Stockholms universitet

Timdata eller Dagsdata - Vad predikterar nästkommande dags volatilitet bäst?

Fredrik Käll

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2017:12
Matematisk statistik
Juni 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2017:12**
http://www.math.su.se/matstat

Timdata eller Dagsdata - Vad predikterar nästkommande dags volatilitet bäst?

Fredrik Käll*

Juni 2017

Sammanfattning

Vi undersöker i den här uppsatsen om det med hjälp av högre frekvens är möjligt att skapa sig en bättre volatilitetsprediktion än med lägre frekvens. De modeller vi har valt att undersöka är ARCH och GARCH som vi undersöker under både normal- och t-fördelningsantaganden. Vi kommer fram till att en högre frekvens inte skapar en bättre prediktion vilket kan förklaras av att de modellerna skapade med data av högre frekvens kräver prediktion av er steg, vilket leder till större osäkerhet.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: fredrik.kall@outlook.com. Handledare: Mathias Lindholm och Filip Lindskog.

Innehåll

1 Förord och tack

Denna uppsats ger mig en kandidatexamen i matematisk statistik på Stockholms Universitet. Jag skulle vilja tacka mina handledare Filip Lindskog och Mathias Lindholm för deras vägledning och hjälp under uppsatsen. Vill även tacka mina klasskamrater som har agerat stöttepelare och bollplank under studietiden, och vars motto har varit ”Så länge man har roligt”.

2 Introduktion

2.1 Bakgrund

Volatilitet är ett riskmått som visar hur stora rörelserna varit på en aktie eller ett index. Då volatilitet är ett riskmått är det sällan intressant att veta vad som tidigare varit, utan är mer intressant att undersöka vad volatiliteten i framtiden kommer att vara. Anledningen till varför framtidens volatilitet är intressant är för att man med hjälp av den kan beräkna eventuella pris på finansiella derivat vars underliggande faktor är till exempel en aktie, men även för att beräkna andra typer av riskmått som till exempel Value at Risk.
Då till exempel aktiepris är kontinuerliga, vilket innebär att deras pris ändras kontinuerligt, väcks tanken om det med hjälp av data av högre frekvens går att skapa sig en bättre modell, vilket leder till en bättre prediktion. Tanken varför det eventuellt skulle vara bättre med prediktion av en högre frekvens är att man då tar del av information man går miste om vid lägre frekvens. Ett problem som kommer uppstå är dock att med en högre frekvens krävs det flera prediktionssteg. Vi kommer därför i den här uppsatsen att undersöka om det med hjälp av timupplöst data går att prediktera nästkommande dags volatilitet och om så är fallet, är då denna prediktion bättre än prediktion gjord med hjälp av dagsupplöst data.
Volatilitet är sällan den samma under längre perioder vilket gör att man använder sig av så kallade heteroskedastiska modeller vid prediktion av detta. Vi kommer i den här uppsatsen att fokusera på två stycken av dessa, ARCH och GARCH. Dessa modeller predikterar variansen som vi senare kommer att använda för att skapa oss prediktionsintervall som jämförs mot de sanna värdena.

2.2 Data

Vi kommer i den här uppsatsen att fokusera på två stycken frekvenser av data, timmar och dagar. Data vi kommer att använda oss av i den här uppsatsen är stängningskursen för det börsnoterade bolaget Johnson&Johnson, vilket är hämtad från det ryska finansbolaget Finam [2], och sträcker sig från 1 mars 2012 till och med 2 mars 2017.
Då data är av olika frekvens innebär detta även att antalet observationer skiljer sig för de olika dataseten. Den timupplösta består av 8634 observationer och dagsupplösningen består av 1249 observationer. Eftersom intresset i den här uppsatsen handlar om att prediktera nästkommande dag kommer detta innebära att prediktionen med hjälp av timmar kommer bestå av fler än ett steg. Antalet steg kommer dock skilja sig då det existerar dagar då börsen enbart är öppen halvdagar, men större delen av timprediktion kommer bestå av sju steg då det är antalet öppna timmar en normal dag.
Ett problem som har uppstått kring data är att det vissa dagar inte existerar information från timupplösningen, närmare bestämt 7 dagar. Då antalet dagar där timupplösning saknas inte är så många kommer detta inte ställa till med några problem. Bara för att data inte existerar kan vi inte bortse från detta,

vilket vi har löst genom att vi har slumpat fram dessa från en normalfördelning med standardavvikelse från tidigare dag, och som väntevärde har vi använt dagsavslutet från dagen innan.

2.3 Johnson & Johnson

Johnson & Johnson, JNJ, är ett amerikanskt multinationellt läkemedels- och medicinteknik-bolag som är noterat på New York Stock Exchange, NYSE. Företaget grundades 1886 och har idag ca 116 200 anställda. Deras marknadsvärde uppgår till ca 345 miljarder dollar. Då JNJ är ett så kallat icke-cykliskt bolag så kommer vi i vår data inte heller se så stora svängningar i deras dagliga avkastning, relativt andra aktier.

3 Teori

I den här sektionen går vi igenom den mest grundläggande teorin som kommer att användas under analysen. Teorin är tagen från Tsay [1] om ingenting annat nämns.

3.1 Tidsserie

En tidsserie är observationer tagna från en slumpvariabel vars värde ändras med tiden. Observationerna är tagna allt som oftast med en fast differens, till exempel en vecka, en månad eller ett år. Exempel på tidsserier är minutligt värde på en växelkurs, dagligt stängningsvärde på en aktie och så vidare.

3.2 Volatilitet och Avkastning

Avkastningen på en finansiell tillgång definieras som

$$R_t = \frac{S_t - S_{t-1}}{S_{t-1}}, \tag{1}$$

där S_t är stängningskursen i tidpunkt t och S_{t-1} är stängningskursen i tiden $t-1$, det vill säga tidpunkten innan. Värt att notera är att det verkliga avståndet mellan t och $t-1$ kommer att skilja sig mellan de olika tidsserierna då de är av olika frekvenser. Då inte avkastningen för en tillgång är addidativ är det lämpligt att använda sig av den logaritmerade avkastningen, $\log(R_t)$. Den logaritmerade avkastningen definieras enligt

$$r_t = \log(\frac{S_t}{S_{t-1}}) = \log(S_t) - \log(S_{t-1}). \tag{2}$$

Volatilitet är ett mått på hur stora svängningar till exempel ett pris på en aktie har. Volatilitet mäts som standardavvikelsen, σ, på tidsserien där

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{3}$$

3.3 Autokorrelationsfunktionen (ACF)

Autokorrelationsfunktionen används för att beräkna korrelationen mellan två observationer Y_t och Y_{t-k} där korrelationen ges av ekvationen

$$\rho = \frac{Cov(Y_t, Y_{t-k})}{\sqrt{Var(Y_t)Var(Y_{t-k})}}, \tag{4}$$

där ρ talar om det linjära sambandet mellan tidpunkterna t och $t-k$.

