Matematisk statistik
Stockholms universitet

Modellering och prediktion av tidsserier gällande sjukförmåner inom socialförsäkringen

Per Johansson

Examensarbete 2006:8

Postal address:

Matematisk statistik
Dept. of Mathematics
Stockholms universitet
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Matematisk statistik
Stockholms universitet
Examensarbete **2006:8**,
http://www.math.su.se/matstat

Modellering och prediktion av tidsserier gällande sjukförmåner inom socialförsäkringen

Per Johansson*

Juni 2006

Sammanfattning

Sjukpenning och rehabiliteringspenning är två av de förmåner inom socialförsäkringen som är dominerande vad gäller inkomstbortfall vid arbetsoförmåga. I det här arbetet analyseras tidsseriedata med antalet utbetalade dagar av de ovan nämnda förmånerna där data erhållits från Försäkringskassan. Prognoser har gjorts för de åtta tidsserierna i analysen, fem stycken gällande sjukpenning samt tre stycken gällande rehabiliteringspenning, baserat på framtagna modeller för transformationer av serierna. Prognoserna för sju serier visar på en minskning av utbetalade dagar medan en serie för rehabiliteringspenningen visar på en ökning av utbetalade dagar.

*Postal address: Matematisk statistik, Stockholms universitet, SE-106 91, Sweden. E-mail: perje75@hotmail.com. Handledare: Joanna Tyrcha.

Abstract

Two of the benefits in the social insurance that are dominant at loss of income by working inability are sickness benefit and the other according to rehabilitation. In this thesis time series data is analysed applied to disbursements of the above described benefits where data has been obtained by Försäkringskassan. For the eight series in the analysis, five according to sickness benefit and three according to rehabilitation, forecasts have been proposed based on models related to transformations of the series. The forecasts of all series but one indicate a decrease of disbursements.

Förord

Det här examensarbetet har utförts under ht05/vt06 omfattande 20 poäng för magisterexamen i matematisk statistik vid Stockholms Universitet. Mitt syfte med det här arbetet har varit att ta fram modeller för tidsserier gällande sjukförmåner.

Jag vill tacka Jan Eriksson på Försäkringskassans huvudkontor för datamaterialet som analysen bygger på. Ett tack till Försäkringskassan för att jag fick en arbetsplats på huvudkontoret att arbeta från samt tack till Ola Rylander med flera för den hjälp jag fått med frågeställningar. Slutligen vill jag tacka min handledare Joanna Tyrcha vid Stockholms Universitet för den handledning jag fått under arbetets gång.

Innehåll

Inledning

Försäkringskassan administrerar den svenska socialförsäkringen och ansvarar för huvuddelen av samhällets ekonomiska skyddsnät. Den 1 januari 2005 inrättades Försäkringskassan genom en sammanslagning av Riksförsäkringsverket (RFV) och de 21 allmänna försäkringskassorna runt om i landet. De föreskrifter, råd och vägledningar som beslutats av RFV gäller även efter Försäkringskassans inrättande. Försäkringskassans föreskrifter och allmänna råd kan delas in i följande områden:

sjukförmåner
barn, familj och handikapp
pension
allmänt

Totalt finns närmare femtio olika förmåner eller bidrag inom socialförsäkringen. I det här examensarbetet analyseras data observerat över tid, tidsseriedata, gällande sjukförmånerna sjukpenning och rehabiliteringspenning.

1. Bakgrund

1.1 Sjukförmåner

Av de förmåner som finns i socialförsäkringen är sjukpenning, rehabiliteringspenning samt aktivitets- och sjukersättning de dominerande ersättningarna för inkomstbortfall vid arbetsoförmåga. För förmånerna gäller att ersättning betalas som en kvarts, en halv, tre kvarts eller en hel förmån beroende på hur pass mycket arbetsförmågan är nedsatt för respektive individ. Inom ohälsostatistiken mäter man utbetalningarna i antal dagar där förmånerna omräknats till ”hela dagar”, s k nettodagar, d.v.s att två halvdagar blir en heldag osv. Vid summering av bruttodagar omvandlas alla förmåner till heldagar och därefter summeras alla heldagar.

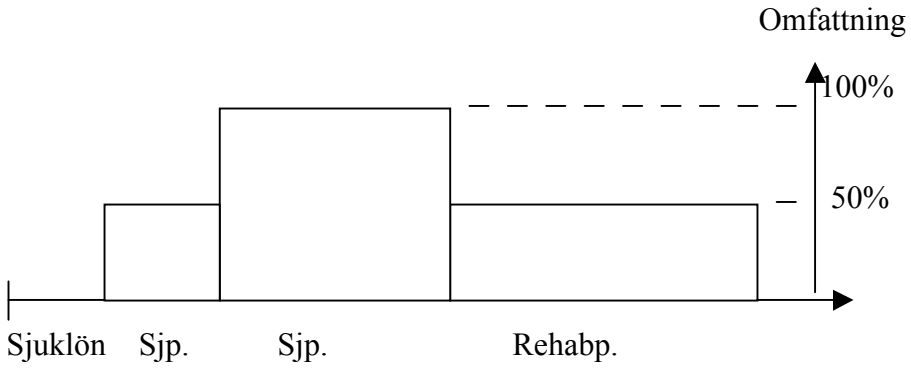
1.2 Sjukfall

Vid ett sjukfall för en arbetstagare betalar arbetsgivaren ut sjuklön de 14 första dagarna, där dag ett utgör karensdag. Fortsätter sjukfallet efter de två första veckorna övertar Försäkringskassan utbetalningen och betalar fortsättningsvis ut sjukpenning. För studenter, egenföretagare m fl betalar Försäkringskassan ut sjukpenning från dag ett. Ett sjukfall definieras i det här arbetet som för en individ på varandra följande utbetalningar av sjukförmånerna sjukpenning och/eller rehabiliteringspenning. Ett sjukfall kan ha mellan en upp till flera hundra

utbetalningar beroende på längden på sjukfallet. En utbetalning har alltid tre datum som identifierar utbetalningen. Dag ett i sjukfallet; första dagen i utbetalningen, s k fr.o.m-datum samt sista dagen i utbetalningen, s k t.o.m-datum. Om tidsskillnaden från t.o.m-datum i en utbetalning till fr.o.m-datum i nästa utbetalning i det genererade sjukfallet överstiger fem arbetsdagar betraktas den nya utbetalningen som den första utbetalningen på ett nytt sjukfall.

1.3 Sjukpenning och rehabiliteringspenning

De sjukförmåner som kan ingå i ett sjukfall är sjukpenning, rehabiliteringspenning och arbetsskadepenning. Antalet utbetalade dagar av arbetsskadepenning är dock så liten i förhållande till de andra två förmånernas utbetalningar så den har ej tagits med i analysen. Nedan visas hur ett sjukfall kan vara uppbyggt med olika förmåner och med olika sjukskrivningsgrad.



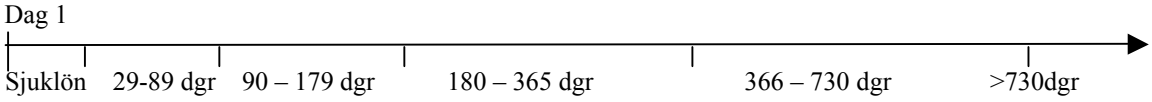
Om en person haft sjukpenning i ca ett år prövas det om det finns möjlighet till att få rehabiliteringspenning i avsikt att kunna komma tillbaka till arbetslivet i full omfattning. För en arbetstagare gäller att arbetsgivaren har förstahandsansvaret för att uppmärksamma och utreda behov av arbetsinriktad rehabilitering. Arbetsgivaren har alltid skyldighet att utreda en arbetstagares behov av rehabiliteringsåtgärder när denne varit sjukskriven över åtta veckor. Försäkringskassan ska bedöma den försäkrades behov av och möjlighet till rehabilitering, detta gäller även för arbetslösa. Rehabiliteringspenning utgår under den tiden som arbetslivsinriktad rehabilitering pågår. Om den försäkrade helt saknar arbetsförmåga utgår hel rehabiliteringspenning. Är arbetsförmågan inte nedsatt helt men är nedsatt med minst trefjärdedelar utgår trefjärdedels rehabiliteringspenning. På samma sätt utgår halv och en fjärdedels rehabiliteringspenning. Den som är sjukskriven på heltid är berättigad till hel rehabiliteringspenning när rehabiliteringsperioden börjar.

Utbetalning av rehabiliteringspenning föregås i samtliga fall av en sjukpenningperiod. Målet med rehabiliteringspenning är ju ett återinträdande till arbetslivet men det är inte ovanligt med sjukfall där sjukpenning- och rehabiliteringsperioder avlöser varandra. Om rehabiliteringsperioden efterföljs av

en period med sjukpenning betraktas det antingen som en fortsättning av det pågående sjukfallet eller som ett nytt sådant. På samma sätt för rehabiliteringsperioder mellan sjukpenningperioder.

1.4 Syfte och mål med arbetet

I det här examensarbetet har jag valt att analysera tidsserier med månadsvisa data innehållande antalet utbetalda nettodagar för hela landet gällande sjukpenning och rehabiliteringspenning. Fem tidsserier består av sjukpenningdagar och tre serier består av rehabiliteringspenningdagar. Ett sjukfall börjar alltid, som tidigare nämnts, med utbetalning av sjuklön från arbetsgivaren och för vissa grupper med sjukpenning från dag ett. För varje sjukfall gäller att de kan delas upp i olika tidsintervall, s k längdklasser, se Figur 1.1. Antalet av Försäkringskassan utbetalda nettodagar summeras månad för månad i en tioårsperiod för de två förmånerna inom respektive längdklass. En utbetalning för en enskild individ som består av sjukdagar 70 – 100 bidrar alltså till den första längdklassen med 20 dagar samt till den andra med 11 dagar. Syftet med uppsatsen är att med hjälp av de tidsserier som datamaterialet består av, alla innehållande 133 månadsvisa observationer från januari 1994 till januari 2005 ta fram en ARMA-modell för respektive tidsserie och utifrån den framtagna modellen prediktera framtida värden för respektive längdklass och sjukförmån.



Figur 1.1 Sjukfall uppdelat i längdklasser.

I de sjukfall som bara innehåller sjukpenning räknas dagarna i längdklasserna från dag ett i sjukfallet. För sjukfall med rehabiliteringspenning (som ju föregås av minst en period med sjukpenning) räknas dagarna i längdklasserna gällande rehabiliteringspenning från den tidpunkt då sjukpenningen övergår i rehabiliteringspenning. För de sjukfall där rehabiliteringspenning efterföljs av en period med ytterligare sjukpenning finns två möjligheter. Antingen ses den nya sjukpenningperioden som en fortsättning av den gamla eller så betraktas den nya perioden som ett nytt sjukfall och då räknas dagarna i längdklasserna om från dag ett igen.

Längdklass	Sjukpenning
1	29 – 89 dagar
2	90 – 179 dagar
3	180 – 365 dagar
4	366 – 730 dagar
5	> 730 dagar

Tabell 1.1 Antal dagar från sjukfallets eller sjukpenningperiodens början.

Längdklass	Rehabiliteringspenning
1	29 – 89 dagar
2	90 – 179 dagar
3	180 – 365 dagar

Tabell 1.2 Antal dagar från rehabiliteringsperiodens början.

1.5 Historik

Nedan en kort historik om ändringar av regler för sjukpenningen:

Kort historik:

Mars 1991: Ersättningsnivån på sjukpenning sänks från 90% till 65% (sjukdag 1-3) och 80% (sjukdag 15 – 90).

Januari 1992: Sjuklön från arbetsgivare till anställda införs under sjukperiodens 14 första dagar. Sjukpenning till anställda betalas från dag 15 i sjukperioden.

April 1993: En karensdag infördes. Ersättningsnivån på sjukpenning sänks från 90% till 80% från sjukdag 91.

Januari 1996: Ersättningsnivån sänks generellt till 75%

Januari 1997: Sjuklön från arbetsgivare till anställda förlängs till sjukperiodens 28 första dagar.

