STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

28 november 2019 9–14

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

Uppgift 1

a) Skriv den centrerade regressionsmodellen som

$$Y_i = \alpha_c + \beta(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, 16,$$

med intercept $\alpha_c = \alpha + \beta\bar{x}$. Minsta kvadrat-skattningarna av α_c och β ges av

$$\begin{array}{rcl} \hat{\alpha}_c & = & \sum_i Y_i/16 = 312.0/16 = 19.5, \\ \hat{\beta} & = & \sum_i Y_i(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2 = 9.0/45.0 = 0.20. \end{array} \tag{1}$$

Det ger en skattad halt

$$\hat{\alpha} = \hat{\alpha}_c - \hat{\beta} \cdot \bar{x} = 19.5 - 0.2 \cdot \frac{168.0}{16} = 17.4 \tag{2}$$

av typ A-bakterier.

b) Eftersom de två skattningarna i (1) är oberoende stokastiska variabler, följer av (2) att

$$\begin{array}{rcl} \mathrm{Var}(\hat{\alpha}) & = & \mathrm{Var}(\hat{\alpha}_c) + \mathrm{Var}(\hat{\beta}) \cdot \bar{x}^2 \\ & = & \frac{\sigma^2}{16} + \frac{\sigma^2 \bar{x}^2}{\sum_i (x_i - \bar{x})^2} \\ & = & \sigma^2 \left(\frac{1}{16} + \frac{(168.0/16)^2}{45.0} \right) \\ & = & 2.5125 \cdot \sigma^2. \end{array} \tag{3}$$

För att skatta feltermernas varians så utnyttjar vi att variationskällan Residual har 16-2=14 frihetsgrader. Av detta följer att

$$\hat{\sigma}^2 = \frac{\mathrm{Kvs(Residual)}}{14} = \frac{19.0}{14} = 1.3571. \tag{4}$$

Genom att kombinera (3) med (4) så får vi ett medelfel

$$d = \sqrt{2.5125} \cdot \hat{\sigma} = \sqrt{2.5125 \cdot 1.3571} = 1.8465 = 1.85.$$

c) Ett 95 % konfidensintervall för halten A-bakterier är

$$\begin{aligned} I_\alpha &= (\hat{\alpha} - t_{0.025}(14) \cdot d, \hat{\alpha} + t_{0.025}(14) \cdot d) \\ &= (17.4 - 2.145 \cdot 1.8465, 17.4 + 2.145 \cdot 1.8465) \\ &= (13.4, 21.4), \end{aligned}$$

där värdet på t-kvantilen fås från tabell ($t_{0.025}(14) = \sqrt{F_{0.05}(1, 14)}$).

Uppgift 2

a) Vi börjar med att fylla i antalet frihetsgrader f i den fullständiga modellens variansanalystabell, med förkortningarna M1, M2 och M3 för de tre metallerna:

Variationskälla	f	Kvs
M1	1	3.1
M2	1	5.1
M3	1	8.9
Residual	16	25.0
Totalt	19	42.1

I första FS-steget testas tre olika grundmodeller, var och en med en förklarande variabel, med ett F-test mot en hypotesmodell som bara innehåller intercept. Eftersom M3 har störst kvadratsumma börjar vi med att undersöka F-kvoten för delmodellen som endast har M3 som förklarande variabel. Om vi använder denna delmodell som grundmodell får vi enligt ledningen

$$\begin{aligned} \text{Kvs(Regression)} &= \|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 \\ &= \text{Kvs(M3)} \\ &= 8.9, \\ \text{Kvs(Residual)} &= \|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\|^2 \\ &= \text{Kvs(M1)} + \text{Kvs(M2)} + \text{Kvs(Residual)}_{\text{fullst}} \\ &= 3.1 + 5.1 + 25.0 \\ &= 33.2, \end{aligned}$$

