



Stockholms
universitet

Hur samverkar bensinpris och råoljepris?

En vektor autoregressiv analys

Shuhua Zhang

Kandidatuppsats 2017:14
Matematisk statistik
Juni 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Hur samverkar bensinpris och råoljepris?

En Vektor Autoregressiv analys

Shuhua Zhang*

Juni 2017

Sammanfattning

I den här uppsatsen undersöker vi förhållandet mellan bensinpris i USA och råoljepris med hjälp av en så kallad *Vektor Autoregressiv modell*. Genom att använda veckovisa bensinpriser (bensinpriset för samtliga kvaliteter och typer i USA) och råoljepriser så kallad "Brent - Europe" mellan 9 juni 2003 och 10 april 2017, upptäcker vi att en vektor autoregressiv modell med lag 2 (VAR(2)) visar sig vara den bäst anpassade modellen för de första 500 observationerna i data. Modellen visar att bensin- och råoljepriset har ett återkopplingsförhållande men påverkas inte av samma lags i deras historiska dataserier. Vi upptäcker också att förklaringsgraden till ekvationen för råolja är mindre än 10%, vilket vittnar om att data över historiska bensinpriser i USA inte är en viktig förklarande faktor för råoljepriset.

Prediktionsförmågan utvärderas med hjälp av de resterande 223 datapunkterna samt genom att beräkna RMSE för våra modeller. Resultaten pekar på att VAR(2) har bäst prediktionsförmåga och att våra vektor autoregressiva modeller predikterar bensinpriset bättre än motsvarande AR modell. Sammanfattningsvis kan vi dra slutsatsen att VAR(2) är en tillräcklig modell, inte bara baserat på anpassning, utan även på prediktion.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: navalniteflora@gmail.com. Handledare: Filip Lindskog, Mathias Millberg Lindholm.

Förord och tack

Denna uppsats utgör ett självständigt arbete om 15 hp vilket leder till en kandidatexamen i matematisk statistik vid Stockholms Universitet.

Jag vill tacka mina handledare, Filip Lindskog, speciellt Mathias Lindholm, vars goda ideer och vägledning har varit av stor betydelse. Jag vill även tacka mina vänner i synnerhet Andreas Rosen Björkhage.

Hur samverkar bensinpris och råoljepris?

En Vektor Autoregressiv analys

Shuhua Zhang

4 juni 2017

Sammanfattning

I den här uppsatsen undersöker vi förhållandet mellan bensinpris i USA och råoljepris med hjälp av en så kallad *Vektor Autoregressiv modell*. Genom att använda veckovisa bensinpriser (bensinpriset för samtliga kvaliteter och typer i USA) och råoljepriser så kallad "Brent - Europe" mellan 9 juni 2003 och 10 april 2017, upptäcker vi att en vektor autoregressiv modell med lag 2 (VAR(2)) visar sig vara den bäst anpassade modellen för de första 500 observationerna i data. Modellen visar att bensin- och råoljepriset har ett återkopplingsförhållande men påverkas inte av samma lags i deras historiska dataserier. Vi upptäcker också att förklaringsgraden till ekvationen för råolja är mindre än 10%, vilket vittnar om att data över historiska bensinpriser i USA inte är en viktig förklarande faktor för råoljepriset.

Prediktionsförmågan utvärderas med hjälp av de resterande 223 datapunkterna samt genom att beräkna RMSE för våra modeller. Resultaten pekar på att VAR(2) har bäst prediktionsförmåga och att våra vektor autoregressiva modeller predikterar bensinpriset bättre än motsvarande AR-modell. Sammanfattningsvis kan vi dra slutsatsen att VAR(2) är en tillräcklig modell, inte bara baserat på anpassning, utan även på prediktion.

Abstract

This paper studies the relationship between gasoline prices in the US and crude oil prices by employing *Vector Autoregressive Models*. Using weekly data between 9th of June 2003 and 10th of April 2017, we find that a vector autoregressive model with order 2 (VAR (2)) proves to be the best fitted model for the first 500 observations in data. The model shows that gasoline prices and crude oil prices have a feedback relationship but are not affected by same lags in their historical data. We also find that the adjusted R square ($\text{Adj-}R^2$) for the crude oil equation is less than 10%, which indicate that data on historical gasoline prices in US is not an important explanatory factor for crude oil prices.

Predictive abilities of models are evaluated by using the remaining 223 observations in data as well as calculating RMSE for our models. Results indicate that VAR (2) has the best predictive ability and that our vector autoregressive models forecast gasoline prices better than the corresponded AR models. In summary we conclude that VAR (2) is a sufficient model not only on the description but also on prediction of gasoline data.

Innehåll

1	Inledning	5
1.1	Bakgrund	5
1.2	Tidigare forskning	5
1.3	Studiemål	6
2	Teori	7
2.1	Tidsserie, Stationäritet	7
2.2	ACF och CCM	8
2.3	Ljung-Box-test	10
2.4	VAR modell	11
2.4.1	VAR modell	11
2.4.2	Metod till lag urval: AIC och M(p)	12
2.4.3	Metod till parameter skattning: OLS och ML	13
3	Data	15
3.1	Databeskrivning	15
3.2	Datahantering	16
4	Metod och Resultat	17
4.1	Undersökning av stationäritet	17
4.2	Korskorrelationsundersökning	18
4.3	Lagspecificering och parameterskattning	20
4.4	Modellgranskning	22
4.5	Prediktion	24
5	Slutsats och Diskussion	27
6	Appendix	30
7	Referens	33

1 Inledning

1.1 Bakgrund

Oljan är sedan ungefär 50 år världens mest använda energiråvara. Det beror på att den innehåller mycket energi och är lätt att lagra och transportera. Trots ökad oro för klimatförändringar och många förnybara energikällors snabb tillväxt under 2000-talet, dominerar fortfarande oljan 35% av totala energiförsörjningen i världen. Den internationella energiorganisationen IEA (International Energy Agency) har gjort bedömningen att efterfrågan på energi i världen kommer att öka med ungefär två procent per år fram till år 2020 och oljans andel av den totala energitillförseln kommer att vara lika stor som idag [1]. Bensin som en oljeprodukt har en helt dominerande roll för motorfordon, båtar och flyg i USA och övriga världen. År 2016 förbrukades totalt 143,37 miljarder gallon bensin i USA och bensinförbrukning omfattade i genomsnitt 5% av U.S. hushållens totala utgifter. Juni i 2016 nådde konsumtionen av bensin i USA ett nytt rekord av 9,7 miljoner gallon per dag, vilket motsvarar en 1,8% ökning jämfört med samma period 2015 (U.S. Energy Information Administration).

Bensin och råoljas särställning inom energiförsörjning leder till stort intresse av hur man kan förbättra bevakningen av prisutvecklingen samt prediktionen av denna. I USA bestämmer bensinpris av fyra huvudliga komponenter: råoljepris, raffineringsekostnader, återförsäljning- och transportkostnader och skatt. Bensinpriset reflekteras av de ovan nämnda komponenterna tillsammans med vinstmål av diverse marknadsaktörer (raffinörer, marknadsförare, stations ägare m.m.). Med det sagt så påverkas bensinpriset i huvudsak av priset på råolja (U.S. Energy Information Administration) medan priset på råolja i sin tur bestäms av global tillgång och efterfrågan. Världsekonomisk tillväxt är den viktigaste faktorn som påverkar efterfrågan. En viktigast faktor i oljeförsörjningen är OPEC:s organisation (the Petroleum Exporting Countries) som utövar betydande inflytande på oljepriset genom att sätta en övre produktionsgräns för sina medlemmar. OPEC producerade cirka 42% av världens totala råolja från 2000 till 2014. OPEC-länderna behåller nästan hela världens oljeproduktionskapacitet och har nästan tre fjärdedelar av världens beräknade råoljereserver. Under de senaste åren har USA ökat oljeproduktion och detta har bidragit till att bromsa ökning av bensinpriset (U.S Energy Information Administration (EIA)).

1.2 Tidigare forskning

På grund av den dominerade ställningen som bensin och råolja har i energiförsörjning och den inverkan dessa har för ekonomisk utveckling, motiverar det en analys av dessa varors prisutveckling samt hur dessa eventuellt samverkar. Det finns

enormt många rapporter som fokuserar på relationen mellan bensin- och råoljepris. Dessa studier å ena sidan fokuserar på linjärt samband mellan bensinpris och andra faktorer, t.ex. råoljepris, KPI, hushållsinkomst m.m. å andra sidan studerar forskare korskorrelationen mellan flera relativa tidsserier. I en rapport "What Drives U.S. Gasoline Prices" [8] visade att ett linjärt samband mellan bensinpriset och råoljepriset är att enhetsökning i råoljepriset resulterar i en ökning motsvarande 0.024 dollar på bensinpriset. Gällande korskorrelationsstudier, kan vi hitta Lon-Mu Liu[2] studerade dynamisk relationen mellan U.S. bensinpriset, råoljepriset och aktien av bensin med hjälp av Box-Jenkins ARIMA metod, [3] använde ett fixt linjärt förhållande att prediktera bränslepriserna i Kalifonia, Afshin Honarvar [4] undersöker i sitt arbete från 2009 huruvida förhållandet mellan bensin och råolja är symmetriskt, dvs om effekten på bensinpriset är lika responsivt för prisökningar som för minskningar i råoljan. Lutz Kilian [5] studerar hur en s.k. "Structure VAR model" kan tillämpas till att undersöka variation i bensinpriset med förhållande av råoljepriset.

1.3 Studiemål

I den här uppsatsen är målsättningen att undersöka förhållande mellan veckovis bensinpris i USA och råoljepris så kallad "Brent-Europe". För att besvara frågeställningen använder vi oss en vektor autoregressiv modell (VAR) vilket i enkla ordlag kan beskrivas som en flerdimensionell autoregressiv modell.

2 Teori

I det här avsnittet kommer de flesta teorier och uttryck från bok "Analysis of Finance Time Series 3rd Edition" av RUEY S. TSAY 2010. Fetstil i texterna används för vektornotation.

2.1 Tidsserie, Stationäritet

TIDSSERIE

En *tidsserie* kan enkelt beskrivas som en sekvens data punkter som mäts i regelbundet i tiden. Det aktuella tidsintervallet mellan mätpunkterna kan exempelvis vara timvis, dagsvis eller veckovis. I den här uppsatsen kommer vi att fokusera på finansiella tidsserier. Vanliga exempel på dessa skulle kunna vara priset på en aktie eller en växlingskurs.

