BOKMÅL

NHH

HJEMMEEKSAMEN MET4

Vår 2020

Dato: 15. mai 2020

Tidsrom: 09:00 - 13:00

Antall timer: 4

BESVARELSEN SKAL LEVERES I WISEFLOW

På våre nettsider finner du informasjon om hvordan du leverer din besvarelse: https://www.nhh.no/for-studenter/eksamen/innlevering-individuelt-og-i-gruppe/

Kandidatnummer blir oppgitt på StudentWeb i god tid før innlevering. Kandidatnummer skal være påført på alle sider øverst i høyre hjørne (ikke navn eller studentnummer). Ved gruppeinnlevering skal alle gruppemedlemmers kandidatnummer påføres.

Samarbeid mellom individer eller grupper om utarbeidelse er ikke tillatt, og utveksling av egenprodusert materiale til andre individer eller grupper skal ikke forekomme. En besvarelse skal bestå av individets, eller gruppens egne vurderinger og analyse. All kommunikasjon under hjemmeeksamen er å anse som fusk. Alle innleverte oppgaver blir behandlet i Urkund, NHHs datasystem for tekst- og plagiatkontroll

UTFYLLENDE BESTEMMELSER OM EKSAMEN

https://www.nhh.no/globalassets/for-studenter/forskrifter/utfyllende-bestemmelser-til-forskrift-om-fulltidsstudiene-ved-nhh.pdf

Antall sider, inkludert forside: 10

Antall vedlegg: 5 (Alle vedlegg følger etter oppgavene)

Oppgave 1

I forbindelse med kommunevalgkampen i 2019 var det fokus på utslipp av klimagasser. To av regjeringspartiene, Høyre og Venstre, ønsket å formidle budskapet om at Norges klimagassutslipp hadde gått ned under deres regjeringstid, og publiserte de to grafene som er vist i **Vedlegg 1a** og **1b**.

Begge disse figurene ble kritisert for å vere *misvisende*. På hvilke(n) måte(r) er de det? Beskriv kort med ord hvordan du tenker at figurene kunne vært mindre misvisende.

Svar: Grafen i vedlegg 1a er misvisende fordi tre år, 2014–2016, er trukket samme til ett år, slik at nedgangen i klimagassutslipp ser mye brattere ut enn det den faktisk er. Hvis man trekker ut x-aksen slik at det er like langt mellom hver år, vil man se at det er en jevn nedgang fra 2012 til 2018. Faktisk så var det i følge SSB et hopp i utslipp i de tre årene som er utelatt fra figuren, les mer her:

`https://www.faktisk.no/artikler/Yo8/pa-hoyres-klimagraf-blir-fire-ar-til-ett`

Grafen i vedlegg 1b (som viser det lille "hoppet" som er utelatt fra forrige figur) ser vi har en kraftig trunkert y-akse. Det gir inntrykk av at nedgangen i klimagassutslipp er markant, mens hvis vi hadde latt y-aksen begynne på null, så ville utviklingen sett ut til å være nærmest helt flat. Dette er et godt eksempel på at vi ved å velge forskjellige grafiske fremstillingsmåte kan legge vekt på ulike aspekter ved datasettet. Med denne figuren ønsker nok avsenderen å rette oppmerksomheten mot en *nedgang*, mens en figur som inkluderer origo legger vekt på *nivået*, og at den relativt sett er ganske stabilt.

Oppgave 2

Lokal luftforurensing er et stort problem i mange byområder. Som ledd i en større kartlegging av luftkvaliteten i en stor europeisk by har myndighetene satt opp en sensor langs en travel innfartsåre som måler en rekke parametre hver time, deriblant konsentrasjonen av nitrogendioksid (NO_2). Dette er en gass som i store doser kan føre til svekket lungefunksjon og forverring av astma og bronkitt.

Vi skal i denne oppgaven undersøke data fra denne måleren, og ser på den gjennomsnittlige daglige konsentrasjonen av NO_2, målt over en periode på 391 dager. I første omgang ønsker vi å se om det er forskjell i forventet NO_2-konsentrasjon mellom helgedager (lørdag og søndag) og ukedager (mandag til fredag). I tabellen under finner vi en deskriptiv statistikk for målingene fordelt på de to kategoriene. Måleenheten er mikrogram per kubikkmeter ($\mu g/m^3$).

	Gj. snitt	St. avvik	Median	Min	Max	N
Ukedager	116.8	31.0	113.2	43.6	223.2	279
Helg	99.0	32.6	91.8	38.0	215.4	112

a) Test om variansen til NO_2-konsentrasjonen er lik mellom ukedager og helgedager.

