STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

25 november 2020 9–16

Examinator: Ola Hössjer, tel. 070/672 12 18, ola@math.su.se

Uppgift 1

a) Låt $\bar{Y}_{j\cdot}$ vara medelvärdet av Y_i för de personer som fått j mg av vaccinet. Vi noterar att

$$\begin{aligned}
\bar{x} &= \textstyle\sum_j \sum_{i;x_i=j} x_i/20 = \sum_{j=0}^4 4j/20 = \sum_{j=0}^4 j/5 = 2,\\
\textstyle\sum_{i=1}^{20}(x_i-\bar{x})^2 &= 4\textstyle\sum_{j=0}^4 (j-2)^2 = 4\cdot 10,\\
\textstyle\sum_{i=1}^{20}(x_i-\bar{x})Y_i &= 4\textstyle\sum_{j=0}^4 (j-2)\bar{Y}_j\\
&= 4[(-2)2+(-1)3.4+0\cdot 4.0+1\cdot 5.2+2\cdot 6.1]\\
&= 4\cdot 10.0.
\end{aligned}$$

Det ger en minsta kvadrat-skattning

$$\hat{\beta} = \frac{\sum_i (x_i-\bar{x})Y_i}{\sum_i (x_i-\bar{x})^2} = \frac{\sum_{j=0}^4 (j-2)\bar{Y}_j}{\sum_{j=0}^4 (j-2)^2} = \frac{10.0}{10} = 1.0.$$

b) Variansen för skattningen av effektparametern β ges av

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{20}(x_i-\bar{x})^2} = \frac{\sigma^2}{4\cdot 10} = \frac{\sigma^2}{40}.$$

c) Låt s_j^2 vara stickprovsvariansen för den grupp av individer som får dosen j mg av vaccinet. Summan av kvadratavvikelserna $(Y_i-\bar{Y}_{j\cdot})^2$ för individerna i i grupp j, ges av $(4-1)s_j^2 = 3s_j^2$. Genom att summera detta uttryck över alla fem grupper får vi den totala kvadratsumman

$$\mathrm{Kvs(Inom\ dos)} = 3\sum_{j=0}^4 s_j^2,$$

för variationskällan Inom dos, med $5(4-1) = 15$ frihetsgrader. Det ger en skattning

$$\begin{aligned}\hat{\sigma}^2 &= \text{Mkvs(Inom dos)} \\ &= \text{Kvs(Inom dos)}/15 \\ &= \textstyle\sum_{j=0}^{4} 3s_j^2/15 \\ &= \textstyle\sum_{j=0}^{4} s_j^2/5 \\ &= (1 + 1.8 + 2.4 + 2.0 + 2.8)/5 \\ &= 2.0\end{aligned}$$

av feltermsvariansen. Medelfelet för skattningen av β blir

$$d = \sqrt{\widehat{\text{Var}}(\hat{\beta})} = \sqrt{\frac{\hat{\sigma}^2}{40}} = \sqrt{\frac{2.0}{40}} = 0.2236.$$

d) Det följer av a) och c) ovan att ett tvåsidigt konfidensinterval för β med konfidensgrad 95.0% ges av

$$\begin{aligned}(\hat{\beta} - t_{0.025}(15)d, \hat{\beta} + t_{0.025}(15)d) &= (1.0 - 2.1314 \cdot 0.2236, 1.0 + 2.1314 \cdot 0.2236) \\ &= (0.523, 1.477),\end{aligned}$$

där $t_{0.025}(15) = \sqrt{F_{0.05}(1, 15)}$ fås ur tabell. Eftersom 0 inte ingår i detta intervall så kan vi förkasta nollhypotesen $\beta = 0$ på nivån 5.0%. Vaccinet har alltså signifikanta biverkningar.

