



Statisk og Dynamisk Modellering af Incidens af Akut Myokardie Infarkt

af Stefan Christensen

**TITEL:**

Statisk og Dynamisk Modellering af Incidens af Akut Myokardie Infarkt

FORFATTER:

Stefan Christensen

VEJLEDER:

Søren Lundbye-Christensen

SEMESTER:

MAT6

PROJEKTPERIODE:

2/9-2004 til 10/1-2006

ANTAL KOPIER:

10

ANTAL SIDER:

117 sider

ANTAL SIDER I ALT:

129 sider

SYNOPSIS Teori-delen af denne speciale-rapport omhandler generaliserede lineære modeller og state space modeller. For den generaliserede lineære model beskrives maksimum-likelihood estimation af parametre, inferens og modelkontrol. For den Gaussiske state space model, også kaldet den dynamiske lineære model, beskrives Kalman-filtrering, Kalman-udglatning og maksimum-likelihood estimation af ukendte parametre. Desuden beskrives en specifik type af state space modeller kaldet strukturelle tidsrække modeller. Derefter udvides teorien til state space modeller med observationsfordeling fra den eksponentielle familie eller dynamiske generaliserede lineære modeller. Der beskrives to forskellige metoder til at foretage Kalman-filtrering og -smoothing i en dynamisk generaliseret lineær model. Anvendelses-delen af specialet beskæftiger sig med analyse af Poisson-fordelte data for antallet af daglige indlæggelser som følge af første-gangs blodpropper i hjertet i perioden 1983-1999 i Nordjyllands amt. Dataene analyseres ved hjælp af en Poisson regressionsmodel, en strukturel tidsrække model og en Poisson state space model. Dataene udviser sæsonvariation med toppunkter forår og efterår/vinter og lavpunkt om sommeren. Sæsonvariationen afhænger af alder men ikke af køn. Analysen med en Poisson state space model viser, at sæsonvariationen har ændret sig svagt i løbet af undersøgelsesperioden.

Forord

Denne speciale-rapport omhandler generaliserede lineære modeller og state space modeller og er udarbejdet i perioden september 2004 til januar 2006 på Aalborg Universitet.

I rapporten er kildehenvisninger markeret med [<kildenummer>], hvor kildenummeret er anført i litteraturlisten. Vektorer er symboliseret ved fed skrift, $\mathbf{x}$, og matricer er repræsenteret ved dobbelt understregning, $\underline{\underline{X}}$.

Jeg vil gerne takke min vejleder Søren Lundbye-Christensen. Jeg har nydt stor gavn af hans store faglige viden, engagement og smittende humør. Jeg vil også gerne takke Claus Dethlefsen for hans velvilje til at besvare mine spørgsmål i forbindelse med min anvendelse af program-pakken *sspir* i data-analysen.

Aalborg den 10. januar 2006.

Stefan Christensen

Indhold

1	Indledning	1
I	Teori-del	3
2	Den Generaliserede Lineære Model	5
2.1	Den eksponentielle familie af fordelinger	5
2.1.1	Poisson-fordelte observationer	6
2.2	Generaliserede lineære modeller	7
2.2.1	Poisson-model	8
2.3	Maksimum-likelihood estimation	8
2.3.1	Maximum-likelihood estimation for Poisson-modellen	10
2.3.2	Fordeling af maksimum-likelihood estimator	10
2.4	Inferens	11
2.4.1	Devians	11
2.4.2	Test baseret på deviansen	12
2.4.3	Wald-test	13

2.5	Modelkontrol	13
2.5.1	Standardiserede residualer	13
2.5.2	Varianshomogenitet	14
2.5.3	Seriel korrelation	17
2.5.4	Fordelingstype	20
3	Den Dynamiske Lineære Model	23
3.1	Den dynamiske lineære model	23
3.2	Betinget uafhængighed i den DLM	25
3.3	Kalman-filtrering	25
3.4	Kalman-udglatning	29
3.5	Maksimum-likelihood estimation af ukendte parametre	30
3.6	Strukturelle tidsrække modeller	32
3.6.1	Filtrering og udglatning	33
3.6.2	Modelkontrol	35
3.6.3	StructTS	36
4	Den Dynamiske Generaliserede Lineære Model	39
4.1	Den dynamiske generaliserede lineære model	39
4.2	Kalman-filtrering med konjugeret a priori fordeling	40
4.2.1	Kalman-filtrering for Poisson-modellen	43
4.3	Extended Kalman-filtrering og -smoothing	45
4.3.1	Extended Kalman-filtrering og -smoothing for Poisson-modellen	47

4.4	sspir	48
4.5	Estimation af parametre	49
II	Anvendelses-del	51
5	Introduktion til anvendelses-delen	53
5.1	Processen bag dannelsen af en AMI	53
5.1.1	Risikofaktorer	54
5.2	Årsager til sæsonvariationen af AMI	54
5.3	Ugedagsvariation af AMI	55
5.4	Beskrivelse af datasæt	56
5.4.1	Registrering af patienter	56
5.4.2	Aldersstandardisering	56
5.4.3	Resultater	57
5.5	Problemformulering for data-analysen	60
6	Data-analyse med Statisk Poisson Model	61
6.1	Opstilling af model	61
6.2	Modelreduktion	62
6.3	Modelkontrol	64
6.3.1	Varianshomogenitet	64
6.3.2	Seriel korrelation	66
6.3.3	Fordelingstype	67

6.3.4	Kontrol ved simulering	68
6.4	Resultater	69
6.4.1	Trend	69
6.4.2	Sæsonvariation	71
6.4.3	Ugedagsvariation	75
6.5	Opsamling	77
7	Data-analyse med Strukturel Tidsrække Model	79
7.1	Opstilling af model	79
7.2	Valg af variansparametre	80
7.3	Modelkontrol	84
7.3.1	Varianshomogenitet	84
7.3.2	Seriel korrelation	85
7.3.3	Normalitet	85
7.4	Resultater	86
7.5	Opsamling	89
8	Data-analyse med Dynamisk Poisson Model	91
8.1	Opstilling af model	91
8.2	Valg af variansparametre	93
8.3	Modelkontrol	95
8.3.1	Seriel korrelation	96
8.3.2	Varianshomogenitet	97

8.3.3	Fordelingstype	98
8.4	Resultater	99
8.4.1	Trend	99
8.4.2	Sæsonvariation	100
8.4.3	Ugedagsvariation	105
8.5	Opsamling	106
9	Diskussion	107
9.1	Opsamling af resultater fra data-analysen	107
9.1.1	Trend	107
9.1.2	Sæsonvariation	107
9.1.3	Ugedagsvariation	108
9.2	Diskussion af resultater	108
9.3	Konklusion	109
A	Appendix	111
A.1	Den multivariable normalfordeling	111
A.2	Linear Bayes' estimation	112
A.3	Store tals lov	113

Kapitel 1

Indledning

Dette speciale består af en teori-del, en anvendelses-del samt et appendix.

Teori-delen består af tre kapitler. Kapitel 2 omhandler den generaliserede lineære model. Modellens opbygning, maksimum-likelihood estimation af parametre, inferens og modelkontrol beskrives.

I kapitel 3 beskrives den dynamiske lineære model, Kalman-filtrering og Kalman-udglatning. Desuden beskrives strukturelle tidsrække modeller, som er et special-tilfælde af den dynamiske lineære model, samt maksimum-likelihood estimation af ukendte parametre.

I kapitel 4 udvides teorien for dynamiske modeller til det generelle ikke-normalfordelte tilfælde. Der beskrives to forskellige metoder til at udføre Kalman-filtrering. Den ene er Bayesiansk orienteret og bygger på teorien om konjugerede fordelinger samt linear Bayes' estimation. Den anden metode, der beskrives, er extended Kalman-filtrering og -smoothing, som er baseret på at udlede en Gaussisk approksimation af modellen.

Anvendelses-delen består af en analyse af et datasæt, der omhandler forekomsten af førstegangs-tilfælde (også kaldet *incidensen*) af *akut myokardie infarkt* (blodprop i hjertet - forkortet AMI) i Nordjyllands amt i perioden 1983-1999. I kapitel 5 beskrives det, hvordan en AMI opstår, og hvilke mulige årsager der kan være til årstidsvariationen af AMI-incidens. Derudover beskrives datasættet. Kapitlet afrundes med en problemformulering for data-analysen. I kapitlerne 6, 7 og 8 anvendes teorien fra teori-delen i en analyse af datasættet med henholdsvis en statisk Poisson regressions-model, en strukturel tidsrække model og en dynamisk Poisson model. Anvendelses-delen afsluttes med en diskussion af resultaterne af data-analysen.

Del I

Teori-del

Kapitel 2

Den Generaliserede Lineære Model

Generaliserede lineære modeller (GLM) blev udviklet af Nelder og Wedderburn, se [18], og er en generalisering af regression ved lineære normale modeller til tilfælde, hvor observationerne ikke nødvendigvis er normalfordelte. Mere præcist kan modellen anvendes til regression med observationer, der har en fordeling fra den *eksponentielle familie*.

2.1 Den eksponentielle familie af fordelinger

Definition 2.1 En fordeling tilhører den eksponentielle familie af fordelinger, hvis den tilhørende tætheds-/sandsynlighedsfunktion kan skrives på formen

$$f(y|\eta, V) = \exp\{[d(y)\eta - a(\eta)]/V\}b(y, V), \quad (2.1)$$

hvor a , b og d er kendte funktioner, og a er to gange differentiabel.

Størrelsen η kaldes den *naturlige* parameter, mens $V \geq 1$ er en *dispersions*-parameter, som er inkluderet for at tage højde for de situationer, hvor den observerede varians af Y er større end forventet udfra den antagede fordeling af Y . Hvis $d(y) = y$ siges (2.1) at være på *kanonisk* form.

Ved at udregne den momentfrembringende funktion kan den betingede middelværdi og varians af $d(Y)$ udledes. Den momentfrembringende funktion for $d(Y)$ er givet ved

$$\begin{aligned}
 \psi(t) &= E\{\exp[t \cdot d(Y)]\} \\
 &= \int \exp[t \cdot d(y)] f(y|\eta, V) dy \\
 &= \int \exp[t \cdot d(y)] \exp\{[d(y)\eta - a(\eta)] / V\} b(y, V) dy \\
 &= \int \exp\{[d(y)(\eta + tV) - a(\eta)] / V\} b(y, V) dy \\
 &= \exp\{[a(\eta + tV) - a(\eta)] / V\} \int \exp\{[d(y)(\eta + tV) - a(\eta + tV)] / V\} b(y, V) dy \\
 &= \exp\{[a(\eta + tV) - a(\eta)] / V\}.
 \end{aligned}$$

Det følger, at

$$\psi'(t) = a'(\eta + tV) \exp\{[a(\eta + tV) - a(\eta)] / V\} = a'(\eta + tV) \psi(t),$$

og

$$\psi''(t) = Va''(\eta + tV) \psi(t) + a'(\eta + tV) \psi'(t) = \psi(t) \{Va''(\eta + tV) + [a'(\eta + tV)]^2\}.$$

Ifølge teorien for momentfrembringende funktioner gælder der, at

$$\mu = E[d(Y)] = \psi'(0) = a'(\eta), \quad (2.2)$$

og

$$E\{[d(Y)]^2\} = \psi''(0) = Va''(\eta) + \{E[d(Y)]\}^2,$$

hvoraf der følger, at

$$\text{Var}[d(Y)] = E\{[d(Y)]^2\} - \{E[d(Y)]\}^2 = Va''(\eta). \quad (2.3)$$

Af fordelinger, der tilhører den eksponentielle familie, kan nævnes gamma, binominal, Poisson og normalfordelingen. Da Poisson-fordelte observationer analyseres i anvendelses-delen af rapporten, vil den generelle teori for den GLM, der præsenteres i dette kapitel, løbende blive udledt for Poisson-tilfældet.

2.1.1 Poisson-fordelte observationer

Lad Y være Poisson-fordelt med parameter, μ , det vil sige, μ angiver både middelværdien og variansen for Y . Da gælder der, at

$$f(y|\mu) = \frac{\mu^y \exp(-\mu)}{y!},$$

for $y = 0, 1, \dots$

For at vise at sandsynlighedsfunktionen for en Poisson-fordeling kan skrives på formen (2.1), omskrives den til

$$f(y|\mu) = \frac{\exp(y \ln \mu - \mu)}{y!}. \quad (2.4)$$

For $d(y) = y$, $\eta = \ln \mu$, $a(\eta) = \exp(\eta)$, $V = 1$ og $b(y, V) = 1/y!$, er (2.4) identisk med (2.1), hvilket vil sige, at Poisson-fordelingen tilhører den eksponentielle familie og er på kanonisk form. Det ses, at $a'(\eta) = a''(\eta) = \exp(\eta) = \mu$ som forventet.

2.2 Generaliserede lineære modeller

Definition 2.2 Den generaliserede lineære model er defineret ved

1. Uafhængige responsvariable $Y_1, Y_2, \dots, Y_n$, der alle har den samme type fordeling fra den eksponentielle familie med tætheds-/sandsynlighedsfunktioner på kanonisk form

$$f(y_i|\eta_i, V_i) = \exp\{[y_i\eta_i - a(\eta_i)]/V_i\}b(y_i, V_i). \quad i = 1, 2, \dots, n.$$

2. Vektorer $\mathbf{x}_1, \dots, \mathbf{x}_n$ af forklarende variable, hvor $\mathbf{x}_i \in \mathbb{R}^p$ er tilknyttet observationen Y_i . Disse vektorer organiseres i en $n \times p$ -dimensional matrix,

$$\underline{\underline{X}} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix},$$

der svarer til designmatricen i en lineær normal model.

3. En parametervektor $\boldsymbol{\beta} \in \mathbb{R}^p$, som består af koefficienterne for de forklarende variable.
4. En kendt, monoton og kontinuert link-funktion, g , for hvilken det gælder, at

$$g(\mu_i) = \lambda_i = \mathbf{x}_i^T \boldsymbol{\beta}, \quad (2.5)$$

hvor $\mu_i = E(Y_i)$.

Det ses, at det ikke kan antages at middelværdiparameteren, μ_i , er en lineær funktion af den lineære regressionskomponent, λ_i , som i den lineære normal model.

2.2.1 Poisson-model

Ved modellering af Poisson-data, som ikke kan være negative, benyttes ofte en log-lineær model således at

$$g(\mu_i) = \ln \mu_i = \mathbf{x}_i^T \boldsymbol{\beta} \Leftrightarrow \mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}).$$

Dermed sikres det, at middelværdierne af observationerne kun antager positive værdier.

2.3 Maksimum-likelihood estimation

I dette afsnit beskrives maksimum-likelihood estimation af parametre i den GLM.

Det følger af (2.1), at den simultane log-likelihoodfunktion for $Y_1, Y_2, \dots, Y_n$ er givet ved

$$l(\boldsymbol{\beta}) = \ln \prod_{i=1}^n f(y_i | \eta_i, V_i) = \sum_{i=1}^n \{ [y_i \eta_i - a(\eta_i)] / V_i + \ln b(y_i, V_i) \} = \sum_{i=1}^n l_i(\boldsymbol{\beta}), \quad (2.6)$$

hvor $l_i(\boldsymbol{\beta})$ er log-likelihoodfunktionen for Y_i .

Score-funktionen er givet ved

$$S(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \frac{\partial l_i(\boldsymbol{\beta})}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}.$$

Det følger af (2.6) og (2.2), at

$$\frac{\partial l_i(\boldsymbol{\beta})}{\partial \eta_i} = \frac{y_i - \mu_i}{V_i}.$$

Ligeledes følger det af (2.2) og (2.3), at

$$\frac{\partial \eta_i}{\partial \mu_i} = 1 \Big/ \frac{\partial \mu_i}{\partial \eta_i} = 1 \Big/ \frac{\partial^2 a(\eta_i)}{\partial \eta_i^2} = \frac{V_i}{\text{Var}(Y_i)}.$$

Endelig ses det af (2.5), at

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial \lambda_i} \cdot \frac{\partial \lambda_i}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial \lambda_i} \mathbf{x}_i.$$

Tilsammen giver det, at

$$S(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{y_i - \mu_i}{\text{Var}(Y_i)} \mathbf{x}_i \frac{\partial \mu_i}{\partial \lambda_i}. \quad (2.7)$$

Opfattes (2.7) som en definition på en stokastisk variabel defineret udfra de stokastiske y_i 'er, ses det, at

$$E[S(\beta)] = \mathbf{0}, \quad (2.8)$$

og at

$$\text{Var}[S(\beta)] = \sum_{i=1}^n \frac{\text{Var}(Y_i) \mathbf{x}_i \mathbf{x}_i^T \left(\frac{\partial \mu_i}{\partial \lambda_i} \right)^2}{\text{Var}(Y_i)^2} = \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^T \left(\frac{\partial \mu_i}{\partial \lambda_i} \right)^2}{\text{Var}(Y_i)}.$$

Den negative afledte af $S(\beta)$ kaldes *informationsmatricen* og er givet ved

$$\underline{\underline{\mathcal{I}}}(\beta) = -\frac{\partial S}{\partial \beta} = -\frac{\partial S}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta} = \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^T \left(\frac{\partial \mu_i}{\partial \lambda_i} \right)^2}{\text{Var}(Y_i)} = \text{Var}[S(\beta)]. \quad (2.9)$$

I nogle tilfælde kan maksimum-likelihood estimatoren, $\hat{\beta}$, udledes ved at sættet udtrykket i (2.7) lig med nul og løse ligningen eksplicit, men ofte er det nødvendigt at beregne estimatoren ved hjælp af numeriske metoder. De fleste algoritmer bygger på *Newton's metode*. Lad $\hat{\beta}^{(0)}$ være det indledende gæt af maksimum-likelihood estimatoren for β . Ved en Taylor-udvikling af $S(\beta)$ omkring $\hat{\beta}^{(0)}$ opnås, at

$$S(\beta) \approx S(\hat{\beta}^{(0)}) - \underline{\underline{\mathcal{I}}}(\hat{\beta}^{(0)}) (\beta - \hat{\beta}^{(0)}).$$

Ved at sætte dette udtryk lig med nul og løse for β opnås en revideret værdi af maksimum-likelihood estimatoren

$$\hat{\beta}^{(1)} = \hat{\beta}^{(0)} + \underline{\underline{\mathcal{I}}}^{-1}(\hat{\beta}^{(0)}) S(\hat{\beta}^{(0)}). \quad (2.10)$$

Derefter Taylor-udvikles $S(\beta)$ omkring $\hat{\beta}^{(1)}$ og proceduren gentages. Algoritmen terminerer, hvis

$$\frac{\|\hat{\beta}^{(k)} - \hat{\beta}^{(k-1)}\|}{\|\hat{\beta}^{(k-1)}\|} \leq \varepsilon,$$

for et på forhånd valgt $\varepsilon > 0$. For at gøre algoritmen mere effektiv indføres ofte en konstant, c , således at opdateringsligningen bliver

$$\hat{\beta}^{(k)} = \hat{\beta}^{(k-1)} + c \cdot \underline{\underline{\mathcal{I}}}^{-1}(\hat{\beta}^{(k-1)}) S(\hat{\beta}^{(k-1)}).$$

Der findes forskellige metoder til at finde den optimale værdi af c , som ifølge [7], s. 143 oftest ligger mellem 0 og 1. Hvis c er for stor, er der risiko for, at maksimumspunktet for likelihoodfunktionen bliver overskredet ved opdatering af $\hat{\beta}^{(k)}$. Problemet med at anvende en for lille værdi af c er naturligvis, at algoritmen skal køre længere, inden den konvergerer. Det kan være en god ide at lade værdien af c aftage for hver iteration af algoritmen.

2.3.1 Maximum-likelihood estimation for Poisson-modellen

For Poisson-modellen i afsnit 2.2.1 bliver udtrykket for scorefunktionen

$$S(\beta) = \sum_{i=1}^n (y_i - \mu_i) x_i, \quad (2.11)$$

idet

$$\frac{\partial \mu_i}{\partial \lambda_i} = \frac{\partial \exp(\lambda_i)}{\partial \lambda_i} = \exp(\lambda_i) = \mu_i,$$

og $\text{Var}(Y_i) = \mu_i$. Det følger, at informationsmatricen er givet ved

$$\underline{\underline{J}}(\beta) = \sum_{i=1}^n x_i x_i^T \mu_i. \quad (2.12)$$

Givet et datasæt kan maksimum-likelihood estimatoren udledes ved at anvende Newton's algoritme.

2.3.2 Fordeling af maksimum-likelihood estimator

Ved en Taylor-udvikling af $S(\beta)$ omkring $\hat{\beta}$ opnås, idet $S(\hat{\beta}) = \mathbf{0}$ pr. definition af maksimum-likelihood estimatoren, at

$$S(\beta) \approx S(\hat{\beta}) - \underline{\underline{J}}(\hat{\beta})(\beta - \hat{\beta}) = \underline{\underline{J}}(\hat{\beta})(\hat{\beta} - \beta).$$

Det følger, at

$$\hat{\beta} - \beta \approx \underline{\underline{J}}^{-1}(\hat{\beta}) S(\beta), \quad (2.13)$$

og da $E[S(\beta)] = \mathbf{0}$, følger det umiddelbart, at

$$E(\hat{\beta}) \approx \beta,$$

hvilket betyder, at $\hat{\beta}$ er en *asymptotisk central* estimator for β . Da $\text{Var}[S(\beta)] = \underline{\underline{J}}(\beta)$ ifølge (2.9), og $\underline{\underline{J}}$ er symmetrisk, ses det, at

$$\text{Var}(\hat{\beta} - \beta) \approx \underline{\underline{J}}^{-1}(\hat{\beta}) \text{Var}[S(\beta)] \underline{\underline{J}}^{-1}(\hat{\beta}) \approx \underline{\underline{J}}^{-1}(\hat{\beta}).$$

Idet $\hat{\beta}$ bliver normalfordelt for $n \rightarrow \infty$ ifølge [9] s. 44, følger det, at det approksimativt gælder, at

$$\hat{\beta} \sim \mathcal{N}_p[\beta, \underline{\underline{J}}^{-1}(\hat{\beta})] \quad (2.14)$$

for stort n .

2.4 Inferens

Dette afsnit omhandler hypotese-test, som anvendes i forbindelse med modelmodifikation.

2.4.1 Devians

Det kan undersøges, hvor godt en model beskriver data ved at sammenligne med den *fulde* model. I den fulde model er dataene fittet perfekt, hvilket vil sige, at $\beta_i = \lambda_i = g(y_i)$, for $i = 1, 2, \dots, n$, og $\underline{\underline{X}} = \underline{\underline{I}}_n$. *Likelihood ratio* størrelsen for en model er givet ved

$$R = \frac{L(\hat{\beta}_f)}{L(\hat{\beta})},$$

hvor $\hat{\beta}_f$ angiver maksimum likelihood estimatoren af β i den fulde model. Det ses, at hvis R er tæt på 1, er det en indikation af, at modellen er næsten lige så god som den fulde, mens store værdier er kritiske. *Deviansen* er defineret ved

$$D = 2 \ln R = 2 \left[l(\hat{\beta}_f) - l(\hat{\beta}) \right], \quad (2.15)$$

hvor $l(\hat{\beta}) = \ln L(\hat{\beta})$. I det følgende udledes fordelingen af deviansen. Ligning (2.15) kan omskrives til

$$D = 2 \{ [l(\hat{\beta}_f) - l(\beta_f)] - [l(\hat{\beta}) - l(\beta)] + [l(\beta_f) - l(\beta)] \}. \quad (2.16)$$

Ved en Taylor-udvikling af $l(\beta)$ omkring $\hat{\beta}$ fås det, at

$$\begin{aligned} 2[l(\hat{\beta}) - l(\beta)] &\approx 2 \left\{ l(\hat{\beta}) - \left[l(\hat{\beta}) + \mathbf{S}(\hat{\beta})(\beta - \hat{\beta}) - \frac{1}{2}(\beta - \hat{\beta})^T \underline{\underline{\mathcal{J}}}(\hat{\beta})(\beta - \hat{\beta}) \right] \right\} \\ &= (\hat{\beta} - \beta)^T \underline{\underline{\mathcal{J}}}(\hat{\beta})(\hat{\beta} - \beta). \end{aligned} \quad (2.17)$$

Det følger af (2.14) og [6] s. 8, at

$$(\hat{\beta} - \beta)^T \underline{\underline{\mathcal{J}}}(\hat{\beta})(\hat{\beta} - \beta) = [\hat{\beta} - \mathbf{E}(\hat{\beta})]^T [\text{Var}(\hat{\beta})]^{-1} [\hat{\beta} - \mathbf{E}(\hat{\beta})] \sim \chi_p^2. \quad (2.18)$$

Det følger, at de to første led i firkantede parenteser i (2.16) er χ^2 -fordelt med henholdsvis n og p frihedsgrader. Det tredje led, $v = [l(\beta_f) - l(\beta)]$, er en positiv konstant, der er tæt på nul, hvis modellen er næsten lige så god som den fulde. Hvis de to første led i (2.16) er uafhængige, og differensen mellem dem ikke er negativ, gælder det ifølge [6] s. 10, at

$$D \sim \chi^2(n - p, v),$$

hvor v er en *ikke-centralitets* parameter.

Devians for Poisson-modellen

Det følger af (2.4), at

$$l(\beta) = \sum_{i=1}^n y_i \ln \mu_i - \sum_{i=1}^n \mu_i - \sum_{i=1}^n \ln y_i!.$$

Idet $\hat{\mu}_i$ angiver maksimum-likelihood estimatet for μ_i , ses det, at

$$l(\hat{\beta}) = \sum_{i=1}^n y_i \ln \hat{\mu}_i - \sum_{i=1}^n \hat{\mu}_i - \sum_{i=1}^n \ln y_i!.$$

I den fulde model er $\hat{\mu}_i = y_i$, hvilket vil sige, at

$$l(\hat{\beta}_f) = \sum_{i=1}^n y_i \ln y_i - \sum_{i=1}^n y_i - \sum_{i=1}^n \ln y_i!,$$

idet $0 \cdot \ln 0$ defineres til at være lig med nul. Deviansen er givet ved

$$D = 2[l(\hat{\beta}_f) - l(\hat{\beta})] = 2 \left[\sum_{i=1}^n y_i \ln(y_i/\hat{\mu}_i) - \sum_{i=1}^n (y_i - \hat{\mu}_i) \right] \sim \chi_{n-p}^2. \quad (2.19)$$

2.4.2 Test baseret på deviansen

Det kan undersøges om en model, m_1 , er acceptabel i forhold til den fulde model ved at udregne deviansen for modellen. Hypotesen, der testes, er

$$H_1 : \quad \beta = \beta_1,$$

hvor $\beta_1 \in \mathbb{R}^p$.

Ud fra χ^2 fordelingen og det valgte signifikansniveau, α , kan det bestemmes, hvilke værdier af deviansen der er kritiske for hypotesen. Hvis der fås en værdi, hvor sandsynligheden for at få en værdi, der er større, er mindre end α , er det en indikation af, at hypotesen er falsk.

Antag at H_1 accepteres. Det kan testes, om m_1 kan reduceres til en simplere model, m_0 , hvor m_0 er en undermodel af m_1 . I dette tilfælde betragtes hypotesen

$$H_0 : \quad \beta = \beta_0,$$

hvor $\beta_0 \in \mathbb{R}^q$, $q < p$, i forhold til hypotesen H_1 . Dette gøres ved at beregne forskellen i maksimum-likelihood mellem de to modeller. Givet at devianserne D_1 og D_0 er uafhængige, gælder der, at

$$2[l(\hat{\beta}_1) - l(\hat{\beta}_0)] = 2[l(\hat{\beta}_f) - l(\hat{\beta}_0)] - 2[l(\hat{\beta}_f) - l(\hat{\beta}_1)] = D_0 - D_1 \sim \chi_{p-q}^2.$$

2.4.3 Wald-test

I afsnit 2.4.2 blev det undersøgt om en model generelt kan reduceres, hvilket betyder, at en eller flere forklarende variable fjernes fra modellen. Dette svarer til at teste, om de tilhørende koefficienter kan sættes lig med nul. Nulhypotesen bliver

$$H_0 : \quad \underline{\underline{C}}\beta = \mathbf{0},$$

hvor $\underline{\underline{C}}$ har fuld række-rang, $s \leq p$, ifølge [9] s. 47. Hvis det skal testes, om $\beta_k = 0$, består $\underline{\underline{C}}$ af ene nuller undtagen i den k 'te søjle, som består af et-taller. Under H_0 er

$$\underline{\underline{C}}\hat{\beta} \sim \mathcal{N}_s[\mathbf{0}, \underline{\underline{C}}\mathcal{I}^{-1}(\hat{\beta})\underline{\underline{C}}^T].$$

Det følger af 2.18, at det approksimativt gælder, at

$$(\underline{\underline{C}}\hat{\beta})^T (\underline{\underline{C}}\mathcal{I}^{-1}(\hat{\beta})\underline{\underline{C}}^T)^{-1} (\underline{\underline{C}}\hat{\beta}) \sim \chi_s^2,$$

Ovenstående er en kvadratisk approksimation af devians-teststørrelsen og benævnes *Wald-teststørrelsen*.

En mere generel definition af Wald-teststørrelsen er givet ved

$$(\underline{\underline{C}}\hat{\beta} - \xi)^T (\underline{\underline{C}}\mathcal{I}^{-1}(\hat{\beta})\underline{\underline{C}}^T)^{-1} (\underline{\underline{C}}\hat{\beta} - \xi) \sim \chi_s^2, \quad (2.20)$$

hvor $\xi \in \mathbb{R}^s$. I dette tilfælde testes hypotesen

$$H : \quad \underline{\underline{C}}\beta = \xi.$$

En yderligere anvendelse af Wald-testen vil blive beskrevet i afsnit 6.4.2.

2.5 Modelkontrol

Modelkontrollen har til formål at undersøge, om de antagelser, modellen er bygget på, er overholdt. I dette afsnit vil forskellige metoder til modelkontrol blive beskrevet.

2.5.1 Standardiserede residualer

Betragt en række observationer $Y_1, Y_2, \dots, Y_n$, hvor Y_i har middelværdi, μ_i , og varians, σ_i^2 . Da er de *standardiserede residualer* givet ved

$$r_i = \frac{y_i - \mu_i}{\sqrt{\sigma_i^2}}.$$

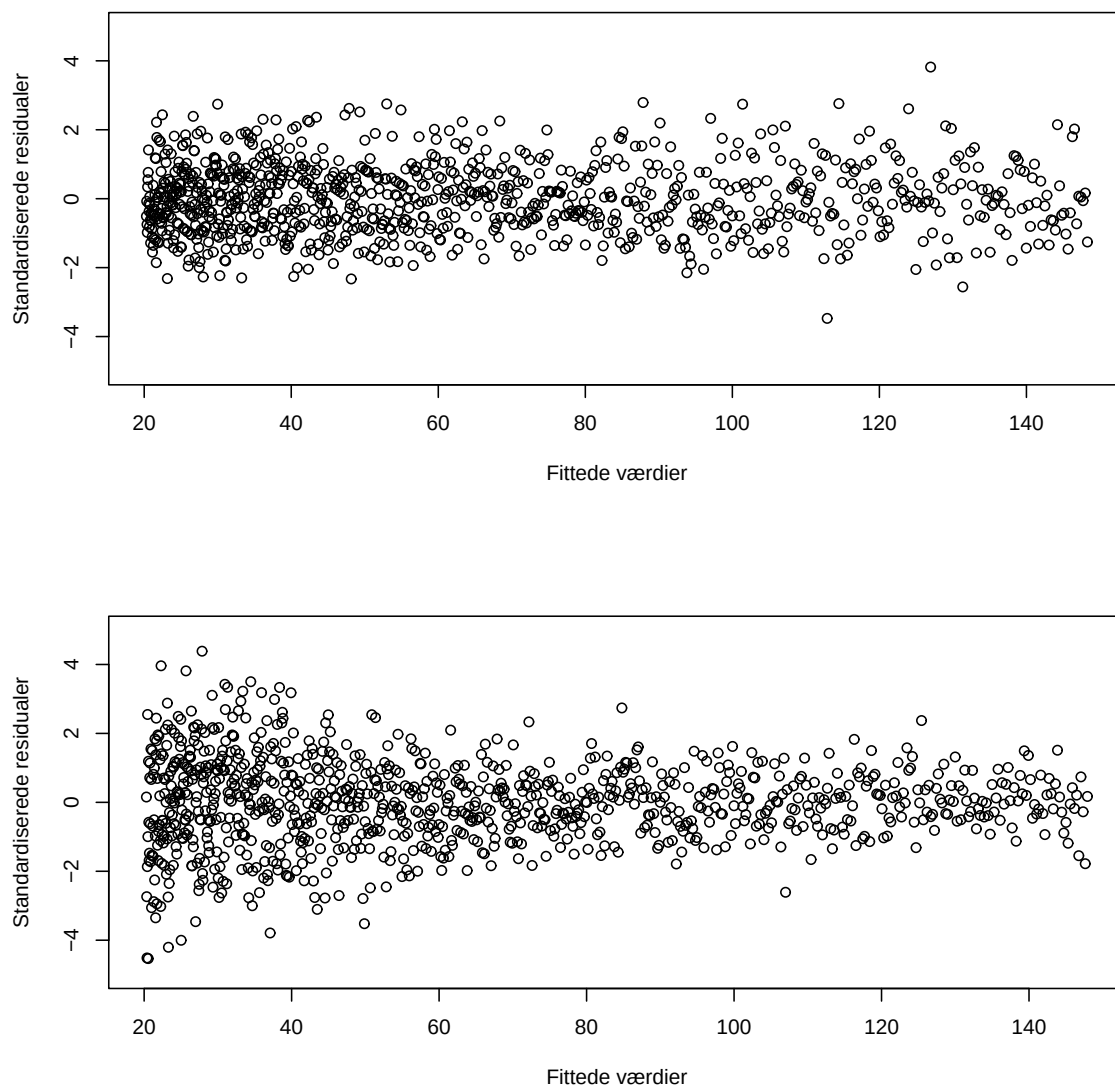
De kan estimeres ved

$$\hat{r}_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\sigma}_i^2}},$$

hvor $\hat{\mu}_i$ og $\hat{\sigma}_i^2$ er maksimum-likelihood estimerne for henholdsvis middelværdien og variansen af Y_i . Hvis modellen er passende, og n er stor, er de estimerede standardiserede residualer tilnærmelsesvis uafhængige, og de har approksimativt middelværdi 0 og varians 1. Plots af de estimerede standardiserede residualer kan anvendes på flere måder i forbindelse med modelkontrol, hvilket vil blive beskrevet og illustreret i det følgende.

