UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK2120 — LØSNINGSFORSLAG
Eksamensdag:	7. juni 2013.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 8 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

a) Vi finner de tallene som er erstattet med spørsmålstegn på følgende måte:

- Det er $I = 3$ nivåer for temperatur og $J = 3$ nivåer for katalysator. Antall frihetsgrader for samspillet (interaksjonen) mellom temperatur og katalysator er da $(I-1)(J-1) = 2 \cdot 2 = 4$. Alternativt kan vi benytte at det totale antall frihetsgrader er $n - 1 = 18 - 1 = 17$, slik at antall frihetsgrader for samspillet er $17 - 2 - 2 - 9 = 4$.
- Kvadratsummen for residualene er $SSE = 27.0$ med 9 frihetsgrader. Middelkvadratsummen for residualene er da $MSE = 27.0/9 = 3.0$.
- F-observatoren for samspill mellom temperatur (faktor A) og katalysator (faktor B) er gitt som $F = MSAB/MSE$ der $MSAB$ er middelkvadratsummen for samspillet. Vi får at $F = 9.6/3.0 = 3.2$.

b) Ved toveis variansanalyse betrakter ved de registrerte reaksjonshastighetene som observerte verdier av stokatiske variabler X_{ijk} som er gitt ved

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}\,; \quad k = 1,2\,; \quad j = 1,2,3\,; \quad i = 1,2,3\,;$$

der ϵ_{ijk}-ene er uavhengige og $N(0,\sigma^2)$-fordelte. For å gjøre parameterene i modellen identifiserbare må de tilfredsstille noen restriksjoner. Det er vanlig å anta at $\sum_{i=1}^{3}\alpha_i = \sum_{j=1}^{3}\beta_j = \sum_{i=1}^{3}\gamma_{ij} = \sum_{j=1}^{3}\gamma_{ij} = 0$.

(Fortsettes på side 2.)

Vi har samspill mellom faktorene temperatur og katalysator hvis effekten av katalysatoren avhenger av temperaturen (eller tilsvarende at effekten av temperaturen avhenger av mengden katalysator). I variansanalysemodellen er det γ_{ij}-ene som beskriver samspillet. Hvis $\gamma_{ij} = 0$ for $i, j = 1, 2, 3$ er det ikke samspill mellom temperatur og katalysator.

c) For å teste nullhypotesen om at det ikke er noe samspill mellom temperatur og mengden av katalysator, dvs. $H_0 : \gamma_{ij} = 0$ for $i, j = 1, 2, 3$, bruker vi testobservatoren $F = MSAB/MSE$. Hvis H_0 er sann, er testobsetvatoren F-fordelt med 4 og 9 frihetsgrader.

I punkt a fant vi $F = 3.2$. Tabellen for F-fordelingene gir at $F_{0.10,4,9} = 2.69$ og $F_{0.05,4,9} = 3.63$. Siden testobservatoren har en verdi mellom disse to persentilene, blir P-verdien for testen mellom 5% og 10%. Vi forkaster dermed ikke nullhypotesen hvis vi holder oss til det vanlige signifkansnivået på 5%. (Siden P-verdien er forholdsvis liten, er det en tendens til samspill. Men det er altså ikke signifikant.)

Oppgave 2

a) I gruppe nummer i er det n_i insekter. Vi lar $p(x_i)$ være sannsynligheten for at et tilfeldig valgt innsekt i denne gruppen vil dø. Hver av innsektene vil enten dø eller overleve forsøket. Hvis insektene dør/overlever uavhengig av hverandre, har vi et binomisk forsøk. Da er antall innsekter som dør binomisk fordelt, $Y_i \sim \text{bin}(n_i, p(x_i))$.

b) Vi antar at $p(x_i)$ er gitt ved den logistiske modellen (2). Da blir likelihood funksjonen

