UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1110 — Statistiske metoder og dataanalyse
Eksamensdag:	Torsdag 8. desember 2022
Tid for eksamen:	09.00 – 13.00
Oppgavesettet er på 6 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Godkjent kalkulator Formelsamling for STK1110

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Nedenfor er det gitt kritiske verdier $t_{\alpha,\nu}$ for t-fordelingen med ν frihetsgrader for noen verdier av α og ν. Du vil få bruk for tabellen i Oppgave 1 og 3.

α :	0.10	0.05	0.025	0.0125	0.01	0.005
$t_{\alpha,13}$	1.350	1.771	2.160	2.533	2.650	3.012
$t_{\alpha,14}$	1.345	1.761	2.145	2.510	2.624	2.977
$t_{\alpha,15}$	1.341	1.753	2.131	2.490	2.602	2.947
$t_{\alpha,16}$	1.337	1.746	2.120	2.473	2.583	2.921
$t_{\alpha,17}$	1.333	1.740	2.110	2.458	2.567	2.898
$t_{\alpha,18}$	1.330	1.734	2.101	2.445	2.552	2.878

Oppgave 1

Under er det vist gjennomsnittlig sjøtemperatur i aprilmåned, målt utenfor Ekofisk-feltet i løpet av 15 år i perioden 1995 til 2022 (målingene er oppgitt i $°C$):

```
6.6 4.6 6.3 7.1 6.8 5.8 6.5 6.9 4.8 7.9 7.7 7.4 7.5 6.4 7.2
```

La X_i være i-te måling av temperaturen. Det antas at $X_1, \ldots, X_{15} \overset{uif}{\sim} N(\mu, \sigma^2)$ (målingene av temperaturen er gjort med nokså lange mellomrom, slik at antakelsen om uavhengighet er rimelig).

a

- Utled et 95% konfidensintervall for forventet sjøtemperatur.
- Beregn intervallet for dataene over, når du får vite at observert snitt og standardavvik er $\bar{x} = (1/15)\sum_{i=1}^{15} x_i = 6.63$ og $s = \sqrt{(1/14)\sum_{i=1}^{15}(x_i - \bar{x})^2} = 0.967$.

(Fortsettes på side 2.)

Temperaturen har tidligere vært på gjennomsnittlig $6.2°C$, men forskere tror den kan ha endret seg. De vil derfor teste hypotesene

$$H_0 : \mu = 6.2 \text{ mot } H_a : \mu \neq 6.2.$$

b

- Utled en hypotesetest for disse hypotesene med signifikansnivå 5%.
- Hva blir konklusjonen ut fra de observasjonene som ble gjort?

c

- Forklar hva en P-verdi generelt betyr, altså definér hva den er.
- Finn et uttrykk for P-verdien til testen i Oppgave **b**.
- Bruk tabellen over kritiske verdier for t-fordelingen til å si noe om størrelsesorden for P-verdien.

Oppgave 2

La $X_1, \ldots, X_n$ være uavhengige stokastiske variabler fra $Gamma(4, \beta)$-fordelingen, dvs. med sannsynlighetstetthet

$$f(x; \beta) = \frac{1}{6\beta^4} x^3 e^{-\frac{x}{\beta}}, \ x > 0,$$

med $\beta > 0$ som ukjent parameter. Anta så at vi har observasjoner $x_1, \ldots, x_n$ av disse.

a

Sett opp likelihood- og log-likelihood-funksjonen, og vis at maksimum likelihood-estimatoren blir $\hat{\beta}_{ML} = \frac{1}{4}\bar{X}$.

b

Vis at Fisher-informasjonen i én observasjon er $I(\beta) = \frac{4}{\beta^2}$.

c

Begrunn at $\hat{\beta}_{ML}$ er tilnærmet normalfordelt $N(\beta, \sigma^2_{\hat{\beta}})$-fordelt for store n, og finn et uttrykk for $\sigma^2_{\hat{\beta}}$.

La nå $\psi = \frac{1}{\beta}$.

d

Begrunn at maksimum likelihood-estimatoren for ψ er $\hat{\psi}_{ML} = 4/\bar{X}$.

(Fortsettes på side 3.)