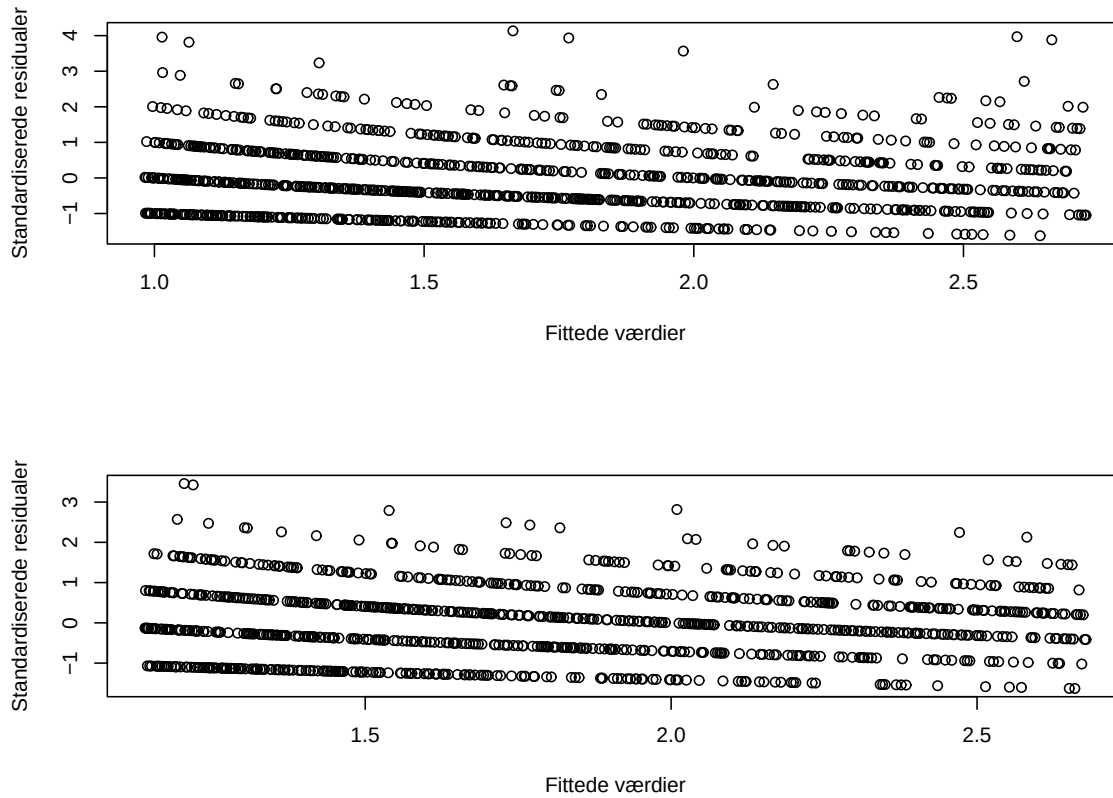
2.5.2 Varianshomogenitet

Det kan undersøges, om de standardiserede residualer har en homogen varians på approksimativt 1 ved at plotte $\hat{\mu}_i$ mod $\hat{r}_i$. Følgende simulering viser et tilfælde af inhomogen varians. Betragt to datasæt med samme middelværdier men med henholdsvis Poisson variable, hvor variansen er lig med middelværdien, og afrundede normalfordelte variable, hvor variansen er konstant. Dette giver ved tilpasning af en Poisson model anledning til følgende to residualplots, hvoraf det baseret på Poisson variable viser et plot med homogen varians, mens det andet viser en inhomogen, "tragtformet" varians, idet variansen af residualerne bliver mindre i takt med, at de forventede værdier vokser.



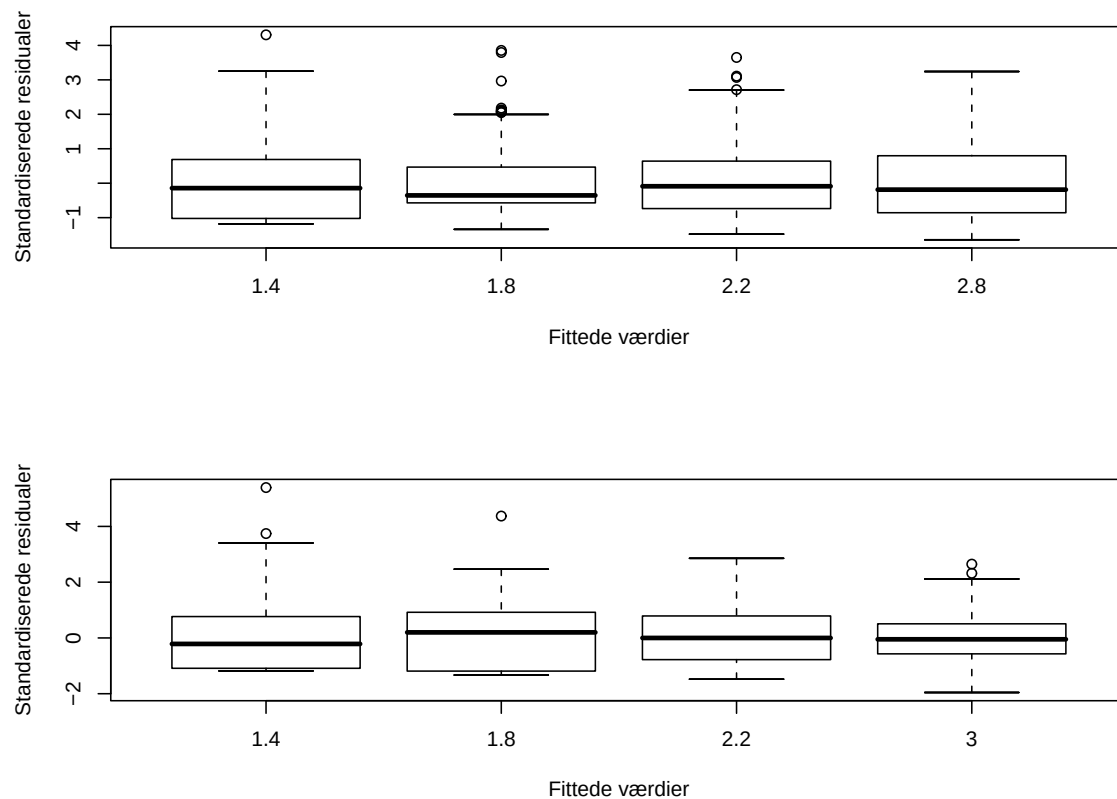
Figur 2.1: Residualplot af to datasæt fittet ved en Poisson-model. Øverst ses et plot af standardiserede residualer for Poisson-fordelte observationer, mens det nederste plot viser standardiserede residualer for normalfordelte observationer.

I tilfælde af observationer med små middelværdier kan det være svært at kontrollere for inhomogen varians, hvilket er illustreret i figur 2.2.



Figur 2.2: Residualplot af Poisson-fordelte observationer (øverst) og normalfordelte observationer (nederst) med lille intensitet fittet ved en Poisson-model.

I tilfælde af en lille intensitet kan et *boxplot*, hvor residualerne inddeles i bokse, være behjælpeligt. Linien, der går igennem boksene, angiver medianen, mens boksens afgrænsninger svarer til første og tredje kvartil af dataene. De to udlæggere fra boksen har en længde på højst halvanden gange boksens længde og definerer omtrent 95% konfidensintervallet for dataene. Boxplottet har den fordel, at det skjuler den fjerlignende struktur i scatterplottet, der opstår ved gruppering af residualer, der svarer til observerede nuller, ettere o.s.v. Herved forhindres det, at øjemålet snydes.



Figur 2.3: Boxplot af standardiserede residualer fra Poisson-fordelte observationer (øverst) og fra normalfordelte observationer (nederst) med lille intensitet fittet ved en Poisson-model. Boksen, der svarer til værdien 2.2 på x -aksen, består af residualer for observationer med fittede værdier i intervallet 1.8-2.2.

Boxplottet i figur 2.3 (nederst) viser, at variansen af de standardiserede residualer for de normalfordelte observationer bliver mindre i takt med, at intensiteten stiger.

2.5.3 Seriel korrelation

En tidsrække består af en række observationer $Y_1, Y_2, \dots$, hvor Y_t angiver observationen til tiden t . Ved tidsrække-data kan det undersøges, hvorvidt antagelsen om ukorrelerede residualer er overholdt ved at udregne *autokorrelationen* for de standardiserede residualer. Dette er dog besværliggjort ved,

at de estimerede residualer aldrig er helt ukorrelerede, hvilket skyldes, at de fittede værdier hver især er beregnet ud fra alle observationerne og derfor er korrelerede. Denne korrelation er dog forsvindende for et stort datasæt. Den serielle korrelation mellem residualerne er givet ved

$$\rho(k) = \text{Cor}(r_t, r_{t+k}) = \frac{\text{Cov}(r_t, r_{t+k})}{\sqrt{\text{Var}(r_t)\text{Var}(r_{t+k})}}$$

og er naturligvis lig med nul i teorien. Den k 'te *empiriske autokovarianskoefficient* for en stikprøve, $y_1, y_2, \dots, y_n$, er givet ved

$$\hat{g}(k) = \sum_{t=1}^{n-k} (\hat{r}_t - \bar{r})(\hat{r}_{t+k} - \bar{r}) / (n - k),$$

hvor $\bar{r}$ er den empiriske middelværdi, $\sum_{t=1}^n \hat{r}_t / n$. Den k 'te *empiriske autokorrelationskoefficient* er da givet ved

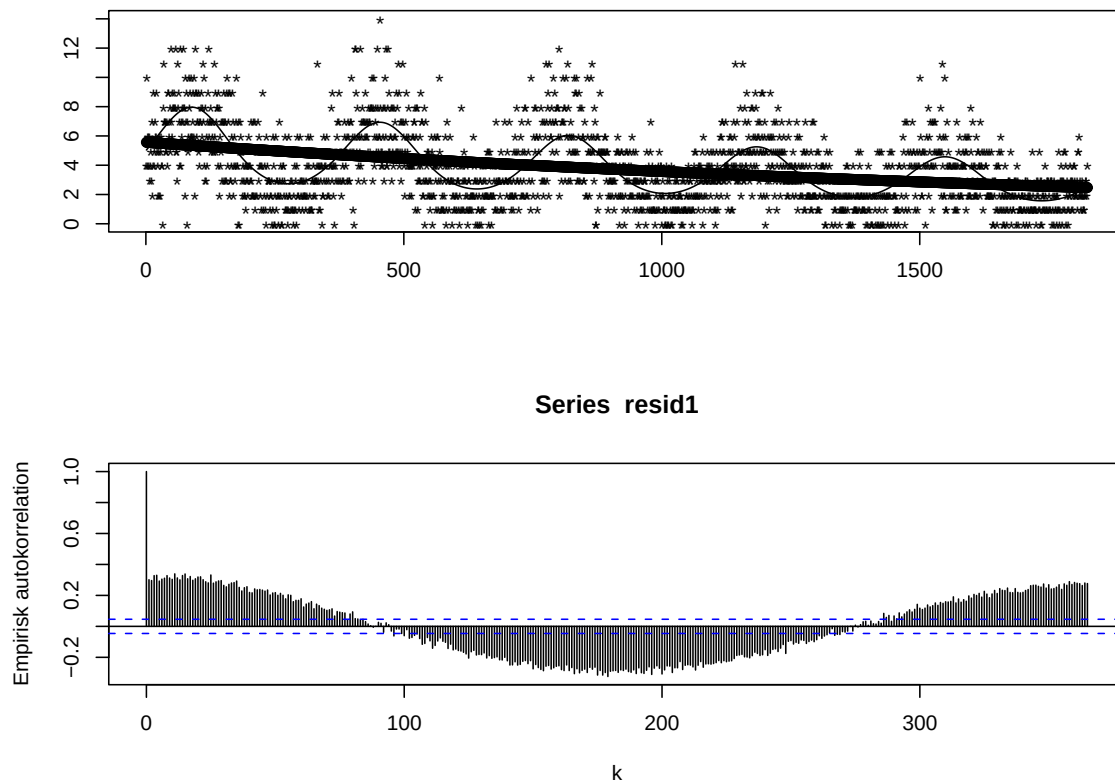
$$\hat{\rho}(k) = \hat{g}_k / \hat{g}_0,$$

idet $\hat{g}_0$ er estimeret for $\text{Var}(r_t)$ for alle t .

Ifølge [5] s. 39 gælder det approksimativt for stort n og for normalfordelte observationer, at $\hat{\rho}(k) \sim \mathcal{N}(0, 1/n)$ under antagelse af, at residualerne er uafhængige. Det følger, at 95% konfidensintervallet for $\rho(k)$ approksimativt er givet ved

$$(-1.96\sqrt{1/n}, 1.96\sqrt{1/n}).$$

På figur 2.4 øverst ses et plot af en simuleret tidsrække af en Poisson-fordeling, hvor logaritmen til intensiteten følger en sinuskurve. Observationerne er fittet ved en log-lineær funktion af tiden. På figur 2.4 nederst ses det, at den serielle korrelation er signifikant.



Figur 2.4: Øverst ses Poisson-fordelte observationer med forventede værdier, der følger en sinuskurve med periode 365 på log-skalaen, fittet ved en log-lineær funktion af tiden. Nederst ses den empiriske autokorrelation for residualerne. De stiplede linier angiver 95%-konfidensintervallet for autokorrelationen.

Udseendet af autokorrelationsplottet i figur 2.4, skyldes, at langt hovedparten af de residualer, der ligger tæt på eller en sinus-svingning fra hinanden, har samme fortegn, hvilket medfører, at de er positivt korrelerede. Modsvarende har residualerne, der ligger en halv sinus-svingning fra hinanden, for det meste modsat fortegn og er dermed negativt korrelerede.

I dette tilfælde skyldes den serielle korrelation ikke, at observationerne i sig selv er korrelerede men nærmere, at den valgte model giver en dårlig repræsentation af data. Serielt korrelerede residualer kan således skyldes en afhængighed mellem responsvariablen og en forklarende variabel, der ikke er erkendt eller målt. I dette tilfælde synes det at være en sæsonvariation i dataene, der ikke er erkendt.

2.5.4 Fordelingstype

I dette afsnit beskrives metoder til at kontrollere, at observationernes fordeling er korrekt specificeret.

Definition 2.3 Den empiriske fordelingsfunktion for uafhængige og ensfordelte stokastiske variable $X_1, X_2, \dots, X_n$ er givet ved

$$F_e(x) = \frac{1}{n} \sum_{i=1}^n 1_{X_i}(x),$$

hvor

$$1_{X_i}(x) = \begin{cases} 1 & \text{for } X_i \leq x \\ 0 & \text{ellers} \end{cases}$$

Sætning 2.4 For uafhængige stokastiske variable $X_1, X_2, \dots, X_n$ med samme fordelingsfunktion $F(x)$ gælder det, at

$$\lim_{n \rightarrow \infty} F_e(x) = F(x).$$

Bevis:

Det følger af definition 2.3, at fordelingen af $1_{X_i}(x)$ er givet ved

$$\begin{aligned} P[1_{X_i}(x) = 0] &= P(X_i > x) = 1 - P(X_i \leq x) = 1 - F(x) \\ P[1_{X_i}(x) = 1] &= P(X_i \leq x) = F(x). \end{aligned}$$

Det følger, at

$$E[1_{X_i}(x)] = 0 \cdot [1 - F(x)] + 1 \cdot F(x) = F(x).$$

Da dette gælder for alle $i \in \{1, 2, \dots, n\}$, følger det af (A.4), at

$$\lim_{n \rightarrow \infty} [P(|F_e(x) - F(x)| < \varepsilon)] = \lim_{n \rightarrow \infty} \left[P\left(\left|\frac{1}{n} \sum_{i=1}^n 1_{X_i}(x) - F(x)\right| < \varepsilon\right) \right] = 1$$

for ethvert $\varepsilon > 0$. Dette er ækvivalent med, at

$$\lim_{n \rightarrow \infty} F_e(x) = F(x).$$

■

Det følger af sætning 2.4, at en metode til kontrol af fordelingstype er at se på de observationer, der har ca. samme middelværdi og varians og kontrollere, om de empiriske middelværdier og varianser svarer

dertil. Det svarer til at bestemme den empiriske middelværdi af de observationer, $y_1, y_2, \dots, y_k$, for hvilke der gælder, at

$$|\hat{\mu}_t - \xi| \leq \varepsilon, \quad t = 1, 2, \dots, k, \quad (2.21)$$

for på forhånd valgte $\varepsilon > 0$ og $\xi \in \mathbb{R}$, under antagelse af at observationerne har approksimativt ens varians, σ^2 . Disse observationer kan antages approksimativt at have samme middelværdi, ξ , og det følger, at den empiriske middelværdi af y 'erne skal ligge tæt på ξ , og at den empiriske varians skal ligge tæt på σ^2 , for stort k .

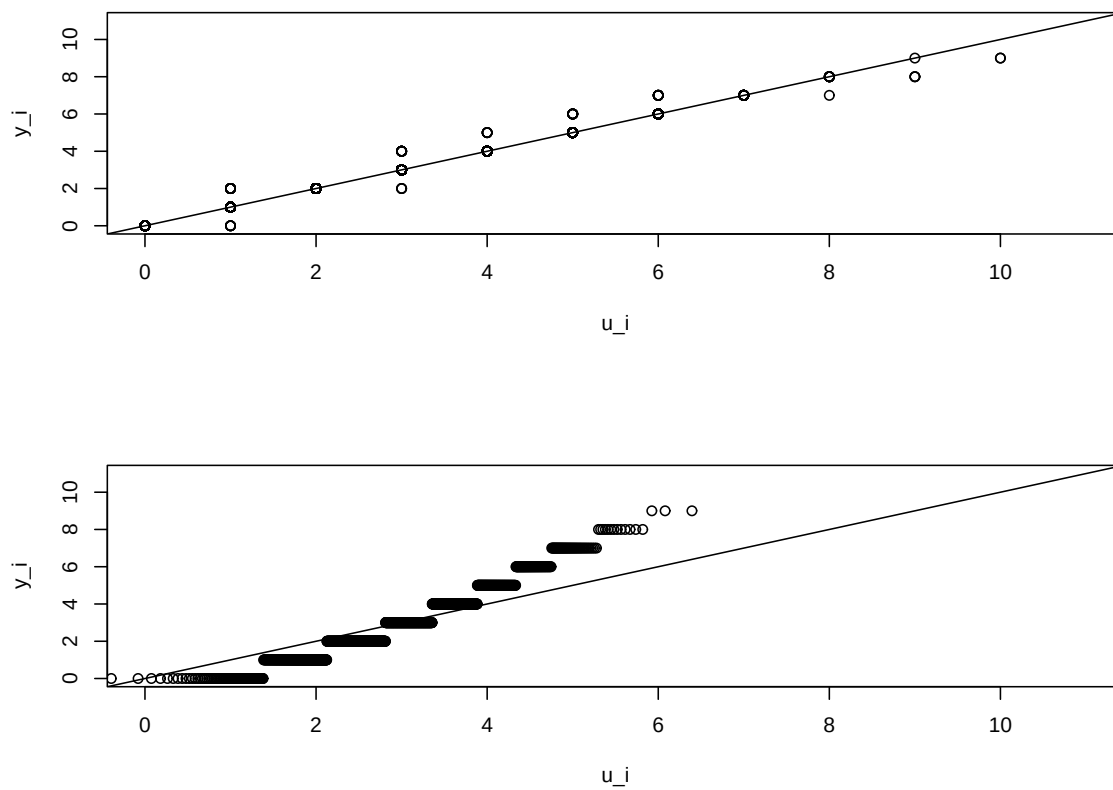
En metode til yderligere kontrol er at foretage et såkaldt *fraktilplot*. Her kontrolleres ikke blot middelværdi og varians men også om den valgte type af fordeling er korrekt.

Lad observationerne $y_1, y_2, \dots, y_k$ opfylde (2.21) for på forhånd valgte ε og ξ . Ved at ordne y_i 'erne efter størrelse opnås at $F_e(y_i) = i/k$. Dermed skal det ifølge sætning 2.4 gælde for stort k , at

$$F(y_i) \approx \frac{i}{k} \Leftrightarrow y_i \approx F^{-1}\left(\frac{i}{k}\right) = u_i,$$

hvor u_i benævner i/k -fraktilen for fordelingen af y_i .

Derved følger det, at et plot af y_i mod u_i skal følge en ret linie med skæring i origo og hældning 1, hvis fordelingen af y_i 'erne er korrekt specificeret. Da $F^{-1}(1) = \infty$, sættes u_i ofte lig med $F^{-1}[(i - 0.5)/k]$. I figur 2.5 øverst ses et fraktilplot af Poisson-fordelte observationer, der er korrekt specificeret. Det nederste plot i figur 2.5 viser et fraktilplot af Poisson-fordelte observationer, der er fejlspecificeret som normalfordelte.



Figur 2.5: Poisson-fraktilplot (øverst) og normal-fraktilplot (nederst) af simulerede Poisson-fordelte observationer.

Kapitel 3

Den Dynamiske Lineære Model

I dette kapitel beskrives den dynamiske lineære model, Kalman-filtrering og Kalman-udglatning samt estimation af ukendte parametre. Desuden beskrives strukturelle tidsrække modeller, som er et special-tilfælde af den dynamiske lineære model.

3.1 Den dynamiske lineære model

En tidsrække består af en serie observationer, $Y_1, Y_2, \dots$, hvor Y_t angiver observationen til tiden t . En tidsrække beskrives sædvanligvis ikke af egentlige forklarende variable men af uobserverede komponenter som trend og sæsonvariation. En sådan model kunne have udseendet

$$Y_t = \mu_t + \gamma_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2),$$

hvor μ_t angiver trenden, γ_t sæsonvariationen og ε_t er et støjled med konstant varians, σ^2 .

Ved analyse af tidsrække-data er det ikke altid tilfredsstillende at anvende en model, hvor komponenterne som trend og sæsonvariation er deterministiske funktioner af tiden. En mere fleksibel og dynamisk model kan opnås ved at lade disse komponenter være stokastiske. Det vil sige, at f.eks. sæsonvariationen ændres gradvist på stokastisk vis fra periode til periode. Denne type modeller, hvor komponenterne er stokastiske, kaldes *dynamiske lineære modeller* (DLM) eller *state space* modeller. I en DLM kædes observationerne sammen med en ikke-observeret process, $\theta_0, \theta_1, \dots$, via en *observationsligning*, en *tilstandsligning* samt noget *startinformation*.

Definition 3.1 Den en-dimensionelle DLM er defineret ved

$$\text{Observationsligning: } Y_t = \mathbf{F}_t^T \boldsymbol{\theta}_t + \nu_t, \quad \nu_t \sim \mathcal{N}(0, V_t)$$

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

$$\text{Startinformation: } (\boldsymbol{\theta}_0 | D_0) \sim \mathcal{N}_p(\mathbf{m}_0, \underline{\underline{C}}_0),$$

hvor elementerne i ν_t og $\boldsymbol{\omega}_t$ er både serielt og indbyrdes uafhængige samt uafhængige af $(\boldsymbol{\theta}_0 | D_0)$.

Informationsmængden til tiden t er givet rekursivt ved

$$D_t = \{D_{t-1}, Y_t\},$$

hvor D_0 er informationsmængden bestående af $\mathbf{F}_t$, $\underline{\underline{G}}_t$, V_t og $\underline{\underline{W}}_t$ for alle t .

Elementerne i modellen er kort beskrevet i nedenstående tabel.

Element	Beskrivelse
Y_t	Observation
$\boldsymbol{\theta}_t$	$1 \times p$ tilstandsvektor
$\mathbf{F}_t$	$1 \times p$ designvektor
$\underline{\underline{G}}_t$	$p \times p$ transitionsmatrix
ν_t	Observationsstøj
$\boldsymbol{\omega}_t$	$1 \times p$ systemfejlsvektor
V_t	Observationsvarians
$\underline{\underline{W}}_t$	$p \times p$ systemvariansmatrix
$\mathbf{m}_t$	$1 \times p$ middelværdivektor
$\underline{\underline{C}}_t$	$p \times p$ variansmatrix

Det antages indtil videre, at $\mathbf{F}_t$, $\underline{\underline{G}}_t$, V_t og $\underline{\underline{W}}_t$ er kendte.

Startinformationen angiver fordelingen af $\boldsymbol{\theta}_t$ til tiden $t = 0$. Dette indikerer, at Y_1 ikke skal ses som begyndelsen på tidsrækken men derimod som fortsættelsen af en tidligere observeret tidsrække. I tilfælde af, at Y_1 rent faktisk er den første tilgængelige observation, er situationen mere vanskelig. I [25], hvor den DLM behandles ud fra en *Bayesiansk* tilgang, anvendes notationen $(\boldsymbol{\theta}_1 | D_0)$ til at benævne a priori fordelingen af den første tilstandsvektor baseret på forudgående viden, der ikke inkluderer tidligere observationer. Set fra en ikke-Bayesiansk/klassisk synsvinkel er det, at $\boldsymbol{\theta}_1$ kan have en fordeling, inden der er observeret nogen observationer meget kontroversielt. I [7], kapitel 5, beskrives metoder til at løse initialiseringsproblemet ud fra en klassisk synsvinkel.

Udfra definition 3.1 ses det, at θ_t kan skrives ud som en linearkombination af systemfejlsvektorer og θ_0 , som alle er multivariabelt normalfordelte. Det ses deraf, at θ_t også er multivariabelt normalfordelt. Det følger, at Y_t er normalfordelt, da den er givet ved en linearkombination af normalfordelte størrelser. Enhver linearkombination af elementerne i Y_t og θ_t er normalfordelt, og deraf følger det, at den fælles fordeling af Y_t og θ_t er multivariabelt normalfordelt. Dette vil blive anvendt i afsnit 3.3.

3.2 Betinget uafhængighed i den DLM

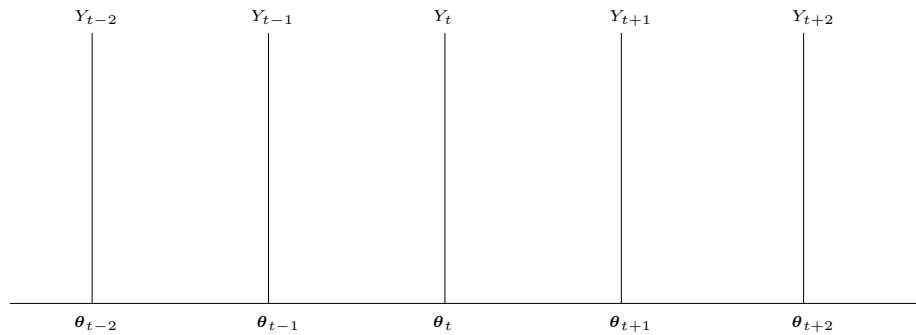
Det ses udfra definition 3.1, at

$$(Y_t | \theta_0, \theta_1, \dots, \theta_t, \dots, Y_1, Y_2, \dots, Y_{t-1}, Y_{t+1}, \dots) \sim \mathcal{N}(\mathbf{F}_t^T \theta_t, V_t), \quad (3.1)$$

og at

$$(\theta_t | \theta_0, \dots, \theta_{t-1}, Y_1, \dots, Y_{t-1}) \sim \mathcal{N}_p(\underline{\underline{G}}_t \theta_{t-1}, \underline{\underline{W}}_t). \quad (3.2)$$

Det følger, at givet θ_t er Y_t uafhængig af θ_s for $s \neq t$, og af Y_r for $r \neq t$, samt at givet θ_{t-1} er θ_t uafhængig af θ_s for $s < t - 1$ og af Y_r for $r < t$. Dette motiverer, at den betingede uafhængighedsstruktur i en DLM kan beskrives ved en graf tilsvarende interaktionsgrafen for en *grafisk model*, se figur 3.1. For mere information om grafiske modeller, se [17].



Figur 3.1: Den betingede uafhængighedsstruktur i en DLM.

3.3 Kalman-filtrering

Formålet med *Kalman-filtreret* er at opdatere fordelingen af tilstandsvektoren i takt med, at nye observationer bliver tilgængelige.

Sætning 3.2 I en en-dimensionel DLM gælder følgende fordelinger.

$$\text{a priori til } t: \quad (\boldsymbol{\theta}_t | D_{t-1}) \sim \mathcal{N}_p(\mathbf{a}_t, \underline{\underline{R}}_t), \quad (3.3)$$

$$\text{hvor} \quad \mathbf{a}_t = \underline{\underline{G}}_t \mathbf{m}_{t-1} \quad \text{og} \quad \underline{\underline{R}}_t = \underline{\underline{G}}_t \underline{\underline{C}}_{t-1} \underline{\underline{G}}_t^T + \underline{\underline{W}}_t.$$

$$\text{Et-trins forudsigelse:} \quad (Y_t | D_{t-1}) \sim \mathcal{N}(f_t, Q_t), \quad (3.4)$$

$$\text{hvor} \quad f_t = \mathbf{F}_t^T \mathbf{a}_t \quad \text{og} \quad Q_t = \mathbf{F}_t^T \underline{\underline{R}}_t \mathbf{F}_t + V_t.$$

$$\text{a posteriori til } t: \quad (\boldsymbol{\theta}_t | D_t) \sim \mathcal{N}_p(\mathbf{m}_t, \underline{\underline{C}}_t), \quad (3.5)$$

$$\text{hvor} \quad \mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t e_t, \quad \underline{\underline{C}}_t = \underline{\underline{R}}_t - \mathbf{A}_t Q_t \mathbf{A}_t^T,$$

$$\mathbf{A}_t = \underline{\underline{R}}_t \mathbf{F}_t / Q_t, \quad \text{og} \quad e_t = Y_t - f_t.$$

Bevis:

Beviset er baseret på [25] s. 104-105.

Det vises ved induktion, at (3.5) er opfyldt, hvorved fordelingerne i (3.3) og (3.4) fremkommer som mellemresultater.

Tilfældet $t = 0$ er opfyldt ud fra definition 3.1. Antag at

$$(\boldsymbol{\theta}_{t-1} | D_{t-1}) \sim \mathcal{N}_p(\mathbf{m}_{t-1}, \underline{\underline{C}}_{t-1}). \quad (3.6)$$

Det følger, at

$$E(\boldsymbol{\theta}_t | D_{t-1}) = E(\underline{\underline{G}}_t \boldsymbol{\theta}_{t-1} | D_{t-1}) + E(\boldsymbol{\omega}_t | D_{t-1}) = \underline{\underline{G}}_t \mathbf{m}_{t-1},$$

og at

$$\text{Var}(\boldsymbol{\theta}_t | D_{t-1}) = \text{Var}(\underline{\underline{G}}_t \boldsymbol{\theta}_{t-1} | D_{t-1}) + \text{Var}(\boldsymbol{\omega}_t | D_{t-1}) = \underline{\underline{G}}_t \underline{\underline{C}}_{t-1} \underline{\underline{G}}_t^T + \underline{\underline{W}}_t,$$

da $\boldsymbol{\theta}_{t-1}$ og $\boldsymbol{\omega}_t$ er uafhængige givet D_{t-1} . Da $\boldsymbol{\theta}_t$ og ν_t ligeledes er uafhængige givet D_{t-1} , udledes fordelingen i (3.4) på tilsvarende vis.

Da

$$\text{Cov}(Y_t, \boldsymbol{\theta}_t | D_{t-1}) = \text{Cov}(\mathbf{F}_t^T \boldsymbol{\theta}_t + \nu_t, \boldsymbol{\theta}_t | D_{t-1}) = \text{Var}(\boldsymbol{\theta}_t | D_{t-1}) \mathbf{F}_t = \underline{\underline{R}}_t \mathbf{F}_t = Q_t \mathbf{A}_t^T,$$

gælder det, at den fælles fordeling for $\boldsymbol{\theta}_t$ og Y_t givet D_{t-1} er givet ved

$$\begin{pmatrix} \boldsymbol{\theta}_t \\ Y_t \end{pmatrix} | D_{t-1} \sim \mathcal{N}_{p+1} \left[\begin{pmatrix} \mathbf{a}_t \\ f_t \end{pmatrix}, \begin{pmatrix} \underline{\underline{R}}_t & \mathbf{A}_t Q_t \\ Q_t \mathbf{A}_t^T & Q_t \end{pmatrix} \right].$$

Udfra den fælles fordeling kan den marginale fordeling for θ_t givet Y_t og D_{t-1} bestemmes ved hjælp af sætning A.1, og (3.5) fremkommer.

Bevis af (3.5) ved hjælp af Bayes' sætning.