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{i=1}^{5} \binom{n_i}{y_i} p(x_i)^{y_i} \left(1 - p(x_i)\right)^{n_i - y_i} \\ &= \prod_{i=1}^{5} \binom{n_i}{y_i} \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}\right)^{y_i} \left(1 - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}\right)^{n_i - y_i} \\ &= \prod_{i=1}^{5} \binom{n_i}{y_i} \frac{e^{\beta_0 y_i + \beta_1 x_i y_i}}{\left(1 + e^{\beta_0 + \beta_1 x_i}\right)^{n_i}} \end{aligned}$$

c) Log-likelihood funksjonen er gitt ved:

$$\begin{aligned} l(\beta_0, \beta_1) &= \log L(\beta_0, \beta_1) \\ &= \sum_{i=1}^{5} \left\{ \log \binom{n_i}{y_i} + \beta_0 y_i + \beta_1 x_i y_i - n_i \log\left(1 + e^{\beta_0 + \beta_1 x_i}\right) \right\} \end{aligned}$$

(Fortsettes på side 3.)

Da er score-funksjonene gitt ved

$$s_0(\beta_0,\beta_1)=\frac{\partial}{\partial\beta_0}l(\beta_0,\beta_1)=\sum_{i=1}^{5}\left\{y_i-n_i\,\frac{e^{\beta_0+\beta_1x_i}}{1+e^{\beta_0+\beta_1x_i}}\right\}$$

og

$$s_1(\beta_0,\beta_1)=\frac{\partial}{\partial\beta_1}l(\beta_0,\beta_1)=\sum_{i=1}^{5}\left\{x_iy_i-n_ix_i\,\frac{e^{\beta_0+\beta_1x_i}}{1+e^{\beta_0+\beta_1x_i}}\right\}$$

Vi merker oss at scorefunksjonene kan skrives som

$$s_j(\beta_0,\beta_1)=\frac{\partial}{\partial\beta_j}l(\beta_0,\beta_1)=\sum_{i=1}^{5}x_i^j\left\{y_i-n_i\,\frac{e^{\beta_0+\beta_1x_i}}{1+e^{\beta_0+\beta_1x_i}}\right\}$$

for $j=0,1$.

d) Ved å ta utgangspunkt i det siste uttrykket i forrige punkt, finner vi for $j,k=0,1$ at

$$\begin{aligned}J_{jk}(\beta_0,\beta_1) &= -\frac{\partial^2}{\partial\beta_j\partial\beta_k}l(\beta_0,\beta_1)=-\frac{\partial}{\partial\beta_k}s_j(\beta_0,\beta_1)\\ &= \sum_{i=1}^{5}n_ix_i^j\,\frac{\partial}{\partial\beta_k}\left(\frac{e^{\beta_0+\beta_1x_i}}{1+e^{\beta_0+\beta_1x_i}}\right)\\ &= \sum_{i=1}^{5}n_i\,x_i^{j+k}\frac{e^{\beta_0+\beta_1x_i}}{\left(1+e^{\beta_0+\beta_1x_i}\right)^2}\end{aligned}$$

e) For å bestemme maksimum likelihood estimatene ved Newton-Raphsons metode, velger vi startverdier $\beta_0^{(0)}$ og $\beta_1^{(0)}$. Deretter beregner vi nye verdier $\beta_0^{(s+1)}$ og $\beta_1^{(s+1)}$ ved algoritmen

$$\begin{bmatrix}\beta_0^{(s+1)}\\ \beta_1^{(s+1)}\end{bmatrix}=\begin{bmatrix}\beta_0^{(s)}\\ \beta_1^{(s)}\end{bmatrix}+\mathbf{J}(\beta_0^{(s)},\beta_1^{(s)})^{-1}\begin{bmatrix}s_0(\beta_0^{(s)},\beta_1^{(s)})\\ s_1(\beta_0^{(s)},\beta_1^{(s)})\end{bmatrix}$$

for $s=0,1,2,\ldots$. Vi stopper iterasjonen når både $|\beta_0^{(s+1)}-\beta_0^{(s)}|<\epsilon$ og $|\beta_1^{(s+1)}-\beta_1^{(s)}|<\epsilon$ for en gitt nøyaktighet ϵ (for eksempel $\epsilon=10^{-6}$). Maksimum likelihood estimatene $\widehat{\beta}_0$ og $\widehat{\beta}_1$ er de verdiene vi har for $\beta_0^{(s+1)}$ og $\beta_1^{(s+1)}$ når vi stopper iterasjonen.