där $\boldsymbol{Y}$ är observationsvektorn, och $\hat{\boldsymbol{\mu}}$, $\hat{\hat{\boldsymbol{\mu}}}$ skattningar av dess väntevärde enligt grund- respektive hypotesmodellen. Eftersom antalet frihetsgrader för Regression och Residual är 1 respektive $1 + 1 + 16 = 18$ får vi en

$$\begin{aligned} \text{F-kvot} &= \frac{\text{Kvs(Regression)}/1}{\text{Kvs(Residual)}/18} \\ &= \frac{8.9}{33.2/18} \\ &= 4.825, \end{aligned}$$

som överstiger $F_{0.05}(1, 18) = 4.41$. Därför förkastas nollhypotesen, svarande mot att M3 ger ett signifikant bidrag till legeringens hållfasthet.

De andra två delmodellerna med bara M1 respektive M2 som förklarande variabler har också 1 frihetsgrad för Regression och 18 frihetsgrader för Residual. Eftersom deras kvadratsummor Kvs(Regression) = Kvs(M1) respektive Kvs(Regression) = Kvs(M2) för regression är lägre och deras kvadratsummor Kvs(Residual) = Kvs(Total)−Kvs(Regression) för residual är högre jämfört med modellen som bara har M3 som förklarande variabel, så kommer F-kvoterna för båda dessa delmodeller ha lägre värden än 4.825. Därför väljer vi M3 i första steget av FS-schemat.

b) Eftersom vi tog med M3 som förklarande variabel i a) så går vi vidare i andra steget av FS-schemat och testar hypotesmodellen med bara M3 som förklarande variabel mot två olika grundmodeller, som även inkluderar M1 respektive M2. Eftersom M2 har större kvadratsumma än M1, och alla prediktorerna är ortogonala, räcker det att undersöka om M2 ska tas med utöver M3, av samma skäl som att det i a) räckte att testa grundmodellen med M3 mot hypotesmodellen med endast intercept. Med M2 och M3 i grundmodellen och bara M3 i hypotesmodellen får vi då på grund av ortogonaliteten mellan prediktorerna (se ledningen)

$$\begin{aligned}
\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 &= \text{Kvs(Regression)}_{\text{M2+M3}} - \text{Kvs(Regression)}_{\text{M3}} \\
&= [\text{Kvs(M2)} + \text{Kvs(M3)}] - \text{Kvs(M3)} \\
&= \text{Kvs(M2)} \\
&= 5.1, \\
\text{Kvs(Residual)} &= \|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\|^2 \\
&= \text{Kvs(M1)} + \text{Kvs(Residual)}_{\text{fullst}} \\
&= 3.1 + 25.0 \\
&= 28.1,
\end{aligned} \tag{5}$$

och en

$$\begin{aligned}
\text{F-kvot} &= \frac{\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2/1}{\text{Kvs(Residual)}/(1+16)} \\
&= \frac{5.1}{28.1/17} \\
&= 3.085,
\end{aligned} \tag{6}$$

som understiger $F_{0.05}(1, 17) = 4.451$. Därför tas inte M2 med i modellen. Slutsatsten blir att FS-schemat inte tar med någon fler förklarande variabel i andra steget utan stannar efter det första steget. Den valda modellen har alltså bara med M3 som förklarande variabel.

Uppgift 3

a) Antalet frihetsgrader för residualerna $Y_{ij} - \bar{Y}_{i\cdot}$ är $6(4-1) = 18$. Det ger en skattning

$$\begin{aligned}
\hat{\sigma}_\varepsilon^2 &= \text{Mvs(Inom blodkropp)} \\
&= \tfrac{1}{18}\text{Kvs(Inom blodkropp)} \\
&= \tfrac{1}{18}\sum_{i=1}^{6}\sum_{j=1}^{4}(Y_{ij} - \bar{Y}_{i\cdot})^2 \\
&= \tfrac{1}{18}\sum_{i=1}^{6} 3s_i^2 \\
&= \tfrac{1}{6}\sum_{i=1}^{6} s_i^2 \\
&= \tfrac{1}{6}(0.020 + 0.015 + 0.025 + 0.018 + 0.022 + 0.014) \\
&= 0.019
\end{aligned}$$