Set udfra et Bayesiansk synspunkt er det mere naturligt at opdatere fordelingen af θ_t ved hjælp af Bayes' sætning. Givet to stokastiske variable, X og Y , siger Bayes' sætning, at

$$f(X|Y) = \frac{f(Y|X)f(X)}{f(Y)}.$$

Betragtet som en funktion af X alene opnås at

$$f(X|Y) \propto f(Y|X)f(X).$$

Det følger, at

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

Det vil sige, at a posteriori tætheden af θ_t er et produkt af to Gaussiske tætheder og er dermed selv Gaussisk. Middelværdi og varians for a posteriori fordelingen af θ_t mangler at blive udledt. Det følger af (3.3), at

$$f(\theta_t|D_{t-1}) \propto \exp[-(\theta_t - \mathbf{a}_t)^T \underline{\underline{R}}_t^{-1}(\theta_t - \mathbf{a}_t)/2].$$

Desuden ses det af definition 3.1, at

$$f(Y_t|\theta_t) \propto \exp[-(Y_t - \mathbf{F}_t^T \theta_t)^2/2V_t].$$

Det følger, at

$$-2 \ln f(\theta_t|D_t) = (\theta_t - \mathbf{a}_t)^T \underline{\underline{R}}_t^{-1}(\theta_t - \mathbf{a}_t) + (Y_t - \mathbf{F}_t^T \theta_t)^2/V_t + k_1,$$

hvor k_1 er en konstant. Betragtet som en funktion af θ_t kan ovenstående udtryk omskrives til

$$-2 \ln f(\theta_t|D_t) = \theta_t^T (\underline{\underline{R}}_t^{-1} + \mathbf{F}_t \mathbf{F}_t^T / V_t) \theta_t - 2 \theta_t^T (\underline{\underline{R}}_t^{-1} \mathbf{a}_t + \mathbf{F}_t Y_t / V_t) + k_2. \quad (3.7)$$

For en normalfordelt vektor $\mathbf{Y}$ med middelværdi, $\boldsymbol{\mu}$, og variansmatrix, $\underline{\underline{V}}$, gælder der, at

$$-2 \ln f(\mathbf{Y}) = (\mathbf{Y} - \boldsymbol{\mu})^T \underline{\underline{V}}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) + c_1 = \mathbf{Y}^T \underline{\underline{V}}^{-1} \mathbf{Y} - 2 \mathbf{Y}^T \underline{\underline{V}}^{-1} \boldsymbol{\mu} + c_2, \quad (3.8)$$

hvor c_1 og c_2 er konstanter. Det følger derfor af (3.7), at

$$\text{Var}(\theta_t|D_t) = \underline{\underline{C}}_t = (\underline{\underline{R}}_t^{-1} + \mathbf{F}_t \mathbf{F}_t^T / V_t)^{-1} = \underline{\underline{R}}_t - \mathbf{A}_t \mathbf{Q}_t \mathbf{A}_t^T,$$

og at

$$\underline{\underline{C}}_t^{-1} \mathbb{E}(\boldsymbol{\theta}_t | D_t) = \underline{\underline{R}}_t^{-1} \mathbf{a}_t + \mathbf{F}_t Y_t / V_t,$$

hvoraf det følger, at

$$\mathbb{E}(\boldsymbol{\theta}_t | D_t) = \underline{\underline{C}}_t \underline{\underline{R}}_t^{-1} \mathbf{a}_t + \underline{\underline{C}}_t \mathbf{F}_t Y_t / V_t = \mathbf{a}_t + \mathbf{A}_t e_t.$$

■

Størrelsen $e_t = Y_t - f_t$ benævnes *et-trins prediktionsfejlen* for observationen Y_t . Der gælder, at e_t er normalfordelt med middelværdi

$$\mathbb{E}(e_t) = \mathbb{E}[\mathbb{E}(e_t | D_{t-1})] = \mathbb{E}[\mathbb{E}(\mathbf{F}_t^T \boldsymbol{\theta}_t + \nu_t - \mathbf{F}_t^T \mathbf{a}_t | D_{t-1})] = \mathbb{E}(\mathbf{F}_t^T \mathbf{a}_t - \mathbf{F}_t^T \mathbf{a}_t) = 0,$$

og varians

$$\text{Var}(e_t) = \text{Var}(e_t | D_{t-1}) = \text{Var}(Y_t - f_t | D_{t-1}) = \text{Var}(Y_t | D_{t-1}) = Q_t.$$

Følgende resultat er nyttigt i forbindelse med modelkontrol.

Korollar 3.3 *Elementerne i følgen af et-trins prediktionsfejl $\{e_t\}$ er indbyrdes uafhængige.*

Bevis:

Der gælder, at $e_t = Y_t - \mathbb{E}(Y_t | D_{t-1})$ er en linearkombination af $Y_1, \dots, Y_t$. Ifølge korollar A.2 er e_t uafhængig af $Y_1, Y_2, \dots, Y_{t-1}$, hvoraf det følger, at e_t er uafhængig af $e_1, e_2, \dots, e_{t-1}$. ■

Størrelsen $\mathbf{A}_t$ betegnes den *adaptive* vektor til tiden t . Da $\mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t e_t$, ses det, at størrelsen af elementerne i $\mathbf{A}_t$ afgør, hvor stor indflydelse prediktionsfejlen har på a posteriori estimatet for $\boldsymbol{\theta}_t$ i forhold til a priori estimatet for tilstandsvektoren. Det ses, at

$$\mathbf{A}_t = \underline{\underline{R}}_t \mathbf{F}_t / Q_t = \underline{\underline{R}}_t \mathbf{F}_t / (\mathbf{F}_t^T \underline{\underline{R}}_t \mathbf{F}_t + V_t). \quad (3.9)$$

Udtrykket viser, at jo større observationsvariansen V_t er i forhold til varianserne og kovarianserne i $\underline{\underline{R}}_t$, jo mindre bliver elementerne i $\mathbf{A}_t$, hvilket vil sige, at indflydelsen af prediktionsfejlen på a posteriori estimatet for $\boldsymbol{\theta}_t$ bliver mindre. Dette stemmer godt overens med den intuitive forestilling, at jo mere usikker observationen Y_t er i forhold til a priori estimatet af tilstandsvektoren, jo mindre indflydelse får Y_t på a posteriori estimatet for tilstandsvektoren.

3.4 Kalman-udglatning

Ofte er det hensigtsmæssigt at gå tilbage i tiden og estimere tidligere systemvektorer på ny. Dette gælder f.eks. i tilfælde, hvor effekten af en indtruffet hændelse eller sæsonvariationen for dataene ønskes analyseret. Fordelingen af θ_t givet al tilgængelig information, $D_n, n > t$, kaldes den *Kalman-udglattede* fordeling. Den udglattede fordeling bygger på mere information end den filtrerede og er derfor mere nøjagtig.

Kalman-udglatning er beskrevet i [25], men fremstillingen her ligger tættere op ad [3] s. 13-14.

Sætning 3.4 Den Kalman-udglattede fordeling af tilstandsvektoren er givet ved

$$(\theta_t | D_n) \sim \mathcal{N}_p(\tilde{\mathbf{m}}_t, \tilde{\underline{\underline{C}}}_t), \quad (3.10)$$

hvor

$$\begin{aligned} \tilde{\mathbf{m}}_t &= \mathbf{m}_t + \underline{\underline{B}}_t(\tilde{\mathbf{m}}_{t+1} - \mathbf{a}_{t+1}), \\ \tilde{\underline{\underline{C}}}_t &= \underline{\underline{C}}_t + \underline{\underline{B}}_t(\tilde{\underline{\underline{C}}}_{t+1} - \underline{\underline{R}}_{t+1})\underline{\underline{B}}_t^T, \\ \tilde{\mathbf{m}}_n &= \mathbf{m}_n, \quad \tilde{\underline{\underline{C}}}_n = \underline{\underline{C}}_n. \end{aligned}$$

og

$$\underline{\underline{B}}_t = \underline{\underline{C}}_t \underline{\underline{G}}_{t+1}^T \underline{\underline{R}}_{t+1}^{-1}.$$

Bevis:

Beviset er pr. induktion. Ifølge Kalman-filteret er

$$(\theta_n | D_n) \sim \mathcal{N}_p(\mathbf{m}_n, \underline{\underline{C}}_n),$$

hvormed basistrinet er opfyldt. Med hensyn til induktionstrinet antag at

$$(\theta_{t+1} | D_n) \sim \mathcal{N}_p(\tilde{\mathbf{m}}_{t+1}, \tilde{\underline{\underline{C}}}_{t+1}). \quad (3.11)$$

Det mangler at blive vist, at (3.10) er opfyldt til tiden t . I det følgende udledes fordelingen af $(\theta_t | \theta_{t+1}, D_n)$, hvilket sammen med antagelsen i (3.11) muliggør, at fordelingen af $(\theta_t | D_n)$ kan bestemmes. Da det er vist i afsnit 3.2, at θ_t er uafhængig af $Y_{t+1}, \dots, Y_n$ givet θ_{t+1} , gælder det, at

$$(\theta_t | \theta_{t+1}, D_n) = (\theta_t | \theta_{t+1}, D_t). \quad (3.12)$$

Herefter udledes den fælles fordeling af θ_t og θ_{t+1} givet D_t . Fra Kalman-filteret følger det, at

$$(\theta_t | D_t) \sim \mathcal{N}_p(\underline{m}_t, \underline{C}_t),$$

og at

$$(\theta_{t+1} | D_t) \sim \mathcal{N}_p(\underline{a}_{t+1}, \underline{R}_{t+1}).$$

Da

$$\text{Cov}(\theta_t, \theta_{t+1} | D_t) = \text{Cov}(\theta_t, \underline{G}_{t+1} \theta_t + \omega_{t+1} | D_t) = \text{Cov}(\theta_t, \underline{G}_{t+1} \theta_t | D_t) = \underline{C}_t \underline{G}_{t+1}^T,$$

opnås at

$$\begin{pmatrix} \theta_t \\ \theta_{t+1} \end{pmatrix} | D_t \sim \mathcal{N}_{2p} \left[\begin{pmatrix} \underline{m}_t \\ \underline{a}_{t+1} \end{pmatrix}, \begin{pmatrix} \underline{C}_t & \underline{C}_t \underline{G}_{t+1}^T \\ \underline{G}_{t+1} \underline{C}_t & \underline{R}_{t+1} \end{pmatrix} \right].$$

Ved at anvende sætning A.1 følger det, at

$$(\theta_t | \theta_{t+1}, D_n) = (\theta_t | \theta_{t+1}, D_t) \sim \mathcal{N}_p(\underline{m}_t + \underline{B}_t(\theta_{t+1} - \underline{a}_{t+1}), \underline{C}_t - \underline{B}_t \underline{R}_{t+1} \underline{B}_t^T). \quad (3.13)$$

Det følger af (3.11), at

$$\tilde{m}_t = E(\theta_t | D_n) = E[E(\theta_t | \theta_{t+1}, D_n) | D_n] = \underline{m}_t + \underline{B}_t(\tilde{m}_{t+1} - \underline{a}_{t+1}),$$

og

$$\begin{aligned} \tilde{C}_t = \text{Var}(\theta_t | D_n) &= E[\text{Var}(\theta_t | \theta_{t+1}, D_n) | D_n] + \text{Var}[E(\theta_t | \theta_{t+1}, D_n) | D_n] \\ &= \underline{C}_t + \underline{B}_t(\tilde{C}_{t+1} - \underline{R}_{t+1})\underline{B}_t^T. \end{aligned}$$

■

3.5 Maksimum-likelihood estimation af ukendte parametre

Hidtil er det blevet antaget, at varianserne V_t og $\underline{W}_t$, transitionsmatricen $\underline{G}_t$ og designvektoren $\underline{F}_t$ alle er kendte. I de fleste tilfælde afhænger disse størrelser delvist af ukendte parametre. Lad de ukendte parametre være indeholdt i vektoren, ψ .

I det følgende opstilles likelihood-funktionen for ψ . Der gælder ved gentagen brug af Bayes' sætning (se afsnit 3.3), at

$$L(\psi) = \prod_{t=1}^n f(y_t | y_{t-1}, \dots, y_1, \psi),$$

hvor n angiver antallet af observationer. Den betingede fordeling af Y_t givet D_{t-1} er ifølge Kalman-filteret givet ved $(Y_t|D_{t-1}) \sim \mathcal{N}(f_t, Q_t)$, hvor f_t og Q_t afhænger af ψ .

Det følger, at

$$L(\psi) = \prod_{t=1}^n (2\pi Q_t)^{-1/2} \exp \left[-\sum_{t=1}^n (Y_t - f_t)^2 / 2Q_t \right].$$

Dermed er

$$l(\psi) = \ln L(\psi) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^n \ln Q_t - \frac{1}{2} \sum_{t=1}^n (Y_t - f_t)^2 / Q_t. \quad (3.14)$$

For at maksimere ovenstående log-likelihoodfunktion benyttes som regel en numerisk optimerings-procedure som f.eks. Newton's algoritme, der er beskrevet i afsnit 2.3. I [7] s. 144-147 vises det, hvordan score-vektoren kan udregnes rekursivt givet en værdi af ψ . De ukendte størrelser bestemmes ved hjælp af Kalman-filteret og -smootheren. Numerisk beregning af informationsmatricen er vanskelig. I stedet kan den approksimeres ved den såkaldte BFGS metode, som er beskrevet kort i [7].

I mange tilfælde kan proceduren gøres simplere ved at estimere en af parametrene separat. Dette gælder i de tilfælde, hvor varianserne er givet ved

$$V_t = \sigma^2 V_t^*, \quad \underline{W}_t = \sigma^2 \underline{W}_t^*, \quad \underline{C}_0 = \sigma^2 \underline{C}_0^*,$$

hvor σ^2 er en af de ukendte parametre i modellen. Det følger fra Kalman-filteret, at $Q_t = \sigma^2 Q_t^*$. Lad $\psi = (\psi^{*T}, \sigma^2)^T$. Det følger, at

$$\begin{aligned} l(\psi^*, \sigma^2) &= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^n \ln(\sigma^2 Q_t^*) - \frac{1}{2} \sum_{t=1}^n (Y_t - f_t)^2 / \sigma^2 Q_t^* \\ &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2} \sum_{t=1}^n \ln Q_t^* - \frac{1}{2\sigma^2} \sum_{t=1}^n (Y_t - f_t)^2 / Q_t^*. \end{aligned} \quad (3.15)$$

Ideen er nu at bestemme maksimum-likelihood estimatet af σ^2 givet en specifik værdi af ψ^* . Derfor differentieres log-likelihoodfunktionen med hensyn til σ^2 .

$$\begin{aligned} \frac{\partial l(\psi^*, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \left(\frac{1}{\sqrt{2}\sigma^2} \right)^2 \sum_{t=1}^n (Y_t - f_t)^2 / Q_t^* \\ &= -\frac{1}{\sigma^2} \left[\frac{n}{2} - \frac{1}{2\sigma^2} \sum_{t=1}^n (Y_t - f_t)^2 / Q_t^* \right]. \end{aligned}$$

Det ses, at ovenstående udtryk er lig med nul for

$$\sigma^2(\psi^*) = \frac{1}{n} \sum_{t=1}^n (Y_t - f_t)^2 / Q_t^* = \widehat{\sigma^2}(\psi^*). \quad (3.16)$$

Ved at indsætte udtrykket for $\widehat{\sigma^2}(\psi)$ i (3.15) fremkommer profil log-likelihoodfunktionen

$$l(\psi^*) = -\frac{n}{2}[\ln(2\pi) + 1] - \frac{1}{2} \sum_{t=1}^n \ln Q_t^* - \frac{n}{2} \ln \widehat{\sigma^2}(\psi^*).$$

Dermed er log-likelihoodfunktionen blevet reduceret med en parameter i forhold til (3.14). Når $\widehat{\psi^*}$ er blevet bestemt, indsættes denne i (3.16), hvorved maksimum-likelihood estimatet for $\sigma^2(\widehat{\psi^*})$ fremkommer.

3.6 Strukturelle tidsrække modeller

Strukturelle tidsrække modeller er special-tilfælde af en DLM og blev først beskrevet af Andrew Harvey, se [13].

Den simpleste strukturelle tidsrække model er en *local level model*, som har observationsligningen

$$Y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2),$$

og hvor niveauet udvikler sig ved en *random walk*

$$\mu_{t+1} = \mu_t + \xi_t, \quad \xi_t \sim \mathcal{N}(0, \sigma_\xi^2).$$

En mere kompliceret model fremkommer ved at tilføje en hældning til niveauet

$$\mu_{t+1} = \mu_t + \kappa_t + \xi_t, \quad \xi_t \sim \mathcal{N}(0, \sigma_\xi^2). \quad (3.17)$$

Hældningen, κ_t , er stokastisk og udvikler sig over tid ved en *random walk*;

$$\kappa_{t+1} = \kappa_t + \zeta_t, \quad \zeta_t \sim \mathcal{N}(0, \sigma_\zeta^2). \quad (3.18)$$

En *local level model* tilføjet med en hældning givet ved ovenstående ligning kaldes for en *local linear trend model*.

En *basic structural model* (BSM) er en *local linear trend model*, der er udvidet med et sæsonvariationsled. Observationsligningen får udseendet

$$Y_t = \mu_t + \gamma_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2),$$

hvor μ_t er beskrevet som i *local linear trend modellen*, og γ_t er sæsonvariations-komponenten, som er givet ved

$$\gamma_{t+1} = -\gamma_t - \gamma_{t-1} - \cdots - \gamma_{t-s+2} + \delta_t, \quad \delta_t \sim \mathcal{N}(0, \sigma_\delta^2). \quad (3.19)$$

Størrelsen, s , angiver perioden for sæsonvariationen, f.eks. antallet af måneder i et år. Det ses, at hvis sæsonvariationen gøres deterministisk ved at sætte σ_δ^2 lig med nul, gælder det, at

$$\sum_{j=0}^{s-1} \gamma_{t-j} = 0.$$

En positiv værdi af γ_t angiver således, at niveauet for årstid nr. t ligger over gennemsnittet, mens en negativ værdi angiver, at det ligger under gennemsnittet. I dette tilfælde er sæsonvariationen altid den samme. Hvis sæsonvariations-komponenten derimod tillades at være dynamisk, adderer de s led til δ_t , og sæsonvariationen er dermed kun konstant i middel og kan gradvist ændre sig.

Sæsonvariationen kunne også beskrives ved en Fourier-række, som f.eks.

$$\gamma_t = \sum_{j=1}^d [\gamma_{c,j} \cos(\tau_j t) + \gamma_{s,j} \sin(\tau_j t)],$$

hvor $\tau_j = 2\pi j/s$, og d angiver graden af Fourier-rækken. Det ses, at

$$\sum_{t=1}^s \gamma_t = \sum_{j=1}^d \left[\gamma_{c,j} \sum_{t=1}^s \cos(\tau_j t) + \gamma_{s,j} \sum_{t=1}^s \sin(\tau_j t) \right] = 0.$$

Det stokastiske element tilføjes ved at lade $\gamma_{c,j}$ og $\gamma_{s,j}$ afhænge af tiden, således at de udvikler sig ved en random walk

$$\gamma_{t+1}^{c,j} = \gamma_t^{c,j} + \delta_t^{c,j}, \quad \gamma_{t+1}^{s,j} = \gamma_t^{s,j} + \delta_t^{s,j},$$

hvor $\delta_t^{c,j}$ og $\delta_t^{s,j}$ er normalfordelte med middelværdi nul og varians σ_δ^2 .

3.6.1 Filtrering og udglatning

Estimering af de uobserverede komponenter i en strukturel tidsrække model håndteres mest effektivt ved at skrive modellen i state space form. Dette gør det muligt at anvende Kalman-filteret og -smootheren på modellen. I dette afsnit udføres dette først for en local linear trend model og derefter for den mere komplicerede BSM.

Definer

$$\boldsymbol{\theta}_t = (\mu_t, \kappa_t)^T.$$

Da svarer (3.17) og (3.18) netop til tilstandsligningen

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

hvor

$$\underline{\underline{G}}_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad \omega_t = (\xi_t, \zeta_t)^T, \quad \underline{\underline{W}}_t = \begin{bmatrix} \sigma_\xi^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix}.$$

Observationsligningen bliver

$$Y_t = \mathbf{F}_t^T \boldsymbol{\theta}_t + \nu_t, \quad \nu_t \sim \mathcal{N}(0, V_t),$$

hvor

$$\mathbf{F}_t = (1, 0)^T, \quad \nu_t = \varepsilon_t, \quad \text{og} \quad V_t = \sigma_\varepsilon^2. \quad (3.20)$$

For en BSM udvides tilstandsvektoren til

$$\boldsymbol{\theta}_t = (\mu_t, \kappa_t, \gamma_t, \gamma_{t-1}, \dots, \gamma_{t-s+2})^T.$$

Transitionsmatricen er blokdiagonal med en blok for henholdsvis trenden og sæsonkomponenten.

$$\underline{\underline{G}}_t = \begin{bmatrix} \underline{\underline{T}} & \\ & \underline{\underline{S}} \end{bmatrix},$$

hvor

$$\underline{\underline{T}} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

og

$$\underline{\underline{S}} = \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

Det ses, at γ_t sættes til minus summen af de foregående $s - 1$ sæsonled, som forskydes en plads i $\boldsymbol{\theta}_t$, således at γ_{t-s+1} ikke længere er et element i vektoren.

Systemfejls-vektoren er givet ved

$$\omega_t = (\xi_t, \zeta_t, \delta_t, 0, \dots, 0)^T$$

og har fordelingen

$$\omega_t \sim \mathcal{N}(\mathbf{0}, \underline{\underline{W}}_t),$$

hvor

$$\underline{\underline{W}}_t = \begin{bmatrix} \sigma_\xi^2 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_\zeta^2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_\delta^2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

I observationsligningen er

$$\mathbf{F}_t = (1, 0, 1, 0, \dots, 0)^T,$$

mens ν_t er uændret i forhold til (3.20).

For tilfældet $s = 4$ opnås tilstandsligningen

$$\boldsymbol{\theta}_t = \begin{bmatrix} \mu_t \\ \kappa_t \\ \gamma_t \\ \gamma_{t-1} \\ \gamma_{t-2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu_{t-1} \\ \kappa_{t-1} \\ \gamma_{t-1} \\ \gamma_{t-2} \\ \gamma_{t-3} \end{bmatrix} + \begin{bmatrix} \xi_t \\ \zeta_t \\ \delta_t \\ 0 \\ 0 \end{bmatrix}$$

og observationsligningen

$$Y_t = [1 \ 0 \ 1 \ 0 \ 0] \boldsymbol{\theta}_t + \nu_t.$$

Dermed er BSM-modellen bragt i state space form og Kalman-filteret og -smootheren kan anvendes på modellen til at estimere den betingede fordeling af tilstandsvektoren $\boldsymbol{\theta}_t$ givet data under forudsætning af, at varianserne for støjledene samt $\mathbf{m}_0$ og $\underline{\underline{C}}_0$ er kendte.

3.6.2 Modelkontrol

Modelkontrol af strukturelle tidsrække modeller kan baseres på de standardiserede prediktionsfejl

$$r_t = \frac{e_t}{\sqrt{Q_t}}, \quad t = 1, \dots, n. \quad (3.21)$$

Det kan kontrolleres, om variansen af r_t er homogen i to lige store dele af data-sættet ved hjælp af teststørrelsen

$$H(k) = \frac{\sum_{t=k/2+1}^k r_t^2}{\sum_{t=1}^{k/2} r_t^2} \sim \mathcal{F}_{k/2, k/2}$$

under antagelse af hypotesen, at variansen er homogen. Størrelsen k er et positivt, lige tal, der er mindre end eller lig med antallet af observationer.

Det er vist i afsnit 3.3, at prediktionsfejlene er serielt ukorrelerede. Det kan undersøges, om dette, samt antagelsen om normalitet, er overholdt.

3.6.3 StructTS

StructTS er en funktion i programmet *R*, som anvendes til at modellere tidsrække-data ved en strukturel tidsrække model. Softwaren er udviklet af Brian Ripley, se [20].

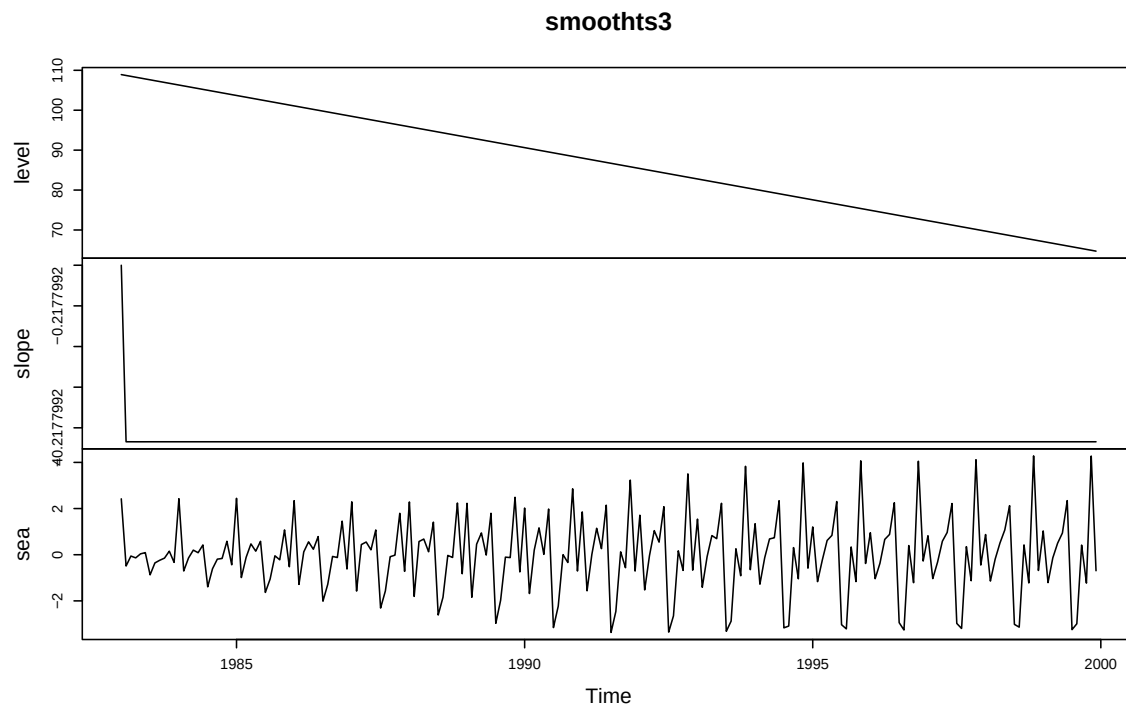
Funktionen har udseendet

```
StructTS(x, type=c("level", "trend", "BSM"), fixed=NULL)
```

Nedenfor følger en forklaring af funktionens argumenter.

- **x** En vektor bestående af observationer fra en tidsrække, hvor en periode er defineret.
- **type** Kan være en local level model, local linear trend model eller en BSM. I en BSM beskrives sæsonvariationen ved (3.19).
- **fixed** En vektor som kan anvendes til at angive varianserne i modellen. Kun varianser, der er specificeret som "NA", estimeres ved maksimum-likelihood. Hvis vektoren ikke er anført som argument, estimeres samtlige varianser.

Ved kald af funktionen foretages Kalman-filtrering af dataene, som derefter kan udglattes ved hjælp af kommandoen *tsSmooth*. Dette giver et output, der viser de estimerede værdier af modellens komponenter for hvert datapunkt i tidsrækken, hvilket også illustreres i et plot. Figur 3.2 viser udviklingen af trenden, hældningen og sæsonvariationen for en BSM. Desuden kan de estimerede varianser for støj-ledene tilgås.

Figur 3.2: Output af *tsSmooth*.

Kapitel 4

Den Dynamiske Generaliserede Lineære Model

I dette afsnit generaliseres teorien for den Gaussiske state space model til også at omfatte modeller, hvor observationerne har en fordeling fra den eksponentielle familie. Der beskrives to forskellige metoder til at foretage Kalman-filtrering og -smoothing i denne type af state space modeller, også kaldet *dynamiske generaliserede lineære modeller* (DGLM).

4.1 Den dynamiske generaliserede lineære model

Notation

Hvis fordelingen af en stokastisk vektor, $\mathbf{X}$, kun er bestemt ved middelværdi og variansmatrix, anvendes notationen

$$\mathbf{X} \sim (\mathbf{m}, \underline{\underline{V}}).$$

Definition 4.1 Den en-dimensionelle dynamiske generaliserede lineære model er defineret ved

$$\text{Observationsmodel:} \quad f(y_t|\eta_t) = \exp\{[y_t\eta_t - a(\eta_t)]/V_t\}b(y_t, V_t) \quad (4.1)$$

$$E(Y_t|\eta_t) = \mu_t = a'(\eta_t), \quad \text{Var}(Y_t|\eta_t) = \sigma_t^2 = V_t a''(\eta_t),$$

$$h(\eta_t) = g(\mu_t) = \lambda_t = \mathbf{F}_t^T \boldsymbol{\theta}_t,$$

hvor g og h er kontinuerte og monotone link-funktioner.

$$\text{Tilstandsligning:} \quad \boldsymbol{\theta}_t = \underline{\underline{G}}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim (\mathbf{0}, \underline{\underline{W}}_t) \quad (4.2)$$

$$\text{Startinformation: } (\boldsymbol{\theta}_0|D_0) \sim (\mathbf{m}_0, \underline{\underline{C}}_0). \quad (4.3)$$

Til forskel fra den almindelige DLM er Y_t og $\boldsymbol{\theta}_t$ ikke nødvendigvis normalfordelte. Observationen Y_t afhænger af tilstandsvektoren, $\boldsymbol{\theta}_t$, gennem den naturlige parameter, η_t , der via linkfunktionen h kædes til en linearkombination af elementerne i $\boldsymbol{\theta}_t$. Tilstandsligningen er som i den DLM, men fordelingen af systemfejlen, ω_t , er kun specificeret ved middelværdi og variansmatrix, idet den præcise fordeling er ukendt. I afsnit 4.3 antages det dog, at tilstandsvektoren er normalfordelt.

4.2 Kalman-filtrering med konjugeret a priori fordeling

West, Harrison og Migon udnyttede teorien for *konjugerede familier* samt *linear Bayes' estimation* til at foretage Kalman-filtrering i en DGLM, se [26].

Da det ikke kan antages, at observationerne er normalfordelte, er fremgangsmåden for Kalman-filtrering lidt anderledes end i den DLM, ligesom de udledte fordelinger ikke er eksakte men approksimationer. Som i afsnit 3.3 udledes a posteriori-fordelingen af $\boldsymbol{\theta}_t$ pr. induktion. Basistrinet er givet i (4.3). Antag at

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

Det følger af (4.2), at

$$(\boldsymbol{\theta}_t|D_{t-1}) \sim (\mathbf{a}_t, \underline{\underline{R}}_t), \quad (4.4)$$

hvor $\mathbf{a}_t = \underline{\underline{G}}_t \mathbf{m}_{t-1}$, og $\underline{\underline{R}}_t = \underline{\underline{G}}_t \underline{\underline{C}}_{t-1} \underline{\underline{G}}_t^T + \underline{\underline{W}}_t$.

Der gælder, at

$$E(\boldsymbol{\theta}_t|D_t) = E[E(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})|D_t], \quad (4.5)$$

og at

$$\text{Var}(\boldsymbol{\theta}_t|D_t) = \text{Var}[E(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})|D_t] + E[\text{Var}(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})|D_t]. \quad (4.6)$$

Størrelserne $E(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})$ og $\text{Var}(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})$ kan estimeres ved hjælp af linear Bayes' estimation udfra den fælles a priori fordeling af $\boldsymbol{\theta}_t$ og λ_t . Det ses af (4.4), at

$$(\lambda_t|D_{t-1}) \sim (f_t, q_t), \quad (4.7)$$

hvor $f_t = \mathbf{F}_t^T \mathbf{a}_t$, og $q_t = \mathbf{F}_t^T \underline{\underline{R}}_t \mathbf{F}_t$. Det følger, at

$$\begin{pmatrix} \boldsymbol{\theta}_t \\ \lambda_t \end{pmatrix} | D_{t-1} \sim \left[\begin{pmatrix} \mathbf{a}_t \\ f_t \end{pmatrix}, \begin{pmatrix} \underline{\underline{R}}_t & \underline{\underline{R}}_t \mathbf{F}_t \\ \mathbf{F}_t^T \underline{\underline{R}}_t & q_t \end{pmatrix} \right].$$

Ud fra afsnit A.2 opnås linear Bayes' estimaterne

$$\hat{\mathbf{E}}(\boldsymbol{\theta}_t | \lambda_t, D_{t-1}) = \mathbf{a}_t + \underline{\underline{R}}_t \mathbf{F}_t (\lambda_t - f_t) / q_t,$$

og

$$\widehat{\text{Var}}(\boldsymbol{\theta}_t | \lambda_t, D_{t-1}) = \underline{\underline{R}}_t - \underline{\underline{R}}_t \mathbf{F}_t \mathbf{F}_t^T \underline{\underline{R}}_t / q_t.$$

Ved at indsætte i (4.5) og (4.6) fremkommer at

$$\mathbf{E}(\boldsymbol{\theta}_t | D_t) = \mathbf{E}[\mathbf{a}_t + \underline{\underline{R}}_t \mathbf{F}_t (\lambda_t - f_t) / q_t | D_t], \quad (4.8)$$

og

$$\text{Var}(\boldsymbol{\theta}_t | D_t) = \text{Var}[\mathbf{a}_t + \underline{\underline{R}}_t \mathbf{F}_t (\lambda_t - f_t) / q_t | D_t] + \underline{\underline{R}}_t - \underline{\underline{R}}_t \mathbf{F}_t \mathbf{F}_t^T \underline{\underline{R}}_t / q_t. \quad (4.9)$$

Fordelingen af λ_t givet D_t mangler at blive udledt. Da observationsfordelingen er defineret ud fra den naturlige parameter, η_t , udledes først a posteriori fordelingen for denne parameter. Der gælder, at

$$f(\eta_t | D_t) = f(\eta_t | y_t, D_{t-1}) \propto f(\eta_t, y_t | D_{t-1}) = f(\eta_t | D_{t-1}) f(y_t | \eta_t, D_{t-1}). \quad (4.10)$$

For at løse denne ligning er det nødvendigt at kende den fuldstændige a priori fordeling for η_t . Beregningsmæssigt er det mest hensigtsmæssigt, hvis $f(\eta_t | D_{t-1})$ er på en bestemt form, som medfører, at a posteriori tætheden, $f(\eta_t | D_t)$, har samme form. Ifølge [25] s. 519 er denne form givet ved

$$f(\eta_t | D_{t-1}) = c(r_t, s_t) \exp[r_t \eta_t - s_t a(\eta_t)], \quad (4.11)$$

hvor r_t og s_t er ukendte parametre, og c er en kendt funktion. Hvis a priori tætheden er på formen i (4.11), siges det, at a priori fordelingen for η_t tilhører den *konjugerede familie*. Det ses af (4.7), at

$$\mathbf{E}[h(\eta_t) | D_{t-1}] = f_t, \quad \text{og} \quad \text{Var}[h(\eta_t) | D_{t-1}] = q_t,$$

hvoraf middelværdien og variansen for a priori fordelingen af η_t enten kan beregnes eksakt eller approksimeres ud fra en Taylor-udvikling af $\eta_t = h^{-1}(\lambda_t)$ omkring f_t . Derefter kan parametrene, r_t , og s_t bestemmes ud fra den udledte middelværdi og varians, hvormed a priori fordelingen for η_t er fuldt ud specificeret.