For å komme i gang med iterasjonen, må vi velge startverdier $\beta_0^{(0)}$ og $\beta_1^{(0)}$. Det kreves ikke at studentene kommenterer dette. Men en mulig måte å gå fram på er følgende. Ved å omforme den logistiske modellen (2) får vi

$$e^{\beta_0+\beta_1x_i}=\frac{p(x_i)}{1-p(x_i)}$$

(Fortsettes på side 4.)

som gir at

$$\beta_0 + \beta_1 x_i = \log\left(\frac{p(x_i)}{1-p(x_i)}\right)$$

Ved å benytte at andelen døde i gruppene 1 og 5 er henholdsvis 6/48 og 44/50 og at log-dosene i disse gruppene er 0.41 og 1.01, kan vi finne startverdier ved å løse likningene

$$\beta_0^{(0)} + 0.41\beta_1^{(0)} = \log\left(\frac{6/48}{1-6/48}\right) = -1.946$$

$$\beta_0^{(0)} + 1.01\beta_1^{(0)} = \log\left(\frac{44/50}{1-44/50}\right) = 1.992$$

Det gir $\beta_0^{(0)} = -4.64$ og $\beta_1^{(0)} = 6.56$.

f) Maksimum likelihood estimatorene er tilnærmet normalfordelte:

$$\widehat{\beta}_j \overset{\text{tiln}}{\sim} N(\beta_j, \sigma^2_{\widehat{\beta}_j})\,;\ j = 0, 1$$

Vi estimerer variansene til estimatorene ved elementene på diagonalen i den inverse informasjonsmatrisen. Spesielt har vi at $\widehat{\sigma}^2_{\widehat{\beta}_1} = 0.782$.

På vanlig måte er et (tilnærmet) 95% konfidensintervall for β_1 gitt ved

$$\widehat{\beta}_1 \pm 1.96\,\widehat{\sigma}_{\widehat{\beta}_1}$$

Det gir intervallet 7.011 ± 1.733, dvs. fra 5.278 til 8.744.

g) LD50 (som står for "lethal dose 50%") er den verdien av log-dosen x som svarer til 50% dødelighet, dvs. $p(x) = 0.50$. Av uttrykket (2) for den logistiske modellen finner vi at LD50 er løsningen av ligningen $\beta_0 + \beta_1 \text{LD50} = 0$, dvs. $\text{LD50} = -\beta_0/\beta_1$. Et estimat for LD50 er dermed

$$\widehat{\text{LD50}} = \frac{\widehat{\beta}_0}{\widehat{\beta}_1} = \frac{-4.792}{7.011} = 0.683$$

h) For gruppe i bestemmer vi forventet antall døde av uttrykket $e_i = n_i p^*(x_i)$, der $p^*(x_i)$ er gitt ved det logistiske uttrykket (2) med maksimum likelihood estimatene 4.792 og 7.011 innsatt for β_0 og β_1. Merk at forventet antall levende i gruppe i er $n_i - e_i$.

For å teste nullhypotesen om at dødssannsynlighetene er gitt ved den logistiske modellen, kan vi bruk den kji-kvadratobservatoren vi får ved å sammenligne observert antall døde og observert antall levende for de fem gruppene med de tilsvarende forventete antallene:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{5}\left\{\frac{(y_i - e_i)^2}{e_i} + \frac{\left[n_i - y_i - (n_i - e_i)\right]^2}{n_i - e_i}\right\} \\ &= \sum_{i=1}^{5}\left\{\frac{(y_i - e_i)^2}{e_i} + \frac{(y_i - e_i)^2}{n_i - e_i}\right\}\end{aligned}$$

(Fortsettes på side 5.)

