisk gennemgang af estimaterne for regressionsparemetrene, sammenlignes de med estimaterne for en tilsvarende GLM analyse, og en tilsvarende analyse ved GLIMMIX modellen. Denne sammenligning viser, at estimaterne ikke er væsentligt forskellige. Dvs. at der i denne sammenhæng lige så godt kan anvendes en GLM til at estimere regressionsparametrene, som den mere komplicerede DGLM eller GLIMMIX model.

Der foretages endvidere en kritisk gennemgang af den anvendte AR(1) model, ved at sammenligne den med den mere komplicerede TOEP model. Ud fra denne analyse ses, at AR(1) strukturen ikke passer helt dårligt. Mere avancerede processer kunne dog tænkes, at være anvendelige, men da en undersøgelse af dette ligger, også uden for rammerne af dette speciale, arbejdes ikke videre med emnet.

Da de betragtede strukturer for en latent proces ikke kan anvendes til mænd,

kunne det være interessant at kigge nærmere på dette emne. Dvs. undersøge om der overhovedet kan påvises en latent proces for mænd, og i givet fald, hvilken struktur den kunne tænkes, at have. Der har været prøvet med en AR struktur, en random walk samt den beskrevne TOEP model, men ingen af dem er anvendelige for mændene. Emnet er dog ikke blevet analyseret nærmere, da tiden løber.

Hypotesetest for dynamiske stokastiske parametre, samt modelkontrol for dynamiske modeller, er generelt et meget lidt udforsket område, hvorved disse emner ikke er indarbejdet yderligere i denne analyse. Det kunne ikke desto mindre være interessant, at arbejde videre med området.

Et andet emne der kunne være interessante at betragte, hvis tiden tillod det, er hvorvidt andre forklarende variabler kunne tænkes at forklare en yderligere del af variationen i data. Forslag til sådanne variabler er opsummeret i afsnit 7.1.2, men det er dog ikke givet, at der eksisterer datamateriale til dem.
Det kunne endvidere være interessant, at undersøge hvor vidt de "drop outs"der er i datasættet, kunne tænkes at have en skizofrenirisiko forskellig fra personerne i datasættet. Da antallet af "drop outs"ikke er særlig stort, og metoder til sådanne analyser ligger uden for rammerne af specialet, er der ikke set nærmere på dette emne.

Del IV

Appendiks

A Fordelinger

I dette appendiks gennemgåes en række fordelinger, samt egenskaber ved disse fordelinger, som anvendes i dette speciale.

A.1 Poissonfordelingen

Poissonfordelingen anvendes, hvis et forsøg med en lille sandsynlighed for, at en hændelse A indtræffer, udføres et stort antal gange, og de enkelte forsøg er uafhængige. Størrelsen X angiver da antallet af forsøg, der resulterer i hændelsen A.

Poissonfordelingen kan defineres formelt ved følgende definition.

Definition A.1 **(Poissonfordelingen)**
Lad X være en stokastisk variabel, og lad $\xi \in \mathbb{R}$, da er X poissonfordelt, hvis den tilhørende tæthedsfunktion er på formen,

$$f(x) = \frac{\exp(-\xi)\,\xi^x}{x!},$$

for $x \in \mathbb{N}$. For alle andre værdier af x er $f(x) = 0$.
Notationen er $X \sim Poi(\xi)$. $\triangle$

Tæthedsfunktionen for poissonfordelingen kan omskrives til den eksponentielle form (3.1), ved

$$f(x|\eta, V) = \exp\Big[x\underbrace{\ln(\xi)}_{\eta} - \underbrace{\xi}_{b(\eta)} \underbrace{-\ln(x!)}_{c(x,V)}\Big], \tag{A.1}$$

hvor $a(V) = 1$.

