Hemuppgift 4

STAH14 - Prediktering

HT 21

Kim Thurow

Inledning

Hemuppgift 4 är en del av kursen tidsserieanalys given på Lunds Universitet. I denna uppgift används två set av data från SMHI för att analysera en korttidsserie samt en långtidsserie. Uppgiften går ut på att ta fram en korttidsprognos, en prognoshoristont samt en långtidsmätning av temperatur i en viss ort inom landet. Jag har valt Abisko mätstation som ligger längst upp i Sverige och gränsar mot Finland.

Korttidsprognos

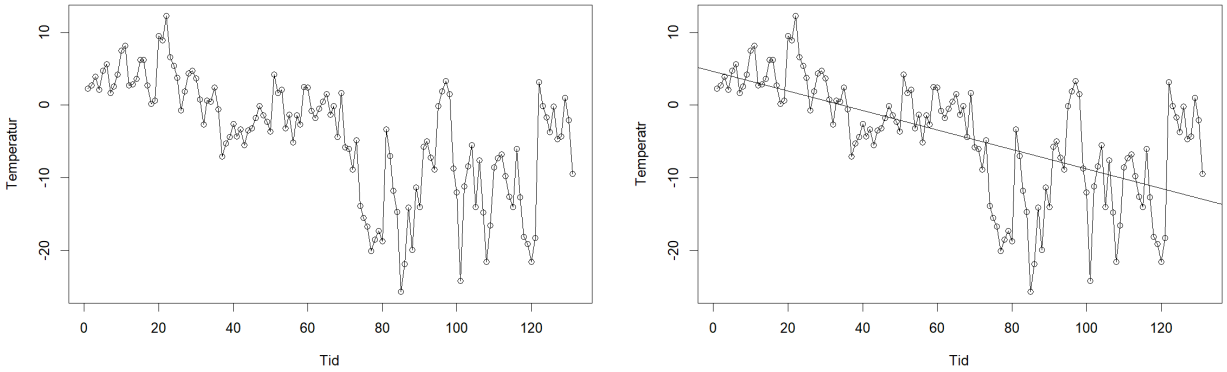

Figure 1: Temperatur över tid t.v samt med linjär modell t.h

En tidsserie över 4 månader används. Temperaturen mäts en gång per timme, jag väljer ut morgontemperaturen 06:00. Därefter sätter jag antal observationer, frekvens, till 1 per dag. I ploten i figur 1 ser vi tidsserien över temperaturen i Abisko. I första delen av ploten kan det röra sig om en random walk, där det i viss mån verkar vara så att positiva värden följs av positiva och negativa följs av negativa. Dock vet vi att emperaturen följer en deterministisk trend, där det från september blir kallare mot slutet av året. I andra hälften av tidsserien är det stora variationer i temperaturen, mellan -20 och 0 gradigt i ett slags cykliskt mönster. Jag provar en linjär modell och får ut att cirka 39% av variationen förklaras av den linjära modellen. Modellen visas i ploten av tidsserien i figur 1 till höger.
Interceptet, det vill säga μ ligger i denna modell på ca 4.6 grader.

Jag prövar även en säsongsmodell där $\mu_t = \mu_{t-5}$, 5 månader som period 1,2,3,4 och 5. Dock visar det sig att här förklaras variationen i tidsserien enbart 18% av modellen. En cyklisk modell prövas också, baserat på en veckoperiodicitet, dock blir förklaringskoefficienten 'Adjusted R'Squared' väldigt låg, $< 0.01\%$.

