UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1000 — Innføring i anvendt statistikk
Eksamensdag:	Mandag 3. desember 2018.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 6 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Lærebok Moore, McCabe & Craig: Introduction to the practice of statistics. Det er tillatt å notere i læreboka, men ikke lov å klistre inn notater. Ordliste. Kapittel 14, som deles ut. Godkjent kalkulator.

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

Det er fire oppgaver med til sammen ti delspørsmål. Hvert delspørsmål teller likt.

Oppgave 1. Vi tenker oss at total endring i global gjennomsnittstemperatur de neste 50 år kan skrives som en sum $W = X + Y$, der $X \sim N(0, \sigma_X)$ er endring som følge av naturlige klimasvingninger og $Y \sim N(\mu_Y, \sigma_Y)$ er endring som følge av menneskeskapt klimagass-utslipp. Endringen som følge av naturlige svingninger har altså forventing 0 og standardavvik σ_X. Parameterne μ_Y og σ_Y representerer forventning og standardavvik for effekten av menneskeskapt klimagass-utslipp.
Vi antar at X og Y er uavhengige tilfeldige variable. Forutsatt at klimagass-utslippene holdes på dagens nivå, er det gitt fra klimamodeller at $\mu_Y = 2°C$ og $\sigma_Y = 1.5°C$. Vi antar videre at $\sigma_X = 0.5°C$.

a) Finn sannsynligheten for at X er større enn $1°C$.

b) Beregn sannsynligheten for at total endring i global gjennomsnittstemperatur på 50 år (W) er større enn $3°C$.

Oppgave 2. *Cyrtodiopsis dalmanni* er en type flue med øyne på stilk ut til hver side av hodet. Hunn-fluene velger hanner å pare seg med på bakgrunn av avstanden mellom øynene.
Et eksperiment utføres for å undersøke om fluenes diett påvirker lengden på øyestilkene, dvs. hvor langt fra hverandre øyene står. En gruppe hann-fluer ble alet opp på mais (diett av høy kvalitet) og en gruppe hann-fluer ble alet opp på bomullsfiber (diett av lav kvalitet). Hver flue vokste opp isolert fra de andre, slik at vi kan betrakte dem som uavhengige.

(Fortsettes på side 2.)

Avstanden mellom øynene (i mm) ble målt for hver voksen hann-flue. Data ble samlet inn for 21 fluer med mais-diett og 24 fluer med bomulls-diett. Resultatene er lagret i vektor x1 (mais) og x2 (bomull), som oppsummeres i R-utskriften nedenfor. Du kan bruke utskriften fra R når du besvarer oppgavene nedenfor.

a) Formuler antakelser og hypoteser du vil bruke til å vurdere om det er grunnlag for å påstå at det er forskjell på den forventede avstanden mellom øynene hos fluer som har spist mais og fluer som har spist bomullsfiber. Skriv ned hvilken test-observator du vil bruke for å teste hypotesene, og begrunn svaret. Hvilken fordeling har test-observatoren når nullhypotesen er riktig?

b) I R-utskriften ligger det resultater fra to hypotese-tester. Velg den som korresponderer med svaret ditt ovenfor, og finn verdien på test-observator og tilhørende p-verdi fra utskriften. Skriv en konklusjon.

c) Beregn så et 95% konfidensintervall for forskjellen mellom forventet avstand mellom øynene hos fluer som spiser mais og fluer som spiser bomullsfiber. Dette intervallet er erstattet med spørsmålstegn i utskriften.

```
(Uskriften er noe redigert)

> summary(x1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.890   2.000   2.050   2.047   2.110   2.150
> summary(x2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.050   1.338   1.545   1.543   1.650   2.120
 > c(sd(x1), var(x1))
[1] 0.07471 0.00558
> c(sd(x2), var(x2))
[1] 0.28493 0.08119

> t.test(x1,x2)

Welch Two Sample t-test

data:  x1 and x2
t = 8.3477, df = 26.568, p-value = 6.666e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 ??? ???
sample estimates:
mean of x mean of y
 2.047143  1.542917
```

(Fortsettes på side 3.)

```
> t.test(x1,x2,var.equal = TRUE)

Two Sample t-test

data:  x1 and x2
t = 7.866, df = 43, p-value = 7.345e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 ??? ???
sample estimates:
mean of x mean of y
 2.047143  1.542917
```

Oppgave 3. Det er vanlig å modellere "potensialet" (sannsynligheten) for å finne et bestemt mineral i en gitt lokasjon ved hjelp av logistisk regresjon. Forklaringsvariable er ulike geologiske karakteriseringer av lokasjonen. For innsamlede data, er responsen 1 hvis mineralet er tilstede, 0 ellers.