Uppgift 2

a) Försöket kan beskrivas med en allmän linjär modell

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

$\boldsymbol{Y} = (Y_1, \ldots, Y_6)^T = (57.4, 58.4, 58.4, 30.2, 28.7, 27.7)^T$ är observationsvektorn, $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$ parametervektorn,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}.$$

designmatrisen samt $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_6)^T$ en vektor av oberoende och $N(0, \sigma^2)$-fördelade feltermer ε_i. Minsta kvadrat-skattningen av parametervektorn ges av

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \frac{1}{5}\boldsymbol{A}^T\boldsymbol{Y} = \begin{pmatrix} 28.4 \\ 29.0 \\ 29.6 \end{pmatrix} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\theta}_3 \end{pmatrix},$$

där vi i andra ledet utnyttjade att designmatrisen har ortogonala kolumner med kvadratsumma 5 (dvs $\boldsymbol{A}^T\boldsymbol{A} = 5\boldsymbol{I}_3$). I de två sista leden användes

$$\begin{aligned}\hat{\theta}_1 &= (Y_1 + Y_2 - Y_4 + Y_5 + Y_6)/5 \\ &= (57.4 + 58.4 - 30.7 + 28.7 + 27.7)/5 \\ &= 28.4\end{aligned}$$

för skattningen av längden av Ark 1, och motsvarande för längderna av de andra två arken.

b) Nollhypotesen kan skrivas som $H_0 : \theta_1 = \theta_2 = \theta_3 = \lambda$. Det svarar mot sambandet

$$\boldsymbol{\theta} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \lambda = \boldsymbol{B}\lambda.$$

c) För att skatta λ så inför vi matrisen $\boldsymbol{C} = \boldsymbol{AB} = (2, 2, 2, 1, 1, 1)^T$ och noterar att under nollhypotesen så gäller $\boldsymbol{Y} = \boldsymbol{C}\lambda + \boldsymbol{\varepsilon}$, dvs $\boldsymbol{C}$ är designmatris under H_0. Det ger minsta kvadrat-skattningen

$$\hat{\lambda} = (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}\boldsymbol{Y} = \frac{1}{15}(2Y_1 + 2Y_2 + 2Y_3 + Y_4 + Y_5 + Y_6) = 29.0.$$

Vi kan nu skatta $\boldsymbol{\mu}$ under grund- och hypotesmodellen som

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= \boldsymbol{A}\hat{\boldsymbol{\theta}} = (57.4, 58.0, 58.6, 30.2, 29.0, 27.8)^T, \\ \hat{\boldsymbol{\mu}} &= \boldsymbol{C}\hat{\lambda} = (58.0, 58.0, 58.0, 29.0, 29.0, 29.0)^T.\end{aligned}$$

Kvadratsumman för differensen mellan dessa två vektorer är

$$\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 = (-0.6)^2 + 0^2 + 0.6^2 + 1.2^2 + 0^2 + (-1.2)^2 = 3.60.$$

Eftersom vi har $N = 6$ observationer, grundmodellen har $k = 3$ parametrar och hypotesmodellen $l = 1$ parameter får vi slutligen en

$$\text{F-kvot} = \frac{\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2/(3-1)}{\text{Kvs(Residual)}/(6-3)} = \frac{3.60/2}{0.3/3} = 18 > F_{0.05}(2, 3) = 9.55.$$

Vi kan alltså på nivån 5% förkasta nollhypotesen att arken har samman längd.

Uppgift 3

a) Vi börjar med att fylla i antal frihetsgrader och medelkvadratsummor i variansanalystabellen, plus en ny kolumn med uttryck för E(Mkvs).