Svar: Vi skal teste $H_0 : \sigma_1^2 = \sigma_2^2$ mot $H_1 : \sigma_1^2 \neq \sigma_2^2$. Vi setter den største variansen øvesrt i brøken, slik at testobservatoren blir

$$F = \frac{S_2^2}{S_1^2} = \frac{32.6^2}{31.0^2} = 1.11.$$

Testobservatoren er under nullhypotesen F-fordelt med $112 - 1 = 111$ og $279 - 1 = 278$ frihetsgrader, som gir en kritisk verdi på $F_{0.025}^{111,278} \approx F_{0.025}^{120,200} = 1.37$. Siden $F < F_{0.025}^{200,100}$ kan vi *ikke* forkaste nullhypotesen om lik varians.

b) Test om forventet NO_2-konsentrasjon er lik mellom ukedager og helgedager.

Svar: Ut fra resultatet i forrige oppgave bruker vi en t-test for to populasjoner med *lik* varians for å teste $H_0 : \mu_1 = \mu_2$ mot $H_1 : \mu_1 \neq \mu_2$. Vi beregner først variansen til differansen av gjennomsnittene:

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(279 - 1) \cdot 31.0^2 + (112 - 1) \cdot 32.6^2}{279 + 112 - 2} = 990.0$$

Testobservatoren for en to-utvalgs t-test er gitt ved

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{116.8 - 99.0}{\sqrt{990.0 \cdot \left(\frac{1}{279} + \frac{1}{112}\right)}} = 5.06,$$

som under nullhypotesen er t-fordelt med $279 + 112 - 2 = 389$ frihetsgrader. Kritisk verdi for en tosidig t-test på 5% nivå med 389 frihetsgrader er 1.96, så vi forkaster nullhypotesen med klar margin.

c) Hvilke forutsetninger gjør vi for å gjennomføre testene i spørsmål a) og b)? Bruk informasjonen du har tilgjengelig til å vurdere om forutsetningene er oppfylt.

Svar: For å gjøre testene i spørsmål a) og b) forutsetter vi normalfordelte og uavhengige observasjoner. Fra tabellen med deskriptiv statitsikk kan vi se at gjennomsnitt og median ligner på hverandre, så fordelingene er i det minste rimelig symmetriske. Uansett har vi såpass mange observsajoner at gjennomsnittene som følge av sentralgenseteoremet er tilnærmet normalfordelte.

Observasjonene er gjort etter hverandre i tid. Det betyr at vi neppe kan anta uavhengighet, da vi kan forvente at det ulike trender (som værforhold og ukesyklusen) har betydning for flere dager etter hverandre.

Dersom konsentrasjonen av NO_2 overstiger 100 $\mu g/m^3$ er det ikke anbefalt at barn, eldre, eller personer med lungesykdommer oppholder seg utendørs i lengre perioder, og myndighetene må utstede et såkalt gult farevarsel. Vi ønsker videre å analysere om det er systematiske forskjeller også mellom ukedagene. Vi får oppgitt kontingenstabellen under, som viser antall dager med gult farevarsel, fordelt på de forskjellige ukedagene.

Ukedag	Antall dager med gult farevarsel	Antall dager uten gult farevarsel
Mandag	31	20
Tirsdag	34	17
Onsdag	38	15
Torsdag	36	16
Fredag	42	9

d) Test om kjennetegnene "gult farevarsel" og "ukedag" er uavhengige. Hva betyr resultatet i praksis? Du får oppgitt at testobservatoren i den aktuelle testen er gitt ved $\chi^2 = 6.14$ (Du trenger altså ikke skrive opp hele utregningen av testobservatoren, det holder at du viser hvordan det kan gjøres).

Svar: I denne oppgaven får vi oppgitt verdien av testobservatoren, så det holder å vise utregningen av denne med symboler, eventuelt at de første par leddene er gitt med tallverdier. Skal man gi den fullstendige utregningen må vi ha de observerte marginale frekvensene, og dernest frekvensene vi forventer ved uavhengighet. Den tabellen ser slik ut, som vi har begynt å fylle inn:

	Med farevarsel	Uten farevarsel	Sum
Mandag	$51 \cdot 181/258 = 35.78$	$51 \cdot 77/258 = 15.22$	\$51
Tirsdag	$\vdots$	$\vdots$	51
Onsdag			53
Torsdag			52
Fredag			51
Sum	181	77	258

Testobservatoren i χ^2-testen er gitt ved

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = \frac{(31 - 35.78)^2}{35.78} + \frac{(20 - 15.22)^2}{15.22} + \cdots = 6.14,$$

som altså er oppgitt i oppgaveteksten. Under nullhypotesen er testobservatoren χ^2-fordelt med $(r-1)(s-1) = (2-1)(5-1) = 4$ frihetsgrader, så kritisk verdi på 5% signifikansnivå er 9.49. Vi kan dermed *ikke* forkaste nullhypotesen om uavhengighet mellom ukedag og farevarsel.