Ud fra den generelle teori for tætheder på eksponentiel form, findes middelværdien og variansen ud fra (3.4) og (3.5), ved

$$\mathbb{E}[X] = \xi \quad \text{og} \quad \mathbb{V}\text{ar}[X] = \xi. \tag{A.2}$$

A.1.1 Den konjugerede fordeling

En konjugeret fordeling til poissonfordelingen, kan findes ved at anvende en fordelingen med en tæthedsfunktion på samme form, som likelihoodfunktionen til X. Ud fra definition A.1 ses, at gammafordelingen, som defineret nedenfor, kan anvendes som konjugeret fordeling til poissonfordelingen.

Hvis den eksponentielle tæthedsfunktion (A.1) for poissonfordelingen anvendes, er parameteren givet ved $\eta = \ln(\xi)$. I dette tilfælde kan gammafordelingen derfor ikke anvendes, som konjugeret fordeling. Istedet anvendes en log-gammafordeling, som gennemgåes efter gammafordelingen.

Gammafordelingen

For at definere gammafordelingen indføres først gammafunktionen. For $\alpha \in \mathbb{R}_+$ er denne givet ved

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt. \tag{A.3}$$

Herefter kan standard gammafordelingen defineres, som følger.

***Definition A.2* (Standard gammafordelingen)**
Lad Z være en stokastisk variabel, og lad $\alpha \in \mathbb{R}_+$ være **formparameter**, *da er Z standard gammafordelt, hvis den tilhørende tæthedsfunktion er på formen*

$$f(z) = \frac{1}{\Gamma(\alpha)} z^{\alpha-1} \exp(-z),$$

for $z \in \mathbb{R}_+$. For $z \notin \mathbb{R}_+$ er $f(z) = 0$.
Notationen er $Z \sim$ Gamma$(\alpha, 1)$. △

Ud fra tæthedsfunktionen i definition A.2 ses, at den kumulantfrembringende funktion for standard gammafordelingen, er givet ved

$$\mathrm{K}(t) = -\alpha \ln(1-t).$$

Dvs.

$$\mathbb{E}[Z] = \mathrm{K}'(0) = \alpha \quad \text{og} \quad \mathbb{V}\mathrm{ar}[Z] = \mathrm{K}''(0) = \alpha. \tag{A.4}$$

En skaleret gammafordeling har en ekstra **skalaparameter** $\beta \in \mathbb{R}_+$. Hvis en stokastisk variabel X er gammafordelt med formparameter α og skalaparameter β, defineres fordelingen af X ved, at $X = \beta^{-1}Z$, hvor $Z \sim \text{Gamma}(\alpha, 1)$. Denne udgave af gammafordelingen har således tæthedsfunktionen

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x),$$

for $x \in \mathbb{R}_+$. For $x \notin \mathbb{R}_+$ er $f(x) = 0$. Notationen er $X \sim \text{Gamma}(\alpha, \beta)$.

Middelværdien og variansen af $X \sim \text{Gamma}(\alpha, \beta)$, kan findes, ud fra middelværdien og variansen til standard gammafordelingen i (A.4), ved

$$\mathbb{E}[X] = \frac{\alpha}{\beta} \quad \text{og} \quad \mathbb{V}\mathrm{ar}[X] = \frac{\alpha}{\beta^2}. \tag{A.5}$$

Log-gammafordelingen

For at definere log-gammafordelingen indføres først digammafunktionen. For $\alpha \in \mathbb{R}_+$ er denne givet ved

$$\gamma(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}. \tag{A.6}$$

Herefter kan log-gammafordelingen defineres, som følger.

Definition A.3 **(Log-gammafordelingen)**
Lad X være en stokastisk variabel med fordelingen $X \sim$ Gamma(α, β). Den stokastiske variabel $Y = \ln(X)$ er da log-gammafordelt, hvilket betegnes $Y \sim$ log-gamma(α, β). △

Da $X = \beta^{-1}Z$, hvor $Z \sim \text{Gamma}(\alpha, 1)$, findes

$$Y_1 = Y + \ln(\beta) = \ln(X) + \ln(\beta) = \ln(Z).$$

For at finde momenterne til log-gammafordelingen anvendes teorien for tæthedsfunktioner på den eksponentielle form. Dette gøres ved, at finde tæthedsfunktionen for Y_1 ud fra tæthedsfunktionen af Z, og observere, at den

opnåede tæthedsfunktion er på den eksponentielle form. Dvs.