Vi ønsker nå i stedet å bruke Bayes-estimatoren til å estimere ψ, og spesifiserer derfor apriorifordelingen $\psi \sim Gamma(\alpha_0, \beta_0)$. Sannsynlighetstettheten til hvert enkelt datapunkt kan da reparametriseres til

$$f(x|\psi) = \frac{\psi^4}{6} x^3 e^{-\psi x}, \ x > 0.$$

e

- Hva vil det si at en apriorifordeling er konjugert med fordelingen til dataene?
- Vis at apriorifordelingen til ψ er konjugert med fordelingen til $X_1, \ldots, X_n|\psi$, og nærmere bestemt at
$[\psi|X_1 = x_1, \ldots, X_n = x_n] \sim Gamma\left(\alpha_0 + 4n, \frac{1}{\frac{1}{\beta_0} + \sum_{i=1}^n x_i}\right)$.
- Finn Bayes-estimatoren for ψ, og sammenlikn denne med maksimum likelihood-estimatoren fra Oppgave **d**.

Oppgave 3

En ønsker å undersøke hvordan nivået på årlige importen til et land avhenger av det årlige forbruket. Dataene nedenfor viser årlig import og forbruk i Frankrike i millioner FRF (franske franc) i løpet av de 18 årene fra 1949 til 1966:

```
import forbruk
15.9   108.1
16.4   114.8
19.0   123.2
19.1   126.9
18.8   132.1
20.4   137.7
22.7   146.0
26.5   154.1
28.1   162.3
27.6   164.3
26.3   167.6
31.1   176.8
33.3   186.6
37.0   199.7
43.3   213.9
49.0   223.8
50.3   232.0
56.6   242.9
```

Vi tilpasser først en enkel lineær regresjonsmodell med import som responsvariabel Y og forbruk som forklaringsvariabel x_1, dvs.

$$Y_i = \beta_0 + \beta_1(x_{i1} - \bar{x}_1) + \epsilon_i, \ i = 1, \ldots, 18, \tag{1}$$

der $\bar{x}_1 = (1/18)\sum_{i=1}^{18} x_{i1}$, og vi antar at $\epsilon_1, \ldots, \epsilon_{18} \overset{uif}{\sim} N(0, \sigma^2)$. Resultatet av denne analysen er gitt i R -utskriften på neste side.

(Fortsettes på side 4.)

R-utskrift fra tilpasning til modell (1):

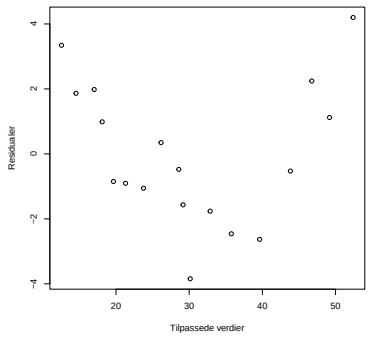
```
Call:
lm(formula = import ~ I(forbruk - mean(forbruk)))

Residuals:
Min      1Q  Median      3Q     Max
-3.8435 -1.4407 -0.5032  1.6780  4.1982

Coefficients:
Estimate  Std. Error  t value Pr(>|t|)
(Intercept)                 30.07778    0.52732   57.04  < 2e-16 ***
I(forbruk - mean(forbruk))  0.29559      0.01305   22.65 1.39e-13 ***
---Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.237 on 16 degrees of freedom
Multiple R-squared:  0.9698,Adjusted R-squared:  0.9679
F-statistic: 513.1 on 1 and 16 DF,  p-value: 1.392e-13
```

a

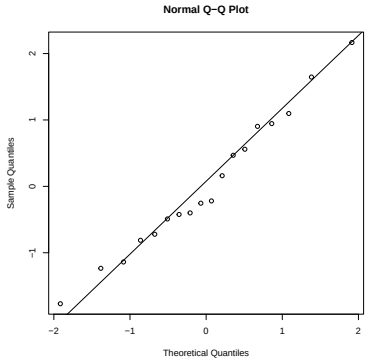
- Gi en fortolkning av estimatene $\hat{\beta}_0$ og $\hat{\beta}_1$.
- Lag så et 95% konfidensintervall for β_1 (du trenger ikke å utlede det, men skriv opp formelen du bruker).
- Kan en forvente at importen øker med omtrent 0.25 millioner FRF når forbruket øker med 1 million FRF?

Residualplott for modell (1) er vist under:

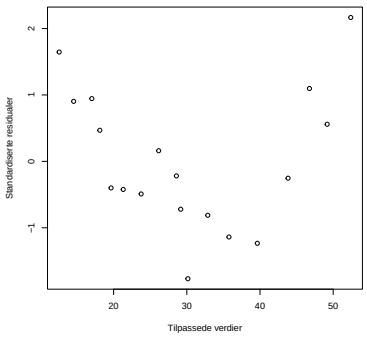
`Residualer mot tilpassede verdier`



`Normalfordelingsplott av standardiserte residualer`



`Standardiserte residualer mot tilpassede verdier`



b

- Hvilke modellantakelser har en gjort i modell (1)?
- Hvordan sjekker en hver av modellantakelsene med ved hjelp av residualplottene over?
- Benytt residualplottene over til å vurdere gyldigheten av modellantakelsene i modell (1).