For å få full pott her må studenten tydelig vise at han/hun vet hvordan dette skal regnes ut! Det står eksplisitt i oppgaveteksten. De som bare skriver opp at $6.14 < 9.49$, så ikke forkast, får maksimalt 3/10 poeng for å finne rett verdi i tabellen.

For å bedre forstå hvilke faktorer som forklarer variasjon i NO_2-konsentrasjon setter vi opp en regresjonsmodell basert på et datasett med følgende forklaringsvariabler:

Variabel	Forklaring
`WeekdayMonday - WeekdaySunday`	Dummyer som angir dag
`Temperature`	Daglig gjennomsnittlig lufttemperatur (°C)
`Humidity`	Daglig gjennomsnittlig relativ luftfuktighet (%-poeng)
`Winter`	Dummy som tar verdien 1 fra oktober t.o.m. mars, 0 ellers

I **Vedlegg 2** finner du en lineær regresjonsmodell med NO_2-konsentrasjonen som responsvariabel i kolonne (1).

e) For en gitt årstid, temperatur og luftfuktighet, hvilken ukedag har i følge den estimerte regresjonsmodellen høyest forventet NO_2-konsentrasjon?

Svar: Vi ser at der er dummyvariabelen for *fredag* som er valgt som referansekategori i denne regresjonsmodellen, og at alle andre ukedagdummyer er negative. Det betyr at alle ukedager estimeres til å ha et lavere nivå av NO_2 enn fredag, og at altså fredag er den dagen som estimeres til å ha det høyeste forventet NO_2-nivå.

f) Gi en *kortfattet* fortolkning av regresjonsmodell (1) i Vedlegg 2.

Svar: Som nevnt i forrige oppgave estimeres alle ukedagene bortsett fra onsdag og torsdag til å ha statistisk signifikant et lavere nivå av NO_2 enn fredag. Av disse er det lørdag og søndag som har det største forventede avviket på hhv -19 og -36 $\mu g/m^3$, som er naturlig siden vi tidligere har etablert at det er en klar forskjell mellom uke- og helgedager. Videre ser vi at varmere dager og dager med høyere luftfuktighet har lavere forventet NO_2. I tillegg til temperaturkoeffisienten, er også koeffisienten til dummyvariabelen som indikerer om vi er i vinterhalvåret statistisk signifikant. Den estimerer at forventet NO_2-nivå i vinterhalvåret er 18 $\mu g/m^3$ høyere enn i sommerhalvåret, selv når vi har kontrollert for temperatur.

Regresjonsmodellen gir en rimielig god tilpasning til observasjonene, da justert R^2 er på 0.416, som betyr at den forklarer 41.6% av variasjonen i NO_2-nivået.

g) Bruk figurene i Vedlegg 3 til å diagnostisere regresjonsmodell (1) i Vedlegg 2. Skriv kortfattet.

Svar: Residualplottet viser antydninger til heteroskedastisitet, ved at variansen til residualene ser ut til å øke med predikert verdi. Videre ser vi at det er autokorrelasjon i residualserien, som er naturlig siden vi har tidsrekkedata. Begge disse forholdene gjør at vi ikke bør tolke den statistiske inferensen i forrige oppgave for bokstavelig. Verken histogrammet eller QQ-plottet viser særlig avvik fra normalitetsantakelsen.

Vi tilpasser en ny regresjonsmodell med de samme forklaringsvariablene, men denne gangen bruker vi logistisk regresjon, og som responsvariabel bruker vi dummyvariabelen `danger_warning`, som indikerer om gjenomsnittskonsentrasjonen av NO_2 den aktuelle dagen oversteg 100 $\mu g/m^3$, og at myndighetene derfor måtte utstede gult farevarsel. Den estimerte

modellen er gitt i kolonne (2) i **Vedlegg 2**.

h) På hvilken måte gir den logistiske regresjonen et annet bilde enn resultatet vi fikk i oppgave d)? Hvordan forklarer du det?

