BOKMÅL

NHH

SKOLEEKSAMEN MET4

Vår 2025

Dato: 29. april 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: 9

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

Del 1 - Dataanalyse med R

I denne delen skal du analysere data fra en kundeundersøkelse gjennomført av et selskap som tilbyr digitale abonnementstjenester. Selskapet vurderer ulike tiltak for å øke kundenes forbruk, blant annet en ny markedsføringskampanje. Du skal i oppgavene under bruke dataene til å beskrive kundene, gjennomføre hypotesetester og estimere regresjonsmodeller.

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

```
load("survey.Rdata")
```

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

Tabell 1: Beskrivelse av variabler i datasettet `survey`

Variabelnavn	Beskrivelse
`age`	Alder i hele år (18–80)
`gender`	Kjønn (`"Mann"` eller `"Kvinne"`)
`income`	Husholdningsinntekt (i 1000 kr per måned)
`ad_exposure`	Opplevd reklameeksponering (skår fra 0 til 10)
`loyalty_program`	1 = Medlem av lojalitetsprogram, 0 = Ikke medlem
`satisfaction`	Skår 1–10 på fornøydhet med tjenesten
`promo_response`	`"Ja"` eller `"Nei"` på om reklamen påvirket kunden
`spending`	Total kr brukt på tjenesten siste 3 måneder

Oppgave 1 – Deskriptiv statistikk

(a) Lag et boksplott som sammenligner `spending` mellom menn og kvinner. Kommenter kort hva figuren viser.

(b) Lag et boksplott som sammenligner `ad_exposure` mellom menn og kvinner. Kommenter resultatet og formuler en mulig årsak til forskjellen i `spending` du så i oppgave (a).

Oppgave 2 – Hypotesetesting

(a) Test om variansen i `spending` er forskjellig mellom menn og kvinner. Formuler null- og alternativhypotese, og tolk resultatet.

(b) Gjennomfør en to-utvalgs t-test for å undersøke om det er en signifikant forskjell i gjennomsnittlig `spending` mellom menn og kvinner. Formuler null- og alternativhypotese, og tolk resultatet.

(c) Lag en krysstabell som viser fordelingen av `promo_response` (Ja/Nei) etter kjønn. Gjennomfør en kjikvadrattest for å undersøke om det er en sammenheng mellom kjønn og reklamerespons. Formuler null- og alternativhypotese, og tolk resultatet.

Oppgave 3 – Regresjonsanalyse

(a) **Estimer en lineær regresjonsmodell der `spending` er responsvariabel og `gender` er forklaringsvariabel. Presenter resultatene og tolk begge koeffisientene.**

(b) **Estimer en ny regresjonsmodell der `spending` forklares av `gender`, `income`, `ad_exposure` og `loyalty_program`.**

(i) **Tolk koeffisienten for `ad_exposure`.**

(ii) **Sammenlign koeffisienten for `gender` i denne modellen med den du fant i oppgave (a). Hva kan være en mulig forklaring på at denne har endret seg?**

(c) **Bruk regresjonsmodellen fra oppgave (b) til å predikere forventet `spending` for en gjennomsnittlig mannlig kunde. Bruk gjennomsnittlige verdier for de numeriske variablene, og sett `gender = "Mann"` og `loyalty_program = 0`.**

Selskapet vurderer å gjennomføre en markedsføringskampanje som innebærer at hver kunde i gjennomsnitt øker sin reklameeksponering med 2 enheter, og samtidig blir med i lojalitetsprogrammet. Hver enhets økning i reklameeksponering koster 5 kroner per kunde, og det koster 100 kroner å rekruttere én ny lojalitetskunde. Selskapet vurderer kampanjen som lønnsom dersom en estimert økningen i spending for en gjennomsnittskunde er minst fem ganger så stor som kostnaden per kunde.

(d) **Bruk regresjonsmodellen fra oppgave (b) og dette kriteriet til å vurdere om kampanjen bør gjennomføres.**

Del 2 - Regneoppgaver

Oppgave 4 – Deskriptiv statistikk

I november 2023 publiserte E24 en sak med overskriften «Solid oppgang i forbruksgjelden kan skape større problemer for folk»[1] . Figur 1 i Vedlegg 1 ble brukt som illustrasjon i artikkelen.

Mener du det er grunnlag for E24s vinkling, basert på informasjonen i figuren? Gi en kort begrunnelse for svaret ditt.