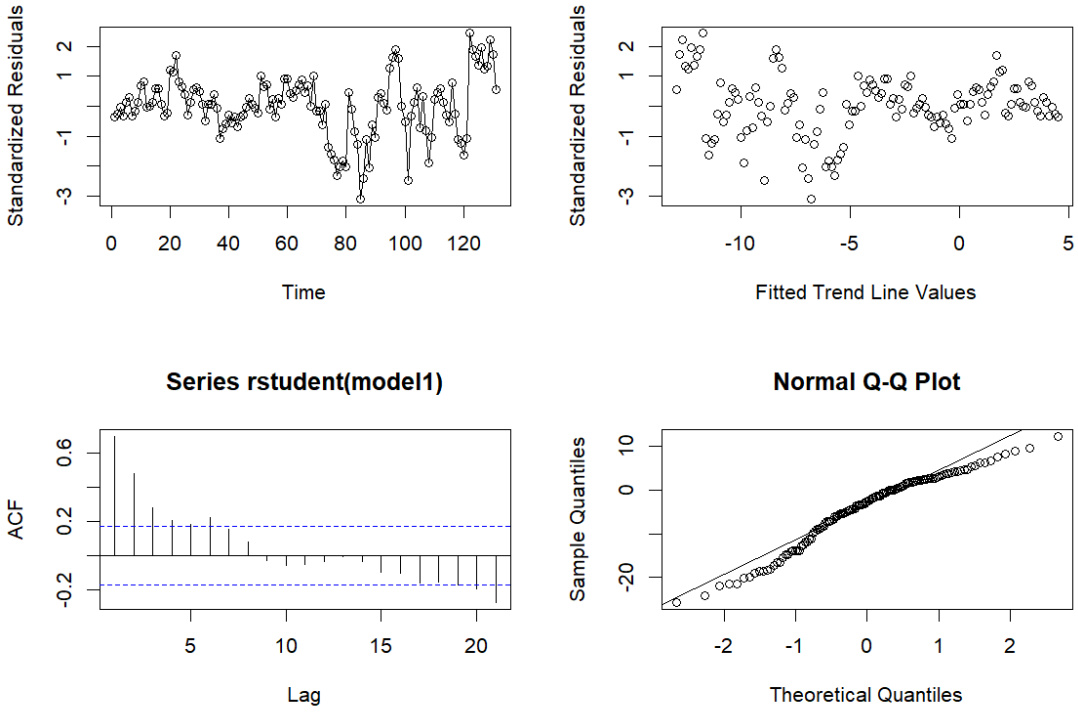

Figure 2: Verifikation av modell

I den första ploten till vänster i figur 2 är de standardiserade residualerna plottade över tid. Vi ser samma mönster som i tidsserien, vilket tyder på att trenden inte är enbart white noise. Till höger ses de standardiserade residualerna plottade mot den linjära modellen. För att modellen ska anses vara korrekt, ska residualerna vara helt slumpmässiga. Här ses ett tydligt mönster, speciellt i andra hälften av ploten. Random walk modellen förklarar inte tidsserien särskilt bra, vilket vi vet sedan innan.

I ACF:en över residualerna i den linjära modellen kan vi se tre signifikanta lags som är exponentiellt avtagande, vilket visar att i de första lagen finns det ett beroende av tid. Därefter ser lagen slumpmässiga ut med några positiva och några negativa lags i ett oscillerande mönster. Det tyder på att tidsserien inte är en random walk. I qqploten, ned till höger i figur 2, plottas själva tidsserien för att undersöka om residualerna är normalfördelade, vilket avvikelserna från den linjära linjen visar att dem inte är. De är alltså inte oberoende. Dock visar ett runstest på den linjära modellen på just oberoende med 38 observerade runs och 66 förväntade.

Ett adf test, där nollhypotesen är icke-stationäritet visar på 0.39 i p-värde och mothypotesen om stationäritet kan accepteras.
En EACF-plot på den linjära modellens residualer visar på MA(2)-MA(6) processer. Ett ARIMA test med MK-metoden och ordningen MA(2) skattar μ= -4.2 samt σ^2=26.9.

Prediktion

I figur 3 ses en prediktion över 14 dagar från en modell med veckoperiodicitet i en MA(2)-modell. Här ses den breda variationen som skattades tidigare i konfidensintervallet kring det förväntade medelvärdet kommande 14 dagar framåt i tiden. Det förväntade medelvärdet ligger runt -10. Att det förväntade medelvärdet ligger på -10 grader och inte är 0, motsäger teorin om stationäritet. För att en tidsserie ska vara stationär ska den inte vara beroende av tid.