Svar: I oppgave d) kunne vi ikke forkaste nullhypotesen om at ukedag og farenivå er uavhengige. I den logistiske regresjonen i kolonne (2) i vedlegg 2 ser det ut til å være forskjeller mellom referansekategorien fredag, og ukedagene mandag og tirsdag. Dersom vi hadde valgt en annen referansedag ville vi kanskje sett avvik mellom andre dager også.

Dette kan vi forklare ved at den logistiske regresjonen kontrollerer for flere forhold enn bare ukedag. I tillegg til å ta med de to helgedagene som egne kategorier, kontrollerer vi også for værforhold og årstid. Det betyr at den uforklarte variasjonen i responsvariabelen (dummy for farevarsel) er blitt *mindre*, slik at eventuelle forskjeller mellom dagene trer klarere frem.

Videre så har vi i denne oppgaven bruk en *statistisk modell* til å forklare sammenhengen mellom ukedag og farenivå, i motsetning til spørsmål d) der vi kun ser på om det er statistisk avhengighet eller ikke. Når vi antar at sammenhengen tar en bestemt form (f.eks at den kan beskrives av logistisk regresjon) så er også det i seg selv med på å forklare den observerte variasjonen. Hvis modellen er korrekt, eller tilnærmet korrekt, er det en god ting. Hvis modellen er feil er det selvsagt ikke bra, siden vi da kan bli lurt til å tro at det er en sammenheng som ikke er reell. **(Dette siste poenget er kanskje litt subtilt, og ikke nødvendig for full pott).**

Myndighetene må bestemme seg for om de skal utstede farevarsel et døgn i forveien. I morgen er det lørdag 16. mai, og i den aktuelle byen er det meldt en gjennomsnittlig temperatur på 19 °C og en gjennomsnittlig relativ luftfuktighet på 47%.

i) Bruk den logistiske regresjonsmodellen til å predikere *sannsynligheten* for at gjennomsnittlig NO_2-konsentrasjon overstiger 100 $\mu g/m^3$. Gi en kort vurdering om myndighetene bør utstede gult farevarsel. (Husk at luftfuktigheten er gitt på skala 0–100, og ikke 0–1)

Svar: Den predikerte log-oddsen får vi ved å sette inn for variablene (lørdag, temperatur, fuktighet, vinterdummyen er null):

$$z = 5.052 + -2.292 - 0.086 * 19 - 0.044 * 47 = -0.942.$$

Den predikerte *sannsynligheten* er gitt ved følgende sammenheng:

$$P(Y = 1|Z = z) = \frac{e^z}{1 + e^z} = \frac{e^{-0.942}}{1 + e^{-0.942}} \approx 0.28.$$

Den predikerte sannsynligheten klart under 50%, som passer godt med den tidlighere analysen vår. Det er snakk om en forholdsvis varm lørdag i sommerhalvåret, og vi vil nok ikke utstede farevarsel.

Det kan også være gode argumenter for at vi ikke nødvendigvis bruker 50% som terskel for farevarsel. Kanskje er det mer alvorlig å *ikke* utstede et farevarsel som burde vært sendt ut fordi det kan være farlig for folk, enn å utstede et unødvendig farevarsel. Føre var osv., og det kan tilsi at vi f.eks. bruker 40% eller 30% sannsynlighet som grense. Det kommer litt an på situasjonen, som vi ikke har full oversikt over her.

j) Svar spørsmål i) ved å benytte den lineære regresjonsmodellen i Vedlegg 2, kolonne (1) i stedet. Se bort fra usikkerhet knyttet til estimering av regresjonskoeffisientene når du svarer på dette spørsmålet.

Svar: Modell (1) i vedlegg 2 forklarer NO_2-nivået direkte, og vi predikerer følgende nivå for i morgen:

$$Y = 172.893 - 19.239 - 1.787 \cdot 19 - 0.517 \cdot 47 = 95.402.$$

I følge modellen vår vil NO_2-nivået i morgen være en trekning fra en normalfordeling med forventning 95.402 of standardavvik 24.640 (se regresjonsutskriften). Vi regner ut snnsynligheten for å overstige et nivå på 100 ved å finne til tihørende sannsynligheten i tabell for standard normalfordeling:

$$\begin{aligned} P(Y > 100) &= P\left(\frac{Y - 95.402}{24.64} > \frac{100 - 95.402}{24.64}\right) \\ &= P(Z > 0.19) = 1 - P(Z < 0.19) = 1 - 0.5753 = 0.42 \end{aligned}$$

Denne sannsynligheten er en del større enn den vi fant ved å bruke logistisk regresjon. Forklaringen på at disse to størrelsene er ulike er at vi har brukt to forskjellige modeller til å regne dem ut, og det vil nok være en god idé å gjøre en nærmere analyse av hvilken modell som passer best med virkeligheten.