Finansmarknaden är inte isolerad och prISRörelser på en marknad kan lätt att sprida sig och påverka andra marknader. Det är denna insikt som medför att vid analys av finansiella tidsserier kan det vara klokt att fokusera på samspelet mellan flera tidsserier för att därigenom få en bättre förståelse om den eventuella dynamiska struktur dem emellan. Denna typ av studie kallas *multivariat tidsserie analys*. En multivariat tidsserie består av en vektor vars rader i sin tur utgörs av flera enkla tidsserier med samma tidsintervall (vilket medför att det har samma kolonndimension).

För tydlighetens skull, kan en multivariat tidsserie $\mathbf{y}_t = (\mathbf{y}_{1t}, \mathbf{y}_{2t}, \dots, \mathbf{y}_{kt})'$ som innehåller k endimensionella tidsserier av längd T skrivas i matrisform som:

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{y}_{1t} \\ \mathbf{y}_{2t} \\ \vdots \\ \mathbf{y}_{kt} \end{bmatrix} = \begin{bmatrix} y_{1,1} & y_{1,2} & y_{1,3} & \dots & y_{1,T} \\ y_{2,1} & y_{2,2} & y_{2,3} & \dots & y_{2,T} \\ \dots & \dots & \dots & \dots & \dots \\ y_{k,1} & y_{k,2} & y_{k,3} & \dots & y_{k,T} \end{bmatrix}$$

STATIONÄRITET

Ett fundamentalt begrepp inom tidsserieanalys är *stationäritet*. En tidsserie y_t sägs vara *strikt stationär* om dess simultana-fördelning är invariant över tid. Det är ett starkt villkor och i praktiken kan detta villkor vara svårt att verifiera (Tsay, sid. 30), vilket leder oss in på begreppet *svag stationäritet*. Svag stationäritet kräver att medelvärdet och kovariansen av en tidsserie är invarianta över tid, d.v.s. medelvärdet $E(y_t) = \mu$ är konstant och $\text{Cov}(y_t, y_{t-l}) = \gamma_l$ beror bara på avståndet l (eller kallas lag l) och inte av tiden t för $t \in \mathbf{Z}$.

En k -dimensionell tidsserie $\mathbf{y}_t = (\mathbf{y}_{1t}, \dots, \mathbf{y}_{kt})'$ är svagt stationär om dess medelvärdesvektor μ och kovariansmatris Γ_0 ej förändras över tid. Definitionen av μ och

Γ_0 presenteras mer utförligt i kommande avsnitt 2.2.

2.2 ACF och CCM

Definition och vidare beskrivning av "Autokorrelationsfunktion" (ACF) kan hittas i kapitel 2, sid. 30 - 32 i Tsay och motsvarande för "Korskorrelationsmatris" (eng. Cross-Correlation Matrices, CCM) kan läsas i kapitel 8, sid. 390-393. i Tsay

AUTOKORRELATIONSFUNKTION(ACF)

För en svagt stationär tidsserie y_t definieras autokorrelationsfunktionen mellan två tidpunkter t och $t-l$ som

$$\rho_l = \frac{\text{Cov}(y_t, y_{t-l})}{\sqrt{\text{Var}(y_t)\text{Var}(y_{t-l})}} = \frac{\text{Cov}(y_t, y_{t-l})}{\text{Var}(y_t)} = \frac{\gamma_l}{\gamma_0}. \quad (1)$$

Likheten mellan det andra och tredje uttrycket i (1) gäller på grund av egenskaper som följer av svagt stationäritet, d.v.s. $\text{Var}(y_t) = \text{Var}(y_{t-l})$. Speciellt är $\rho_0 = 1$, $\rho_l = \rho_{-l}$ och $-1 \leq \rho_l \leq 1$. För en okorrelerad tidsserie mellan t och $t-l$ följer det naturligtvis då att $\rho_l = 0$. För en given tidsserie kan vi skatta autokorrelationen mellan observationen i tid t och $t-l$ med hjälp av följande funktion

$$\hat{\rho}_l = \frac{\sum_{t=l+1}^T (y_t - \bar{y})(y_{t-l} - \bar{y})}{\sum_1^T (y_t - \bar{y})^2}, \quad 0 \leq l < T-1 \quad (2)$$

där $\bar{y}$ är det aritmetiska medelvärde av tidsserien och beräknas enligt $\bar{y} = \sum_1^T y_t / T$.

KORSKORRELATIONSMATRIS(CCM)

Det följer att för en k -dimensionell tidsserie studeras fler än en tidsserie samtidigt. Därmed undersöker vi korskorrelationsmatrisen (eng. Cross-Correlation Matrix (CCM)) vilket möjliggör analys av korrelation mellan olika tidsserier för olika lags. För en tidsserievektor $\mathbf{y}_t$, definierar vi medelvärdesvektor och kovariansmatris som

$$\boldsymbol{\mu} = E(\mathbf{y}_t), \quad \boldsymbol{\Gamma}_0 = E(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_t - \boldsymbol{\mu})'. \quad (3)$$

I den multivariata analysen fokuserar vi främst på relationen mellan alla komponenter i tidsserievektorn, dvs korrelationen mellan de olika serierna och mellan lag l . Detta medför att en lag- l korskovariansmatris (eng. cross-covariance matrix)

Γ_l kan definieras som

$$\Gamma_l \equiv [\Gamma_{ij}(l)] = E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-l} - \boldsymbol{\mu})']. \quad (4)$$

Korrelationskoefficienten mellan y_{it} och $y_{j,t-l}$ definieras som

$$\rho_{ij}(l) = \frac{\Gamma_{ij}(l)}{\sqrt{\Gamma_{ii}(0)\Gamma_{jj}(0)}} = \frac{\text{Cov}(y_{it}, y_{j,t-l})}{\sqrt{\text{Var}(y_{it})\text{Var}(y_{jt})}}. \quad (5)$$

Låt $\mathbf{D}$ vara en $k \times k$ diagonalmatris vars diagonal består av alla komponenternas standardavvikelser i $\mathbf{y}_t$, d.v.s. $\mathbf{D} = \text{diag}(\sqrt{\Gamma_{11}(0)}, \dots, \sqrt{\Gamma_{kk}(0)})$, då kan korskorrelationen mellan samtliga y_{it} och $y_{j,t-l}$, också kallas lag- l korskorrelationsmatris (CCM), skrivas som

$$\boldsymbol{\rho}_l \equiv [\rho_{ij}(l)] = \mathbf{D}^{-1} \Gamma_l \mathbf{D}^{-1} \quad (6)$$

Speciellt kan korrelationen mellan tidsserier i och j i $\mathbf{y}_t$ under samma tidpunkt t betecknas som

$$\rho_{i,j}(0) = \frac{\Gamma_{i,j}(0)}{\sqrt{\Gamma_{ii}(0)\Gamma_{jj}(0)}} = \frac{\text{Cov}(y_{it}, y_{jt})}{\sqrt{\text{Var}(y_{it})\text{Var}(y_{jt})}} \quad (7)$$

$$\boldsymbol{\rho}_0 = \mathbf{D}^{-1} \Gamma_0 \mathbf{D}^{-1}. \quad (8)$$

Den kallas lag-noll eller på engelska "concurrent cross-correlation" matris.

Från (5) – (8) vet vi att komponenterna i korskorrelationsmatrisen beskriver linjära samband mellan y_{it} och $y_{j,t-l}$ ($0 \leq l \leq t$) som inträffas i tidpunkt t . Om $\rho_{ij}(l) \neq 0$ för något $l > 0$, säger vi att serien y_{jt} *leder* (eng. lead) serien y_{it} i lag l . Analogt är $\rho_{ji}(l)$ ett linjärt sambandsmått mellan y_{jt} och $y_{i,t-l}$. Eftersom $\rho_{ij}(l)$ och $\rho_{ji}(l)$ för $i \neq j$ beskriver olika linjära samband mellan serierna $\{y_{it}\}$ och $\{y_{jt}\}$, i de flesta fall är $\rho_{ij}(l) \neq \rho_{ji}(l)$, d.v.s. korskorrelationsmatrisen $\boldsymbol{\rho}_l$ och korskovariansmatrisen Γ_l inte är symmetriska.

Enlig $\rho_{ij}(l)$ och $\rho_{ji}(l)$, innehåller korskorrelationsmatris $\{\boldsymbol{\rho}_l | l = 0, 1, \dots\}$ av en svagt stationär tidsserievektor följande information:

1. Komponenterna på diagonalen $\{\rho_{ii}(l) | l = 0, 1, \dots\}$ är autokorrelationsfunktion av y_{it} .

2. Off-diagonalelementet $\rho_{ij}(0)$ mäter samtidigt (eng. concurrent) linjärt samband mellan y_{it} och y_{jt} .
3. För $l > 0$, mäter den off-diagonalelementet $\rho_{ij}(l)$ linjärt samband mellan y_{it} och det förflutna värdet $y_{j,t-l}$.

Därför således om $\rho_{ij}(l) = 0$ för alla $l > 0$, då är tidsserien y_{it} oberoende av något förflutet värde $y_{j,t-l}$ i serien y_{jt} .

Generellt kan linjärt samband mellan två tidsserier $\{y_{it}\}$ och $\{y_{jt}\}$ sammanfattas som nedan:

1. y_{it} och y_{jt} är linjärt okorrelerade om $\rho_{ij}(l) = \rho_{ji}(l) = 0$ för alla $l \geq 0$.
2. y_{it} och y_{jt} är samverkande (eng. concurrently correlated) men inte *led-lag* förhållande (eng. lead-lag relationship) om olikheten $\rho_{ij}(l), \rho_{ji}(l) \neq 0$ gäller bara för $l = 0$.
3. y_{it} och y_{jt} är enkelriktad korrelerade (eng. unidirectional relationship) om $\rho_{ij}(l) = 0$ för alla $l > 0$ men $\rho_{ji}(v) \neq 0$ för något $v > 0$ eller tvärtom.
4. y_{it} och y_{jt} har återkopplingsförhållande, om $\rho_{ij}(l) \neq 0$ och $\rho_{ji}(v) \neq 0$ för något l och v .

För en given multivariat tidsserie kan vi skatta korskovariansmatrisen $\mathbf{\Gamma}_l$ och korskorrelationsmatrisen $\mathbf{\rho}_l$ enligt

$$\hat{\mathbf{\Gamma}}_l = \frac{1}{T} \sum_{t=l+1}^T (\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_{t-l} - \bar{\mathbf{y}})' \quad (9)$$

respektiv

$$\hat{\mathbf{\rho}}_l = \hat{\mathbf{D}}^{-1} \hat{\mathbf{\Gamma}}_l \hat{\mathbf{D}}^{-1} \quad (10)$$

där $\bar{\mathbf{y}}$ är medelvärdevektor och beräknas som $\bar{\mathbf{y}} = (\sum_1^T \mathbf{y}_t) / T$ och $\hat{\mathbf{D}}$ är en $k \times k$ diagonalmatris bestående av de skattade standardavvikelserna.