Ved at indsætte (4.1) og (4.11) i (4.10) opnås at

$$\begin{aligned} f(\eta_t|D_t) &\propto c(r_t, s_t) \exp[r_t\eta_t - s_t a(\eta_t)] \exp\{\phi_t[y_t\eta_t - a(\eta_t)]\} b(y_t, V_t) \\ &= c(r_t, s_t) b(y_t, V_t) \exp[(r_t + \phi_t y_t)\eta_t - (\phi_t + s_t)a(\eta_t)], \end{aligned} \quad (4.12)$$

hvor $\phi_t = 1/V_t$. Der gælder, at

$$\begin{aligned} f(y_t|D_{t-1}) &= \int f(y_t, \eta_t|D_{t-1}) d\eta_t \\ &= \int f(y_t|\eta_t, D_{t-1}) f(\eta_t|D_{t-1}) d\eta_t \\ &= \int \exp\{\phi_t[y_t\eta_t - a(\eta_t)]\} b(y_t, V_t) c(r_t, s_t) \exp[r_t\eta_t - s_t a(\eta_t)] d\eta_t \\ &= b(y_t, V_t) c(r_t, s_t) \int \exp[(r_t + \phi_t y_t)\eta_t - (\phi_t + s_t)a(\eta_t)] d\eta_t \\ &= \frac{b(y_t, V_t) c(r_t, s_t)}{c(r_t + \phi_t y_t, \phi_t + s_t)} \int c(r_t + \phi_t y_t, \phi_t + s_t) \exp[(r_t + \phi_t y_t)\eta_t - (\phi_t + s_t)a(\eta_t)] d\eta_t \\ &= \frac{b(y_t, V_t) c(r_t, s_t)}{c(r_t + \phi_t y_t, \phi_t + s_t)}. \end{aligned} \quad (4.13)$$

Det følger af (4.12), at

$$\begin{aligned} f(\eta_t|D_t) &\propto f(y_t|D_{t-1}) c(r_t + \phi_t y_t, s_t + \phi_t) \exp[(r_t + \phi_t y_t)\eta_t - (s_t + \phi_t)a(\eta_t)] \\ &\propto c(r_t + \phi_t y_t, s_t + \phi_t) \exp[(r_t + \phi_t y_t)\eta_t - (s_t + \phi_t)a(\eta_t)]. \end{aligned} \quad (4.14)$$

Det vil sige, at a posteriori fordelingen for η_t har samme form som a priori fordelingen med parametre opdateret til $r_t^* = r_t + \phi_t y_t$, og $s_t^* = s_t + \phi_t$.

Endelig kan a posteriori middelværdi og varians for λ_t udledes. Der gælder, at

$$(\lambda_t|D_t) \sim (f_t^*, q_t^*),$$

hvor

$$f_t^* = E[h(\eta_t)|D_t] \quad \text{og} \quad q_t^* = \text{Var}[h(\eta_t)|D_t]$$

enten kan beregnes eksakt eller approksimeres ved en Taylor-udvikling af λ_t omkring den estimerede middelværdi af $(\eta_t|D_t)$, som er givet udfra r_t^* og s_t^* . Ved indsættelse i (4.8) og (4.9) fremkommer at

$$(\theta_t|D_t) \sim (\mathbf{m}_t, \underline{\underline{C}}_t),$$

hvor

$$\mathbf{m}_t = \mathbf{a}_t + \underline{\underline{R}}_t \mathbf{F}_t (f_t^* - f_t)/q_t, \quad (4.15)$$

og

$$\underline{\underline{C}}_t = \underline{\underline{R}}_t - \underline{\underline{R}}_t \mathbf{F}_t \mathbf{F}_t^T \underline{\underline{R}}_t (1 - q_t^*/q_t)/q_t. \quad (4.16)$$

Udtrykkene for den Kalman-udglattede middelværdi og varians for θ_t bestemmes som i det Gaussiske tilfælde, idet linear Bayes's estimation anvendes i stedet for sætning A.1.

En svaghed ved West, Harrison og Migons metode er, at den ikke kan generaliseres til det multivariable tilfælde.

4.2.1 Kalman-filtrering for Poisson-modellen

Da Poisson-fordelte observationer analyseres i anvendelses-delen af rapporten, vil der i dette afsnit blive foretaget Kalman-filtrering for en Poisson state space model. Det vil sige, at beregningerne ovenfor her gentages i Poisson-tilfældet. I afsnit 2.1.1 blev det udledt, at den naturlige parameter, $\eta_t = \ln \mu_t$ i Poisson-fordelingen. Det antages, at link-funktionen $h(\eta_t)$ er identiteten, så at $\eta_t = \lambda_t = \mathbf{F}_t^T \theta_t$.

A priori for μ_t

Da sandsynlighedsfunktionen i (2.4) er defineret ud fra μ_t , er det mere behændigt at udlede fordelingerne af μ_t i stedet for fordelingerne af η_t . For at estimere middelværdi og varians for $(\mu_t|D_{t-1})$ foretages en Taylor-udvikling af μ_t omkring middelværdien af $(\lambda_t|D_{t-1})$.

$$\mu_t = \exp(\lambda_t) \approx \exp(f_t) + \exp(f_t)(\lambda_t - f_t).$$

Det følger, at

$$(\mu_t|D_{t-1}) \sim [\exp(f_t), \exp(f_t)^2 q_t]. \quad (4.17)$$

Da det er mest hensigtsmæssigt, antages det, at a priori fordelingen for μ_t er den konjugerede til Poisson-fordelingen.

Lemma 4.2 *Lad $(Y_t|\mu_t) \sim \text{Poisson}(\mu_t)$. Da gælder der, at den konjugerede a priori fordeling for μ_t er givet ved*

$$(\mu_t|D_{t-1}) \sim \text{gamma}(\alpha_t, \beta_t).$$

Bevis:

Ved at indsætte i (4.11) fremkommer at

$$f(\mu_t|D_{t-1}) = c(r_t, s_t) \exp(r_t \ln \mu_t - s_t \mu_t) = c(r_t, s_t) \mu_t^{r_t} \exp(-s_t \mu_t),$$

hvilket er tæthedsfunktionen for en gamma-fordeling med parametre r_t og s_t og normaliseringskonstant $c(r_t, s_t) = s_t^{r_t} / \Gamma(r_t)$. ■

Parametrene r_t og s_t kan bestemmes ud fra (4.17). Der gælder for en gamma-fordeling med parametre α og β , at middelværdien er givet ved α/β og variansen ved α/β^2 . Det følger, at

$$E(\mu_t | D_{t-1}) = \frac{r_t}{s_t} = \exp(f_t),$$

og at

$$\text{Var}(\mu_t | D_{t-1}) = \frac{r_t}{s_t^2} = \exp(f_t)^2 q_t.$$

Ved at løse ligningerne opnås at

$$(\mu_t | D_{t-1}) \sim \text{gamma}(r_t, s_t),$$

hvor

$$\begin{aligned} r_t &= \frac{1}{q_t}, \\ s_t &= \frac{1}{\exp(f_t) q_t}. \end{aligned}$$

A posteriori for μ_t

Det ses af (4.14), at

$$(\mu_t | D_t) \sim \text{gamma}(r_t^*, s_t^*),$$

hvor

$$r_t^* = r_t + y_t = \frac{1}{q_t} + y_t,$$

og

$$s_t^* = s_t + 1 = \frac{1}{\exp(f_t) q_t} + 1.$$

A posteriori for λ_t

For at estimere middelværdi og varians for λ_t givet D_t , Taylor-udvikles λ_t omkring den estimerede middelværdi for $(\mu_t | D_t)$, som er givet ved r_t^* / s_t^* .

$$\lambda_t = \ln \mu_t \approx \ln \left(\frac{r_t^*}{s_t^*} \right) + \frac{s_t^*}{r_t^*} \left(\mu_t - \frac{r_t^*}{s_t^*} \right).$$

Det følger, at

$$(\lambda_t | D_t) \sim \left[\ln \left(\frac{r_t^*}{s_t^*} \right), \left(\frac{s_t^*}{r_t^*} \right)^2 \frac{r_t^*}{(s_t^*)^2} \right] = \left[\ln \left(\frac{r_t^*}{s_t^*} \right), \frac{1}{r_t^*} \right] = (f_t^*, q_t^*),$$

hvor

$$f_t^* = \ln \left(\frac{1/q_t + y_t}{1/\exp(f_t)q_t + 1} \right),$$

og

$$q_t^* = \frac{1}{1/q_t + y_t}.$$

Ved at indsætte f_t^* og q_t^* i (4.15) og (4.16) fremkommer a posteriori fordelingen af θ_t i en Poisson state space model.

4.3 Extended Kalman-filtrering og -smoothing

Extended Kalman-filtrering og -smoothing er en approksimativ metode i modeller med Gaussisk tilstandsligning men ikke nødvendigvis Gaussisk observationsfordeling. Overordnet benyttes Taylorudvikling til at approksimere tætheden af observationerne ved en normalfordeling, således at a posteriori fordelingen af tilstandsvektoren bliver Gaussisk. Forskellen fra den DGLM, der blev anvendt i afsnit 4.2, er, at det her antages, at tilstandsvektoren er normalfordelt.

Extended Kalman-filtrering og smoothing blev beskrevet i [8] og [9], men fremstillingen her ligger tættere op ad [3].

Definer $\mathbf{y}^n = (y_1, y_2, \dots, y_n)$. Fordelingen af $\theta_0, \theta_1, \dots, \theta_n$ givet data skal bestemmes. Der gælder ifølge Bayes' sætning (se afsnit 3.3), at

$$\begin{aligned} f(\theta_0, \theta_1, \dots, \theta_n | \mathbf{y}^n) &\propto f(\theta_0, \theta_1, \dots, \theta_n, \mathbf{y}^n) \\ &= f(y_n | \theta_0, \dots, \theta_n, \mathbf{y}^{n-1}) f(\theta_0, \dots, \theta_n, \mathbf{y}^{n-1}) \\ &= f(y_n | \theta_0, \dots, \theta_n, \mathbf{y}^{n-1}) f(\theta_n | \theta_0, \dots, \theta_{n-1}, \mathbf{y}^{n-1}) f(\theta_0, \dots, \theta_{n-1}, \mathbf{y}^{n-1}). \end{aligned}$$

Det følger af uafhængigheds-betragtningerne i afsnit 3.2, at

$$f(\theta_0, \theta_1, \dots, \theta_n | \mathbf{y}^n) \propto \prod_{t=1}^n f(y_t | \theta_t) \prod_{t=1}^n f(\theta_t | \theta_{t-1}) f(\theta_0). \quad (4.18)$$

I tilfælde af ikke-Gaussiske observationer er ovenstående udtryk oftest umuligt at beregne eksakt, hvilket yderligere betyder, at det er meget vanskeligt at udlede a posteriori middelværdi og varians for θ_t , $t = 0, 1, \dots, n$. Ideen med den udvidede Kalman-filtrerer og -smoother er at approksimere $\ln f(y_t | \theta_t)$ ved en kvadratisk funktion af θ_t , idet det medfører, at (4.18) bliver en Gaussisk tæthed.

For en model fra den eksponentielle familie gælder der parallelt med resultaterne i afsnit 2.3, at

$$\frac{\partial \ln f(y_t|\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t} = \frac{\partial \ln f}{\partial \eta_t} \frac{\partial \eta_t}{\partial \mu_t} \frac{\partial \mu_t}{\partial \lambda_t} \frac{\partial \lambda_t}{\partial \boldsymbol{\theta}_t} = k'(\lambda_t) \frac{1}{\sigma_t^2} (y_t - \mu_t) \mathbf{F}_t, \quad (4.19)$$

hvor k er den inverse funktion til g . Hvis observationerne er normalfordelte, gælder der, at

$$\ln f(y_t|\boldsymbol{\theta}_t) = -\frac{(y_t - \mathbf{F}_t^T \boldsymbol{\theta}_t)^2}{2\sigma_t^2} + c = -\frac{y_t^2 + \boldsymbol{\theta}_t^T \mathbf{F}_t \mathbf{F}_t^T \boldsymbol{\theta}_t}{2\sigma_t^2} + \frac{y_t \mathbf{F}_t^T \boldsymbol{\theta}_t}{\sigma_t^2} + c. \quad (4.20)$$

Da det første led i (4.20) er en kvadratisk funktion af $\boldsymbol{\theta}_t$, følger det, at differentialet med hensyn til $\boldsymbol{\theta}_t$ er en lineær funktion af $\boldsymbol{\theta}_t$. Denne egenskab skal approksimationen af $f(y_t|\boldsymbol{\theta}_t)$ også have. I (4.19) afhænger både $k'(\lambda_t)$, σ_t^2 og μ_t af $\boldsymbol{\theta}_t$. Da de to første størrelser antages at afhænge af $\boldsymbol{\theta}_t$ i mindre grad end den sidste, approksimeres de ved konstanter. Der indsættes en fast værdi, $\tilde{\lambda}_t = \mathbf{F}_t^T \tilde{\boldsymbol{\theta}}_t$, hvor $\tilde{\boldsymbol{\theta}}_t$ er et estimat af $\boldsymbol{\theta}_t$. Definer

$$\tilde{\sigma}_t^2 = \sigma_t^2(\tilde{\lambda}_t).$$

Da $\mu_t = k(\lambda_t)$ ikke kan antages at være en lineær funktion af $\boldsymbol{\theta}_t$, foretages en Taylor-udvikling af $k(\lambda_t)$ omkring $\tilde{\lambda}_t$.

$$\mu_t = k(\lambda_t) \approx k(\tilde{\lambda}_t) + k'(\tilde{\lambda}_t)(\lambda_t - \tilde{\lambda}_t).$$

Det følger, at

$$\frac{\partial \ln f(y_t|\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t} \approx k'(\tilde{\lambda}_t) \frac{1}{\tilde{\sigma}_t^2} \{y_t - [k(\tilde{\lambda}_t) + k'(\tilde{\lambda}_t)(\lambda_t - \tilde{\lambda}_t)]\} \mathbf{F}_t \quad (4.21)$$

$$= k'(\tilde{\lambda}_t) \frac{1}{\tilde{\sigma}_t^2} [y_t - k(\tilde{\lambda}_t) + k'(\tilde{\lambda}_t)\tilde{\lambda}_t - k'(\tilde{\lambda}_t)\mathbf{F}_t^T \boldsymbol{\theta}_t] \mathbf{F}_t, \quad (4.22)$$

som er en lineær funktion af $\boldsymbol{\theta}_t$.

Dermed er der udledt en lineær approksimation til (4.19), og en Gaussisk approksimation af tætheden $f(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n|\mathbf{y}^n)$ kan for et observeret $\mathbf{y}^n$ i princippet bestemmes ved at udregne (4.18). Det er dog simplere at bestemme den tilsvarende Gaussiske state space model, hvorefter den sædvanlige Kalman-filtrerer og -smoother kan anvendes på modellen. Definer den stokastiske variabel

$$\tilde{Y}_t = Y_t - k(\tilde{\lambda}_t) + k'(\tilde{\lambda}_t)\tilde{\lambda}_t \quad (4.23)$$

og definer en state space model med observationsligning

$$\tilde{Y}_t = k'(\tilde{\lambda}_t) \mathbf{F}_t^T \boldsymbol{\theta}_t + \nu_t, \quad \nu_t \sim \mathcal{N}(0, \tilde{\sigma}_t^2). \quad (4.24)$$

Da $(\tilde{Y}_t|\boldsymbol{\theta}_t) \sim \mathcal{N}[k'(\tilde{\lambda}_t) \mathbf{F}_t^T \boldsymbol{\theta}_t, \tilde{\sigma}_t^2]$, følger det, at

$$\frac{\partial \ln f(\tilde{y}_t|\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t} = k'(\tilde{\lambda}_t) \frac{1}{\tilde{\sigma}_t^2} (\tilde{y}_t - k'(\tilde{\lambda}_t) \mathbf{F}_t^T \boldsymbol{\theta}_t) \mathbf{F}_t.$$

Ved at indsætte $\tilde{y}_t$ ses det, at ligningen er identisk med (4.21). Dermed er det vist, at state space modellen med observationsligning (4.24) er en Gaussisk approksimation af en DGLM. Kalman-filteret og -smootheren kan anvendes på denne tilnærmende model til at udregne middelværdi og varians for θ_t givet observationerne. Approksimationen kan gøres bedre ved at indsætte det Kalman-udglattede estimat af θ_t i (4.23), hvormed $\tilde{Y}_t$ opdateres. Derefter anvendes Kalman-filteret og -smootheren igen på de reviderede observationer. Denne procedure, kaldet *iterative extended Kalman filtrering og -smoothing*, udføres, indtil de udledte fordelinger af θ_t givet observationerne er konvergeret. Det kan vises, at $f(\theta_0, \theta_1, \dots, \theta_n | \mathbf{y}^n)$ og $f(\theta_0, \theta_1, \dots, \theta_n | \tilde{\mathbf{y}}^n)$ har maksimum for samme værdier af $\theta_0, \theta_1, \dots, \theta_n$, når processen er konvergeret.

4.3.1 Extended Kalman-filtrering og -smoothing for Poisson-modellen

Lad $(Y_t | \eta_t) \sim \text{Poisson}(\mu_t)$. I Poisson-tilfældet er forholdet mellem den betingede middelværdi og den naturlige parameter givet ved $\ln \mu_t = \eta_t$. Det antages som tidligere, at $\eta_t = \lambda_t = \mathbf{F}_t^T \theta_t$.

Der gælder, at

$$\frac{\partial \ln f(y_t | \theta_t)}{\partial \theta_t} = \frac{\partial \ln f}{\partial \eta_t} \frac{\partial \eta_t}{\partial \theta_t} = (y_t - \mu_t) \mathbf{F}_t.$$

For at udlede en Gaussisk approksimation Taylor-udvikles μ_t omkring $\tilde{\mu}_t = \exp(\tilde{\eta}_t)$, hvor $\tilde{\eta}_t$ er et estimat af η_t .

$$\mu_t \approx \exp(\tilde{\eta}_t) + \exp(\tilde{\eta}_t)(\eta_t - \tilde{\eta}_t) = \tilde{\mu}_t + \tilde{\mu}_t(\mathbf{F}_t^T \theta_t - \tilde{\eta}_t).$$

Det følger, at

$$\frac{\partial \ln f(y_t | \theta_t)}{\partial \theta_t} \approx [y_t - (\tilde{\mu}_t + \tilde{\mu}_t \mathbf{F}_t^T \theta_t - \tilde{\mu}_t \tilde{\eta}_t)] \mathbf{F}_t = \tilde{\mu}_t \left(\frac{y_t - \tilde{\mu}_t}{\tilde{\mu}_t} + \tilde{\eta}_t - \mathbf{F}_t^T \theta_t \right) \mathbf{F}_t,$$

som er en lineær funktion af θ_t . Definer

$$\tilde{Y}_t = \frac{Y_t - \tilde{\mu}_t}{\tilde{\mu}_t} + \tilde{\eta}_t$$

og definer state space modellen med observationsligning

$$\tilde{Y}_t = \mathbf{F}_t^T \theta_t + \nu_t, \quad \nu_t \sim \mathcal{N}(0, 1/\tilde{\mu}_t). \quad (4.25)$$

Det følger, at

$$\frac{\partial \ln f(\tilde{y}_t | \theta_t)}{\partial \theta_t} = \tilde{\mu}_t \left(\frac{y_t - \tilde{\mu}_t}{\tilde{\mu}_t} + \tilde{\eta}_t - \mathbf{F}_t^T \theta_t \right) \mathbf{F}_t.$$

Dermed er modellen defineret udfra (4.25) en normalfordelt approksimation til Poisson-modellen, og Kalman-filteret og -smootheren kan anvendes på modellen.

Det indledende estimat af η_t kan bestemmes udfra observationen, y_t . Da $Y_t \sim \text{Poisson}(\exp(\eta_t))$, gælder der, at $\ln y_t \approx \eta_t$. Da y_t kan være nul, estimeres η_t ved

$$\tilde{\eta}_t = \ln(y_t + 1).$$

4.4 `sspir`

`sspir` er en pakke i *R*, som indeholder funktioner til at analysere data ved en DGLM. Pakken er udviklet af Claus Dethlefsen og Søren Lundbye-Christensen, se [4].

En state space model defineres i `sspir` ved funktionen `ssm`, som har følgende syntaks.

```
ssm(formula, family=gaussian, data, subset, time)
```

Syntaksen ligger tæt op ad *glm*-funktionens, som anvendes til at analysere data med en GLM. Her følger en forklaring af funktionens argumenter.

- **formula** Til forskel fra *glm*-funktionen kan parametre være dynamiske, hvilket vil sige, at de udvikler sig over tid som en random walk. Hvis variabelen x har en dynamisk parameter, skrives det

```
tvar(x).
```

Det er muligt at definere et polynomium til beskrivelse af trenden med funktionen

```
polytime(time, degree=p)
```

Sæsonvariation kan beskrives som i *StructTS* med funktionen

```
sumseason(time, period=m)
```

eller ved en Fourier-række

```
polytrig(time, period=m, degree=d)
```

- **family** angiver hvilken fordeling indenfor den eksponentielle familie, observationerne har, samt hvilken link-funktion modellen har.
- **data** Der er mulighed for at angive et datasæt, som indeholder værdier for variable specificeret under *formula*.
- **subset** giver mulighed for kun at analysere en delmængde af observationerne fra datasættet.
- **time** angiver tidspunkterne, t_k , for observationerne.

Til forskel fra *StructTS* er det i *sspir* et krav, at varianserne, bortset fra observations-variansen, er specificeret. Der er dermed ikke mulighed for automatisk at estimere variansparametre.

Modellen fittes ved hjælp af funktionen *kfs*, som anvender itereret extended Kalman-filter og -smoother på dataene. Funktionen har udseendet

```
kfs(ssm, maxiter=50, epsilon=1e-06, debug=FALSE) .
```

Nedenfor følger en beskrivelse af argumenterne.

- **ssm** En state space model.
- **maxiter** Et positivt tal der angiver det maksimale antal iterationer af Kalman-filteret.
- **epsilon** Et positivt, lille tal der angiver konvergenskriteriet.
- **debug** Har værdierne sandt eller falsk. Giver mulighed for at følge hvert skridt i algoritmen.

Outputtet er de udglattede estimater for tilstandsvektoren, $\tilde{\mathbf{m}}_t$, med varians-estimat, $\underline{\underline{\tilde{C}}}_t$. Derudover beregnes de fittede værdier på logaritmeskalaen, $\mathbf{F}_t^T \tilde{\mathbf{m}}_t$, observationsvariansen, V_t , samt de reviderede observationer $\tilde{y}_t$. Endelig kan det ses, hvor mange iterationer af algoritmen der er kørt samt log-likelihooden for modellen.

4.5 Estimation af parametre

Jeg har valgt ikke at beskæftige mig med estimation af ukendte parametre i en DGLM, som er noget vanskeligere end i det Gaussiske tilfælde. Der findes anvendelige metoder til estimation af parametre

i en DGLM som f.eks. *Markov Chain Monte Carlo* (MCMC), se [2], og *Importance sampling*, se [7]. I *sspir* beregnes log-likelihooden for den approksimerende Gaussiske model som nævnt i afsnit 4.4. For et datasæt af en begrænset størrelse er det derfor muligt at anvende numeriske estimeringsmetoder i *sspir*. Dette vil blive beskrevet nærmere i kapitel 8.

Del II

Anvendelses-del

Kapitel 5

Introduktion til anvendelses-delen

Anvendelses-delen i dette speciale omhandler en analyse af et datasæt, der består af antal daglige tilfælde af førstegangs-AMI i perioden 1983-1999 i Nordjyllands amt. Analysen vil blive beskrevet i de tre følgende kapitler, men først vil det blive forklaret nærmere, hvad en AMI er, og hvordan den opstår. Derudover vil det blive beskrevet, hvilke risikofaktorer der er for AMI, samt hvordan disse faktorer har indflydelse på sæsonvariationen af AMI-incidensen. Desuden vil datasættet blive beskrevet nærmere, og en problemformulering for anvendelsesdelen vil blive præsenteret.

5.1 Processen bag dannelsen af en AMI

Dette afsnit er baseret på [27].

En AMI opstår indirekte på grund af for højt indhold af kolesterol i blodet. Kolestrolen aflejres på indersiden af blodåren, og når laget bliver tykt nok, kan dele af kolesterol-laget gå i stykker og rive lidt af blodårevæggen med, hvormed der opstår et lille sår på indersiden af blodåren.

Herefter går processen omkring sårheling i gang. Blodpladerne klumper sig sammen og danner en sårskorpe, og blodtilførslen til området omkring såret hæmmes. En AMI dannes, hvis sårskorpen bliver så stor, at den blokerer blodåren. I tilfælde af en AMI vil det væv, der har fået blokeret tilførslen af blod, dø i løbet af et par timer. I nogle tilfælde river sårskorpen sig løs og sætter sig fast et andet sted i åren, typisk hvor der er en indsnævring, hvorved AMI'en opstår på et nyt sted i åren.

5.1.1 Risikofaktorer

Det vides, at en for fed kost kan forhøje kolesterol-indholdet i blodet. AMI'er opstår på grund af kolesterol-aflejringer i blodårene, men sandsynligheden, for at det udvikler sig til en AMI, er markant større for rygere, idet rygning får blodet til at klumpe sig mere sammen end normalt. Man kan desuden være arvelig disponibel for at udvikle AMI'er. Andre risikofaktorer er forhøjet blodtryk, manglende motion, diabetes, stress og depression. Desuden er alder en risikofaktor, idet ca. halvdelen af alle hjertepatienter er 65 år eller derover. Yngre mænd har højere risiko for AMI end yngre kvinder, men til gengæld har ældre kvinder en større dødelighed af hjertesygdomme end ældre mænd.

Dette er de væsentligste risikofaktorer. I næste afsnit vil yderligere, årstidsbestemte faktorer blive beskrevet.

5.2 Årsager til sæsonvariationen af AMI

En amerikansk undersøgelse, [23], af 259891 AMI-tilfælde i perioden 1. Juli 1994 til 31. juli 1996 viste, at der blev registreret 53% flere AMI'er om vinteren (21. december til 19. marts) end om sommeren (21. juni til 22. september). I f.eks. [28] og [12] beskrives årsagerne til, at der er årstidsvariation for AMI'er. Forklaringen er, at bestemte risikofaktorer forekommer hyppigere om vinteren.

- **Lav temperatur:** Koldt vejr betyder, at blodpladerne klumper sig mere sammen, hvilket forøger risikoen for en AMI. Desuden forårsager kulde, at blodkarene trækker sig sammen og bliver tyndere, hvilket får blodtrykket til at stige.
- **Infektioner:** Der er størst risiko for at pådrage sig en infektion om vinteren. Infektioner forårsager betændelse, hvilket øger risikoen for, at kolesterol-aflejringer på indersiden af årene river sig løs. Desuden får patienter med influenza eller lungebetændelse ofte høj feber, hvilket er belastende for hjertet.
- **Mangel på motion:** Folk dyrker mindre motion, når der er kulde og sne udenfor. I varme klimaer menes det dog, at der dyrkes mere motion om vinteren. Desuden må det forventes, at mindre motion ikke lige med det samme vil resultere i en øget risiko for AMI.
- **Hård fysisk aktivitet i kulde:** Det menes, at hård fysisk anstrengelse i koldt vejr, som f.eks. sneskovling, kan medføre AMI for personer, der ikke er i god fysisk form. Snescovling er specielt risikabelt, da det som regel finder sted om morgenen, og undersøgelser har vist, at på netop det tidspunkt af dagen er mennesket mest sårbar overfor AMI.

- **Dårlig kost:** I forbindelse med julen bliver der holdt adskillige sammenkomster, og der fortæres ofte en større mængde mad end normalt. Da der tilmed ofte er tale om måltider med et stort fedtindhold, menes der at være en risiko, for at det kan udløse en AMI. Desuden indtages ofte en større mængde alkohol ved højtider. Det menes, at små mængder alkohol kan reducere risikoen for AMI, men indtages der for meget, kan det give forhøjet blodtryk og en abnormal hjerterytme.
- **Depression og stress:** Mange mennesker føler sig ensomme i julen, hvilket kan udløse depression. Omvendt er der også mange, der føler sig stressede i julen, fordi de har for mange aftaler.
- **Rygning:** Undersøgelser har vist, at tobak-salget toppe i julen.
- **Kolesterol:** Indholdet af kolesterol i blodet toppe i december for mænd og i januar for kvinder.
- **Sol-lys:** Det er dokumenteret, at sol-lys på flere områder er en vigtig faktor for mennesket. Muligvis kan mangel på sol-lys forøge risikoen for AMI. Derudover er manglende sol-lys en kilde til depression.

Der er ikke noget entydigt svar på spørgsmålet, om der også i varme klimaer er en øget forekomst af AMI om vinteren. Ifølge [28] er der på Hawaii registreret flest AMI' er i januar, og en undersøgelse fra Malta, se [11], viste samme billede. På den anden side viste en taiwanesisk undersøgelse, se [16], af 540 tilfælde ingen sæsonvariation i områder med subtropisk klima. En undersøgelse fra Kuwait, se [1], viste tilmed en større hyppighed af AMI om sommeren end om vinteren, hvilket indikerer, at et meget varmt og tørt klima kan forøge risikoen for AMI. I [11] rapporteres om større sæsonvariation af AMI hos kvinder end hos mænd.

5.3 Ugedagsvariation af AMI

Ifølge [28] viser amerikanske undersøgelser, at der er 20% større risiko for AMI om mandagen end for de øvrige dage i ugen. Dette understøttes af [16], hvor der rapporteres, at risikoen er 9% større om mandagen. Årsagen til det større antal AMI-tilfælde om mandagen menes at være stress. En anden forklaring kunne være, at personer, der får symptomer for AMI i weekenden, ofte venter til om mandagen med at gå til lægen.

5.4 Beskrivelse af datasæt

I dette afsnit beskrives datasættet, som analyseres i anvendelsesdelen af specialet. Datasættet er tidligere blevet analyseret i [10], og det er herfra, at de følgende oplysninger er hentet.

5.4.1 Registrering af patienter

Nordjyllands amts registreringssystem af personer indlagt på de syv nordjyske hospitaler blev etableret i 1977. Informationer, som CPR-nummer, indlæggelsesdato, diagnose, og eventuelle kirurgiske indgreb, blev registreret i en database. En komplet registrering blev etableret d. 1. januar 1980. AMI-datasættet består af det daglige antal AMI-patienter inddelt efter alder og køn i perioden fra 1. januar 1980 til 31. december 2001 og indeholder kun personer, der ikke tidligere havde været indlagt med en AMI. Det vil sige, at det kun er antallet af nye AMI-tilfælde, også kaldet *incidensen*, der registreres. Idet registreringen af AMI startede systematisk i 1980, er der en større risiko, for at denne betingelse ikke er fuldstændig overholdt i årene 1980-1982, og derfor er disse år udeladt af analysen. Efter introduktionen af medikamentet *troponin T* i 2000 blev der observeret en markant stigning i antallet af erkendte tilfælde. Det blev antaget i [10], at dette skyldes en ændring i diagnostiseringsproceduren, og derfor er årene 2000 og 2001 ikke inkluderet i analysen.

5.4.2 Aldersstandardisering

For at kunne vurdere udviklingen i incidensen af AMI er det nødvendigt at tage højde for befolknings-sammensætningen, idet risikoen for at få en AMI stiger med alderen. Der vil derfor helt naturligt være en højere incidens i en gruppe med høj gennemsnitsalder end i en gruppe, hvor gennemsnitsalderen er lav. En ændret alderssammensætning over tid giver anledning til en ændret forekomst, uanset om incidensen pr. 100000 (incidensraten) er konstant i hver aldersgruppe. Da befolknings-sammensætningen ændrer sig over tid, er det nødvendigt at beregne den *aldersstandardiserede* incidensrate. Det vil sige, at befolkningssammensætningen på en bestemt dag vælges som standard. For mere information om aldersstandardisering og beslægtede emner, se [21].

Incidensraten pr. person på dag nr. t er givet ved y_t/N_t , hvor y_t angiver incidensen på dag nr. t , og N_t angiver antallet af personer på dag nr. t . Det følger, at incidensen på dag nr. t , standardiseret efter

befolknings-sammensætningen den 1. januar 1983, er givet ved

$$A(t) = \sum_{i=1}^m \frac{y_t^{(i)}}{N_t^{(i)}} N_{83}^{(i)},$$

hvor $y_t^{(i)}$ angiver incidensen på dag nr. t for aldersgruppe i , $N_t^{(i)}$ angiver antallet af personer i aldersgruppe i på dag nr. t , og $N_{83}^{(i)}$ angiver antallet af personer i aldersgruppe i den 1. januar 1983, mens m er antallet af aldersgrupper.

Den aldersstandardiserede incidensrate pr. 100000 er da givet ved

$$\lambda(t) = \frac{\sum_{i=1}^m \frac{y_t^{(i)}}{N_t^{(i)}} N_{83}^{(i)}}{N_{83}} \cdot 100000 = \frac{\sum_{i=1}^m \lambda(t)^{(i)} N_{83}^{(i)}}{N_{83}},$$

hvor N_{83} angiver det totale antal personer den 1. januar 1983, og $\lambda(t)^{(i)}$ angiver incidensraten på dag nr. t for aldersgruppe i .

5.4.3 Resultater

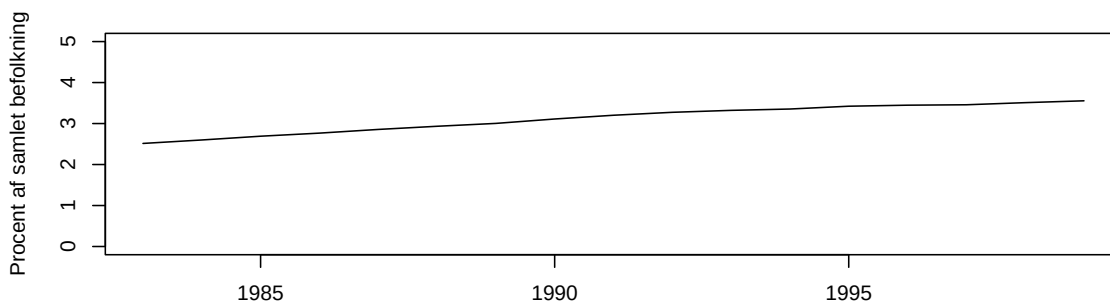
Datasættet omfatter i alt 17989 patienter, som er inddelt i fire aldersgrupper (0-59 år, 60-69 år, 70-79 år og 80 år eller derover) og efter køn. Nedenfor ses fordelingen af antallet af patienter i grupperne.