3.4 Partiella Auto Korrelations Funktionen (PACF)

Även den partiella autokorrelationsfunktionen, PACF, visar korrelationen mellan två olika lags. Det som skiljer ACF och PACF är att i PACF så beräknas autokorrelationen betingat på de kortare lagsen. Beräkning av PACF kan göras med hjälp av en AR(p) modell vilket är taget från Tsay [1], se sid 36. En AR modell är en modell som representerar en tidsserie med ett linjärt samband, det vill säga att tidsserien i tidpunkt t beror linjärt på de observerade värdena innan. En AR(p) modell beror således på p stycken dagar innan.
Vi låter oss anta följande AR modeller på varandra följande ordning

$$r_t = \phi_{0,1} + \phi_{1,1} r_{t-1} + e_{1t}$$

$$r_t = \phi_{0,2} + \phi_{1,2} r_{t-1} + \phi_{2,2} r_{t-2} + e_{2t}$$

$$r_t = \phi_{0,3} + \phi_{1,3} r_{t-1} + \phi_{2,3} r_{t-2} + \phi_{3,3} r_{t-3} + e_{3t}$$

$$r_t = \phi_{0,4} + \phi_{1,4} r_{t-1} + \phi_{2,4} r_{t-2} + \phi_{3,4} r_{t-3} + \phi_{4,4} r_{t-4} + e_{4t}$$

$$\vdots$$

där $\phi_{0,j}, \phi_{i,j}$ och $\{e_{jt}\}$ är konstanttermen, koefficienten för r_{t-i} samt feltermen för en AR(j) modell. Dessa modeller är multilinjära funktion och kan bli estimerade med hjälp av minstakvadrat metoden. Vi har vidare att $\phi_{i,i}$ är den partiella autokorrelationen vid lag i. Vi ser även i ekvationerna ovan att $\phi_{i,i}$ är hur mycket r_{t-i} tillför till r_t.

3.5 Tidsseriemodeller

Vid beräkningar och undersökningar av volatilitet använder man så kallade betingande heteroskedastiska modeller. Att en modell är heteroskedastisk menas att data har en förändrad varians med tiden, det vill säga att storleken på svängningarna inte är konstant. Att den är betingad menas med att man betingar den på tidigare tidsperioder. Den enklaste volatilitets modellen är den så kallade ARCH modellen, Autoregressive Conditional Heteroskedasticity modellen.
Innan vi börjar förklara modeller ansätter vi en modell för den logaritmerade avkastningen,

$$r_t = \mu_t + a_t = \mu_t + \sigma_t \epsilon_t \tag{5}$$

där ϵ_t är feltermen som är oberoende och likafördelade N(0,1) eller t-fördelade och

$$\mu_t = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} - \sum_{i=1}^{q} \theta_i a_{t-i} \tag{6}$$

är en medelvärdesekvation med $p, q \in Z$, och är större eller lika med 0, och a_t är så kallade chocker. Varför den här modellen ansätts är för att få bort eventuell korrelation mellan avkastningarna.
När vi nu har ansatt en modell för den logaritmerade avkastningen kan vi titta

på de två olika variansmodeller som kommer ligga som grund för den här uppsatsen.

3.5.1 ARCH

En ARCH modell av graden m definieras som

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 \tag{7}$$

där $a_t = \sigma_t \epsilon_t$ vilket kommer ifrån r_t, ekvation 7. Vi sätter restriktionerna att $a_0 > 0$ samt att $a_i \geq 0$ då variansen inte kan vara negativ eller lika med noll, ty den logaritmerade avkastningen skulle då vara en konstant. Vad vi kan se i ekvation 9 är att en stor chock från tidigare tidpunkter genererar en stor chock i nuvarande tidpunkt. Detta är en av fördelarna med ARCH modellen, att under turbulenta tider så förblir ofta variansen hög under en längre tid, inte bara vid en tidpunkt.

3.5.2 GARCH

GARCH modellen är snarlik ARCH modellen, skillnaden är att GARCH modellen inte bara tar hänsyn till tidigare chocks utan tar även hänsyn till tidigare volatilitet. Vi får därför att GARCH(m,s) definieras

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2. \tag{8}$$

Vad vi då kan se i ekvation 10 är att i GARCH modell är det inte bara höga chockar sedan tidigare som påverkar volatiliteten, utan även hög volatilitet föder hög volatilitet, även kallat volatilitets clustering. I GARCH-modellen ansätts samma restriktioner för α_0 och α_i som i ARCH, det vill säga $\alpha_0 > 0$ och $\alpha_i \geq 0$. Utöver dessa ansätter vi även restriktionen $\beta_i \geq 0$ och $\sum_{i=1}^{max(m,s)} (\alpha_i + \beta_i) < 1$. Den senare restriktionen ansätts för att variansen ska vara ändlig.

3.6 Estimering av parametrar

Estimeringsmetoden som kommer att användas är maximum likelihood vilket är en metod som skattar parametrarna så att sannolikheten att det givna utfallet är maximalt. Maximum likelihood definieras enligt

$$L(x) = \prod_{i=1}^{N} f(x_i), \tag{9}$$

där N är antalet observationer och $f(x_i)$ är sannolikhetsfunktionen för den stokastiska variabeln, X, i oberservation i. Vi kommer i den här uppsatsen att undersöka prediktionen under antagandet om att feltermerna kommer från normal- eller t-fördelning. Normalfördelningens sannolikhetsfunktion definieras

$$f(x) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{\sigma^2}} \tag{10}$$

vilket ger likelihoodfunktionen

$$L(x) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\sigma_i^2\pi}} e^{-\frac{(x_i-\mu)^2}{\sigma_i^2}} \tag{11}$$

och t-fördelningens sannolikhetsfunktion definieras

$$f(x) = \frac{\gamma(\frac{v+1}{2})}{\gamma(\frac{v}{2})\sqrt{\pi v\sigma^2}} (1 + \frac{1}{v}\frac{(x-\mu)^2}{\sigma^2})^{-\frac{v+1}{2}} \tag{12}$$

vilket ger likelihoodfunktionen vid t-fördelning

$$L(x) = \prod_{i=1}^{N} \frac{\gamma(\frac{v+1}{2})}{\gamma(\frac{v}{2})\sqrt{\pi v\sigma_i^2}} (1 + \frac{1}{v}\frac{(x_i-\mu)^2}{\sigma_i^2})^{-\frac{v+1}{2}}, \tag{13}$$

där x_i är det observerade värdet, σ_i^2 är variansen i tidpunkt i och v är antalet frihetsgrader.
Vid estimering av parametrar i ARCH och GARCH används dock en betingad maximumlikelihood, detta då den innehåller värden från tidigare dagar. Vi får således att maximumlikelihooden kommer att definieras enligt

$$L(x) = \prod_{i=1}^{N} f(x_i|x_{i-1}, \ldots, x_{i-n}), \tag{14}$$

där $n = \max\{m, s\}$ och x_i är här i vårt fall den logaritmerade avkastningen.

3.7 Tester

3.7.1 Ljung-Box

Ljung-Box testet[1] är ett test där man testar om autokorrelationen på tidsserien är skild från 0. Ljung-Box följer av

$$Q = n(n+2)\sum_{k=1}^{h} \frac{\rho_k^2}{n-k}, \tag{15}$$

där n är stickprovsstorleken, ρ_k är autokorrelationen under lag k och h är antalet lags man testar. Vi har även att Q är $\chi^2(h)$ under H_0 = data är okorrelerat.

[1] Teori är hämtad från [4]

3.7.2 Skevhet och Kurtosis

Skevhet och kurtosis[2] är två mått som talar om hur asymmetrisk respektive hur tjocka svansar en fördelning har. Ett positivt skevhet värde betyder att fördelningen ofta genererar ett högre värde än väntevärdet, det vill säga mer massa till höger om väntevärdet. Ett negativt värde blir således motsatsen, att fördelningen har mer massa till vänster om väntevärdet. Skevhet beräknas med hjälp av

$$S = \frac{\mu_3}{\mu_2^{3/2}}, \tag{16}$$

där μ_2 och μ_3 står för andra och tredje centralmomentet. Kurtosis är ett mått som talar om tjockleken på svansarna. En normalfördelning har som standard ett kurtosis på tre. Ett kurtosis över tre i normalfördelningsfallet betyder att svansarna är tjocka, det vill säga att sannolikheten för att få ett extremare värde är högre än normalt. Ett lågt värde betyder att sannolikheten att få ett extremt värde är lägre än normalt. Vi använder i den här uppsatsen en fördefinierad kurtosisfunktion i R som defineras enligt

$$C = \frac{\mu_4}{\mu_2^2} - 3, \tag{17}$$

där även här indexeringen på μ står för graden av centralmoment.