Oppgave 3

Vi tenker oss at konsentrasjonen av NO_2 ved tidspunkt t kan skrives på følgende måte:

$$\mathrm{NO2}_t = T_t + S_t + R_t,$$

der T_t er en trendkomponent, S_t er en sesongkomponent og R_t er residualserien, altså det som ikke fanges opp av trend- og sesongkomponentene. I **Vedlegg 4** ser vi et plott av $\mathrm{NO2}_t$, samt estimerte trend- og sesongkomponenter, og et autokorrelasjonsplott for residualtidsrekken.

a) Forklar *kort* hva vi lærer av å se på de estimerte trend- og sesongkomponentene.

Svar: Trendkomponenten viser tydelig at NO_2-nivået er lavere i sommerhalvåret enn i vinterhalvåret. Sesongkomponenten har en periode på 7 dager, og inneholder en ukentlig svingning på ca 30 $\mu\mathrm{g/m}^3$.

I **Vedlegg 5** har vi tilpasset to ulike tidrekkemodeller til residualtidsrekken R_t.

b) Hvilke to tidsrekkemodeller har vi tilpasset? Hvilken av de to modellene passer best til datasettet? Begrunn svaret, både ved hjelp av utskriftene i Vedlegg 5 og en av figurene i Vedlegg 4.

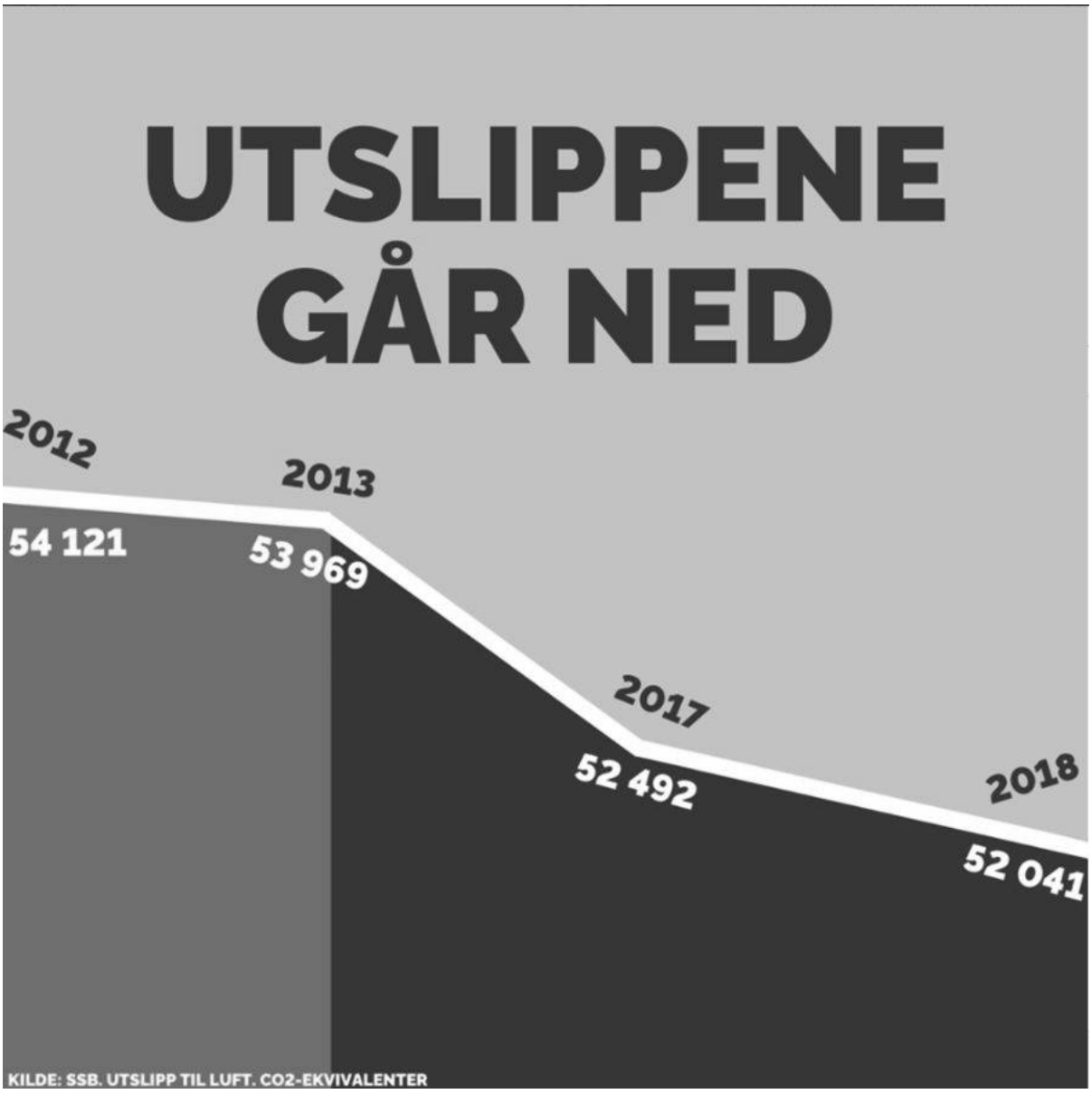
Svar: I de to utskriftene ser vi at vi har tilpasset en AR(1)- og en MA(1)-modell. Vi ser fra utskriftene at det er AR(1)-modellen som passr best, for den har lavest AIC og lavest prediksjonsfeil. Dette kan vi også begrunne ut fra autokorrelasjonsplottet i vedlegg 4, som ligner mer på en AR(1)-prosess (som har en eksponensielt avtagende autokorrelasjonsfunksjon) enn en MA(1)-prosess (der autokorrelasjonen er null for lag 2 og oppover).