av σ_ε^2. Eftersom radmedelvärdena

$$\bar{Y}_{i\cdot} = \mu + \delta_i + \bar{\varepsilon}_{i\cdot} \sim N(\mu, \sigma_\delta^2 + \frac{\sigma_\varepsilon^2}{4})$$

är sinsemellan oberoende, kan vi använda deras stickprovsvarians för att skatta $\sigma_\delta^2 + \sigma_\varepsilon^2/4$:

$$\begin{aligned}
\hat{\sigma}_\delta^2 + \tfrac{1}{4}\hat{\sigma}_\varepsilon^2 &= \tfrac{1}{6-1}\sum_{i=1}^{6}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\
&= \tfrac{1}{5}\big[(7.5-7.6)^2 + (7.1-7.6)^2 + (8.2-7.6)^2 + (7.4-7.6)^2 \\
&\quad +(7.9-7.6)^2 + (7.5-7.6)^2\big] \\
&= 0.152.
\end{aligned}$$

Det ger i sin tur en väntevärdesriktig skattning

$$\hat{\sigma}_\delta^2 = 0.152 - \frac{0.019}{4} = 0.147$$

av σ_δ^2.

b) Minsta kvadrat-skattningen av μ är $\hat{\mu} = \bar{Y}_{\cdot\cdot} = 7.6$. Eftersom

$$\text{Var}(\hat{\mu}) = \frac{\sigma_\delta^2}{6} + \frac{\sigma_\varepsilon^2}{24}$$

kan vi utnyttja räkningarna i deluppgift b) för att erhålla ett medelfel

$$\begin{aligned}
d &= \sqrt{\widehat{\text{Var}}(\hat{\mu})} \\
&= \sqrt{\tfrac{1}{6}(\hat{\sigma}_\delta^2 + \tfrac{1}{4}\hat{\sigma}_\varepsilon^2)} \\
&= \sqrt{\tfrac{0.152}{6}} \\
&= 0.1592,
\end{aligned}$$

och ett konfidensintervall

$$\begin{aligned}
I_\mu &= \hat{\mu} \pm t_{0.025}(5)d \\
&= 7.6 \pm \sqrt{F_{0.05}(1,5)} \cdot 0.1592 \\
&= 7.6 \pm \sqrt{6.608} \cdot 0.1592 \\
&= (7.19, 8.01)
\end{aligned}$$

för μ med konfidensgrad 95%.

Uppgift 4

a) Inflytandet av respektive faktor skattas till

$$\begin{aligned}
\hat{G} &= (\bar{Y}_{+\cdot\cdot} - \bar{Y}_{-\cdot\cdot})/2 \\
&= (-Y_{---} + Y_{+--} - Y_{-+-} + Y_{++-} - Y_{--+} + Y_{+-+} - Y_{-++} + Y_{+++})/8 \\
&= 0.3125, \\
\hat{S} &= (\bar{Y}_{\cdot+\cdot} - \bar{Y}_{\cdot-\cdot})/2 \\
&= (-Y_{---} - Y_{+--} + Y_{-+-} + Y_{++-} - Y_{--+} - Y_{+-+} + Y_{-++} + Y_{+++})/8 \\
&= 0.4375, \\
\hat{M} &= (\bar{Y}_{\cdot\cdot+} - \bar{Y}_{\cdot\cdot-})/2 \\
&= (-Y_{---} - Y_{+--} - Y_{-+-} - Y_{++-} + Y_{--+} + Y_{+-+} + Y_{-++} + Y_{+++})/8 \\
&= -0.1875.
\end{aligned} \tag{7}$$