2.3 Ljung-Box-test

Det finns en mängd olika metoder som kan användas för att testa om residualerna är korrelerade. Ljung-Box-testet är ett diagnostiskt verktyg som används för att utvärdera huruvida modellen anpassningen är bra. Testet tillämpas genom att testa korrelationen av residualerna efter det att modellen applicerats på data. Teststatistikan $Q(m)$ kan formuleras enligt:

$$Q(m) = T(T+2) \sum_{k=1}^m \frac{\hat{r}_k^2(\epsilon_t^2)}{T-k} \quad (11)$$

där $\hat{r}_k$ betecknar den skattade autokorrelationskoefficienten av residualserien i lag k , och m är antalet lag som ska testas. Det testas med hypotesen $H_0: r_1 = \dots = r_m = 0$ mot $H_a: r_k \neq 0$ för något $k = 1, 2, \dots, m$. Under nollhypotesen är $Q(m)$ asymptotisk χ^2 fördelad med frihetsgrader m [7]. Välj ett antal lag (m) korrelation, om korrelationen $\hat{r}_1, \dots, \hat{r}_m$ är tillräcklig liten, d.v.s. de är inte signifikanta i en bestämd signifikansnivå, då ska vi inte förkasta nollhypotesen och drar vi slutsatsen att residualerna är okorrelerade och detta innebär att modellen med lag m redan har lyckats att fånga in alla korrelationer och modellen anpassar bra. Simuleringsstudier föreslår att valet av $m \approx \ln(T)$ ger bättre prestanda (sid. 33, Tsay).

2.4 VAR modell

Beskrivning av VAR modell kan hittas i kapitel 8 avsnitt 2 sid. 399 och 403 i Tsay.

2.4.1 VAR modell

Om vi ska prediktera en svagt stationär tidsserie som inte enbart är baserade på sin tidigare värden utan även andra svaga stationära tidsserier kan vi modellera dessa serier med hjälp av en så kallad *Vektor Autoregressiv (VAR) Modell*. VAR modellen utvidgas från en enkel autoregressiv modell genom att lägga till andra seriers historiska data som förklarande variabler. Till exempel, För en två-dimensionell tidsserie $\mathbf{y}_t = (y_{1t}, y_{2t})'$, kan en vektor autoregressiv modell med lag 1 (VAR(1)) beskrivas enligt:

$$y_{1t} = \Phi_{10} + \Phi_{11}y_{1,t-1} + \Phi_{12}y_{2,t-1} + \epsilon_{1t} \quad (12)$$

$$y_{2t} = \Phi_{20} + \Phi_{21}y_{1,t-1} + \Phi_{22}y_{2,t-1} + \epsilon_{2t} \quad (13)$$

där ϵ_{1t} och ϵ_{2t} betecknar modellens feltermen och antas vara okorrelerade. Denna kan allmänt skrivas på matrisform som

$$\mathbf{y}_t = \Phi_0 + \Phi_1 \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t. \quad (14)$$

Det är inte svårt att utvidga modellen genom att binda mer än två serier och mer än en lag. Generellt för en k -dimensionell tidsserie, definieras en vektor autoregressiv modell med lag p (VAR(p)) som

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{kt} \end{bmatrix} = \begin{bmatrix} \Phi_1^{(0)} \\ \Phi_2^{(0)} \\ \vdots \\ \Phi_k^{(0)} \end{bmatrix} + \begin{bmatrix} \Phi_{11}^{(1)} & \Phi_{12}^{(1)} & \Phi_{13}^{(1)} & \dots & \Phi_{1k}^{(1)} \\ \Phi_{21}^{(1)} & \Phi_{22}^{(1)} & \Phi_{23}^{(1)} & \dots & \Phi_{2k}^{(1)} \\ \dots & \dots & \dots & \dots & \dots \\ \Phi_{k1}^{(1)} & \Phi_{k2}^{(1)} & \Phi_{k3}^{(1)} & \dots & \Phi_{kk}^{(1)} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{k,t-1} \end{bmatrix} + \dots$$

$$+ \begin{bmatrix} \Phi_{11}^{(p)} & \Phi_{12}^{(p)} & \Phi_{13}^{(p)} & \dots & \Phi_{1k}^{(p)} \\ \Phi_{21}^{(p)} & \Phi_{22}^{(p)} & \Phi_{23}^{(p)} & \dots & \Phi_{2k}^{(p)} \\ \dots & \dots & \dots & \dots & \dots \\ \Phi_{k1}^{(p)} & \Phi_{k2}^{(p)} & \Phi_{k3}^{(p)} & \dots & \Phi_{kk}^{(p)} \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{k,t-p} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \vdots \\ \epsilon_{kt} \end{bmatrix}$$

Och i matrisnotation

$$\mathbf{y}_t = \Phi_0 + \Phi_1 \mathbf{y}_{t-1} + \dots + \Phi_p \mathbf{y}_{t-p} + \boldsymbol{\epsilon}_t \quad (15)$$

där $\mathbf{y}_t, \dots, \mathbf{y}_{t-p}$, Φ_0 och $\boldsymbol{\epsilon}_t$ är $k \times 1$ vektorer och $\Phi_1, \dots, \Phi_p$ är $k \times k$ matriser vars komponenter förklarar korskorrelation mellan serierna.

En grundläggande förutsättning till VAR(p) modell är att residualvektor $\boldsymbol{\epsilon}_t$ är vitt brus, d.v.s residualerna till varje ekvation ska vara oberoende och likfördelad stokastiska variabler med $E(\epsilon_{it}) = 0$ och varians σ_{it} . Men det är tillåtet att kovariansmatrisen eller korrelationsmatrisen till residualvektorn inte är diagonala. Det vill säga korrelation mellan residualerna från olika ekvationer föreligger.

För en given svagt stationär tidsserievektor, ska vi använda det iterativa förfarandet för lagspecifiering, parameterskattning och modellgranskning för att bygga en VAR modell.

2.4.2 Metod till lag urval: AIC och M(p)

Två mest användbara kriterier till att bestämma antalet lag är Akaike Informationskriterium (AIC) och χ^2 teststatistik M(p).

AKAIKE INFORMATIONSKRITERIUM (AIC)

Under förutsättningen att residualvektor i modellen är oberoende, likfördelad med väntevärdevektorn $\mathbf{0}$ och kovariansvektor $\boldsymbol{\Sigma}_t$, kan parametrarna i modellen skattas med hjälp av maximum likelihood (ML). Låt L beteckna den maximerade likelihood-funktionen till parametrarna, då definieras AIC som

$$AIC := 2k - 2\log L \quad (16)$$

där k är antal parametrar i modellen och $\log L$ är log-likelihoodfunktion. Modellen som har minst AIC-värde anses vara den bäst anpassade och därmed den som rekommenderas. AIC kriteriet kräver inte att residualerna måste vara normalfördelade och är därför ett användbart mått för vår fortsatta analys.

χ^2 TESTSTATISTIK M(p)

Ett annat kriterium till att identifiera antalet lag är att göra ett χ^2 test. För en godtycklig VAR(p) modell som anpassats till en k - dimensionell tidsserie med tidslängd T , kan residualvektorn skattas enligt

$$\hat{\boldsymbol{\epsilon}}_t = \mathbf{y}_t - \hat{\Phi}_0 - \hat{\Phi}_1 \mathbf{y}_{t-1} - \dots - \hat{\Phi}_{t-p} \mathbf{y}_{t-p} \quad (17)$$

och motsvarande kovariansmatris kan skattas som

$$\hat{\Sigma}_p = \frac{1}{T-2p-1} \sum_{t=p+1}^T \hat{\epsilon}_t^{(p)} (\hat{\epsilon}_t^{(p)})' \quad (18)$$

För att specificera lag p , kan vi formulera testhypotesen enligt : $H_0 : \Phi_l = \mathbf{0}$ som därmed ställs mot alternativhypotesen $H_a : \Phi_l \neq \mathbf{0}$ för $l = 1, 2, \dots$. Därefter kan vi använda $\hat{\Sigma}$ i modell p och $p-1$ för att konstruera teststatistiken $M(p)$ som används för att avgöra om vi kan förkasta testhypotesen eller inte. Matematiskt kan vi uttrycka $M(p)$ enligt

$$M(p) = -(T - k - p - \frac{3}{2}) \log\left(\frac{|\hat{\Sigma}_p|}{|\hat{\Sigma}_{p-1}|}\right) \quad (19)$$

där $|\hat{\Sigma}_p|$ betecknar determinanten av den skattade kovariansmatrisen i modell VAR(p). Under förutsättningen att residualerna är oberoende, likfördelade samt normalfördelade är denna teststatistika χ^2 fördelad med k^2 frihetsgrader. Vid en förutbestämd signifikansnivå α , om $p(M(p) \geq \chi_\alpha(k^2)) \leq \alpha$ innebär det att vi kan förkasta H_0 , dvs vi accepterar att modell med lag p annars väljer vi modell med lag $p-1$.

2.4.3 Metod till parameter skattning: OLS och ML

MINSTA-KVADRATMETODEN (OLS)

Minsta-kvadratmetoden(OLS) användas bland annat vid regressionsanalys för minimera felet i en funktion som ska anpassas utifrån observerade värden [15]. För en VAR(p) modell till en k -dimensionell tidsserie $\mathbf{y}_t$, kan modellens parametrar skattas med hjälp av den OLS metoden. Den kallas skattning till multivariat linjär regression i multivariat statistisk analys[14]. Låt oss anta n oberoende observationer. För dessa observationer kan en multipel linjär regression modell ansättas, som har följande form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (20)$$

där $\mathbf{Y}$ och $\boldsymbol{\epsilon}$ är $n \times 1$ vektorer, $\mathbf{X}$ är en $n \times (r+1)$ matris och $\boldsymbol{\beta}$ är en $(r+1) \times 1$ vektor. Att skatta parametervektorn $\boldsymbol{\beta}$ med hjälp av OLS är det samma som att hitta lösningen som minimerar kvadratsumman av residualerna $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Lös ekvationen till första derivatan av funktionen kan vi få $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Denna skattningsmetod kallas Minsta-kvadratmetoden. OLS skattningen som resulterar utifrån detta är en konsistent och väntevärdesriktig skattning. Notera att om residualerna är normalfördelade, överensstämmer skattningarna som kan beräknas via OLS med de som kan beräknas med hjälp av så kallad maximum likelihood metod (ML).