Anta at den beste modellen fra spørsmål **b)** representerer den *sanne* modellen for residualtidsrekken R_t.

c) Er $\mathbf{R_t}$ stasjonær? Begrunn svaret.

Svar: Koeffisienten i denne AR(1)-prosessen er $\phi = 0.6245$. En AR(1)-prossess er stasjonær dersom $|\phi| < 1$, altså er dette en stasjonær tidsrekke.

Vedlegg 1a: Graf publisert av Høyre



Bakgrunn: Denne grafen ble publisert av Høyre på Facebook 1. november 2019, og viser norske CO_2-utslipp (i 1000 tonn CO_2-ekvivalenter) som funksjon av tid. Høyre overtok regjeringsmakten sammen med Fremskrittspartiet etter stortingsvalget i 2013.

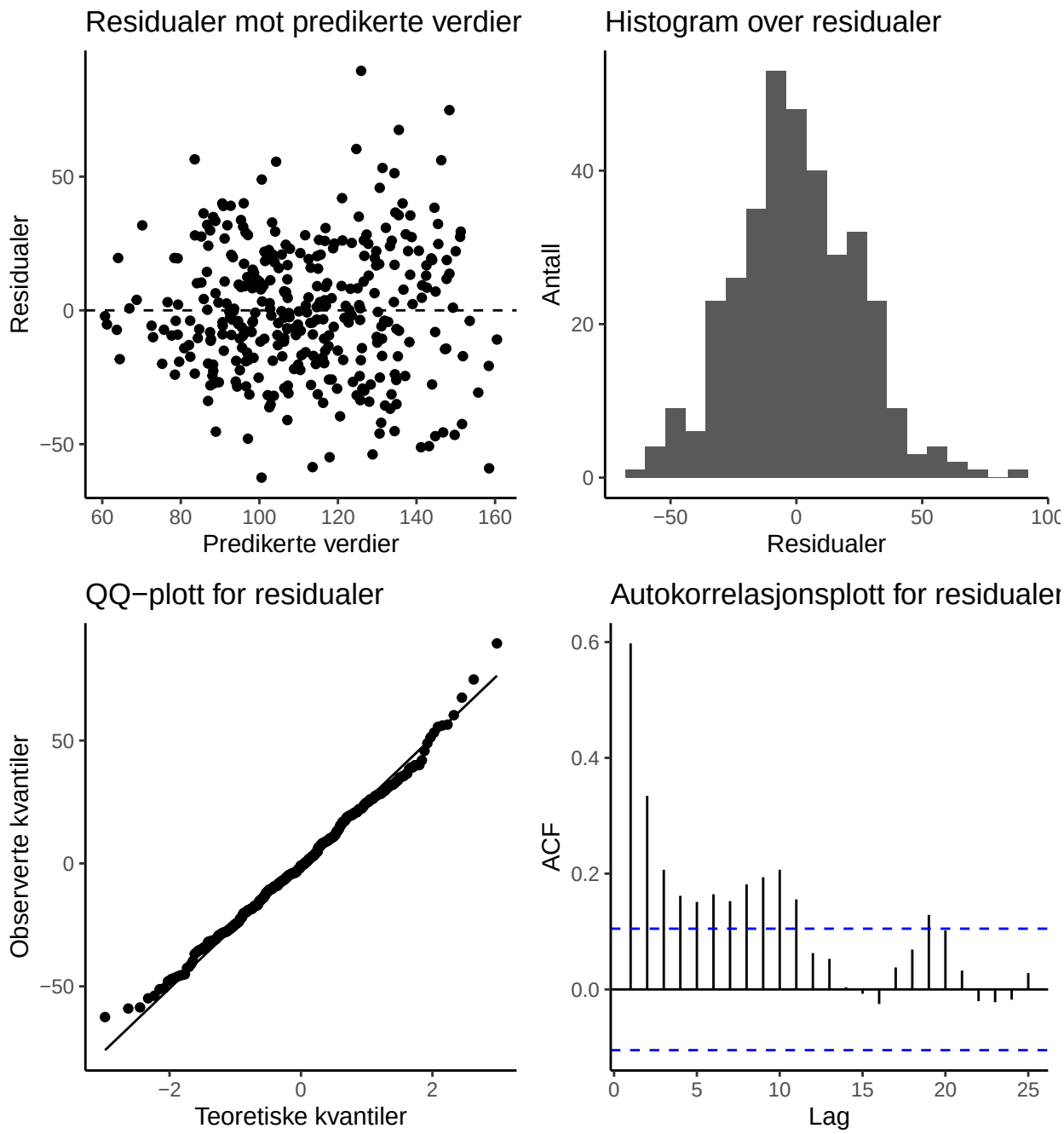
Vedlegg 1b: Graf publisert av Venstre

Bakgrunn: Denne grafen ble publisert av klima- og miljøminister Ola Elvestuen fra Venstre på Twitter 1. november 2019 (men senere tatt bort), og viser norske CO_2-utslipp (i 1000 tonn CO_2-ekvivalenter) som funksjon av tid. Venstre gikk inn i regjering sammen med Høyre og Fremskrittspartiet i januar 2018.

