UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK2120 — Statistiske metoder og dataanalyse 2.
Eksamensdag:	Mandag 13. juni 2016.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 7 sider.	
Vedlegg:	Tabell over normalfordeling, t-fordeling, χ^2-fordeling og F-fordeling.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamlinger for STK1100/STK1110 og STK2120.

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

Oppgave 1

Menneskene har hatt terninger i flere tusen år. De eldste terningene var skåret ut av et bein fra sauefoten, eller de var lagd av stein, metall eller keramikk. Vi regner i dag med at sannsynligheten er 1/6 for hvert av de seks mulige utfalene når vi kaster en terning. Men var det tilfellet også for de terningene som ble brukt i tidligere tider?

På British Museum i London fins det flere terninger fra romertiden, blant annet en terning av marmor og en terning av jern. Det har blitt gjort forsøk med å kaste disse terningene[1]. Da marmorterningen ble kastet 204 ganger, ble resultatet som gitt i tabellen nedenfor. Her er N_i antall kast som ga i øyne på terningen.

Antall øyne (i):	1	2	3	4	5	6
Antall kast (N_i):	27	28	23	47	25	54

La p_i være sannsynligheten for i øyne. Vi vil teste nullhypotesen

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6 \tag{1}$$

mot den alternative hypotesen at ikke alle p_i-ene er like.

[1] David, F. N. (1955). Studies in the history of probability and statistics. I. Dicing and gaming (a note on the history of probability). *Biometrika* **42**, 1-15.

(Fortsettes på side 2.)

For å teste (1), kan vi bruke kji-kvadrat observatoren

$$\chi^2 = \sum_{i=1}^{6} \frac{(N_i - E_i)^2}{E_i}, \tag{2}$$

der E_i-ene er de forventete antall under nullhypotesen.

a) Forklar hvordan du kan finne E_i-ene og bestem dem for marmorterningen. For marmorterningen får kji-kvadrat observatoren (2) verdien 25.17. Hvilken konklusjon kan du trekke av det?

Terningen av jern har også blitt kastet 204 ganger. Det ga følgende resultat:

Antall øyne (i):	1	2	3	4	5	6
Antall kast (N_i):	35	39	30	21	37	42

b) For jernterningen får kji-kvadrat observatoren (2) verdien 8.35. Bestem P-verdien så godt det lar seg gjøre fra vedlagte tabell. Hvilken konklusjon vil du trekke for jernterningen?

For å teste nullhypotesen (1), er det vanlig å bruke kji-kvadrat observatoren (2). Men det er ikke den eneste muligheten.

c) Beskriv en annen test du kan bruke for å teste nullhypotesen (1). Det holder her at du gir en formel for testobservatoren og angir hvilken fordeling den har under nullhypotesen. Du trenger ikke bestemme verdien av testobservatoren for de to terningene.

Oppgave 2

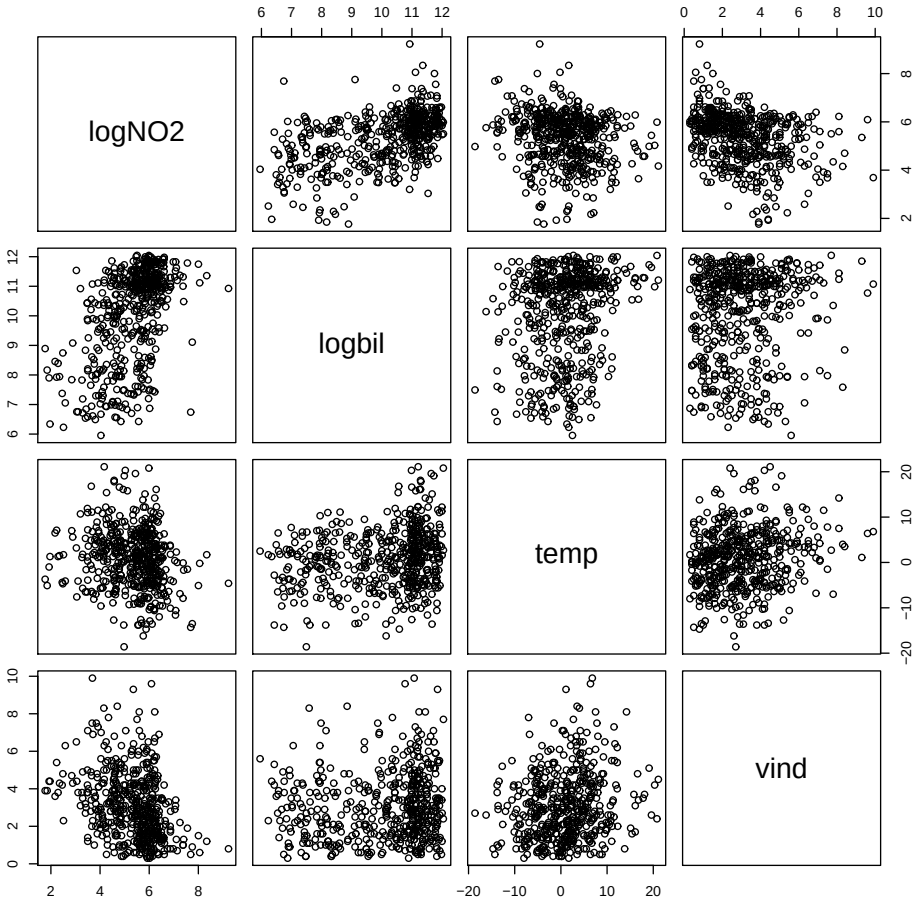
I denne oppgaven vil vi se nærmere på hvordan luftforurensning avhenger av trafikktetthet og værforhold. Konkret vil vi studere hvordan konsentrasjonen av nitrogendioksid (NO_2) ved en målestasjon på Alnabru i Oslo avhenger biltrafikk, temperatur og vindstyrke.