MAXIMUM LIKELIHOOD METOD (ML)

För en given tidsserie y_t , vars betingade fördelning som kan uttryckas enligt $f(y_t|y_{t-1}, \dots, y_1; \boldsymbol{\theta})$ är normalfördelad med väntevärde μ_t och varians σ_t^2 , kan likelihoodfunktion till data uttryckas som

$$f(y_1, \dots, y_t; \boldsymbol{\theta}) = f(y_1, \boldsymbol{\theta}) \prod_{t=2}^T \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left[-\frac{(y_t - \mu_t)^2}{2\sigma_t^2}\right]. \quad (21)$$

där $f(y_1; \boldsymbol{\theta})$ betecknar den marginella täthetsfunktionen för den första observationen y_1 . De värden till $\boldsymbol{\theta}$ som maximerar den likelihoodfunktionen kallas maximum likelihood skattare (MLE) av $\boldsymbol{\theta}$. Eftersom log-funktion är en monotont växande funktion, kan MLE fås av att maximera motsvarande log-likelihoodfunktion

$$\log(f(y_1, \dots, y_t; \boldsymbol{\theta})) = \log(f(y_1, \boldsymbol{\theta})) - \frac{1}{2} \sum_{t=2}^T [\log(2\pi) + \log(\sigma_t^2) + \frac{(y_t - \mu_t)^2}{\sigma_t^2}] \quad (22)$$

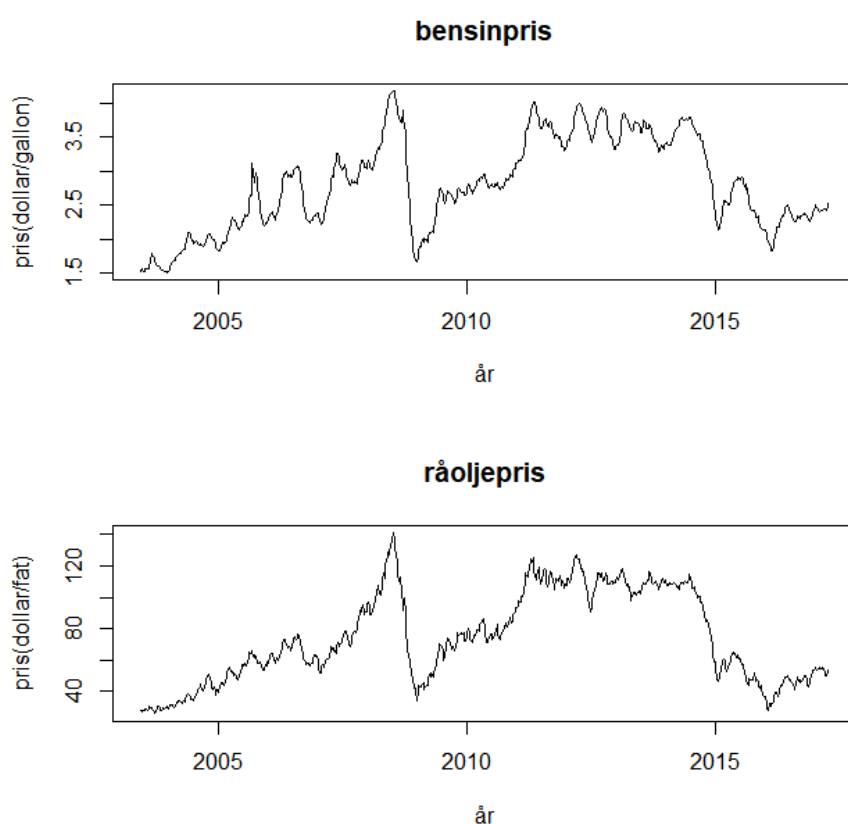
som beräkningsmässigt är lättare att hantera.

3 Data

3.1 Databeskrivning

Som vi nämnde tidigare bestäms bensinpriset i USA av fyra komponenter: 1) råoljepris, 2) raffineringkostnader, 3) återförsäljning- och transportkostnader och 4) skatt. De tredje och fjärde komponenterna utgör detaljhandelssegmentet och tenderar att vara relativt stabil.

Figur 1: Plottar av veckovisa bensin- och råoljepris från 9 juni 2003 till 10 april 2017. Bensinpriset i enhet \$/gallon, råoljepris i enhet \$/fat.



För att analysera hur bensinpriset förändrar över tiden, ska vi fokusera på de första två komponenterna som står för det mesta av variationen på bensinpriset. Vi ska begränsa oss i denna uppsats att inte ta den andra komponenten som studieobjekt. Anledning till detta är att raffineringkostnader skiljer sig mellan regioner i USA och det kommer att vara stor variation och av liten betydelse om vår studie baserar på nationell nivå. Det finns två stora internationella spotpriser på råolja: Brent

Crude oil och West Texas Intermediate (WTI). I en rapport av EIA [8] visade sig att Brent är mer signifikant än WTI för bestämning av bensinpris i USA, varför vi väljer spotpris av Brent Crude oil som data till råolja.

Data i denna studie omfattar veckovisa bensin- och råoljepriser mellan 9 juni 2003 och 10 april 2017 och sträcker sig över totalt 723 veckor. Bensinpriset beräknas som priset av bensin med alla kvaliteter och alla typer i enhet \$/gallon. För data över råoljepris har vi valt data så kallad "Brent - Europe", med tillhörande enhet \$/fat (1 fat motsvarar 31.5 gallon). Bensinpriset har stickprovsmedelvärde 2.81 och standardavvikelse 0.679 och råoljas stickprovsmedelvärde och standardavvikelse är 74.12 respektive 28.80. Om vi jämför de två tidsserierna i figur 1 ser vi grovt att spridningen mellan bensinpriserna och råoljepriserna är relativt liten, vilket innebär att vi kan förvänta oss en positiv korrelation. Vid beräkning av korrelationskoefficient får vi värdet 0.578 vilket därmed inte är förvånande. Den höga korrelationskoefficienten kan vara en indikator att en vektor autogressiv modell skulle kunna tillämpas.

3.2 Datahantering

Tidsserie av bensinpris och råoljepris har inte samma enhet och stickprovsstandardavvikelserna är också stora, så vi ska transformera priserna till logavkastningar r_{it} för $i = 1, 2$. Logavkastning fås genom att räkna $\log(\frac{y_t}{y_{t-1}})$. En fördel med att göra om priserna till logavkastningar är att dessa är skalfria, vilket betyder att logavkastningen inte beror på någon enhet. Vi betecknar logavkastningstidsserien över bensinpris som r_{1t} och logavkastningstidsserien över råoljepris som r_{2t} . För enkelhetens skull kommer vi i fortsättningen att kalla r_{1t} för **bensin** och r_{2t} för **råolja** i denna studie. (Plottar av logavkastningstidsserier finns i figur 7 i Appendix).

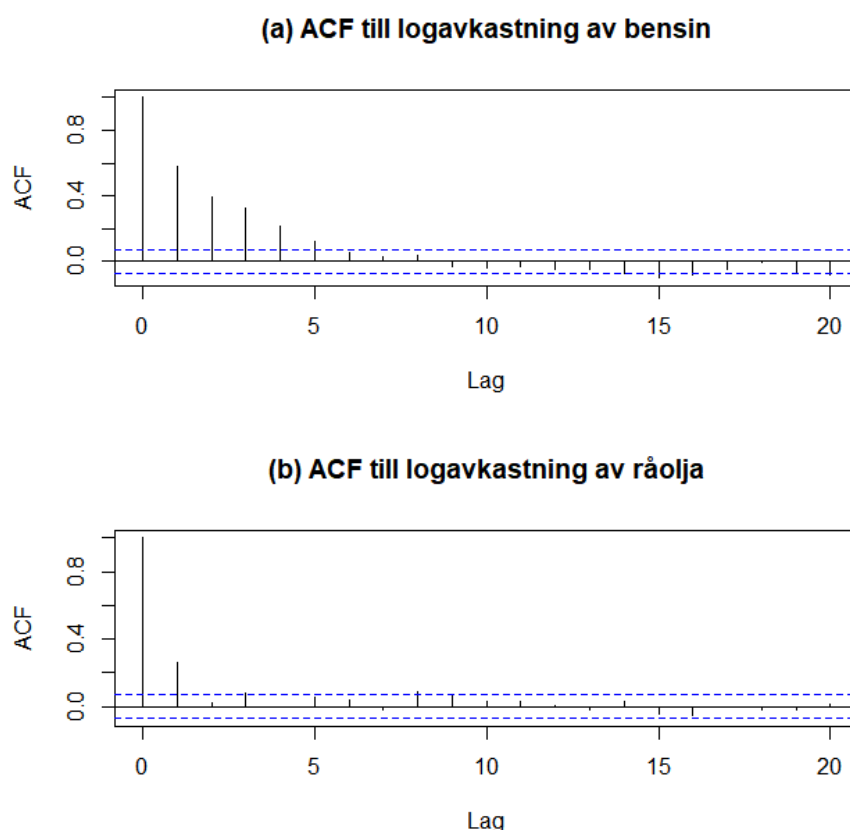
Utöver att endast identifiera den bäst anpassande modellen vill vi även utvärdera modellens prediktionsförmåga. För att göra det väljer vi att använda de första 500 tidpunkterna till modellanpassning (tidsperioder: 2003-06-09 - 2013-01-02) och resterande 223 tidpunkter (tidsperioder: 2013-01-07 - 2017-04-10) till att bedöma modellens prediktionsförmåga. Stickprovsmedelvärdet och standardavvikelsen för de första 500 behandlade tidpunkterna för bensin resp. råolja är 0.0015, 0.0237 respektive 0.00277 och 0.040, kovariansen för de två tidsserierna är 0.0006 och korrelationskoefficienten har beräknat till 0.592.

4 Metod och Resultat

4.1 Undersökning av stationäritet

För att data ska kunna tillämpas till en tidsseriemodell behöver vi först kontrollera om tidsserien har icke-stationära komponenter, t.ex säsongsbeteende, kalenderefekter m.m.. I figur 7 "plottar av logavkastningstidsserier" ser det inte synlig trend eller periodicitet. Vi vill styrka detta genom att även studera autokorrelationsfunktionen (ACF). Autokorrelationsfunktion till bensin och råolja visas i figur 2. Det visar sig att ACF till bensin sjunker exponentiellt till det icke-signifikant intervallet då lag m ökar och inga signifikanta säsong- eller kalenderstrender föreligger, så att bensin kan antas vara en svagt stationär tidsserie. För ACF till råolja är bara den första autokorrelationskoefficienten signifikant, således kan råolja också antas vara svagt stationär.