Januari 1998: Ersättningsnivån höjs generellt till 80%.

April 1998: Sjuklön från arbetsgivare till anställda förkortas till sjukperiodens 14 första dagar.

Juli 2003: Sjuklön från arbetsgivare till anställda förlängs till sjukperiodens 21 första dagar. Ersättningsnivån på sjukpenning sänks till 77,6%.

Januari 2005: Sjuklön från arbetsgivare till anställda förkortas till sjukperiodens 14 första dagar. Ersättningsnivån på sjukpenning höjs till 80%.

I och med att sjuklöneperioden ändrats över tid, med längd som skiftat mellan 14 och 28 dagar, inom de tio åren som analysen baseras på har dataanalys och modellering för sjukpenningen gjorts från och med längdklassen 29 – 89 dagar som ju är den första längdklassen där de månadsvisa observationerna är jämförbara. Även för rehabiliteringspenningen är det fr.o.m denna längdklass som analysen görs.

2 Teori

2.1 Stokastiska processer

En stokastisk process kan definieras som en sekvens av stokastiska variabler $X(t), t \in T,$ där T är de tidpunkter för vilken processen är definierad. Inom tidsserieanalys är det allt för oftast en observation per tidsenhet, kallad $x(t)$ om observationerna är kontinuerliga och den betecknas x_t om observationerna är diskreta. Nedan beskrivs teori om diskreta observationer inom tidsserieanalys som data i det här arbetet baseras på. En tidsseriemodell för observerade data $\{x_t\}$ är en specifikation av fördelningsfunktionerna av en sekvens av stokastiska variabler $\{X_t\}$ där $\{x_t\}$ är de stokastiska variablernas observerade värden (utfall):

$$P(X_1 \le x_1, \ldots, X_n \le x_n) \quad -\infty < x_1, \ldots, x_n < \infty \quad n = 1, 2, \ldots \tag{1}$$

En sådan specificering är ofta för komplex då den kommer att innehålla för många parametrar att skatta utifrån befintliga data. Istället används första och andramomenten av fördelningsfunktionerna d.v.s väntevärdet $E(X_t)$, variansen $Var(X_t)$ och $Cov(X_{t+h}, X_t) = \gamma_X(h), t = 1, 2, \ldots, h = 0, 1, 2, \ldots$. I det speciella fallet då alla fördelningsfunktioner är multivariat normalfördelade bestäms fördelningen helt och hållet av de två momenten.

2.2 Stationäritet

Definition 1

Låt $\{X_t\}$ vara en tidsserie med $E(X_t^2) < \infty$. Väntevärdet av $\{X_t\}$ vid tiden *t* är $\mu_X(t) = E(X_t)$. Kovariansfunktionen av $\{X_t\}$ är $\gamma_X(r,s) = Cov(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))]$ för alla heltal *r* och *s*.

Definition 2

$\{X_t\}$ är (svagt) stationär om
(i) $\mu_X(t)$ är oberoende av *t*, och
(ii) $\gamma_X(t+h,t)$ är oberoende av *t* för varje *h*.

Strikt stationäritet av en tidsserie innebär att $(X_1, \ldots, X_n)$ och $(X_{1+h}, \ldots, X_{n+h})$ har samma fördelningsfunktioner för alla heltal *h* och $n > 0$.

Definition 3

Låt $\{X_t\}$ vara en stationär tidsserie. Autokovariansfunktionen (ACVF) av $\{X_t\}$ vid tidsförskjutning *h* är $\gamma_X(h) = Cov(X_{t+h}, X_t)$. Autokorrelationsfunktionen (ACF) av $\{X_t\}$ vid tidsförskjutning h är $\rho_X(\mathrm{h}) = \frac{\gamma_X(h)}{\gamma_X(0)} = Cor(X_{t+h}, X_t)$.

Definition 4

$\{X_t\}$ kallas vitt brus om det är en sekvens av okorrelerade stokastiska variabler, alla med väntevärde noll och varians σ^2. Vitt brus beskrivs nedan som $\mathrm{WN}(0, \sigma^2)$.

Definition 5

$\{X_t\}$ är en AR(p)-process om $\{X_t\}$ är stationär och om det för varje t gäller att $X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t$ där $Z_t \sim \mathrm{WN}(0, \_^2)$, $t = 0, \pm 1, \ldots$ och där $\phi_1, ..., \phi_p$ är konstanter.

Definition 6

$\{X_t\}$ är en MA(q)-process om $\{X_t\}$ är stationär och om det för varje t gäller att $X_t = Z_t + \__1 Z_{t-1} + \ldots + \__q Z_{t-q}$ där $Z_t \sim \mathrm{WN}(0, \_^2)$, $t = 0, \pm 1, \ldots$ och där $\theta_1, ..., \theta_q$ är konstanter.

2.3 Kausalitet

Definition 7

Tidsserien $\{X_t\}$ är en linjär process om den kan uttryckas som

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j} \quad \text{för alla t,} \tag{2}$$

där $\{Z_t\} \sim \mathrm{WN}(0, \_^2)$ och $\{\psi_j\}$ är en sekvens av konstanter med $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

Proposition 1

Låt $\{Y_t\}$ vara en stationär tidsserie med väntevärde noll och kovariansfunktion γ_Y. Om $\sum_{j=-\infty}^{\infty}|\psi_j| < \infty$, så är tidsserien $X_t = \sum_{j=-\infty}^{\infty}\psi_j Y_{t-j}$ stationär med väntevärde noll och ACVF

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\psi_j\psi_k\gamma_Y(h+k-j). \tag{3}$$

I det speciella fallet då $\{X_t\}$ är en linjär process gäller:

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty}\psi_j\psi_{j+h}\sigma^2 \tag{4}$$

Om $\psi = 0$ för alla j < 0 dvs om $X_t = \sum_{j=0}^{\infty}\psi_j Z_{t-j}$ så kallas den linjära processen för en MA(∞). Observera att X_t endast beror på tidigare värden av Z_t.

Låt oss titta närmare på AR(1)-processen $X_t - \phi X_{t-1} = Z_t$
där $\{Z_t\} \sim WN(0,\sigma^2)$, $|\phi| < 1$, och $\{Z_t\}$ är okorrelerad med X_s för varje $s < t$.
Betrakta den linjära processen

$$X_t = \sum_{j=0}^{\infty}\phi^j Z_{t-j} \tag{5}$$

Eftersom (5) är en lösning till AR(1)-processen ovan och Proposition 1 ger att den också är stationär med väntevärde noll och ACVF fås:

$$\gamma_x(h) = \sum_{j=0}^{\infty}\phi^j\phi^{j+h}\sigma^2 = \frac{\sigma^2\phi^h}{1-\phi^2} \qquad \text{för} \quad h \geq 0.$$

Nedan visas att (5) är den enda stationära lösningen av AR(1)-processen.

Låt $\{Y_t\}$ vara en godtycklig stationär lösning. Vi har

$$Y_t = \phi Y_{t-1} + Z_t = Z_t + \phi Z_{t-1} + \phi^2 Y_{t-2} = \ldots = Z_t + \phi Z_{t-1} + \ldots + \phi^k Z_{t-k} + \phi^{k+1} Y_{t-k-1}$$

Om $\{Y_t\}$ är stationär är $E(Y_t^2)$ ändlig och oberoende av *t* vilket medför att

$$E\left(Y_t - \sum_{j=0}^{k} \phi^j Z_{t-j}\right)^2 = \phi^{2k+2} E(Y_{t-k-1})^2 \to 0 \quad \text{då} \quad k \to \infty.$$

Detta ger att Y_t är lika med medelkvadratgränsvärdet $\sum_{j=0}^{\infty} \phi^j Z_{t-j}$ och att (5) är en unik lösning av AR(1)-processen nämnd ovan.
Serien i (5) konvergerar inte om $|\phi| > 1$. Genom att skriva om $X_t - \phi X_{t-1} = Z_t$ i formen $X_t = -\phi^{-1} Z_{t+1} + \phi^{-1} X_{t+1}$ fås efter några ekvationer att $X_t = -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j}$.
Denna lösning ses allmänt som onaturlig då X_t är korrelerad med framtida värden av Z_t. En viktig del av modellering av tidsserie är att prediktera framtida värden. Att då göra prognoser utifrån framtida värden innebär en motsägelse. Att X_t är representerad som (5) innebär att den är kausal, d.v.s. endast korrelerad med tidigare värden, i det här fallet enbart av Z_t.

2.4 ARMA-processer

En viktig parametrisk familj av stationära tidsserier är ARMA-processer.

Definition 8

$\{X_t\}$ är en ARMA(p,q)-process om $\{X_t\}$ är stationär och om det för varje *t* gäller att
$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}$
där $\{Z_t\} \sim WN(0,\sigma^2)$, samt att $(1 - \phi_1 z - \ldots - \phi_p z^p)$ och $(1 + \theta_1 z + \ldots + \theta_q z^q)$ inte har några minsta gemensamma nämnare.

För ARMA-processer gäller följande:

Definition 9

En stationär lösning $\{X_t\}$ av ARMA-processen i Definition 8 existerar (och är också en unik stationär lösning) om och endast om

$\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p \neq 0$ för alla $|z| = 1$.

Definition 10

En ARMA(p,q)-process $\{X_t\}$ är kausal, eller en kausal funktion av $\{Z_t\}$, om det finns konstanter $\{\psi_j\}$ så att $\sum_{j=0}^{\infty} |\psi_j| < \infty$ och $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ för alla *t*.
Kausalitet är ekvivalent med villkoret

$$\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p \neq 0 \qquad \text{för alla } |z| \leq 1.$$

Utgångspunkten för det här examensarbetet är prediktion av framtida värden för de olika tidsserierna och teorin för prediktion bygger på minsta kvadratmetoden. Det hela bygger på en minimering av $E(X_{n+h} - a_o - a_1 X_n - ... - a_n X_1)^2$ där $a_o, a_1, ..., a_n$ är konstanter och *h* är positivt heltal. För en MA(q)-process är det generellt så att det för motsvarande tidsseriens ACF gäller att korrelationer för tidsförskjutningar större än q är lika med noll. Det finns en korrelationsfunktion, PACF, som är relaterad till AR-processer på samma sätt som ACF är relaterad till MA-processer. Båda dessa funktioner kan vara ett bra verktyg då en modell ska tas fram genom att man tittar på tidsseriens båda dessa korrelationer, då seriens korrelationer för tidsförskjutningar upp till en fjärdedel av antalet observationer i serien väl kan approximeras med modellens ACF och PACF.

2.5 Akaikes informationskriterium

Vilken metod är lämpligast att använda då prognoser ska göras? För att kunna avgöra om en ARMA-process passar data används olika test som beskrivs nedan. Ett kriterium som visat sig vara användbart vid framtagande av en modell är det s k Akaikes informations-kriterium, AIC. Den bygger på en minimering av AIC-statistikan som består av maximumlikelihoodskattningen med avseende på parametrarna ϕ, _ och _^{2}. För att få en förståelse för hur ML-funktionen ser ut bör algoritmen i Appendix, nedan kallad algoritmen, först läsas.