Vi vil se på 500 målinger av følgende variabler:

- `logNO2`: logaritmen *med grunntall 2* til konsentrasjonen av NO_2 ($\mu g/m^3$)
- `logbil`: logaritmen *med grunntall 2* til antall biler per time
- `temp`: temperatur to meter over bakken (grader Celsius)
- `vind`: vindstyrke (m/s)

(Fortsettes på side 3.)

Matriseplottet gir en oversikt over dataene:



Nedenfor er det gitt resultatet av en lineær regresjonanalyse med `logNO2` som responsvariabel og `logbil`, `temp` og `vind` som forklaringsvariabler:

```
             Estimate Std. Error t value
(Intercept)   1.899296   0.240993   7.881
logbil        0.409026   0.023384  17.492
temp         -0.038155   0.005673       ?
vind         -0.211491        ?  -10.359

Residual standard error: 0.8007 on ? degrees of freedom
Multiple R-squared:  0.4566
```

a) Beskriv de antagelsene regresjonsanalysen bygger på. Ser du noen problemer med noen av disse antagelsene for studien av luftforurensningsdataene?

b) Forklar hva de ulike delene av utskriften ovenfor står for, og fyll inn de tallene som skal stå der det er gitt spørsmålstegn i utskriften.

(Fortsettes på side 4.)

c) Beskriv hvilken betydning biltrafikk, temperatur og vindstyrke har for konsentrasjonen av NO_2. (Det er ikke nok å angi om en forklaringsvariabel er signifikant eller ikke. Du må også angi hvor stor effekt hver av forklaringsvariablene har.)

Vi ønsker å finne en modell som best mulig beskriver hvordan luftforurensningen avhenger av biltrafikk, temperatur og vindstyrke. Vi gjør det ved å tilpasse regresjonsmodeller der forklaringsvariablene velges blant:

(i) de tre variablene `logbil`, `temp` og `vind`,

(ii) kvadratene av de tre variablene, gitt som `I(logbil^2)`, `I(temp^2)` og `I(vind^2)` i R-utskriften nedenfor,

(iii) interaksjonene mellom de tre variablene, gitt som `logbil:temp`, `logbiler:vind` og `temp:vind` i R-utskriften nedenfor.

Tilsammen gir dette ni mulige forklaringsvariabler.

For hver $m = 1, 2, \ldots, 9$ tilpasser vi alle mulige regresjonsmodeller med m forklaringsvariabler og velger den av dem som har minst residualkvadratsum (kvadratsum for feil).

