BOKMÅL

NHH

SKOLEEKSAMEN
MET4

Høst 2025

Dato: 25. november 2025

Tidsrom: 09:00 - 15:00

Antall timer: 6

Foreleser/kursansvarlig kan kontaktes av eksamensvakt på telefon: *99385583*

TILLATTE HJELPEMIDLER:
Kalkulator Ja ☒ Nei □
Ordbok: én tospråklig ordbok (kategori 1)
Alle trykte/egenskrevne hjelpemidler (kategori 3).

Antall sider, inkludert forside og vedlegg: 11

Antall digitale vedlegg: 3 (claims.Rdata, MET4Formler.pdf, relevante_r_kommandoer.pdf)

Del 1 - Dataanalyse med R

I denne delen skal du analysere data fra et forsikringsselskap som tilbyr bilforsikringer til privatkunder. Selskapet ønsker å forstå hvilke faktorer som kan forklare hvorfor enkelte nye kunder melder skade kort tid etter at forsikringen er tegnet ("early claims"). Målet er å kunne identifisere risikokunder tidlig og sette inn forebyggende tiltak, for eksempel rådgivning eller justering av egenandel.

Datasettet `claims` inneholder informasjon om et utvalg av 1500 kunder som nylig har kjøpt bilforsikring. I oppgavene under skal du bruke dataene til å beskrive kundene, gjennomføre relevante tester og estimere modeller som kan brukes til å forutsi en tidlig skade.

Last inn datasettet ved navn `claims` i R-minnet ved hjelp av følgende kommando, der vi antar at filen ligger i arbeidsmappen din:

```
load("claims.Rdata")
```

En oversikt over variablene i datasettet finner du i Tabell 1.

Tabell 1: Beskrivelse av variabler i datasettet `claims`.

Variabelnavn	Beskrivelse
`id`	Unik identifikator for hver kunde (1–1500)
`early_claim`	`"Yes"` = Kunde meldte skade innen 90 dager, `"No"` = Ingen skade
`driver_age`	Alder på hovedfører (18–85 år)
`years_license`	Antall år med førerkort (0–60 år)
`car_age`	Bilens alder i hele år (0–20)
`mileage_km`	Deklarert årlig kjørelengde (3 000–50 000 km)
`engine_power_hp`	Motoreffekt i hestekrefter (60–350 hk)
`channel`	Kjøpskanal (`"Web"` (Internett), `"Agent"` (Kundebehandler) eller `"Phone"` (telefon))
`garage`	`"Yes"` = Garasje, `"No"` = Ingen garasje
`region`	Bostedsområde (`"Urban"` (I by) eller `"Rural"` (landlig))
`multi_policy`	`"Yes"` = Flere forsikringer i selskapet, `"No"` = Kun bilforsikring
`payment_method`	Betalingsmåte (`"AvtaleGiro"` eller `"Faktura"`)

Oppgave 1 – Deskriptiv statistikk

(a) Lag en tabell med deskriptiv statistikk for følgende variabler: `driver_age`, `years_license`, `car_age`, `mileage_km`. Basert på tabellen, kommenter kort sentraltendens og spredning for variablene.

(b) Lag et histogram av `mileage_km`. Lag et boksplott som sammenligner fordelingen av `driver_age` for kundene med og uten `early_claim`. Kommenter kort hva du ser i figurene.

(c) Lag søylediagram for variabelen `channel` og frekvenstabeller for variablene `garage`, `region` og `multi_policy`. Kommenter kort hva figuren og tabellene viser om kundene i utvalget.

Oppgave 2 – Hypotesetesting

(a) **Undersøk om kunder med tidlig skade har høyere deklarert kjørelengde (`mileage_km`) enn kunder uten tidlig skade ved å gjennomføre en to-utvalgs t-test. Formuler null- og alternativhypotese, og tolk resultatet.**

(b) **Undersøk om de tre kjøpskanalene er *like mye brukt* ved å gjennomføre en kji-kvadrat goodness-of-fit-test på frekvensene for `channel`. Formuler null- og alternativhypotese, og tolk resultatet.**

Oppgave 3 – Logistisk regresjon og KNN

Del datasettet inn i et treningssett (70%) og et testsett (30%) med følgende R-kode:

```
treningssett <- claims[1:1050, ]
testsett <- claims[1051:nrow(claims), ]
```

(a) **Bruk treningssettet til å tilpasse en *logistisk regresjonsmodell* der `early_claim` er responsvariabel og følgende variabler brukes som forklaringsvariabler:**

`driver_age`, `car_age`, `mileage_km`, `region`, `payment_method`, `garage`.