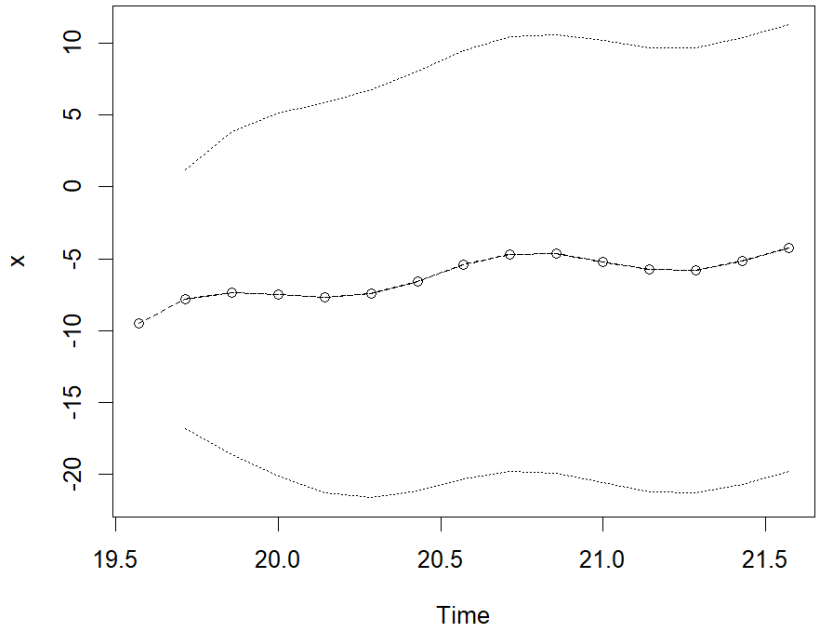

Figure 3: Prediktion 14 dgr

Korsvalidering

Pressvärdet, det vill säga skillnaden mellan skattade och verkliga residualer i kvadrat blev i en cyklisk modell värdet 765.6. Den linjära modellen gav pressvärdet 706,7.

Det är alltså den linjära modellen som har den lägsta skillnaden mellan skattade och verkliga värden.

Prognoshorisont

Tre plots för olika längd på tidsserien, från 131 observationer till 50. Vi kan se att de två lags som tas med i MA(2) processen, är mycket flackare med fler observationer och blir brantare med färre.