Anta att $\{X_t\}$ är en Gaussiansk tidsserie med väntevärde noll och ACVF $\kappa(i,j) = E(X_i X_j)$. Låt $\boldsymbol{X_n} = (X_1, \ldots, X_n)'$ och låt $\hat{\boldsymbol{X}}_{\boldsymbol{n}} = (\hat{X}_1, \ldots, \hat{X}_n)$ där $\hat{X}_1 = 0$ och $\hat{X}_j = E(X_j \mid X_1, \ldots, X_{j-1}) = P_{j-1} X_j$, $j \geq 2$. Låt $\Gamma_{\boldsymbol{n}}$ stå för kovariansmatrisen $\Gamma_n = \mathrm{E}(\boldsymbol{X_n X_n'})$ och anta vidare att Γ_n är icke-singulär. Likelihooden av $\boldsymbol{X_n}$ är då:

$$L(\Gamma_{\boldsymbol{n}}) = (2\pi)^{-n/2} (\det \Gamma_{\boldsymbol{n}})^{-1/2} \exp(-(1/2)\, \boldsymbol{X_n'}\, \Gamma_{\boldsymbol{n}}^{-1}\, \boldsymbol{X_n})$$

En beräkning av det $\Gamma_{\boldsymbol{n}}$ och $\Gamma_{\boldsymbol{n}}^{-1}$ kan undvikas genom att uttrycka dessa i termer av enstegsprediktionsfelen $X_j - \hat{X}_j$ med varianser v_{j-1}, $j = 1, \ldots, n$ som kan beräknas rekursivt med algoritmen. Låt θ_{ij}, $j = 1, \ldots, i$, $i = 1, 2, \ldots$ vara

koefficienterna som fås då algoritmen tillämpats gällande ACVF κ av $\{X_t\}$. Med $\boldsymbol{C_n}$ såsom i algoritmen i Appendix fås att $\boldsymbol{X_n} = \boldsymbol{C_n}(\boldsymbol{X_n} - \hat{\boldsymbol{X}}_{\boldsymbol{n}})$. Det kan visas att komponenterna i $\boldsymbol{X_n} - \hat{\boldsymbol{X}}_{\boldsymbol{n}}$ är okorrelerade som ger att $\boldsymbol{X_n} - \hat{\boldsymbol{X}}_{\boldsymbol{n}}$ har diagonalkovariansmatrisen $\boldsymbol{D_n}$ = diag($v_0,...,v_{n-1}$). Vi har nu att Γ_n = $\boldsymbol{C_n D_n C_n'}$ samt att

$\boldsymbol{X_n'}\Gamma_n^{-1}\boldsymbol{X_n} = (\boldsymbol{X_n} - \hat{\boldsymbol{X}}_{\boldsymbol{n}})'\boldsymbol{D_n^{-1}}(\boldsymbol{X_n} - \hat{\boldsymbol{X}}_{\boldsymbol{n}}) = \sum_{j=1}^{n}(X_j - \hat{X}_j)^2 / v_{j-1}$ och

det $\Gamma_n = (\det \boldsymbol{C_n})^2(\det \boldsymbol{D_n}) = v_0 v_1 \ldots v_{n-1}$. Nu fås att likelihooden kan skrivas som:

$$\mathrm{L}(\Gamma_n) = \frac{1}{\sqrt{(2\pi)^n v_0 \ldots v_{n-1}}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(X_j - \hat{X}_j)^2 / v_{j-1}\right\} \tag{7}$$

Likelihooden för data från en ARMA(p,q)-process fås från algoritmen genom att räkna fram enstegsprediktionsfelen $\hat{X}_{i+1}$ och motsvarande medelkvadratfel v_i.

$$\hat{X}_{n+1} = \begin{cases} \sum_{j=1}^{n}\theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}) & 1 \le n < m \\ \phi_1 X_n + \ldots \phi_p X_{n+1-p} + \sum_{j=1}^{q}\theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), & n \ge m \end{cases} \tag{8}$$

samt $E(X_{n+1} - \hat{X}_{n+1})^2 = \sigma^2 E(W_{n+1} - \hat{W}_{n+1})^2 = \sigma^2 r_n$ där värdena på θ_{nj} och r_n bestäms från algoritmen.
Nu har vi följande:

$$L(\boldsymbol{\phi},\boldsymbol{\theta}\ \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n r_0 \ldots r_{n-1}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\frac{(X_j - \hat{X}_j)^2}{r_{j-1}}\right\} \tag{9}$$

Genom att partiellt differentiera $lnL(\boldsymbol{\phi},\boldsymbol{\theta},\ \sigma^2)$ med avseende på σ^2 samt att använda att $\hat{X}_j$ och r_j är oberoende av σ^2 fås ML-skattningarna $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\theta}}$ och $\hat{\sigma}^2$ till följande: $\hat{\sigma}^2 = n^{-1}\boldsymbol{S}(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}})$ där $\boldsymbol{S}(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}}) = \sum_{j=1}^{n}(X_j - \hat{X}_j)^2 / r_{j-1}$ och $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\theta}}$ är värdena av $\boldsymbol{\phi}$ och $\boldsymbol{\theta}$ som minimerar $l(\boldsymbol{\phi},\boldsymbol{\theta}) = ln(n^{-1}\boldsymbol{S}(\boldsymbol{\phi},\boldsymbol{\theta})) + n^{-1}\sum_{j=1}^{n} lnr_{j-1}$.

AIC-kriteriet: Välj $p,q,\boldsymbol{\phi}_p,\boldsymbol{\theta}_q$ som minimerar

$$\mathrm{AIC} = -2lnL(\boldsymbol{\phi}_p,\boldsymbol{\theta}_q, \boldsymbol{S}(\boldsymbol{\phi}_p,\boldsymbol{\theta}_q)/n) + 2(p+q+1) \tag{10}$$

Den sista termen är en ”straffterm” som baseras på att det vid prediktion av tidsseriedata ej är bra med för höga värden på p och q. En modell med höga värden på p och q ger generellt en liten skattad WN-varians men när den framtagna modellen används för att prediktera framtida värden beror medelkvadratfelet av prediktionerna ej endast på WN-variansen utan också på fel som uppstår från skattning av modellens parametrar.

2.6 Residualer

När en modell väl tagits fram är nästa steg att kolla hur pass bra den framtagna modellen passar tidsserien. För att göra detta betraktas residualerna

$$\hat{W}_t = (X_t - \hat{X}_t(\hat{\phi},\hat{\theta}))/(r_{t-1}(\hat{\phi},\hat{\theta}))^{1/2}, \quad t = 1,\ldots,n. \tag{11}$$

Den typ av residualer som används för att testa en modells duglighet fås genom att dividera residualerna med den skattade standardavvikelsen för WN d.v.s:

$$\hat{R}_t = \hat{W}_t / \hat{\sigma} \quad \text{där} \quad \hat{\sigma} = \sqrt{(\sum_{t=1}^{n} W_t^2)/n}.$$

Om den framtagna modellen är väl anpassad för tidsseriedata bör dessa residualer ha egenskaper liknande WN(0,1).

Det första signifikanstestet avgör om den stationära tidsserien är vitt brus. Skulle den vara det finns det inte någon ARMA- process som representerar den givna tidsserien. Ett annat test avgör om parametrarna i den framtagna modellen är signifikanta. Vi betraktar återigen residualerna

$$\hat{W}_t = (X_t - \hat{X}_t(\hat{\phi},\hat{\theta}))/(r_{t-1}(\hat{\phi},\hat{\theta}))^{1/2} \quad t = 1,\ldots n.$$

Dessa ska, om modellen är adekvat, ha egenskaper liknande

$$W_t(\phi,\theta) = (X_t - \hat{X}_t(\phi,\theta))/(r_{t-1}(\phi,\theta))^{1/2} \quad t = 1,\ldots n. \tag{12}$$

.

W_t är en approximation av WN-termen i ARMA-processen på så sätt att $E(W_t(\phi,\theta) - Z_t)^2 \to 0$ då $t \to \infty$. Alltså ska $\hat{W}_t$ ha egenskaper liknande Z_t i den framtagna ARMA-processen. En rad tester med nollhypotesen att residualerna är WN kan användas för att se om residualerna är WN(0,1). Med signifikansnivån 0,05 bör p-vädret vara åtminstone högre än 0,05, då ju nollhypotesen visar på att den framtagna modellen väl överens stämmer med tidsserien.

Om den modell med det optimala värdet på AIC-statistikan inte går genom alla signifikanstester letar man efter modeller med AIC-värden nära den lokala minipunkten med avseende på AIC-statistikan och stannar när en modell hittas där samtliga tester går inom.

3 Analys

3.1 Transformationer och differenciering

För att kunna göra prognoser framåt baserat på en modell för tidsseriedata, är det första delmomentet att från ursprungsdata ta fram en stationär tidsserie. Det är ju utifrån en stationär tidsserie som en möjlig kausal ARMA-modell kan tas fram och från denna modell, om den passar data tillräckligt bra, som prediktion av framtida värden kan göras. Vilket är då tillvägagångssättet för att ta fram en stationär tidsserie? För att få fram så mycket information som möjligt om en tidsserie plottas data över tid. De kännetecken som främst är viktiga är outliers, trend, säsongskomponent samt variation över tid.

Outliers kan försvåra analysen. Antingen betraktas den avvikande observationen som sann vilket medför att den ska vara med vid modellering av tidsserien. Betraktas observationen som en felaktig observation bör observationen få värdet av tidsseriens väntevärde alternativt om möjligt tas bort. Samtliga tidsserier i den här analysen har dock väl sammanhängande grafer.

En trend kan definieras som ”förändring av medelvärde över lång tid” och kan ha stor betydelse för modellering av data. Trenden som begrepp är relativ och beror på antalet observationer i tidsserien. Det som ses som en trend i ett delintervall av data kan i ett större sammanhang vara mindre betydelsefullt.

Det finns olika typer av transformationer som kan göras av ickestationära data för att få fram en ny serie som är mer kompatibel vad gäller stationäritet. Att transformera data kan ses som ett filter där den nya transformerade serien är nödvändig för att kunna hitta en modell. När väl en modell tagits fram kan t.ex. prediktion av modellen göras och via den prediktion av ursprungliga data. En förändring av variansen över tid kan reduceras genom att ta naturliga logaritmen av ursprungsdata. Den transformerade tidsserien blir då mer jämn och är bättre anpassad vid modellering.

I den här analysen har differenciering använts som metod för att eliminera trend och säsongskomponenter och det bygger som namnet antyder på att subtrahera observationer inom ett visst tidsintervall. Med observationer uppbyggda som en additiv modell enligt $X_t = m_t + s_t + Y_t$ där m_t är trendkomponenten i observationen, s_t är säsongskomponenten samt Y_t är en slumpmässig komponent fungerar differenciering som följer:

Anta att säsongskomponenten har period d , dvs att $s_t = s_{t+d}$. Genom att definiera differencieringsoperatorn $\nabla_d X_t = X_t - X_{t-d} = (1 - B^d)X_t$ fås att $\nabla_d X_t = m_t - m_{t-d} +$

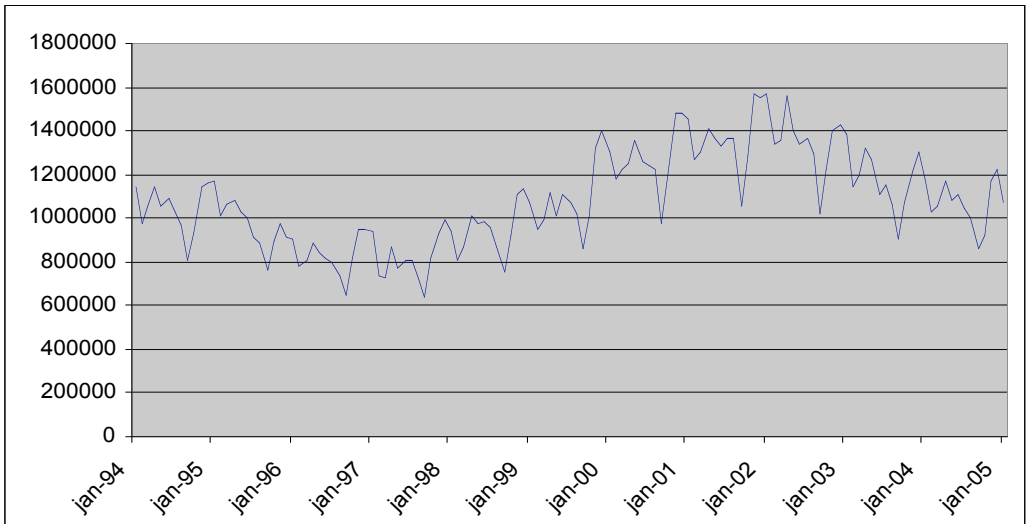
$Y_t - Y_{t-d}$ som endast har en trendkomponent och en slumpkomponent. För att därefter ta bort trenden, $m_t - m_{t-d}$, kan ett polynom av ∇ användas.