Presenter en utskrift av modellen og gi en kort fortolkning av koeffisienten til `driver_age` og `garage`.

(b) **Klassifiser kundene i testsettet som `"Yes"` dersom sannsynligheten fra modellen i (a) overstiger henholdsvis 0.10 og 0.30 (ellers `"No"`). Lag kontingenstabell for begge grensene som viser klassifiseringsresultatene og kommenter hvilke fordeler og ulemper de to grenseverdiene har.**

(c) **Bruk treningssettet til å tilpasse en KNN-modell for å klassifisere tidlig skade. La antall naboer k være 5 og bruk de samme forklaringsvariablene som i (a). Bruk modellen til å klassifisere kundene i testsettet. Lag en kontingenstabell som viser klassifiseringsresultatene, rapporter andel riktige klassifiseringer, og kommenter kort hvordan KNN-modellen presterer basert på disse resultatene.**

Et korrekt `"Yes"` gir stor gevinst ved at tidlig skade oppdages og dyre utbetalinger unngås, mens et korrekt `"No"` gir en liten gevinst ved å unngå unødvendige tiltak. Feilklassifiseringer gir tap: falske alarmer har en administrativ kostnad, mens oversette skader er klart dyrest:

- Kunden har faktisk tidlig skade `"Yes"` og klassifiseres `"Yes"`: **+20 000 kr**
- Kunden har faktisk ingen tidlig skade `"No"` og klassifiseres `"No"`: **+300 kr**
- Kunden har faktisk ingen tidlig skade `"No"` og klassifiseres `"Yes"`: **-500 kr**
- Kunden har faktisk tidlig skade `"Yes"` og klassifiseres `"No"`: **-5000 kr**

(d) **Sammenlign alle tre klassifiseringsmetoder (logistisk med grenseverdiene 0.10 og 0.30, samt KNN) ut fra gevinstene over. Beregn forventet gevinst per kunde for alle metodene, og angi hvilken metode som gir størst forventet gevinst. Gi en kort kommentar til resultatet.**

Del 2 - Regneoppgaver

Mange har nok hørt om **Norgespris**, den nye støtteordningen for strømkostnader som Regjeringen innførte fra 1. oktober 2025. Dette er en fastprisordning som gir husholdningene en forutsigbar strømpris på 50 øre/kilowatt-time (kWh). Det er opp til hver enkelt husholdning om de ønsker å binde seg til denne fastprisen eller beholde det gamle strømstøtteordningen som dekker 90% av strømprisen som overstiger 75 øre/kWh.

Fra nettsiden elhub.no kan man studere historisk forbruk og sammenligne hvordan månedsprisene på strøm for din husstand ville vært med Norgespris eller med den vanlige strømstøtteordningen. I regnedelen av eksamen har vi hentet ut denne informasjonen for boligen til en postdoktor ved NHH - la oss kalle han Sondre - og vi skal se på aspekter knyttet til om han skal binde seg til Norgespris eller ikke. Datasettet fra elhub dekker perioden september 2022-september 2025 (n = 37). Merk at Norgespris ikke har eksistert i den historiske perioden. For hver måned, regner vi ut differansen i strømkostnader for Sondres husholdning (kostnad med vanlige strømstøtteordning - kostnad med Norgespris).

Variabel	Enhet	Beskrivelse
`norgespris`	NOK	Strømkostnaden for Sondre med Norgespris per måned.
`stromstotte`	NOK	Strømkostnaden for Sondre med Strømstøtteordningen per måned.
`differanse` (Y_t)	NOK	Differansen: `stromstotte` - `norgespris`.
`maned`	År-måned	Hvilket år og måned de andre variablene er målt på.
`temperatur`	°C	Månedlig gjennomsnittstemperatur i Bergen.
`nedbor`	mm	Månedlig total nedbørsmengde i Bergen.
`forbruk`	kWh	Månedlig strømforbruk for Sondre.