La oss for enkelthets skyld si at vi har bare en forklaringsvariabel, x. ML-estimering fra data gir at estimatet for β_1 i den logistiske modellen er $b_1 = 0.78$ og standardfeilen er $SE_{b_1} = 0.09$. Finn et estimat for odds-ratioen for å finne mineralet for en enhets økning i x, og gi en kort tolkning av resultatet. Finn så et 95% konfidensintervall for odds-ratioen.

Oppgave 4. Lysten på iskrem varierer muligens med været. Et datasett fra en stat i USA består av observasjoner samlet gjennom 30 ulike uker, spredd over et par år. For hver av ukene har vi en observasjon av gjennomsnittlig is-konsum per innbygger (i liter per uke), sammen med gjennomsnittlig temperatur i samme uke (i grader Celsius). De 30 dataparene er plottet i figuren nedenfor, sammen med en minste kvadraters regresjonslinje beregnet ved hjelp av R.

La y_i være gjennomsnittlig is-konsum i uke i, og x_i gjennomsnittstemperaturen i uke i, $i = 1, 2, \cdots, 30$. En enkel lineær regresjonsmodell for sammenhengen mellom disse variablene er

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

$i = 1, 2, \cdots, 30$. Vi antar at ϵ_i-ene er uavhengige og normalfordelte med forventning 0 og samme ukjente standardavvik σ.

Resultatet av en enkel lineær regresjonsanalyse i R og to diagnostiske plott finner du til slutt i denne oppgaven. En av verdiene i utskriften er slettet og erstattet med spørsmålstegn. Du kan bruke utskriften fra R når du besvarer oppgavene nedenfor.

a) Hva er korrelasjonen mellom ukentlig gjennomsnittstemperatur x og gjennomsnittlig is-konsum y? Finn minste kvadraters estimater b_0 og b_1 for β_0 og β_1. Forklar hvordan du tolker verdien b_1. Finn et estimat for standardavviket σ.

(Fortsettes på side 4.)

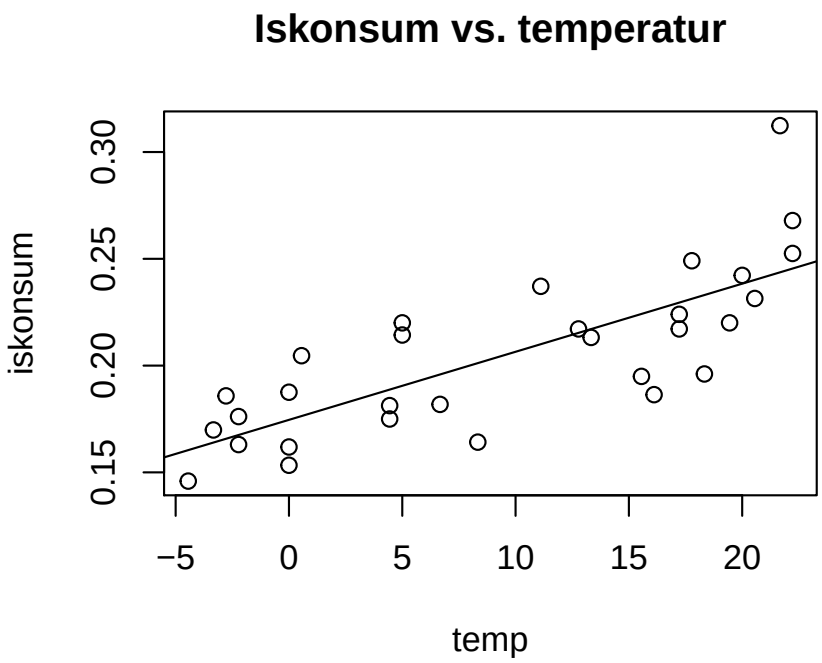
b) Diskuter kort hvor godt modellen passer til data på bakgrunn av utskrift og figurer nedenfor. Hva betyr det at størrelsen Multiple R-squared er 0.6016?