Alternativt härleder man (7) från den allmänna formeln för minsta kvadratskattningar. Försöket svarar mot ju en allmän linjär modell med observationsvektor

$$\boldsymbol{Y} = (Y_{---}, Y_{+--}, Y_{-+-}, Y_{++-}, Y_{--+}, Y_{-++}, Y_{+-+}, Y_{+++})^T,$$

parametervektor $\boldsymbol{\theta} = (\mu, \bar{G}, \bar{S}, \bar{M})^T$ och designmatris

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \Longrightarrow \boldsymbol{A}^T\boldsymbol{A} = 8\boldsymbol{I}_4,$$

där $\boldsymbol{I}_4$ är identitetsmatrisen av ordning 4. Det medför att minsta kvadratskattningarna $\hat{G}$, $\hat{S}$ och $\hat{M}$ svarar mot komponenterna 2,3,4 hos

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\mu} \\ \hat{G} \\ \hat{S} \\ \hat{M} \end{pmatrix} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \frac{1}{8}\boldsymbol{A}^T\boldsymbol{Y}, \tag{8}$$

vilket överensstämmer med (7).

b) Vi kan antingen utnyttja (8) och dra slutsatsen att $\hat{\boldsymbol{\theta}}$ har en fyrdimensionell normalfördelning

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}) = N(\boldsymbol{\theta}, \frac{\sigma^2}{8}\boldsymbol{I}_4).$$

Speciellt följer att $\hat{G}, \hat{S}, \hat{M}$ är väntevärdesriktiga, oberoende och normalfördelade skattningar med samma varians $\sigma^2/8$. Alternativt kan man motivera detta genom att utgå från de explicita formlerna för de tre skattningarna i (7), och att alla Y_{ijk} är oberoende och normalfördelade med väntevärden

$$\mu_{ijk} = \mu + \bar{G}i + \bar{S}j + \bar{M}k$$

och samma varians σ^2.

c) För att testa hypotesmodellen mot grundmodellen ska vi först hitta kvadratsumman för Avvikelse från hollhypotes (eller Regression). Enligt ledningen får vi denna kvadratsumma genom att addera ihop kvadratsummorna för de tre ingående faktorerna G, S och M, dvs

$$\begin{aligned} \text{Kvs(Regression)} &= 8(\hat{G}^2 + \hat{S}^2 + \hat{M}^2) \\ &= 8\left[0.3125^2 + 0.4375^2 + (-0.1875)^2\right] \\ &= 2.5938, \end{aligned}$$

och

$$\begin{aligned} \text{Kvs(Residual)} &= \text{Kvs(Total)} - \text{Kvs(Regression)} \\ &= 3.0687 - 2.5938 \\ &= 0.4750. \end{aligned}$$

Eftersom antalet frihetsgrader för Regression och Residual är 3 respektive $8 - 1 - 3 = 4$, ger det en

$$\text{F-kvot} = \frac{\text{Kvs(Regression)}/3}{\text{Kvs(Residual)}/4} = \frac{2.5938/3}{0.4750/4} = 7.28,$$

som överstiger $F_{0.05}(3, 4) = 6.59$. Således förkastar vi nollhypotesen och konstaterar att små variationier av ingredienserna har en signifikant inverkan på kundnöjdheten.

Uppgift 5

a) Vi skriver om den allmänna linjära modellen för datamaterialet på matrisform $\boldsymbol{Y} = \boldsymbol{A\theta} + \boldsymbol{\varepsilon}$, där $\boldsymbol{Y} = (Y_1, \ldots, Y_6)^T$ är observationsvektorn, $\boldsymbol{\theta} = (\beta_1, \beta_2)^T$ parametervektorn, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_6)^T$ feltermsvektorn och