Under nullhypotesen er testobservatoren kji-kvadratfordelt. Antall frihetsgrader er lik antall parametere i apriori modellen (hvor dødssannsynlighetene i de fem gruppene kan variere fritt) minus antall parametere under den logistiske modellen, dvs. $5 - 2 = 3$.

Vi finner at

$$\chi^2 = \frac{(6-6.15)^2}{6.15} + \frac{(6-6.15)^2}{48-6.15} + \cdots + \frac{(44-45.40)^2}{45.40} + \frac{(44-45.40)^2}{50-45.40} = 1.32$$

Av tabellen for kji-kvadratfordelingene ser vi at $\chi^2_{0.90,3} = 0.584$ og $\chi^2_{0.10,3} = 6.251$. P-verdien for testen blir dermed mellom 90% og 10% (og nærmere 90% enn 10%). Det betyr at vi ikke forkaster nullhypotesen, og vi kan konkludere med at den logistiske modellen gir en god beskrivelse av dataene. (For å forkaste nullhypotesen på 5% nivå måtte testobservatoren ha vært større enn $\chi^2_{0.05,3} = 7.815$.)

i) En annen test vi kan bruke er sannsynlighetskvotetesten (likelihood ratio testen). For å utføre denne testen bestemmer vi maksimum likelihood estimatene for den fulle modellen, dvs. for modellen der dødssannsynlighetene $p(x_i)$ ikke (nødvendigvis) er gitt ved den logistiske modellen (2). Disse maksimum likelihood estimatene er gitt som $\widehat{p}(x_i) = y_i/n_i$.

Sannsynlighetskvoten (likelihood ratio) kan da skrives som

$$\begin{aligned}
\Lambda &= \frac{\prod_{i=1}^5 \binom{n_i}{y_i} p^*(x_i)^{y_i}(1-p^*(x_i))^{n_i-y_i}}{\prod_{i=1}^5 \binom{n_i}{y_i} \widehat{p}(x_i)^{y_i}(1-\widehat{p}(x_i))^{n_i-y_i}} \\
&= \prod_{i=1}^5 \left\{ \left(\frac{p^*(x_i)}{\widehat{p}(x_i)}\right)^{y_i} \left(\frac{1-p^*(x_i)}{1-\widehat{p}(x_i)}\right)^{n_i-y_i} \right\} \\
&= \prod_{i=1}^5 \left\{ \left(\frac{n_i p^*(x_i)}{n_i \widehat{p}(x_i)}\right)^{y_i} \left(\frac{n_i - n_i p^*(x_i)}{n_i - n_i \widehat{p}(x_i)}\right)^{n_i-y_i} \right\} \\
&= \prod_{i=1}^5 \left\{ \left(\frac{e_i}{y_i}\right)^{y_i} \left(\frac{n_i - e_i}{n_i - y_i}\right)^{n_i-y_i} \right\}
\end{aligned}$$

Vi forkaster nullhypotesen når sannsynlighetskvoten er tilstrekkelig liten, eller ekvivalent når

$$-2\log\Lambda = 2\sum_{i=1}^5 \left\{ y_i \log\left(\frac{y_i}{e_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - e_i}\right) \right\}$$

er tilstrekkelig stor. Under nullhypotesen er $-2\log\Lambda$ tilnærmet kjikvadratfordelt med $5-2=3$ frihetsgrader, så vi forkaster nullhypotesen på nivå 5% hvis $-2\log\Lambda > \chi^2_{0.05,3} = 7.815$.

(Fortsettes på side 6.)

Oppgave 3

a) La $\mathbf{0} = [0, 0, \ldots, 0]'$ være den n-dimensjonale nullvektoren. Da er forventningsvektoren $E(\mathbf{Y}) = [E(Y_1), E(Y_2), \ldots, E(Y_n)]'$ til $\mathbf{Y}$ gitt ved

$$E(\mathbf{Y}) = E\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{0} = \mathbf{X}\boldsymbol{\beta}$$

Det følger at $\widehat{\boldsymbol{\beta}}$ har forventningsvektor

$$E(\widehat{\boldsymbol{\beta}}) = E\left(\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Y}\right) = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'E\left(\mathbf{Y}\right) = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

Det viser at $\widehat{\boldsymbol{\beta}}$ er forventningsrett.