Oppgave 5: Hypotesetesting

Et forskerteam[2] ønsker å undersøke om et moralsk dytt (nudge) kan redusere sannsynligheten for at personer lyver i en incentivert situasjon. Deltakerne ble tilfeldig fordelt i to grupper:

- **Kontrollgruppen** fikk ingen tilleggsinformasjon
- **Tillitsnudge-gruppen** ble møtt med budskapet «Vi stoler på deg» før de ble bedt om å rapportere resultatet

Alle deltakerne skulle først gjette utfallet av et terningskast, deretter gjennomføre kastet, og til slutt rapportere om de hadde gjettet riktig. Dersom de rapporterte at de gjettet riktig, mottok de en pengepremie. Deltakere som gjettet feil, kunne likevel være uærlige og rapportere at de hadde gjettet riktig for å få premien. Det var ikke mulig å etterprøve sannheten i de individuelle rapporteringene.

Tabell 2 viser hvor mange deltakere i hver gruppe som rapporterte at de hadde gjettet riktig.

Tabell 2: Antall riktige gjetninger per gruppe

gruppe	Antall deltakere som rapporterte at de gjettet riktig	Antall deltakere
Kontroll	181	400
Tillitsnudge	116	400

La p_1 og p_2 være de **sanne andelene** som ville rapportert at de gjettet riktig i henholdsvis tillitsnudge- og kontrollgruppen, dersom vi kunne observert hele populasjonen av mulige eksperimentdeltakere.

(a) (i) **Anta at deltakerne i begge gruppene er helt ærlige. Hvilke verdier av p_1 og p_2 forventer vi?**

(ii) **Anta at deltakerne i begge gruppene er like ærlige. Hvilken relasjon forventer vi å se mellom p_1 og p_2?**

(iii) **Anta at tillitsnudge-gruppen er mer ærlig enn kontroll gruppen. Hvilken relasjon forventer vi å se mellom p_1 og p_2?**

[1] https://e24.no/norsk-oekonomi/i/0Qg2k2/solid-oppgang-i-forbruksgjelden-kan-skape-stoerre-problemer-for-folk
[2] Eksperimentet er basert på Ekström et al. (2023), se https://doi.org/10.1257/aer.20211128.

Forskerne har en hypotese om at personer som mottar en tillitsnudge på denne måten er mer ærlige enn de som ikke mottar en tillitsnudge.

(b) Bruk svaret ditt i (a) til å formulere en passende null- og alternativ hypotese om p_1 og p_2. Utfør hypotesetesten med 1% signifikansnivå.

(c) Drøft kort i hvilken grad resultatet i (b) gir støtte til den kausale påstanden om at tillitsnudgen fører til mer ærlighet. Gi et kort eksempel på en realistisk situasjon der en tilsvarende "nudge" kan være nyttig.

Oppgave 6: Tidsrekker

Figuren i Vedlegg 2 viser antall registrerte skader på grunn av skred i Norge fra 1980 til 2023. Dataene er hentet fra norsk naturskadepool. Figuren viser også en glatting av tidsrekken beregnet ved hjelp av enkel eksponentiell glatting med vekting $w = 0.3$.

(a) Vurder om tidsrekken fremstår som stasjonær.

(b) Ved slutten av tidsserien, i 2023, ble det registrert 906 skader. Anta at den forrige glattede verdien var 800. Beregn den nye glattede verdien for 2023. Gi en prediksjon for 2024.

(c) Drøft kort svakheten ved å bruke enkel eksponentiell glatting til prediksjon når tidsserien har en jevn stigende trend.

Oppgave 7: KNN og logistisk regresjon

Et revisjonsteam i en større virksomhet ønsker å identifisere bilag som kan være mistenkelige, og som derfor bør gjennomgås manuelt. Basert på historiske data har de informasjon om flere bilagsrelaterte variabler, samt om bilaget i ettertid ble vurdert som mistenkelig av revisor.

Tabell 3 i Vedlegg 3 viser en oversikt over variablene i datasettet. Vedlegg 3 viser også resultatene fra en logistisk regresjon der den avhengige variabelen er `mistenkelig`, og alle de øvrige variablene i datasettet brukes som forklaringsvariabler.

(a) Skriv opp den estimerte logistiske regresjonsmodellen som vist i Vedlegg 3. Gi en fortolkning av koeffisienten for `vedlegg` i modellen.

Et konkret bilag har følgende verdier: `beløp` = 30, `endringer` = 2, `bilagsnummer` = 800, `vedlegg` = 1 og `bilagstype` = `faktura`.