3.2 Univariata och multivariata metoder

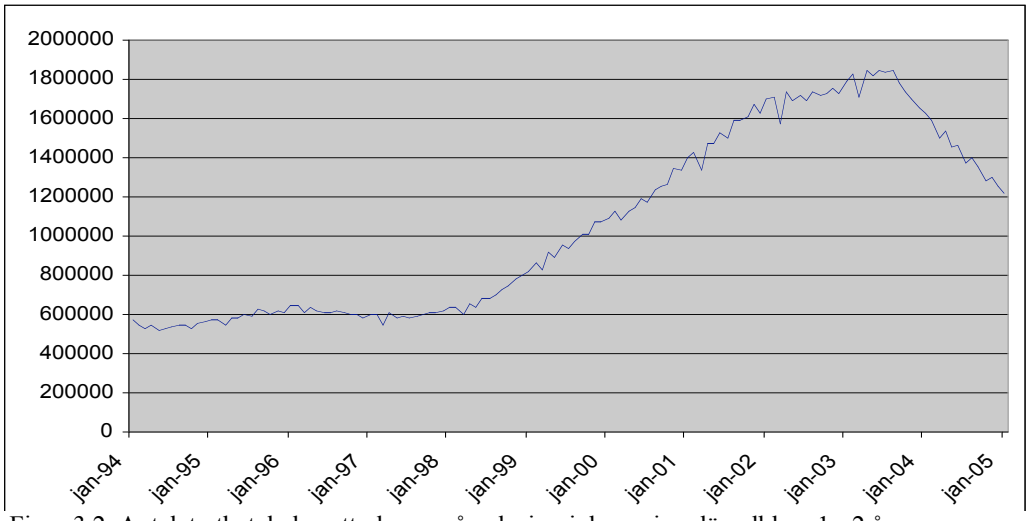
Analys av tidsserier kan delas upp i antingen en univariat eller multivariat metod. Den univariata metoden bygger på att man tittar på korrelation och beroende inom tidsserien som analyseras. Alternativet hade varit en multivariat tidsserie som är en vektorvärd tidsserie där analysen bygger på att korrelation och beroende inte bara analyseras inom en tidsserie utan även mellan serierna. De kompletterande tidsserierna är ofta av den karaktären att man antar en viss korrelation mellan var och en av dessa och den tidsserie som står till grund för analysen. En faktor som skulle kunna vara med i analysen är ifall antalet sjukskrivningar tenderar att ändras när ersättningsnivåerna för sjukförmånerna ändras. En annan faktor skulle kunna vara regionala skillnader i utgifterna för ohälsan sett över tid. På grund av examensarbetets omfattning har jag inte gjort någon multivariat analys. I det här arbetet har den univariata metoden använts men en faktor som behöver justeras i den univariata metoden är att antalet individer i åldrarna 16 – 64 år som ju analysen baseras på inte är konstant. Antalet individer i åldrarna 16 – 64 år är monotont stigande under tioårsperioden som analyseras. I januari 2005 är det ca 5 % fler i åldrarna 16 – 64 år än det var i januari 1994. En korrigering av antalet dagar bör göras då det inte är helt orimligt att anta att ju fler individer det är i populationen, desto fler utbetalade nettodagar som ju inte ger en jämförbar analys. Genom att göra en standardisering där antalet individer i populationen januari 1994 får index ett divideras antalet utbetalade nettodagar för övriga månader i analysen med indextalet för just den månaden. På så sätt kan antalet individer som variabel reduceras i analysen. Modellering görs alltså för befolkningen motsvarande den i januari 1994.

3.3 Grafer

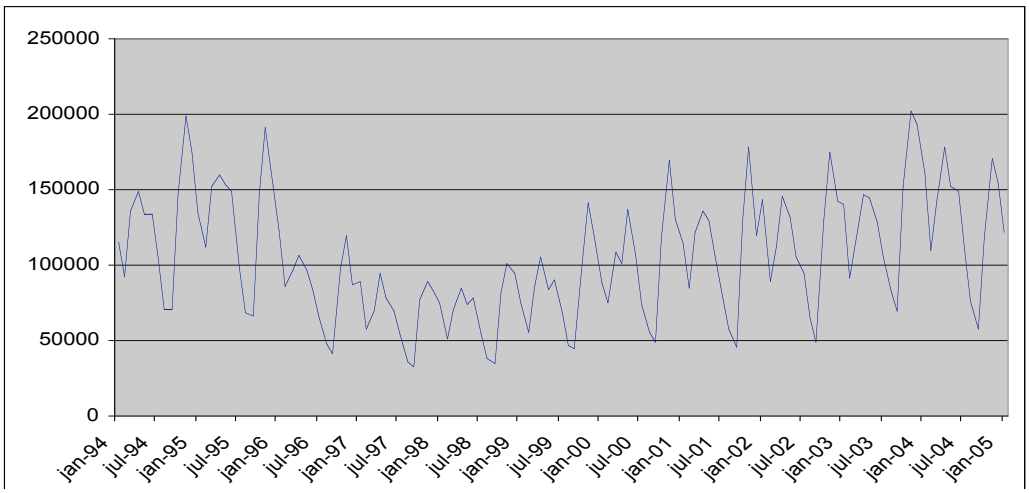
Efter att samtliga serier befolkningskorrigerats kan vi nu ta hjälp av de transformationer som beskrivits ovan för att ta fram stationära tidsserier. För sjukpenningen gäller att längdklasserna ett, två och tre påminner om varandra vad gäller utseende. Längdklasser fyra och fem påminner om varandra. Alla tre längdklasser inom rehabiliteringspenningen har ungefär samma egenskaper. Nedan visas befolkningskorrigerade tidsserier för längdklass ett och fyra gällande sjukpenningen samt serien för längdklass ett för rehabiliteringspenningen.



Figur 3.1 Antalet utbetalade nettodagar månadsvis, sjukpenning, längdklass 29 – 89 dagar. Befolkningskorrigerad.



Figur 3.2 Antalet utbetalade nettodagar månadsvis, sjukpenning, längdklass 1 - 2 år. Befolkningskorrigerad.



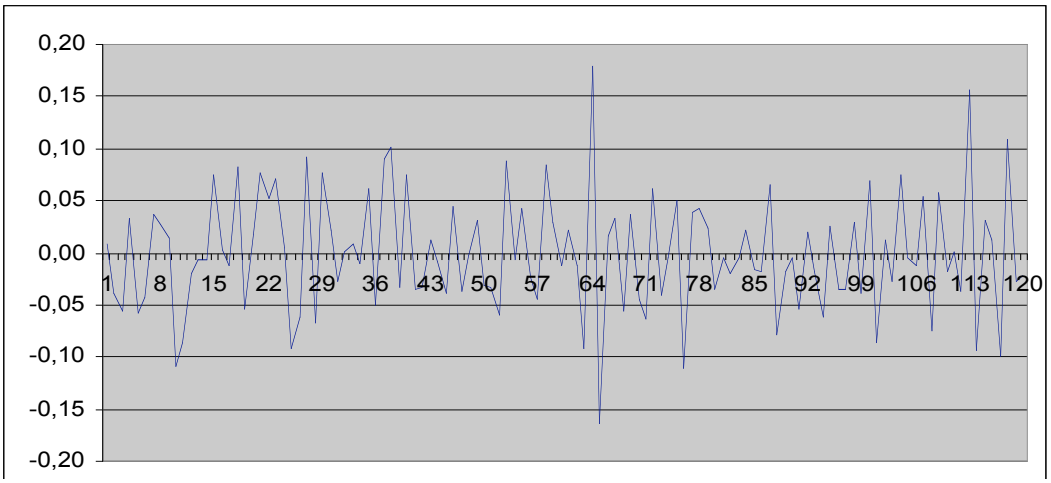
Figur 3.3 Antalet utbetalade nettodagar månadsvis, rehabiliteringspenning, längdklass 29 – 89 dagar. Befolkningskorrigerad.

3.4 Ändring av varians

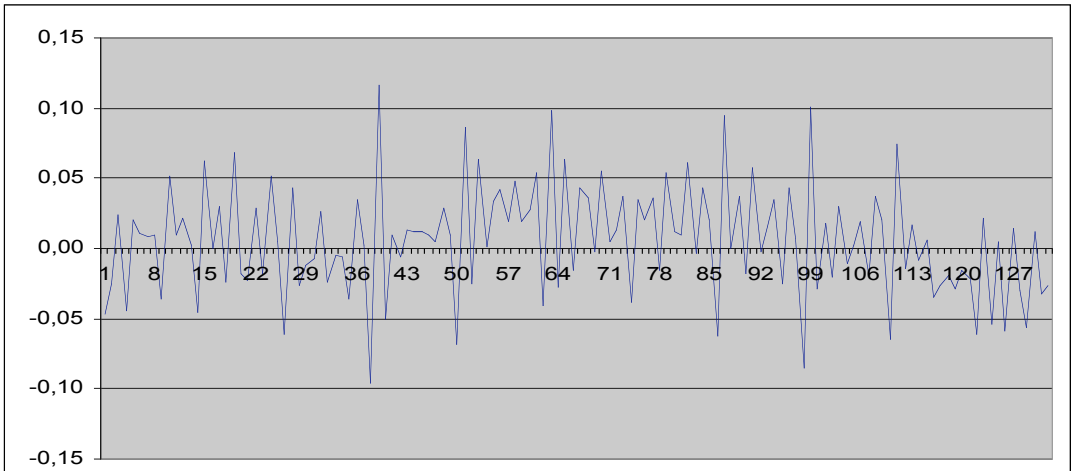
Låt oss först titta på variansens ändring över tid. För de fem tidsserier inom sjukpenningen som har analyserats har naturliga logaritmen av ursprungsdata tagits då variansen tenderar att öka över tid åtminstone från millennieskiftet och framåt för att sedan plana ut lite vid årsskiftet 2003-2004. Detta gäller främst för de tre första längdklasserna (tidsserierna). För de två sista längdklasserna gäller att variansen ökar från 2001 till mitten av 2003. För de tre tidsserierna inom rehabiliteringspenningen är variansen störst i början och slutet av tidsperioden så även för dessa har den naturliga logaritmen av ursprungsdata tagits för att komma steget närmare en stationär tidsserie.

3.5 Differenciering

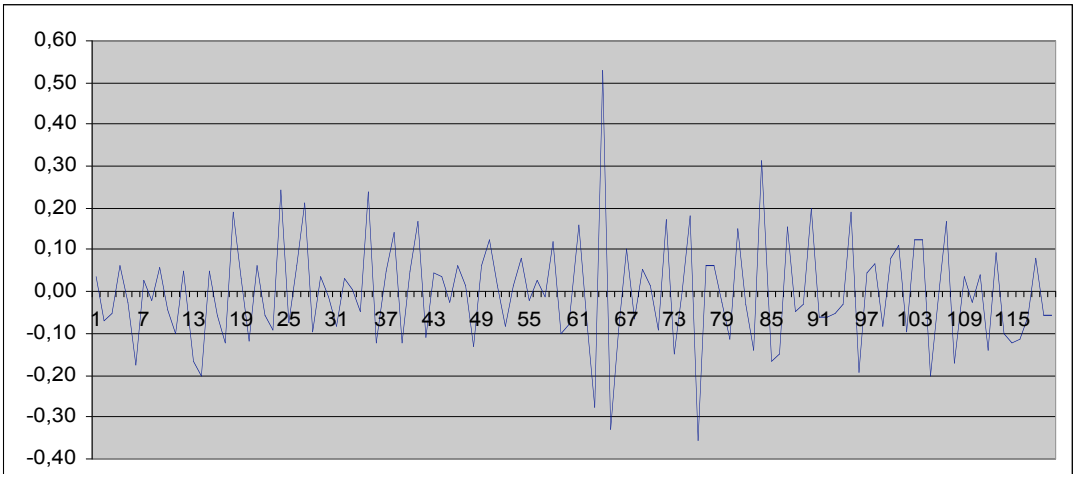
De tre första tidsserierna inom sjukpenningen samt de tre tidsserierna för rehabiliteringspenningen visar en klar periodicitet på 12 månader med relativt lika utseende inom varje period för respektive tidsserie. De två sista serierna inom sjukpenningen har inte samma klara periodicitet. Efter att ha transformerat serierna för att få en jämnare varians differencierar vi respektive serie som har periodicitet 12 med operatorn ∇_{12}. Vi får för alla serier utom de två sista för sjukpenningen de nya serierna $\nabla_{12}X_t = X_t - X_{t-12}$, $t = 13,14,\ldots$. De två övriga serierna blir $\nabla X_t = X_t - X_{t-1,}$ $t = 2,3,\ldots$. Det kan visas att serien $\nabla\nabla_d X_t$ är en stationär tidsserie om den ursprungliga serien har period d. Vi differencierar alltså de sex serierna med ursprunglig period 12 en gång till med operatorn ∇ och får serier som är ”tillräckligt stationära”. Genom att titta på ACF samt PACF för respektive stationär serie kan man se om de behöver differencieras ytterligare någon gång. Ifall de båda korrelationsfunktionerna inte avtar snabbt för förskjutningar bakåt i tiden bör serien differencieras ytterligare. För samtliga bearbetade serier i den här analysen är korrelationsfunktionerna snabbt avtagande varmed serierna betraktas som tillräckligt stationära. För de tre serier som presenterats i figurer ovan visas nedan resultatet. Samtliga serier är nu klara att analyseras för att ta fram en ARMA-modell.



