UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK2120 — Statistiske metoder og dataanalyse 2
Eksamensdag:	Torsdag 3. juni 2010
Tid for eksamen:	09.00 – 12.00
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabell over normalfordeling og t-fordeling
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og STK1120/STK2120

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

Oppgave 1

Dataene i tabell 1 nedenfor viser sammenhengen mellom alder i år og lengde i millimeter for et utvalg fisk av et bestemt fiskeslag i en amerikansk innsjø.

Tabell 1: Alder og lengde for 12 fisk

Observasjon	1	2	3	4	5	6	7	8	9	10	11	12
Alder	1	1	2	2	3	3	4	4	4	5	5	6
Lengde i mm	67	62	83	91	140	150	150	140	160	160	150	170

Utskriften og plottet nedenfor viser resultatet av å tilpasse en polynomisk modell der den forventede lengden er et 2. grads polynom i alder, dvs.

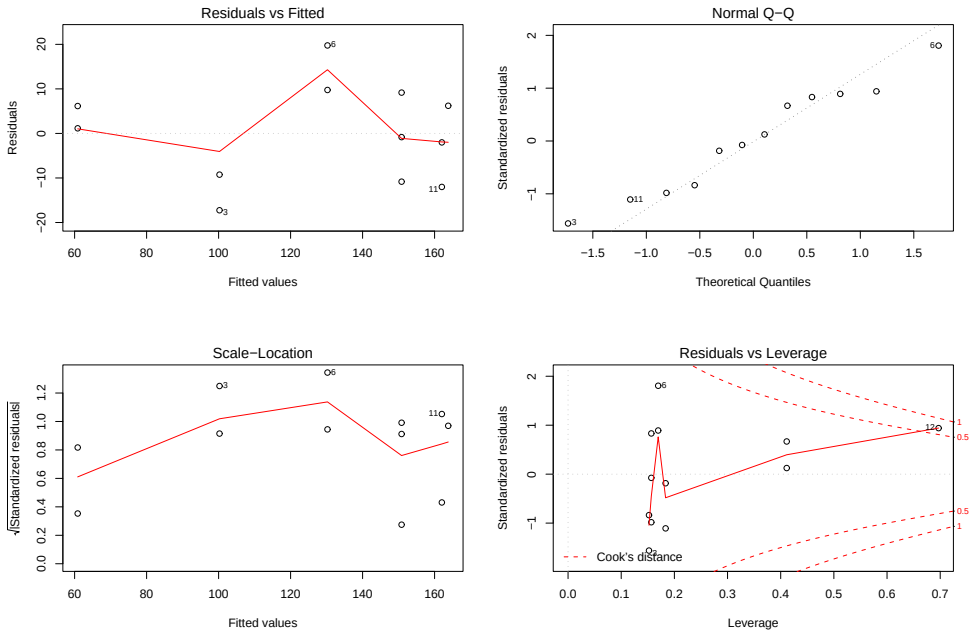
$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,\ i = 1, \ldots, 12.$$

Lengden y er her responsvariabelen og alderen x kovariaten.

```
Call:
lm(formula = Length ~ Age + I(Age^2), data = llmary)

Residuals:
     Min       1Q   Median       3Q      Max
-17.2508  -9.6459   0.1585   6.9513  19.7559

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   12.046     15.326   0.786 0.452058
Age           53.508     10.217   5.237 0.000537 ***
```

(Fortsettes på side 2.)

```
I(Age^2)      -4.703      1.504  -3.127 0.012181 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12 on 9 degrees of freedom
Multiple R-Squared: 0.9238,     Adjusted R-squared: 0.9068
F-statistic: 54.52 on 2 and 9 DF,  p-value: 9.33e-06
```



a) Hva er de vanlige forutsetningene for en modell av denne typen? Kommenter modelltilpasningen.

b) Hvordan kan hypotesen $H_0 : \beta_2 = 0$ testes mot alternativet $H_A : \beta_2 \neq 0$? Hva blir p-verdien?