3.7.3 MSEP

För att evaluera en prediktion kan man använda MSEP, Mean Square Error Prediction, som definieras enligt följande ekvation,

$$M = \frac{1}{m} \sum_{N-m+1}^{N} (x_t - \hat{x}_t)^2, \tag{18}$$

där N är antalet observationer, m antalet predikterade värden, x_t det sanna värdet och $\hat{x}_t^2$ är det skattade värdet. Då vi predikterar volatilitet som inte är direkt observerbar kommer vi istället använda MSEP för att jämföra de modeller vars andel överträdelser ligger nära det önskade. Detta görs då det rent teoretiskt går att dra en rät linje där andelen överträdelser stämmer överens med den önskade andelen. Detta är dock inte speciellt önskvärt då detta inte är en verklig prediktion. Vårt $\hat{x}_t$ kommer således att vara det övre prediktionsintervallet som vi jämför emot $|x_t|$.

3.7.4 Jarque-Bera Test

För att undersöka om data kommer ifrån en normalfördelning kan man använda sig av ett Jarque-Bera test. JB testet definieras som

$$JB = \frac{n}{6}(\hat{S}^2 + \frac{1}{4}(\hat{C} - 3)^2), \tag{19}$$

[2]Teorin är hämtad från [3]

där n är antalet observationer, $\hat{S}$ är stickprovsskevheten och $\hat{C}$ är stickprovskurtosis. Då data kommer från en normalfördelning så är JB statistiskan χ^2 med två frihetsgrader, vilket leder till att H_0 är att data kommer från en normalfördelning. En normalfördelning har normalt en skevhet på 0 och ett kurtosis på 3, vilket gör att avvikelse från detta påverkar testet.

4 Metod

Vi har i den här uppsatsen valt att dela upp data i två delar, där första delen används för att skapa oss en modell vi sedan tillämpar på den andra delen av data. Då de olika dataseten är av olika frekvens, vilket betyder olika antal observationer, så kommer även de två modellerna att skattas på olika antal observationer. Modellen för dagar kommer att skattas med hjälp av 749 observationer, medan modellen för timmar kommer att skattas på 5189 observationer. Vi kommer därefter alltid att prediktera enbart en dag i taget, vilket för modellen för timmar innebär flera steg. Antalet prediktionssteg för timmodellen kommer att variera då antalet öppna börstimmar kan variera från dag till dag. Antalet dagar som predikteras totalt kommer dock vara densamma.
Som nämnt tidigare kommer vi i den här uppsatsen att undersöka de två heteroskedastiska modeller ARCH och GARCH. Då graden av GARCH är svår att ta fram på egen hand så kommer vi här att enbart använda oss utav GARCH(1,1). Metoden som kommer ligga till grund för framtagandet av ARCH modell är tagen från Tsay [2] vars tillvägagångssätt består av tre stycken huvudsteg, se sid 86.
(1) Bygg en ekonometrisk modell (t ex en ARMA) för avkastningsserien för att ta bort eventuell korrelation. Använd sedan modellens residualer för att testa för ARCH effekt.
(2) Specificera graden på ARCH modellen och utför sen skattningar.
(3) Kontrollera den anpassade modellen och förfina om nödvändigt.

I samband med steg ett, det vill säga ansätta en ekonometrisk modell kommer vi även undersöka vilken eventuell fördelning feltermerna kommer ifrån. Inom finansiella tidsserier gör man ofta antagandet om att data kommer från de symmetriska fördelningarna normal- eller t-fördelning. Vilken av fördelningarna man antar spelar, som tidigare nämnt, roll med tanke på att detta kommer att påverka hur vi skattar parametrarna vilket i sin tur påverkar prediktionen. Den här delen av analysen kommer att utföras med hjälp av diverse plottar och tester, till exempel QQ-plottar och Jarque-Bera Test. Vi kommer även här att undersöka om det existerar någon ARCH-effekt, vilket även är en del av Tsays tillvägagångssätt. När vi har tagit fram de olika modellerna kommer vi att undersöka vilken av modellerna som predikterar bäst. Detta kommer ske med hjälp av prediktionsintervall som vi senare drar slutsatser kring med hjälp av bland annat tester.

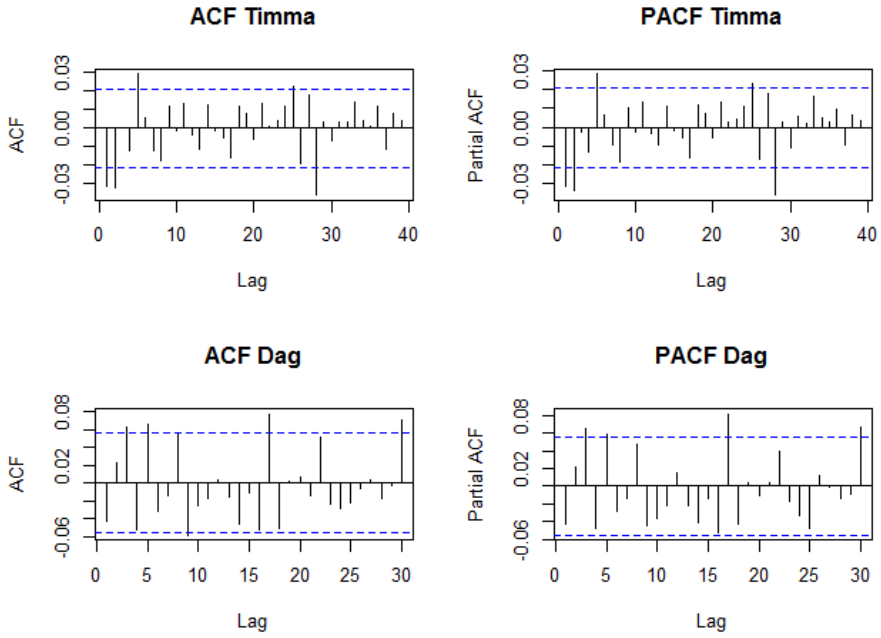
4.1 Modellanpassning

Som nämnt inledningsvis i teorin så ansätter vi modellen

$$r_t = \mu + a_t = \mu + \sigma_t \epsilon_t, \tag{20}$$

och det första steget blir då att ansätta en medelvärdesmodell, det vill säga en modell för μ. Medelvärdesmodellen är av typen ARMA, och som även nämnts i teorin, används för att ta bort eventuell korrelation mellan observationerna. För att undersöka vilken grad av AR samt grad av MA som skall användas undersöker vi ACF och PACF för de olika dataseten, vilka visas i Figur 1 nedan.

I Figur 1 ser vi att det är lag 1, 2 och 4 för timma som passerar prediktionsintervallet för både ACF och PACF, vilket innebär att dessa är signifikanta. Vad vi däremot kan se är att dessa värden är så pass små, vilket innebär att de inte har någon speciellt stor inverkan och väljer därför att bortse från dem. Tittar vi då istället på ACF och PACF för dagsdata så ser vi att det även där finns några signifikanta lags, men även dessa väldigt små, vilket gör att vi även bortser från dessa. Eftersom det inte existerar några signifikanta lags på någon av graferna i Figur 1 så ansätter vi att medelvärdesmodellen enbart är en konstant.



Figur 1: AFC och PACF för logaritmerade avkastnignen för timma och dag.