Figur 3.4 Den transformerade och differencierade serien i Figur 1.1, sjukpenning, längdklass 29 – 89 dagar.



Figur 3.5 Den transformerade och differencierade serien i Figur 1.2, sjukpenning, längdklass 1-2 år.



Figur 3.6 Den transformerade och differencierade serien i Figur 1.3, rehabiliteringspenning, längdklass 29 - 89 dagar.

Att serien är ” tillräckligt stationär” kan sammanfattas med att det inte är någon systematisk ändring i väntevärde för tidsserien, att variansen inte ändras nämnvärt över tid och att serien är säsongsrensad.

3.6 Test av modell

Nu har vi åtta stycken stationära serier som vi via AIC-kriteriet, beskrivet i teoriavsnittet, ska hitta åtta ARMA-modeller för. Den modell som klarar modelltesten och har det mest optimala AIC-värdet får representera motsvarande tidsserie.

Det första testet vid framtagande av en adekvat modell är ifall den stationära serien transformerad från ursprungsdata är vitt brus. Testet bygger på antagandet att det för stora n gäller att autokorrelationerna av en sekvens oberoende likafördelade stokastiska variabler med ändlig varians är approximativt i.i.d N(0, 1/n). Testet bygger på Ljung-Boxstatistikan enligt följande:

$$Q_{LB} = n(n+2)\sum_{j=1}^{n} \hat{\rho}^2(j)/(n-j)$$

Med nollhypotesen att den stationära serien är i.i.d förkastas hypotesen om $Q_{LB} > \chi^2_{1-\alpha}(h)$.

Om hypotesen inte förkastas är analysen klar då det inte finns någon modell att ta fram, dvs serien består av okorrelerade stokastiska variabler. Om hypotesen förkastas är nästa test huruvida parametrarna i den framtagna modellen är signifikanta. Testet bygger på t-testet med den skattade standardavvikelsen för respektive parameter. Nollhypotesen är här att parametern inte tillför något till modellen. Nästa steg är att analysera korrelationsmatrisen mellan de ingående parametrarna. Höga korrelationsvärden mellan två parametrar tyder på att åtminstone en av dem bör tas bort från modellen. Slutligen testar vi om residualerna beskrivna i kapitel 2 har egenskaper som vitt brus med väntevärde noll och varians ett. Här är nollhypotesen att residualerna är vitt brus. Om modellen är giltig bör p-värdena vara relativt höga, åtminstone högre än 0,05.

4. Resultat

För att få fram samtliga resultat i det här kapitlet har jag använt PROC ARIMA-proceduren i SAS. En rad olika modeller med olika antal parametrar har testats och av de modeller som klarat modelltesterna har den med lägst AIC-värde fått representera respektive tidsserie. Nedan redovisas de framtagna ARMA-modellerna för de åtta stationära tidsserierna i kapitel 3. En första utgångspunkt har varit att analysera korrelationsfunktionerna ACF samt PACF upp till tidsförskjutningar som är ca en fjärdedel av totala antalet observationer, dvs i det här fallet upp till förskjutning 30 för respektive tidsserie.

Modellen som representerar den stationära tidsserien för längdklass 1, sjukpenning:

$$(X_t + 0{,}3305X_{t-1} + 0{,}2462X_{t-4} - 0{,}2645X_{t-6} - 0{,}2294X_{t-9})(X_t + 0{,}3891X_{t-12}) = Z_t$$

Detta är ju en ren AR-modell men den ser lite annorlunda ut jämfört med definitionen av en AR-modell i kapitel 2. Modellen ovan är kallad en multiplikativ modell då den är en produkt av enklare modeller, i det här fallet av två AR-processer. Multiplikativa modeller används ofta då det är relativt starka säsongsbetonade mönster i responsserien. Tidsförskjutningen i den andra faktorn ovan har mycket riktigt period 12.

Nedan följer resultaten för modellen:

AIC-värde: -396,21

Test av autokorrelationer för vitt brus

till förskj.	Chi2	DF	p-värde
6	35,38	6	<0,0001
12	67,61	12	<0,0001
18	70,73	18	<0,0001
24	87,93	24	<0,0001
30	119,45	30	<0,0001

Skattning av parametrar

parametrar	skattning	SD	t-värde	p-värde
AR1,1	-0,3305	0,0822	-4,02	0,0001
AR1,2	-0,2462	0,0841	-2,93	0,0041
AR1,3	0,2645	0,0840	3,15	0,0021
AR1,4	0,2294	0,0867	2,65	0,0093
AR2,1	-0,3891	0,0921	-4,23	<0,0001

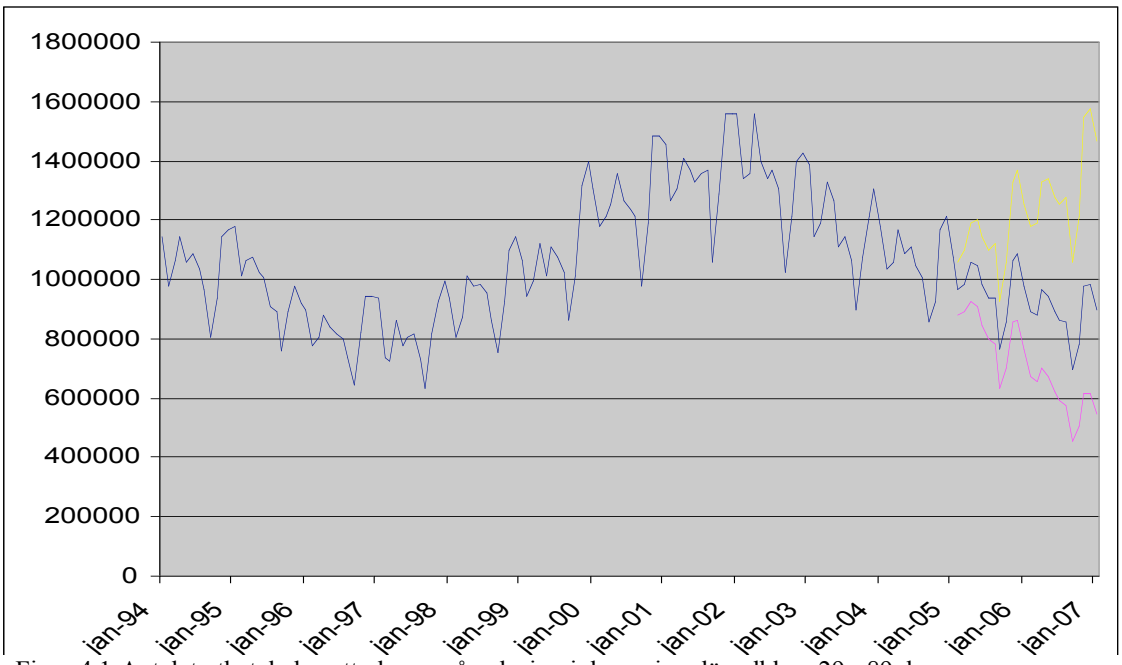
Korrelation mellan parametrar

parametrar	AR1,1	AR1,2	AR1,3	AR1,4	AR2,1
AR1,1	1,000	-0,294	-0,074	-0,088	-0,073
AR1,2	-0,294	1,000	0,085	-0,043	-0,059
AR1,3	-0,074	0,085	1,000	-0,268	-0,089
AR1,4	-0,088	-0,043	-0,268	1,000	0,023
AR2,1	-0,073	-0,059	-0,089	0,023	1,000

Test av residualer

till förskj.	Chi2	DF	p-värde
6	3,39	1	0,0656
12	6,64	7	0,4678
18	11,32	13	0,5839
24	19,20	19	0,4442
30	29,69	25	0,2362

Samtliga p-värden klarar 5%-nivån. Nedan visas grafen över den ursprungliga tidsserien med predikterade värden 24 månader framåt inklusive 95% konfidensintervall för de predikterade värdena.



Figur 4.1 Antalet utbetalade nettodagar månadsvis, sjukpenning, längdklass 29 - 89 dagar samt 24 predikterade månader med 95%-igt konfidensintervall.

Modellen som representerar den stationära tidsserien för längdklass 2, sjukpenning:

$$(X_t - 0{,}1740X_{t-6} - 0{,}2247X_{t-9} + 0{,}2531X_{t-10}) = (Z_t + 0{,}2415Z_{t-2})(Z_t - 0{,}5660Z_{t-12})$$

Här har vi en ARMA-modell, d.v.s både p och q är större än noll. Även denna modell är multiplikativ (med period 12 i den andra faktorn) men här består den multiplikativa modellen av två MA-modeller. Resultaten för modellen:

AIC-värde: -525,85

Test av autokorrelationer för vitt brus

till förskj.	Chi2	DF	p-värde
6	13,64	6	0,0339
12	36,84	12	0,0002
18	49,98	18	<0,0001
24	64,09	24	<0,0001
30	77,51	30	<0,0001

Skattning av parametrar

parametrar	skattning	SD	t-värde	p-värde
MA1,1	-0,2415	0,0933	-2,59	0,0109
MA2,1	0,5660	0,0831	6,81	<0,0001
AR1,1	0,1740	0,0897	1,94	0,0548
AR1,2	0,2247	0,0927	2,42	0,0170
AR1,3	-0,2531	0,0943	-2,69	0,0083