Vi har at ε_i-ene er uavhengige og $N(0, \sigma^2)$-fordelte. Derfor har $\mathbf{Y}$ kovariansmatrise $\mathrm{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$, der $\mathbf{I}$ er identitetsmatrisen av dimensjon $n \times n$. Av dette følger det at kovariansmatrisen til $\widehat{\boldsymbol{\beta}}$ blir:

$$\begin{aligned}
\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) &= \mathrm{Cov}\left(\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Y}\right) \\
&= \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\,\mathrm{Cov}\left(\mathbf{Y}\right)\left(\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right)' \\
&= \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\,\left(\sigma^2\mathbf{I}\right)\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1} \\
&= \sigma^2\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1} \\
&= \sigma^2\mathbf{C}
\end{aligned}$$

der $\mathbf{C} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}$.

b) Vi ser nå på forventningen svarende til verdiene $x_1^*, \ldots, x_k^*$ av forklaringsvariablene, dvs.

$$\mu(x_1^*, \ldots, x_k^*) = \beta_0 + \beta_1 x_1^* + \cdots + \beta_k x_k^* = (\mathbf{x}^*)'\boldsymbol{\beta}$$

der $\mathbf{x}^* = [1, x_1^*, \ldots, x_k^*]'$. Vi estimerer denne ved

$$\widehat{\mu}(x_1^*, \ldots, x_k^*) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \cdots + \widehat{\beta}_k x_k^* = (\mathbf{x}^*)'\widehat{\boldsymbol{\beta}}$$

Vi har at

$$E\left(\widehat{\mu}(x_1^*, \ldots, x_k^*)\right) = E\left((\mathbf{x}^*)'\widehat{\boldsymbol{\beta}}\right) = (\mathbf{x}^*)'E(\widehat{\boldsymbol{\beta}}) = (\mathbf{x}^*)'\boldsymbol{\beta} = \mu(x_1^*, \ldots, x_k^*),$$

så estimatoren er forventningsrett.

Variansen til estimatoren blir:

$$\begin{aligned}
V(\widehat{\mu}(x_1^*, \ldots, x_k^*)) &= V\left((\mathbf{x}^*)'\widehat{\boldsymbol{\beta}}\right) = (\mathbf{x}^*)'\mathrm{Cov}(\widehat{\boldsymbol{\beta}})\mathbf{x}^* \\
&= (\mathbf{x}^*)'\left(\sigma^2\mathbf{C}\right)\mathbf{x}^* = \sigma^2\,(\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*
\end{aligned}$$

(Fortsettes på side 7.)

c) Vi har at $\widehat{\mu}(x_1^*,\ldots,x_k^*) = (\mathbf{x}^*)'\widehat{\boldsymbol{\beta}} = (\mathbf{x}^*)'\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Y}$. Det betyr at $\widehat{\mu}(x_1^*,\ldots,x_k^*)$ er en lineærkombinasjon av de normalfordelte Y_i-ene, så $\widehat{\mu}(x_1^*,\ldots,x_k^*)$ er normalfordelt. Det følger at den standardiserte variabelen

$$Z = \frac{\widehat{\mu}(x_1^*,\ldots,x_k^*) - \mu(x_1^*,\ldots,x_k^*)}{\sqrt{\sigma^2(\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*}},$$

er standardnormalfordelt. Videre er det kjent at $U = [n-(k+1)]S^2/\sigma^2$ er kji-kvadrat fordelt med $\nu = n-(k+1)$ frihetsgrader og at S^2 er uavhengig av $\widehat{\boldsymbol{\beta}}$. Da er Z uavhengig av U, og det følger at