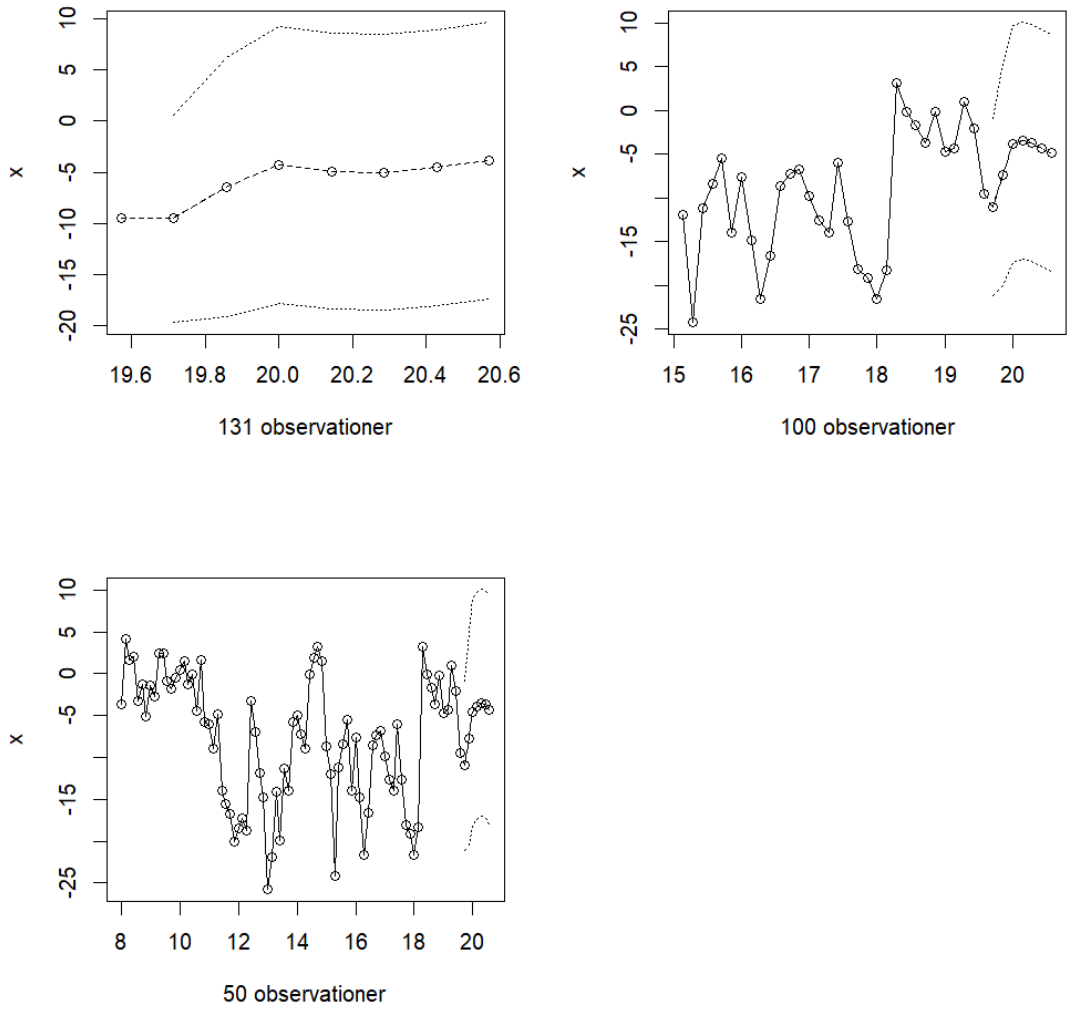

Figure 4: Prognoshorisont

Långtidsmätning av temperaturen

I den längre tidsserien finns observationer tagna tre gånger om dagen, med start 1 januari 1966 till 31 januari 2020.
En tidsserie skapas där enbart morgontemperaturen tagits med och plottas därefter, se figur 5. Ett cykliskt mönster kan urskiljas.