Det er variabelen `differanse` vi vil bruke som responsvariabel i våre modeller og bruker notasjonen Y_t for differansen målt ved tidspunkt (måned) t. Merk at $Y_t > 0$ betyr at Norgespris ville lønnet seg denne måneden (lavere kostnader med Norgepris enn uten), mens $Y_t < 0$ vil være i disfavør Norgespris (høyere kostnader med Norgespris enn uten).

I tillegg har vi informasjon om strømforbruket (`forbruk`), hva gjennomsnittstemperaturen (`temperatur`) og nedbørsmegden (`nedbor`) var i Bergen for hver måned[1].

Oppgave 4 – Hypotesetesting

Hypotesen vi ønsker å teste er om det vil lønne seg for Sondre å tegne en Norgesprisavtale. I tabellen under finner du oppsummerende statistikk for datakolonnene `norgespris`, `stromstotte` og `differanse`.

	Oppsummerende statistikk					
	median	mean	sd	q25	q75	n
`norgespris`	594	622.59	215.15	428	830	37
`stromstotte`	767	898.08	577.30	421	1193	37
`differanse`	192	275.49	397.06	-19	564	37

[1] Meterologiske data fra https://seklima.met.no/.

(a) Hvilken hypotesetest og tilhørende null- og alternativehypotese bør man bruke i dette tilfellet? Argumenter for hvor du mener signifikansnivået bør ligge for denne hypotesetesten.

(b) Utfør hypotesetesten. Hva er din konklusjon?

Oppgave 5 - Linear regresjon

I spørsmålet om Norgespris vil lønne seg er det essensielle hva strømprisen vil være det neste året. I Norge er det velkjent at temperatur er en veldig viktig prediktor for strømpris. Når det er kaldere, øker strømforbruket fordi vi trenger å varme opp husene våre. Nedbør er også positivt for lavere strømpriser, siden det medfører mer tilsig i vannmagasinene til de store strømprodusentene.

I **Vedlegg 1** har vi tilpasset en linear regresjon for `differanse` forklart av `temperatur`.

(a) Skriv opp den estimerte modellen. Finn temperaturintervallet hvor Norgespris er mer lønnsom enn vanlig strømstøtte i følge modellen.

Vi tar inn forklaringsvariablene `nedbor` og `forbruk`, og lager noen modellkandidater som vi sammenligner med Akaikes informasjonskriterium (AIC) (se **Vedlegg 2**).

(b) Hvilken modell vil du velge basert på AIC? Forbedrer forklaringsvariabelen `forbruk` modellene? Kommenter.

(c) Bruk modellen `lm2` og prediker `differanse` for Oktober 2025, gitt at langtidsvarselet predikerer en gjennomsnittstemperaturen på 10.4°C og nedbørsmengde på 225.8 mm for denne måneden. Gi en fortolkning av koeffisienten for nedbør. Hvilke mekanismer kan ligge bak fortegnet på denne koeffisienten?

Oppgave 6 – Tidsrekker

Differansen i strømkostnad for de to støtteordningene i måned t, Y_t, er en tidsrekke for $t = 1, \ldots, T$, hvor Y_1 er differansen i september 2022 og Y_T er den i september 2025. Vi har plottet den faktiske tidsrekken i **Vedlegg 3**.

(a) Indiker hvilken støtteordning som ville vært mest økonomisk gunstig for i) januar 2025 og ii) juli 2025.

(b) Forklar hvorfor tidsrekken ikke er stasjonær.

Siden tidsrekken ikke er stasjonær, dekomponerer vi den i tre deler:

$$Y_t = T_t + S_t + R_t,$$

hvor S_t er en deterministisk sesongkomponent, T_t er en trend-komponent og R_t er en residual-komponent.