$$\begin{aligned} f_{Y_1}(y_1) &= \frac{d}{dy_1} F_Z(\exp(y_1)) \\ &= \exp(y_1)\, f_Z(\exp(y_1)) \\ &= \exp(y_1)\, \frac{1}{\Gamma(\alpha)} \big(\exp(y_1)\big)^{\alpha-1} \exp(-\exp(y_1)) \\ &= \exp\Big[y_1 \underbrace{\alpha}_{\eta} - \underbrace{\ln(\Gamma(\alpha))}_{b(\eta)} \underbrace{-\exp(y_1)}_{c(y_1,V)}\Big], \end{aligned} \tag{A.7}$$

hvor $a(V) = 1$.

Ved at betragte (A.1) ses, at tæthedsfunktion for log-gammafordelingen, stemmer ganske godt over ens med tæthedsfunktionen for $(\eta|x)$, hvor x er den naturlige parameter.

Ud fra den generelle teori for tætheder på eksponentiel form, findes middelværdien og variansen for Y_1, ud fra (3.4) og (3.5), ved

$$\mathbb{E}[Y_1] = \gamma(\alpha) \quad \text{og} \quad \mathbb{V}\text{ar}[Y_1] = \gamma'(\alpha), \tag{A.8}$$

hvor $\gamma(\cdot)$ er digammafunktionen givet ved (A.6).

Herved kan middelværdien og variansen for Y findes ved

$$\mathbb{E}[Y] = \gamma(\alpha) - \ln(\beta) \quad \text{og} \quad \mathbb{V}\text{ar}[Y] = \gamma'(\alpha). \tag{A.9}$$

Ud fra en udledning ækvivalent med (A.7), er tæthedsfunktionen for Y givet ved

$$f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} \exp(\alpha y - \beta \exp(y)). \tag{A.10}$$

De ovenstående resultater for log-gammafordelingen er ligeledes gennemgået i en sætning i Larsson [2001, s. 165-167], men udledningen gennemgås dog også her, da resultaterne kan vises noget kortere på denne form.

A.2 Binomialfordelingen

Lad et forsøg med sandsynligheden p, for at en hændelse A indtræffer, være gentaget n gange. Er de enkelte forsøg uafhængige af hinanden, og angiver den stokastiske variabel X antallet af forsøg, der resulterer i hændelsen A, da betegnes X som værende binomialfordelt.

Binomialfordelingen kan defineres formelt ved følgende definition.

Definition A.4 **(Binomialfordelingen)**
Lad X være en stokastisk variabel, lad $n \in \mathbb{Z}_+$, og lad $p \in \mathbb{R}$ opfylde at $0 < p < 1$, da er X binomialfordelt, hvis den tilhørende tæthedsfunktion er på formen

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x},$$

for $x = 0, 1, \ldots, n$. For alle andre værdier af x er $f(x) = 0$.
Notationen er $X \sim Bi(n, p)$. Δ

Tæthedsfunktionen for binomialfordelingen kan omskrives til den eksponentielle form (3.1), ved

$$f(y|\eta, V) = \exp\Big[n\big(y \underbrace{\operatorname{logit}(p)}_{\eta} - \underbrace{\ln(1+\exp(\eta))}_{b(\eta)}\big) + \underbrace{\ln\left(\binom{n}{y}\right)}_{c(y,V)}\Big], \quad \text{(A.11)}$$

hvor $a(V) = 1/n$, og $y = x/n$.

Ud fra den generelle teori for tætheder på eksponentiel form, findes middelværdien og variansen ud fra (3.4) og (3.5), ved

$$\mathbb{E}[X] = np \quad \text{og} \quad \mathbb{V}\text{ar}[X] = np(1-p). \quad \text{(A.12)}$$

A.2.1 Den konjugerede fordeling

En konjugeret fordeling til binomialfordelingen kan findes ved, at anvende en fordelingen med en tæthedsfunktion på samme form, som likelihoodfunktionen til X. Ud fra definition A.4 ses, at betafordelingen, som defineret nedenfor, kan anvendes som konjugeret fordeling til binomialfordelingen.