Figur 2: Autokorrelationsfunktion till bensin och råolja

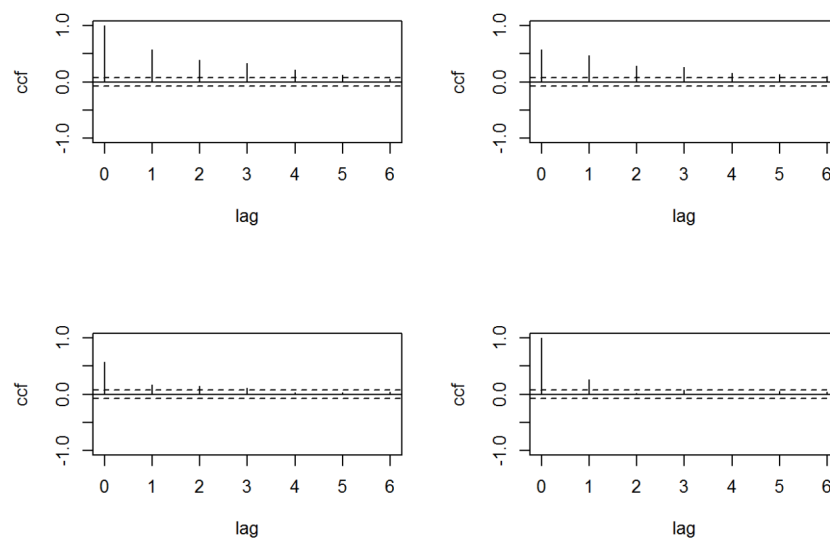


4.2 Korskorrelationsundersökning

För att upptäcka vilket förhållande bensin och råolja har och hur de två tidsserierna påverkas av deras historiska värden ska vi studera korskorrelationsmatrisen.

För att återkoppla till diskussionen i avsnitt 2.2 gällande korskorrelationsmatriser, minns vi att för en bivariat tidsserie kan det finnas olika förhållande status mellan två tidsserier. Figur 3 visar stickprovskorskorrelationsfunktionen mellan bensin och råolja. Ovan till vänster i figuren visas autokorrelationsfunktionen för bensin, vilket är samma bild som visas i figur 2 (a). Motsvarande ACF för råolja visas till höger i den nedre delen av figuren. Ovan till höger är korskorrelationsfunktion mellan bensin och råolja, från vilket det visar sig att korrelationskoefficienterna $\rho_{12}(1)$, $\rho_{12}(2)$, $\rho_{12}(3)$, $\rho_{12}(4)$ och $\rho_{12}(5)$ är signifikant nollskilda. Detta innebär att tidsserien "bensin" har ett *led*-lag förhållande gentemot råolja i lag 1, 2, 3, 4 och 5.

Figur 3: Stickprovsauto- och korskorrelationsfunktion (CCF) av bensin och råolja: (a) ACF av bensin (ovan till vänster); (b) korskorrelation mellan bensin och råolja (ovan till höger); (c) korskorrelation mellan råolja och bensin (nere till vänster); (d) ACF av råolja. Streckade linjer anger 95% konfidensintervall.



Nere till vänster i figur 3 representerar korskorrelationsfunktion mellan råolja och bensin. Vi ser att korskorrelationen mellan råolja och bensin inte blir detsamma som korskorrelationen mellan bensin och råolja. Det visar sig att korrelationskoefficienter mellan råolja och bensin i lag 1, 2 och 3 är signifikant nollskilda, vilken betyder att råolja r_{2t} enbart korrelerar med bensins historiska värden $r_{1,t-1}$, $r_{1,t-2}$ samt $r_{1,t-3}$. Enligt definition om olika typ linjärt samband av två tidsserier konsta-

Tabell 1: Korskorrelationsmatris (a) och teckentabell (b) av bensin och råolja: i var och en matris betyder siffran i plats (1,1) autokorrelationskoefficienten av bensin; (1,2) korskorrelationskoefficienten mellan bensin och råolja; (2,1) korskorrelationskoefficienten mellan råolja och bensin och (2,2) autokorrelationskoefficienten av råolja. "+": signifikant positiv korrelerad; "-": signifikant negativ korrelerad; ".": icke signifikant korrelerad; signifikansgräns: $\pm 2/\sqrt{T} \approx \pm 0.09$

<i>Korskorrelationsmatris (a)</i>											
lag1		lag2		lag3		lag4		lag5		lag6	
0.58	0.47	0.39	0.29	0.33	0.26	0.22	0.16	0.12	0.13	0.05	0.08
0.17	0.26	0.14	0.02	0.11	0.08	0.03	-0.00	0.03	0.05	0.04	0.04

<i>teckenvisning (b)</i>											
+	+	+	+	+	+	+	+	+	+	.	.
+	+	+	.	+	.	.	.	.	.	.	.

terar vi att bensin och råolja har ett återkopplingsförhållande men de har inte samma "led-lag". Råolja *leder* bensin i lag 1, 2, 3, 4 samt 5 medan bensin har förhållande med råolja endast i lag 1, 2 samt 3.

Korskorrelationsmatriser kan vara ett annat alternativt studieobjekt i en korskorrelationsundersökning för en multivariat tidsserie. Dessa är hjälpsamma eftersom det medför att vi kan kvantifiera korrelationen av två tidsserier. I tabell 1(a) presenterar vi korrelationsmatriser upp till lag 6. Varje lag motsvaras av en 2×2 matris med fyra komponenter. Till exempel, 0.58 i lag 1 är korrelationskoefficient mellan $r_{1,t}$ och $r_{1,t-1}$ (är också autokorrelationskoefficient av bensin i lag 1). Värdet på plats 1,2 (0.47) och 2,1 (0.17) i lag 1 motsvarar varsin korskorrelationskoefficient $\rho_{12}(1)$ och $\rho_{21}(1)$ och autokorrelationskoefficient för råolja är 0.26.

Tabell 1(b) visar en förenklad korskorrelationsmatris som i sin tur kan lätt identifiera de signifikanta korskorrelationerna. Signifikansgränsen beräknas enligt $|\pm 2/\sqrt{T}|$ (sid. 394 i Tsay) [10]. Om korrelationskoefficienten är större (mindre) eller lika med $2/\sqrt{T}$ ($-2/\sqrt{T}$), i vårt fall $\approx \pm 0.09$, betecknar vi det som "+" (" -").

Från tabell 1(a) upptäcker vi:

- uppenbarligen är matriserna inte symmetriska, vilket innebär en icke-symmetrisk förhållande mellan de två tidsserierna.
- beroende av bensin på råolja är starkare än råolja på bensin ty $\rho_{12}(l) > \rho_{21}(l)$ för $l = 1, \dots, 6$;
- bensins historiska data är mer inflytelserika till att beräkna bensin i jämförelse med förhållande mellan råoljas historiska data och råolja.

Läs av teckenvisningen i tabell 1(b) kan vi direkt få slutsats att alla signifikanta korskorrelationer är positiva, vilket innebär att de två tidsserierna interagerar i samma riktning.

4.3 Lagspecifiering och parameterskattning

En modell anses vara den mest välanpassande om den lyckas fånga in alla korrelationer, d.v.s. residualerna från modellen är då okorrelerade och kan likna vitt brus. Vidare är det fördelaktigt att finna en så enkel modell som möjligt och som därmed har minimalt antal lags. Det motiverar frågan: hur många lags anses vara lämpligt? Vid denna frågeställning behöver vi specificera antalet lags med hjälp av vissa bedömningskriterier. I vår rapport väljer vi AIC och $M(p)$ som underlag för våra beslut gällande antalet lags. I tabell 2 har vi samlat dessa värden där de fetstilta siffrorna representerar de minsta värdena.

Enligt AIC-kriteriet sker det minst AIC-värdet i lag 3, så om vi använder AIC-värde som urvalskriterium bör vi därmed välja en VAR(3) modell. Om vi jämföra p-värde från χ^2 -teststatistiken $M(p)$, erbjuder två modellkandidater beroende på vilken signifikansnivå vi använder, en VAR(3) modell rekommenderas när vi sätter signifikansnivå 1% medan en VAR(8) modell föreslås om vi sätter istället signifikansnivå 5%. I så fall bestämmer vi att kombinera två resultaten och då ska en VAR(3) modell väljas.

Tabell 2: Lag urvals resultat

lag	AIC	p(M(p))
0	-14.4708	0
1	-14.9900	0
2	-15.0221	0
3	-15.0402	0.004
4	-15.0358	0.3231
5	-15.0355	0.1131
6	-15.0306	0.3673
7	-15.0262	0.6646
8	-8.8731	0.0414

Parametrarna i den specificerade VAR(3)-modellen skattas med OLS metod och dessa vi har tur kan lätt fås med hjälp av ett R-paket "VAR" (inbyggt i RStudio). Parameterskattningarna för VAR(3) visas till vänster i tabell 3 (a) och (b). VAR(3) innehåller två karakteristiska ekvationer som beskriver de linjära relationerna till varsin tidsserie ((a) bensin och (b) råolja i tabellen). Varje ekvation består av sju paramet-

rar varav en konstant och sex lutningskoefficienter. Parameter "Bensin.11" betyder exempelvis $r_{1,t-1}$ i modellen, "råolja.11" betecknar $r_{2,t-1}$, o.s.v.

Tabell 3: Parameterskattning för VAR(3) och VAR(2)

(a) bensin

	VAR(3)			VAR(2)		
Parameter	Skattning	Std. avvikelse	Signif. nivå	Skattning	Std. avvikelse	Signif. nivå
Konstant	0.0003	0.0008		0.0005	0.0008	
Bensin.11	0.3280	0.0553	***	0.3455	0.0554	***
Råolja.11	0.1413	0.0274	***	0.1397	0.0273	***
Bensin.12	0.0933	0.0586		0.1824	0.0505	***
Råolja.12	-0.0134	0.0279		-0.0313	0.0269	
Bensin.13	0.0952	0.0511				
Råolja.13	0.0345	0.0270				
Residual Std. avvikelse	0.0174			0.0176		
Justerad R^2	0.3644			0.3550		

(b) Råolja

	VAR(3)			VAR(2)		
Parameter	Skattning	Std. avvikelse	Signif. nivå	Skattning	Std. avvikelse	Signif. nivå
Konstant	0.002	0.002		0.002	0.002	
Bensin.11	-0.052	0.113		-0.054	0.112	
Råolja.11	0.258	0.056	***	0.249	0.055	***
Bensin.12	0.405	0.119	***	0.388	0.102	***
Råolja.12	-0.170	0.057	***	-0.158	0.054	***
Bensin.13	-0.085	0.104				
Råolja.13	0.05	0.055				
Residual Std. avvikelse	0.0385			0.0385		
Justerad R^2	0.0777			0.0785		

Notera: *** betyder att parametern är signifikant nollskild vid 5% signifikansnivå.