För att bestämma vilken konstant vi ska ansätta μ till så inleder vi med att titta på grafen över den logaritmerade avkastningen, Figur 10 i Appendix, och kan då se att medelvärdet tycks ligga kring noll. Vi väljer dock även att ta ett medelvärde för datat och ser då att medelvärdet är skiljt från noll, om så ändå väldigt lite. Vi utför därför ett t-test för att undersöka om medelvärdet är signifikant, där H_0 är att medelvärdet är lika med 0. Vi kan i Tabell 1 nedan se värdena för datas medelvärde och standardavvikelse, samt p-värdet för t-testet, och kan då se att medelvärdet tycks vara signifikant. Detta medför att vi kommer att arbeta vidare med modellen

$$r_t = 5.16676 \cdot 10^{-4} + a_t, \tag{21}$$

för dagsdata, och för timdata kommer vi att använda

$$r_t = 7.45923 \cdot 10^{-5} + a_t. \tag{22}$$

Dessa förenklar vi därefter genom att vi drar medelvärdet från respektive observation.

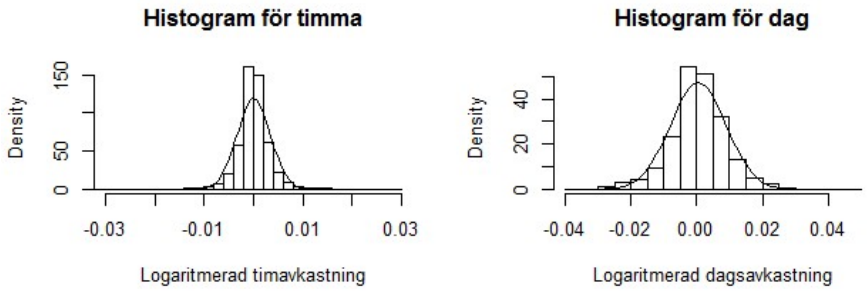
	Timma	Dag
μ	$7.45923 \cdot 10^{-5}$	$5.16676 \cdot 10^{-4}$
σ	$3.35565 \cdot 10^{-3}$	$8.46393 \cdot 10^{-3}$
p-värde	0.03892	0.03123

Tabell 1: Stickprovs-medelvärde, -standardavvikelse och p-värde på t-test, $H_0 : \mu = 0$

När vi nu har ansatt modellen till

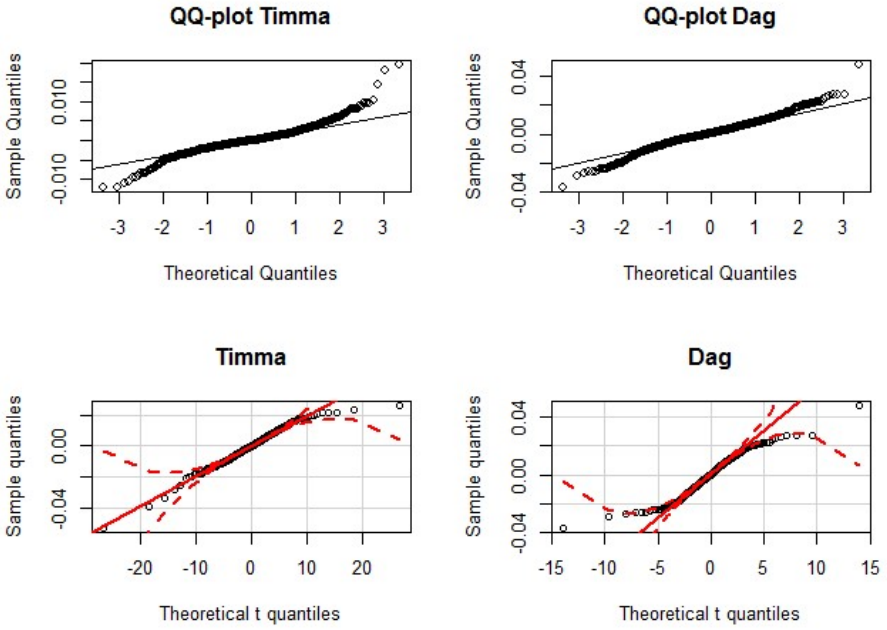
$$\tilde{r}_t = a_t = \sigma_t \epsilon_t,$$

där $\tilde{r}_t = r_t - \mu$, så kan vi undersöka residualernas eventuella fördelning men även undersöka för ARCH-effekt.
Vi har i Figur 2 nedan plottat residualerna för respektive dataset i ett histogram. Vad vi kan se i figuren är som önskat att residualerna tycks komma från en symmetrisk fördelning. Vi har även på histogrammen lagt en normalfördelningskurva med μ = stickprovsmedelvärdet och σ^2 = stickprovsvarianen. Vi kan då se att residualerna tycks följa normalfördelningskurvan ganska väl.



Figur 2: Histogram över residualerna för log avkastnigen för timma och dag

Eftersom residualerna inte ser helt främmande ut för en normalfördelning så plottar vi även data i en QQ-plot, vilket visar de teoretiska kvantilerna mot de observerade. Vi ser då i Figur 3 på den övre raden en QQ-plot mot normalfördelningens kvantiler, där vi kan notera att, som vi även såg i histogrammen, att data ser ut att komma från en symmetrisk fördelning. Vad vi däremot kan se i QQ-plotten som vi inte såg i histogrammen är att data tycks ha avvikande svansar. Vi ser även att dagsdata har mindre avvikande svansar än timdata, vilket kan bero på den centrala gränsvärdessatsen.
Då residualerna tycks ha avvikande svansar så väljer vi även att plotta data mot de teoretiska kvantilerna från en t-fördelning, vilket vi kan se i den undre raden av Figur 3. Vi kan här se timdata tycks vara bättre anpassat till en t-fördelning, medan dagsdata tycks vara sämre anpassat på en t-fördelning än en normalfördelning.



Figur 3: QQ-plot för timma och dag. Övre raden för normalfördelning och den undre för t-fördelning

Då vi har sett att data tycks ha avvikande svansar, men anser att det är symmetriskt väljer vi även att utföra test för precis detta, skevhet och kurtosis. Vi väljer även att göra ett JB-test vilket testar om data är från en normalfördelning, där vi under H_0 antar att data är normalfördelat. Vi kan då i Tabell 2 nedan avläsa resultaten och ser då att båda data tycks vara symmetriskt, men har båda ett större kurtosis än 0, vilket betyder att extrema utfall är mer sannolika än i det normala fallet. Vad som dock är intressant är att JB-testet ger väldigt låga p-värden vilket gör att vi kan förkasta nollhypotesen, vilket innebär att data inte kommer från en normalfördelning. Trots detta resultat kommer vi välja att vid fortsatta analyser även undersöka under antagandet om normalfördelning då både histogrammen och QQ-plottarna tyder på att data kan komma från denna fördelning.