Vi får da følgende sekvens av modeller:

```
> fit.exhaustive=regsubsets(logNO2~logbil+temp+vind+I(logbil^2)+I(temp^2)+I(vind^2)+
                 logbil:temp+logbil:vind+temp:vind,data=no2,nvmax=9,method="exhaustive")
> summary.exhaustive=summary(fit.exhaustive)
> summary.exhaustive

.... (redigert output) ....

Selection Algorithm: exhaustive
    logbil temp vind I(logbil^2) I(temp^2) I(vind^2) logbil:temp logbil:vind temp:vind
1   " "    " "  " "  "*"         " "       " "       " "         " "         " "
2   " "    " "  "*"  "*"         " "       " "       " "         " "         " "
3   " "    "*"  "*"  "*"         " "       " "       " "         " "         " "
4   " "    "*"  "*"  "*"         " "       "*"       " "         " "         " "
5   "*"    "*"  "*"  "*"         " "       "*"       " "         " "         " "
6   "*"    "*"  "*"  "*"         " "       "*"       " "         "*"         " "
7   "*"    "*"  "*"  "*"         " "       "*"       "*"         "*"         " "
8   "*"    "*"  "*"  "*"         " "       "*"       "*"         "*"         "*"
9   "*"    "*"  "*"  "*"         "*"       "*"       "*"         "*"         "*"
```

For disse ni modellene får vi følgende verdier av R^2, justert R^2 og Mallows C_p:

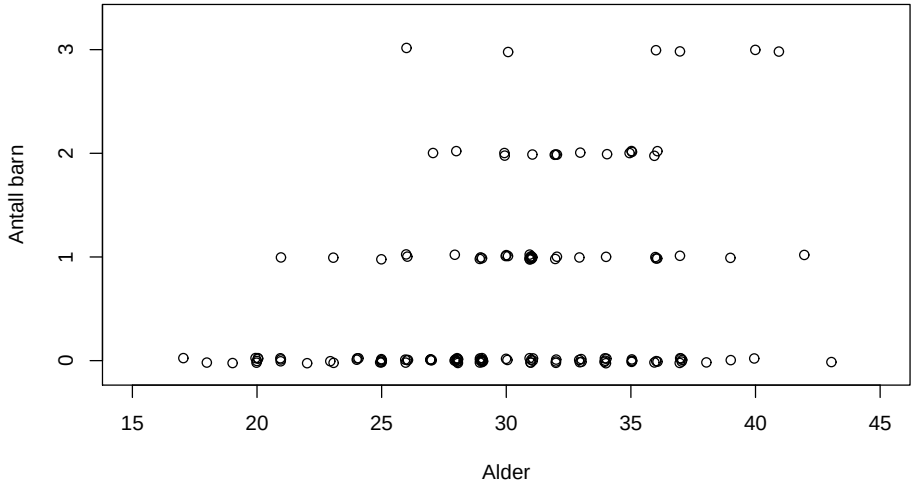
```
> summary.exhaustive$rsq
      0.2662  0.4137  0.4644  0.4925  0.4987  0.5057  0.5069  0.5078  0.5078
> summary.exhaustive$adjr2
      0.2648  0.4113  0.4612  0.4884  0.4936  0.4997  0.4999  0.4998  0.4988
> summary.exhaustive$cp
      234.52   89.70   41.21   15.28   11.08    6.07    6.94    8.04   10.00
```

d) Bruk resultatene ovenfor til å avgjøre hvilke forklaringsvariabler du vil ta med i regresjonsmodellen. Begrunn svaret ditt.

(Fortsettes på side 5.)

Oppgave 3

Ved et sykehus har en registrert hvor mange barn 140 gravide kvinner har fra før og alderen til kvinnene (i hele år). Figuren nedenfor viser et plott av dataene. (For å unngå at mange punkter kommer helt oppå hverandre, er et lite tilfeldig tall lagt til aldrene og antall barn før de er plottet.)



Vi ønsker å studere hvordan antall barn avhenger av alderen til kvinnene. Vi vil da anta at antall barn en kvinne har, er Poisson fordelt med en parameter som avhenger av alderen til kvinnen. Men før vi ser nærmere på sammenhengen mellom kvinnenes alder og antall barn, vil vi se generelt på en regresjonsmodell for Poissonfordelte data.