Betafordelingen

For at definere betafordelingen indføres først betafunktionen. Denne er defineret ved

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}, \quad \text{(A.13)}$$

hvor gammafunktionen $\Gamma(\cdot)$ er givet ved (A.3).

Betafordelingen kan herefter defineres, som følger.

Definition A.5 **(Betafordelingen)**
Lad X være en stokastisk variabel, og lad $\alpha, \beta \in \mathbb{R}_+$, da er X betafordelt,

hvis den tilhørende tæthedsfunktion er på formen

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

for $0 < x < 1$. *For* $x \leq 0$ *eller* $x \geq 1$ *er* $f(x) = 0$.
Notationen er $X \sim Beta(\alpha, \beta)$. △

Middelværdi og varians har formlerne

$$\mathbb{E}[X] = \frac{\alpha}{\alpha + \beta}, \tag{A.14}$$

og

$$\mathbb{V}\text{ar}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

[Bernardo and Smith, 2000, s. 116-117].

B Profil likelihood

Dette appendiks indeholder en kort introduktion til **profil likelihood** metoder. En dybere indførsel i teorien er bla. givet i Barndorff-Nielsen and Cox [1994].

Likelihood

Da profil likelihood metoder bygger på maksimum likelihood estimation (MLE), introduceres dette først ganske kort.

Lad $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ være realisationer af en stokastisk variabel Y, med tæthedsfunktion $f(y|\boldsymbol{\beta})$. Likelihoodfunktionen $L(\boldsymbol{\beta}|\boldsymbol{y})$ for parametervektoren $\boldsymbol{\beta}$, er da givet ved den simultane tæthedsfunktion $f(y_1, \ldots, y_n|\boldsymbol{\beta})$, taget som en funktion af $\boldsymbol{\beta}$. Værdien af $\boldsymbol{\beta}$, hvor $L(\boldsymbol{\beta}|\boldsymbol{y})$ antager sit maksimum, betegnes MLE, og har notationen $\hat{\boldsymbol{\beta}}$.
Hvis der arbejdes med en stokastisk vektorer er likelihoodfunktionen tilsvarende givet ved den tilhørende simultane tæthedsfunktion til realisationerne af den stokastiske vektor.

Maksimum likelihood metodens anvendelighed bygger på, at $\hat{\boldsymbol{\beta}}$ under generelle regularitetsbetingelser, er approksimativt normalfordelt ved

$$\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}_p\left(\boldsymbol{\beta}, \boldsymbol{\mathcal{I}}^{-1}\left(\hat{\boldsymbol{\beta}}\right)\right), \tag{B.1}$$

for $n \to \infty$, hvor $\hat{\boldsymbol{\beta}}$ er maksimum likelihoodestimatet for $\boldsymbol{\beta}$, og $\boldsymbol{\mathcal{I}}\left(\hat{\boldsymbol{\beta}}\right)$ er den

tilhørende informationsmatrix givet ved (3.7). [Fahrmeir and Tutz, 2001, s. 44-45].

I mange sammenhænge anvendes **log-likelihood funktionen** $l(\cdot) = \ln(L(\cdot))$ i stedet for likelihoodfunktionen. Da logaritmefunktionen er bijektiv, og monoton voksende, har log-likelihoodfunktionen de samme egenskaber, som ovenfor er beskrevet for likelihoodfunktionen.

Hvis realisationerne $(y_1, \ldots, y_n)^{\mathrm{T}}$ er uafhængige, er likelihoodfunktionen givet ved

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}),$$

hvorved log-likelihoodfunktionen er givet ved

$$l(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{n} \ln(f(y_i|\boldsymbol{\beta})).$$