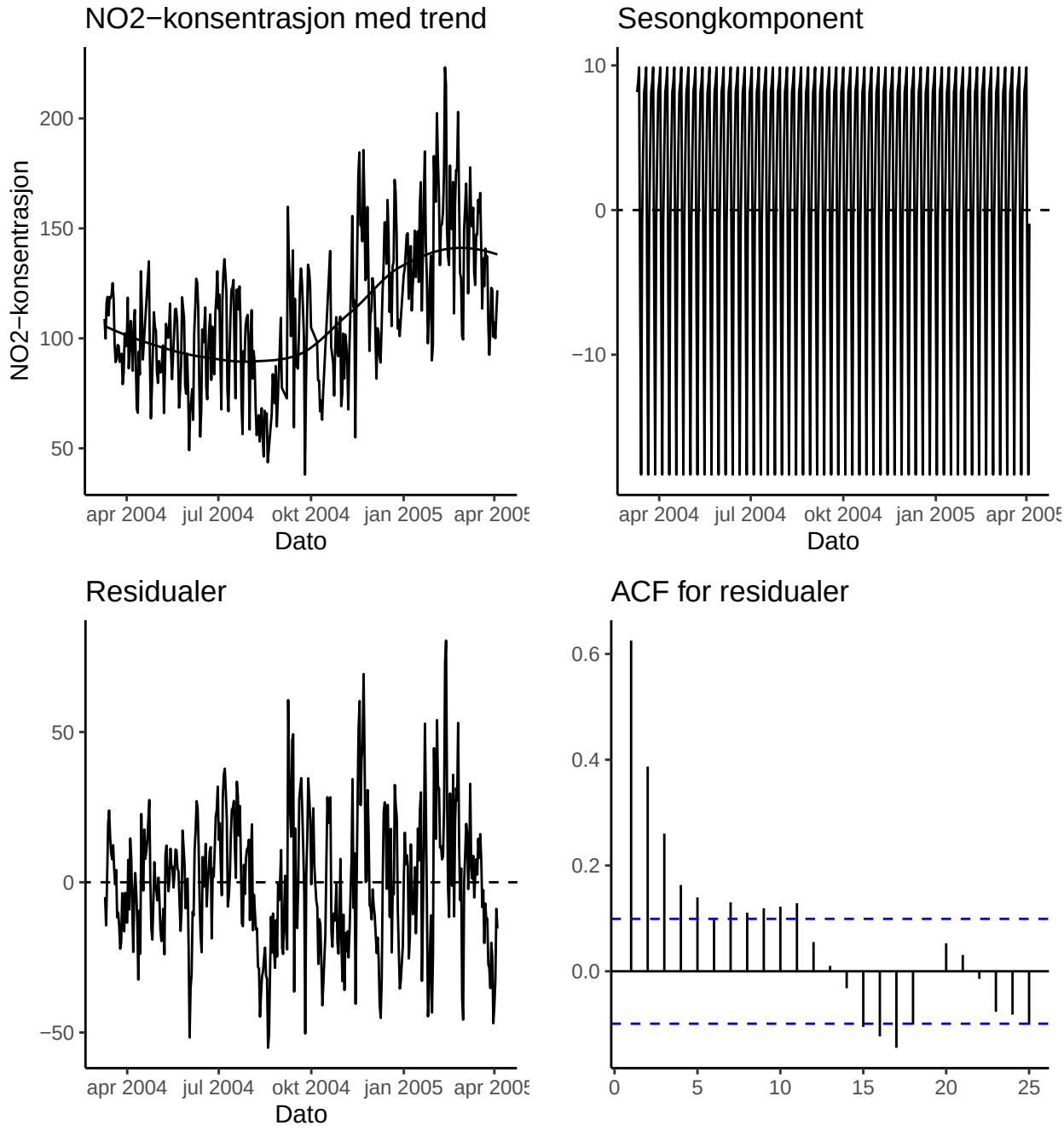
Vedlegg 2: Regresjonsutskrifter

	Dependent variable:	
	no2	danger_warning
	OLS	logistic
	(1)	(2)
WeekdayMonday	-16.562***	-1.587***
	(4.958)	(0.521)
WeekdaySaturday	-19.239***	-2.292***
	(4.889)	(0.519)
WeekdaySunday	-35.619***	-2.730***
	(5.050)	(0.545)
WeekdayThursday	-4.949	-0.915*
	(4.930)	(0.521)
WeekdayTuesday	-11.389**	-1.147**
	(4.946)	(0.523)
WeekdayWednesday	-5.857	-0.825
	(4.909)	(0.525)
Temperature	-1.787***	-0.086***
	(0.252)	(0.025)
Humidity	-0.517***	-0.044***
	(0.121)	(0.013)
Winter	17.688***	1.289***
	(4.187)	(0.393)
Constant	172.893***	5.052***
	(9.485)	(1.043)
Observations	349	349
R2	0.431	
Adjusted R2	0.416	
Log Likelihood		-185.570
Akaike Inf. Crit.		391.141
Residual Std. Error	24.640 (df = 339)	
F Statistic	28.489*** (df = 9; 339)	
Note:		*p<0.1; **p<0.05; ***p<0.01

Vedlegg 3: Diagnoseplott til regresjon (1) i Vedlegg 2



Vedlegg 4: Dekomponering av tidsrekke



Forklaring:

- Oppe til venstre er NO_2-konsentrasjonen plottet gjennom observasjonsperioden sammen med en estimert trendkomponent.
- Oppe til høyre har vi plottet sesongkomponenten, som har en periode på 7 dager.
