UNIVERSITETET I OSLO

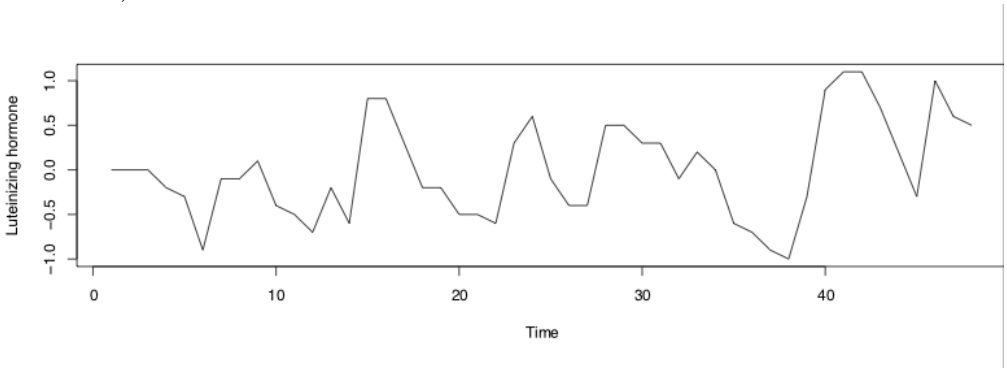
Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1120 — Statistiske metoder og dataanalyse 2.
Eksamensdag:	Fredag 2. juni 2006.
Tid for eksamen:	09.00 – 12.00.
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabeller for χ^2, t og F fordelingene.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/ STK1110 og for STK1120.

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

Oppgave 1.

Datasettet plottet nedenfor viser "luteinizing hormone" i blodsampler på 10 minutters intervaller fra en kvinne, 48 observasjoner (fratrukket gjennomsnittsverdi).



Det som er spesielt med disse data er at påfølgende observasjoner er *korrelerte*. Hvis vi lar x_i være observasjon i tidsintervall i, så er den empiriske korrelasjonen mellom x_i og x_{i+1} lik 0.58. For å analysere slike data må en da lage en modell som tar hensyn til korrelasjonene mellom observasjonene.
En mulig slik modell er den *autoregressive* modellen av orden 2:

$x_i = a_1 x_{i-1} + a_2 x_{i-2} + \varepsilon_i, i = 1, ..., n$

(Fortsettes side 2.)

der $\{\varepsilon_i, i = 1, ..., n\}$ er uavhengige normalfordelte variable med forventning 0 og varians σ^2. En annen måte å formulere dette på er at gitt tidligere observasjoner så er x_i normalfordelt med forventning $a_1 x_{i-1} + a_2 x_{i-2}$ og varians σ^2. Dette er et eksempel på en mer generell klasse av modeller som vi kaller *tidsseriemodeller*.
Vår interesse vil ligge i å gjøre inferens mhp a_1, a_2 og σ^2. For enkelthets skyld vil vi anta at $x_0 = x_{-1} = 0$.

(a) Vis at likelihood funksjonen er gitt ved

$$L(a_1, a_2, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i - a_1 x_{i-1} - a_2 x_{i-2})^2}$$

og bruk dette til å vise at maksimering av L mhp (a_1, a_2) er ekvivalent med minimering av

$$S(a_1, a_2) = \sum_{i=1}^{n} (x_i - a_1 x_{i-1} - a_2 x_{i-2})^2$$

Hint: Du kan bruke at sannsynlighetstettheten $f(x_1, x_2, x_3, ..., x_n)$ for data $x_1, ..., x_n$ kan skrives som

$$f(x_1, x_2, x_3, ..., x_n) = f(x_1) f(x_2|x_1) f(x_3|x_1, x_2) \cdots f(x_n|x_1, ..., x_{n-1}).$$

(b) Sett opp et sett av likninger som maksimum likelihood estimatene $(\hat{a}_1, \hat{a}_2)$ for (a_1, a_2) må tilfredsstille.

Argumentér for at disse estimatene kan beregnes direkte uten å ty til iterative metoder som Newton-Raphson (du behøver ikke å utføre selve utregningene).

Finn også maksimum likelihood estimatet for σ^2 uttrykt som en funksjon av data og $(\hat{a}_1, \hat{a}_2)$.

Basert på de foreliggende data, får en at

$\hat{a}_1 = 0.7110, \hat{a}_2 = -0.2220, \hat{\sigma}^2 = 0.1880$

og

$$\bar{J}(\hat{a}_1, \hat{a}_2, \hat{\sigma}^2)^{-1} = \begin{pmatrix} 0.0199 & -0.0115 & 0.0000 \\ -0.0115 & 0.0204 & 0.0000 \\ 0.0000 & 0.0000 & 0.0015 \end{pmatrix}.$$

der $\bar{J}(\hat{a}_1, \hat{a}_2, \hat{\sigma}^2)^{-1}$ er den inverse matrisen til den observerte informasjonsmatrisen.