Gruppe	Incidens
Mænd: 0-59 år	3040
Mænd: 60-69 år	3297
Mænd: 70-79 år	3604
Mænd: over 79 år	1607
Kvinder: 0-59 år	852
Kvinder: 60-69 år	1479
Kvinder: 70-79 år	2300
Kvinder: over 79 år	1810

Den årlige incidens er faldet fra 1352 i 1983 til 804 i 1999, hvilket svarer til et fald på lidt over 40%. Figur 5.1 viser udviklingen af incidensraten pr. 100000 med og uden aldersstandardisering. Figuren viser, at den aldersstandardiserede incidensrate er faldet mest, hvilket indikerer, at andelen af ældre mennesker er vokset i løbet af undersøgelsesperioden. Dette understøttes af figur 5.2, som viser, at den procentvise andel af personer over 79 år i Nordjyllands amt er steget med ca. 40% i løbet af perioden.



Figur 5.1: Den optrukne linie angiver den årlige incidenrate af AMI pr. 100000 i Nordjyllands amt i perioden 1983-1999, mens den stiplede linie angiver den årlige incidensrate af AMI pr. 100000 aldersstandardiseret efter befolknings-sammensætningen i 1983.

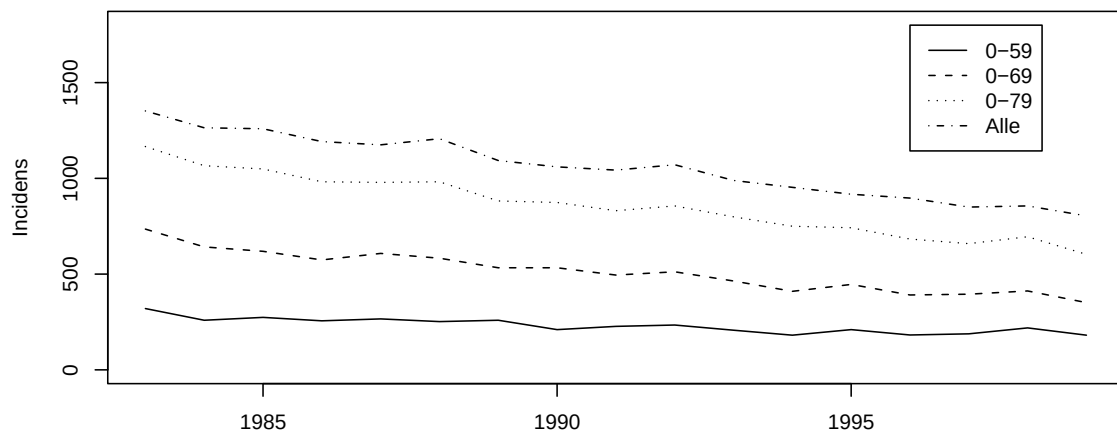


Figur 5.2: Den procentvise andel af personer over 79 år i Nordjyllands amt i perioden 1983-1999.

Det ses på figur 5.3, at der er sket et markant større fald i incidensen for mænd end for kvinder. Det hører dog med, at mændene i starten af perioden havde en næsten dobbelt så stor incidens som kvinderne.



Figur 5.3: Den optrukne linie angiver den årlige incidens af AMI for mænd og den stiplede linie den årlige incidens af AMI for kvinder i perioden 1983-1999.



Figur 5.4: Årlig incidens af AMI afhængig af alder.

Figur 5.4 viser, at incidensen for personer over 79 år har været stort set uændret gennem undersøgelsesperioden til forskel fra incidensen i de øvrige aldersgrupper, som er faldet. Dette skyldes stigningen

i andelen af personer i denne aldersgruppe som illustreret i figur 5.2. I næste kapitel vil det blive beregnet, hvor meget incidensraten pr. 100000 falder i gennemsnit pr. år for de forskellige grupper.

5.5 Problemformulering for data-analysen

Som beskrevet i afsnit 5.4.3 er incidensen af AMI faldet ca. 40% i løbet af undersøgelsesperioden. Data-analysen i dette speciale vil først og fremmest søge at besvare spørgsmålet, om der samtidig er sket en ændring af årstidsvariationen af dataene. Mere præcist er det interessant at undersøge, om faldet i incidens har været mere markant for nogle årstider end for andre, da det kan være med til at give et fingerpeg om, hvilket faktorer der især har indflydelse på incidensen. Hvis det f.eks. blev observeret, at der i december er sket et større fald af incidensen end i de øvrige måneder, kunne det hænge sammen med, at nordjyderne er begyndt at spise sundere i julen, og at kosten har stor betydning for incidensen.

Det vil desuden blive undersøgt, om der er forskelle i sæsonvariation mellem aldersgrupperne og kønnene. Dette kan give en ide om, hvilke risikofaktorer der er centrale for en specifik gruppe. Derudover vil trenden og ugedagsvariationen af dataene blive analyseret.

I det følgende kapitel vil AMI-dataene blive analyseret med en statisk Poisson regressionsmodel med det formål at efterprøve de opnåede resultater i [10]. Dataene analyseres ved hjælp af en basic structural model i kapitel 7, som skal ses som en indledning til den dynamiske data-analyse. De estimerede varianser herfra vil blive benyttet i kapitel 8, hvor dataene analyseres med en dynamisk Poisson regressionsmodel. Herefter vil resultaterne fra den statiske og dynamiske data-analyse blive sammenlignet og diskuteret.

Kapitel 6

Data-analyse med Statisk Poisson Model

I dette kapitel gennemføres en analyse af AMI-datasættet, der er beskrevet i afsnit 5.4, med en GLM, hvilket har til formål at efterprøve resultaterne fra [10]. Der vil blive foretaget en grundig modelkontrol med henblik på at finde argumenter for anvendelsen af en dynamisk model. Dataene analyseres ved hjælp af funktionen *glm* i programmet *R* v. 2.1.1.

6.1 Opstilling af model

Der skal opstilles en model, der kan beskrive trenden samt sæson- og ugedagsvariationen af AMI-dataene.

Antallet af personer i Nordjyllands amt, der bliver indlagt på grund af en førstegangs-AMI på en given dag, t , antages at være binominal-fordelt med meget stor antalsparameter, $n(t)$, og meget lille sandsynlighedsparameter, $p(t)$, (som naturligvis er ukendt). Der gælder, at en binominal-fordeling med stort n og lille p kan approksimeres ved en Poisson-fordeling med intensitet, np , se f.eks. [24] for et bevis. Under antagelse af at observationerne er uafhængige kan dataene analyseres med en GLM. Incidensen af AMI på dag nr. t , $y(t)$, kan derfor approksimeres ved en Poisson-fordeling med midelværdiparameter, $\mu(t) = n(t)p(t)$. For at kunne måle incidensen i forhold til befolkningsstørrelse omskrives dette til

$$\mu(t) = \left(\frac{n(t)}{100000} \right) \lambda(t),$$

hvor $\lambda(t)$ er den forventede incidensrate pr. 100000 for dag nr. t . De systematiske variationer i $\lambda(t)$ opdeles i tre komponenter. En basiskomponent, $\alpha(t)$, der afhænger af ugedagen, en komponent, $\gamma(t)$,

der beskriver sæsonvariationen, og en komponent, $q(t)$, der beskriver trenden af dataene. I overensstemmelse med praksis for analyse med en GLM benyttes en log-lineær model, der er givet ved

$$\ln \lambda(t) = \alpha(t) + \gamma(t) + q(t). \quad (6.1)$$

Ugedagsvariationen beskrives som i (3.19), idet $\sigma_\delta^2 = 0$.

$$\alpha(t) = -\alpha(t-1) - \alpha(t-2) - \dots - \alpha(t-6).$$

I [10] blev trenden modelleret lineært, men her ønskes det undersøgt, om trenden kan beskrives bedre ved et 9. grads polynomium. Det vil sige, at trendkomponenten indledningsvis er givet ved

$$q(t) = a_0 + a_1 t + a_2 t^2 + \dots + a_9 t^9.$$

Sæsonvariationen, $\gamma(t)$, er givet ved en Fourier-række (se afsnit 3.6), der i den indledende model har grad fire.

$$\gamma(t) = \sum_{j=1}^4 \gamma_{c,j} \cos(\tau_j t) + \gamma_{s,j} \sin(\tau_j t),$$

hvor $\tau_j = 2\pi j / 365.25$. At beskrive sæsonvariationen ved en Fourier-række er oplagt, idet det giver mulighed for at undersøge om sæsonvariationen mellem to grupper er ens. Dette gøres ved at teste hypotesen, at koefficienterne for de trigonometriske led er ens. Andre modeller til beskrivelse af sæsonvariation er f.eks. *generaliserede additive modeller* (GAM), som til forskel fra den almindelige GLM kan bestå af glatte funktioner af en eller flere variable, se f.eks. [14]. Analyse af data ved hjælp af en GAM er eksempelvis beskrevet i [22].

Da

$$\ln \mu(t) = \ln \left(\frac{n(t)}{100000} \right) + \ln \lambda(t),$$

ses det af (6.1), at

$$\ln \mu(t) = \alpha(t) + \gamma(t) + q(t) + \ln \left(\frac{n(t)}{100000} \right). \quad (6.2)$$

Denne model vil blive anvendt i analysen af dataene. Udover en analyse af den samlede befolkning vil der blive foretaget stratificeret analyse efter køn og alder.

6.2 Modelreduktion

Det undersøges, om modellen i (6.2) kan reduceres ved at foretage Wald- og devians-tests. Der benyttes et signifikansniveau på 5% for begge teststørrelser, som er beskrevet i afsnit 2.4. Funktionen

summary i R, der har en model som argument, giver et output, der for hver af de forklarende variable viser p -værdien fra en Wald-test af, om variabelen kan udelades fra modellen. Wald-testen benyttes til at give en indikation af, hvor meget graden af henholdsvis polynomiet og Fourier-rækken kan reduceres, samt om ugedagsvariationen er signifikant. Modellen reduceres på baggrund af p -værdierne, og der foretages en devianstest af den gamle model mod den nye for at sikre, at der er grundlag for at reducere modellen.

Modelreduktionen af modellen for den samlede befolkning gav resultatet, at sæsonvariations-komponenten kan reduceres til en Fourier-række af grad to, og at polynomiet kan reduceres til et lineært led. Derved fremkommer følgende model.

$$\ln \mu(t) = \alpha(t) + \gamma(t) + \beta t + \ln \left(\frac{n(t)}{100000} \right), \quad (6.3)$$

hvor

$$\gamma(t) = \gamma_{c,1} \cos(\tau_1 t) + \gamma_{s,1} \sin(\tau_1 t) + \gamma_{c,2} \cos(\tau_2 t) + \gamma_{s,2} \sin(\tau_2 t).$$

Modellen adskiller sig fra den anvendte model i [10], hvor sæsonvariations-komponenten bestod af en Fourier-række af grad fire. Stratificeret på køn og alder gav modelreduktionen helt tilsvarende modeller bortset fra de 70-79 årige, hvor sæsonvariationen ikke er signifikant.

Nedenstående tabel giver et billede af, i hvor høj grad modellens tre komponenter hver især forklarer variationen af AMI-incidensraten.

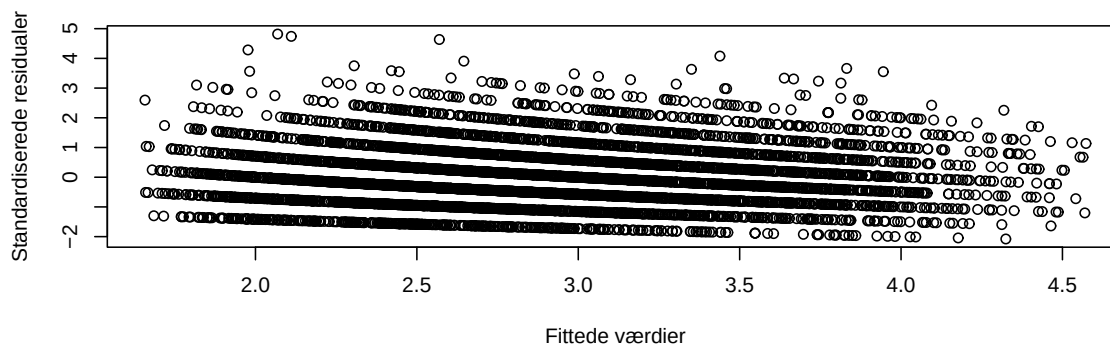
Komponent	Reduktion i devians
Sæson	36
Ugedag	256
Trend	448

Udgangspunktet er en *nul-model*, der fitter data ved det empiriske gennemsnit. Tabellen viser, hvor meget deviansen reduceres ved inkludering af henholdsvis sæson-, ugedags- og trend-komponenten i modellen. Det vil sige, at i en model, hvor kun sæson-komponenten indgår, reduceres deviansen med 36 i forhold til nul-modellen. Det ses, at trend-komponenten giver den største forklaring på variationen af AMI-dataene, mens sæson-komponenten forklarer mindst.

6.3 Modelkontrol

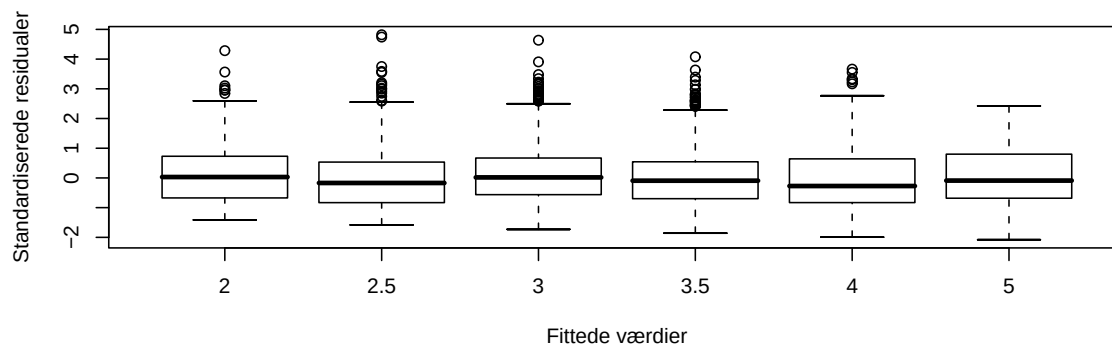
Inden resultaterne for modellen i (6.3) beskrives, kontrolleres det, om de antagelser, som modellen bygger på, er opfyldt. Kontrollen er baseret på teorien, der er beskrevet i afsnit 2.5.

6.3.1 Varianshomogenitet



Figur 6.1: Fittede værdier plottet mod standardiserede residualer. Plottet er svært at aflæse, da der er tale om Poisson-observationer med lille intensitet.

Figur 6.1 viser fittede værdier for modellen plottet mod standardiserede residualer. Plottet er svært at aflæse på grund af den lille intensitet. I stedet opdeles de fittede værdier i intervaller, og fordelingen af de tilsvarende residualer angives ved et boxplot, se figur 6.2.

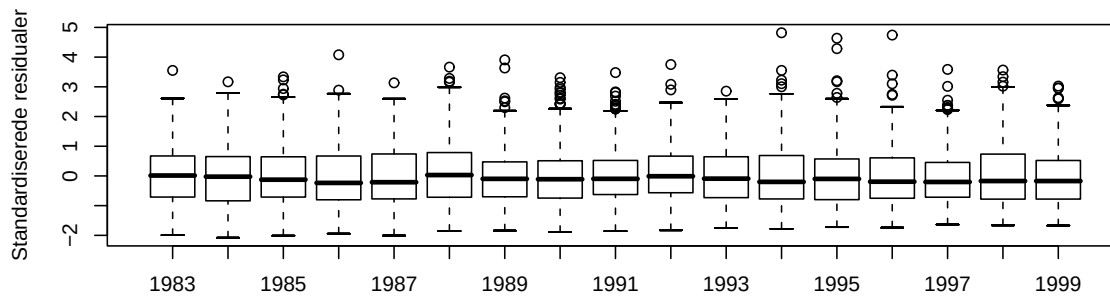


Figur 6.2: Fittede værdier plottet mod standardiserede residualer. Boksen, der svarer til værdien 3 på x -aksen, beskriver fordelingen af residualerne for observationerne med fittede værdier i intervallet 2.5-3.0.

Plottet indikerer ikke, at variansen er inhomogen. Plottets asymmetriske udseende skyldes de lave intensiteter. Som en ekstra kontrol udregnes de empiriske varianser af de standardiserede residualer svarende til forskellige intervaller af fittede værdier.

$\hat{\mu}_t$	$\text{Var}(\hat{r}_t)$
≤ 2.0	1.09
2.0-2.5	1.05
2.5-3.0	1.00
3.0-3.5	1.03
3.5-4.0	0.97
> 4.0	1.02

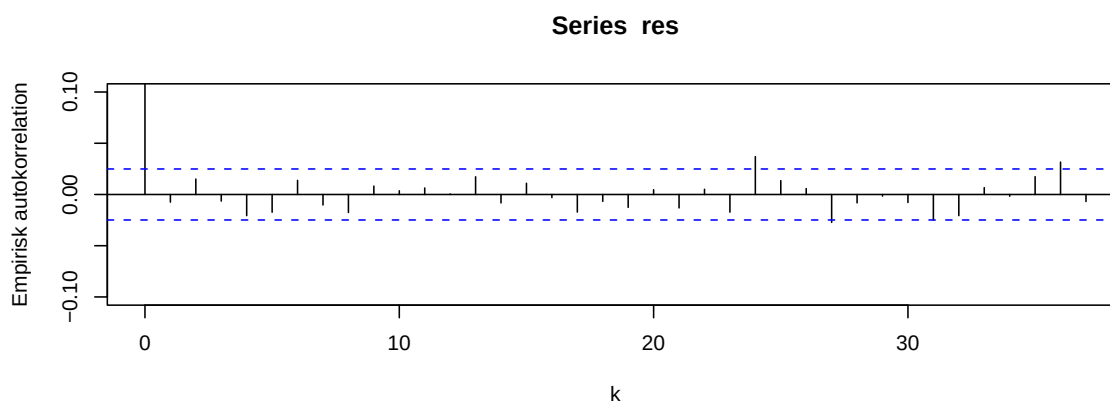
Det ses, at varianserne er lidt for store for små forventede værdier, men det er ikke noget, der indikerer inhomogen varians.



Figur 6.3: Fordeling af standardiserede residualer for alle år i undersøgelsesperioden.

For at undersøge om modellen beskriver dataene lige godt gennem hele undersøgelsesperioden, samles residualerne sammen på år og undersøges ved et box-plot. Figur 6.3 viser, at de standardiserede residualer har nogenlunde konstant fordeling gennem undersøgelsesperioden.

6.3.2 Seriel korrelation



Figur 6.4: Plot af empirisk autokorrelations-funktion for forskellige k .

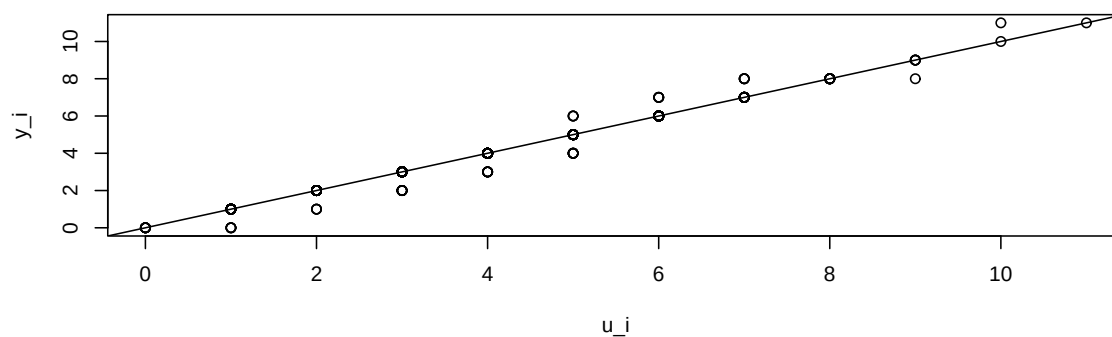
På figur 6.4 ses et plot, der angiver autokorrelationen mellem de standardiserede residualer. Det ses, at der ikke er tale om seriel korrelation, idet autokorrelations-funktionen befinder sig indenfor 95% konfidensintervallet for næsten alle k .

6.3.3 Fordelingstype

Nedenfor ses en tabel over de empiriske middelværdier og varianser af observationerne, hvis forventede værdier ligger i intervallet $(\xi - 0.2, \xi + 0.2)$.

ξ	Empirisk middelværdi	Empirisk varians
2.0	1.99	2.14
2.5	2.52	2.67
3.0	3.02	3.05
3.5	3.49	3.42
4.0	3.95	4.32

Det ses, at observationerne med approksimativt middelværdi 4 har en lidt for stor varians. Derfor foretages et fraktilplot for at undersøge situationen nærmere.



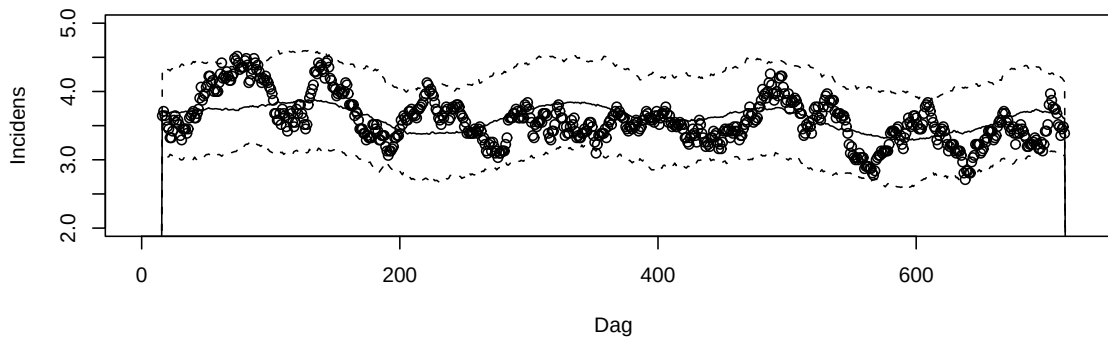
Figur 6.5: Fraktilplot af observationer med forventet værdi 4.

Figur 6.5 viser et fraktilplot af observationer med forventet værdi 4. Det ses, at observationerne tilnærmelsesvis ligger på en ret linie gennem origo med hældning 1, hvilket indikerer, at deres fordeling

er korrekt specificeret.

6.3.4 Kontrol ved simulering

Det kan undersøges, om modellen beskriver data godt ved at simulere en række Poisson-fordelinger udfra de fittede værdier. For hvert datapunkt simuleres 100 Poisson-fordelte observationer med middelværdi $\hat{\mu}_t$. Derved fås 100 Poisson-fordelte data-sæt med samme intensiteter. For hvert datasæt beregnes et glidende gennemsnit og derudfra bestemmes den empiriske middelværdi og varians for hvert datapunkt baseret på de 100 udglattede observationer. Det kan kontrolleres, om AMI-dataene er plausible i forhold til deres estimerede middelværdier ved at plotte dem sammen med 95% konfidensintervallet af de udglattede, simulerede Poisson-fordelte observationer.



Figur 6.6: Den optrukne linie angiver den empiriske middelværdi af de udglattede, simulerede observationer med intensitet $\hat{\mu}_t$ for de første to år, mens den prikkede linie angiver en approksimation af 95% konfidensintervallet for observationerne. Prikkerne angiver glidende gennemsnit af AMI-dataene.

Det ses på figur 6.6, at langt størstedelen af AMI-dataene ligger indenfor 95% konfidensintervallet for de simulerede data, hvorfor det ikke kan konkluderes, at modellen er upassende.

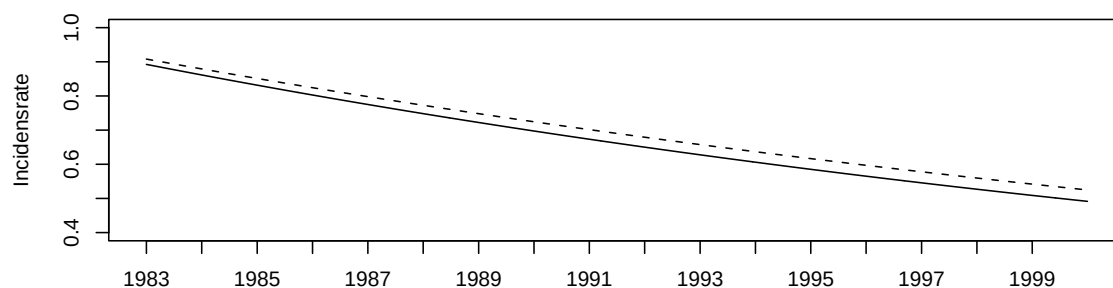
Modelkontrollen har dermed vist, at modellen er anvendelig. Stratificeret på køn og alder gav modelkontrollen samme resultat.

6.4 Resultater

I dette afsnit beskrives resultaterne af data-analysen med en GLM for henholdsvis trend, sæson- og ugedagsvariation.

6.4.1 Trend

Samtlige modeller kunne reduceres til at have en lineær trendkomponent. Den aldersstandardiserede trend viser udviklingen af den aldersstandardiserede incidensrate, som er defineret i afsnit 5.4.2.



Figur 6.7: Den optrukne linie angiver trenden for AMI-incidensraten pr. 100000 standardiseret efter befolknings-sammensætningen den 1. januar 1983, mens den stiplede linie angiver trenden uden aldersstandardisering.

Det ses i figur 6.7, at den aldersstandardiserede incidensrate er faldet mere end den ikke-standardiserede, hvilket også var resultatet ved analysen af de rå data i afsnit 5.4.3. Dette skyldes som tidligere vist, at andelen af personer over 79 år er vokset i løbet af undersøgelsesperioden.



Figur 6.8: Den optrukne linie angiver den forventede årlige incidensrate, aldersstandardiseret efter befolkningssammensætningen i 1983, mens den stiplede linie angiver den observerede årlige incidensrate, aldersstandardiseret efter befolkningssammensætningen i 1983.

Figur 6.8 viser den forventede aldersstandardiserede incidensrate plottet sammen med den observerede. Det ses, at modellen giver en forholdsvis god beskrivelse af trenden for AMI-dataene. Det ses dog, at incidensenraten var større end forventet i 1988 og 1992.

Forventet relativt fald af incidensraten

I tilfælde af en log-lineær trend kan det forventede relative fald af incidensraten pr. år beregnes på simpel vis. Lad β benævne koefficienten for trendkomponenten. Da gælder der, at

$$365\beta = \ln[\lambda(t + 365)] - \ln[\lambda(t)] \Rightarrow \exp(365\beta) = \frac{\lambda(t + 365)}{\lambda(t)}.$$

Det følger, at det relative fald af incidensraten pr. år er givet ved

$$\frac{\lambda(t) - \lambda(t + 365)}{\lambda(t)} = 1 - \exp(365\beta)$$

og kan estimeres ved $1 - \exp(365\hat{\beta})$.

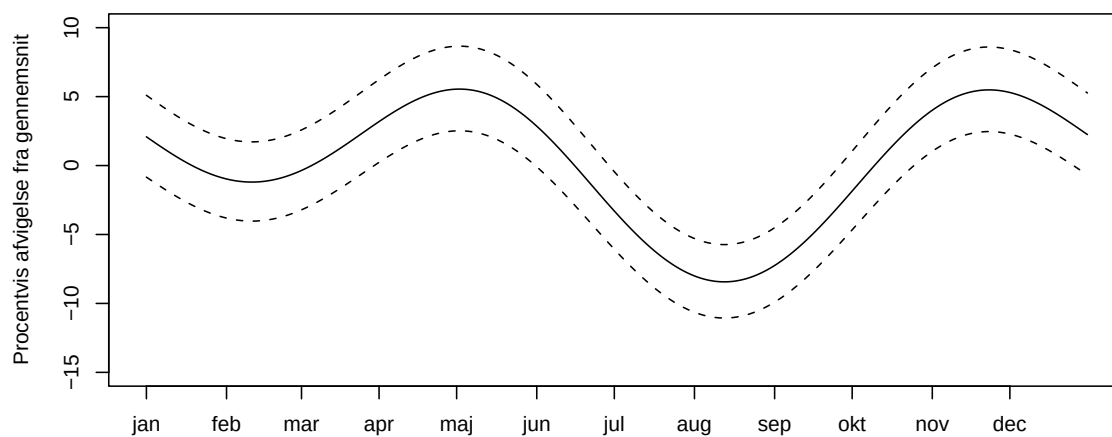
Nedenfor ses en tabel, der viser det forventede relative årlige fald af incidensraten inklusive 95% konfidensinterval for den samlede befolkning og de køns- og alders-opdelte grupper.

Gruppe	Fald pr. år (%)	95% KI
Samlet	3.2	2.9-3.5
Mænd	3.6	3.3-4.0
Kvinder	2.3	1.8-2.8
0-59 år	3.1	2.5-3.7
60-69 år	4.1	3.5-4.6
70-79 år	3.5	3.0-4.0
Over 79 år	2.7	2.0-3.4

Det ses, at der har været et betydeligt større fald i incidensraten blandt mænd end blandt kvinder, og at de 60-69 årige er den aldersgruppe, der har oplevet det største fald. For flere grupper er der en afvigelse på 0.1 procentpoint i forhold til resultaterne i [10], hvilket muligvis skyldes forskellen i modellering af sæsonvariation.

6.4.2 Sæsonvariation

Sæsonvariationen for AMI-incidensraten er vist i figur 6.9.



Figur 6.9: Sæsonvariation for AMI-incidensraten pr. 100000 med 95% konfidensinterval.

Det ses, at incidensraten er størst i april/maj med et toppunkt på ca. 5.5% over gennemsnittet og i november/december ca. 5% over gennemsnittet. Incidensraten er mindst i august ca. 8.5% under gennemsnittet. En tilsvarende undersøgelse af sæsonvariationen i [10] viste toppunkter i maj og november/december ca. 6% over gennemsnittet og et lavpunkt i juli/august ca. 8% under gennemsnittet. Derudover er kurven i figur 6.9 mere glat end i [10], hvilket skyldes det færre antal trigonometriske led.

Ændring af sæsonvariation over tid

Et af hovedformålene for data-analysen er at undersøge, om sæsonvariationen har ændret sig gennem undersøgelsesperioden.

Det afprøves, om der kan inkluderes en vekselvirkning mellem tidsperiode og sæsonvariations-komponenten, idet dataene inddeles i perioderne 1983-1986, 1987-1990, 1991-1994 og 1995-1999. En test af forskellen i devians mellem modellerne med og uden vekselvirkning gav en p -værdi på 0.1. Idet signifikansniveauet er sat til 5%, indikerer dette, at der ikke er en signifikant sammenhæng mellem sæsonvariation og tidsperiode.

Derfor kan det på grundlag af denne type modeller konkluderes, at der ikke er sket en signifikant ændring af sæsonvariationen gennem undersøgelsesperioden. I de to følgende kapitler vil dette blive undersøgt ved hjælp af dynamiske modeller.

Sæsonvariation efter alder og køn

Til at undersøge om der er signifikant forskel på sæsonvariationen af incidensraten for de fire aldersgrupper samt kønnene foretages en Wald-test.

Lad β_i være $1 \times p$ parametervektoren for sæsonvariations-komponenten i modellen for aldersgruppe i . Hypotesen, der skal testes, er

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4,$$

hvilket svarer til

$$H_0 : \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_1 - \beta_3 \\ \beta_1 - \beta_4 \end{bmatrix} = \mathbf{0}.$$

Under antagelse af H_0 følger det af (2.14), og af at $\hat{\beta}_i$ 'erne er uafhængige, at

$$\begin{bmatrix} \hat{\beta}_1 - \hat{\beta}_2 \\ \hat{\beta}_1 - \hat{\beta}_3 \\ \hat{\beta}_1 - \hat{\beta}_4 \end{bmatrix} \sim \mathcal{N}_{3p}(\mathbf{0}, \underline{\underline{V}}),$$

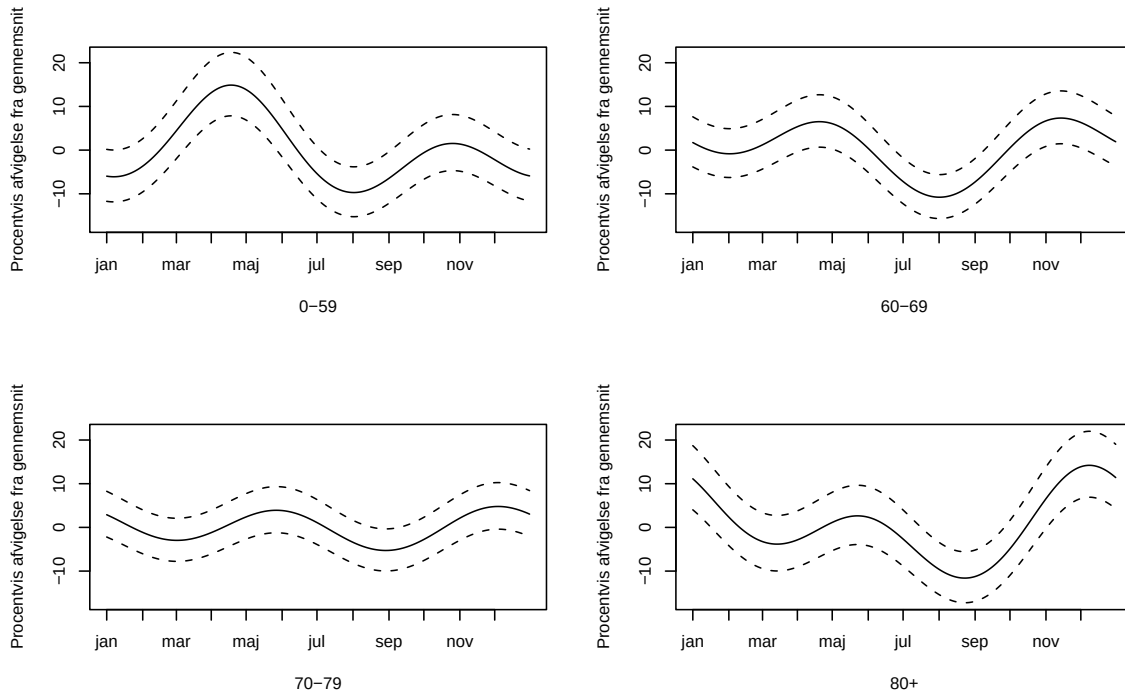
hvor

$$\underline{\underline{V}} = \begin{bmatrix} \underline{\underline{\gamma}}^{-1}(\hat{\beta}_1) + \underline{\underline{\gamma}}^{-1}(\hat{\beta}_2) & \underline{\underline{\gamma}}^{-1}(\hat{\beta}_1) & \underline{\underline{\gamma}}^{-1}(\hat{\beta}_1) \\ \underline{\underline{\gamma}}^{-1}(\hat{\beta}_1) & \underline{\underline{\gamma}}^{-1}(\hat{\beta}_1) + \underline{\underline{\gamma}}^{-1}(\hat{\beta}_3) & \underline{\underline{\gamma}}^{-1}(\hat{\beta}_1) \\ \underline{\underline{\gamma}}^{-1}(\hat{\beta}_1) & \underline{\underline{\gamma}}^{-1}(\hat{\beta}_1) & \underline{\underline{\gamma}}^{-1}(\hat{\beta}_1) + \underline{\underline{\gamma}}^{-1}(\hat{\beta}_4) \end{bmatrix}.$$

Det følger, at

$$\begin{bmatrix} \hat{\beta}_1 - \hat{\beta}_2 \\ \hat{\beta}_1 - \hat{\beta}_3 \\ \hat{\beta}_1 - \hat{\beta}_4 \end{bmatrix}^T \underline{\underline{V}}^{-1} \begin{bmatrix} \hat{\beta}_1 - \hat{\beta}_2 \\ \hat{\beta}_1 - \hat{\beta}_3 \\ \hat{\beta}_1 - \hat{\beta}_4 \end{bmatrix} \sim \chi_{3p}^2.$$

Som beskrevet i afsnit 6.2 kunne sæsonvariations-komponenten reduceres til en Fourier-række af grad to for tre af de fire aldersgrupper ligesom for den samlede befolkning. For de 70-79 årige er sæsonvariationen ikke signifikant, hvilket allerede er en indikation af, at sæsonvariationen afhænger af alder. For at kunne teste H_0 inkluderes alligevel en Fourier-række af grad to til beskrivelse af sæsonvariationen i modellen for de 70-79 årige. Wald-testen viste, at forskellene i sæsonvariation mellem aldersgrupperne er signifikante på 2%-niveauet. Sæsonvariationen for de fire aldersgrupper er vist i figur 6.10.