Vi upptäcker från tabell 3 att inte alla parametrarna i VAR(3) är signifikant nollskilda vid 5% nivå. Till exempel har ekvationen till bensin i VAR(3) (i tabell 3 (a)) bara två signifikant nollskilda parametrar, nämligen "Bensin.11" och "Råolja.11" som ligger i lag 1 och ekvationen till råolja i sin tur har tre signifikant nollskilda parametrar, "Råolja.11" i lag 1 samt "Bensin.12" och "Råolja.12" i lag 2, därmed har inga para-

metrar i lag 3 i varsin ekvation signifikant effekt. Detta leder oss att misstänka om parametrarna i lag 3 går att klippa bort. Med andra ord, vi skulle kunna undersöka om en VAR(2) modell kan vara en reducerad men lämplig modell.

Innan vi jämför dessa två modeller, ska vi utvärdera om VAR(2) är godtagbar, d.v.s. om modellen lyckas fånga in alla korrelationerna för vardera serie. Detta löses genom att studera autokorrelationen för residualerna. ACF-plot av residualerna av varsin ekvation presenteras i figur 4. Figurerna ser bra ut i den bemärkelsen att få stolpar sträcker sig utanför konfidensbanden. Detta innebär att residualerna kan antas vara okorrelerade och modellen anses fånga beroendet mellan observationerna i tisserievektorn. För att styrka denna slutsats kan vi även göra ett Ljung-Box-test av residualerna. Vid signifikansnivån 5% får vi p-värde 0.9011 till ekvationen av bensin resp. 0.7266 till ekvationen av råolja som bekräftar att modellen anpassar data bra.

För att avgöra om VAR(2) kan beskriva data bättre än VAR(3), ska vi göra en antal jämförelser. Först jämför vi parameterskattningarna i vardera modell. Vi upptäcker att de skattade parametrarna från de två modellerna är snarlika, (se exempelvis parametern "bensin.11" i tabell 3 (a)). Den parametern med störst skillnad mellan modellerna är "Bensin.12" i ekvationen för bensin. Parametern är inte signifikant nollskild i VAR(3) men i fallet för VAR(2)-modellen är den signifikant på 5%. Ytterligare upptäcker vi att de justerade förklaringsgraderna i VAR(2) (0.3644 resp. 0.0777) inte har förändrat stort i jämförelse med justerade förklaringsgraderna i VAR(3) (0.3550 resp. 0.0785). Om vi jämför AIC-värdena för de två modellerna, är skillnaden liten. (-15.0221 för VAR(2) och -15.0402 för VAR(3)). Sammanfattningsvis visar detta på att VAR(2) inte beskriver data sämre än VAR(3), alltså godtar vi VAR(2).

Den specificerade VAR(2)-modellen med de skattade parametrarna kan skrivas i matrisform enligt:

$$\begin{bmatrix} r_{1t} \\ r_{2t} \end{bmatrix} = \begin{bmatrix} 0.00046 \\ 0.00213 \end{bmatrix} + \begin{bmatrix} 0.3455 & 0.1397 \\ -0.0540 & 0.2487 \end{bmatrix} \begin{bmatrix} r_{1,t-1} \\ r_{2,t-1} \end{bmatrix} + \begin{bmatrix} 0.1824 & -0.0313 \\ 0.3876 & -0.1580 \end{bmatrix} \begin{bmatrix} r_{1,t-2} \\ r_{2,t-2} \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix}$$

Korrelationsmatris av e_{1t} och e_{2t} är :

$$\begin{bmatrix} 1.000 & 0.587 \\ 0.587 & 1.000 \end{bmatrix}$$

4.4 Modellgranskning

Från den identifierade VAR(2)-modellen får vi följande observationer: (a) den samtidiga korrelationskoefficienten mellan e_{1t} och e_{2t} är 0.587 som är nära den stickprovskorrelationskoefficienten mellan serierna r_{1t} och r_{2t} (0.592). (b) Båda logavkastningstidsserier har marginellt positiva medelvärden (0.0015 resp. 0.0028) som

ger de två serierna milt uppåtgående trender vid tidsförskjutningarna. (c) Om vi omskriver modellen med endast de signifikant nollskilda koefficienterna, får vi :

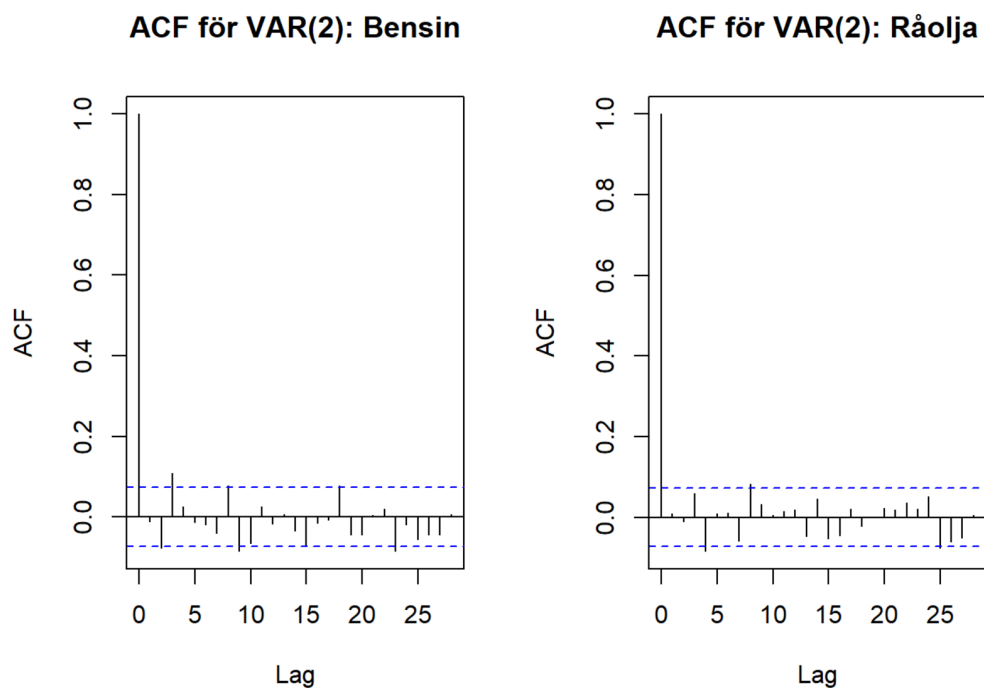
$$r_{1t} = \hat{\Phi}_{11}^{(1)} r_{1,t-1} + \hat{\Phi}_{12}^{(1)} r_{2,t-1} + \hat{\Phi}_{11}^{(2)} r_{1,t-2} + e_{1t} \quad (23)$$

$$r_{2t} = \hat{\Phi}_{22}^{(1)} r_{2,t-1} + \hat{\Phi}_{21}^{(2)} r_{1,t-2} + \hat{\Phi}_{22}^{(2)} r_{2,t-2} + e_{2t}. \quad (24)$$

Från ekvationer (23) och (24) konkluderar vi att bensin och råolja har ett återkopplingsförhållande men inte reagerar i samma led-lag. Enligt ekvationen (23), bestäms veckovis bensin av sina två senaste veckornas värden samt det sista veckans värdet av råolja medan enligt (24) påverkas veckovis råolja av sina två senaste veckornas värden samt den näst senaste mätpunkten av bensin.

Vi märker att den justerade förklaringsgraden till ekvationen för råoljan är väldigt låg (0.079), detta vittnar om att data över historiska bensinpriser i USA inte är en viktig förklarande faktor för råolja trots att korrelation av dem är hög. Därför är det inte meningsfullt att studera råolja med den tillämpade VAR(2)-modellen.

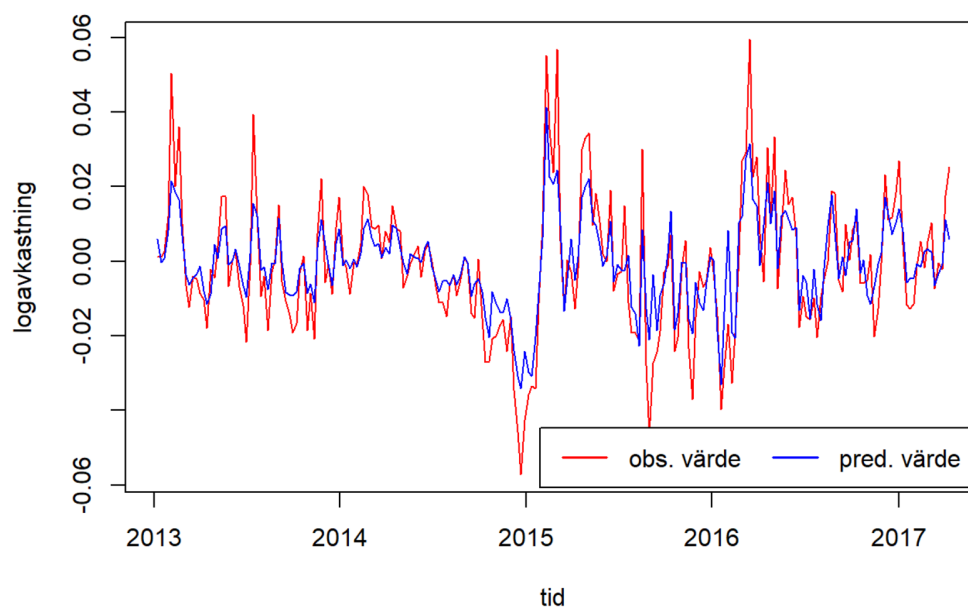
Figur 4: ACF av residualer av VAR(2)



4.5 Prediktion

Vid den ekvationen för bensin i den identifierade VAR(2)-modellen kan vi prediktera logavkastningar av bensin. För en VAR (p) modell, beräknas 1-steps-prognos enligt $\mathbf{r}_{t+1} = \Phi_0 + \sum_{i=1}^p \Phi_i \mathbf{r}_{t+1-i}$. Med hjälp av RStudio predikterar vi logavkastningarna för bensin under tidsperioden: 2013-01-07 - 2017-04-10. Dessa predikterade värden kommer sedan att jämföras med de faktiska för att undersöka modellens prediktionsförmåga. I figur 5 illustrerar den röda linjen de observerade värdena och blåa linjen de predikterade resultaten. Från figuren finner vi det värt att nämna tre observationer: (a) predikterade logavkastningarna skapar ett mindre förväntat rörelseområde. (b) modellen har en bättre prediktionsförmåga för observationer som inte är vändpunkter, det illustreras tydligt då de största observerade skillnaderna (mellan predikterat och faktiskt värde) är vid vändpunkter. (c) Modellens prediktionsförmåga för mer volatila perioder är förhållandevis låg.