(c) Sesongkomponenten her kan tolkes som en erstatter for temperaturen vi brukte i Oppgave 5. Drøft denne påstanden kort. Hvorfor er det fornuftig å

bruke en sesong-komponent i stedet for å direkte bruke temperaturen når hensikten med modellen er å predikere langt frem i tid?

Vi tilpasser en ARIMA(p,d,q)-modell til $T_t + R_t$. Utskrift fra den estimerte modellen er gitt i **Vedlegg 4**.

(d) Skriv opp ARIMA modellen med estimerte koeffisienter. Hint: Det kan være nyttig å definere $X_t = T_t + R_t$ og sette opp modellen for X_t.

Det er interessant hvordan kostnadsdifferansen summerer seg opp ett år frem i tid. En måte å gjøre dette på er å simulere ett år frem i tid, $\{Y_{T+1}, \ldots, Y_{T+12}\}$, fra den estimerte modellen i Vedlegg 3, si 1000 ganger. Deretter kan vi for hver av de 1000 simuleringene regne ut total strømkostnadsdifferansen for hele året: $W = \sum_{h=1}^{12} Y_{T+h}$. I **Vedlegg 5** har vi plottet de 1000 simuleringene av $\{Y_{T+1}, \ldots, Y_{T+12}\}$, samt et histogram av de tilhørende 1000 simulerte verdiene av W. Gjennomsnittet av de simulerte verdiene av W er **5909 NOK** og standardavviket er **3840 NOK**.

(e) Hvorfor ser det rimelig ut at normalfordeling (den røde linjen) passer godt til histogrammet i Vedlegg 5? Approksimer sannsynligheten for at Norgespris vil lønne seg for Sondre det neste året.

Kritikere av innføringen av Norgespris-ordningen mener at den fjerner insentivene til strømsparing når tilgangen på strøm er knapp. Det at (per 26.10.2025) 1 074 808 husholdninger og fritidsboliger har tegnet Norgespris-avtale[2], gjør at en betydelig andel av strømkunder, spesielt i Sør-Norge, ikke behøver å tilpasse forbruket sitt til prissvinginger. Modellen vår betinger på at utviklingen i kostnadsdifferansen med- og uten Norgespris fortsetter inn i det neste året.

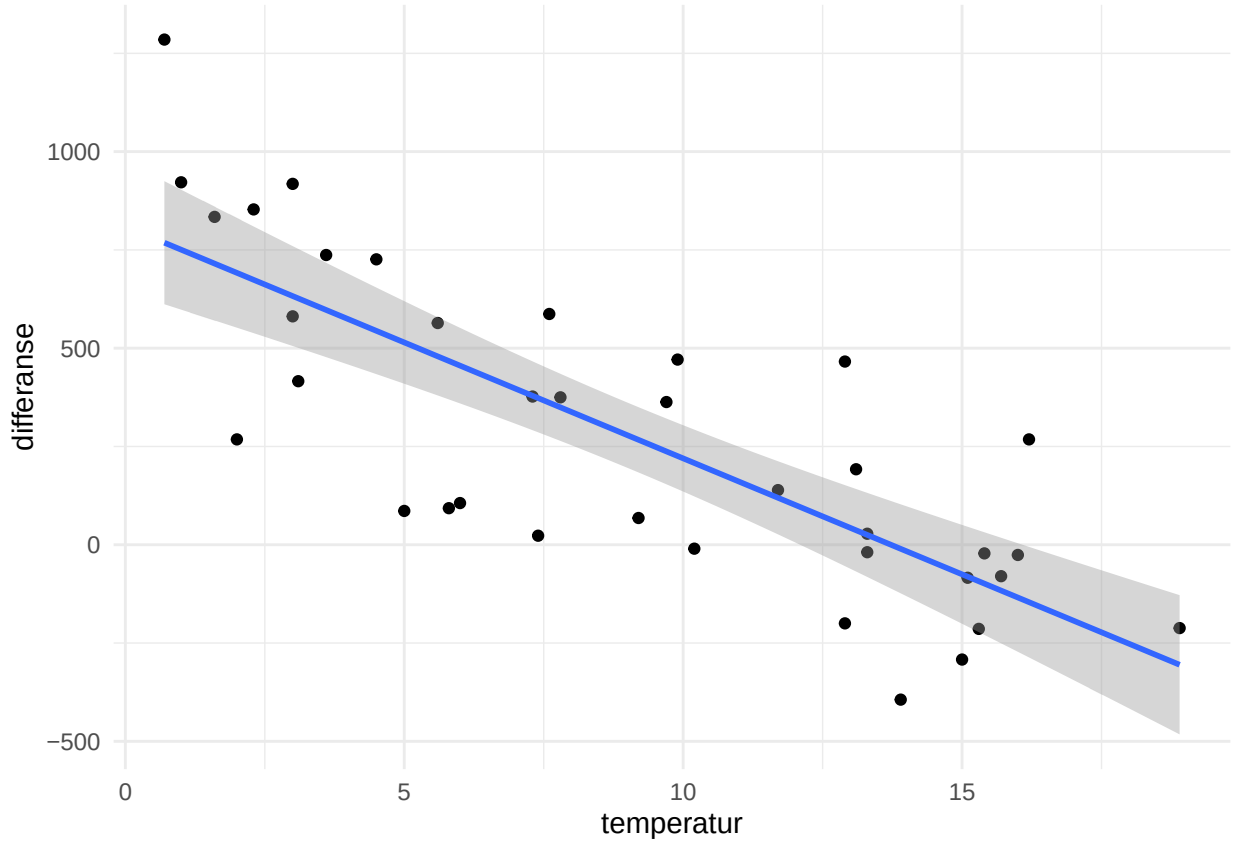
(f) Vil disse aspektene ha innvirkning på prediksjonen ifra vår modell? I så fall, hvordan ser du for deg det vil påvirke? Drøft kort.