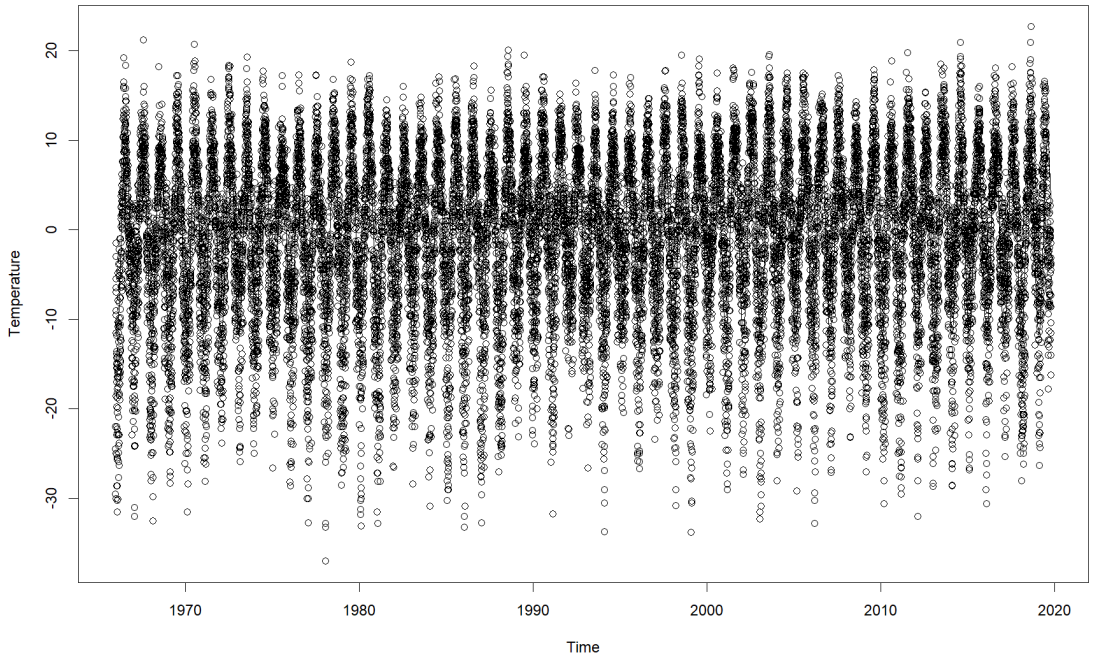

Figure 5: Plot temperatur 1966-2020

En cyklisk modell provas och variansen som förklaras av modellen uppgår till 66%. Även en säsongsmodell provas, den förklarar variationen upp till 67%.
För att skatta parametrarna används EACF och arima(), där det framkommer att tidsserien är en ARIMA(4,4) med $\sigma^2 = 16.24$, μ= -0.8.
Ett Ljung-box test ger ett signifikant resultat, vilket ger en indikation att modellen är korrekt. Ett Augmented Dickey-Fuller test visar att stationäritet inte kan antas.

I figur 6 ses en ACF plot uppe till vänster. Den visar 40 lags (ADF.testet visade enbart 26), och visar en avtagande autokorrelation som verkar öka igen efter 20 lags. I den partiella ACF:en är lag 1 tydligt signifkant med många lags därefter runt noll, vilket tyder på att autokorrelationen försvinner efter derivering av modellen. I ploten nere till vänster ses residualerna i modellen återge samma mönster som i tidsserien, vilket inte visar på några konstigheter. QQploten däremot visar på en avvikelse från normalitet i början av tidsserien.