Figur 6.10: Sæsonvariation af AMI-incidensraten for de fire aldersgrupper.

Det ses, at for alle fire aldersgrupper når sæsonvariationen et toppunkt om foråret. For de to yngste grupper ligger toppunktet i april/maj, mens det for de ældste indtræffer i maj/juni. Desuden bliver toppunktet markant mindre med stigende alder. Sæsonvariationen når et nyt toppunkt i oktober/november for den yngste aldersgruppe. Med stigende alder flytter dette toppunkt sig længere hen på året, således at det indtræffer i december for den ældste aldersgruppe. Der er en tendens til, at toppunktet bliver markant større med stigende alder bortset fra de 70-79 årige, som jo ikke har den store sæsonvariation. For de to yngste aldersgrupper når sæsonvariationen sit lavpunkt i juli/august, mens det for de ældste sker i august/september.

Undersøgelsen i [10] viste også en signifikant forskel på sæsonvariationen for aldersgrupperne. Der blev udledt de samme overordnede karakteristika for gruppernes sæsonvariation som her, men der er dog små forskelle, hvilket kan ses i tabellen nedenfor. Tabellen angiver med cirka-tal de procentvise afvigelser fra gennemsnittet for yderpunkterne i sæsonvariationen for de to undersøgelser.

	Gammel undersøgelse	Ny undersøgelse
0-59 år: Forårs-toppunkt	13%	15%
0-59 år: Sommer-lavpunkt	-14%	-10%
0-59 år: Efterårs-toppunkt	1%	2%
60-69 år: Forårs-toppunkt	8%	7%
60-69 år: Sommer-lavpunkt	-14%	-11%
60-69 år: Efterårs-toppunkt	8%	7%
70-79 år: Forårs-toppunkt	6%	4%
70-79 år: Sommer-lavpunkt	-8%	-5%
70-79 år: Efterårs-toppunkt	10%	5%
over 80 år: Forårs-toppunkt	2%	2%
over 80 år: Sommer-lavpunkt	-13%	-12%
over 80 år: Efterårs-toppunkt	15%	14%

En Wald-test viste, at der ikke er en signifikant forskel mellem sæsonvariationen for mænd og kvinder ($p = 0.19$), hvilket også var resultatet i [10].

6.4.3 Ugedagsvariation

I dette afsnit analyseres ugedagsvariationen for incidensraten, hvilket ikke er beskrevet i [10]. Først beskrives en generel metode til estimering af konfidensintervaller.

Lad $\hat{\alpha} \sim \mathcal{N}(\alpha, V)$ være en central estimator for en parameter, α , og lad g være en differentiabel funktion. Ved en Taylor-udvikling af $g(\hat{\alpha})$ omkring $g(\alpha)$ fremkommer at

$$g(\hat{\alpha}) \approx g(\alpha) + g'(\alpha)(\hat{\alpha} - \alpha).$$

Det følger, at $g(\hat{\alpha})$ approksimativt er normalfordelt med middelværdi $g(\alpha)$ og varians

$$\text{Var}[g(\hat{\alpha})] \approx [g'(\alpha)]^2 V. \quad (6.4)$$

Ved at indsætte $\hat{\alpha}$ i stedet for α følger det, at 95% konfidensintervallet for $g(\alpha)$ kan estimeres ved

$$\left(g(\hat{\alpha}) - 1.96|g'(\hat{\alpha})|\sqrt{V}, g(\hat{\alpha}) + 1.96|g'(\hat{\alpha})|\sqrt{V} \right).$$

Denne metode vil blive anvendt til at generere konfidensintervaller for plots af ugedagsvariation.

Incidensraten for den k -te ugedag er i procent givet ved

$$m(u_1, \dots, u_7) = 100 \cdot \frac{\exp(u_k)}{\exp(u_1) + \dots + \exp(u_7)},$$

hvor $\exp(u_i)$ angiver den forventede incidensrate for ugedag nr. i , idet der ses bort fra trend og sæsonvariation. Definer $\mathbf{u} = (u_1, \dots, u_7)^T$ og lad dens maksimum-likelihood estimator, $\hat{\mathbf{u}}$, have variansmatrix $\underline{\underline{V}}$. Da antallet af observationer er stort, gælder der ifølge teorien for GLM approksimativt, at

$$\hat{\mathbf{u}} \sim \mathcal{N}_7(\mathbf{u}, \underline{\underline{V}}).$$

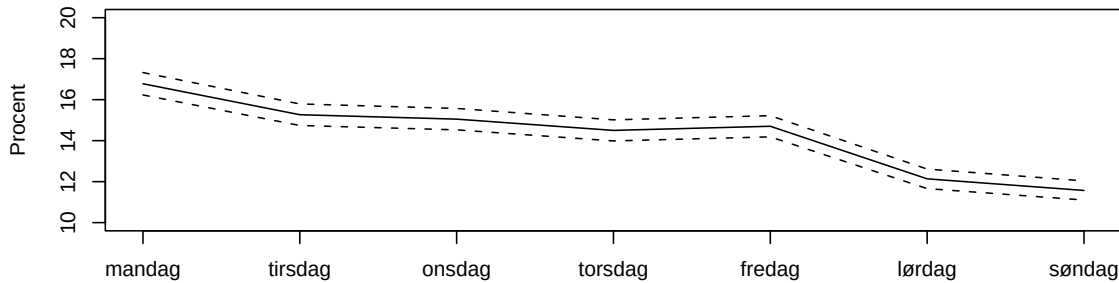
Definer $\mathbf{h}(\mathbf{u}) = \left(\frac{\partial m}{\partial u_1}, \dots, \frac{\partial m}{\partial u_7} \right)^T$. Det følger af (6.4), at

$$\text{Var}[m(\mathbf{u})] \approx \mathbf{h}(\mathbf{u})^T \underline{\underline{V}} \mathbf{h}(\mathbf{u}),$$

Det følger, at 95% konfidensintervallet for $m(\mathbf{u})$ approksimativt er givet ved

$$\left(m(\hat{\mathbf{u}}) - 1.96 \sqrt{\mathbf{h}(\hat{\mathbf{u}})^T \underline{\underline{V}} \mathbf{h}(\hat{\mathbf{u}})}, m(\hat{\mathbf{u}}) + 1.96 \sqrt{\mathbf{h}(\hat{\mathbf{u}})^T \underline{\underline{V}} \mathbf{h}(\hat{\mathbf{u}})} \right).$$

Figur 6.11 viser incidensraten af AMI fordelt på ugedag. Det ses, at incidensraten er størst om mandagen, hvor ca. 17% af tilfældene forventes at indtræffe. Incidensraten er forholdsvis konstant fra tirsdag til fredag men falder i weekenden til ca. 12% af tilfældene.



Figur 6.11: Incidensraten i procent fordelt på ugedag med 95% konfidensintervaller.

Ugedagsvariation efter alder og køn

Ligesom for sæsonvariationen viser en devianstest, at ugedagsvariationen ikke har ændret sig signifikant gennem undersøgelsesperioden ($p = 0.2$). To Wald-tests viste ingen signifikant forskel på ugedagsvariationen for kønnene ($p = 0.28$) eller aldersgrupperne ($p = 0.87$).

6.5 Opsamling

I dette kapitel er AMI-datasættet blevet analyseret med en statisk Poisson regressionsmodel. Modelkontrollen gav ikke anledning til at forkaste modellen. Analysen viste, at trenden af AMI-incidensraten er lineær på log-skalaen både for den samlede befolkning og stratificeret på alder og køn. Det forventede relative fald af incidensraten pr. år er på 3.2% for den samlede befolkning. Faldet har været størst for mænd og for de 60-69 årige, hvilket også var resultatet i [10].

Incidensraten udviste sæsonvariation med toppunkter i april/maj og november/december og lavpunkt i august. Med mindre afvigelser var sæsonvariationen som udledt i [10], idet de foretagne hypotesetest gav samme resultater. Sæsonvariationen afhænger af alder men ikke af køn, og den har ikke ændret sig signifikant gennem undersøgelsesperioden. En undersøgelse af ugedagsvariationen viste størst incidens om mandagen og mindst incidens i weekenden.

Kapitel 7

Data-analyse med Strukturel Tidsrække Model

I dette kapitel analyseres AMI-dataene ved hjælp af en basic structural model, som blev defineret i afsnit 3.6. Dette sker med funktionen *StructTS* i programmet *R*.

7.1 Opstilling af model

I forbindelse med opstilling af en BSM er det første problem, der skal tages hensyn til, at AMI-dataene ikke opfylder antagelsen om, at observationerne er normalfordelte. Derfor er data aggregeret pr. måned, idet en måned på 30 dage bruges som standard. Datapunkterne er dermed givet ved

$$Y_t = 30 \cdot X_t,$$

hvor X_t angiver incidensen i måned nr. t divideret med antallet af dage i måneden. Dermed opnås en større intensitet, og observationerne bliver approksimativt normalfordelte ifølge den *centrale grænseværdi sætning*, se f.eks. [19]. Det vil sige, at hvis Y_t har intensitet m_t , gælder det approksimativt, at

$$Y_t \sim \mathcal{N}(m_t, m_t).$$

Desuden foretages der en logaritme-transformation af observationerne. Dette medfører dels, at midelværdien svarer til skalaen for GLM-analysen og dels, at de estimerede varianser kan benyttes i

forbindelse med den efterfølgende data-analyse ved en dynamisk Poisson-model. Ved en Taylor udvikling af $\ln Y_t$ omkring $\ln m_t$

$$\ln Y_t \approx \ln m_t + \frac{1}{m_t}(Y_t - m_t),$$

ses det, at det approksimativt gælder, at

$$\ln Y_t \sim \mathcal{N}\left(\ln m_t, \frac{1}{m_t}\right).$$

Det vil sige, at for de logaritme-transformerede observationer er variansen omvendt proportional med intensiteten og dermed ikke konstant, som den skal være ifølge definitionen af en BSM.

En konstant varians kan opnås ved i stedet at foretage en kvadratrods-transformation af Y_t . En Taylor-udvikling af $\sqrt{Y_t}$ omkring $\sqrt{m_t}$ giver

$$\sqrt{Y_t} \approx \sqrt{m_t} + \frac{1}{2\sqrt{m_t}}(Y_t - m_t).$$

Det følger, at det approksimativt gælder, at

$$\sqrt{Y_t} \sim \mathcal{N}\left(\sqrt{m_t}, \frac{1}{4}\right).$$

Det vil senere blive vist, at resultaterne for modellerne med henholdsvis logaritme- og kvadratrods-transformerede observationer er næsten ens. Derfor benyttes den log-lineære model på grund af de førnævnte fordele derved.

Modellen får følgende udseende.

$$\begin{aligned} \ln Y_t &= \mu_t + \gamma_t + \varepsilon_t, & \varepsilon_t &\sim \mathcal{N}(0, \sigma_\varepsilon^2) \\ \mu_{t+1} &= \mu_t + \kappa_t + \xi_t, & \xi_t &\sim \mathcal{N}(0, \sigma_\xi^2) \\ \kappa_{t+1} &= \kappa_t + \zeta_t, & \zeta_t &\sim \mathcal{N}(0, \sigma_\zeta^2) \\ \gamma_{t+1} &= -\gamma_t - \dots - \gamma_{t-10} + \delta_t, & \delta_t &\sim \mathcal{N}(0, \sigma_\delta^2). \end{aligned} \tag{7.1}$$

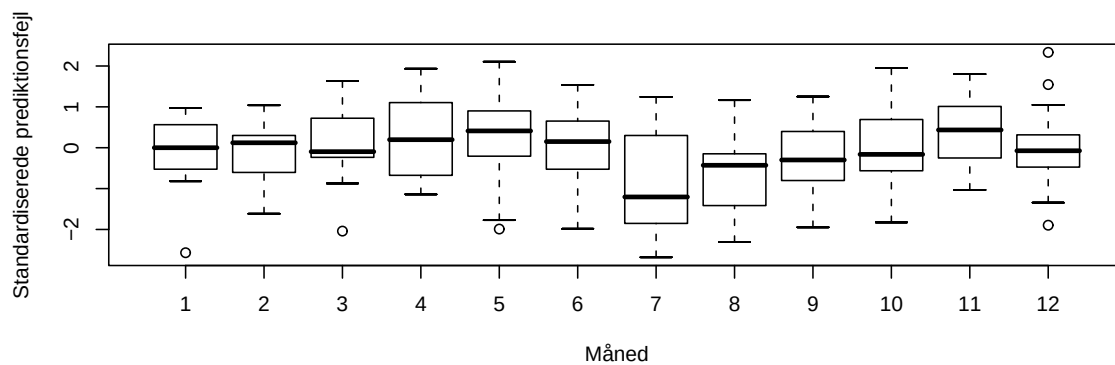
7.2 Valg af variansparametre

Ved at anvende *StructTS* på modellen i (7.1) fremkommer følgende estimer af varianserne.

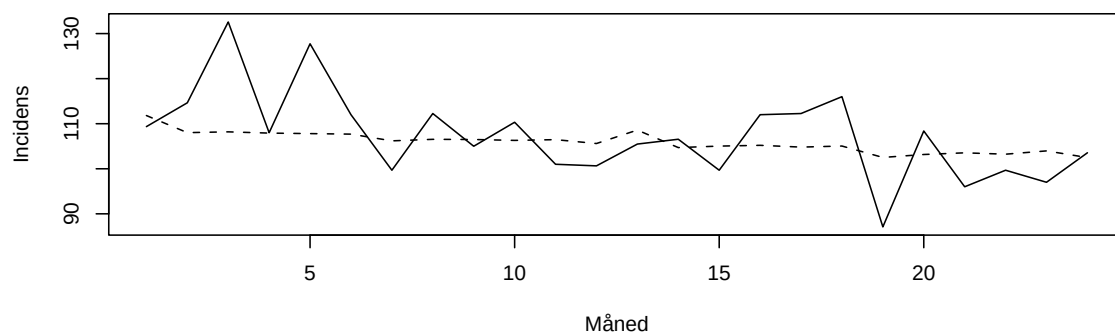
Niveau	σ_ξ^2	$7 \cdot 10^{-7}$
Hældning	σ_ζ^2	0
Sæson	σ_δ^2	$7 \cdot 10^{-5}$
Observationsstøj	σ_ε^2	$1.145 \cdot 10^{-2}$

Det ses, at hældningskomponenten er konstant.

For disse varians-estimerer viser et plot af de standardiserede prediktionsfejl mod måned imidlertid et mønster, der kunne ligne sæsonvariationen for dataene, se figur 7.1.



Figur 7.1: Standardiserede prediktionsfejl plottet mod måned for model med variansparametre opnået fra *StructTS*.



Figur 7.2: Den optrukne linie angiver data, mens den stiplede linie angiver den forventede incidens for de første to år for model med variansparametre opnået fra *StructTS*.

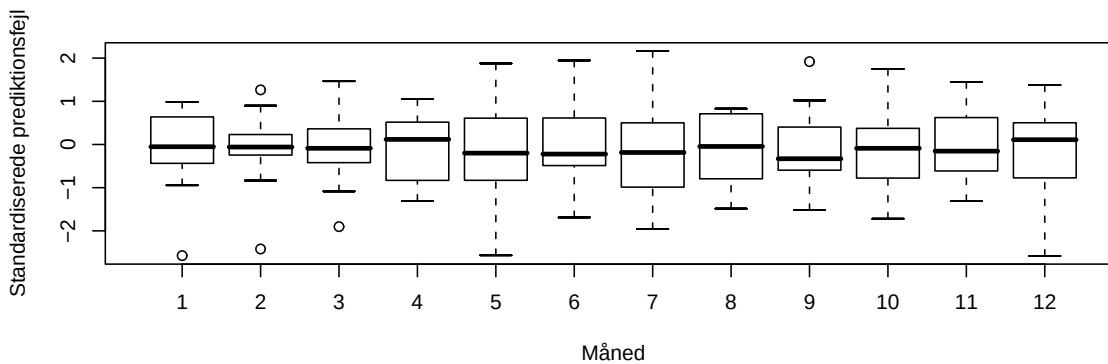
For at undersøge dette nærmere plottes den forventede månedlige incidens sammen med de månedsaggregerede data for de første to år, se figur 7.2.

Det ses på figur 7.2, at modellen slet ikke beskriver sæsonvariationen i dataene tilstrækkeligt. Dette indikerer, at varians-estimatet for sæsonvariations-komponenten er for lille, og dermed er modellen for ufleksibel. Der fremkommer det samme resultat for en model med kvadratrods-transformerede observationer og for en model uden transformation. Det forsøges derfor at fitte en model, hvor variansen for sæsonvariationen er fastsat til en ny værdi, ca. 70 gange større, idet σ_δ^2 sættes lig med 0.005. Denne værdi er opnået ved inspicering af residualplots for forskellige værdier af σ_δ^2 . Dette giver følgende varians-estimer for den nye model.

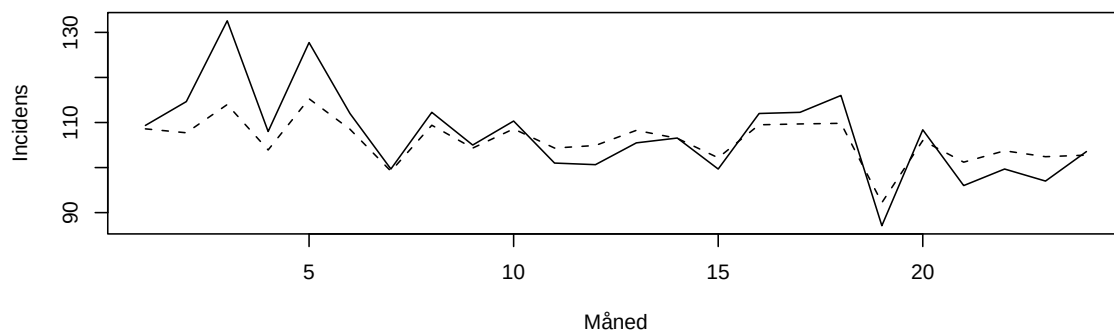
Niveau	σ_ξ^2	$2 \cdot 10^{-6}$
Hældning	σ_ζ^2	0
Sæson	σ_δ^2	$5 \cdot 10^{-3}$
Observationsstøj	σ_ε^2	$7.50 \cdot 10^{-3}$

Det ses, at ændringen medførte en stigning af σ_ξ^2 samt et fald i varians-estimatet for observationsstøjen, ε .

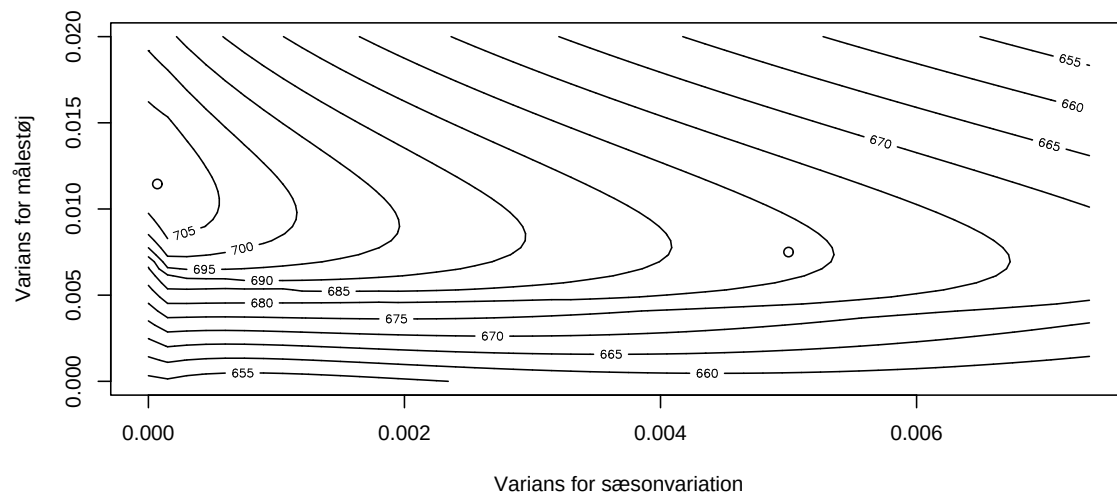
Det ses i figur 7.3, at de standardiserede prediktionsfejl er blevet renset for sæsonvariation i den reviderede model, og figur 7.4 viser, at modellen giver en langt bedre beskrivelse af data.



Figur 7.3: Standardiserede prediktionsfejl plottet mod måned for model med variansparametre korregeret i forhold til *StructTS*.



Figur 7.4: Den optrukne linie angiver data, mens den stiplede linie angiver den forventede incidens for de første to år for model med variansparametre korrigeret i forhold til *StructTS*.



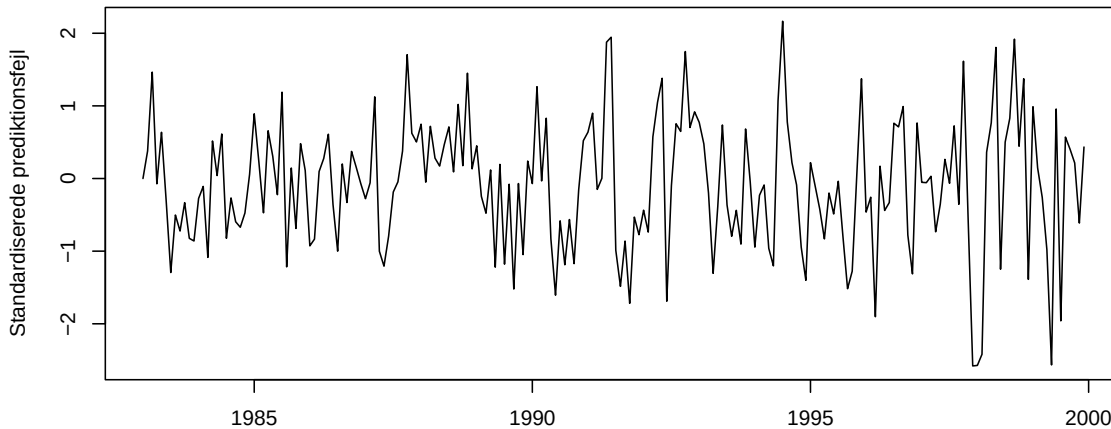
Figur 7.5: Plot af log-likelihood som funktion af σ_δ^2 og σ_ε^2 . Maksimum likelihood estimatet og det anvendte estimat er angivet med punkter.

Den nye model har en log-likelihood på 681.4 mod tidligere 709.7 på trods af, at den tydeligvis beskriver data bedre. Figur 7.5 viser log-likelihooden som en funktion af σ_δ^2 og σ_ε^2 .

7.3 Modelkontrol

I dette afsnit foretages en modelkontrol af den netop udledte model. Kontrollen bygger på de standardiserede prediktionsfejl, der er defineret i afsnit 3.6.2.

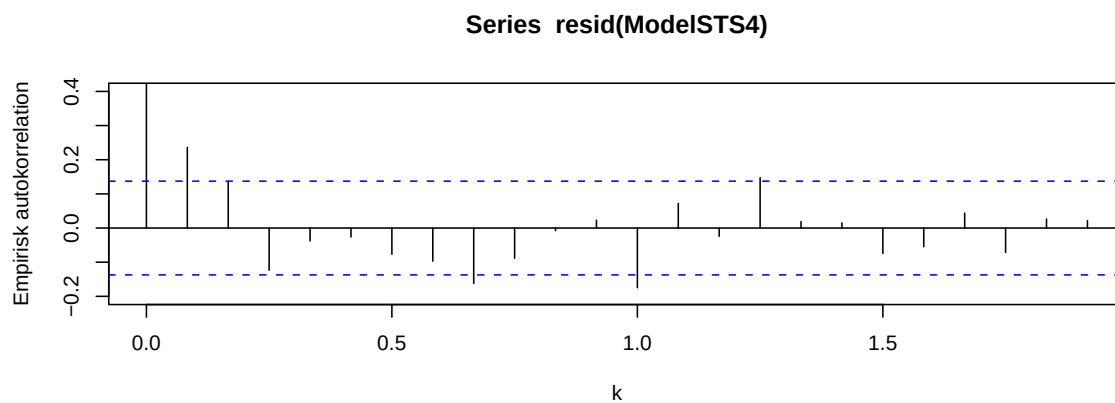
7.3.1 Varianshomogenitet



Figur 7.6: Plot af standardiserede prediktionsfejl over tid.

Figur 7.6 indikerer, at variansen af de standardiserede prediktionsfejl forøges gennem undersøgelsesperioden. Dette skyldes, at middelværdien, m_t , af de utransformererede data er en aftagende funktion af t , hvoraf det følger, at $\text{Var}(\ln Y_t) = 1/m_t$ bliver større i løbet af perioden. For $k = 204$ har teststørrelsen defineret i afsnit 3.6.2 en p -værdi på 0.0009, hvilket som forventet viser, at der er tale om inhomogen varians.

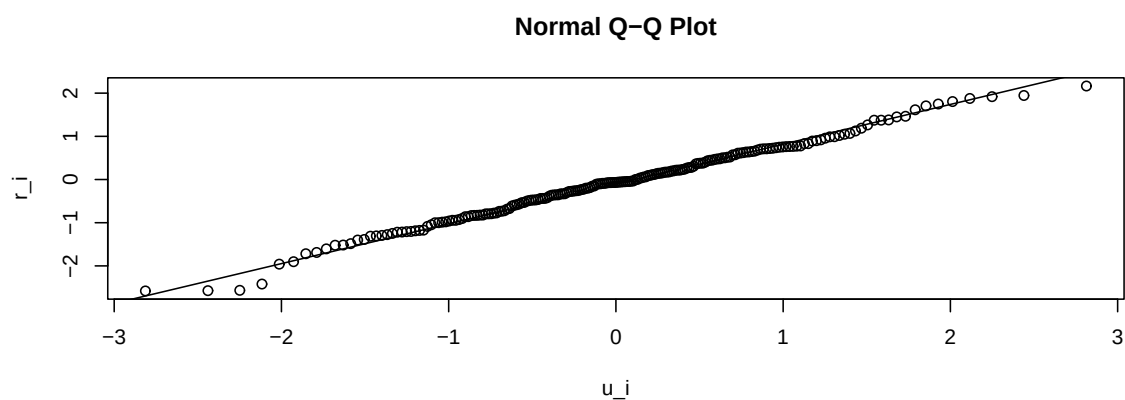
7.3.2 Seriel korrelation



Figur 7.7: Empirisk autokorrelation af standardiserede prediktionsfejl.

Det undersøges, om antagelsen om ukorrelerede prediktionsfejl er opfyldt ved hjælp af et autokorrelationsplot. Figur 7.7 indikerer, at r_t 'erne er positivt korrelerede for $k = 1$.

7.3.3 Normalitet



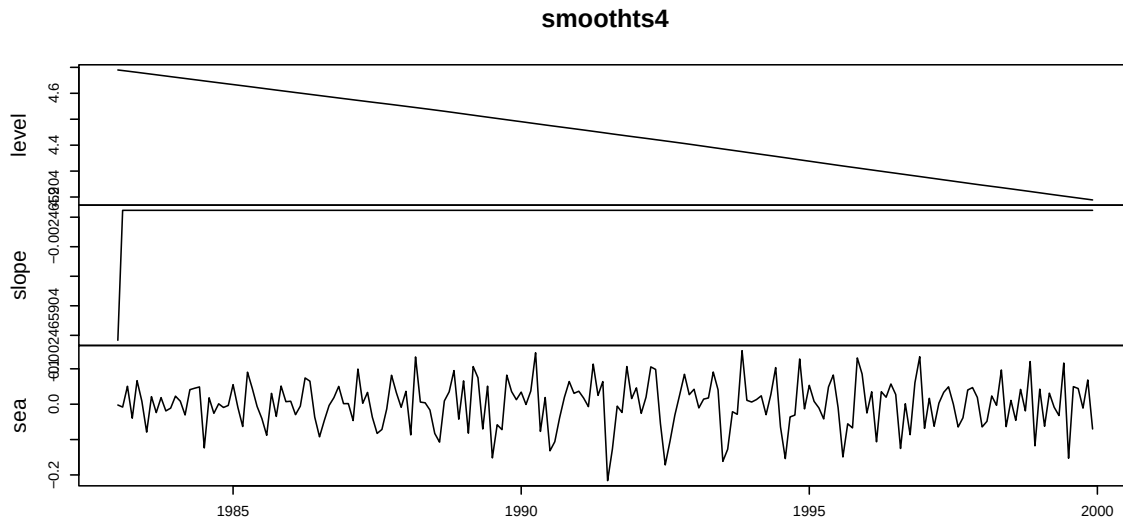
Figur 7.8: Normal-fraktilplot af standardiserede prediktionsfejl.

Antagelsen, om at Y_t er normalfordelt, er ikke oplagt opfyldt. Dels afhænger den af den centrale grænseværdi sætning (at de måneds-aggregerede data er approksimativt normalfordelte), dels af at der ikke er for stor spændvidde mellem dataene, således at log-transformationen bliver for krum. Det kan kontrolleres, om antagelsen om normalitet er overholdt ved at undersøge fordelingen af de standardiserede prediktionsfejl. Kvartilplottet i figur 7.8 viser, at de standardiserede prediktionsfejl approksimativt er normalfordelte.

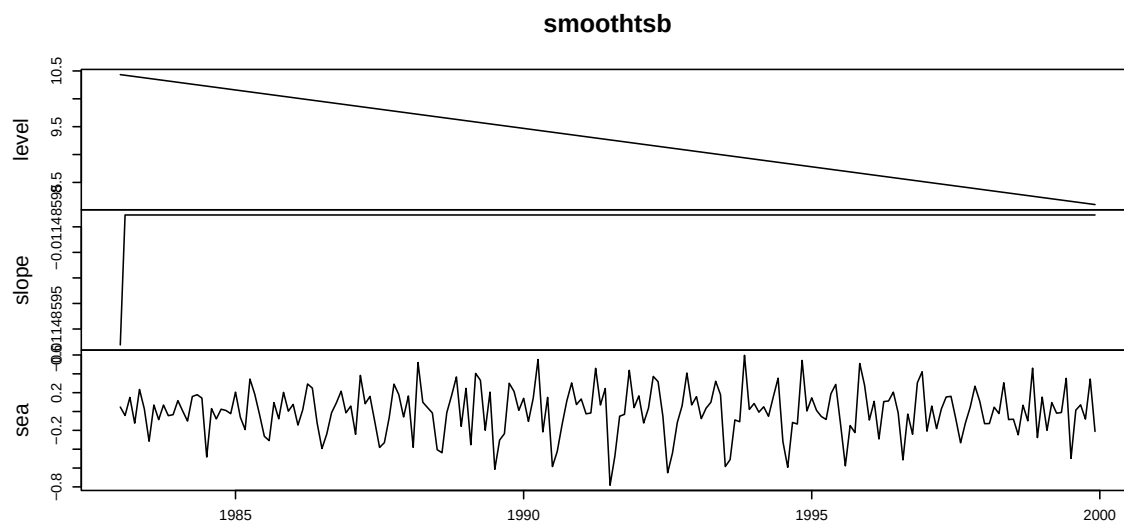
7.4 Resultater

I dette afsnit beskrives resultaterne af data-analysen med en BSM. Outputtet fra funktionen *tsSmooth* (se afsnit 3.6.3) er vist i figur 7.9. Det ses på figurene 7.9 og 7.10, at modellerne med henholdsvis logaritme- og kvadratrods-transformation af observationerne giver stort set det samme resultat. Som tidligere nævnt benyttes log-modellen, da det giver mulighed for at benytte de estimerede varianser som initiale gæt i analysen med en dynamisk Poisson model.