Variationskälla	Kvs	f	Mkvs	E(Mkvs)
Tjocklek	18.0	4-1=3	18.0/3=6.0	$\sigma_\epsilon^2 + 2\sigma_\gamma^2 + 2 \cdot 3\sigma_\alpha^2$
Färg	18.4	3-1=2	18.4/2=9.2	$\sigma_\epsilon^2 + 2\sigma_\gamma^2 + 2 \cdot 4\sigma_\beta^2$
Samspel	7.2	(4-1)(3-1)=6	7.2/6 = 1.2	$\sigma_\epsilon^2 + 2\sigma_\gamma^2$
Inom celler	6.0	$4 \cdot 3(2-1) = 12$	6.0/12=0.5	σ_ϵ^2
Totalt	52.8	$4 \cdot 3 \cdot 2 - 1 = 23$		

Genom att kombinera informationen från de två högra kolumnerna får vi väntevärdesriktiga skattningar

$$\begin{aligned}
\hat{\sigma}^2_\epsilon &= \text{Mkvs(Inom celler)} = 0.5, \\
\hat{\sigma}^2_\gamma &= [\text{Mkvs(Samspel)} - \text{Mkvs(Inom celler)}]/2 = (1.2 - 0.5)/2 = 0.35, \\
\hat{\sigma}^2_\alpha &= [\text{Mkvs(Tjocklek)} - \text{Mkvs(Samspel)}]/6 = (6.0 - 1.2)/6 = 0.80, \\
\hat{\sigma}^2_\beta &= [\text{Mkvs(Färg)} - \text{Mkvs(Samspel)}]/8 = (9.2 - 1.2)/8 = 1.00
\end{aligned}$$

av de fyra varianskomponenterna.

b) Genom att utnyttja den givna definitionen av Y_{ijk} för en tvåvägs variansanalysmodell typ II, och medelvärdesbilda över alla 24 observationer ijk fås uttrycket

$$\hat{\mu} = \bar{Y}_{...} = \mu + \bar{\alpha}_{.} + \bar{\beta}_{.} + \bar{\gamma}_{..} + \bar{\varepsilon}_{...} \tag{1}$$

för totalmedelvärdet av alla uppmätta färgnyanser. Eftersom alla stokastiska termer i (1) är oberoende så följer att

$$\begin{aligned}
\text{Var}(\hat{\mu}) &= \text{Var}(\bar{\alpha}_{.}) + \text{Var}(\bar{\beta}_{.}) + \text{Var}(\bar{\gamma}_{..}) + \text{Var}(\bar{\varepsilon}_{...}) \\
&= \sigma^2_\alpha/4 + \sigma^2_\beta/3 + \sigma^2_\gamma/(4 \cdot 3) + \sigma^2_\epsilon/(4 \cdot 3 \cdot 2).
\end{aligned}$$

Av detta följer i sin tur att medelfelet ges av

$$\begin{aligned}
d &= \sqrt{\hat{\sigma}^2_\alpha/4 + \hat{\sigma}^2_\beta/3 + \hat{\sigma}^2_\gamma/12 + \hat{\sigma}^2_\epsilon/24} \\
&= \sqrt{0.8/4 + 1.0/3 + 0.35/12 + 0.5/24} \\
&= 0.764.
\end{aligned}$$

Uppgift 4

a) Vi börjar med utöka den givna tabellen med en till kolumn, där de fyra cellmedelvärdena $\bar{Y}_{ij.} = (Y_{ij1} + Y_{ij2})/2$ och totalmedelvärdet $\bar{Y}_{...}$ anges.

S	M	Y_{ij1}	Y_{ij2}	$\bar{Y}_{ij.}$
-	-	6.4	6.0	6.2
+	-	8.0	7.6	7.8
-	+	5.6	5.8	5.7
+	+	7.0	7.2	7.1
Medel				6.7

Minsta kvadrat-skattningar av de två huvudeffekterna, och av samspelseffekten, ges av

$$\begin{aligned}
\hat{S} &= (-\bar{Y}_{--} + \bar{Y}_{+-} - \bar{Y}_{-+} + \bar{Y}_{++})/4 = 0.75, \\
\hat{M} &= (-\bar{Y}_{--} - \bar{Y}_{+-} + \bar{Y}_{-+} + \bar{Y}_{++})/4 = -0.30, \\
\widehat{SM} &= (+\bar{Y}_{--} - \bar{Y}_{+-} - \bar{Y}_{-+} + \bar{Y}_{++})/4 = -0.05.
\end{aligned} \tag{2}$$