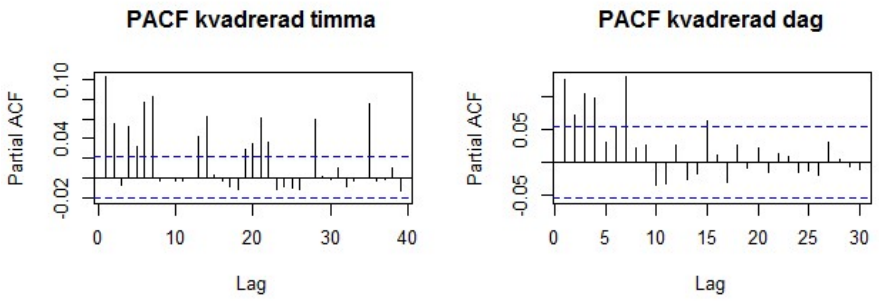
	Skevhet	Kurtosis	JB
Timma	0.5262	5.5835	p-value $<$ 2.2e-16
Dag	-0.0841	1.8052	p-value $<$ 2.2e-16

Tabell 2: Skevhet, Kurtsis och p-värde för Jarque-Bera test

Vidare ska vi undersöka om det existerar någon ARCH-effekt vilket undersöks dels genom att titta på de kvadrerade feltermerna i en ACF men även med hjälp av ett LB-test. Vi vill här se att det existerar signifikanta lags då detta betyder att svängningarna är korrelerade. Vi kan i Figur 11 i appendix se om det där existerar en ARCH-effekt. Vad vi även kan se i ACFen för de kvadrerade feltermerna för timmar är att det tycks existera någon typ av säsongseffekt. Då det inte finns någon heteroskedastisk modell som tar hänsyn till säsongseffekt så är det här någonting vi bortser ifrån men ha det i åtanke, för att se om det ställer till med problem för vidare analys. Vidare kan vi i Tabell 3 nedan se resultaten från LB-testen vilka visar på att vi kan förkasta H_0 = att det inte existerar någon ARCH effekt, vilket bekräftar det vi såg i ACFerna.

Serie	**P-värde**
Timma	$2.975 \cdot 10^{-4}$
Dag	$1.638 \cdot 10^{-3}$

Tabell 3: Ljung-Box test för timma och dag, H_0 = data är okorrelerat



Figur 4: PACF för kvadrerade logaritmerade avkastningen för timma och dag.

När vi har bekräftat att det existerar en ARCH effekt så ska vi bestämma vilken grad av ARCH som skall användas. Graden av ARCH bestämmer vi genom att se hur många signifikanta lags det existerar i PACF för de kvadrerade residualerna. Vi kan i Figur 4 ovan se att PACF för de båda dataseten visar på sju stycken signifikanta lags, vilket då även innebär att båda modellerna kommer att vara ARCH(7).

4.2 Modellvalidering

Vi har i tidigare sektioner kommit fram till att vi ska undersöka de två modellerna ARCH(7) samt GARCH(1,1) under antaganden om att feltermen är normal- eller t-fördelad. Vi fortsätter därför med att skatta parametrarna för de åtta olika modellerna, för att därefter undersöka dess signifikans. Parametrarna är som nämnt tidigare skattade på den första delen av respektive data,

och skattningarna för ARCH(7) kan avläsas i Tabell 4 och för GARCH(1,1) kan parameterskattningarna avläsas i Tabell 5. Vi har i tabellen valt att markera de parameterskattningar som är signifikanta på 5% nivån med *, de som är signifikanta på 1% nivån med ** och de som inte är signifikanta på varken 1% eller 5% nivå utan någonting.

	Timma Norm	Timma Std	Dag Norm	Dag Std
ω	$4.474 \cdot 10^{-6} **$	$4.221 \cdot 10^{-6} **$	$3.250 \cdot 10^{-5} **$	$2.970 \cdot 10^{-5} **$
α_1	$1.004 \cdot 10^{-1} **$	$1.590 \cdot 10^{-1} **$	$2.077 \cdot 10^{-1} **$	$2.023 \cdot 10^{-1} **$
α_2	$9.209 \cdot 10^{-3}$	$4.276 \cdot 10^{-2} **$	$1.104 \cdot 10^{-1} *$	$1.103 \cdot 10^{-1}$
α_3	$1.000 \cdot 10^{-8}$	$6.876 \cdot 10^{-4}$	$7.530 \cdot 10^{-2}$	$1.078 \cdot 10^{-1}$
α_4	$1.000 \cdot 10^{-8}$	$1.972 \cdot 10^{-3}$	$1.136 \cdot 10^{-1} *$	$1.345 \cdot 10^{-1} *$
α_5	$1.000 \cdot 10^{-8}$	$1.000 \cdot 10^{-8}$	$1.000 \cdot 10^{-8}$	$1.715 \cdot 10^{-2}$
α_6	$1.361 \cdot 10^{-1} **$	$1.203 \cdot 10^{-1} **$	$1.443 \cdot 10^{-2}$	$1.934 \cdot 10^{-2}$
α_7	$3.696 \cdot 10^{-1} **$	$4.262 \cdot 10^{-1} **$	$1.000 \cdot 10^{-8}$	$1.000 \cdot 10^{-8}$

Tabell 4: Parameterskattningar, ARCH(7), ** innebär signifikanta på 1% och * att de är signifikanta på 5%

	Timma Norm	Timma Std	Dag Norm	Dag Std
ω	$1.455 \cdot 10^{-7} **$	$1.879 \cdot 10^{-7} **$	$8.371 \cdot 10^{-6} *$	$7.841 \cdot 10^{-6} *$
α_1	$2.981 \cdot 10^{-2} **$	$4.101 \cdot 10^{-2} **$	$1.773 \cdot 10^{-1} **$	$1.926 \cdot 10^{-1} **$
β_1	$9.555 \cdot 10^{-1} **$	$9.483 \cdot 10^{-1} **$	$7.018 \cdot 10^{-1} **$	$7.021 \cdot 10^{-1} **$

Tabell 5: Parameterskattningar, GARCH(1,1), ** innebär signifikanta på 1% och * att de är signifikanta på 5%

Vi kan i tabellerna avläsa att det existerar många parameterskattningar i ARCH(7) modellerna som inte är signifikanta och väljer därför att ta bort dessa för att skatta om parametrarna utan dessa. Vi kan även i Tabell 5 se att alla parameterskattningar är signifikanta i de fyra GARCH(1,1) modellerna. I Tabell 6 presenteras de nya skattningarna efter exkludering av de insignifikanta. Vid en första exkludering visade det sig även att α_2 blev insignifikant för timdata under antagandet om t-fördelning, och det är därför som även den har försvunnit. Vi ser då även att nästan alla parametrar nu är signifikanta på 1% nivån.

	Timma Norm	Timma Std	Dag Norm	Dag Std
ω	$3.610 \cdot 10^{-6} **$	$3.617 \cdot 10^{-6} **$	$3.493 \cdot 10^{-5} **$	$3.896 \cdot 10^{-5} **$
α_1	$9.197 \cdot 10^{-2} **$	$9.201 \cdot 10^{-2} **$	$2.460 \cdot 10^{-1} **$	$2.836 \cdot 10^{-1} **$
α_2	-	-	$1.122 \cdot 10^{-1} *$	-
α_3	-	-	-	-
α_4	-	-	$1.450 \cdot 10^{-1} **$	$1.667 \cdot 10^{-1} **$
α_5	-	-	-	-
α_6	$8.341 \cdot 10^{-2} **$	$8.291 \cdot 10^{-2} **$	-	-
α_7	$3.691 \cdot 10^{-1} **$	$3.692 \cdot 10^{-1} **$	-	-

Tabell 6: Parameterskattningar efter exkludering av insignifikanta parametrar, ARCH(7), ** innebär signifikanta på 1% och * att de är signifikanta på 5%

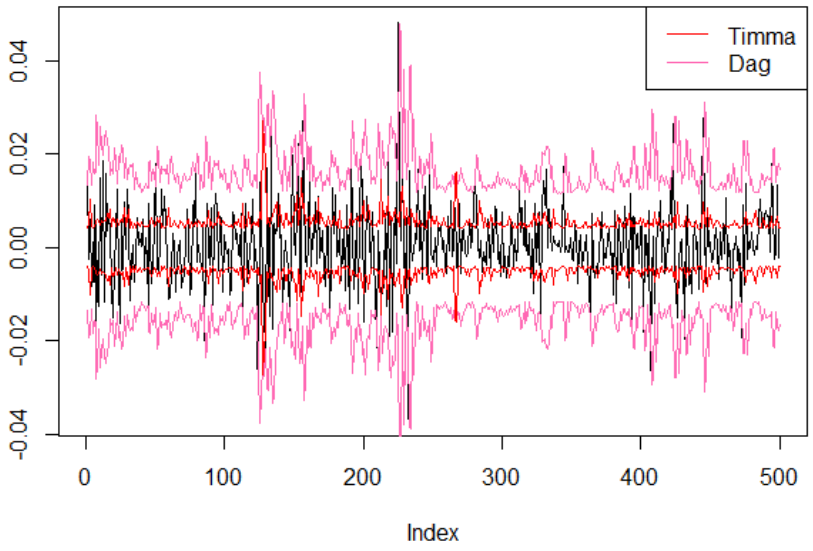
Vi får således 8 stycken modeller vars prediktionsförmåga vi kommer att undersöka.