Figur 5: Predikterade och observerade veckovisa logavkastningar av bensin: 2013-01-07 - 2017-04-10



En användbar metod för att jämföra prediktionsförmåga mellan olika modeller är att beräkna och jämföra *kvadratroten ur medelkvadratfelet* (eng. root mean square error (RMSE)). Detta kan beräknas enligt

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^T (r_t - \hat{r}_t)^2} \quad (25)$$

där r_t betecknar den observerade logavkastningen och $\hat{r}_t$ betecknar den predikterade logavkastningen. Ett mindre RMSE innebär ett mindre avstånd mellan predikterat och faktiskt värde och därmed en bättre prediktionsförmåga.

Med hjälp av RMSE kan vi jämföra prediktionsförmågan inte bara mellan VAR(2) och VAR(3) utan även mellan en VAR modell och en enkel autoregressiv (AR) modell. En enkel autoregressiv modell med lag p (AR(p)) är en linjär funktion som endast inkluderar sina historiska data fram till och med lag p som förklarande variabler. För tydlighets skull kan en AR(2) modell skrivas enligt:

$$r_t = \Phi_0 + \Phi_1 r_{t-1} + \Phi_2 r_{t-2} + \epsilon_t. \quad (26)$$

medan en vektor autoregressiv modell av ordning 2 (VAR(2)) modell kan skrivas enligt:

$$r_{1t} = \Phi_0 + \Phi_{11}^{(1)} r_{1,t-1} + \Phi_{12}^{(1)} r_{2,t-1} + \Phi_{11}^{(2)} r_{1,t-2} + \Phi_{12}^{(2)} r_{2,t-2} + \epsilon_t. \quad (27)$$

Poängen med att jämföra RMSE mellan våra VAR modeller och motsvarande AR modell är att detta kan hjälpa oss att svara på om en modell som tar in andra tidsserier som förklarande variabler kan prediktera responsen bättre. I vårt fall om RMSE till ekvation för bensin i VAR-modellen är lägre än den motsvarande AR-modellen för bensin innebär det att VAR modellen har bättre prediktionsförmåga för bensin, d.v.s. genom att inkludera råoljas historiska data i modellen erhåller vi bättre prediktioner av bensinpriset relativt en modell som endast inkluderar historiska data över bensin.

Tabell 4: RMSE till VAR modell och AR modell

lag	VAR modell	AR modell
2	0.00094	0.00169
3	0.00096	0.00103

Så vi ska beräkna RMSE till fyra modellkandidater: VAR(2), VAR(3), AR(2) och AR(3). Resultatet samlas vi i tabell 4. Det visar sig att VAR(2)-modellen har marginellt lägre RMSE än VAR(3), vilket innebär att VAR(2) har något bättre prediktionsförmåga än VAR(3). Om vi jämför RMSE mellan respektive VAR modell och motsvarande AR modell, har båda VAR modellerna lägre värde än motsvarande AR modell. Detta innebär att råoljas historiska data bidrar till att prediktera bensin bättre. Av de fyra

modellerna har VAR(2) det minsta RMSE-värdet vilket medför att VAR(2) predikterar bensin bäst.

Vi har även beräknat prisprognoser under år 2017 till och med vecka 15 (2017-4-10) som finns i tabell 5 i Appedix. I tabellen visar även prisskillnaden mellan predikterat värdet och faktiskt värdet.

5 Slutsats och Diskussion

Syftet med den här uppsatsen är att undersöka det eventuella förhållandet mellan de veckovisa amerikanska bensin- och råoljepriserna så kallad "Brent - Europa" med hjälp av en vektor autoregressiv process. För att undersöka vilken VAR modell som är att föredra vad gäller anpassning till data såväl som vid prediktion, delar vi tidsintervallet i två perioder. Den första bestående av 500 tidpunkter inom perioden 2003-06-09 - 2013-01-02 använder vi till att specificera en vektor autoregressiv modell och de resterande 223 tidpunkterna för perioden 2013-01-07 - 2017-04-10 används för att undersöka modellens prediktionsförmåga.

En förutsättning för att användandet av en vektor autoregressiv modell är att det ska existera korskorrelation mellan komponenterna i tidsserievektorn. För att undersöka detta använder vi oss av så kallade korskorrelationsfunktioner och korskorrelationsmatriser. Resultaten visar att bensin och råolja har ett återkopplingsförhållande men att dem inte reagerar i samma "led-lag". Som vi tidigare nämnt så avser vi logavkastningstidsserien för bensinpris då vi skriver "bensin" och motsvarande serie för "råolja". Resultaten för korskorrelationsmatrisundersökningen i avsnitt 4.2 visar att vid signifikansgräns ≈ 0.90 , ser vi att råolja *leder* bensin i lag 1, 2, 3, 4 samt 5 medan råolja har signifikant förhållande med bensin enbart i lag 1, 2 samt 3.

Vidare använder vi oss av AIC-kriterium och χ^2 statistikan $M(p)$ som lagspecificeringskriterier. Detta resulterar i att en VAR(3) anses vara en lämplig modell enligt AIC-kriteriet och vid χ^2 statistikan $M(p)$ får vi två modellkandidater beroende på vilken signifikansnivå vi använder. d.v.s. en VAR(3) modell rekommenderas när vi använder oss av signifikansnivå 1% medan en VAR(8) modell föreslås om vi sätter signifikansnivå 5%. På så sätt kombinerar vi resultaten och väljer en VAR(3) modell. Parametrarna skattas med hjälp av OLS metoden. Men när vi studerar skattade parametrarna av VAR(3)-modellen, upptäcker vi att vissa lags parametrar inte är signifikant nollskilda, vilket leder vi oss till att undersöka om modellen kan reduceras till en VAR(2). För att VAR(2) kunna anses vara en lämplig modell, först kontrollerar vi om denna modell kan beskriva alla beroende relationer av data. Med hjälp av ACF-plot och Ljung-Box-testet beslutar vi att VAR(2) anses vara en lämplig modell. Vid ett antal jämförelser t.ex. att jämföra parametrarna, förklaringsgraderna och AIC-värden mellan två modellerna godtar vi slutligen den VAR(2)-modellen.

Från den identifierade VAR(2)-modellen observerar vi att residualerna mellan två ekvationer är korrelerade med samtidig korrelationskoefficient 0.587 som är nära den stickprovskorrelationskoefficienten av två tidsserierna (0.592). Det visar sig också att bensin och råolja har ett återkopplingsförhållande men påverkas inte i samma lags av deras historiska dataserier. Om vi endast tittar på de signifikant nollskilda parametrarna i modellen kan vi konstatera att bensin påverkas av de två senaste datapunkterna i samma serie och den senaste punkten i motsvarande råoljeserien (se ekvation (23)) medan råolja i sin tur beror av de två senaste mätpunkterna

i samma serie och den näst senaste mätpunkten i bensinserien (se ekvation (24)).

Vi märker också att den justerade förklaringsgraden till ekvationen för råoljan är låg (0.079), detta vittnar om att data över historiska bensinpriser i USA inte är en viktig förklarande faktor för råolja. Från tidigare forskningsrapporter som undersökt vilka faktorer som påverkar råolja [11], [12] vet vi att råoljepriset påverkas väsentligen makroekonomiska faktorer, t.ex. ekonomisk utveckling runt om världen, tillgång och efterfrågan om råolja och däremot har bensinpris i USA troligen liten betydelse till prissprognoser för råoljan. Det kan möjligen vara en förklaring att varför förklaringsgraden till råoljas ekvation i modellen är så låg. I så fall bestämmer vi att inte göra vidare prisprognoser för råoljan via den identifierade VAR(2)-modellen. Emellertid har vi predikerat ett-steps-prognos för bensin under tidsperioder 2013-01-07 - 2017-04-10. Predikerade priserna rör sig i ett mer konserverat intervall, d.v.s. prediktionsintervallet är precis som förväntat mindre än det faktiskt logavkastningsintervallet. Om vi tittar på de "outliers" (de volatila perioderna) är prisprognoserna lång ifrån de faktiska värdena. En möjlig förklaring till detta är att vi inte har involverat alla möjliga förklaringsfaktorer i en VAR modell, vilka möjligt kan vara t.ex. dollarkurs, hushållsinkomst i USA, andra sorters energiutveckling m.fl.

I den här rapporten har vi också beräknat kvadratroten ur medelkvadratfelet (eng. root mean square error (RMSE)) till fyra modeller: VAR(2), VAR(3), AR(2) och AR(3). Syftet med dessa beräkningar är att å ena sidan undersöka vilken av de två jämförande VAR modellerna har bättre prediktionsförmåga, å andra sidan studera om VAR modellen kan prediktera bensin bättre än AR modellen, d.v.s. genom att inkludera råoljas historiska data i modellen erhåller vi bättre prediktioner av bensinpriset relativt en modell som endast inkluderar historiska data över bensin. Resultaten visar att VAR(2) har marginellt lägre RMSE-värde än VAR(3), vilket innebär att VAR(2) predikterar bensinpriset något bättre än VAR(3). Så sammanfattningsvis kan vi dra slutsats att vid en vektor autoregressiv process är VAR(2) en tillräcklig modell, inte bara baserat på anpassning, utan även på prediktion.

Om vi jämför RMSE mellan respektive VAR modell och motsvarande AR modell, har båda VAR modellerna lägre värde än motsvarande AR modell. Detta innebär att råoljas historiska data bidrar till att prediktera bensin bättre. Av de fyra modellerna har VAR(2) det minsta RMSE-värdet vilket medför att VAR(2) predikterar bensin bäst.

I denna studie har vi inte beräknat prediktionsintervall utan endast den förväntade ett-steps-prognosen. Anledning till detta är att residualerna i modellen inte anses att vara normalfördelade (se Appendix figur 6 QQ-plot) och således kan inte övre- och nedrekvantilen direkt beräknas. Därmed har vi inte heller provat att göra t-fördelningsanpassning. Denna studie skulle kunna vara en vidare studie.