(c) Konstruer (tilnærmede) 95% konfidensintervaller for a_1 og a_2.

En alternativ modell til den vi har antatt så langt er

$x_i = a_1 x_{i-1} + \varepsilon_i, i = 1, ..., n.$

Basert på de konfidensintervall du har konstruert, vil en slik forenkling av modellen være rimelig?

(Fortsettes side 3.)

Oppgave 2.

Dyr kategoriseres ofte etter egenskaper de har. Vi vil se på et datasett der 197 dyr er kategorisert i forhold til 2 ulike egenskaper, hver med to mulige kategorier. Tilsammen får vi da 4 ulike kombinasjoner av egenskaper. For det gitte datasettet er dyrene fordelt over de 4 kombinasjonene som følger:

$Y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$.

Vi vil anta observasjonene er multinomisk fordelt med cellesannsynligheter gitt ved

$(p_1, p_2, p_3, p_4) = \left(\frac{1}{2} + \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta\right)$

(en rimelig modell basert på genetisk teori der en av kategoriene er dominant i forhold til den andre for hver av egenskapene).

(a) Vis at maksimum likelihood estimatet for θ må tilfredsstille likningen

$$n\theta^2 - (y_1 - 2(y_2 + y_3) - y_4)\theta - 2y_4 = 0.$$

For de observerte data gir dette to mulige løsninger, -0.55068 og 0.62682. Argumentér for at det er kun den siste løsningen som er aktuell i dette tilfellet.

For å si noe om egenskapene til maksimum likelihood estimatet $\hat{\theta}$ for θ, ble 1000 bootstrap simuleringer utført. En bootstrap-simulering av $\hat{\theta}$ ble utført ved å trekke med tilbakelegging fra de 197 dyrene og deretter telle hvor mange dyr som falt innenfor de 4 ulike kombinasjone. Dette ga

$$\begin{aligned}
\bar{\theta}^* &= \frac{1}{1000}\sum_{b=1}^{1000}\theta_b^* = 0.62822 \\
s_{\theta^*} &= \sqrt{\frac{1}{1000-1}\sum_{b=1}^{1000}(\theta_b^* - \bar{\theta}^*)^2} = 0.0516 \\
\theta^*_{(25)} &= 0.5240 \quad \text{(dvs den 25. minste verdien av } \theta^*\text{-ene)} \\
\theta^*_{(975)} &= 0.7264 \quad \text{(dvs den 975. minste verdien av } \theta^*\text{-ene)}
\end{aligned}$$

der θ_b^* er bootstrap simulering nr. b av $\hat{\theta}$.

(b) Basert på disse resultatene, er det grunn til å tro at $\hat{\theta}$ er veldig forventningsskjev?

Konstruer 95% konfidensintervaller for θ både basert på normaltilnærmingen og ved standard Bootstrap intervaller.

Kommenter resultatene.

(Fortsettes side 4.)

(c) Definer nå nullhypotesen

$$H_0 : p = (p_1, p_2, p_3, p_4) = \left(\tfrac{1}{2} + \tfrac{1}{4}\theta, \tfrac{1}{4}(1-\theta), \tfrac{1}{4}(1-\theta), \tfrac{1}{4}\theta\right)$$

Beskriv hvordan du kan teste denne hypotesen.

Utfør testen og skriv en konklusjon.

Hint: Du kan her bruke at $(n\hat{p}_1, n\hat{p}_2, n\hat{p}_3, n\hat{p}_4) = (129.37, 18.38, 18.38, 30.87)$ der $\hat{p}_j$ er p_j med $\hat{\theta}$ innsatt som estimat for θ.

Oppgave 3.

Box, Hunter og Hunter (1976) beskriver et to-faktor eksperiment der alle kombinasjoner av 3 gifter (`poison`) og 4 motgifter (`antidote`) blir studert. Fire replikasjoner (på laboratoriedyr) er tilfeldig allokert til hver av behandlingskombinasjonene. Overlevelsestidene er vist i tabellen på slutten av oppgaven. En ANOVA analyse av disse dataene gå følgende resultat:

```
                Df  Sum Sq Mean Sq F value
poison           2 1.03708 0.51854 23.3314
antidote         3 0.92012 0.30671 13.8000
poison:antidote  6 0.25027 0.04171  1.8768
Residuals       36 0.80010 0.02222
```

Sett opp den modell som ligger bak denne analysen.
Formuler aktuelle hypoteser som kan testes på disse dataene.
Utfør testene og gi en konklusjon.

	Antidote			
Poison	A	B	C	D
1	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
2	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
3	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.35
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

SLUTT