5 Analys

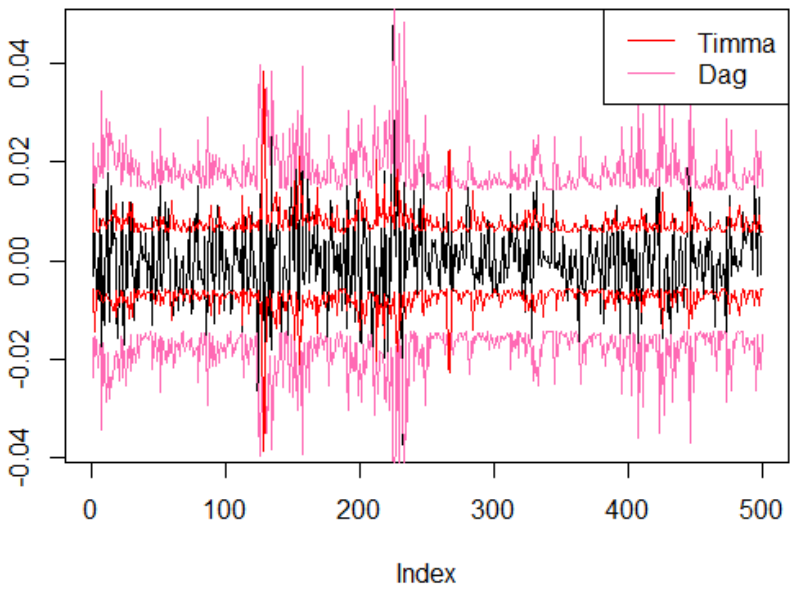
När vi nu har tagit fram åtta stycken modeller ska vi se hur väl dessa predikterar. Som nämnts inledningsvis så ligger intresset i den här uppsatsen i att undersöka om man med hjälp av en högre frekvens kan skapa en bättre prediktion än med en lägre frekvens, men vi kommer även att se vilken modell som bedöms mest lämplig att använda sig av. Utöver detta så måste modellerna undersökas om de tar bort eventuellt beroende. Detta gör vi genom att lösa ut för ϵ_t i ekvation 22 som vi sedan plottar i ACFer.

Då vi arbetar med volatilitetsmodellerna ARCH och GARCH som predikterar nästkommande dags varians kommer vi använda oss av ett 95%-igt prediktionsintervall. Vi kommer att använda prediktionsintervallet för att se hur många av de sanna värdena som överträder detta, samt undersöka hur stort avståndet är mellan de sanna värdena och prediktionsintervallet. Detta gör vi för att se hur pass nära vårt prediktionsintervall ligger.

Vad som åter är värt att nämna är att vi har skapat modeller på den första delen av datasetet, det vill säga 749 observationer för dagsmodellerna och 5189 observationer för timmodellerna. Detta kan vara intressant att ha i åtanke då vi utvärderar modellen eftersom att evalueringen sker över ca två års period, vilket även innebär att den bäst lämpade modellen kan vara en annan i slutet av evaluerings datat.
Vi inleder med att undersöka de olika ARCH modellerna, vars prediktionsintervall är plottat mot de sanna värdena i Figur 5 samt Figur 6. Figur 5 visar prediktion gjorda med ARCH under normalfördelnings antagandet, medan Figur 6 visar prediktion gjord under antagande om t-fördelade feltermer. Prediktion gjord med hjälp av dagsmodeller är plottade i rosa och prediktion gjord med timmodeller representeras av de röda graferna.

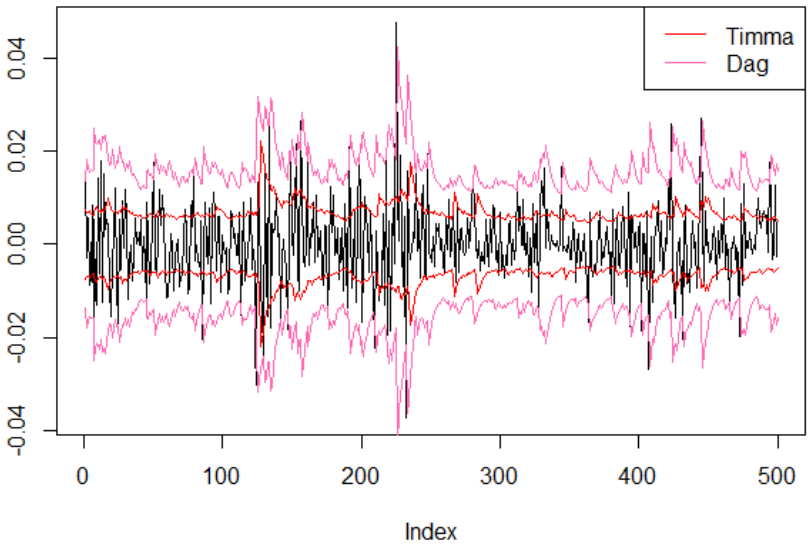


Figur 5: Sanna värden mot prediktionsintervall med ARCH(7) där feltermerna är normalfördelade

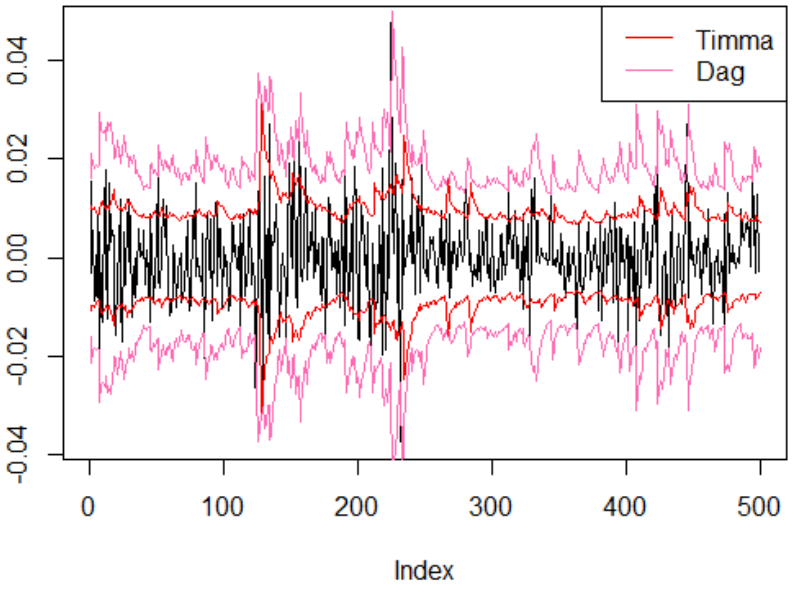


Figur 6: Sanna värden mot prediktionsintervall med ARCH(7) där feltermerna är t-fördelade

Vad vi i Figur 5 och 6 kan se är att såväl den normal- och t-fördelade dagsmodellen tycks fånga upp de största svängningarna vilket medför att prediktionsintervallen följer de sanna värdena bättre. Av enbart graferna är det svårt att avgöra om någon av dessa skulle kunna anses bättre än den andra, men vad som går att se är att andelen överträdelser inte tycks vara speciellt stor. Tittar vi istället på prediktionen gjord med hjälp av timmar kan vi se att både under normal- och t-fördelningsantagande tycks dessa underprediktera. Vad vi även kan se är att de t-fördelade feltermerna, som presenteras i Figur 6, tycks fånga upp de större svängningarna en aning bättre än de normalfördelade, om så ändå väldigt lite. Även dessa är svårt att se om det är någon är bättre än den andra. Vad vi däremot kan se är att vid modellering med en ARCH-modell så tycks dagsdata prediktera bättre än timdata, både vid normal- och t-fördelning.