Lon-Mu Liu [2] demonstrerar att "outlier" (som allmänt påpekar de extrema värdena) medför problem till tidsseriens modellidentifiering samt påverkar prispro-

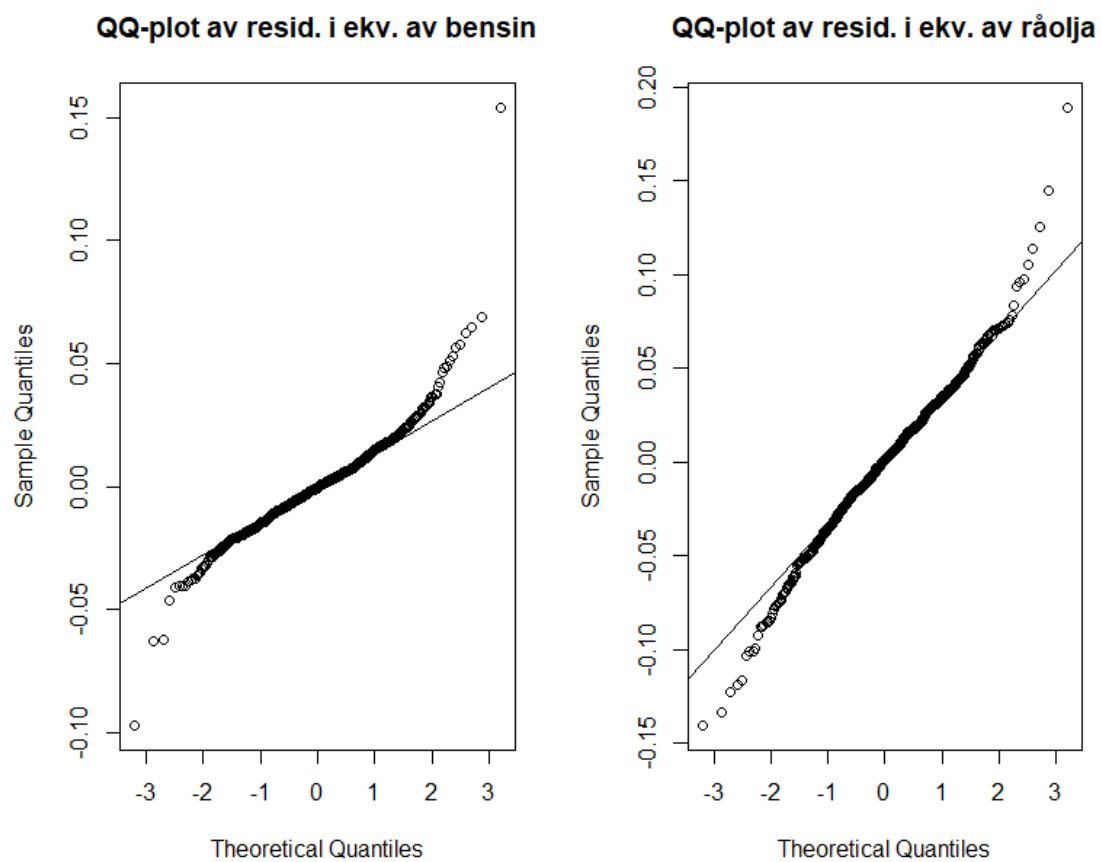
gnos och prediktionsintervall. I vår datas tidsintervall inkluderar bl.a. den ekonomisk krisen i 2008 som antas i flesta analysen en stor volatil period och denna härleder ett väntevärdesfel av de skattade parametrarna och ytterligare medför s.k. "carry-over effect" på prediktionen[13]. Dessa undersökningar har vi inte gått genom i denna studien. Att jämföra modellens beskrivningsförmåga för data med "outlier" och utan "outlier" kan också vara ett intressant ämne.

6 Appendix

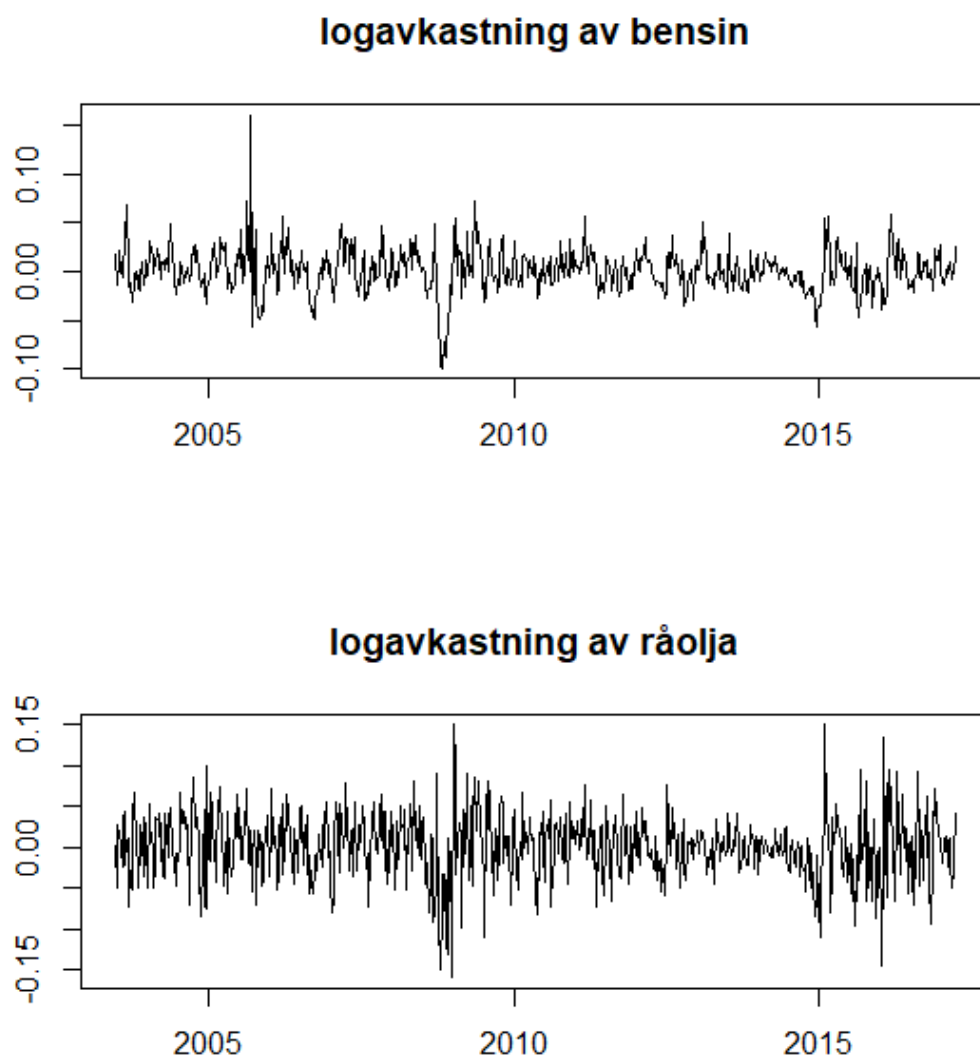
Tabell 5: Predikterade priser vs. faktiska priser (2017-01-02 - 2017-04-10)

Vecka	Dag	Obs. pris	Pred. pris	Skillnad
1	2017-01-02	2.485	2.520	-0.035
2	2017-01-09	2.496	2.516	-0.020
3	2017-01-16	2.467	2.453	0.014
4	2017-01-23	2.436	2.425	0.011
5	2017-01-30	2.408	2.397	0.011
6	2017-02-06	2.405	2.403	0.002
7	2017-02-13	2.418	2.414	0.004
8	2017-02-20	2.414	2.420	-0.006
9	2017-02-27	2.427	2.435	-0.008
10	2017-03-06	2.452	2.458	-0.006
11	2017-03-13	2.434	2.418	0.016
12	2017-03-20	2.433	2.425	0.008
13	2017-03-27	2.428	2.427	0.001
14	2017-04-03	2.471	2.499	-0.028
15	2017-04-10	2.534	2.549	-0.015

Figur 6: QQ-plot av residualer i VAR(2)-modellen



Figur 7: Plottar av logavkastningsserier



7 Referens

- [1] *Välkommen till oljans värld: från källa till konsument.*
Svenska petroleum institutet (SPI), 2004
spbi.se
- [2] Lon-Mu Liu, *Dynamic relationship analysis of U.S. gasoline and crude oil prices.*
Journal of FORECASTING, Vol 10, Issue 5, Sept. 1991. sid. 521-547
- [3] California Energy Commission, *TRANSPORTATION FUEL PRICE AND DEMAND FORECASTS: Inputs and Methods for the 2009 Integrated Energy Policy Report*
<http://www.energy.ca.gov/2009publications/>
- [4] Afshin Honarvar, *Asymmetry in retail gasoline and crude oil price movements in the United States: An application of hidden cointegration technique*
ScienceDirect, Energy Economics, Volum 31, problem 3, maj 2009, sid. 395-402
- [5] Lutz Kilian, *Explaining Fluctuations in Gasoline Prices: A Joint Model of the Global Crude Oil Market and the U.S. Retail Gasoline Market*
University of Michigan and CEPR, 2 sep 2009
<http://www-personal.umich.edu/~lkilian/ej090309final.pdf>
- [6] Wayne A. Fuller, *Introduction to Statistical Time Series, 2nd Edition*, avsnitt 8
John Wiley & Sons, Inc. (1996)
- [7] Anderson T.W., Walker A.M., *On the Asymptotic Distribution of the Autocorrelations of a Sample From a Linear Stochastic Process.* Annals of Mathematical Statistics (1964), 35, 3, sid. 1296-1303.
- [8] U.S. Department of Energy, *What Drives U.S. Gasoline Prices?*
Independent Statistics & Analysis (2014), Eneergy Information Administration
<https://www.eia.gov/analysis/studies/gasoline/>
- [9] Tsay R.S., *Analysis of Financial Time Series 3rd ed.*
John Wiley & Sons, Inc. (2010). avsnitt 2.2, sid 31
- [10] Tiao, George C., Bisgaard, Søren, Box, George E. P., *Box on quality and discovery: with design, control and robustness*
Wiley, c2000.

- [11] Ron Alquist, Lutz Kilian, and Robert J. Vigfusson, *Forecasting the Price of Oil*. Board of Governors of the Federal Reserve System, juli, 2011, No.1022
www.federalreserve.gov/pubs/ifdp/
- [12] CRAIG ADEYANJU, *Top Factors that Move the Price of Brent Crude Oil*
www.futuresknowledge.com/news-and-analysis/energy/
- [13] Chung Chen, Lon-Mu Liu, *Forecasting time series with outliers*
Journal of FORECASTING, Vol 12, Issue 1, Jan. 1993. sid. 13-35
- [14] Archard A. Johnson, Dean W. Wichern, *Applied Multivariate Statistical Analysis, sixth edition*
©2007 Pearson Education, Inc. Upper Saddle River, NJ 07458. sid. 364
- [15] Patrik Andersson, Joanna Tyrcha, *Notes in Econometrics*
Stockholms universitet, Matematiska institutionen, Avdelningen för Matematisk Statistik, Tryckning Mars 2016. sid. 17