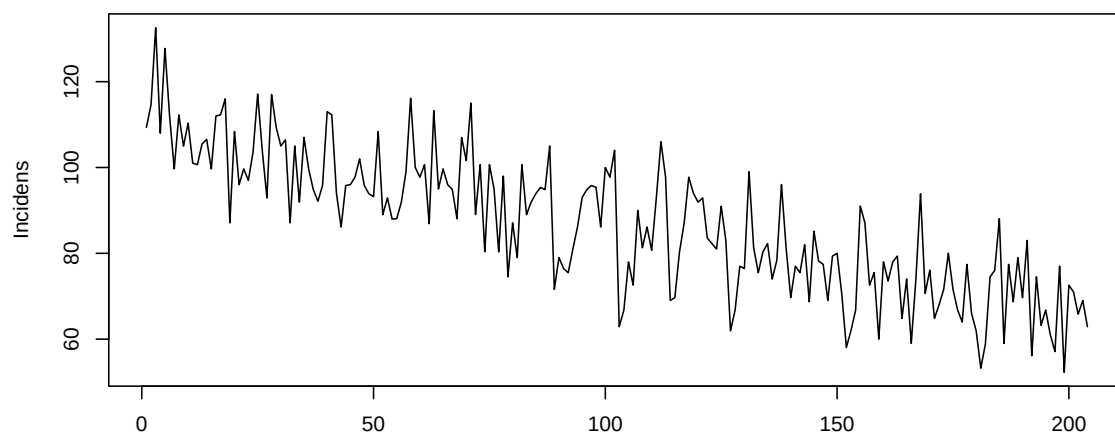
Det ses på figur 7.9, at trenden er lineær, og at udsvingene i sæsonvariation ikke er konstante gennem undersøgelsesperioden, idet de er mindst i starten og størst i midten af perioden. Dette stemmer godt overens med figur 7.11, som viser de måneds-aggregerede data.



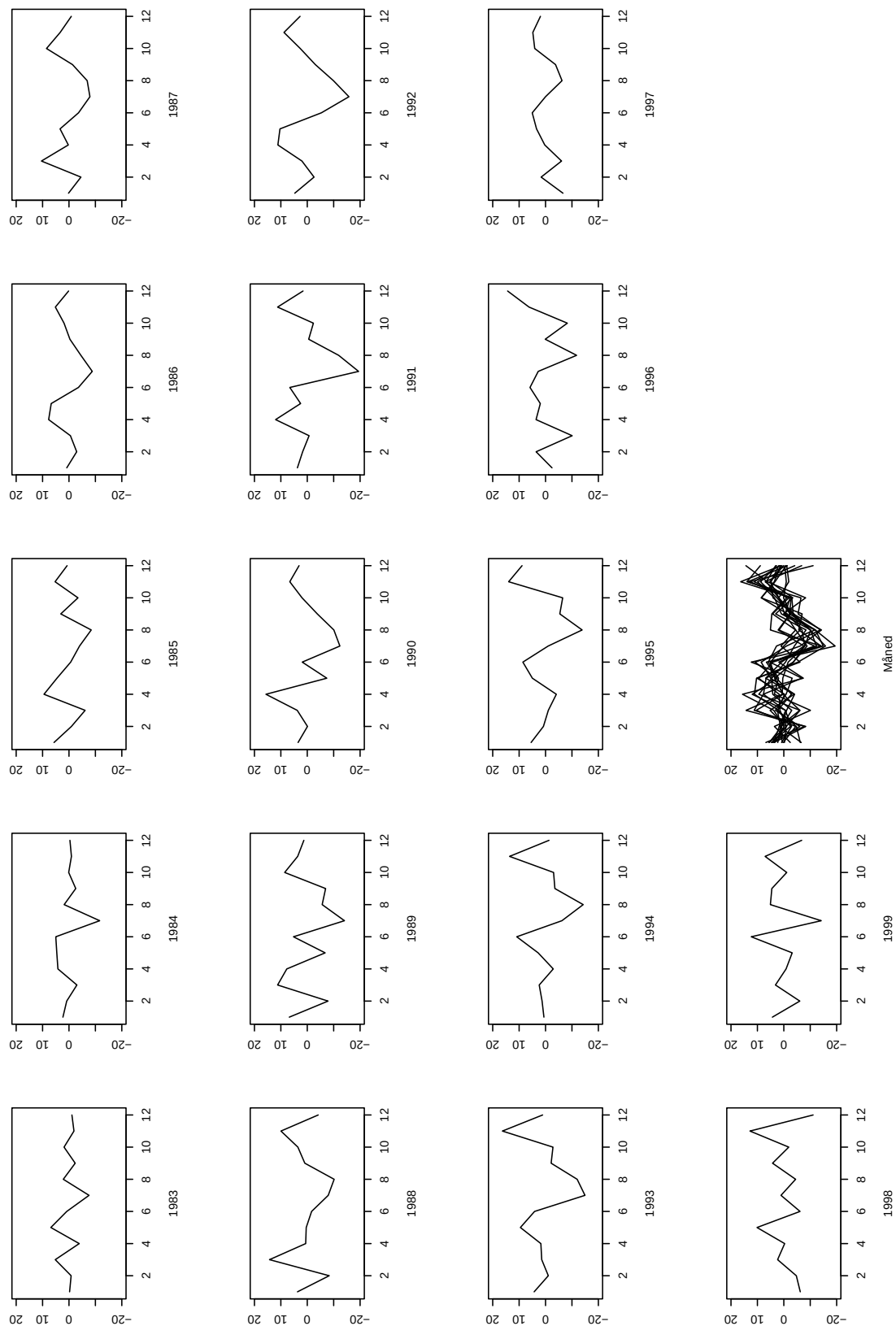
Figur 7.9: Kalman-udglattede værdier af logaritme-transformerede observationer.



Figur 7.10: Kalman-udglattede værdier af kvadratrods-transformerede observationer.



Figur 7.11: Observeret AMI-incidens pr. måned.



Figur 7.12: Sæsonvariation af AMI-incidens for hvert år og for alle årene plottet på en gang.

Figur 7.12 viser sæsonvariationen for hvert år, og det ses, at variationen ændrer sig betydeligt fra år til år. Det ses også, at størrelsen af udsvingene varierer, mens det er svært at udlede en generel tendens for sæsonvariationen på grund af de relativt store forskelle fra år til år. Som i den statiske analyse er der toppunkter forår og efterår/vinter samt lavpunkt om sommeren.

7.5 Opsamling

I data-analysen med en BSM viste maksimum-likelihood estimaterne for varianserne sig at være uhen-sigtsmæssige med hensyn til at beskrive sæsonvariationen i dataene. Dette indikerer, at der kan være problemer med at anvende *StructTS*. Fittet blev forbedret ved at forøge variansen for sæsonvariations-komponenten, selvom likelihooden for modellen herved blev mindre.

Analysen viste, at trenden er lineær på log-skalaen, så deri var der ingen forskel i forhold til den statiske analyse, men i modsætning til denne kan det ses, at sæsonvariationen ændrer sig i løbet af undersøgelsesperioden. Dog påviste modelkontrollen inhomogen varians og nogen korrelation for prediktionsfejlene, hvorfor resultaterne skal tages med forbehold.

Der er ikke foretaget stratificeret analyse efter alder og køn. Dette vil blive gjort i forbindelse med data-analysen ved en Poisson state space model, hvor observationerne udgøres af daglig AMI-incidens som i den statiske analyse.

Kapitel 8

Data-analyse med Dynamisk Poisson Model

I dette kapitel vil AMI-dataene blive analyseret med en Poisson state space model. Dette vil ske ved hjælp af R-pakken *sspir*, som er beskrevet i afsnit 4.4.

8.1 Opstilling af model

Modellen har følgende udseende.

$$\begin{aligned} \text{Observationsmodel:} \quad f(y_t|\mu_t) &= \frac{\mu_t^{y_t} \exp(-\mu_t)}{y_t!} \\ E(Y_t|\mu_t) &= \text{Var}(Y_t|\mu_t) = \mu_t \\ \ln \mu_t &= \lambda_t = \mathbf{F}_t^T \boldsymbol{\theta}_t \\ \text{Tilstandsligning:} \quad \boldsymbol{\theta}_t &= \underline{\underline{G}}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim \mathcal{N}_p(\mathbf{0}, \underline{\underline{W}}) \\ \text{Startinformation:} \quad (\boldsymbol{\theta}_0|D_0) &\sim \mathcal{N}_p(\mathbf{m}_0, \underline{\underline{C}}_0). \end{aligned}$$

Det bemærkes, at systemvarianserne er konstante. Tilstandsvektoren, $\boldsymbol{\theta}_t$, er givet ved

$$\boldsymbol{\theta}_t = (\beta_t, \kappa_t, \gamma_t^{c,1}, \gamma_t^{s,1}, \gamma_t^{c,2}, \gamma_t^{s,2}, \alpha_t, \alpha_{t-1}, \dots, \alpha_{t-5})^T,$$

hvor β_t angiver trenden, κ_t er hældningen, $\gamma_t^{c,1}, \dots, \gamma_t^{s,2}$ er sæsonvariationsparametre, mens de resterende parametre beskriver ugedagsvariationen. Sæsonvariationen beskrives ved en Fourier-række af grad to, se afsnit 3.6.

$$\gamma_t = \gamma_t^{c,1} \cos(\tau_1 t) + \gamma_t^{s,1} \sin(\tau_1 t) + \gamma_t^{c,2} \cos(\tau_2 t) + \gamma_t^{s,2} \sin(\tau_2 t),$$

hvor $\tau_j = 2\pi j/365.25$, og

$$\begin{aligned}\gamma_{t+1}^{c,i} &= \gamma_t^{c,i} + \delta_t^{c,i}, & \delta_t^{c,i} &\sim \mathcal{N}(0, \sigma_\delta^2) \\ \gamma_{t+1}^{s,i} &= \gamma_t^{s,i} + \delta_t^{s,i}, & \delta_t^{s,i} &\sim \mathcal{N}(0, \sigma_\delta^2),\end{aligned}$$

for $i = 1, 2$.

Ugedagskomponenten er givet ved

$$\alpha_{t+1} = -\alpha_t - \dots - \alpha_{t-5} + \rho_t, \quad \rho_t \sim \mathcal{N}(0, \sigma_\rho^2).$$

Middelværdi og variansmatrix for den initiale tilstandsvektor, θ_0 , defineres i *sspir* ved

$$\mathbf{m}_0 = (0, 0, \dots, 0)^T$$

og

$$\underline{\underline{C}}_0 = 10^6 \cdot \underline{\underline{I}}_p.$$

Transitionsmatricen er blokdiagonal med en blok for henholdsvis trenden, sæsonkomponenten og ugedagskomponenten.

$$\underline{\underline{G}}_t = \begin{bmatrix} \underline{\underline{T}} & & \\ & \underline{\underline{S}} & \\ & & \underline{\underline{U}} \end{bmatrix},$$

hvor

$$\underline{\underline{T}} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad \underline{\underline{S}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

og

$$\underline{\underline{U}} = \begin{bmatrix} -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Systemfejlsvektoren er givet ved

$$\omega_t = (\xi_t, \zeta_t, \delta_t^{c,1}, \delta_t^{s,1}, \delta_t^{c,2}, \delta_t^{s,2}, \rho_t, 0, \dots, 0)^T,$$

mens designvektoren er givet ved

$$\mathbf{F}_t = (1, 0, \cos(\tau_1 t), \sin(\tau_1 t), \cos(\tau_2 t), \sin(\tau_2 t), 1, 0, \dots, 0)^T.$$

8.2 Valg af variansparametre

Der mangler at blive estimeret variansparametre for modellen. For AMI-datasættet tager det i *sspir* ca. en halv time at beregne log-likelihooden for givne værdier af varianserne. Det vil sige, at applikationen af en numerisk maksimerings-algoritme beregningsmæssigt vil kræve et uoverstigeligt tidsforbrug. Som beskrevet i afsnit 4.5 findes der andre metoder til estimation af varianserne i en DGLM som MCMC og Importance sampling. En undersøgelse af disse metoders anvendelighed i denne sammenhæng ville være en oplagt udvidelse, hvis arbejdet skulle fortsættes efter specialeskrivningen. I stedet fastlægges varianserne ved mere *ad hoc*-prægede metoder, som beskrives i det følgende.

Ved at dividere de månedlige varianser, fra data-analysen med en BSM i forrige kapitel, med 30 opnås foreløbige estimer af varianserne for trend, hældning og sæsonvariation, mens variansen for ugedagsvariationen indledningsvis er sat til 10^{-4} . Det følger, at systemvarianserne er givet ved diagonal-elementerne i en blokmatrix med følgende tre blokke for henholdsvis trend, sæson- og ugedagskomponenten.

$$8 \cdot 10^{-8} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \frac{0.005}{30} \cdot \underline{I}_4,$$

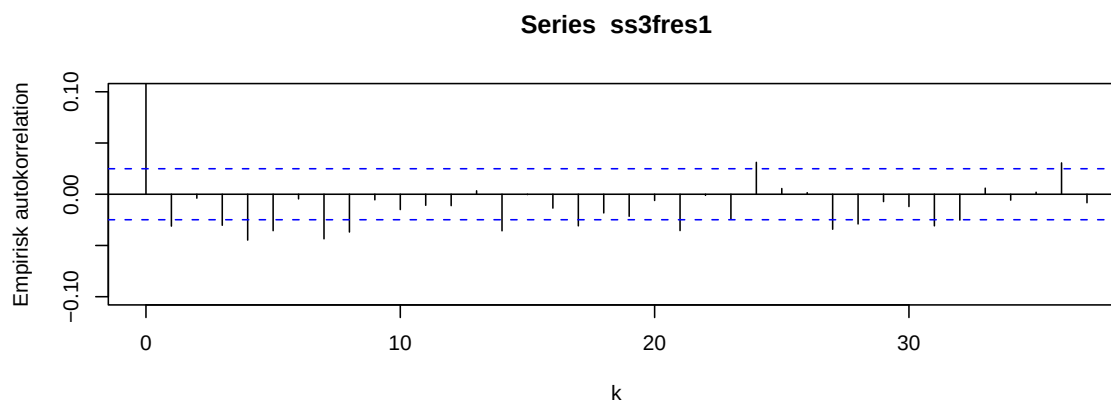
og

$$10^{-4} \cdot \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

For at undersøge om denne model er anvendelig foretages en modelkontrol. Da det er stort set umuligt at udregne prediktionsfejlene i *sspir*, anvendes i stedet de standardiserede residualer

$$\frac{Y_t - \exp(\mathbf{F}_t^T \tilde{\mathbf{m}}_t)}{\sqrt{\exp(\mathbf{F}_t^T \tilde{\mathbf{m}}_t)}},$$

hvor $\tilde{\mathbf{m}}_t$ er det Kalman-udglattede estimat af tilstandsvektoren. Et autokorrelationsplot viste en signifikant seriel korrelation mellem de standardiserede residualer, hvilket er illustreret i figur 8.1.

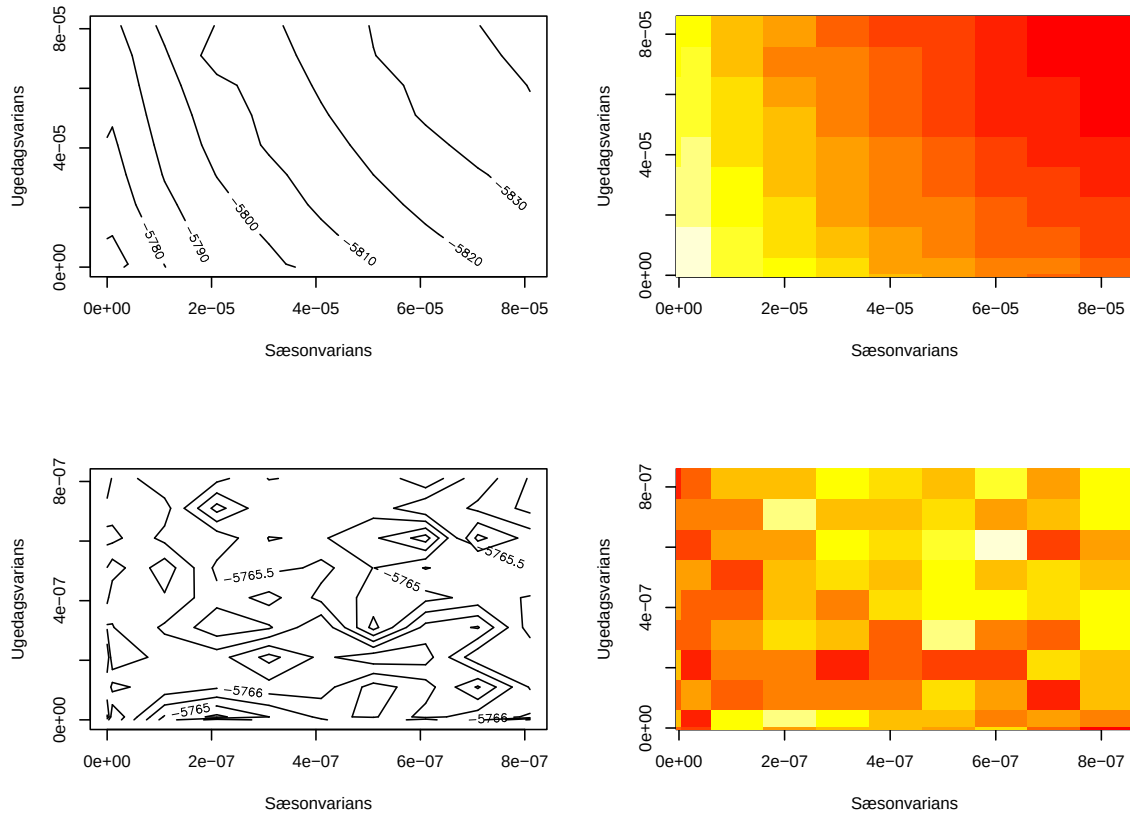


Figur 8.1: Empirisk autokorrelationsplot af standardiserede residualer fra Poisson state space model initialiseret med variansparametre opnået fra *StructTS*.

Ifølge [15] er residualerne også teoretisk set svagt negativt korrelerede. Derfor skal figur 8.1 tages med forbehold. Alligevel forsøges det at udlede en model, hvor den serielle korrelation er mindre.

For den Gaussiske local level model (se afsnit 3.6) vises det i [25], at negativ seriel korrelation mellem prediktionsfejlene forekommer, når systemvariansen er specificeret for højt i forhold til observationsvariansen. Derfor forsøges det, at reducere varianserne i $\underline{W}$. Den optimale model vil blive fundet ved at variere på variansparametrene og sammenligne modellernes likelihood.

Der opstod beregningsmæssige problemer med at foretage itereret extended Kalman-filtrering og -udglatning i *sspir*, når variansen for trendkomponenten var angivet til mindre end 10^{-6} (dette viste sig at skyldes, at der var for stort spring mellem varianserne i $\underline{C}_0$, som er sat til 10^6 , og $\underline{W}$). Det kan derfor anbefales i stedet at sætte $\underline{C}_0$ lig med identitetsmatricen). Det blev derfor besluttet at fastsætte $\text{Var}(\xi_t)$ til at være lig med 10^{-6} og maksimere log-likelihooden som funktion af sæson- og ugedagsvarianserne, idet variansen for hændingskomponenten fortsat blev sat til at være lig med nul.



Figur 8.2: Log-likelihood som funktion af sæson- og ugedagsvariansen. For plottene til højre gælder det, at jo lysere et område er, jo større er log-likelihooden.

Figur 8.2 viser log-likelihooden som funktion af sæson- og ugedagsvariansen. Det ses, at likelihood-funktionen approksimativt har maksimum, når begge varianser antager værdien $6 \cdot 10^{-7}$. Det vises i næste afsnit, at for disse variansparametre, er der noget mindre negativ seriel korrelation mellem residualerne. Derfor vælges en model, hvor

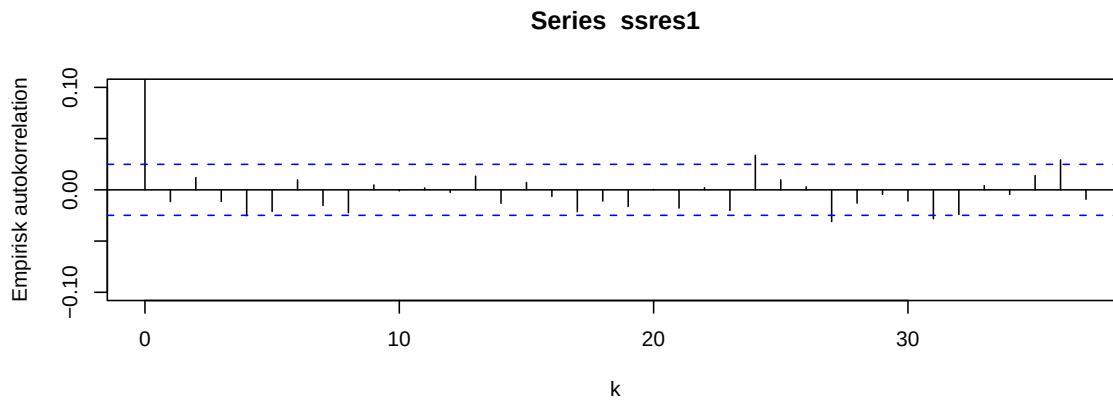
$$\text{Var}(\xi_t) = 10^{-6}, \quad \text{Var}(\zeta_t) = 0, \quad \text{Var}(\delta_t) = 6 \cdot 10^{-7}, \quad \text{Var}(\rho_t) = 6 \cdot 10^{-7}. \quad (8.1)$$

8.3 Modelkontrol

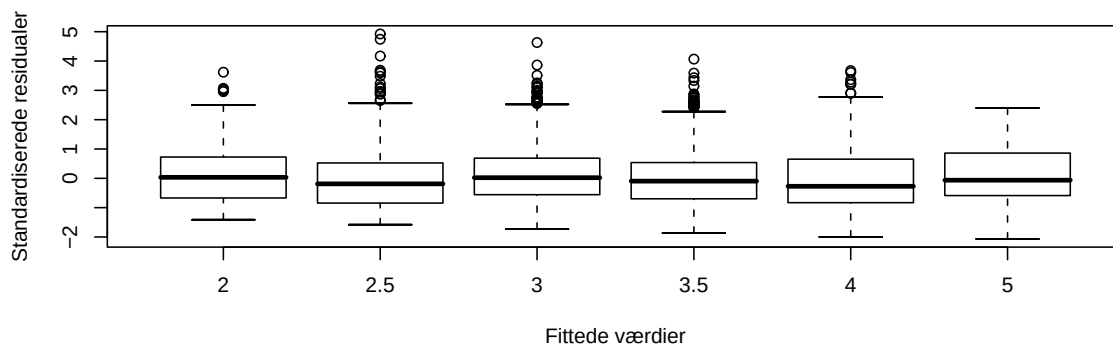
I dette afsnit foretages en modelkontrol af modellen med variansparametre givet ved 8.1.

8.3.1 Seriel korrelation

Det ses i figur 8.3, at de standardiserede residualer ikke er signifikant serielt korrelerede, og at den negative korrelation er noget mindre markant end i figur 8.1.



Figur 8.3: Empirisk autokorrelationsplot af standardiserede residualer fra Poisson state space model, initialiseret med variansparametre korrigeret i forhold til *StructTS*.

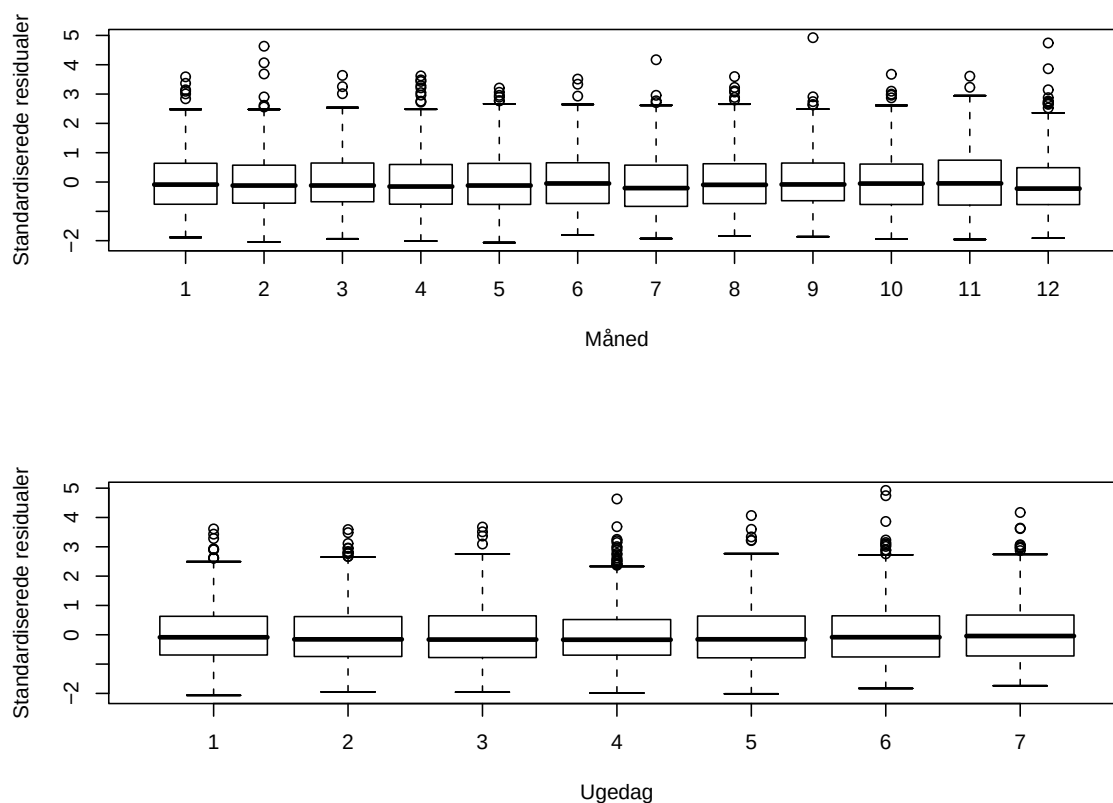


Figur 8.4: Standardiserede residualer plottet mod fittede værdier. Boksen, der svarer til værdien 3 på x-aksen, beskriver fordelingen af residualerne svarende til fittede værdier i intervallet 2.5-3.0.

8.3.2 Varianshomogenitet

I figur 8.4 ses et plot af fittede værdier mod standardiserede residualer. Som i afsnit 6.3.1 anvendes et boxplot, da et almindeligt scatterplot er vanskeligt at fortolke på grund af de lave intensiteter. Plottet indikerer ikke, at der er en inhomogen varians for residualerne i forhold til middelværdien.

Det undersøges, om modellen afspejler de systematiske variationer i dataene ved at plote de standardiserede residualer mod henholdsvis måned og ugedag, se figur 8.5.



Figur 8.5: Standardiserede residualer plottet mod måned (øverst) og ugedag (nederst).

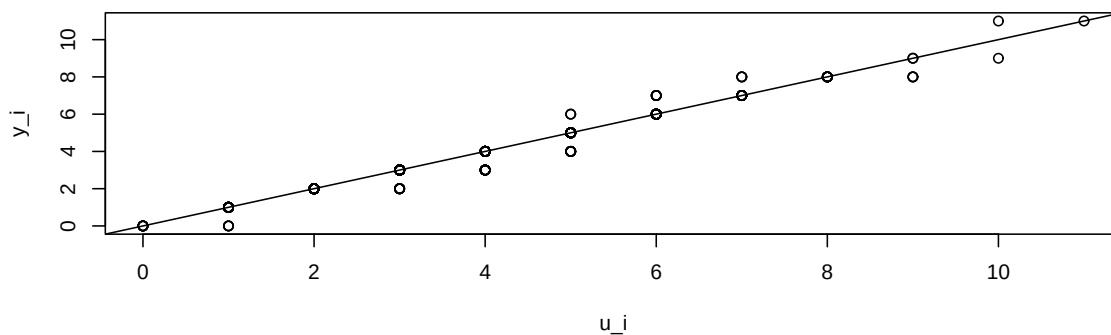
Det ses, at fordelingen af de standardiserede residualer er nogenlunde konstant i forhold til ugedag og måned, hvilket indikerer, at modellen giver en god beskrivelse af variationerne i data.

8.3.3 Fordelingstype

Det undersøges, om de empiriske middelværdier og varianser for observationer med approksimativt samme fordeling svarer til det forventede.

ξ	Empirisk middelværdi	Empirisk varians
2.0	1.96	2.08
2.5	2.51	2.62
3.0	2.97	3.03
3.5	3.50	3.46
4.0	3.94	4.19

Det ses, at de empiriske varianser er lidt for store, men at de passer bedre end i den statiske analyse, se afsnit 6.3.3. Figur 8.6 viser et fraktil-plot for observationer med intensitet på approksimativt 4. Det ses, at fordelingen af observationerne tilsyneladende er korrekt specificeret.

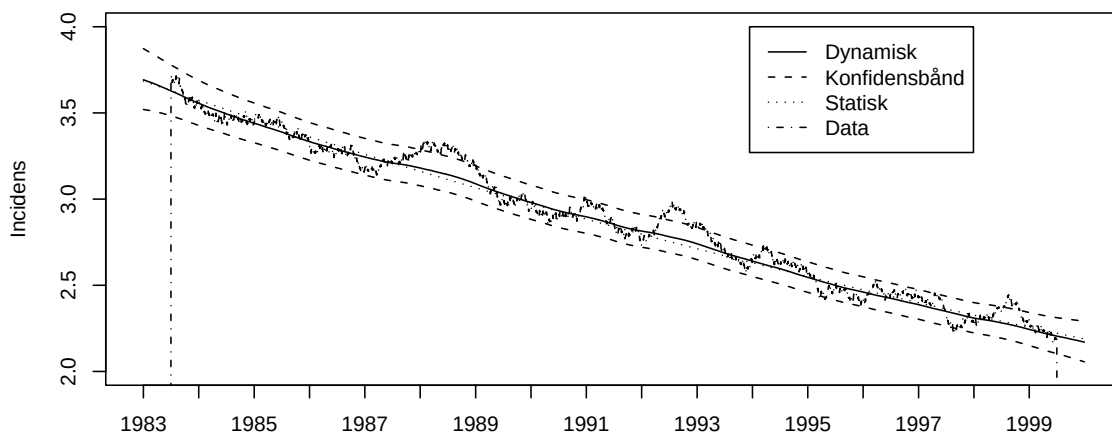


Figur 8.6: Fraktilplot af observationer med intensitet approksimativt 4.

8.4 Resultater

8.4.1 Trend

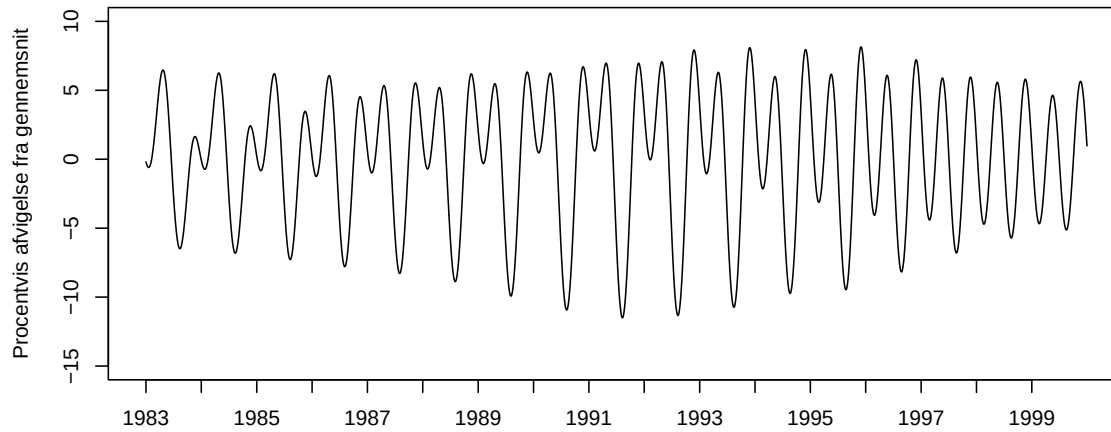
I analysen med en statisk model var konklusionen, at en retlinet trend var tilstrækkelig, hvilket blev bekræftet af analysen med en BSM. For en DGLM fås en kurve for trenden, der snor sig omkring den statiske trendlinie, hvilket kan ses i figur 8.7, der også viser 95% konfidensintervallet for den dynamiske trend samt glidende gennemsnit af data.



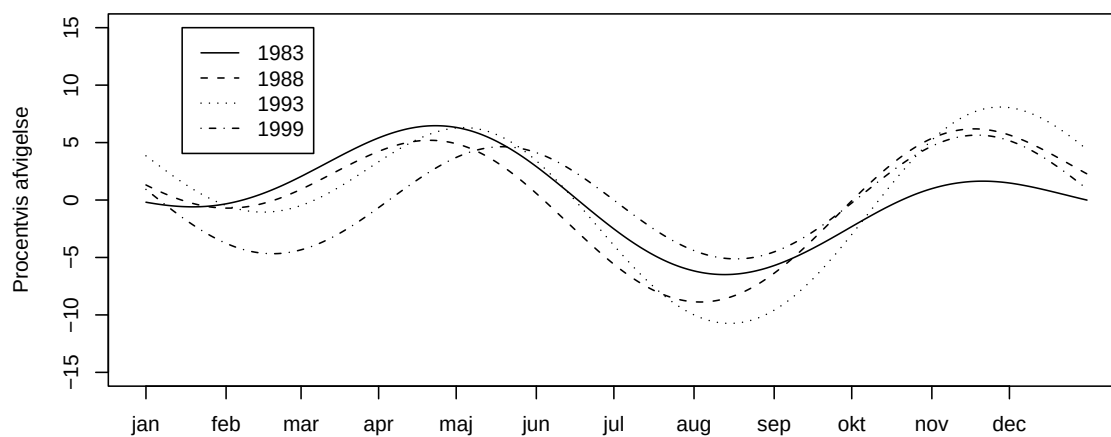
Figur 8.7: Trenden af AMI-incidens fra Poisson state space model med 95% konfidensinterval plottet sammen med den statiske trend og glidende gennemsnit af data af orden 365.

Det ses, at den dynamiske trend som forventet følger udviklingen i dataene i lidt højere grad end den statiske. Der er dog perioder, hvor dataene ligger uden for 95% konfidensintervallet for trenden. Den statiske trend ligger klart indenfor konfidensintervallet af den dynamiske, hvilket indikerer, at ændringerne i trenden er forholdsvis små. De to modeller har stort set samme likelihood.

8.4.2 Sæsonvariation



Figur 8.8: Sæsonvariationen af AMI-incidens.