Språklig presisering (ikke relevant for løsning av eksamen)

I denne oppgaven har vi brukt de dagligdagse begrepene *strømpris*, *strømkostnad* og *strømforbruk*. Vi burde bruke begrepene *energipris*, *energikostnad* og *energiforbruk*, men da må vi hele tiden presisere at det er elektrisk energi vi snakker om. Men så lenge det er snakk om kilowatt-timer (kWh) er dette **energi** og ikke **strøm** (som måles i ampere).

[2] https://elhub.no/data-og-innsikt/statistikk-for-norgespris

Vedlegg 1 - Lineær regresjon med temperatur




```r
lm1 <- lm(differanse ~ temperatur, data = data) # Model 1
summary(lm1)
```

```
##
## Call:
## lm(formula = differanse ~ temperatur, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -428.63 -210.70   25.54  155.18  516.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  809.566     81.813   9.895 1.12e-11 ***
## temperatur   -58.988      7.827  -7.536 7.87e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 248.7 on 35 degrees of freedom
## Multiple R-squared:  0.6187, Adjusted R-squared:  0.6078
## F-statistic: 56.79 on 1 and 35 DF,  p-value: 7.871e-09
```

Vedlegg 2 - Lineær regresjoner

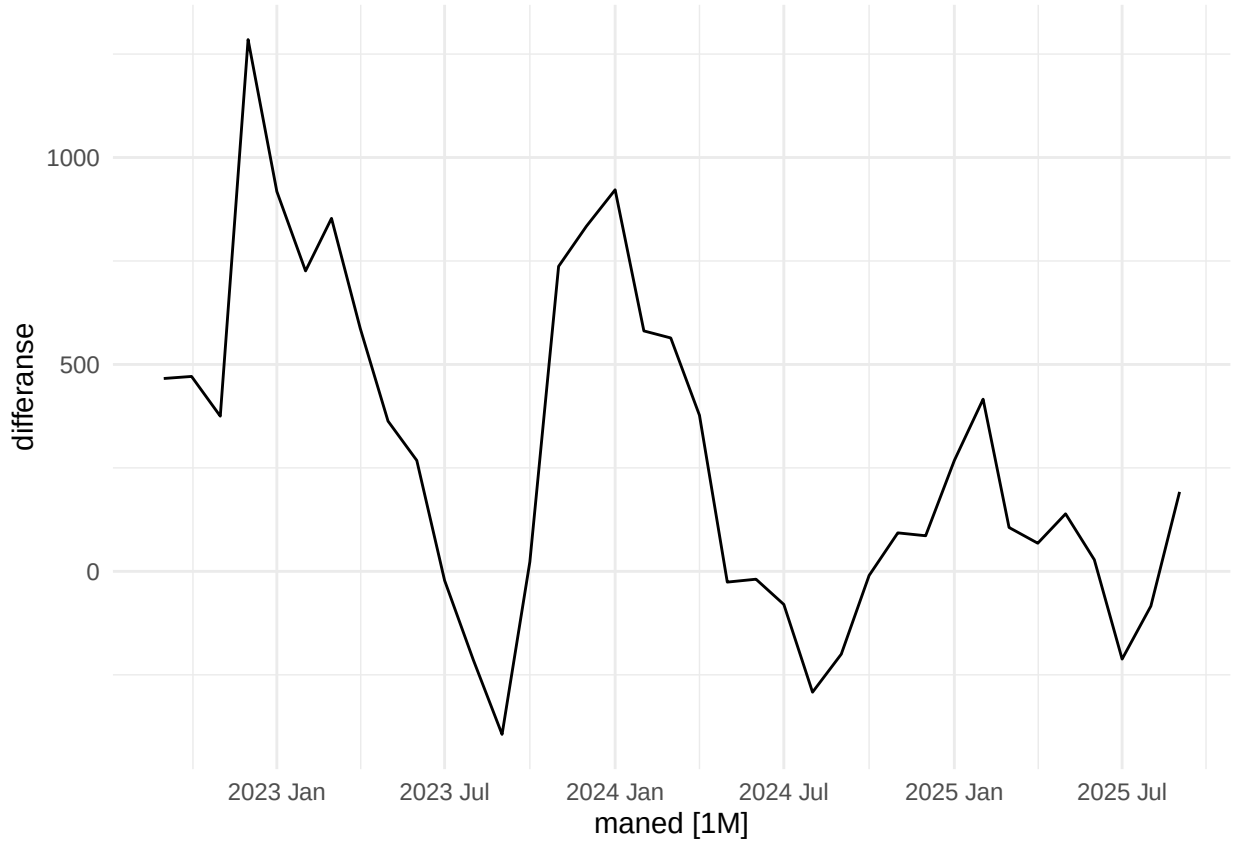
```
lm2 <- lm(differanse ~ temperatur + nedbor, data = data) # Model 2
lm3 <- lm(differanse ~ temperatur + nedbor + forbruk, data = data) # Model 3
