UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK2120 — Statistiske metoder og dataanalyse 2.
Eksamensdag:	Fredag 7. juni 2013.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 5 sider.	
Vedlegg:	1) Tabell for standardnormalfordelingen. 2) Tabell for t-fordelingene. 3) Tabell for kji-kvadrat fordelingene. 4) Tabell for F-fordelingene.
Tillatte hjelpemidler:	Godkjent lommeregner og formelsamlinger for STK1100/STK1110 og STK2120.

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

Oppgave 1

En kjemiker har undersøkt hvordan reaksjonshastigheten ved dannelsen av en polymer avhenger av temperaturen og mengden av en katalysator. Reaksjonshastigheten ble målt to ganger for hver kombinasjon av temperatur og mengden av katalysator, og resultatet ble som gitt i tabellen nedenfor.

	Katalysator		
Temperatur	0.5%	0.6%	0.7%
50 °C	38, 41	45, 47	57, 59
60 °C	44, 43	56, 57	70, 69
70 °C	44, 47	56, 60	70, 67

Dataene har blitt analysert med toveis variansanalyse der temperatur (`temp`) og mengden av katalysator (`kat`) ble betraktet som kategoriske forklaringsvariabler, hver med tre nivåer. Resultatet av variansanalysen er gitt i R-utskriften nedenfor (tre av tallene i tabellen er erstattet med spørsmålstegn).

```
                          Df     Sum Sq    Mean Sq    F value
factor(temp)               2      332.1      166.1     55.352
factor(kat)                2     1520.1      760.1    253.352
factor(temp):factor(kat)   ?       38.6        9.6          ?
Residuals                  9       27.0          ?
```

(Fortsettes på side 2.)

a) Forklar hvordan du kan bestemme de tallene i variansanalysetabellen som er erstattet med spørsmålstegn.

b) Beskriv den statistiske modellen som ligger til grunn for variansanalysen og forklar hva det betyr at det er samspill (interaksjon) mellom temperatur og mengden av katalysator.

c) Beskriv hvordan vi kan teste nullhypotesen om at det ikke er noe samspill mellom temperatur og mengden av katalysator. Utfør testen og formuler hvilken konklusjon du kan trekke av den.

Oppgave 2

I denne oppgaven skal vi se nærmere på et forsøk der en ønsket å studere giftigheten av kjemikalet rotenon. Fem grupper med omtrent femti insekter i hver gruppe ble utsatt for ulike doser av rotenon, og en registrerte hvor mange av insektene som døde i hver gruppe. Resultatet av forsøket er gitt i tabellen nedenfor.

Gruppe nummer	10-logaritmen av giftdosen (x_i)	Antall insekter (n_i)	Antall døde (y_i)
1	0.41	48	6
2	0.58	48	16
3	0.71	46	24
4	0.89	49	42
5	1.01	50	44

La n_i være antall insekter i gruppe nummer i og la y_i være antallet av dem som døde. La videre x_i være 10-logaritmen til giftdosen for den i-te gruppen.

a) Forklar hvorfor det er rimelig å anta at $y_1, y_2, \ldots, y_5$ er observerte verdier av uavhengige stokastiske variabler $Y_1, Y_2, \ldots, Y_5$ som er binomisk fordelte:

$$Y_i \sim \text{bin}(n_i, p(x_i)) \,,$$

der $p(x_i)$ er sannsynligheten for at et insekt som får log-dosen x_i vil dø.

I punktene b - g vil vi anta at sammenhengen mellom $p(x_i)$ og x_i er gitt ved den logistiske regresjonsmodellen

$$p(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \tag{1}$$

b) Vis at likelihood funksjonen kan skrives som

$$L(\beta_0, \beta_1) = \prod_{i=1}^{5} \binom{n_i}{y_i} \frac{e^{\beta_0 y_i + \beta_1 x_i y_i}}{(1+e^{\beta_0+\beta_1 x_i})^{n_i}}$$

(Fortsettes på side 3.)

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

$$s_0(\beta_0, \beta_1) = \frac{\partial}{\partial \beta_0} l(\beta_0, \beta_1) \quad \text{og} \quad s_1(\beta_0, \beta_1) = \frac{\partial}{\partial \beta_1} l(\beta_0, \beta_1)\,.$$

c) 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

$$J_{jk}(\beta_0, \beta_1) = -\frac{\partial^2}{\partial \beta_j \partial \beta_k} l(\beta_0, \beta_1) \quad \text{for} \quad j, k = 0, 1.$$

d) Vis at

$$J_{jk}(\beta_0, \beta_1) = \sum_{i=1}^{5} n_i \, x_i^{j+k} \frac{e^{\beta_0 + \beta_1 x_i}}{(1 + e^{\beta_0 + \beta_1 x_i})^2} \quad \text{for} \quad j, k = 0, 1.$$

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

Vi kan ikke finne eksplisitte uttrykk for maksimum likelihood estimatorene $\widehat{\beta}_0$ og $\widehat{\beta}_1$, så derfor må estimatene finnes ved numerisk optimering av likelihooden.