$$\frac{\widehat{\mu}(x_1^*,\ldots,x_k^*) - \mu(x_1^*,\ldots,x_k^*)}{S\sqrt{(\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*}} = \frac{\frac{\widehat{\mu}(x_1^*,\ldots,x_k^*)-\mu(x_1^*,\ldots,x_k^*)}{\sqrt{\sigma^2(\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*}}}{\sqrt{\frac{[n-(k+1)]S^2/\sigma^2}{n-(k+1)}}} = \frac{Z}{\sqrt{\frac{U}{\nu}}}$$

er t-fordelt med $\nu = n-(k+1)$ frihetsgrader.

Av dette har vi at

$$P\left(-t_{\alpha/2,n-(k+1)} \le \frac{\widehat{\mu}(x_1^*,\ldots,x_k^*) - \mu(x_1^*,\ldots,x_k^*)}{S\sqrt{(\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*}} \le t_{\alpha/2,n-(k+1)}\right) = 1-\alpha$$

På vanlig måte gir dette følgende $100(1-\alpha)\%$ konfidensintervall for $\mu(x_1^*,\ldots,x_k^*)$:

$$\widehat{\mu}(x_1^*,\ldots,x_k^*) \pm t_{\alpha/2,n-(k+1)}\, S\sqrt{(\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*}$$

d) Vi har

$$Y^* = \beta_0 + \beta_1 x_1^* + \cdots + \beta_k x_k^* + \varepsilon^* = \mu(x_1^*,\ldots,x_k^*) + \varepsilon^*$$

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

$$Y^* - \widehat{\mu}(x_1^*,\ldots,x_k^*) = \mu(x_1^*,\ldots,x_k^*) + \varepsilon^* - \widehat{\mu}(x_1^*,\ldots,x_k^*)$$

er normalfordelt med forventning lik 0 og

$$\begin{aligned} V\left(Y^* - \widehat{\mu}(x_1^*,\ldots,x_k^*)\right) &= V\left(\mu(x_1^*,\ldots,x_k^*) + \varepsilon^* - \widehat{\mu}(x_1^*,\ldots,x_k^*)\right) \\ &= V(\varepsilon^*) + V\left(\widehat{\mu}(x_1^*,\ldots,x_k^*)\right) \\ &= \sigma^2(\mathbf{x}^*)'\mathbf{C}\mathbf{x}^* + \sigma^2 \\ &= \sigma^2\left\{1 + (\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*\right\} \end{aligned}$$

Dermed er

$$\frac{Y^* - \widehat{\mu}(x_1^*,\ldots,x_k^*)}{\sqrt{\sigma^2\left\{1 + (\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*\right\}}}$$

(Fortsettes på side 8.)

standardnormalfordelt. Ved et tilsvarende resonnement som i forrige punkt følger det at

$$\frac{Y^* - \widehat{\mu}(x_1^*, \ldots, x_k^*)}{S\sqrt{1 + (\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*}}$$

er t-fordelt med $n - (k+1)$ frihetsgrader. Dermed har vi at

$$P\left(-t_{\alpha/2,n-(k+1)} \leq \frac{Y^* - \widehat{\mu}(x_1^*, \ldots, x_k^*)}{S\sqrt{1 + (\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*}} \leq t_{\alpha/2,n-(k+1)}\right) = 1 - \alpha$$

Av dette følger det at

$$P\left(\widehat{\mu}(x_1^*, \ldots, x_k^*) - t_{\alpha/2,n-(k+1)} S\sqrt{1 + (\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*} \leq Y^* \leq \right.$$
$$\left. \widehat{\mu}(x_1^*, \ldots, x_k^*) + t_{\alpha/2,n-(k+1)} S\sqrt{1 + (\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*}\right) = 1 - \alpha$$

Dermed er

$$\widehat{\mu}(x_1^*, \ldots, x_k^*) \pm t_{\alpha/2,n-(k+1)}\, S\sqrt{1 + (\mathbf{x}^*)'\mathbf{C}\mathbf{x}^*}$$

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