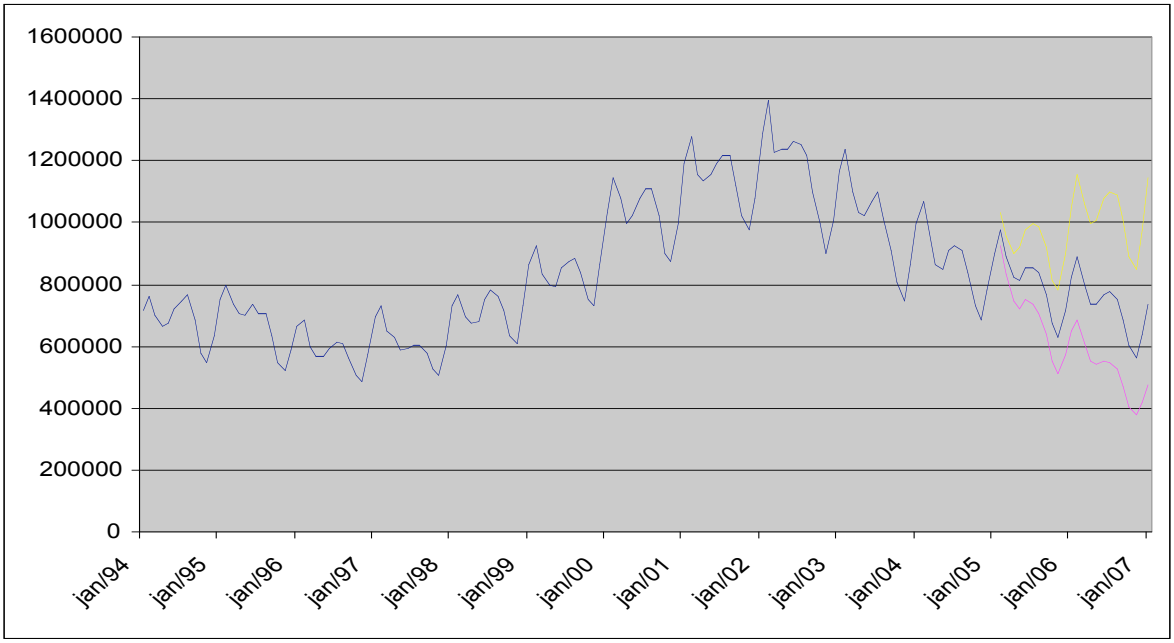
Korrelation mellan parametrar

parametrar	MA1,1	MA2,1	AR1,1	AR1,2	AR1,3
MA1,1	1,000	-0,042	-0,015	-0,174	0,151
MA2,1	-0,042	1,000	0,056	0,103	-0,007
AR1,1	-0,015	0,056	1,000	-0,166	0,151
AR1,2	-0,174	0,103	-0,166	1,000	-0,156
AR1,3	0,151	-0,007	0,151	-0,156	1,000

Test av residualer

till förskj.	Chi2	DF	p-värde
6	6,75	1	0,0094
12	9,54	7	0,2164
18	15,06	13	0,3037
24	21,75	19	0,2966
30	26,07	25	0,4039

P-värdet fram till tidsförskjutning 6 i testet av residualer är något lågt, annars är värdena enligt 5%-nivån.



Figur 4.2. Antalet utbetalade nettodagar månadsvis, sjukpenning, längdklass 90 - 179 dagar samt 24 predikterade månader med 95%-igt konfidensintervall.

Enligt prognosen minskar antalet av Försäkringskassan utbetalade nettodagar där nedgången hållit i sig sedan år 2002.

Modellen som representerar den stationära tidsserien för längdklass 3, sjukpenning:

$X_t = (Z_t + 0{,}2629Z_{t-3} + 0{,}2718Z_{t-5} + 0{,}2804Z_{t-6})(Zt - 0{,}2225Z_{t-10})$

Här visar det sig att den optimala modellen är multiplikativ uppbyggd av två MA-modeller där den senare faktorn har period 10. Grafen visar dock på en klar periodicitet på 12 månader.

Resultat:

AIC-värde: -537,84

Test av autokorrelationer för vitt brus

till förskj.	Chi2	DF	p-värde
6	27,05	6	0,0001
12	43,01	12	<0,0001
18	47,91	18	0,0002
24	58,97	24	<0,0001
30	69,77	30	<0,0001

Skattning av parametrar

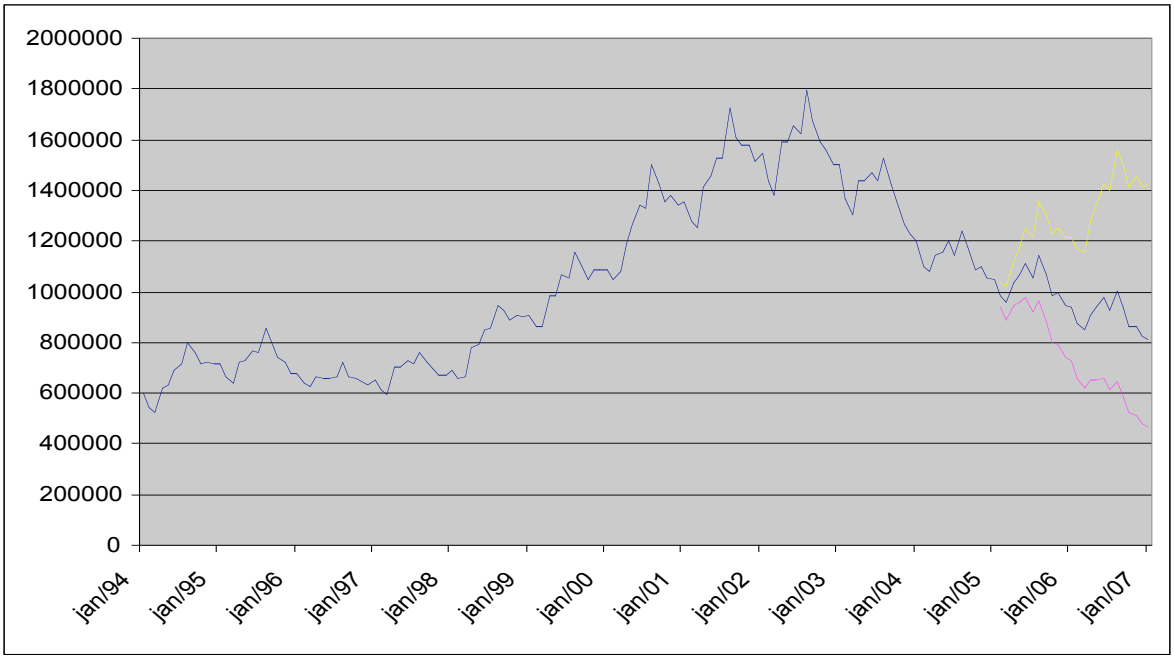
parametrar	skattning	SD	t-värde	p-värde
MA1,1	-0,2629	0,0905	-2,90	0,0044
MA1,2	-0,2718	0,0903	-3,01	0,0032
MA1,3	-0,2804	0,0887	-3,16	0,0020
MA2,1	0,2225	0,0987	2,25	0,0261

Korrelation mellan parametrar

parametrar	MA1,1	MA1,2	MA1,3	MA2,1
MA1,1	1,000	-0,203	0,245	0,019
MA1,2	-0,203	1,000	-0,049	0,227
MA1,3	0,245	-0,049	1,000	0,004
MA2,1	0,019	0,227	0,004	1,000

Test av residualer

till förskj.	Chi2	DF	p-värde
6	4,15	2	0,1255
12	12,73	8	0,1214
18	15,45	14	0,3478
24	21,82	20	0,3502
30	27,37	26	0,3903



Figur 4.3. Antalet utbetalade nettodagar månadsvis, sjukpenning, längdklass 180 - 365 dagar samt 24 predikterade månader med 95%-igt konfidensintervall.

Prognosen visar på att de utbetalade dagarna minskar för hela prognosperioden på ungefärligen samma sätt som för längdklassen innan.

Modellen som representerar den stationära tidsserien för längdklass 4, sjukpenning:

$(X_t - 0{,}2630X_{t-2} - 0{,}3203X_{t-3})(X_t - 0{,}7708X_{t-12}) = (Z_t - 0{,}3336Z_{t-1} + 0{,}3830Z_{t-5})$

Multiplikativ ARMA-modell där produkten består av två AR-modeller där den senare har period 12. Den stationära serien i Figur 3.5 som modellen baseras på skulle möjligtvis kunna differencieras en gång till för att få en serie med väntevärde mer nära noll. Dock avtar seriens ACF och PACF snabbt. Inom litteraturen påvisas att man ej bör ”överdifferenciera” en serie. Detta sammantaget gör att det är serien i Figur 3.5 som används vid modellframtagande.
Nedan följer resultaten:

AIC-värde: -620,63

Test av autokorrelationer för vitt brus

till förskj.	Chi2	DF	p-värde
6	49,57	6	<0,0001
12	156,42	12	<0,0001
18	204,00	18	<0,0001
24	280,18	24	<0,0001
30	326,63	30	<0,0001

Skattning av parametrar

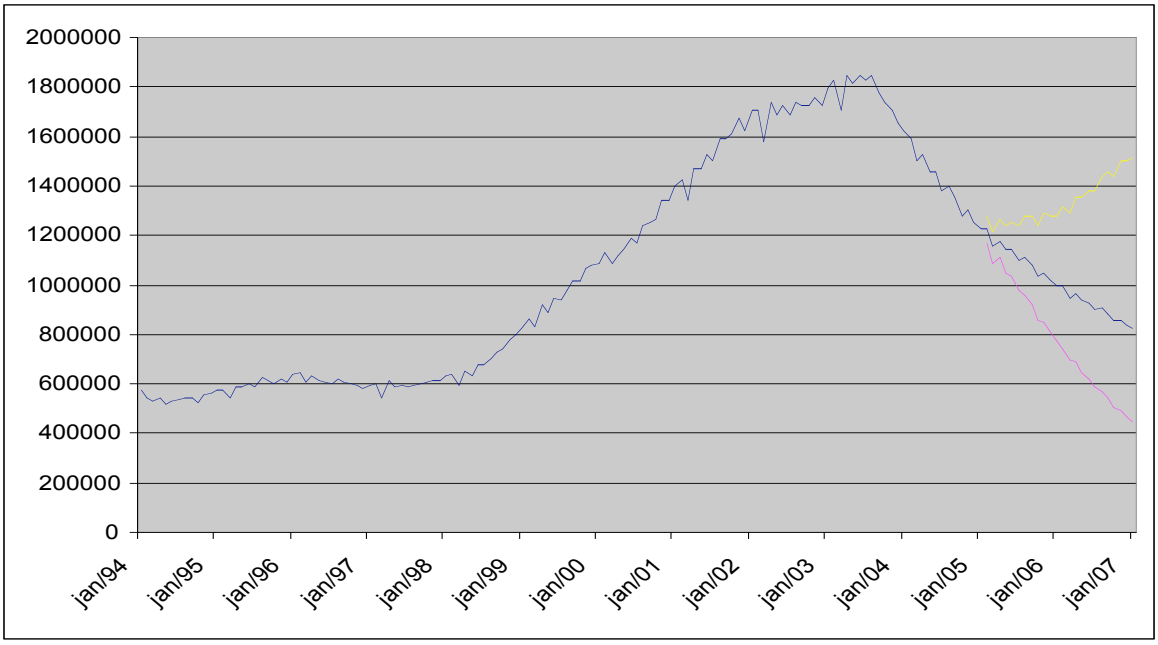
parametrar	skattning	SD	t-värde	p-värde
MA1,1	0,3336	0,0796	4,19	<0,0001
MA1,2	-0,3830	0,0833	-4,60	<0,0001
AR1,1	0,2630	0,0870	3,02	0,0030
AR1,2	0,3203	0,0841	3,81	0,0002
AR2,1	0,7708	0,0623	12,37	<0,0001

Korrelation mellan parametrar

parametrar	MA1,1	MA1,2	AR1,1	AR1,2	AR2,1
MA1,1	1,000	0,223	0,293	-0,002	-0,092
MA1,2	0,223	1,000	0,249	0,233	-0,207
AR1,1	0,293	0,249	1,000	-0,132	-0,143
AR1,2	-0,002	0,233	-0,132	1,000	-0,075
AR2,1	-0,092	-0,207	-0,143	-0,075	1,000

Test av residualer

till förskj.	Chi2	DF	p-värde
6	1,19	1	0,2745
12	10,06	7	0,1853
18	14,37	13	0,3484
24	24,69	19	0,1709
30	30,97	25	0,1900



Figur 4.4. Antalet utbetalade nettodagar månadsvis, sjukpenning, längdklass 366 - 730 dagar samt 24 predikterade månader med 95%-igt konfidensintervall.

Även här minskar antalet dagar från år 2003 för att enligt prognosen fortsätta minska under hela den predikterade perioden.

Modellen som representerar den stationära tidsserien för längdklass 5, sjukpenning:

$(X_t - 0{,}4568X_{t-3} - 0{,}3622X_{t-5})(X_t - 0{,}8417X_{t-12}) = (Z_t - 0{,}2375Z_{t-7})$

Multiplikativ ARMA-modell där produkten består av två AR-modeller, den senare med period 12.