Profil likelihood

Lad parametervektoren være givet ved $\boldsymbol{\beta} = (\alpha, \beta_1, \ldots, \beta_k)^{\mathrm{T}}$, hvor $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$. Hvis der ønskes en likelihoodfunktion for α, kan profil likelihooden anvendes. Denne findes ved, for enhver fast værdi af α, at maksimere $l(\boldsymbol{\beta})$ med hensyn til $\boldsymbol{\beta}_1$, hvorved der opnås en værdi $\hat{\boldsymbol{\beta}}(\alpha)$, som afhænger af α. Hvis denne værdi indsættes i log-likelihoodfunktionen $l(\cdot)$, opnås profil likelihoodfunktionen for α, givet ved

$$l^p(\alpha) = l\left(\begin{bmatrix} \alpha \\ \hat{\boldsymbol{\beta}}(\alpha) \end{bmatrix}\right).$$

Maksimum likelihood estimatet for $\boldsymbol{\beta}$ opnås herefter ved, at maksimere $l^p(\alpha)$, hvorved estimatet $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\alpha})$ opnås.

C Gennemgang af analysetabeller

De betegnelser der er anvendes i tabellerne over udskrifterne fra SAS, er gennemgået i den følgende liste.

Koefficient er koefficienterne til forklarende variable, der indgår i analysen (β'erne).

Estimat er estimatet af den tilhørende koefficient (β-værdi).

Standard afvigelse er standard afvigelsen på estimatet, forkortet std.afv.

LR 95% konfidensgrænser er grænserne for et 95% konfidensinterval, opnået ud fra χ^2-værdien til en statistik baseret på deviansen (se afsnit 3.2.2).

95 % interval er grænserne for et 95% interval baseret på covariansmatricen, der er estimeret ved det udvidede Kalmanfilter. Værdierne er udregnet ved den tilhørende Wald statistik.

P-værdi er P-værdien for hypotesen, at den givne parameter kan udelades. I GLM analysen bygger den anvendte test på χ^2-værdien fundet ved deviansen (se afsnit 3.2.2). I DGLM analysen er det Wald's test, der anvendes (denne er ligelides introduceret i afsnit 3.2.2).
I **type 1 analysen** starter testen nedefra i listen af forklarende variable, og kører opefter. Således er den hypotese der testes, af modellen indeholdende de ovenstående forklarende variabler, mod modellen

indeholdende den givne forklarende variabel, og de ovenstående forklarende variabler. Herved testes de forklarende variabler væk i rækkefølge, i denne analyse.

I **type 3 analysen** testes modellen indeholdene alle de forklarende variabler undtagen den angivne, mod modellen indeholdene alle forklarende variabler. Dette er herved en test af den enkelte forklarende variabel.

NOTE: Bemærk af estimationsprocedurerne anvender "corner point estimation"for at opnå regulærere designmaticer. Dvs. den sidste værdi i hver forklarende variabel anvendes som referenceværdi. En dybere indførsel i denne metode, og andre tilsvarende metoder er bla. givet i [McCullagh and Nelder, 1989, s. 65-67] △

D Diverse

Dette appendiks indeholder forskellige sætninger, der anvendes i specialet.

Sætning D.1
For to stokastiske vektorer $\boldsymbol{X}$ og $\boldsymbol{Y}$ gælder, at

$$\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]] \quad og \quad \mathbb{V}ar[\boldsymbol{Y}] = \mathbb{E}[\mathbb{V}ar[\boldsymbol{Y}|\boldsymbol{X}]] + \mathbb{V}ar[\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]] .$$

[West and Harrison, 1997, s. 635]

Δ

Sætning D.2 (Bayes sætning)
Lad $f(\boldsymbol{y}|\boldsymbol{\theta})$ være tæthedsfunktionen til fordelingen af $(\boldsymbol{Y}|\boldsymbol{\theta})$, da gælder at

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\, f(\boldsymbol{\theta})}{f(\boldsymbol{y})} \propto f(\boldsymbol{y}|\boldsymbol{\theta})\, f(\boldsymbol{\theta}) .$$

Δ

Sætning D.3 (Lineært Bayesiansk estimation)
Lad fordelingen af den stakkede stokastiske vektor af $\boldsymbol{\theta}$ og λ være delvist specificeret således, at

$$\begin{bmatrix} \boldsymbol{\theta} \\ \lambda \end{bmatrix} \sim \left[\begin{bmatrix} \boldsymbol{\mu}_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^{\mathrm{T}} & \Sigma_{22} \end{bmatrix} \right] .$$