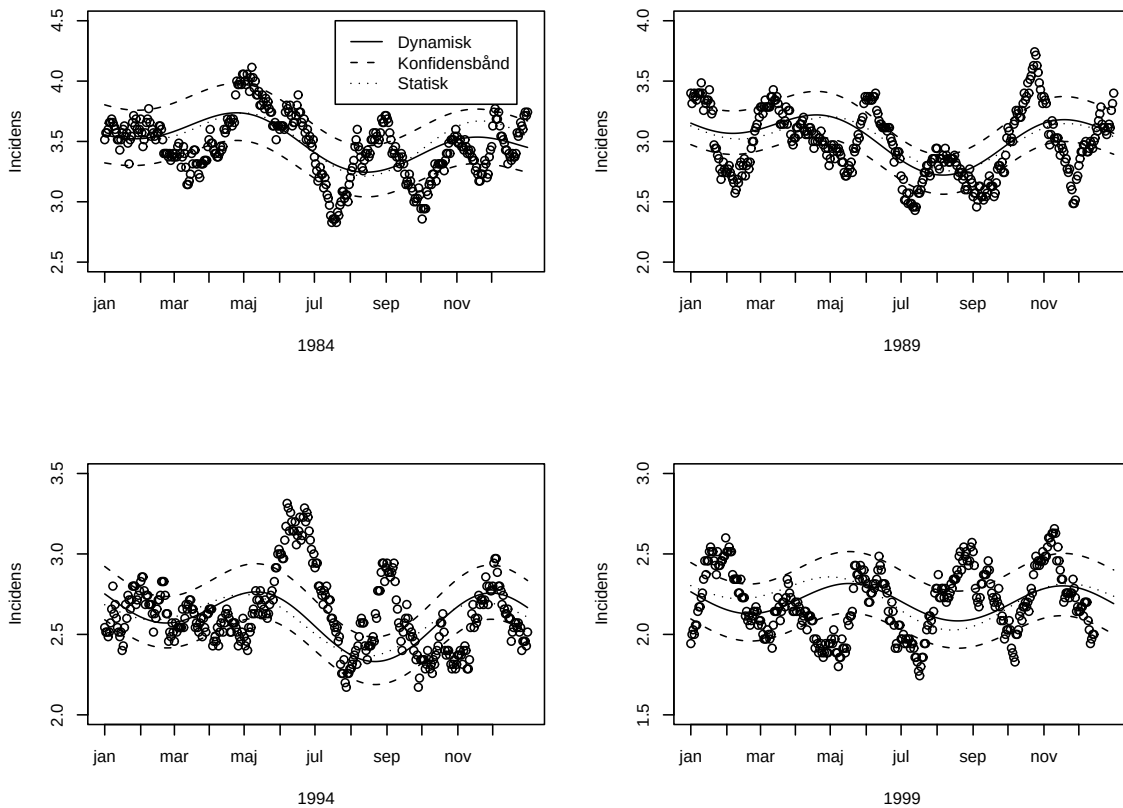
Figur 8.9: Sæsonvariation for årene 1983, 1988, 1993 og 1999.

Det ses i figur 8.8, at udsvingene i sæsonvariationen voksede frem til midten af perioden, hvorefter de aftog igen, hvilket også var resultatet af analysen med en BSM i forrige kapitel.

Figuren viser, at forårstopunktet faldt fra ca. 6.5% til ca. 5% over gennemsnittet i perioden 1983-1988. Derefter steg det til ca. 7% over gennemsnittet i 1992, hvorefter det igen faldt til ca. 4.5% over gennemsnittet i 1999. Sommerlavpunktet faldt fra ca. 6.5% til ca. 12% under gennemsnittet i 1991, hvorefter det steg igen til ca. 5.5% under gennemsnittet i 1999. Efterårs/vinter-toppunktet steg frem til 1995 fra ca. 1.5% til ca. 8% over gennemsnittet, hvorefter det faldt til ca. 5.5% over gennemsnittet i 1999. Desuden er der fra midten af halvfemserne blevet dannet et lavpunkt i februar, der i 1999 er vokset til ca. 5% under gennemsnittet.

Sæsonvariationen for årene 1983, 1988, 1993 og 1999 er vist i figur 8.9, hvor det fremgår, at sæsonvariationen har forrykket sig i løbet af undersøgelsesperioden. Mest markant er det, at forårstoppunktet har flyttet sig til ca. en måned senere på året i løbet af perioden.

I figur 8.10 ses den forventede incidens baseret på henholdsvis den dynamiske og den statiske model samt den observerede incidens. Ændringerne i sæsonvariation er forholdsvis små, idet de forventede incidenser for den statiske model ligger indenfor konfidensintervallerne for de forventede incidenser fra den dynamiske model. Som det kan ses i figur 8.2, falder likelihooden kun meget lidt, når sæson- og ugedags-komponenterne er statiske. Det ses, at en stor del af dataene ligger udenfor konfidensintervallerne.



Figur 8.10: Forventet AMI-incidens uden ugedagsvariation for dynamisk model med 95% konfidens-interval plottet sammen med den forventede incidens for en statisk model og glidende gennemsnit af data for årene 1984, 1989, 1994 og 1999.

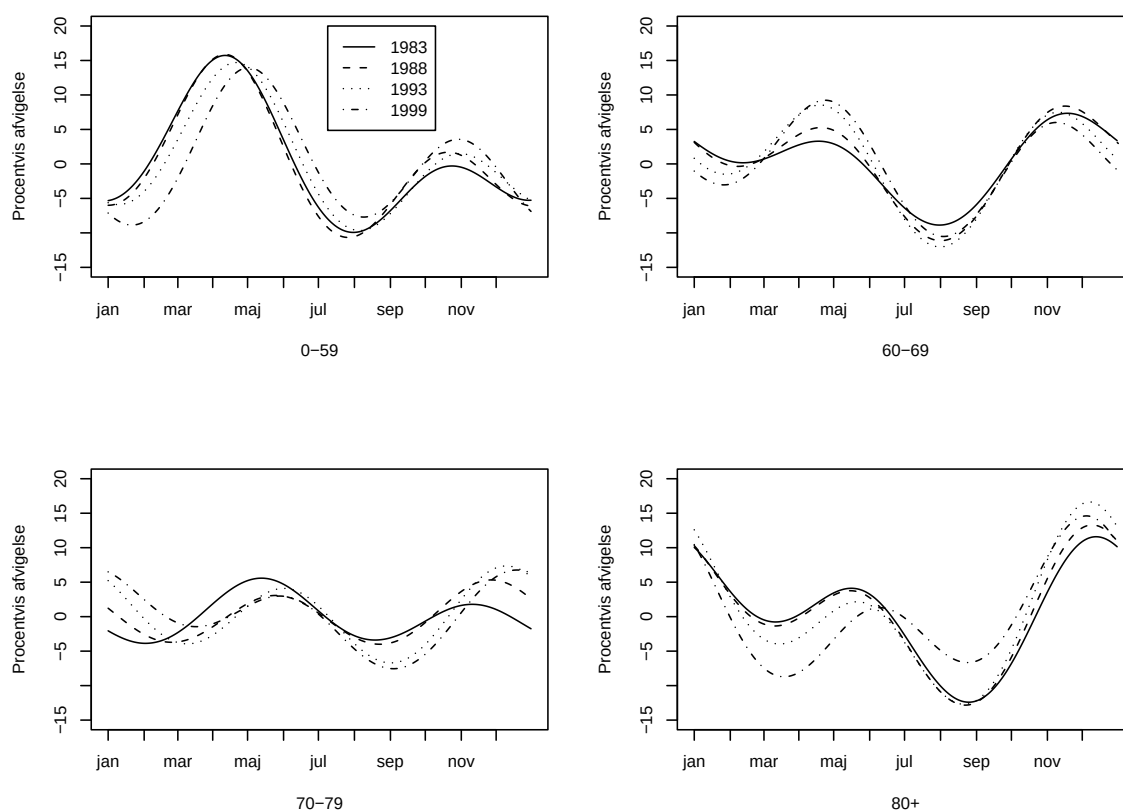
Sæsonvariation for aldersgrupper

I analysen stratificeret efter alder og køn benyttes de samme variansparametre som i modellen for den samlede befolkning.

Den statiske analyse viste, at forårstoppunktet først og fremmest viste sig ved analyse af de yngste AMI-patienter, mens efterårs/vinter-toppunktet hovedsageligt skyldtes de ældste patienter. Dette billede ses også i den dynamiske analyse ligesom de øvrige forskelle i sæsonvariation, som blev dokumenteret i afsnit 6.4.2, hvilket er illustreret i figur 8.11, som viser udviklingen i sæsonvariation for

de fire aldersgrupper, 0-59 år, 60-69 år, 70-79 år og 80 år eller derover.

For de 0-59 årige opstår der i løbet af perioden et lavpunkt i januar ca. 9.5% under gennemsnittet. Desuden ses det, at sæsonvariationen for de 60-69 årige er modsat den generelle tendens, idet forårstoppunktet er vokset fra ca. 3% til ca. 9% over gennemsnittet, mens incidensen er blevet relativt mindre om sommeren, og efterårs/vintertoppunktet er aftaget lidt i løbet af undersøgelsesperioden.



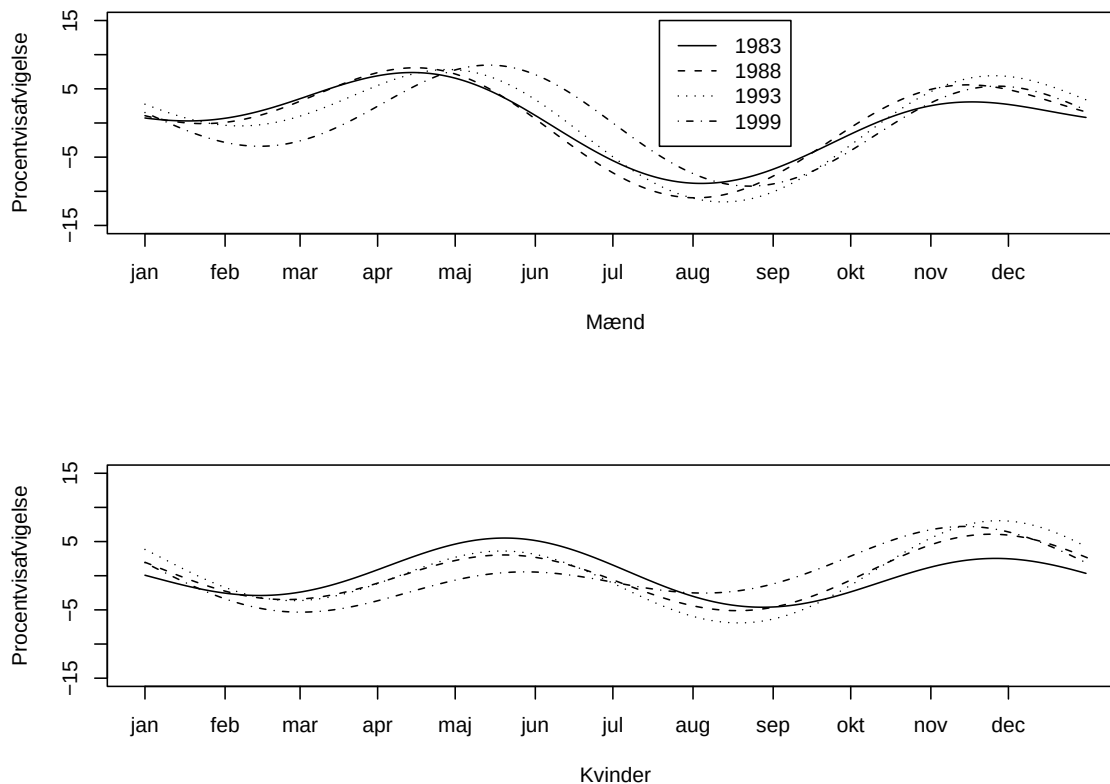
Figur 8.11: Udviklingen i sæsonvariation over tid for de fire aldersgrupper.

Som det blev udledt i den statiske analyse, har de 70-79 årige ikke den store sæsonvariation, men det ses på figur 8.11, at den er blevet større i løbet af perioden. Sommerlavpunktet har flyttet sig et par uger længere hen på året, og stik mod den generelle tendens er det faldet fra ca. 3.5% til ca. 8% under gennemsnittet. Derudover er det bemærkelsesværdigt, at efterårs/vinter-toppunktet for de 70-79 årige har flyttet sig fra starten af november til slutningen af december i løbet af perioden. For gruppen af

patienter over 79 år er det mest interessante, at der er opstået et markant lavpunkt i marts ca. 9% under gennemsnittet.

For alle fire aldersgrupper gælder det dog, at sæsonvariationen for en statisk model ligger inde for 95% konfidensintervallet af sæsonvariationen for den dynamiske model, hvilket indikerer, at ændringerne i sæsonvariation er relativt små.

Sæsonvariation for køn



Figur 8.12: Udviklingen i sæsonvariation for mænd og kvinder.

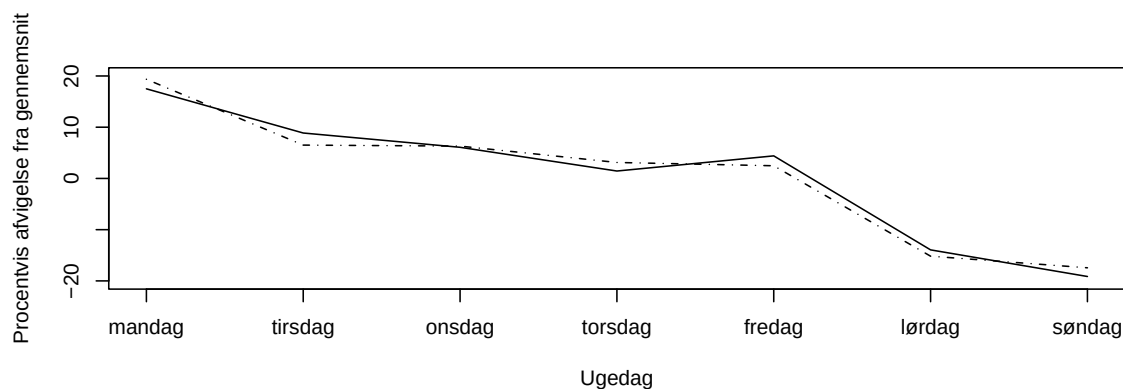
Figur 8.12 viser, hvordan sæsonvariationen har udviklet sig for de to køn i perioden 1983-1999. Det ses, at der er større sæsonvariation for mænd end for kvinder. Ellers er der ikke de helt store forskelle, hvilket også var resultatet af den statiske analyse.

Det er bemærkelsesværdigt, at det hovedsageligt er hos mændene, at forårstoppunktet har flyttet sig ca. en måned senere hen på året, så det nu indtræffer på ca. samme tid for begge køn. Hos kvinderne har sommerlavpunktet flyttet sig, så det indtræffer et par uger tidligere på året, mens det for mændene har flyttet sig til lidt senere på året, hvilket betyder, at det stort set har været konstant for den samlede befolkning, som det ses i figur 8.9.

Heller ikke for nogen af kønnene har der været en markant ændring af sæsonvariationen gennem undersøgelsesperioden, idet sæsonvariationen for en statisk model befinder sig indenfor 95% konfidensintervallet af sæsonvariationen for den dynamiske model.

8.4.3 Ugedagsvariation

Figur 8.13 viser udviklingen af ugedagsvariationen gennem undersøgelsesperioden. Som i den statiske analyse er incidensen størst om mandagen og mindst i weekenden. Det bemærkes, at incidensen om mandagen er steget til næsten 20% over gennemsnittet, men der er ikke sket markante ændringer.



Figur 8.13: Den fuldt optrukne linie angiver ugedagsvariationen for den første uge, mens den stiplede linie angiver ugedagsvariationen for den sidste uge i undersøgelsesperioden.

8.5 Opsamling

Analysen med en dynamisk Poisson model har vist de samme overordnede karakteristika for trend, sæson- og ugedagsvariation af AMI-incidens, som blev udledt i analysen med en GLM. Der er påvist små ændringer i sæsonvariation over tid både med hensyn til størrelsen af udsvingene, og hvornår på året ekstremumpunkterne af sæsonvariationen indtræffer.

Variansparametrene i de anvendte modeller blev estimeret ved *ad hoc*-metoder. Det ville være interessant at forfølge MCMC eller lignende metoder til estimation af parametre i en DGLM for at se, om det gav andre resultater.

Kapitel 9

Diskussion

Trenden, sæson- og ugedagsvariationen af incidensen af AMI i Nordjyllands amt i perioden 1983-1999 er blevet beskrevet ved hjælp af tre forskellige modeller i kapitlerne 6, 7 og 8. I dette kapitel vil resultaterne af data-analysen kort blive opsummeret og diskuteret.

9.1 Opsamling af resultater fra data-analysen

9.1.1 Trend

Data-analysen med en statisk Poisson model viste, at trenden af AMI-incidensraten er lineær på log-skalaen både for den samlede befolkning og stratificeret på køn og alder. Analyserne med henholdsvis en basic structural model og en dynamisk Poisson-model gav også resultatet, at trenden er approksimativt lineær. Det forventede relative fald af incidensraten pr. år for den samlede befolkning er på 3.2%. Faldet har været størst for mænd og for de 60-69 årige.

9.1.2 Sæsonvariation

Analysen af sæsonvariationen af AMI-incidensen viste toppunkter i april/maj og november/december samt et lavpunkt i august. Desuden er det vist, at sæsonvariationen afhænger af alder. For de yngste er incidensen klart størst om foråret, mens den topper i december for de ældste. Til gengæld er der ikke signifikant forskel på sæsonvariationen mellem kvinder og mænd.

Sæsonvariationen har ifølge den statiske analyse ikke ændret sig i løbet af undersøgelsesperioden, mens den dynamiske analyse viste, at sæsonvariationen bevæger sig. Fra 1983 til 1999 er incidensen blevet relativt mindre i forårs månederne, mens den er blevet relativt større om sommeren og i november/december. Desuden er udsvingene i sæsonvariation størst i midten af undersøgelsesperioden. Ændringerne er dog forholdsvis små, idet sæsonvariationen for den statiske model ligger indenfor 95% konfidensintervallet af sæsonvariationen for den dynamiske model.

9.1.3 Ugedagsvariation

Analysen af ugedagsvariationen af AMI-incidens viste, at incidensen er størst om mandagen og mindst i weekenden. Den dynamiske analyse viste kun små ændringer i ugedagsvariationen.

9.2 Diskussion af resultater

Den statiske analyse gav med små afvigelser de samme resultater, som blev udledt i [10] med hensyn til trend, sæson- og ugedagsvariation.

Den markant lavere incidens af AMI om sommeren stemmer godt overens med den amerikanske undersøgelse, [23], af ca. 260000 AMI-tilfælde i perioden 1. juli 1994 til 31. juli 1996. Dog blev der i den undersøgelse observeret det største antal AMI'er i perioden 21. december til 19. marts, mens analysen i dette speciale har vist et toppunkt allerede i november/december. Dog bør det holdes for øje, at den amerikanske undersøgelse ikke begrænsede sig til registrering af førstegangs-AMI'er, som det er tilfældet for datasættet fra Nordjylland, hvilket kan have en betydning for sæsonvariationen. Årsagen til den større risiko om vinteren menes at være, at en lang række risikofaktorer forekommer hyppigere på denne årstid som beskrevet i afsnit 5.2. En mulig forklaring på november/december toppunktet kunne være, at kroppen har svært ved at klare det markante fald i temperaturen forårsaget af den pludselige vinterkulde, som gerne forekommer i november måned i Danmark.

Som vist i analysen er det først og fremmest personer over 79 år, der har en forhøjet risiko for at blive ramt af en førstegangs-AMI i november/december. Dette indikerer, at risikofaktorerne, der forekommer på denne årstid, har størst betydning for denne aldersgruppe. Toppunktet om foråret for de yngste blev ikke udledt i [23] og er umiddelbart svært at forklare. Muligvis er stress en vigtig risikofaktor for personer i den erhvervs-aktive alder, hvilket dog ikke umiddelbart forklarer den udledte sæsonvariation for denne gruppe.

Der er så vidt det vides ikke tidligere beskrevet analyser af AMI-data med en dynamisk model. Som beskrevet i afsnit 9.1.2 skete der en relativ stigning i AMI-incidens i november/december fra ca. 1.5% over gennemsnittet i 1983 til ca. 8% over gennemsnittet i 1995, hvorefter den begyndte at falde igen. Dette kunne eventuelt hænge sammen med, at der blev spist mere usund mad i disse måneder i midten af halvfemserne end i midten af firserne. Den udledte dynamiske Poisson-model viser således, hvordan sæsonvariationen udvikler sig på længere sigt. Til gengæld er det ikke muligt at se ændringerne på kort sigt, hvilket kunne være interessant med henblik på at udlede en eventuel sammenhæng mellem incidens og f.eks. temperatur. Eksempelvis kunne en markant højere incidens end normalt i en usædvanlig kold november måned indikere, at der er en sådan sammenhæng. Til det formål ville en model med større sæsonvariations-varians, som den der blev udledt i analysen med en BSM, være at foretrække.

Den udledte ugedagsvariation af incidensen stemmer godt overens med [16] og [28], der rapporterer om henholdsvis 9 og 20% større risiko for en AMI om mandagen i forhold til de øvrige ugedage.

9.3 Konklusion

Hovedformålet med anvendelsesdelen af dette speciale var at undersøge, om der samtidig med, at AMI-incidensen er faldet fra 1983-1999, er sket en ændring af sæsonvariationen. Den statiske data-analyse viste, at dette ikke var tilfældet og en grundig modelkontrol gav ikke anledning til at forkaste den anvendte model. I analysen med en dynamisk Poisson model blev der udledt relativt små ændringer i sæsonvariationen, blandt andet en øget incidens i november/december set i forhold til de øvrige måneder. Årsagen kunne være, at kostvanerne er blevet dårligere i disse måneder.

Som beskrevet i kapitel 8 er den anvendte dynamiske model blevet udledt ved hjælp af *ad hoc*-metoder, og det kunne være interessant på et senere tidspunkt at afprøve mere stringente metoder til estimation af variansparametre. Men der er ingen tvivl om, at jeg har opnået en masse erfaring med data-analyse med dynamiske modeller. Anvendelsen af helt nyt software har naturligt nok budt på vanskeligheder undervejs, hvilket har affødt en række forbedringsforslag til den næste udgave af *sspir*.

Bilag A

Appendix

A.1 Den multivariable normalfordeling

Følgende vigtige resultat kan findes i [25] s. 637-638.

Sætning A.1 Lad $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \underline{\underline{V}})$ og lad

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \mathbf{m} = \begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{pmatrix}, \quad \underline{\underline{V}} = \begin{pmatrix} \underline{\underline{V}}_{11} & \underline{\underline{V}}_{12} \\ \underline{\underline{V}}_{21} & \underline{\underline{V}}_{22} \end{pmatrix},$$

hvor $\mathbf{X}_i \sim \mathcal{N}(\mathbf{m}_i, \underline{\underline{V}}_{ii})$, $i = 1, 2$. Der gælder, at den betingede fordeling af $\mathbf{X}_1$ givet $\mathbf{X}_2$ er givet ved

$$(\mathbf{X}_1 | \mathbf{X}_2) \sim \mathcal{N}[\mathbf{m}_1(\mathbf{X}_2), \underline{\underline{V}}_{11}(\mathbf{X}_2)], \quad (\text{A.1})$$

hvor

$$\mathbf{m}_1(\mathbf{X}_2) = \mathbf{m}_1 + \underline{\underline{V}}_{12} \underline{\underline{V}}_{22}^{-1} (\mathbf{X}_2 - \mathbf{m}_2),$$

og

$$\underline{\underline{V}}_{11}(\mathbf{X}_2) = \underline{\underline{V}}_{11} - \underline{\underline{V}}_{12} \underline{\underline{V}}_{22}^{-1} \underline{\underline{V}}_{21}.$$

Korollar A.2 $\mathbf{X}_1 - \mathbf{m}_1(\mathbf{X}_2)$ er uafhængig af $\mathbf{X}_2$.

Bevis:

Der gælder, at

$$\begin{aligned}
 \text{Cov}[\mathbf{X}_1 - \mathbf{m}_1(\mathbf{X}_2), \mathbf{X}_2] &= \text{Cov}\{\mathbf{X}_1 - [\mathbf{m}_1 + \underline{\underline{V}}_{12}\underline{\underline{V}}_{22}^{-1}(\mathbf{X}_2 - \mathbf{m}_2)], \mathbf{X}_2\} \\
 &= \text{Cov}(\mathbf{X}_1, \mathbf{X}_2) - \underline{\underline{V}}_{12}\underline{\underline{V}}_{22}^{-1}\text{Cov}(\mathbf{X}_2 - \mathbf{m}_2, \mathbf{X}_2) \\
 &= \text{Cov}(\mathbf{X}_1, \mathbf{X}_2) - \underline{\underline{V}}_{12}\underline{\underline{V}}_{22}^{-1}\underline{\underline{V}}_{22} \\
 &= \underline{\underline{0}}.
 \end{aligned}$$

Da $\mathbf{X}_1 - \mathbf{m}_1(\mathbf{X}_2)$ og $\mathbf{X}_2$ er multivariabelt normalfordelte, følger det, at de er uafhængige. ■

A.2 Linear Bayes' estimation

Sætning A.1 gælder kun i det multivariabelt normalfordelte tilfælde. Hvis det ikke kan antages, at $\mathbf{X}_1$ og $\mathbf{X}_2$ er multivariabelt normalfordelte, kan middelværdi og variansmatrix for den betingede fordeling af $\mathbf{X}_1$ givet $\mathbf{X}_2$ estimeres ved *linear Bayes' estimation*. Følgende resultat benyttes i de efterfølgende udregninger.

Sætning A.3 Lad $\underline{\underline{A}}$ være en $n \times m$ matrix og lad $\underline{\underline{B}}$ være en $m \times n$ matrix. Der gælder, at

$$\text{trace}(\underline{\underline{AB}}) = \text{trace}(\underline{\underline{BA}}).$$

Bevis:

Der gælder, at

$$\text{trace}(\underline{\underline{AB}}) = \sum_{i=1}^n (\underline{\underline{AB}})_{ii} = \sum_{i=1}^n \sum_{k=1}^m A_{ik} B_{ki} = \sum_{k=1}^m \sum_{i=1}^n B_{ki} A_{ik} = \sum_{k=1}^m (\underline{\underline{BA}})_{kk} = \text{trace}(\underline{\underline{BA}}).$$

■

Antag at $\mathbf{X}_1$ er en $n \times 1$ vektor, og at $\mathbf{X}_2$ er en $p \times 1$ vektor. Lad $\mathbf{d}$ være et vilkårligt estimat af $\mathbf{X}_1$ og lad præcisionen af $\mathbf{d}$ være givet ved en *tabs-funktion* $T(\mathbf{X}_1, \mathbf{d})$, der er defineret ved

$$T(\mathbf{X}_1, \mathbf{d}) = (\mathbf{X}_1 - \mathbf{d})^T (\mathbf{X}_1 - \mathbf{d}) = \text{trace}(\mathbf{X}_1 - \mathbf{d})^T (\mathbf{X}_1 - \mathbf{d}) = \text{trace}(\mathbf{X}_1 - \mathbf{d})(\mathbf{X}_1 - \mathbf{d})^T.$$

Lad $r(\mathbf{d})$ benævne det forventede tab for $\mathbf{d}$ givet $\mathbf{X}_2$. Der gælder, at

$$r(\mathbf{d}) = E[T(\mathbf{X}_1, \mathbf{d} | \mathbf{X}_2)] = \text{trace } E\{[\mathbf{X}_1 - \mathbf{d}(\mathbf{X}_2)][\mathbf{X}_1 - \mathbf{d}(\mathbf{X}_2)]^T\}. \quad (\text{A.2})$$

Givet $\mathbf{X}_2$ minimeres (A.2) for $\mathbf{d}(\mathbf{X}_2) = \mathbf{m} = E(\mathbf{X}_1|\mathbf{X}_2)$, og det følger, at

$$r(\mathbf{m}) = \text{trace } \text{Var}(\mathbf{X}_1|\mathbf{X}_2). \quad (\text{A.3})$$

Problemet er, at $E(\mathbf{X}_1|\mathbf{X}_2)$ ikke kan beregnes, fordi den præcise fordeling af $\mathbf{X}_1$ og $\mathbf{X}_2$ ikke kendes.

Ideen bag linear Bayes' estimation er at bestemme det $\mathbf{d}(\mathbf{X}_2)$, kaldet *linear Bayes' estimatet*, der minimerer (A.2), men hvor mængden af mulige estimater er indskrænket til lineære estimater på formen

$$\mathbf{d}(\mathbf{X}_2) = \mathbf{h} + \underline{\underline{H}}\mathbf{X}_2,$$

hvor $\mathbf{h}$ er en $n \times 1$ vektor, og $\underline{\underline{H}}$ er en $n \times p$ matrix. Linear Bayes' estimatet er dermed en lineær approksimation af $E(\mathbf{X}_1|\mathbf{X}_2)$. Det vises i [25] s. 124, at linear Bayes' estimatet er identisk med det eksakte resultat for det normalfordelte tilfælde. Det følger, at

$$\widehat{E}(\mathbf{X}_1|\mathbf{X}_2) = \mathbf{m}_1 + \underline{\underline{V}}_{12}\underline{\underline{V}}_{22}^{-1}(\mathbf{X}_2 - \mathbf{m}_2),$$

og

$$\widehat{\text{Var}}(\mathbf{X}_1|\mathbf{X}_2) = \underline{\underline{V}}_{11} - \underline{\underline{V}}_{12}\underline{\underline{V}}_{22}^{-1}\underline{\underline{V}}_{21}.$$

A.3 Store tals lov

Følgende sætning har betegnelsen *de store tals lov*. Et bevis for sætningen kan findes i [24] s. 210.

Sætning A.4 For ukorrelerede stokastiske variable, $X_1, \dots, X_n$, med samme middelværdi og varians gælder for ethvert $\varepsilon > 0$, at

$$\lim_{n \rightarrow \infty} \left[P \left(\left| \frac{X_1 + \dots + X_n}{n} - \mu \right| < \varepsilon \right) \right] = 1, \quad (\text{A.4})$$

hvor $\mu = E(X_i)$, $i = 1, \dots, n$.

Litteratur

- [1] A. R. Al-Yusuf, J. Kolar, S. K. Bhatnagar, A. Hudak, J. Smid. Seasonal variation in the incidence of unstable angina and acute myocardial infarction: effect of dry hot climate on the occurrence of complications following acute myocardial infarction. *Journal of Tropical Medicine and Hygiene*, 89(4):157-61, 1986.
- [2] B. P. Carlin, N. G. Polson, D. S. Stoffer. A Monte Carlo Approach to Nonnormal and Non-linear State-Space Modeling. *Journal of the American Statistical Association*, 87(418):493-500, 1992.
- [3] C. Dethlefsen. *Space time problems and applications*. PhD, Aalborg Universitet, 2002.
- [4] C. Dethlefsen, S. L. Christensen. Formulating state space models in R with focus on longitudinal regression models, R-21, 2005 (to appear in J. Stat. Software).
<http://cran.md.tsukuba.ac.jp/src/contrib/Descriptions/sspir.html>.
- [5] P. J. Diggle. *A Biostatistical Introduction*. Oxford, 1990.
- [6] A. J. Dobson. *An Introduction To Generalized Linear Models*. 2. edition. Chapman and Hall, 2001.
- [7] J. Durbin, S. J. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.
- [8] L. Fahrmeir. Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, 87(418):501-509, 1992.
- [9] L. Fahrmeir, G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. 2.edition. Springer, 2001.

- [10] T. Fischer, S. L. Christensen, S. P. Johnsen, H. C. Schønheyder, H. T. Sørensen. Secular trends and seasonality in first-time hospitalization for acute myocardial infarction - a Danish population-based study. *International Journal of Cardiology*, 2004.
- [11] V. Grech, O. Aquilina, J. Pace. Gender differences in seasonality of acute myocardial infarction admissions and mortality in a population-based study. *Journal of Epidemiology and Community Health*, 55:147-148, 2001.
- [12] C. Haran. The Lion in Winter: Why Heart Attacks Peak During the Holidays.

http://healthology.healthology.com/focus_article.asp?b=healthology&f=cardio&c=cardio_winterattack&spg=SCH.

17/12 2004.
- [13] A. C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, 1989.
- [14] T. J. Hastie, R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.
- [15] B. Jørgensen, S. L. Christensen, P. X. Song, L. Sun. A State Space Model for Multivariate Longitudinal Count Data. *Biometrika*, 86(1):169-181, 1999.
- [16] C. Ku, C. Yang, W. Lee, H. Chiang, C. Liu, S. Lin. Absence of a Seasonal Variation in Myocardial Infarction Onset in a Region without Temperature Extremes. *Cardiology*, 89(4), 1998.
- [17] S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.
- [18] J. A. Nelder, R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370-384, 1972.
- [19] V. V. Petrov. *Limit Theorems of Probability Theory*. Oxford University Press, 1995.
- [20] B. D. Ripley. Time Series in R 1.5.0. *R News*, 2/2, 2002.
- [21] K. J. Rothman. *Epidemiology - An Introduction*. Oxford, 2002.
- [22] E. Samoli, J. Schwartz, B. Wojtyniak, G. Touloumi, C. Spix, F. Balducci, S. Medina, G. Rossi, J. Sunyer, L. Bacharova, H. R. Anderson, K. Katsouyanni. Investigating Regional Differences in Short-Term Effects of Air Pollution on Daily Mortality in the APHEA Project: A Sensitivity Analysis for Controlling Long-Term Trends and Seasonality. *Environmental Health Perspectives*, 109(4), 2001.

- [23] F. A. Spencer, R. J. Goldberg, R. C. Becker, J. M. Gore. Seasonal distribution of acute myocardial infarction in the second National Registry of Myocardial Infarction. *Journal of the American College of Cardiology*, 31(6):1226-33, 1998.
- [24] M. Sørensen. *En introduktion til sandsynlighedsregning*. 3. edition. Afdeling for Teoretisk Statistik, Københavns Universitet, 2002.
- [25] M. West, P. J. Harrison. *Bayesian Forecasting and Dynamic Models*. 2. edition. Springer, 1997.
- [26] M. West, P. J. Harrison, H. S. Migon. Dynamic Generalized Linear Models and Bayesian Forecasting. *Journal of the American Statistical Association*, 80(389):73-83, 1985.
- [27] <http://www.hjerteforeningen.dk/sw6497.asp>. 13/4 2004.
- [28] Timing is everything: Fluctuations in cardiac risk. *Harvard Men's Health Watch*, december 2004.