Resultaten för modellen:

AIC-värde: -588,64

Test av autokorrelationer för vitt brus

till förskj.	Chi2	DF	p-värde
6	35,44	6	<0,0001
12	118,60	12	<0,0001
18	164,39	18	<0,0001
24	224,71	24	<0,0001
30	258,22	30	<0,0001

Skattning av parametrar

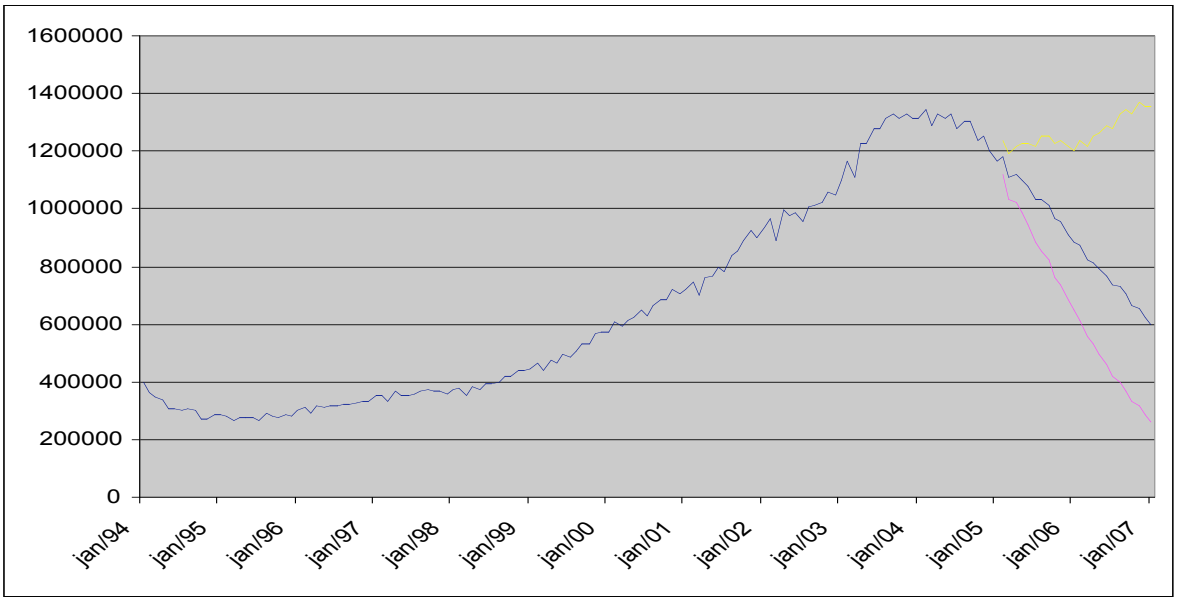
parametrar	skattning	SD	t-värde	p-värde
MA1,1	0,2375	0,0879	2,70	0,0078
AR1,1	0,4568	0,0712	6,42	<0,0001
AR1,2	0,3622	0,0714	5,08	<0,0001
AR2,1	0,8417	0,0519	16,21	<0,0001

Korrelation mellan parametrar

parametrar	MA1,1	AR1,1	AR1,2	AR2,1
MA1,1	1,000	0,078	0,026	0,100
AR1,1	0,078	1,000	-0,322	-0,078
AR1,2	0,026	-0,322	1,000	-0,109
AR2,1	0,100	-0,078	-0,109	1,000

Test av residualer

till förskj.	Chi2	DF	p-värde
6	2,52	2	0,2835
12	7,62	8	0,4712
18	15,88	14	0,3205
24	19,42	20	0,4948
30	23,84	26	0,5852



Figur 4.5. Antalet utbetalade nettodagar månadsvis, sjukpenning, längdklass > 730 dagar samt 24 predikterade månader med 95%-igt konfidensintervall.

För den här längdklassen dalar kurvan senare än för de tidigare klasserna. Prognosperioden börjar några observationer efter toppen på kurvan och prognosen visar på en relativt snabb minskning av utbetalade nettodagar för den två år långa prognosperioden.

Modellen som representerar den stationära tidsserien för längdklass 1, rehabiliteringspenning:

$(X_t + 0{,}3785X_{t-1} - 0{,}3432X_{t-3} - 0{,}2255X_{t-20} + 0{,}2067X_{t-22}) = (Z_t - 0{,}5121Z_{t-12})$

AIC-värde: -232,68

Test av autokorrelationer för vitt brus

till förskj.	Chi2	DF	p-värde
6	81,26	6	<0,0001
12	98,88	12	<0,0001
18	117,63	18	<0,0001
24	189,61	24	<0,0001
30	228,69	30	<0,0001

Skattning av parametrar

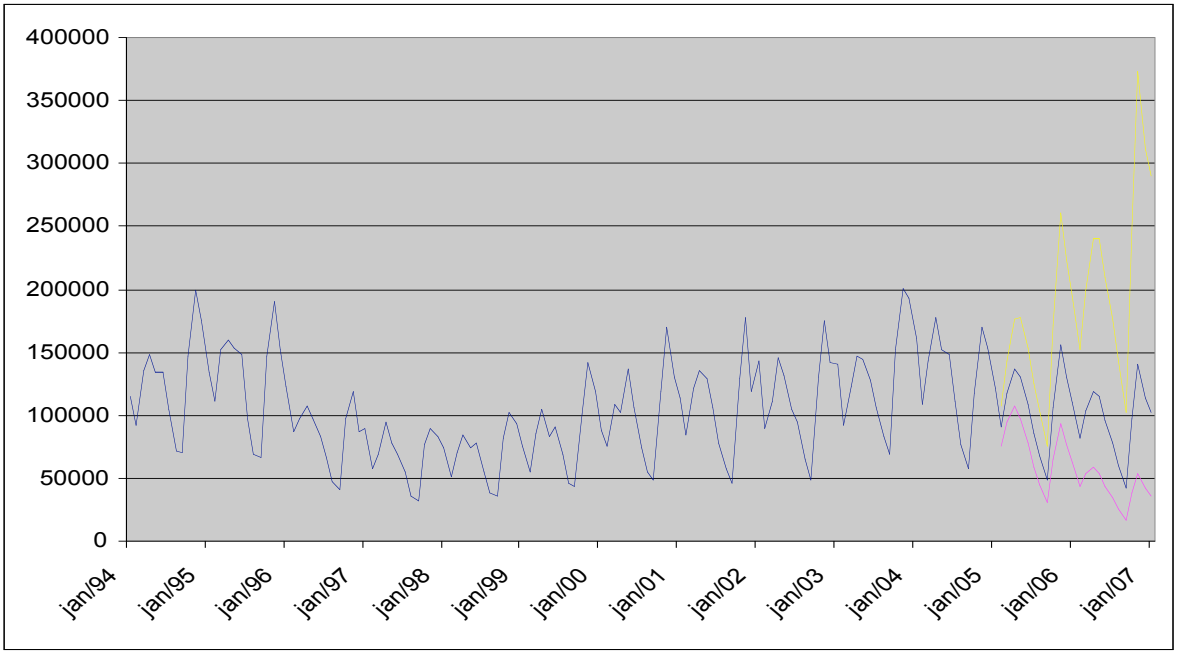
parametrar	skattning	SD	t-värde	p-värde
MA1,1	0,5121	0,0912	5,61	<0,0001
AR1,1	-0,3785	0,0772	-4,91	<0,0001
AR1,2	0,3432	0,0752	4,56	<0,0001
AR1,3	0,2255	0,0805	2,80	0,0060
AR1,4	-0,2067	0,0815	-2,53	0,0126

Korrelation mellan parametrar

parametrar	MA1,1	AR1,1	AR1,2	AR1,3	AR1,4
MA1,1	1,000	0,175	-0,044	0,051	0,059
AR1,1	0,175	1,000	0,185	0,238	0,374
AR1,2	-0,044	0,185	1,000	-0,246	0,226
AR1,3	0,051	0,238	-0,246	1,000	0,208
AR1,4	0,059	0,374	0,226	0,208	1,000

Test av residualer

till förskj.	Chi2	DF	p-värde
6	6,59	1	0,0103
12	12,42	7	0,0876
18	16,26	13	0,2352
24	20,21	19	0,3820
30	23,57	25	0,5443



Figur 4.6. Antalet utbetalade nettodagar månadsvis, rehabiliteringspenning, längdklass 29 - 89 dagar samt 24 predikterade månader med 95%-igt konfidensintervall.

Detta är den första längdklassen som analyseras gällande rehabiliteringspenningen. Ursprungliga serien ovan visar på en ganska stor ändring av variansen över tid och precis innan prognosperioden skönjas en viss minskning av antalet utbetalade nettodagar. Minskningen håller i sig hela den predikterade perioden.

ARMA-modellen som representerar den stationära tidsserien för längdklass 2, rehabiliteringspenning:

$(X_t + 0{,}3069X_{t-1} - 0{,}3700X_{t-3} + 0{,}2992X_{t-16}) = (Z_t - 0{,}6310Z_{t-12})$

Resultat:

AIC-värde: -212,06

Test av autokorrelationer för vitt brus

till förskj.	Chi2	DF	p-värde
6	68,66	6	<0,0001
12	92,09	12	<0,0001
18	108,98	18	<0,0001
24	171,43	24	<0,0001
30	215,50	30	<0,0001

Skattning av parametrar

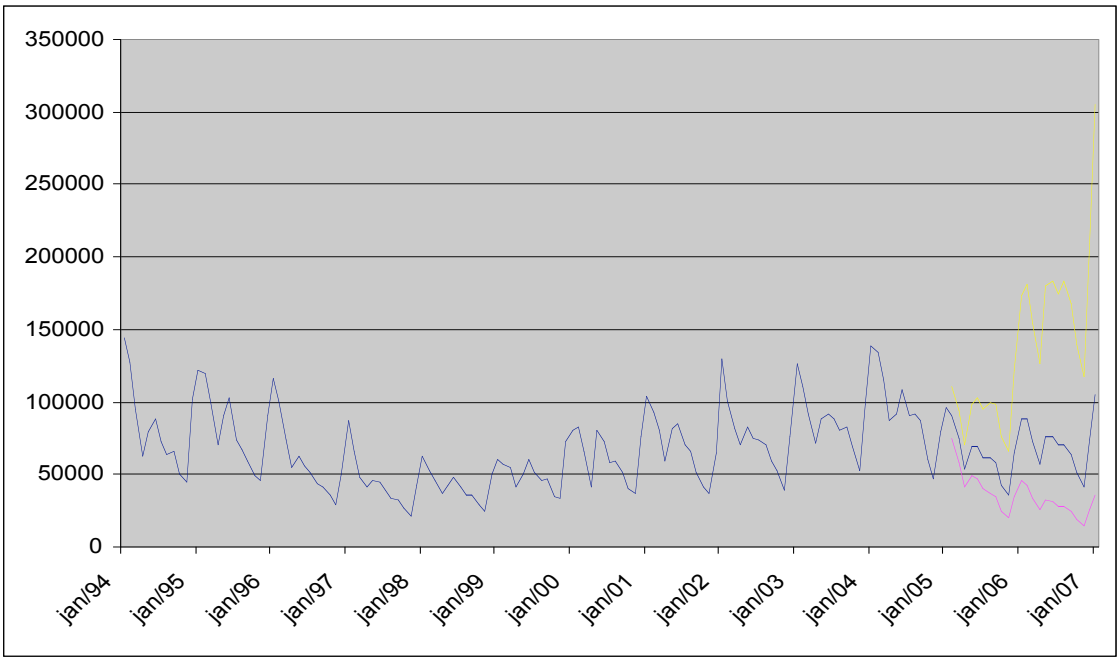
parametrar	skattning	SD	t-värde	p-värde
MA1,1	0,6310	0,0847	7,45	<0,0001
AR1,1	-0,3069	0,0745	-4,12	<0,0001
AR1,2	0,3700	0,0771	4,80	<0,0001
AR1,3	-0,2992	0,0801	-3,74	0,0003

Korrelation mellan parametrar

parametrar	MA1,1	AR1,1	AR1,2	AR1,3
MA1,1	1,000	0,091	-0,016	-0,083
AR1,1	0,091	1,000	0,141	-0,025
AR1,2	-0,016	0,141	1,000	0,259
AR1,3	-0,083	-0,025	0,259	1,000

Test av residualer

till förskj.	Chi2	DF	p-värde
6	5,53	2	0,0631
12	8,69	8	0,3693
18	18,29	14	0,1939
24	24,97	20	0,2025
30	30,24	26	0,2576



Figur 4.7. Antalet utbetalade nettodagar månadsvis, rehabiliteringspenning, längdklass 90 - 179 dagar samt 24 predikterade månader med 95%-igt konfidensintervall.