lm4 <- lm(differanse ~ temperatur + forbruk, data = data) # Model 4
AIC(lm1,lm2, lm3, lm4)
```

```
##     df      AIC
## lm1  3 517.1343
## lm2  4 512.0334
## lm3  5 512.7456
## lm4  4 518.4659
```

```
summary(lm2)
```

```
##
## Call:
## lm(formula = differanse ~ temperatur + nedbor, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -450.58 -157.19    0.84  128.32  507.19
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1058.5930   119.6167   8.850 2.42e-10 ***
## temperatur   -62.9260     7.3628  -8.546 5.54e-10 ***
## nedbor        -1.0452     0.3897  -2.682   0.0112 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.2 on 34 degrees of freedom
## Multiple R-squared:  0.6853, Adjusted R-squared:  0.6668
## F-statistic: 37.02 on 2 and 34 DF,  p-value: 2.913e-09
```

Vedlegg 3 - Tidsrekken `differanse`



Vedlegg 4 - Tidsrekkemodellen

```
## Series: differanse
## Model: STL decomposition model
## Combination: season_adjust + season_year
##
## ========================================
##
## Series: season_adjust
## Model: ARIMA(0,1,1)
##
## Coefficients:
##           ma1
##       -0.4823
## s.e.   0.1683
##
## sigma^2 estimated as 40372:  log likelihood=-241.61
## AIC=487.23   AICc=487.59   BIC=490.39
##
## Series: season_year
## Model: SNAIVE
##
## sigma^2: 31.8384
```

Vedlegg 5 - Fremskrivning av `differanse`