Figur 7: Sanna värden mot prediktionsintervall med GARCH(1,1) där felter-merna är normalfördelade



Figur 8: Sanna värden mot prediktionsintervall med GARCH(1,1) där felter-merna är t-fördelade

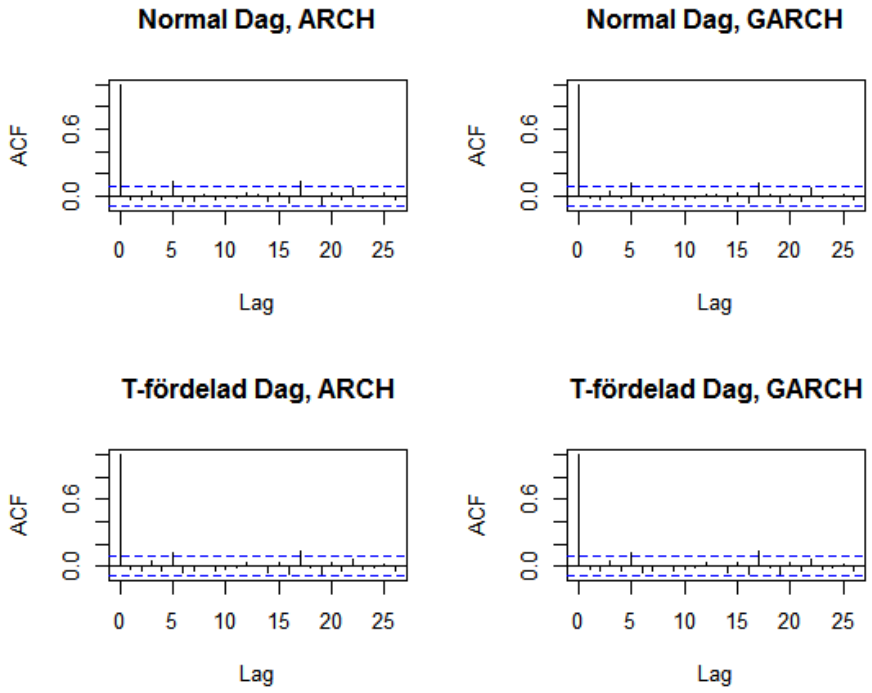
Studerar vi vidare Figur 7 och Figur 8 som visar de sanna värdena mot ett prediktionsintervall från GARCH med normal- respektive t-fördelning så ser vi att dessa intervall uppträder ganska likt prediktionsintervallen skapade med hjälp av ARCH modellerna. Vad vi kan se är att även här tycks dagsdata prediktera ganska väl medan timdata återigen underpredikterar. Vad man mer kan se är att timprediktionen inte får speciellt stora svängningar över lag utan tycks vara ganska raka, men även att det bara tycks vara de större verkliga svängningarna som tycks påverka timprediktionen.
Då det är svårt att se antalet överträdelser i graferna så är dessa presenterade i Tabell 7 nedan, där vi även har beräknat andelen överträdelser och * betyder att det är en GARCH modell. Då vi använder oss av ett 95%-igt prediktionsintervall vill vi även att andelen överträdelser skall ligga kring 5%. Vi kan då i Tabell 7 se att alla prediktionsintervall gjorda med dagsdata tycks ligga väldigt nära den önskade andelen överträdelser. Vad vi även kan se i tabellen, som vi även noterade i graferna, är att timmodellerna underpredikterade då deras andel ligger mellan 20 och 45%. Vi kan även se att det är de t-fördelade dagsmodellerna som tycks vara bättre av dessa fyra.

	Antal	Andel	Frihetsgrader
Timma Norm	231	0.462	-
Timma Std	163	0.326	4
Dag Norm	41	0.082	-
Dag Std	27	0.054	9
*Timma Norm	203	0.406	-
*Timma Std	113	0.226	4
*Dag Norm	39	0.078	-
*Dag Std	26	0.052	8

Tabell 7: Antalet överträdelser, andel överträdelser och antal frihetsgrader för de skattade prediktionsintervallen, där * representerar att det är en GARCH-modell

Utifrån att enbart studera graferna och beräkningar är det svårt att dra slutsatsen om vilken modell som är bäst lämpad för prediktion. Slutsatsen vi kan dra av graferna och Tabell 7 är däremot att dagsdatas prediktionsintervall framstår betydligt bättre än timdatas. Varför vi anser dessa bättre är delvis för, som nämnts ovan, att prediktionsintervallen följer de sanna värdena bättre vid användning av dagsdata, men även att andelen överträdelser för dagsdata ligger nära 5% vilket är den önskade andelen. Timmodellerna har som ovan nämnt en överträdelse på över 20%, vilket är en bit över det önskade.

I och med att vi enbart väljer att gå vidare med dagsdata undersöks även bara dessa om dom rensar för beroende. Vi kan då se i Figur 9 nedan som visar de kvadrerade feltermerna efter modellanpassning att det inte existerar några signifikanta lags vilket också innebär att det inte existerar någon korrelation mellan feltermerna. Då vi kollade efter ARCH-effekt såg vi dock att det tycktes existera en säsongseffekt för feltermerna i timdata. För att se ifall detta kunde vara någonting som låg till grund till att timprediktionen är sämre än dagsdata valde vi därför att undersöka de kvadrerade feltermerna efter modellanpassning på dessa också. Vi kan i Figur 12 i Appendix se att modellerna tycks ha rensat för säsongseffekt då det enbart tycks existera ett fåtal lags som överstiger konfidensintervallet. Då dessa är så pass låga och få kan vi dock göra antagandet om att det inte är dessa som påverkar prediktionen med timdata.



Figur 9: AFC för de kvadrerade feltermerna för dagar efter modellanpassning

När vi nu har dragit slutsatsen att dagsmodellerna är de modeller som tycks prediktera bäst och även sett att dessa modeller rensar för eventuellt beroende ska vi undersöka om man kan säga om någon av modellerna är bättre än dom andra. För att undersöka detta utför vi ett MSEP test där vi jämför det övre prediktionsintervallet mot absolutbeloppet av de sanna värdena. Den modell som kan anses bäst är den vars MSEP värde är det lägsta, vilket betyder att modellens prediktionsintervall ligger närmast de sanna värdena. Vi kan i Tabell 8 se respektive modells MSEP. Då vi arbetar med den logaritmerade avkastningen så kommer värden vara väldigt små, vilken innebär att vi inte kan titta på dom enskilt utan bör jämföra dem mot varandra. Vad vi kan se när vi jämför dessa är att de som framstår bäst är de modeller med normalfördelade feltermer, men då vi sen tidigare vet att GARCH modellerna har en överträdelseandel nära den önskade så är det trots detta svårt att avgöra om någon modell är bättre än den andra.