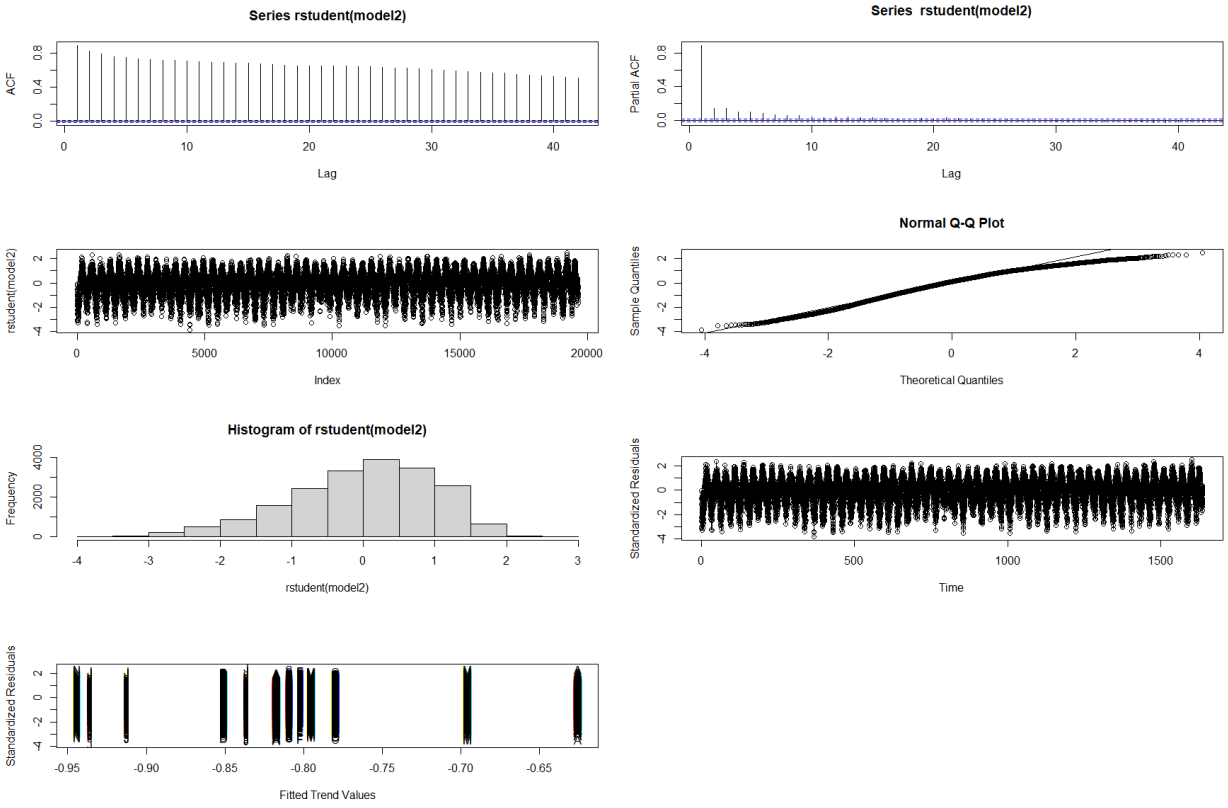

Figure 6: Säsongsmodell, verifikation modell

För att avgöra om en signifikant höjning har skett i slutet av mätningen, används konfidensintervallet från korttidsdatan 2021 och tillämpas på 1966.
Tidsserien delades upp i två delar, där åren 1966 - 1990 ingår i model2 och 1991-2020 ingår i model3.

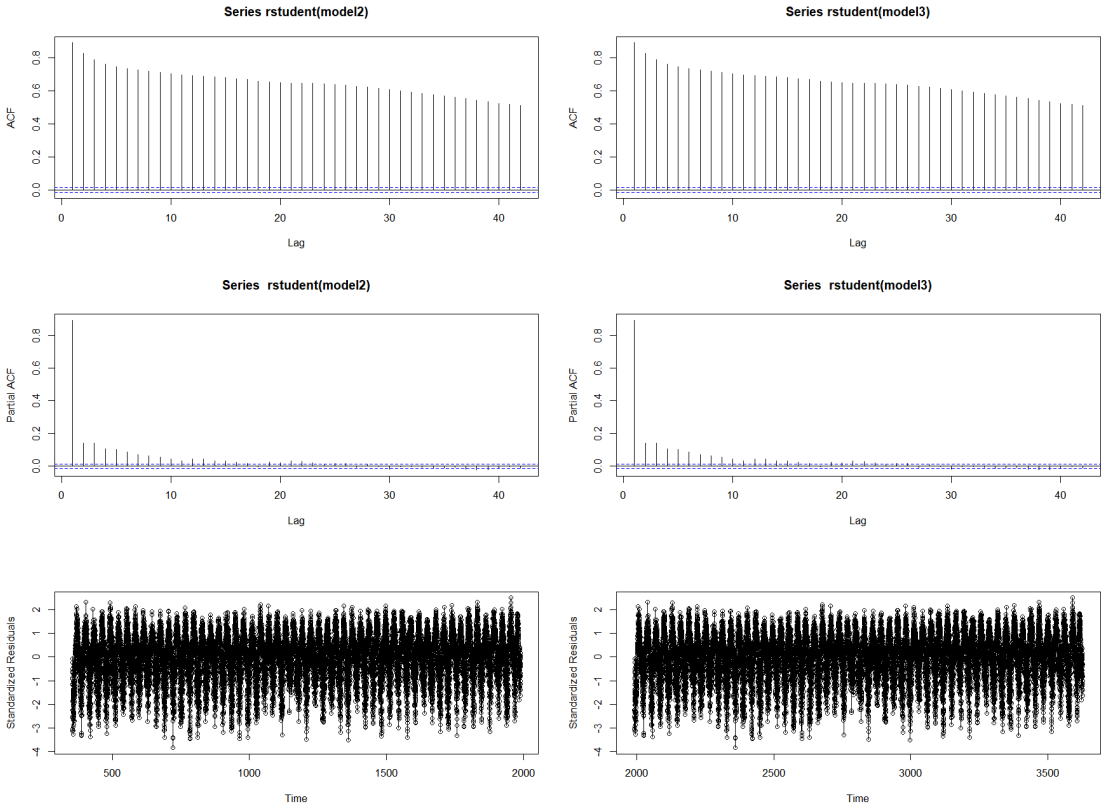

Figure 7: Jämförelse av två modeller, där model2 är åren 66-90 och model 3 är åren 91-20.

Graferna i figur 7 indikerar att det inte finns skillnader i de två modellerna. Summeringen av modellerna gav också identiska resultat.
Det tyder på att det är samma modell i början som i slutet av mätningen.