UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK2120 — Løsningsforslag
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 STK1120/2120

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

Oppgave 1

a) Forventet respons er en lineær funksjon i parametrene som er koeffesienter til ikke-tilfeldige kovariater. Feilleddene er additive, og ukorrelerte med forventning null og samme varians. Hvis man i tillegg antar at feilleddene er normalfordelte gir det eksakte resultater for fordelingen til estimatorene for koeffesientene og variansen til restleddene.

En R^2 på 0.92 indikerer at en betydelig del av variasjonen i responsvariabelen er forklart ved regresjonsligningen.

Det er relativt få observasjoner, men panel 1 og 3 gir ikke noen spesiell indikasjon på at variansen øker med forventet respons. Normalfordelingsplottet virker OK. Plottet av leverage viser at den 12'te observasjonen har størst leverage, noe som ikke er overraskende gitt at observasjonene med lavest og høyest alder kan ha stor innflytelse på tilpasningen av et 2. grads polynom. Den 12'te observasjonen har også høy Cook distanse, noe som viser at hvis den blir utelatt har det stor betydning for tilpasningen.

b) For testen $H_0 : \beta_2 = \beta_{2,0}$ mot alternativet $H_A : \beta_2 \neq \beta_{2,0}$ er testobservatoren $t = (\hat{\beta}_2 - \beta_{2,0})/s_{\hat{\beta}_2}$ der $s_{\hat{\beta}_2}$ er et estimat for standardfeilen til $\hat{\beta}_2$ og H_0 forkastes for store verdier av $|t|$.

Utskriften gir at for $\beta_{2,0} = 0$ er $\hat{\beta}_2 = -4.703, s_{\hat{\beta}_2} = 1.504, t = -3.127$ og p-verdien 0.012181 som gir forkastning på nivå 0.025, men ikke nivå 0.01.

For $\beta_{2,0} = -3$ er $t = (-4.703 - (-3.0))/1.504 = -1.132$. 97.5% kvantilen i t-fordelinen med 9 f.g. er 2.262, dvs ingen forkastning med nivå 0.05.

c)

$$\tilde{Y} = X\tilde{\beta} + \tilde{\epsilon}$$

(Fortsettes på side 2.)

der $\tilde{Y} = (Y_1, \dots, Y_{12})'$, $\tilde{\beta} = (\beta_0, \beta_1, \beta_2)'$ og $\tilde{\epsilon} = (\epsilon_1, \dots, \epsilon_{12})'$. Designmatrisen X er 12×3 matrisen

$$X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 4 & 16 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \end{pmatrix}$$

Minste kvadraters estimatorer for $\tilde{\beta}$ er $\hat{\tilde{\beta}} = (X'X)^{-1}X'\tilde{Y}$ og $E[\hat{\tilde{\beta}}] = E[(X'X)^{-1}X'\tilde{Y}] = E[(X'X)^{-1}X'(X\tilde{\beta} + \tilde{\epsilon}] = (X'X)^{-1}X'X\tilde{\beta} + (X'X)^{-1}E[\tilde{\epsilon}] = \tilde{\beta}$, siden restleddene har forventning null.

d) Hvis A er matrise er $Cov(A\tilde{Y}) = ACov(\tilde{Y})A'$. Siden vektoren av residualer er $\tilde{e} = \tilde{Y} - X\hat{\tilde{\beta}}$, er ved innsetting $\tilde{e} = \tilde{Y} - X(X'X)^{-1}X'\tilde{Y} = (I - X(X'X)^{-1}X')\tilde{Y} = (I - H)\tilde{Y}$ der hattematrisen H tilfredstiller $H^2 = H$ og $H' = H$. Da er $Cov(\tilde{e}\tilde{e}') = Cov[(I-H)\tilde{Y}\tilde{Y}'(I-H)'] = (I-H)Cov[\tilde{Y}\tilde{Y}'](I-H)' = \sigma^2(I-H)(I-H)' = \sigma^2(I-H-H'+HH') = (I-H)\sigma^2$.