Alternativt minsta kvadrat-skattas parametervektorn $\boldsymbol{\theta} = (\mu, \bar{S}, \bar{M}, \overline{SM})^T$ med hjälp av formeln $\hat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y}$, där $\boldsymbol{Y} = (\bar{Y}_{--}, \bar{Y}_{+-}, \bar{Y}_{-+}, \bar{Y}_{++})^T$

är observationsvektorn för den reducerade modellen med cellmedelvärden, och där

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

är designmatrisen.

b) Vi har att

$$\begin{aligned} \text{Kvs(Inom celler)} &= \textstyle\sum_{i,j}[(Y_{ij1} - \bar{Y}_{ij\cdot})^2 + (Y_{ij1} - \bar{Y}_{ij\cdot})^2] \\ &= \textstyle\sum_{i,j}(Y_{ij1} - Y_{ij2})^2/2 \\ &= [(6.4-6.0)^2 + (8.0-7.6)^2 + (5.6-5.8)^2 + (7.0-7.2)^2]/2 \\ &= 0.2. \end{aligned}$$

Eftersom antal frihetsgrader för variationskällan Inom celler är $2 \cdot 2(2-1) = 4$, så följer att

$$\hat{\sigma}^2 = \text{Mkvs(Inom celler)} = \frac{\text{Kvs(Inom celler)}}{4} = 0.05$$

är en väntevärdesriktig skattning av feltermsvariansen σ^2.

c) Låt $\Delta = 2\bar{M}$ vara den parameter vi söker konfidensintervall för. Det följer av (2) att $\hat{\Delta} = 2(-0.30) = -0.60$ och

$$\begin{aligned} \text{Var}(\hat{\Delta}) &= 2^2[\text{Var}(\bar{Y}_{--}) + \text{Var}(\bar{Y}_{+-}) + \text{Var}(\bar{Y}_{-+}) + \text{Var}(\bar{Y}_{++})]/4^2 \\ &= 4[4(\sigma^2/2)]/16 \\ &= \sigma^2/2. \end{aligned}$$

Det ger ett medelfel $d = \sqrt{\hat{\sigma}^2/2} = \sqrt{0.05/2} = 0.1581$. Eftersom antalet frihetsgrader för variationskällan Inom celler är $4(2-1) = 4$ får vi ett konfidensintervall

$$\begin{aligned} I_\Delta &= (\hat{\Delta} - t_{0.025}(4)d, \hat{\Delta} + t_{0.025}(4)d) \\ &= (-0.6 - 2.7764 \cdot 0.158, -0.6 + 2.7764 \cdot 0.158) \\ &= (-1.039, -0.161) \end{aligned}$$

med täckningsgrad 0.95. Eftersom 0 inte tillhör detta intervall så sänks rehabiliteringstiden signifikant då fler skruvar används för att fästa protesen.

Uppgift 5

a) Hattmatrisen definieras enligt $\boldsymbol{H} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$.

b) Eftersom minsta kvadrat-skattningen av parametervektorn ges av $\hat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y}$ kan vi skriva om residualvektorn som

$$\begin{aligned}\boldsymbol{e} &= \boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\theta}} \\ &= \boldsymbol{Y} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} \\ &= \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} \\ &= (\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{Y}, \\ &= (\boldsymbol{I}_N - \boldsymbol{H})(\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}) \\ &= (\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{\varepsilon},\end{aligned}$$

där vi i fjärde ledet införde enhetsmatrisen $\boldsymbol{I}_N$ av ordning N och i sjätte ledet utnyttjade att $\boldsymbol{H}\boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\theta}$, vilket följer av definitionen av hattmatrisen. Kovariansmatrisen för residualvektorn ges av