Det lineære Bayesianske estimat for middelværdivektoren $\boldsymbol{m}$*, og covariansmatricen* $\boldsymbol{C}$*, til* $(\boldsymbol{\theta}|\lambda)$*, er da givet ved henholdsvis*

$$\boldsymbol{m} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\frac{\lambda - \mu_2}{\Sigma_{22}} \quad \textit{og} \quad \boldsymbol{C} = \boldsymbol{\Sigma}_{11} - \frac{\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{12}^{\mathrm{T}}}{\Sigma_{22}}.$$

Δ

[West and Harrison, 1997, s. 122-124]. Dette er endvidere gennemgået fint i West et al. [1985, s. 76-77].

E Notation

Den nedenstående liste er en gennemgang af den **notation**, som er anvendt i dette speciale. Notationerne for stokastiske variable, stokastiske vektorer, realisationer af disse, samt skalarer, faste vektorer og faste matricer, har i visse tilfælde overlap. Det må derfor ses ud af konteksten hvad der er tale om.

Y	Stokastisk variabel.
y	Realisation af en stokastisk variabel.
$\boldsymbol{Y}$	Stokastisk vektorer.
$\boldsymbol{y}$	Realisation af en stokastisk vektor.
$\{y_t\}$	Tidsrække.
$\{\boldsymbol{y}_t\}$	Longitudinelle data.
$\bar{\boldsymbol{y}}$	Gennemsnittet af stokastiske vektorer.
$\hat{\cdot}$	Estimat.
$\boldsymbol{A}$	Ikke stokastisk matrix.
$\boldsymbol{A}_t(k)$	En k-trins prædiktion til tiden t, af matricen $\boldsymbol{A}$.
$\boldsymbol{I}$	Identitets matrix.
$\widetilde{m}$	Værdi opnået ved Kalmanudglatning.
$\boldsymbol{A}^T$	Den transponerede af en matrix.
$\boldsymbol{A}^{-1}$	Den inverse af en matrix.

$\mathbb{E}[\cdot]$	Middelværdioperator, anvendt på stokastiske variable, stokastiske vektorer og stokastiske matricer.
$\mathbb{V}\mathrm{ar}[\cdot]$	Variansoperator, anvendt på stokastiske variable, og stokastiske vektorer.
$\mathcal{V}(\cdot)$	Variansfunktion.
$\mathrm{Cov}[\cdot,\cdot]$	Covariansoperator, anvendt på stokastiske variable, og stokastiske vektorer.
$\mathrm{D}(\cdot)$	Deviansoperator.
D_t	Informationen til tiden t, for en given model.
$\mathrm{M}(\cdot)$	Momentfembringende funktion.
$\mathrm{K}(\cdot)$	Kumulantfrembringende funktion.
$r_\cdot^{(d)}$	Deviansresidual.
$r_\cdot^{(p)}$	Pearsonresidual.
$\mathrm{Bi}(n,p)$	Binomialfordeling med parametrene n og p.
$\mathrm{B}(\alpha,\beta)$	Betafunktionen med parametrene α og β.
$\mathrm{Beta}(\alpha,\beta)$	Betafordelingen med parametrene α og β.
$\mathcal{N}(\mu,\sigma^2)$	Normalfordeling med parametrene μ and σ^2.
$\mathcal{N}_d(\boldsymbol{\mu},\boldsymbol{\Sigma})$	Den flerdimensionale normalfordeling med d dimensioner, og parametrene $\boldsymbol{\mu}$ og $\boldsymbol{\Sigma}$.
$\mathrm{Poi}(\mu)$	Poissonfordeling med parameteren μ.
$\Gamma(\nu)$	Gammafunktion med parameter ν.
$\gamma(\nu)$	Digammafunktion med parameter ν.
$\mathrm{Gamma}(\alpha,\beta)$	Gammafordeling med formparameteren α og skalaparameteren β.
$\text{log-gamma}(\alpha,\beta)$	Log-gammafordeling med parametrene α og β.
$\chi^2(n)$	En χ^2 fordeling med n frihedsgrader.
$\mathrm{F}(n,m)$	En F-fordelingen med n tællerfrihedsgrader. og m nævnerfrihedsgrader.
$[\mu,V]$	Delvist specificeret fordeling med første moment μ, og andet moment V.
$\mathcal{CP}(r,s)$	Den konjugerede fordeling med parametrene r og s.
M_i	Model i.
$P(\cdot)$	Sandsynligheden for det angivne udtryk.
$F_Y(y)$	Fordelingsfunktionen for den stokastiske variabel Y.
$f(y)$	Tæthedsfunktionen for den stokastisk variabel Y. Andre bogstaver end f anvendes nogle steder.
$f_X(y)$	Tæthedsfunktion for den stokastisk variabel X, med værdier givet ved y.
$f(\cdot,\cdot)$	Simultan tæthedsfunktion for to stokastiske variable.
$f(\cdot\|\cdot)$	Tæthedsfunktion for en betinget fordeling af en stokastisk variabel.
$\mathcal{G}(\cdot)$	Tæthedsfunktion for en Gaussisk fordeling af en stokastisk variabel.
$\ln(\cdot)$	Den naturlige logaritme.
$\mathrm{logit}(\cdot)$	Logit-funktionen, defineret ved $\mathrm{logit}(p) = \ln(p/(1-p))$, hvor $0 < p < 1$.