e) Diagonalelementene i hattematrisen kalles leverage. Fra uttrykket for kovariansmatrisen til residualene utledet i punkt d) ser vi at høye verdier av leverage svarer til små varianser for residualene, dvs. god tilpasning i disse punktene. Dette kan bety at de tilsvarende observasjonene er viktige for tilpasningen. For andregradspolynom kan observasjonene som svarer til små og store verdier av kovariaten ha denne egenskapen siden de kan være avgjørende for krumningen. I dette tilfellet er det observasjonene 1,2 og 12, som også har de høyeste leverage verdiene.

Oppgave 2

a)
```
Analysis of Variance Table

Response: reaksjonsrate
                                          Df Sum Sq Mean Sq F value  Pr(>F)
factor(reaksjonsmiddel)                    1     12      12    3.00 0.13397
factor(katalysator)                        2     86      43   10.75 0.01039 *
factor(reaksjonsmiddel):factor(katalysator) 2    14       7    1.75 0.25193
Residuals                                  6     24       4
```

Siste linje i tabellen gir opplysningene som trengs for å teste om inklusjon av samspill er nødvendig. Testobservatoren er 1.75, som skal sammenlignes med en Fisher-fordeling med 2 og 6 frihetsgrader.

(Fortsettes på side 3.)

Dette gir en p-verdi på 0.25, slik at nullhypotesen om at samspillet er signifikant ikke forkastes hvis nivåer til testen er mindre 0.25.

b)

$$X = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{pmatrix}$$

Kolonnene svarer henholdsvis til μ, α_1, β_1 og β_2, og observasjonene er ordnet leksiografisk, dvs. første rad 111, andre 112, tredje 121,..., siste (og tolvte) 232.

c) På grunn av ortogonalitet blir matrisen $X'X$ blokkdiagonal

$$X'X = \begin{pmatrix} 12 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 \\ 0 & 0 & 8 & 4 \\ 0 & 0 & 4 & 8 \end{pmatrix}$$

Den inverse finnes ved å invertere blokkene.

$$(X'X)^{-1} = \begin{pmatrix} 1/12 & 0 & 0 & 0 \\ 0 & 1/12 & 0 & 0 \\ 0 & 0 & 8/48 & -4/48 \\ 0 & 0 & -4/48 & 8/48 \end{pmatrix}$$

Vectoren $X'\tilde{y}$ er

$$X'\tilde{y} = \begin{pmatrix} y_{\cdots} \\ y_{1\cdot\cdot} - y_{2\cdot\cdot} \\ y_{\cdot 1\cdot} - y_{\cdot 3\cdot} \\ y_{\cdot 2\cdot} - y_{\cdot 3\cdot} \end{pmatrix}$$

slik at minste kvadraters estimatorer $(X'X)^1 X'\tilde{y}$ er $y_{\cdots}/12, (y_{1\cdot\cdot} - y_{2\cdot\cdot})/12, (y_{\cdot 1\cdot} - y_{\cdot 3\cdot})/6 - (y_{\cdot 2\cdot} - y_{\cdot 3\cdot})/12, -(y_{\cdot 1\cdot} - y_{\cdot 3\cdot})/12 + (y_{\cdot 2\cdot} - y_{\cdot 3\cdot})/6$.

Tallene i oppgaven gir $y_{1\cdot\cdot} = 42, y_{2\cdot\cdot} = 54, y_{\cdot 1\cdot} = 20, y_{\cdot 2\cdot} = 46, y_{\cdot 3\cdot} = 30$, slik at $\hat{\mu} = 96/12 = 8$, $\hat{\alpha}_1 = (42-54)/12 = -1, \hat{\beta}_1 = (20-30)/6 - (46-30)/12 = -3$ og $\hat{\beta}_2 = (46-30)/6 - (20-30)/12 = 3.5$. Oppsumeringsbetingelen gir at $\hat{\alpha}_2 = 1$ og $\hat{\beta}_3 = -(\hat{\beta}_1 + \hat{\beta}_2) = -(-3+3.5) = -0.5$.