	MSEP
Dag Norm	$1.453 \cdot 10^{-4}$
Dag Std	$2.010 \cdot 10^{-4}$
*Dag Norm	$1.429 \cdot 10^{-4}$
*Dag Std	$2.138 \cdot 10^{-4}$

Tabell 8: MSEP värde för respektive dagsmodell där * representerar GARCH-modell

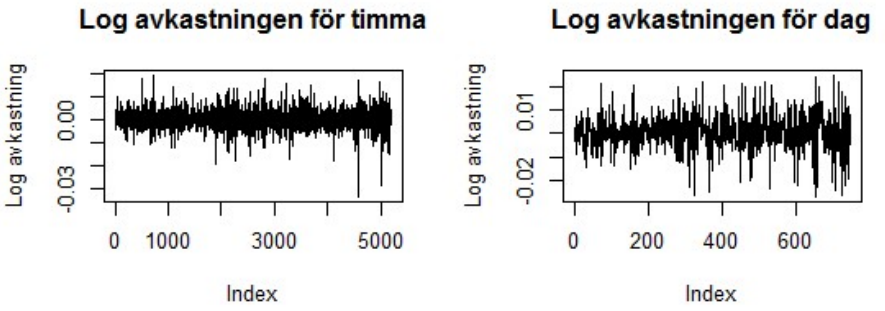
6 Slutsats och Diskussion

Uppsatsens huvudsyfte var att undersöka om det med hjälp av data av högre frekvens går att prediktera volatilitet bättre än med lägre frekvens. Den bakomliggande idén var att med högre frekvens får man tillgång till fler observationer under samma tidsperiod.

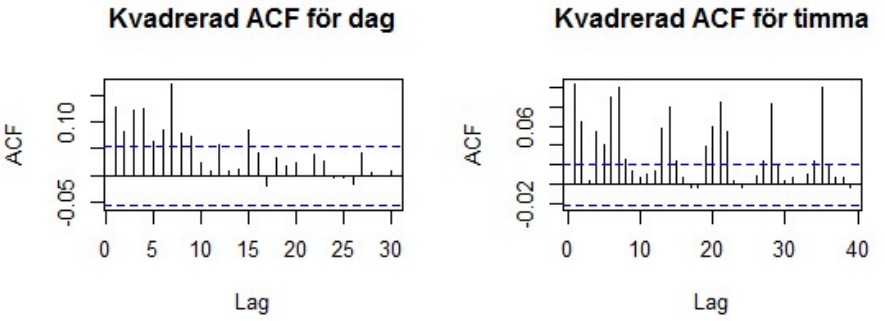
Vi skapade oss fyra modeller från respektive frekvens under två olika antaganden om feltermernas fördelning. Dessa åtta modellers prediktionsförmåga undersöktes senare via jämförelse av prediktionsintervall mot de sanna värdena, där vi kom fram till att de modeller som var bäst lämpad för volatilitetsprediktion var modellerna skapade av dagsdata. Då alla tester för att utvärdera prediktion var bra för alla dagsmodeller och även snarlika var det svårt att avgöra om någon av dessa modeller var bättre än någon annan. Anledningen till att modellen med lägre frekvens, dagsdata, predikterar bättre än den med högre frekvens, timdata, ligger troligtvis i att den med högre frekvens behöver prediktera flersteg, givet att prediktionen av samma punkt är eftersökt. Problemet som uppstår är att prediktion längre än ett steg kommer att innehålla minst ett predikterat värde, det vill säga ett värde som inte är exakt. Vid ökat antalet steg kommer detta skapa mer och mer osäkerhet i prediktionen. Som ett exempel så innehöll timprediktion med hjälp av ARCH(7) ofta enbart ett observerat värde samt sex stycken predikterade.

Då det existerar fler än bara två heteroskedastiska modeller skulle man som vidare studie kunna undersöka om man hjälp av någon annan modell skulle kunna skapa en bättre prediktion för någon av frekvenserna, eventuellt bara den ena. Man skulle även kunna undersöka andra frekvenser än de undersökta i denna uppsats. Varför andra frekvenser skulle vara av intresse är för att vid dagsprediktion med hjälp av timmar så krävs det upp till och med sju prediktionssteg, till skillnad från till exempel månadsprediktion med hjälp av veckor vilket bara kräver som maximalt fem stegs prediktion. Då antalet prediktionssteg minskar så minskar även antalet fel som uppstår. Ett annat sätt att undersöka samma typ av problem är att man istället för att ansätta en modell på första delen av data så kan man efter varje prediktion skatta om parametrarna vilket gör att modellen blir bättre anpassad. Att skatta om parametrarna skulle eventuellt få de timbaserade modellerna att prediktera bättre, men då dagsmodellerna fungerade väl är det inte säkert att denna metod skulle tillföra speciellt mycket.

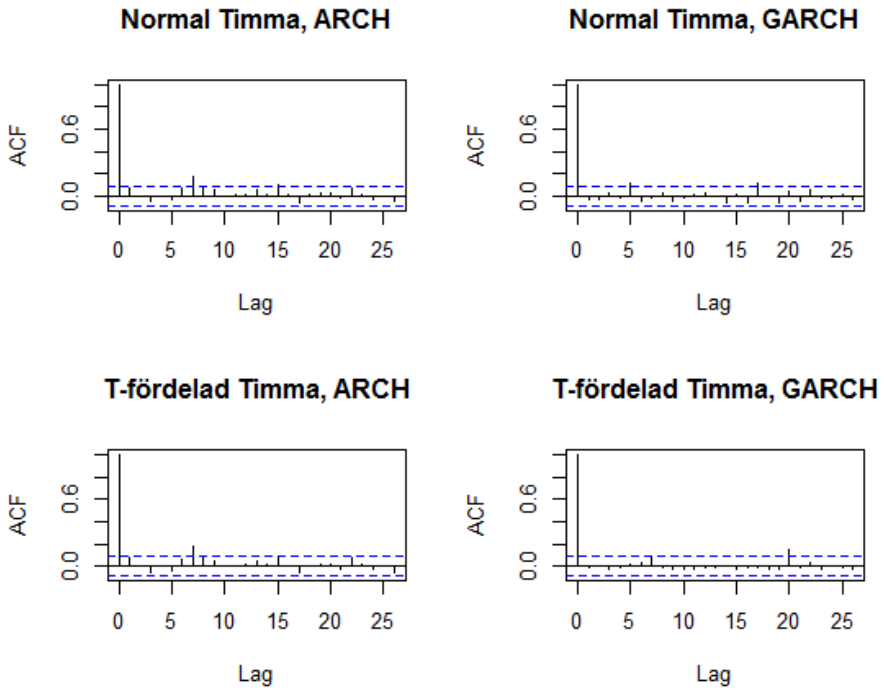
7 Appendix



Figur 10: Den logaritmerade avkastningen för de timmar och dagar som används för modellanpassning



Figur 11: ACF för de kvadrerade feltermerna för timmar och dagar före modellanpassning



Figur 12: ACF för de kvadrerade feltermerna för timmar efter modellanpassning

8 Referenser

[1] Tsay R.S. *Analysis of Financial Time Series 3rd ed.* John Wiley and Sons, Inc., (2010).

[2] Finam
`https://www.finam.ru/`, Apr 2017

[3] R-tutor
`http://www.r-tutor.com/`, Maj 2017

[4] Peter J. Brockwell, Richard A. Davis *Introduction to Time Series and Forecasting Volume 1* Taylor Francis, 2002
`https://books.google.se/books?id=VHB4OSAmwcUCpg=PA35redir`$_esc = yv = onepageq = ljungf = false, Maj$2017

[5] Kronman Sanna *The volatility of tomorrow - Comparison of GARCH and EGARCH models applied to Texas Instruments stock returns* . (2015).

[6] Zhi Li *Estimating dynamic volatility of returns for Deutsche Bank* . (2015).

[7] Avanza
`https://www.avanza.se/`, Apr 2017