$\mathbb{N}$	De naturlige tal, dvs. de positive heltal og nul.
$\mathbb{Z}$	Heltallene.
$\mathbb{Z}_+$	De positive heltal.
$\mathbb{R}$	De reelle tal.
$\mathbb{R}_+$	De positive reelle tal.
$\mathbb{R}^n$	Et n-dimensionalt Euklidisk vektorrum af reelle tal.
$L(\cdot)$	Likelihoodfunktionen.
$l(\cdot)$	Log-likelihoodfunktionen.
$l^p(\cdot)$	Profil likelihoodfunktionen.
$\sim$	Fordelt som.
$\dot{\sim}$	Approksimativt fordelt som.
$\mid$	Givet.
$\doteq$	Approksimativt lig med.
$\approx$	Cirka lig med.
$\equiv$	Identisk fordelt.
$\propto$	Proportional med.
$\dot{\propto}$	Approksimativt proportional med.
$\square$	Bevis slut.
$\vartriangle$	Definition, sætning, korollar, note, model eller metode slut.

Forkortelser

LM	Lineær normal model.
GLM	Generaliseret lineær model.
DLM	Dynamisk lineær model.
DGLM	Dynamisk generaliseret lineær model.
MLE	Anvendt for Maximum Likelihood Estimat eller Maximum Likelihood Estimator, afhængigt af konteksten.

Litteratur

O. E. Barndorff-Nielsen and D. R. Cox. *Inference and Asymptotics.* Monographs on Statistics and Applied Probability 52. Chapman and Hall, 1994.

J. Bernardo and A. Smith. *Bayesian Theory.* Wiley, 2000.

C. Dethlefsen. *Space Time Problems and Applications.* PhD thesis, Aalborg University, 2001.

P. J. Diggle. *Time Series, A Biostatistical Introduction.* Oxford Statistical Science Series 5. Oxford University Press, first edition, 1990.

A. J. Dobson. *An Introduction to Generalized Linear Models.* Chapman and Hall, first edition, 1990.

J. Durbin and S. J. Koopman. Time series analysis of non-gaussian observations based on state space models from both classical and bayesian perspectives. *Journal of Royal Statistical Society*, 62:3–56, 2000.

J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods.* Oxford Statistical Science Series. Oxford University Press, New York, first edition, 2001.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer-Verlag, first edition, 1994.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer-Verlag, second edition, 2001.