$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \\ -1 & 1 \\ -1 & -1 \\ 1 & -1 \end{pmatrix} = (\boldsymbol{x}_1, \boldsymbol{x}_2).$$

designmatrisen. Speciellt ser vi att designmatrisens två kolumner $\boldsymbol{x}_1$ och $\boldsymbol{x}_2$ är ortogonala, så att

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$$

är diagonal. Minsta kvadrat-skattaren av effektparametrarna kan skrivas som

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \frac{1}{4}\boldsymbol{A}^T\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{x}_1^T\boldsymbol{Y}/4 \\ \boldsymbol{x}_2^T\boldsymbol{Y}/4 \end{pmatrix}. \tag{9}$$

Vidare ges den simultana normalfördelningen för $\hat{\beta}_1$ och $\hat{\beta}_2$ av

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}\right) = N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \begin{pmatrix} \sigma^2/4 & 0 \\ 0 & \sigma^2/4 \end{pmatrix}\right). \tag{10}$$

b) Vi utnyttjar ledningen och skriver om prediktionsfelet som

$$Y - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 = \varepsilon - (\hat{\beta}_1 - \beta_1)x_1 - (\hat{\beta}_2 - \beta_2)x_2.$$

Enligt a) är $\hat{\beta}_1$ och $\hat{\beta}_2$ sinsemellan oberoende, eftersom kovariansmatrisen i (10) är diagonal. Eftersom $\hat{\beta}_1$ och $\hat{\beta}_2$ endast beror av feltermerna $\varepsilon_1, \ldots, \varepsilon_6$, är dessa skattningar oberoende av feltermen ε för den nya observationen. Vi får därför att

$$\begin{aligned} \text{MSEP} &= \text{Var}(\varepsilon) + x_1^2\text{Var}(\hat{\beta}_1) + x_2^2\text{Var}(\hat{\beta}_2) \\ &= \sigma^2\left[1 + \tfrac{x_1^2+x_2^2}{4}\right]. \end{aligned} \tag{11}$$

c) Den förenklade modellen, där bara kovariat 1 ingår, kan skrivas som $\boldsymbol{Y} = \boldsymbol{A}_1\beta_1 + \boldsymbol{\epsilon}$, där designmatrisen $\boldsymbol{A}_1 = \boldsymbol{x}_1$ svarar mot första kolumnen i $\boldsymbol{A}$ och där $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_6)^T$. Vi får därför en minsta kvadrat-skattare

$$(\boldsymbol{A}_1^T\boldsymbol{A}_1)^{-1}\boldsymbol{A}_1\boldsymbol{Y} = \frac{1}{4}\boldsymbol{x}_1^T\boldsymbol{Y} = \hat{\beta}_1,$$

som överensstämmer med (9), eftersom kolumnerna i $\boldsymbol{A}$ är ortogonala. Prediktionsfelet för den nya observationen, när det första datasetet anpassats till den mindre modellen med kovariat 1, blir därför

$$Y - \hat{\beta}_1 x_1 = \varepsilon - (\hat{\beta}_1 - \beta_1)x_1 + \beta_2 x_2.$$

På motsvarande sätt som i a) får vi ett prediktionsfel

$$\begin{aligned} \text{MSEP}_1 &= \text{Var}(\varepsilon) + x_1^2\text{Var}(\hat{\beta}_1) + x_2^2\beta_2^2 \\ &= \sigma^2\left[1 + \tfrac{x_1^2}{4}\right] + x_2^2\beta_2^2. \end{aligned} \tag{12}$$

Genom att bilda differensen mellan (11) och (12) ser vi att

$$\text{MSEP} - \text{MSEP}_1 = x_2^2(\frac{\sigma^2}{4} - \beta_2^2).$$

Alltså ger den mindre modellen en bättre prediktion av Y om $x_2 \neq 0$ och $\sigma^2 > 4\beta_2^2$. Det lönar sig alltså inte att skatta β_2 om denna parameter är liten i förhållande till feltermsvariansen σ^2.