Här ändras variansen precis innan prognosperioden jämfört med observationerna dessförinnan. Detta ger en viss effekt på hur variationen i prognosperioden blir. Grafen är någorlunda konstant under prognosperioden.

ARMA-modellen som representerar den stationära tidsserien för längdklass 3, rehabpenning:

$(X_t + 0{,}4178X_{t-1} - 0{,}3847X_{t-3} + 0{,}2088X_{t-22}) = (Z_t - 0{,}3145Z_{t-12})$

Resultat:

AIC-värde: -182,62

Test av autokorrelationer för vitt brus

till förskj.	Chi2	DF	p-värde
6	73,96	6	<0,0001
12	101,34	12	<0,0001
18	118,06	18	<0,0001
24	171,42	24	<0,0001
30	199,85	30	<0,0001

Skattning av parametrar

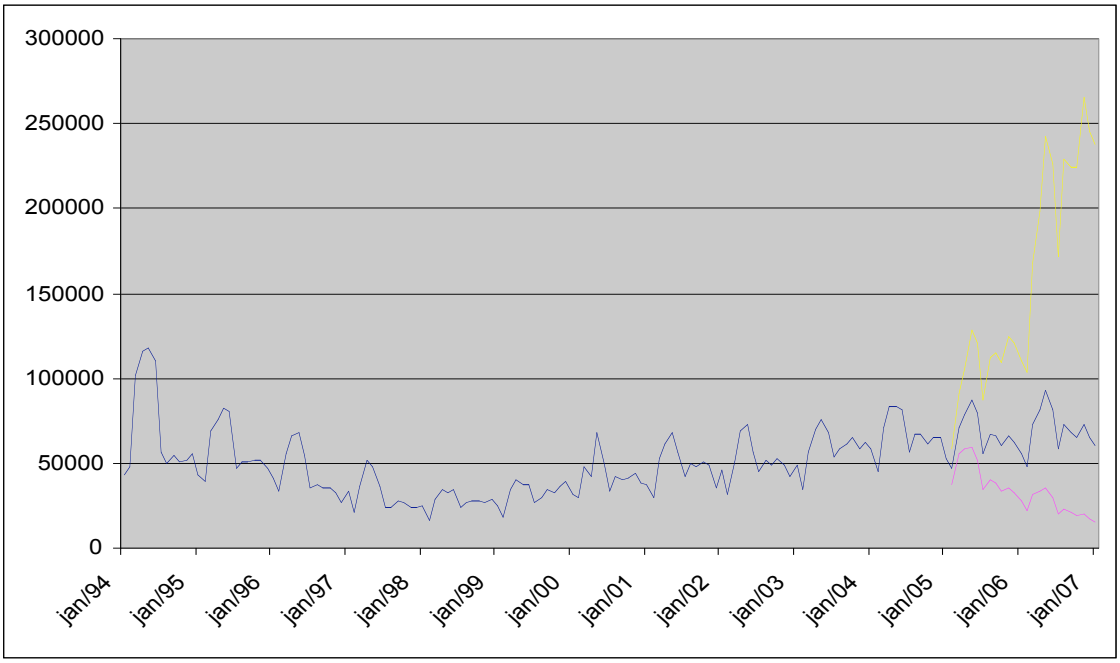
parametrar	skattning	SD	t-värde	p-värde
MA1,1	0,3145	0,0942	3,34	0,0011
AR1,1	-0,4178	0,0766	-5,45	<0,0001
AR1,2	0,3847	0,0769	5,00	<0,0001
AR1,3	-0,2088	0,0815	-2,56	0,0117

Korrelation mellan parametrar

parametrar	MA1,1	AR1,1	AR1,2	AR1,3
MA1,1	1,000	0,173	-0,007	-0,013
AR1,1	0,173	1,000	0,189	0,225
AR1,2	-0,007	0,189	1,000	0,282
AR1,3	-0,013	0,225	0,282	1,000

Test av residualer

till förskj.	Chi2	DF	p-värde
6	5,13	2	0,0769
12	7,53	8	0,4802
18	9,47	14	0,7995
24	19,58	20	0,4843
30	23,38	26	0,6112



Figur 4.8. Antalet utbetalade nettodagar månadsvis, rehabiliteringspenning, längdklass 180 - 365 dagar samt 24 predikterade månader med 95%-igt konfidensintervall.

Den här prognosen är den enda som indikerar på en ökning av antalet utbetalda nettodagar. Tidsserien har ett visst mönster med stigande värden ca 60 observationer innan prognosperioden.

5. Slutsats

Mitt syfte med det här examensarbetet har varit att ta fram modeller och göra prognoser för tidsserier gällande sjukförmåner. Datamaterialet som jag erhållit från Försäkringskassans huvudkontor består av månadsvisa observationer med antalet av Försäkringskassan utbetalade nettodagar uppdelade i längdklasser för förmånerna sjukpenning och rehabiliteringspenning över hela landet. Att prediktera framtida värden för en tidsserie är ofta en komplex uppgift då flera faktorer kan påverka responsserien. T.ex. kan ändringar i ersättningsnivåer påverka hur många dagar som det betalas ut ersättning för. I det här arbetet har jag valt att göra en univariat analys p.g.a arbetets omfattning. I motsats till multivariata metoder tittar man i den univariata på korrelationer mellan värden enbart inom responsserien som analyseras. Via transformationer av ursprungsserierna till stationära serier har ARMA-modeller tagits fram och utifrån dessa kan prediktioner av ursprungsserierna göras där minsta kvadratmetoden används.

Samtliga serier gällande sjukpenningen har relativt lika utseende med en markant ökning av utbetalade dagar i slutet av nittiotalet samt en minskning några år därefter. Alla dessa fem serier visar enligt prognosen på en fortsatt minskning av antalet utbetalade nettodagar.

För rehabiliteringspenningen gäller att de två första tidsserierna liknar de för sjukpenningen vad gäller ökning och minskning av antalet dagar. Däremot är variationen inte lika omfattande. Prognosen för den första serien indikerar på en minskning medan prognosen för den andra visar på relativt konstanta värden. Egenskaperna för den tredje serien skiljer sig från de övriga. Här har antalet utbetalade dagar ökat sedan några år och prognosen visar på en fortsatt ökning.

Appendix

Anta att $\{X_t\}$ är en tidsserie med väntevärde noll och $E|X_t|^2 < \infty$ för varje t och $E(X_i X_j) = \kappa(i,j)$. Vi inför följande notering för enstegsprediktorn och dess medelkvadratfel:

$$\hat{X}_n = \begin{cases} 0, & om \quad n = 1, \\ P_{n-1}X_{n,} & om \quad n = 2, 3, ..., \end{cases}$$

samt

$$v_n = E(X_{n+1} - P_n X_{n+1})^2$$

där $P_n X_{n+h} = a_0 + a_1 X_n + ... + a_n X_1$ där $a_0, ..., a_n$ är konstanter som fås då $E(X_{n+h} - a_0 - a_1 X_n - ... - a_n X_1)^2$ minimeras enligt minsta kvadratmetoden gällande prediktion av framtida värden i serien, samt *h* är ett positivt heltal.

Vi inför även enstegsprediktionsfelen

$$\boldsymbol{U_n} = \boldsymbol{X_n} - \hat{\boldsymbol{X}}_{\boldsymbol{n}} \tag{a}$$

Via $\boldsymbol{U_n} = (U_1, ..., U_n)'$ samt $\boldsymbol{X_n} = (X_1, ..., X_n)'$ kan den sista ekvationen skrivas som $\boldsymbol{U_n} = \boldsymbol{A_n X_n}$ där $\boldsymbol{A_n}$ har formen

$$\boldsymbol{A_n} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ a_{11} & 1 & 0 & \dots & 0 \\ a_{22} & a_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ a_{n-1,n-1} & a_{n-1,n-2} & a_{n-1,n-3} & \dots & 1 \end{bmatrix}$$

Detta ger att $\boldsymbol{A_n}$ är icke-singulär, med invers $\boldsymbol{C_n}$:

$$\boldsymbol{C_n} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \theta_{11} & 1 & 0 & \dots & 0 \\ \theta_{22} & \theta_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \dots & 1 \end{bmatrix}$$

Vektorn med enstegsprediktorvärdena $\hat{\boldsymbol{X}}_{\boldsymbol{n}} := (X_1, P_1X_2, ..., P_{n-1}X_n)'$ kan uttryckas som

$$\hat{\boldsymbol{X}}_{\boldsymbol{n}} = \boldsymbol{X_n} - \boldsymbol{U_n} = \boldsymbol{C_n} - \boldsymbol{U_n} = \Theta_{\boldsymbol{n}}(\boldsymbol{X_n} - \hat{\boldsymbol{X}}_{\boldsymbol{n}}), \tag{b}$$

där

$$\Theta_n = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \theta_{11} & 1 & 0 & \dots & 0 \\ \theta_{22} & \theta_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \dots & 1 \end{bmatrix}$$

och $\boldsymbol{X_n}$ satisfierar

$$\boldsymbol{X_n} = \boldsymbol{C_n}(\boldsymbol{X_n} - \hat{\boldsymbol{X}}_{\boldsymbol{n}}) \tag{c}$$

Ekvationen i (c) ovan kan skrivas om som

$$\hat{X}_{n+1} = \begin{cases} 0, & om \quad n = 0, \\ \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}) & om \quad n = 1,2,..., \end{cases} \tag{d}$$

från vilken enstegsprediktorvärdena $\hat{X}_1, \hat{X}_2, ...$ kan beräknas rekursivt efter att koefficienterna θ_{ij} har bestämts. Följande algoritm genererar dessa koefficienter och dess medelkvadratfel $v_i = E(X_{i+1} - \hat{X}_{i+1})^2$:

$$v_o = \kappa(1,1),$$

$$\theta_{n,n-k} = v_k^{-1}(\kappa(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j), \quad 0 \le k < n,$$

samt

$$v_n = \kappa(n+1, n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j.$$

För kausala processer kan algoritmen användas för en transformerad process av $\{X_t\}$:

$$\begin{cases} W_t = \sigma^{-1} X_t, & t = 1,\ldots,m, \\ W_t = \sigma^{-1}\phi(B) X_t, & t > m, \end{cases}$$

där m=max(p,q), se Definition 8, kapitel 2.

Med användande av algoritmen ovan för processen $\{W_t\}$ fås:

$$\begin{cases} \hat{W}_{n+1} = \sum_{j=1}^{n} \theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}), & 1 \le n < m, \\ \hat{W}_{n+1} = \sum_{j=1}^{q} \theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}), & n \ge m, \end{cases}$$

där koefficienterna θ_{nj} och medelkvadratfelen $r_n = E(W_{n+1} - \hat{W}_{n+1})^2$ fås rekursivt från algoritmen.

Referenser

C. Chatfield: The analysis of time series: An introduction

Robert H. Shumway och David S. Stoffer: Time series analysis and its applications

Jianqing Fan och Qiwei Yao: Nonlinear timeseries; nonparametric and parametric methods

William W. S. Wei: Time series analysis; univariate and multivariate methods

Peter J. Brockwell och Richard A. Davis: Introduction to timeseries and forecasting

www.forsakringskassan.se