Fra lignende undersøkelser finner en ofte at koeffesienten foran andregradsleddet er -3.0. Er det grunnlag for å tro noe annet i dette tilfellet? Formuler spørsmålet som et hypotesetestings-problem, og utfør testen med et signifikansnivå på 0.05.

c) Skriv modellen på matriseform. Angi spesielt hva designmatrisen, X, blir i dette tilfellet. Forklar hvorfor minste kvadraters estimatoren for $(\beta_0, \beta_1, \beta_2)'$ er forventningersrett.

d) Utled et uttrykk for kovariansmatrisen til residualene, og vis hvordan hattematrisen (the hat matrix) inngår?

e) Forklar hva sammenhengen mellom hatte-matrisen og leverage verdiene er. De beregnede leverage verdiene for den polynomiske modellen ovenfor er gjengitt i tabell 2. Kommenter disse verdiene.

(Fortsettes på side 3.)

Tabell 2: Leverage verdier for den polynomiske modellen.

Observasjon	1	2	3	4	5	6	7	8	9	10	11	12
Leverage	0.41	0.41	0.15	0.15	0.17	0.17	0.16	0.16	0.16	0.18	0.18	0.70

Oppgave 2

Tabellen 3 nedenfor inneholder et datasett om reaksjonsrater fra en kjemisk fabrikk. På en produksjonslinje har man variert bruken av katalysator og reagensmiddel for å studere stabiliteten i produksjonen. Faktoren *middel* har to nivåer og faktoren *katalysator* har tre. For hver kombinasjon av nivåer er det to observasjoner eller replikasjoner. Forsøkene er gjennomført i tilfeldig rekkefølge.

Tabell 3: Reaksjonsrater.

		Katalysator		
		I	II	III
Reagens-	I	4, 6	11, 7	5, 9
middel	II	6, 4	13, 15	9, 7

Det er nedenfor gjengitt noen verdier fra en toveis variansanalysetabell for observasjonene.

```
Analysis of Variance Table

Response: reaksjonsrate
                                    Df Sum Sq Mean Sq F value  Pr(>F)
factor(middel)                       *     12       *    3.00 0.13397
factor(katalysator)                  *     86      43    *    0.01039
factor(middel):factor(katalysator)   *     14       *    1.75 0.25193
Residuals                            6     24       4
```

a) Fyll ut de manglende verdiene i tabellen. Forklar hvorfor en modell uten samspill er tilstrekkelig.

b) En lineær modell uten samspill kan formuleres som

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},\ i = 1, 2,\ j = 1, 2, 3,\ k = 1, 2.$$

der $\sum_i \alpha_i = \sum_j \beta_j = 0$ og restleddene $\epsilon_{ijk},\ i = 1, 2,\ j = 1, 2, 3,\ k = 1, 2$ er uavhengige $N(0, \sigma^2)$ fordelte variable. Vis hvordan modellen kan uttrykkes på matriseform. Hva er designmatrisen i en slik formulering? Vær omhyggelig med å forklare hvilken kolonne som svarer til hvilken parameter, og hvilken rad som svarer til hvilken observasjon.

c) Finn minste kvadraters estimater for parametrene $\mu, \alpha_i, \beta_j,\ i = 1, 2,\ j = 1, 2, 3$, i modellen fra punkt b).

(Fortsettes på side 4.)

Oppgave 3

Den geometriske fordelingen har punktsannsynlighet

$$p(j) = (1-p)p^{j-1},\ j = 1, 2, \ldots$$

der p er en parameter, $0 < p < 1$.

Anta at $X_1, \ldots, X_n$ er uavhengige tilfeldige variable som alle er geometrisk fordelte med samme parameter p.

a) Finn sannsynlighetsmaksimeringsestimatoren (SME) for p basert på $X_1, \ldots, X_n$. Finn informasjonen, og angi den tilnærmede fordelingen til SME nå antallet observasjoner er stort..

En annen fordeling på de positive heltallene $1, 2, \ldots$ har punktsannsynlighet som kan skrives

$$p(j) = \frac{\Gamma(k+j-1)}{(j-1)!\Gamma(k)}(1+\frac{m}{k})^{-k}(\frac{m}{m+k})^{j-1},\ j = 1, 2, \ldots$$

der $m > 0$ og $k > 0$ er parametre.

b) Sett opp ligningene til bestemmelse av sannsynlighetsmaksimeringsestimatoren (SME) for m og k basert på et tilfeldig utvalg $X_1, \ldots, X_n$, og finn et uttrykk for SME for m. Forklar hvorfor denne fordelingen kan betraktes som en generalisering av den geometriske fordelingen.

SLUTT