G. R. Grimmet and D. R. Stirzaker. *Probability and Random Proceses.* Oxford University Press, second edition, 1992.

K. Gronemann. *Psykiatri -en grundbog.* Kava, 1995. Fjerde udgave.

H. Häfner and W. Heiden. Epidemiology af schizophrenia. *The Canadian Journal of Psychiatry*, 42(2):139–145, 1997.

A. Jablensky. The 100-year epidemiology of schizophrenia. *Schizophrenia Research*, (28):111–125, 1997.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 35-45(82), 1960.

R. E. Kalman. New methods in wiener filtering theory. In J.L. Bogdanoff and F. Kozin, editors, *First Symposium of Engineering Applications of Random Function Theory and Probability*. Wiley, New York, 1963.

R. E. Kendel and W. Adams. Unexplained fluctuations in the risk for schizophrenia by month and year of birth. *British Journal of Psychiatry*, 158:758–763, 1991.

K. N. Jensen & H. J. Larsson. Dynamiske modeller for tidsrækker -med focus på poissonfordelte observationer. Master thesis, Aalborg University, June 2001.

S. L. Lauritzen. Time series analysis in 1880: A discussion of contributions made by T. N. Thiele. *International Statistical Review*, 49:319–331, 1981.

E. L. Lehmann and G. Casella. *Theory og Point Estimation*. Springer, second edition, 1998.

J. K. Lindsey. *Applying Generalized Linear Models*. Springer Texts in Statistics. Springer, New York, 1997.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman and Hall, second edition, 1989.

P. B. Mortensen. Hvorfor er skizofrenirisikoen større i byer end på landet? *Ugeskrift for Læger*, 163/35(27. august 2001):4717–4720, 2001.

P. B. Mortensen, C. B. Pedersen, T. Westergaard, J. Wohlfahrt, H. Ewald, O. Mors, P. K. Andersen, and M. Melbye. Effects of family history and place and season of birth on the risk of schizophrenia. *The New England Journal of Medicine*, 340(February 25, 1999):603–608, 1999.

J. Nelder and R. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society*, A 135(3):370–384, 1972.

T. Obel and M. B. Pedersen. Statistical Models for Longitudinal Data - featuring Tweedie class state space models. Master thesis, ir-96-2005, Aalborg University, July 1996.

SAS institute inc. *SAS/STAT User's Guide, Version 8*, 1999. Vol. 1,2 and 3.

G. A. F. Seber. *Multivariate Observations*. Wiley series in Probability and Mathematical Statistics. John Wiley & Sons, first edition, 1984.

P. Vestergaard & T. Sørensen. *Psykiatri, En lærebog om voksnes psykiske sygdomme*. Foreningen af danske lægestuderenes forlag, 2000. Første udgave, andet oplag.

T. N. Thiele. Om anvendelse af mindste kvadraters metode i nogle tilfælde, hvor en komplikation af visse slags uensartede tilfældige fejlkilder giver fejlene en "systematisk"karakter. *Videnskabeligt Selskabs Skrift*, 1880. 5. Række, naturvidenskabelig og matematisk Afd. XII. 5.

E. F. Torrey, P. B. Mortensen, C. B. Pedersen, J. Wohlfahrt, and M. Melbye. Risk factors and confounders in the geographical clustering of schizophrenia. *Schizophrenia Research*, 49 (2001):295–299, 2000.

W. Venables and B. Ripley. *Modern Applied Statistics with S-PLUS*. Statistics and Computing. Springer-Verlag, second edition, 1997.

M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, second edition, 1997.

M. West, J. Harrison, and H. Migon. Dynamic generalized linear models and bayesian forecasting. *Journal of American Statistical Association*, 80 (Martch):73–83, 1985.

T. Westergaard, P. B. Mortensen, C. B. Pedersen, J. Wohlfahrt, and M. Melbye. Exposure to prenatal and childhood infrections and the risk of schizophrenia. *Arch Gen Psychiatry*, 56(Nov 1999):993–998, 1999.

World Health Organization. *Manual of the International Classification of Diseases (ICD-8)*, 1967.

World Health Organization. *ICD-10, The ICD-10 Classification of Mental and Behavioural Disorder*, 1992.

Indeks