Vi har uavhengige stokastiske variabler $Y_1, Y_2, \ldots, Y_n$ og tilhørende forklaringsvariabler $x_1, x_2, \ldots, x_n$. Vi vil anta at Y_i er Poissonfordelt med en parameter $\lambda(x_i) = E(Y_i \,|\, x_i)$ som avhenger x_i; $i = 1, 2, \ldots, n$. Spesielt vil vi anta at $\lambda(x_i)$ har formen

$$\lambda(x_i) = e^{\beta_0 + \beta_1 x_i}. \tag{3}$$

a) Vis at likelihood funksjonen kan skrives som

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{i=1}^{n} \frac{\lambda(x_i)^{y_i}}{y_i!} e^{-\lambda(x_i)} \\ &= \left(\prod_{i=1}^{n} y_i! \right)^{-1} \exp\left\{ \beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} e^{\beta_0 + \beta_1 x_i} \right\}, \end{aligned}$$

der $y_1, y_2, \ldots, y_n$ er de observerte verdiene av Y_i-ene.

(Fortsettes på side 6.)

La $l(\beta_0, \beta_1) = \log L(\beta_0, \beta_1)$ være log-likelihood funksjonen. Da er vektoren av score-funksjoner gitt ved

$$\mathbf{s}(\beta_0, \beta_1) = \begin{bmatrix} s_0(\beta_0, \beta_1) \\ s_1(\beta_0, \beta_1) \end{bmatrix},$$

der

$$s_j(\beta_0, \beta_1) = \frac{\partial}{\partial \beta_j} l(\beta_0, \beta_1) \quad \text{for} \quad j \in \{0, 1\}.$$

b) Uttrykk score-funksjonene ved hjelp av y_i-ene, x_i-ene, β_0 og β_1.

Den observerte informasjonsmatrisen er gitt ved

$$\mathbf{J}(\beta_0, \beta_1) = \begin{bmatrix} J_{00}(\beta_0, \beta_1) & J_{01}(\beta_0, \beta_1) \\ J_{10}(\beta_0, \beta_1) & J_{11}(\beta_0, \beta_1) \end{bmatrix},$$

der

$$\begin{aligned} J_{00}(\beta_0, \beta_1) &= -\frac{\partial^2}{\partial \beta_0^2} l(\beta_0, \beta_1), \\ J_{01}(\beta_0, \beta_1) &= J_{10}(\beta_0, \beta_1) = -\frac{\partial^2}{\partial \beta_1 \partial \beta_0} l(\beta_0, \beta_1), \\ J_{11}(\beta_0, \beta_1) &= -\frac{\partial^2}{\partial \beta_1^2} l(\beta_0, \beta_1). \end{aligned}$$

c) Vis at

$$J_{jk}(\beta_0, \beta_1) = \sum_{i=1}^{n} x_i^{j+k} \, e^{\beta_0 + \beta_1 x_i} \quad \text{for} \quad j, k \in \{0, 1\}.$$

Merk at $\mathbf{J}(\beta_0, \beta_1)$ ikke avhenger av y_i-ene. Fishers informasjonsmatrise $\mathbf{I}(\beta_0, \beta_1)$ er derfor lik den observerte informasjonsmatrisen.

Vi kan ikke finne eksplisitte uttrykk for maksimum likelihood estimatene $\hat{\beta}_0$ og $\hat{\beta}_1$, så vi må finne dem ved numerisk optimering av log-likelihooden.

d) Forklar hvordan du kan finne maksimum likelihood estimatene ved å bruke Newton-Raphsons metode.

Vi ser så på dataene om hvordan antall barn avhenger av alderen til kvinnene.
Vi tilpasser modellen (3) med

x_i = alder til den i-te kvinnen $- 30$.

(Fortsettes på side 7.)

Vi får da maksiumum likelihood estimatene:

$$\hat{\beta}_0 = -0.7006 \quad \text{og} \quad \hat{\beta}_1 = 0.0899,$$

mens den inverse informasjonsmatrisen blir

$$\mathbf{J}(\beta_0, \beta_1)^{-1} = \begin{bmatrix} 0.015883 & -0.001238 \\ -0.001238 & 0.000500 \end{bmatrix}.$$

e) Bestem de estimerte standardfeilene til maximum likelihood estimatene og lag et tilnærmet 95% konfidensintevall for β_1. Kommenter spesielt hvilket resultat om fordelingen av maksimum likelihood estimatorene du benytter deg av når du lager konfidensintervallet.

f) Vis at for modellen (3) er $\{\lambda(x+1)-\lambda(x)\}/\lambda(x) = e^{\beta_1}-1$. Det betyr at $\theta = e^{\beta_1}-1$ angir den *relative* økningen i forventet antall barn når alderen til en kvinne øker med ett år. Bestem et 95% konfidensintervall for θ.

SLUTT