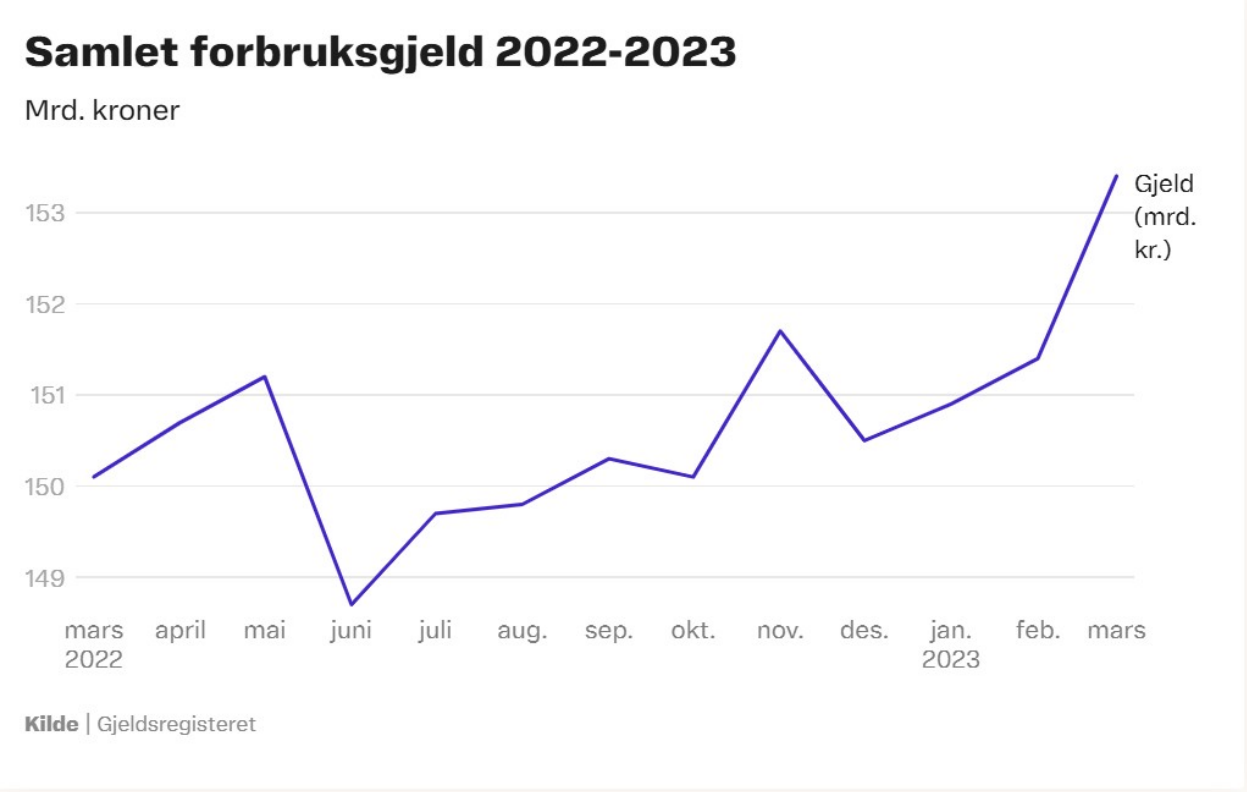
(b) Beregn sannsynligheten for at bilaget er `mistenkelig`, og klassifiser det ved hjelp av en klassifiseringsgrense på 0.5.

Tabellen under viser verdien av variabelen `mistenkelig` for de 10 bilagene i datasettet som ligger nærmest bilaget i b), målt ved euklidisk avstand mellom de standardiserte forklaringsvariablene til det aktuelle bilaget og hver av de 10 bilagene. Denne avstanden er også oppgitt i tabellen.

mistenkelig	Avstand
1	0.42
0	0.48
1	0.53
1	0.60
0	0.61
1	0.65
0	0.70
1	0.72
0	0.76
0	0.78