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

Ut fra dataene gitt først i oppgaven finner vi maksimum likelihood estimatene

$$\widehat{\beta}_0 = -4.792 \quad \text{og} \quad \widehat{\beta}_1 = 7.011\,,$$

mens den inverse av informasjonsmatrisen blir

$$\mathbf{J}(\widehat{\beta}_0, \widehat{\beta}_1)^{-1} = \begin{bmatrix} 0.410 & -0.548 \\ -0.548 & 0.782 \end{bmatrix}.$$

f) Bestem et (tilnærmet) 95% konfidensintervall for β_1. Kommenter spesielt hvilket resultat om fordelingen av maksimum likelihood estimatorene du benytter deg av ved konstruksjon av intervallet.

g) LD50 er den giftdosen som svarer til 50% dødelighet for insektene. Finn et estimat for LD50.

(Fortsettes på side 4.)

Vi har ovenfor antatt at sannsynligheten for at et insekt som får log-dosen x_i vil dø er gitt ved den logistiske regresjonsmodellen (1). Vi vil til slutt i oppgaven undersøke om dette er en rimelig antagelse.

Tabellen nedenfor gir antall døde insekter som ble observert for hver av de fem gruppene og det antall døde insekter vi ville forvente å få hvis den logistiske regresjonsmodellen er riktig.

Gruppe nummer	Antall døde (y_i)	Forventet antall døde (e_i)
1	6	6.15
2	16	15.65
3	24	25.13
4	42	39.67
5	44	45.40

h) Forklar hvordan vi har beregnet de forventede antallene i den siste kolonnen av tabellen. Bruk tallene i tabellen til å teste nullhypotesen om at sannsynlighetene $p(x_i)$ er gitt ved den logistiske regresjonsmodellen (1).

i) Beskriv en annen test du kan bruke for å teste nullhypotesen i forrige punkt. Det holder her at du gir en formel for testobservatoren og angir hvilken fordeling den har under nullhypotesen. Du trenger ikke å bestemme verdien av testobservatoren.

Oppgave 3

Vi vil i denne oppgaven se nærmere på den multiple lineære regresjonsmodellen:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \; ; \quad i = 1, 2, \ldots, n; \tag{2}$$

der x_{ij}-ene er gitte forklaringsvariabler og ε_i-ene er uavhengige og $N(0, \sigma^2)$-fordelte.

Det er velkjent at (2) kan skrives på formen

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,,$$

der $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_n]'$, $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]'$, $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]'$ og $\mathbf{X}$ er $n \times (k+1)$ matrisen der i-te rad har elementene $1, x_{i1}, \ldots, x_{ik}$. Det er også velkjent at minste kvadraters estimator for $\boldsymbol{\beta}$ er gitt ved

$$\widehat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\mathbf{Y}\,.$$

(Fortsettes på side 5.)

a) Vis at $\widehat{\boldsymbol{\beta}}$ er forventningsrett og at kovariansmatrisen til $\widehat{\boldsymbol{\beta}}$ er gitt ved $\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{C}$, der $\mathbf{C} = [\mathbf{X}'\mathbf{X}]^{-1}$.

La $x_1^*, x_2^*, \ldots, x_k^*$ være gitte verdier av forklaringsvariablene. Vi er interessert i å estimere forventet respons svarende til disse verdiene, dvs.

$$\mu(x_1^*, x_2^*, \ldots, x_k^*) = \beta_0 + \beta_1 x_1^* + \beta_1 x_2^* + \cdots + \beta_k x_k^* \tag{3}$$

Merk at hvis vi innfører vektoren $\mathbf{x}^* = [1, x_1^*, \ldots, x_k^*]'$, kan (3) gis på formen $\mu(x_1^*, x_2^*, \ldots, x_k^*) = (\mathbf{x}^*)'\boldsymbol{\beta}$.

b) Angi en forventningsrett estimator $\widehat{\mu}(x_1^*, x_2^*, \ldots, x_k^*)$ for (3) og vis at variansen til estimatoren kan skrives som $V(\widehat{\mu}(x_1^*, x_2^*, \ldots, x_k^*)) = \sigma^2(\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*$, der $\mathbf{C}$ er gitt i punkt a.

c) Bestem fordelingen til

$$\frac{\widehat{\mu}(x_1^*, x_2^*, \ldots, x_k^*) - \mu(x_1^*, x_2^*, \ldots, x_k^*)}{S\sqrt{(\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*}},$$

der S^2 er den vanlige forventningsrette estimatoren for σ^2. Utled et $100(1-\alpha)\%$ konfidensintervall for (3).

La Y^* være en ny observasjon fra den lineære regresjonsmodellen (2) svarende til verdiene $x_1^*, x_2^*, \ldots, x_k^*$ av forklaringsvariablene. Vi har altså at

$$Y^* = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_k x_k^* + \varepsilon^*\,,$$

der ε^* er $N(0, \sigma^2)$-fordelt og uavhengig av $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$.

d) Bestem et $100(1-\alpha)\%$ prediksjonsintervall for Y^*.

SLUTT