(Fortsettes på side 5.)

Ut fra resultatene i residualplottene i Oppgave **b**, velger vi å tilpasse følgende multiple lineære regresjonsmodell i stedet:

$$Y_i = \gamma_0 + \gamma_1(x_{i1} - \bar{x}_1) + \gamma_2 x_{i2} + \varepsilon_i, \ i = 1, \ldots, 18, \tag{2}$$

der $x_{i2} = (x_{i1} - \bar{x}_1)^2$ og $\varepsilon_1, \ldots, \varepsilon_{18} \overset{uif}{\sim} N(0, \tau^2)$.

c

- Vis at forventet import nå er en kvadratisk funksjon av forbruket.
- Hvor mye øker forventet import når forbruket øker med én enhet fra x_1^* (altså fra x_1^* til $x_1^* + 1$)?
- Hva er nå fortolkningen av konstantleddet γ_0, og hva med fortolkningen av γ_1?

Under og på neste side vises

- R-utskriften fra en test av modell (2) mot modell (1).
- R-utskriften fra tilpasningen til modell (2)
- residualplott for modell (2)

d

- Hvilke hypoteser er det som testes i R-utskriften under?
- Hvilken av modellene (1) og (2) vil du velge basert på hypotesetesten og R-utskriften fra tilpasningen? Begrunn svaret ditt.
- Bruk residualplottene til å sjekke modellantakelsene for modell (2), og sammenlikn med Oppgave **b**.

R-utskrift fra hypotesetest av modell (2) mot modell (1):

```
Analysis of Variance Table

Model 1: import ~ I(forbruk - mean(forbruk))
Model 2: import ~ I(forbruk - mean(forbruk)) + I((forbruk - mean(forbruk))^2)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     16 80.082
2     15 16.397  1    63.685 58.261 1.528e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

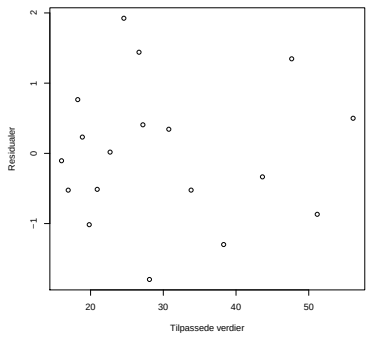
(Fortsettes på side 6.)

R-utskrift fra tilpasning til modell (2):

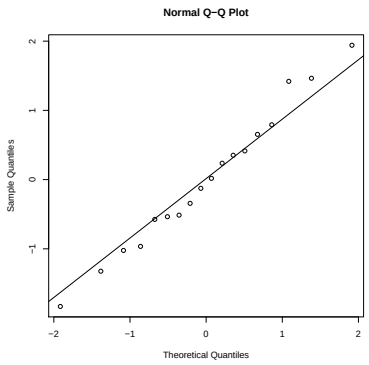
```
Call:
lm(formula = import ~ I(forbruk - mean(forbruk)) + I((forbruk -
    mean(forbruk))^2))

Residuals:
     Min       1Q   Median       3Q      Max
-1.79590 -0.52376 -0.04364  0.47666  1.92453

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    28.034259   0.363875  77.044  < 2e-16 ***
I(forbruk - mean(forbruk))      0.277111   0.006561  42.233  < 2e-16 ***
I((forbruk - mean(forbruk))^2)  0.001251   0.000164   7.633 1.53e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.046 on 15 degrees of freedom
Multiple R-squared:  0.9938,Adjusted R-squared:  0.993
F-statistic:  1204 on 2 and 15 DF,  p-value: < 2.2e-16
```

Residualplott for modell (2):

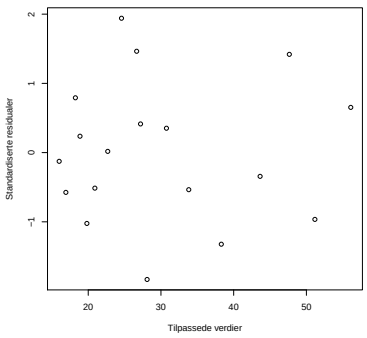
Residualer mot tilpassede verdier



Normalfordelingsplott av standardiserte residualer



Standardiserte residualer mot tilpassede verdier



END