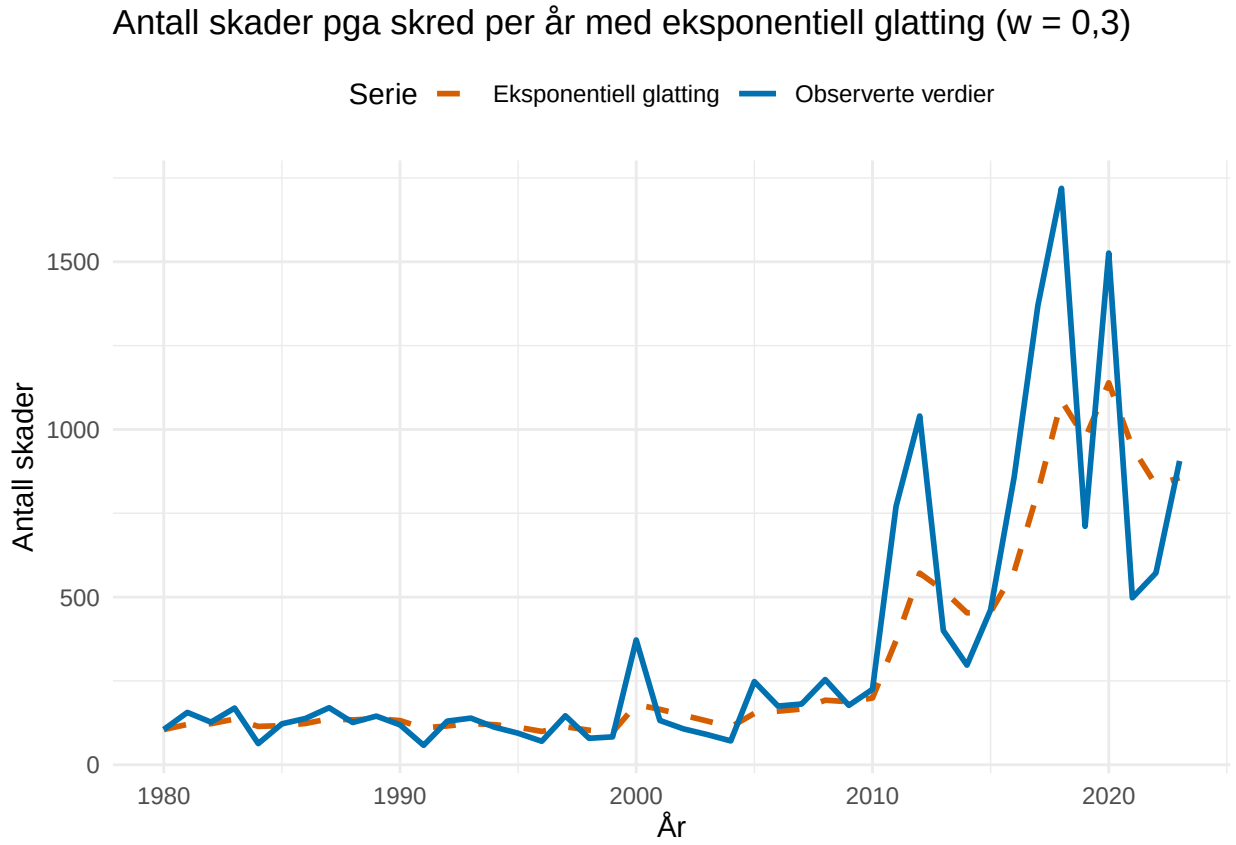
(c) **Bruk en KNN-modell med $K = 5$ til å klassifisere bilaget som `mistenkelig` eller `ikke mistenkelig`. Hvilken modell (KNN eller logistisk regresjon) vil du anbefale dersom revisjonsteamet ønsker å forstå hva som kjennetegner juks med regnskapet?**

Vedlegg 1 - Figur fra E24-artikkel



Figur 1: Brukt i E24-artikkelen "Solid oppgang i forbruksgjelden" (november 2023)

Vedlegg 2 - Skader på grunn av skred



Vedlegg 3 - Regnskapsdata

Tabell 3: Variabelbeskrivelse for datasettet `bilagdata`

Variabel	Beskrivelse
`mistenkelig`	1 = bilaget ble senere flagget som mistenkelig av revisor, 0 = ikke flagget
`beløp`	Verdien av bilaget i 1000 kr
`endringer`	Antall ganger bilaget har blitt redigert
`bilagsnummer`	Løpende nummer for bilaget i bokføringssystemet
`vedlegg`	Antall vedlegg knyttet til bilaget
`bilagstype`	Type bilag: faktura, refusjon eller intern postering

```
##
## Call:
## glm(formula = mistenkelig ~ belop + endringer + bilagsnummer +
##     vedlegg + bilagstype, family = binomial, data = bilagdata)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.1123211  0.3595787  -0.312  0.75476
## belop               0.0249895  0.0078719   3.175  0.00150 **
## endringer           0.5954645  0.1251387   4.758 1.95e-06 ***
## bilagsnummer        0.0008273  0.0004593   1.801  0.07163 .
## vedlegg            -0.3325868  0.1057391  -3.145  0.00166 **
## bilagstypeintern    0.8702731  0.3845355   2.263  0.02362 *
## bilagstyperefusjon  0.0505083  0.2875427   0.176  0.86057
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 433.82  on 399  degrees of freedom
## Residual deviance: 373.15  on 393  degrees of freedom
## AIC: 387.15
##
## Number of Fisher Scoring iterations: 5
```