- Nede til venstre har vi plottet NO_2-konsentrasjonen etter at vi har trukket ut trend- og sesongkomponentene fra tidsrekken ("residualene").
- Nede til høyre ser vi den estimerte autokorrelasjonsfunksjonen til residualtidsrekken.

Vedlegg 5: To estimerte tidsrekkemodeller for Rt

```
Call:
arima(x = airquality_ts$residuals, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.6245    -0.0821
s.e.  0.0393     2.3389

sigma^2 estimated as 304.1:  log likelihood = -1672.83,  aic = 3351.65

Training set error measures:
                      ME     RMSE      MAE       MPE    MAPE      MASE
Training set 0.009145292 17.43983 13.27842 -133.9213 426.148 0.9166486
                    ACF1
Training set 0.00558858
```

```
Call:
arima(x = airquality_ts$residuals, order = c(0, 0, 1))

Coefficients:
         ma1  intercept
      0.5536    -0.0151
s.e.  0.0391     1.4523

sigma^2 estimated as 342.3:  log likelihood = -1695.86,  aic = 3397.73

Training set error measures:
                       ME     RMSE      MAE      MPE    MAPE      MASE
Training set -0.004784311 18.50129 14.31871 14.36084 241.087 0.9884633
                   ACF1
Training set 0.1395284
```