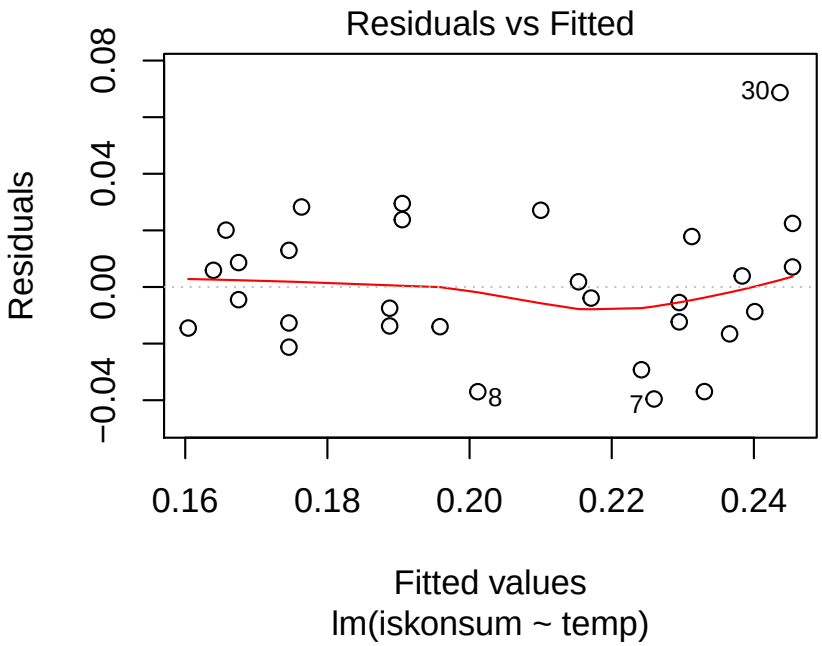
c) Finn estimert standardavvik for b_1 ved hjelp av andre verdier i utskriften og bruk dette til å beregne et 95% konfidensintervall for β_1. Hvordan kan du bruke konfidensintervallet til å teste hypotesene $H_0 : \beta_1 = 0$ mot $H_a : \beta_1 \neq 0$?

d) Hvor mye is predikerer vi at vil konsumeres per innbygger i en uke med gjennomsnittstemperatur på 27 grader Celsius i USA-staten vi har data for? Finn et 95% prediksjonsintervall for predikert is-konsum i en uke med 27 grader for å svare på dette. Kommenter kort om gyldigheten av prediksjon ved en gjennomsnittstemperatur på 27 grader basert på de foreliggende data.

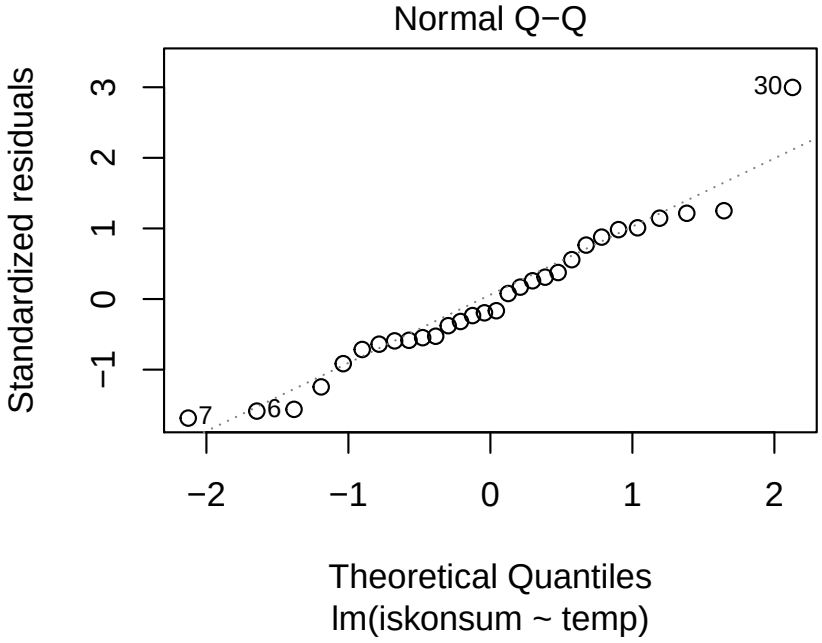
```
(Utskriften er noe redigert)

Call: lm(formula = iskonsum ~ temp)

Residuals:
   Min        1Q    Median      3Q     Max
-0.0396  -0.0139  -0.0042   0.0166   0.0686

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.17459   0.00641  27.253  < 2e-16 ***
temp        0.00319     ?????   6.502 4.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0241 on 28 degrees of freedom
Multiple R-squared:  0.6016,Adjusted R-squared:  0.5874
F-statistic: 42.28 on 1 and 28 DF,  p-value: 4.789e-07

 > predict(lm(iskonsum~temp),newdata=data.frame(temp=27), interval='confidence')
        fit      lwr       upr
1 0.2607 0.2409 0.2804

> predict(lm(iskonsum~temp),newdata=data.frame(temp=27), interval='prediction')
        fit       lwr       upr
 1 0.2607 0.2075 0.3138
```

(Fortsettes på side 5.)





(Fortsettes på side 6.)


Normal Q−Q
Standardized residuals
−1
0
1
2
3
30
7
6
−2
−1
0
1
2
Theoretical Quantiles
lm(iskonsum ~ temp)

SLUTT