Oppgave 3

a) Likelihooden er $L(p) = \prod_{i=1}^n p^{x_i - 1}(1-p) = p^{\sum x_i - n}(1-p)^n$ slik at

$$\log(L(p)) = n\log(1-p) + (\sum x_i - n)\log(p)$$

(Fortsettes på side 4.)

og

$$\frac{d\log(L(p))}{dp} = -n/(1-p)+(\sum x_i - n)/p = ((1-p)\sum x_i - n)/(p(1-p))$$

Første ordens betingelsen $\frac{d\log(L(p))}{dp} = 0$ gir at SME er $\hat{p} = 1 - \frac{n}{\sum x_i}$. Informasjonene finnes av $Var(\frac{\log(L(p))}{dp}) = \frac{n}{(1-p)^2 p}$ siden $Var(X_i) = p/(1-p)^2$ fra formelsamlingen når parameteriseringen av den geometriske fordelingen tas i betraktning. Alternativt kan første og andre ordens moment i fordelingen beregnes direkte. Den tilnærmede fordelingen til $(\hat{p} - p)$ er derfor $N(0, (1-p)^2 p/n)$ eller $\sqrt{n}(\hat{p} - p)$ er tilnærmet $N(0, p(1-p)^2)$.

b) Likelihooden er $L(m,k) = \prod_{i=1}^{n} \frac{\Gamma(k+x_i-1)}{(x_i-1)!\Gamma(k)}(1+\frac{m}{k})^{-k}(\frac{m}{m+k})^{x_i-1}$ slik at

$$\log(L(m,k)) = \sum \log(\Gamma(k+x_i-1)) - \sum \log((x_i-1)!) - n\log(\Gamma(k))$$
$$-nk\log(m+k) + nk\log(k) + \sum (x_i-1)(\log(m) - \log(m+k)).$$

Følgelig

$$\frac{d\log(L(m,k))}{dk} = \sum \frac{\Gamma'(k+x_i-1)}{\Gamma(k+x_i-1)} - n\frac{\Gamma'(k)}{\Gamma(k)}$$
$$-n\log(m+k) - \frac{nk}{m+k} + n + n\log(k) - \sum (x_i-1)/(m+k)$$

og

$$\frac{d\log(L(m,k))}{dm} = -nk/(m+k) + \sum (x_i-1)(1/m - 1/(m+k)) = \frac{-mnk + (m+k-m)\sum(x_i-1)}{(m+k)m}.$$

Fra førsteordensbetingelsen $\frac{d\log(L(m,k)}{dm} = 0$ følger at SME, $\hat{m} = \bar{x} - 1$. Ved innsetting følger da at SME for k finnes som løsningen av den ikke-lineære ligningen

$$\sum \frac{\Gamma'(k+x_i-1))}{\Gamma(k+x_i-1)} + n\frac{\Gamma'(k)}{\Gamma(k)} - nk\log(\bar{x}-1+k) - \frac{nk}{\bar{x}-1+k} + n + n\log(k) - \sum(x_i-1)/(\bar{x}-1+k) = 0.$$

Den kan løses ved den endimensjonale Newton-Raphson algoritmen for eksempel.

La $k = 1$. Da er punktsannsynligheten $p(j) = \frac{\Gamma(1+j-1)}{(j-1)!\Gamma(1)}(1 + \frac{m}{1})^{-1}(\frac{m}{m+1})^{j-1} = \frac{\Gamma(j)}{(j-1)!\Gamma(1)}(1+m)^{-1}(\frac{m}{m+1})^{j-1} = (1+m)^{-1}(1 - \frac{1}{m+1})^{j-1}$, siden $\Gamma(x+1) = x\Gamma(x), \Gamma(1) = 1$. Sett så $p = m/(m+1)$, slik at $m = p/(1-p)$ og $1 + m = 1/(1-p)$.