$$\begin{aligned}\mathrm{Var}(\boldsymbol{e}) &= \mathrm{Var}[(\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{\varepsilon}] = (\boldsymbol{I}_N - \boldsymbol{H})\mathrm{Var}(\boldsymbol{\varepsilon})(\boldsymbol{I}_N - \boldsymbol{H})^T \\ &= \sigma^2(\boldsymbol{I}_N - \boldsymbol{H})(\boldsymbol{I}_N - \boldsymbol{H})^T = \sigma^2(\boldsymbol{I}_N - \boldsymbol{H}),\end{aligned} \tag{3}$$

där vi i tredje ledet utnyttjade att $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}_N$ och i fjärde ledet att hattmatrisen är idempotent ($\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$) och symmetrisk ($\boldsymbol{H}^T = \boldsymbol{H}$). Notera att diagonalelement i av (3) svarar mot identiteten $\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$. Eftersom $E(e_i) = 0$ följer av detta att

$$E[\mathrm{Kvs(Residual)}] = \sum_{i=1}^{N} E(e_i^2) = \sum_{i=1}^{N} \mathrm{Var}(e_i) = \sigma^2 \sum_{i=1}^{N}(1 - h_{ii}) = \sigma^2(N - k),$$

där vi i sista ledet utnyttjade $\sum_{i=1}^{N} h_{ii} = k$.

c) Prediktionsfelsvektorn kan skrivas som

$$\boldsymbol{Z} - \hat{\boldsymbol{\mu}} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\epsilon} - \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\epsilon} - \boldsymbol{H}(\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}) = \boldsymbol{\epsilon} - \boldsymbol{H}\boldsymbol{\varepsilon}.$$

Dess kovariansmatris är

$$\begin{aligned}\mathrm{Var}(\boldsymbol{Z} - \hat{\boldsymbol{\mu}}) &= \mathrm{Var}(\boldsymbol{\epsilon} - \boldsymbol{H}\boldsymbol{\varepsilon}) = \mathrm{Var}(\boldsymbol{\epsilon}) + \mathrm{Var}(\boldsymbol{H}\boldsymbol{\varepsilon}) \\ &= \sigma^2\boldsymbol{I}_N + \sigma^2\boldsymbol{H}\boldsymbol{H}^T = \sigma^2(\boldsymbol{I}_N + \boldsymbol{H}),\end{aligned} \tag{4}$$

där vi andra ledet utnyttjade att feltermsvektorerna $\boldsymbol{\varepsilon}$ och $\boldsymbol{\epsilon}$ är oberoende, i femte ledet att $\mathrm{Var}(\boldsymbol{\epsilon}) = \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}_N$ och i sista ledet att $\boldsymbol{H}\boldsymbol{H}^T = \boldsymbol{H}$. Genom att identifiera diagonalelementen i höger- och vänsterleden av (4), ser vi att

$$E[\mathrm{Kvs(Prediktion)}] = \sum_{i=1}^{N} E[(Z_i - \hat{\mu}_i)^2] = \sigma^2 \sum_{i=1}^{N}(1 + h_{ii}) = \sigma^2(N + k).$$

d) Den predikterade kvadratsumman är väntevärdersriktig

$$\begin{aligned}E[\widehat{\mathrm{Kvs}}(\mathrm{Prediktion})] &= E[\mathrm{Kvs(Residual)}] + CE(\hat{\sigma}^2) \\ &= \sigma^2(N - k) + C\sigma^2 \\ &= \sigma^2(N + k)\end{aligned}$$

enligt resultatet i c), om $C = 2k$. Anta att vi använder $\widehat{\text{Kvs}}$(Prediktion) för att välja mellan olika modeller, genom att minimera detta uttryck. Vi tolkar $C\hat{\sigma}^2 = 2k\hat{\sigma}^2$ som en term som bestraffar modeller med många parametrar k för att undvika överanpassning, eftersom Kvs(Residual) är mindre ju större en modeller är.